



Université catholique de Louvain
Institute of Statistics, Biostatistics and Actuarial Sciences

Non-life Insurance Analytics

Risk segmentation, synthetic data, and fair pricing

Charlotte JAMOTTON

A thesis submitted to the
Université catholique de Louvain
in partial fulfillment of the
requirements for the degree of
DOCTOR OF SCIENCES

Thesis Committee:

Prof. **Donatien Hainaut**
Prof. **Michel Denuit**
Prof. **Olivier Caelen**
Prof. **Katrien Antonio**
Prof. **Marie-Pier Côté**

UCLouvain
UCLouvain
UCLouvain
KU Leuven & University of Amsterdam
Université Laval

Supervisor
President
Secretary

August 2026

There's always gonna be another mountain.

J. Alexander and J. Mabe

Acknowledgements

I kept thinking that these pages would be the easiest ones to write. Then the write-up happened. Somewhere between the final edits, the last figures, and the never-ending list of things still to fix, I wasn't quite sure what to write here anymore.

Memories are strange. We forget equations. We forget reviewer comments. We even forget why we spent three days debugging something that turned out to be a missing comma. But somehow, we remember exactly who was there. Who filled the empty spaces between chapters with some of our happiest PhD memories. Because writing a thesis is one thing. But reaching the end of it surrounded by an absurdly talented, mildly chaotic, and occasionally questionable group of humans is a whole other story. These people were found somewhere between coffee breaks (which may have gotten a little out of hand consumption-wise), conference seasons, sport classes (with questionable levels of coordination), shared miles (and injuries duh), hiking weekends (where the km estimation was about as good as my orientation), lake loops (turned into Strava art), movie nights (Disney included), beers (for research purposes obviously), endless yapping (the very best kind), a few objectively legendary stories, and conversations that shaped me far more than this thesis. Most of the people mentioned here probably don't know how much of a difference they made. I can, at least, try to thank them all. So here we are.

The acknowledgements that follow switch between French, English, en een beetje Vlaams, much like my days did during these four years.

Avant tout, un merci tout particulier à mon superviseur Prof. Donatien Hainaut. Donatien, merci pour ton temps tout au long de ces quatre années, pour ta franchise, et pour chacune de tes remarques qui m'ont aidé à ajuster les voiles quand le vent changeait, tout en me laissant naviguer par moi-même.

Je souhaite également remercier les membres de mon jury, Prof. Katrien Antonio, Prof. Marie-Pier Côté, Prof. Michel Denuit, et Prof. Olivier Caelen, pour le temps consacré à la lecture de ce travail et pour leurs précieux commentaires. Dank je wel, Katrien, voor je scherpe blik, je rustige en warme aanwezigheid tijdens de conferenties, voor de glazen, en voor het feit dat je er altijd op toezag dat ik veilig terug naar mijn hotel geraakte (tot grote vreugde van de KU Leuven-

jongens, uiteraard). Marie-Pier, merci d'avoir partagé ton regard sur l'équité avec tant d'ouverture, ainsi que ta jovialité lors de nos quelques échanges. Michel, merci pour ton enseignement, et tout particulièrement pour ton cours d'assurance dommages, qui restera mon préféré de toute ma scolarité. Olivier, merci pour ton remarquable investissement en tant que membre de mon jury.

À tous ceux du SMCS, de l'IT, et des RH, merci pour votre gentillesse, dans les couloirs ou à la cafétéria, qui ont rendu certains jours plus humain.

Because a place is only as good as the people who fill it, a heartfelt thank you to the colleagues and friends with whom I share some of my dearest PhD memories. À Arnaud (la drama queen avec son humour décalé), Aurèle, Benjamin (et ses Buddy, et sa Dr. of fine arts préf'), Caroline, Edouard (pour ces quatre années de doctorat partagées), Ensiyeh (for that karaoke night with Mengxue and Vanessa), Fanny, Gabriel, Guillaume (pour tous ces kms parcourus ensemble, Belle-île-en-mer, et tout le reste sauf le step), Hortense, Hugo (pour les soirées Disney, et les conversations 10e8 au soir au bord du lac, c'était gucci), Hugues (VAE>GAN), Jean-Loup, José, Keivan (the OG running buddy), Kenrick, Lara (la queen (pour rester dans la royauté et ne pas utiliser le mot en "b") des PMB), Lisa (et son accent toulousain), Lise (et son calme légendaire avec le planning pendant que mon dernier neurone "chill" court-circuitait), Madeline (pour avoir écouté tous mes plans sur la comète, mes plots twists, pour le sucre, les bike & run, Mechelen sous la neige, et les dingeries qui m'arrivaient qu'avec toi), Mathias (trail?), Mathilde, Meng (for all our lost-in-translation moments, our sunset walks, our tea-time, and four years of shared memories that quietly anchored me), Mirco (pour avoir trouvé parfaitement normal de me déposer à 5h du mat' à la gare un samedi), Morine (pour tes matching outfits, les tea times qui finissaient systématiquement en discussions sur la vie, et pour avoir rendu cette dernière année tellement plus lumineuse), Patricia (qué maja, gracias for the fortune cookies, for introducing me to chocolate coffee and running, hope we both find the proof of that uncorrelated model of yours), Philippe (aka le papa de l'institut), Phuphu (forever associated with those sweet potato gummies, sorryaintsorry), Stéphane, Tommaso, Torben, and Vanessa (for the pat of encouragement before my first talk in Valencia).

To the many colleagues I met in conferences. A particular shout-out to Gayani (for that church-pub night) and Mélanie (the kindest hearted yes-queen), à Olivier (pour les photos) et Oscar (ton timing était toujours iconique, je dois te le laisser), to Peter (for Lausanne high dives), and Zahara (for our shared bond over a Belgian beer in Bruxelles). A special thanks to the KU Leuven team for all the conference memories, from Valencia to Seoul. Thank you for the walks, the (beach) runs, the bike rides, the (karaoke, planes, and real) saunas, the mosquito bites, the shared meals, flights, buses, and train rides. To Dries and his tornados, for the Liverpool karaoke, and for the London bike ride. To Eva and her bubbly personality, for Edinburgh and Zaventem passport control line. To Freek and his hearty laugh, for handling my sliding-door crisis in Stockholm. To Jens and his sparky-eyed bril-

liance, for his Strava traces that made me think I could follow along (I could not). To Paul and his beer smile, for yapping through the entire flight back from Valencia, voor die compleet maffe conferencesessie in Edinburgh (honorable mention to your burned arm hair while rocking that hole), and for bringing your running shoes to Liverpool. To Rune and his beer knowledge, for the duikje in the Emajõgi, the museum bike ride, and all the shared beers (besides that Capri-Sun one).

À Charles et Renée Baltus, pour les petits pains du dimanche matin, la glace, les pommes, et les biscuits à la Saint-Nicolas.

To the three biggest yappers in my life, Morgane, Nicolas, and Perrine. Thank you for filling my quietness with your smartness or stupidity. Un merci tout particulier à Sumathi pour son amitié ces dernières années, pour les cafés à la récré, les cours de step, nos tours de lac, nos bike & run, nos goûts musicaux, nos soirées ciné, et tout le sel que tu m'as laissée déverser devant toi.

À ma famille. À mes grands-parents. À mes soeurs, Camille et Sophie (2/3).

*À mes parents,
avec gratitude, et le plus profond respect,
merci pour tout, mais surtout et avant tout,
pour votre léift au fil des saisons de la vie.*

Contents

I	Modeling risk heterogeneity	7
1	Insurance analytics with clustering techniques	9
1.1	Introduction	9
1.2	K-means clustering	12
1.2.1	A holistic overview	12
1.2.2	Burt distance for categorical data	14
1.2.3	Burt distance for mixed data types	21
1.3	Burt-adapted K-means variants	28
1.3.1	Mini-batch K-means	28
1.3.2	Fuzzy K-means	30
1.4	Burt-adapted spectral clustering	31
1.5	Conclusion	39
1.A	Tables	41
1.B	Figures	41
1.C	Algorithms	42
1.D	Proofs	44
2	Variational AutoEncoder for synthetic insurance data	45
2.1	Introduction	46
2.2	Related works	48
2.3	Variational AutoEncoder	50
2.3.1	Introduction	50
2.3.2	Variational approximation of the posterior	53
2.3.3	Variational lower bound	54
2.3.4	Reparameterization trick	56
2.4	Heterogeneous data	57
2.4.1	Dummy encoding and softmax activation	58
2.4.2	Quantile transformation, min-max scaling, and tanh activation	61

2.4.3	Heterogeneous reconstruction loss	64
2.5	Model training	67
2.5.1	Motor insurance portfolio	67
2.5.2	Model architecture and hyper-parameters set up	69
2.5.3	Reconstruction quality	72
2.5.4	Synthetic data generation	74
2.6	Conclusion	80
2.A	Proofs	81
2.B	Algorithms	82
2.C	Figures	85
2.D	Tables	87
3	Latent Dirichlet Allocation for tabular insurance data	89
3.1	Introduction	89
3.2	Methods	93
3.2.1	Notations	94
3.2.2	Generative process	95
3.2.3	Variational posterior and ELBO	98
3.3	Results and discussion	100
3.3.1	Experimental design	100
3.3.2	Datasets	101
3.3.3	LDA soft clustering for structured data	103
3.3.4	Burt adapted fuzzy (or soft) K-means	109
3.4	Conclusions	114
3.A	Notations	116
3.B	Sampled distributions	116
3.C	ELBO derivation	116
3.D	LDA soft clustering	118
3.E	Burt-adapted soft clustering	120
II	Fairness in insurance pricing	123
4	A multivariate Energy-based fairness adjuster for premiums	125
4.1	Introduction	125
4.2	Notations and definitions	128
4.3	Measures of disparities	129
4.3.1	Classical distances and divergences	129
4.3.2	Energy distance	131
4.4	Energy gradient fairness adjuster	133
4.4.1	Gradient-based fairness correction	133
4.4.2	New policyholders	135
4.4.3	Post-hoc autocalibration	136
4.5	Results	137

4.5.1	Datasets	137
4.5.2	Best-estimates	139
4.5.3	Fairness adjustment $G=2$	142
4.5.4	Fairness adjustment $G>2$	149
4.6	Discussion	153
4.A	Proofs	156
4.B	Datasets	159
4.C	Best-estimates	161
4.D	Best-estimates autocalibration	161
4.E	Simulated dataset	164
4.F	Real-world dataset	167
4.F.1	Best-estimates	167
4.F.2	Fairness adjustment $G=2$	169
4.F.3	Fairness adjustment $G>2$	174
4.G	Wasserstein and MMD benchmarks	177
4.G.1	Wasserstein (OT)	177
4.G.2	MMD-based gradient fairness adjustment	178
4.G.3	Results	179
5	A multi-criteria fair Gaussian regressor for insurance pricing	185
5.1	Introduction	185
5.2	bGGPR for log-link counts	190
5.2.1	Poisson frequency model with exposure	190
5.2.2	Posterior inference via Laplace approximation	191
5.2.3	Prediction and premium estimation	192
5.2.4	Portfolio balance	194
5.3	Multi-criteria FairGGPR	197
5.3.1	Fairness families	198
5.3.2	Multi-criteria FairGGPR for log-link counts	200
5.4	Results	205
5.4.1	Regression models	206
5.4.2	FairGGPR models	210
5.5	Discussion	217
5.A	Proofs	219
5.A.1	GGPR	219
5.A.2	FairGGPR	226
5.B	Results	232
5.C	Market simulation scenarios	240
5.C.1	Settings	241
5.C.2	Results	242

Introduction

Insurance pricing and risk management are fundamentally data-driven activities. Insurers rely on large portfolios of policyholders to assess risk, estimate expected losses, and set premiums that ensure portfolio balance and profitability. Each policy is typically described by a combination of numerical variables (such as age, income, or claim history) and categorical attributes (such as gender, geographic location, or vehicle characteristics). These datasets are inherently structured and heterogeneous, combining mixed data types, high-cardinality categorical variables, non-Gaussian marginals, and missing values. Dependencies between features are rarely linear, and interactions between risk factors can be complex.

While modern machine learning methods are well suited to model complex, non-linear relationships within such data, their direct application to insurance raises significant challenges. Black-box predictions may achieve high accuracy but provide limited insight into how risk is distributed across the portfolio, complicating both regulatory compliance and strategic decision-making. In particular, actuarial models must be interpretable in the sense that the contribution of individual risk factors to the predicted premium must be identifiable and defensible, so that actuaries can identify unusual risk patterns, and justify pricing decisions to regulators and management.

In practical actuarial work, modeling choices are driven by multiple, often competing objectives. Beyond predictive accuracy, actuaries must ensure that models are consistent with established principles of risk sharing. Premiums must remain adequate at the portfolio level, ensuring that expected claims and expenses are covered while maintaining competitiveness. Certain policyholder characteristics may be legally restricted, even if they improve accuracy, and even when excluded from the model, correlated variables can implicitly reproduce their effects. These constraints become more pronounced as insurance data grow richer. Telematics generate increasingly granular behavioral information. While such data can refine risk estimation, they also amplify segmentation, potentially leading to extreme premium dispersion and weakening the mutualization principle that underpins insurance. At the same time, regulatory frameworks such as the RGPD impose re-

strictions on data usage, and retention, reinforcing the need for transparent and accountable models. The tension between personalization and solidarity thus emerges as a central structural challenge in modern actuarial science. This tension is further reflected in the lack of consensus between fairness criteria. Individual fairness requires that similar policyholders receive similar premiums, emphasizing consistency at the micro level and coherence in the local geometry of the risk space. Group fairness, by contrast, focuses on aggregate outcomes across predefined protected groups, seeking to prevent systematic disparities in pricing distributions.

Actuarial modeling is therefore inherently normative. It reflects not only statistical regularities observed in data, but also societal choices about risk pooling, mutualization, and fairness. The challenge is to reconcile statistical rigor with these ethical and regulatory constraints, creating models that are simultaneously predictive, explainable, and aligned with the principles of equitable risk allocation.

This thesis argues that many of the current difficulties (i.e., how insurance data are represented, how risk heterogeneity is structured, and how fairness constraints are imposed on pricing models) in actuarial machine learning are too often addressed independently. Representation learning is developed without regard to fairness, fairness corrections are applied without reconsidering the underlying data structure, and segmentation is treated as a purely exploratory task disconnected from pricing objectives. This fragmentation limits both the interpretability of actuarial models and their effectiveness for portfolio management, risk monitoring, and regulatory reporting.

Outline of the thesis

This thesis investigates how modern statistical and machine learning methodologies can be adapted to address three tightly connected actuarial problems: (i) the representation and segmentation of heterogeneous insurance data into meaningful latent risk profiles, (ii) the limitations imposed by data availability and privacy regulations, such as the RGPD, and (iii) the design of pricing models that remain compatible with legal and ethical fairness requirements while maintaining portfolio-level balance.

The thesis adopts a progressive, problem-driven approach. Each chapter addresses a specific limitation encountered in the analysis of insurance portfolios and motivates the methodological developments of the subsequent chapters. In doing so, it aims to reconcile predictive accuracy, interpretability, and fairness within a coherent actuarial framework. By positioning the work at the intersection of data science, actuarial methodology, and regulatory compliance, the thesis highlights both the technical and societal relevance of modern insurance modeling.

The thesis is divided into two conceptual blocks: representation of risk heterogeneity, and fair insurance pricing.

I. Modeling risk heterogeneity

This first part addresses a foundational actuarial question: how is risk heterogeneity structured within large and complex insurance portfolios before pricing takes place? Rather than treating heterogeneous data as a preprocessing inconvenience, the following chapters consider representation as a primary modeling object. Through distance-based clustering, continuous latent embeddings, and probabilistic mixture modeling, this part develops progressively richer frameworks for uncovering latent risk profiles in mixed-type insurance data. The objective is not only dimensionality reduction, but the construction of meaningful and interpretable representations that can serve as a basis for segmentation and downstream pricing decisions.

Chap 1. Heterogeneity as a first-order actuarial problem

A defining characteristic of insurance data is their heterogeneity. Portfolios combine continuous variables with a variety of marginal distributions, and high-cardinality categorical attributes, often with missing values and complex dependency structures. Despite this, much of the actuarial literature continues to rely on preprocessing pipelines and modeling choices that are inherited from numerical settings and only imperfectly adapted to mixed-type data.

The thesis begins by treating heterogeneity not as a nuisance to be engineered away, but as a first-order modeling challenge. Before prediction or pricing can take place, insurers must understand how policies relate to one another and how risk is structured across the portfolio.

Chapter 1 addresses this foundational problem through unsupervised clustering of mixed-type insurance data. By introducing a novel dissimilarity measure based on Burt distance, the chapter demonstrates how classical clustering algorithms can be systematically adapted to heterogeneous portfolios while preserving dependencies between categorical modalities. This contribution establishes a geometric framework for comparing insurance policies, scalable to realistic portfolio sizes and compatible with multiple clustering paradigms.

While Burt clustering reveals group-level structure, it remains inherently limited. It often produces discrete assignments, and does not offer a continuous notion of similarity suitable for downstream modeling. These limitations motivate a shift toward latent representations.

Chap 2. Latent representations

To move beyond explicit distances and hard partitions, the thesis next considers representation learning approaches that allow insurance data to be embedded into lower-dimensional latent spaces.

Chapter 2 introduces Variational AutoEncoders (VAEs) as a tool for learning continuous latent representations of heterogeneous insurance portfolios. This

chapter positions representation learning as more than a technical convenience. The latent space becomes a modeling object in its own right, encoding similarities between policies and enabling the generation of synthetic insurance data. The proposed framework addresses practical actuarial concerns, including mixed data types, and the need for preprocessing non-Gaussian marginal distributions through quantile transformations.

By enabling both dimensionality reduction and data synthesis, this chapter responds directly to emerging challenges in actuarial practice related to data availability, privacy regulation, and portfolio simulation. At the same time, it exposes a key limitation of purely latent representations. They often lack direct actuarial interpretability and do not naturally translate into identifiable risk profiles.

Chap 3. From representations to interpretable risk profiles

Interpretability is not an optional feature in actuarial modeling, it is a requirement. Risk segmentation must be explainable.

Chapter 3 addresses this requirement by reframing insurance segmentation as a probabilistic mixture problem. By adapting Latent Dirichlet Allocation (LDA) to structured insurance data, the chapter introduces a soft clustering framework in which each policy is represented as a mixture of latent risk profiles. This approach captures overlapping risk characteristics, accommodates missing data, and produces directly interpretable outputs in the form of modality distributions.

In contrast to both hard clustering and neural embeddings, this framework explicitly models uncertainty in segmentation and aligns naturally with actuarial concepts of partial risk membership. Importantly, it establishes a probabilistic bridge between representation and pricing. Policies are no longer assigned to a single class, but characterized by weighted exposure to multiple latent risk profiles.

II. Fairness in insurance pricing

Building on the risk representations developed in Part I, this second part turns to the central actuarial task of premium allocation under legal and ethical constraints. As pricing models become increasingly complex, the tension between predictive accuracy, regulatory compliance, and fairness intensifies. The chapters in this part investigate how fairness considerations can be integrated into actuarial modeling, first through post-processing distributional corrections and subsequently through the direct incorporation of competing fairness paradigms within a unified Bayesian framework. Together, they examine how different fairness interventions reshape portfolio outcomes and how these interventions can be reconciled with the principles of mutualization and actuarial balance.

Chap 4. Pricing under fairness constraints

Once risk has been represented and structured, the central actuarial task remains the allocation of premiums. However, modern pricing models must satisfy anti-discrimination regulations.

Chapter 4 confronts this challenge directly by developing a solidarity-based post-processing framework for premium prediction models. Focusing on demographic parity as a group fairness criterion, the chapter introduces a gradient-based correction using the Energy distance to align premium distributions across protected groups. This approach allows multiple sensitive attributes to be handled simultaneously and proposes a post-hoc autocalibration step to preserve portfolio-level balance.

Starting from unconstrained premium estimates, fairness is introduced through an iterative correction procedure whose effect can be quantified at each step. The resulting premium adjustments are therefore directly observable and attributable to the fairness intervention, allowing the impact of demographic parity on individual and portfolio-level premiums to be explicitly analysed.

This chapter makes a normative claim. In insurance pricing, fairness cannot be reduced to the exclusion of sensitive variables. Instead, it must be addressed at the level of outcome distributions, in a manner consistent with the mutualization principles underlying insurance. Nevertheless, post-hoc group-level corrections raise rightful deeper questions about coherence with the existing regulation, generalization to the insurance market, and the interaction between competing fairness notions.

Chap 5. A unified framework for competing fairness notions

The final chapter addresses these questions by embedding fairness interventions directly into the structure of the pricing model.

Chapter 5 develops a Bayesian non-parametric framework based on Generalized Gaussian Process Regression that integrates multiple fairness paradigms within a single model. Rather than advocating for a single definition of fairness, or apply successive post-hoc corrections, the chapter provides a unified setting in which actuarial, individual-level, and group-level fairness criteria can be implemented, compared, and evaluated jointly. By modifying kernel structures and objective functions, the framework accommodates best-estimate, unaware, aware, (hyper-aware), and corrective pricing.

Unlike the post-processing framework of Chapter 4, fairness interventions are incorporated directly within the estimation procedure. As a result, fairness affects not only the premium estimate itself but the entire predictive distribution, including both predictive means and predictive uncertainty. The framework therefore allows the impact of fairness interventions on risk assessment and uncertainty quantification to be studied jointly. Furthermore, multiple fairness notions can be incorporated simultaneously within a single constrained learning procedure, allow-

ing their interactions to be analysed directly rather than through a sequence of successive corrective adjustments.

This contribution moves beyond the prevailing fairness-accuracy trade-off narrative. It demonstrates that fairness interventions reshape portfolio outcomes in distinct ways and that their actuarial implications cannot be understood in isolation from model structure and uncertainty.

Thesis contributions

Collectively, the chapters of this thesis establish a structured approach to actuarial modeling that integrates data representation, segmentation, uncertainty, and fairness. By systematically connecting unsupervised learning, probabilistic modeling, and fair regression, this work provides methodological tools for actuaries navigating the increasing complexity of data, regulation, and societal expectations in modern insurance markets.

The chapters of this thesis are based on a series of peer-reviewed journal articles and ongoing research contributions. Chapter 1 is published as “Insurance Analytics with Clustering Techniques” by Jamotton, Hainaut, and Hames (2024) in *Risks*. Chapter 2 is published as “Variational AutoEncoder for Synthetic Insurance Data” by Jamotton and Hainaut (2024) in *Intelligent Systems with Applications*. Chapter 3 is published as “Latent Dirichlet Allocation for Structured Insurance Data” by Jamotton and Hainaut (2026c) in *Annals of Data Science*. Chapter 4 is published as “A Multivariate Energy-based Fairness Adjuster for Premiums” in the *North American Actuarial Journal*. Chapter 5 titled “A Multi-Criteria Fair Gaussian Regressor for Insurance Pricing” represents ongoing work and is included in this thesis as a work in progress.

Part I

Modeling risk heterogeneity

Chapter 1

Insurance analytics with clustering techniques

The K-means algorithm and its variants are well-known clustering techniques. In actuarial applications, these partitioning methods can identify clusters of policies with similar attributes. The resulting partitions provide an actuarial framework for creating maps of dominant risks and unsupervised pricing grids. This research aims to adapt established clustering methods to complex insurance datasets containing both categorical and numerical variables. To this end, we propose a novel approach based on Burt distance. We begin by reviewing the K-means algorithm to establish the foundation for our Burt distance-based framework. We then extend the scope of application of the mini-batch and fuzzy K-means variants to heterogeneous insurance data. Additionally, we adapt spectral clustering, a technique based on graph theory that accommodates non-convex cluster shapes. To mitigate the computational complexity associated with spectral clustering's $O(n^3)$ runtime, we introduce a data-reduction method for large-scale datasets using our Burt distance-based approach.

Charlotte Jamotton, Donatien Hainaut, and
Thomas Hames (Sept. 2024). “Insurance Analytics with Clustering Techniques”. In: *Risks* 12.9, p. 141. ISSN: 2227-9091. DOI: [10.3390/risks12090141](https://doi.org/10.3390/risks12090141)

1.1 Introduction

Cluster analysis is among the most widely used techniques in statistical data analysis and machine learning. It aims to uncover group structures within a dataset by grouping objects (e.g., insurance policies) into distinct clusters. These clusters are formed so that they are as heterogeneous as possible, while objects within a

cluster are as homogeneous as possible. In actuarial applications, clustering can help identify dominant groups of policies. The resulting clusters can be further used to create maps of insured risks and unsupervised pricing grids.

Although clustering techniques have been extensively studied in fields such as image processing, interest in their application to actuarial research was slower to emerge. Early notable contributions include G. J. Williams et al. (1997), who applied K-means to identify policyholders with high claim ratios in a motor vehicle insurance portfolio, and Hsu et al. (1999), who introduced self-organising maps (SOMs) for knowledge discovery in insurance policy data. More recently, Hainaut (2018b) compared the effectiveness of K-means and SOMs in discriminating motorcycle insurance policies, based on the analysis of joint frequencies of categorical variables.

Originally designed for numerical settings, conventional clustering methods are not well suited to the mixed-type nature of insurance datasets. Specifically, K-means relies on Euclidean distance, a dissimilarity measure for vectors in $\mathbb{R}^n$. To address this limitation, Z. Huang (1997) introduced the K-modes algorithm. It substitutes Euclidean distance with Hamming, a dissimilarity measure for binary-coded categorical variables. Z. Huang (1998) later extended their approach to mixed-type data with the K-prototypes algorithm, a hybrid method combining K-means and K-modes using a weighted sum of Euclidean and Hamming distances. In the actuarial literature, Abdul-Rahman et al. (2021) used K-modes as a benchmark for their decision tree classifier in customer segmentation and profiling for life insurance. Similarly, Gan (2013) benchmarked K-prototypes against their Kriging method for valuing variable annuity guarantees. Yin et al. (2021) extended K-prototypes to spatial data in life insurance, while Zhuang et al. (2018) adapted it to account for missing data and reduce the method's sensitivity to prototype initialisation. More recent adaptations have incorporated Gower (1971) distance, an alternative to Hamming, into K-medoids (Gentle et al. 1991) to handle mixed data types. For applications in insurance, see also Campo et al. (2024), Debener et al. (2023), and Gan and Valdez (2019).

Clustering techniques are not limited to centroid-based methods such as K-means. In Weiss (1999), Weiss had already stressed the appeal of eigenvector-based methods for segmentation. The simplicity and stability of eigendecomposition algorithms are often cited as advantages of spectral clustering. For further details, readers may refer to J. Shi et al. (2000), A. Y. Ng et al. (2001), and Belkin et al. (2001). Luxburg (2007) also provides a concise step-by-step tutorial on the graph-theory foundations of spectral clustering. Recently, Mbuga et al. (2021) extended spectral clustering to heterogeneous datasets by defining a global dissimilarity measure using a weighted sum.

Addressing mixed-type datasets remains challenging. M. Wei et al. (2015) discuss the difficulties of developing linear combinations of distances to handle such data effectively. Rather than adapting the dissimilarity measure in K-means for categorical data, P. Shi et al. (2022) proposed to first project categorical data into

Euclidean space via categorical embeddings and then apply standard K-means clustering. Xiong et al. (2011) adopted a different approach with their divisive hierarchical clustering algorithm specifically designed for categorical objects. Their method starts with an initialisation based on Multiple Correspondence Analysis (MCA), and employs the chi-square χ^2 distance between individual objects and sets of objects to define the objective function.

The aim of this chapter is to adapt well-established clustering methods (such as K-means, its variants, and spectral clustering) to complex insurance datasets containing both categorical and numerical variables. The primary challenge lies in defining an appropriate distance measure between insurance policies characterised by mixed data types. To achieve this, we propose a novel approach based on Burt distance.

This study contributes to the literature by introducing Burt distance as a novel dissimilarity measure for clustering heterogeneous data. We evaluate our approach by benchmarking it against standard clustering methods, including K-means, K-modes, K-prototypes, and K-medoids, using a motor insurance portfolio. Unlike Hamming distance, our Burt distance-based framework accounts for potential dependencies between categorical classes (modalities). Moreover, our approach offers a more scalable alternative to the computationally expensive Gower distance, particularly when handling insurance portfolios with thousands of policies described by multiple attributes. We also empirically validate the applicability of our Burt approach to K-means variants (mini-batch and fuzzy) and spectral clustering. Finally, we introduce a data-reduction method, based on our Burt-adapted K-means, to mitigate the computational complexity associated with spectral clustering $O(n^3)$ runtime.

The outline of the chapter is as follows. In Section 1.2, we provide a holistic overview of the unsupervised K-means algorithm and its K-means++ heuristic to establish the foundational concepts and notations for our Burt distance-based framework. We empirically assess the performance of our proposed Burt distance-based K-means, comparing it with K-modes (Hamming distance) for categorical data, K-means (Euclidean distance) for numeric data, K-prototypes (Euclidean and Hamming distances), and K-medoids (Gower distance) for mixed-type data, using a motor insurance portfolio. In Sections 1.3.1 and 1.3.2, we adapt the mini-batch and fuzzy K-means variants to our Burt approach. The mini-batch extension is particularly well-suited to address computational challenges associated with large-scale insurance datasets, while fuzzy K-means, a soft form of clustering in which each object can belong to multiple clusters, is better aligned with the actuarial perspective on policy classification. Finally, in Section 1.4, we adapt spectral clustering to our Burt framework and propose a prior data-reduction method using Burt-adapted K-means to mitigate the computational complexity of spectral clustering’s eigendecomposition for large-scale datasets.

1.2 K-means clustering

1.2.1 A holistic overview

Consider a set of n numeric objects, denoted as $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, where each $\mathbf{x}_i \in \mathbb{R}^m$, and let $K \leq n$ be an integer. The K-means algorithm (Hartigan 1975) seeks to partition the dataset $\mathcal{X}$ into K disjoint clusters, $\{S_u\}_{u=1, \dots, K}$. The optimal partition is obtained by minimising the intraclass inertia I_a in Eq. (1.2.1) while maximising the interclass inertia I_c in Eq. (1.2.5).

The intraclass inertia I_a is the sum of the individual cluster inertia, defined as I_u in Eq. (1.2.2), weighted by their relative size $\frac{1}{n}|S_u|$:

$$I_a = \sum_{u=1}^K \frac{|S_u|}{n} I_u = \frac{1}{n} \sum_{u=1}^K \sum_{\mathbf{x}_i \in S_u} d(\mathbf{x}_i, \mathbf{c}_u). \quad (1.2.1)$$

The inertia of cluster S_u , denoted I_u , is the sum of the dissimilarity measure $d(\cdot)$ between each point $\mathbf{x}_i \in S_u$ and the cluster centroid $\mathbf{c}_u$ normalised by the cluster size $|S_u|$:

$$I_u = \sum_{\mathbf{x}_i \in S_u} \frac{1}{|S_u|} d(\mathbf{x}_i, \mathbf{c}_u). \quad (1.2.2)$$

The centroid (or centre of gravity) of cluster S_u is a m -vector $\mathbf{c}_u = (c_{u,1}, \dots, c_{u,m})$ defined as the geometric centre of the objects in S_u :

$$\mathbf{c}_u = \frac{1}{|S_u|} \sum_{\mathbf{x}_i \in S_u} \mathbf{x}_i. \quad (1.2.3)$$

In K-means, the dissimilarity measure $d(\cdot)$ is the squared Euclidean distance:

$$d(\mathbf{x}_i, \mathbf{c}_u) = \sum_{j=1}^m (x_{i,j} - c_{u,j})^2. \quad (1.2.4)$$

The interclass inertia I_c is defined as the inertia of the cloud of centers of gravity:

$$I_c = \sum_{u=1}^K \frac{|S_u|}{n} d(\mathbf{c}_u, \mathbf{c}) \quad \text{with } \mathbf{c} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i. \quad (1.2.5)$$

Optimising such an objective function is known to be computationally challenging (NP-hard). Nonetheless, efficient heuristics exist that rapidly converge to a local optimum. The most widely used approach is an iterative refinement technique, known as K-means, detailed in Algorithm 1. Given an initial set of K random centroids $\mathbf{c}_1(0), \dots, \mathbf{c}_K(0)$ and a given distance $d(\cdot)$, Algorithm 1 seeks to partition the dataset into K clusters $\{S_1(0), \dots, S_K(0)\}$ according to the following rule:

$$S_u = \left\{ \mathbf{x}_i \in \mathcal{X} \mid u = \arg \min_{u \in \{1, \dots, K\}} d(\mathbf{x}_i, \mathbf{c}_u) \right\}.$$

Algorithm 1 alternates between two steps:

1. Assignment step: At iteration e , each observation $\mathbf{x}_i$ is assigned to the cluster $S_u(e)$ whose centroid $\mathbf{c}_u(e)$ is closest.
2. Update step: The K centroids $(\mathbf{c}_u)_{u=1:K}$ are updated to the new cluster means $(\mathbf{c}_u(e+1))_{u=1:K}$.

At each iteration, the intraclass inertia decreases. This iterative process continues until convergence, that is, until cluster and/or centroid assignments no longer change.

Algorithm 1: K-means clustering.

Data: Data set $\mathcal{X}$, number of clusters K

Initialisation

- └ Random initial centroids $\mathbf{c}_1(0), \dots, \mathbf{c}_K(0)$
- └ Set iteration counter $e \leftarrow 0$

while *not converged* **do**

for $i = 1$ **to** n **do**

 Assign point $\mathbf{x}_i$ to closest cluster:

$$S_u(e) = \left\{ \mathbf{x}_i \in \mathcal{X} \mid u = \arg \min_{u \in \{1, \dots, K\}} d(\mathbf{x}_i, \mathbf{c}_u(e)) \right\}$$

for $u = 1$ **to** K **do**

 Update centroid:

$$\mathbf{c}_u(e+1) = \frac{1}{|S_u(e)|} \sum_{\mathbf{x}_i \in S_u(e)} \mathbf{x}_i$$

 Compute total distance (objective function):

$$d^{\text{total}} = \sum_{u=1}^K \sum_{\mathbf{x}_i \in S_u(e)} d(\mathbf{x}_i, \mathbf{c}_u(e+1))$$

$e \leftarrow e + 1$

The K-means++ algorithm (Arthur et al. 2007), detailed in Algorithm 2, offers an alternative to the centroids' random initialisation. The centroids obtained from Algorithm 2 are subsequently used as the starting values for the standard K-means iterative optimization procedure detailed in Algorithm 1. K-means++ has been shown to improve both the running time and the quality of the final clustering solution. However, K-means remains an heuristic. There is no guarantee of finding a global optimum, and the seed value may influence the results.

Algorithm 2: K-means++ initialisation of centroids.**Data:** Data set $\mathcal{X}$, number of clusters K **Initialisation**└ Select first centroid $\mathbf{c}_1 \in \mathcal{X}$ uniformly at random. Set $\mathcal{C} = \{\mathbf{c}_1\}$ **for** $u = 2$ **to** K **do**└ **for** $i = 1$ **to** n **do**└└ Compute distance of $\mathbf{x}_i$ to the set of chosen centroids:

$$d(\mathbf{x}_i, \mathcal{C}) = \min_{c \in \mathcal{C}} d(\mathbf{x}_i, c)$$

└└ Select next centroid $\mathbf{c}_u$ at random from $\mathcal{X}$ with probability:

$$p_i = \frac{d(\mathbf{x}_i, \mathcal{C})}{\sum_{i=1}^n d(\mathbf{x}_i, \mathcal{C})}, \quad i = 1, \dots, n$$

└└ Update set of centroids: $\mathcal{C} \leftarrow \mathcal{C} \cup \{\mathbf{c}_u\}$

As mentioned in Section 1.1, standard K-means cannot handle categorical data because its Euclidean measure of dissimilarity in Eq. (1.2.4) is appropriate only for numerical variables. The following Section 1.2.2 introduces Burt distance as a measure of dissimilarity for clustering non-numeric data.

1.2.2 Burt distance for categorical data

Let us first introduce a non-life insurance portfolio from the Swedish insurance company Wasa. The dataset is available on the companion website of the book by Ohlsson et al. (2010) and contains motorcycle insurance data from 1994 to 1998. Each policy is characterised by two quantitative and three qualitative rating factors. The quantitative variables are the insured's age and the age of their vehicle. The categorical variables, described in Table 1.A.1 (in Appendix), are the policyholder's gender, geographic location, and vehicle power class (determined by the power-to-weight ratio in $\text{KW} \times 100 / \text{vehicle weight in kg} + 75$, rounded to the nearest integer). The database also contains for each policy the number of claims, the total claim costs, and the contract period. After removing contracts with null duration, the database counts 62 436 insurance policies.

The categorical rating factors (in character-based format) must be pre-processed into a machine-processable (numeric) form. Let us denote by m_{cat} the number of categorical variables, and by m_{01} the total number of modalities across all categorical variables. The categorical rating factors can be one-hot encoded in an $n \times m_{01}$ super-indicator matrix:

$$\mathbf{D} = (d_{i,j})_{i=1,\dots,n; j=1,\dots,m_{01}},$$

where $d_{i,j} = 1$ if the i -th policy is characterised by the j -th modality, and $d_{i,j} = 0$

otherwise.

Example 1.1. For illustration, consider policies described by the gender (M = male or F = female) and education level (H = high school, C = college, or U = university), such that $m_{\text{cat}} = 2$ and $m_{01} = 5$. If the first policyholder is a man with an undergraduate degree and the second is a woman with a graduate degree, the first two rows of matrix $\mathbf{D}$ are represented as in the disjunctive Table 1.1.

Policy	Gender		Degree		
	M	F	H	C	U
1	1	0	0	1	0
2	0	1	0	0	1

Table 1.1: Example of a disjunctive table with a binary variable (Gender) and an ordinal variable (Degree, with three modalities).

As mentioned in Section 1.1, Z. Huang (1997) addressed the limitation of K-means by introducing K-modes, which substitutes Euclidean distance in the K-means algorithm with Hamming distance, a dissimilarity measure specifically designed for binary-coded categorical variables. This distance measures dissimilarity by counting the number of different bit positions in two bit strings, specifically the u -th centroid and the i -th policy bit strings:

$$d(\mathbf{x}_i, \mathbf{c}_u) = \sum_{j=1}^{m_{01}} \mathbf{1}\{x_{i,j} \neq c_{u,j}\}. \quad (1.2.6)$$

Hamming distance is a simple measure of discordance between observations. Therefore, it fails to discriminate between observations that are relatively close or far from each other. Our proposed Burt distance addresses this issue by studying joint frequencies in modalities. Specifically, Burt distance considers the co-occurrence of modalities by using a contingency table (Burt 1950). In this table, defined as an $m_{01} \times m_{01}$ Burt matrix $\mathbf{B}$, each entry represents the number of policies, $n_{r,c}$, that share both modalities r and c :

$$\mathbf{B} = (n_{r,c})_{r,c=1,\dots,m_{01}}.$$

This symmetric Burt matrix, illustrated in Table 1.2, is directly related to the disjunctive table $\mathbf{D}$ through the following mathematical relationship:

$$\mathbf{B} = \mathbf{D}^\top \mathbf{D}.$$

By definition, $\mathbf{B}$ is composed of $m_{\text{cat}} \times m_{\text{cat}}$ blocks $\mathbf{B}_{b,b'}$ for $b, b' = 1, \dots, m_{\text{cat}}$ and the sum of elements within a block $\mathbf{B}_{b,b'}$ is equal to the number of policies, n ,

inside the insurance portfolio $\mathcal{X}$. The row-wise sum of $n_{r,c}$ is equal to:

$$n_{r,.} = \sum_{c=1,\dots,m_{01}} n_{r,c} = m_{\text{cat}} n_{r,r}. \quad (1.2.7)$$

Because of the symmetry of the Burt matrix, we infer that:

$$n_{.,c} = \sum_{r=1,\dots,m_{01}} n_{r,c} = m_{\text{cat}} n_{c,c}. \quad (1.2.8)$$

By construction, blocks $\mathbf{B}_{b,b}$ for $b = 1, \dots, m_{\text{cat}}$ are diagonal matrices. Diagonal entries contain the number of policies, respectively, described by the modalities $1, \dots, m_b$ of the b -th categorical variable.

Example 1.2. In our example, we have $n_{1,1} + n_{2,2} = n$ and $n_{3,3} + n_{4,4} + n_{5,5} = n$. Here, $n_{1,1}$ and $n_{2,2}$ count the total number of men and women in the portfolio, whereas $n_{3,3}$, $n_{4,4}$ and $n_{5,5}$ count the number of policyholders with, respectively, a high school, college, or university degree.

		M	F	H	C	U
Gender	M	$n_{1,1}$	0	$n_{1,3}$	$n_{1,4}$	$n_{1,5}$
	F	0	$n_{2,2}$	$n_{2,3}$	$n_{2,4}$	$n_{2,5}$
Degree	H	$n_{3,1}$	$n_{3,2}$	$n_{3,3}$	0	0
	C	$n_{4,1}$	$n_{4,2}$	0	$n_{4,4}$	0
	U	$n_{5,1}$	$n_{5,2}$	0	0	$n_{5,5}$

Table 1.2: Burt matrix for the disjunctive Table 1.1.

In the same manner as Hainaut (2018b), we evaluate the level of dependence between modalities with a chi-square χ^2 distance (Burt 1950; Nishisato 1985). Intuitively, the distance between two modalities is measured by the sum of weighted gaps between joint frequencies with respect to all modalities. In that respect, the chi-square distance between rows r and r' of $\mathbf{B}$ is defined as:

$$\chi^2(r, r') = \sum_{c=1}^{m_{01}} \frac{n}{n_{.,c}} \left(\frac{n_{r,c}}{n_{r,.}} - \frac{n_{r',c}}{n_{r',.}} \right)^2 \quad r, r' \in \{1, \dots, m_{01}\}. \quad (1.2.9)$$

We recall that K-means is a centroid-based clustering algorithm that relies on minimizing dissimilarity measures to determine cluster assignments. Bregman divergences are the only distortion functions for which such centroid-based clustering schemes are possible (Banerjee et al. 2004). As we favour the Euclidean divergence for its simplicity, we replace the original joint frequency values $n_{r,c}$ in $\mathbf{B}$ with their weighted values $n_{r,c}^W$:

$$n_{r,c}^W := \frac{n_{r,c}}{\sqrt{n_{r,.} n_{.,c}}} \quad \text{for } r, c = 1, \dots, m_{01}.$$

Given the row-wise and column-wise sums in Eq. (1.2.7) and Eq. (1.2.8),

$$n_{r,c}^W := \frac{n_{r,c}}{m_{\text{cat}} \sqrt{n_{r,r} n_{c,c}}} \quad \text{for } r, c = 1, \dots, m_{01}.$$

With the diagonal matrix $\mathbf{C} = \text{diag} \left(n_{11}^{-\frac{1}{2}} \dots n_{m_{01}m_{01}}^{-\frac{1}{2}} \right)$, the weighted Burt matrix, $\mathbf{B}^W$, is equal to:

$$\mathbf{B}^W = \frac{1}{m_{\text{cat}}} \mathbf{C} \mathbf{B} \mathbf{C}.$$

Since each modality $j = 1, \dots, m_{01}$ is represented by its corresponding row $r = 1, \dots, m_{01}$ in $\mathbf{B}^W$, that is, by a vector in $\mathbb{R}^{m_{01}}$, each policy can be projected in Burt space by taking the centre of gravity $\mathbf{D}_{i,\cdot} \mathbf{B}^W / m_{\text{cat}}$ of its modalities in $\mathbf{D}_{i,\cdot}$, whose Burt coordinates are stored in the corresponding rows of $\mathbf{B}^W$. This is illustrated in Fig. 1.1 in the case of three modalities.

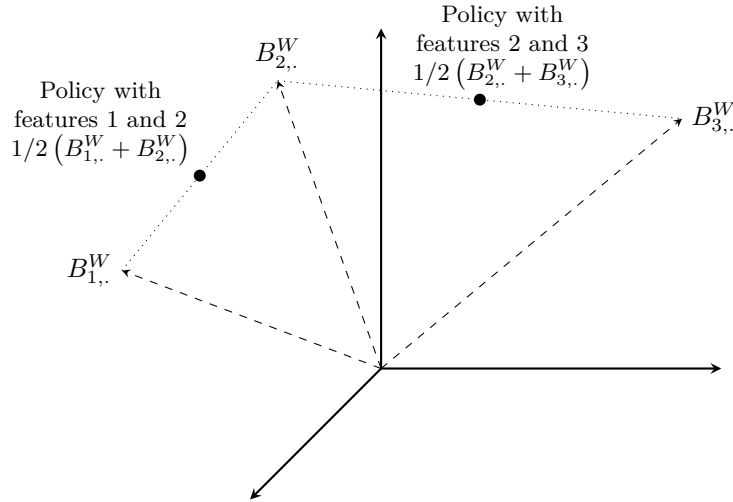


Figure 1.1: Illustration of two policies in Burt space with three modalities.

Burt distance between the i -th and i' -th policies is then given by:

$$d(i, i') = \left\| \mathbf{D}_{i,\cdot} \mathbf{B}^W / m_{\text{cat}} - \mathbf{D}_{i',\cdot} \mathbf{B}^W / m_{\text{cat}} \right\|_2. \quad (1.2.10)$$

Note that the proposed approach is closely related to Multiple Correspondence Analysis (MCA), as both methods are based on the disjunctive matrix $\mathbf{D}$, the Burt matrix $\mathbf{B}$, and the chi-square geometry associated with categorical data. However, the two methodologies target different objectives. In MCA, the objective is to construct a low-dimensional Euclidean representation of categorical modalities

that preserves (as closely as possible) the chi-square geometry of the full disjunctive representation. The resulting coordinates are typically obtained via a spectral decomposition and are used for dimensionality reduction and interpretation of associations between modalities. Whereas, the approach adopted here does not perform any dimensionality reduction. Instead, each policy is represented directly in the full weighted Burt space through the centre of gravity of its associated modalities (obtained by averaging the corresponding rows of the weighted Burt matrix). Clustering is then implemented on these policy-level representations, to construct a partition based on Euclidean proximity in this constructed feature space rather than recovering the principal axes of association between modalities.

We adapt the K-means Algorithm 1 to our Burt distance-based framework as follows. Each policy is projected into Burt space and identified by its centre of gravity $\mathbf{D}_{i,:} \mathbf{B}^W / m_{\text{cat}}$. The new centroids $\mathbf{c}_u(e+1)$ of $S_u(e)$, previously defined as in Eq. (1.2.3), are now defined as the following m_{01} -vector:

$$\mathbf{c}_u(e+1) = \frac{1}{|S_u(e)|} \sum_{\mathbf{x}_i \in S_u(e)} \mathbf{D}_{i,:} \mathbf{B}^W / m_{\text{cat}}.$$

The Euclidean distance, defined in Eq. (1.2.4), is replaced by the Burt distance, defined in Eq. (1.2.10). Note that the Burt-based distance in Eq. (1.2.10) measures similarity between policies through the co-occurrence patterns of their associated modalities. Two policies are considered close when their modalities exhibit similar co-occurrence structures with all other modalities in the dataset. This notion of similarity is inherently relational and does not encode any intrinsic ordering between modalities of the same categorical variable, since such modalities are mutually exclusive and therefore never co-occur. As a consequence, ordinal structure within a categorical variable cannot be directly captured in a row or column-wise interpretation of the Burt matrix. While ordinal distances could in principle be introduced through alternative encodings, such approaches become problematic in a multivariate setting, as they may induce inconsistent scaling across variables, generate artificial co-occurrences, and distort the interpretation of Burt entries. Mixed approaches combining ordinal distances with Burt-based representations are possible but typically require additional weighting schemes to ensure comparability across heterogeneous metrics. In this work, we chose to retain a purely Burt-based representation in order to preserve cross-variable co-occurrence information, which is empirically important in the insurance context.

To benchmark this newly defined Burt distance-based K-means, we first implement K-modes on the categorical data (Gender, Zone, and Class) from the Wasa insurance dataset. To implement K-modes, the dissimilarity measure $d(\cdot)$ in K-means Algorithm 1 is simply replaced by Hamming distance defined in Eq. (1.2.6).

Table 1.3 reports the policy allocation, the dominant modalities, as well as the average claim frequency associated with each cluster. The dominant modalities are identified as the dominant representative(s) in each cluster. A quick analysis reveals the most and least risky driver profiles. The riskiest category, with an average

claim frequency of 1.94%, contains a majority of vehicles with the highest ratio of power (EV ratio 25–). The lowest claim frequencies (0.6–0.785%) correspond to a majority of women living in small towns or in the countryside (zone 3 and 4).

Cluster	% of policies	Gender	Zone	Class	Frequency (%)
6	2.82	K	4	5	0.6003
1	5.73	K	4	3	0.6270
5	2.77	K	3	[3,4]	0.7855
10	13.39	M	[2,4]	4	0.8559
3	17.03	M	4	3	0.8919
2	20.61	M	3	3	0.9374
9	19.99	M	[3,4]	5	1.0532
7	2.73	K	2	[3,4]	1.2236
8	5.06	M	2	1	1.8256
4	9.88	M	[2,4]	6	1.9443

Table 1.3: Summary of K-modes clustering (with Hamming distance) results applied to policy data characterised by three categorical variables. The table reports cluster sizes, dominant features, and average claim frequencies per cluster.

If we use the average motor claim frequency per cluster as a predictor $\hat{\lambda}$ we can estimate the goodness of fit of the partition with the Poisson deviance defined in Eq. (1.2.11). The deviance is the difference between the saturated and partitioned model's log-likelihood. If N_i and ν_i are, respectively, the number of claims and the duration (exposure) of the i -th contract, the Poisson deviance is defined as:

$$\mathcal{D}_{\text{Poisson}} = 2 \sum_{i=1}^n N_i \left(\frac{\nu_i}{N_i} \hat{\lambda}_i - \left(\log \frac{\hat{\lambda}_i \nu_i}{N_i} + 1 \right) I_{\{N_i \geq 1\}} \right). \quad (1.2.11)$$

The Poisson deviance, AIC, and BIC are reported in Table 1.4. Both AIC and BIC are computed with a number of degrees of freedom equal to 10×16 (# of clusters $\times$ # of modalities).

Deviance	6576.31
AIC	8244.88
BIC	9691.58

Table 1.4: Statistics of goodness of fit obtained by partitioning the categorical data into 10 clusters with K-modes.

We now implement our Burt distance-based K-means on the same set of categorical variables (Gender, Zone, and Class). Table 1.5 reports the policy allocation,

the dominant modalities, and the claim frequency associated with each cluster. For each cluster S_u , the dominant modalities are identified either as the cluster's most common representatives, as in Table 1.3, or as the ones whose Burt coordinates are closest, in terms of the ℓ^2 -norm, to the cluster centre:

$$\arg \min_{j \in \{1, \dots, m_b\}} \left\| c_u - B_{(j + \sum_{b'=1}^{b-1} m_{b'})_{\cdot}}^W \right\|_2, \quad \text{for } b = 1, \dots, m_{\text{cat}},$$

where m_b are the number of modalities of the b -th categorical variable.

The average claim frequency per cluster is more heterogeneous (from 0.38% up to 3.08%) than in Table 1.3 (0.62–1.94%). The most and least risky driver profiles appear more clearly defined. The riskiest category, with an average claim frequency of about 3%, much higher than with Hamming distance, contains a majority of vehicles in the central and semi-central parts of Sweden's three largest cities (zone 1). The cluster with the second-highest claim frequency (on average) also has a majority of vehicles with high power (class 4–6) driven in suburbs plus middle-size cities (zone 2), whereas the lowest claim frequencies are characterised by vehicles in small towns and the countryside (zone 4 and 6).

Cluster	% of policies	Gender	Zone	Class	Frequency (%)
2	8.21	M	4	3	0.3811
9	12.27	M	4	[4, 5]	0.5051
1	8.95	M	6	[3, 5]	0.6808
4	6.43	K	4	[3, 5]	0.6813
7	4.58	M	2	3	0.8621
6	11.82	M	4	[1, 6]	0.9237
10	8.75	K	[2, 3]	[3, 4]	1.0144
5	16.81	M	3	[3, 5]	1.0825
3	10.94	M	2	[4, 6]	2.1482
8	11.23	M	1	[3, 4]	3.0878

Table 1.5: Table summary of Burt distance-based K-means clustering results (policy allocation, dominant features, and average claim frequencies per cluster) applied to policy data characterised by three categorical variables.

The deviance, AIC, and BIC are reported in Table 1.6. Both AIC and BIC are computed with a number of degrees of freedom equal to 10×16 (# of clusters $\times$ # of modalities). The goodness of fit is clearly better than the one obtained using K-modes with Hamming distance in Table 1.4.

Deviance	6350.04
AIC	8018.61
BIC	9465.31

Table 1.6: Statistics of goodness of fit obtained by partitioning the categorical data into 10 clusters with Burt distance-based K-means.

1.2.3 Burt distance for mixed data types

We have both theoretically and empirically demonstrated that Burt distance is a favourable alternative to Hamming distance for clustering categorical data. In the following, we show that our Burt distance-based approach also improves partition quality when handling heterogeneous datasets, outperforming K-prototypes and K-medoids. We first illustrate empirically, using the standard K-means algorithm, that the unsupervised discretisation of continuous variables, required for our Burt approach, does not negatively affect the method’s goodness of fit.

We start by applying the standard K-means algorithm to the numeric rating factors (owner’s age and vehicle age) in the Wasa dataset. Fig. 1.2 shows the 10 clusters in a two-dimensional space with the owner’s and vehicle ages on the axes.

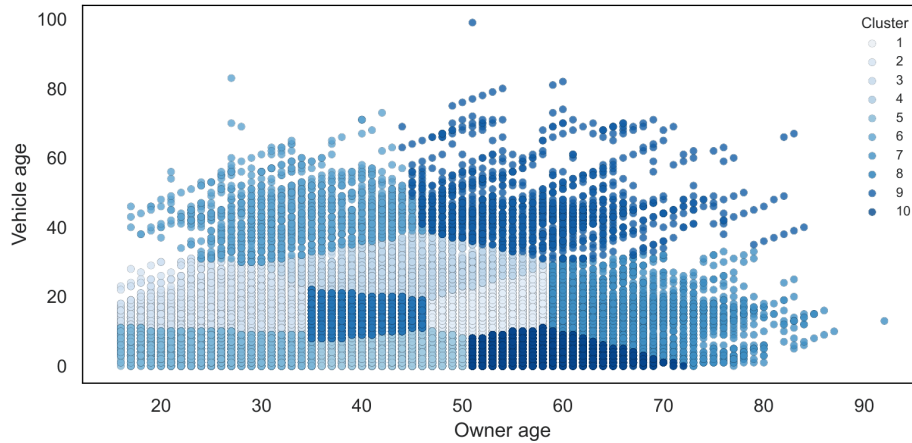


Figure 1.2: Illustration of the partitioning of the numeric data into 10 clusters with the K-means algorithm (using Euclidean distance).

The deviance, AIC, and BIC are reported in Table 1.7. Both AIC and BIC are computed with a number of degrees of freedom equal to 10×2 ($\#$ of clusters $\times$ $\#$ of numeric variables).

Deviance	6097.18
AIC	7485.75
BIC	7666.59

Table 1.7: Statistics of goodness of fit obtained by partitioning the numeric data into 10 clusters with the K-means algorithm (using Euclidean distance).

Table 1.8 reports the claim frequency, the percentage of policies, and the dominant modalities associated with each cluster. The riskiest category, with an average claim frequency of around 4%, is characterised by young or less experienced drivers (aged 23–29) of recent vehicles with high ratio powers (class 3–6), whereas the cluster with the lowest claim frequency (around 0%) is made up of middle-aged males (48–62 years old) driving classic cars (38+ years old) with low ratio power (class 1). We also observe that clusters with the lowest claim frequencies have a majority of insureds living in smaller towns or the countryside (zones 3 and 4), while clusters with the highest claim frequencies additionally include a majority living in suburbs or middle-sized cities (zone 2).

Cluster	% of policies	Gender	Zone	Class	Owner's age	Vehicle age	Frequency (%)
5	2.00	M	[3, 4]	1	[48, 62]	[38, 48]	0
9	4.85	M	[3, 4]	[3, 4]	[40, 48]	[21, 26]	0.2035
2	15.16	M	[3, 4]	[3, 5]	[48, 53]	[13, 17]	0.3698
10	6.46	M	[3, 4]	[3, 5]	[60, 66]	[12, 18]	0.45
3	14.94	M	[3, 4]	[3, 5]	[42, 47]	[13, 17]	0.51
7	2.09	M	[2, 4]	1	[28, 43]	[37, 48]	0.5106
6	13.54	M	[2, 4]	[3, 4]	[44, 49]	[1, 5]	0.8667
8	11.65	M	[2, 4]	[3, 4]	[50, 55]	[1, 5]	0.9183
4	14.42	M	[2, 4]	[3, 5]	[25, 32]	[12, 16]	1.5317
1	14.89	M	[2, 4]	[3, 6]	[23, 29]	[1, 7]	4.1332

Table 1.8: Table summary (policy allocation, dominant features, and average claim frequencies per cluster) of K-means applied to data with two continuous variables.

To apply our Burt-based K-means framework, we first discretise the continuous variables using univariate K-means clustering. The owner's age and vehicle age are partitioned into four and three clusters, respectively, resulting in the intervals [16, 35), [35, 47), [47, 58), [58, 93] for owner's age and [0, 11), [11, 29), [29, 100] for vehicle age. The use of an unsupervised discretisation is motivated by the skewness and presence of extreme values, particularly in vehicle age. While univariate discretisation ignores potential interactions between rating factors, embedding the resulting categorical (discretised) representation into Burt space allows us to capture pairwise associations between variables in the transformed space.

The deviance reported in Table 1.9 is similar to that obtained with non-discretised numeric data in Table 1.7. AIC and BIC are computed with 10×7 degrees of free-

dom ($\#$ of clusters $\times$ $\#$ of modalities). For comparison, the residual deviance of a GLM model fitted to the same categorical classes is equal to 6 094.49.

Fig. 1.3 displays the 10 clusters in a two-dimensional space, with the hard-edge cutoffs reflecting the discretisation process.

Deviance	6097.51
AIC	7586.08
BIC	8219.01

Table 1.9: Statistics of goodness of fit of the Burt distance-based K-means applied to data with two discretised continuous variables.

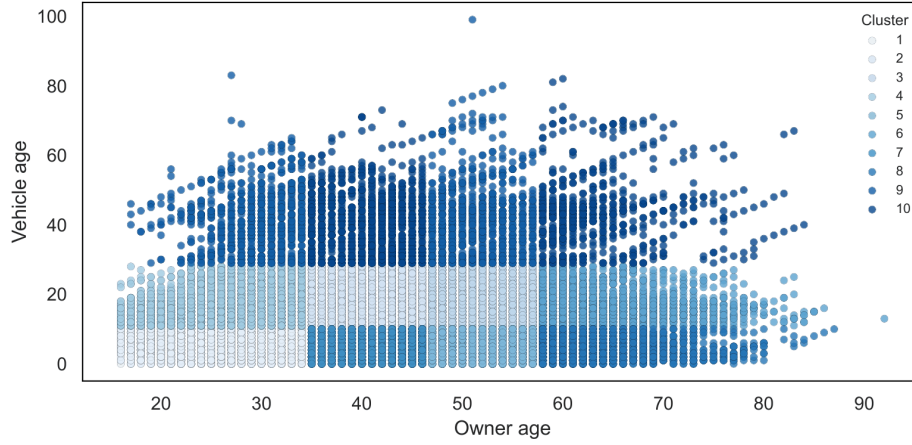


Figure 1.3: Illustration of the partitioning of the discretised numeric data (projected in Burt space) into 10 clusters with the Burt distance-based K-means.

Table 1.10 summarises the claim frequency, percentage of policies, and dominant modalities for each cluster. Policy allocation and the two riskiest categories are broadly consistent with Table 1.8, though small differences are observed. Notably, Burt-adapted K-means better identifies the link between low ratio power and smaller claim frequencies. The two lowest-frequency clusters are the only ones labelled with class 1 (lowest ratio power), which was not the case in Table 1.8.

Cluster	% of policies	Gender	Zone	Class	Owner's age	Vehicle age	Frequency (%)
9	2.66	M	[3,4]	1	[35,46]	[29,46]	0.2207
10	2.76	M	[2,4]	1	[29,52]	[29,45]	0.3083
3	15.31	M	[3,4]	[3,5]	[47,52]	[13,17]	0.3257
2	15.23	M	[3,4]	[3,5]	[42,46]	[12,17]	0.4080
7	6.41	M	[3,4]	[3,5]	[58,63]	[13,17]	0.4560
6	12.96	M	[2,4]	[3,5]	[47,52]	[1,6]	0.9060
5	10.46	M	[2,4]	[3,4]	[42,46]	[1,5]	0.9150
8	4.69	M	[2,4]	[3,4]	[58,62]	[1,5]	0.9444
4	14.32	M	[2,4]	[3,5]	[25,32]	[12,16]	1.5324
1	15.21	M	[2,4]	[3,6]	[23,29]	[1,7]	4.0664

Table 1.10: Table summary (policy allocation, dominant features, and average claim frequencies per cluster) of Burt distance-based K-means applied to two discretised continuous variables.

We assess the discretisation required by our Burt framework on a simulated dataset to further validate our findings. We generate 1000 observations from two continuous variables X_1 and X_2 . The simulated data points, illustrated in Fig. 1.4, are such that we can visually identify two distinct clusters. We categorize each variable, as required by our Burt approach, with the standard K-means algorithm (unsupervised approach) in 10 categories. We then apply our Burt distance-based K-means to the discretised data. Fig. 1.4 shows that our Burt approach was effectively able to identify the two clusters.

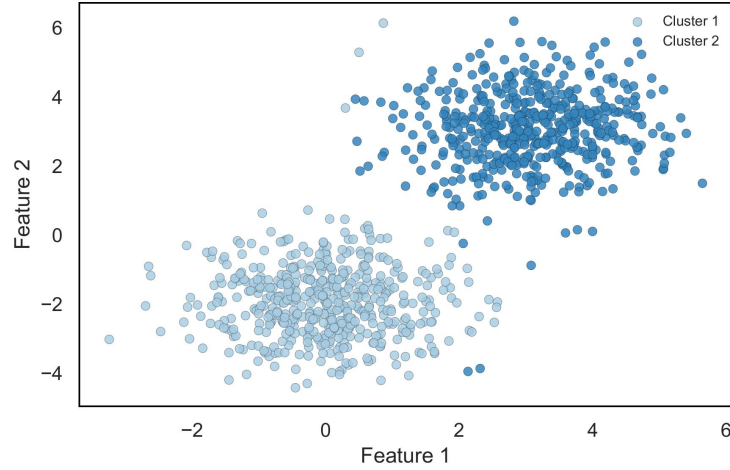


Figure 1.4: Clustering results obtained with the Burt distance-based K-means algorithm on a simulated dataset.

We can now benchmark our Burt distance-based K-means against K-prototypes (linear combination of Euclidean and Hamming distances) and K-medoids (Gower distance). Z. Huang (1998) extended K-modes (Z. Huang 1997) to handle mixed data types with the K-prototypes algorithm. This hybrid method combines K-means and K-modes by using a weighted sum of Euclidean distance for numeric data and Hamming distance for categorical data:

$$d(\mathbf{x}_i, \mathbf{c}_u) = \sum_{j \in \text{numeric}} (x_{i,j} - c_{u,j})^2 + \beta \left(\sum_{j \in \text{binary}} \mathbf{1}\{x_{i,j} \neq c_{u,j}\} \right),$$

where $\beta \in \mathbb{R}^+$ is a weighted factor that controls the relative importance of categorical variables with respect to numerical ones. Table 1.12 presents the clustering results obtained by applying K-prototypes with $\beta = 1$ to the Wasa insurance portfolio, which contains two continuous and three categorical features. As reported in Table 1.11, the riskiest cluster is once again characterised by younger drivers of vehicles with high power-to-weight ratios (classes 5 and 6), while the cluster with the lowest claim frequencies consists mainly of ancient vehicles (23+). Nevertheless, no clear patterns seem to emerge from the cluster analysis.

Cluster	% of policies	Gender	Zone	Class	Owner's age	Vehicle age	Frequency (%)
13	4.45	M	4	[1,4]	[39,51]	[23,30]	0.1754
5	3.50	M	[3,4]	1	[30,53]	[38,46]	0.2280
15	7.94	M	3	3	[43,53]	[13,19]	0.3010
11	9.28	M	4	3	[51,62]	[1,16]	0.3856
14	9.24	M	4	5	[43,55]	[12,17]	0.3888
4	5.07	M	4	2	[45,60]	[13,19]	0.5943
12	5.82	K	4	[3,4]	[33,46]	[12,18]	0.5985
7	9.01	M	[2,3]	4	[43,52]	[1,17]	0.6748
2	4.56	K	[1,2]	3	[24,48]	[1,16]	1.2525
6	4.85	M	[2,6]	6	[45,54]	[1,8]	1.3813
9	5.30	M	4	[1,4]	[18,29]	[12,17]	1.5103
1	7.37	M	1	3	[39,50]	[0,15]	1.6583
3	8.41	M	4	3	[21,29]	[1,6]	2.2660
8	8.57	M	2	5	[24,34]	[1,17]	2.5845
10	6.63	M	3	6	[23,29]	[6,13]	3.8295

Table 1.11: Table summary of K-prototypes clustering results (dominant features and average claim frequencies per cluster).

Moreover, the deviance, in Table 1.12, is worse than the one obtained when clustering only quantitative variables. One can eventually consider adjusting β in order to minimize the deviance, but we failed to find a value of β that significantly improved the results for the Wasa dataset.

Deviance	6161.47
AIC	8050.04
BIC	10491.35

Table 1.12: Statistics of goodness of fit obtained by partitioning the mixed-type dataset into 15 clusters with the K-prototypes algorithm.

We further benchmark Burt-distance based K-means for mixed data types with K-medoids (Gentle et al. 1991). Unlike K-means, it selects actual data points (medoids) as the cluster centers, making it more robust to noise and outliers. When dealing with mixed data types (numerical and categorical), Gower distance is commonly used as the K-medoids dissimilarity measure. It computes the distance between two objects as a combination of dissimilarities across all variable types. Formally, Gower distance is defined as:

$$d(\mathbf{x}_i, \mathbf{x}_{i'}) = \frac{1}{m} \sum_{j=1}^m \delta_{i,i'}^{(j)} d_{i,i'}^{(j)},$$

where m denotes the number of features, $d_{i,i'}^{(j)}$ is the normalised distance for the j -th feature, and $\delta_{i,i'}^{(j)}$ is an indicator function that equals 1 if the j -th feature is comparable for both observations, and 0 otherwise.

Gower distance requires computing pairwise dissimilarities between all n observations. This results in the calculation of $\frac{1}{2}n(n-1)$ pairwise distances, which scales quadratically with the size of the dataset. When working with an insurance portfolio, this rapidly becomes computationally prohibitive. On the Wasa dataset, computing the Gower distance matrix took approximately 25 minutes. The exhaustive search for optimal medoids, and the complexity of handling mixed data types, led to an additional 26 minutes to run the K-medoids algorithm with the pre-computed distances. Moreover, the resulting deviance, in Table 1.13, is even worse than the one obtained with K-prototypes in Table 1.12.

Deviance	6290.92
AIC	8179.49
BIC	10620.80

Table 1.13: Statistics of goodness of fit obtained by partitioning the mixed-type dataset into 15 clusters with the K-medoids algorithm (using Gower distance).

The disappointing results obtained with K-prototypes and K-medoids lead us to apply our Burt distance-based K-means on the Wasa dataset with two discretised continuous variables and three categorical variables. As reported in Table 1.14, we obtain a greater heterogeneity between clusters, with average claim frequencies

ranging from 0.25% up to 8%. The first four riskiest categories are characterised by a majority of young drivers (22–29 years old). The differences in the vehicle ages are also more defined between clusters, with the oldest vehicles assigned to the cluster with the smallest claim frequency, and more recent vehicles assigned to clusters with higher claim frequencies. Vehicles with the lowest power ratio (class 1) are also exclusively represented in the cluster with the smallest claim frequency (see cluster 16), whereas vehicles with the highest power ratio (class 6) are only represented in clusters with higher claim frequencies (see clusters 11, 6, and 14). The cluster with the highest claim frequency (see cluster 1) is also exclusively labelled with vehicles in central and semi-central parts of Sweden’s three largest cities (zone 1). The under-representation of female drivers in the portfolio is also reflected in the dominant gender (mostly men) associated with each cluster.

Cluster	% of policies	Gender	Zone	Class	Owner’s age	Vehicle age	Frequency (%)
13	4.83	M	[3,4]	1	[35,55]	[29,46]	0.2512
2	9.86	M	4	[4,5]	[40,52]	[12,17]	0.3141
3	9.52	M	[1,3]	[3,5]	[47,52]	[13,17]	0.3225
10	7.60	M	[2,3]	[3,5]	[42,46]	[12,17]	0.4842
9	8.09	K	[3,4]	[3,5]	[25,51]	[13,17]	0.5647
5	5.77	M	4	[3,5]	[47,55]	[1,6]	0.5936
7	9.19	M	[3,4]	[3,5]	[58,62]	[1,17]	0.6645
8	4.23	M	4	[3,5]	[25,34]	[11,16]	0.8449
4	6.46	M	[1,3]	[3,4]	[42,46]	[1,5]	1.0119
12	6.83	K	[3,4]	[3,4]	[24,52]	[1,6]	1.2278
15	7.23	M	[2,3]	[3,4]	[47,52]	[1,5]	1.2402
11	5.80	M	[2,3]	[3,6]	[24,31]	[11,15]	2.0485
6	4.38	M	4	[3,6]	[22,29]	[1,7]	2.3001
14	6.56	M	[2,3]	[3,6]	[23,29]	[1,8]	4.2849
1	3.66	M	1	[3,4]	[25,31]	[1,9]	8.0840

Table 1.14: Table summary of Burt distance-based K-means on mixed data types (dominant features and average claim frequencies per cluster).

The deviance, AIC, and BIC are reported in Table 1.15. Both AIC and BIC are computed with 15×23 degrees of freedom ($\#$ of clusters $\times$ $\#$ of modalities). The deviance, measuring our model’s goodness of fit, is clearly better than the one obtained with K-prototypes or K-medoids. In comparison, the deviance of a GLM model fitted to the same dataset is around 5 824.

Deviance	5978.96
AIC	8017.53
BIC	11136.99

Table 1.15: Statistics of goodness of fit obtained by partitioning the mixed-type dataset into 15 clusters with the Burt distance-based K-means algorithm.

Note that we chose the number of clusters K based on the marginal gain of intra-class inertia and the marginal reduction in deviance. As illustrated in Fig. 1.B.1 (in Appendix), the marginal gain of inertia and the marginal reduction in deviance (averaged over fifteen seed values) become limited above a certain number of clusters. Given the size of our motorcycle portfolio (60 000+ policies), we chose $K = 15$ in order to get closer to the GLM deviance (5 823.66) while still effectively summarizing the large amount of information in our data.

1.3 Burt-adapted K-means variants

In this Section, we extend the scope of application of our Burt framework to K-means variants, namely the mini-batch and fuzzy K-means. The mini-batch extension is particularly suited to address computational challenges associated with large-scale data, a common issue with insurance datasets, whereas fuzzy K-means is a soft form of clustering (each object can belong to multiple clusters).

1.3.1 Mini-batch K-means

The K-means algorithm stores the entire dataset in main memory, which can become computationally costly when handling large datasets. The mini-batch K-means algorithm was proposed as a solution by Sculley (2010). This extension uses small random batches of data with a fixed size. In each iteration, a new random sample from the dataset is used to update the clusters. This process continues until convergence. The specifics of how we adapted the mini-batch variant to our Burt framework are outlined in Algorithm 3 (in Appendix).

Empirical evidence suggests a substantial reduction in computational time, though often at the expense of partition quality. Fig. 1.5 compares the runtime, with respect to the number of clusters, of our Burt-adapted mini-batch K-means.

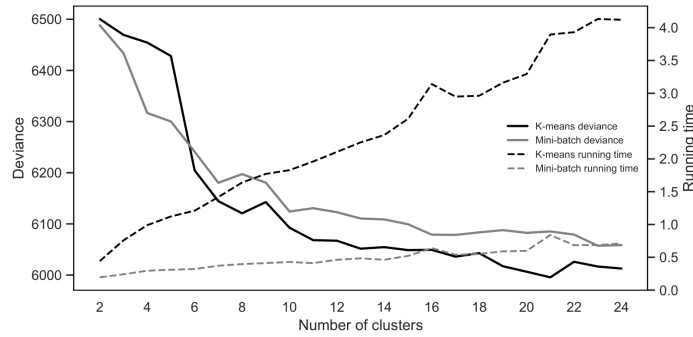


Figure 1.5: Partition quality (solid lines) measured by the average deviance over 10 random seeds, and runtime (dashed lines) with respect to the number of clusters.

While the discrepancy in partition quality (between the standard K-means and its mini-batch variant) remains small, the difference in computational time grows noticeably. The mini-batch algorithm exhibits a relatively stable runtime, while K-means becomes increasingly time-consuming. However, the overall gain in computation time is still limited (both algorithms complete within a few seconds). To observe significant gains in computation time, mini-batch K-means would need to be tested on a larger dataset than the Wasa dataset used here.

We applied the mini-batch algorithm to the Wasa dataset using Burt distance, defined in Eq. (1.2.10), with batch sizes of 5 000 policies. As described in Section 1.2.3, continuous variables are discretised. Table 1.16 presents the results for 15 centroids. The mini-batch K-means runs faster and gives approximately the same results. We observe more or less the same dominant modalities discovered in Table 1.14.

Cluster	% of policies	Gender	Zone	Class	Owner's age	Vehicle age	Frequency (%)
6	3.71	M	4	1	[35, 64]	[13, 48]	0.2625
11	7.94	M	4	[3,4]	[47, 59]	[12, 17]	0.2759
2	3.31	M	[1,3]	1	[29, 60]	[29, 45]	0.4523
10	6.91	M	4	[5,6]	[22, 46]	[12, 16]	0.5052
15	6.77	M	[1,4]	[3,4]	[42, 46]	[12, 17]	0.5295
8	8.06	K	[3,4]	[3,5]	[25, 51]	[13, 17]	0.5661
5	9.85	M	4	3	[47, 55]	[1, 5]	0.6560
9	8.58	M	[1,2]	[3,6]	[47, 54]	[13, 17]	0.7156
3	5.44	M	3	3	[23, 63]	[13, 17]	0.7496
14	8.26	M	[2,4]	[3,5]	[42, 46]	[1, 6]	0.9482
12	5.58	M	[2,3]	5	[24, 52]	[13, 17]	1.0196
7	6.82	K	[3,4]	[3,4]	[24, 52]	[1, 6]	1.2301
4	5.81	M	[2,3]	[4,5]	[47, 53]	[1, 7]	1.6504
1	4.89	M	4	[4,6]	[22, 28]	[5, 10]	3.2768
13	8.07	M	[1,3]	[3,5]	[23, 29]	[1, 6]	5.9861

Table 1.16: Burt-adapted mini-batch clustering. Policy allocation, dominant features, and average claim frequencies per cluster.

The deviance, reported in Table 1.17, is slightly higher than that of a segmentation based on the Burt-adapted K-means algorithm. Note that this minor reduction in cluster quality may be due to the increased presence of noise in the mini-batch algorithm.

Deviance	6049.68
AIC	8088.25
BIC	11207.71

Table 1.17: Statistics of goodness of fit obtained by partitioning the dataset into 15 clusters with the Burt-adapted mini-batch K-means.

1.3.2 Fuzzy K-means

The K-means algorithm, often referred to as hard clustering, attempts to partition a dataset into K distinct clusters. Each data point is assigned to the cluster whose centre is the nearest and may only belong to one cluster. In contrast, the fuzzy K-means algorithm proposed by Dunn (1973) and later improved by Bezdek et al. (1984) is known as soft clustering, as it allows data points to belong to more than one cluster. Given a finite set of data, the algorithm returns two outputs. First, a set of K cluster centers $\{\mathbf{c}_1, \dots, \mathbf{c}_K\}$, and second, a partition matrix W , which contains the membership values $w_{i,u}$ for $i = 1, \dots, n$ and $u = 1, \dots, K$. Like K-means, fuzzy K-means aims to minimize the intra-cluster variance:

$$\arg \min \sum_{i=1}^n \sum_{u=1}^K w_{i,u}^\alpha d(\mathbf{x}_i, \mathbf{c}_u),$$

where

$$w_{i,u} = \frac{1}{\sum_{k=1}^K \left(\frac{d(\mathbf{x}_i, \mathbf{c}_u)}{d(\mathbf{x}_i, \mathbf{c}_k)} \right)^{\frac{2}{\alpha-1}}}.$$

The hyper-parameter $\alpha \in \mathbb{R}^+$ determines the level of cluster fuzziness. A large α results in smaller membership grades $w_{i,u}$, and hence fuzzier clusters. In the limit $\alpha = 1$, the memberships, $w_{i,u}$, converge to 0 or 1, which amounts to crisp partitioning.

The arbitrary centroid initialisation and the pre-specified number of clusters both retain their significant influence over the final partition. The Burt-adapted fuzzy K-means is detailed in Algorithm 4 (in Appendix).

We apply the Burt-adapted fuzzy K-means to the Wasa insurance dataset. We once again discretise the continuous variables as explained in Section 1.2.3. We run the algorithm with a fuzziness parameter of $\alpha = 1.20$. The policies are assigned to the most likely cluster (highest $w_{i,u}$ for $u = 1, \dots, K$). Note that, if we set $\alpha = 2$, which is a standard level of fuzziness in the literature, several clusters are not assigned any policies, but some policies have a non-null probability of belonging to them.

Table 1.18 reports the most frequent features and the average claim frequency in each cluster. We find similar profiles to those found in Table 1.14 with more recent (older) vehicles with higher (lower) ratio power driven by younger (older) policyholders in clusters with higher (lower) claim frequencies on average.

The deviance, in Table 1.19, is better than that of the mini-batch K-means, and remains quite close to the one obtained with our Burt-adapted K-means.

Cluster	% of policies	Gender	Zone	Class	Owner's age	Vehicle age	Frequency (%)
2	5.16	M	4	[3,5]	[47,52]	[12,17]	0.2604
7	4.72	M	[3,4]	1	[35,60]	[29,46]	0.2992
14	8.21	M	[2,3]	[3,5]	[47,52]	[13,17]	0.3552
6	4.92	M	4	[3,5]	[41,46]	[12,18]	0.3586
9	6.02	M	[3,4]	[3,5]	[58,63]	[13,17]	0.4490
12	7.60	M	[2,3]	[3,5]	[42,46]	[12,17]	0.4842
4	3.78	M	[2,4]	3	[47,52]	[1,5]	0.6654
10	4.40	M	[2,4]	[3,4]	[58,62]	[1,5]	0.8285
1	4.23	M	4	[3,5]	[25,34]	[11,16]	0.8449
11	14.88	K	[3,4]	[3,4]	[24,51]	[1,17]	0.8598
3	8.47	M	[2,4]	[3,5]	[42,46]	[1,6]	0.9340
5	7.29	M	[2,4]	[4,5]	[47,52]	[1,7]	1.0891
13	4.44	M	4	[3,6]	[22,29]	[1,7]	2.2784
15	7.08	M	[2,3]	[3,5]	[25,32]	[11,15]	2.4298
8	8.80	M	[2,3]	[3,6]	[23,29]	[1,7]	5.7983

Table 1.18: Burt-adapted fuzzy clustering. Policy allocation, dominant features, and average claim frequencies per cluster.

Deviance	6026.25
AIC	8064.82
BIC	11184.27

Table 1.19: Statistics of goodness of fit obtained by partitioning the dataset into 15 clusters with the Burt-adapted fuzzy K-means.

1.4 Burt-adapted spectral clustering

Since K-means relies on Euclidean distance, each iteration partitions the space into (hyper)spherical clusters. This makes K-means unsuitable for detecting clusters with non-convex shapes. For example, in Fig. 1.6, K-means fails to distinguish the inner and outer rings or the upper and lower moons of the circular and moon-shaped data. This poor clustering performance stems from the fact that the natural clusters do not form convex regions. By exploiting a deeper data geometry, spectral clustering (J. Shi et al. 2000; A. Y. Ng et al. 2001; Belkin et al. 2001) addresses the lack of flexibility of K-means in cluster boundaries. While K-means attempts to associate each cluster with a hard-edged sphere, spectral clustering embeds the data into a new space derived from a graphical representation of the dataset. Applying the K-means algorithm to this representation bypasses the spherical limitation, and results in the intended non-convex cluster shapes.

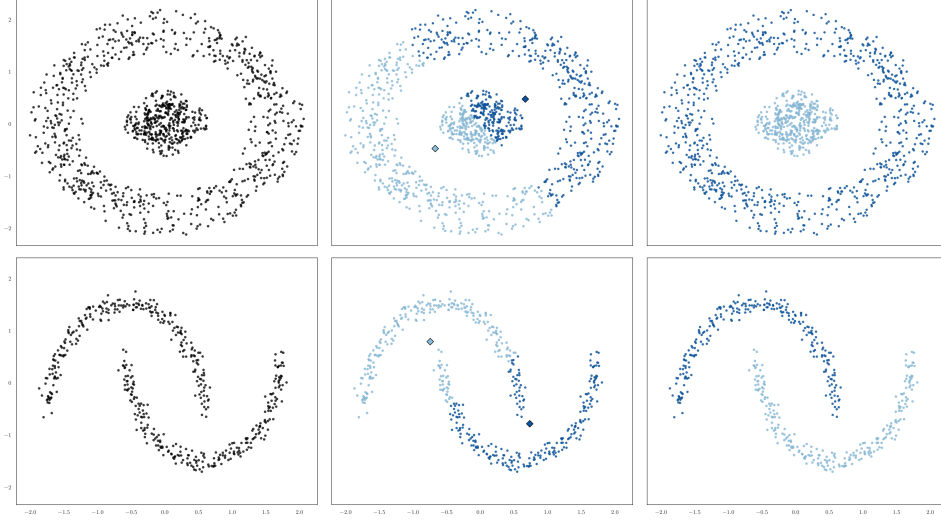


Figure 1.6: Illustration of the partition of non-convex data (left) with the K-means (middle) and spectral clustering (right) algorithms. Each cluster is identified by a color. The centroids are represented with a diamond shape.

The graph theory is the core component in the graph representation of the dataset. As illustrated in Fig. 1.7, an undirected graph $G = (V, E, W)$ is defined by three elements:

1. a set of vertices $V = \{v_i\}_{i=1, \dots, n}$ where each vertex v_i represents one of the n data points in the dataset;
2. a set of edges $E = \{e_{i,i'} : v_i \longleftrightarrow v_{i'}\}$ whose entries are equal to 1 if two vertices v_i and $v_{i'}$ are connected by an undirected edge $e_{i,i'}$, and 0 otherwise;
3. a set of weights $W = \{w_{ii'} : w_{ii'} \neq 0 \text{ if } v_i \longleftrightarrow v_{i'}\}$ that contains the similarity between two vertices linked by an edge.

E and W can both be represented as $n \times n$ matrices. The matrix W is often referred to as the weighted adjacency matrix A .

The dataset $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, where $\mathbf{x}_i \in \mathbb{R}^m$, can thence be represented as a graph $G = (V, E, W)$ by first associating each data point $\mathbf{x}_i$ with a vertex v_i . A measure of similarity is then defined between two data points i and i' . A common choice is the Gaussian kernel:

$$S(i, i') = \exp\left(-\frac{d(i, i')}{\alpha}\right), \quad (1.4.1)$$

where $\alpha \in \mathbb{R}^+$ is a tuning parameter. Two highly similar data points are connected by an edge $e_{i,i'}$ with a weight $w_{ii'}$ equal to their similarity measure, while data

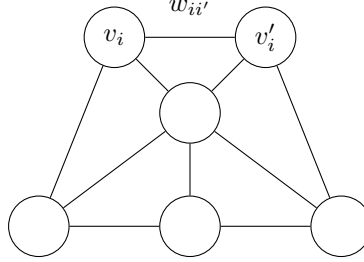


Figure 1.7: Vertices, edges, and weights of a graph.

points with low similarity are considered disconnected. Various methods exist for constructing a pairwise similarity graph. Notably, the ε -neighborhood graph, the (mutual) k -nearest neighbor graph, and the fully connected graph (where all points in the dataset are connected).

In the ε -neighborhood graph, points are connected if the pairwise distance between them is smaller than a threshold ε . In practice, we keep edges with similarity above $\exp(-\varepsilon/\alpha)$, and set all other similarities to zero.

In the (mutual) k -nearest neighbor graph, the vertex v_i is connected to vertex $v_{i'}$ if $v_{i'}$ is among the k -nearest neighbors of v_i . Since the neighborhood relationship is not symmetric, and the graph must be symmetric, we need to enforce the symmetry. The graph is made symmetric by ignoring the directions of the edges, i.e., by connecting v_i and $v_{i'}$ if either v_i is among the k -nearest neighbors of $v_{i'}$ or $v_{i'}$ is among the k -nearest neighbors of v_i , resulting in the k -nearest neighbor graph. Alternatively, the connection is made if v_i and $v_{i'}$ are mutual k -nearest neighbors, resulting in the mutual k -nearest neighbor graph.

