LATENT DIRICHLET ALLOCATION FOR STRUCTURED INSURANCE DATA

Charlotte Jamotton, Donatien Hainaut

LIDAM Discussion Paper ISBA
2024 / 08

UCLouvain

ISBA | LIDAM
Louvain Institute of Data Analysis and Modeling in economics and statistics

ISBA

Voie du Roman Pays 20 - L1.04.01

B-1348 Louvain-la-Neuve

Email : lidam-library@uclouvain.be

https://uclouvain.be/en/research-institutes/lidam/isba/publication.html

Latent Dirichlet Allocation for structured insurance data

Charlotte Jamotton* and Donatien Hainaut†

Université catholique de Louvain (UCLouvain)
Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA)
Louvain Institute of Data Analysis and Modeling (LIDAM)
Louvain-la-Neuve, Belgium

Abstract

This article explores the application of Latent Dirichlet Allocation (LDA) to structured tabular insurance data. LDA is a probabilistic topic modelling 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 mixture of latent (hidden) topics, and topics as mixtures of words. This study introduces the LDA's document-topic distribution as a soft clustering tool for unsupervised learning tasks in the actuarial field. By defining each topic as a risk profile, and by treating insurance policies as documents and the modalities of categorical covariates as words, we show how LDA can be extended beyond textual data and can offer a framework to uncover underlying structures within insurance portfolios. 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.

Keywords: latent dirichlet allocation, topic modelling, soft clustering, insurance data, risk profile, natural language processing

1. Introduction

The Latent Dirichlet Allocation (LDA) is a probabilistic topic modelling approach introduced by Blei, Ng, and Jordan (2003) as an extension to Hofmann (1999) probabilistic Latent Semantic Analysis (pLSA). It is widely employed in unsupervised Natural Language Processing (NLP) for topic modelling and document classification. Since machine learning models require numerical input, textual documents must first be represented as vectors of identifiers, i.e., in

*charlotte.jamotton@uclouvain.be
†donatien.hainaut@uclouvain.be

a machine processable form. In this vector space model (VSM) and under the Bag-of-Words (BoW) assumptions, the sequential arrangement of words within a document, as well as the order of documents in 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, Ng, and Jordan 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.

In actuarial sciences, we often work with insurance portfolios in a tabular format (i.e., rows representing different policies and columns representing different covariates) that may contain both structured and unstructured data. Structured data include numeric and categorical data, whereas unstructured data usually refer to the claims textual description. Wang and Xu (2018) demonstrated that applying LDA to (unstructured) insurance text data improved the predictive accuracy of frauds for automobile insurance claims. By combining the LDA's topical numeric features with the original numerical and one-hot encoded categorical covariates, they managed to boost the predictive accuracy of their supervised neural network. In their survey on categorical data processing, Hancock and Khoshgoftaar (2020) recognised the research potential of applying Wang and Xu (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 discretized numeric rating factors.

Categorical variables are pervasive in insurance datasets where policies are typically characterised by the policyholder's geographic location, gender, or age group, to name a few. The qualitative (non-numeric) and diverse nature of categorical variables (ordinal, nominal, and/or with high cardinality) makes their transformation to a machine processable form challenging. One-hot encoding remains the go-to pre-processing procedure when dealing with categorical data (see e.g., Jose et al. (2024)), although embedding has gained some traction in recent years (see Richman (2021) for a review of its application within actuarial science).

Like embedding, LDA originates from Natural Language Processing. Using the terminology specific to the LDA model, each policy $\mathbf{x}$ characterised by N_d categorical (and discretized numeric) variables can be viewed as a set of words or modalities $\{x_1, x_2, \ldots, x_{N_d}\}$. An insurance portfolio (corpus) X could thus be interpreted as a collection of D policies (documents) $\{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_D}\}$ depicting K different risk profiles (topics) denoted $\mathbf{z_1}, \mathbf{z_2}, \ldots, \mathbf{z_K}$. 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 $\boldsymbol{\beta}_k$ and document-topic $\boldsymbol{\theta}_d$ distributions could be respectively renamed risk profile-modality and policy-risk profile distributions. The BoW's exchangeability still holds for both policies (documents) and modalities (words within a document); 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).