Note that the choice of rule for constructing the graph influences the edge matrix and may affect the resulting clusters. In our analysis, we chose to construct the similarity graph using a symmetric k -nearest neighbor graph specified in Python through the spectral clustering implementation from the scikit-learn package.

To work with the graph representation, we apply the graph Fourier transform, which is based on the Laplacian representation, denoted L , of a graph G , given by the difference between its diagonal degree matrix D and adjacency matrix A :

$$L = D - A, \quad (1.4.2)$$

where the degree matrix D is the diagonal matrix defined by:

$$D_{ii} = \sum_{i'} w_{i,i'}, \quad D_{ii} = 0 \quad \text{for } i \neq i',$$

and the adjacency matrix is given by:

$$A_{ii'} = \begin{cases} w_{i,i'} & \text{if vertices } v_i \text{ and } v_{i'} \text{ are connected,} \\ 0 & \text{otherwise.} \end{cases}$$

The quantity D_{ii} represents the degree of a vertex v_i , that is, the total weight of edges connected to v_i . In the unweighted case, this reduces to the number of neighboring vertices connected to v_i .

The matrix L , defined in Eq. (1.4.2), is called a graph Laplacian. This terminology comes from the analogy with the classical Laplace operator in continuous spaces, which measures how a function varies locally around a point. Similarly, the graph Laplacian measures how a function defined on the vertices of a graph varies across neighboring vertices.

To better understand this, consider a function $f : V \rightarrow \mathbb{R}$ assigning a real value to each vertex v_i of the graph. If two connected vertices have very different function values, the graph Laplacian produces a large variation term. Conversely, neighboring vertices with similar values produce a small variation.

This analogy becomes particularly intuitive for discrete periodic functions. Consider a function taking values at times $1, 2, \dots, N$ with periodic boundary conditions. Such a structure can naturally be represented by a ring graph, as illustrated in Fig. 1.8, where the first and last vertices are connected to reflect the periodicity.

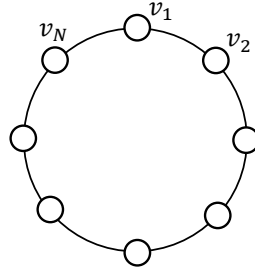


Figure 1.8: Ring representation of a period with N steps.

In Fig. 1.8, none of the vertices have edges to themselves, so diagonal elements of A are 0. Because our graph is undirected, $E_{ii'} = E_{i'i}$. In the resulting Laplacian matrix, the diagonal entries are the degrees of the vertices, and the off-diagonal elements are the negative edge weights:

$$L = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & \vdots \\ 0 & 0 & 2 & 0 \\ 0 & \dots & 0 & 2 \end{pmatrix} - \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & \vdots \\ 0 & 1 & 0 & 1 \\ 0 & \dots & 1 & 0 \end{pmatrix} = \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & \vdots \\ 0 & -1 & 2 & -1 \\ 0 & \dots & -1 & 2 \end{pmatrix}.$$

The role of the graph Laplacian becomes clearer when interpreting it as an operator acting on functions defined on the vertices of the graph. Let

$$\mathbf{f} = (f(v_j))_{j=1, \dots, N},$$

represent the vector of function values evaluated at the vertices of a graph G . Its product with the graph Laplacian $L\mathbf{f}$, corresponds precisely to the second-order finite difference derivative of the function $f(\cdot)$. This operation plays a key role in analysing the graph's structure, as the partitioning of the graph (including determining the number of clusters) can be inferred from the eigenvalues and eigenvectors of the graph Laplacian matrix, a process known as spectral analysis.

Since L is a symmetric matrix, it can be decomposed using eigenvalue decomposition $L = U\Sigma U^\top$ where U is the matrix of eigenvectors, and Σ is a diagonal matrix containing the eigenvalues. As the Laplacian is positive semi-definite, all its eigenvalues are nonnegative. Moreover, because each row of L sums to 0, the constant vector $\mathbf{1}$ satisfies $L\mathbf{1} = \mathbf{0}$, implying that 0 is always the smallest eigenvalue of L .

The analysis of the eigenvalues provides useful insights into the graph's structure. Specifically, the multiplicity of the zero eigenvalue reflects the number of disconnected components in the graph, giving a way to partition the graph. For instance, if all vertices are disconnected, all eigenvalues will be zero. As edges are added, some eigenvalues become non-zero, and the number of zero eigenvalues corresponds to the number of connected components in the graph. To illustrate this, consider the graph in Fig. 1.9, composed of K disconnected sub-graphs. In this case, the Laplacian matrix will have K zero eigenvalues, indicating that the graph consists of K independent components.

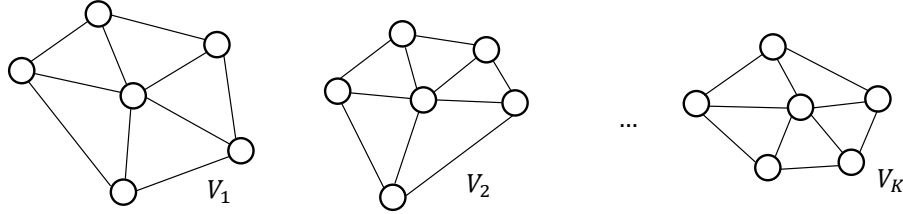


Figure 1.9: Graph with K sub-graphs.

Beyond identifying the number of connected components, the eigenvalues of the graph Laplacian provide information about the density of the connections. The smallest non-zero eigenvalue, known as the spectral gap, indicates the overall connectivity density of the graph. A large spectral gap suggests a densely connected graph, while a small spectral gap implies that the graph is close to being disconnected. Thus, the first positive eigenvalue offers a continuous measure of the graph's connectedness. Moreover, the second smallest non-zero eigenvalue, often referred to as the Fiedler value or the algebraic connectivity, is particularly useful for partitioning the graph. This value approximates the minimum graph cut required to divide the graph into two connected components. For example, if the graph consists of two groups of vertices, V_1 and V_2 , connected by an additional

edge (as shown in Fig. 1.10), vertices v_1 and v_2 on either side of the cut can be assigned to V_1 or V_2 by analyzing the values in the Fiedler vector.

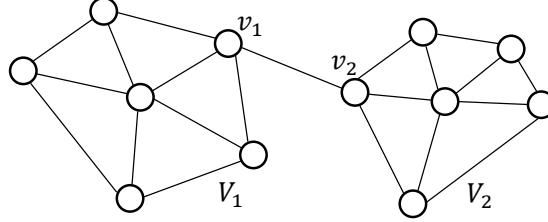


Figure 1.10: Two weakly connected sub-graphs.

For graphs consisting of K well-separated clusters (i.e., weak inter-cluster connectivity), the first K eigenvectors associated with the zero eigenvalue are approximately piecewise constant on the connected components (Paccanaro 2006). In other words, vertices belonging to the same component (i.e., points belonging to the same cluster) have identical coordinates in the spectral embedding. Intuitively, since the eigenvectors of zero tell us how to partition the graph, the first K columns of U (the eigenvectors' coordinates) correspond to the indicator vectors of the connected components (i.e., clusters).

Proposition 1.1 (Piecewise constancy of eigenvectors.). *Let $G = (V, E)$ be an undirected weighted graph with non-negative weights ($w_{ii'} \geq 0$) and let $L = D - A$ denote its graph Laplacian. Suppose that G consists of K connected components $V_1, \dots, V_K$. Let $U \in \mathbb{R}^{n \times K}$ contain an orthonormal basis of the eigenspace corresponding to the eigenvalue 0. Then, Meilă et al. (2001) showed that the eigenvalue 0 has multiplicity K , its eigenspace is spanned by the indicator vectors $\mathbf{1}_{V_1}, \dots, \mathbf{1}_{V_K}$, and the spectral embedding defined by the rows of U ,*

$$v_i \mapsto U_{i,:} \in \mathbb{R}^K,$$

is constant on each connected component, i.e.

$$U_{i,:} = U_{i',:} \quad \text{for all } v_i, v_{i'} \in V_k.$$

Consequently, the connected components of G can be recovered by clustering the row vectors of U . The proof is given in Appendix 1.D

In practice, the component structure is unknown (i.e., the K clusters are not identified). For graphs consisting of K well-separated clusters, the Laplacian can be interpreted as a perturbation of the ideal block-diagonal case. In this setting, the first K eigenvectors remain approximately piecewise constant (see, e.g., Meilă et al. (2001)). Hence, the rows of the truncated eigenvector matrix provide low-dimensional representations of the vertices, and clustering these representations

yields an approximation of the underlying partition. This motivates the spectral clustering procedure described in Algorithm 5 in Appendix.

As demonstrated by Luxburg (2007), spectral clustering’s performance is superior to other widely used clustering algorithms, particularly due to its ability to accommodate for non-convex cluster shapes. To illustrate this, we apply spectral clustering to the circular dataset introduced earlier in this Section, which consists of 1 200 data points (800 in the outer ring, and 400 in the inner circle).

For this analysis, we construct a graph using the mutual k -nearest neighbors with $k = 20$ and set the similarity parameter to $\alpha = 1$. The left plot of Fig. 1.11 clearly shows that spectral clustering effectively identifies the distinct inner and outer rings of the dataset. In the right plot, we present the coordinates of all pairs of eigenvectors, $(U_{i,1}, U_{i,2})_{i=1,\dots,1200}$. Consistent with the above property, we observe in the right plot of Fig. 1.11 that the eigenvector coordinates for points within the same cluster are identical (they are superimposed).

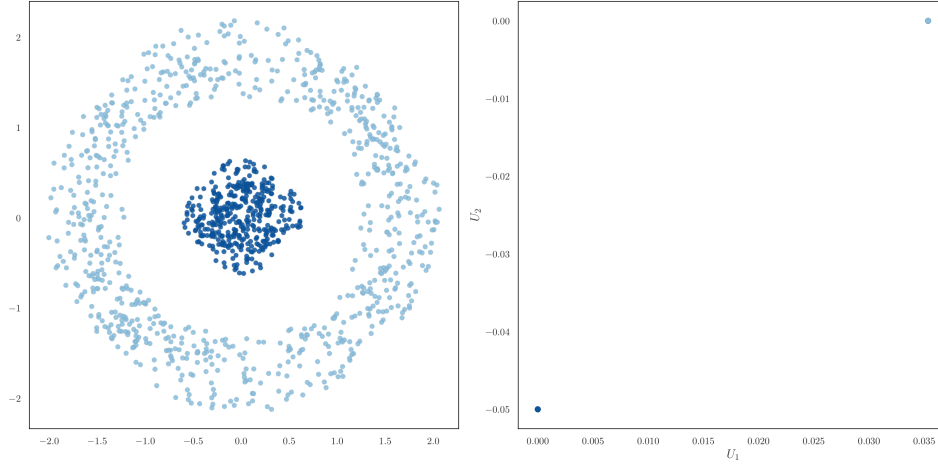


Figure 1.11: (Left) Partition of a non-convex dataset with spectral clustering. (Right) Pairs of eigenvectors’ coordinates $(U_{i,1}, U_{i,2})_{i=1,\dots,n}$.

In real-world applications, performing spectral clustering on large datasets presents significant challenges, primarily due to the rapid growth in the size of the edge and weight matrices (E, W) . When only few vertices are connected, it is possible to represent (E, W) as a sparse matrix, where most elements are zero, thus saving memory and computational resources. An alternative approach to manage this complexity consists of first using the K-means algorithm to reduce the number of data points to a manageable number of centroids, and then perform spectral clustering on these centroids. Fig. 1.12 demonstrates the effectiveness of this approach. The figure shows the 100 centroids obtained from applying K-means to the original dataset of 1 200 points, along with their corresponding cluster. The

right plot illustrates the coordinates of the pairs of eigenvectors, $(U_{i,1}, U_{i,2})$, for i ranging from 1 to 100, reflecting the results of the spectral clustering applied to these centroids.

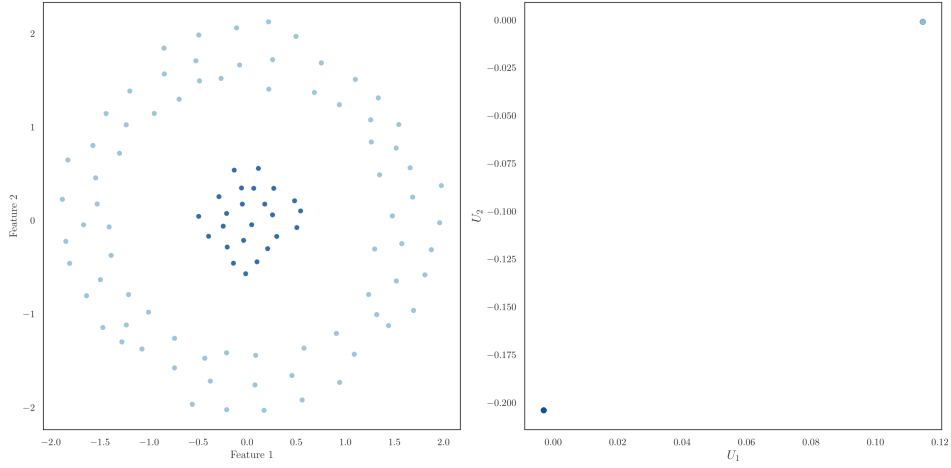


Figure 1.12: (Left) Spectral clustering partitioning of a non-convex dataset that has been preliminarily reduced with the K-means algorithm. (Right) Pairs of eigenvectors' coordinates $(U_{i,1}, U_{i,2})_{i=1,\dots,n}$.

To conclude this Section, we apply the spectral clustering algorithm to the Wasa insurance dataset. We first convert the variables driver's age and vehicle age into categorical variables, as described in Section 1.2.3.

We then compute the disjunctive table $\mathbf{D}$ and the weighted Burt matrix $\mathbf{B}^{\mathbf{W}}$ using Burt distance. Following the procedure outlined in Section 1.2.2, each insurance policy $i = 1, \dots, N$ characterised by multiple modalities is represented by its centre of gravity in Burt space, i.e., $\mathbf{x}_i \rightarrow \mathbf{D}_{i,\cdot} \mathbf{B}^{\mathbf{W}} / m_{\text{cat}}$.

The dataset contains 62 436 contracts. We reduce its dimension by applying our Burt distance-based K-means algorithm with 1 500 centroids. We then construct a graph based on these centroids using the mutual k-nearest neighbors (with $k = 70$) and a similarity parameter $\alpha = 1$. We run the spectral clustering algorithm on this reduced dataset with $K = 15$ clusters.

Table 1.20 summarizes the results, including the policy allocation, the dominant features, and the average claim frequency for each cluster. The application of spectral clustering to the reduced dataset, processed through our Burt distance-based K-means, demonstrates its effectiveness in discriminating between drivers with different risk profiles.

Notably, unlike K-means (and its mini-batch and fuzzy variants), spectral clustering can identify a greater number of clusters predominantly consisting of female drivers, who are underrepresented in the insurance portfolio. Additionally, the

allocation of policies differs significantly from previous results, with the cluster exhibiting the lowest claim frequency containing the majority of policies (approximately 40%). This indicates that spectral clustering effectively identifies the rarity of claim events. Table 1.21 further supports our method's efficacy, achieving reasonable goodness of fit in terms of deviance.

Cluster	% policies	Gender	Zone	Class	Owner age	Vehicle age	Frequency (%)
5	40.18	M	4	[3,5]	[42,53]	[12,18]	0.3785
4	2.28	M	[3,4]	1	[60,66]	[1,45]	0.4424
7	4.04	K	[2,4]	[3,5]	[23,33]	[13,17]	0.6854
9	6.04	K	[3,4]	[3,4]	[42,50]	[1,16]	0.7040
2	1.78	K	[3,4]	[3,6]	[23,54]	[6,11]	1.1766
14	10.34	M	[1,3]	[3,5]	[44,52]	[1,5]	1.2747
1	5.43	M	[1,4]	3	[24,60]	[1,5]	1.4537
6	9.56	M	[2,3]	[4,6]	[25,55]	[12,16]	1.5390
12	4.89	M	4	[4,5]	[21,58]	[2,11]	1.5788
3	4.71	M	[2,4]	6	[28,57]	[7,11]	1.7313
10	1.64	K	[1,4]	3	[24,54]	[1,5]	2.1993
15	1.56	M	[2,3]	[4,5]	[28,58]	[2,3]	2.6608
13	1.66	M	[2,4]	[2,4]	[23,57]	1	3.1749
11	1.49	M	1	[3,4]	[28,33]	[1,5]	6.5044
8	4.40	M	[2,3]	[4,6]	[24,27]	[4,11]	6.6864

Table 1.20: Burt-adapted spectral clustering on the Wasa dataset preliminarily reduced with the Burt-adapted K-means algorithm. Dominant features and average claim frequencies per cluster.

Deviance	6069.01
----------	---------

Table 1.21: Statistic of the goodness of fit obtained by partitioning the dataset into 15 clusters with the Burt-adapted spectral algorithm.

1.5 Conclusion

This research chapter explores a Burt distance-based approach to extend the scope of well-established clustering techniques to actuarial applications. The main challenge consists in defining Burt distance as a dissimilarity measure between observations characterised by both numerical rating factors and categorical classes.

Our empirical analysis reveals that our Burt distance-based K-means is able to identify relevant clusters of policies. Numerical tests and benchmarking against K-means, K-modes, K-prototypes, and K-medoids demonstrate the robustness of our proposed method on categorical, continuous, and mixed data types. We further

assess the applicability of our Burt approach to K-means variants. The mini-batch variant was shown to be particularly suitable for large-scale insurance portfolios. The second variant, based on fuzzy logic, allows for an observation to belong to multiple clusters. Our numerical experiment reveals that both Burt-adapted variants give consistent results on our dataset.

In the last part of this chapter, we adapt spectral clustering to our Burt approach. It requires a graphical representation that is particularly costly in terms of computer resources when analyzing large datasets. We address this drawback by introducing a data reduction method using our Burt-adapted K-means algorithm with a large number of centroids. Numerical tests reveal that this approach is also competitive in terms of deviance.

By leveraging these different clustering techniques, we intend for practitioners to gain deeper and more meaningful insights into the underlying structure of complex insurance datasets.

1.A Tables

Rating factor	Class	Class description
Gender	M	Male (ma)
	K	Female (kvinnor)
Geographic area	1	Central and semi-central parts of Sweden's three largest cities
	2	Suburbs plus middle-sized cities
	3	Lesser towns, except those in classes 5 or 7
	4	Small towns and countryside
	5	Northern towns
	6	Northern countryside
	7	Gotland (Sweden's largest island)
Vehicle class	1	EV ratio 1–5
	2	EV ratio 6–8
	3	EV ratio 9–12
	4	EV ratio 13–15
	5	EV ratio 16–19
	6	EV ratio 20–24
	7	EV ratio 25+

Table 1.A.1: Description of the modalities of the categorical rating factors Zone, Class, and Gender, from the motorcycle insurance dataset by Ohlsson et al. (2010).

1.B Figures

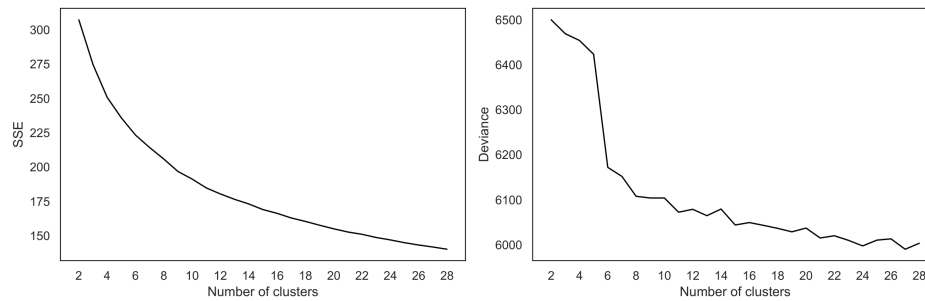


Figure 1.B.1: Burt distance-based K-means metrics. Left plot: evolution of the total intra-class inertia. Right plot: evolution of the deviance.

1.C Algorithms

Algorithm 3: Mini-batch K-means algorithm.

Data: Dataset $\mathcal{X}$, number of clusters K , batch size n_{batch}

Initialisation

Initialize centroids $\{\mathbf{c}_u(0)\}_{u=1}^K$ (e.g., randomly from $\mathcal{X}$)
Initialize cluster counts $N_u \leftarrow 0$, $u = 1, \dots, K$
Compute Burt transformed observations $\mathbf{x}_i = \mathbf{D}_{i,\cdot} \mathbf{B}^{\mathbf{W}} / m_{\text{cat}}$
Set iteration counter $e \leftarrow 0$

while *not converged* **do**

Sample a mini-batch $M \subset \mathcal{X}$ of size n_{batch}

Initialize empty batch clusters $S_u \leftarrow \emptyset$, $u = 1, \dots, K$

for $\mathbf{x}_i \in M$ **do**

Assign point to nearest centroid:

$$u(i) = \arg \min_{u \in \{1, \dots, K\}} d(\mathbf{x}_i, \mathbf{c}_u(e))$$

Update assignment:

$$S_{u(i)} \leftarrow S_{u(i)} \cup \{\mathbf{x}_i\}$$

for $u = 1$ **to** K **do**

if $|S_u| > 0$ **then**

Compute mini-batch centroid:

$$\mathbf{c}_u^{\text{batch}} = \frac{1}{|S_u|} \sum_{\mathbf{x}_i \in S_u} \mathbf{x}_i$$

Update cluster count:

$$N_u \leftarrow N_u + |S_u|$$

Update centroid using incremental moving average:

$$\mathbf{c}_u(e+1) = (1 - \eta_u(e)) \mathbf{c}_u(e) + \eta_u(e) \mathbf{c}_u^{\text{batch}}$$

where

$$\eta_u(e) = \frac{|S_u|}{N_u}$$

else

$\mathbf{c}_u(e+1) \leftarrow \mathbf{c}_u(e)$

$e \leftarrow e + 1$

Algorithm 4: Fuzzy clustering.**Data:** Dataset $\mathcal{X}$, number of clusters K , fuzziness $\alpha > 1$ **Initialisation**

- └ Randomly select K initial centroids $\mathbf{c}_1(0), \dots, \mathbf{c}_K(0)$
- └ Set iteration counter $e \leftarrow 0$

while *not converged* **do****for** $i = 1$ *to* n **do**Update membership of $\mathbf{x}_i$ to each cluster $u = 1, \dots, K$:

$$w_{i,u}^\alpha(e) = \frac{1}{\sum_{v=1}^K \left(\frac{d(\mathbf{x}_i, \mathbf{c}_v(e))}{d(\mathbf{x}_i, \mathbf{c}_u(e))} \right)^{\frac{2}{\alpha-1}}}$$

for $u = 1$ *to* K **do**Update centroid of cluster u :

$$\mathbf{c}_u(e+1) = \frac{\sum_{i=1}^n w_{i,u}^\alpha(e) \mathbf{D}_{i,\cdot} \mathbf{B}^W / m_{\text{cat}}}{\sum_{i=1}^n w_{i,u}^\alpha(e)}$$

Compute total weighted distance:

$$d^{\text{total}}(e+1) = \sum_{u=1}^K \sum_{i=1}^n w_{i,u}^\alpha(e) d(\mathbf{x}_i, \mathbf{c}_u(e+1))$$

└ $e \leftarrow e + 1$ **Algorithm 5:** Spectral clustering.**Data:** Dataset $\mathcal{X}$ **Initialisation**

- └ Represent $\mathcal{X}$ as a graph $G = (V, E, W)$

Calculate the Laplacian matrix $L = D - A$ Compute the eigendecomposition $L = U \Sigma U^\top$ Fix k and construct $U^{(k)} \in \mathbb{R}^{n \times k}$ from the eigenvectors corresponding to the k smallest eigenvaluesApply K-means (Algorithm 1) to the rows $U_{i,\cdot}^{(k)}$, for $i = 1, \dots, n$ **Output:** Assign each data point to the cluster corresponding to $U_{i,\cdot}^{(k)}$

1.D Proofs

Proof of Proposition 1.1. If G is connected (with $K = 1$), then for any u satisfying $Lu = 0$ (i.e., u is an eigenvector with 0 eigenvalue),

$$u^\top Lu = \sum_{i,i'=1}^n w_{ii'}(u_i - u_{i'})^2.$$

Since all weights $w_{ii'}$ are non-negative, the quadratic form equals zero if and only if:

$$w_{ii'}(u_i - u_{i'})^2 = 0 \quad \text{for all } i, i'.$$

This condition is met whenever the two vertices $u_i, u_{i'}$ are disconnected (i.e., $w_{ii'} = 0$). Thus whenever the two vertices are connected (i.e., $w_{ii'} > 0$), we must have $u_i = u_{i'}$. By connectivity, this implies that u is constant on V . Hence, for $K = 1$, the nullspace of L is one-dimensional and spanned by $\mathbf{1}$.

If G consists of $K > 1$ connected components $V_1, \dots, V_K$, then after ordering vertices according to component (i.e., sub-graph) membership, the Laplacian takes block-diagonal form:

$$L = \begin{pmatrix} L_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & L_K \end{pmatrix}.$$

Each block L_k corresponds to a connected graph and therefore has a one-dimensional nullspace spanned by the constant vector on V_k . The nullspace of L is the direct sum of these subspaces and thus has dimension K . A basis is given by the indicator vectors $\mathbf{1}_{V_k}$.

Let U contain an orthonormal basis of this eigenspace. Each column of U is constant on every V_k . Hence, for any two vertices $v_i, v_{i'} \in V_k$, their row representations coincide:

$$U_{i,.} = U_{i',.}.$$

Therefore, the spectral embedding maps all vertices of the same connected component to the same point in $\mathbb{R}^K$. Clustering the row vectors of U thus exactly recovers the connected components. □

Chapter 2

Variational AutoEncoder for synthetic insurance data

This chapter explores the application of Variational AutoEncoders (VAEs) to insurance data. Previous research has demonstrated the successful implementation of generative models, especially VAEs, across diverse domains such as image recognition, text classification, and recommender systems. However, their application to insurance data, particularly heterogeneous insurance portfolios with mixed continuous and discrete attributes, remains largely unexplored. This study provides novel insights into the use of VAEs for unsupervised learning tasks in actuarial science, including dimensionality reduction and synthetic data generation. We propose a VAE model with a quantile transformation for continuous (latent) variables, a reconstruction loss that combines categorical cross-entropy and mean squared error, and a KL divergence-based regularisation term. The architecture of our VAE circumvents the need to pre-train or fine-tune a neural network to encode categorical variables into n -dimensional representative vectors within a continuous vector space of dimension $\mathbb{R}^n$. We assess the model's ability to reconstruct complex insurance data and generate synthetic insurance policies using a motor portfolio. Our experimental results demonstrate the potential of VAEs to address challenges associated with data availability in the insurance industry.

Charlotte Jamotton and Donatien Hainaut (2024).
“Variational AutoEncoder for Synthetic Insurance
Data”. In: *Intelligent Systems with Applications*
24, p. 200455. ISSN: 2667-3053. DOI: [10.1016/
j.iswa.2024.200455](https://doi.org/10.1016/j.iswa.2024.200455)

2.1 Introduction

The insurance industry relies heavily on data to accurately assess its risk exposure, which ultimately determines the success of any business model built on risk management. However, insurance data often contains sensitive personal information, such as the policyholder’s name, address, or gender, which are subject to regulatory requirements¹. Simply removing these identification factors does not guarantee complete anonymisation of the data. Moreover, the more data we conceal, the less valuable information is left for future analysis. This data anonymity issue has sparked interest in synthetic data and its associated deep learning-based generative models.

A generative model, or synthesis, is trained to generate artificial data that replicate the structure and characteristics of the original dataset. Therefore, statistical analysis performed on authentic or synthetic data should yield sensibly comparable results. In actuarial science, such generative processes could provide access to unlimited data that mimics the risk profiles of actual policyholders. Compliance with privacy data regulations becomes less of a concern, as computer-generated data is no longer personally identifiable information or associated with a data subject. Synthetic data could also serve as a valuable insurance tool in situations where actual data is unavailable (as is often the case when designing a new product) or to correct biased training datasets. For example, skewed representations with respect to gender, race, or other minority groups could be mitigated by generating additional representative policies.

The two most prominent approaches to generative modelling are Generative Adversarial Networks (GANs) (I. J. Goodfellow et al. 2014) and Variational Autoencoders (VAEs) (Kingma et al. 2013). Both have demonstrated success in generating realistic images, text, and audio. While GANs are often praised for their ability to generate high-quality visual samples, they are also singled out for their time-consuming iterative optimisation. VAEs, on the other hand, are known for their relatively straightforward implementation and hyper-parameters setup, as well as their explicit inference network and tractable likelihood, though they tend to generate blurry or fuzzy outputs (see, e.g., Zamany et al. (2022)). The bottleneck architecture, unique to autoencoders, led us to prioritise VAEs over the computationally expensive GANs in this research chapter. The appeal of a bottleneck layer stems from the manifold hypothesis, which suggests that although data may have hundreds of dimensions, its intrinsic structure can be sufficiently described in a lower-dimensional space. This manifold learning approach is analogous to human perception, where we focus on core features rather than all intricate details (e.g., we do not recall every aspect of a dog appearance to recognise or distinguish it from a cat, we focus on core characteristics instead). This hypothesis is the motivation

¹The EU General Data Protection Regulation (GDPR) on data protection and privacy, and the European Court of Justice ruling in the Test-Achats case (C-236/09) concerning sex discrimination came into effect in 2018 and 2011, respectively.

behind dimensionality reduction techniques such as Principal Component Analysis (Abdi et al. 2010), t-distributed Stochastic Neighbour Embedding (Maaten et al. 2008), or autoencoders (often defined as non-linear PCA). Ideally, when applied to insurance data, an autoencoder should learn some compressed representation of policies’ descriptive attributes, and policyholders with similar risk profiles should be mapped to nearby points in the latent space, facilitating clustering or the generation of new, meaningful policies.

In actuarial literature, Neural Networks (NNs) have predominantly been used for supervised learning, with only a few exceptions. For example, Hainaut (2018b) and Denuit, Hainaut, et al. (2019a) explored the predictive potential of unsupervised self-organising maps for non-life insurance data. Early works on supervised actuarial applications Hainaut (2018a), Noll et al. (2018), and Ferrario et al. (2018) trained networks to learn a mapping $f : \mathcal{X} \rightarrow \mathcal{Y}$, where $\mathcal{X}$ is an insurance portfolio defined by n policies and m rating factors (e.g., driver age, years of driving experience, vehicle brand), and $\mathcal{Y}$ is either a continuous (claim frequency) or discrete (number of claims) target. In contrast, the unsupervised learning approach adopted here seeks to learn and understand the structure of the inputs themselves, rather than trying to map inputs to target.

This chapter aims to find a low-dimensional representation of heterogeneous tabular insurance data (i.e., to reduce and blur the rating factors space) and to use the decoder network to generate synthetic insurance policies. Our contributions are as follows:

1. The application of a Variational AutoEncoder, a Bayesian extension of the classical autoencoder (Kramer 1991), to insurance data.
2. A quantile transformation to pre-process multi-modal, non-Gaussian data.
3. An all-in-one training of a VAE for heterogeneous data (no need to pre-train layers to fine-tune a numerical representation for the categorical attributes) with a loss function that combines a reconstruction loss, mixing categorical cross-entropy and mean squared error, with a Kullback-Leibler (KL) divergence regularisation term.
4. The generation of high-quality synthetic insurance policies using a quantile transformation of the latent space.

The chapter is structured as follows. Section 2.2 reviews the state of the art. Section 2.3 introduces the VAE model and motivates its use as a generative process. Section 2.4 presents our main contributions and adjustments to the classical VAE. Section 2.5 details the experimental setup and results. Subsection 2.5.1 describes the heterogeneous insurance portfolio used to evaluate the quality of our VAE model. Subsection 2.5.2 reviews the model architecture and hyperparameters setup, evaluates the reconstructed portfolio’s quality and latent space structure, and illustrates the impact of introducing the quantile transformation. Subsection 2.5.4

investigates whether our VAE can generate high-quality synthetic insurance portfolios characterised by discrete and continuous variables. Section 2.6 concludes. To the best of our knowledge, the use of quantile transformations and Variational AutoEncoders for such unsupervised learning tasks (i.e., dimensionality reduction and synthetic data generation) in actuarial science has not been previously explored.

2.2 Related works

Generative modelling from learned latent representations is a central application of Variational AutoEncoders (VAEs) (Kingma et al. 2013). This technique plays a crucial role in many sectors, including insurance, where accurate risk assessment, modelling, and decision-making require access to high-quality, diverse, and representative data. Over the years, researchers have developed various data augmentation techniques to expand datasets while preserving the underlying structure of the original data.

In natural language processing, J. Wei et al. (2019) proposed text augmentation techniques such as replacing or inserting random words in a sentence to create new contexts and improve text classification models. Similarly, Xie et al. (2020) used unsupervised data augmentation techniques for text classification, including back translation and paraphrase generation, to improve model robustness through consistency training. In computer vision, Krizhevsky et al. (2017) employed simple data augmentation techniques, such as horizontal flipping, random cropping, and zooming, to create mirror images and simulate positional variations, and improve the generalisation performance of deep Convolutional Neural Networks (CNNs). Similarly, Simard et al. (2003) discuss elastic distortions, such as rotation (to simulate different viewing angles), scaling, and translation to expand image datasets applied to CNNs. While these traditional methods can increase dataset size, they often fail to capture complex structure and dependencies within the data. Advanced generative approaches, such as Generative Adversarial Networks (GANs) (I. J. Goodfellow et al. 2014) and VAEs (Kingma et al. 2013), have shown promise in overcoming these limitations. For comparative reviews of deep generative models, including VAEs, GANs, normalising flows, and energy-based, and autoregressive models, we refer to Bond-Taylor et al. (2022) and Chakraborty et al. (2024), which also cover GAN variants, latent architectures, validation metrics, and application domains.

In recent years, numerous VAE variants tailored to specific applications have been introduced. Real-world applications of VAEs span healthcare, finance, natural language processing, and computer vision. In medicine, VAEs have been used for disease diagnosis (García-Ordás et al. 2020; Mansour et al. 2021), detection of abnormal lesions or pathology in brain MR images (Steinmann et al. 2021; Baur et al. 2019), patient monitoring (Miotto et al. 2016), drug discovery (Lim et al. 2018; Song et al. 2022), and discrete molecule data generation (Kusner, Paige, et al. 2017). In finance, they have been used for fraud detection (Alazizi

et al. 2020; Tingfei et al. 2020), risk assessment (Mancisidor et al. 2021), and portfolio optimisation (Duan et al. 2022; C. Zhang et al. 2020). In natural language processing, VAEs are used for text generation (Semeniuta et al. 2017; W. Wang et al. 2019; X. Zhang et al. 2019), text classification (W. Xu et al. 2017), sentiment analysis (Wu et al. 2019; Lu et al. 2020), recommender systems (X. Li et al. 2017), and machine translation (Pagnoni et al. 2018). In computer vision, applications include image synthesis (H. Huang et al. 2018), object recognition (B. Li et al. 2019), and image captioning (Shen et al. 2020).

A direction of research of particular interest for actuarial applications is the integration of heterogeneous data into the VAE architecture. This typically involves using separate encoding pathways for each data type before merging them in the latent space, often with constraints to enforce domain-specific properties. By integrating such domain knowledge, VAEs can learn more meaningful representations and capture relevant dependencies between diverse data types. Despite significant progress, challenges remain when applying VAEs to real-world insurance datasets due to their disparate nature (i.e., continuous, (positive) numerical, binary, multinomial, or ordinal) with a variety of marginal distributions (e.g., multi-modal or long-tailed).

The presence of mixed-type values, and more specifically categorical rating factors, has already been addressed in the actuarial literature by Noll et al. (2018) and Ferrario et al. (2018) through entity embeddings (Guo et al. 2016) or joint embeddings (Delong et al. 2023). However, these approaches require pre-training layers to fine-tune numerical representations for categorical features, which increases model calibration time as the number of categorical variables grows. This does not align with our intention to train an all-in-one unsupervised autoencoder.

Several works (Nazábal et al. 2020; Ögretir et al. 2022; Antelmi et al. 2019) have instead integrated heterogeneous and incomplete data directly into VAE designs using various likelihood models. Foundational studies (L. Xu and Veeramachaneni 2018; L. Xu, Skoularidou, et al. 2019) explored the generation of (medical and educational records) mixed-type tabular data. They address imbalanced categorical variables with a conditional generator and training-by-sampling. They further tackle the challenges associated with multi-modal non-Gaussian continuous variables via mode-specific normalisation using Gaussian mixture models (Bishop et al. 2006). While effective, this normalisation scheme increases input dimensionality by the number of Gaussians used to approximate each variable’s distribution.

Nazábal et al. (2020) and Ma et al. (2020) proposed hierarchical two-stage VAES for supervised tasks, modelling heterogeneous likelihoods at the bottom level with independent Deep Neural Networks (DNNs) and capturing statistical dependencies at the top level with a shared DNN. Similarly, Suh et al. (2016) use Gaussian copulas to model local dependencies. Ögretir et al. (2022) took the idea a step further by designing a VAE model for time-series and longitudinal datasets, modelling temporal dynamics with auxiliary covariates. Antelmi et al. (2019) tackled the interpretability issues of the latent space by introducing constraints on

latent representations to disentangle single-channel contributions while accounting for latent relationships across heterogeneous features.

Despite these advances, dealing with multi-modal marginal distributions and the associated mode-seeking behaviour of the KL divergence remains largely unaddressed, with the exception of Bishop et al. (2006) who proposed the mode-specific normalisation mentioned above. In this chapter, we propose a quantile transformation of the continuous variables as well as a novel synthesising approach (that also makes use of the quantile transformation) to tackle the collapse of the approximate posterior distribution towards a single mode and the biased generation of samples towards this specific mode. By applying the quantile transformation at both input and latent levels, we not only succeeded to capture the multiple modes of the original continuous variables distributions in the reconstructed and synthetic datasets, but we also maintain the deviance performance of a regression model (GLM) trained on the synthetic data for predicting the claim frequencies of the original data.

2.3 Variational AutoEncoder

2.3.1 Introduction

An autoencoder (Kramer 1991) is a pair of two connected feed-forward Neural Networks (NNs), the encoder and the decoder. The encoder network takes in an input (e.g., insurance policies) and maps it into a compressed representation, the latent encoding. This encoding contains enough relevant information for the decoder to convert it back to the original input. Formally, the encoder network learns a non-linear transformation,

$$e_{\phi} : \mathcal{X} \rightarrow \mathcal{Z},$$

parameterised by ϕ that maps each input vector $\mathbf{x}$ of dimension m from the original high-dimensional input space $\mathcal{X}$ to a lower-dimensional latent space $\mathcal{Z}$. The latent encoding,

$$\mathbf{z} = e_{\phi}(\mathbf{x}),$$

is a low-dimensional representation of the original input. Conversely, the decoder network learns a non-linear transformation,

$$d_{\theta} : \mathcal{Z} \rightarrow \mathcal{X},$$

parameterised by θ that projects the latent codes $\mathbf{z}$ of dimension $d < m$ back into the original high-dimensional input space $\mathcal{X}$. This transformation takes the latent vector $\mathbf{z}$ as input to reconstruct the original input data,

$$\hat{\mathbf{x}} = d_{\theta}(\mathbf{z}).$$

The autoencoder,

$$f(\mathbf{x}) = d_{\theta}(e_{\phi}(\mathbf{x})) ,$$

illustrated in Fig. 2.3.1, is trained as a whole to minimise the difference between each input $\mathbf{x}$ and its reconstruction $\hat{\mathbf{x}}$.

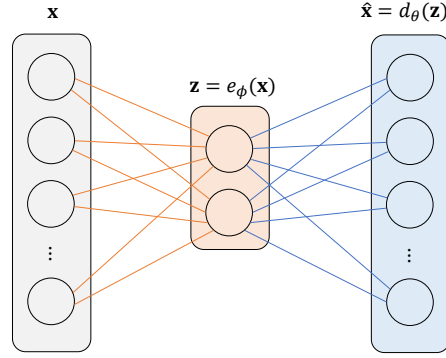


Figure 2.3.1: Architecture of a shallow autoencoder.

The traditional encoder network projects each input vector $\mathbf{x} \in \mathcal{X}$ (e.g., each insurance policy) to a single latent value z_j for each encoding dimension $j = 1, \dots, d$. Therefore, each sample $\mathbf{x}$ is encoded independently of other samples, even if they are from the same class or distribution (e.g., have similar vehicle characteristics and risk profiles). This makes the latent space unlikely to be continuous, where policies with similar risk profiles may be mapped to very different or distant latent encodings. It becomes problematic when the decoder network is used as a generative model. If the latent space has discontinuities or gaps and a variation is sampled from there (i.e., from outside the manifold of the latent codes), the decoder will not be able to generate meaningful or realistic outputs because it was never trained with encoded vectors coming from that region of the latent space so it has no idea how to deal with it. The regularity that is expected from the latent space of a generative model can be expressed through two properties:

1. Continuity: The transformation $e_{\phi} : \mathcal{X} \rightarrow \mathcal{Z}$ should ensure that (dis)similar samples $\mathbf{x}$ have (dis)similar latent vectors $\mathbf{z}$. Similarly, the transformation $d_{\theta} : \mathcal{Z} \rightarrow \mathcal{X}$ should ensure that two close points in the latent space have similar contents once decoded. In actuarial terms, this ensures that policies that depict the same vehicle characteristics or risk profile have very close representations in the latent space and similar contents once decoded.
2. Completeness: A point sampled from the latent space should give meaningful and realistic content once decoded.

Kingma et al. (2013) and Rezende, Mohamed, and Wierstra (2014) introduced the Variational AutoEncoder (VAE) as a Bayesian extension to the autoencoder (Kramer 1991) to address this shortcoming. The NNs of their Bayesian model, illustrated in Fig. 2.3.2, learn a probabilistic (instead of deterministic) representation of the data points in the latent space. This approach should improve the decoder's robustness to small changes in $\mathbf{z}$.

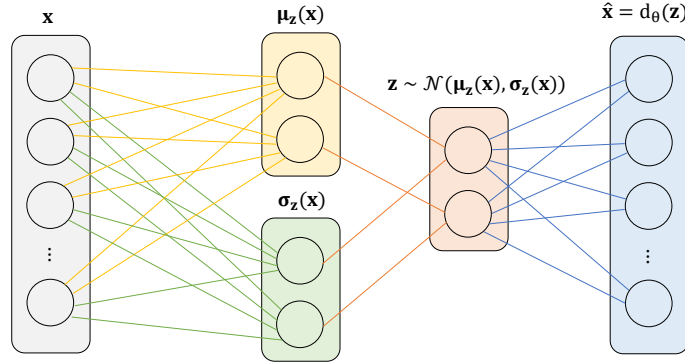


Figure 2.3.2: Architecture of a shallow Variational AutoEncoder with a Gaussian prior.

Nevertheless, such a probabilistic approach comes with its fair share of setbacks, and their solutions. Modelling the true probability distribution $p(\mathbf{x})$ over the limited set of insurance policies $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ is a tedious and challenging task (which involves, e.g., modelling complex correlations between the rating factors). Instead of directly learning $p_{\theta}(\mathbf{x})$ (where θ is the true, but unknown, set of parameters), we can introduce (unobserved) vectors $\mathbf{z}$ and define a conditional distribution, denoted $p_{\theta}(\mathbf{x}|\mathbf{z})$, from which we assume the data points $\mathbf{x}_i$ are generated.

We can further introduce a prior distribution over $\mathbf{z}$, denoted as $p_{\theta}(\mathbf{z})$. The latent codes $\mathbf{z}$ will be sampled from that distribution, which is assumed to follow an isotropic standard multivariate normal distribution:

$$p_{\theta}(\mathbf{z}) = \mathcal{N}(\mathbf{z} | 0, \mathbf{I}), \quad (2.3.1)$$

where $\mathbf{I}$ is the identity matrix of dimension $d \times d$.

We can only observe $\mathbf{x}$, but we would like to infer the characteristics of $\mathbf{z}$. In other words, we want to compute the (true) posterior distribution $p_{\theta}(\mathbf{z}|\mathbf{x})$. Using Bayes theorem,

$$p_{\theta}(\mathbf{x}|\mathbf{z}) p_{\theta}(\mathbf{z}) = p_{\theta}(\mathbf{x}, \mathbf{z}) = p_{\theta}(\mathbf{z}|\mathbf{x}) p_{\theta}(\mathbf{x}), \quad (2.3.2)$$

we have

$$p_{\theta}(\mathbf{z}|\mathbf{x}) = \frac{p_{\theta}(\mathbf{x}|\mathbf{z}) p_{\theta}(\mathbf{z})}{p_{\theta}(\mathbf{x})}, \quad (2.3.3)$$

where $p_{\theta}(\mathbf{x}, \mathbf{z})$ denotes the joint distribution. To obtain the true data distribution $p_{\theta}(\mathbf{x})$, we need to marginalise over the latent codes $\mathbf{z}$. The marginal likelihood of $\mathbf{x}$ for a candidate set of parameters θ is defined (Kingma et al. 2013) as the integral over the generative model $p_{\theta}(\mathbf{x}|\mathbf{z}) p_{\theta}(\mathbf{z})$:

$$p_{\theta}(\mathbf{x}) = \int p_{\theta}(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \int p_{\theta}(\mathbf{x}|\mathbf{z}) p_{\theta}(\mathbf{z}) d\mathbf{z}. \quad (2.3.4)$$

In practice, the use of non-linear deep neural nets makes the integral of the marginal likelihood intractable (i.e., no analytical solution) and the true posterior distribution $p_{\theta}(\mathbf{z}|\mathbf{x})$ unknown. While Generative Adversarial Networks avoid the integral in Eq. (2.3.4) altogether, Variational AutoEncoders use an approximation of the true, but unknown, posterior with a recognition model and a maximisation of the log-likelihood lower bound (found by variational inference).

2.3.2 Variational approximation of the posterior

The exact posterior $p_{\theta}(\mathbf{z}|\mathbf{x})$ is approximated with a tractable distribution known as the variational or approximate posterior $q_{\phi}(\mathbf{z}|\mathbf{x})$. In machine learning, a probabilistic encoder network e_{ϕ} (also called recognition model) is used to learn the set of variational parameters ϕ such that the tractable $q_{\phi}(\mathbf{z}|\mathbf{x})$ is used to perform approximate inference of the intractable posterior distribution $p_{\theta}(\mathbf{z}|\mathbf{x})$. The parameters ϕ of the recognition model e_{ϕ} are learned jointly with the parameters θ of the generative model d_{θ} . Since the prior $p_{\theta}(\mathbf{z})$ is assumed to follow an isotropic Gaussian distribution, the encoder network will output two parameters, the mean $\boldsymbol{\mu}_z$ and covariance matrix $\boldsymbol{\Sigma}_z$. Formally, the approximate posterior is assumed to follow an isotropic multivariate Gaussian distribution of dimension d :

$$q_{\phi}(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z} | \boldsymbol{\mu}_z(\mathbf{x}), \boldsymbol{\Sigma}_z(\mathbf{x})), \quad (2.3.5)$$

where $\boldsymbol{\mu}_z(\mathbf{x}) \in \mathbb{R}^d$ and $\boldsymbol{\Sigma}_z(\mathbf{x}) = \sigma_z^2 I \in \mathbb{R}^{d \times d}$ are such that:

$$\boldsymbol{\mu}_z(\mathbf{x}) = \begin{pmatrix} \mu_{z,1}(\mathbf{x}) \\ \vdots \\ \mu_{z,d}(\mathbf{x}) \end{pmatrix}, \quad \boldsymbol{\Sigma}_z(\mathbf{x}) = \begin{pmatrix} \sigma_{z,1}^2(\mathbf{x}) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_{z,d}^2(\mathbf{x}) \end{pmatrix}.$$

The variational parameters $\boldsymbol{\mu}_z(\mathbf{x})$ and $\boldsymbol{\Sigma}_z(\mathbf{x})$ are the outputs of a single layer neural network:

$$\begin{aligned} \mathbf{h}_z &= U(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1), \\ \boldsymbol{\mu}_z(\mathbf{x}) &= \mathbf{W}_2 \mathbf{h}_z + \mathbf{b}_2, \\ \ln \boldsymbol{\Sigma}_z(\mathbf{x}) &= \text{diag}(\mathbf{W}_3 \mathbf{h}_z + \mathbf{b}_3), \end{aligned}$$

where U is an element-wise activation function (e.g., sigmoid, rectified linear unit, scaled exponential linear unit, etc.), $\mathbf{W}$ is a weight matrix, and $\mathbf{b}$ is a bias vector. To ensure the positivity of σ_z^2 , the encoder network is asked to output its logarithm (the network may otherwise learn (by back-propagation) negative values for σ_z^2).

2.3.3 Variational lower bound

Recall that we would like to approximate some posterior distribution $p_\theta(\mathbf{z}|\mathbf{x})$ with a tractable distribution $q_\phi(\mathbf{z}|\mathbf{x})$ and make this approximate inference model $q_\phi(\mathbf{z}|\mathbf{x})$ as close as possible to the true, but intractable, posterior $p_\theta(\mathbf{z}|\mathbf{x})$. Mathematically, we can find such an approximation by deriving a variational lower bound that will make the likelihood function, in Eq. (2.3.4), tractable. Such a scheme involves the Kullback-Leibler (KL) divergence. This divergence is a measure of dissimilarity between two probability distributions. If $p(y)$ and $q(y)$ are two densities, their reverse KL divergence from the approximate q to the true p , written as $D_{\text{KL}}(q||p)$, is defined as:

$$D_{\text{KL}}(q||p) = \mathbb{E}_{Y \sim q} \left[\ln \frac{q(Y)}{p(Y)} \right] = \int q(y) \ln \left(\frac{q(y)}{p(y)} \right) dy.$$

Proposition 2.1. *The marginal log-likelihood is equal to the following sum:*

$$\ln(p_\theta(\mathbf{x})) = D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) || p_\theta(\mathbf{z}|\mathbf{x})) + \mathcal{L}(\mathbf{x}; \theta, \phi), \quad (2.3.6)$$

where $\mathcal{L}$ is called the variational lower bound:

$$\mathcal{L}(\mathbf{x}; \theta, \phi) = \mathbb{E}_{q_\phi} [-\ln q_\phi(\mathbf{z}|\mathbf{x}) + \ln p_\theta(\mathbf{x}, \mathbf{z})]. \quad (2.3.7)$$

The proof (Kingma et al. 2013) is given in Appendix 2.A.

The first term in Eq. (2.3.6) is the KL divergence of the approximate from the true posterior. It is intractable due to $p_\theta(\mathbf{z}|\mathbf{x})$. However, since the KL divergence is a measure of the difference between two probability densities, it is always positive. Therefore, the marginal log-likelihood (or evidence) is lower bounded by the variational lower bound $\mathcal{L}$:

$$\ln p_\theta(\mathbf{x}) \geq \mathcal{L}(\mathbf{x}; \theta, \phi) = \mathbb{E}_{q_\phi} [-\ln q_\phi(\mathbf{z}|\mathbf{x}) + \ln p_\theta(\mathbf{x}, \mathbf{z})].$$

Therefore, minimising the KL divergence between the exact posterior $p_\theta(\mathbf{z}|\mathbf{x})$ and the variational $q_\phi(\mathbf{z}|\mathbf{x})$ is equivalent to maximising this variational lower bound $\mathcal{L}$ during the training process. As a result, the complex inference problem is reduced to a simpler optimisation problem:

$$\hat{\theta}, \hat{\phi} = \arg \max_{\theta, \phi} \sum_{\mathbf{x} \in \mathcal{X}} \mathcal{L}(\mathbf{x}; \theta, \phi).$$

This is often referred to as the Evidence Lower Bound Optimisation, denoted as ELBO $(\mathbf{x}; \theta, \phi)$. The closer the evidence's lower bound is to the marginal likelihood,

the closer the variational approximation will be to the true posterior distribution. For optimisation, it is convenient to rewrite this variational (tractable) lower bound $\mathcal{L}(\mathbf{x}; \theta, \phi)$ in the following way:

$$\begin{aligned}\mathcal{L}(\mathbf{x}; \theta, \phi) &= \mathbb{E}_{q_\phi} [-\ln q_\phi(\mathbf{z}|\mathbf{x}) + \ln p_\theta(\mathbf{z}) + \ln p_\theta(\mathbf{x}, \mathbf{z}) - \ln p_\theta(\mathbf{z})] \\ &= -\mathbb{E}_{q_\phi} \left[\ln \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z})} \right] + \mathbb{E}_{q_\phi} \left[\ln \frac{p_\theta(\mathbf{x}, \mathbf{z})}{p_\theta(\mathbf{z})} \right] \\ &= -D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) || p_\theta(\mathbf{z})) + \mathbb{E}_{q_\phi} [\ln p_\theta(\mathbf{x}|\mathbf{z})] .\end{aligned}\quad (2.3.8)$$

The first term in Eq. (2.3.8) is the KL divergence of the approximate posterior (i.e., the distribution of the encoder outputs) from the prior distribution $p_\theta(\mathbf{z}) = \mathcal{N}(\mathbf{z}|0, \mathbf{I})$. Both of these functions are assumed to follow an isotropic Gaussian distribution and admit a closed-form solution.

Proposition 2.2. *The KL divergence involved in the definition of the variational lower bound is equal to:*

$$D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) || p_\theta(\mathbf{z})) = \frac{1}{2} \sum_{k=1}^d \left(\sigma_{z,k}(\mathbf{x})^2 + \mu_{z,k}(\mathbf{x})^2 - (1 + \ln \sigma_{z,k}^2(\mathbf{x})) \right) ,$$

where d is the dimension of the latent space. The proof is given in Appendix 2.A.

The KL divergence term in Eq. (2.3.8) may be interpreted as a regulariser or encoding error. It penalises the learned distribution $q_\phi(\mathbf{z}|\mathbf{x})$ when it diverges too much from the true prior distribution $p_\theta(\mathbf{z})$. This penalty term enables the use of the decoder network as a generative model. Because the probabilistic autoencoder model maps each data point onto a distribution instead of a single point in the latent space, the variational encoding covers an area whose centre and size are controlled by the means and standard deviations.

However, if the outputted means are significantly different and the standard deviations are too small, there may still be some gaps in the latent space, leaving the decodable latent space discontinuous, which is precisely what we want to avoid. By introducing the KL divergence as a regulariser, we force the encoder to find latent vectors that approximately follow a standard Gaussian distribution, encouraging the latent vectors to occupy a more centralised and uniform location. This should help to meet the regularity requirements of the latent space.

While the KL divergence should ensure that the latent distribution is similar to the prior distribution, we do not want to end up describing every observation using the same unit Gaussian as this would treat every observation as having the same characteristics and fail to describe the original data. The second term in Eq. (2.3.8) prevents that by encouraging the decoder's output $\hat{\mathbf{x}}$ to reconstruct the original input $\mathbf{x}$. Formally, it represents the cross-entropy of $p_\theta(\mathbf{x}|\mathbf{z})$ w.r.t. q_ϕ :

$$\mathbb{E}_{q_\phi} [\ln p_\theta(\mathbf{x}|\mathbf{z})] = \int \ln p_\theta(\mathbf{x}|\mathbf{z}) q_\phi(\mathbf{z}|\mathbf{x}) d\mathbf{z} .$$

The cross-entropy of a random variable is the level of uncertainty inherent in the variable possible outcome and may be interpreted as a reconstruction loss. Statistically speaking, the decoder should parameterise a likelihood distribution that places enough probability mass on the true data. While the KL divergence in Eq. (2.3.8) admits an analytical expression (see Proposition 2.2), the reconstruction loss in Eq. (2.3.8) is usually approximated by L simulations:

$$\mathbb{E}_{q_\phi} [\ln p_\theta(\mathbf{x}|\mathbf{z})] \approx \frac{1}{L} \sum_{l=1}^L \ln p_\theta(\mathbf{x}|\mathbf{z}^{(l)}), \quad (2.3.9)$$

where $\mathbf{z}^{(l)} \sim q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_z(\mathbf{x}), \boldsymbol{\Sigma}_z(\mathbf{x}))$. This random sampling of the latent codes $\mathbf{z}^{(l)}$, illustrated in Fig. 2.3.3, requires some extra attention.

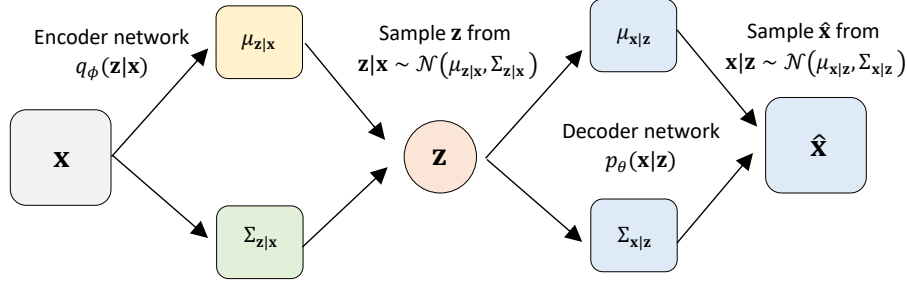


Figure 2.3.3: Illustration of the sampling process for a Gaussian encoder and decoder. The random node is represented by a circle.

2.3.4 Reparameterization trick

During the training process, each network parameter is evaluated with respect to the final output loss using a technique known as back-propagation. The random sampling of the latent code $\mathbf{z}^{(l)} \sim q_\phi(\mathbf{z}|\mathbf{x})$ is making the lower-dimensional space stochastic. However, NNs cannot back-propagate through stochastic nodes. Luckily, there exists a reparameterization trick, illustrated in Fig. 2.3.4, which suggests that instead of randomly sampling $\mathbf{z}$ from a Gaussian probability density, the latent codes can be sampled from an auxiliary random noise $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{I})$ shifted by the latent means vectors $\boldsymbol{\mu}_z(\mathbf{x})$ and scaled by the covariance matrix $\boldsymbol{\Sigma}_z(\mathbf{x})$:

$$\mathbf{z} = \boldsymbol{\mu}_z(\mathbf{x}) + \sqrt{\boldsymbol{\Sigma}_z(\mathbf{x})} \odot \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{I}).$$

While the error term $\boldsymbol{\varepsilon}$ introduced is stochastic, the node $\mathbf{z}$ is now deterministic, enabling the back-propagation through it. Mathematically, we can now optimise $\mathcal{L}$ with respect to the parameters $\phi = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ of the variational posterior using the

back-propagation technique. θ and ϕ are obtained by maximising,

$$\sum_{\mathbf{x} \in \mathcal{X}} \tilde{\mathcal{L}}(\mathbf{x}; \theta, \phi) = \sum_{\mathbf{x} \in \mathcal{X}} \left[-D_{\text{KL}}(q_{\phi}(\mathbf{z}|\mathbf{x}) \parallel p_{\theta}(\mathbf{z})) + \frac{1}{L} \sum_{l=1}^L \ln p_{\theta}(\mathbf{x}|\mathbf{z}^{(l)}) \right], \quad (2.3.10)$$

where $\mathbf{z}^{(l)} = \boldsymbol{\mu}_{\mathbf{z}}(\mathbf{x}) + \sqrt{\boldsymbol{\Sigma}_{\mathbf{z}}(\mathbf{x})} \odot \boldsymbol{\varepsilon}^{(l)}$ and $\boldsymbol{\varepsilon}^{(l)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

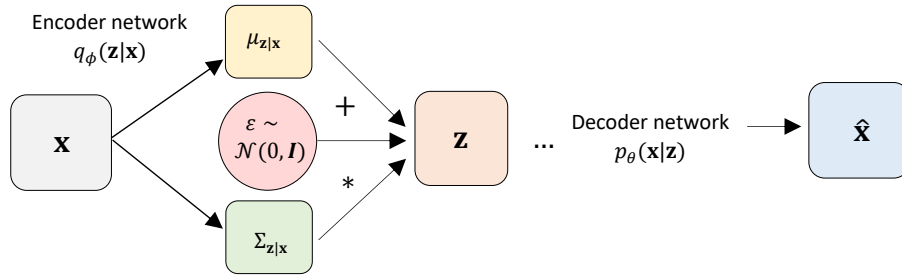


Figure 2.3.4: Illustration of the reparameterization trick. The random node is represented by a circle.

This random generation of $\mathbf{z}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ implies the encoding varies, for the same input $\mathbf{x}$, in every single pass through the sampling layer, even if the means and standard deviations remain the same. This allows the decoder to be exposed to a range of encoding variations for the same input during training. Consequently, the decoder learns that all nearby points in latent space should refer to a sample of the same class. By doing so, we are essentially enforcing a continuously meaningful latent space representation. As a result, every point close to the encoded location of one of the input policies will be decoded to something similar to that original policy, thereby making the VAE highly suitable for the generation of new samples. The number of simulations L in the gradient estimator in Eq. (2.3.9) and Eq. (2.3.10) can also be set to 1 as the decoder is already exposed to variations (and the mini-batch size will be set large enough).

2.4 Heterogeneous data

The VAE model we have just defined was first designed by Kingma et al. (2013) with continuous variables in mind, for applications like image generation where the input and output variables are continuous pixel values. This creates challenges

for the model application to real-world datasets (e.g., insurance portfolios), which contain various types of features (i.e., (positive) numerical, binary, multinomial, and ordinal) with a variety of marginal distributions (e.g., multi-modal, long-tail). In the following subsections, we will address and discuss the challenges related to mapping categorical variables to numeric vectors, the multi-modal distributions of continuous variables, and the need for the reconstruction loss to reflect the heterogeneity of the input data. In summary, we propose to handle such heterogeneous datasets as follow:

1. For categorical variables (binary, ordinal, or multinomial):
 - (a) Dummy encoding before inputting them into the encoder network.
 - (b) Each set of dummy variables corresponding to the same categorical variable is passed through a softmax activation function in the output layer of the decoder network.
2. For continuous variables:
 - (a) Quantile (non-parametric) transformation to ensure a uni-modal normal distribution before inputting them into the encoder network.
 - (b) Min-max scaling to the range $(-1, 1)$ to match the range of the `tanh` activation function used in the output layer of the decoder network.
3. Reconstruction loss: We introduce a weighted combination of mean squared error and categorical cross-entropy (minimised during training) to account for the different distributions of the pre-processed continuous and categorical variables.

2.4.1 Dummy encoding and softmax activation

Deep learning models require all inputs to be numeric. However, insurance policies are known to also contain information in categorical forms (e.g., geographic location, vehicle power, brand, or colour). Whether nominal (e.g., colour) or ordinal (e.g., vehicle power), categorical variables must be encoded to numbers to train any deep learning model. Methods that map a set of alternative values are very few. Dummy encoding is one of the most common and simplest approaches to represent categorical variables numerically. It maps each modality to a binary vector, as illustrated in the following example.

Example 2.1. For each distinct modality in the categorical column `Zone`, a new dummy column is created, such that:

Zone	→	Zone ₁	Zone ₂	Zone ₃	Zone ₄	Zone ₅	Zone ₆	Zone ₇
3		0	0	1	0	0	0	0
7		0	0	0	0	0	0	1
1		1	0	0	0	0	0	0
4		0	0	0	1	0	0	0
⋮		⋮	⋮	⋮	⋮	⋮	⋮	⋮
6		0	0	0	0	0	1	0

Table 2.4.1: Dummy encoding of the nominal variable Zone with seven modalities.

One of the main issues with dummy encoding is that the dimension of the resulting sparse binary vectors scales linearly with the number of categories present in the data. Additionally, introducing new categories requires retraining the model completely as the new binary vectors and their corresponding input nodes should also be introduced. Moreover, since binary vectors lack semantic significance at the encoder’s input layer, the model may struggle to capture the data underlying structure and understand relationships or similarities between different categories based on their dummy representations.

Entity embedding (Guo et al. 2016) has been gaining momentum in the data field for addressing some of the weaknesses of dummy encoding. By learning low-dimensional dense vector representations of categories, embeddings aim to capture semantic relationships between categories and further offer the benefits of a uniform mapping and a smaller vector space than dummy encoding. Yet, to produce the best numerical representation of a categorical variable, entity embeddings are computationally intensive, and fine-tuning the NNs hyper-parameters can quickly become time-consuming compared to the ease of implementation of dummy encoding. Furthermore, even entity embedding takes the detour of an ordinal or dummy encoder before passing the so-encoded input into a set of layers. As pointed out by Richman (2020), the first layer of an autoencoder learns a representation of all of the input attributes that is comparable to an embedding layer. For more details, we refer to DeLong et al. (2023), who recently proposed a joint embedding to learn the numerical representation of categorical rating factors and fed it, together with the original numerical features, as inputs of a neural network for a supervised learning task on insurance data.

Vector quantisation via codebooks (Wu et al. 2019) is conceptually related to entity embedding, as both approaches involve learning low-dimensional representations of categorical variables. However, they differ in their underlying objectives to address the weaknesses of dummy encoding. While entity embedding overcomes the inability of dummy encoding to capture semantic relationships between categories based on their dummy representations, vector quantisation via codebooks focuses on the scalability issue of dummy encoding. By using a fixed set of representative vectors, known as a codebook, the model generalises well to unseen

categories by mapping them to the nearest vectors in the codebook and, therefore, does not require retraining the model when new categories are introduced, unlike dummy encoding.

Although alternative encoding methods like entity embedding or codebook quantisation may enhance learning with VAEs, they add complexity to the model architecture and training procedure. Moreover, these methods do not address at once all the weaknesses associated with dummy encoding. When considering the trade-offs between simplicity, interpretability, and performance, dummy encoding is still a reasonable choice for representing categorical variables in machine learning models, especially when the number of categories is relatively small. Furthermore, the interpretability of the binary vectors can be valuable for understanding and analysing the latent space of the VAE.

It is worth mentioning that the choice of the encoding method (dummy encoding, entity embedding, or vector quantisation) is part of a pre-processing step applied to the data before inputting it into the encoder network of the VAE. Some authors (Oord et al. 2017; Jang et al. 2016) took the idea of handling categorical variables one step further by introducing the Vector Quantised Variational AutoEncoder (VQ-VAE) and Gumbel-softmax trick approach, both of which involve major changes in the latent space assumptions. Given the exclusively categorical nature of their input data, they propose discrete latent representations with categorical posterior and prior distributions instead of the traditional continuous learning representations (with a Gaussian prior). Although the use of discrete latent representations has been shown (Oord et al. 2017; Jang et al. 2016) to be a more natural fit for many applications inherently discrete (e.g., language, speech, images), there is no theoretical evidence that switching to a categorical latent representation with categorical prior and posterior distributions could supersede the traditional continuous latent representations when dealing with heterogeneous datasets containing both categorical and continuous data.

Our intention to develop a straightforward and all-in-one unsupervised autoencoder led us to select dummy encoding as a baseline approach to map our categorical variables to numeric vectors and to focus on the traditional continuous latent representation (with a Gaussian prior) to study the impact of introducing a quantile transformation (see following subsection 2.4.2) on the reconstructed and synthetic data.

Given the choice of dummy encoding to transform the categorical variables, we opted for softmax activation functions for the decoder. Because our VAE architecture jointly models the categorical and continuous variables in the input data, the decoder's output layer will be sliced to extract the values corresponding to all modalities of the same categorical variable. Each slice will be individually passed through a softmax activation function, as illustrated in the following Fig. 2.4.1.

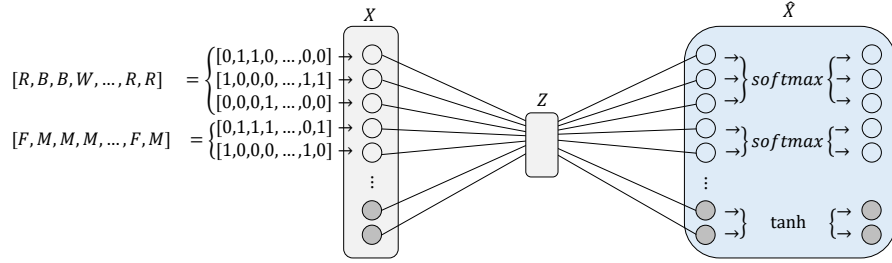


Figure 2.4.1: Example of the slicing technique applied on the decoder output of the Variational AutoEncoder with two categorical variables **Colour** and **Gender** whose categories Red (R), Black (B), White (W), and Female (F), Male (M), are dummy encoded. The darker nodes represent continuous variables (see subsection 2.4.2).

2.4.2 Quantile transformation, min-max scaling, and tanh activation

In addition to the categorical variables pre-processing, scaling continuous features has been shown to benefit NNs optimisation by increasing the speed of convergence of the gradient descent and preventing the optimisation from getting stuck in a local optimum (see, e.g., LeCun et al. (2012) and I. Goodfellow et al. (2016)). Standardisation and normalisation are the two most common approaches to scale continuous variables. However, in real-world datasets, continuous variables come in many forms (i.e., skewed, long-tailed, multi-modal). The challenges of training NNs on multi-modal distributions have already been highlighted in the literature. Among others, Pascanu et al. (2013) discussed how the vanishing gradient problem is exacerbated by multi-modal distributions, leading to gradient saturation and slow convergence. Non-normal, skewed, and long-tailed distributions can further affect the distributional assumptions of our model and the balance of our training samples, which can eventually impact the model's performance and generalisation. In summary, if the input distributions are multi-modal and/or non-Gaussian, the learned latent space might fail to effectively capture all the modes or the complexities of the data, resulting in incomplete, blurred, and/or biased reconstructions of the original input.

Moreover, if at least one of the latent encoding dimensions were to capture and reproduce all the modes present in the data distribution, another issue related to the use of the KL divergence would arise. If the true posterior distribution of the latent variables given the input data, denoted as $p(\mathbf{z}|\mathbf{x})$, is also multi-modal (i.e., it has multiple distinct peaks or modes), minimising the reverse KL divergence (during the training process) can lead to the collapse of the approximate posterior distribution towards a single mode. This will result in generated samples being biased towards a specific mode, neglecting the other modes of the true data

distribution. This mode-seeking behaviour of the KL divergence can further impact the generation of new data points. Researchers have proposed various techniques to address this issue and encourage the model to capture and generate samples from multiple modes. Some of these techniques include using alternative divergence measures like the Jensen-Shannon divergence (Deasy et al. 2020), incorporating adversarial training (Mescheder et al. 2017), employing more flexible generative models like Normalising Flows (Rezende and Mohamed 2015), or introducing a mode-specific normalisation (L. Xu and Veeramachaneni 2018). The representation of a continuous column after applying this mode-specific normalisation becomes the concatenation of a normalised vector $\alpha_{i,j} = (x_{i,j} - \eta_k)/4\phi_k$ and m_{01} binary vectors indicating which of the k Gaussian $\mathcal{N}(\eta_k, \phi_k)$ each value $x_{i,j}$ belongs to. Such a normalisation further increases the (already significant) input dimension by the number of Gaussians used to approximate the multi-modal distributions.

To address these challenges, we propose a quantile transformation to compel continuous variables to follow an uni-modal normal distribution. Compared to the aforementioned techniques, our non-parametric quantile transformer does not require significant changes in the VAE architecture, maintains the numeric input dimension, tends to spread out the most frequent values, smooth out unusual distributions, and is less sensitive to outliers than more traditional scaling methods.

Definition 2.1 (Quantile transformation to a standard Normal distribution). Let X be a continuous random variable with cumulative distribution function (CDF):

$$F_X(x) = \mathbb{P}(X \leq x).$$

The quantile transformation of X to a standard Normal random variable is defined by:

$$X' = T(X) = \Phi^{-1}(F_X(X)),$$

where Φ^{-1} denotes the quantile function (inverse cumulative distribution function) of the standard Normal distribution.

The transformation $T(X) = \Phi^{-1}(F_X(X))$ is based on the probability integral transform. Specifically, if X is a continuous random variable with cumulative distribution function F_X , then the transformed variable $U = F_X(X)$ follows a uniform distribution on the unit interval $U \sim \mathcal{U}(0, 1)$. Indeed, for every $u \in [0, 1]$,

$$\mathbb{P}(U \leq u) = \mathbb{P}(F_X(X) \leq u) = \mathbb{P}(X \leq F_X^{-1}(u)) = F_X(F_X^{-1}(u)) = u,$$

which is precisely the cumulative distribution function of a uniform random variable on $[0, 1]$.

The second step of the transformation maps the uniform variable U into Gaussian space through the inverse standard Normal cumulative distribution function:

$$X' = \Phi^{-1}(U).$$

Since Φ is the cumulative distribution function of the standard Normal distribution,

$$\mathbb{P}(X' \leq x') = \mathbb{P}(\Phi^{-1}(U) \leq x') = \mathbb{P}(U \leq \Phi(x')) = \Phi(x'),$$

for all $x' \in \mathbb{R}$. Consequently,

$$X' \sim \mathcal{N}(0, 1).$$

Therefore, the overall transformation,

$$X \mapsto F_X(X) \mapsto \Phi^{-1}(F_X(X)),$$

maps any continuous random variable X , regardless of skewness, heavy tails, or multimodality, into a random variable with standard Normal distribution while preserving the ordering of the observations.

In practice, this non-linear transformation is applied independently to each feature $X_j \in \mathcal{X}$. Since the true distribution F_{X_j} is generally unknown, it is estimated empirically from the training data using sample quantiles and empirical percentiles. More precisely, the observed feature values are first ranked in ascending order in order to construct an empirical cumulative distribution function $\hat{F}_{X_j}$. Each observation is then mapped to its corresponding cumulative probability, $U = \hat{F}_{X_j}(X_j)$, thereby transforming the original feature into an approximately uniform random variable on the unit interval $U \sim \mathcal{U}(0, 1)$. In a second step, the uniformly distributed variable U is mapped to a target distribution through the corresponding quantile function. In the case of a Gaussian target distribution, the transformation is given by

$$X'_j = \Phi^{-1}(U) = \Phi^{-1}(\hat{F}_{X_j}(X_j)),$$

where Φ^{-1} denotes the inverse cumulative distribution function of the standard Normal distribution. Consequently, the transformed feature approximately follows $\mathcal{N}(0, 1)$. More generally, the same framework may be used to map the data to any prescribed target distribution by replacing Φ^{-1} with the corresponding quantile function. Note that, because the empirical cumulative distribution function is estimated on a finite sample, observations lying below or above the fitted data range are mapped to the lower and upper bounds of the target distribution, respectively.

The complete quantile transformation procedure towards a Gaussian distribution is detailed in Algorithm 6 in the Appendix.

After applying the quantile transformation, we further apply min-max scaling to each normally transformed variable X'_j . This scaling ensures that the transformed values of each variable fall within the range $(-1, 1)$ across all policies. Mathematically, for each x'_{ij} :

$$x''_{ij} = \frac{x'_{ij} - \min(X'_j)}{\max(X'_j) - \min(X'_j)} \times (b_s - b_i) + b_i, \quad \text{for } b_i = -1, b_s = 1,$$

where $\min(X'_j)$ and $\max(X'_j)$ respectively denote the minimum and maximum values of the j -th continuous variable across all policies.

The range $(-1, 1)$ was chosen based on the activation function used in the output layer of our decoder network. While a min-max scaling to the range $(0, 1)$ could be considered together with a sigmoid activation function, the hyperbolic tangent ($\tanh$), whose range is from -1 to 1, is often recommended over the sigmoid function for having stronger (and non-biased) gradients (see, e.g., LeCun et al. (2012)). Because our VAE architecture jointly models the numerical and categorical variables contained in the input data, the decoder's output layer will be sliced to extract the values corresponding to the continuous variables before passing them through a $\tanh$ activation function (see darker nodes in Fig. 2.4.1).

2.4.3 Heterogeneous reconstruction loss

Recall that the parameters of the VAE are trained using a negative log-likelihood (reconstruction loss) with a regulariser (KL divergence):

$$\sum_{\mathbf{x} \in \mathcal{X}} \tilde{\mathcal{L}}(\mathbf{x}; \theta, \phi) = \sum_{\mathbf{x} \in \mathcal{X}} \left(-D_{\text{KL}}(q_{\phi}(\mathbf{z}|\mathbf{x}) \parallel p_{\theta}(\mathbf{z})) + \frac{1}{L} \sum_{l=1}^L \ln p_{\theta}(\mathbf{x}|\mathbf{z}^{(l)}) \right).$$

During the training, the error calculated by this loss function $\tilde{\mathcal{L}}(\mathbf{x}; \theta, \phi)$ is propagated backward through the network to update the network and variational parameters (θ, ϕ) and eventually find the optimal values (i.e., the values that minimise the error). Therefore, $\tilde{\mathcal{L}}(\mathbf{x}; \theta, \phi)$ plays a crucial role in the model's convergence. Its purpose is dual. First, the KL divergence regulariser term encourages the VAE to learn a continuously meaningful latent space, which enables the generation of new meaningful policies. Second, the reconstruction loss $\ln p_{\theta}(\mathbf{x}|\mathbf{z}^{(l)})$ encourages the VAE to learn a latent space that captures the salient features of the heterogeneous insurance portfolio which enables its reconstruction given its latent representation. In other words, optimising the reconstruction log-likelihood and regulariser simultaneously encourages the latent state to have distributions close to the prior $\mathcal{N}(0, 1)$ but deviating when necessary. The β -VAE model proposed by Higgins et al. (2017) offers a way to control the balance between reconstruction accuracy and latent space disentanglement by introducing a hyper-parameter $\beta > 1$ to increase the weight on the KL divergence term. While this hyper-parameter β can easily be introduced into our loss function, choosing an appropriate value for it would introduce additional tuning challenges, and better disentanglement may come at the cost of blurrier reconstructions (Higgins et al. 2017).

While the KL divergence admits an analytical expression (see Proposition 2.2), the reconstruction loss $\ln p_{\theta}(\mathbf{x}|\mathbf{z}^{(l)})$ can easily be transformed depending on the distribution assumption of the decoder outputs, and the choice of the decoder depends upon the nature of our input $\mathcal{X}$. Because we assume $\mathcal{X}$ to be a real-world dataset containing various types of features with various marginal distributions, the decoder output, and the associated reconstruction loss, should reflect its heterogeneity.

After pre-processing (normal quantile transform), each continuous variable is assumed to follow a uni-modal Gaussian distribution. Maximising the likelihood of a model whose predictions are assumed to follow a Gaussian distribution is equivalent to minimising mean squared error. In addition, each dummy encoded categorical variable is treated as a multi-class classification problem. Maximising the likelihood of a model whose predictions are assumed to follow a multinomial distribution is equivalent to minimising categorical cross-entropy (also called soft-max loss). To this end, we propose a reconstruction loss that combines the mean squared error and the categorical cross-entropy.

Mathematically, if $\mathbf{x}$ is a vector in $\mathbb{R}^{m_{\text{num}}}$, a natural choice for the decoder is the multivariate Gaussian distribution with, e.g., a diagonal covariance structure (to reduce the number of parameters to estimate),

$$p_{\theta}(\mathbf{x}|\mathbf{z}) = N(\mathbf{x} | \boldsymbol{\mu}_x(\mathbf{z}), \boldsymbol{\Sigma}_x(\mathbf{z})) ,$$

where $\boldsymbol{\mu}_x(\mathbf{z}) \in \mathbb{R}^{m_{\text{num}}}$ and $\boldsymbol{\Sigma}_x(\mathbf{z}) = \boldsymbol{\sigma}_x^2(\mathbf{z}) \mathbf{I} \in \mathbb{R}^{m_{\text{num}} \times m_{\text{num}}}$ are the outputs of a single layer neural network with weights $\mathbf{W}$ and biases $\mathbf{b}$:

$$\begin{aligned} \mathbf{h}_x &= \text{SELU}(\mathbf{W}_1 \mathbf{z} + \mathbf{b}_1) , \\ \boldsymbol{\mu}_x(\mathbf{z}) &= \tanh(\mathbf{W}_2 \mathbf{h}_x + \mathbf{b}_2) , \\ \ln \boldsymbol{\Sigma}_x(\mathbf{z}) &= \text{diag}(\tanh(\mathbf{W}_3 \mathbf{h}_x + \mathbf{b}_3)) , \end{aligned}$$

where

$$\boldsymbol{\mu}_x(\mathbf{z}) = \begin{pmatrix} \mu_{x,1}(\mathbf{z}) \\ \vdots \\ \mu_{x,m_{\text{num}}}(\mathbf{z}) \end{pmatrix} , \quad \boldsymbol{\Sigma}_x(\mathbf{z}) = \begin{pmatrix} \sigma_{x,1}^2(\mathbf{z}) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{x,m_{\text{num}}}^2(\mathbf{z}) \end{pmatrix} ,$$

and the Scaled Exponential Linear Unit (SELU) (Klambauer et al. 2017) and hyperbolic tangent (tanh) activation functions are respectively defined as:

$$\begin{cases} \text{SELU}(x) = \lambda x & \text{if } x > 0 \quad \text{with } \lambda \approx 1.05 \\ \text{SELU}(x) = \lambda \alpha (e^x - 1) & \text{if } x \leq 0 \quad \text{with } \alpha \approx 1.67 \end{cases} , \quad \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} .$$

In this case, the probabilistic decoder has density:

$$p_{\theta}(\mathbf{x}|\mathbf{z}) = \frac{\exp\left(-\frac{1}{2} \sum_{k=1}^{m_{\text{num}}} \frac{(x_k - \mu_{x,k}(\mathbf{z}))^2}{\sigma_{x,k}^2(\mathbf{z})}\right)}{(2\pi)^{\frac{1}{2}m_{\text{num}}} \sqrt{\prod_{k=1}^{m_{\text{num}}} \sigma_{x,k}^2(\mathbf{z})}} ,$$

with a log-likelihood equal to:

$$\ln p_{\theta}(\mathbf{x}|\mathbf{z}) = -\frac{m_{\text{num}}}{2} \ln(2\pi) - \frac{1}{2} \sum_{k=1}^{m_{\text{num}}} \ln \sigma_{x,k}^2(\mathbf{z}) - \frac{1}{2} \sum_{k=1}^{m_{\text{num}}} \frac{(x_k - \mu_{x,k}(\mathbf{z}))^2}{\sigma_{x,k}^2(\mathbf{z})} .$$

If we both consider a Gaussian encoder and decoder, network parameters are obtained by maximising:

$$\begin{aligned}
\sum_{\mathbf{x} \in \mathcal{X}} \tilde{\mathcal{L}}(\mathbf{x}; \theta, \phi) &= \sum_{\mathbf{x} \in \mathcal{X}} \left(-D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z})) + \frac{1}{L} \sum_{l=1}^L \ln p_\theta(\mathbf{x}|\mathbf{z}^{(l)}) \right) \\
&= \sum_{\mathbf{x} \in \mathcal{X}} \frac{1}{2} \sum_{k=1}^d (1 + \ln \sigma_{z,k}^2(\mathbf{x}) - \sigma_{z,k}^2(\mathbf{x}) - \mu_{z,k}^2(\mathbf{x})) \\
&\quad + \sum_{\mathbf{x} \in \mathcal{X}} \frac{1}{L} \sum_{l=1}^L \left[-\frac{m_{\text{num}}}{2} \ln(2\pi) - \frac{1}{2} \sum_{k=1}^{m_{\text{num}}} \ln \sigma_{x,k}^2(\mathbf{z}^{(l)}) \right. \\
&\quad \left. - \frac{1}{2} \sum_{k=1}^{m_{\text{num}}} \frac{(x_k - \mu_{x,k}(\mathbf{z}^{(l)}))^2}{\sigma_{x,k}^2(\mathbf{z}^{(l)})} \right].
\end{aligned}$$

where

$$\mathbf{z}^{(l)} = \boldsymbol{\mu}_z(\mathbf{x}) + \sqrt{\boldsymbol{\Sigma}_z(\mathbf{x})} \boldsymbol{\epsilon}^{(l)}.$$

Maximising the log-likelihood $\ln p_\theta(\mathbf{x}|\mathbf{z})$ of a model whose predictions are assumed to follow a Gaussian distribution is equivalent to minimising the mean squared error:

$$L_{\text{MSE}} = \sum_{i=1}^n \sum_{j=1}^{m_{\text{num}}} \|x_{i,j} - \hat{x}_{i,j}\|^2.$$

In parallel, let $\mathbf{x}$ denote the dummy-encoded vector of a categorical variable with c modalities. A natural choice for the decoder distribution is the multinomial model:

$$p(\mathbf{x}|\mathbf{z}) = \prod_{j=1}^c y_j^{x_j}(\mathbf{z}),$$

where $y(\mathbf{z}) = (y_1(\mathbf{z}), \dots, y_c(\mathbf{z}))$ are the outputs of a single layer neural network:

$$\begin{aligned}
\mathbf{h}_x &= \text{SELU}(\mathbf{W}_1 \mathbf{z} + \mathbf{b}_1), \\
y(\mathbf{z}) &= \text{softmax}(\mathbf{W}_2 \mathbf{h}_x + \mathbf{b}_2),
\end{aligned}$$

where the softmax activation function,

$$\text{softmax}(x) = \left(\frac{e^{x_{j'}}}{\sum_{j=1}^c e^{x_j}} \right)_{j'=1, \dots, c}, \quad (2.4.1)$$

is applied to the output scores inferred by the NN for each class $i = 1, \dots, c$ in the multi-class variable X .

The likelihood is then equal to the categorical cross-entropy:

$$\ln p(\mathbf{x}|\mathbf{z}) = \sum_{j=1}^c x_j \ln(y_j(\mathbf{z})). \quad (2.4.2)$$

Since the categorical variable is dummy encoded, there is only one element, denoted x^* , in the target vector $\mathbf{x} = \{x_1, \dots, x_c\}$ which is not zero. We can rewrite Eq. (2.4.2) as follow:

$$\ln p(\mathbf{x}|\mathbf{z}) = \ln \left(\frac{e^{x^*}}{\sum_{j=1}^c e^{x_j}} \right).$$

Maximising the likelihood of a model whose predictions are assumed to follow a multinomial distribution is equivalent to minimising this categorical cross-entropy.

Note that while softmax activations provide a convenient mechanism for reconstructing categorical variables, they treat ordinal variables as nominal categories and therefore do not explicitly account for the ordering between modalities. Future research could investigate ordinal-specific decoder formulations, such as cumulative logit decoders (Gutiérrez et al. 2015), CORAL-type architectures (Cao et al. 2020), or distance-based reconstruction losses, which incorporate the ordinal structure directly into the learning process.

The architecture of our Variational AutoEncoder is detailed in Algorithm 7 in Appendix. For the probabilistic encoder, we use a single hidden layer with Gaussian output. For the probabilistic decoder, we propose a combination of Gaussian and multinomial outputs, depending on the data type (continuous or discrete). The algorithm starts drawing batches by taking the first M instances (first batch) from the randomly shuffled dataset, calculates the average error, and updates the parameters. This completes one training step (i.e., one iteration or gradient update). When all batches have been drawn from the training dataset (i.e., the model has been trained on the entire dataset), that concludes one epoch.

2.5 Model training

2.5.1 Motor insurance portfolio

In our application, each policy in the Swedish insurance company database (Ohlsson et al. 2010) is described by two continuous variables, the policyholder’s age and its vehicle age, and by three categorical variables, the gender (with two modalities), zone (with seven modalities), and class (with seven modalities), detailed in Table 1.A.1. The portfolio further contains, for each policy $i = 1, \dots, n$, an ordinal variable reporting the number of claims $N_i \in \{0, 1, 2\}$, and a continuous variable ν_i indicating the contract duration (exposure). After selecting contracts with a duration $\nu_i \in]0, 12]$ months, the dataset counts $n = 62\,289$ policies. As expected for a motor portfolio it is highly imbalanced, with zero claims accounting for 98.95% of the policies, one claim for 1%, and two claims for 0.05%.

The age, vehicle age and contract duration all follow a multi-modal distribution (see Fig. 2.5.1). We use the quantile (non-parametric) transformation, detailed in Algorithm 6, to map the values of each continuous variable to an uni-modal normal distribution. We then apply a min-max transformation to scale the normally

distributed data to the range $(-1, 1)$. The distributions before and after scaling are showed in Fig. 2.5.1.

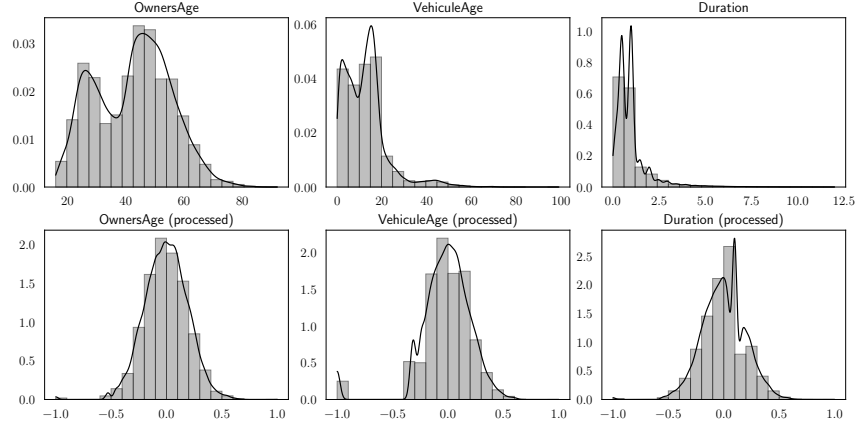


Figure 2.5.1: Distributions of OwnersAge, VehicleAge, and Duration before (top row) and after (bottom row) preprocessing. The preprocessing consists of a normal quantile transformation followed by min-max scaling to the range $(-1, 1)$. Histogram bars are shown in gray with overlaid black density curves.

The categorical variables Zone, Class, Gender and NumberClaims are dummy encoded with, respectively, 7, 7, 2, and 3 dummy columns. A quick analysis of our insurance tabular data reveals (see Fig. 2.5.2) a highly imbalanced dataset in which the number of policies in one modality (e.g., male policyholders) greatly outnumbers the policies in another (e.g., female policyholders). Vehicles with the highest power ratio (i.e., Class 7), as well as Northern geographic locations (Zone 5, 6, and 7) are also strongly under-represented.

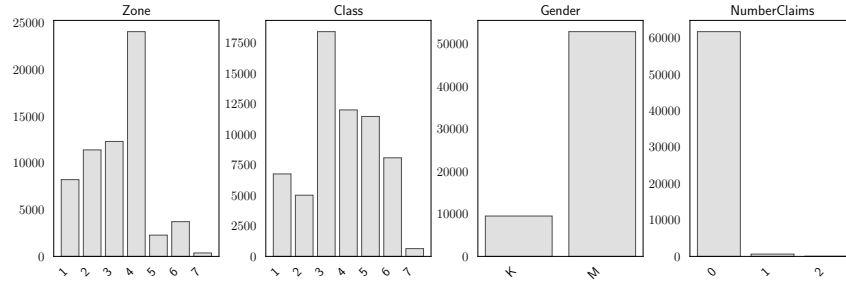


Figure 2.5.2: Distributions of the categorical variables in the dataset. Each bar represents the count of observations in the corresponding category.

When dealing with imbalanced data, controlling the model architecture and hyper-parameters setup like the batch size can play an important role in the performance of the model (i.e., reconstruction accuracy, and synthetic data quality) and its ability to learn a meaningful latent space. If certain classes are under-represented in the training subsets, whose size are determined by the batch size hyper-parameter, the latent space representation may not be well-distributed or separated for the different classes, leading to challenges in generating diverse and high-quality samples for the minority classes. To ensure each training subset includes enough representative examples from the minority classes, we recommend the use of smaller batch sizes as larger batch sizes might inadvertently amplify the influence of the majority class(es) and lead to poor performance on the minority class(es). In the following subsections, we find that choosing a moderate batch size of 256 ensures minority classes representation while still remaining within acceptable computation time (around 20 minutes). The ability of our VAE to handle imbalanced classes, that is, its ability to reconstruct and generate new policies representing all classes, including those with fewer observations, is illustrated in the following subsections in Fig. 2.5.4 and Fig. 2.5.9. The following subsection 2.5.2 further details the hyper-parameters set up and model architecture that lead to the model performance.

2.5.2 Model architecture and hyper-parameters set up

We train our Variational AutoEncoder model (see Algorithm 7) on the pre-processed insurance portfolio with $n = 62\,289$ rows, $m_{\text{num}} = 3$ quantile and min-max transformed columns, and $m_{01} = 19$ dummy columns. The architecture of our encoder and decoder networks, as well as the values of the key hyper-parameters involved in the training of our VAE are listed in Tables 2.5.1 and 2.5.2.

Experimentation and tuning are often necessary to find optimal values for these hyper-parameters depending on the input data and its complexity, and the desired properties of the learned latent space. Since our VAE is trained to learn to reconstruct the input, the input and output space are both of dimension $m = 22$ (i.e., the first layer of the encoder and the last layer of the decoder both have 22 neurons). Our intention to build a simple and straightforward autoencoder led us to choose a shallow architecture with a single hidden layer in the encoder NN. The mapping of the original input to the latent space is done through the use of a SELU activation function. By definition, an activation function determines whether a neuron should be activated by combining the sum of the weighted inputs and a bias term. SELUs are activation functions that induce self-normalisation (i.e., automatically converge to a zero mean and unit variance) and are a common choice for NNs, not only because they help alleviate the vanishing gradient problem (like ReLUs), but also because they cannot die (contrary to ReLUs). SELUs can also get below 0, allowing the system to have a zero average output, which can improve the speed of convergence (Glorot et al. 2011).

The input portfolio is thence mapped to its latent representation through a single hidden layer with a SELU activation function and a Lecun normal initializer. The variational parameters are initialised with zeros (for the means), and a small constant value set to 0.1 (for the log variances).

	Encoder network	Bottleneck layer	Decoder network
Number of layers	1	1	2
Neurons per layer	22	15	(22, 22)
Activation function(s)	SELU	–	(SELU, tanh / softmax)
Weight initialisation	LeCun normal	$\mu = 0, \sigma = 0.1$	(LeCun normal, Xavier)

Table 2.5.1: Neural network architecture and configuration of the encoder, bottleneck, and decoder components (see Fig. 2.B.1 in Appendix).

The latent (hidden) representation space is typically referred to as a bottleneck layer, as it contains fewer neurons than the layer below or above it (Tishby et al. 2000). This encourages the NN to learn a compact but meaningful representation of the data.

In the instance of motorcycle insurance policies, the latent encoding may contain a hidden representation of the vehicle characteristics and its owner’s risk profile.

The choice of setting the latent space dimension to 15, thereby reducing the dimension of about 1/3, is based on an extensive analysis of the quality of both the reconstructed and synthetic data. The metrics and visualisations used to choose the latent space dimension are detailed in the following two subsections for the optimal value of $d = 15$.

Figure 2.5.3 gives some insights into this choice by showing the evolution of the Poisson deviance with respect to the latent space dimension d . This deviance metric evaluates the quality of the reconstructed portfolio by estimating the goodness of fit of a GLM model trained on the reconstructed insurance portfolio to predict the claim frequencies of the authentic policies. To train our regression model (GLM) on the reconstructed portfolio, we extracted the covariates Zone, Class, Gender, OwnersAge, and VehicleAge from the decoder’s output. We first discretised the continuous variables OwnersAge and VehicleAge before dummy encoding all covariates. We regressed the target variable NumberClaims (extracted from the decoder’s output) given the dummy encoded covariates with a sample weight equal to the Duration (also extracted from the decoder’s output). We can then use this trained GLM as an estimator to predict the claim frequencies $\hat{\lambda}_i$ of the original portfolio and compute the Poisson deviance, defined as:

$$\mathcal{D}_{\text{Poisson}} = 2 \sum_{i=1}^n N_i \left(\frac{\nu_i}{N_i} \hat{\lambda}_i - \left(\log \frac{\hat{\lambda}_i \nu_i}{N_i} + 1 \right) I_{\{N_i \geq 1\}} \right). \quad (2.5.1)$$

where N_i/ν_i are the observed claim frequencies of the authentic policies (obtained by dividing NumberClaims by Duration in the original dataset) and $\hat{\lambda}_i$ are the

predicted frequencies of the original policies estimated by the GLM trained on the reconstructed policies. The closer this deviance is to the deviance of a GLM trained on the original portfolio, i.e., 5961.84, the better our reconstruction quality is.

To give some insights into the impact of the latent space $\mathcal{Z}$ dimension d on the reconstruction quality (measured by the Poisson deviance in Fig. 2.5.3), we provide a brief analysis of its impact on the quality of the reconstructed categorical variables.

With $d = 1$, only the most represented modality in each categorical variable (see Fig. 2.5.2) is reconstructed. When we increase the $\mathcal{Z}$ dimension to two, our model is able to reconstruct additional categories (Zone 2, Class 1, 3, 5, and 7).

With $d = 3$, while almost all modalities of the variable Class are reconstructed (except Class 2, with the second lowest number of observations), only Zone 4, Male, and 0 reported claims are reconstructed.

With $d = 4$, we start to see real improvements in the categorical variables reconstruction (all modalities of the variable Class, Gender, and Zone, except Zone 7 with the lowest number of observations, are reconstructed). As we increase the $\mathcal{Z}$ dimension to five, six, and seven, we improve the miss-classification error, yet, we still fail to reconstruct Class 2, with the second lowest number of observations.

As of $d = 8$, the model is able to reconstruct all modalities. However, the reconstructed portfolio still deviates too much from the authentic as the GLM trained on the reconstructed fails to accurately predict the claim frequencies of the authentic portfolio as illustrated in Fig. 2.5.3 (increase in deviance of 51.55%). We first observe a deviance drop, with a deviance closer to the reference value of 5961.84, with $d = 10, 11$, and 13. Starting from $d = 15$ (chosen as the optimal value), the deviance stabilises.

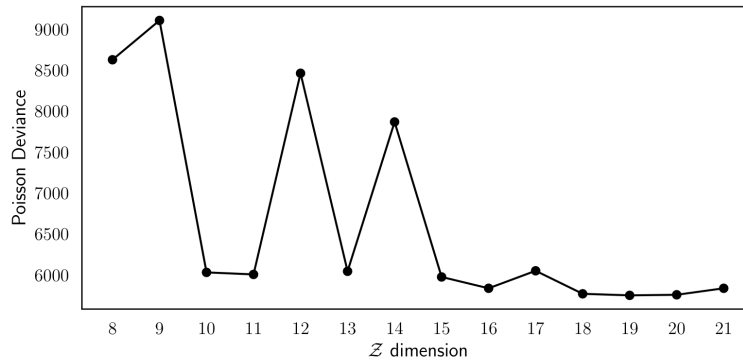


Figure 2.5.3: Evolution of the Poisson deviance of a GLM model trained on the reconstructed insurance portfolio to predict the claim frequencies of the authentic policies, with respect to the latent space $\mathcal{Z}$ dimension $d \in (8, 21)$.

Our decoder mirrors the architecture of the encoder but operates in reverse,

generating outputs that match the input data. The latent space is mapped back to the initial space with a SELU activation function and a Lecun normal initializer. Because of the heterogeneity in the distributions of our input data (see Section 2.4.3), the decoding layer outputs are sliced before being passed through tanh and softmax activation functions with Xavier initializers in the final layer of our decoder (as illustrated in Fig. 2.4.1 and Fig. 2.B.1).

Hyperparameter	Value
Optimiser (RMSprop) learning rate	0.001
Batch size	256
Number of epochs	75

Table 2.5.2: Choice of the optimiser, its learning rate, and the gradient descent hyper-parameters.

The RMSprop optimiser with default learning rate value $\alpha = 0.001$ (which determines the speed of the optimiser’s descent on the error curve) gave better results on our training data than the often-recommended Adam optimiser. The main hyper-parameters of the (mini-batch) gradient descent optimisation algorithm are the batch size and the number of epochs. The larger the datasets, the more time-consuming and computationally expensive each training step becomes. In Keras, batches of the randomly shuffled dataset are used for each gradient update. Smaller batch sizes (e.g., 16, 32, 64) introduce more noise and randomness to the (more frequent) parameter updates, which can have regularising effects and improve the model’s generalisation and, therefore, the quality of synthetic policies. However, training with smaller batches requires more iterations and may take longer overall. Conversely, larger batch sizes (e.g., 128, 256, 512, 1024) can accelerate training but may result in less noisy weight updates and potentially hinder the model’s generalisation. After experimenting with different batch sizes and monitoring the model’s performance (i.e., reconstruction accuracy, latent space properties, and synthetic data quality), we find that a batch size of 256 and 75 epochs ensure that each batch provides enough representative examples from the minority classes while remaining within acceptable computation time (around 20 minutes).

2.5.3 Reconstruction quality

Thanks to the KL divergence regularizer term in the loss function, our VAE was able to learn multivariate independent latent variables. As shown in Fig. 2.C.1 in Appendix, the linear correlations between the latent encodings are close to 0. The latent variables are also centred around 0 with standard deviations close to 1 (see Table 2.D.1 and Fig. 2.C.2 in Appendix).

Thanks to the reconstruction loss, which allows for deviations from the unit Gaussian prior to capture the data most salient features, we also observe that

policies which reported 1 or 2 motor claims are concentrated towards a lower region of the latent space and that their mean and variance also slightly deviate from the standard values of 0 and 1 as illustrated in Fig. 2.C.2 in Appendix. This enabled our VAE model to reconstruct the original data fairly accurately.

To evaluate the quality of the reconstructed data, a reverse scaling is applied to the decoder's output $\hat{\mathbf{x}}$. A reverse dummy encoding is applied to the reconstructed categorical classes, and a reverse min-max transformation followed by a reverse quantile transformation is applied to the reconstructed continuous variables. As shown in Fig. 2.5.4, the frequency of each observed category in the reconstructed data matches that of the original data. Fig. 2.5.5 also shows that the densities of the original and reconstructed continuous data are almost perfectly overlapping.

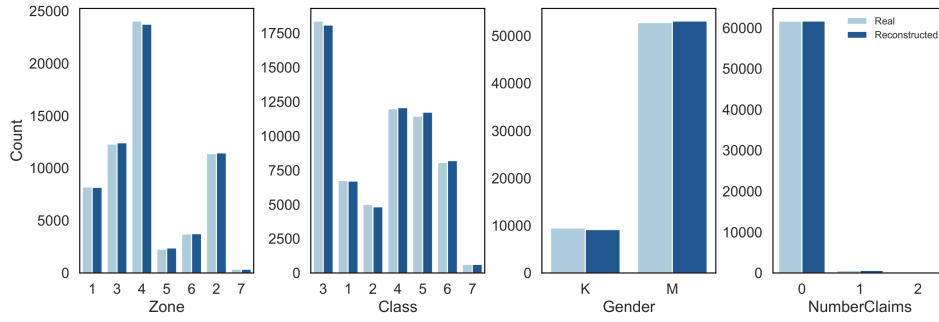


Figure 2.5.4: Comparison of real (light blue) and reconstructed (dark blue) categorical data distributions (frequency of each observed category).

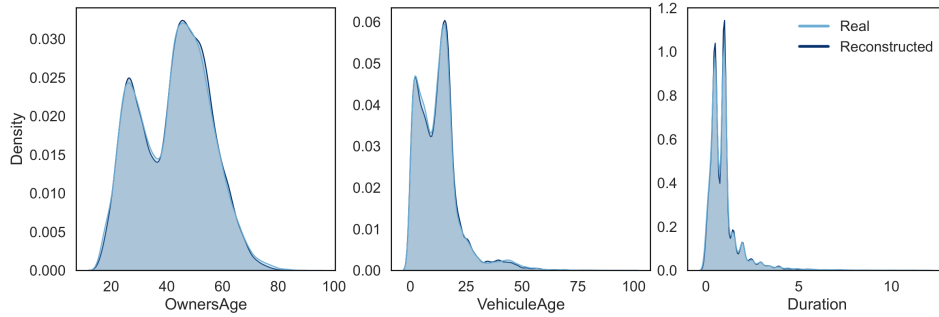


Figure 2.5.5: Comparison of real (light blue) and reconstructed (dark blue) continuous data distributions.

The visual comparison between Fig. 2.5.5 and Fig. 2.5.6 illustrates the positive impact of introducing and applying a quantile transformation on the continu-

ous variables (following non-Gaussian multi-modal distributions), before inputting them into the encoder network. Without applying our normal quantile transformation on the inputted continuous variables, the distributions of the reconstructed variables (in dark blue) clearly fail to capture the two modes of the original VehicleAge’s distribution and the multiple modes of the original Variable Duration.

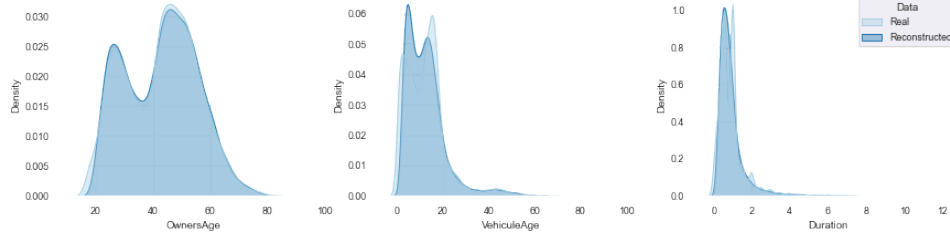


Figure 2.5.6: Reconstructed (in dark blue) Duration, OwnersAge and VehicleAge without quantile transformation compared to their real distributions (in light blue).

The impact of introducing a quantile transformation on the multi-modal input variables is even more pronounced when comparing the synthetic data distributions in Fig. 2.5.7 and Fig.2.5.10 (obtained in subsection 2.5.4 by feeding random noises to the decoder’s network). When training our VAE on our pre-processed data without a normal quantile transformation scaler, the synthetic distributions (in dark blue) in Fig. 2.5.7 completely fail to reproduce the modes of the original distributions (in light blue). As illustrated in Fig.2.5.10, the normal quantile transformation we proposed to scale our continuous input data greatly improves the distributions of the synthetic continuous data.

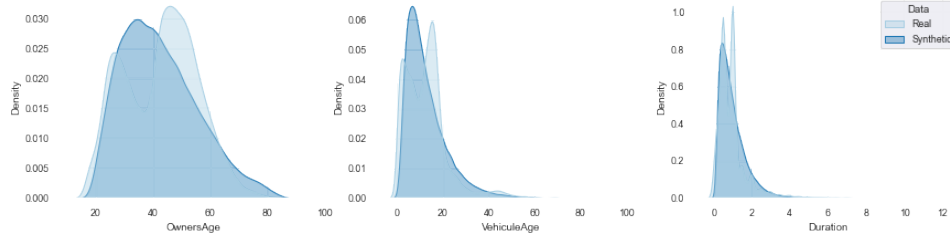


Figure 2.5.7: Synthetic (in dark blue) Duration, OwnersAge and VehicleAge without quantile transformation compared to their real distributions (in light blue).

2.5.4 Synthetic data generation

As motivated in Section 2.3, the trained decoder network can be used as an unsupervised generative model to create new insurance policies that mimic the real

policies patterns. This can become a valuable tool when real-world data is difficult to collect. Whether the data is real or artificial is irrelevant; what matters are the structure, statistical characteristics, and patterns inside the data. Ideally, synthetic data should preserve the data distributions and correlation structure.

In their seminal papers, neither Kingma et al. (2013) nor Rezende, Mohamed, and Wierstra (2014) seem to specify the best way to generate new data points. In their research on β -VAE, Higgins et al. (2017) do mention that the most informative latent units of β -VAE have the highest KL divergence from the unit Gaussian prior, confirming at least that the posterior distribution does not perfectly match the Gaussian prior $p(\mathbf{z}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and the difference matters. As illustrated in the left plot of Fig. 2.5.8, although the latent variables are roughly centred around 0 with standard deviations close to 1 (see Table 2.D.1 in Appendix), they do not necessarily fit neatly into a standard bell curve. Applying a normal quantile transformation $\Phi^{-1}(F(Z))$ to each latent variable Z ensures their densities are bell-shaped (see right plot of Fig. 2.5.8).

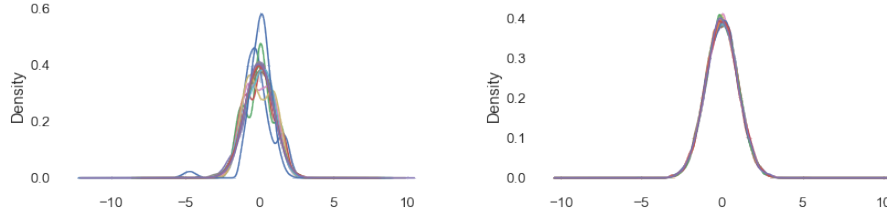


Figure 2.5.8: Densities of the $d = 15$ latent variables, before and after normal quantile transformation.

With that knowledge in mind, we generate new policies by first sampling from the prior distribution $p(\mathbf{z}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. By doing so, samples $\mathbf{z}$ from the whole data distribution are drawn. Because of the non-bell shaped distributions of the latent codes $\mathbf{z}$, sampling from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ might fail and under-sample or omit some regions of the true distribution, over-sample others, and even provide inputs not covered by the aggregate posterior. Therefore, we propose to apply the inverse quantile transformations of the latent variables to the normally distributed samples $\mathbf{z}$.

The new samples $\mathbf{z}$ are then fed as input to the decoder. The quality of the so-generated synthetic policies is evaluated with respect to the statistical characteristics of the original portfolio. Fig. 2.5.9 and Fig. 2.5.10 show that the distributions of the synthetic data match those of the original data. The ability of our VAE model to generate new policies representing all classes, including those with fewer observations (e.g., Class 7 or Zone 7), was achieved thanks to the model architecture (including the heterogeneous reconstruction loss with a categorical cross-entropy for categorical variables) and the hyper-parameters setup (including the choice of a moderate batch size to include enough representative examples from the minority

classes in the training subsets) detailed in the previous subsection 2.5.2.

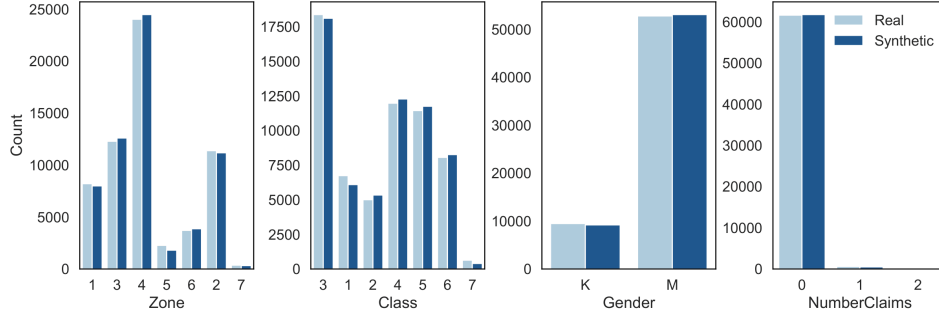


Figure 2.5.9: Comparison of real (light blue) and synthetic (dark blue) categorical data distributions (frequency of each observed category).

As mentioned at the end of the previous subsection 2.5.2 and illustrated in Fig. 2.5.10, the normal quantile transformation we introduced and applied to scale our continuous variables (before inputting them into the encoder network) greatly improved the similarity between the synthetic continuous variables distributions and the original ones.

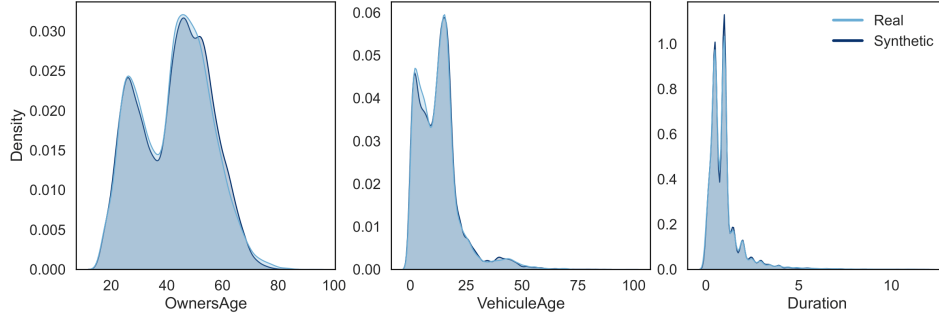


Figure 2.5.10: Comparison of real (light blue) and synthetic (dark blue) continuous data distributions.

This visual analysis of the quality of our synthetic data is further supported and verified by a wide range of measures. We first evaluate the ability of our VAE model to capture pairwise correlations between continuous variables by computing Pearson and Spearman correlations. Table 2.5.3 shows that the original correlations between continuous variables are preserved in the synthetic dataset. The scores in Table 2.5.3 were obtained by computing $1 - 0.5 |\text{corr}^{\text{real}} - \text{corr}^{\text{synthetic}}|$.

		OwnersAge		VehicleAge	
		Pearson ρ	Spearman r_s	Pearson ρ	Spearman r_s
VehicleAge	Real	0.0755	0.0671	–	–
	Synthetic	0.0770	0.0777	–	–
	Score (%)	99.9243	99.4729	–	–
Duration	Real	0.1173	0.1694	–0.0278	0.0787
	Synthetic	0.1334	0.1689	0.0526	0.0608
	Score (%)	99.1948	99.9766	95.9783	99.1068

Table 2.5.3: Pearson and Spearman correlations between real and synthetic continuous data.

The synthetic and original continuous data distributions are further compared using the Kolmogorov-Smirnov (KS) statistic-based metric. The results are reported in Table 2.5.4. The output values are pretty close to 1, indicating that the distributions (the real CDFs and the synthetic CDFs values) are almost identical, verifying our visual analysis.

Variable	KS metric
OwnersAge	0.9779
VehicleAge	0.9856
Duration	0.9863

Table 2.5.4: Kolmogorov-Smirnov statistic-based similarity metric between real and synthetic continuous variables.

The quality of the synthetic categorical data is checked using the Chi-Squared test-based metric. The results reported in Table 2.5.5 are once again an indication of the high-quality of our synthetic data. Note, however, that due to the low number of policies (25 out of 62 289) with two reported claims, we often fail to generate qualitative artificial policies with two claims (hence the 67% for the NumberClaims modalities coverage).

	Zone	Class	Gender	NumberClaims
Chi-squared test statistic	0.9852	0.9815	0.9954	0.9977
Modality coverage (%)	100.0000	100.0000	100.0000	67.6667

Table 2.5.5: Chi-squared test statistics and modality coverage between real and synthetic categorical variables.

The boundaries adherence and statistical similarities reported in Table 2.5.6 are also verified. All synthetic data are within the minimum and maximum bounds of the real data, and we observe highly similar values for the mean, standard deviations, and different quantiles.

		OwnersAge		VehicleAge		Duration	
		Real	Synthetic	Real	Synthetic	Real	Synthetic
Statistical similarity	mean	42.5438	42.8047	12.5584	12.6940	1.0079	0.9938
	std	12.9549	12.9445	9.6846	9.6192	1.0621	0.9759
	min	16.0000	16.0000	0.0000	0.0000	0.0027	0.0027
	25%	31.0000	31.0000	5.0000	6.0000	0.4904	0.4959
	50%	44.0000	44.0000	12.0000	12.0000	0.8877	0.9178
	75%	52.0000	53.0000	16.0000	17.0000	1.0000	1.0000
	max	92.0000	83.0000	99.0000	87.0000	11.9945	11.7653
Statistical score	mean	0.9966		0.9986		0.9988	
	std	0.9999		0.9993		0.9928	
	median	1.0000		1.0000		0.9975	

Table 2.5.6: Statistical similarity between real and synthetic continuous variables. For each statistic t , the associated score is defined as $1 - \|t^{\text{real}} - t^{\text{synthetic}}\| / (\max^{\text{real}} - \min^{\text{synthetic}})$.

We further compare the contingency similarity of the categorical variables in Table 2.5.7. These scores were obtained by means of a contingency table, which displays frequencies for combinations of two categorical columns.

	Zone	Class	Gender
Class	0.9581	–	–
Gender	0.9796	0.9738	–
NumberClaims	0.9843	0.9797	0.9931

Table 2.5.7: Contingency score between real and synthetic categorical variables.

To evaluate whether our synthetic insurance portfolio could be used as a substitute for the original portfolio in actuarial analyses, we compare the Poisson deviance in Eq. (2.5.1) obtained with two different GLMs.

The first GLM is trained on the original insurance portfolio X and is used to predict the claim frequencies $\hat{\lambda}_i^{\text{GLM}_1}$ of the original policies $i = 1, \dots, n$. By replacing in Eq. (2.5.1) the observed claim frequencies N_i/ν_i by the number of claims in the original variable NumberClaims divided by the policy's exposure in the original variable Duration, and $\hat{\lambda}_i$ by the predicted claim frequencies $\hat{\lambda}_i^{\text{GLM}_1}$, we obtain a Poisson deviance equal to 5 691.01.

The second GLM is trained on the synthetic portfolio X^s and is used to predict

the claim frequencies $\hat{\lambda}_i^{GLM_2}$ of the original policies $i = 1, \dots, n$. We were able to fit a GLM to the synthetic data because both the continuous variable Duration and the variable NumberClaims were fed as input to our VAE. By replacing in Eq. (2.5.1) the observed claim frequencies N_i/ν_i by the synthetic variable NumberClaims divided by the synthetic variable Duration (extracted from the generative decoder’s output), and $\hat{\lambda}_i$ by the predicted claim frequencies $\hat{\lambda}_i^{GLM_2}$ of the original policies, we obtain a Poisson deviance equal to 5961.84. Table 2.5.8 reports the Poisson deviances and the difference (in %) from the reference Poisson deviance obtained with the first GLM. The smaller this variation (%), the more our synthetic insurance portfolio can be considered a strong substitute for the real one. Fig. 2.C.4 in Appendix illustrates the aforementioned results; subject to the same estimator (GLM), predicted claim frequencies from the authentic and synthetic portfolios should be close to the first bisector.

We further assess the quality of the reconstructed portfolio by fitting a third GLM on the reconstructed portfolio X^r and by predicting the claim frequencies $\hat{\lambda}_i^{GLM_3}$ of the original policies $i = 1, \dots, n$. By replacing in Eq. (2.5.1) the observed claim frequencies N_i/ν_i by the reconstructed variable NumberClaims divided by the reconstructed variable Duration (extracted from the decoder’s output), and $\hat{\lambda}_i$ by the predicted claim frequencies $\hat{\lambda}_i^{GLM_3}$ of the original policies, we obtain a Poisson deviance equal to 5713.09, which is relatively close to the Poisson deviance obtained with the first GLM. This strengthens our previous quality analysis of the reconstructed portfolio.

Data type	GLM model	Poisson deviance	% difference
Real data	1	5691.01	–
Synthetic data	2	5961.84	4.76 %
Reconstructed data	3	5713.09	0.39 %

Table 2.5.8: Poisson deviances for GLM models on real, synthetic, and reconstructed data.

While a low reconstruction error indicates that the VAE has successfully learned the underlying data distribution, it does not in itself guarantee privacy protection. In particular, generative models may still exhibit memorisation effects, where a fraction of synthetic records are exact or near-exact replicas of real observations.

In our experiment, the fake and original portfolio both contain 62 289 policies. In the fake portfolio, 2.55% of the fake policies are duplicates and 97.45% of the policies are unique. Among the unique fake policies, 1.82% were identified as direct copies of real policies, which corresponds to a potential risk of re-identification and therefore constitutes a privacy concern.

It is important to distinguish between memorisation risk and formal privacy guarantees. The former can be assessed empirically using nearest-neighbour com-

parisons, distance-based matching criteria, or membership inference attacks, while the latter requires explicit mechanisms such as differential privacy. Several privacy-preserving extensions of VAEs have been proposed, including differentially private VAEs (DP-VAEs) (Weggenmann et al. 2022), which inject calibrated noise into the training procedure or modify the optimisation objective to ensure formal privacy guarantees. However, such approaches typically introduce a trade-off between privacy and utility, as stronger privacy constraints tend to reduce the fidelity of the generated data and may impact downstream predictive performance.

2.6 Conclusion

This chapter explored the application of VAEs for multi-type insurance data. The study’s objective was to leverage the power of VAEs to extract a meaningful and low-dimensional representation of an insurance portfolio to generate artificial data. Thanks to our quantile transformations and NNs architecture, we successfully generated realistic and diverse samples of synthetic insurance policies characterised by multinomial and multi-modal non-Gaussian variables. The ability of our VAE to uncover the underlying structure in the input data offers a promising avenue for insurance companies to enhance their decision-making processes by facilitating tasks such as risk assessment and claim prediction while overcoming privacy concerns associated with the use of authentic policyholder information. Although VAEs assure a bright outlook for insurance analytics, there are still challenges to overcome. In particular, the research highlighted the significance of appropriate model architecture and hyper-parameters tuning to achieve optimal results. Careful consideration should be given to the selection of activation functions, layer sizes, regularisation techniques, and optimisation algorithms to ensure the effective training and performance of VAEs on insurance data.

Note that, given the moderate dimensionality of the portfolio analysed in this study (22 encoded variables and approximately 60 000 observations), a standard VAE appears sufficiently flexible. Future work could nevertheless investigate non-parametric extensions such as stick-breaking VAEs, not because of sample size limitations, but because they allow the complexity of the latent representation to adapt automatically to heterogeneous insurance portfolios. This may be particularly beneficial when modelling portfolios containing numerous latent risk segments or richer sources of information such as telematics, text, or claims histories.

For future research, the latent codes obtained from our VAE model could also serve as a valuable input for a subsequent unsupervised regression step (e.g., clustering), provided that the target variable (e.g., NumberClaims) is withdrawn from the encoder’s input. By utilising the blurred latent representation of the original rating factors in a downstream task, the direct and explicit usage of sensitive attributes, such as gender or ethnicity, in the subsequent regression task could be avoided. Finally, our generative model could also be adapted and conditioned on a modality or a target value to generate new additional representative policies.

2.A Proofs

Proof of Proposition 2.1. As $\int q_\phi(\mathbf{z}|\mathbf{x})d\mathbf{z} = 1$, the marginal log-likelihood is rewritten as follows:

$$\ln p_\theta(\mathbf{x}) = \int q_\phi(\mathbf{z}|\mathbf{x}) \ln p_\theta(\mathbf{x}) d\mathbf{z}.$$

Given that $p_\theta(\mathbf{x}) = \frac{p_\theta(\mathbf{x}, \mathbf{z})}{p_\theta(\mathbf{z}|\mathbf{x})}$ (Bayes theorem), we can develop the integral in the following manner:

$$\begin{aligned} \ln p_\theta(\mathbf{x}) &= \int q_\phi(\mathbf{z}|\mathbf{x}) (\ln p_\theta(\mathbf{x}, \mathbf{z}) - \ln p_\theta(\mathbf{z}|\mathbf{x})) d\mathbf{z} \\ &= \int q_\phi(\mathbf{z}|\mathbf{x}) (\ln p_\theta(\mathbf{x}, \mathbf{z}) - \ln q_\phi(\mathbf{z}|\mathbf{x}) + \ln q_\phi(\mathbf{z}|\mathbf{x}) - \ln p_\theta(\mathbf{z}|\mathbf{x})) d\mathbf{z} \\ &= \int q_\phi(\mathbf{z}|\mathbf{x}) \ln \left(\frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z}|\mathbf{x})} \right) d\mathbf{z} + \int q_\phi(\mathbf{z}|\mathbf{x}) \ln \left(\frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} \right) d\mathbf{z} \\ &= D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) || p_\theta(\mathbf{z}|\mathbf{x})) + \mathcal{L}(\mathbf{x}; \theta, \phi). \end{aligned}$$

□

Proof of Proposition 2.2. By definition, the KL divergence is:

$$D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) || p_\theta(\mathbf{z})) = \mathbb{E}_{q_\phi}[\ln q_\phi(\mathbf{z}|\mathbf{x})] - \mathbb{E}_{q_\phi}[\ln p_\theta(\mathbf{z})],$$

where the densities are equal to:

$$\begin{cases} p_\theta(\mathbf{z}) &= \frac{\exp(-\frac{1}{2}\mathbf{z}^\top \mathbf{z})}{(2\pi)^{d/2}}, \\ q_\phi(\mathbf{z}|\mathbf{x}) &= \frac{\exp\left(-\frac{1}{2}\sum_{k=1}^d \frac{(z_k - \mu_{z,k}(\mathbf{x}))^2}{\sigma_{z,k}(\mathbf{x})^2}\right)}{(2\pi)^{d/2} \sqrt{\prod_{k=1}^d \sigma_{z,k}^2(\mathbf{x})}}. \end{cases}$$

A direct calculation allows to infer that for $\mathbf{z} \in \mathbb{R}^d$:

$$\begin{aligned} \mathbb{E}_{q_\phi}(\ln q_\phi(\mathbf{z}|\mathbf{x})) &= \mathbb{E}_{q_\phi} \left[-\frac{d}{2} \ln(2\pi) - \frac{1}{2} \sum_{k=1}^d \ln \sigma_{z,k}^2(\mathbf{x}) - \frac{1}{2} \sum_{k=1}^d \frac{(z_k - \mu_{z,k}(\mathbf{x}))^2}{\sigma_{z,k}(\mathbf{x})^2} \right] \\ &= -\frac{d}{2} \ln(2\pi) - \frac{1}{2} \sum_{k=1}^d \ln \sigma_{z,k}^2(\mathbf{x}) - \frac{1}{2} \mathbb{E}_{q_\phi} \left[\sum_{k=1}^d \frac{(z_k - \mu_{z,k}(\mathbf{x}))^2}{\sigma_{z,k}(\mathbf{x})^2} \right] \\ &= -\frac{d}{2} \ln(2\pi) - \frac{1}{2} \sum_{k=1}^d \ln \sigma_{z,k}^2(\mathbf{x}) - \frac{1}{2} \sum_{k=1}^d \mathbb{E}_{q_\phi} \left[\frac{(z_k - \mu_{z,k}(\mathbf{x}))^2}{\sigma_{z,k}(\mathbf{x})^2} \right] \\ &= -\frac{d}{2} \ln(2\pi) - \frac{1}{2} \sum_{k=1}^d (1 + \ln \sigma_{z,k}^2(\mathbf{x})), \end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}_{q_\phi}(\ln p_\theta(\mathbf{z})) &= \mathbb{E}_{q_\phi} \left[-\frac{d}{2} \ln(2\pi) - \frac{1}{2} \mathbf{z}^\top \mathbf{z} \right] \\
&= -\frac{d}{2} \ln(2\pi) - \frac{1}{2} \sum_{k=1}^d \mathbb{E}_{q_\phi} [z_k^2] \\
&= -\frac{d}{2} \ln(2\pi) - \frac{1}{2} \sum_{k=1}^d \left(\sigma_{z,k}(\mathbf{x})^2 + \mu_{z,k}(\mathbf{x})^2 \right).
\end{aligned}$$

□

2.B Algorithms

Algorithm 6: Quantile transformation of a continuous feature to a standardized Normal distribution.

Input: Observed values $\{x_1, x_2, \dots, x_n\}$ of a continuous variable X

Initialisation

Sort the observations in ascending order: $x_{(1)} \leq \dots \leq x_{(n)}$

Choose the number of quantiles n_q

for $j = 1$ **to** n_q **do**

Define evenly spaced probability levels: $u_j \leftarrow \frac{j-1}{n_q-1}$

Compute empirical index: $k_j \leftarrow u_j(n-1)$

$\ell_j \leftarrow \lfloor k_j \rfloor$

$\delta_j \leftarrow k_j - \ell_j$

Define empirical quantile by linear interpolation:

$q_j \leftarrow (1 - \delta_j)x_{(\ell_j+1)} + \delta_j x_{(\ell_j+2)}$

for $i = 1$ **to** n **do**

Find j such that $q_j \leq x_i \leq q_{j+1}$

$u_i \leftarrow u_j + \frac{x_i - q_j}{q_{j+1} - q_j} (u_{j+1} - u_j)$

$u_i \leftarrow \min(1, \max(0, u_i))$ (clipping)

$z_i \leftarrow \Phi^{-1}(u_i)$ where Φ^{-1} denotes the quantile function of $\mathcal{N}(0, 1)$

Set $T(X) = \frac{z - \bar{z}}{\sigma_z}$, with $\bar{z} = \frac{1}{n} \sum_{i=1}^n z_i$, $\sigma_z = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (z_i - \bar{z})^2}$

Output: Transformed variable $T(X)$ with approximately standard Normal distribution

Algorithm 7: Variational AutoEncoder for a heterogeneous insurance portfolio.

Input: Insurance portfolio $\mathcal{X} \in \mathbb{R}^{n \times m}$ consisting of n policies with m_{num} continuous and m_{cat} categorical rating factors: $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$

Pre-processing

for $j = 1, \dots, m_{\text{cat}}$ **do**
 Dummy-encode categorical variable j :

$$\mathbf{x}^{(j)} \leftarrow \{x_1^{(j)}, \dots, x_{c_j}^{(j)}\}, \quad x_{j'}^{(j)} \in \{0, 1\}.$$

for $k = 1, \dots, m_{\text{num}}$ **do**
 Apply quantile transformation and min-max scaling to numeric variable j :

$$X'_j \leftarrow \Phi^{-1}(F(X_j)), \quad X''_j \leftarrow 2 \frac{X'_j - \min(X'_j)}{\max(X'_j) - \min(X'_j)} - 1.$$

Initialisation

Set hyperparameters: learning rate α , batch size M , epochs N
 Initialise encoder and decoder parameters (ϕ, θ)

while *parameters (θ, ϕ) not converged* **do**

Sample a mini-batch $\mathcal{M} = \{\mathbf{x}_1, \dots, \mathbf{x}_M\} \in \mathcal{X}$;
for $\mathbf{x} \in \mathcal{M}$ **do**
 Compute encoder outputs: $(\boldsymbol{\mu}, \log \boldsymbol{\sigma}^2) \leftarrow e_\phi(\mathbf{x})$
for $l = 1$ **to** L **do**
 Sample latent variables: $\mathbf{z}^{(l)} \leftarrow \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}^{(l)}, \quad \boldsymbol{\epsilon}^{(l)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
 Compute numeric reconstruction loss:

$$\mathcal{L}^{\text{num}} \leftarrow \frac{1}{L} \sum_{l=1}^L \left\| \mathbf{x}_{\text{num}} - \tanh(d_\theta(\mathbf{z}^{(l)})_{\text{num}}) \right\|^2$$

Compute categorical reconstruction loss:

$$\mathcal{L}^{\text{cat}} \leftarrow -\frac{1}{L} \sum_{l=1}^L \sum_{j=1}^{m_{\text{cat}}} \sum_{j'=1}^{c_j} x_{j'}^{(j)} \log \text{softmax}(d_\theta(\mathbf{z}_{c_j}^{(l)})_{j'})$$

Compute KL divergence: $\mathcal{L}_{\text{KL}} \leftarrow -\frac{1}{2} \sum_{d=1}^D (1 + \log \sigma_d^2 - \sigma_d^2 - \mu_d^2)$

Compute gradients: $\nabla_{\theta, \phi} \tilde{\mathcal{L}}^M(\mathcal{M}; \theta, \phi)$

Update parameters:

$$\begin{aligned} \theta &\leftarrow \theta - \alpha \nabla_\theta \sum_{\mathbf{x} \in \mathcal{M}} (\mathcal{L}^{\text{num}} + 2\mathcal{L}^{\text{cat}}) \\ \phi &\leftarrow \phi - \alpha \nabla_\phi \sum_{\mathbf{x} \in \mathcal{M}} (\mathcal{L}^{\text{num}} + 2\mathcal{L}^{\text{cat}} + \beta \mathcal{L}_{\text{KL}}) \end{aligned}$$

Output: Trained encoder and decoder parameters (ϕ, θ) .

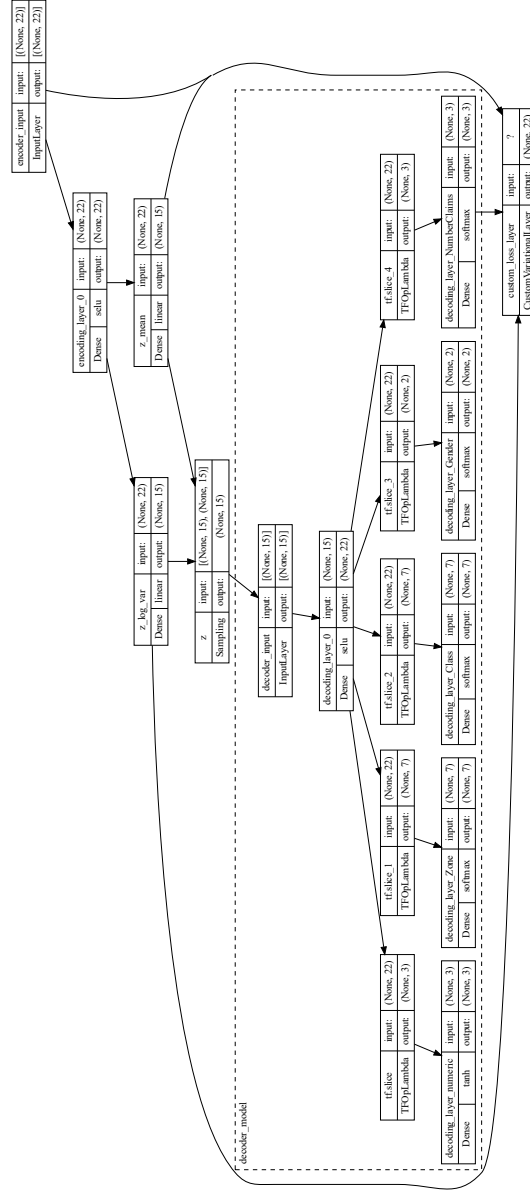


Figure 2.B.1: Model architecture.

2.C Figures

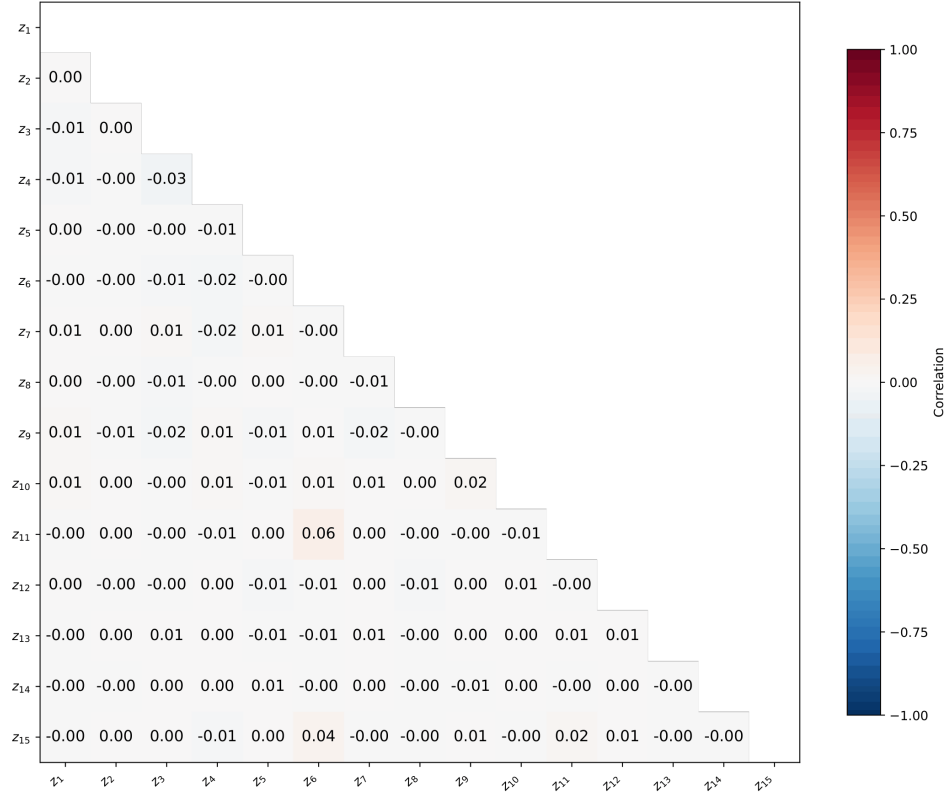


Figure 2.C.1: Pearson's correlation between the latent codes $(Z_j)_{j=1,\dots,d}$.

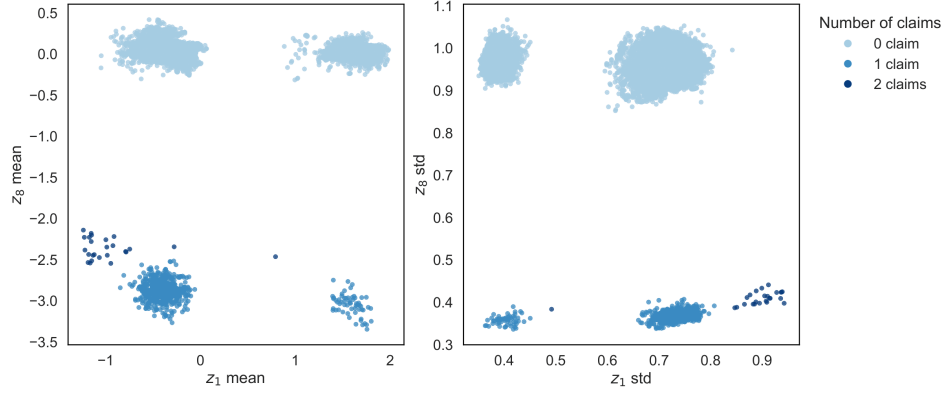


Figure 2.C.2: Scatter plot of the latent means (left) and standard deviations (right) between the observed values of two latent variables Z_j and $Z_{j'}$.

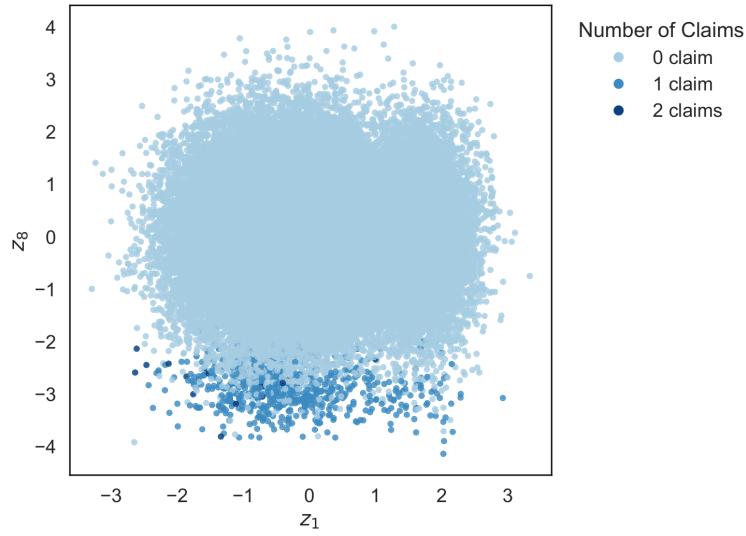


Figure 2.C.3: Scatter plot between the observed values of two latent variables Z_j and $Z_{j'}$.

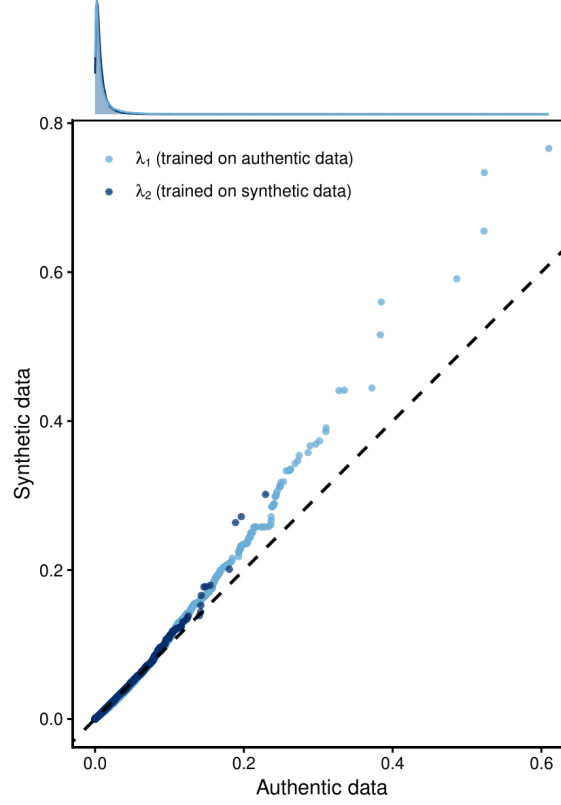


Figure 2.C.4: Proximity to the 1st bisector of the synthetic and authentic predicted claim frequencies, where λ_1 and λ_2 denote the average claim numbers estimators respectively calibrated on the authentic $x_i^{\text{authentic}}$ and synthetic $x_k^{\text{synthetic}}$ data (where i and k denote the i -th and k -th policyholder in the authentic and synthetic portfolio). The light blue dots refer to the pair of points $(\lambda_1(x^{\text{authentic}}), \lambda_1(x^{\text{synthetic}}))$ and the darker blue dots refer to the pair of points $(\lambda_2(x^{\text{authentic}}), \lambda_2(x^{\text{synthetic}}))$.

2.D Tables

	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9	Z_{10}	Z_{11}	Z_{12}	Z_{13}	Z_{14}	Z_{15}
Mean	-0.04	0.00	0.01	0.02	-0.01	-0.00	-0.02	0.01	0.03	-0.01	0.00	-0.00	0.00	-0.00	0.00
Standard deviation	1.00	1.00	0.99	1.00	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 2.D.1: Latent means and standard deviations.

Chapter 3

Latent Dirichlet Allocation for tabular insurance data

This chapter explores the application of Latent Dirichlet Allocation (LDA) to structured tabular insurance data. LDA is a probabilistic approach initially developed in Natural Language Processing (NLP) to uncover the underlying structure of (unstructured) textual data. It was designed to represent textual documents as a mixture of latent topics, and topics as mixtures of words. This study introduces LDA's document-topic distribution as a soft clustering tool for unsupervised learning tasks in actuarial science. By defining each topic as a risk profile, each insurance policy as a document, and each modality of categorical covariates as a word, we extend LDA's application beyond textual data. Our experimental results and analysis highlight how the modelling of policies based on topic cluster membership, and the identification of dominant modalities within each risk profile, can give insights into the prominent risk factors contributing to higher or lower claim frequencies.

Charlotte Jamotton and Donatien Hainaut (Feb. 2026c). “Latent Dirichlet Allocation for Structured Insurance Data”. In: *Annals of Data Science*. ISSN: 2198-5812. DOI: [10.1007/s40745-026-00674-8](https://doi.org/10.1007/s40745-026-00674-8)

3.1 Introduction

Insurance companies manage large portfolios of policies, each described by numerous policyholder features and policy information. These datasets are typically structured in tabular form, where rows represent individual policies and columns correspond to covariates. Tabular insurance data often consist of a heterogeneous

mix of numerical variables (e.g., insured age, policy duration) and categorical variables (e.g., geographic location, gender, vehicle type). This mixture presents several challenges. Numerical variables may require scaling, discretisation, or standardization, depending on the model’s distributional assumptions. Categorical variables also pose unique challenges due to their often qualitative, and high-cardinality nature Avanzi et al. (2024). Missing values further complicate preprocessing. Accurately representing these variables in machine-processable form is critical, as they directly influence risk assessment and pricing decisions.

Current practice in actuarial machine learning often relies on one-hot encoding to pre-process categorical variables. While simple and interpretable, this approach can lead to high-dimensional sparse representations, making models harder to interpret and potentially limiting predictive performance. Embeddings (Richman 2020) and other dimensionality reduction methods have been proposed in actuarial applications as an alternative to encode categorical variables directly into numeric, machine-processable representations (P. Shi et al. 2022). These methods also address challenges of high cardinality by capturing similarities between categories in a lower-dimensional space. For example, Campo et al. (2024) first addressed the challenges of hierarchically structured nominal risk factors by proposing a top-down partitioning algorithm that groups similar categories and generates features for each hierarchical level. Their approach reduces the effective number of categories, and improves differentiation between high and low risk groups. This method has been extended by Richman and Mario V. Wüthrich (2024) using multi-dimensional, multi-level random-effects embeddings, regularized according to the hierarchical structure of the risk factor. Similarly, Wilsens et al. (2024) apply entity embeddings in a hierarchical setting combined with a top-down clustering algorithm to reduce both within-level dimensionality and overall granularity.

In parallel, clustering and segmentation methods have been applied to identify groups of similar policyholders and define homogeneous risk profiles. For instance, Gan and Valdez (2019) applied clustering to telematics-based motor insurance data to uncover latent driving behavior segments, Campbell (1986) proposed an integrated clustering approach combining policy attributes and claims experience for risk differentiation, and Yao (2016) developed a clustering framework to identify natural groupings of policyholders for improved predictive modelling. More recent studies have explored clustering for mixed-type insurance data, including Yin et al. (2021), who applied k-prototypes to life insurance portfolios with numerical, categorical, and spatial variables, and Zhuang et al. (2018), who compared multiple clustering algorithms for auto insurance customer segmentation. However, traditional approaches such as K-means or hierarchical clustering assign each policy to a single cluster, failing to model uncertainty in cluster membership and ignoring the overlapping nature of policyholders’ risk profiles. Probabilistic and soft clustering approaches have been applied to actuarial problems to address this. For example, nonparametric approaches for actuarial data compression (O’Hagan et al. 2016), topic-based finite-mixture models that jointly cluster textual and loss

information (Hou et al. 2024), and fuzzy C-means variants adapted for fraud detection or semi-supervised segmentation (Koc et al. 2024; Hashemi et al. 2023; Oskouei et al. 2024).

The Latent Dirichlet Allocation (LDA) introduced by Blei et al. (2003) as an extension to Hofmann (1999) probabilistic Latent Semantic Analysis (pLSA), offers a promising framework that accommodates heterogeneous portfolios, variable-length policies, missing values, and soft cluster membership to generate interpretable latent risk profiles. Originally developed for topic modelling, LDA represents each document as a mixture of topics, where each topic is a distribution over words. Under the Bag-of-Words (BoW) assumption, the sequential arrangement of words within a document, as well as the order of documents within the corpus can be disregarded Xuan et al. (2014). According to de Finetti’s theorem, the distribution of any set of exchangeable random variables can be represented as a mixture of distributions of independent and identically distributed random variables. Therefore, assuming exchangeability for documents and words within documents amounts to considering a mixture model for both. The three-level Bayesian LDA generative model (Blei et al. 2003) is built around this exchangeability assumption. By modelling each document as a mixture of latent (hidden) topics and each topic as a mixture of words, LDA attempts to capture the underlying distributional structure of textual documents. This probabilistic, soft clustering framework allows documents to simultaneously belong to multiple topics in varying proportions.

While Yibo Wang et al. (2018) demonstrated that applying LDA to unstructured textual claim descriptions improves fraud detection in automobile insurance, its potential for structured categorical insurance data remains largely unexplored. In their survey on categorical data processing, Hancock et al. (2020) recognised the research potential of applying Yibo Wang et al. (2018) technique to other types of insurance data. The ongoing research to develop more efficient and effective encoding methods for handling categorical data in machine learning lead us to consider the application of the LDA model on categorical and discretised numeric rating factors.

Using the terminology specific to the LDA model, each policy $\mathbf{x}$ characterised by N_d categorical (and discretised numeric) variables can be viewed as a set of words or modalities $\{x_1, x_2, \dots, x_{N_d}\}$. An insurance portfolio (corpus) $\mathcal{X}$ could thus be interpreted as a collection of D policies (documents) $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_D\}$ depicting K different risk profiles (topics) denoted $\{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_K\}$. Instead of modelling topics over words as in text data, LDA would model risk profiles over modalities. A risk profile (topic) would then be defined as a probability distribution over modalities (words). The so-called topic-word β_k and document-topic θ_d distributions could be respectively renamed risk profile-modality and policy-risk profile distributions. The order of policies within a portfolio is inconsequential, and the modalities characterising each policy are also exchangeable (it does not matter if the gender of the insured appears before or after its geographic location). Therefore, the BoW’s exchangeability still holds for both policies (documents) and

modalities (words within a document).

The policy-topic θ_d distribution can be used to summarise and organise insurance policies into a smaller set of prominent risk profiles (topics), defined by a list of most-likely (or dominant) modalities, e.g., young drivers of high-power vehicles in the riskiest profile. This corresponds to the definition of cluster analysis (Jain et al. 1999). LDA is to textual documents what soft clustering is to numeric observations. In actuarial literature, clustering has been applied to identify groups of policyholders with similar technical attributes (e.g., their car model) and build predictive models (see e.g., Gan and Valdez (2019), Campbell (1986), and Yao (2016)). However, when dealing with insurance policies, the crisp cluster assignments of standard clustering algorithms amounts to classifying each policy into a single group. This approach fails to reflect the policy attributes that may overlap with different groups. Whereas LDA's probability distribution θ_d over K risk profiles, given a policy d , assigns cluster membership weights to each policy topic. This is the definition of fuzzy (or soft) clustering (M.-S. Yang 1993).

Compared to deep learning embeddings, LDA offers several practical advantages for actuarial applications. While neural embeddings (Richman 2020) can capture complex interactions between categories, they are typically optimized for predictive performance in downstream tasks. LDA, by contrast, provides interpretable topic distributions, requires fewer hyperparameters, can be applied directly to categorical and discretised numeric variables without requiring additional encoding or neural network architectures, and naturally supports soft clustering, which aligns with the overlapping risk characteristics often observed in insurance portfolios. LDA is therefore directly suited for policy segmentation.

Moreover, in LDA the number of words (modalities) in each document (policy) can differ. This property allows LDA to be flexible enough to model documents (policies) of varying lengths, while still effectively discovering topics (risk profiles) that are common across the entire corpus (insurance portfolio). This characteristic of LDA's hierarchical model is of interest to model policies containing missing values. While incomplete data analysis has already been addressed in the literature by K. W. Ng et al. (2008), who introduced grouped Dirichlet distributions to analyse incomplete categorical data, or Dinh et al. (2021), who proposed a clustering algorithm to group mixed numerical and categorical data with missing values, to name a few, the management of missing values in actuarial applications is not strongly developed.

In this chapter, we contribute to the actuarial literature by (i) extending LDA to structured categorical (instead of unstructured textual) insurance data, (ii) introducing the LDA document-topic distribution as an unsupervised soft clustering tool for policy segmentation, and (iii) demonstrating its empirical performance on realistic insurance portfolio scenarios. To the best of our knowledge, it is the first application of LDA to structured, non-textual, insurance data.

The chapter is organised as follows. Section 3.2 introduces the LDA model as a soft clustering tool for structured insurance data. Section 3.2.1 sets the notations.

Section 3.2.2 and Section 3.2.3 detail the generative process and the variational inference method (Hoffman et al. 2010) used to derive the model’s approximate posterior. Section 3.3 presents our experiment settings and results and compares our approach with the Burt-adapted fuzzy clustering (Jamotton, Hainaut, and Hames 2024). Section 3.4 concludes with insights and potential extensions.

3.2 Methods

Fig. 3.2.1 illustrates the workflow of the LDA modelling framework. First, the insurance portfolio $\mathcal{X}$ is represented as a collection of policies, each expressed as a sequence of modalities. Second, LDA is applied to uncover latent risk profiles (topics) explaining the distribution of observed modalities. Finally, the posterior distributions of the latent variables are approximated using variational Bayes (VB) optimisation, yielding interpretable per-policy risk profile (topic) mixtures. The framework produces a soft clustering of policies in a low-dimensional latent risk space, which can be used for predictive premium modelling.

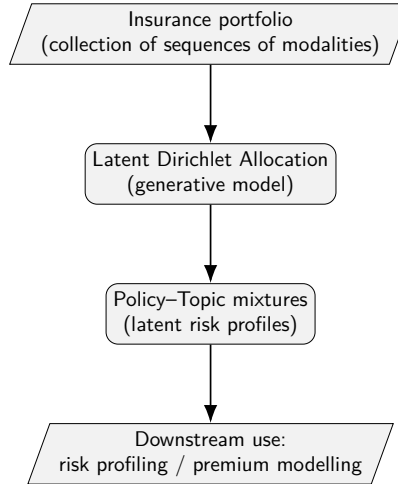


Figure 3.2.1: Overview of the proposed modelling framework. The insurance portfolio is represented as categorical policy data, encoded into modalities, processed through the LDA generative model, and inferred via VB. The resulting topic mixtures define latent risk profiles for downstream actuarial use.

Note that LDA is closely related to classical credibility theory (Bühlmann et al. 2005). Note that LDA is closely related to classical credibility theory (Bühlmann et al. 2005). In both frameworks, policyholders are modelled through latent risk profiles that create partial pooling across observations. In LDA, the policy-specific mixture plays a role analogous to a latent risk profile, with modality occurrences

conditionally independent given this latent structure. Similarly, in credibility models, individual risk parameters are treated as random effects drawn from a population distribution, and posterior estimates combine individual experience and collective information. This shrinkage effect is structurally similar to the Dirichlet-multinomial updating mechanism in LDA, where posterior topic proportions combine observed evidence (data) with prior population-level structure. LDA can be viewed as a multivariate generalisation of credibility-style hierarchical modelling, where dependence comes from shared latent risk components rather than explicit correlation structures.

3.2.1 Notations

Before introducing the Latent Dirichlet Allocation (LDA) framework, we define the main notations and concepts used throughout this chapter. For reference, Table 3.A.1 summarises the key mathematical symbols used throughout the chapter, together with their actuarial interpretation.

Let $\mathcal{X}$ denote a tabular insurance portfolio (corpus) of dimension $D \times C$ consisting of a collection of D insurance policies (documents) and C categorical variables (e.g., policyholder's gender, geographic location, vehicle colour, etc.):

$$\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_D\} .$$

Each policy (document) $\mathbf{x}$ in the insurance portfolio can be described as a finite sequence of N_d words, representing the categorical or discretised numeric rating factors of the policy, often referred to as modalities (e.g., female, west, grey, etc.):

$$\mathbf{x} = \{x_1, \dots, x_{N_d}\} .$$

The insurance portfolio (corpus) can further be characterised by its vocabulary size, $V = \sum_{c=1}^C m_c$, which is equal to the number of unique modalities, m_c , across all C categorical variables.

Each policy $\mathbf{x}$ is modelled as a mixture $\theta^{(1 \times K)}$ over the K latent risk profiles (topics), each capturing a co-occurrence pattern of modalities across policies.

We denote x_{dn} the n th modality of the d th policy. θ_d denotes the topic distribution (or mixture of risk profiles) for policy d . Similarly, z_{dn} is a latent variable indicating the risk profile (topic) assigned to the n th modality of the d th policy. Each risk profile $k = 1, \dots, K$ will be described by a mixture $\beta_k^{(1 \times V)}$ over all V modalities (words). In other words, $\beta_k = (\beta_{k1}, \dots, \beta_{k|V|})$ denotes the multinomial distribution of modalities under risk profile k , where $\sum_v \beta_{kv} = 1$. By doing so, LDA estimates which latent topics (risk profiles) are important for a particular policy, and which modalities (words) are important for a particular risk profile (topic).

Under these notations, the goal of the LDA model is to estimate the posterior distributions of the latent variables $\mathbf{z}$ and $\boldsymbol{\theta}$ given the observed modalities and

hyperparameters (α, η) . α and η are the Dirichlet hyperparameters controlling, respectively, the concentration of θ_d (policy–topic mixtures) and β_k (topic–modality distributions). The next Section formalizes the corresponding generative process.

3.2.2 Generative process

The LDA generative process, illustrated in Fig. 3.2.2, works as follows. For K pre-specified risk profiles (topics), LDA assumes the d th policy $\mathbf{x}_d \in \mathcal{X}$ with length N_d is generated in the following steps (Yi Wang 2011; Carpenter 2010; Darling 2011):

1. For each risk profile $k = 1, \dots, K$, $\beta_k \sim \mathcal{D}_V(\eta)$
2. For each policy $d = 1, \dots, D$
 - (a) $\theta_d \sim \mathcal{D}_K(\alpha)$
 - (b) for each modality x_{dn}
 - i. $z_{dn} \sim \mathcal{M}_K(1, \theta_d)$
 - ii. $x_{dn} | z_{dn} \sim \mathcal{M}_V(1, \beta_{z_{dn}})$

In the first step, a risk profile-modality (topic-word) distribution β_k is drawn for each risk profile k over the vocabulary V from a V -dimensional Dirichlet distribution $\mathcal{D}_V(\eta)$ with hyperparameter η . Each element $\beta_{k,n}$ in β_k indicates a probability of a modality n for this topic k . The higher the η value, the more likely each risk profile (topic) is to contain a mixture of most of the modalities and not a single modality specifically. Note that a slight drawback of LDA is that we have to specify upfront the number of topics to group the set of documents into. Setting a small number of topics K typically leads to fairly general risk profiles that are mixtures of more specific sub-risk profiles (Krestel et al. 2009). Fukuyama et al. (2022) addressed this by introducing a topic alignment method to find an optimal number of topics.

In the second step, the risk profiles proportions θ_d are sampled for each policy d over the available K risk profiles from a K -dimensional Dirichlet distribution $\mathcal{D}_K(\alpha)$ with corpus-level hyperparameter α . Each element θ_{dk} in θ_d indicates a probability of a risk profile k for this policy d . The higher the α value, the more likely each policy is to contain a mixture of most of the risk profiles and not a single risk profile specifically. This policy-specific risk profile matrix $\theta = (\theta_1, \dots, \theta_D) \in \mathbb{R}^{D \times K}$ can be further interpreted as a low-dimensional representation of each policy in the risk profiles space, and ultimately as a soft clustering tool.

A latent topic (risk profile) assignment $z_{dn} \in \{1, \dots, K\}$ is then sampled for each modality x_{dn} , $n = 1, \dots, N_d$ in the d th policy from a K -dimensional Multinomial distribution $\mathcal{M}_K(1, \theta_d)$ over the K topics parameterised by the topics weights θ_d .

Finally, the observed modality x_{dn} is sampled given the sampled risk profile z_{dn} from a V -dimensional Multinomial distribution $\mathcal{M}_V(1, \beta_{z_{dn}})$ parameterised by

$\beta_{z_{dn}}$, following the standard LDA generative process. In insurance data, modalities within a categorical variable are mutually exclusive. Each modality x_{dn} should then be sampled from a subset $V^{(n)} \subset V$ containing only the unique modalities of the corresponding categorical variable. Formally, this can be written as:

$$x_{dn}|z_{dn} \sim \mathcal{M}_{V^{(n)}} \left(1, \beta_{z_{dn}}^{V^{(n)}} \right). \quad (3.2.1)$$

For example, if the first categorical variable, $n = 1$, is the geographic location (West, East, or Central) of the policyholder, $V^{(1)}$ should only contain the identifiers of the modalities West, East, and Central. While Eq. (3.2.1) reflects the practical restriction of sampling for insurance data, LDA is not used in this work as a generative model for data synthesis but rather as a latent representation framework to encode categorical insurance variables and derive risk profiles. Consequently, we never actually sample x_{dn} from the model. Therefore, all subsequent joint and marginal distribution formulas are defined over the full vocabulary V , ensuring formal consistency with the standard LDA framework.

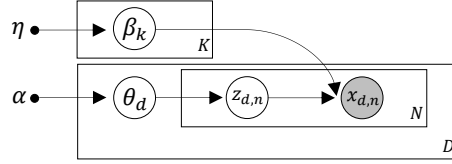


Figure 3.2.2: Representation of the smoothed LDA generative process. The top rectangle defines the topic-level parameters, the inner rectangle defines the modality-level parameters, and the outer rectangle defines the policy-level parameters. We can only observe x (filled with grey).

The LDA generative process described and illustrated above results in the following joint distribution (Yi Wang 2011; Carpenter 2010; Darling 2011):

$$p(x, z, \theta, \beta | \alpha, \eta) = \prod_{k=1}^K p(\beta_k | \eta) \prod_{d=1}^D p(\theta_d | \alpha) \prod_{n=1}^{N_d} p(x_{dn} | \theta_d, \beta), \quad (3.2.2)$$

where the modalities x are the only observable data and z, θ, β are the latent (hidden) variables. When marginalising over topic assignments z , the modalities (words) distribution is given by a mixture of weights $p(z_{dn} | \theta_d)$ and components $p(x_{dn} | z_{dn}, \beta)$:

$$p(x_{dn} | \theta_d, \beta) = \sum_z p(x_{dn} | z_{dn}, \beta) p(z_{dn} | \theta_d),$$

such that the joint distribution in Eq. (3.2.2) becomes:

$$p(x, z, \theta, \beta | \alpha, \eta) = \prod_{k=1}^K p(\beta_k | \eta) \prod_{d=1}^D p(\theta_d | \alpha) \prod_{n=1}^{N_d} p(x_{dn} | z_{dn}, \beta) p(z_{dn} | \theta_d), \quad (3.2.3)$$

where β_k and θ_d are sampled using Dirichlet probability distributions (see Eq. (3.B.1) and (3.B.2) in Appendix 3.B for detailed formulas) and z_{dn} and $x_{dn}|z_{dn}$ are sampled using Multinomial probability distributions (see Eq. (3.B.3) and (3.B.4) in Appendix 3.B).

While LDA's goal is to uncover the latent topics z , we can only observe x . Therefore, we need to reverse the generative process described above to learn the posterior distribution $p(z, \theta, \beta|x, \alpha, \eta)$ of the latent variables z_d , θ_d , and β_k given the evidence (i.e., the observed data x) and the Dirichlet hyperparameters α, η . This amounts to solving the following equation (Darling 2011):

$$p(z, \theta, \beta|x, \alpha, \eta) = \frac{p(x, z, \theta, \beta|\alpha, \eta)}{\int_{\theta, \beta} \sum_z p(x, z, \theta, \beta|\alpha, \eta)}. \quad (3.2.4)$$

The numerator of this conditional Eq. (3.2.4) can be expanded given Eq. (3.B.1), (3.B.2), (3.B.3), and (3.B.4) as follows:

$$\begin{aligned} p(x, z, \theta, \beta|\alpha, \eta) &= \prod_{k=1}^K p(\beta_k|\eta) \prod_{d=1}^D p(\theta_d|\alpha) \prod_{n=1}^{N_d} p(x_{dn}|z_{dn}, \beta_{1:K}) p(z_{dn}|\theta_d) \\ &= c \prod_{k=1}^K \left(\prod_{v=1}^V \beta_{kv}^{\eta_v-1} \right) \prod_{d=1}^D \left(\prod_{k=1}^K \theta_{dk}^{\alpha_k-1} \right) \prod_{n=1}^{N_d} \prod_{k=1}^K \prod_{v=1}^V (\beta_{kv} \theta_{dk})^a, \end{aligned}$$

where,

$$c = \frac{\Gamma\left(\sum_{v=1}^V \eta_v\right)}{\prod_{v=1}^V \Gamma(\eta_v)} \frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)},$$

is a constant, and $a = \mathbb{I}\{x_{dn} = v, z_{dn} = k\}$ is an indicator function.

The denominator of the conditional Eq. (3.2.4) is intractable due to the coupling of the latent variables θ and β in the summation over topics in the marginal likelihood:

$$\int_{\beta} \int_{\theta} \sum_z p(x, z, \theta, \beta|\alpha, \eta) = c \int_{\beta} \prod_{v=1}^V \beta_{kv}^{\eta_v-1} \int_{\theta} \prod_{k=1}^K \theta_{dk}^{\alpha_k-1} \prod_{n=1}^{N_d} \prod_{k=1}^K \prod_{v=1}^V (\beta_{kv} \theta_{dk})^a.$$

There are two families of techniques to derive a (tractable) approximate posterior distribution, that are, the Markov Chain Monte Carlo (MCMC) sampling approaches and the VB optimisation approaches (Yi Wang 2011; Carpenter 2010; Darling 2011). Although the choice of a simplified parametric distribution introduces bias, VB is empirically shown to be faster than and as accurate as MCMC whose stochastic nature and iterative sampling process often require longer convergence time (Hoffman et al. 2010). To efficiently fit models to larger collections of documents, Hoffman et al. (2010) developed an online VB algorithm based on online stochastic optimisation. It is implemented in the Scikit-learn library in Python (Pedregosa et al. 2011) and detailed hereafter.

3.2.3 Variational posterior and ELBO

A variational posterior distribution $q(z, \theta, \beta | \phi, \gamma, \lambda)$, illustrated in Fig. 3.2.3, is introduced to approximate the intractable posterior $p(z, \theta, \beta | x, \alpha, \eta)$. Under the mean-field assumption, this joint distribution of latent variables $q(z, \theta, \beta | \phi, \gamma, \lambda)$ can be factorised into three mutually independent marginal distributions:

$$q(z, \theta, \beta | \gamma, \lambda) = \prod_{k=1}^K q(\beta_k | \lambda_k) \prod_{d=1}^D q(\theta_d | \gamma_d) \prod_{n=1}^{N_d} q(z_{dn} | \phi_{dn}), \quad (3.2.5)$$

parameterised by three variational parameters:

- $\vec{\lambda}_k$ of length V for topic $k = 1, \dots, K$;
- non-negative (and does not sum to 1) $\vec{\gamma}_d$ of length K for policy $d = 1, \dots, D$;
- non-negative ϕ of size $D \times V \times K$ with $\sum_{k=1}^K \phi_{dvk} = 1, \forall dv$.

β_k and θ_d are sampled using Dirichlet probability distributions (see Eq. (3.B.5) and (3.B.6) in Appendix 3.B for detailed formulas) and z_{dn} is sampled using a Multinomial probability distribution (see Eq. (3.B.7) in Appendix 3.B).

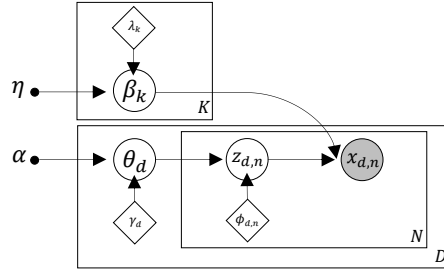


Figure 3.2.3: Representation of the variational distribution to approximate the posterior in the smoothed LDA model. We can only observe x (filled with grey). The variational parameters are framed with lozenge-shapes.

Note that de Finetti's theorem states that an infinitely exchangeable sequence can be represented as a mixture of conditionally i.i.d. random variables. The key implication is that exchangeability does not imply marginal independence. Instead, it guarantees the existence of a latent structure such that observations become independent once conditioned on it. In LDA, this structure is represented through the policy-topic mixture θ_d and the topic-modality distributions β . Each modality x_{dn} is generated by first sampling a latent topic assignment z_{dn} , and then sampling x_{dn} . As a result, the z_{dn} are conditionally i.i.d. given θ_d , and the observed modalities x_{dn} are conditionally i.i.d. given the latent structure (θ_d, β) . Conditional

independence holds in the same spirit as random-effects or credibility models. Any remaining dependence reflects potential model misspecification, meaning that conditional independence should be interpreted as a modeling approximation rather than a literal property of the data-generating process.

The variational parameters are estimated by minimising the Kullback-Leibler (KL) divergence $\text{KL}(q(z, \theta, \beta | \phi, \gamma, \lambda) || p(z, \theta, \beta | x, \alpha, \eta))$ between the variational approximation q and the true posterior distribution p :

$$\min_q \left\{ \mathbb{E}_q \left[\log \frac{q(z, \theta, \beta | \phi, \gamma, \lambda)}{p(z, \theta, \beta | x, \alpha, \eta)} \right] \right\}. \quad (3.2.6)$$

Minimising Eq. (3.2.6) is equivalent to maximising the evidence lower bound (ELBO):

$$\mathcal{L}(\phi, \gamma, \lambda; \alpha, \eta) = \mathbb{E}_q[\ln p(x, z, \theta, \beta | \alpha, \eta)] - \mathbb{E}_q[\ln q(z, \theta, \beta | \phi, \gamma, \lambda)],$$

derived from the marginal log-likelihood $\ln p(x | \alpha, \eta)$ using Jensen's inequality (see Appendix 3.C for detailed derivations).

To maximise the ELBO with respect to the variational parameters ϕ, γ and λ , we first maximise $\mathcal{L}$ with respect to ϕ_{dvk} . This is a constrained maximisation since $\sum_k \phi_{dvk} = 1 \forall dv$. We form the Lagrangian by isolating the terms which contain ϕ_{dvk} and by adding the appropriate Lagrange multipliers:

$$\begin{aligned} \mathcal{L}_{[\phi_{dvk}]} &= \phi_{dvk} (\mathbb{E}_q[\ln \beta_{kv}] + \mathbb{E}_q[\ln \theta_{dk}] - \ln \phi_{dvk}) + \lambda_k^{\text{Lagrange}} \left(\sum_{k=1}^K \phi_{dvk} - 1 \right) \\ \frac{\partial \mathcal{L}}{\partial \phi_{dvk}} &= \mathbb{E}_q[\ln \beta_{kv}] + \mathbb{E}_q[\ln \theta_{dk}] - (\ln \phi_{dvk} + 1) + \lambda_k^{\text{Lagrange}}. \end{aligned}$$

Setting the gradient to zero and solving for ϕ_{dvk} , we get the update rule for ϕ_{dvk} :

$$\phi_{dvk} \propto \exp(\mathbb{E}_q[\ln \beta_{kv}] + \mathbb{E}_q[\ln \theta_{dk}]). \quad (3.2.7)$$

We then maximise the ELBO with respect to γ_{dk} :

$$\begin{aligned} \mathcal{L}_{[\gamma_{dk}]} &= \left(\sum_{v=1}^V n_{dv} \phi_{dvk} + (\alpha - \gamma_{dk}) \right) \left(\Psi(\gamma_{dk}) - \Psi \left(\sum_{k=1}^K \gamma_{dk} \right) \right) + a(\gamma_{dk}) + \text{cst} \\ \frac{\partial \mathcal{L}}{\partial \gamma_{dk}} &= \left(\Psi'(\gamma_{dk}) - \Psi' \left(\sum_{k=1}^K \gamma_{dk} \right) \right) \left(\sum_{v=1}^V n_{dv} \phi_{dvk} + (\alpha - \gamma_{dk}) \right), \end{aligned}$$

where the digamma function $\Psi(\cdot)$ is the derivative of the logarithm of the Gamma function, and is used to compute expectations of logarithms of Dirichlet-distributed variables in closed form.

Setting the gradient to zero and solving for γ_{dk} , we get the update rule for γ_{dk} :

$$\gamma_{dk} = \alpha + \sum_{v=1}^V n_{dv} \phi_{dvk}. \quad (3.2.8)$$

The number of times policy d is assigned to topic k is thus equal to the Dirichlet prior α plus the weighted sum of the number of times each modality in d is assigned to topic k , where the weight ϕ_{dvk} is the probability that modality v in policy d belongs to topic k .