The policy-topic $\boldsymbol{\theta}_d$ distribution could theoretically be used as a data mining tool 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, say young drivers of high-power vehicle in the riskiest profile. This is nothing else than the definition of cluster analysis (Jain, Murty, and Flynn 1999). In one respect, LDA is to textual documents what soft clustering is to numeric observations. In the actuarial literature, clustering has already 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 (2020), 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 policies attributes that may overlap with different groups. Hence the interest of having LDA's probability distribution θ_d over K risk profiles, given a policy d. This amounts to assign to each policy topic cluster membership weights, which is consistent with the definition of fuzzy (or soft) clustering (Yang 1993).

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 could turn into an interesting prospect to model policies containing missing values. While incomplete data analysis has already been addressed in the literature by Ng et al. (2008), who introduced grouped Dirichlet distributions to analyse incomplete categorical data, or Dinh, Huynh, and Sriboonchitta (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.

Our article contributes to the actuarial literature by introducing the application of LDA to categorical (instead of textual) data and by introducing LDA document-topic distribution as an unsupervised soft clustering tool. To the best of our knowledge, Wang and Xu (2018) paper titled *Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud* is the only LDA case study on (textual) insurance data.

The article is organised as follows. Section 2.1 introduces the LDA generative process and motivates its use in the context of categorical insurance data. Section 2.2 details the variational inference developed by Hoffman, Bach, and Blei (2010) to derive the LDA model's approximate posterior distribution. Section 3 presents our experiment settings and results with a comparison to the Burt-adapted fuzzy clustering (Jamotton, Hainaut, and Hames 2023). Section 4 gives our conclusions.

2. Methods

Let 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.):

$$X = \{\mathbf{x}_1, \ldots, \mathbf{x}_D\} \ . \tag{1}$$

The 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 across all categorical variables.

Each policy (document) $\mathbf{x}$ in the insurance portfolio (corpus) can be represented as a finite sequence of N_d words or modalities (e.g., female, west, grey, etc.):

$$\mathbf{x} = \left\{x_1, \ldots, x_{N_d}\right\}, \tag{2}$$

and will be modelled as a mixture $\theta^{(1\times K)}$ over the K latent risk profiles. We will denote $x_{d,n}$ the nth modality of the dth policy and θ_d the topic distribution of the dth policy. Similarly, $z_{d,n}$ will denote the latent risk profile (topic) assigned to the nth modality of the dth policy. Each risk profile $k = 1, \ldots, K$ will be described by a mixture $\beta_k^{(1\times V)}$ over all V modalities (words). 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).

2.1 Generative Process

The LDA generative process, illustrated in Figure 1, works as follows. For K prespecified risk profiles (topics), LDA assumes the dth policy $\mathbf{x}_d \in X$ with length N_d is generated in the following steps (Wang 2011; Carpenter 2010; Darling 2011):

1. For $k = 1, \ldots, K$, $\beta_k \sim \mathcal{D}_V(\eta)$
2. For $d = 1, \ldots, D$
   (a) $\theta_d \sim \mathcal{D}_K(\alpha)$
   (b) for each $x_{d,n}$
      i. $z_{d,n} \sim \mathcal{M}_K(1, \theta_d)$
      ii. $x_{d,n}|z_{d,n} \sim \mathcal{M}_V\left(1, \beta_{z_{d,n}}\right)$

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 hyper-parameter 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, Fankhauser, and Nejdl 2009). Fukuyama, Sankaran, and Symul (2021) 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 hyper-parameter α. Each element $\theta_{d,k}$ 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 $\theta = (\theta_1, \ldots, \theta_D) \in \mathbb{R}^{D\times K}$ 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 $z_{d,n} \in \{1, \ldots, K\}$ is then sampled for each modality $x_{d,n},\ n = 1, \ldots, N_d$ in the dth 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 $x_{d,n}$ is sampled given the sampled risk profile $z_{d,n}$ from a V-dimensional Multinomial distribution $\mathcal{M}_V\left(1, \beta_{z_{d,n}}\right)$ parameterised by $\beta_{z_{d,n}}$. This Multinomial sampling could be challenged in the context of insurance data as modalities within a categorical variable are mutually exclusive. In theory, each modality $x_{d,n}$ should be sampled from a subset $V^{(n)}$ of the vocabulary V that should only contain the unique modalities of the corresponding categorical variable:

$$x_{d,n}|z_{d,n} \sim \mathcal{M}_{V^{(n)}}\left(1, \beta_{z_{d,n}}^{V^{(n)}}\right) . \tag{3}$$

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.