We finally maximise the ELBO with respect to λ_{kv} :

$$\begin{aligned}\mathcal{L}_{[\lambda_{kv}]} &= \left(\sum_{d=1}^D n_{dv} \phi_{dvk} + (\eta - \lambda_{kv}) \right) \left(\Psi(\lambda_{kv}) - \Psi\left(\sum_{v=1}^V \lambda_{kv}\right) \right) + b(\lambda_{kv}) + \text{cst} \\ \frac{\partial \mathcal{L}}{\partial \lambda_{kv}} &= \left(\sum_{d=1}^D n_{dv} \phi_{dvk} + (\eta - \lambda_{kv}) \right) \left(\Psi'(\lambda_{kv}) - \Psi'\left(\sum_{v=1}^V \lambda_{kv}\right) \right).\end{aligned}$$

Setting the gradient to zero and solving for λ_{kv} , we get the update rule for λ_{kv} :

$$\lambda_{kv} = \eta + \sum_{d=1}^D n_{dv} \phi_{dvk}. \quad (3.2.9)$$

Similarly to the interpretation of γ_{dk} , the count of modality v in topic k is equal to the Dirichlet prior η plus the weighted sum of modality count v in each policy d , where the weight ϕ_{dvk} is the probability that modality v in document d belongs to topic k .

Given the updates rules for the variational parameters ϕ, γ and λ (see Eq. (3.2.9), (3.2.8), and (3.2.7)), the online VB EM steps for solving LDA are detailed in Algorithm 8 (Hoffman et al. 2010).

3.3 Results and discussion

3.3.1 Experimental design

We aim to assess the validity, consistency, and interpretability of the proposed LDA-based soft clustering approach for structured insurance data, and to benchmark its performance against both a Generalized Linear Model (GLM) and a Burt adapted fuzzy K-means. The evaluation focuses on two distinct insurance portfolios differing in size, target distribution, and data structure.

We consider a loan default portfolio containing 25 917 policies, and a motor insurance portfolio containing 62 289 policies. Each portfolio includes a combination of categorical, binary, and continuous covariates, as detailed in Tables 3.3.1 and 3.3.1. Continuous variables are discretized into four categories using their empirical quartiles (25th, 50th, and 75th percentiles). This transformation converts continuous features into categorical modalities, which is required by both the Burt matrix representation and the LDA framework, as these methods operate on discrete categorical data. The text descriptions in the loan dataset are not used for LDA training.

For both datasets, the observations are randomly partitioned into an 80% training subset and a 20% test subset (Table 3.3.2). All model fitting is performed exclusively on the training subset, and the test subset is used to evaluate the out-of-sample goodness of fit.

For each portfolio, we train multiple LDA models with the number of topics K varying from 2 to 40. Each model is initialized using symmetric Dirichlet priors on the topic–word and document–topic distributions. The hyperparameters are held constant across runs to ensure comparability. For reference, a GLM is fitted using the same training data, where categorical (and discretised) variables are one-hot encoded. The Burt adapted fuzzy K-means is trained under the same conditions, with K set to match the LDA topic count.

Model performance is evaluated on the test subsets using the deviance of the assumed target distribution, that is, Binomial deviance for the loan portfolio and Poisson deviance for the motor portfolio. Each observation is assigned to the topic (or cluster) of maximum posterior probability $\max(\theta_d)$, from which the predicted frequency $\hat{y}_d$ is derived as the empirical mean of targets within the same cluster. The resulting cluster-level predictions are compared across models to assess fit quality.

To assess interpretability, the top modalities and average numeric values per topic are analyzed and visualized. All results (deviance values, boxplots, and cluster compositions) are reported on the test sets, unless otherwise stated.

3.3.2 Datasets

We use two distinct insurance portfolios to evaluate the validity and consistency of the LDA model across various types of covariates with a variety of marginal distributions and different assumptions regarding the target variable’s distribution. The first dataset is a loan default portfolio with 25 917 policies described by four continuous variables (discretised into four distinct groups based on their 25th, 50th, and 75th percentile), two binary variables, six categorical variables (i.e., structured data) and a claim description provided by the borrower (i.e., unstructured data) detailed in Table 3.3.1. The target variable takes two possible values (Charged Off or Fully Paid) and is assumed to follow a Binomial distribution.

Continuous variables	Min	Max	Mean	Std
Loan amount	500	35 000	11 369.9	7 274.1
Interest rate	5.42	24.4	11.95	3.60
Annual income	400	3 900 000	68 854	59 173
Revolving utilization	0	99.9	48.49	28.22

Table 3.3.1: Description of the loan portfolio variables with summary statistics.

Categorical variables	Modalities (proportion)
Inquiries	0 (48%), 1 (27%), 2 (15%), ≥ 3 (10%)
Region	East (51%), West (29%), Central (20%)
Purpose	Debt consolidation (47%), Other (19%), Credit card (13%), Home improvement (7%), Major purchase (5%), Small business (5%)
Grade	1–10 (56%), 11–20 (34%), 21–35 (10%)
Home ownership	Rent (48%), Mortgage (44%), Own (8%)
Verification status	Not verified (45%), Verified (32%), Source verified (23%)
Term	36 months (76%), 60 months (24%)
Public records	0 (95%), 1–3 (5%)
Loan status (target)	Fully Paid (85%), Charged Off (15%)

Table 3.3.1: Description of the loan portfolio variables with summary statistics (continued).

To validate the results obtained with this first database and check their reproducibility, we use a second insurance portfolio that contains information about motorcycle insurance policies over the period 1994–1998 (Ohlsson et al. 2010). Table 3.3.1 gives the description and basic summary statistics of the two quantitative variables and the three categorical variables describing each policy. The target variable (number of claims) is assumed to follow a Poisson distribution. Given the exposure of each policy, we will predict its claim frequency (i.e., the ratio between the number of claims and the contract duration). In the following analyses, the policyholder’s age and its vehicle’s age are categorised into respectively nine (≤ 19 , $20 - 29$, $30 - 39$, $\dots$, $80 - 89$, ≥ 90) and four ($0 - 4$, $5 - 10$, $11 - 20$, ≥ 21) distinct groups. This motorcycle portfolio contains 62 289 insurance policies, which is double the number of observations in the loan dataset.

Categorical variables	Modalities (proportion)
Gender	M (85%), K (15%)
Geographic area	1 (13%), 2 (18%), 3 (20%), 4 (39%), 5 (4%), 6 (6%), 7 (1%)
Vehicle power	-5–1 (11%), 6–8 (8%), 9–12 (30%), 13–15 (19%), 16–19 (18%), 20–24 (13%), 25+ (1%)
Continuous variables	Min Max Mean Std
Owner age	16 92 42.54 12.95
Vehicle age	0 99 12.56 9.68

Table 3.3.1: Description of the motorcycle insurance portfolio attributes with summary statistics (Ohlsson et al. 2010).

Target variables				
Claim frequency	0	182.51	0.028	0.947
Exposure	0.002	11.99	1.008	1.06
Number of claims	0 (99.89%), 1 (0.101%), 2 (0.0004%)			

Table 3.3.1: Description of the motorcycle insurance portfolio attributes with summary statistics (Ohlsson et al. 2010) (continued).

3.3.3 LDA soft clustering for structured data

As explained in Section 3.2, by considering each distinct value of the categorical (and discretised numeric) rating factors as a word, we can extend the scope of application of the LDA model to structured data and use the document-topic matrix $\theta^{(D \times K)}$ as a soft clustering tool where $k = 1, \dots, K$ are the topic clusters and $d = 1, \dots, D$ are the observations we want to cluster.

For each insurance portfolio introduced above, a LDA model is trained using 80% of the dataset. The model's goodness of fit is then evaluated on the remaining 20% of the dataset. This train-test split, displayed in Table 3.3.2, is essential for assessing how well the model generalises to unseen data. Given the learned parameters from the trained model (i.e., the topic-word and policy-topic distributions), we can infer the risk profiles distributions $\theta_{d^{new}}$ of these unseen (new) insurance policies (in the test set) using the optimised posterior distribution $q(\theta)$.

	Train set (80%)	Test set (20%)
Loan	20 733	5 184
Motor	49 831	12 458

Table 3.3.2: Number of observations (insurance policies) in the training-testing subsets.

To evaluate the goodness of fit of the models, each policy is assigned to the topic for which it has the highest relative probability $\max(\theta_d)$. The quality of the topic clustering can then be evaluated with the Binomial deviance for the loan dataset and with the Poisson deviance for the motor dataset. The Binomial deviance is defined as:

$$\mathcal{D}_{\text{Binomial}}(y, \hat{y}) = 2 \sum_{d=1}^D \left(y_d \ln \left(\frac{y_d}{\hat{y}_d} \right) + (1 - y_d) \ln \left(\frac{1 - y_d}{1 - \hat{y}_d} \right) \right), \quad (3.3.1)$$

where y_d is the observed loan default probability of the d th policy, and $\hat{y}_d$ is the predicted loan default probability of the d th policy (which is equal to the average

number of defaults of all policies assigned to the same (topic) cluster as the d th policy). The Poisson deviance is defined as:

$$\mathcal{D}_{\text{Poisson}}(y, \hat{y}) = 2 \sum_{d=1}^D N_d \left(\frac{\nu_d}{N_d} \hat{\lambda}_d - \left(\ln \left(\frac{\nu_d}{N_d} \hat{\lambda}_d \right) + 1 \right) \mathbb{I}_{\{N_d \geq 1\}} \right), \quad (3.3.2)$$

where N_d and ν_d are respectively the number of claims and the exposure of the d th policy, and $\hat{\lambda}_d$ is the predicted claim frequency (which is equal to the average number of claim frequencies of all policies assigned to the same (topic) cluster as the d th policy).

For each portfolio, we trained around 40 different LDA models (by increasing the number of topics K) on the structured covariates in the respective training subset (80%) of each portfolio. Fig. 3.3.1 shows the evolution of the Binomial (left plot) and Poisson (right plot) deviances with respect to the number of topics $K \in (2, 40)$ used to fit these LDA models. The orange lines represent the evolution of the deviances on the train sets, whereas the green lines represent the evolution of the deviances on the test sets. We observe a general (but not smooth) decrease in the deviances when the number of topics K increases, especially for the motor portfolio (see right plot of Fig. 3.3.1) which contains more observations (see Table 3.3.2). In the following analyses, all results (including goodness of fit or deviance values) are evaluated on the test sets, unless otherwise specified.

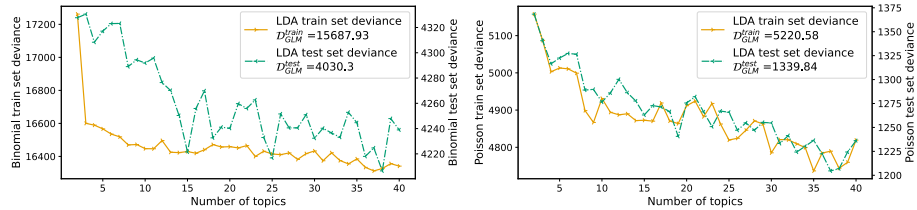


Figure 3.3.1: Evolution of the LDA Binomial (left plot) and Poisson (right plot) deviances with respect to the number of LDA topics $K \in (2, 40)$.

If we chose to train our LDA model with $K = 15$ topics on a subset (80%) of the loan portfolio (in Table 3.3.1), we get a Binomial deviance equal to 4221.52 (with 15 distinct $\hat{y}_d$). In comparison, a GLM trained on the same subset (80%) of the loan portfolio (with the difference that the discretised and categorical covariates are one-hot encoded) gives a lower (and better) deviance equal to 4030.3. The granularity of the GLM (4809 distinct $\hat{y}_d$) is however much higher than LDA clustering (15 distinct $\hat{y}_d$). Note that a separate validation set or cross-validation procedure would typically be used to tune K . However, in this structure discovery application, K was chosen as a compromise to be sufficiently large to capture meaningful structure in the data, while remaining small enough to ensure a concise analysis of cluster

composition. Fig. 3.3.1 illustrates how the low number of observations in the test set of the loan portfolio leads to non-negligible variability in the predictive accuracy out-of-sample.

While the use of LDA’s document-topic distribution as an unsupervised soft clustering tool is not as competitive in terms of deviance as a standard GLM on our loan portfolio, we still observe in Table 3.3.3 that the LDA approach gives some insights into the prominent risk factors contributing to higher or lower claim frequencies. This table 3.3.3 gives a summary of some of the most common modalities (in columns titled **term** and **sub_grade**) and average numeric values (in columns **int_rate** and **revol_util**) in each topic cluster. The topic clusters (as well as the corresponding figures) are ordered by increasing (average) number of loan defaults (see **Frequency** column). The categorical variables **home_ownership**, **verification_status**, **purpose**, **region**, **inq_last_6mths**, and **pub_rec** are omitted from Table 3.3.3 as none of their respective categories (modalities) has a consistent association with higher or lower loan default rates (see Fig. 3.D.1 in Appendix). The continuous variables **loan_amnt** and **annual_inc** are also omitted as no increasing (or decreasing) trend with the default rate is observed (see Fig. 3.D.2 in Appendix).

The column titled **% of policies** displays the size of each cluster (i.e., number of policies relative to the portfolio size). We observe that nearly 50% of the total observations are in the first three clusters with the lowest number of defaults. Insurance claims are often considered as rare events due to their low probability of occurrence. In the context of clustering, the data may naturally exhibit an imbalanced distribution, where the majority of policies are assigned to lower-risk clusters with a lower likelihood of defaults.

Topic	% of policies	Term	Sub-grade	Interest rate	Revolving utilization	Frequency (%)
1	0.77	36 months	1-10	6.86	37.0	0.0500
2	37.21	36 months	1-10	8.55	36.61	0.1032
3	11.94	36 months	1-10	9.53	45.77	0.1066
4	2.37	36 months	1-10	9.64	30.24	0.1138
5	4.53	36 months	1-10	9.61	46.26	0.1404
6	2.22	60 months	1-10	10.53	43.29	0.1565
7	3.49	36 months	11-20	11.68	55.09	0.1602
8	10.07	36 months	1-10, 11-20	12.78	67.04	0.1686
9	1.58	60 months	11-20	11.86	51.17	0.1829
10	2.72	36 months	21-35	15.04	62.26	0.1844
11	7.99	36 months	1-10, 11-20	12.34	57.26	0.1932
12	9.43	36 months	11-20	12.54	51.81	0.2331
13	0.33	36 months	11-20	15.17	63.44	0.2353
14	1.60	60 months	11-20, 21-35	15.22	61.08	0.2530
15	3.74	60 months	21-35	15.33	72.80	0.2835

Table 3.3.3: Some of the most common modalities and average numeric values in each topic cluster obtained with the LDA soft clustering model on the test set (20%) of the loan portfolio. The table is ordered by increasing (average) number of loan defaults.

We observe in Table 3.3.3 and in the left plot of Fig. 3.3.2 that the two topic clusters with the highest number of defaults have the highest number of long-term (60 months) loans. Moreover, the topic cluster characterised by the lowest number of loan defaults exclusively consists of short-term loans. We also observe in the right plot of Fig. 3.3.2 a decline in the number of policies characterised by lower grades (1–10) and an increase in the number of policies characterised by higher grades (11–20, 21–35), when the default rate increases.



Figure 3.3.2: Topic cluster distribution in the categorical variables `term` and `sub_grade`. The topic clusters are ordered by increasing (average) number of loan defaults, as in Table 3.3.3.

With regard to numeric covariates, we observe in the left plot of Fig. 3.3.3 an upward trend where higher interest rates are associated with topic clusters characterised by higher default rates. This increase with the probability of loan default is also observable, but less pronounced, in the right plot of Fig. 3.3.3 for the revolving line utilisation rate.

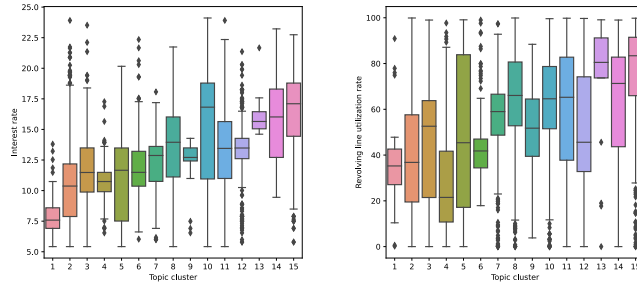


Figure 3.3.3: Boxplot of the numeric variables `int_rate` and `revol_util` by topic cluster. The topic clusters are ordered by increasing (average) number of loan defaults, as in Table 3.3.3.

To confirm LDA's ability to model the underlying structure of structured insurance data, we train our LDA model on a subset (80%) of the motor portfolio introduced above (in Table 3.3.1). To strike a balance between capturing the underlying data structure and maintaining a manageable number of clusters for analysis, we chose to set the number of topic clusters K to 10. With $K = 10$, the Poisson deviance is equal to 1 276.51 (with 10 distinct $\hat{y}_d$). In comparison, a GLM trained on the same subset (80%) of the motor portfolio (with the difference that the discretised and categorical covariates are one-hot encoded) gives a deviance equal to 1 339.84 (with more than 1 269 distinct $\hat{y}_d$). With a quite lower level of granularity (10 versus 1 269 distinct $\hat{y}_d$), the LDA clustering model results in a significantly lower deviance than the one obtained with a GLM, which was not the case in our loan portfolio analysis. This suggests that LDA is better at accurately capturing the underlying structures of larger insurance portfolios. The following analysis of the topic clusters composition, summarised in Table 3.3.4, further highlights LDA's ability to give insights into the prominent risk factors contributing to higher or lower claim frequencies. This Table 3.3.4 gives a summary of some of the most common modalities (in columns titled **Zone**, **Class** and **Gender**) and average numeric values (in columns **OwnersAge** and **VehicleAge**) in each topic cluster. The topic clusters (as well as the corresponding figures) are ordered by increasing (average) motor claim frequencies (see **Frequency** column).

Topic	% of policies	Zone	Class	Gender	OwnersAge	VehicleAge	Frequency (%)
1	7.53	4	1	M	48.17	32.19	0.13
2	14.44	4	3, 4	M	46.20	14.50	0.29
3	6.01	4, 1	3	K	45.00	6.62	0.52
4	15.61	4	6, 3	M	56.46	14.00	0.61
5	8.35	4, 2	4	M	46.50	13.00	0.79
6	8.01	4, 1	3, 4	M	32.80	15.00	1.20
7	13.42	2, 3	3	M	50.79	2.00	1.38
8	7.71	3, 6	5	M	47.18	15.00	1.54
9	10.05	4, 3	6, 3	M	35.69	6.00	2.03
10	8.87	4, 2	5, 4	M	26.75	8.00	4.89

Table 3.3.4: Most common modalities and average numeric values with LDA soft clustering (motor portfolio). The table is ordered by increasing (average) motor claim frequencies.

Cluster 3, with the third lowest number of claims, has an interesting gender make-up. While all other clusters have a male dominating gender, cluster 3 has the largest proportions of females (see right plot of Fig. 3.3.4). It is also the cluster with the largest number of insureds located in Sweden's three largest cities (see the proportion of **Zone=1** in the left plot of Fig. 3.3.4).

The left plot of Fig. 3.3.5 also highlights the differences in age distribution across topic clusters. It is interesting to note that the topic cluster associated with

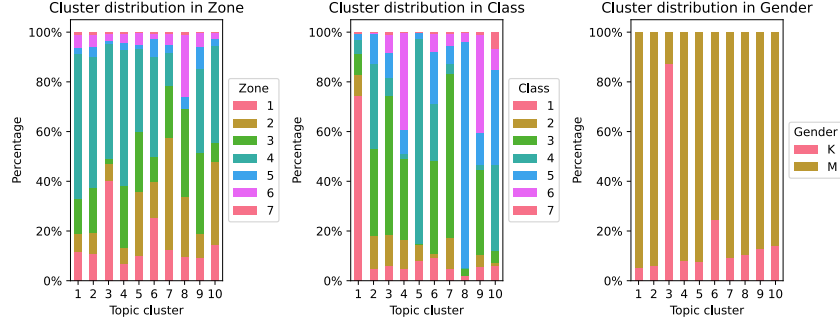


Figure 3.3.4: Topic cluster distribution in the categorical variables **Zone**, **Class** and **Gender**. The topic clusters are ordered by increasing (average) motor claim frequencies, as in Table 3.3.4.

the highest claim frequency is exclusively characterised by younger (less experienced) policyholders (26–27 years old), and is amongst the most concentrated age distribution. Insurance premiums for less experienced drivers are typically higher due to their perceived higher risk profile. We also observe in the right plot of Fig. 3.3.5 that the topic cluster associated with the lowest claim frequency consists of policies with the oldest vehicle’s ages (can refer to collection vehicle) with a very large majority of low EV ratio ≤ 5 (see middle plot of Fig. 3.3.4). Collection vehicles might be driven less frequently, reducing the likelihood of accidents.

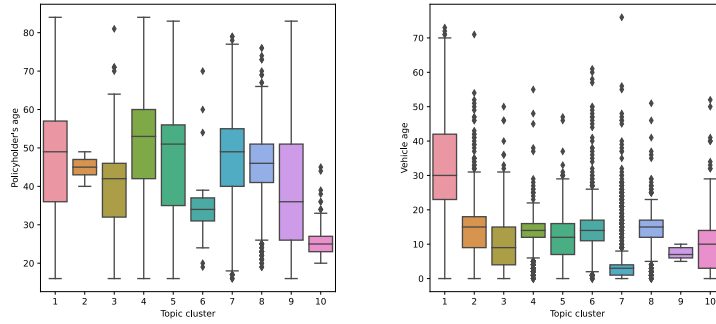


Figure 3.3.5: Boxplot of the numeric variables **OwnersAge** and **VehicleAge** by topic cluster. The topic clusters are ordered by increasing (average) motor claim frequencies, as in Table 3.3.4.

The following Table 3.3.5 provides a summary comparison of the goodness of fit

obtained with our LDA soft clustering approach on both (test sets of) the insurance portfolios in Table 3.3.1 (loan portfolio) and Table 3.3.1 (motor portfolio).

	Loan	Motor
# of topic clusters K	15	10
# of policies in the test set	5 184	12 458
Deviance type	Binomial	Poisson
LDA deviance	4 221.52	1 276.51
GLM deviance	4 030.30	1 339.84

Table 3.3.5: Summary of the LDA goodness of fit results on the test sets.

3.3.4 Burt adapted fuzzy (or soft) K-means

In this Section, we compare our LDA soft clustering approach to the Burt adapted fuzzy (or soft) K-means (Jamotton, Hainaut, and Hames 2024). In this Burt framework, categorical data are mapped to numeric vectors as follow:

$$\mathbf{B}^{\mathbf{W}} = \frac{1}{q} \mathbf{C} \mathbf{B} \mathbf{C}, \quad (3.3.3)$$

where q is equal to the number of categorical (and discretised numeric) variables, $\mathbf{B}$ is equal to the product of the disjunctive matrix $\mathbf{D}$ and its transpose, and the diagonal weight matrix $\mathbf{C}$ contains on its diagonal the number of policies characterised by the i th modality, denoted by n_{ii} . For more details, we refer to Jamotton, Hainaut, and Hames (2024). For a survey of other state-of-the-art mixed data clustering algorithms, see Ahmad et al. (2019). The following Fig. 3.3.6 shows the evolution of the Binomial and Poisson deviances with respect to the number of fuzzy K-means clusters. The decrease in deviance appears smoother with the Burt approach than with LDA for the loan portfolio (see left plot of Fig. 3.3.1), despite the low number of observations for the loan portfolio out-of-sample.

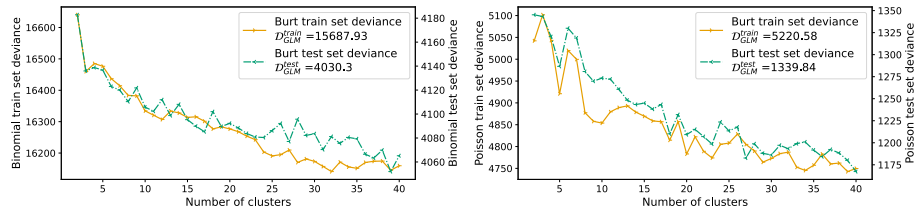


Figure 3.3.6: Evolution of the Binomial (left plot) and Poisson (right plot) deviances with respect to the number of K-means clusters $K \in (2, 40)$ in our Burt adapted fuzzy K-means.

If we train our Burt adapted fuzzy K-means on the same subset (80%) of the loan insurance portfolio (in Table 3.3.1) as in the LDA approach, with a number of fuzzy cluster $K = 15$ (equals to the number of topic clusters in the LDA approach), we get a Binomial deviance (on the test set) equal to 4095.31 (with 15 distinct $\hat{y}_d$). In comparison, the GLM deviance was equal to 4030.3 (with 4809 distinct $\hat{y}_d$) and the LDA deviance was equal to 4221.52 (with 15 distinct $\hat{y}_d$). The following Table 3.3.6 reports some of the most common modalities and average numeric values in each topic cluster.

While the Burt and GLM approaches outperform LDA in terms of deviance in this loan portfolio analysis, we observe that the Burt and LDA models give similar insights into the prominent risk factors contributing to higher or lower claim frequencies. Similarly to LDA, the left plot of Fig. 3.3.7 highlights the majority (minority) of long-term loans in the clusters associated with higher (lower) default rates. Unlike what we observed in the LDA approach, the first cluster, associated with the lowest default rate, is not exclusively characterised by long-term loans (but still consists of a large majority of long-term loans). Conversely, Burt appears to discriminate better than LDA between different grades. The column `sub_grade` in Table 3.3.6 displays a consistent evolution of the grade sub-classes with the increase of default rate. This majority of lower (higher) grades in clusters associated with lower (high) default rate is visually confirmed by the right plot of Fig. 3.3.7.

Topic	% of policies	Term	Sub_grade	Int_rate	Revol_util	Frequency (%)
1	7.14	36 months	1-10	6.42	17.85	0.0459
2	6.17	36 months	1-10	6.64	33.28	0.0531
3	8.20	36 months	1-10	7.53	19.30	0.0847
4	4.98	36 months	1-10	9.16	36.56	0.0930
5	5.59	60 months	1-10	9.91	33.14	0.1034
6	8.16	36 months	1-10	10.31	51.08	0.1064
7	6.67	36 months	1-10	9.86	53.32	0.1098
8	7.64	36 months	1-10	9.88	49.55	0.1111
9	8.55	36 months	11-20	13.09	71.45	0.1580
10	7.79	36 months	11-20	12.79	55.82	0.1683
11	6.96	36 months	11-20	13.64	61.82	0.1717
12	6.23	36 months	11-20	12.82	53.21	0.2415
13	5.63	60 months	11-20	15.62	70.47	0.2432
14	6.15	60 months	21-35	17.71	71.01	0.3009
15	4.13	60 months	21-35	16.52	53.79	0.3178

Table 3.3.6: Dominant modalities and average numeric values (from the loan database) of the continuous variables with K-means Burt adapted soft clustering. The table is ordered by increasing (average) number of loan defaults.



Figure 3.3.7: Cluster distribution in the categorical variables `term` and `sub_grade`. The clusters are ordered by increasing (average) number of loan defaults, as in Table 3.3.3.

In Fig. 3.3.8, the upward trend in both the interest rate and revolving line utilisation rate appears to be more consistent than the one observed in the LDA approach in Fig. 3.3.3.

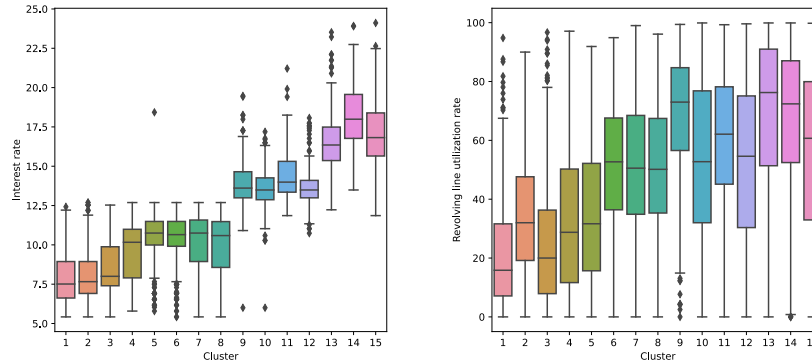


Figure 3.3.8: Boxplot of the numeric variables `int_rate` and `revol_util` by cluster. The clusters are ordered by increasing (average) number of loan defaults, as in Table 3.3.3.

However, note that the size of the different clusters is more or less the same (see column titled `% of policies` in Table 3.3.1). Whereas in the LDA approach, the first three topic clusters already accounted for 50% of the entire portfolio (and reflected the rarity of insurance claims).

To be consistent with the LDA analysis in Section 3.3.3, we also train our

Burt adapted fuzzy K-means model on a subset (80%) of the motor portfolio (in Table 3.3.1), with the same number of fuzzy clusters $K = 10$ as the number of LDA topic clusters. It results in a Poisson deviance equal to 1273.37 (with 10 distinct $\hat{y}_d$), which is quite close to the one obtained with the LDA approach (1276.51 with 10 distinct $\hat{y}_d$) and still better than the one obtained with the GLM (1339.84 with more than 1269 distinct $\hat{y}_d$). While both approaches seem to be able to distinguish high-risk groups from the low-risk groups, the LDA approach produces a higher degree of heterogeneity between the clusters with significantly different claim frequencies. While the average claim frequency of the first four clusters are quite similar in the Burt approach, ranging from 0.42% up to 0.49% (see Table 3.3.7), we observe in the LDA approach in Table 3.3.4 an average claim frequency of 0.13%, 0.29%, 0.52%, and 0.61% in the first four clusters.

Cluster	% of policies	Zone	Class	Gender	OwnersAge	VehicleAge	Frequency
1	9.35	4	1	M	42.94	29.46	0.42
2	14.54	4, 3	3, 4	K	41.53	11.25	0.45
3	7.59	4	3, 4	M	46.20	7.50	0.46
4	13.78	4	5, 3, 4	M	47.00	14.00	0.49
5	7.87	3	5, 3	M	44.94	15.00	0.57
6	6.62	4, 2	3, 4	M	53.50	6.00	1.01
7	15.77	2, 1	3, 5	M	47.40	15.50	1.16
8	7.12	4, 3, 2	3	M	46.24	2.67	1.20
9	9.91	4, 3, 2	4, 5	M	36.50	2.00	3.35
10	7.45	4, 3	6, 5	M	25.50	7.00	4.80

Table 3.3.7: Dominant modalities and average numeric values of the continuous variables (from the motor database) with K-means Burt adapted soft clustering. The table is ordered by increasing (average) motor claim frequencies.

Nevertheless, the dominant categories and average numeric values obtained across the hardened fuzzy clusters in Table 3.3.7 are highly similar to our LDA topic clusters analysis in Table 3.3.4. The evolution of the numeric variables with respect to the clusters average claim frequency in Fig. 3.3.9, as well as the cluster distribution in the categorical variables in Fig. 3.3.10 are also quite similar to that of the LDA approach. For example, the first cluster (associated with the smallest claim frequency) is once again characterised by older vehicles (see right plot of Fig. 3.3.10) with the lowest power class (see middle plot of Fig. 3.3.9). Whereas, the cluster with the highest claim frequency exclusively consists of young (less experienced) drivers (around 25), as illustrated in the left plot of Fig. 3.3.10, and has the largest proportions of higher power vehicles (see middle plot of Fig. 3.3.9). The cluster distribution in the **Gender** variable (see right plot of the following Fig. 3.3.9) is particularly interesting as one cluster (with the second lowest claim frequency on average) contains the majority of the female policyholders in the portfolio, while all other clusters predominantly (if not exclusively) consist of male policyholders.

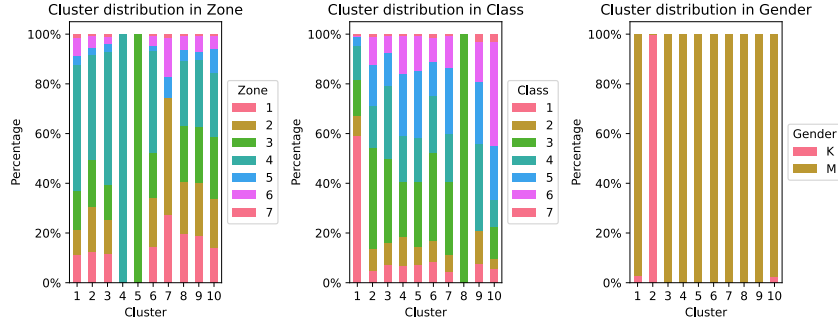


Figure 3.3.9: Cluster distribution in the categorical variables **Zone**, **Class** and **Gender**. The clusters are ordered by increasing (average) motor claim frequencies, as in Table 3.3.7.

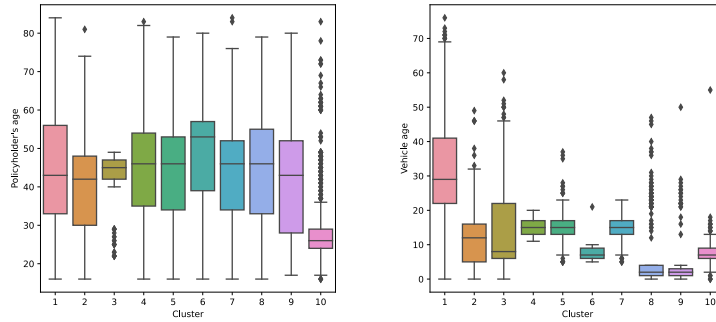


Figure 3.3.10: Boxplot of the numeric variables **OwnersAge** and **VehicleAge** by cluster. The clusters are ordered by increasing (average) motor claim frequencies, as in Table 3.3.7.

The following Table 3.3.8 provides a summary of the goodness of fit obtained in our LDA and Burt soft clustering approaches on both (test sets of) our insurance portfolios.

While the LDA-based soft clustering approach does not systematically outperform classical models such as the GLM or the Burt adapted fuzzy K-means in terms of deviance, its practical relevance resides in the interpretability and dimensional reduction it provides. For instance, in the loan portfolio, the LDA model exhibits a higher Binomial deviance than the GLM. However, it reduces the prediction space to only 15 interpretable topic clusters. This compression facilitates the identification of high and low-risk groups and the visualisation of prominent risk

	Loan	Motor
# of topic clusters K	15	10
# of policies in the test set	5 184	12 458
Deviance	Binomial	Poisson
LDA deviance	4 221.52	1 276.51
Burt deviance	4 095.31	1 273.37
GLM deviance	4 030.30	1 339.84

Table 3.3.8: Summary of the goodness of fit results on the test sets.

factors, providing insights into the latent structure of insurance portfolios. Moreover, while Burt clustering was initially designed for constructing discrete pricing grids, LDA provides a more detailed representation of portfolio structure, focusing on explainability of underlying risk drivers rather than purely segmentation-based pricing performance.

The residual heatmaps Figures 3.D.3 and 3.E.3 in Appendix further illustrate model performance beyond overall deviance. Fig. 3.D.3 corresponds to LDA residuals, and Fig. 3.E.3 to Burt residuals. The marginal histograms (top and right) show similar marginal distributions (e.g., policyholder’s age, and vehicle age), meaning both models are evaluated on the same range of input features. Differences in the residual pattern stem from the model behavior, not from data imbalance. In the visualizations, the left plot corresponds to the loan portfolio and the right plot to the motor portfolio, with the continuous predictors densities highlighting regions with more concentrated data where residuals are smaller and more stable. No systematic patterns in residual $y - \hat{y}$ magnitude are apparent in either portfolio across covariate combinations, indicating that predictions remain well-calibrated across the feature space. Larger residuals are observed mainly in low-density regions, reflecting the reduced number of observations rather than model misspecification. These visual diagnostics support the quantitative deviance comparisons reported in Table 3.3.8.

3.4 Conclusions

In this chapter, we introduced LDA’s document-topic distribution as an unsupervised soft clustering tool and extended LDA’s probabilistic framework, initially developed for unstructured text data, to tabular insurance data. To model (structured) discrete covariates in tabular insurance portfolios, we defined each topic as a risk profile, each insurance policy as a document and each modality of the categorical and discretised numeric rating factors as a word. Much like the original fuzzy K-means, the soft clustering ability of LDA allows to model each policy by a topic cluster membership. Through the identification of dominant modalities

within each topic cluster, insurers can gain insights into risk factors contributing to higher or lower claim frequencies (e.g., default rates).

Our experimental results and analysis showed that the use of LDA’s document-topic distribution as an unsupervised soft clustering tool is effective at organising insurance policies characterised by categorical (and discretised numeric) variables into a smaller set of prominent risk profiles (topics). While it did not show to always be as competitive in terms of deviance, comparative analyses with the Burt adapted soft clustering approach still supports the effectiveness of LDA in capturing the underlying structure within structured insurance data.

Building on the findings of this study, several prospects for future research are apparent. First, hybrid approaches that integrate LDA with embedding-based methods (e.g., word or entity embeddings) could enhance the representation of complex interactions among risk factors, capturing subtle dependencies across modalities that pure LDA may overlook. Such approaches may allow insurers to leverage both probabilistic topic structures and continuous latent representations for more accurate risk profiling. From a predictive modelling perspective, LDA provides a latent simplex embedding θ for each policy, while β defines the semantic basis of that embedding. Downstream regression models could be built either on θ directly or on θ projected through β into a risk-weighted feature space. Finally, the inherent flexibility of LDA in handling varying document lengths suggests potential for addressing heterogeneous datasets with missing values, or combining discrete and continuous variables via hybrid discretisation or embedding techniques. Moreover, dynamic or sequential LDA variants could be applied to better understand how risk profiles develop over time, capturing evolving risk patterns. Collectively, these extensions could strengthen LDA’s utility for predictive modelling and actuarial insights.

3.A Notations

Symbol	Mathematical definition	Actuarial interpretation
$\mathcal{X}$	Set of all policies $\{\mathbf{x}_1, \dots, \mathbf{x}_D\}$	Insurance portfolio
$\mathbf{x}_d$	d th policy with N_d modalities	Individual policy characteristics
x_{dn}	n th modality of policy d	Value of a risk factor (e.g., driver age group)
$\mathcal{V}$	Set of all unique modalities	Collection of all covariate levels
K	Number of latent topics	Number of latent risk profiles
β_k	Topic–modality distribution	Distribution of risk factor combinations in risk profile k
θ_d	Policy–topic mixture	Policy’s mixture of latent risk profiles
z_{dn}	Latent topic assignment	Risk profile linked to modality x_{dn}
α	Dirichlet prior for θ_d	Regularity of portfolio-level risk heterogeneity
η	Dirichlet prior for β_k	Smoothness of risk factor representation within a profile

Table 3.A.1: Summary of notation and actuarial interpretation.

3.B Sampled distributions

$$p(\theta_d|\alpha) = \mathcal{D}(\theta_d|\alpha) = \frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \theta_{dk}^{\alpha_k-1}, \quad (3.B.1)$$

$$p(\beta_k|\eta) = \mathcal{D}(\beta_k|\eta) = \frac{\Gamma\left(\sum_{v=1}^V \eta_v\right)}{\prod_{v=1}^V \Gamma(\eta_v)} \prod_{v=1}^V \beta_{kv}^{\eta_v-1}, \quad (3.B.2)$$

$$p(x_{dn}|z_{dn}, \beta_{1:K}) = \mathcal{M}(x_{dn}|z_{dn}, \beta_{1:K}) = \beta_{z_{dn}, x_{dn}}, \quad (3.B.3)$$

$$p(z_{dn}|\theta_d) = \mathcal{M}(z_{dn}|\theta_d) = \theta_{d, z_{dn}}. \quad (3.B.4)$$

$$p(\theta_d|\gamma_d) = \mathcal{D}(\theta_d|\gamma_d) = \frac{\Gamma\left(\sum_{k=1}^K \gamma_{dk}\right)}{\prod_{k=1}^K \Gamma(\gamma_{dk})} \prod_{k=1}^K \theta_{dk}^{\gamma_{dk}-1}, \quad (3.B.5)$$

$$q(\beta_k|\lambda_k) = \mathcal{D}(\beta_k|\lambda_k) = \frac{\Gamma\left(\sum_{v=1}^V \lambda_{kv}\right)}{\prod_{v=1}^V \Gamma(\lambda_{kv})} \prod_{v=1}^V \beta_{kv}^{\lambda_{kv}-1}, \quad (3.B.6)$$

$$q(z_{dn} = k) = \mathcal{M}_K\left(\vec{\phi}_{dn}\right) = \phi_{dx_{dn}k}. \quad (3.B.7)$$

3.C ELBO derivation

The marginal log-likelihood is defined as:

$$\ln p(x|\alpha, \eta) = \ln \int \sum_z p(x, z, \theta, \beta|\alpha, \eta) d\theta, \quad (3.C.1)$$

where $p(x, z, \theta, \beta | \alpha, \eta)$ is the joint distribution of observed variables x and latent variables z, θ, β . We introduce the variational distribution $q(z, \theta, \beta | \phi, \gamma, \lambda)$ defined in Eq. (3.2.5) and obtain the lower bound from Jensen's inequality:

$$\begin{aligned}
\ln p(x | \alpha, \eta) &= \ln \int \sum_z p(x, z, \theta, \beta | \alpha, \eta) \frac{q(z, \theta, \beta | \phi, \gamma, \lambda)}{q(z, \theta, \beta | \phi, \gamma, \lambda)} d\theta \\
&= \ln \mathbb{E}_q \left[\frac{p(x, z, \theta, \beta | \alpha, \eta)}{q(z, \theta, \beta | \phi, \gamma, \lambda)} \right] \\
&\geq \mathbb{E}_q \left[\ln \frac{p(x, z, \theta, \beta | \alpha, \eta)}{q(z, \theta, \beta | \phi, \gamma, \lambda)} \right] \\
&= \mathbb{E}_q [\ln p(x, z, \theta, \beta | \alpha, \eta)] - \mathbb{E}_q [\ln q(z, \theta, \beta | \phi, \gamma, \lambda)] \\
&= \mathcal{L}(\phi, \gamma, \lambda; \alpha, \eta) ,
\end{aligned} \tag{3.C.2}$$

where the under-script q is a short form for $q(z, \theta, \beta | \phi, \gamma, \lambda)$. From Bayesian rules, we have:

$$\mathcal{L}(\phi, \gamma, \lambda; \alpha, \eta) = -KL(q(z, \theta, \beta | \phi, \gamma, \lambda) || p(z, \theta, \beta | x, \alpha, \eta)) + \ln p(x | \alpha, \eta) .$$

The variational form of the marginal likelihood in Eq. (3.C.1) is then equal to:

$$\ln p(x | \alpha, \eta) = \mathcal{L}(\phi, \gamma, \lambda; \alpha, \eta) + KL(q(z, \theta, \beta | \phi, \gamma, \lambda) || p(z, \theta, \beta | x, \alpha, \eta)) .$$

To facilitate calculations (and coding), the tractable lower bound $\mathcal{L}(\phi, \gamma, \lambda; \alpha, \eta)$ in Eq. (3.C.2) is reorganised using the complete model in Eq. (3.2.3) and the variational distribution in Eq. (3.2.5):

$$\begin{aligned}
\mathcal{L}(x; \phi, \gamma, \lambda; \alpha, \eta) &= \mathbb{E}_q [\ln p(\beta | \eta)] + \mathbb{E}_q [\ln p(\theta | \alpha)] + \mathbb{E}_q [\ln p(x | z, \beta)] + \mathbb{E}_q [\ln p(z | \theta)] \\
&\quad - \mathbb{E}_q [\ln q(\beta | \lambda)] - \mathbb{E}_q [\ln q(\theta | \gamma)] - \mathbb{E}_q [\ln q(z | \phi)] .
\end{aligned}$$

We now expand the expectations above to get the complete objective function (ELBO) as functions of the variational parameters. In Hoffman et al. (2010), the per-corpus terms are brought into the summation over policies, and are thus divided by the number of policies D , such that the ELBO is equal to:

$$\begin{aligned}
\mathcal{L} &= \sum_{d=1}^D \left\{ \sum_{v=1}^V n_{dv} \sum_{k=1}^K \phi_{dvk} (\mathbb{E}_q [\ln \theta_{dk}] + \mathbb{E}_q [\ln \beta_{kv}] - \ln \phi_{dvk}) \right\} \\
&\quad + \sum_{d=1}^D \left\{ \sum_{k=1}^K a(\gamma_{dk}) + (\alpha - \gamma_{dk}) \mathbb{E}_q [\ln \theta_{dk}] \right\} \\
&\quad + \sum_{d=1}^D \left\{ \sum_{k=1}^K \left(\sum_{v=1}^V b(\lambda_{kv}) + (\eta - \lambda_{kv}) \mathbb{E}_q [\ln \beta_{kv}] \right) / D \right\} \\
&\quad + \sum_{d=1}^D \{ (\ln \Gamma(K\alpha) - K \ln \Gamma(\alpha)) + K (\ln \Gamma(V\eta) - V \ln \Gamma(\eta)) / D \} ,
\end{aligned}$$

where $a(\gamma_{dk}) = \ln \Gamma(\gamma_{dk}) - \ln \Gamma\left(\sum_{k'=1}^K \gamma_{dk'}\right)$, $b(\lambda_{kv}) = \ln \Gamma(\lambda_{kv}) - \ln \Gamma\left(\sum_{v'=1}^V \lambda_{kv'}\right)$, $\mathbb{E}_q[\ln \theta_{dk}] = \Psi(\gamma_{dk}) - \Psi\left(\sum_{k'=1}^K \gamma_{dk'}\right)$, and $\mathbb{E}_q[\ln \beta_{kv}] = \Psi(\lambda_{kv}) - \Psi\left(\sum_{v'=1}^V \lambda_{kv'}\right)$.

3.D LDA soft clustering

Algorithm 8: Online variational Bayes for LDA (Hoffman et al. 2010)

Initialisation

Define iteration weight:

$$\rho_t \triangleq (\tau_0 + t)^{-\kappa},$$

where $\tau_0 > 1$ is the learning offset, t is the EM iteration number, and $\kappa \in (0.5, 1.0)$ is the learning decay.

Define convergence condition:

$$\frac{1}{K} \sum_{k=1}^K |\text{change in } \gamma_{tk}| < 10^{-5}.$$

Randomly initialize variational parameter λ .

while *not converged* **do**

 // E-step

 Initialize $\gamma_{tk} = 1$ for all k ;

repeat

 Compute

$$\phi_{tvk} \propto \exp\left(\mathbb{E}_q[\ln \beta_{kv}] + \mathbb{E}_q[\ln \theta_{tk}]\right)$$

 and normalize over k ;

 Update

$$\gamma_{tk} = \alpha + \sum_{v=1}^V n_{tv} \phi_{tvk}.$$

until *convergence*;

 // M-step

 Compute

$$\tilde{\lambda}_{kv} = \eta + D n_{tv} \phi_{tvk}.$$

 Update variational parameter:

$$\lambda \leftarrow (1 - \rho_t) \lambda + \rho_t \tilde{\lambda}.$$

Output: β, θ

 Compute $\mathcal{L}$ for diagnostics.

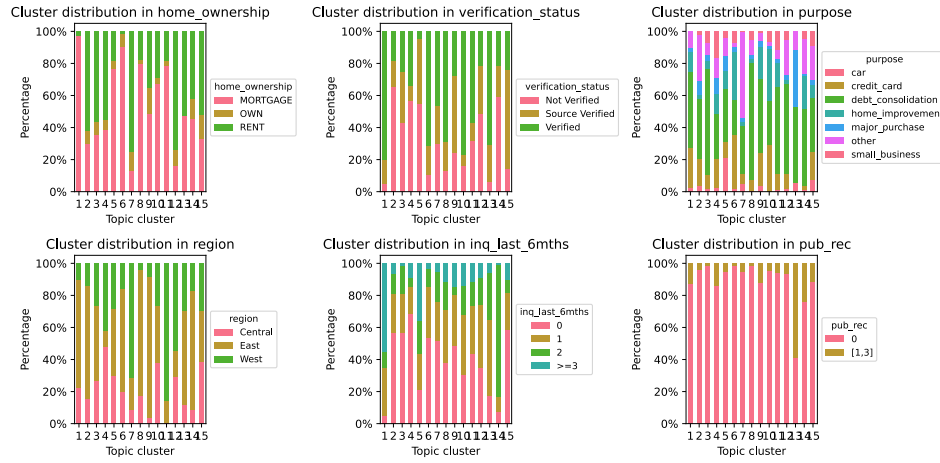


Figure 3.D.1: Topic cluster distribution (from the LDA approach) in the categorical variables `home_ownership`, `verification_status`, `purpose`, `region`, `inq_last_6mths`, and `pub_rec`. The topic clusters are ordered by increasing (average) number of loan defaults, as in Table 3.3.3.

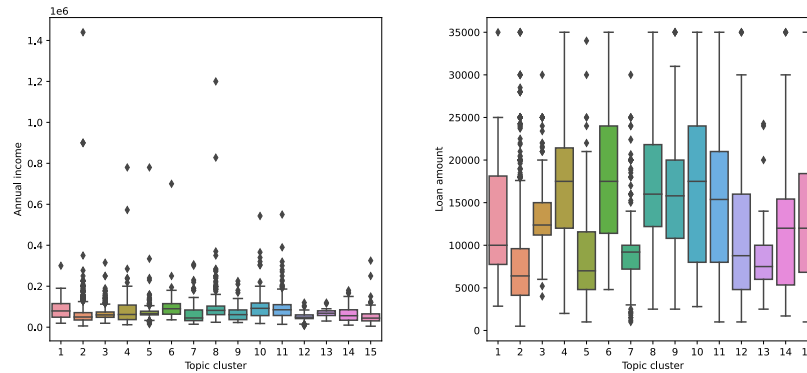


Figure 3.D.2: Boxplot of the numeric variables `annual_inc` and `loan_amnt` by topic cluster (LDA clustering approach). The topic clusters are ordered by increasing (average) number of loan defaults, as in Table 3.3.3.

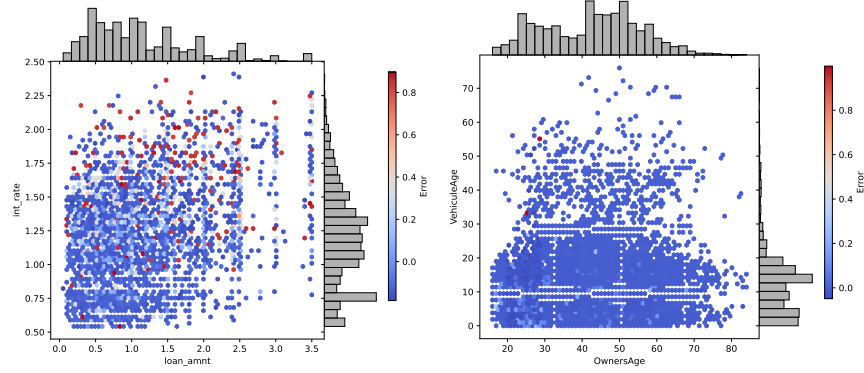


Figure 3.D.3: Heatmap of residuals ($y - \hat{y}$) across two continuous predictors, overlaid with the density of the predictors. No systematic pattern in residual magnitude is apparent. Larger residuals are more frequent in low-density regions due to fewer observations. Left plot: loan portfolio. Right plot: motor portfolio.

3.E Burt-adapted soft clustering

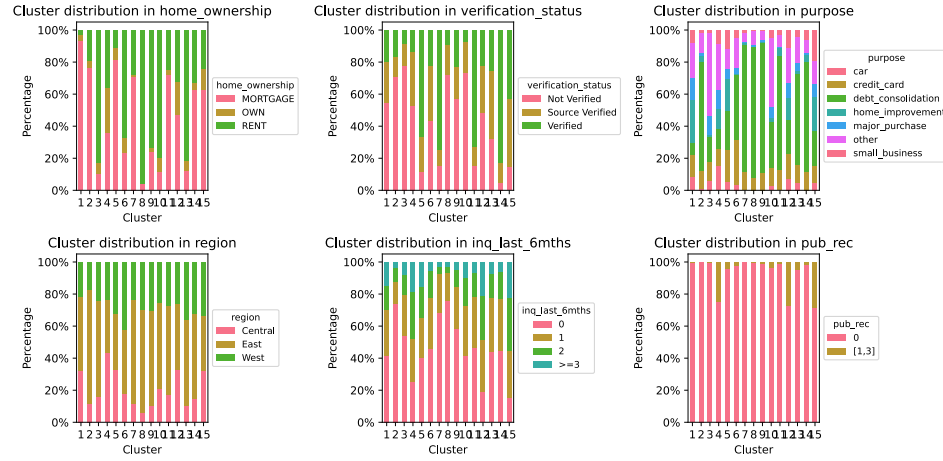


Figure 3.E.1: Cluster distribution (from the Burt approach) in the categorical variables `home_ownership`, `verification_status`, `purpose`, `region`, `inq_last_6mths`, and `pub_rec`. The clusters are ordered by increasing (average) number of loan defaults, as in Table 3.3.6.

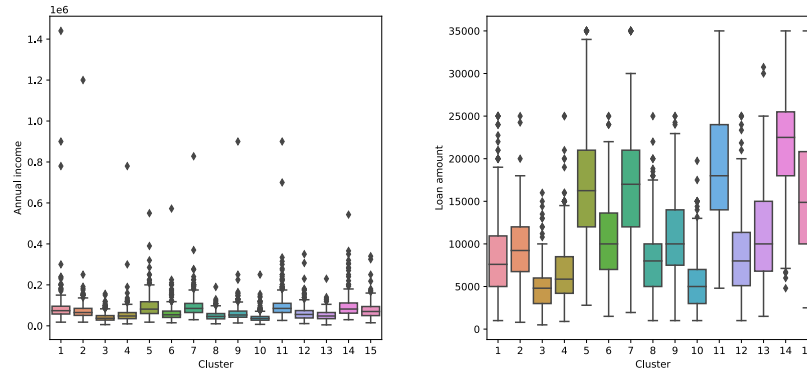


Figure 3.E.2: Boxplot of the numeric variables `annual_inc` and `loan_amnt` by cluster (Burt adapted clustering approach). The clusters are ordered by increasing (average) number of loan defaults, as in Table 3.3.6.

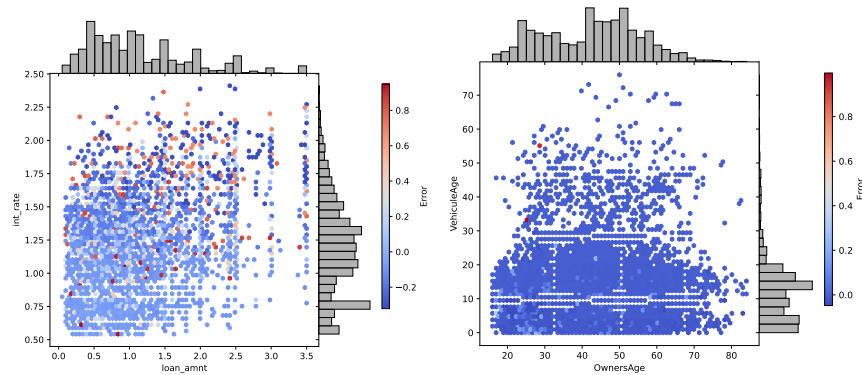


Figure 3.E.3: Heatmap of residuals ($y - \hat{y}$) across two continuous predictors, overlaid with the density of the predictors. No systematic pattern in residual magnitude is apparent. Larger residuals are more frequent in low-density regions due to fewer observations. Left plot: loan portfolio. Right plot: motor portfolio.

Part II

Fairness in insurance pricing

Chapter 4

A multivariate Energy-based fairness adjuster for premiums

Fairness in insurance pricing has received increasing attention, particularly in light of regulations calling for non-discriminatory premium estimation such as the EU Gender Directive (2012). This research focuses on removing structural bias in predicted premiums as a solidarity-driven fairness concept. We propose a gradient-based distributional adjustment designed to mitigate disparities across protected groups. The method relies on the Energy distance, a multivariate metric that enables the joint alignment of predicted premium distributions across multiple sensitive attributes (including non-binary ones) simultaneously. To maintain overall prediction balance, a post-hoc autocalibration step corrects biases in the total predicted number of claims. In addition, the method supports fairness adjustments for new policyholders without requiring retraining of the underlying fairness model. We evaluate the proposed methodology in a car insurance pricing setting, where demographic factors such as age and gender are commonly used for risk assessment and premium determination. Results show that the method effectively reduces group-level disparities while retaining a significant share of predictive accuracy.

Charlotte Jamotton and Donatien Hainaut (2026b).
“A Multivariate Energy-based Fairness Adjuster
for Premiums”. In: *North American Actuarial
Journal*. ISSN: 1092-0277. DOI: [10.1080/10920277.
2026.2705183](https://doi.org/10.1080/10920277.2026.2705183)

4.1 Introduction

A key characteristic of the insurance industry is its inverted production cycle. Ideally, premiums should be a function of the insured’s risk, but this risk is un-

observable when the contract is signed. Therefore, to set the premium that must be paid upfront, insurers instead rely on salient (and observable) features that correlate with the insured's risk. However, some of these rating factors, such as gender and ethnicity, are prohibited under anti-discrimination laws. While policy-makers seek to ensure that insurers do not make distinctions based on these legally prohibited characteristics (Avraham 2017, p.2), the law of insurance underwriting inherently grants insurers the authority to discriminate because it makes statistical sense to do so (Gandy 2016). This phenomenon, often described as an injustice by generalisation (Schanze 2013), is rooted in the assumption that individuals meet the average characteristics of the risk group to which they belong. This assumption lies at the core of the mutualization mechanism in insurance pricing, where risks are segmented into homogeneous rate-making classes or risk pools, imposing a form of solidarity among policyholders with similar risk profiles. Nonetheless, as Charpentier (2023, p.11, p.241–242) argues, individuals should not be treated differently because of their group membership, particularly when that membership is assigned at birth, as in the case of gender or ethnicity.

The seemingly incompatible notions of actuarial pricing and fairness present significant challenges for actuaries, who must comply with anti-discrimination regulations while still differentiating risk based on relevant characteristics. Moreover, the concept of discrimination itself is complex, multifaceted, and subject to interpretation (Edwards 2007), making its identification in insurance particularly difficult. As Charpentier (2023, p.1–10) explains, achieving fairness in insurance requires integrating justice and ethics into data-driven pricing models. For a broader discussion on the legal, economic, and ethical aspects of discrimination in actuarial pricing, we refer to Frees et al. (2021) and Charpentier (2023).

In recent years, the rise of big data and advances in machine learning techniques have further exacerbated these ongoing questions about fairness in insurance. While insurers are prohibited from directly discriminating based on specific policyholder attributes, simply removing these sensitive variables from pricing models does not entirely eliminate the risk of discrimination. This raises a critical question (Tschantz 2022). If insurers are not allowed to directly discriminate based on gender, why are proxy variables, such as vehicle power which correlate with and indirectly infer gender, still permissible?

In the actuarial literature, recent studies by Lindholm, Richman, Tsanakas, and Mario V. Wüthrich (2022b) and Charpentier et al. (2023) have proposed different post-processing approaches to mitigate both direct and indirect discrimination in insurance pricing. Lindholm, Richman, Tsanakas, and Mario V. Wüthrich (2022b) and Lindholm, Richman, Tsanakas, and Mario V. Wüthrich (2024b) individual fairness method aims to address proxy discrimination, commonly referred to as confounding bias in causal theory (O. Côté et al. 2025). Inspired by post-stratification techniques, it consists of averaging over sensitive information and a change of measure to prevent implicit inference of protected characteristics. Conversely, group fairness methodologies, such as those using Wasserstein barycenters

proposed by Charpentier et al. (2023) and extended by Hu et al. (2023) and Hu et al. (2024), aim to align outcome distributions across protected groups so that no association exists between sensitive information and premiums. According to O. Côté et al. (2025, p.10), these so-called corrective premiums offer the most stringent mitigation of indirect discrimination. A comprehensive review and classification of existing fairness methodologies in insurance can be found in Xin et al. (2023) and O. Côté et al. (2025).

In this chapter, we focus on Demographic Parity (DP), a solidarity-driven group fairness criterion, to remove structural bias in predicted premiums. Although EU regulations on gender-neutral pricing align more closely with an individual fairness perspective, we argue that DP is more consistent with the risk pooling mechanism inherent in insurance pricing. In practice, insurance contracts are typically priced using regression models (Denuit, Hainaut, et al. 2019d), such as Generalized Linear Models (GLMs), which assign a limited set of unique premiums to a much larger pool of policies. In other words, the risk assessment and premium allocation are not individualised but structured around groups with shared characteristics. By adopting DP, we ensure that the probability of receiving a given premium level becomes independent of demographic attributes such as gender, promoting fair treatment across groups in line with the risk pooling principle. We acknowledge that different theoretical and practical considerations may support alternative fairness criteria.

The main contributions of this paper are threefold. First, we propose a gradient-based fairness correction using the Energy distance to reduce disparities across multiple sensitive attributes simultaneously, including non-binary variables. The multivariate nature of the Energy distance is a key advantage of our approach. It allows for simultaneous fairness adjustments across protected variables with different numbers of sensitive classes, such as gender (binary) and discretised age (multi-class), and is therefore order-independent. Second, we introduce a post-hoc autocalibration step (Denuit, Charpentier, et al. 2021) to correct biases in the total predicted claim count after fairness adjustments. Third, we design a simple out-of-sample mapping mechanism to apply fairness corrections to new policyholders. Note that, while several recent approaches, including individual fairness and counterfactual fairness methods (Kusner, Loftus, et al. 2017; Grari et al. 2020; Fahrenwaldt et al. 2024), can generalise fairness adjustments to new observations, our contribution lies in providing a straightforward mapping mechanism that enables the out-of-sample transformation of premiums without re-optimising fairness corrections for each new policyholder. In the machine-learning literature, post-hoc corrections have similarly been shown to enforce group-fairness constraints without retraining (Hardt et al. 2016; Pleiss et al. 2017).

The chapter is structured as follows. Section 4.2 introduces notations and defines Demographic Parity (DP) as a group fairness criterion. Section 4.3 presents the theoretical justification for choosing the Energy distance and its relationship to other statistical distances. Section 4.4 details the gradient-based adjustment, the autocalibration procedure, and the out-of-sample mapping. Section 4.5 evaluates

the methodology in a car insurance pricing setting, demonstrating its effectiveness in reducing group-level disparities while preserving a substantial share of predictive accuracy.

4.2 Notations and definitions

Consider an insurance portfolio represented as a tabular dataset $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ with n policies (rows) and m rating factors (columns). Each policy $i = 1, \dots, n$ is described by a vector of characteristics $\mathbf{x}_i = (x_{i1}, \dots, x_{im}) \in \mathbb{R}^m$. We partition the m rating factors into two disjoint sets. The set of allowed covariates $\mathcal{A} \subseteq \{1, \dots, m\}$, and the set of prohibited (or discriminatory) covariates $\mathcal{D} \subseteq \{1, \dots, m\}$, such that $\mathcal{A} \cup \mathcal{D} = \{1, \dots, m\}$ and $\mathcal{A} \cap \mathcal{D} = \emptyset$. For each policy i , we denote by $\mathbf{A}_i = (x_{ij})_{j \in \mathcal{A}}$, and $\mathbf{D}_i = (x_{ij})_{j \in \mathcal{D}}$ the subvectors of allowed and prohibited rating factors, respectively.

For every policy i , we observe the number of claims N_i (count data) reported during its contract duration (or exposure) ν_i . Although we present the notations using claim counts N , the methodology applies to any real-valued insurance response. Throughout the analysis, we will often refer to the claim frequency per unit of exposure, defined as the ratio $\lambda_i = N_i/\nu_i$. Counts are low-support integers (typically 0–3), while λ is continuous on $[0, N/\min(\nu)]$, more stable, and more interpretable for illustrating fairness adjustments.

In actuarial pricing, the objective is to approximate the (unknown) conditional expectation of the response variable Y given the available rating information $(\mathbf{A}, \mathbf{D})$:

$$\mu(\mathbf{A}, \mathbf{D}) = \mathbb{E}[Y | (\mathbf{A}, \mathbf{D})] . \quad (4.2.1)$$

In the empirical analysis presented in Section 4.5, we use a Light Gradient Boosting machine (LightGBM) to model the premium best-estimates $\hat{\pi}(\mathbf{A}, \mathbf{D})$ for the (true) pure premium $\mu(\mathbf{A}, \mathbf{D})$ defined in Eq. (4.2.1). This best-estimate predictor uses all available features, including the protected attributes $\mathbf{D}$. Although it is statistically optimal (because it exploits all predictive information), it is also directly discriminatory by construction. In this chapter, we focus on Demographic Parity (DP), a solidarity-driven group fairness criterion, to remove structural bias from predicted premiums.

Definition 4.1 (Demographic Parity). Demographic Parity (DP) is a group fairness criterion that requires the conditional distribution of the premiums to be identical across all protected demographic groups. For two groups a and b in the support of $\mathbf{D}$,

$$\mathbb{P}(\hat{\pi}(\mathbf{A}, \mathbf{D}) | \mathbf{D} = a) = \mathbb{P}(\hat{\pi}(\mathbf{A}, \mathbf{D}) | \mathbf{D} = b) . \quad (4.2.2)$$

In practice, DP ensures that no protected group receives systematically different pricing. See Xin et al. (2023) for alternative group fairness criteria definitions.

In a regression setting, violations of Eq. (4.2.2) can be quantified by a statistical distance between the conditional distributions $\hat{\pi}(\mathbf{A}, \mathbf{D}) \mid (\mathbf{D} = a)$ and $\hat{\pi}(\mathbf{A}, \mathbf{D}) \mid (\mathbf{D} = b)$. Many distances may be used to quantify disparities across protected groups, including divergences (e.g., Kullback-Leibler) or optimal-transport metrics (e.g., Wasserstein). As detailed in Section 4.4, we propose the use of the Energy distance, defined in Section 4.3, to align the group-conditional distributions of predicted premiums in order to satisfy the DP condition in Eq. (4.2.2).

4.3 Measures of disparities

In this section, we review several distances and divergences commonly used to quantify discrepancies between probability distributions, with a particular emphasis on their suitability for gradient-based fairness correction. We discuss the Kullback-Leibler divergence, the Wasserstein distance, the Cramér distance, and its multivariate extension i.e., the Energy distance. Their respective properties and limitations motivate our choice of the Energy distance for enforcing DP in premium predictions.

4.3.1 Classical distances and divergences

In machine learning, Kullback et al. (1951) divergence is among the most widely used measures of dissimilarity between two distributions.

Definition 4.2 (Kullback-Leibler Divergence). For probability distributions P and Q with densities p and q , the Kullback-Leibler (KL) divergence is defined as:

$$d_{\text{KL}}(P \parallel Q) = \int_{-\infty}^{\infty} p(u) \log \left(\frac{p(u)}{q(u)} \right) dx.$$

As a measure of relative entropy, the KL divergence does not account for the geometric proximity between outcomes. For example, consider an actuarial pricing context where $p(u)$ assigns a 0.01% probability to a premium level $u = 100$, while $q(u)$ assigns the same probability to $u = 101$. Despite the two premiums being very close, the KL divergence treats them as entirely different outcomes. To address this limitation, researchers (Peyré et al. 2019; Panaretos et al. 2019) have turned to alternative distance metrics that consider the underlying geometry between outcomes, such as Wasserstein.

Definition 4.3 (Wasserstein Distance). For two probability distributions P and Q , the Wasserstein distance of order p , also known as the Earth Mover distance when $p = 1$, is defined as (Kantorovich et al. 1958):

$$W_p(P, Q) = \left(\inf_{\gamma \in \Gamma(P, Q)} \int \|u - v\|^p d\gamma(u, v) \right)^{1/p}, \quad (4.3.1)$$

where $\Gamma(P, Q)$ denotes the set of all joint distributions for the pair (U, V) with marginals P and Q , respectively. The infimum in Eq. (4.3.1) is taken over all possible couplings (defined below) $\gamma \in \Gamma(P, Q)$ ensuring that the optimal transport plan minimises the cost of transporting mass from u to v (Villani 2009).

Definition 4.4 (Coupling). A coupling of two probability distributions P and Q is any joint distribution γ on (U, V) such that $U \sim P$ and $V \sim Q$. The set of all such couplings is denoted $\Gamma(P, Q)$ (Villani 2009; Santambrogio 2015).

The Wasserstein distance accounts for the underlying geometry between outcomes through the transport cost $\|u - v\|^p$. However, a key drawback for optimisation is that W_p does not satisfy the unbiased sample gradients property (Bellemare et al. 2017), one of three key properties (the other two being sum invariance and scale sensitivity) in machine learning when using a distance as a loss function.

Definition 4.5 (Unbiased Sample Gradients). Let $d(P, Q)$ be a statistical divergence. We say d has unbiased sample gradients if, for all m independent samples $\mathbf{u}_m$ drawn from P , the expected gradient of the sample loss $\theta \mapsto d(\hat{P}_m, Q_\theta)$ satisfies (Bellemare et al. 2017, p.3):

$$\mathbb{E}_{\mathbf{u}_m \sim P} \left[\nabla_\theta d(\hat{P}_m, Q_\theta) \right] = \nabla_\theta d(P, Q_\theta) .$$

Bellemare et al. (2017) show that the Wasserstein distance lacks unbiased sample gradients, leading to convergence issues and motivating Cramér (1928) distance as an unbiased alternative. This problem was further addressed by Genevay et al. (2017) via the Sinkhorn divergence, and confirmed in practice for minibatch optimisation by Fatras et al. (2019).

Definition 4.6 (Cramér Distance). For one-dimensional distributions P and Q with cumulative distribution functions (CDFs) F_P and F_Q , Cramér (1928) distance is defined as the ℓ_2^2 -distance:

$$\ell_2^2(P, Q) = \int_{-\infty}^{\infty} |F_P(u) - F_Q(u)|^2 dx . \quad (4.3.2)$$

Its square root is a member of the ℓ_p family of metrics:

$$\ell_p(P, Q) = \left(\int_{-\infty}^{\infty} |F_P(u) - F_Q(u)|^p dx \right)^{1/p} . \quad (4.3.3)$$

Interestingly, for one-dimensional distributions P and Q , the p -Wasserstein distance in Eq. (4.3.1) can be expressed in terms of the quantile functions F_P^{-1} and F_Q^{-1} (Panaretos et al. 2019):

$$W_p(P, Q) = \left(\int_0^1 |F_P^{-1}(u) - F_Q^{-1}(u)|^p du \right)^{1/p} . \quad (4.3.4)$$

For $p = 1$, Eq. (4.3.4) simplifies to the ℓ_1 -distance, a member of the ℓ_p family defined in Eq. (4.3.3) like the square-root of the Cramér distance:

$$\begin{aligned} W_1(P, Q) &= \int_0^1 \left| F_P^{-1}(u) - F_Q^{-1}(u) \right| dx \\ &= \int_{-\infty}^{+\infty} |F_P(u) - F_Q(u)| dx \\ &= \ell_1(P, Q). \end{aligned}$$

Besides accounting for the underlying geometry between outcomes (like Wasserstein and unlike KL), the Cramér distance also possesses desirable optimisation properties, including unbiased sample gradients (unlike Wasserstein), sum invariance, and scale sensitivity, making it more suitable for gradient-based learning than the Wasserstein metric. The following property establishes that the squared ℓ_2 -distance admits unbiased sample gradients, as demonstrated in Appendix 4.A.

Proposition 4.1 (Unbiased sample gradient of the squared ℓ_2 -distance). *Let P be a probability measure and let $\{Q_\theta\}_{\theta \in \Theta}$ denote a parameterized family of probability measures, with cumulative distribution functions $F_P(u)$ and $F_{Q_\theta}(u)$, respectively. The gradient of the squared ℓ_2 -distance between P and Q_θ satisfies the following identity (Bellemare et al. 2017, p.15):*

$$\nabla_\theta \ell_2^2(P, Q_\theta) = \mathbb{E}_{\mathbf{u}_m \sim P} \left[\nabla_\theta \ell_2^2(\hat{P}_m, Q_\theta) \right],$$

where $\hat{P}_m$ denotes the empirical measure based on m independent samples drawn from P .

4.3.2 Energy distance

Although the Cramér distance combines the best of KL and Wasserstein, it is limited to univariate distributions. In the context of actuarial fairness, real-world scenarios often involve multiple sensitive attributes. A key observation by Székely (2003) is that $\ell_2^2(P, Q)$ in Eq. (4.3.2) can be rewritten as:

$$\ell_2^2(P, Q) = \mathbb{E}[|U - V|] - \frac{1}{2}\mathbb{E}[|U - U'|] - \frac{1}{2}\mathbb{E}[|V - V'|],$$

where $|\cdot|$ denotes the absolute value for scalars $u, v \in \mathbb{R}$, and U' and V' are independent and identically distributed copies of U and V . As shown by Székely (2003), replacing absolute values by Euclidean norms $\|\cdot\|$ generalises Eq. (4.3.2) to higher dimensions, known as the Energy distance.

Definition 4.7 (Energy Distance). For d -dimensional independent random variables U, V , the Energy distance (for the Euclidean metric) is defined as:

$$\begin{aligned} \mathcal{E}(P, Q) &= 2\mathbb{E}_{U \sim P, V \sim Q} [\|U - V\|] \\ &\quad - \mathbb{E}_{U, U' \sim P} [\|U - U'\|] - \mathbb{E}_{V, V' \sim Q} [\|V - V'\|]. \end{aligned} \quad (4.3.5)$$

In each expectation the relevant draws are independent (independent pairs for the within-distribution terms and independent cross draws for the first term).

For univariate distributions, the equivalence $\mathcal{E}(P, Q) = 2\ell_2^2(P, Q)$ is proved in Appendix 4.A. Like Cramér distance, the $\mathcal{E}$ -distance admits unbiased sample gradients, as formally demonstrated in Appendix 4.A.

Proposition 4.2 (Unbiased sample gradient of the $\mathcal{E}$ -distance). *Let $U \sim P$, $V \sim Q_\theta$, and let $\hat{U}_m \sim \hat{P}_m$ denote m independent samples drawn from P . The gradient ∇ of the Energy distance $\mathcal{E}(P, Q_\theta)$ with respect to θ satisfies the following identity (Bellemare et al. 2017, p.15):*

$$\nabla_\theta \mathcal{E}(P, Q_\theta) = \mathbb{E}_{\mathbf{u}_m \sim P} \left[\nabla_\theta \mathcal{E}(\hat{P}_m, Q_\theta) \right],$$

showing that $\mathcal{E}$ admits unbiased sample gradients.

The $\mathcal{E}$ -distance also retains a strong connection to optimal transport. The Wasserstein distance in Eq. (4.3.1) can be defined in terms of expectations over coupled random variables $U \sim P$ and $V \sim Q$:

$$\begin{aligned} W_p(P, Q) &= \left(\inf_{\gamma \in \Gamma(P, Q)} \int \|u - v\|^p d\pi(u, v) \right)^{1/p} \\ &= \inf_{\gamma \in \Gamma(P, Q)} \mathbb{E}_{(U, V) \sim \pi} \left[\|U - V\|^p \right]^{1/p}. \end{aligned}$$

As discussed in Sejdinović et al. (2013), the Energy distance can be equivalently represented in the integral form. The first expectation in Eq. (4.3.5) may be written as the iterated integral:

$$\mathbb{E}_{U \sim P, V \sim Q} [\|U - V\|] = \int_{\mathcal{U}} \left(\int_{\mathcal{V}} \|u - v\| Q(dy) \right) P(dx).$$

Given that U and V are independent, the joint law of (U, V) is the product measure $P \otimes Q$. By Fubini/Tonelli the iterated integral equals the integral with respect to the product measure $P \otimes Q$:

$$\int_{\mathcal{U}} \int_{\mathcal{V}} \|u - v\| Q(dy) P(dx) = \int_{\mathcal{U} \times \mathcal{V}} \|u - v\| (P \otimes Q)(dx dy) = \mathbb{E}_{(U, V) \sim P \otimes Q} [\|U - V\|].$$

Consequently, Eq. (4.3.5) is equivalently written as:

$$\mathcal{E}(P, Q) = 2 \mathbb{E}_{(U, V) \sim P \otimes Q} [\|U - V\|] - \mathbb{E}_{P \otimes P} [\|U - U'\|] - \mathbb{E}_{Q \otimes Q} [\|V - V'\|].$$

The last two expectations are already written as expectations under $(U, U') \sim P \otimes P$ and $(V, V') \sim Q \otimes Q$ to emphasize that those terms depend only on the marginals P and Q . The product measure (joint law) $P \otimes Q$ is an element of the set of couplings

$\Gamma(P, Q)$. The usual Energy distance is closely related to the 1-Wasserstein distance evaluated at the independent/product coupling $\gamma_\perp = P \otimes Q \in \Gamma(P, Q)$:

$$\mathcal{E}(P, Q) = 2 \mathbb{E}_{(U, V) \sim \gamma_\perp} [\|U - V\|] - \mathbb{E}_{P \otimes P} [\|U - U'\|] - \mathbb{E}_{Q \otimes Q} [\|V - V'\|].$$

As noted by Duong et al. (2022), both the Wasserstein and Energy distances can be interpreted as the energy required to move one probability distribution P to another Q . The Wasserstein distance measures the cost of the optimal transport plan, whereas the Energy distance considers the cost of a random (maximum-entropy) transport plan. In the following Section 4.4, we introduce a gradient-based fairness correction mechanism based on the Energy distance.

4.4 Energy gradient fairness adjuster

As outlined in Section 4.2, our goal is to mitigate group-level disparities across $G \geq 2$ protected groups by enforcing DP. More specifically, we seek to align the conditional distribution of predicted premiums $\hat{\pi}(\mathbf{A}, \mathbf{D})$ across all values of the sensitive attribute(s) $\mathbf{D}$. The Energy distance provides a differentiable and geometrically meaningful measure of disparity, making it particularly suitable for this type of gradient-based fairness correction.

4.4.1 Gradient-based fairness correction

Let $U, V \in \mathbb{R}^d$ be two independent random variables drawn from distributions P and Q , respectively. Given samples $\{\mathbf{u}_i\}_{i=1}^{n_u}$ and $\{\mathbf{v}_j\}_{j=1}^{n_v}$ in $\mathbb{R}^d$, the Energy distance, defined in Eq. (4.3.5), admits the empirical estimate:

$$\begin{aligned} \mathcal{E}(P, Q) = & \frac{2}{n_u n_v} \sum_{i=1}^{n_u} \sum_{j=1}^{n_v} \|\mathbf{u}_i - \mathbf{v}_j\| \\ & - \frac{1}{n_u^2} \sum_{i=1}^{n_u} \sum_{j \neq i}^{n_u} \|\mathbf{u}_i - \mathbf{u}_j\| - \frac{1}{n_v^2} \sum_{i=1}^{n_v} \sum_{j \neq i}^{n_v} \|\mathbf{v}_i - \mathbf{v}_j\|, \end{aligned} \quad (4.4.1)$$

with the pairwise Euclidean distance defined as:

$$\|\mathbf{u}_i - \mathbf{v}_j\| = \sqrt{\sum_{k=1}^d (u_{ik} - v_{jk})^2}.$$

We propose a fairness correction based on the Energy gradient that iteratively nudges the distributions of U and V toward each other. For an observed pair $(\mathbf{u}_i, \mathbf{v}_j)$, the gradients of the pairwise norm with respect to a component u_{ik} or v_{jk} are:

$$\frac{\partial \|\mathbf{u}_i - \mathbf{v}_j\|}{\partial u_{ik}} = \frac{u_{ik} - v_{jk}}{\|\mathbf{u}_i - \mathbf{v}_j\|}, \quad \text{and} \quad \frac{\partial \|\mathbf{u}_i - \mathbf{v}_j\|}{\partial v_{jk}} = -\frac{v_{jk} - u_{ik}}{\|\mathbf{v}_j - \mathbf{u}_i\|}.$$

Hence, the gradient of the Energy distance with respect to individual observations $\{\mathbf{u}_i\}_{i=1}^{n_u}$ and $\{\mathbf{v}_j\}_{j=1}^{n_v}$ is:

$$\begin{aligned}\nabla_{\mathbf{u}_i}\mathcal{E}(P, Q) &= \frac{2}{n_u n_v} \sum_{j=1}^{n_v} \frac{\mathbf{u}_i - \mathbf{v}_j}{\|\mathbf{u}_i - \mathbf{v}_j\|} - \frac{2}{n_u^2} \sum_{j=1}^{n_u} \frac{\mathbf{u}_i - \mathbf{u}_j}{\|\mathbf{u}_i - \mathbf{u}_j\|}, \quad \text{and} \\ \nabla_{\mathbf{v}_j}\mathcal{E}(P, Q) &= -\frac{2}{n_u n_v} \sum_{i=1}^{n_u} \frac{\mathbf{v}_j - \mathbf{u}_i}{\|\mathbf{v}_j - \mathbf{u}_i\|} + \frac{2}{n_v^2} \sum_{i=1}^{n_v} \frac{\mathbf{v}_j - \mathbf{v}_i}{\|\mathbf{v}_j - \mathbf{v}_i\|}.\end{aligned}$$

By iteratively updating U and V via gradient descent, we progressively align their distributions:

$$\begin{pmatrix} U^{(e+1)} \\ V^{(e+1)} \end{pmatrix} = \begin{pmatrix} U^{(e)} \\ V^{(e)} \end{pmatrix} - \eta \begin{pmatrix} \nabla_U \mathcal{E}(P, Q) \\ \nabla_V \mathcal{E}(P, Q) \end{pmatrix}, \quad (4.4.2)$$

where η is the learning rate and e indexes the iteration (epoch).

Example 4.1 (Two sensitive groups). For $G = 2$ groups (e.g., gender), let $U = \hat{\pi}_{\text{female}}$ and $V = \hat{\pi}_{\text{male}}$, with $d = 1$. Eq. (4.4.2) can be rewritten as:

$$\hat{\pi}^{\text{fair}(e+1)} = \begin{pmatrix} \hat{\pi}_{\text{female}}^{(e)} \\ \hat{\pi}_{\text{male}}^{(e)} \end{pmatrix} - \eta \begin{pmatrix} \nabla_{\hat{\pi}_{\text{female}}} \mathcal{E}(P, Q) \\ \nabla_{\hat{\pi}_{\text{male}}} \mathcal{E}(P, Q) \end{pmatrix}. \quad (4.4.3)$$

For $G > 2$ groups, the premium vector can be partitioned as:

$$\hat{\pi} = \{\hat{\pi}_g\}_{g=1}^G = \begin{pmatrix} \{\hat{\pi}_g\}_{g=1}^{\lceil G/2 \rceil} \\ \{\hat{\pi}_g\}_{g=\lceil G/2 \rceil+1}^G \end{pmatrix} = \begin{pmatrix} U \\ V \end{pmatrix}. \quad (4.4.4)$$

Note that the ordering of dimensions is not meaningful. Dimensions can be treated as exchangeable.

Example 4.2 (More than two sensitive groups). Consider the protected attributes $\mathbf{D}$ include both the policyholder's gender $D_1 \in \{\text{female}, \text{male}\}$ and his age group $D_2 \in \{\text{young}, \text{middle-aged}, \text{senior}\}$, leading to $G = 6$ distinct groups (e.g., young female, middle-aged male, etc.). Given that $G = 6$, U and V are both 3-dimensional variables in $\mathbb{R}^3$, with $d = 3$, where $\{u_{i,1}\}_{i=1}^{n_1}$ represents the young female premiums, $\{v_{j,1}\}_{j=1}^{n_{\lceil G/2 \rceil+1}}$ the young male premiums, $\{u_{i,2}\}_{i=1}^{n_2}$ the middle-aged female premiums, etc.

Since the vectors $\hat{\pi}_g$ sizes n_g may differ across groups $g \in (1, \dots, G)$, we account for unequal group sizes by sampling equal-sized subsets from U and V at each iteration (see Proposition 4.2). Moreover, to ensure pairwise comparisons are made not only within the same age category (e.g., comparing young female with young male, and middle-aged female with middle-aged male, etc.) but across different pairs (e.g., comparing young female with senior male), a uniform random permutation (Knuth 1981), denoted τ , is applied to U (or V equivalently) at each iteration e .

Algorithm 9 below is a pseudo-algorithm outlining the procedure. The premium predictions $\hat{\pi}$ are adjusted iteratively using gradient descent (via backpropagation) over multiple epochs until the fairness loss function is minimised. The gradient $\nabla \mathcal{L}_{\text{fair}} = \nabla \mathcal{E}$ is computed via automatic differentiation (using `loss.backward` in PyTorch).

Algorithm 9: $\mathcal{E}$ -based fairness adjuster

Data: $\hat{\pi} = \{\hat{\pi}_g\}_{g=1}^G$
Initialisation: $\hat{\pi}^{\text{fair}}(0) \leftarrow \{\hat{\pi}_g\}_{g=1}^G$
for $e = 1$ **to** E **do**
 Partition predictions as in Eq. (4.4.4):
 $U^\tau(e-1) \leftarrow \{\hat{\pi}_g^{\text{fair}}(e-1)\}_{g=1}^{\lceil G/2 \rceil} \in \mathbb{R}^{n_u \times d}$
 $V(e-1) \leftarrow \{\hat{\pi}_g^{\text{fair}}(e-1)\}_{g=\lceil G/2 \rceil+1}^G \in \mathbb{R}^{n_v \times d}$
 Compute pairwise distances from Eq. (4.4.1):
 $\text{dist}_{UV} \leftarrow \frac{1}{n_u n_v} \sum_{i=1}^{n_u} \sum_{j=1}^{n_v} \sqrt{\sum_{k=1}^d (u_{ik}(e-1) - v_{jk}(e-1))^2}$
 $\text{dist}_{UU} \leftarrow \frac{1}{n_u^2} \sum_{i=1}^{n_u} \sum_{j \neq i} \sqrt{\sum_{k=1}^d (u_{ik}(e-1) - u_{jk}(e-1))^2}$
 $\text{dist}_{VV} \leftarrow \frac{1}{n_v^2} \sum_{i=1}^{n_v} \sum_{j \neq i} \sqrt{\sum_{k=1}^d (v_{ik}(e-1) - v_{jk}(e-1))^2}$
 Compute Energy distance fairness loss:
 $\mathcal{L}_{\text{fair}}(e) \leftarrow 2 \cdot \text{dist}_{UV} - \text{dist}_{UU} - \text{dist}_{VV}$
 Compute gradient $\nabla \mathcal{L}_{\text{fair}}(e)$ w.r.t. $U(e-1), V(e-1)$ using Eq. (4.4.2)
 via automatic differentiation and update predictions:
 $\hat{\pi}^{\text{fair}}(e) \leftarrow \begin{pmatrix} U(e) \\ V(e) \end{pmatrix} = \begin{pmatrix} U(e-1) \\ V(e-1) \end{pmatrix} - \eta \begin{pmatrix} \nabla_{U(e-1)} \mathcal{L}_{\text{fair}}(e) \\ \nabla_{V(e-1)} \mathcal{L}_{\text{fair}}(e) \end{pmatrix}$
end
return $\hat{\pi}^{\text{fair}}(E)$

4.4.2 New policyholders

We further propose a mechanism to adjust predictions for new policyholders, so that we do not need to retrain the entire fairness adjuster. For each policyholder in the set used to train the fairness adjuster model, an individual fairness correction factor γ_i is defined as:

$$\gamma_i = \frac{\hat{\pi}_i^{\text{fair}}}{\hat{\pi}_i},$$

where $\hat{\pi}_i$ is the original predicted premium before fairness adjustment, and $\hat{\pi}_i^{\text{fair}}$ is the fairness-adjusted premium. This correction factor γ_i quantifies the individual adjustment required for fairness. We train a random forest regressor f_θ to predict γ_i based on the policyholder's features $(\mathbf{A}_i, \mathbf{D}_i)$:

$$\hat{\gamma}_i = f_\theta(\mathbf{A}_i, \mathbf{D}_i).$$

This model learns how different features influence fairness adjustments. For a new policyholder with features $(\mathbf{A}_{\text{new}}, \mathbf{D}_{\text{new}})$, the fairness correction factor is predicted as:

$$\hat{\gamma}_{\text{new}} = f_{\theta}(\mathbf{A}_{\text{new}}, \mathbf{D}_{\text{new}}).$$

The initial regression model (e.g., LightGBM) predictions $\hat{\pi}_{\text{new}}$ are then adjusted using this factor:

$$\hat{\pi}_{\text{new}}^{\text{fair}} = \hat{\pi}_{\text{new}} \cdot \hat{\gamma}_{\text{new}}.$$

As showed in Section 4.5.3, this mechanism assigns to new policyholders premiums that align with the fairness correction learned during the fairness adjuster training.

4.4.3 Post-hoc autocalibration

Finally, note that the regression models (particularly tree-based boosting) used to predict the premium best-estimates $\hat{\pi}(\mathbf{A}, \mathbf{D})$ often exhibit imbalances between predicted and actual claims total. Such discrepancies can adversely affect the financial stability of an insurance portfolio.

A model is globally unbiased when the expected value of the predicted premiums matches the expected value of the actual outcomes Y :

$$\mathbb{E}[\hat{\pi}(\mathbf{A}, \mathbf{D})] = \mathbb{E}[Y].$$

A model is locally unbiased when, for every specific value s of the predicted premiums:

$$\mathbb{E}[Y | \hat{\pi}(\mathbf{A}, \mathbf{D}) = s] = s.$$

Deviations from these conditions indicate that certain segments may be underpriced or overpriced, with substantial implications for insurance pricing strategies.

Denuit, Charpentier, et al. (2021) address these calibration issues by proposing a local regression technique to correct model bias. The expected actual loss $\mathbb{E}[Y | \hat{\pi}(\mathbf{A}, \mathbf{D})]$ is first estimated using local regression techniques on $\{(\hat{\pi}(\mathbf{a}_i, \mathbf{d}_i), y_i)\}$. The predictions $\hat{\pi}$ are adjusted with a multiplicative factor $\lambda_{\alpha}(s)$ derived from the local regression estimates:

$$\lambda_{\alpha}(s) = \frac{\mathbb{E}[Y | \hat{\pi}(\mathbf{A}, \mathbf{D}) = s]}{s}.$$

The balance corrected premiums $\hat{\pi}^{\text{BC}}$ are then given by:

$$\hat{\pi}^{\text{BC}} = \lambda_{\alpha}(\hat{\pi}(\mathbf{a}, \mathbf{d})) \cdot \hat{\pi}(\mathbf{a}, \mathbf{d}).$$

As illustrated in Section 4.D, the balance correction is applied after the fairness adjustment, thereby ensuring that the final adjusted premiums accurately reflect the underlying risk across different segments:

$$\hat{\pi}^{\text{fair, BC}} = \lambda_{\alpha}^{\text{fair}}(\hat{\pi}^{\text{fair}}) \cdot \hat{\pi}^{\text{fair}}.$$

4.5 Results

In this Section 4.5, we empirically evaluate the approach proposed in Section 4.4. In Section 4.5.1, we first introduce the datasets and their key characteristics. In Section 4.5.2, we model the premium best-estimates $\hat{\pi}(\mathbf{A}, \mathbf{D})$ using a tree-based boosting model (LightGBM). In Sections 4.5.3 and 4.5.4, we adjust the premium best-estimates for fairness across two and six protected demographic groups, and correct for the balance bias.

4.5.1 Datasets

To evaluate the proposed methodology, we first simulate a motor insurance portfolio $\mathcal{X}$ of $n = 100\,000$ policies. Table 4.5.1 displays a sample of five policies, where each column corresponds to a rating factor. In this case study, the policyholder's gender and age, two key risk factors in motor insurance, are considered as prohibited variables $\mathbf{D} = (D_1, D_2)$. For each policy, we observe the number of claims $N_i \in \{0, 1, 2, 3\}$ reported during its contract duration (or exposure) $\nu_i \in (1, 365)$ measured in days. As the majority of policies report zero claims (see Table 4.B.1 in Appendix), the count variable N is highly zero-inflated.

Throughout the analysis, we work with the claim frequency per unit of exposure $\lambda_i = N_i/\nu_i$. Note that we assume that the models predictions $\hat{\lambda}$ (per unit of exposure), are the so-called premium best-estimates $\hat{\pi}(\mathbf{A}, \mathbf{D})$.

	Income	Age	Power	Education	Location	Type	Gender	Ethnicity
1	37047.85	27.48	Low	High School	Urban	Sports	M	American
2	14426.15	24.31	High	Bachelor	Suburban	SUV	M	Asian
3	25834.59	28.24	Medium	High School	Suburban	SUV	F	European
4	51670.00	32.62	Low	Master	Urban	Sports	M	American
5	28547.69	26.45	High	PhD	Rural	Sedan	F	African
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Table 4.5.1: Illustrative sample of five policies from the simulated motor insurance portfolio.

To gain further insight into the dataset, Fig. 4.B.1 and 4.B.2 in Appendix illustrate the distributions of the rating factors, presented in Table 4.5.1, with respect to the variable of interest N . Fig. 4.B.1 reports average claim counts across various categories, while Fig. 4.B.2 shows the empirical distributions of age and income across $N \in \{0, 1, 2, 3\}$.

Note that the synthetic portfolio was generated using a combination of marginal distributions and dependency structures. For example, vehicle power class and vehicle type were generated using gender-dependent probabilities, while geographic location was generated conditionally on ethnicity. Income was constructed as a

function of age, education, gender, ethnicity, location, vehicle characteristics, and their associated multiplicative effects, creating a network of dependencies among variables.

Interaction effects were also deliberately introduced when generating claim frequencies. Examples include Gender $\times$ Age, Age $\times$ Vehicle Type, Vehicle Power Class $\times$ Age, Geographic Location $\times$ Vehicle Power Class, and Education $\times$ Income. These interactions create non-linear relationships between protected and non-protected variables, such that seemingly legitimate rating factors become statistically associated with protected attributes. As a result, several risk factors act as proxy variables for protected characteristics. For instance, vehicle type and power class indirectly encode information about gender, while geographic location partially reflects ethnicity.

Because claim frequencies depend on these interacting factors, discrimination can emerge indirectly even when protected attributes are excluded from the pricing model. This design provides a realistic setting for evaluating fair actuarial models, as it reproduces both direct and indirect pathways through which sensitive information can influence premium predictions.

To evaluate whether the gradient-based fairness correction maintain its effectiveness beyond controlled simulation settings, we also consider a real-world car insurance portfolio publicly available at Roy (2021). After removing observations with missing values, the working dataset consists of 8 149 complete policies. Each contract is assigned a unit exposure $\nu_i = 1$, and the observed number of past accidents is used as the claim count, denoted N_i . Accordingly, we define the claim frequency as $\lambda_i = N_i/\nu_i$, which we model under a Poisson specification consistent with classical actuarial frequency modelling.

Table 4.5.2 presents four representative observations to illustrate the structure of the dataset. Each policy is described by two continuous variables (credit score and annual mileage) and a set of categorical attributes capturing demographic, socio-economic, and vehicle-related characteristics. These include the policyholder’s gender and race (the sensitive attributes in our analysis), age, driving experience, education level, income class, household composition, postal code, vehicle type, and traffic violations (speeding and DUI).

	Age	Gender	Race	Driving Exp.	Education	Income	Credit Score	...
1	65+	female	majority	30y+	university	upper class	0.592	...
2	40–64	female	majority	20–29y	high school	middle class	0.654	...
3	26–39	female	minority	10–19y	none	poverty	0.422	...
4	16–25	female	majority	0–9y	high school	poverty	0.344	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Table 4.5.2: Illustrative sample of four policies from the car insurance portfolio publicly available at Roy (2021).

For each dataset, we split the observations into three disjoint sets. A GLM and a LightGBM are trained on 60% of the data (in-sample training set) to obtain baseline predictive models for discriminatory premiums. The trained GLM and LightGBM models are used to predict discriminatory premiums on a first out-of-sample test set (with 20% observations). These discriminatory premiums are then corrected for fairness and balance (post-hoc autocalibration). A Random Forest model is finally trained on the first out-of-sample set to approximate the mapping from features to corrected premiums. This allows prediction of fairness-adjusted premiums for a held-out second out-of-sample test set (also with 20% observations), representing new policyholders not used in the training or correction stages.

The analysis presented in the following sections on the simulated data is reproduced for the real-world dataset in Appendix 4.F. As in the simulated study, enforcing DP leads to a clear fairness-accuracy trade-off. While the adjustment successfully reduces dependence between premiums and protected characteristics, it deteriorates predictive performance and increases claim count overestimation. The loss in accuracy is particularly pronounced when DP is enforced jointly across gender and ethnicity, as the (already limited) number of observations are further partitioned into smaller protected groups, reducing the reliability of the estimated group-specific distributions. The subsequent autocalibration step partially recovers actuarial balance and predictive accuracy without affecting DP.

4.5.2 Best-estimates

We consider a Generalized Linear Model (GLM) and a boosting algorithm (LightGBM) to model the premium best-estimates $\hat{\pi}$. The boosting model is tuned via `lightgbm` and `optuna`, with optimal hyper-parameters reported in Table 4.C.1.

Fig. 4.5.1 compares predicted and observed claim frequencies across 10 quantile bins. Bars represent total exposure within each bin, while the solid and dashed lines show the predicted $\hat{\lambda}$ and observed λ frequencies, respectively. A closer alignment between the two lines across bins indicates superior model performance. The GLM (left plot) systematically under- and over-predicts frequencies in several bins, whereas LightGBM (right plot) aligns much more closely with the empirical frequencies across all bins.

The Poisson deviance (a likelihood-based measure of predictive accuracy) is widely used to evaluate claim count regression models. It is defined as:

$$\mathcal{D}_{\text{Poisson}}(\lambda, \hat{\lambda}) = 2 \sum_{i=1}^n N_i \left(\frac{\nu_i}{N_i} \hat{\lambda}_i - \left(\ln \left(\frac{\nu_i}{N_i} \hat{\lambda}_i \right) + 1 \right) \mathbb{I}_{\{N_i \geq 1\}} \right), \quad (4.5.1)$$

where N_i , ν_i , and $\hat{\lambda}_i$ are respectively the i th policyholder's observed number of claims, observed exposure, and estimated claim frequency (per unit of exposure).

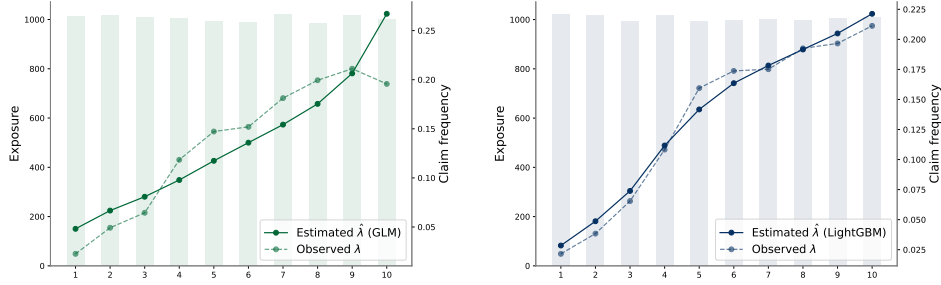


Figure 4.5.1: Comparison of predicted and observed claim frequencies per unit exposure across 10 quantile-based bins of predicted frequency. The solid line shows the estimated frequency $\hat{\lambda}$ from the model, while the dashed line shows the observed frequency λ computed from the data. The bars indicate the total exposure in each bin. Left plot: GLM; right plot: LightGBM.

Table 5.4.1 confirms that LightGBM has a lower deviance than the GLM, reflecting its ability to better capture the variation in the data by modelling nonlinear relationships and interactions more effectively than GLM (which is restricted to linear relationships unless explicitly specified).

Model	Poisson deviance
GLM	6625.6248
LightGBM	6539.1550

Table 4.5.3: Comparison of Poisson deviance values on the first out-of-sample test set for the GLM and the LightGBM. Lower deviance indicates a better fit to the observed claim frequencies.

Fig. 4.5.2 further illustrates the distribution of predicted premiums and highlight differences in risk allocation between the two models.

Given the models predictions $\hat{\lambda}$, the probability of observing exactly c claim(s) is given by Poisson probability mass function (pmf):

$$f(c, \lambda_i) = \mathbb{P}(N_i = c) = \frac{(\lambda_i)^c e^{-\lambda_i}}{c!},$$

with $\mathbb{P}(N_i > 0) = 1 - \mathbb{P}(N_i = 0) = 1 - e^{-\lambda_i}$. Summing these probabilities over all policies yields the expected total claim count for each claim level c . Table 4.5.4 compares observed and predicted counts. Although LightGBM better fits the quantile patterns (see Fig. 4.5.1) and deviance metrics (see Table 5.4.1), it overestimates by

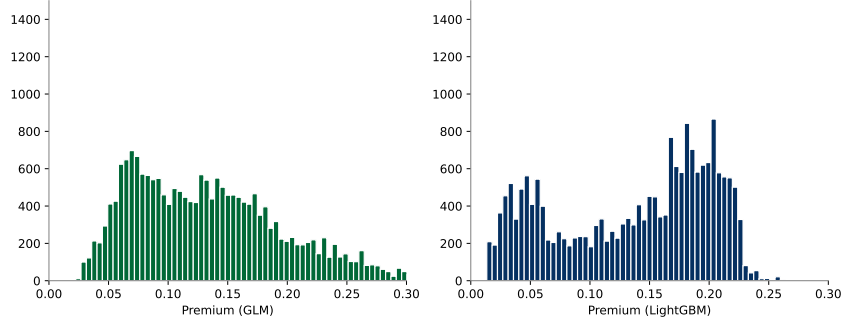


Figure 4.5.2: Histogram comparison of predicted premiums $\hat{\pi}$ for the GLM (left) and LightGBM (right). The plots illustrate the distribution of predicted premiums and highlight differences in risk allocation between the two models.

13 the total number of policies with at least one claim $c > 0$. This overestimation bias is a well-known consequence of minimising the Poisson deviance outside the canonical GLM setting (Denuit, Charpentier, et al. 2021).

c	Observed	GLM	LightGBM	GLM gap	LightGBM gap
0	18719	18717.5133	18705.6169	-1.4867	-13.3831
1	1221	1213.9650	1224.7363	-7.0350	3.7363
2	56	65.2688	66.5456	9.2688	10.5456
3	4	3.1122	2.9848	-0.8878	-1.0152
> 0	1281	1282.4867	1294.3831	1.4867	13.3831

Table 4.5.4: Comparison between observed and predicted claim counts for the LightGBM and GLM. The column c denotes the number of claims per policy, and the gap columns show the difference between predicted and observed counts within each group. The aggregated row ($c > 0$) summarises the total predicted versus observed claim counts for policies with at least one claim. Highlighted value indicates the total gap.

We can further check the estimation bias by running a simple Poisson regression of the intercept ($\hat{\beta}_0 = -2.0023$) and consider its prediction per unit of exposure $\hat{\lambda} = \bar{\pi} = 0.13502$ as the true premium mean. In Table 4.5.5, we compare it with the GLM and LightGBM average premium values $\bar{\pi} = \frac{1}{n} \sum_{i=1}^n \hat{\lambda}_i$ (Denuit, Charpentier, et al. 2021). While both models are close to the true premium mean $\bar{\pi} = 0.13502$, LightGBM does slightly overestimate the average premium.

The negative bias (or overestimation) observed in the LightGBM predictions can be corrected with a multiplicative factor $\mathbb{E}[Y|\hat{\pi}(\mathbf{X}) = F_{\hat{\pi}}^{-1}(u)] / F_{\hat{\pi}}^{-1}(u)$ pro-

posed by Denuit, Charpentier, et al. (2021). This balance correction, detailed in Appendix 4.D, is as key post-hoc step of the fairness adjustment presented in the following Sections 4.5.3 and 4.5.4.

Statistic	GLM _{β_0}	GLM	LightGBM
Mean $\bar{\pi}$	0.1350	0.1350	0.1363
10% Quantile	–	0.0602	0.0391
90% Quantile	–	0.2288	0.2115

Table 4.5.5: Summary statistics of the expected premiums $\hat{\pi}$ for the baseline intercept-only GLM (GLM _{β_0}), the full GLM, and the LightGBM. Missing quantiles for GLM _{β_0} are denoted by – because it produces a constant expected premium across policies. The highlighted value indicates that the LightGBM’s mean prediction diverges from the baseline mean.

The best-estimate premiums $\hat{\pi}(\mathbf{A}, \mathbf{D})$, presented in Section 4.5.2 and in Appendix 4.D, are discriminatory with respect to sensitive attributes such as gender.

Following the theoretical framework outlined in Section 4.4, our goal is to enforce DP by aligning the distributions of predicted premiums across $G \geq 2$ demographic groups. We first illustrate this mechanism for a binary sensitive attribute in Section 4.5.3, and extend the approach to settings where $G > 2$ in Section 4.5.4.

4.5.3 Fairness adjustment $G=2$

In the binary setting $G = 2$, the predicted (discriminatory) premiums $\hat{\pi}$ are partitioned into two demographic groups $\{\hat{\pi}_g\}_{g=1}^G$ corresponding to $g = \{\text{Female}, \text{Male}\}$. These premiums are iteratively adjusted using algorithm 9 to equalise not only the means, but also the dispersion and overall shape of their distributions.

Fig. 4.5.3 illustrates the distributions of predicted premiums before and after fairness adjustment across the two demographic groups. Solid lines correspond to the original discriminatory premiums, while dashed lines represent the premiums after fairness correction.

Prior to adjustment, the distributions clearly differ between genders, with females (in red) concentrated in the lower-premium range and males (in blue) in the higher range, reflecting gender-based pricing differences.

After fairness adjustment, the distributions for both groups nearly overlap, demonstrating that our Energy-based fairness correction effectively mitigates gender-based disparities.

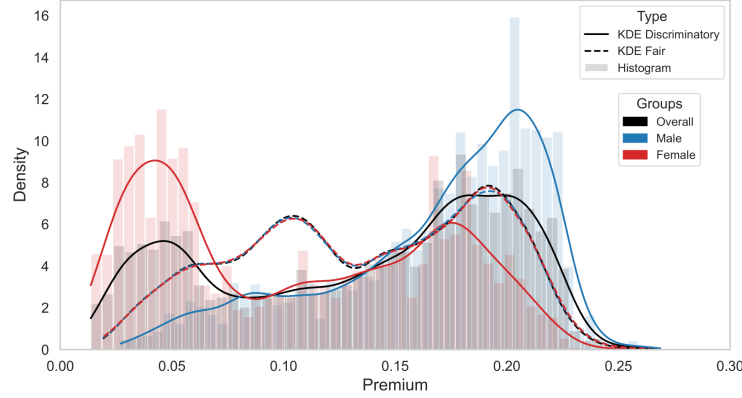


Figure 4.5.3: Kernel density estimates (KDEs) of predicted premiums by gender. Solid lines represent the original LightGBM discriminatory premiums, while dashed lines correspond to the fairness-adjusted premiums. The male group is shown in blue, the female group in red, and the overall portfolio in black.

Enforcing independence between premiums and gender (a key risk factor) constrains the predictions and prevents the model from fully exploiting gender-related risk information. As illustrated in Fig. 4.5.4, the model inevitably sacrifices some predictive accuracy to enforce DP. As later detailed in Table 4.5.10, the deterioration in predictive accuracy remains however modest, with a +0.62% increase in deviance relative to the GLM benchmark.

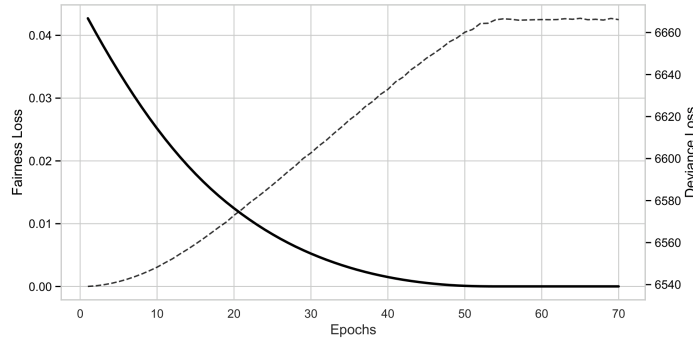


Figure 4.5.4: Evolution of the fairness loss (in Eq. (4.3.5)) and the Poisson deviance loss (in Eq. (4.5.1)) over training epochs.

While the fairness adjustment inevitably alters the original premium distribution, Fig. 4.E.1 in Appendix illustrates the crucial role of the intra-group terms

$\mathbb{E}[|U - U'|]$ and $\mathbb{E}[|V - V'|]$ in our fairness adjustment. Omitting these terms collapses the distributional spread (see last column), and leads to highly concentrated premium with unnaturally sharp and narrow peaks. This results in a poor model accuracy, as the fairness-adjusted premiums do not respect the natural distributional spread of the initial (discriminatory) premiums $\hat{\pi}$. Including these intra-group terms (see middle column) preserves the predictions dispersion, reflecting the model's capacity to actuarially discriminate between policyholders, thereby ensuring a meaningful DP adjustment, and highlighting the suitability of the Energy distance for this type of distributional fairness correction.

Tables 4.5.7 and 4.5.6 further reveal that the fairness adjustment also leads to an increase in the overestimation of the overall claim count (from 13 to 19), and to a larger deviation from the true mean $\mu = 0.13502$.

Statistic	GLM $_{\beta_0}$	LightGBM	Fair LightGBM
Mean $\bar{\pi}$	0.1350	0.1363	0.1366
10% Quantile	—	0.0391	0.0587
90% Quantile	—	0.2115	0.2050

Table 4.5.6: Summary statistics of expected premiums $\hat{\pi}$ for the baseline intercept-only GLM (GLM $_{\beta_0}$), the standard LightGBM, and the fairness-adjusted LightGBM. Highlighted values indicate divergence from the target mean premium, indicating the effect of fairness adjustment.

c	Observed	LightGBM	Fair LightGBM	LightGBM gap	Fair LightGBM gap
0	18719	18705.6169	18699.0905	-13.3831	-19.9095
1	1221	1224.7363	1234.2958	3.7363	13.2958
2	56	66.5456	63.7865	10.5456	7.7865
3	4	2.9848	2.7249	-1.0152	-1.2751
> 0	1281	1294.3831	1300.9095	13.3831	19.9095

Table 4.5.7: Comparison of observed claim counts with predictions from standard LightGBM and fairness-adjusted LightGBM. The gap columns show the difference between predicted and observed counts. Highlighted values indicate the aggregated total gaps, illustrating the impact of the fairness adjustment on total balance.

In Fig. 4.5.5, a noticeable drop below the diagonal is observed (right plot) for higher levels of s , which may explain the increased negative bias detected after fairness adjustment in Table 4.5.6. A detailed view of the local bias is provided in Fig. 4.E.2 (in Appendix).

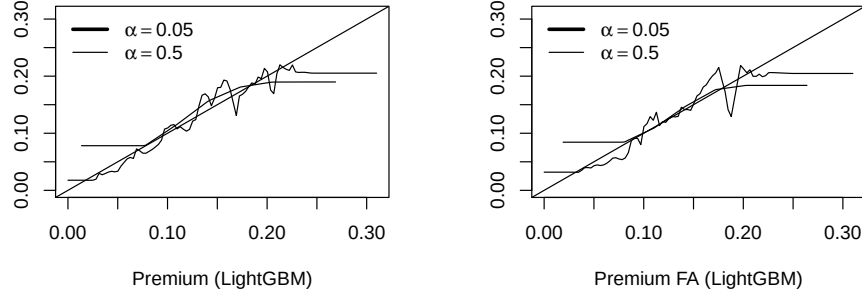


Figure 4.5.5: Visualisation of $s \mapsto \mathbb{E}[Y | \hat{\pi}(\mathbf{X}) = s]$ before and after fairness adjustment (FA) across the gender variable.

The negative bias (or overestimation) observed in the fair LightGBM predictions can be corrected using the multiplicative bias-correction factor proposed by Denuit, Charpentier, et al. (2021) detailed in Appendix 4.D:

$$\frac{\mathbb{E}[Y | \hat{\pi}(\mathbf{X}) = F_{\hat{\pi}}^{-1}(x)]}{F_{\hat{\pi}}^{-1}(x)},$$

and estimated using a local GLM smoother with bandwidth $\alpha = 0.16$.

Table 4.F.5 reports summary statistics before and after bias correction. The average premium shifts from 0.1366 to 0.1351, bringing it close to the reference mean 0.13502.

Statistic	GLM _{β_0}	LightGBM	Fair LightGBM	Fair BC LightGBM
Mean $\bar{\pi}$	0.1350	0.1363	0.1366	0.1351
10% Quantile	—	0.0391	0.0587	0.0461
90% Quantile	—	0.2115	0.2050	0.2035

Table 4.5.8: Summary statistics of expected premiums $\hat{\pi}$ for the baseline intercept-only GLM (GLM _{β_0}), standard LightGBM, fairness-adjusted LightGBM, and bias-corrected fairness-adjusted LightGBM. Highlighted values indicate divergence from the target mean premium, illustrating the effect of fairness adjustment and subsequent bias correction.

Table 4.5.9 shows the improvement in predicted claim counts. The gap in the number of policies with at least one claim drops from 19.91 to 5.55.

c	Observed	Fair LightGBM	Fair BC LightGBM	Fair LightGBM gap	Fair BC LightGBM gap
0	18719	18699.0905	18713.4489	-19.9095	-5.5511
1	1221	1234.2958	1221.2914	13.2958	0.2914
2	56	63.7865	62.5397	7.7865	6.5397
3	4	2.7249	2.6244	-1.2751	-1.3756
> 0	1281	1300.9095	1286.5511	19.9095	5.5511

Table 4.5.9: Comparison of observed claim counts with predicted claims from the fairness-adjusted LightGBM before and after bias correction. The gap columns show the difference between predicted and observed counts. Highlighted values indicate the aggregated total gaps, illustrating the reduction of prediction bias after applying balance correction.

Table 4.5.10 summarises the Poisson deviance values. We note that the fairness-accuracy trade-off manifests at two distinct levels. First, a slight degradation in predictive accuracy (individual deviance fit) from 6539.155 to 6666.717 immediately after the fairness adjustment. The fairness correction inevitably reduces predictive accuracy to some extent, as the optimisation process prioritizes DP over likelihood-based fit. Enforcing independence between premiums and gender (a key risk factor) further prevents the model from exploiting gender-related risk information. Second, a minor imbalance in aggregate calibration (alignment between predicted and observed total claims shown in Table 4.5.7). To address the latter, we introduced an additional post-hoc step, which partially restores, not only the actuarial balance, but also the predictive accuracy (6654.98) that was temporarily reduced by the fairness correction, without affecting fairness, as shown in Fig. 4.5.6.

Model	Poisson Deviance	Δ GLM	Δ LightGBM
GLM	6625.6248		
LightGBM	6539.1550	-1.3051%	
LightGBM BC	6542.6450		
LightGBM fair (gender)	6666.7170	+0.6202%	+1.9507%
LightGBM fair BC (gender)	6654.9800	+0.4430%	+1.7713%

Table 4.5.10: Poisson deviance comparison at the same evaluation stage (out-of-sample 1) for the generalised linear model (GLM), bias-corrected (BC), and fairness-adjusted versions of the LightGBM model. Lower deviance indicates a better fit to the observed claim frequencies. Δ percentages compare each fairness (and balance) corrected model relative deterioration with the GLM and LightGBM.

Fig. 4.5.6 shows the local redistribution of density after balance correction

(dotted lines). It further confirms that, even after balance correction, fairness is maintained as the conditional premium densities for both gender groups remain well-aligned.

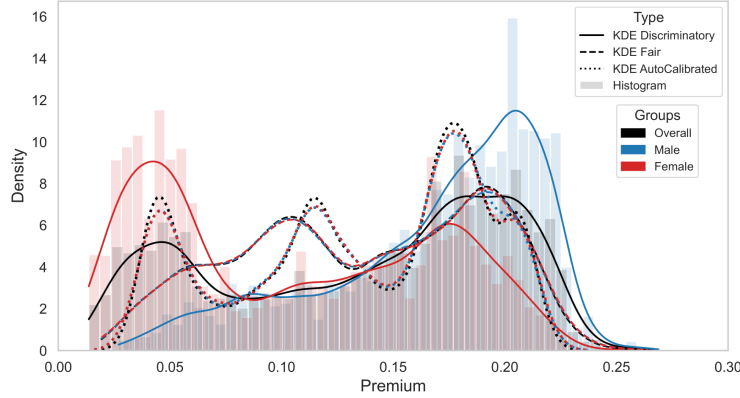


Figure 4.5.6: Kernel density estimates (KDEs) of predicted premiums by gender. Solid lines represent the original LightGBM discriminatory premiums, dashed lines the fairness-adjusted premiums, and dotted lines the fairness-adjusted and balance-corrected premiums. The male group is shown in blue, the female group in red, and the overall portfolio in black.

Table 4.5.11 summarises several key fairness metrics discussed in the literature. These include the Wasserstein distance $W_p(U, V)$ extensively studied by Hu et al. (2023), Hu et al. (2024), Charpentier (2023), and Charpentier et al. (2023), and inspired by Chzhen et al. (2020), the MMD and Sinkhorn divergence $S_\epsilon(U, V)$ proposed by Oneto et al. (2020), as well as widely adopted metrics in the statistical and causal theory (see e.g., Charpentier (2023) and Deka et al. (2023)) such as the Group Fairness in Expectation Δ_{GFE} and Statistical Parity Δ_{SP} . We further provide the Wasserstein barycenter and MMD-based adjustments benchmarks for comparison in Appendix 4.G. The comparison with the Wasserstein benchmark highlights important differences in how DP is enforced. Although both methods reduce group disparities, the Wasserstein correction achieves parity through quantile-level redistribution, resulting in larger distortions among high-risk policyholders and weaker rank preservation. In contrast, the proposed energy-based adjustment results in larger distortions among lower-risk policyholders, better preserves the original risk ranking, and achieves stronger conditional parity for high-risk policyholders.

Metric with $U = \hat{\pi}_{\text{Female}}$ and $V = \hat{\pi}_{\text{Male}}$	$\hat{\pi}$	$\hat{\pi}^{\text{fair, BC}}$
$\mathcal{E}(U, V) = 2\mathbb{E}[\ U - V\] - \mathbb{E}[\ U - U'\] - \mathbb{E}[\ V - V'\]$	0.042704	8e-06
$W_p(U, V) = (\inf_{\gamma \in \Gamma(U, V)} \mathbb{E}_{(U, V) \sim \gamma} \ U - V\ ^p)^{1/p}$	($p = 1$) 0.064760 ($p = 2$) 0.070531	0.001529 0.002113
$\text{MMD}^2(U, V) = \mathbb{E}[k(U, U')] + \mathbb{E}[k(V, V')] - 2\mathbb{E}[k(U, V)]$	0.205325	2e-06
$S_{\epsilon=0.1}(U, V) = W_{\epsilon}(U, V) - \frac{1}{2}W_{\epsilon}(U, U) - \frac{1}{2}W_{\epsilon}(V, V)$	0.028611	2.2e-05
$\Delta_{GFE} = \mathbb{E}[U] - \mathbb{E}[V] $	0.062959	0.000155
$\Delta_{SP} = \sup_s P(U \leq s) - P(V \leq s) $	0.396203	0.030354

Table 4.5.11: Comparison of key fairness metrics before and after $\mathcal{E}$ -based fairness adjustment and balance correction. MMD is computed using a Gaussian RBF kernel $k(u, v) = \exp(-\|u - v\|^2 / (2\sigma^2))$ with fixed bandwidth $\sigma = 0.1$.

For new policyholders, we extend this methodology by applying an individual correction factor γ_i , as outlined in Section 4.4.2. A random forest regressor learns the mapping from policyholder characteristics to correction factors, generalising the fairness adjustment mechanism beyond the dataset (out-of-sample 1) used to train the adjuster.

Fig. 4.5.7 shows that the adjusted predictions for new policyholders (out-of-sample 2) remain consistent across sensitive groups with the results observed in Fig. 4.5.6, with only minor deviations.

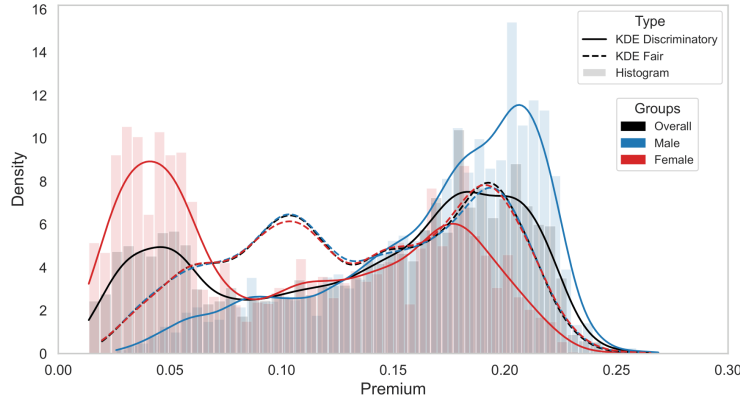


Figure 4.5.7: Kernel density estimates (KDEs) of predicted premiums by gender for new policyholders. Solid lines represent the LightGBM discriminatory premiums, and dashed lines the fairness-adjusted premiums. The male group is shown in blue, the female group in red, and the overall portfolio in black.

Table 4.5.12 summarises the Poisson deviance results before and after fairness correction on the new test set (out-of-sample 2) with 20 000 policies. The fair

LightGBM model exhibits slightly higher deviance than the unconstrained LightGBM on both out-of-sample sets (+1.95% and +2.60% relative to LightGBM), consistent with typical accuracy–fairness trade-offs. All models, including GLM and baseline LightGBM, deteriorate when predicting genuinely new policyholders (Δ of +3.8–4.13%), indicating that a portion of the deviance increase for the new policyholders (from +1.95% to +2.60%) stems from intrinsic out-of-sample difficulty rather than instability introduced by the fairness mechanism (Hainaut 2025).

	in-sample	out-of-sample 1	out-of-sample 2	Δ
GLM	19997.9223	6625.6248	6880.0621	+3.8102%
LightGBM	19539.5379	6539.1550 (-1.3051%)	6809.2512 (-1.0292%)	+4.1304%
LightGBM fair	–	6666.7170 (+1.9507%) (+0.6202%)	6986.1686 (+2.5982%) (+1.5422%)	+4.7917%

Table 4.5.12: Poisson deviance comparison across models: (i) in-sample (training) set ($n = 60,000$) to fit the discriminatory GLM and LightGBM models, (ii) a first out-of-sample (test) set ($n = 20,000$) to predict discriminatory premiums and correct for fairness and autocalibration, and (iii) a second out-of-sample (test) set ($n = 20,000$) of new policyholders, for which corrected premiums were predicted using a Random Forest trained on the first test set. Percentages in parentheses compare each model with the GLM and LightGBM at the same evaluation stage. Δ measures the relative deterioration when moving from the first to the second out-of-sample (new policyholders).

4.5.4 Fairness adjustment $G > 2$

As discussed in Section 4.4, our $\mathcal{E}$ -adjuster can simultaneously adjust premiums across more than two sensitive groups. For $G > 2$, the predicted premiums $\hat{\pi}$ are partitioned into G vectors $\{\hat{\pi}_g\}_{g=1,\dots,G}$ of sizes n_g organized in d -dimensional variables U and V as in Eq. (4.4.4). In this setting, equal-sized samples are drawn from each group in U , and for each group in V . A uniform random permutation is also applied to ensure the pairwise comparisons span all group combinations (see algorithm 9).

Fig. 4.5.8 illustrates the fairness adjustment of $\hat{\pi}$ across six demographic groups, determined by the policyholders’ gender (female or male) and age (young, middle-aged, or senior). The solid lines correspond to the initial predicted premiums partitioned into G demographic groups $\{\hat{\pi}_g\}_{g=1}^6$. The overlapping dashed lines correspond to the corresponding fairness adjusted premiums $\{\hat{\pi}_g^{\text{fair}}\}_{g=1}^6$. We observe that after adjustment, the dashed lines largely overlap within each gender-age pairing, indicating that the $\mathcal{E}$ -adjuster has effectively reduced disparity across these six groups, while preserving the spread of the initial distribution (in black).

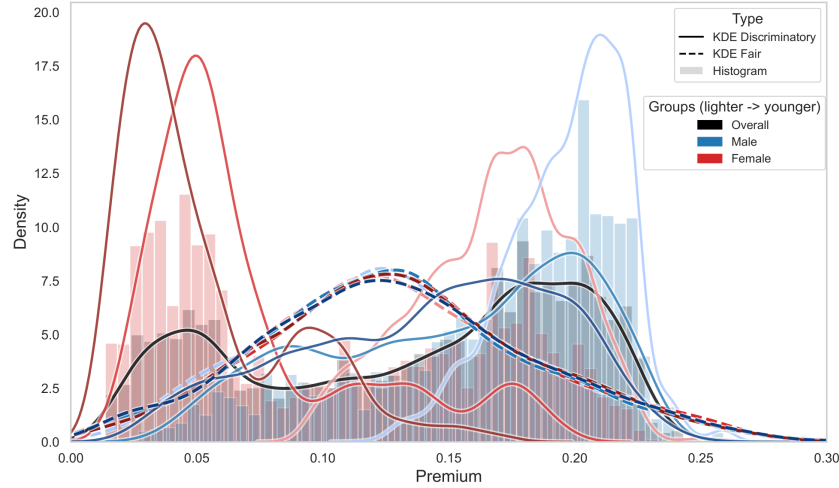


Figure 4.5.8: Kernel density estimates (KDEs) of predicted premiums by gender and age group. Solid lines represent the LightGBM discriminatory premiums, and dashed lines the fairness-adjusted premiums. The male groups are shown in blue, the female groups in red, and the overall portfolio in black. The color shades correspond to age groups. Histogram correspond to the overall empirical frequencies of female, male, and overall discriminatory premiums.

Tables 4.5.13 and 4.5.14 reveal that the fairness adjustment leads to a significant underestimation (around -42) of the overall number of claims, and a larger deviation from the mean value.

c	Observed	LightGBM	Fair LightGBM	LightGBM gap	Fair LightGBM gap
0	18719.0	18705.6169	18761.1463	-13.3831	42.1463
1	1221.0	1224.7363	1178.2408	3.7363	-42.7592
2	56.0	66.5456	58.1117	10.5456	2.1117
3	4.0	2.9848	2.4107	-1.0152	-1.5893
> 0	1281.0	1294.3831	1238.8537	13.3831	-42.1463

Table 4.5.13: Comparison of observed claim counts with predictions from standard LightGBM and fairness-adjusted LightGBM. The gap columns show the difference between predicted and observed counts. Highlighted values indicate the aggregated total gaps, illustrating the impact of the fairness adjustment on total balance.

Statistic	GLM $_{\beta_0}$	LightGBM	Fair LightGBM
Mean $\bar{\pi}$	0.13502	0.136334	0.129824
10% Quantile		0.039125	0.061754
90% Quantile		0.211547	0.203506

Table 4.5.14: Summary statistics of expected premiums $\hat{\pi}$ for the baseline intercept-only GLM (GLM $_{\beta_0}$), standard LightGBM, and fairness-adjusted LightGBM. Highlighted values indicate divergence from the target mean premium, illustrating the effect of fairness adjustment.

For $\alpha = 0.02$, we manage to significantly improve the underestimation after balance correction. In Table 4.5.15 we observe a smaller divergence from the true mean $\mu = 0.13502$, from 0.1298 to 0.1341, and in Table 4.5.16 we observe a decrease in the overall overestimation from -42 to 6.94 claims.

Statistic	GLM $_{\beta_0}$	LightGBM	Fair LightGBM	Fair BC LightGBM
Mean $\bar{\pi}$	0.13502	0.136334	0.129824	0.134132
10% Quantile		0.039125	0.061754	0.107966
90% Quantile		0.211547	0.203506	0.160423

Table 4.5.15: Summary statistics of expected premiums $\hat{\pi}$ for the baseline intercept-only GLM (GLM $_{\beta_0}$), standard LightGBM, fairness-adjusted LightGBM, and bias-corrected fairness-adjusted LightGBM. Highlighted values indicate divergence from the target mean premium, illustrating the effect of fairness adjustment and subsequent bias correction.

c	Observed	Fair LightGBM	Fair BC LightGBM	Fair LightGBM gap	Fair BC LightGBM gap
0	18719.0	18761.1463	18712.0574	42.1463	-6.9426
1	1221.0	1178.2408	1230.1127	-42.7592	9.1127
2	56.0	58.1117	55.8166	2.1117	-0.1834
3	4.0	2.4107	1.9556	-1.5893	-2.0444
> 0	1281.0	1238.8537	1287.9426	-42.1463	6.9426

Table 4.5.16: Comparison of observed claim counts with predicted claims from the fairness-adjusted LightGBM before and after bias correction. The gap columns show the difference between predicted and observed counts. Highlighted values indicate the aggregated total gaps, illustrating the reduction of prediction bias after applying balance correction.

Table 4.5.17 summarises the deviance loss before and after fairness adjustment (for $G = 2$ and $G = 6$) and balance correction. As expected, the deviance deteriorates much more when adjusting across six demographic groups (gender and age jointly) compared to two (gender only). This reflects the increasing difficulty of aligning multiple conditional premium distributions simultaneously, especially when these attributes are strong predictors. The post-hoc autocalibration step does recover a non-negligible portion of predictive accuracy. For example, for the six-group setting, the deviance decreases from 7341.93 to 6891.32 after balance correction (i.e., +4% relative to the GLM benchmark), showing that the post-hoc calibration mitigates a substantial share of the initial accuracy loss.

Model	Poisson Deviance	Δ GLM	Δ LightGBM
GLM	6625.6248		
LightGBM	6539.1550	-1.3051%	
LightGBM BC	6542.6450		
LightGBM fair (gender)	6666.7170	+0.6202%	+1.9507%
LightGBM fair BC (gender)	6654.9800	+0.4430%	+1.7713%
LightGBM fair (age & gender)	7341.9258	+10.8111%	+12.2764%
LightGBM fair BC (age & gender)	6891.3200	+4.0101%	+5.3855%

Table 4.5.17: Poisson deviance comparison across models at the same evaluation stage (out-of-sample 1). Δ percentages compare each fairness (and balance) corrected model relative deterioration with the GLM and LightGBM.

Insurers should be aware that the fairness adjustment may disproportionately affect smaller or extreme subgroups. Table 4.5.18 presents the Poisson deviance for each gender-age subgroup, comparing the baseline LightGBM predictions with the fairness-adjusted and balance-corrected models. The relative change (%) in Poisson deviance quantifies the trade-off between enforcing DP and preserving predictive accuracy across demographic group. Fairness adjustment alone deteriorates deviance for all groups, with the largest relative increases observed for young males (23.36%) and senior females (24.71%). This reflects that substantial redistribution of predicted premiums is needed in these groups with higher-frequencies to achieve fairness, which naturally affects predictive accuracy.

Applying the post-hoc balance correction substantially mitigates these increases. For example, young males' deviance increase drops from 23.36% to 2.37%, and young females' from 21.10% to 1.35%. Similarly, middle-aged and senior male groups also see their relative increases fall below 5%, indicating that the BC step effectively restores predictive accuracy (while retaining the fairness adjustment).

Interestingly, the post-hoc balance correction slightly amplifies the increase in deviance for middle-aged and senior female policyholders. This highlights the difficulty of retaining predictive accuracy for lower claim frequency groups (see darker

red KDEs in Fig. 4.5.8 where a substantial mass of the initial premium distributions is around 0).

Overall, the results highlight that, although fairness enforcement inevitably perturbs the original predictions, especially for groups with higher claim frequencies (i.e., young policyholders), the post-hoc autocalibration mitigates much of the associated loss.

Gender	Age group	n	LightGBM	Fair LightGBM	Fair BC LightGBM	% Fair	% Fair BC
Female	Middle-aged	3,391	708.7603	812.7800	820.6330	14.6763%	15.7843%
Female	Senior	2,636	451.4856	563.0406	572.7071	24.7084%	26.8495%
Female	Young	4,049	1,631.4034	1,975.6220	1,653.4445	21.0995%	1.3510%
Male	Middle-aged	3,328	1,184.7231	1,384.3420	1,204.5646	16.8494%	1.6748%
Male	Senior	2,644	849.8874	974.0947	886.5005	14.6146%	4.3080%
Male	Young	3,952	1,712.8952	2,113.0673	1,753.4705	23.3623%	2.3688%

Table 4.5.18: Poisson deviance by demographic group for LightGBM, fairness-adjusted LightGBM, and fairness-adjusted LightGBM with balance correction (BC). Percentage increases relative to LightGBM are shown for both adjustments.

Calibration tables in Appendix 4.E further show the impact of fairness corrections by subgroup. For senior policyholders, females (Table 4.E.3) experience a substantial degradation in calibration after fairness adjustment, with deviations of -103.00 ($c = 0$), whereas males (Table 4.E.6), despite being similarly represented, show negligible changes (2.78). This also reflects that substantial redistribution of predicted premiums is needed in the lower claim frequencies groups (see senior and middle-aged KDEs in Fig. 4.5.8) to achieve fairness, which naturally affects calibration.

4.6 Discussion

In this chapter, we introduce a novel gradient-based group fairness correction using the multivariate Energy distance to enforce demographic parity and reduce group-level disparities across multiple sensitive attributes simultaneously. The multivariate nature of the Energy distance is a key advantage of our approach. It allows for simultaneous fairness adjustments across protected variables with different numbers of sensitive classes, such as gender (binary) and discretised age (multi-class), and is therefore order-independent.

As expected, the loss in accuracy increases when the number of demographic groups grows, reflecting the greater difficulty of simultaneously aligning multiple conditional premium distributions, especially when enforcing fairness for key risk factors, such as gender and age in car pricing. However, the predictive accuracy is partially recovered through a post-hoc autocalibration step (Denuit, Charpentier,

et al. 2021) introduced to correct biases in the total predicted claim count after fairness adjustments.

We further introduce a simple out-of-sample mapping mechanism to apply fairness corrections to new policyholders. While the random forest used to map policyholder characteristics to individual correction factors does require training, it is important to distinguish this lightweight post-processing step from retraining the underlying pricing model. The LightGBM estimator and gradient-based fairness correction are not refitted once the fairness-adjusted premiums have been produced. As with any predictive model, some degradation is observed when moving to genuinely new policyholders, but our results indicate that this deterioration is comparable to the baseline in/out-of-sample gap observed for both GLM and unconstrained LightGBM models.

While gradient descent introduces an optimisation layer, the process remains transparent and explainable to regulators for several reasons.

First, the LightGBM model that provides the initial discriminatory premium predictions is itself a standard, interpretable model in actuarial practice, with feature importances and partial dependence plots readily available to explain its drivers.

Second, the fairness adjustment is applied as a post-hoc correction, rather than as an in-processing constraint embedded inside the model fitting. We believe this distinction is important for transparency. In-processing methods typically modify the loss function or introduce latent constraints during training, making each change in the predictions less interpretable. In contrast, in our post-processing setup, the correction is defined by a single, well-specified quantitative objective, that is, minimising the Energy distance between conditional premium distributions across sensitive groups. At every iteration, the magnitude and direction of the adjustment can be quantified as a deterministic transformation of the original predictions, allowing practitioners to track (and justify) fairness improvements step by step.

Moreover, this Energy distance itself has a well-established statistical interpretation as a measure of disparity between distributions. The between-group term encourages the alignment of premiums across groups (reducing disparities between U and V), while the within-group terms preserve the internal variability of each group. This structure ensures that the correction smooths inter-group differences (improving inter-group fairness) without collapsing intra-group heterogeneity/diversity, thereby maintaining portfolio diversity and the credibility structure of the premiums.

Third, the post-hoc autocalibration step (following Denuit, Charpentier, et al. (2021)) adds a local GLM correction to the adjusted premiums. GLMs are a well-documented and interpretable regression framework in actuarial practice, and this step systematically restores balance with actual claims while preserving actuarial soundness.

We believe that the combination of these sequential steps makes the contribu-

tion of each fairness or calibration correction transparent.

Besides this transparency concern, another practical challenge in fair insurance pricing arises when protected attributes are not available. Recent works have addressed this issue explicitly. For instance, Gabric et al. (2024) propose a variational Bayesian framework for discrimination-free pricing in the absence of observed sensitive attributes, while T. Zhang et al. (2025) consider settings where sensitive attributes are available only in a noisy or privatized form. These approaches highlight an important distinction between fairness under full observability and fairness under partial access to protected information.

In the context of the proposed energy-distance-based framework, the case of noisy or privatized sensitive attributes can in principle be accommodated by replacing exact group membership with its observed corrupted counterpart, leading to fairness constraints defined over induced noisy group partitions. However, the resulting fairness penalty is then aligned with the distribution of the observed (noisy) attribute rather than the true latent protected attribute, introducing an additional layer of measurement error.

By contrast, the fully unobserved case is fundamentally more challenging. In the absence of any direct information on protected attributes, enforcing group-wise distributional alignment requires introducing additional modelling assumptions, such as latent group structure or the use of proxy variables correlated with the protected attribute. This includes approaches where sensitive attributes are inferred from observable features (e.g., name or location based inference as in the Colorado BIFSG framework), but such strategies raise well-known concerns regarding identifiability, stability, and the normative validity of the resulting fairness guarantees, since the fairness constraint becomes dependent on an estimated rather than observed protected structure, which may itself embed bias and introduce additional layers of uncertainty.

4.A Proofs

Proof of Proposition 4.1. The Cramér distance is unique among ℓ_p^p distances as its sample gradient is an unbiased estimator of the population gradient. This property, established by Bellemare et al. (2017, p. 16), is crucial in stochastic gradient-based optimisation when probability measures are used as loss functions.

Let P be a probability distribution, and let $\{Q_\theta\}_{\theta \in \Theta}$ denote a differentiable parametric family of probability measures with cumulative distribution functions $F_P(u)$ and $F_{Q_\theta}(u)$. The squared ℓ_2 distance admits the representation:

$$\ell_2^2(P, Q_\theta) = \int_{-\infty}^{+\infty} (F_{Q_\theta}(u) - F_P(u))^2 du.$$

By definition, differentiating with respect to θ gives:

$$\nabla_\theta \ell_2^2(P, Q_\theta) = \nabla_\theta \int_{-\infty}^{+\infty} (F_{Q_\theta}(u) - F_P(u))^2 du.$$

Under the hypothesis that the convergence of the square follows from the convergence of the ordinary values (see e.g., the Leibniz rule under dominated convergence), we may swap differentiation and integration:

$$= \int_{-\infty}^{+\infty} \nabla_\theta (F_{Q_\theta}(u) - F_P(u))^2 du.$$

Application of the chain rule of differentiation then gives:

$$= \int_{-\infty}^{+\infty} 2(F_{Q_\theta}(u) - F_P(u)) \nabla_\theta F_{Q_\theta}(u) du.$$

By Lemma 1 in Bellemare et al. (2017, p. 14), we replace the true CDF $F_P(u)$ by its expected empirical CDF $\mathbb{E}_{\mathbf{U}_m \sim P}[F_{\hat{P}_m}(u)]$:

$$= \int_{-\infty}^{+\infty} 2(F_{Q_\theta}(u) - \mathbb{E}_{\mathbf{U}_m \sim P}[F_{\hat{P}_m}(u)]) \nabla_\theta F_{Q_\theta}(u) du.$$

By linearity of expectation:

$$= \int_{-\infty}^{+\infty} 2 \mathbb{E}_{\mathbf{U}_m \sim P} (F_{Q_\theta}(u) - F_{\hat{P}_m}(u)) \nabla_\theta F_{Q_\theta}(u) du.$$

By Fubini's theorem (Royden et al. 1988), we interchange the order of integration and expectation:

$$\begin{aligned} &= \mathbb{E}_{U_m \sim P} \int_{-\infty}^{+\infty} 2 \left(F_{Q_\theta}(u) - F_{\hat{P}_m}(u) \right) \nabla_\theta F_{Q_\theta}(u) du \\ &= \mathbb{E}_{U_m \sim P} \left[\nabla_\theta \ell_2^2(\hat{P}_m, Q_\theta) \right]. \end{aligned}$$

□

Proof of Proposition 4.2. Following Bellemare et al. (2017, p. 16), the Energy distance between two probability measures P and Q_θ is defined as:

$$\mathcal{E}(P, Q) = 2\mathbb{E}\|U - V\| - \mathbb{E}\|U - U'\| - \mathbb{E}\|V - V'\|,$$

where $X \sim P$, $Y \sim Q_\theta$, and U', V' are independent copies of U and V , respectively. For the empirical measure $\hat{P}_m$, we write $\hat{U}_m \sim \hat{P}_m$. The gradient of $\mathcal{E}(P, Q_\theta)$ w.r.t. θ is:

$$\begin{aligned} \nabla_\theta \mathcal{E}(P, Q_\theta) &= \nabla_\theta \left(2\mathbb{E}\|U - V\| - \mathbb{E}\|U - U'\| - \mathbb{E}\|V - V'\| \right) \\ &= 2\nabla_\theta \mathbb{E}\|U - V\| - \nabla_\theta \mathbb{E}\|U - U'\| - \nabla_\theta \mathbb{E}\|V - V'\|. \end{aligned}$$

Since $\mathbb{E}\|U - U'\|$ does not depend on θ , this simplifies to:

$$= 2\nabla_\theta \mathbb{E}\|U - V\| - \nabla_\theta \mathbb{E}\|V - V'\|.$$

Similarly, for the empirical measure $\hat{P}_m$, we have:

$$\begin{aligned} \nabla_\theta \mathcal{E}(\hat{P}_m, Q_\theta) &= \nabla_\theta \left(2\mathbb{E}\|\hat{U}_m - V\| - \mathbb{E}\|\hat{U}_m - \hat{U}'_m\| - \mathbb{E}\|V - V'\| \right) \\ &= 2\nabla_\theta \mathbb{E}\|\hat{U}_m - V\| - \nabla_\theta \mathbb{E}\|V - V'\|. \end{aligned}$$

All together, we have:

$$\begin{aligned} \nabla_\theta \mathcal{E}(P, Q_\theta) &= \mathbb{E}_{U_m \sim P} \nabla_\theta \mathcal{E}(\hat{P}_m, Q_\theta) \\ 2\nabla_\theta \mathbb{E}\|U - V\| - \nabla_\theta \mathbb{E}\|V - V'\| &= 2\mathbb{E}_{U_m \sim P} \nabla_\theta \mathbb{E}\|\hat{U}_m - V\| - \nabla_\theta \mathbb{E}\|V - V'\| \\ \nabla_\theta \mathbb{E}\|U - V\| &= \mathbb{E}_{U_m \sim P} \nabla_\theta \mathbb{E}\|\hat{U}_m - V\|. \end{aligned}$$

Under the assumption that ∇_θ and the expectation over U_m commute, we have:

$$\nabla_\theta \mathbb{E}\|U - V\| = \nabla_\theta \mathbb{E}_{U_m \sim P} \mathbb{E}\|\hat{U}_m - V\|.$$

Under the independence of the variables with $\hat{P}_m$ and Q_θ , we have:

$$= \nabla_\theta \mathbb{E}_{U_m \sim P} \mathbb{E}_{u \sim \hat{P}_m} \|u - V\|.$$

Finally, given that the expected empirical distribution is P , we find:

$$\begin{aligned} &= \nabla_\theta \mathbb{E}_{u \sim P} \mathbb{E} \|u - V\| \\ &= \nabla_\theta \mathbb{E} \|U - V\|. \end{aligned}$$

□

Proof of equivalence with the Energy distance. In the univariate case, Székely (2003) established that the Energy distance and the ℓ_2^2 distance are linearly related:

$$\ell_2^2(P, Q) = \frac{1}{2} \mathcal{E}(P, Q).$$

Let P and Q be probability distributions on $\mathbb{R}$ with cumulative distribution functions F_P and F_Q , respectively. In one dimension, the Euclidean norm $\|x - y\|$ reduces to the absolute value $|x - y|$ for scalars $x, y \in \mathbb{R}$. Therefore, the Energy distance can be expressed as:

$$\mathcal{E}(P, Q) = 2 \mathbb{E}[|U - V|] - \mathbb{E}[|U - U'|] - \mathbb{E}[|V - V'|],$$

where $U, U' \sim P$ i.i.d. and $V, V' \sim Q$ i.i.d., with all pairs independent. By definition of the expectation of functions of independent random variables:

$$\begin{aligned} \mathbb{E}[|U - V|] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |u - v| dP(u) dQ(v). \\ \mathbb{E}[|U - U'|] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |u - u'| dP(u) dP(u'). \\ \mathbb{E}[|V - V'|] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |v - v'| dQ(v) dQ(v'). \end{aligned}$$

It is well-known that (Székely 2003), in one dimension, the expected absolute difference can be expressed in terms of the squared difference of the CDFs:

$$\begin{aligned} \mathbb{E}[|U - V|] &= \int_{-\infty}^{\infty} (F_P(u) + F_Q(u) - 2F_P(u)F_Q(u)) du = \int_{-\infty}^{\infty} (F_P(u) - F_Q(u))^2 du \\ \mathbb{E}[|U - U'|] &= \int_{-\infty}^{\infty} (F_P(u) + F_P(u) - 2F_P(u)^2) du = \int_{-\infty}^{\infty} (2F_P(u) - 2F_P(u)^2) du \\ \mathbb{E}[|V - V'|] &= \int_{-\infty}^{\infty} (2F_Q(u) - 2F_Q(u)^2) du. \end{aligned}$$

Substituting these results into the definition of Energy distance gives:

$$\begin{aligned}
\mathcal{E}(P, Q) &= 2\mathbb{E}[|U - V|] - \mathbb{E}[|U - U'|] - \mathbb{E}[|V - V'|] \\
&= \int_{-\infty}^{\infty} \left[2F_P(u) + 2F_Q(u) - 4F_P(u)F_Q(u) \right. \\
&\quad \left. - (2F_P(u) - 2F_P(u)^2) - (2F_Q(u) - 2F_Q(u)^2) \right] du \\
&= \int_{-\infty}^{\infty} \left[-4F_P(u)F_Q(u) + 2F_P(u)^2 + 2F_Q(u)^2 \right] du \\
&= 2 \int_{-\infty}^{\infty} \left[-2F_P(u)F_Q(u) + F_P(u)^2 + F_Q(u)^2 \right] du \\
&= 2 \int_{-\infty}^{\infty} (F_P(u) - F_Q(u))^2 du \\
&= 2\ell_2^2(P, Q).
\end{aligned}$$

□

4.B Datasets

Number of claims	Count	Percentage
0	93,535	93.53%
1	6,109	6.11%
2	335	0.34%
3	21	0.02%

Table 4.B.1: The table reports the number of policies and the corresponding percentage for each observed claim count $N_i \in \{0, 1, 2, 3\}$ in the synthetic motor insurance portfolio. As shown, the majority of policies have zero claims, highlighting the rarity of the event.

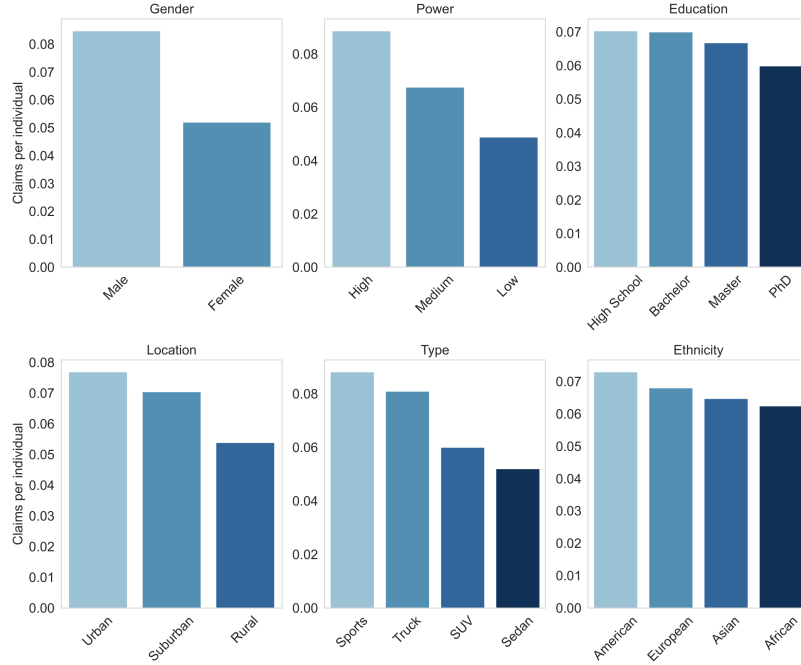


Figure 4.B.1: Average number of claims per policyholder for each category of the discrete rating factors in the simulated motor insurance portfolio. Categories within each factor are sorted in descending order of claim frequency to highlight the most at-risk segments.

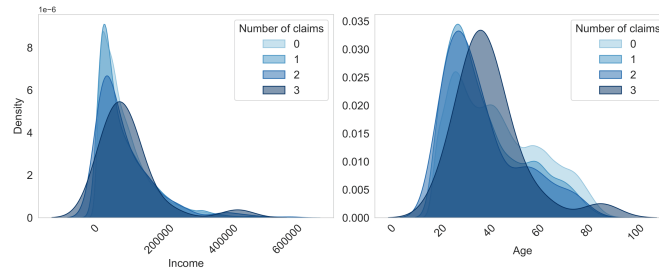


Figure 4.B.2: Density distributions of the continuous rating factors (age and income) stratified by the number of claims $N_i \in \{0, 1, 2, 3\}$ in the simulated motor insurance portfolio. These plots show how the distribution of continuous covariates varies across different claim counts.

4.C Best-estimates

Parameter	Value	Description
learning_rate	0.0099	Step size shrinkage used in each boosting step to prevent overfitting. Smaller values make learning more conservative.
feature_fraction	0.9368	Fraction of features randomly selected for each tree (column subsampling), used to prevent overfitting.
bagging_fraction	0.6151	Fraction of data randomly selected for each bagging iteration (row subsampling), helps reduce variance.
bagging_freq	4	Frequency of bagging; LightGBM performs data subsampling every 4 boosting iterations.
lambda_l2	2.5348e-05	L2 regularization term on weights to prevent overfitting (ridge penalty).
min_sum_hessian_in_leaf	0.7570	Minimum sum of instance Hessians (second-order gradients) in a leaf; larger values can make the model more conservative.
max_depth	6	Maximum depth of each tree; limits model complexity and helps control overfitting.
min_gain_to_split	0.4139	Minimum gain required to make a further partition on a leaf; acts as a regularization term to control splitting.
min_data_in_leaf	74	Minimum number of data points allowed in a leaf; prevents overly specific leaf nodes.
num_leaves	51	Maximum number of leaves per tree; controls model complexity and expressiveness.
objective	poisson	Objective function optimized by the model; here, a Poisson regression suited for count data.

Table 4.C.1: Optimized LightGBM hyperparameters using `optuna` and their descriptions.

4.D Best-estimates autocalibration

The negative bias (or overestimation) observed in the LightGBM predictions in Table 4.5.5 can be corrected using the multiplicative bias-correction factor proposed by Denuit, Charpentier, et al. (2021):

$$\frac{\mathbb{E}[Y|\hat{\pi}(\mathbf{X}) = F_{\hat{\pi}}^{-1}(x)]}{F_{\hat{\pi}}^{-1}(x)}, \quad (4.D.1)$$

estimated using a local smoother with bandwidth $\alpha = 0.25$. This correction reduces global bias while preserving the Poisson deviance (barely increases from 6539.15 to 6542.64).

Table 4.D.1 reports summary statistics before and after bias correction. The average premium shifts from 0.1363 to 0.1346, bringing it close to the reference mean.

Statistic	GLM $_{\beta_0}$	LightGBM	BC
			LightGBM
Mean $\bar{\pi}$	0.1350	0.1363	0.1346
10% Quantile	–	0.0391	0.0374
90% Quantile	–	0.2115	0.2003

Table 4.D.1: Summary statistics of the expected premiums $\hat{\pi}$ for the baseline intercept-only GLM (GLM $_{\beta_0}$), the LightGBM, and the bias-corrected LightGBM. Highlighted value indicates the shift in deviation from the baseline mean premium, illustrating the effect of bias correction.

Table 4.D.2 shows the improvement in predicted claim counts. The gap in the number of policies with at least one claim drops from 13.38 to -1.37 , effectively removing the systematic bias.

c	Observed	LightGBM	BC	LightGBM gap	BC
			LightGBM		LightGBM gap
0	18719	18705.6169	18720.3746	-13.3831	1.3746
1	1221	1224.7363	1211.9909	3.7363	-9.0091
2	56	66.5456	64.7117	10.5456	8.7117
3	4	2.9848	2.8178	-1.0152	-1.1822
> 0	1281	1294.3831	1279.6254	13.3831	-1.3746

Table 4.D.2: Comparison of observed claim counts with predicted claims from LightGBM before and after bias correction. The columns gap show the difference between predicted and observed counts. Highlighted value indicates the reduction of prediction bias after correction.

Fig. 4.D.1 displays histograms of premium predictions before and after bias correction.

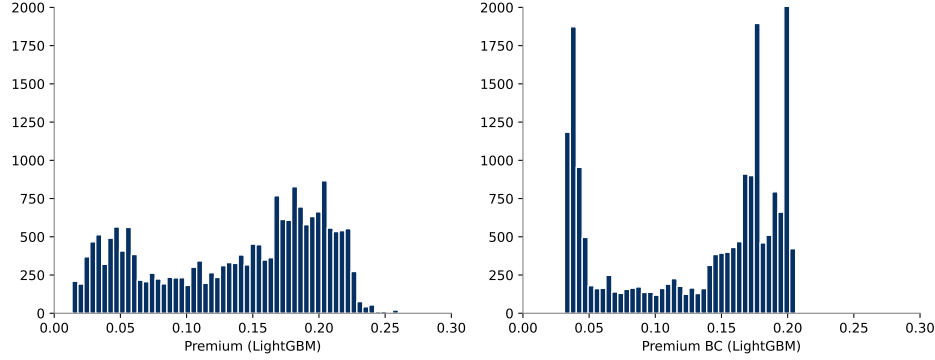


Figure 4.D.1: Histogram comparison of predicted premiums $\hat{\pi}$ for the LightGBM before (left) and after (right) balance correction.

Fig. 4.D.2 illustrates how the mapping $s \mapsto \mathbb{E}[Y|\hat{\pi}(\mathbf{X}) = s]$ moves closer to the first bisector (i.e., true mean) $\mu(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ after correction.

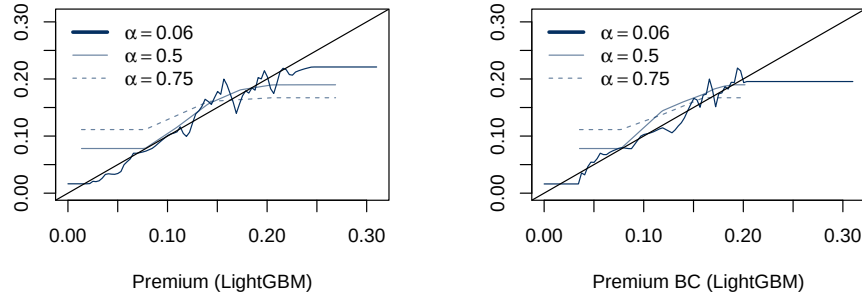


Figure 4.D.2: Visualisation of $s \mapsto \mathbb{E}[Y|\hat{\pi}(\mathbf{X}) = s]$ before (left plot) and after (right plot) bias correction (BC). α is a smoothing hyper-parameter in the `locfit` package in R, and the diagonal corresponds to $\hat{\pi} = \mu$.

Overall, the autocalibration step successfully realigns the LightGBM predictions with the true mean while maintaining predictive performance (measured by the Poisson deviance). This balance correction is a key post-hoc step of the fairness adjustment.

4.E Simulated dataset

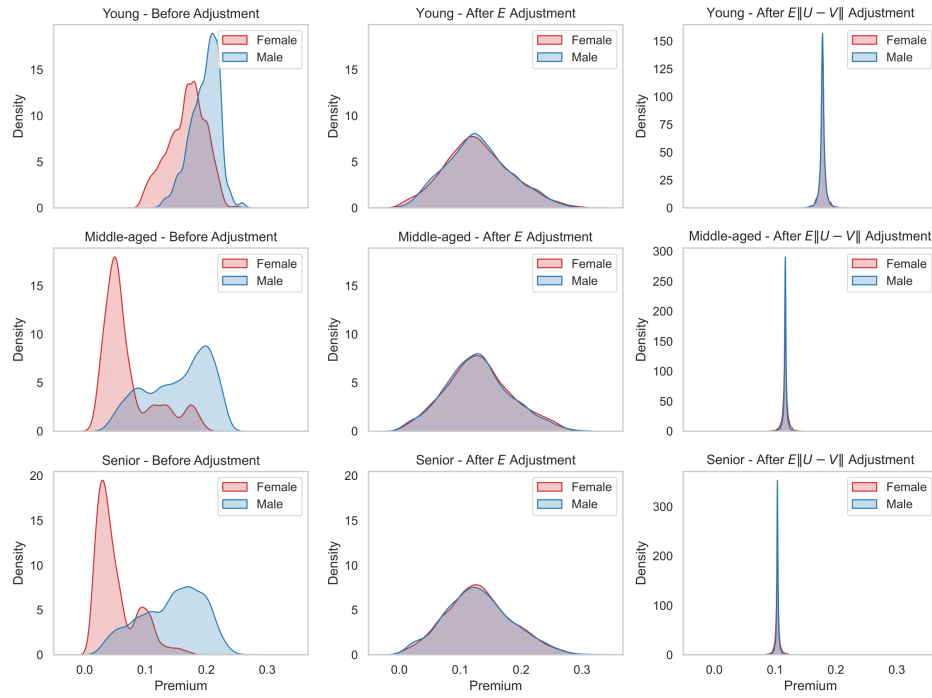


Figure 4.E.1: Visualisation of the role of the intra-group variability terms $\mathbb{E}[\|U - U'\|]$ and $\mathbb{E}[\|V - V'\|]$ in the fairness adjuster.

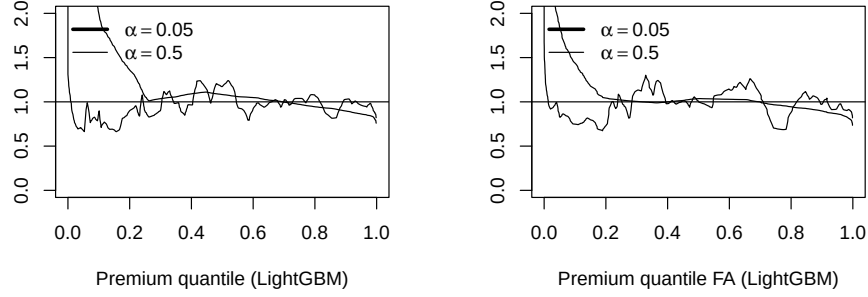


Figure 4.E.2: Visualisation of the local bias $\mathbb{E}[Y|\hat{\pi}(\mathbf{X}) = F_{\hat{\pi}}^{-1}(x)] / F_{\hat{\pi}}^{-1}(x)$ before and after fairness adjustment (FA).

c	Observed	LightGBM	Fair LightGBM	Fair BC LightGBM	Δ	Δ Fair	Δ Fair BC
0	3711	3723.96	3795.55	3786.66	12.96	84.55	75.66
1	324	306.64	241.02	250.56	-17.36	-82.98	-73.44
2	14	17.58	11.92	11.37	3.58	-2.08	-2.63
3	0	0.78	0.49	0.40	0.78	0.49	0.40

Table 4.E.1: Young-females calibration table showing the observed number of claims and the predicted counts from the LightGBM, the fairness-adjusted LightGBM, and the bias-corrected and fairness-adjusted LightGBM, along with the differences between predicted and observed counts.

c	Observed	LightGBM	Fair LightGBM	Fair BC LightGBM	Δ	Δ Fair	Δ Fair BC
0	3281	3271.76	3180.63	3172.47	-9.24	-100.37	-108.53
1	105	115.35	200.10	208.72	10.35	95.10	103.72
2	3	3.77	9.85	9.47	0.77	6.85	6.47
3	2	0.12	0.41	0.33	-1.88	-1.59	-1.67

Table 4.E.2: Middle-aged-females calibration table showing the observed number of claims and the predicted counts from the LightGBM, the fairness-adjusted LightGBM, and the bias-corrected and fairness-adjusted LightGBM, along with the differences between predicted and observed counts.

c	Observed	LightGBM	Fair LightGBM	Fair BC LightGBM	Δ	Δ Fair	Δ Fair BC
0	2576	2569.34	2473.00	2465.42	-6.66	-103.00	-110.58
1	56	65.09	155.14	162.94	9.09	99.14	106.94
2	4	1.54	7.54	7.38	-2.46	3.54	3.38
3	0	0.04	0.31	0.26	0.04	0.31	0.26

Table 4.E.3: Senior-females calibration table showing the observed number of claims and the predicted counts from the LightGBM, the fairness-adjusted LightGBM, and the bias-corrected and fairness-adjusted LightGBM, along with the differences between predicted and observed counts.

c	Observed	LightGBM	Fair LightGBM	Fair BC LightGBM	Δ	Δ Fair	Δ Fair BC
0	3579	3584.58	3707.25	3697.33	5.58	128.25	118.33
1	355	343.67	232.74	243.22	-11.33	-122.26	-111.78
2	17	22.57	11.51	11.05	5.57	-5.49	-5.95
3	1	1.13	0.48	0.39	0.13	-0.52	-0.61

Table 4.E.4: Young-males calibration table showing the observed number of claims and the predicted counts from the LightGBM, the fairness-adjusted LightGBM, and the bias-corrected and fairness-adjusted LightGBM, along with the differences between predicted and observed counts.

c	Observed	LightGBM	Fair LightGBM	Fair BC LightGBM	Δ	Δ Fair	Δ Fair BC
0	3093	3090.95	3122.93	3116.95	-2.05	29.93	23.95
1	224	224.27	194.96	201.64	0.27	-29.04	-22.36
2	10	12.22	9.69	9.08	2.22	-0.31	-0.92
3	1	0.54	0.41	0.32	-0.46	-0.59	-0.68

Table 4.E.5: Middle-aged-males calibration table showing the observed number of claims and the predicted counts from the LightGBM, the fairness-adjusted LightGBM, and the bias-corrected and fairness-adjusted LightGBM, along with the differences between predicted and observed counts.

c	Observed	LightGBM	Fair LightGBM	Fair BC LightGBM	Δ	Δ Fair	Δ Fair BC
0	2479	2465.02	2481.78	2473.23	-13.98	2.78	-5.77
1	157	169.71	154.28	163.03	12.71	-2.72	6.03
2	8	8.88	7.61	7.47	0.88	-0.39	-0.53
3	0	0.38	0.32	0.26	0.38	0.32	0.26

Table 4.E.6: Senior-males calibration table showing the observed number of claims and the predicted counts from the LightGBM, the fairness-adjusted LightGBM, and the bias-corrected and fairness-adjusted LightGBM, along with the differences between predicted and observed counts.

4.F Real-world dataset

The analysis presented in Section 4.5 on simulated data is reproduced for the real-world dataset introduced in Section 4.5.1, which is partially synthetically generated according to its Kaggle description.

4.F.1 Best-estimates

We consider a boosting algorithm (LightGBM) to model the premium best-estimates $\hat{\pi}$. The boosting model is tuned via `lightgbm` and `optuna` with optimal hyperparameters reported in Table 4.F.1.

Parameter	Value	Parameter	Value
learning_rate	0.0164	min_sum_hessian_in_leaf	0.0002
feature_fraction	0.8795	max_depth	7
bagging_fraction	0.8946	min_gain_to_split	0.3384
bagging_freq	1	min_data_in_leaf	35
lambda_l2	1.1452e-09	num_leaves	30

Table 4.F.1: Optimized LightGBM (with poisson objective) hyperparameters using `optuna`. Their descriptions are detailed in Table 4.C.1.

Fig. 4.5.1 compares predicted and observed claim frequencies across 10 quantile bins. Bars represent total exposure within each bin, while the solid and dashed lines show the predicted $\hat{\lambda}$ and observed λ frequencies, respectively. A closer alignment between the two lines across bins indicates superior model performance.

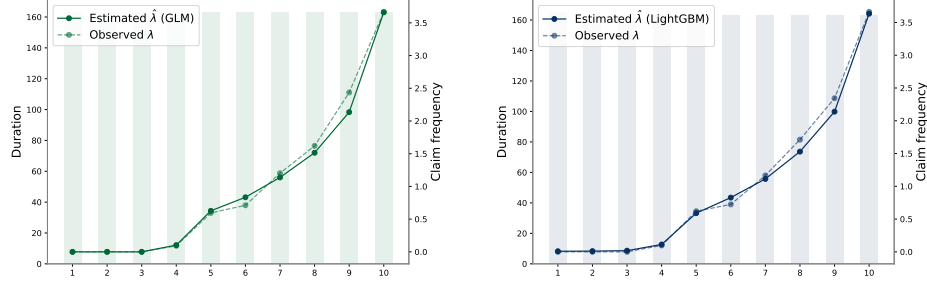


Figure 4.F.1: Comparison of predicted and observed claim frequencies per unit exposure across 10 quantile-based bins. The solid line shows the estimated frequency $\hat{\lambda}$ from the model, while the dashed line shows the observed frequency λ . The bars indicate the total exposure in each bin. Left plot: GLM; right plot: LightGBM.

As observed in Table 4.F.2, LightGBM has a slightly higher deviance than the GLM. In the analysis on simulated data, we observed the opposite.

Model	Poisson deviance
GLM	1436.7176
LightGBM	1476.3945

Table 4.F.2: Comparison of Poisson deviance values, defined in Eq. (4.5.1), on the first out-of-sample test set for the GLM and the LightGBM. Lower deviance indicates a better fit to the observed claim frequencies.

Fig. 4.F.2 illustrates the distribution of predicted premiums and highlights differences in risk allocation between the two models. Since the predicted premiums correspond to Poisson count estimates, many values (in the first histogram bin) are close to zero (but strictly positive). For readability, all visualisations are based on the transformed scale $\log(1 + \hat{\lambda}_i)$.

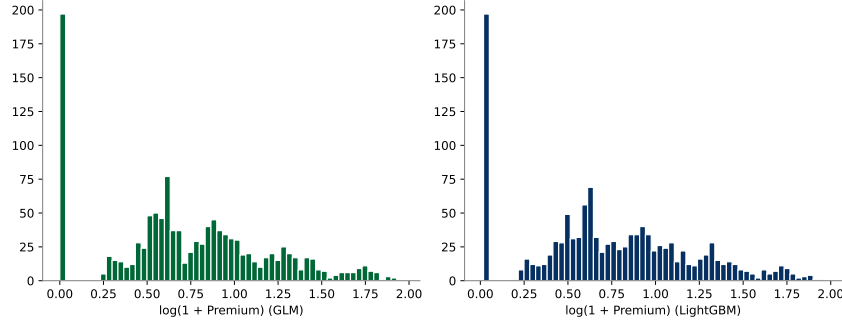


Figure 4.F.2: Histogram comparison of predicted premiums $\hat{\pi}$ for the GLM (left) and LightGBM (right). The plots illustrate the distribution of predicted premiums and highlight differences in risk allocation between the two models.

4.F.2 Fairness adjustment $G=2$

In the binary setting $G = 2$, the predicted (discriminatory) premiums $\hat{\pi}$ are partitioned into two demographic groups $\{\hat{\pi}_g\}_{g=1}^G$ corresponding to $g = \{\text{female}, \text{male}\}$. These premiums are iteratively adjusted using algorithm 9 to equalise not only the means, but also the dispersion and overall shape of their distributions.

Fig. 4.F.3 presents the cumulative distributions of predicted premiums for the overall portfolio and for the male and female subgroups, before (solid lines) and after (dashed lines) the fairness adjustment. All predictions are visualised on the transformed scale $\log(1 + \hat{\lambda}_i)$. While the analysis on the simulated portfolio in Section 4.5.3 relied on standard (non-cumulative) kernel density estimates, the cumulative KDEs in Fig. 4.F.3 allow for a better comparison of the distributional shift across groups, particularly in the lower premium range where many values are close to zero.

Before the adjustment, the cumulative distributions differ noticeably between genders. The cumulative curve for females lies above that for males (i.e., female policyholders have lower predicted premiums on average than males). After the fairness adjustment, the cumulative distributions for both groups (dashed lines) nearly coincide, indicating that our Energy-based fairness correction effectively mitigates gender-based pricing disparities.

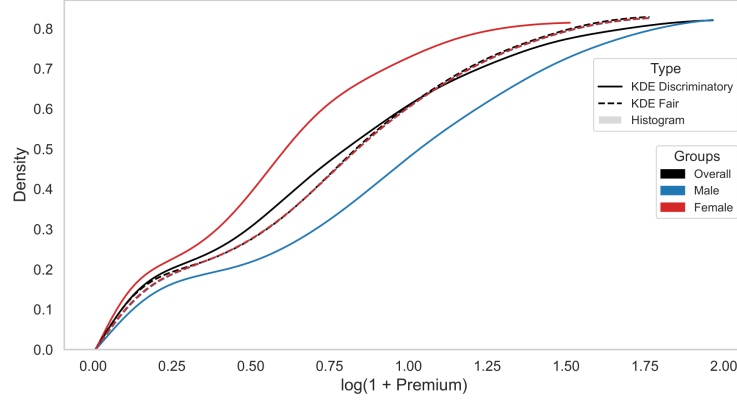


Figure 4.F.3: Cumulative kernel density estimates (KDEs) of predicted premiums (on the transformed scale $\log(1 + \hat{\lambda}_i)$) by gender. Solid lines represent the original LightGBM discriminatory premiums, while dashed lines correspond to the fairness-adjusted premiums. The male group is shown in blue, the female group in red, and the overall portfolio in black.

Enforcing independence between premiums and gender (a key risk factor) constrains the predictions and prevents the model from fully exploiting gender-related risk information. As illustrated in Fig. 4.F.4, the model inevitably sacrifices some predictive accuracy to enforce DP.

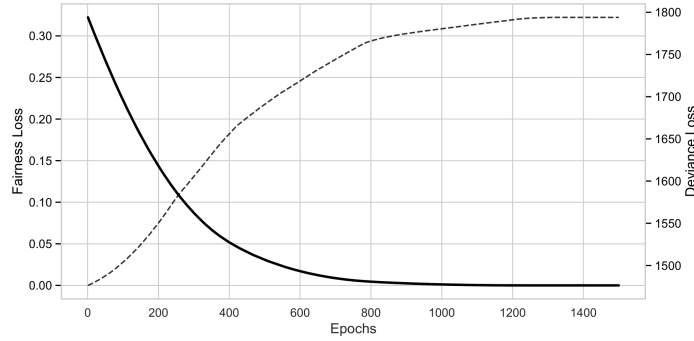


Figure 4.F.4: Evolution of the fairness loss (i.e., the Energy distance in Eq. (4.3.5)) and the Poisson deviance loss (in Eq. (4.5.1)) over training epochs.

Table 4.F.3 shows that the fairness adjustment also leads to an increase in the overestimation of the overall claim count (from 44 to 81). However, we also observe in Table 4.F.4 that the fairness adjustment leads to a smaller deviation from the

true mean $\mu = 1.0119$.

c	Observed	LightGBM	LightGBM _{fair}	LightGBM gap	Fair LightGBM gap
0	951	906.9875	869.6073	-44.0125	-81.3927
1	272	304.0601	318.1396	32.0601	46.1396
2	154	188.3684	208.1326	34.3684	54.1326
3	101	106.6544	117.8240	5.6544	16.8240
4	62	58.4609	61.4248	-3.5391	-0.5752
5	36	31.6624	30.2977	-4.3376	-5.7023
6	28	16.8910	14.1690	-11.1090	-13.8310
7	12	8.7817	6.2518	-3.2183	-5.7482
8	8	4.3992	2.5920	-3.6008	-5.4080
9	4	2.1040	1.0082	-1.8960	-2.9918
> 0	679.0	723.0125	760.3927	44.0125	81.3927

Table 4.F.3: Comparison of observed claim counts with predictions from standard LightGBM and fairness-adjusted LightGBM. The gap columns show the difference between predicted and observed counts. Highlighted values indicate the aggregated total gaps, illustrating the impact of the fairness adjustment on total balance.

Statistic	GLM _{β_0}	LightGBM	Fair LightGBM
Mean $\bar{\pi}$	1.0609	0.9983	1.0119
10% Quantile	—	0.0099	0.0099
90% Quantile	—	2.6672	2.4771

Table 4.F.4: Summary statistics of expected premiums $\hat{\pi}$ for the baseline intercept-only GLM (GLM _{β_0}), the standard LightGBM, and the fairness-adjusted LightGBM. Highlighted values indicate divergence from the target mean premium, indicating the effect of fairness adjustment.

The negative bias (or overestimation) observed in the fair LightGBM predictions can be corrected using the multiplicative bias-correction factor proposed by Denuit, Charpentier, et al. (2021):

$$\frac{\mathbb{E}[Y|\hat{\pi}(\mathbf{X}) = F_{\hat{\pi}}^{-1}(x)]}{F_{\hat{\pi}}^{-1}(x)}, \quad (4.F.1)$$

estimated using a local smoother with bandwidth $\alpha = 0.16$. Table 4.F.5 reports summary statistics before and after bias correction. The average premium shifts from 1.0119 to 1.0397, bringing it close to the reference mean.

Statistic	GLM $_{\beta_0}$	LightGBM	Fair LightGBM	Fair BC LightGBM
Mean $\bar{\pi}$	1.0609	0.9983	1.0119	1.0397
10% Quantile	—	0.0099	0.0099	0.0000
90% Quantile	—	2.6672	2.4771	2.5625

Table 4.F.5: Summary statistics of expected premiums $\hat{\pi}$ for the baseline intercept-only GLM (GLM $_{\beta_0}$), standard LightGBM, fairness-adjusted LightGBM, and bias-corrected fairness-adjusted LightGBM. Highlighted values indicate divergence from the target mean premium, illustrating the effect of fairness adjustment and subsequent bias correction.

Table 4.F.6 summarises the Poisson deviance values. We note that the fairness–accuracy trade-off manifests at two distinct levels. The fairness correction inevitably reduces predictive accuracy, as the optimisation process prioritises DP over likelihood-based fit. The post-hoc autocalibration step partially restores, not only the actuarial balance, but also a share of the predictive accuracy (1667.6080) that was initially reduced by the fairness correction, without affecting fairness, as shown in Fig. 4.F.5.

Model	Poisson Deviance	Δ GLM	Δ LightGBM
GLM	1436.7176		
LightGBM	1476.3944	+0.0276	
LightGBM fair (gender)	1793.9142	+0.2486	+0.2151
LightGBM fair BC (gender)	1667.6080	+0.1607	+0.1295

Table 4.F.6: Poisson deviance comparison at the same evaluation stage (out-of-sample 1) for the generalised linear model (GLM), bias-corrected (BC), and fairness-adjusted versions of the LightGBM model. Lower deviance indicates a better fit to the observed claim frequencies. $\Delta = (a - b)/b$ values compare each fairness (and balance) corrected model relative deterioration with the GLM and LightGBM.

Fig. 4.F.5 shows the local redistribution of density after balance correction (dotted lines). Even after balance correction, fairness is maintained as the conditional premium densities for both gender groups remain well-aligned.

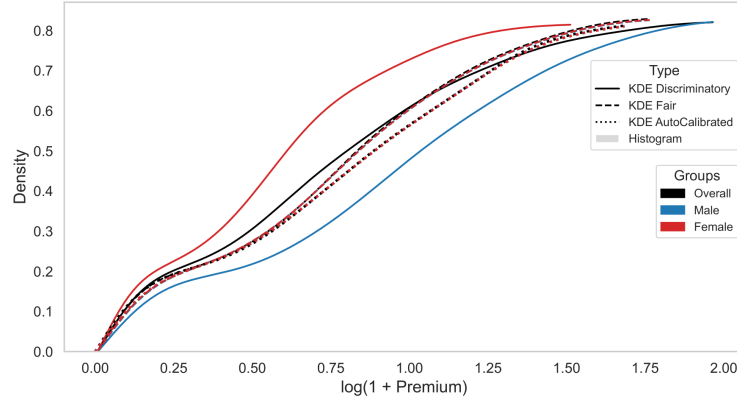


Figure 4.F.5: Cumulative kernel density estimates (KDE) of predicted premiums by gender. Solid lines represent the original LightGBM discriminatory premiums, dashed lines the fairness-adjusted premiums, and dotted lines the fairness-adjusted and balance-corrected premiums. The male group is shown in blue, the female group in red, and the overall portfolio in black.

Table 4.F.7 summarises several keys fairness metrics discussed in the literature (see dicussion related to Table 4.5.11).

Metric with $U = \hat{\pi}_{\text{female}}$ and $V = \hat{\pi}_{\text{male}}$	$\hat{\pi}$	$\hat{\pi}^{\text{fair, BC}}$
$W_1(U, V)$	0.883082	0.175620
$W_2(U, V)$	1.264377	0.306100
$\text{MMD}^2(U, V)$	0.037794	8.2×10^{-5}
$\Delta_{\text{GFE}} = \mathbb{E}[U] - \mathbb{E}[V] $	0.760072	0.001787
$\Delta_{\text{SP}} = \sup_s P(U \leq s) - P(V \leq s) $	0.335124	0.017543
$S_{\epsilon=0.1}(U, V)$	0.701744	0.000864

Table 4.F.7: Comparison of key fairness metrics before and after $\mathcal{E}$ -based fairness adjustment and balance correction. MMD is computed using a Gaussian RBF kernel $k(u, v) = \exp(-\|u - v\|^2 / (2\sigma^2))$ with fixed bandwidth $\sigma = 0.1$.

For new policyholders, a random forest regressor learns the mapping from policyholder characteristics to correction factors, as outlined in Section 4.4.2. Fig. 4.F.6 shows that the adjusted predictions for new policyholders (out-of-sample 2) remain consistent across sensitive groups with the results observed in Fig. 4.F.5.

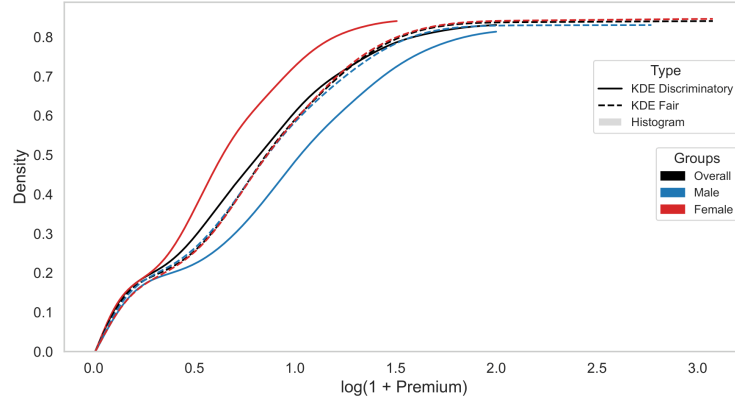


Figure 4.F.6: Cumulative kernel density estimates (KDEs) of predicted premiums by gender for new policyholders. Solid lines represent the LightGBM discriminatory premiums, dashed lines the fairness-adjusted premiums, and dotted lines the fairness-adjusted and balance-corrected premiums. The male group is shown in blue, the female group in red, and the overall portfolio in black.

Table 4.F.8 summarises the Poisson deviance results before and after fairness correction on the new test (out-of-sample 2) set with 1 630 policies. The transition from out-of-sample 1 to out-of-sample 2 reveals a nearly negligible change in loss (from 1793.91 to 1794.33). Both out-of-sample sets contain only $n = 1,630$ observations, which naturally induces some instability. With such small test samples, even moderate changes in the composition of policyholders between the two out-of-sample sets can produce nontrivial fluctuations in deviance after fairness correction.

4.F.3 Fairness adjustment $G > 2$

For $G > 2$, the predicted premiums $\hat{\pi}$ are partitioned into G vectors $\{\hat{\pi}_g\}_{g=1}^G$ of sizes n_g organized in d -dimensional variables U and V as in Eq. (4.4.4). In this setting, equal-sized samples are drawn from each group in U , and for each group in V . A uniform random permutation is also applied to ensure the pairwise comparisons span all group combinations (see algorithm 9).

Fig. 4.F.7 illustrates the fairness adjustment of $\hat{\pi}$ across four demographic groups, determined by the policyholders' gender (female or male) and ethnicity (minority or majority). The solid lines correspond to the initial predicted premiums partitioned into G demographic groups $\{\hat{\pi}_g\}_{g=1}^4$. The overlapping dashed lines correspond to the corresponding fairness adjusted premiums $\{\hat{\pi}_g^{\text{fair}}\}_{g=1}^4$. We observe that after adjustment, the dashed lines largely overlap within each gender-

	in-sample	out-of-sample 1	out-of-sample 2	Δ
GLM	4019.5312	1436.7176	1380.0193	-3.9464%
LightGBM	3814.3971	1476.3945 (+0.0276)	1440.2732 (+0.0437)	-2.4466%
LightGBM fair	—	1793.9142 (+0.2151)	1794.3334 (+0.2458)	-0.0234%
		(+0.2486)	(+0.3002)	

Table 4.F.8: Poisson deviance comparison across models: (i) in-sample (training) set ($n = 4,889$) to fit the discriminatory GLM and LightGBM models, (ii) a first out-of-sample (test) set ($n = 1,630$) to predict discriminatory premiums and correct for fairness and autocalibration, and (iii) a second out-of-sample (test) set ($n = 1,630$) of new policyholders, for which corrected premiums were predicted using a Random Forest trained on the first test set. Values in parentheses compare each model relatively (i.e., $(a - b)/b$) with the GLM and LightGBM at the same evaluation stage. $\Delta = (a - b)/b$ measures the relative deterioration when moving from the first to the second out-of-sample (new policyholders).

ethnicity pairing, indicating that the $\mathcal{E}$ -adjuster has effectively reduced disparity across these four groups.

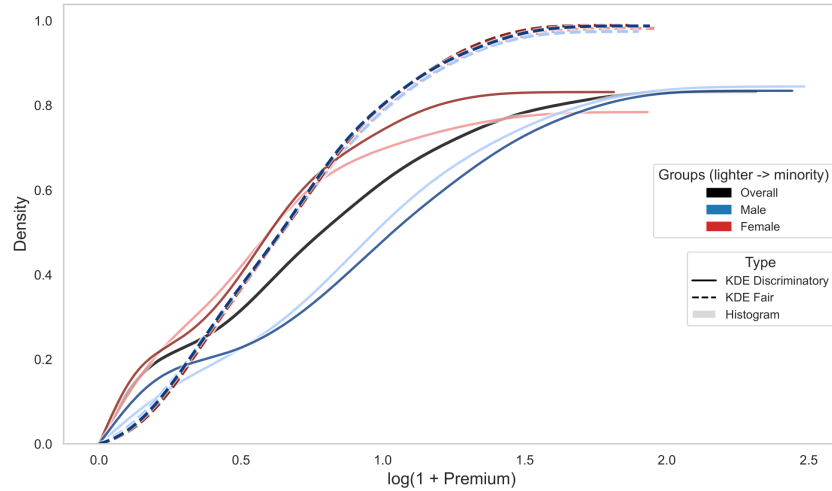


Figure 4.F.7: Kernel density estimates (KDEs) of predicted premiums by gender and age group. Solid lines represent the LightGBM discriminatory premiums, and dashed lines the fairness-adjusted premiums. The male groups are shown in blue, the female groups in red, and the overall portfolio in black. The lighter color shade corresponds to the minority group.

Table 4.F.9 reveals that the fairness adjustment actually mitigated the deviation from the mean value.

Statistic	GLM $_{\beta_0}$	LightGBM	Fair LightGBM
Mean $\bar{\pi}$	1.0609	1.0119	1.0597
10% Quantile		0.0099	0.2556
90% Quantile		2.4771	2.2170

Table 4.F.9: Summary statistics of expected premiums $\hat{\pi}$ for the baseline intercept-only GLM (GLM $_{\beta_0}$), standard LightGBM, and fairness-adjusted LightGBM. Highlighted values indicate divergence from the target mean premium, illustrating the effect of fairness adjustment.

Table 4.F.10 summarises the deviance loss before and after fairness adjustment (for $G = 2$ and $G = 4$). We observe overall a relatively stark deterioration in predictive accuracy following fairness adjustment, especially when adjusting across four protected groups. In the simulated study, model adjustment relied on 100 000 policies (with 60% for training and 20% for testing). Here, however, fairness adjustments are derived from only 1 630 policies in the first out-of-sample set, further split across two or four sensitive groups, leaving roughly 400 premium predictions per group. Using such small group-specific samples to estimate conditional distributions and to enforce DP clearly leads to poor predictive accuracy (in terms of Poisson deviance).

Model	Poisson Deviance	Δ GLM	Δ LightGBM
GLM	1436.7176		
LightGBM	1476.3944	+0.0276	
LightGBM fair (gender)	1793.9142	+0.2486	+0.2151
LightGBM fair BC (gender)	1667.6080	+0.1607	+0.1295
LightGBM fair BC (gender & ethnicity)	4718.2180	+2.284	+2.1958

Table 4.F.10: Poisson deviance comparison at the same evaluation stage (out-of-sample 1) for the generalised linear model (GLM), bias-corrected (BC), and fairness-adjusted versions of the LightGBM model. Lower deviance indicates a better fit to the observed claim frequencies. Δ compare each fairness (and balance) corrected model relative deterioration with the GLM and LightGBM.

4.G Wasserstein and MMD benchmarks

To further benchmark the proposed gradient-based fairness correction, we compare it against two post-processing baselines. First, against a distributional alignment method based on kernel density estimation, optimal transport barycenters, and quantile remapping. Second, against a maximum mean discrepancy (MMD)-based gradient adjustment applied directly on the predictions. Both approaches enforce DP by reducing distributional discrepancies between the conditional predictive distributions of men and women after training (here, LightGBM predictions).

4.G.1 Wasserstein (OT)

Because predictions lie in $[0, 1]$, we estimate bounded densities using a logit-transformed kernel density estimator:

$$f_G(\pi) = \exp(\text{KDE}(\text{logit}(\pi))) (\pi(1 - \pi))^{-1}, \quad G \in W, M.$$

From these KDEs we compute empirical PDF f_W, f_M and CDF F_W, F_M estimates.

Before applying optimal transport correction, a simple baseline consists of scaling each group's predictions to match the global mean:

$$\hat{\pi}_G^{\text{scaled}} = \frac{\bar{\pi}}{\bar{\pi}_G} \hat{\pi}_G.$$

This ensures equal means but leaves distributional differences largely untouched, precisely the limitation the barycenter method addresses.

Let w_W, w_M be empirical group weights. For each group CDF (F_G), define the barycentric target distribution F^* through a pair of reference-specific barycenters:

$$\begin{aligned} F_W^*(\pi^*) &= F_W(\pi), & \pi^* &= w_W \pi + w_M F_M^{-1}(F_W(\pi)), \\ F_M^*(\pi^*) &= F_M(\pi), & \pi^* &= w_M \pi + w_W F_W^{-1}(F_M(\pi)). \end{aligned}$$

Both constructions lead to the same Wasserstein barycenter F^* , whose density f^* is obtained by differentiating the monotone transport map. This F^* acts as the demographically fair predictive distribution toward which both groups are transported.

For each individual prediction $\hat{\pi}_i$, we first compute the group-specific empirical quantile:

$$u_i = F_G(\hat{\pi}_i).$$

We map it through the opposite group's inverse CDF:

$$T_G(\hat{\pi}_i) = F_G^{-1}(u_i).$$

We then construct the barycentric transport map:

$$\hat{\pi}_i^{\text{fair}} = w_G \hat{\pi}_i + w_{\bar{G}} T_G(\hat{\pi}_i).$$

This yields a set of adjusted predictions $\hat{\pi}^{\text{fair}}$ whose distribution matches the barycenter F^* . In practice, this approach nearly eliminates discrepancy:

$$W_1(F_W^{\text{fair}}, F_M^{\text{fair}}) \approx 0, \quad F_W^{\text{fair}} \simeq F_M^{\text{fair}} \simeq F^*.$$

4.G.2 MMD-based gradient fairness adjustment

As a second baseline, we implement a gradient-based fairness correction driven by the Maximum Mean Discrepancy (MMD) between group-conditional predictions. Both the Wasserstein-OT approach and the MMD gradient method aim at enforcing DP at the distributional level, but while the former constructs an explicit transport map between empirical distributions, the latter performs iterative RKHS alignment in prediction space.

Let $M = \{\hat{\pi}_i : S_i = M\}$ and $W = \{\hat{\pi}_j : S_j = W\}$. The squared MMD between the two groups under an RBF kernel k_σ is:

$$\text{MMD}^2(W, M) = \mathbb{E}[k_\sigma(W, W')] + \mathbb{E}[k_\sigma(M, M')] - 2\mathbb{E}[k_\sigma(W, M)].$$

We minimize this quantity by directly updating predictions via gradient descent:

$$\hat{\pi} \leftarrow \hat{\pi} - \eta \nabla_{\hat{\pi}} \text{MMD}^2(W, M),$$

where η is a step size.

The corresponding gradients for individual samples in group M and W are given by:

$$\begin{aligned} \nabla_{\hat{\pi}_i \in M} &= -\frac{2}{n_M^2 \sigma^2} \sum_{i'} (\hat{\pi}_i - \hat{\pi}_{i'}) k_\sigma(\hat{\pi}_i, \hat{\pi}_{i'}) + \frac{2}{n_M n_W \sigma^2} \sum_j (\hat{\pi}_i - \hat{\pi}_j) k_\sigma(\hat{\pi}_i, \hat{\pi}_j), \\ \nabla_{\hat{\pi}_j \in W} &= -\frac{2}{n_W^2 \sigma^2} \sum_{j'} (\hat{\pi}_j - \hat{\pi}_{j'}) k_\sigma(\hat{\pi}_j, \hat{\pi}_{j'}) + \frac{2}{n_M n_W \sigma^2} \sum_i (\hat{\pi}_j - \hat{\pi}_i) k_\sigma(\hat{\pi}_j, \hat{\pi}_i). \end{aligned}$$

This procedure iteratively adjusts predictions such that the RKHS embeddings of both groups become aligned, yielding:

$$\text{MMD}(\hat{\pi}_W^{\text{fair}}, \hat{\pi}_M^{\text{fair}}) \rightarrow 0.$$

Note that while MMD is a well-established discrepancy measure, the empirical MMD requires all pairwise kernel evaluations between samples across and within groups, resulting in $\mathcal{O}(n^2)$ complexity per iteration. Although approximations such as mini-batch MMD or random Fourier features can reduce computational cost, they introduce additional stochastic variance in the estimated discrepancy. In practice, these approaches require careful tuning of batch size, learning rate, and kernel bandwidth. These parameters are known to strongly affect both convergence stability and the quality of the resulting alignment.

Given that our goal is not to design a large-scale stochastic optimization procedure, but rather to evaluate MMD as a deterministic distributional alignment baseline, we adopt the full-batch formulation to ensure numerical stability and reproducibility.

We further fix the kernel bandwidth to 0.0622. It is a median heuristic bandwidth choice, which is one of the most standard (and surprisingly effective) ways to set the kernel scale in MMD.

We first evaluated the MMD gradient adjustment using a learning rate η of 0.01. This configuration required approximately 143 minutes for 500 iterations and resulted in only marginal changes in the group distributions, which exhibited only a slight convergence toward each other. Increasing the learning rate to 0.1 still did not yield meaningful overlap between the marginal distributions after 500 iterations. For comparison, the energy-based gradient method converged in fewer than 70 iterations under the same experimental setting with $\eta = 0.001$. Further increasing the learning rate to 1.0 led to faster displacement of the distributions (124 minutes for 500 iterations), with the marginals beginning to overlap. However, the resulting alignment remained insufficient to achieve DP at a level comparable to the energy-gradient method. Finally, we chose to set the learning rate to 1.5, to improve distributional alignments.

In this setting, MMD-based gradient alignment exhibits slower convergence and higher sensitivity to optimization hyperparameters compared to energy-distance-based corrections.

4.G.3 Results

As shown in Table 4.G.1, the discriminatory model exhibits a positive group mean difference, indicating that men tend to receive higher premium than women. The energy-based adjustment removes the mean disparity exactly, outperforming both MMD and Wasserstein. Wasserstein slightly over-corrects the gap and produces a negative mean difference (-0.0054), whereas MMD slightly undercorrects it ($+0.0008$). Although the Wasserstein barycenter enforces equality of marginal distributions by construction, it requires the numerical inversion of empirical CDFs. This step is inherently less stable, especially in sparse regions or in the presence of heavy tails (common in insurance data). The fact that MMD is kernel-based may also explain the residual differences, as further reductions in marginal discrepancies become increasingly difficult.

Model	Group mean difference (men – women)
LightGBM	0.0630
Energy LightGBM	0.0000
Wasserstein LightGBM	-0.0054
MMD LightGBM	0.0008

Table 4.G.1: Group mean differences (men – women) for predicted claim frequencies. Positive values indicate that men have higher predicted frequencies than women.

Fig. 4.G.1 shows the different effects of the fairness adjustments per group. Both methods reduce premiums for men while increasing premiums for women. In the lower and middle quantiles, the Wasserstein-corrected premium (solid line, with crosses) are overall slightly closer to the observed discriminatory predictions (dotted line). Wasserstein adjustment, however, introduces a stronger distortion in higher frequencies bins. MMD adjustments induce similar stronger distortions than Energy ones in the upper tails. Overall, women see substantially higher premiums than originally, while men receive lower premiums. This highlights a trade-off between overall distributional parity and extreme per-group distortions, particularly for high-risk policyholders.

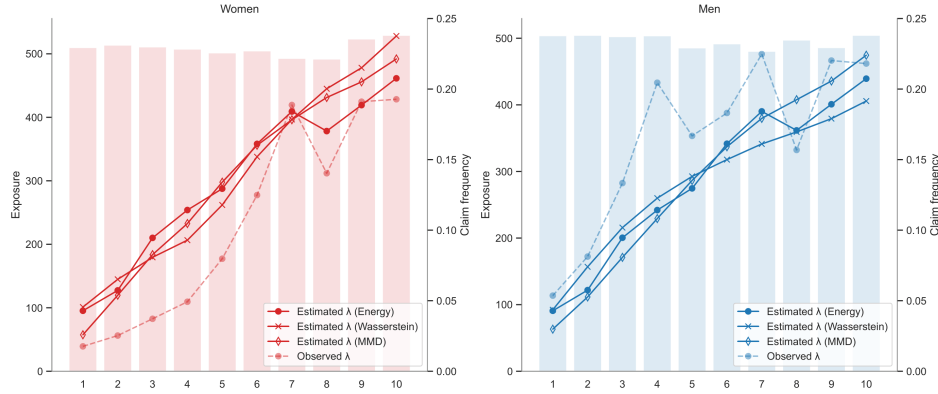


Figure 4.G.1: Comparison of predicted and observed claim frequencies by gender per unit exposure across 10 LightGBM quantile-based bins. The dashed line shows the observed frequency λ computed from the data. The solid line shows the estimated frequency $\hat{\lambda}$. Energy-based adjustments are indicated by circles, Wasserstein adjustments by crosses, and MMD by diamonds. Colors differentiate gender.

Fig. 4.G.2 further shows a comparison of original and adjusted predicted premiums for male and female policyholders. The energy-based adjustment leads to only

mild distortion in the tail, and its group-specific effects look stronger for men (further from the diagonal). MMD clearly leads to the strongest adjustments for men (further from the diagonal) in the lower and mid-tail. In contrast, the Wasserstein adjustment leads to a stronger deformation of the upper tail, reflecting a heavier redistribution of extreme values. Unlike Wasserstein-based corrections, which explicitly couple quantiles across groups and induce visible redistribution in the tails, MMD and Energy do not enforce a structured cross-group matching of quantiles, leading to less pronounced adjustments in the extremes of the distribution.

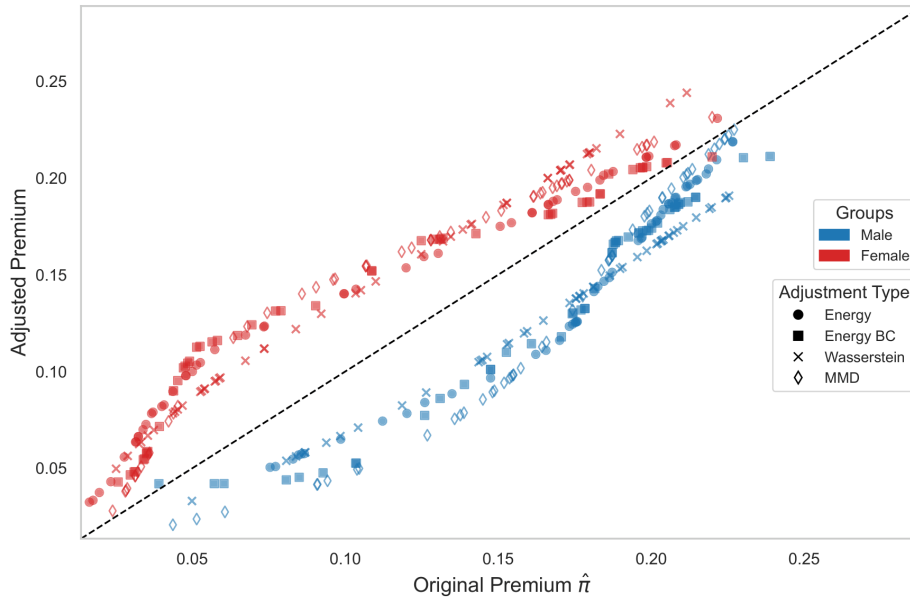


Figure 4.G.2: Scatter plot of original discriminatory premiums and adjusted premiums for male and female policyholders. Each point represents an individual prediction, with the x-axis showing the original LightGBM-predicted premium $\hat{\pi}$ and the y-axis showing the corresponding adjusted premium. Energy-based adjustments are indicated by circles, Energy-based and BC-corrected by squares, Wasserstein barycenter by crosses, and MMD by diamonds. Colors differentiate gender: blue for male and red for female. The dashed black identity line indicates perfect preservation of the original discriminatory predictions. Subsampling (10% of points) is applied for visualization clarity.

Table 4.G.2 provides a quantitative assessment of how disruptive each correction is with respect to the original risk differentiation. Fairness adjustments should mitigate unfair dependence on protected attributes without disrupting the core risk ranking. High-risk policyholders should still receive higher premiums than

low-risk ones. Otherwise, this would break actuarial credibility, tariff stability, and portfolio steering, leading to anti-selection and misaligned underwriting incentives.

The Spearman (1904) rank correlation between the discriminatory and adjusted premiums in Table 4.G.2 show that all corrections affect the initial ordering of risks. The energy-based adjustment preserves it slightly more effectively ($\rho = 0.8604$) than the Wasserstein barycentric correction ($\rho = 0.8266$). Wasserstein barycenters enforces equality at each quantile level. This can locally reorder individuals, especially in distributional tails, leading to more pronounced rank distortions.

Model	Spearman ρ
Energy LightGBM	0.8604
Wasserstein LightGBM	0.8266
MMD LightGBM	0.8615

Table 4.G.2: Spearman rank correlation between discriminatory and adjusted premium.

Table 4.G.3 further reports the rank correlation by gender. Because Spearman correlation only depends on ranks, values close to 1 indicate that the transformation applied by each method is almost monotone within each group. This is exactly what happens here. All methods act essentially as group-specific monotone recalibrations, especially Wasserstein, so they do not meaningfully reorder individuals within the same gender.

Model	Men	Women
Energy LightGBM	0.9998	0.9998
Wasserstein LightGBM	1.0000	1.0000
MMD LightGBM	0.9999	1.0000

Table 4.G.3: Spearman rank correlation between discriminatory and adjusted premium, by gender.

Fig. 4.G.3 evaluates conditional DP:

$$\mathbb{E}[\hat{\pi}_{\text{fair}}|\text{men}, q_{\hat{\pi}}] = \mathbb{E}[\hat{\pi}_{\text{fair}}|\text{women}, q_{\hat{\pi}}], \quad (4.G.1)$$

where $\hat{\pi}_{\text{fair}}$ are the Energy, Wasserstein, or MMD corrected premium, and $q_{\hat{\pi}}$ are the quantiles of the original discriminatory premium.

This shows local fairness, not just global parity.

The Wasserstein and Energy-corrected premium show mild differences in the middle quantiles. Wasserstein stays consistently negative across all upper quantiles, showing persistent conditional disparity across premium levels. This indicates that

women tend to receive slightly higher predicted premiums than men, but the deviation is stable across the risk spectrum. Whereas Energy returns toward zero in the upper tail, improving conditional parity. Energy seems to better enforce parity for higher-risk policyholders. The MMD curve shows stronger non-monotonicity across the risk spectrum, with relatively large negative disparities in mid-quantiles (down to approximately -0.10) followed by partial recovery in the upper tail (similarly to Energy).

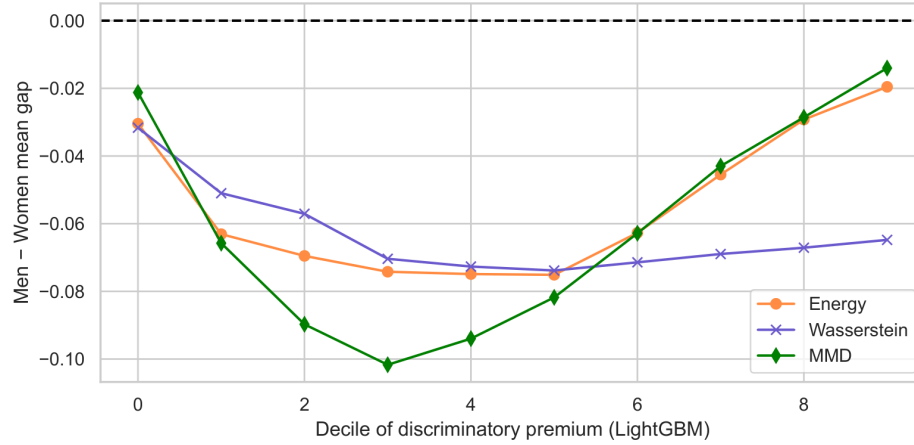


Figure 4.G.3: Conditional demographic parity gaps (men – women) by discriminatory premium quantile for each fairness method. Negative values indicate higher predicted premiums for women.

These empirical differences are consistent with the theoretical properties of the metrics.

The Energy distance acts as a global discrepancy measure. It integrates distributional differences over the entire support. Since the upper tail typically represents only a small fraction of the probability mass (5% here), its contribution to the Energy distance gradient is small, implying that the optimisation primarily aligns the bulk of the distribution.

The Wasserstein distance, in contrast, operates directly on quantile differences, giving each quantile, including extreme ones, equal weight in its objective. As a result, any misalignment in the far right tail contributes proportionally to the objective, forcing the correction to act more aggressively on large predicted values.

The Maximum Mean Discrepancy (MMD) sensitivity is strongly governed by the choice of kernel bandwidth. Small bandwidth translates to very local, almost point matching, alignments. Whereas large bandwidth translates to very global, almost mean matching alignments. Consequently, MMD can induce stronger lo-

calised adjustments than Energy, particularly in low-density or mid-tail regions, while remaining less disruptive in the extremes than Wasserstein.

Hence, from a distributional standpoint, the Energy distance preserves the extremal structure more faithfully, the Wasserstein distance enforces stricter quantile-level parity at the cost of stronger distortions in the tail, and MMD provides a kernel-based comparison that can capture both local and global discrepancies without explicit quantile coupling.

Note that both energy-based and MMD-based gradient adjustments are unconstrained functional perturbations of the prediction surface, meaning they do not inherently enforce positivity of the output. Since premiums are updated via gradient-driven corrections, sufficiently large updates (especially in sparse regions or tails) can in principle push some corrected values below zero unless an explicit positivity constraint is imposed (e.g., floor constraint or a log-link parameterisation in the case of count data). In our experiments, we verified that no negative premiums were produced after applying the corrections.

Chapter 5

A multi-criteria fair Gaussian regressor for insurance pricing

This chapter studies how multiple notions of fairness can be incorporated into a single Bayesian non-parametric regression framework for insurance pricing, with a focus on claim frequency modeling under a log-link. We consider a Generalized Gaussian Process Regression (GGPR) model for count data with risk exposure and introduce fairness interventions in its architecture. Specifically, we address notions of individual fairness by altering the kernel structure to control the similarity between policies (e.g., to mitigate omitted variable bias). We also address group-level fairness by enforcing demographic parity through constraints affecting the posterior. This modified GGPR architecture allows us to jointly enforce multiple fairness notions within a single probabilistic model. We empirically explore trade-offs with portfolio balance to prevent systematic overpricing or underpricing at the portfolio level, and how different fairness criteria interact when combined. The results highlight the importance of adopting a multi-criteria, context-aware approach to fairness in insurance pricing.

Charlotte Jamotton and Donatien Hainaut (2026a).
“A Multi-Criteria Fair Gaussian Regressor for Insurance Pricing”. In: *working paper*. DOI: [None](#)

5.1 Introduction

Insurance pricing is fundamentally concerned with the accurate assessment of risk. Modern actuarial practice increasingly relies on big data and flexible machine learning methods to capture complex non-linear interactions in claim data (Barry et al. 2020). While these models often improve predictive accuracy (Y. Yang et al. 2017), they may rely on complex interactions with proxy variables, raising concerns about

discrimination. In response, a growing body of actuarial research (Fahrenwaldt et al. 2024) has borrowed concepts from algorithmic fairness and causal theory, including demographic parity, counterfactual fairness, and omitted variable bias (often referred to as proxy discrimination in the actuarial literature), to question whether algorithmic pricing models amplify, perpetuate, or mitigate existing societal inequalities. As noted by O. Côté et al. (2025, p. 2), the insurance literature struggles with a lack of conceptual consensus, with fairness terms frequently used inconsistently across normative, regulatory, and modeling assumptions (Dolman et al. 2021).

A key difficulty is that fairness in insurance pricing is a multidimensional concept, spanning legal compliance, ethical acceptability, statistical properties, and market outcomes. As a result, fairness notions are not merely normative choices about what constitutes unjust discrimination, but technical constraints with direct implications for insurance markets, affecting pricing strategies, portfolio composition, and market share. Moreover, competing definitions and criteria, such as individual versus group fairness, causal versus statistical notions of discrimination, proxy versus direct discrimination, are often mutually incompatible. Therefore, different fairness criteria and methodologies are defended as the correct solution, yet there is still no consensus on a single approach that aligns with both regulatory constraints and actuarial soundness.

This chapter adopts a deliberately neutral stance. Rather than committing to a single notion of fairness, we propose a framework that embeds multiple fairness paradigms into the mathematical structure of a single regression model. Our objective is not to resolve the fairness debate normatively, but to provide an experimental setting in which the practical implications and trade-offs of competing fairness notions can be studied side by side.

We focus in particular on the tension between actuarial, individual, and group notions of fairness, which correspond to different levels at which equity can be defined and enforced.

Classical actuarial fairness is rooted in the principle of risk equivalence. Policyholders with identical risk profiles should, in expectation, pay identical premiums. This notion is closely tied to portfolio balance, mutualization, and insurer solvency, as it ensures that premiums remain aligned with underlying risk. However, actuarially fair pricing does not address fair treatment across protected groups.

A prominent strand of the literature adopts individual-level notions of fairness, focusing on the elimination of proxy discrimination. Lindholm, Richman, Tsanakas, and Mario V. Wüthrich (2024b) and related work propose a discrimination-free pricing based on inverse probability weighting and sensitivity-based corrections. In such approach to fairness, individuals with identical observable characteristics receive identical premiums, regardless of their group membership. This eliminates direct discrimination, and mitigates omitted variable bias. Nevertheless, it does not fully address indirect discrimination, as the observed covariates still encode structural inequalities. Consequently, group-level disparities in premiums may persist.

In other words, it fixes the statistical bias, but not the societal bias embedded in the data.

To explicitly control for such disparities, group-level notions of fairness must be considered, such as demographic parity. Early actuarial contributions by Hu et al. (2023) and Hu et al. (2024) use optimal transport tools, such as Wasserstein barycenters, to enforce group fairness. Demographic parity imposes equality of outcomes across protected groups, independently of their risk profiles. While often viewed as ill-suited to insurance pricing (Lindholm, Richman, Tsanakas, and Mario V. Wüthrich 2024b), as it directly discriminates, it can be interpreted as a solidarity-driven criterion or a form of positive discrimination, introducing intentional redistribution across groups. From an actuarial perspective, enforcing demographic parity aligns with principles of collective risk sharing and induces cross-subsidization, as low-risk groups may overpay while high-risk groups are subsidized. This weakens the direct link between premiums and risk, creating a trade-off between fairness and actuarial adequacy. Such distortions may impact profitability and generate adverse selection, as mispriced risks self-select across competing insurers.

A series of papers by O. Côté et al. (2025) and M.-P. Côté et al. (2026) emphasize the importance of causal theory and market-level implications of fairness corrections, showing that portfolio-specific fairness does not necessarily extend to market-wide equity and may exacerbate adverse selection in competitive settings.

Together, these contributions highlight fundamental disagreements not only about what constitutes fairness, but also about at which level (individual, portfolio, or market) it should be enforced.

Fair pricing is further complicated by regulatory constraints. Under the Affordable Care Act in the United States, health insurers are restricted to a small set of rating variables, prioritizing solidarity over risk-based pricing. In jurisdictions such as Colorado, insurers are prohibited from using ethnicity data while still being required to prove the absence of discriminatory outcomes, creating an implicit demand for inference and counterfactual fairness audits.

These regulatory paradoxes underscore the impracticality of committing to a single fairness notion. Instead, insurers must navigate a complex landscape where fairness is enforced differently depending on legal, social, and market contexts.

Despite the variety of fairness approaches proposed in the literature, implementations typically rely on a single criterion. Moreover, fairness is usually enforced through post-processing adjustments rather than embedded within the pricing model, limiting transparency into how fairness constraints interact with model uncertainty, nonlinearity, and predictive structure. These limitations motivate a unified modeling framework in which multiple fairness criteria can be jointly incorporated.

To structure this landscape, we adopt the typology introduced by O. Côté et al. (2025), which classifies fairness interventions into five families (best-estimate, unaware, aware, hyperaware, and corrective). Additional insights are drawn from Xin

et al. (2023), who systematically contrast individual and group fairness criteria under different regulatory regimes, and from sensitivity-based analyses (Lindholm, Richman, Tsanakas, and Mario V Wüthrich 2026) which further sharpen the disagreement regarding proxy discrimination and causal relevance.

To study these competing fairness notions, we require a predictive model that can capture non-linear dependencies of the response (such as claim amounts, claim counts, or claim frequencies) on multiple rating factors and that can also embed multiple fairness interventions directly into its architecture.

Traditional regression models often rely on parametric assumptions, where the actuary specifies a fixed functional form of the mapping relating the covariates to the target variable. For instance, Generalized Linear Models (GLMs) are widely used in pricing motor and health insurance, where claim frequency is modeled using a Poisson distribution with a log-link function to ensure portfolio-balance (Denuit, Hainaut, et al. 2019c). Similarly, Generalized Additive Models (GAMs) extend GLMs by allowing smooth, non-linear effects of predictors, such as driver age or vehicle power, using spline-based components (Denuit, Hainaut, et al. 2019b). While these approaches are interpretable and statistically efficient, they require substantial prior knowledge about the data-generating process and involve model selection procedures to ensure the most appropriate model is retained.

When the true relationship between features and the outcome is complex or unknown, non-parametric regression offers a more flexible alternative. Methods such as splines or kernel regression can capture arbitrary non-linear patterns by defining local smoothing over the covariate space. In insurance, kernel methods have been used for tasks like claim cost reserving (Shim et al. 2011; Hiabu 2016) and mortality modeling (Risk et al. 2024). However, such methods depend on design choices like knot placement or bandwidth selection, which are often heuristic and sensitive to input dimensionality. More recently, regression trees, random forests, and gradient boosting machines (e.g., XGBoost, LightGBM) have gained popularity in actuarial modeling due to their ability to model complex interactions without explicit specification of the functional form (Denuit, Hainaut, et al. 2020; Chevalier et al. 2025). These ensemble methods have proven effective in various insurance tasks, including pricing (Henckaerts et al. 2021), mortality forecasting (Bjerre 2022), and fraud detection (Taha et al. 2020).

A third alternative is to adopt a Bayesian non-parametric strategy, and directly model the unknown underlying function. Bayesian non-parametric methods like Gaussian processes (GPs) place a prior distribution directly over the space of functions, allowing the data to inform both the shape and complexity of the model. In other words, rather than assuming a fixed functional form, we formulate the uncertainty about the mapping into a prior over mappings and then use Bayes rule to reach the posterior. As a distribution over functions, a GP does not require an explicit parametric form and yields a posterior that reflects both prior beliefs and empirical evidence. This makes it particularly suitable for capturing complex, non-linear relationships in insurance data where model uncertainty is a central concern.

In insurance applications, GPs have been used in mortality modeling (Ludkovski et al. 2018), claim reserving (Ang et al. 2022), and option pricing (Hainaut and Vrans 2024; Hainaut 2026).

We contribute to the actuarial literature as follow. First, we extend the applicability of GPR to realistic insurance settings by building a balance-corrected claim count model with exposure, relying on Generalized Gaussian Process Regression (GGPR) for non-Gaussian data. In particular, we build on existing work on GGPR for Poisson and Tweedie-type likelihoods with offset terms, which has been shown to provide a flexible alternative to GLMs in frequency modeling for insurance portfolios (see, e.g., B. Wang et al. (2014), Lloyd et al. (2015), Rigby et al. (2005), and Hainaut and Denuit (2025)). This balance-corrected extension (bGGPR) is essential to ensure that fairness corrections remain compatible with standard actuarial practice. Second, rather than committing to a single criterion a priori, we show how a wide range of fairness criteria, spanning both individual and group-level notions, can be incorporated into a single FairGGPR model through targeted modifications of the kernel structure and objective function. Third, we assess these fairness interventions both individually and jointly, comparing their effects side by side on standard fairness and accuracy metrics, and on the overall balance between collected premiums and incurred claim. Lastly, we examine a strategic question that has received comparatively little attention in the literature (M.-P. Côté et al. 2026). Can fairness enforcement benefit insurers financially? Using simulated portfolios, we show how fairness interventions and the proportion of protected policyholders influence the aggregate balance between collected premiums and incurred claims in distinct ways, with potential implications for policyholder retention and market share. This perspective moves beyond the traditional fairness-accuracy trade-off by highlighting how fairness enforcement may shape portfolio and market dynamics rather than acting solely as a regulatory constraint.

To the best of our knowledge, this is the first work that integrates multiple fairness criteria within a single Bayesian non-parametric model. By doing so, we provide a framework through which fairness definitions can be compared, and evaluated according to the specific objectives and constraints faced by insurers and regulators, challenging the common binary framing that one fairness notion must be chosen over another.

The remainder of the chapter is organized as follows. Section 5.2 introduces the balance-corrected Generalized Gaussian Process Regression (bGGPR) model for Poisson count data with risk exposure and sets up the notations. Section 5.3 presents the multi-criteria FairGGPR model and details how different fairness criteria are integrated within its Bayesian architecture. Section 5.4 reports empirical results, focusing on fairness-accuracy trade-offs, and portfolio balance. Section 5.5 concludes by discussing implications for insurers and regulators, and by arguing for a context-dependent application of fairness criteria in insurance pricing.

5.2 bGGPR for log-link counts

This Section introduces the balance-corrected Generalized Gaussian Process Regression (bGGPR) for Poisson count data with exposure, which is particularly relevant for modeling insurance claim frequencies. The GGPR was initially proposed by Chan et al. (2011) where only similarity between risk profiles matters. GGPR applies to response obeying a distribution in the Exponential Dispersion (ED) family, exactly as with GLMs and GAMs. The GGPR was recently adapted by Hainaut and Denuit (2025) to account for policies with different exposures to risk, like the coverage duration. This is important for actuarial applications where exposures often play a major role, as they reflect the volume of the risk. We further extend it to ensure aggregate balance at the portfolio level.

5.2.1 Poisson frequency model with exposure

Consider an insurance portfolio $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, where each policy $i = 1, \dots, n$ is described by a vector of rating factors $\mathbf{x}_i = (x_{i1}, \dots, x_{im})$. These covariates are partitioned into a set of allowed covariates $\mathcal{A}$ and a set of prohibited or sensitive covariates $\mathcal{D}$, with $\mathcal{A} \cup \mathcal{D} = \{1, \dots, m\}$ and $\mathcal{A} \cap \mathcal{D} = \emptyset$. We write $\mathbf{a}_i = (x_{ij})_{j \in \mathcal{A}}$ and $\mathbf{d}_i = (x_{ij})_{j \in \mathcal{D}}$ the corresponding subvectors. In this Section, the (b)GGPR model is formulated using the full set of available covariates $\mathbf{X} = (\mathbf{A}, \mathbf{D}) \in \mathbb{R}^{n \times m}$. In the next Section, we explicitly distinguish between allowed and protected attributes in order to introduce fairness modeling constraints.

Proposition 5.1. *Let N denote the number of claims reported by a policy over an exposure period $\nu > 0$. Conditionally on the covariates $\mathbf{x}$, we assume a Poisson model with exposure:*

$$N \mid \mathbf{x} \sim \mathcal{P}(\nu\lambda(\mathbf{x})),$$

where $\lambda(\mathbf{x}) > 0$ is the claim intensity per unit of exposure. Define the key ratio:

$$Y := \frac{N}{\nu},$$

which represents the claim frequency per unit of exposure. Since $N = \nu Y$ takes values in $\mathbb{N}$, the conditional probability mass function of Y is induced by that of N :

$$p(y \mid \mathbf{x}) = \mathbb{P}(N = \nu y \mid \mathbf{x}) = \exp(-\nu\lambda(\mathbf{x})) \frac{(\nu\lambda(\mathbf{x}))^{\nu y}}{(\nu y)!},$$

which is a density in the ED family (Wuthrich et al. 2021) as shown in Proof 5.A.1 in Appendix.

The objective is to approximate the (unknown) conditional expectation of the response $\nu Y \sim \mathcal{P}(\lambda(\mathbf{x})\nu)$ given the covariates to derive the amount of pure premium:

$$\mu(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{x}] = \gamma'(\theta(\mathbf{x})) = \lambda(\mathbf{x}),$$

with $\theta(\mathbf{x}) = \ln \lambda(\mathbf{x})$ and $\gamma'(\cdot) = \exp(\cdot)$.

Remark 5.1 (Generalization to the ED family). While we present the notations using count data, the methodology can be generalized to Gamma or Inverse Gaussian distribution for claim severities.

To model the conditional mean $\mu(\mathbf{x})$, we introduce a latent function $g(\cdot)$ defined on the log scale:

$$\ln \mu(\mathbf{x}) = g(\mathbf{x}) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma_n^2) .$$

This formulation induces nonlinear dependence of the claim frequency on the covariates through a kernel function (similarity matrix), while retaining the interpretability of a log-link model.

Proposition 5.2. *Let $\mathbf{g} = (g(\mathbf{x}_1), \dots, g(\mathbf{x}_n))^\top \in \mathbb{R}^n$ denote the latent function values at the n training inputs. Under the Gaussian Process (GP) prior, $\mathbf{g}$ follows a multivariate Gaussian distribution with n -mean vector $\mathbf{0}$ and $n \times n$ Gram matrix:*

$$C(\mathbf{X}, \mathbf{X}) = K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I}_n,$$

where $K(\mathbf{X}, \mathbf{X}) \in \mathbb{R}^{n \times n}$ is the kernel matrix with entries $[K(\mathbf{X}, \mathbf{X})]_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, $i, j = 1, \dots, n$, and σ_n^2 is a small scale variability. See Proof 5.A.1 in Appendix.

5.2.2 Posterior inference via Laplace approximation

Given observations (key ratios) $\mathbf{y} = (y_1, \dots, y_n)^\top$ and exposures $\boldsymbol{\nu} = (\nu_1, \dots, \nu_n)^\top$, the conditional log-likelihood takes the Poisson form:

$$\ln p(\mathbf{y} \mid \mathbf{g}) = \sum_{i=1}^n \nu_i (y_i g_i - \exp(g_i)) + c(y_i, \nu_i),$$

where $c(\cdot)$ collects terms independent of $\mathbf{g}$. The posterior density of the latent vector $\mathbf{g}$ is then:

$$p(\mathbf{g} \mid \mathbf{y}, \mathbf{X}) \propto p(\mathbf{y} \mid \mathbf{g}) p(\mathbf{g} \mid \mathbf{X}) .$$

Because the key-ratio Y follows a non-Gaussian distribution, the Poisson likelihood is not conjugate to the Gaussian prior. Therefore, the posterior distribution of the latent function $\mathbf{g}$ is analytically intractable (since it is no longer Gaussian) and must be approximated. Following C. K. Williams et al. (2006), we approximate the posterior using a Laplace approximation. Specifically, the posterior is approximated by a Gaussian distribution centered at its mode:

$$q(\mathbf{g} \mid \mathbf{y}, \mathbf{X}) \approx \mathcal{N}(\hat{\mathbf{g}}, \Sigma_{\mathbf{g}}).$$

The n -dimensional mean vector $\hat{\mathbf{g}}$ is equal to (see Proof 5.A.1):

$$\hat{\mathbf{g}} = \arg \max_{\mathbf{g}} \{\psi(\mathbf{g})\}, \quad (5.2.1)$$

with

$$\psi(\mathbf{g}) = \boldsymbol{\nu}^\top (\mathbf{y} \odot \mathbf{g} - \exp(\mathbf{g})) - \frac{1}{2} \mathbf{g}^\top C(\mathbf{X}, \mathbf{X})^{-1} \mathbf{g},$$

where $\odot$ denotes the Hadamard product. The $n \times n$ covariance matrix $\Sigma_{\mathbf{g}}$ is given by the inverse negative Hessian of the log-posterior evaluated at $\hat{\mathbf{g}}$ (see Proof 5.A.1):

$$\Sigma_{\mathbf{g}} = \left(C(\mathbf{X}, \mathbf{X})^{-1} + H(\hat{\mathbf{g}}) \right)^{-1}, \quad (5.2.2)$$

with the $n \times n$ diagonal matrix $H(\hat{\mathbf{g}}) = \text{diag}(\boldsymbol{\nu} \odot \exp(\hat{\mathbf{g}}))$.

The optimization problem defining $\hat{\mathbf{g}}$ in Eq. (5.2.1) is typically solved using Newton-Raphson iterations with $\mathbf{g}^{(0)} = \ln(\frac{1}{n} \sum_i y_i)$:

$$\hat{\mathbf{g}}^{(j)} = \hat{\mathbf{g}}^{(j-1)} - \left(\nabla^2 \psi(\hat{\mathbf{g}}^{(j-1)}) \right)^{-1} \nabla \psi(\hat{\mathbf{g}}^{(j-1)}), \quad (5.2.3)$$

where the gradient $\nabla \psi(\hat{\mathbf{g}}^{(j-1)})$ takes the following simple form:

$$\nabla \psi(\hat{\mathbf{g}}^{(j-1)}) = \boldsymbol{\nu} \left(\mathbf{y} - \exp(\hat{\mathbf{g}}^{(j-1)}) \right) - C(\mathbf{X}, \mathbf{X})^{-1} \hat{\mathbf{g}}^{(j-1)},$$

and the Hessian is equal to:

$$\begin{aligned} \nabla^2 \psi(\hat{\mathbf{g}}^{(j-1)}) &= -\text{diag}(\boldsymbol{\nu} \odot \exp(\hat{\mathbf{g}}^{(j-1)})) - C(\mathbf{X}, \mathbf{X})^{-1} \\ &= -H(\hat{\mathbf{g}}^{(j-1)}) - C(\mathbf{X}, \mathbf{X})^{-1}. \end{aligned}$$

5.2.3 Prediction and premium estimation

Proposition 5.3 (New Policy). *Let $\mathbf{x}_*$ denote a new policy (out-of-sample) with $k_{**} = k(\mathbf{x}_*, \mathbf{x}_*) \in \mathbb{R}$. Let $k_* = k(\mathbf{x}_*, \mathbf{X}) \in \mathbb{R}^{1 \times n}$ denote the vector of similarities between $\mathbf{x}_*$ and the training inputs $\mathbf{X}$:*

$$k_* = k(\mathbf{x}_*, \mathbf{X}) = (k(\mathbf{x}_*, \mathbf{x}_1), \dots, k(\mathbf{x}_*, \mathbf{x}_n)),$$

and $k_*^\top \in \mathbb{R}^{n \times 1}$ its transpose. Under the Laplace approximation, the predictive distribution of the latent variable $g_* = g(\mathbf{x}_*)$ is approximately Gaussian,

$$g_* \mid \mathbf{x}_*, \mathbf{X}, \mathbf{y} \approx \mathcal{N}(m_*, s_*^2),$$

with

$$\begin{aligned} m_* &= k_* C(\mathbf{X}, \mathbf{X})^{-1} \hat{\mathbf{g}}, \\ s_*^2 &= k_{**} - k_* (C(\mathbf{X}, \mathbf{X}) + H(\hat{\mathbf{g}}))^{-1} k_*^\top. \end{aligned}$$

See Proof 5.A.1 in Appendix.

Remark 5.2 (Interpretation of the predictive equations in Proposition 5.3). The predictive mean m_* can be understood as a regularised similarity-weighted average of the latent risk values $\hat{\mathbf{g}}$ (estimated from the training portfolio). The vector $k_* = k(\mathbf{x}_*, \mathbf{X})$ determines how strongly each past observation influences the prediction at $\mathbf{x}_*$, while the matrix $C(\mathbf{X}, \mathbf{X})^{-1}$ corrects for dependencies between training points induced by the kernel, preventing redundant or highly correlated observations from being over-weighted. In this sense, the prediction at $\mathbf{x}_*$ is obtained by aggregating information from training policies that are most similar to $\mathbf{x}_*$. The predictive variance s_*^2 quantifies the remaining uncertainty. It can be interpreted as the prior uncertainty k_{**} , reduced by the amount of information explained by similar training points. Hence, the variance decreases when $\mathbf{x}_*$ is well represented in the training set and increases when the policy lies in a sparsely populated region of the feature space.

Transforming back to the original scale, the predictive mean, that is the estimate of the key ratio Y_* , satisfies:

$$\mathbb{E}[Y_* \mid \mathbf{x}_*, \mathbf{X}, \mathbf{y}] \approx \mathbb{E}[\exp(g_*)] = \exp\left(m_* + \frac{1}{2}s_*^2\right),$$

which corresponds to the log-normal moment correction induced by the exponential link. This quantity provides an estimate of the pure premium per unit of exposure. The corresponding expected claim count is:

$$\mathbb{E}[N_* \mid \mathbf{x}_*, \mathbf{X}, \mathbf{y}] \approx \nu_* \exp\left(m_* + \frac{1}{2}s_*^2\right).$$

This transformation back to the original scale introduces a log-normal correction term $\exp(m_* + \frac{1}{2}s_*^2)$, which implies that both the estimated risk level and its uncertainty contribute positively to the final premium. In particular, higher epistemic uncertainty leads to higher expected premiums, reflecting a risk-loading effect induced by the probabilistic model.

Remark 5.3 (Pure premium). In the remainder of the article, we write $\hat{\pi}(\mathbf{x}_*) = \mathbb{E}[Y_* \mid \mathbf{x}_*, \mathbf{X}, \mathbf{y}]$ for the pure premium, i.e., the expected loss per unit of exposure, excluding any loading factors (such as expenses, risk margins, or profit loadings).

In Algorithm 10 below, we detail the pseudo-code corresponding to the GGPR model for count data under a log-link specification. Note that, in practice, the Hessian matrix is never explicitly inverted. Instead, all linear systems are solved using Cholesky decompositions together with Woodbury-type identities to ensure numerical stability and scalability.

Algorithm 10: GGPR for log-link count data

Data: Training data X , frequencies y , exposures ν , kernel $k(\cdot, \cdot)$
 Compute Gram matrix: $C = K(X, X) + \sigma_n^2 I_n$
 Initialize latent function: $\hat{\mathbf{g}}^{(0)} = \ln\left(\frac{1}{n} \sum_i y_i\right)$
for $j = 1, 2, \dots$ *until convergence* **do**
 Compute gradient: $\nabla\psi(\hat{\mathbf{g}}^{(j-1)}) = \nu \odot (y - \exp(\hat{\mathbf{g}}^{(j-1)})) - C^{-1}\hat{\mathbf{g}}^{(j-1)}$
 Compute Hessian: $H^{(j-1)} = -\text{diag}(\nu \odot \exp(\hat{\mathbf{g}}^{(j-1)})) - C^{-1}$
 Update latent mode using Newton-Raphson:
 $\hat{\mathbf{g}}^{(j)} = \hat{\mathbf{g}}^{(j-1)} - (H^{(j-1)})^{-1} \nabla\psi(\hat{\mathbf{g}}^{(j-1)})$
 Compute posterior covariance: $\Sigma_g = (C^{-1} + \text{diag}(\nu \odot \exp(\hat{\mathbf{g}})))^{-1}$
 Predict new policy x_* : $k_* = k(x_*, X)$, $k_{**} = k(x_*, x_*)$
 Compute predictive mean and variance of latent function:

$$m_* = k_*^\top C^{-1} \hat{\mathbf{g}}, \quad s_*^2 = k_{**} - k_*^\top (C + H^{-1})^{-1} k_*$$

Compute predictive premium (key ratio): $\hat{\pi}_* = \mathbb{E}[Y_*] \approx \exp\left(m_* + \frac{1}{2}s_*^2\right)$
 Compute expected claim count: $\mathbb{E}[N_*] \approx \nu_* \hat{\pi}_*$

Remark 5.4 (Kernel choice). Throughout this work, we use the squared exponential kernel due to its smoothness and universal approximation properties. While more elaborate kernel constructions may be preferable for mixed continuous and categorical data, kernel engineering is not the focus of this study and is left for future work. It is defined as:

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{1}{2\ell^2} \|\mathbf{x} - \mathbf{x}'\|^2\right), \quad (5.2.4)$$

where $\|\cdot\|^2$ denotes the squared Euclidean distance, σ_f^2 controls the marginal variance of the latent function, and ℓ represents the characteristic length scale governing smoothness. Larger values of ℓ induce smoother functions, while smaller values allow for more rapid local variation. The optimal set of hyper-parameters $\Theta = \{\sigma_f^2, \ell\}$ is estimated by minimizing the Poisson deviance with the Nelder-Mead algorithm (Nelder et al. 1965) as further detailed in Section 5.4.

5.2.4 Portfolio balance

In insurance pricing, portfolio balance (or portfolio-level calibration) requires that the total predicted number of claims across all policies matches the total observed number of claims. Formally, if N_i denotes the observed claims and ν_i the exposure for policy i , a model is portfolio-balanced if:

$$\sum_{i=1}^n \nu_i \hat{Y}_i = \sum_{i=1}^n N_i,$$

where $\hat{Y}_i$ is the predicted claim frequency. This actuarial property ensures that the insurer's total expected losses are consistent with the realized experience, and is critical for regulatory and solvency compliance (Lindholm and Mario V Wüthrich 2025).

Proposition 5.4 (Portfolio balance in GGPR). *Let N_i denote the number of claims for policy i , ν_i its exposure, and $Y_i = N_i/\nu_i$ the claim frequency. Under the GGPR for log-link counts, let $\hat{\mathbf{g}}$ denote the Laplace posterior mode. Define the predictive claim frequency using the log-normal correction:*

$$\hat{Y}_i := \exp\left(m_i + \frac{1}{2}s_i^2\right),$$

where m_i and s_i^2 are the posterior predictive mean and variance of the latent g_i . Actuarial fairness at the posterior mode holds if $C^{-1}\hat{\mathbf{g}}$ sums to zero:

$$\sum_{i=1}^n \nu_i \exp(\hat{g}_i) = \sum_{i=1}^n N_i \quad \text{s.t.} \quad \mathbf{1}_n^\top C^{-1}\hat{\mathbf{g}} = 0. \quad (5.2.5)$$

It follows that the GGPR predictions $\hat{Y}_i$ do not satisfy this equality:

$$\sum_{i=1}^n \nu_i \hat{Y}_i \neq \sum_{i=1}^n N_i.$$

See Proof 5.A.2 in Appendix.

Proposition 5.5 (Balance-corrected GGPR (bGGPR) at training points). *Let $\hat{\mathbf{g}}$ denote the Laplace posterior mode and define the predictive premiums at the training points:*

$$\hat{\boldsymbol{\pi}}(\hat{\mathbf{g}}) = \exp\left(\hat{\mathbf{g}} + \frac{1}{2} \text{diag}(\Sigma_{\hat{\mathbf{g}}})\right), \quad \Sigma_{\hat{\mathbf{g}}} = (C^{-1} + \text{diag}(\boldsymbol{\nu} \odot e^{\hat{\mathbf{g}}}))^{-1}.$$

Define the portfolio balance functional:

$$\phi(\hat{\mathbf{g}}) = \sum_{i=1}^n \nu_i \hat{\pi}_i(\hat{\mathbf{g}}) - \sum_{i=1}^n N_i.$$

Starting from $\hat{\mathbf{g}}^{(0)} = \hat{\mathbf{g}}$, consider the iterative updates:

$$\hat{\mathbf{g}}^{(k+1)} = \hat{\mathbf{g}}^{(k)} + \Delta \mathbf{g}^{(k)},$$

where the correction is given by the Gaussian process projection:

$$\Delta \mathbf{g}^{(k)} = \Sigma_{\hat{\mathbf{g}}} \mathbf{u}^{(k)} \left(\mathbf{u}^{(k)\top} \Sigma_{\hat{\mathbf{g}}} \mathbf{u}^{(k)} \right)^{-1} b^{(k)},$$

with:

$$\mathbf{u}^{(k)} = \boldsymbol{\nu} \odot \hat{\boldsymbol{\pi}}(\hat{\mathbf{g}}^{(k)}), \quad b^{(k)} = \sum_{i=1}^n N_i - \sum_{i=1}^n \nu_i \hat{\pi}_i(\hat{\mathbf{g}}^{(k)}).$$

Then, under mild regularity conditions, the sequence $(\hat{\mathbf{g}}^{(k)})$ converges locally to a fixed point satisfying:

$$\phi(\mathbf{g}^*) = 0,$$

i.e., the predictive premiums satisfy portfolio-level calibration:

$$\sum_{i=1}^n \nu_i \hat{\pi}_i(\mathbf{g}^*) = \sum_{i=1}^n N_i.$$

See Proof 5.A.2 in Appendix.

Note that this is not a post-hoc correction. It can be interpreted as an approximate inference procedure in a constrained Gaussian process. Each update step k corresponds to conditioning the Gaussian process $\mathbf{g} \sim \mathcal{N}(\hat{\mathbf{g}}, \Sigma_{\hat{\mathbf{g}}})$ on a linear pseudo-observation of the form:

$$\mathbf{u}^{(k)\top} \mathbf{g} = b^{(k)},$$

In this sense, the iterative projection:

$$\Delta \mathbf{g}^{(k)} = \Sigma_{\hat{\mathbf{g}}} \mathbf{u}^{(k)} (\mathbf{u}^{(k)\top} \Sigma_{\hat{\mathbf{g}}} \mathbf{u}^{(k)})^{-1} b^{(k)},$$

is exactly the posterior mean of the GP conditioned on the linear observation. Equivalently, the augmented Gram matrix:

$$\tilde{C}^{(k)} = \begin{bmatrix} \Sigma_{\hat{\mathbf{g}}} & \Sigma_{\hat{\mathbf{g}}} \mathbf{u}^{(k)} \\ \mathbf{u}^{(k)\top} \Sigma_{\hat{\mathbf{g}}} & \mathbf{u}^{(k)\top} \Sigma_{\hat{\mathbf{g}}} \mathbf{u}^{(k)} \end{bmatrix},$$

represents the joint covariance of $(\mathbf{g}, \mathbf{u}^{(k)\top} \mathbf{g})$, and conditioning on $\mathbf{u}^{(k)\top} \mathbf{g} = b^{(k)}$ produces the same projection. Thus, each iterative step minimally adjusts the GP posterior in Mahalanobis distance while enforcing the linearized portfolio-level calibration.

In Algorithm 11 below, we detail the pseudo-code corresponding to the bGGPR model for count data under a log-link specification satisfying portfolio balance.

Algorithm 11: Balance-corrected GGPR (bGGPR) for log-link count data

Data: Training data X , frequencies y , exposures ν , kernel $k(\cdot, \cdot)$
Output: Predictive premium $\hat{\pi}_*$ satisfying portfolio-level calibration
 Compute Gram matrix: $C = K(X, X) + \sigma_n^2 I_n$
 Initialize latent function: $\hat{\mathbf{g}}^{(0)} = \log(\frac{1}{n} \sum_{i=1}^n y_i)$
for $j = 1, 2, \dots$ *until convergence* **do**
 Compute gradient: $\nabla\psi(\hat{\mathbf{g}}^{(j-1)}) = \nu \odot (y - e^{\hat{\mathbf{g}}^{(j-1)}}) - C^{-1} \hat{\mathbf{g}}^{(j-1)}$
 Compute Hessian: $H^{(j-1)} = -\text{diag}(\nu \odot e^{\hat{\mathbf{g}}^{(j-1)}}) - C^{-1}$
 Newton update: $\hat{\mathbf{g}}^{(j)} = \hat{\mathbf{g}}^{(j-1)} - (H^{(j-1)})^{-1} \nabla\psi(\hat{\mathbf{g}}^{(j-1)})$
 Compute Laplace posterior covariance: $\Sigma_{\hat{\mathbf{g}}} = (C^{-1} + \text{diag}(\nu \odot e^{\hat{\mathbf{g}}}))^{-1}$
if *balance correction* **then**
 Compute training-point premiums: $\hat{\pi} = \exp(\hat{\mathbf{g}} + \frac{1}{2} \text{diag}(\Sigma_{\hat{\mathbf{g}}}))$
 Compute balance residual: $b = \sum_{i=1}^n N_i - \sum_{i=1}^n \nu_i \hat{\pi}_i$
 Define constraint direction: $\mathbf{u} = \nu \odot \hat{\pi}$
 Compute GP projection: $\Delta \mathbf{g} = \Sigma \mathbf{u} (\mathbf{u}^\top \Sigma \mathbf{u})^{-1} b$
 Update latent mode: $\hat{\mathbf{g}} \leftarrow \hat{\mathbf{g}} + \Delta \mathbf{g}$
 Update covariance: $\Sigma_{\hat{\mathbf{g}}} = (C^{-1} + \text{diag}(\nu \odot e^{\hat{\mathbf{g}}}))^{-1}$
 For a new policy x_* , compute: $k_* = k(x_*, X)$, $k_{**} = k(x_*, x_*)$
 Predictive mean and variance:
 $m_* = k_*^\top C^{-1} \hat{\mathbf{g}}$, $s_*^2 = k_{**} - k_*^\top (C + H^{-1})^{-1} k_*$
 Predictive balance corrected premium:
 $\hat{\pi}_* = \exp(m_* + \frac{1}{2} s_*^2)$, $\mathbb{E}[N_*] = \nu_* \hat{\pi}_*$

5.3 Multi-criteria FairGGPR

Fairness in actuarial pricing involves multiple, often incompatible notions, and most existing approaches address them in isolation. From an actuarial perspective, premiums must additionally satisfy portfolio balance to prevent systematic overpricing or underpricing at the portfolio level (see Section 5.2.4). Fairness interventions generally violate this condition unless explicit rebalancing constraints are imposed (see e.g., Jamotton and Hainaut (2026b)).

In this Section, we detail the introduction of fairness notions within the GGPR for log-link count data presented in Section 5.2. We first review existing fairness definitions organized into five families (best-estimate, unaware, aware, hyperaware, and corrective) by O. Côté et al. (2025). We control notions of individual fairness through the similarity matrix, and impose group-level fairness as expectations or covariance restrictions on posterior moments of g (similarly to balance correction in Section 5.2.4).

5.3.1 Fairness families

Let $Y \in \mathbb{R}_+$ denote the target loss variable (i.e., the claim frequency per unit of exposure), $A \in \mathcal{A}$ the vector of permitted rating factors, and $D \in \mathcal{D}$ the vector of protected attributes. We assume a latent risk function $g(\mathbf{x})$, with $\mathbf{x} = (\mathbf{a}, \mathbf{d})$, and a pricing rule $\pi(\cdot)$ derived from the predictive distribution of g . The different fairness families correspond to alternative ways of satisfying fairness in the pricing rule.

Definition 5.1 (Best-estimate premium). A best-estimate premium uses all available information (A, D) and is sufficient for both A and D :

$$\pi^{\text{BE}}(A, D) = \mathbb{E}[Y \mid A, D].$$

The best-estimate premium corresponds to the classical actuarial principle of risk classification, where premiums reflect the expected loss conditional on all observable risk characteristics. From a predictive standpoint, this rule is statistically optimal. However, the explicit use of protected attributes introduces direct discrimination whenever D is causally or associatively linked to Y . Such directly discriminating pricing is typically prohibited by fairness regulations.

Definition 5.2 (Unaware premium). An unaware premium excludes protected attributes D :

$$\pi^{\text{U}}(A) = \mathbb{E}[Y \mid A].$$

Although unaware pricing appears fair by construction, it generally suffers from omitted variable bias. When causal structures such as $A \leftarrow D \rightarrow Y$ are present, the permitted variables A may act as proxies for D , inducing indirect discrimination. This phenomenon has been extensively documented in both econometrics and algorithmic fairness (see e.g. Barocas et al. (2016) and Kusner, Loftus, et al. (2017)).

Definition 5.3 (Aware premium). An aware premium uses only permitted features A while correcting for the confounding effect of D :

$$\pi^{\text{A}}(A) = \mathbb{E}_D[\mathbb{E}[Y \mid A, D]].$$

Aware premiums eliminate direct discrimination and correct for omitted variable bias by marginalising over D . The discrimination-free pricing framework of Lindholm, Richman, Tsanakas, and Mario V Wüthrich (2022a), Lindholm, Richman, Tsanakas, and Mario V. Wüthrich (2024b), and Lindholm, Richman, Tsanakas, and Mario V Wüthrich (2024a) provides an actuarial formulation of this approach. However, $\pi^{\text{A}}(A)$ may still be statistically associated with D , leading to residual disparate impact through A (O. Côté et al. 2025, p.57).

Definition 5.4 (Hyperaware premium). A hyperaware premium predicts only the component of Y that is independent of D :

$$\pi^{\text{H}}(A) = \mathbb{E}[Y^{\perp D} \mid A],$$

where $Y^{\perp D}$ denotes a component of Y statistically independent of D .

Hyperaware pricing aims to remove both direct and indirect pathways from D to Y by predicting only the component of the target variable Y that is conditionally independent of the sensitive attribute D , while restricting inputs to allowed features A . Unlike the aware model, which adjusts for omitted variable bias in a predictive sense, the hyperaware model attempts to target the fair signal in Y , i.e., the variation in Y that is invariant under changes in D . Formally, we consider the decomposition:

$$Y = Y^{\perp D} + Y^{\parallel D},$$

where $Y^{\perp D}$ is the component of Y orthogonal (independent) to D , and $Y^{\parallel D}$ is the component of Y explained by D .

However, since $Y^{\perp D}$ is not directly observable, this decomposition is only well-defined under additional (strong) assumptions, and fairness improvements are difficult to quantify robustly. For these reasons, we chose not to model a hyperaware premium in this work. Hyperaware models nevertheless serve as an important conceptual bridge toward corrective pricing.

Definition 5.5 (Corrective premium). A corrective premium uses only the component of A independent of D :

$$\pi^C(A^{\perp D}) = \mathbb{E}[Y \mid A^{\perp D}], \quad A^{\perp D} \perp D.$$

Corrective premiums explicitly enforce independence between premiums and protected attributes, typically achieving demographic parity:

$$\pi^C(A^{\perp D}) \perp D.$$

This may require corrective direct discrimination, which remains controversial from a legal standpoint but may be justified when addressing structural inequalities. One can say that this corrective premium is fair because it avoids historically accurate direct discrimination and disparate impact through A . Recent approaches include distributional alignment via optimal transport and Wasserstein barycenters Hu et al. (2023) and Hu et al. (2024).

Fig. 5.3.1 summarizes the different fairness families positions along a trade-off surface between predictive accuracy, individual fairness, and group-level parity constraints.

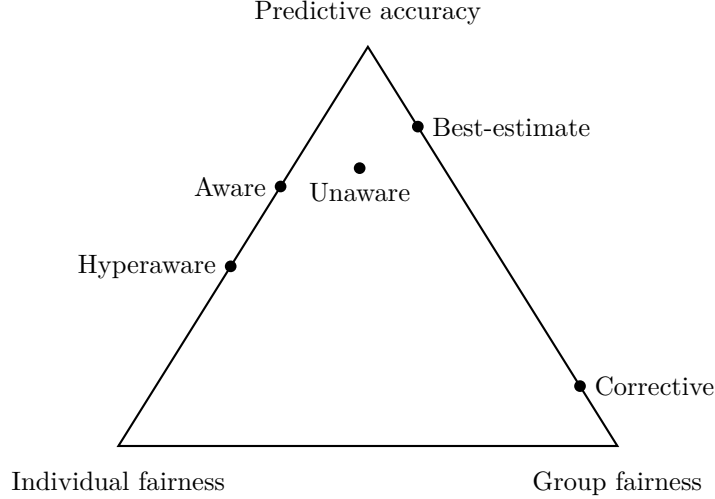


Figure 5.3.1: Positioning of fairness families within the trade-off between predictive accuracy, individual fairness, and group fairness. Best-estimate models maximize predictive accuracy but rely on sensitive attributes, leading to low fairness. Unaware models remove the sensitive attribute but suffer from information loss and provide limited fairness guarantees. Aware and hyperaware approaches improve individual fairness while maintaining good predictive performance, although they do not fully eliminate group disparities. Corrective methods enforce group fairness through redistribution, at the cost of predictive accuracy and individual-level consistency.

5.3.2 Multi-criteria FairGGPR for log-link counts

These different fairness notions are enforced by altering the kernel architecture or by introducing constraints on the predictive distribution.

Remark 5.5 (Use of the non-balance corrected GGPR for presentation). For clarity and to simplify the notation, the fairness interventions are presented in this Section using the non-balance corrected GGPR, but they can also be directly applied to the balance-corrected GGPR, as shown in the empirical results in Section 5.4.

Proposition 5.6 (Best-estimate kernel). *The best-estimate kernel is defined as:*

$$k^{\text{BE}}(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp \left(-\frac{1}{2\ell^2} \|\mathbf{x} - \mathbf{x}'\|^2 \right), \quad (5.3.1)$$

where $\mathbf{x}$ contains all available rating factors, including protected attributes, to encode similarity between policies.

Proposition 5.7 (Best-estimate GGPR premium). *Let $\mathbf{x}_*$ denote a new policy with best-estimate kernel vector:*

$$k_*^{\text{BE}} = k(\mathbf{x}_*, \mathbf{X}), \quad k_{**}^{\text{BE}} = k(\mathbf{x}_*, \mathbf{x}_*).$$

Under the Laplace approximation (with canonical link), the best-estimate GGPR premium corresponds to the baseline, unconstrained GGPR for log-counts:

$$\pi^{\text{BE}}(\mathbf{x}_*) := \mathbb{E}[Y_* \mid \mathbf{x}_*, \mathbf{X}, \mathbf{y}] \approx \exp(m_*^{\text{BE}} + \tfrac{1}{2}s_*^2{}^{\text{BE}}),$$

with:

$$\begin{aligned} m_*^{\text{BE}} &= k_*^{\text{BE}} (C^{\text{BE}})^{-1} \hat{\mathbf{g}}^{\text{BE}}, \\ s_*^2{}^{\text{BE}} &= k_{**}^{\text{BE}} - k_*^{\text{BE}} \left(C^{\text{BE}} + H(\hat{\mathbf{g}}^{\text{BE}})^{-1} \right)^{-1} k_*^{\text{BE}\top}, \end{aligned}$$

where the best-estimate Gram matrix is:

$$C^{\text{BE}} = \sigma_n^2 \mathbf{I}_n + k^{\text{BE}}.$$

If policies $\mathbf{x}$ and $\mathbf{x}'$ are similar then the corresponding $\pi(\mathbf{x})$ and $\pi(\mathbf{x}')$ will be strongly similar. Since the best-estimate kernel, in Eq. (5.3.1), uses all attributes, even small changes in $\mathbf{d}$ can lead to large changes in $k(\mathbf{x}, \mathbf{x}')$, explicitly discriminating.

The kernel architecture can be modified to enforce fairness, especially individual fairness, which requires that similar individuals should be treated similarly (i.e., receive similar premium). Translated into the kernel language, if individuals $\mathbf{x}$ and $\mathbf{x}'$ are similar, except in the sensitive variables space, then $k(\mathbf{x}, \mathbf{x}')$ should remain high. This leads us to the unaware and aware kernels.

Proposition 5.8 (Unaware kernel). *The unaware kernel is defined as:*

$$k^{\text{U}}(\mathbf{a}, \mathbf{a}') = \sigma_f^2 \exp\left(-\frac{1}{2\ell^2} \|\mathbf{a} - \mathbf{a}'\|^2\right), \quad (5.3.2)$$

where $\mathbf{a}$ contains allowed rating factors.

The unaware kernel ensures that, different sensitive values do not affect similarity, that is, regardless of $\mathbf{d} \neq \mathbf{d}'$, $\mathbf{a} = \mathbf{a}' \Rightarrow k = 1$.

Proposition 5.9 (Unaware GGPR premium). *Let $\mathbf{a}_*$ denote a new policy with unaware kernel vector:*

$$k_*^{\text{U}} = k(\mathbf{a}_*, \mathbf{A}), \quad k_{**}^{\text{U}} = k(\mathbf{a}_*, \mathbf{a}_*).$$

Under the Laplace approximation (with canonical link), the unaware GGPR premium corresponds to the baseline best-estimate GGPR for log-counts, but restricted to the allowed (non-sensitive) input variables:

$$\pi^{\text{U}}(\mathbf{a}_*) := \mathbb{E}[Y_* \mid \mathbf{a}_*, \mathbf{A}, \mathbf{y}] \approx \exp(m_*^{\text{U}} + \tfrac{1}{2}s_*^2{}^{\text{U}}),$$

with:

$$\begin{aligned} m_*^U &= k_*^U (C^U)^{-1} \hat{\mathbf{g}}^U, \\ s_*^{2U} &= k_{**}^U - k_*^U \left(C^U + H(\hat{\mathbf{g}}^U)^{-1} \right)^{-1} k_*^{U\top}, \end{aligned}$$

where the unaware Gram matrix is:

$$C^U = \sigma_n^2 \mathbf{I}_n + k^U.$$

While the unaware GGPR removes direct dependence and discrimination on sensitive attributes D , through the kernel architecture in Eq. (5.3.2), it does not fully address discrimination. This is because ignoring sensitive attributes does not eliminate their indirect influence on non-sensitive features, a phenomenon known as proxy discrimination or omitted variable bias.

Proposition 5.10 (Aware kernel). *Let $\mathbf{x} = (\mathbf{a}, \mathbf{d})$ denote the full feature vector, where $\mathbf{a}$ are allowed attributes and $\mathbf{d}$ sensitive attributes. The aware kernel is defined as the population-averaged kernel:*

$$k^A(\mathbf{a}, \mathbf{a}') := \mathbb{E}_{\mathbf{d}, \mathbf{d}'} [k^{\text{BE}}((\mathbf{a}, \mathbf{d}), (\mathbf{a}', \mathbf{d}'))], \quad (5.3.3)$$

where the expectation is taken with respect to the empirical distribution of the sensitive attributes.

The aware kernel measures similarity between two individuals by averaging the best-estimate kernel over the distribution of sensitive attributes.

In practice, the expectation in Eq. (5.3.3) is approximated by Monte-Carlo sampling. Let $\{\mathbf{d}^{(m)}\}_{m=1, \dots, M}$ be independent and identically distributed samples from the empirical distribution of $\mathbf{D}$. The aware kernel admits the approximation:

$$k^A(\mathbf{a}, \mathbf{a}') \approx \frac{1}{M} \sum_{m=1}^M k^{\text{BE}} \left((\mathbf{a}, \mathbf{d}^{(m)}), (\mathbf{a}', \mathbf{d}^{(m)}) \right).$$

Proposition 5.11 (Aware GGPR premium). *Let $\mathbf{a}_*$ denote a new policy with aware kernel vector*

$$k_*^A = k^A(\mathbf{a}_*, \mathbf{A}), \quad k_{**}^A = k^A(\mathbf{a}_*, \mathbf{a}_*).$$

Under the Laplace approximation (with canonical link), the aware GGPR premium corresponds to the baseline best-estimate GGPR for log-counts, but with the kernel averaged over the sensitive attributes:

$$\pi^A(\mathbf{a}_*) := \mathbb{E}[Y_* \mid \mathbf{a}_*, \mathbf{A}, \mathbf{y}] \approx \exp(m_*^A + \frac{1}{2} s_*^{2A}),$$

with

$$\begin{aligned} m_*^A &= k_*^A (C^A)^{-1} \hat{\mathbf{g}}^A, \\ s_*^{2A} &= k_{**}^A - k_*^A \left(C^A + H(\hat{\mathbf{g}}^A)^{-1} \right)^{-1} k_*^{A\top}, \end{aligned}$$

where the aware Gram matrix is defined as:

$$C^A = \sigma_n^2 \mathbf{I}_n + k^A.$$

The approach is inspired by the discrimination-free insurance pricing proposed by Lindholm, Richman, Tsanakas, and Mario V Wüthrich (2022a). In their framework, fairness is enforced by averaging predicted premiums over sensitive attributes,

$$\pi^{\text{fair}}(\mathbf{a}) = \mathbb{E}_{\mathbf{d}}[\pi^{\text{BE}}(\mathbf{a}, \mathbf{d})],$$

which can be interpreted as a post-stratification step applied after model estimation. Post-hoc averaging, as considered here by Lindholm, Richman, Tsanakas, and Mario V Wüthrich (2022a), removes explicit dependence on $\mathbf{d}$ at the level of point predictions (by averaging pointwise expected premiums over the sensitive attributes) but it understates predictive uncertainty by ignoring variability arising from marginalization over sensitive attributes.

In contrast, the aware GGPR integrates out sensitive information at the kernel level (into the similarity measure), prior to posterior inference. This affects the covariance structure of the latent function and, consequently, both posterior mean and uncertainty. As a consequence, predictive uncertainty under aware GGPR reflects both the conditional variability given sensitive attributes and the additional uncertainty arising from marginalizing over them. This distinction can be made explicit using the law of total variance:

$$\text{Var}^A(g_*) = \mathbb{E}_{\mathbf{d}}[\text{Var}(g_* \mid \mathbf{a}_*, \mathbf{d})] + \text{Var}_{\mathbf{d}}[\mathbb{E}(g_* \mid \mathbf{a}_*, \mathbf{d})].$$

Proposition 5.12 (Corrective GGPR). *Under the Laplace approximation with canonical log-link, the corrective GGPR premium is defined as:*

$$\pi^C(\mathbf{x}_*) := \mathbb{E}[Y_* \mid \mathbf{x}_*, \mathbf{X}, \mathbf{y}] \approx \exp\left(m_*^C + \frac{1}{2}s_*^{2C}\right),$$

with:

$$\begin{aligned} m_*^C &= k_*^\top C^{-1} \hat{\mathbf{g}}^C, \\ s_*^{2C} &= k_{**} - k_*^\top (C + H(\hat{\mathbf{g}}^C)^{-1})^{-1} k_*. \end{aligned}$$

Let $\hat{\mathbf{g}}$ denote the Laplace posterior mode and define the predictive premiums at the training points:

$$\hat{\pi}(\hat{\mathbf{g}}) = \exp\left(\hat{\mathbf{g}} + \frac{1}{2} \text{diag}(\Sigma_{\hat{\mathbf{g}}})\right), \quad \Sigma_{\hat{\mathbf{g}}} = (C^{-1} + \text{diag}(\boldsymbol{\nu} \odot e^{\hat{\mathbf{g}}}))^{-1}.$$

Let $D \in \mathbb{R}^{n \times d}$ denote the matrix of d sensitive attributes, and define the centered matrix:

$$D_c = C_n D, \quad C_n = I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top.$$

Define the corrective function:

$$\phi(\hat{\mathbf{g}}) = D_c^\top \hat{\boldsymbol{\pi}}(\hat{\mathbf{g}}).$$

Starting from $\hat{\mathbf{g}}^{(0)} = \hat{\mathbf{g}}$, consider the iterative updates:

$$\hat{\mathbf{g}}^{(k+1)} = \hat{\mathbf{g}}^{(k)} + \Delta \mathbf{g}^{(k)},$$

where the correction is given by the Gaussian projection:

$$\Delta \mathbf{g}^{(k)} = \Sigma_{\hat{\mathbf{g}}} J^{(k)\top} \left(J^{(k)} \Sigma_{\hat{\mathbf{g}}} J^{(k)\top} \right)^{-1} \left(-\phi(\hat{\mathbf{g}}^{(k)}) \right),$$

with:

$$J^{(k)} = D_c^\top \text{diag}(\hat{\boldsymbol{\pi}}(\hat{\mathbf{g}}^{(k)})).$$

Then, under mild regularity conditions, the sequence $(\hat{\mathbf{g}}^{(k)})$ converges locally to a fixed point $\mathbf{g}^C$ satisfying:

$$\phi(\mathbf{g}^C) = 0, \quad \text{i.e.,} \quad D_c^\top \hat{\boldsymbol{\pi}}(\mathbf{g}^C) = 0.$$

This ensures that the empirical covariance between the predictive premiums and the sensitive attributes is zero. See Proof 5.A.2 in Appendix.

Remark 5.6 (Multiple sensitive variables). The construction naturally extends to multiple sensitive variables ($d > 1$).

Remark 5.7 (Linear independence). The corrective intervention can be interpreted as a second-order relaxation of demographic parity. It removes only linear dependence between the protected attribute and the predictions, and is therefore weaker than full demographic parity, which requires equality of the conditional distributions. See Definition 5.5.

Remark 5.8 (Joint calibration and fairness). Both portfolio-level calibration and corrective fairness can be combined in one step, without iterative sequential projections. The full derivation and the associated augmented Gram matrix representation are provided in Proposition 5.15 in Appendix.

In Algorithm 12 below, we detail the pseudo-code corresponding to the FairGGPR model with corrective intervention for count data under a log-link specification. It can be combined with best-estimate, unaware, or aware kernel specification, and balance correction.

Algorithm 12: Corrective FairGGPR for log-link count data

Data: Training data X , frequencies y , exposures ν , sensitive attributes D , kernel $k(\cdot, \cdot)$

Output: Predictive premium $\hat{\pi}_*$ satisfying $\text{Cov}(D, \hat{\pi}) \approx 0$

Compute Gram matrix: $C = K(X, X) + \sigma_n^2 I_n$

Initialize latent function: $\hat{\mathbf{g}}^{(0)} = \log(\frac{1}{n} \sum_{i=1}^n y_i)$

for $j = 1, 2, \dots$ *until convergence* **do**

 Compute gradient: $\nabla\psi(\hat{\mathbf{g}}^{(j-1)}) = \nu \odot (y - e^{\hat{\mathbf{g}}^{(j-1)}}) - C^{-1}\hat{\mathbf{g}}^{(j-1)}$

 Compute Hessian: $H^{(j-1)} = -\text{diag}(\nu \odot e^{\hat{\mathbf{g}}^{(j-1)}}) - C^{-1}$

 Newton update: $\hat{\mathbf{g}}^{(j)} = \hat{\mathbf{g}}^{(j-1)} - (H^{(j-1)})^{-1} \nabla\psi(\hat{\mathbf{g}}^{(j-1)})$

Compute Laplace posterior covariance: $\Sigma_{\hat{\mathbf{g}}} = (C^{-1} + \text{diag}(\nu \odot e^{\hat{\mathbf{g}}}))^{-1}$

if *corrective* **then**

 Compute training-point premiums: $\hat{\pi} = \exp(\hat{\mathbf{g}} + \frac{1}{2} \text{diag}(\Sigma_{\hat{\mathbf{g}}}))$

 Center sensitive attributes: $D_c = D - \mathbb{E}[D]$

 Compute fairness violations: $\mathbf{b} = -D_c^\top \hat{\pi}$

 Define constraint directions: $J = \hat{\pi} \odot D_c$

 Compute GP projection: $\Delta\mathbf{g} = \Sigma_{\hat{\mathbf{g}}} J (J^\top \Sigma_{\hat{\mathbf{g}}} J)^{-1} \mathbf{b}$

 Update latent function: $\hat{\mathbf{g}} \leftarrow \hat{\mathbf{g}} + \Delta\mathbf{g}$

 Update posterior covariance: $\Sigma_{\hat{\mathbf{g}}} = (C^{-1} + \text{diag}(\nu \odot e^{\hat{\mathbf{g}}}))^{-1}$

For a new policy x_* , compute:

$$k_* = k(x_*, X), \quad k_{**} = k(x_*, x_*)$$

Predictive mean and variance:

$$m_* = k_*^\top C^{-1} \hat{\mathbf{g}}, \quad s_*^2 = k_{**} - k_*^\top (C + H^{-1})^{-1} k_*$$

Predictive corrective premium:

$$\hat{\pi}_* = \exp\left(m_* + \frac{1}{2} s_*^2\right), \quad \mathbb{E}[N_*] = \nu_* \hat{\pi}_*$$

5.4 Results

In this Section, we evaluate our approach on a non-life insurance portfolio, available on the companion website of the book by Ohlsson et al. (2010).

First, we assess the predictive performance of the balance-corrected Generalized Gaussian Process Regression (bGGPR) model for claim frequency by benchmarking it against well-established regression approaches in the actuarial literature, namely the Poisson Generalized Linear Model (GLM) and a Poisson gradient boosting

model implemented with LightGBM. This comparison focuses on the goodness-of-fit and calibration of the resulting pure premiums. Each model is trained on a set with 20 000 observations randomly sampled from the portfolio, while predictive performance is evaluated on a disjoint test set with 15 000 policies to assess the models' ability to generalize to unseen risks.

Second, we analyze the impact of introducing and combining different fairness notions into the (b)GGPR in terms of predictive accuracy, and fairness. We benchmark our FairGGPR with a post-processing fairness approach where GGPR predictions are adjusted a posteriori using Wasserstein barycenters.

Finally, we extend the analysis to a set of simulated insurance portfolios representing alternative insurers. These portfolios share the same risk characteristics and data-generating process, but differ in their proportions of sensitive attributes. This controlled setting allows us to examine the robustness of the different fairness criteria to changes in portfolio composition and to assess whether certain fairness definitions are more advantageous than others (e.g., in terms of portfolio retention) depending on the underlying demographic structure of the protected population.

5.4.1 Regression models

We consider a non-life insurance portfolio with $n = 47\,839$ individual policies. For each policy i , we observe a vector of two continuous (policyholder's age and vehicle age) and three categorical (gender, geographic area, and vehicle power) rating characteristics $\mathbf{x}_i$, an exposure period $\nu_i \in (0, 1]$, and an observed claim count $N_i \in \mathbb{N}$. We assume that, conditionally on $\mathbf{x}_i$, the claim count follows a Poisson-type structure $N_i \mid \mathbf{x}_i \sim \text{Poisson}(\nu_i \lambda_i)$ where λ_i denotes the claim intensity per unit of exposure. To estimate $\hat{\lambda}_i$, we consider three regression models. A classical Generalized Linear Model (GLM), a gradient-boosted decision tree model (LightGBM), and the (b)GGPR adapted to Poisson data with different risk exposures.

The baseline model is a Poisson GLM with log-link function. The conditional mean is specified as:

$$\mathbb{E}[N_i \mid \mathbf{x}_i] = \nu_i \exp(\mathbf{x}_i^\top \boldsymbol{\beta}),$$

or equivalently,

$$\ln \hat{\lambda}_i = \mathbf{x}_i^\top \boldsymbol{\beta}.$$

The exposure ν_i enters the model as an offset, ensuring interpretability in terms of frequency per exposure unit. The parameter vector $\boldsymbol{\beta}$ is estimated by maximum likelihood.

While a GLM is actuarially fair by construction on the training set (i.e., the total predicted claim count equals the total observed count), it has limited flexibility. To capture nonlinearities and higher-order interactions between rating factors, we also fit a Poisson gradient boosting model using LightGBM. The model assumes $\ln \lambda_i = f(\mathbf{x}_i)$ where f is an additive ensemble of decision trees trained under a Poisson loss function, with exposure incorporated as an offset. Hyperparameters

(including tree depth, learning rate, number of leaves, and regularization terms) are tuned using an optuna study (see Table 5.B.1), with the Poisson deviance on a disjoint validation set as the optimization criterion.

As a flexible non-parametric alternative, we consider the (b)GGPR model detailed in Section 5.2. The kernel hyperparameters Θ , defined in Eq. (5.2.4), are tuned on the baseline GGPR model (Algorithm 10) by minimizing the Poisson deviance on a validation set using the Nelder et al. (1965) algorithm, and then held fixed for all FairGGPR variants. The evolution of the Poisson deviance on the validation set with respect to the kernel hyper-parameter values is reported in Fig. 5.4.1. The resulting best-estimates are $\log \ell \approx 2.69$ and $\log \sigma_f^2 \approx 4.61$.

The length-scale ℓ controls the smoothness of the latent function (i.e., the rate at which dependence between policies decreases as their covariates differ). A large ℓ indicates that policies with similar characteristics remain strongly correlated, implying a smoothly varying claim frequency surface without abrupt changes in risk. Whereas, the relatively large variance parameter σ_f^2 reflects substantial heterogeneity in the latent log-intensity, allowing the Gaussian process to capture meaningful nonlinear differences in risk across the portfolio. In actuarial terms, premiums vary smoothly but flexibly across the feature space.

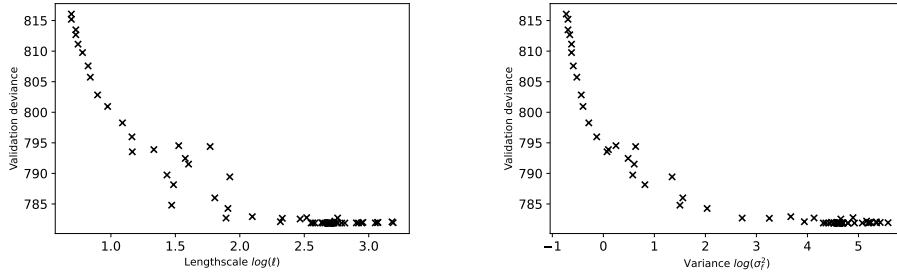
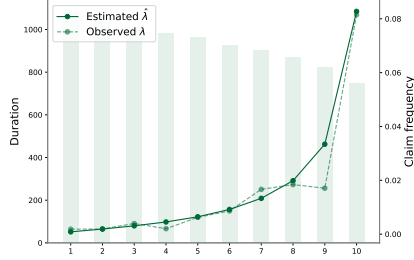
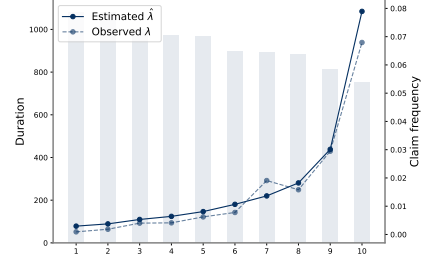


Figure 5.4.1: Evolution of the Poisson deviance on the validation set with respect to the kernel (left) length-scale $\log \ell$ and (right) variance $\log \sigma_f^2$ hyper-parameters for the baseline GGPR model applied to log-count data.

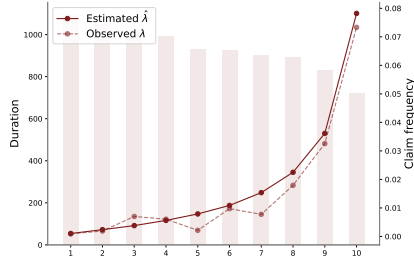
Fig. 5.4.2 assesses local calibration across the risk spectrum. Policies are partitioned into ten quantile-based groups according to their predicted claim frequency $\hat{\lambda}_i$ on the test set (see Fig. 5.B.1 in Appendix for the training set). A well-calibrated model should be characterized by close alignment between predicted (solid lines) and empirical (dashed lines) frequencies across all bins.



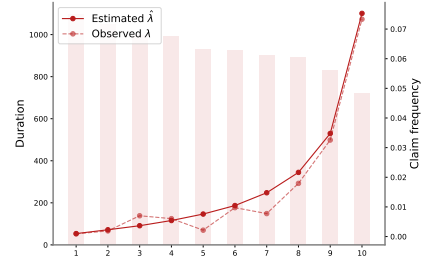
(a) GLM (out-of-sample)



(b) LightGBM (out-of-sample)



(c) GGPR (out-of-sample)



(d) bGGPR (out-of-sample)

Figure 5.4.2: Observed versus predicted claim frequencies (per unit exposure) by risk score bins on the test set (out-of-sample). For each quantile-bin, bars represent the total risk exposure (contract duration), while lines show the predicted claim frequency and the corresponding empirical frequency computed from observed claims.

Overall goodness-of-fit is assessed in Table 5.4.1 using the Poisson deviance:

$$D = 2 \sum_{i=1}^n \left[N_i \log \left(\frac{N_i}{\nu_i \hat{\lambda}_i} \right) - (N_i - \nu_i \hat{\lambda}_i) \right],$$

with the observed number of claims N_i , the observed risk exposure ν_i and the predicted claim rate $\hat{\lambda}_i$ (i.e., predicted claim frequency per unit exposure). Lower deviance indicates a better fit to the observed claim counts. Overall, the deviance values are of comparable magnitude across models, indicating that all approaches achieve a similar level of global goodness-of-fit. In-sample, LightGBM attains the lowest deviance, reflecting its greater flexibility and ability to closely adapt to the training data. The GGPR and its balance-corrected variant have a slightly higher deviance, consistent with the stronger regularisation imposed by the Gaussian process prior. Out-of-sample, bGGPR achieves the lowest deviance, followed closely by its non-balance corrected version and the GLM.

	in-sample	out-of-sample
GLM	1651.3378	1151.6595
LightGBM	1638.0187	1160.5905
GGPR	1663.3374	1152.8628
bGGPR	1663.0289	1151.4493

Table 5.4.1: Poisson deviance values on the training set (in-sample) and test set (out-of-sample). Lower deviance indicates a better fit to the observed claim counts.

We further assess calibration by comparing observed and predicted numbers of claims both in-sample in Table 5.4.2 and out-of-sample in Table 5.4.3. It highlights over or under-prediction across claim counts and overall totals. As established in Proposition 5.4, portfolio balance is not guaranteed in the GGPR.

In-sample, Table 5.4.2 shows that the Poisson GLM reproduces the total number of observed claims exactly, because of maximum likelihood estimation under a log-link which enforces portfolio-level balance. The bGGPR satisfies the same property. Despite this global constraint, both models exhibit misallocation across claim counts. All models overpredict single-claim policies ($N = 1$) and underpredict double-claim policies ($N = 2$). The deviation is stronger for GGPR and LightGBM, suggesting higher flexibility leads to over-dispersion in the tail allocation. The presence of small positive mass at 3 claims (unobserved in data) reflects model-implied tail risk, with magnitude decreasing from GLM to bGGPR.

N	Observed	GLM	LightGBM	GGPR	bGGPR
1	172.00	183.18	189.49	191.47	184.44
2	18.00	6.55	5.86	5.81	5.39
3	0.00	0.26	0.21	0.18	0.16
Total	190.00	190.00 (0.00)	195.57 (+5.57)	197.46 (+7.46)	190.00 (0.00)

Table 5.4.2: Observed vs predicted (in-sample) claim counts. Differences between predicted and observed total counts are in parentheses.

Out-of-sample, Table 5.4.3 shows that none of the models preserve total balance (including GLM, despite the fact that it is, by designed, actuarially fair in-sample). All models systematically overpredict total claims, with the magnitude seemingly increasing with model flexibility. The GLM exhibits the smallest deviation (+13.79), followed by the bGGPR (+15.15). The same structural pattern observed in-sample persists out-of-sample (i.e., overprediction of single-claim policies, underprediction of double-claims, and small but non-zero mass assigned to $N = 3$ despite no observed occurrences).

N	Observed	GLM	LightGBM	GGPR	bGGPR
0	0.00	0.00	0.00	0.00	0.00
1	118.00	138.70	143.67	146.28	140.91
2	12.00	4.89	4.41	4.41	4.10
3	0.00	0.19	0.15	0.13	0.11
Total	130.00	143.79 (+13.79)	148.24 (+18.24)	150.82 (+20.82)	145.12 (+15.12)

Table 5.4.3: Observed vs predicted (out-of-sample) claim counts. Differences between predicted and observed total counts are in parentheses.

5.4.2 FairGGPR models

In the previous section, we benchmarked (b)GGPR for log-count data against GLM and LightGBM. We now move on to the empirical evaluation of FairGGPR models, where kernel specifications (best-estimate, unaware, or aware) can be combined with objective specifications (corrective and/or balance-corrected). We analyse their impact on the distribution of predicted claim frequencies, predictive accuracy, portfolio balance, and measures of fairness. We benchmark these FairGGPR models against a post-hoc Wasserstein barycenter correction. For simplicity, we consider the binary gender as the single sensitive variable. However, FairGGPR can work with continuous or multi-categorical sensitive attributes, as well as with multiple sensitive variables simultaneously.

In Appendix, Fig. 5.B.2 (in-sample) and Fig. 5.B.3 (out-of-sample) show the actual versus predicted claim frequency by the FairGGPR models using the three kernel specifications (best-estimate, unaware, and aware), each estimated with and without balance correction. Fig. 5.B.4 (in-sample) and Fig. 5.B.5 (out-of-sample) show the plots for the corrective FairGGPR models, again considering the three kernel specifications and both with and without balance correction.

Table 5.4.4 reports the Poisson deviance for all models. Using the best-estimate GGPR as the baseline, most FairGGPR models show a moderate degradation in predictive performance out-of-sample (but sometimes improve in-sample). Despite strong in-sample performance, aware models exhibit poorer generalization out-of-sample. Whereas unaware and corrective variants appear more stable. The balance correction (bGGPR) has a very marginal impact on deviance. Note that Wasserstein method relies on a post-hoc transformation, so it does not yield a meaningful notion of out-of-sample generalization, and is therefore not evaluated on the test set.

	GGPR		bGGPR	
	in-sample	out-of-sample	in-sample	out-of-sample
best-estimate	1663.3374	1152.8628	1663.0289	1151.4493
unaware	1669.5150	1166.5473	1669.2486	1165.2124
aware	1650.3395	1176.5964	1644.1029	1165.7397
corrective	1672.2900	1167.7510	1672.2847	1167.6140
unaware corrective	1683.6704	1164.5599	1683.6548	1164.3945
aware corrective	1655.9780	1176.4261	1648.4006	1180.8334
post-hoc Wasserstein	1671.6923	-	-	-

Table 5.4.4: Poisson deviance values on the training (in-sample) and test (out-of-sample) sets across all FairGGPR models.

Tables 5.4.5 and 5.4.6 report the empirical covariance between predicted frequency rates and gender. The observed empirical covariance (49.06 in-sample and 48.11 out-of-sample), serves as the baseline, quantifying the linear dependence of claim frequency on gender in the raw data. All predictive models reduce this dependence to varying degrees.

Best-estimate models in Table 5.4.5 reduce covariance relative to observed but still retain significant gender dependence. Specifically, in-sample, GGPR achieves 24.82, slightly lower than GLM, and its balance-corrected variant (bGGPR) further reduces covariance to 23.91. Out-of-sample, these reductions persist, though covariance values drop slightly more for LightGBM (12.14) and bGGPR (17.79). The lower empirical covariance observed for LightGBM may be interpreted as a consequence of its non-linear, piecewise-constant prediction structure and the aggregation of multiple weak learners, which may mitigate the linear dependence between predictions and the protected attributes. In contrast, GLMs impose an additive linear structure that can directly induce linear dependence on the protected attribute by construction, while Gaussian process regression induces smoother functional dependence through the kernel. However, this ranking is not necessarily intrinsic to the models and may depend on the dataset and on the fact that covariance captures only linear association.

	in-sample	out-of-sample
Observed	49.0608	48.1125
GLM	26.4114	19.4085
LightGBM	16.3529	12.1426
GGPR	24.8253	18.4702
bGGPR	23.9070	17.7878

Table 5.4.5: Empirical covariance $dC_n\hat{\lambda}$ between predicted frequency rates and gender for baseline and best-estimate models.

FairGGPR models in Table 5.4.6 show a drastic reduction in gender dependence. The unaware and aware models reduce in-sample covariance to 4.97 and 5.38, respectively. Their balance-corrected (bGGPR) variants provide marginal additional improvement. Out-of-sample, unaware and aware covariances drop further to 2.46 and 2.80 (with balance-corrected variants remaining similar), showing that these FairGGPR models generalize well to unseen data in reducing gender bias. As expected, corrective models all achieve zero in-sample covariance (up to numerical precision), therefore removing gender dependence. Balance correction does not alter these values in-sample because both corrective and balance adjustments are enforced simultaneously in the latent function space via the augmented Gram matrix. Out-of-sample, covariance slightly deviates from zero.

	GGPR		bGGPR	
	in-sample	out-of-sample	in-sample	out-of-sample
best-estimate	24.8253	18.4702	23.9070	17.7878
unaware	4.9706	2.4643	4.7954	2.3769
aware	5.3799	2.8004	4.7129	2.8036
corrective	0.0000	-1.3438	0.0000	-1.3385
unaware corrective	0.0000	-0.7581	0.0000	-0.7556
aware corrective	0.0000	-0.8950	0.0000	-0.5026
post-hoc Wasserstein	4.4572	-	-	-

Table 5.4.6: Empirical covariance $dC_n\hat{\lambda}$ between predicted frequency rates and gender for baseline (b)GGPR and FairGGPR models.

Tables 5.4.7 and 5.4.8 report in-sample calibration (see Tables 5.B.2 and 5.B.3 for the corresponding out-of-sample analysis). In Table 5.4.7, all non-balance corrected models (BE, U, A) exhibit systematic overestimation of the total number of claims. The deviation is moderate for BE and U (+7.46 and +6.93, respectively), but substantially larger for the aware model (+15.02). This overestimation is primarily driven by the $N = 1$ class, where predicted counts exceed observed

values across all models, while the $N = 2$ class is consistently underestimated. This imbalance suggests a misallocation of mass toward more frequent claim counts (already observed in GLM and LightGBM). As expected, the balance-corrected (BC) models fully remove the aggregate bias, as totals exactly match observed counts. This correction works as a global rescaling and does not fully eliminate class-level discrepancies. The overestimation at $N = 1$ and underestimation at $N = 2$ persist (at reduced magnitude). Hence, the BC models improve marginal calibration but do not fundamentally alter the distributional shape.

N	Obs	BE	BE, BC	U	U, BC	A	A, BC
1	172	191.47	184.44	191.12	184.58	198.61	184.46
2	18	5.81	5.39	5.64	5.27	6.21	5.38
3	0	0.18	0.16	0.17	0.15	0.19	0.16
Total	190	197.46	190	196.93	190	205.02	190
Δ		+7.46	0.00	+6.93	0.00	+15.02	0.00

Table 5.4.7: In-sample calibration by number of claims of the best-estimate (BE), unaware (U), and aware (A) models, with and without balance correction (BC). Δ shows deviations from the total observed counts.

Table 5.4.8 shows that corrective models (C) substantially improve calibration already prior to BC. The total deviations become negligible (+0.81 for C and +0.95 for U,C), indicating that the corrective fairness intervention actually mitigates miscalibration (e.g., with the aware kernel specification). As before, BC enforces exact total calibration for all corrective variants (without affecting independence).

N	Obs	C	C, BC	U, C	U, C, BC	A, C	A, C, BC
1	172	185.53	184.77	186.3978	185.4894	194.0641	185.2648
2	18	5.13	5.09	4.46	4.42	5.01	4.62
3	0	0.14	0.14	0.09	0.09	0.12	0.11
Total	190	190.81	190	190.95	190	199.19	190
Δ		+0.81	0.00	+0.95	0.00	+9.19	0.00

Table 5.4.8: In-sample calibration by number of claims of the corrective (C) best-estimate (BE), unaware (U), and aware (A) models, with and without balance correction (BC). Δ shows deviations from the total observed counts.

Tables 5.B.2 and 5.B.3 in Appendix show a degradation in calibration when moving out-of-sample (as also observed with the GLM, despite the fact that it is, by designed, actuarially fair in-sample).

We further analyse miscalibration across gender to assess whether certain models disproportionately affect one protected group relative to the other.

Table 5.4.9 reports in-sample calibration (absolute) deviations from the observed counts across gender for best-estimate models. Note that absolute deviations are systematically larger for males than for females, both in overestimation ($N = 1$) and underestimation ($N = 2$). This is expected given the substantially larger size of the male subgroup.

At the aggregate level, the GLM achieves perfect calibration for both genders. LightGBM and GGPR both exhibit imbalances, but driven by different deviations. LightGBM overestimates females (+6.58), less represented in the portfolio, while slightly underestimating males (-1.01). GGPR, on the other hand, overestimates both groups, with a larger deviation for males (+5.22) than for females (+2.25). Aggregate miscalibration are then largely explained by errors in the female subgroup for the LightGBM, and male subgroup for the GGPR model.

The bGGPR improves aggregate calibration, reducing deviations to +1.40 for females and -1.40 for males. This symmetric structure reflects the global rescaling induced by BC. Note that the balance correction operates as a global rescaling that enforces aggregate calibration at the portfolio level. It does not explicitly depend on the sensitive attribute (gender), but instead adjusts predictions proportionally across all observations. As a result, the impact of BC is driven by the dominant (male) subgroup, both because it represents the majority of the portfolio and because it exhibits the largest absolute deviations. Consequently, the global correction induced by BC primarily reduces errors for males in our example. This is clearly visible when comparing the deviations with the basic GGPR. For instance, at $N = 1$, the deviation for males decreases substantially (from +15.58 to +9.38), while the reduction for females is more limited (from +3.89 to +3.07). A similar pattern holds at the aggregate level, where the total deviation for males shifts from +5.22 to -1.40, whereas the adjustment for females is smaller (+2.25 to +1.40). Hence, although BC is group-agnostic, it effectively corrects the dominant source of miscalibration, which in this case corresponds to the male subgroup.

N	Observed	GLM		LightGBM		GGPR		bGGPR	
		Female	Male	Female	Male	Female	Male	Female	Male
1	18 / 154	+1.62	+9.56	+8.05	+9.44	+3.89	+15.58	+3.07	+9.38
2	2 / 16	-1.63	-9.82	-1.49	-10.65	-1.65	-10.54	-1.67	-10.93
3	0	+0.008	+0.255	+0.011	+0.195	+0.006	+0.173	+0.005	+0.155
Total	20 / 170	0.00	0.00	+6.58	-1.01	+2.25	+5.22	+1.40	-1.40

Table 5.4.9: In-sample deviations from observed counts by number of claims and gender. Female values are shown in blue, male values in red.

Out-of-sample results are presented in Table 5.4.10. The comparison between the GLM and bGGPR is particularly meaningful because, by design, the GLM and

bGGPR are both calibrated in-sample. However, this total balance property does not generalize out-of-sample, but the bGGPR is comparable to the GLM level of miscalibration out-of-sample.

N	Observed	GLM		LightGBM		GGPR		bGGPR	
		Female	Male	Female	Male	Female	Male	Female	Male
1	6 / 112	+9.85	+10.85	+14.50	+11.18	+11.67	+16.61	+11.00	+11.91
2	0	+0.33	+0.18	+0.43	+0.14	+0.31	+0.12	+0.29	+0.11
3	0	+0.008	+0.185	+0.010	+0.140	+0.006	+0.121	+0.005	+0.108
Total	6 / 124	+10.18	+3.61	+14.93	+3.30	+11.99	+8.84	+11.30	+3.82

Table 5.4.10: Out-of-sample deviations from observed counts by number of claims and gender. Female values are shown in blue, male values in red.

Tables 5.4.11, 5.4.12, and 5.4.13 report the in-sample deviations by gender across the different FairGGPR models. In the basic GGPR, males exhibit the largest overestimation, driving most of the aggregate miscalibration in-sample (see Table 5.4.9). In Table 5.4.11, we observe that unaware and aware models (without BC) shift deviations toward females (especially at $N = 1$), so the largest absolute errors are no longer with males (initially with higher claim frequencies on average in the discriminatory models). The aware variants generally produce slightly higher overestimation for females compared to the unaware variants, reflecting the marginalization over the gender in the kernel structure. As expected, the BC variants rebalance the miscalibration between the two protected groups.

N	Observed	U		U, BC		A		A, BC	
		Female	Male	Female	Male	Female	Male	Female	Male
1	18 / 154	+14.64	+4.48	+13.52	-0.94	+15.78	+10.83	+13.45	-0.99
2	2 / 16	-1.20	-11.16	-1.25	-11.48	-1.13	-10.66	-1.23	-11.39
3	0	+0.019	+0.147	+0.017	+0.133	+0.023	+0.173	+0.020	+0.137
Total	20 / 170	+13.46	-6.52	+12.28	-12.28	+14.68	+0.35	+12.24	-12.24

Table 5.4.11: In-sample deviations by number of claims and gender for GGPR unaware (U), unaware BC (U, BC), aware (A), and aware BC (A, BC) variants. Female values are shown in blue, male values in red.

Table 5.4.12 shows that corrective models miscalibration is quite balanced across gender at the aggregate level, except for the aware corrective model that overestimates female (because of the kernel marginalization), but BC in Table 5.4.13 perfectly rebalance these deviations.

N	Observed	C		U, C		A, C	
		Female	Male	Female	Male	Female	Male
1	18 / 154	+16.26	-2.73	+16.24	-1.84	+17.66	+4.41
2	2 / 16	-1.15	-11.72	-1.25	-12.29	-1.15	-11.83
3	0	+0.020	+0.118	+0.014	+0.080	+0.018	+0.100
Total	20 / 170	+15.14	-14.33	+14.82	-14.05	+16.52	-7.32

Table 5.4.12: In-sample deviations by number of claims and gender for non-BC corrective variants: best-estimate corrective (C), unaware corrective (U, C), and aware corrective (A, C). Female values in blue, male values in red.

N	Observed	C, BC		U, C, BC		A, C, BC	
		Female	Male	Female	Male	Female	Male
1	18 / 154	+16.12	-3.35	+16.07	-2.58	+15.99	-2.72
2	2 / 16	-1.15	-11.76	-1.26	-12.33	-1.22	-12.16
3	0	+0.020	+0.117	+0.014	+0.079	+0.017	+0.092
Total	20 / 170	+14.99	-14.99	+14.82	-14.82	+14.78	-14.78

Table 5.4.13: In-sample deviations by number of claims and gender for balance-corrected (BC) corrective variants: corrective BC (C, BC), unaware corrective BC (U, C, BC), and aware corrective BC (A, C, BC). Female in blue, male in red.

To further illustrate the impact of fairness adjustments, we report in Fig. 5.B.6 and Fig. 5.B.7 in Appendix scatter plots of predicted premiums against the GGPR baseline. These figures highlight the deviations induced by the different fairness interventions, including visual comparison with the Wasserstein benchmark.

Finally, we provide in Appendix 5.C additional analyses illustrating the economic effects of fairness interventions. We simulate multiple insurers with different proportion of protected policyholders to evaluate the impact of fairness interventions on market share, adverse selection, and the relative profitability of insurers under competition. Overall, these simulations reveal that fairness strategies, particularly aware interventions, can become economically advantageous as the protected group represents a larger share of the portfolio, whereas corrective or unaware strategies are often costly or unstable.

Note that the three corrective kernel formulations correspond to different fairness interpretations and lead to different trade-offs between risk prediction and fairness. The best-estimate kernel preserves the most actuarial information. The unaware kernel removes the protected attribute from the similarity structure, which prevents direct discrimination on the protected attribute at the kernel level, that is, no direct dependence on D but proxy pathways through A still exist. The aware

kernel instead marginalises over the protected attribute within the kernel construction, aiming to mitigate proxy effects while preserving more risk structure. Overall, the kernel specification governs the statistical learning of risk.

In all cases, when combined with the corrective intervention enforced at the posterior level, protected information is reintroduced to define the projection that enforces zero linear dependence between predictions and D . The final predictor is therefore explicitly shaped using protected group membership. As a result, the final predictive outcome reflects a combination of the kernel-induced fairness structure and the posterior fairness constraint that aims to satisfy demographic parity, and these two levels should be interpreted jointly rather than independently as they interact through the posterior inference process and jointly determine the final premium estimates.

Which kernel is most appropriate therefore depends on the regulatory and market context, as well as the desired balance between actuarial accuracy and fairness requirements. The framework is intended to compare these options in a consistent Gaussian process setting, rather than to prescribe a single optimal choice.

5.5 Discussion

This work proposes a framework for a multi-criteria Fair Generalized Gaussian Process Regression (FairGGPR) model adapted to count data with heterogeneous risk exposure, as commonly encountered in insurance pricing and frequency modeling. By embedding several fairness notions directly into the probabilistic learning architecture, the proposed approach moves beyond single-criterion corrections and post-hoc adjustments.

A key contribution lies in the adaptation of fairness constraints to the generalized GP setting, where the response variable follows a Poisson distribution and the risk exposure enters through the canonical log-link. Fairness is enforced on the latent intensity function.

Within this modeling choice, multiple fairness criteria are formulated, either as kernel architecture modifications, or as constraints on the posterior predictive distribution. This allows multiple fairness criteria to be combined within a single model, yielding a multi-criteria FairGGPR that trades-off predictive accuracy, smoothness, and several fairness objectives simultaneously. These constraints are enforced during training, rather than applied post-hoc, and can be generalized out-of-sample.

The empirical results show several key findings. First, fairness interventions can be introduced without overly affecting predictive performance. Poisson deviance remains comparable across best-estimate and FairGGPR models, both in-sample and out-of-sample. Second, unaware and aware kernels reduce the empirical covariance between predicted premiums and gender relative to best-estimate models, while corrective models achieve zero covariance by construction. These effects persist

out-of-sample, indicating that fairness corrections generalize beyond the training data.

From an actuarial perspective, the proposed framework offers several advantages. First, fairness corrections are interpretable as controlled deviations from the best-estimate posterior. Second, each fairness criterion corresponds to a clearly defined mathematical object (e.g., covariance with a sensitive attribute, or distributional alignment across groups). Moreover, the ability to combine multiple fairness criteria and sensitive attributes within a single probabilistic model reflects the multi-dimensional nature of discrimination in real-world pricing systems.

Overall, this chapter demonstrates that fairness can be embedded directly into the probabilistic learning mechanism. By jointly accounting for count data, exposure, and multiple fairness interventions within a single model, this framework provides a flexible foundation for fair insurance pricing.

Future research may extend this framework to severity modeling, and further dynamic portfolio settings. In particular, we have explored dynamic portfolio settings in which fairness is enforced under different portfolio composition. Preliminary results suggest that different fairness interventions can have diverse impacts on adverse selection and market share, depending on the proportion of protected policyholders.

Another interesting direction for future work concerns model simplification and interpretability. The complexity of GP models may limit their practical adoption. It would be valuable to investigate whether the corrected predictions obtained from the proposed framework can be approximated by simpler models, such as generalized linear or additive models. Such approximations could provide interpretable surrogates that retain the fairness and calibration properties of the original model while being easier to implement and communicate in practice.

5.A Proofs

5.A.1 GGPR

Proof of Proposition 5.1. Recall that a density in the ED family admits the representation

$$p(y \mid \mathbf{X}) = \exp\left(\frac{y\theta(\mathbf{X}) - \gamma(\theta(\mathbf{X}))}{\phi/\nu} + c(y, \phi, \nu)\right),$$

where θ is the canonical parameter, γ the cumulant function, ϕ the dispersion parameter, and c a normalizing term independent of θ .

For the Poisson model, the dispersion parameter is $\phi = 1$ and the canonical link is logarithmic. Define:

$$\theta(\mathbf{X}) := \ln \lambda(\mathbf{X}).$$

The cumulant function corresponding to the Poisson exponential family is:

$$\gamma(\theta) = \exp(\theta), \quad \gamma'(\theta) = \exp(\theta), \quad \gamma'^{-1}(\mu) = \ln \mu.$$

Substituting $\theta(\mathbf{X}) = \ln \lambda(\mathbf{X})$ and $\phi = 1$ into the ED form yields:

$$\frac{y\theta(\mathbf{X}) - \gamma(\theta(\mathbf{X}))}{\phi/\nu} = \nu y \ln \lambda(\mathbf{X}) - \nu \lambda(\mathbf{X}).$$

Choosing the normalizing term:

$$c(y, \phi, \nu) = -\ln((\nu y)!) + \nu y \ln \nu,$$

we obtain:

$$\begin{aligned} p(y \mid \mathbf{X}) &= \exp(\nu y \ln \lambda(\mathbf{X}) - \nu \lambda(\mathbf{X}) - \ln((\nu y)!) + \nu y \ln \nu) \\ &= \exp(-\nu \lambda(\mathbf{X})) \frac{(\nu \lambda(\mathbf{X}))^{\nu y}}{(\nu y)!}. \end{aligned}$$

□

Proof of Proposition 5.2. Let $\mathbf{X} \in \mathbb{R}^{n \times m}$ be the training set and let Φ be a (design) kernel matrix:

$$\Phi = [\phi(\mathbf{x}_1), \dots, \phi(\mathbf{x}_n)]^\top \in \mathbb{R}^{n \times m}.$$

The Bayesian linear regression model can be expressed as:

$$\mathbf{g} = \Phi \mathbf{w} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_n^2 \mathbf{I}_n),$$

with a Gaussian prior on the weights:

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \Sigma_m),$$

Conditioned on $\mathbf{w}$, the likelihood is Gaussian,

$$\mathbf{g} \mid \mathbf{X}, \mathbf{w} \sim \mathcal{N}(\Phi \mathbf{w}, \sigma_n^2 \mathbf{I}_n).$$

The marginal mean of $\mathbf{g} \mid \mathbf{X}$ is obtained by iterated expectation:

$$\begin{aligned} \mathbb{E}[\mathbf{g} \mid \mathbf{X}] &= \mathbb{E}_{\mathbf{w}}[\mathbb{E}[\mathbf{g} \mid \mathbf{X}, \mathbf{w}]] \\ &= \mathbb{E}_{\mathbf{w}}[\Phi \mathbf{w}] = \Phi \mathbb{E}[\mathbf{w}] = \mathbf{0}. \end{aligned}$$

The marginal covariance follows from the law of total covariance:

$$\begin{aligned} \text{Cov}[\mathbf{g} \mid \mathbf{X}] &= \text{Cov}_{\mathbf{w}}[\mathbb{E}[\mathbf{g} \mid \mathbf{X}, \mathbf{w}]] + \mathbb{E}_{\mathbf{w}}[\text{Cov}[\mathbf{g} \mid \mathbf{X}, \mathbf{w}]] \\ &= \text{Cov}_{\mathbf{w}}[\Phi \mathbf{w}] + \mathbb{E}_{\mathbf{w}}[\sigma_n^2 \mathbf{I}_n] \\ &= \Phi \text{Cov}[\mathbf{w}] \Phi^\top + \sigma_n^2 \mathbf{I}_n \\ &= \Phi \Sigma_m \Phi^\top + \sigma_n^2 \mathbf{I}_n \\ &= K + \sigma_n^2 \mathbf{I}_n \\ &= C(\mathbf{X}, \mathbf{X}). \end{aligned}$$

Since the marginal mean and covariance fully characterize a Gaussian distribution, it follows that:

$$\mathbf{g} \mid \mathbf{X} \sim \mathcal{N}(\mathbf{0}, K + \sigma_n^2 \mathbf{I}_n), \quad K := \Phi \Sigma_m \Phi^\top,$$

or equivalently,

$$\mathbf{g} \mid \mathbf{X} \sim \mathcal{N}(\mathbf{0}, C(\mathbf{X}, \mathbf{X})).$$

□

Proof of Eq. (5.2.1). Assume the latent vector satisfies $\mathbf{g} \mid \mathbf{X} \sim \mathcal{N}(\mathbf{0}, C(\mathbf{X}, \mathbf{X}))$, where $C(\mathbf{X}, \mathbf{X}) = K + \sigma_n^2 \mathbf{I}_n$. Its density is:

$$p(\mathbf{g} \mid \mathbf{X}) = \frac{1}{(2\pi)^{n/2} |C(\mathbf{X}, \mathbf{X})|^{1/2}} \exp\left(-\frac{1}{2} \mathbf{g}^\top C(\mathbf{X}, \mathbf{X})^{-1} \mathbf{g}\right).$$

Taking logarithms and dropping terms independent of $\mathbf{g}$ yields:

$$\ln p(\mathbf{g} \mid \mathbf{X}) = -\frac{1}{2} \mathbf{g}^\top C(\mathbf{X}, \mathbf{X})^{-1} \mathbf{g} + \text{const.}$$

By Bayes' rule, the posterior mode $\hat{\mathbf{g}}$ satisfies:

$$\begin{aligned} \hat{\mathbf{g}} &= \arg \max_{\mathbf{g}} \{\ln p(\mathbf{y} \mid \mathbf{g}) + \ln p(\mathbf{g} \mid \mathbf{X})\} \\ &= \arg \max_{\mathbf{g}} \left\{ \ln p(\mathbf{y} \mid \mathbf{g}) - \frac{1}{2} \mathbf{g}^\top C(\mathbf{X}, \mathbf{X})^{-1} \mathbf{g} \right\}. \end{aligned}$$

For a Poisson likelihood with exposure vector $\boldsymbol{\nu}$, the log-likelihood is:

$$\ln p(\mathbf{y} \mid \mathbf{g}) = \boldsymbol{\nu}^\top (\mathbf{y} \odot \mathbf{g} - \exp(\mathbf{g})) + \text{const.}$$

Defining: $\psi(\mathbf{g}) = \boldsymbol{\nu}^\top (\mathbf{y} \odot \mathbf{g} - \exp(\mathbf{g})) - \frac{1}{2} \mathbf{g}^\top C(\mathbf{X}, \mathbf{X})^{-1} \mathbf{g}$, we obtain:

$$\hat{\mathbf{g}} = \arg \max_{\mathbf{g}} \psi(\mathbf{g}).$$

□

Proof of Eq. (5.2.2). Under the Laplace approximation, the inverse covariance matrix is given by the negative Hessian of the log-posterior evaluated at the mode $\hat{\mathbf{g}}$:

$$\begin{aligned} \Sigma_{\mathbf{g}}^{-1} &= -\nabla_{\mathbf{g}}^2 \ln p(\mathbf{g} \mid \mathbf{y}, \mathbf{X}) \big|_{\mathbf{g}=\hat{\mathbf{g}}} \\ &= -\nabla^2 \psi(\hat{\mathbf{g}}). \end{aligned}$$

For a Poisson likelihood with exposure vector $\boldsymbol{\nu}$, the objective function ψ satisfies:

$$\psi(\hat{\mathbf{g}}) = \boldsymbol{\nu}^\top \exp(\hat{\mathbf{g}}) - \boldsymbol{\nu}^\top (\mathbf{y} \odot \hat{\mathbf{g}}) + \frac{1}{2} \hat{\mathbf{g}}^\top C(\mathbf{X}, \mathbf{X})^{-1} \hat{\mathbf{g}}.$$

Taking the second derivative yields:

$$-\nabla^2 \psi(\hat{\mathbf{g}}) = \text{diag}(\boldsymbol{\nu} \odot \exp(\hat{\mathbf{g}})) + C(\mathbf{X}, \mathbf{X})^{-1}.$$

Therefore, the $n \times n$ covariance matrix is:

$$\Sigma_{\mathbf{g}} = \left(C(\mathbf{X}, \mathbf{X})^{-1} + \text{diag}(\boldsymbol{\nu} \odot \exp(\hat{\mathbf{g}})) \right)^{-1},$$

where $H(\hat{\mathbf{g}}) = \text{diag}(\boldsymbol{\nu} \odot \exp(\hat{\mathbf{g}}))$ denotes the Hessian of the negative log-likelihood. □

Proof of Proposition 5.3. Let $\mathbf{g} = (g(\mathbf{x}_1), \dots, g(\mathbf{x}_n))^\top$ denote the latent function values at the training inputs $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$. Under the Gaussian process prior, the joint distribution of $(\mathbf{g}, g_*)$ is:

$$\begin{pmatrix} \mathbf{g} \\ g_* \end{pmatrix} \sim \mathcal{N} \left(\mathbf{0}, \begin{pmatrix} C(\mathbf{X}, \mathbf{X}) & k_*^\top \\ k_* & k_{**} \end{pmatrix} \right),$$

where $k_* = k(\mathbf{x}_*, \mathbf{X})$ and $k_{**} = k(\mathbf{x}_*, \mathbf{x}_*)$.

For non-Gaussian likelihoods, the posterior distribution of $\mathbf{g}$ is analytically intractable. Under the Laplace approximation, the posterior is approximated by a Gaussian:

$$q(\mathbf{g} \mid \mathbf{y}, \mathbf{X}) \approx \mathcal{N}(\hat{\mathbf{g}}, \Sigma_{\mathbf{g}}),$$

where $\hat{\mathbf{g}}$ is the posterior mode and:

$$\Sigma_{\mathbf{g}} = \left(C(\mathbf{X}, \mathbf{X})^{-1} + H(\hat{\mathbf{g}}) \right)^{-1},$$

with $H(\hat{\mathbf{g}})$ denoting the Hessian matrix of the negative log-likelihood evaluated at $\hat{\mathbf{g}}$. The predictive distribution of g_* is obtained by marginalising over the approximate posterior $q(\mathbf{g} \mid \mathbf{y}, \mathbf{X})$:

$$p(g_* \mid \mathbf{y}, \mathbf{X}) = \int p(g_* \mid \mathbf{g}) q(\mathbf{g} \mid \mathbf{y}, \mathbf{X}) d\mathbf{g}.$$

From the joint Gaussian prior, the conditional distribution of g_* given $\mathbf{g}$ is (see Proposition 5.13 in Appendix):

$$p(g_* \mid \mathbf{g}) = \mathcal{N}\left(k_* C(\mathbf{X}, \mathbf{X})^{-1} \mathbf{g}, k_{**} - k_* C(\mathbf{X}, \mathbf{X})^{-1} k_*^\top\right).$$

The predictive mean is therefore given by:

$$\begin{aligned} m_* &= \mathbb{E}_{q(\mathbf{g} \mid \mathbf{y}, \mathbf{X})} [\mathbb{E}(g_* \mid \mathbf{g})] \\ &= \mathbb{E}_{q(\mathbf{g} \mid \mathbf{y}, \mathbf{X})} \left[k_* C(\mathbf{X}, \mathbf{X})^{-1} \mathbf{g} \right] \\ &= k_* C(\mathbf{X}, \mathbf{X})^{-1} \mathbb{E}_{q(\mathbf{g} \mid \mathbf{y}, \mathbf{X})} [\mathbf{g}] \\ &= k_* C(\mathbf{X}, \mathbf{X})^{-1} \hat{\mathbf{g}}. \end{aligned}$$

Using the law of total variance:

$$s_*^2 = \mathbb{E}_{q(\mathbf{g} \mid \mathbf{y}, \mathbf{X})} [\text{Var}(g_* \mid \mathbf{g})] + \text{Var}_{q(\mathbf{g} \mid \mathbf{y}, \mathbf{X})} [\mathbb{E}(g_* \mid \mathbf{g})].$$

The first term is constant with respect to $\mathbf{g}$ and equals:

$$\mathbb{E}_{q(\mathbf{g} \mid \mathbf{y}, \mathbf{X})} [\text{Var}(g_* \mid \mathbf{g})] = k_{**} - k_* C(\mathbf{X}, \mathbf{X})^{-1} k_*^\top.$$

The second term is:

$$\begin{aligned} \text{Var}_{q(\mathbf{g} \mid \mathbf{y}, \mathbf{X})} [\mathbb{E}(g_* \mid \mathbf{g})] &= \text{Var}_{q(\mathbf{g} \mid \mathbf{y}, \mathbf{X})} \left[k_* C(\mathbf{X}, \mathbf{X})^{-1} \mathbf{g} \right] \\ &= k_* C(\mathbf{X}, \mathbf{X})^{-1} \Sigma_{\mathbf{g}} C(\mathbf{X}, \mathbf{X})^{-1} k_*^\top. \end{aligned}$$

Combining both terms yields:

$$s_*^2 = k_{**} - k_* C(\mathbf{X}, \mathbf{X})^{-1} k_*^\top + k_* C(\mathbf{X}, \mathbf{X})^{-1} \Sigma_{\mathbf{g}} C(\mathbf{X}, \mathbf{X})^{-1} k_*^\top.$$

Substituting $\Sigma_{\mathbf{g}} = (C^{-1} + H)^{-1}$ and applying the Woodbury identity gives:

$$C^{-1} - C^{-1} \Sigma_{\mathbf{g}} C^{-1} = (C + H^{-1})^{-1}.$$

Therefore:

$$s_*^2 = k_{**} - k_* \left(C(\mathbf{X}, \mathbf{X}) + H(\hat{\mathbf{g}})^{-1} \right)^{-1} k_*^\top.$$

The predictive distribution of the latent function value at $\mathbf{X}_*$ is therefore:

$$g_* \mid \mathbf{x}_*, \mathbf{X}, \mathbf{y} \approx \mathcal{N}(m_*, s_*^2),$$

with m_* and s_*^2 as stated in Proposition 5.3.

□

Proposition 5.13 (Predictive distribution in kernel form). *Let $k(\mathbf{x}_*, \mathbf{X}) = \phi(\mathbf{x}_*)^\top \Sigma_c \Phi^\top$ denote the cross-kernel between a new input $\mathbf{x}_*$ and the training data, and let $K = \Phi \Sigma_c \Phi^\top$. Then the predictive distribution for a new output g_* is:*

$$g_* \mid \mathbf{x}_*, \mathbf{X}, \mathbf{y}, \mathbf{g} \sim \mathcal{N}(\mu_*, \sigma_*^2),$$

with:

$$\begin{aligned} \mu_* &= k(\mathbf{x}_*, \mathbf{X})(\sigma_n^2 \mathbf{I}_n + K)^{-1} \mathbf{g}, \\ \sigma_*^2 &= k(\mathbf{x}_*, \mathbf{x}_*) - k(\mathbf{x}_*, \mathbf{X})(\sigma_n^2 \mathbf{I}_n + K)^{-1} k(\mathbf{X}, \mathbf{x}_*). \end{aligned}$$

Proof of Proposition 5.13. Starting from the posterior predictive mean in feature space (see Proposition 5.14 in Appendix):

$$\phi(\mathbf{x}_*)^\top \bar{\mathbf{w}} = \sigma_n^{-2} \phi(\mathbf{x}_*)^\top A^{-1} \Phi^\top \mathbf{g}.$$

Applying the Woodbury identity:

$$A^{-1} = (\sigma_n^{-2} \Phi^\top \Phi + \Sigma_m^{-1})^{-1} = \Sigma_m - \Sigma_m \Phi^\top (\sigma_n^2 \mathbf{I}_n + \Phi \Sigma_m \Phi^\top)^{-1} \Phi \Sigma_m,$$

yields:

$$\begin{aligned} \phi(\mathbf{x}_*)^\top \bar{\mathbf{w}} &= \sigma_n^{-2} \phi(\mathbf{x}_*)^\top A^{-1} \Phi^\top \mathbf{g} \\ &= \sigma_n^{-2} \phi(\mathbf{x}_*)^\top \left(\Sigma_m - \Sigma_m \Phi^\top (\sigma_n^2 \mathbf{I}_n + \Phi \Sigma_m \Phi^\top)^{-1} \Phi \Sigma_m \right) \Phi^\top \mathbf{g} \\ &= \sigma_n^{-2} \phi(\mathbf{x}_*)^\top \Sigma_m \Phi^\top \mathbf{g} - \sigma_n^{-2} \phi(\mathbf{x}_*)^\top \Sigma_m \Phi^\top (\sigma_n^2 \mathbf{I}_n + \Phi \Sigma_m \Phi^\top)^{-1} \Phi \Sigma_m \Phi^\top \mathbf{g}. \end{aligned}$$

We introduce the $1 \times n$ cross-kernel k and $n \times n$ kernel K matrices:

$$k(\mathbf{x}_*, \mathbf{X}) = \phi(\mathbf{x}_*)^\top \Sigma_m \Phi^\top, \quad K = \Phi \Sigma_m \Phi^\top.$$

The predictive mean becomes:

$$\phi(\mathbf{x}_*)^\top \bar{\mathbf{w}} = \sigma_n^{-2} k(\mathbf{x}_*, \mathbf{X}) \mathbf{g} - \sigma_n^{-2} k(\mathbf{x}_*, \mathbf{X}) (\sigma_n^2 \mathbf{I}_n + K)^{-1} K \mathbf{g}.$$

We factor out the cross-kernel term $k(\mathbf{x}_*, \mathbf{X})$:

$$\phi(\mathbf{x}_*)^\top \bar{\mathbf{w}} = k(\mathbf{x}_*, \mathbf{X}) \sigma_n^{-2} \left[\mathbf{I}_n - (\sigma_n^2 \mathbf{I}_n + K)^{-1} K \right] \mathbf{g}.$$

Using the identity $(\sigma_n^2 \mathbf{I}_n + K)^{-1} (\sigma_n^2 \mathbf{I}_n + K) = \mathbf{I}_n$, we simplify the brackets:

$$\begin{aligned} \mathbf{I}_n - (\sigma_n^2 \mathbf{I}_n + K)^{-1} K &= (\sigma_n^2 \mathbf{I}_n + K)^{-1} (\sigma_n^2 \mathbf{I}_n + K - K) \\ &= (\sigma_n^2 \mathbf{I}_n + K)^{-1} \sigma_n^2 \mathbf{I}_n \\ &= \sigma_n^2 (\sigma_n^2 \mathbf{I}_n + K)^{-1}. \end{aligned}$$

We get the final predictive mean in kernel form:

$$\phi(\mathbf{x}_*)^\top \bar{\mathbf{w}} = k(\mathbf{x}_*, \mathbf{X})(\sigma_n^2 \mathbf{I}_n + K)^{-1} \mathbf{g}.$$

This is the standard Bayesian linear regression predictive mean expressed in terms of kernels.

Similarly, the predictive covariance $\phi(\mathbf{x}_*)^\top A^{-1} \phi(\mathbf{x}_*)$ is equal to:

$$\begin{aligned} \phi(\mathbf{x}_*)^\top A^{-1} \phi(\mathbf{x}_*) &= \phi(\mathbf{x}_*)^\top \left(\Sigma_m - \Sigma_m \Phi^\top (\sigma_n^2 \mathbf{I}_n + \Phi \Sigma_m \Phi^\top)^{-1} \Phi \Sigma_m \right) \phi(\mathbf{x}_*) \\ &= \phi(\mathbf{x}_*)^\top \Sigma_m \phi(\mathbf{x}_*) - \phi(\mathbf{x}_*)^\top \Sigma_m \Phi^\top (\sigma_n^2 \mathbf{I}_n + \Phi \Sigma_m \Phi^\top)^{-1} \Phi \Sigma_m \phi(\mathbf{x}_*) \\ &= k(\mathbf{x}_*, \mathbf{x}_*) - k(\mathbf{x}_*, \mathbf{X})(\sigma_n^2 \mathbf{I}_n + K)^{-1} k(\mathbf{X}, \mathbf{x}_*), \end{aligned}$$

where $k(\mathbf{X}, \mathbf{x}_*) = k(\mathbf{x}_*, \mathbf{X})^\top$. The final predictive covariance in kernel form is equal to:

$$\phi(\mathbf{x}_*)^\top A^{-1} \phi(\mathbf{x}_*) = k(\mathbf{x}_*, \mathbf{x}_*) - k(\mathbf{x}_*, \mathbf{X})(\sigma_n^2 \mathbf{I}_n + K)^{-1} k(\mathbf{X}, \mathbf{x}_*).$$

This is the Bayesian linear regression predictive variance in kernel form. The predictive distribution of scalar g_* for a new policyholder $\mathbf{x}_*$ in Gaussian Process is formally defined as:

$$\begin{aligned} p(g_* \mid \mathbf{x}_*, \mathbf{X}, \mathbf{y}, \mathbf{g}) &= \mathcal{N}(k(\mathbf{x}_*, \mathbf{X})(\sigma_n^2 \mathbf{I}_n + K)^{-1} \mathbf{g}, \\ &\quad k(\mathbf{x}_*, \mathbf{x}_*) - k(\mathbf{x}_*, \mathbf{X})(\sigma_n^2 \mathbf{I}_n + K)^{-1} k(\mathbf{X}, \mathbf{x}_*)). \end{aligned}$$

□

Proposition 5.14 (Posterior distribution in feature space). *Let $g(x) = \Phi(x)^\top \mathbf{w}$ be a finite-dimensional feature representation of the latent function, with design matrix $\Phi \in \mathbb{R}^{n \times m}$ evaluated at the inputs $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$. Assume the Gaussian prior:*

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \Sigma_m),$$

and the Gaussian likelihood:

$$\mathbf{y} \mid \Phi, \mathbf{w} \sim \mathcal{N}(\Phi \mathbf{w}, \sigma_n^2 \mathbf{I}_n).$$

Then the posterior distribution of $\mathbf{w}$ is Gaussian,

$$\mathbf{w} \mid \Phi, \mathbf{y} \sim \mathcal{N}(\bar{\mathbf{w}}, A^{-1}),$$

with

$$A = \sigma_n^{-2} \Phi^\top \Phi + \Sigma_m^{-1}, \quad \bar{\mathbf{w}} = A^{-1} \mathbf{b} = A^{-1} \sigma_n^{-2} \Phi^\top \mathbf{y}.$$

Proof of Proposition 5.14. We condition on the feature matrix $\Phi = \Phi(\mathbf{X})$ rather than directly on the inputs $\mathbf{X}$. Recall that the density of an m -dimensional multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma_m)$ is given by:

$$p(\mathbf{w}) = (2\pi)^{-m/2} |\Sigma_m|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{w} - \boldsymbol{\mu})^\top \Sigma_m^{-1}(\mathbf{w} - \boldsymbol{\mu})\right).$$

We assume a zero-mean Gaussian prior on the weights $\mathbf{w} \in \mathbb{R}^m$:

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \Sigma_m),$$

with density:

$$\begin{aligned} p(\mathbf{w}) &= (2\pi)^{-m/2} |\Sigma_m|^{-1/2} \exp\left(-\frac{1}{2}\mathbf{w}^\top \Sigma_m^{-1}\mathbf{w}\right) \\ &\propto \exp\left(-\frac{1}{2}\mathbf{w}^\top \Sigma_m^{-1}\mathbf{w}\right). \end{aligned}$$

The likelihood is similarly assumed Gaussian:

$$\mathbf{y} \mid \Phi, \mathbf{w} \sim \mathcal{N}(\Phi\mathbf{w}, \sigma_n^2 \mathbf{I}_n),$$

with density:

$$\begin{aligned} p(\mathbf{y} \mid \Phi, \mathbf{w}) &= (2\pi)^{-n/2} |\sigma_n^2 \mathbf{I}_n|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{y} - \Phi\mathbf{w})^\top (\sigma_n^2 \mathbf{I}_n)^{-1}(\mathbf{y} - \Phi\mathbf{w})\right) \\ &\propto \exp\left(-\frac{1}{2\sigma_n^2}(\mathbf{y} - \Phi\mathbf{w})^\top (\mathbf{y} - \Phi\mathbf{w})\right). \end{aligned}$$

Ignoring terms not depending on $\mathbf{w}$, the posterior density is proportional to:

$$\begin{aligned} p(\mathbf{w} \mid \Phi, \mathbf{y}) &\propto p(\mathbf{y} \mid \Phi, \mathbf{w}) p(\mathbf{w}) \\ &\propto \exp\left(\left(-\frac{1}{2\sigma_n^2}(\mathbf{y} - \Phi\mathbf{w})^\top (\mathbf{y} - \Phi\mathbf{w}) - \frac{1}{2}\mathbf{w}^\top \Sigma_m^{-1}\mathbf{w}\right)\right). \end{aligned}$$

Expanding the quadratic in the exponential, we get:

$$\begin{aligned} p(\mathbf{w} \mid \Phi, \mathbf{y}) &\propto \exp\left(-\frac{1}{2\sigma_n^2}(\mathbf{y}^\top \mathbf{y} - 2\mathbf{w}^\top \Phi^\top \mathbf{y} + \mathbf{w}^\top \Phi^\top \Phi \mathbf{w}) - \frac{1}{2}\mathbf{w}^\top \Sigma_m^{-1}\mathbf{w}\right) \\ &= \exp\left(-\frac{1}{2\sigma_n^2}\mathbf{y}^\top \mathbf{y} + \sigma_n^{-2}\mathbf{w}^\top \Phi^\top \mathbf{y} - \frac{1}{2}\mathbf{w}^\top (\sigma_n^{-2}\Phi^\top \Phi + \Sigma_m^{-1})\mathbf{w}\right) \\ &= \exp\left(\text{const}_y + b^\top \mathbf{w} - \frac{1}{2}\mathbf{w}^\top A\mathbf{w}\right), \end{aligned}$$

where

$$A := \sigma_n^{-2}\Phi^\top \Phi + \Sigma_m^{-1} \in \mathbb{R}^{m \times m}, \quad b := \sigma_n^{-2}\Phi^\top \mathbf{y} \in \mathbb{R}^m.$$

We drop const_y , since it collects terms not depending on w . The posterior is now proportional to:

$$p(\mathbf{w} \mid \Phi, \mathbf{y}) \propto \exp\left(-\frac{1}{2}\mathbf{w}^\top A \mathbf{w} + \mathbf{b}^\top \mathbf{w}\right).$$

To identify the posterior distribution, we complete the square. For symmetric positive-definite A ,

$$-\frac{1}{2}\mathbf{w}^\top A \mathbf{w} + \mathbf{b}^\top \mathbf{w} = -\frac{1}{2}(\mathbf{w} - A^{-1}\mathbf{b})^\top A(\mathbf{w} - A^{-1}\mathbf{b}) + \frac{1}{2}\mathbf{b}^\top A^{-1}\mathbf{b}.$$

Defining:

$$\bar{\mathbf{w}} := A^{-1}\mathbf{b} = (\sigma_n^{-2}\Phi^\top \Phi + \Sigma_m^{-1})^{-1}\sigma_n^{-2}\Phi^\top \mathbf{y},$$

we obtain:

$$\begin{aligned} p(\mathbf{w} \mid \Phi, \mathbf{y}) &\propto \exp\left(-\frac{1}{2}\mathbf{w}^\top A \mathbf{w} + \mathbf{b}^\top \mathbf{w}\right) \\ &\propto \exp\left(-\frac{1}{2}(\mathbf{w} - \bar{\mathbf{w}})^\top A(\mathbf{w} - \bar{\mathbf{w}})\right). \end{aligned}$$

Hence the posterior $\mathbf{w} \mid \mathbf{y}$ is Gaussian with covariance $A^{-1} \in \mathbb{R}^{m \times m}$ and mean $\bar{\mathbf{w}} \in \mathbb{R}^m$:

$$\mathbf{w} \mid \Phi, \mathbf{y} \sim \mathcal{N}(\bar{\mathbf{w}}, A^{-1}).$$

□

5.A.2 FairGGPR

Proof of Proposition 5.4. The posterior mode $\hat{\mathbf{g}}$ maximizes the log-posterior:

$$\psi(\mathbf{g}) = \boldsymbol{\nu}^\top (\mathbf{y} \odot \mathbf{g} - \exp(\mathbf{g})) - \frac{1}{2}\mathbf{g}^\top C^{-1}\mathbf{g}.$$

The first-order condition (gradient of ψ) at the mode is:

$$\nabla \psi(\hat{\mathbf{g}}) = \boldsymbol{\nu} \odot (\mathbf{y} - \exp(\hat{\mathbf{g}})) - C^{-1}\hat{\mathbf{g}} = 0.$$

Summing over all policies and left-multiplying by $\mathbf{1}_n^\top$:

$$\underbrace{\sum_{i=1}^n \nu_i (y_i - \exp(\hat{g}_i))}_{\text{term 1}} - \underbrace{\mathbf{1}_n^\top C^{-1} \hat{\mathbf{g}}}_{\text{term 2}} = 0. \quad (5.A.1)$$

In Eq. (5.A.1), the first term can be interpreted as the force applied by the portfolio experience (claims), which may deform the latent risk surface so that predictions match observed claims.

Whereas, the second term, can be interpreted as the force applied by the prior to instore smoothness in the latent risk surface since the matrix C encodes the covariance structure between policies, so that policies with similar characteristics have similar latent risks. The precision matrix C^{-1} therefore penalizes deviations from this smooth structure.

A useful geometric analogy, illustrated in Fig. 5.A.1, is to view the GP prior as a network of elastic springs linking similar policies. If the latent risk surface fitted by the claim experience assigns very different risks to policies with similar characteristics, the springs exert forces pulling the surface back toward a smoother configuration.

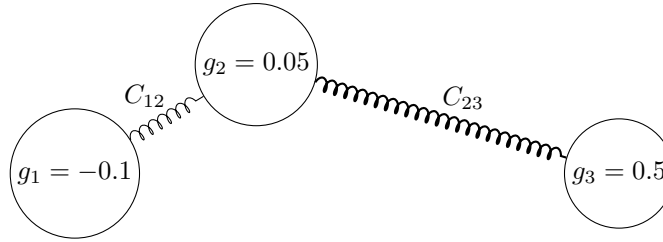


Figure 5.A.1: Illustration of Eq. (5.A.1). Each node represents a policy's latent log-frequency g_i . Data forces nodes toward observed claims. Prior forces similar policies toward each other according to the kernel C , enforcing smoothness. Springs indicate similarity (thicker = more similar characteristic, stronger pull).

Exact calibration at the posterior mode follows if $C^{-1}\hat{\mathbf{g}}$ sums to zero:

$$\sum_{i=1}^n \nu_i \exp(\hat{g}_i) = \sum_{i=1}^n N_i \quad \text{s.t.} \quad \mathbf{1}_n^\top C^{-1} \hat{\mathbf{g}} = 0. \quad (5.A.2)$$

Even if Eq. (5.A.2) holds, the predictive frequency at each point nonetheless includes the posterior variance term:

$$\hat{Y}_i = \exp(m_i + \frac{1}{2}s_i^2) \geq \exp(\hat{g}_i),$$

where $m_i = \hat{g}_i$ and $s_i^2 = \text{diag}(C^{-1} + \text{diag}(\nu \odot \exp(\hat{\mathbf{g}})))^{-1}$ at training points. So summing over the portfolio:

$$\sum_{i=1}^n \nu_i \hat{Y}_i \geq \sum_{i=1}^n \nu_i \exp(\hat{g}_i) \approx \sum_{i=1}^n N_i.$$

This shows that the GGPR predictions break portfolio-level calibration. □

Proof of Proposition 5.5. Let $\hat{\mathbf{g}}$ denote the Laplace posterior mode and define the predictive premiums at the training points:

$$\hat{\pi}(\hat{\mathbf{g}}) = \exp\left(\hat{\mathbf{g}} + \frac{1}{2} \text{diag}(\Sigma_{\hat{\mathbf{g}}})\right),$$

where:

$$\begin{aligned} \hat{\mathbf{g}} &= \arg \max_{\mathbf{g}} \{\psi(\mathbf{g})\} \quad \text{with } \psi(\mathbf{g}) = \sum_{i=1}^n (N_i g_i - \nu_i e^{g_i}) - \frac{1}{2} \mathbf{g}^\top C^{-1} \mathbf{g}, \\ \Sigma_{\hat{\mathbf{g}}} &= (C^{-1} + \text{diag}(\nu \odot e^{\hat{\mathbf{g}}}))^{-1}. \end{aligned}$$

We define the portfolio-level calibration function:

$$\phi(\hat{\mathbf{g}}) = \sum_{i=1}^n \nu_i \hat{\pi}_i(\hat{\mathbf{g}}) - \sum_{i=1}^n N_i,$$

where $\hat{\pi}_i(\hat{\mathbf{g}}) = \exp(\hat{g}_i + \frac{1}{2}(\Sigma_{\hat{\mathbf{g}}})_{ii})$. This is nonlinear in $\hat{\mathbf{g}}$ because of the exponential link and the covariance dependence. Therefore, we need to solve $\phi(\hat{\mathbf{g}}) = 0$ iteratively. We thus do a first-order Taylor expansion of ϕ around $\hat{\mathbf{g}}^{(k)}$:

$$\phi(\hat{\mathbf{g}}^{(k)} + \Delta \mathbf{g}) \approx \phi(\hat{\mathbf{g}}^{(k)}) + \nabla \phi(\hat{\mathbf{g}}^{(k)})^\top \Delta \mathbf{g}.$$

Setting $\phi = 0$ gives the linearized constraint:

$$\nabla \phi(\hat{\mathbf{g}}^{(k)})^\top \Delta \mathbf{g} = -\phi(\hat{\mathbf{g}}^{(k)}).$$

By the chain rule:

$$\frac{\partial \phi}{\partial g_j} = \sum_{i=1}^n \nu_i \frac{\partial \hat{\pi}_i}{\partial g_j},$$

with:

$$\frac{\partial \hat{\pi}_i}{\partial g_j} = \hat{\pi}_i \left(\delta_{ij} + \frac{1}{2} \frac{\partial (\Sigma_{\mathbf{g}})_{ii}}{\partial g_j} \right),$$

where δ_{ij} denotes the Kronecker delta. Differentiating the covariance matrix gives:

$$\frac{\partial (\Sigma_{\mathbf{g}})_{ii}}{\partial g_j} = -(\Sigma_{\mathbf{g}})_{ij}^2 \nu_j e^{g_j}.$$

This expression is exact but dense and costly to evaluate for large n . Using the chain rule and neglecting higher-order dependence of $\Sigma_{\mathbf{g}}$ on $\mathbf{g}$, we approximate:

$$\nabla \phi(\hat{\mathbf{g}}^{(k)}) \approx \mathbf{u}^{(k)} = \nu \odot \hat{\pi}(\hat{\mathbf{g}}^{(k)}).$$

We thus seek the minimal correction satisfying:

$$\mathbf{u}^{(k)\top} \Delta \mathbf{g} = -\phi(\hat{\mathbf{g}}^{(k)}).$$

If all elements of the vector $\mathbf{g}$ were independent and had unit variance, we could just use the usual Euclidean norm for the size of a change $\Delta\mathbf{g}$:

$$\|\Delta\mathbf{g}\|^2 = \sum_i (\Delta g_i)^2$$

However, in a GP, elements of $\mathbf{g}$ are correlated, so changing one might affect others. Moreover, some elements have larger uncertainty than others (i.e., larger diagonal entries in Σ), so a change Δg_i in a component with larger variance is somehow less significant than the same change in a small variance element. Intuitively, it is easier to shift the predictions of points with high posterior variance (e.g., sparse data regions) than points with low posterior variance (e.g., well-supported regions). To account for both correlation and variance, corrections $\Delta\mathbf{g}$ are instead measured using the posterior uncertainty Σ :

$$\|\Delta\mathbf{g}\|_{\Sigma^{-1}}^2 = \Delta\mathbf{g}^\top \Sigma^{-1} \Delta\mathbf{g}.$$

Hence, $\Delta\mathbf{g}^{(k)}$ is the solution to:

$$\min_{\Delta\mathbf{g}} \quad \Delta\mathbf{g}^\top \Sigma^{-1} \Delta\mathbf{g} \quad \text{s.t.} \quad \mathbf{u}^{(k)\top} \Delta\mathbf{g} = -\phi(\hat{\mathbf{g}}^{(k)}). \quad (5.A.3)$$

This is the smallest possible linear correction along the gradient direction to satisfy the portfolio-level calibration in the Laplace approximation. It is such that the balance correction preserves the original structure in regions where the model is most confident.

The solution to Eq. (5.A.3) is the classical projection (solved using Lagrange multiplier):

$$\Delta\mathbf{g}^{(k)} = \Sigma \mathbf{u}^{(k)} \left(\mathbf{u}^{(k)\top} \Sigma \mathbf{u}^{(k)} \right)^{-1} \left(-\phi(\hat{\mathbf{g}}^{(k)}) \right).$$

□

Proof of Proposition 5.12. Let $\hat{\mathbf{g}}$ denote the Laplace posterior mode and define the predictive premiums at the training points:

$$\hat{\boldsymbol{\pi}}(\hat{\mathbf{g}}) = \exp\left(\hat{\mathbf{g}} + \frac{1}{2} \text{diag}(\Sigma_{\hat{\mathbf{g}}})\right),$$

We define the corrective function with respect to the sensitive variables D :

$$\phi(\hat{\mathbf{g}}) = D_c^\top \hat{\boldsymbol{\pi}}(\hat{\mathbf{g}}),$$

where $D_c = C_n D$ with centered matrix $C_n = I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top$. This is nonlinear in $\hat{\mathbf{g}}$ because of the exponential link and the covariance dependence. Therefore, we need to solve $\phi(\hat{\mathbf{g}}) = 0$ iteratively. We thus do a first-order Taylor expansion of ϕ around $\hat{\mathbf{g}}^{(k)}$:

$$\phi(\hat{\mathbf{g}}^{(k)} + \Delta\mathbf{g}) \approx \phi(\hat{\mathbf{g}}^{(k)}) + \nabla \phi(\hat{\mathbf{g}}^{(k)})^\top \Delta\mathbf{g}.$$

Setting $\phi = 0$ (to impose zero empirical covariance) gives the linearized constraint:

$$\nabla\phi(\hat{\mathbf{g}}^{(k)})^\top \Delta\mathbf{g} = -\phi(\hat{\mathbf{g}}^{(k)}).$$

By the chain rule:

$$\frac{\partial\phi}{\partial g_j} = D_c^\top \frac{\partial\hat{\pi}}{\partial g_j}.$$

Neglecting higher-order dependence of the posterior covariance on $\mathbf{g}$, the Jacobian simplifies to a diagonal sensitivity structure, yielding:

$$\nabla\phi(\hat{\mathbf{g}}^{(k)}) \approx J^{(k)} = D_c^\top \text{diag}(\hat{\pi}(\hat{\mathbf{g}}^{(k)})).$$

We thus seek the minimal correction satisfying:

$$J^{(k)} \Delta\mathbf{g} = -\phi(\hat{\mathbf{g}}^{(k)}).$$

As in the GP setting, corrections are measured using the posterior uncertainty:

$$\|\Delta\mathbf{g}\|_{\Sigma^{-1}}^2 = \Delta\mathbf{g}^\top \Sigma^{-1} \Delta\mathbf{g}.$$

Hence, $\Delta\mathbf{g}^{(k)}$ is the solution to:

$$\min_{\Delta\mathbf{g}} \quad \Delta\mathbf{g}^\top \Sigma^{-1} \Delta\mathbf{g} \quad \text{s.t.} \quad J^{(k)} \Delta\mathbf{g} = -\phi(\hat{\mathbf{g}}^{(k)}).$$

The solution is the generalized projection:

$$\Delta\mathbf{g}^{(k)} = \Sigma J^{(k)\top} \left(J^{(k)} \Sigma J^{(k)\top} \right)^{-1} \left(-\phi(\hat{\mathbf{g}}^{(k)}) \right).$$

□

Proposition 5.15 (Joint calibration and fairness as constrained GP projection). *Portfolio balance and independence from sensitive attributes have linearized moment constraints of the form:*

$$\phi(\mathbf{g}) = 0,$$

where $\phi(\mathbf{g}) \in \mathbb{R}^{1+d}$ stacks the calibration and fairness functionals:

$$\phi(\mathbf{g}) = \begin{bmatrix} \phi_{\text{cal}}(\mathbf{g}) \\ \phi_{\text{fair}}(\mathbf{g}) \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n \nu_i \hat{\pi}_i(\mathbf{g}) - \sum_{i=1}^n N_i \\ D_c^\top \hat{\pi}(\mathbf{g}) \end{bmatrix}.$$

Linearizing around $\hat{\mathbf{g}}^{(k)}$ yields the joint constraint:

$$\tilde{J}^{(k)} \Delta\mathbf{g} = -\phi(\hat{\mathbf{g}}^{(k)}),$$

where the augmented Jacobian is given by:

$$\tilde{J}^{(k)} = \begin{bmatrix} \mathbf{u}^{(k)\top} \\ J^{(k)} \end{bmatrix} \in \mathbb{R}^{(1+d) \times n},$$

with

$$\mathbf{u}^{(k)} = \boldsymbol{\nu} \odot \hat{\boldsymbol{\pi}}(\hat{\mathbf{g}}^{(k)}), \quad \mathbf{J}^{(k)} = D_c^\top \text{diag}(\hat{\boldsymbol{\pi}}(\hat{\mathbf{g}}^{(k)})).$$

The minimal Mahalanobis correction is then obtained by solving:

$$\min_{\Delta \mathbf{g}} \quad \Delta \mathbf{g}^\top \Sigma_{\hat{\mathbf{g}}}^{-1} \Delta \mathbf{g} \quad \text{s.t.} \quad \tilde{\mathbf{J}}^{(k)} \Delta \mathbf{g} = -\phi(\hat{\mathbf{g}}^{(k)}),$$

which yields the projection:

$$\Delta \mathbf{g}^{(k)} = \Sigma_{\hat{\mathbf{g}}} \tilde{\mathbf{J}}^{(k)\top} (\tilde{\mathbf{J}}^{(k)} \Sigma_{\hat{\mathbf{g}}} \tilde{\mathbf{J}}^{(k)\top})^{-1} (-\phi(\hat{\mathbf{g}}^{(k)})).$$

Equivalently, this update corresponds to conditioning the Gaussian posterior:

$$\mathbf{g} \sim \mathcal{N}(\hat{\mathbf{g}}, \Sigma_{\hat{\mathbf{g}}})$$

on the joint linear pseudo-observation:

$$\tilde{\mathbf{J}}^{(k)}(\mathbf{g} - \hat{\mathbf{g}}^{(k)}) = -\phi(\hat{\mathbf{g}}^{(k)}).$$

This admits a GP interpretation through the augmented Gram matrix:

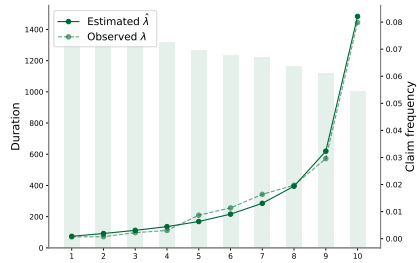
$$\tilde{C}^{(k)} = \begin{bmatrix} \Sigma_{\hat{\mathbf{g}}} & \Sigma_{\hat{\mathbf{g}}} \tilde{\mathbf{J}}^{(k)\top} \\ \tilde{\mathbf{J}}^{(k)} \Sigma_{\hat{\mathbf{g}}} & \tilde{\mathbf{J}}^{(k)} \Sigma_{\hat{\mathbf{g}}} \tilde{\mathbf{J}}^{(k)\top} \end{bmatrix},$$

which represents the joint covariance of $(\mathbf{g}, \phi(\mathbf{g}))$. Conditioning on $\phi(\mathbf{g}) = 0$ yields the same posterior mean.

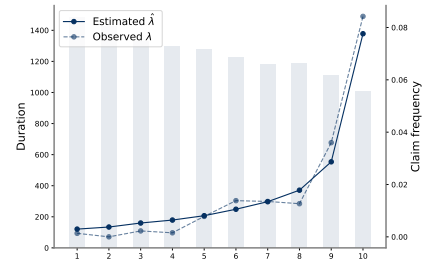
5.B Results

Parameter	Value	Description
Best value	0.05169	Objective metric (e.g., validation loss) achieved in the best trial
Std. Dev.	0.00607	Standard deviation of the objective metric across cross-validation folds
learning_rate	0.00315	Step size applied at each boosting iteration
feature_fraction	0.8925	Fraction of features randomly selected for building each tree
bagging_fraction	0.9662	Fraction of data randomly selected for each iteration (subsampling)
bagging_freq	6	Number of iterations between each bagging operation
lambda_l2	0.00155	L2 regularization term applied to leaf weights
min_sum_hessian_in_leaf	48.31	Minimum sum of second-order gradients (Hessians) in a leaf to allow a split
max_depth	8	Maximum depth allowed for each tree
min_gain_to_split	0.296	Minimum gain required for a split to be performed
min_data_in_leaf	43	Minimum number of data points required in a leaf
num_leaves	233	Maximum number of leaves in a single tree

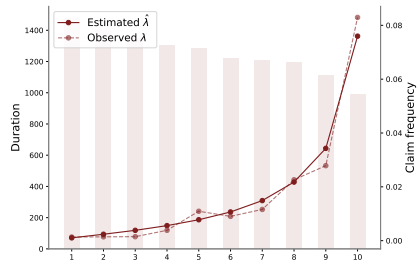
Table 5.B.1: Best LightGBM hyperparameter configuration and performance from the optimization study.



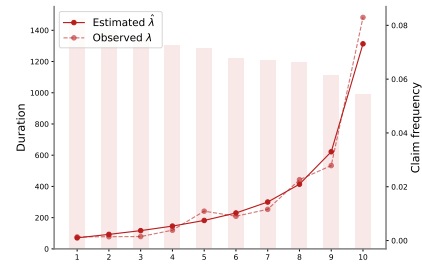
(a) GLM (in-sample)



(b) LightGBM (in-sample)

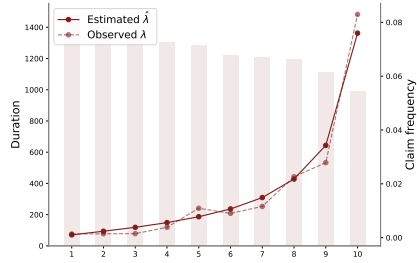


(c) GGPR (in-sample)

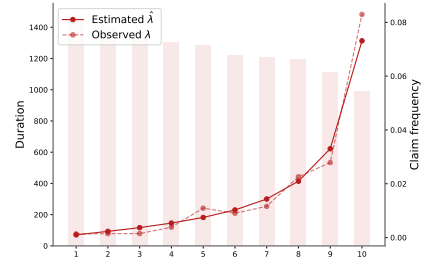


(d) bGGPR (in-sample)

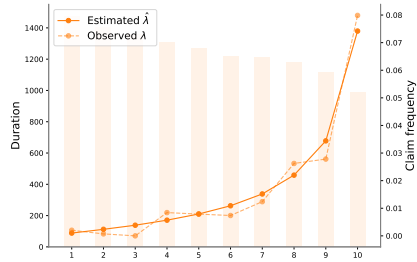
Figure 5.B.1: Observed versus predicted claim frequencies (per unit exposure) by risk score bins on the training set (in-sample). For each quantile-bin, bars represent the total risk exposure (contract duration), while lines show the predicted claim frequency and the corresponding empirical frequency computed from observed claims.



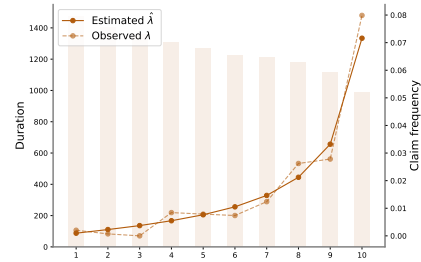
(a) GGPR (in-sample)



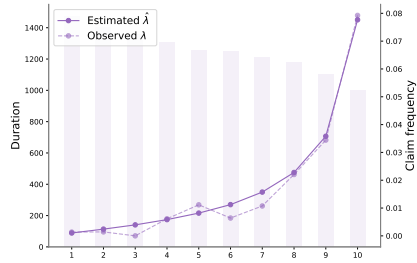
(b) bGGPR (in-sample)



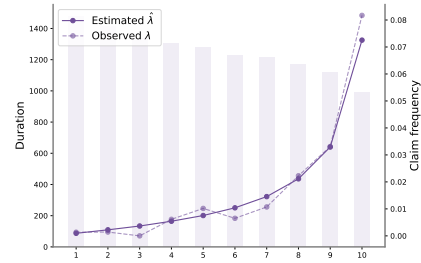
(c) Unaware GGPR (in-sample)



(d) Unaware bGGPR (in-sample)

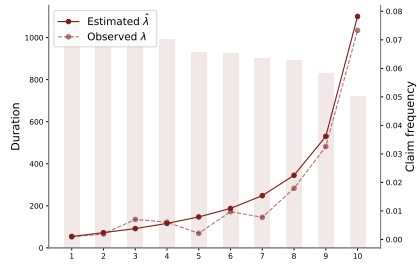


(e) Aware GGPR (in-sample)

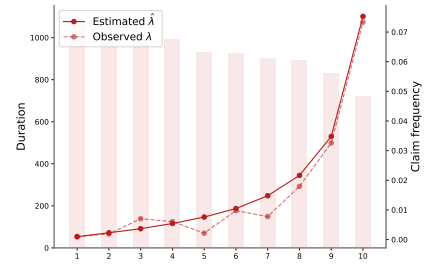


(f) Aware bGGPR (in-sample)

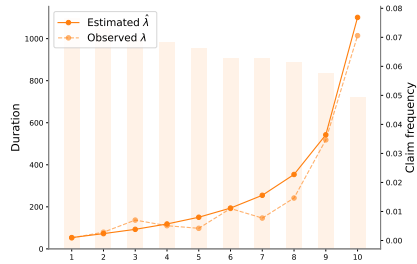
Figure 5.B.2: Actual vs predicted claim frequency for best-estimate, unaware, and aware FairGGPR models with and without balance correction.



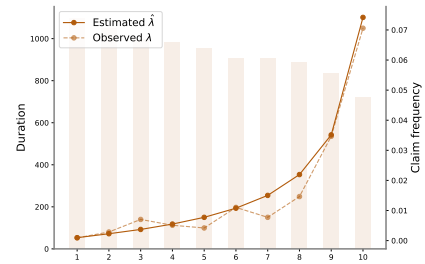
(a) GGPR (out-of-sample)



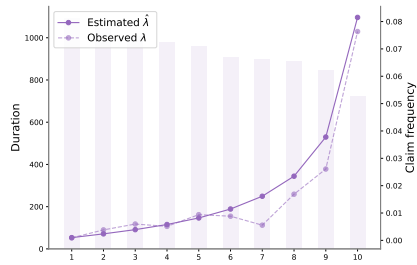
(b) bGGPR (out-of-sample)



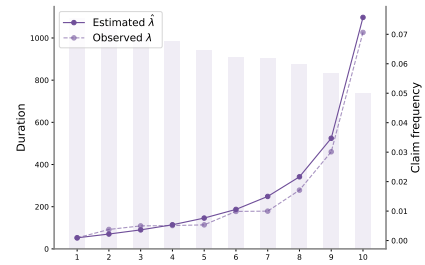
(c) Unaware GGPR (out-of-sample)



(d) Unaware bGGPR (out-of-sample)

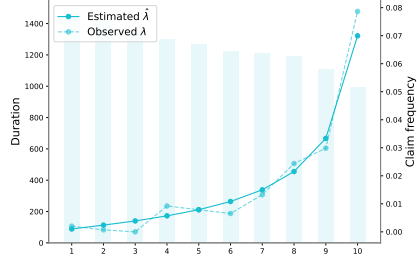


(e) Aware GGPR (out-of-sample)

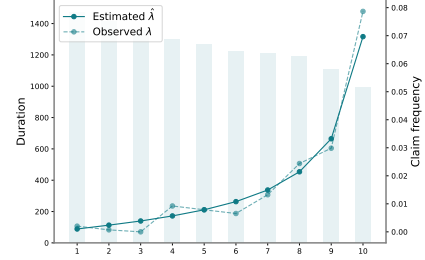


(f) Aware bGGPR (out-of-sample)

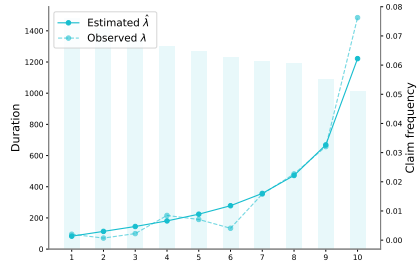
Figure 5.B.3: Actual vs predicted claim frequency (out-of-sample) for best-estimate, unaware, and aware FairGGPR models with and without balance correction.



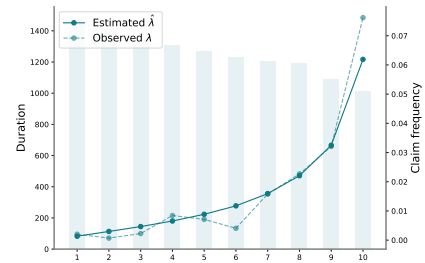
(a) Corrective GGPR with best-estimate kernel (in-sample)



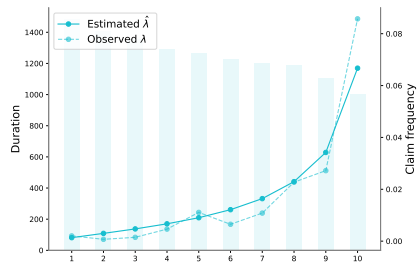
(b) Corrective bGGPR with best-estimate kernel (in-sample)



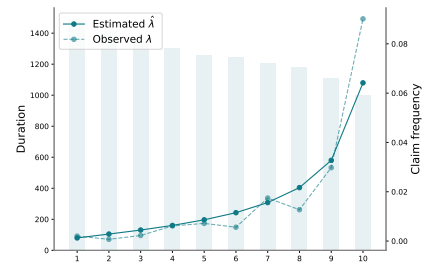
(c) Corrective GGPR with unaware kernel (in-sample)



(d) Corrective bGGPR with unaware kernel (in-sample)

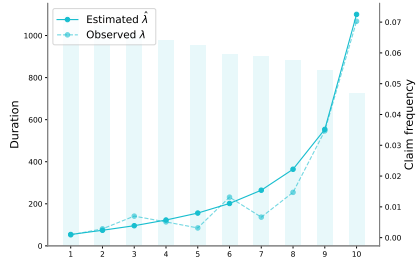


(e) Corrective GGPR with aware kernel (in-sample)

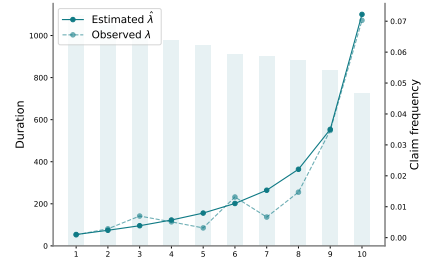


(f) Corrective bGGPR with aware kernel (in-sample)

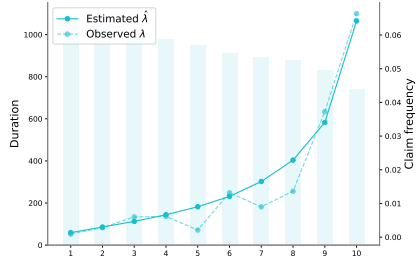
Figure 5.B.4: Actual vs predicted claim frequency (in-sample) for corrective FairGGPR models, with best-estimate, unaware, and aware kernel, with and without balance correction.



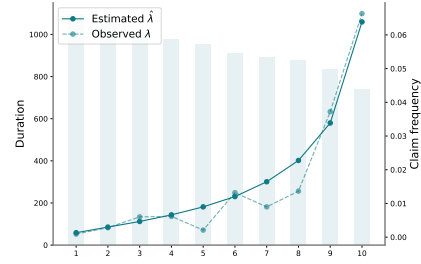
(a) Corrective GGPR with best-estimate kernel (out-of-sample)



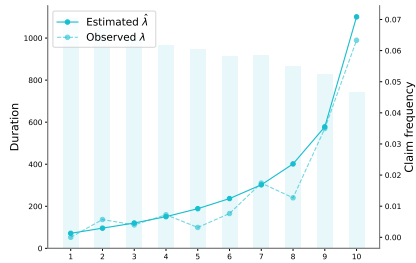
(b) Corrective bGGPR with best-estimate kernel (out-of-sample)



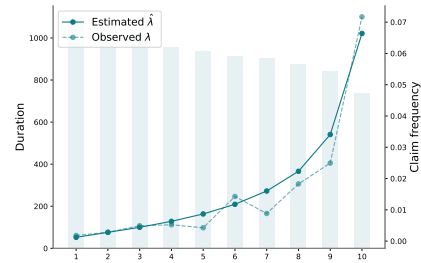
(c) Corrective GGPR with unaware kernel (out-of-sample)



(d) Corrective bGGPR with unaware kernel (out-of-sample)



(e) Corrective GGPR with aware kernel (out-of-sample)



(f) Corrective bGGPR with aware kernel (out-of-sample)

Figure 5.B.5: Actual vs predicted claim frequency (out-of-sample) for corrective FairGGPR models, with best-estimate, unaware, and aware kernel, with and without balance correction.

N	Obs	BE	BE, BC	U	U, BC	A	A, BC
1	118	146.28	140.91	146.58	141.57	152.34	141.66
2	12	4.41	4.09	4.35	4.06	4.86	4.28
3	0	0.13	0.11	0.12	0.11	0.15	0.13
Total	130	150.82	145.1236	151.0582	145.74	157.3533	146.08
Δ		+20.82	+15.12	+21.06	+15.74	+27.35	+16.08

Table 5.B.2: Out-of-sample calibration by number of claims of the best-estimate, unaware, and aware GGPR models (with and without balance correction). Δ shows deviations from the total observed counts.

N	Observed	C	C, BC	U, C	U, C, BC	A, C	A, C, BC
1	118	142.31	141.72	142.42	141.73	149.42	141.83
2	12	3.97	3.94	3.4352	3.40	4.05	3.60
3	0	0.10	0.10	0.0711	0.07	0.10	0.08
Total	130	146.38	145.76	145.93	145.20	153.58	145.52
Δ		+16.38	+15.76	+15.93	+15.20	+23.58	+15.52

Table 5.B.3: Out-of-sample calibration by number of claims of the corrective best-estimate, unaware, and aware GGPR models (with and without balance correction). Δ shows deviations from the total observed counts.

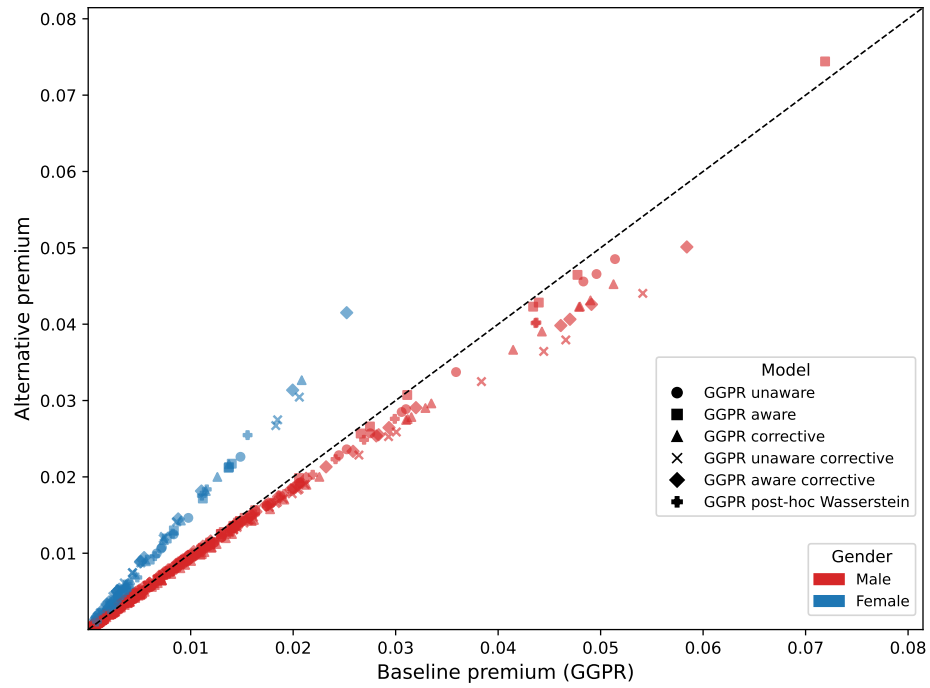


Figure 5.B.6: Scatter plot of predicted premiums (in-sample) for each FairGGPR model against the GGPR baseline (no fairness intervention). Each point corresponds to an individual policy. Deviations from the identity line indicate the magnitude and direction of fairness adjustments. The Wasserstein benchmark is included for comparison.

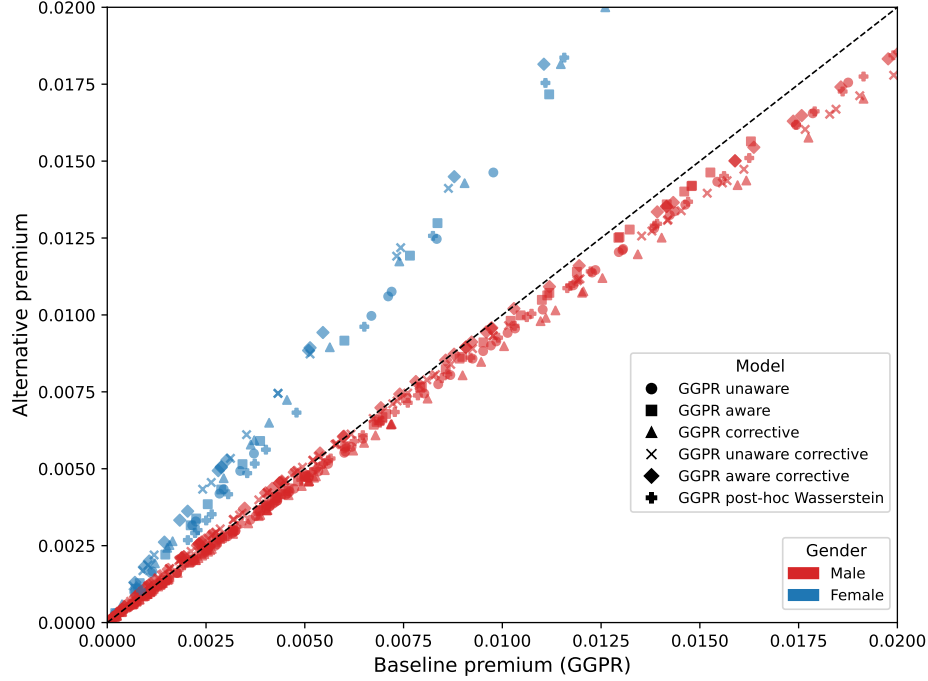


Figure 5.B.7: Zoomed-in version of Fig. 5.B.6 restricted to the region of low premiums. This view highlights the concentration of the distribution near zero and allows for a more precise visualization of small deviations between models.

5.C Market simulation scenarios

Insurers are concerned that enforcing fairness in pricing may affect their profit, solvency, portfolio retention, and overall competitiveness. In particular, if fairness-adjusted premiums lead to higher prices for low-risk policyholders (and lower premiums for previously overcharged groups), competitors with different portfolio composition (and eventually different fairness strategy) could offer lower premiums and attract these lower-risk policyholders.

The idea of this Section is to introduce a decision layer that evaluates economic performance of fairness interventions under different portfolio compositions. In other words, when does a fairness intervention become commercially advantageous for an insurance company (over another) depending on the portfolio proportion of protected policyholders?

To do so, we simulate different insurers with different portfolio compositions. That is, we chose a different value for the proportion $p \in [0, 1]$ of the protected

class in their respective portfolio.

We evaluate them under different fairness strategies, comparing their profit and risk. This allows us to study when a fairness strategy outperforms another (e.g., when the protected group is larger) and adverse selection effects. This is a key insurance insight as an unfair model may offer better pricing and therefore attract profitable risks whereas a fair model may lead to pooled pricing and attract worse risks.

5.C.1 Settings

We simulate a non-life insurance portfolio of arbitrary size $n = 10\,000$ using a Poisson frequency model. Each policyholder is characterized by a set of allowed pricing features $\mathbf{A}$ and a binary protected attribute $D \in \{0, 1\}$. The features are the age of the policyholder $\text{Age} \sim \mathcal{U}(18, 80)$, its vehicle power $\text{Power} \sim \mathcal{N}(100, 20^2)$, its geographic location $\text{Region} \in \{0, 1, 2\}$ with probabilities $(0.5, 0.3, 0.2)$, and the protected attribute $D \sim \text{Bernoulli}(0.3)$. To introduce indirect discrimination, we induce dependence between D and the risk profile by shifting the region variable $\text{Region} \leftarrow \min(\text{Region} + D, 2)$. The continuous variables (age, power) are standardized, and the categorical region variable is one-hot encoded (with one level dropped). The latent log-intensity is defined as:

$$f(\mathbf{A}, D) = -3.0 + 0.02(\text{Age} - 40) + 0.01(\text{Power} - 100) + 0.3\text{Region} + 0.4D.$$

The claim frequency is then:

$$\lambda = \exp(f(\mathbf{A}, D)).$$

This specification incorporates a direct effect of the protected attribute ($0.4D$), and an indirect effect through correlation with region. Each policy is associated with an exposure $\nu \sim \mathcal{U}(0.5, 1.5)$ and the number of claims is generated as:

$$N \sim \text{Poisson}(\nu\lambda), \quad y = \frac{N}{\nu}.$$

To simulate distinct insurers, we resample the portfolio with a chosen proportion $p = \mathbb{P}(D = 1) \in \{0.2, 0.5, 0.8\}$ of protected individuals ($p = 0.2$ corresponds to a portfolio where the protected group is underrepresented, $p = 0.5$ to a balanced portfolio, and $p = 0.8$ to a portfolio where the protected group is the majority). From a market perspective, varying p reflects insurers specializing in different segments, which is key to studying competitive dynamics and adverse selection. From a fairness modeling perspective, the choice of these proportions is intended to capture the effects of different fairness strategies (i.e., best-estimate, unaware, aware, each with and without corrective adjustments) on risk retention, adverse selection, profit, and overall competitiveness.

Remark 5.9. All models are estimated without balance correction, so differences in profitability directly reflect social fairness constraints rather than portfolio-level recalibration adjustments.

For each strategy, we compute the insurer's profit $\sum_i \nu_i \pi_i (1 + \ell) - \nu_i y_i$ with loading factor $\ell = 0.1$, its loss ratio $\text{LR} = \sum_i \nu_i y_i / \sum_i \nu_i \pi_i$, and the training deviance (which measures the quality of the model fit).

We simulate a competitive market in which multiple insurers, each applying a different pricing strategy, compete over a common pool of policyholders. For a given portfolio composition p , all insurers evaluate the same set of risks, but compute premiums according to their respective fairness strategy. Each policyholder is assumed to select the insurer offering the lowest premium:

$$\text{choice}_i = \arg \min_j \nu_i \pi_{ij} (1 + \ell).$$

This allows us to isolate the impact of different fairness strategies and portfolio composition on the insurer's market share and profit under competition.

5.C.2 Results

Table 5.C.1 summarizes profit and loss ratio by strategy and portfolio composition. The best-estimate model (highlighted in grey) corresponds to actuarial pricing without any fairness intervention. The last column reports relative profit differences with respect to this baseline.

We first observe that the aware kernel strategy achieves, quite consistently, the highest profit across all portfolio compositions. Second, enforcing corrective fairness is costly. The impact is particularly pronounced when combined with the unaware kernel and $p = 0.5$, where profitability drops sharply and the loss ratio exceeds 1. Third, the effect (and cost) of fairness depends on the portfolio composition. As the proportion of protected individuals increases (from $p = 0.2$ to $p = 0.8$), all strategies become more profitable. For instance, the aware model with correction yields only an 8.6% gain at $p = 0.5$, but a 32.3% gain at $p = 0.8$. The gap between corrective and non-corrective strategies also narrows as p increases, but the relative ranking remains stable across p . The aware model dominates, followed by the best-estimate and unaware models.

Overall, these results highlight how the relative profitability of fairness strategies depends on market composition. In particular, strategies enforcing fairness may become advantageous when the protected group represents a large share of the portfolio.

Strategy	Profit			Loss Ratio		
	$p = 0.2$	$p = 0.5$	$p = 0.8$	$p = 0.2$	$p = 0.5$	$p = 0.8$
Aware	233.10	260.15	293.93	0.884	0.894	0.905
Aware corrective	181.40	195.14	264.03	0.924	0.938	0.921
Best-estimate	160.49	179.74	199.52	0.942	0.949	0.960
Best-estimate corrective	130.48	130.57	179.64	0.968	0.986	0.972
Unaware	145.76	163.76	185.80	0.954	0.961	0.968
Unaware corrective	96.38	99.32	154.08	0.999	1.011	0.988

Table 5.C.1: Profit and loss ratio by strategy across different portfolio compositions p (no balance correction). The baseline best-estimate GGPR is highlighted in grey.

Table 5.C.2 summarizes the average performance of each fairness strategy across all portfolio compositions. First, the aware model achieves the best overall performance, with the highest average profit, lowest deviance, and lowest loss ratio. Second, corrective fairness interventions systematically deteriorate performance. For all kernel types, introducing correction reduces profit, increases deviance, and raises the loss ratio. This confirms that enforcing fairness through demographic parity limits the model’s ability to differentiate risks. Third, the unaware model performs worse than the best-estimate benchmark, both in terms of profitability and model fit. This highlights the cost of excluding relevant information. Finally, the variability of profits (as measured by the standard deviation) is higher for corrective strategies, suggesting it may also introduce greater instability in pricing outcomes across different portfolio compositions.

Overall, these results reinforce the distinction between individual fairness by kernel design and group fairness by corrective projection, with the former achieving superior economic and statistical performance.

Strategy	Profit (mean)	Profit (std)	Deviance	Loss ratio
Aware	262.39	27.35	4482.27	0.894
Aware corrective	213.53	38.84	4543.50	0.928
Best-estimate	179.92	17.62	4673.01	0.950
Best-estimate corrective	146.90	27.14	4743.80	0.975
Unaware	165.11	18.04	4733.61	0.961
Unaware corrective	116.59	29.12	4804.73	0.999

Table 5.C.2: Average performance across simulations and portfolio compositions (no balance correction). The unconstrained best-estimate model is highlighted in grey.

Table 5.C.3 summarizes profit under competition as a function of the proportion of protected policyholders. Each line corresponds to a pricing strategy. Each block

corresponds to a different portfolio composition.

In this specific simulation setting, the best-estimate model achieves relatively high market shares but generates low profit under competition (probably because it underprices risk and attract higher-risk policyholders). Fairness strategies, particularly the aware interventions, become increasingly profitable as the proportion of protected policyholders increases. Corrective fairness strategies display a non-linear behavior. They may reduce profitability when the protected group is small, but become economically dominant when the protected group constitutes a large fraction of the portfolio. Those are great illustrations of how the economic viability of fairness interventions depends on portfolio composition. Finally, unaware strategies (i.e., simply ignoring the sensitive variables) systematically underperform, both in terms of profitability and stability, confirming that ignoring relevant structure in the data can lead to severe mispricing and unsustainable portfolios.

Strategy	Profit (comp.)			Market share		
	$p = 0.2$	$p = 0.5$	$p = 0.8$	$p = 0.2$	$p = 0.5$	$p = 0.8$
Aware	74.34	65.18	127.66	0.140	0.110	0.175
Aware corrective	74.69	94.37	156.79	0.140	0.189	0.262
Best-estimate	-4.28	23.67	12.66	0.261	0.240	0.103
Best-estimate corrective	-32.13	-56.76	-61.63	0.081	0.147	0.156
Unaware	-18.72	-18.68	-2.72	0.121	0.075	0.035
Unaware corrective	-184.11	-228.37	-267.46	0.257	0.239	0.268

Table 5.C.3: Profit and market share under competition (averaged over simulations). The baseline best-estimate GGPR is highlighted in grey.

These results illustrate the emergence of adverse selection effects. Pricing strategies that are more granular may attract lower-risk individuals, while fairness-constrained strategies may attract higher-risk portfolios. The simulation highlights a key trade-off between fairness and profitability. While unconstrained models achieve better individual risk differentiation, fairness strategies can become economically competitive depending on the market composition. Overall, the results demonstrate that fairness is not only a regulatory or ethical constraint, but also a strategic variable influencing market share and competitive positioning.

Discussion

This thesis set out to model heterogeneous insurance tabular data in a way that is interpretable, statistically robust, and compliant with evolving fairness requirements. Instead of treating clustering, representation learning, and fairness interventions as separate problems, the thesis argues that they should form a single modeling pipeline. Across five research chapters, this work progressively builds a framework linking data heterogeneity, probabilistic modeling, and fair pricing.

Chapters 1 to 3 focus on the latent representation and segmentation of heterogeneous insurance portfolios.

Chapter 1 introduces a Burt distance-based extension of classical clustering algorithms, to better measure dissimilarity for mixed categorical and numerical insurance data. The successful adaptation of K-means variants, fuzzy clustering, and spectral clustering demonstrates that classical unsupervised methods can be made actuarially meaningful.

Chapter 2 moves onto continuous latent representations using Variational AutoEncoders. Unlike distance-based clustering, VAEs provide a generative and probabilistic embedding of insurance data. The ability to generate realistic synthetic policies addresses a practical constraint of actuarial modeling, that is, data availability under privacy regulation. At the same time, this chapter highlights an important methodological trade-off. While neural latent representations can capture complex interactions, they are sensitive to design choices and require hyperparameter tuning to remain stable and interpretable.

Chapter 3 brings back the importance of interpretability by adapting Latent Dirichlet Allocation to tabular insurance data. By modeling policies as mixtures of latent risk profiles, the LDA framework reconciles soft clustering with probabilistic interpretability. Compared to neural embeddings, this approach yields directly interpretable modality distributions, aligning segmentation outputs with actuarial reasoning about rating factors.

Taken together, these three chapters establish a first key insight of the thesis, that is, risk heterogeneity must be structured before it can be regulated. Without a representation of mixed-type data, any downstream adjustment (e.g., fairness

interventions and pricing adjustments) becomes difficult to justify or control.

Chapters 4 and 5 address fairness in the pricing mechanism.

Chapter 4 introduces a gradient-based fairness correction using the multivariate Energy distance to enforce demographic parity across multiple sensitive attributes simultaneously. The post-hoc method aligns premium distributions between groups while preserving variation within each group, and it does so in a way that does not depend on the order of corrections. The addition of autocalibration ensures that fairness adjustments did not break actuarial balance.

Post-hoc and group-level correction remain restrictive choices. Chapter 5 therefore goes one step further by embedding multiple fairness notions directly into a single model in a Poisson setting. Using a Bayesian non-parametric approach based on Generalized Gaussian Process Regression, different fairness criteria are included as part of the learning process itself. The transition from post-processing correction to in-model fairness embedding reflects a conceptual progression. Fairness is not merely a constraint to be satisfied, but a modeling choice that reshapes the risk landscape.

The thesis contributes to actuarial science in five ways. First, it proposes a scalable clustering framework, based on Burt distance, to identify risk profiles in mixed-type insurance data. Second, it develops a VAE architecture adapted to multi-modal non-Gaussian actuarial data for representation learning and synthetic data generation. Third, it extends the use of LDA to structured insurance data as a probabilistic soft clustering framework for interpretable risk profiling. Fourth, it designs a multivariate, order-independent gradient-based fairness correction using Energy distance with autocalibration. And finally, it builds a multi-criteria fair Generalized Gaussian Process Regression model for log-link count data that embeds fairness interventions directly in the Bayesian learning mechanism.

Whether through Burt distance, latent neural embeddings, probabilistic topic mixtures, or kernel functions in GGPR, the structure imposed on the data fundamentally determines downstream pricing and fairness behavior. Representation is not neutral. It encodes assumptions about similarity, heterogeneity, and interaction between rating factors.

Several limitations must be acknowledged. First, the empirical analyses focus mainly on claim frequency modeling. Extending, for example, fairness integration to joint frequency-severity modeling may introduce additional trade-offs. Second, hyperparameter sensitivity in deep generative models can be an issue, especially in practice where reproducibility and stability are critical. Finally, fairness definitions remain normative choices. This thesis does not try to resolve the philosophical debate between competing fairness notions. Instead, it provides tools to implement and compare them in a consistent way.

Bibliography

- Abdi, Hervé and Lynne J. Williams (July 2010). “Principal component analysis”. In: *WIREs Computational Statistics* 2.4, pp. 433–459. ISSN: 1939-0068. DOI: [10.1002/wics.101](https://doi.org/10.1002/wics.101).
- Abdul-Rahman, Shuzlina et al. (2021). “Customer Segmentation and Profiling for Life Insurance using K-Modes Clustering and Decision Tree Classifier”. In: *International Journal of Advanced Computer Science and Applications* 12.9. ISSN: 2158-107X. DOI: [10.14569/ijacsa.2021.0120950](https://doi.org/10.14569/ijacsa.2021.0120950).
- Ahmad, Amir and Shehroz S. Khan (2019). “Survey of State-of-the-Art Mixed Data Clustering Algorithms”. In: *IEEE Access* 7, pp. 31883–31902. ISSN: 2169-3536. DOI: [10.1109/access.2019.2903568](https://doi.org/10.1109/access.2019.2903568).
- Alazizi, Ayman et al. (2020). “Dual Sequential Variational Autoencoders for Fraud Detection”. In: *Advances in Intelligent Data Analysis XVIII*. Springer International Publishing, pp. 14–26. ISBN: 9783030445843. DOI: [10.1007/978-3-030-44584-3_2](https://doi.org/10.1007/978-3-030-44584-3_2).
- Ang, Zi Qing and See Keong Lee (July 2022). “Hierarchical Bayesian Gaussian process regression model for loss reserving using combinations of squared exponential kernels”. In: *Insurance: Mathematics and Economics* 105, pp. 54–63. ISSN: 0167-6687. DOI: [10.1016/j.insmatheco.2022.03.008](https://doi.org/10.1016/j.insmatheco.2022.03.008).
- Antelmi, Luigi et al. (2019). “Sparse Multi-Channel Variational Autoencoder for the Joint Analysis of Heterogeneous Data”. In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 302–311. URL: <https://proceedings.mlr.press/v97/antelmi19a.html>.
- Arthur, David and Sergei Vassilvitskii (2007). “k-means++: the advantages of careful seeding”. In: *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA '07. New Orleans, Louisiana: Society for Industrial and Applied Mathematics, pp. 1027–1035. ISBN: 9780898716245. URL: <https://dl.acm.org/doi/10.5555/1283383.1283494>.

- Avanzi, Benjamin et al. (Apr. 2024). “Machine Learning with High-Cardinality Categorical Features in Actuarial Applications”. In: *ASTIN Bulletin* 54.2, pp. 213–238. ISSN: 1783-1350. DOI: [10.1017/asb.2024.7](https://doi.org/10.1017/asb.2024.7).
- Avraham, Ronen (2017). “Discrimination and Insurance”. In: *The Routledge handbook of the ethics of discrimination*, pp. 335–347. DOI: [10.2139/ssrn.3089946](https://doi.org/10.2139/ssrn.3089946).
- Banerjee, Arindam et al. (Apr. 2004). “Clustering with Bregman Divergences”. In: *Proceedings of the 2004 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, pp. 234–245. DOI: [10.1137/1.9781611972740.22](https://doi.org/10.1137/1.9781611972740.22).
- Barocas, Solon and Andrew D Selbst (2016). “Big data’s disparate impact”. In: *Calif. L. Rev.* 104, p. 671. DOI: [10.15779/Z38BG31](https://doi.org/10.15779/Z38BG31).
- Barry, Laurence and Arthur Charpentier (Jan. 2020). “Personalization as a promise: Can Big Data change the practice of insurance?”. In: *Big Data and Society* 7.1, p. 205395172093514. ISSN: 2053-9517. DOI: [10.1177/2053951720935143](https://doi.org/10.1177/2053951720935143).
- Baur, Christoph et al. (2019). “Deep Autoencoding Models for Unsupervised Anomaly Segmentation in Brain MR Images”. In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Springer International Publishing, pp. 161–169. ISBN: 9783030117238. DOI: [10.1007/978-3-030-11723-8_16](https://doi.org/10.1007/978-3-030-11723-8_16).
- Belkin, Mikhail and Partha Niyogi (2001). “Laplacian eigenmaps and spectral techniques for embedding and clustering”. In: *Proceedings of the 15th International Conference on Neural Information Processing Systems: Natural and Synthetic*. NIPS’01. Vancouver, British Columbia, Canada: MIT Press, pp. 585–591. URL: <https://dl.acm.org/doi/10.5555/2980539.2980616>.
- Bellemare, Marc G. et al. (2017). “The Cramer Distance as a Solution to Biased Wasserstein Gradients”. In: *arXiv preprint*. URL: <https://arxiv.org/abs/1705.10743>.
- Bezdek, James C., Robert Ehrlich, and William Full (Jan. 1984). “FCM: The fuzzy c-means clustering algorithm”. In: *Computers and Geosciences* 10.2–3, pp. 191–203. ISSN: 0098-3004. DOI: [10.1016/0098-3004\(84\)90020-7](https://doi.org/10.1016/0098-3004(84)90020-7).
- Bishop, Christopher M and Nasser M Nasrabadi (2006). *Pattern recognition and machine learning*. 4. Springer. URL: <https://dl.acm.org/doi/10.5555/1162264>.
- Bjerre, Dorethe Skovgaard (May 2022). “Tree-based machine learning methods for modeling and forecasting mortality”. In: *ASTIN Bulletin* 52.3, pp. 765–787. ISSN: 1783-1350. DOI: [10.1017/asb.2022.11](https://doi.org/10.1017/asb.2022.11).
- Blei, David M., Andrew Y. Ng, and Michael I. Jordan (Mar. 2003). “Latent dirichlet allocation”. In: *Journal of machine Learning research* 3.null, pp. 993–1022. ISSN: 1532-4435. DOI: [10.5555/944919.944937](https://doi.org/10.5555/944919.944937).
- Bond-Taylor, Sam et al. (Nov. 2022). “Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.11, pp. 7327–7347. ISSN: 1939-3539. DOI: [10.1109/tpami.2021.3116668](https://doi.org/10.1109/tpami.2021.3116668).

- Bühlmann, Hans and Alois Gisler (2005). *A course in credibility theory and its applications*. Springer. DOI: [10.1007/3-540-29273-X](https://doi.org/10.1007/3-540-29273-X).
- Burt, Cyril (Nov. 1950). “The factorial analysis of qualitative data”. In: *British Journal of Statistical Psychology* 3.3, pp. 166–185. ISSN: 0950-561X. DOI: [10.1111/j.2044-8317.1950.tb00296.x](https://doi.org/10.1111/j.2044-8317.1950.tb00296.x).
- Campbell, Malcolm (1986). “An Integrated System for Estimating the Risk Premium of Individual Car Models in Motor Insurance”. In: *ASTIN Bulletin* 16.2, pp. 165–183. DOI: [10.2143/AST.16.2.2015006](https://doi.org/10.2143/AST.16.2.2015006).
- Campo, Bavo D.C. and Katrien Antonio (Feb. 2024). “On clustering levels of a hierarchical categorical risk factor”. In: *Annals of Actuarial Science* 18.3, pp. 540–578. ISSN: 1748-5002. DOI: [10.1017/s1748499523000283](https://doi.org/10.1017/s1748499523000283).
- Cao, Wenzhi, Vahid Mirjalili, and Sebastian Raschka (2020). “Rank consistent ordinal regression for neural networks with application to age estimation”. In: *Pattern Recognition Letters* 140, pp. 325–331. DOI: [10.1016/j.patrec.2020.11.008](https://doi.org/10.1016/j.patrec.2020.11.008).
- Carpenter, Bob (2010). “Integrating out multinomial parameters in latent dirichlet allocation and naive bayes for collapsed gibbs sampling”. In: *Rapport Technique* 4, p. 464. URL: <https://api.semanticscholar.org/CorpusID:2104516>.
- Chakraborty, Tanujit et al. (Jan. 2024). “Ten years of generative adversarial nets (GANs): a survey of the state-of-the-art”. In: *Machine Learning: Science and Technology* 5.1, p. 011001. ISSN: 2632-2153. DOI: [10.1088/2632-2153/ad1f77](https://doi.org/10.1088/2632-2153/ad1f77).
- Chan, Antoni B. and Daxiang Dong (June 2011). “Generalized Gaussian process models”. In: *CVPR 2011*. IEEE, pp. 2681–2688. DOI: [10.1109/cvpr.2011.5995688](https://doi.org/10.1109/cvpr.2011.5995688).
- Charpentier, Arthur (2023). *Insurance, Biases, Discrimination and Fairness*. DOI: [10.1007/978-3-031-49783-4](https://doi.org/10.1007/978-3-031-49783-4).
- Charpentier, Arthur, Emmanuel Flachaire, and Ewen Gallic (2023). “Optimal Transport for Counterfactual Estimation: A Method for Causal Inference”. In: *Optimal Transport Statistics for Economics and Related Topics*. Springer Nature Switzerland, pp. 45–89. ISBN: 9783031357633. DOI: [10.1007/978-3-031-35763-3_3](https://doi.org/10.1007/978-3-031-35763-3_3).
- Chevalier, Dominik and Marie-Pier Côté (Aug. 2025). “From point to probabilistic gradient boosting for claim frequency and severity prediction”. In: *European Actuarial Journal* 15.3, pp. 707–752. ISSN: 2190-9741. DOI: [10.1007/s13385-025-00428-5](https://doi.org/10.1007/s13385-025-00428-5).
- Chzheng, Evgenii et al. (2020). “Fair regression with wasserstein barycenters”. In: *Advances in Neural Information Processing Systems* 33, pp. 7321–7331. URL: <https://arxiv.org/abs/2006.07286>.
- Côté, Marie-Pier, Olivier Côté, and Arthur Charpentier (2026). “Recoverability of market-wide fair insurance premiums under selection bias”. In: *Insurance: Mathematics and Economics* 129, p. 103268. ISSN: 0167-6687. DOI: <https://doi.org/10.1016/j.insmatheco.2026.103268>.

- Côté, Olivier, Marie-Pier Côté, and Arthur Charpentier (2025). “A fair price to pay: Exploiting causal graphs for fairness in insurance”. In: *Journal of Risk and Insurance* 92.1, pp. 33–75. ISSN: 1539-6975. DOI: [10.1111/jori.12503](https://doi.org/10.1111/jori.12503).
- Cramér, Harald (1928). “On the composition of elementary errors”. In: *Scandinavian Actuarial Journal* 1928.1, pp. 13–74. DOI: [10.1080/03461238.1928.10416862](https://doi.org/10.1080/03461238.1928.10416862).
- Darling, William M (2011). “A theoretical and practical implementation tutorial on topic modeling and gibbs sampling”. In: *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 642–647. URL: <https://api.semanticscholar.org/CorpusID:5790714>.
- Deasy, Jacob, Nikola Simidjievski, and Pietro Liò (2020). “Constraining variational inference with geometric jensen-shannon divergence”. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS ’20*. Vancouver, BC, Canada: Curran Associates Inc. ISBN: 9781713829546. URL: <https://dl.acm.org/doi/10.5555/3495724.3496617>.
- Debener, Jörn, Volker Heinke, and Johannes Kriebel (May 2023). “Detecting insurance fraud using supervised and unsupervised machine learning”. In: *Journal of Risk and Insurance* 90.3, pp. 743–768. ISSN: 1539-6975. DOI: [10.1111/jori.12427](https://doi.org/10.1111/jori.12427).
- Deka, Namrata and Danica J Sutherland (2023). *MMD-B-Fair: Learning Fair Representations with Statistical Testing*. PMLR. URL: <https://arxiv.org/abs/2211.07907>.
- Delong, Łukasz and Anna Kozak (Apr. 2023). “The use of autoencoders for training neural networks with mixed categorical and numerical features”. In: *ASTIN Bulletin* 53.2, pp. 213–232. ISSN: 1783-1350. DOI: [10.1017/asb.2023.15](https://doi.org/10.1017/asb.2023.15).
- Denuit, Michel, Arthur Charpentier, and Julien Trufin (2021). “Autocalibration and Tweedie-dominance for insurance pricing with machine learning”. In: *Insurance: Mathematics and Economics* 101, pp. 485–497. ISSN: 0167-6687. DOI: [10.1016/j.insmatheco.2021.09.001](https://doi.org/10.1016/j.insmatheco.2021.09.001).
- Denuit, Michel, Donatien Hainaut, and Julien Trufin (2019a). *Effective Statistical Learning Methods for Actuaries III: Neural Networks and Extensions*. Springer International Publishing. ISBN: 9783030258276. DOI: [10.1007/978-3-030-25827-6](https://doi.org/10.1007/978-3-030-25827-6).
- (2019b). “Generalized Additive Models (GAMs)”. In: *Effective Statistical Learning Methods for Actuaries I: GLMs and Extensions*. Springer, pp. 253–327.
- (2019c). “Generalized Linear Models (GLMs)”. In: *Effective Statistical Learning Methods for Actuaries I: GLMs and Extensions*, pp. 97–196.
- (2019d). “Insurance Risk Classification”. In: *Effective Statistical Learning Methods for Actuaries I: GLMs and Extensions*, pp. 3–26. ISSN: 2523-3270. DOI: [10.1007/978-3-030-25820-7_1](https://doi.org/10.1007/978-3-030-25820-7_1).
- (2020). “Bagging Trees and Random Forests”. In: *Effective Statistical Learning Methods for Actuaries II: Tree-Based Methods and Extensions*. Springer, pp. 107–130.

- Dinh, Duy-Tai, Van-Nam Huynh, and Songsak Sriboonchitta (Sept. 2021). “Clustering mixed numerical and categorical data with missing values”. In: *Information Sciences* 571, pp. 418–442. ISSN: 0020-0255. DOI: [10.1016/j.ins.2021.04.076](https://doi.org/10.1016/j.ins.2021.04.076).
- Dolman, Chris, Edward Frees, and Fei Huang (Nov. 2021). “Multidisciplinary collaboration on discrimination – not just “Nice to Have””. In: *Annals of Actuarial Science* 15.3, pp. 485–487. ISSN: 1748-5002. DOI: [10.1017/s174849952100021x](https://doi.org/10.1017/s174849952100021x).
- Duan, Yitong et al. (June 2022). “FactorVAE: A Probabilistic Dynamic Factor Model Based on Variational Autoencoder for Predicting Cross-Sectional Stock Returns”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.4, pp. 4468–4476. ISSN: 2159-5399. DOI: [10.1609/aaai.v36i4.20369](https://doi.org/10.1609/aaai.v36i4.20369).
- Dunn, J. C. (Jan. 1973). “A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters”. In: *Journal of Cybernetics* 3.3, pp. 32–57. ISSN: 0022-0280. DOI: [10.1080/01969727308546046](https://doi.org/10.1080/01969727308546046).
- Duong, Lyndon R. et al. (2022). “Representational dissimilarity metric spaces for stochastic neural networks”. In: *arXiv preprint*. DOI: [10.48550/ARXIV.2211.11665](https://doi.org/10.48550/ARXIV.2211.11665).
- Edwards, A. (2007). “Wouter Vandenhoe, Non-Discrimination and Equality in the View of the UN Human Rights Treaty Bodies”. In: *Human Rights Law Review* 7.1, pp. 267–269. ISSN: 1744-1021. DOI: [10.1093/hrlr/ngl032](https://doi.org/10.1093/hrlr/ngl032).
- Fahrenwaldt, Matthias et al. (June 2024). “Fairness: plurality, causality, and insurability”. In: *European Actuarial Journal* 14.2, pp. 317–328. ISSN: 2190-9741. DOI: [10.1007/s13385-024-00387-3](https://doi.org/10.1007/s13385-024-00387-3).
- Fatras, Kilian et al. (2019). *Learning with minibatch Wasserstein : asymptotic and gradient properties*. DOI: [10.48550/ARXIV.1910.04091](https://doi.org/10.48550/ARXIV.1910.04091).
- Ferrario, Andrea, Alexander Noll, and Mario V. Wuthrich (2018). “Insights from Inside Neural Networks”. In: *SSRN Electronic Journal*. ISSN: 1556-5068. DOI: [10.2139/ssrn.3226852](https://doi.org/10.2139/ssrn.3226852).
- Frees, Edward W. and Fei Huang (2021). “The Discriminating (Pricing) Actuary”. In: *North American Actuarial Journal* 27.1, pp. 2–24. ISSN: 2325-0453. DOI: [10.1080/10920277.2021.1951296](https://doi.org/10.1080/10920277.2021.1951296).
- Fukuyama, Julia, Kris Sankaran, and Laura Symul (June 2022). “Multiscale analysis of count data through topic alignment”. In: *Biostatistics* 24.4, pp. 1045–1065. ISSN: 1468-4357. DOI: [10.1093/biostatistics/kxac018](https://doi.org/10.1093/biostatistics/kxac018).
- Gabric, Lydia, Shuang Zhou, and Kenneth Q. Zhou (2024). “A Bayesian Approach to Discrimination-free Insurance Pricing”. In: *SSRN Electronic Journal*. ISSN: 1556-5068. DOI: [10.2139/ssrn.4785927](https://doi.org/10.2139/ssrn.4785927).
- Gan, Guojun (Nov. 2013). “Application of data clustering and machine learning in variable annuity valuation”. In: *Insurance: Mathematics and Economics* 53.3, pp. 795–801. ISSN: 0167-6687. DOI: [10.1016/j.insmatheco.2013.09.021](https://doi.org/10.1016/j.insmatheco.2013.09.021).
- Gan, Guojun and Emiliano A. Valdez (June 2019). “Data Clustering with Actuarial Applications”. In: *North American Actuarial Journal* 24.2, pp. 168–186. ISSN: 2325-0453. DOI: [10.1080/10920277.2019.1575242](https://doi.org/10.1080/10920277.2019.1575242).

- Gandy, Oscar H (2016). *Coming to Terms with Chance: Engaging Rational Discrimination and Cumulative Disadvantage*. Routledge.
- García-Ordás, María Teresa et al. (Feb. 2020). “Detecting Respiratory Pathologies Using Convolutional Neural Networks and Variational Autoencoders for Unbalancing Data”. In: *Sensors* 20.4, p. 1214. ISSN: 1424-8220. DOI: [10.3390/s20041214](https://doi.org/10.3390/s20041214).
- Genevay, Aude, Gabriel Peyré, and Marco Cuturi (2017). *Learning Generative Models with Sinkhorn Divergences*. DOI: [10.48550/ARXIV.1706.00292](https://arxiv.org/abs/1706.00292).
- Gentle, J. E., L. Kaufman, and P. J. Rousseuw (June 1991). “Finding Groups in Data: An Introduction to Cluster Analysis.” In: *Biometrics* 47.2, p. 788. ISSN: 0006-341X. DOI: [10.2307/2532178](https://doi.org/10.2307/2532178).
- Glorot, Xavier, Antoine Bordes, and Yoshua Bengio (2011). “Deep Sparse Rectifier Neural Networks”. In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Geoffrey Gordon, David Dunson, and Miroslav Dudík. Vol. 15. Proceedings of Machine Learning Research. Fort Lauderdale, FL, USA: PMLR, pp. 315–323. URL: <https://proceedings.mlr.press/v15/glorot11a.html>.
- Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep learning*. MIT press. URL: <https://dl.acm.org/doi/10.5555/3086952>.
- Goodfellow, Ian J. et al. (2014). “Generative adversarial nets”. In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*. NIPS’14. Montreal, Canada: MIT Press, pp. 2672–2680. URL: <https://dl.acm.org/doi/10.5555/2969033.2969125>.
- Gower, J. C. (Dec. 1971). “A General Coefficient of Similarity and Some of Its Properties”. In: *Biometrics* 27.4, p. 857. ISSN: 0006-341X. DOI: [10.2307/2528823](https://doi.org/10.2307/2528823).
- Grari, Vincent, Sylvain Lamprier, and Marcin Detyniecki (2020). *Adversarial Learning for Counterfactual Fairness*. DOI: [10.48550/ARXIV.2008.13122](https://arxiv.org/abs/10.48550/ARXIV.2008.13122).
- Guo, Cheng and Felix Berkhahn (2016). *Entity Embeddings of Categorical Variables*. DOI: [10.48550/ARXIV.1604.06737](https://arxiv.org/abs/10.48550/ARXIV.1604.06737).
- Gutiérrez, Pedro Antonio et al. (2015). “Ordinal regression methods: survey and experimental study”. In: *IEEE Transactions on Knowledge and Data Engineering* 28.1, pp. 127–146. DOI: [10.1109/TKDE.2015.2457911](https://doi.org/10.1109/TKDE.2015.2457911).
- Hainaut, Donatien (Jan. 2018a). “A neural-network analyzer for mortality forecast”. In: *ASTIN Bulletin* 48.02, pp. 481–508. ISSN: 1783-1350. DOI: [10.1017/asb.2017.45](https://doi.org/10.1017/asb.2017.45).
- (Dec. 2018b). “A self-organizing predictive map for non-life insurance”. In: *European Actuarial Journal* 9.1, pp. 173–207. ISSN: 2190-9741. DOI: [10.1007/s13385-018-0189-z](https://doi.org/10.1007/s13385-018-0189-z).
- (2025). “In-processing of actuarial and equity fairness constraints for Neural networks”. In: *Available at SSRN 5652150*. DOI: [10.2139/ssrn.5652150](https://ssrn.com/abstract=5652150).
- (Mar. 2026). “American option pricing with model constrained Gaussian process regressions”. In: *Applied Mathematics and Computation* 512, p. 129747. ISSN: 0096-3003. DOI: [10.1016/j.amc.2025.129747](https://doi.org/10.1016/j.amc.2025.129747).

- Hainaut, Donatien and Michel Denuit (2025). *Insurance risk classification with Generalized Gaussian Process Regression models*. Tech. rep. URL: https://dial.uclouvain.be/pr/boreal/object/boreal%3A299662/datastream/PDF_01/view.
- Hainaut, Donatien and Frederic Daniel Vrans (2024). “European option pricing with model constrained Gaussian process regressions”. In: *Available at SSRN 4982344*. DOI: [10.2139/ssrn.4982344](https://doi.org/10.2139/ssrn.4982344).
- Hancock, John T. and Taghi M. Khoshgoftaar (2020). “Survey on categorical data for neural networks”. In: *Journal of Big Data* 7.1. ISSN: 2196-1115. DOI: [10.1186/s40537-020-00305-w](https://doi.org/10.1186/s40537-020-00305-w).
- Hardt, Moritz, Eric Price, and Nathan Srebro (2016). *Equality of Opportunity in Supervised Learning*. URL: <https://arxiv.org/abs/1610.02413>.
- Hartigan John A., 1937- (1975). *Clustering algorithms*. eng. Wiley series in probability and mathematical statistics. New York: Wiley. ISBN: 047135645X.
- Hashemi, Seyed Emadedin, Fatemeh Gholian-Jouybari, and Mostafa Hajiaghahi-Keshteli (Oct. 2023). “A fuzzy C-means algorithm for optimizing data clustering”. In: *Expert Systems with Applications* 227, p. 120377. ISSN: 0957-4174. DOI: [10.1016/j.eswa.2023.120377](https://doi.org/10.1016/j.eswa.2023.120377).
- Henckaerts, Roel et al. (July 2021). “Boosting Insights in Insurance Tariff Plans with Tree-Based Machine Learning Methods”. In: *North American Actuarial Journal* 25.2, pp. 255–285. ISSN: 2325-0453. DOI: [10.1080/10920277.2020.1745656](https://doi.org/10.1080/10920277.2020.1745656).
- Hiabu, M. (Oct. 2016). “On the relationship between classical chain ladder and granular reserving”. In: *Scandinavian Actuarial Journal* 2017.8, pp. 708–729. ISSN: 1651-2030. DOI: [10.1080/03461238.2016.1240709](https://doi.org/10.1080/03461238.2016.1240709).
- Higgins, Irina et al. (2017). “beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework”. In: *International Conference on Learning Representations*. URL: <https://openreview.net/forum?id=Sy2fzU9gl>.
- Hoffman, Matthew D., David M. Blei, and Francis Bach (2010). “Online learning for Latent Dirichlet Allocation”. In: *Proceedings of the 24th International Conference on Neural Information Processing Systems - Volume 1*. NIPS’10. Vancouver, British Columbia, Canada: Curran Associates Inc., pp. 856–864. DOI: [10.5555/2997189.2997285](https://doi.org/10.5555/2997189.2997285).
- Hofmann, Thomas (Aug. 1999). “Probabilistic latent semantic indexing”. In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. SIGIR99. ACM, pp. 50–57. DOI: [10.1145/312624.312649](https://doi.org/10.1145/312624.312649).
- Hou, Yanxi, Xiaolan Xia, and Guangyuan Gao (2024). *Combining Structural and Unstructured Data: A Topic-based Finite Mixture Model for Insurance Claim Prediction*. DOI: [10.48550/ARXIV.2410.04684](https://doi.org/10.48550/ARXIV.2410.04684).
- Hsu, W.H. et al. (1999). “Self-organizing systems for knowledge discovery in large databases”. In: *IJCNN’99. International Joint Conference on Neural Networks*.

- Proceedings (Cat. No.99CH36339)*. Vol. 4. IJCNN-99. IEEE, pp. 2480–2485. DOI: [10.1109/ijcnn.1999.833461](https://doi.org/10.1109/ijcnn.1999.833461).
- Hu, François, Philipp Ratz, and Arthur Charpentier (2023). “Fairness in Multi-Task Learning via Wasserstein Barycenters”. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 295–312. DOI: [10.48550/ARXIV.2306.10155](https://doi.org/10.48550/ARXIV.2306.10155).
- (2024). “A Sequentially Fair Mechanism for Multiple Sensitive Attributes”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 11, pp. 12502–12510. DOI: [10.48550/ARXIV.2309.06627](https://doi.org/10.48550/ARXIV.2309.06627).
- Huang, Huaibo et al. (2018). “IntroVAE: introspective variational autoencoders for photographic image synthesis”. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS’18. Montréal, Canada: Curran Associates Inc., pp. 52–63. URL: <https://dl.acm.org/doi/10.5555/3326943.3326949>.
- Huang, Zhexue (1997). “A fast clustering algorithm to cluster very large categorical data sets in data mining.” In: *Data mining and knowledge discovery* 3.8, pp. 34–39.
- (Sept. 1998). “Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values”. In: *Data Mining and Knowledge Discovery* 2.3, pp. 283–304. ISSN: 1573-756X. DOI: [10.1023/a:1009769707641](https://doi.org/10.1023/a:1009769707641).
- Jain, A. K., M. N. Murty, and P. J. Flynn (Sept. 1999). “Data clustering: a review”. In: *ACM Computing Surveys* 31.3, pp. 264–323. ISSN: 1557-7341. DOI: [10.1145/331499.331504](https://doi.org/10.1145/331499.331504).
- Jamotton, Charlotte and Donatien Hainaut (2024). “Variational AutoEncoder for Synthetic Insurance Data”. In: *Intelligent Systems with Applications* 24, p. 200455. ISSN: 2667-3053. DOI: [10.1016/j.iswa.2024.200455](https://doi.org/10.1016/j.iswa.2024.200455).
- (2026a). “A Multi-Criteria Fair Gaussian Regressor for Insurance Pricing”. In: *working paper*. DOI: [None](#).
- (2026b). “A Multivariate Energy-based Fairness Adjuster for Premiums”. In: *North American Actuarial Journal*. ISSN: 1092-0277. DOI: [10.1080/10920277.2026.2705183](https://doi.org/10.1080/10920277.2026.2705183).
- (Feb. 2026c). “Latent Dirichlet Allocation for Structured Insurance Data”. In: *Annals of Data Science*. ISSN: 2198-5812. DOI: [10.1007/s40745-026-00674-8](https://doi.org/10.1007/s40745-026-00674-8).
- Jamotton, Charlotte, Donatien Hainaut, and Thomas Hames (Sept. 2024). “Insurance Analytics with Clustering Techniques”. In: *Risks* 12.9, p. 141. ISSN: 2227-9091. DOI: [10.3390/risks12090141](https://doi.org/10.3390/risks12090141).
- Jang, Eric, Shixiang Gu, and Ben Poole (2016). *Categorical Reparameterization with Gumbel-Softmax*. DOI: [10.48550/ARXIV.1611.01144](https://doi.org/10.48550/ARXIV.1611.01144).
- Kantorovich, Leonid Vasilevich and SG Rubinshtein (1958). “On a space of totally additive functions”. In: *Vestnik of the St. Petersburg University: Mathematics* 13.7, pp. 52–59.
- Kingma, Diederik P and Max Welling (2013). *Auto-Encoding Variational Bayes*. DOI: [10.48550/ARXIV.1312.6114](https://doi.org/10.48550/ARXIV.1312.6114).

- Klambauer, Günter et al. (2017). “Self-normalizing neural networks”. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS’17. Long Beach, California, USA: Curran Associates Inc., pp. 972–981. ISBN: 9781510860964. URL: <https://dl.acm.org/doi/10.5555/3294771.3294864>.
- Knuth, Donald (1981). “Seminumerical algorithms”. In: *The art of computer programming 2*. DOI: [10.1002/spe.4380120909](https://doi.org/10.1002/spe.4380120909).
- Koc, Oguz, Furkan Baser, and A. Sevtap Selcuk-Kestel (Dec. 2024). “Clustering based fuzzy classification with a noise cluster in detecting fraud in insurance”. In: *Applied Soft Computing* 167, p. 112430. ISSN: 1568-4946. DOI: [10.1016/j.asoc.2024.112430](https://doi.org/10.1016/j.asoc.2024.112430).
- Kramer, Mark A. (Feb. 1991). “Nonlinear principal component analysis using autoassociative neural networks”. In: *AIChE Journal* 37.2, pp. 233–243. ISSN: 1547-5905. DOI: [10.1002/aic.690370209](https://doi.org/10.1002/aic.690370209).
- Krestel, Ralf, Peter Fankhauser, and Wolfgang Nejdl (Oct. 2009). “Latent dirichlet allocation for tag recommendation”. In: *Proceedings of the third ACM conference on Recommender systems*. RecSys ’09. ACM, pp. 61–68. DOI: [10.1145/1639714.1639726](https://doi.org/10.1145/1639714.1639726).
- Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton (May 2017). “ImageNet classification with deep convolutional neural networks”. In: *Communications of the ACM* 60.6, pp. 84–90. ISSN: 1557-7317. DOI: [10.1145/3065386](https://doi.org/10.1145/3065386).
- Kullback, S. and R. A. Leibler (1951). “On Information and Sufficiency”. In: *The Annals of Mathematical Statistics* 22.1, pp. 79–86. DOI: [10.1214/aoms/1177729694](https://doi.org/10.1214/aoms/1177729694).
- Kusner, Matt J., Joshua R. Loftus, et al. (2017). *Counterfactual Fairness*. DOI: [10.48550/ARXIV.1703.06856](https://doi.org/10.48550/ARXIV.1703.06856).
- Kusner, Matt J., Brooks Paige, and José Miguel Hernández-Lobato (2017). “Grammar variational autoencoder”. In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML’17. Sydney, NSW, Australia: JMLR.org, pp. 1945–1954. URL: <https://dl.acm.org/doi/10.5555/3305381.3305582>.
- LeCun, Yann A. et al. (2012). “Efficient BackProp”. In: *Neural Networks: Tricks of the Trade*. Springer Berlin Heidelberg, pp. 9–48. ISBN: 9783642352898. DOI: [10.1007/978-3-642-35289-8_3](https://doi.org/10.1007/978-3-642-35289-8_3).
- Li, Bo, Zhengxing Sun, and Yuqi Guo (July 2019). “SuperVAE: Superpixelwise Variational Autoencoder for Salient Object Detection”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01, pp. 8569–8576. ISSN: 2159-5399. DOI: [10.1609/aaai.v33i01.33018569](https://doi.org/10.1609/aaai.v33i01.33018569).
- Li, Xiaopeng and James She (Aug. 2017). “Collaborative Variational Autoencoder for Recommender Systems”. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’17. ACM, pp. 305–314. DOI: [10.1145/3097983.3098077](https://doi.org/10.1145/3097983.3098077).

- Lim, Jaechang et al. (July 2018). “Molecular generative model based on conditional variational autoencoder for de novo molecular design”. In: *Journal of Cheminformatics* 10.1. ISSN: 1758-2946. DOI: [10.1186/s13321-018-0286-7](https://doi.org/10.1186/s13321-018-0286-7).
- Lindholm, Mathias, Ronald Richman, Andreas Tsanakas, and Mario V Wüthrich (2022a). “Discrimination-free insurance pricing”. In: *ASTIN Bulletin: The Journal of the IAA* 52.1, pp. 55–89. DOI: [10.1017/asb.2021.23](https://doi.org/10.1017/asb.2021.23).
- (2024a). “A Multi-Task Network Approach for Calculating Discrimination-Free Insurance Prices”. In: *European Actuarial Journal* 14.2, pp. 329–369. DOI: [10.2139/ssrn.4155585](https://doi.org/10.2139/ssrn.4155585).
- (2026). “Sensitivity-based measures of discrimination in insurance pricing”. In: *European Journal of Operational Research*. DOI: [10.1016/j.ejor.2026.01.021](https://doi.org/10.1016/j.ejor.2026.01.021).
- (2022b). “A Discussion of Discrimination and Fairness in Insurance Pricing”. In: URL: <https://arxiv.org/abs/2209.00858>.
- (2024b). “What is fair? Proxy discrimination vs. demographic disparities in insurance pricing”. In: *Scandinavian Actuarial Journal* 2024.9, pp. 935–970. ISSN: 1651-2030. DOI: [10.1080/03461238.2024.2364741](https://doi.org/10.1080/03461238.2024.2364741).
- Lindholm, Mathias and Mario V Wüthrich (2025). “The balance property in insurance pricing”. In: *Scandinavian Actuarial Journal*, pp. 1–38.
- Lloyd, Chris et al. (July 2015). “Variational Inference for Gaussian Process Modulated Poisson Processes”. In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by Francis Bach and David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, pp. 1814–1822. DOI: [10.48550/ARXIV.1411.0254](https://doi.org/10.48550/ARXIV.1411.0254).
- Lu, Guangquan et al. (Apr. 2020). “Multi-task learning using variational auto-encoder for sentiment classification”. In: *Pattern Recognition Letters* 132, pp. 115–122. ISSN: 0167-8655. DOI: [10.1016/j.patrec.2018.06.027](https://doi.org/10.1016/j.patrec.2018.06.027).
- Ludkovski, Mike, Jimmy Risk, and Howard Zail (Aug. 2018). “Gaussian process models for mortality rates and improvement factors”. In: *ASTIN Bulletin* 48.3, pp. 1307–1347. ISSN: 1783-1350. DOI: [10.1017/asb.2018.24](https://doi.org/10.1017/asb.2018.24).
- Luxburg, Ulrike von (Aug. 2007). “A tutorial on spectral clustering”. In: *Statistics and Computing* 17.4, pp. 395–416. ISSN: 1573-1375. DOI: [10.1007/s11222-007-9033-z](https://doi.org/10.1007/s11222-007-9033-z).
- Ma, Chao et al. (2020). “VAEM: a deep generative model for heterogeneous mixed type data”. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS ’20. Vancouver, BC, Canada: Curran Associates Inc. ISBN: 9781713829546. URL: <https://dl.acm.org/doi/abs/10.5555/3495724.3496667>.
- Maaten, Laurens van der and Geoffrey E. Hinton (2008). “Visualizing Data using t-SNE”. In: *Journal of Machine Learning Research* 9, pp. 2579–2605. URL: <https://api.semanticscholar.org/CorpusID:5855042>.

- Mancisidor, Rogelio A. et al. (Feb. 2021). "Learning latent representations of bank customers with the Variational Autoencoder". In: *Expert Systems with Applications* 164, p. 114020. ISSN: 0957-4174. DOI: [10.1016/j.eswa.2020.114020](https://doi.org/10.1016/j.eswa.2020.114020).
- Mansour, Romany F. et al. (Nov. 2021). "Unsupervised Deep Learning based Variational Autoencoder Model for COVID-19 Diagnosis and Classification". In: *Pattern Recognition Letters* 151, pp. 267–274. ISSN: 0167-8655. DOI: [10.1016/j.patrec.2021.08.018](https://doi.org/10.1016/j.patrec.2021.08.018).
- Mbuga, Felix and Cristina Tortora (Dec. 2021). "Spectral Clustering of Mixed-Type Data". In: *Stats* 5.1, pp. 1–11. ISSN: 2571-905X. DOI: [10.3390/stats5010001](https://doi.org/10.3390/stats5010001).
- Meilă, Marina and Jianbo Shi (2001). "A Random Walks View of Spectral Segmentation". In: *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics*. Ed. by Thomas S. Richardson and Tommi S. Jaakkola. Vol. R3. Proceedings of Machine Learning Research. Reissued by PMLR on 31 March 2021. PMLR, pp. 203–208. URL: <https://proceedings.mlr.press/r3/meila01a.html>.
- Mescheder, Lars, Sebastian Nowozin, and Andreas Geiger (2017). "Adversarial variational bayes: unifying Variational Autoencoders and Generative Adversarial Networks". In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML'17. Sydney, NSW, Australia: JMLR.org, pp. 2391–2400. URL: <https://dl.acm.org/doi/10.5555/3305890.3305928>.
- Miotto, Riccardo et al. (May 2016). "Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records". In: *Scientific Reports* 6.1. ISSN: 2045-2322. DOI: [10.1038/srep26094](https://doi.org/10.1038/srep26094).
- Nazábal, Alfredo et al. (Nov. 2020). "Handling incomplete heterogeneous data using VAEs". In: *Pattern Recognition* 107, p. 107501. ISSN: 0031-3203. DOI: [10.1016/j.patcog.2020.107501](https://doi.org/10.1016/j.patcog.2020.107501).
- Nelder, J. A. and R. Mead (Jan. 1965). "A Simplex Method for Function Minimization". In: *The Computer Journal* 7.4, pp. 308–313. ISSN: 1460-2067. DOI: [10.1093/comjnl/7.4.308](https://doi.org/10.1093/comjnl/7.4.308).
- Ng, Andrew Y., Michael I. Jordan, and Yair Weiss (2001). "On spectral clustering: analysis and an algorithm". In: *Proceedings of the 15th International Conference on Neural Information Processing Systems: Natural and Synthetic*. NIPS'01. Vancouver, British Columbia, Canada: MIT Press, pp. 849–856. URL: <https://dl.acm.org/doi/10.5555/2980539.2980649>.
- Ng, Kai Wang et al. (Mar. 2008). "Grouped Dirichlet distribution: A new tool for incomplete categorical data analysis". In: *Journal of Multivariate Analysis* 99.3, pp. 490–509. ISSN: 0047-259X. DOI: [10.1016/j.jmva.2007.01.010](https://doi.org/10.1016/j.jmva.2007.01.010).
- Nishisato, Shizuhiko (Sept. 1985). "Michael J. Greenacre Theory and Applications of Correspondence Analysis. London: Academic Press, 1984." In: *Psychometrika* 50.3, pp. 376–377. ISSN: 1860-0980. DOI: [10.1017/s0033312300500404](https://doi.org/10.1017/s0033312300500404).
- Noll, Alexander, Robert Salzmänn, and Mario V. Wuthrich (2018). "Case Study: French Motor Third-Party Liability Claims". In: *SSRN Electronic Journal*. ISSN: 1556-5068. DOI: [10.2139/ssrn.3164764](https://doi.org/10.2139/ssrn.3164764).

- O'Hagan, Adrian and Colm Ferrari (Nov. 2016). "Model-Based and Nonparametric Approaches to Clustering for Data Compression in Actuarial Applications". In: *North American Actuarial Journal* 21.1, pp. 107–146. ISSN: 2325-0453. DOI: [10.1080/10920277.2016.1234398](https://doi.org/10.1080/10920277.2016.1234398).
- Öğretir, Mine et al. (Dec. 2022). "A Variational Autoencoder for Heterogeneous Temporal and Longitudinal Data". In: *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, pp. 1522–1529. DOI: [10.1109/icmla55696.2022.00239](https://doi.org/10.1109/icmla55696.2022.00239).
- Ohlsson, Esbjörn and Björn Johansson (2010). *Non-Life Insurance Pricing with Generalized Linear Models*. Springer Berlin Heidelberg. ISBN: 9783642107917. DOI: [10.1007/978-3-642-10791-7](https://doi.org/10.1007/978-3-642-10791-7).
- Oneto, Luca et al. (2020). "Exploiting mmd and sinkhorn divergences for fair and transferable representation learning". In: *Advances in Neural Information Processing Systems* 33, pp. 15360–15370.
- Oord, Aaron van den, Oriol Vinyals, and Koray Kavukcuoglu (2017). "Neural discrete representation learning". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., pp. 6309–6318. ISBN: 9781510860964. URL: <https://dl.acm.org/doi/10.5555/3295222.3295378>.
- Oskouei, Amin Golzari et al. (Sept. 2024). "SSFCM-FWCW: Semi-Supervised Fuzzy C-Means method based on Feature-Weight and Cluster-Weight learning". In: *Software Impacts* 21, p. 100678. ISSN: 2665-9638. DOI: [10.1016/j.simpa.2024.100678](https://doi.org/10.1016/j.simpa.2024.100678).
- Paccanaro, A. (Mar. 2006). "Spectral clustering of protein sequences". In: *Nucleic Acids Research* 34.5, pp. 1571–1580. ISSN: 1362-4962. DOI: [10.1093/nar/gkj515](https://doi.org/10.1093/nar/gkj515).
- Pagnoni, Artidoro, Kevin Liu, and Shangyan Li (2018). *Conditional Variational Autoencoder for Neural Machine Translation*. DOI: [10.48550/ARXIV.1812.04405](https://doi.org/10.48550/ARXIV.1812.04405).
- Panaretos, Victor M. and Yoav Zemel (2019). "Statistical Aspects of Wasserstein Distances". In: *Annual Review of Statistics and Its Application* 6. Volume 6, 2019, pp. 405–431. ISSN: 2326-831X. DOI: [10.1146/annurev-statistics-030718-104938](https://doi.org/10.1146/annurev-statistics-030718-104938).
- Pascanu, Razvan, Tomas Mikolov, and Yoshua Bengio (2013). "On the difficulty of training recurrent neural networks". In: *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*. ICML'13. Atlanta, GA, USA: JMLR.org, III–1310–III–1318. URL: <https://dl.acm.org/doi/10.5555/3042817.3043083>.
- Pedregosa, Fabian et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12.85, pp. 2825–2830. URL: <http://jmlr.org/papers/v12/pedregosa11a.html>.

- Peyré, Gabriel and Marco Cuturi (2019). “Computational Optimal Transport: With Applications to Data Science”. In: *Foundations and Trends® in Machine Learning* 11.5-6, pp. 355–607. ISSN: 1935-8237. DOI: [10.1561/22000000073](https://doi.org/10.1561/22000000073).
- Pleiss, Geoff et al. (2017). *On Fairness and Calibration*. URL: <https://arxiv.org/abs/1709.02012>.
- Rezende, Danilo Jimenez and Shakir Mohamed (2015). “Variational inference with normalizing flows”. In: *Proceedings of the 32nd International Conference on Machine Learning - Volume 37*. ICML’15. Lille, France: JMLR.org, pp. 1530–1538. URL: <https://dl.acm.org/doi/10.5555/3045118.3045281>.
- Rezende, Danilo Jimenez, Shakir Mohamed, and Daan Wierstra (2014). “Stochastic backpropagation and approximate inference in deep generative models”. In: *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*. ICML’14. Beijing, China: JMLR.org, II-1278–II-1286. URL: <https://dl.acm.org/doi/10.5555/3044805.3045035>.
- Richman, Ronald (Aug. 2020). “AI in actuarial science – a review of recent advances – part 1”. In: *Annals of Actuarial Science* 15.2, pp. 207–229. ISSN: 1748-5002. DOI: [10.1017/s1748499520000238](https://doi.org/10.1017/s1748499520000238).
- Richman, Ronald and Mario V. Wüthrich (Feb. 2024). “High-cardinality categorical covariates in network regressions”. In: *Japanese Journal of Statistics and Data Science* 7.2, pp. 921–965. ISSN: 2520-8764. DOI: [10.1007/s42081-024-00243-4](https://doi.org/10.1007/s42081-024-00243-4).
- Rigby, R. A. and D. M. Stasinopoulos (Apr. 2005). “Generalized Additive Models for Location, Scale and Shape”. In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 54.3, pp. 507–554. ISSN: 1467-9876. DOI: [10.1111/j.1467-9876.2005.00510.x](https://doi.org/10.1111/j.1467-9876.2005.00510.x).
- Risk, Jimmy and Mike Ludkovski (Feb. 2024). “Expressive mortality models through Gaussian process kernels”. In: *ASTIN Bulletin* 54.2, pp. 327–359. ISSN: 1783-1350. DOI: [10.1017/asb.2023.39](https://doi.org/10.1017/asb.2023.39).
- Roy, Sagnik (2021). *Car Insurance Data*. Kaggle Dataset. URL: <https://www.kaggle.com/datasets/sagnik1511/car-insurance-data/data>.
- Royden, Halsey Lawrence and Patrick Fitzpatrick (1988). *Real analysis*. Vol. 32. Macmillan New York.
- Santambrogio, Filippo (2015). *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs and Modeling*. Cham: Springer. DOI: [10.1007/978-3-319-20828-2](https://doi.org/10.1007/978-3-319-20828-2).
- Schanze, Erich (2013). “Injustice by Generalization: Notes on the Test-Achats Decision of the European Court of Justice”. In: *German Law Journal* 14.2, pp. 423–433. DOI: [10.1017/S2071832200001863](https://doi.org/10.1017/S2071832200001863).
- Sculley, D. (Apr. 2010). “Web-scale k-means clustering”. In: *Proceedings of the 19th international conference on World wide web*. WWW ’10. ACM, pp. 1177–1178. DOI: [10.1145/1772690.1772862](https://doi.org/10.1145/1772690.1772862).
- Sejdinović, Dino et al. (2013). “Equivalence of distance-based and RKHS-based statistics in hypothesis testing”. In: *Annals of Statistics* 41.5, pp. 2263–2291. DOI: [10.1214/13-AOS1140](https://doi.org/10.1214/13-AOS1140).

- Semeniuta, Stanislau, Aliaksei Severyn, and Erhardt Barth (2017). “A Hybrid Convolutional Variational Autoencoder for Text Generation”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 627–637. DOI: [10.18653/v1/d17-1066](https://doi.org/10.18653/v1/d17-1066).
- Shen, Xiangqing et al. (Sept. 2020). “Remote sensing image captioning via Variational Autoencoder and Reinforcement Learning”. In: *Knowledge-Based Systems* 203, p. 105920. ISSN: 0950-7051. DOI: [10.1016/j.knosys.2020.105920](https://doi.org/10.1016/j.knosys.2020.105920).
- Shi, Jianbo and J. Malik (2000). “Normalized cuts and image segmentation”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.8, pp. 888–905. ISSN: 0162-8828. DOI: [10.1109/34.868688](https://doi.org/10.1109/34.868688).
- Shi, Peng and Kun Shi (Nov. 2022). “Non-Life Insurance Risk Classification Using Categorical Embedding”. In: *North American Actuarial Journal* 27.3, pp. 579–601. ISSN: 2325-0453. DOI: [10.1080/10920277.2022.2123361](https://doi.org/10.1080/10920277.2022.2123361).
- Shim, Jooyong and Changha Hwang (Mar. 2011). “Kernel Poisson regression machine for stochastic claims reserving”. In: *Journal of the Korean Statistical Society* 40.1, pp. 1–9. ISSN: 1226-3192. DOI: [10.1016/j.jkss.2010.01.004](https://doi.org/10.1016/j.jkss.2010.01.004).
- Simard, P.Y., D. Steinkraus, and J.C. Platt (2003). “Best practices for convolutional neural networks applied to visual document analysis”. In: *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings*. Vol. 1. ICDAR-03. IEEE Comput. Soc, pp. 958–963. DOI: [10.1109/icdar.2003.1227801](https://doi.org/10.1109/icdar.2003.1227801).
- Song, Tao et al. (Aug. 2022). “DeepFusion: A deep learning based multi-scale feature fusion method for predicting drug-target interactions”. In: *Methods* 204, pp. 269–277. ISSN: 1046-2023. DOI: [10.1016/j.ymeth.2022.02.007](https://doi.org/10.1016/j.ymeth.2022.02.007).
- Spearman, C. (Jan. 1904). “The Proof and Measurement of Association between Two Things”. In: *The American Journal of Psychology* 15.1, p. 72. ISSN: 0002-9556. DOI: [10.2307/1412159](https://doi.org/10.2307/1412159).
- Steinmann, Luca et al. (Apr. 2021). “Variational Autoencoder based Novelty Detection for Real-World Time Series”. In: *2021 3rd International Conference on Management Science and Industrial Engineering*. MSIE 2021. ACM, pp. 1–7. DOI: [10.1145/3460824.3460825](https://doi.org/10.1145/3460824.3460825).
- Suh, Suwon and Seungjin Choi (2016). *Gaussian Copula Variational Autoencoders for Mixed Data*. DOI: [10.48550/ARXIV.1604.04960](https://doi.org/10.48550/ARXIV.1604.04960).
- Székely, Gábor J (2003). “E-statistics: The energy of statistical samples”. In: *Bowling Green State University, Department of Mathematics and Statistics Technical Report* 3.05, pp. 1–18. DOI: [10.13140/RG.2.1.5063.9761](https://doi.org/10.13140/RG.2.1.5063.9761).
- Taha, Altyeb Altaher and Sharaf Jameel Malebary (2020). “An Intelligent Approach to Credit Card Fraud Detection Using an Optimized Light Gradient Boosting Machine”. In: *IEEE Access* 8, pp. 25579–25587. ISSN: 2169-3536. DOI: [10.1109/access.2020.2971354](https://doi.org/10.1109/access.2020.2971354).

- Tingfei, Huang, Cheng Guangquan, and Huang Kuihua (2020). “Using Variational Auto Encoding in Credit Card Fraud Detection”. In: *IEEE Access* 8, pp. 149841–149853. ISSN: 2169-3536. DOI: [10.1109/access.2020.3015600](https://doi.org/10.1109/access.2020.3015600).
- Tishby, Naftali, Fernando C. Pereira, and William Bialek (2000). *The information bottleneck method*. DOI: [10.48550/ARXIV.PHYSICS/0004057](https://arxiv.org/abs/10.48550/ARXIV.PHYSICS/0004057).
- Tschantz, Michael Carl (2022). “What is Proxy Discrimination?”. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1993–2003. URL: <https://arxiv.org/abs/2205.05265>.
- Villani, Cédric (2009). *Optimal Transport*. Springer Berlin Heidelberg. DOI: [10.1007/978-3-540-71050-9](https://doi.org/10.1007/978-3-540-71050-9).
- Wang, Bo and Jian Qing Shi (July 2014). “Generalized Gaussian Process Regression Model for Non-Gaussian Functional Data”. In: *Journal of the American Statistical Association* 109.507, pp. 1123–1133. ISSN: 1537-274X. DOI: [10.1080/01621459.2014.889021](https://doi.org/10.1080/01621459.2014.889021).
- Wang, Wenlin et al. (2019). “Topic-Guided Variational Auto-Encoder for Text Generation”. In: *Proceedings of the 2019 Conference of the North. Association for Computational Linguistics*. DOI: [10.18653/v1/n19-1015](https://arxiv.org/abs/10.18653/v1/n19-1015).
- Wang, Yi (2011). *Distributed gibbs sampling of latent dirichlet allocation: The gritty details*. URL: <https://api.semanticscholar.org/CorpusID:6015396>.
- Wang, Yibo and Wei Xu (Jan. 2018). “Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud”. In: *Decision Support Systems* 105, pp. 87–95. ISSN: 0167-9236. DOI: [10.1016/j.dss.2017.11.001](https://doi.org/10.1016/j.dss.2017.11.001).
- Wegenmann, Benjamin et al. (2022). “DP-VAE: Human-readable text anonymization for online reviews with differentially private variational autoencoders”. In: *Proceedings of the ACM Web Conference 2022*, pp. 721–731. DOI: [10.1145/3485447.351223](https://doi.org/10.1145/3485447.351223).
- Wei, Jason and Kai Zou (2019). “EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, pp. 6381–6387. DOI: [10.18653/v1/d19-1670](https://arxiv.org/abs/10.18653/v1/d19-1670).
- Wei, Min, Tommy Chow, and Rosa Chan (Mar. 2015). “Clustering Heterogeneous Data with k-Means by Mutual Information-Based Unsupervised Feature Transformation”. In: *Entropy* 17.3, pp. 1535–1548. ISSN: 1099-4300. DOI: [10.3390/e17031535](https://doi.org/10.3390/e17031535).
- Weiss, Y. (1999). “Segmentation using eigenvectors: a unifying view”. In: *Proceedings of the Seventh IEEE International Conference on Computer Vision*. IEEE, 975–982 vol.2. DOI: [10.1109/iccv.1999.790354](https://doi.org/10.1109/iccv.1999.790354).
- Williams, Christopher KI and Carl Edward Rasmussen (2006). *Gaussian processes for machine learning*. Vol. 2. 3. MIT press Cambridge, MA.
- Williams, Graham J. and Zhexue Huang (1997). “Mining the knowledge mine: The hot spots methodology for mining large real world databases”. In: *Advanced*

- Topics in Artificial Intelligence*. Springer Berlin Heidelberg, pp. 340–348. ISBN: 9783540696490. DOI: [10.1007/3-540-63797-4_87](https://doi.org/10.1007/3-540-63797-4_87).
- Wilsens, Paul, Katrien Antonio, and Gerda Claeskens (Nov. 2024). “Reducing the dimensionality and granularity in hierarchical categorical variables”. In: *Advances in Data Analysis and Classification* 19.4, pp. 1087–1118. ISSN: 1862-5355. DOI: [10.1007/s11634-024-00614-5](https://doi.org/10.1007/s11634-024-00614-5).
- Wu, Chuhan et al. (Feb. 2019). “Semi-supervised dimensional sentiment analysis with variational autoencoder”. In: *Knowledge-Based Systems* 165, pp. 30–39. ISSN: 0950-7051. DOI: [10.1016/j.knosys.2018.11.018](https://doi.org/10.1016/j.knosys.2018.11.018).
- Wuthrich, Mario V. and Michael Merz (2021). “Statistical Foundations of Actuarial Learning and its Applications”. In: *SSRN Electronic Journal*. ISSN: 1556-5068. DOI: [10.2139/ssrn.3822407](https://doi.org/10.2139/ssrn.3822407).
- Xie, Qizhe et al. (2020). “Unsupervised data augmentation for consistency training”. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS ’20. Vancouver, BC, Canada: Curran Associates Inc. ISBN: 9781713829546. URL: <https://dl.acm.org/doi/10.5555/3495724.3496249>.
- Xin, Xi and Fei Huang (2023). “Antidiscrimination Insurance Pricing: Regulations, Fairness Criteria, and Models”. In: *North American Actuarial Journal* 28.2, pp. 285–319. ISSN: 2325-0453. DOI: [10.1080/10920277.2023.2190528](https://doi.org/10.1080/10920277.2023.2190528).
- Xiong, Tengke et al. (May 2011). “DHCC: Divisive hierarchical clustering of categorical data”. In: *Data Mining and Knowledge Discovery* 24.1, pp. 103–135. ISSN: 1573-756X. DOI: [10.1007/s10618-011-0221-2](https://doi.org/10.1007/s10618-011-0221-2).
- Xu, Lei, Maria Skoularidou, et al. (2019). “Modeling tabular data using conditional GAN”. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc. URL: <https://dl.acm.org/doi/10.5555/3454287.3454946>.
- Xu, Lei and Kalyan Veeramachaneni (2018). *Synthesizing Tabular Data using Generative Adversarial Networks*. DOI: [10.48550/ARXIV.1811.11264](https://doi.org/10.48550/ARXIV.1811.11264).
- Xu, Weidi et al. (2017). “Variational autoencoder for semi-supervised text classification”. In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI’17. San Francisco, California, USA: AAAI Press, pp. 3358–3364. URL: <https://dl.acm.org/doi/10.5555/3298023.3298056>.
- Xuan, Junyu et al. (2014). “Release ‘Bag-of-Words’ Assumption of Latent Dirichlet Allocation”. In: *Foundations of Intelligent Systems*. Springer Berlin Heidelberg, pp. 83–92. ISBN: 9783642549243. DOI: [10.1007/978-3-642-54924-3_8](https://doi.org/10.1007/978-3-642-54924-3_8).
- Yang, M.-S. (Dec. 1993). “A survey of fuzzy clustering”. In: *Mathematical and Computer Modelling* 18.11, pp. 1–16. ISSN: 0895-7177. DOI: [10.1016/0895-7177\(93\)90202-a](https://doi.org/10.1016/0895-7177(93)90202-a).
- Yang, Yi, Wei Qian, and Hui Zou (May 2017). “Insurance Premium Prediction via Gradient Tree-Boosted Tweedie Compound Poisson Models”. In: *Journal of Business and Economic Statistics* 36.3, pp. 456–470. ISSN: 1537-2707. DOI: [10.1080/07350015.2016.1200981](https://doi.org/10.1080/07350015.2016.1200981).

- Yao, Ji (July 2016). “Clustering in General Insurance Pricing”. In: *Predictive Modeling Applications in Actuarial Science*. Cambridge University Press, pp. 159–179. DOI: [10.1017/cbo9781139342681.007](https://doi.org/10.1017/cbo9781139342681.007).
- Yin, Shuang et al. (Mar. 2021). “Applications of Clustering with Mixed Type Data in Life Insurance”. In: *Risks* 9.3, p. 47. ISSN: 2227-9091. DOI: [10.3390/risks9030047](https://doi.org/10.3390/risks9030047).
- Zamany, Siamak et al. (Aug. 2022). “Towards Deeper Understanding of Variational Auto-encoders for Binary Collaborative Filtering”. In: *Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval*. IC-TIR ’22. ACM, pp. 254–263. DOI: [10.1145/3539813.3545145](https://doi.org/10.1145/3539813.3545145).
- Zhang, Chi et al. (Sept. 2020). “Stock-Index Tracking Optimization Using Auto-Encoders”. In: *Frontiers in Physics* 8. ISSN: 2296-424X. DOI: [10.3389/fphy.2020.00388](https://doi.org/10.3389/fphy.2020.00388).
- Zhang, Tianhe, Suhan Liu, and Peng Shi (2025). *Discrimination-free Insurance Pricing with Privatized Sensitive Attributes*. DOI: [10.48550/ARXIV.2504.11775](https://doi.org/10.48550/ARXIV.2504.11775).
- Zhang, Xinyuan et al. (2019). “Syntax-Infused Variational Autoencoder for Text Generation”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. DOI: [10.18653/v1/p19-1199](https://doi.org/10.18653/v1/p19-1199).
- Zhuang, Kai, Sen Wu, and Xiaonan Gao (2018). “Auto insurance business analytics approach for customer segmentation using multiple mixed-type data clustering algorithms”. In: *Tehnički vjesnik* 25.6, pp. 1783–1791. DOI: [10.17559/tv-20180720122815](https://doi.org/10.17559/tv-20180720122815).