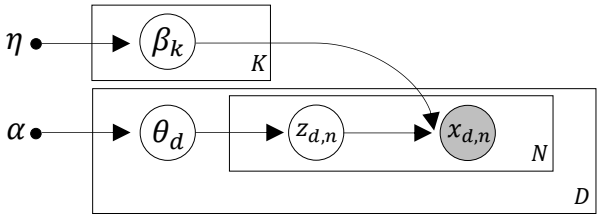

Figure 1. Representation of the smoothed LDA generative process. The top rectangle defines the topic-level parameters, the inner rectangle defines the modality-level parameters, and the outer rectangle defines the policy-level parameters. We can only observe x (filled with grey).

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

$$p\left(x, z, \theta, \beta|\alpha, \eta\right) = \prod_{k=1}^{K} p\left(\beta_k|\eta\right) \prod_{d=1}^{D} p\left(\theta_d|\alpha\right) \prod_{n=1}^{N_d} p\left(x_{d,n}|\theta_d, \beta\right) , \tag{4}$$

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 $p\left(z_{d,n}|\theta_d\right)$ and components $\mathrm{p}\big(x_{d,n}|z_{d,n}, \beta\big)$:

$$p\left(x_{d,n}|\theta_d, \beta\right) = \sum_{z} p\left(x_{d,n}|z_{d,n}, \beta\right) p\left(z_{d,n}|\theta_d\right) , \tag{5}$$

such that the joint distribution in Equation 4 becomes:

$$p\left(x, z, \theta, \beta|\alpha, \eta\right) = \prod_{k=1}^{K} p\left(\beta_k|\eta\right) \prod_{d=1}^{D} p\left(\theta_d|\alpha\right) \prod_{n=1}^{N_d} p\left(x_{d,n}|z_{d,n}, \beta\right) p\left(z_{d,n}|\theta_d\right) , \tag{6}$$

where β_k and θ_d are sampled using Dirichlet probability distributions:

$$p\left(\theta_d|\alpha\right) = \mathcal{D}\left(\theta_d|\alpha\right) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma\left(\alpha_k\right)} \prod_{k=1}^{K} \theta_{d,k}^{\alpha_k - 1} , \tag{7}$$

$$p\left(\beta_k|\eta\right) = \mathcal{D}\left(\beta_k|\eta\right) = \frac{\Gamma\left(\sum_{v=1}^{V}\eta_v\right)}{\prod_{v=1}^{V}\Gamma\left(\eta_v\right)}\prod_{v=1}^{V}\beta_{k,v}^{\eta_v-1}\,, \tag{8}$$

and $z_{d,n}$ and $x_{d,n}|z_{d,n}$ are sampled using Multinomial probability distributions:

$$p\left(x_{d,n}|z_{d,n},\beta_{1:K}\right) = \mathcal{M}\left(x_{d,n}|z_{d,n},\beta_{1:K}\right) = \beta_{z_{d,n},x_{d,n}}\,, \tag{9}$$

$$p\left(z_{d,n}|\theta_d\right) = \mathcal{M}\left(z_{d,n}|\theta_d\right) = \theta_{d,z_{d,n}}\,. \tag{10}$$

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 $p\left(z,\theta,\beta|x,\alpha,\eta\right)$ of the latent variables z_d, θ_d, and β_k given the evidence (i.e., the observed data x) and the Dirichlet hyper-parameters α,η. This amounts to solving the following equation (Darling 2011):

$$p\left(z,\theta,\beta|x,\alpha,\eta\right) = \frac{p\left(x,z,\theta,\beta|\alpha,\eta\right)}{\int_{\theta,\beta}\sum_z p\left(x,z,\theta,\beta|\alpha,\eta\right)}\,. \tag{11}$$

The numerator of this conditional Equation 11 can be expended given Equations 7,8, 9, and 10.

$$\begin{aligned} p\left(x,z,\theta,\beta|\alpha,\eta\right) &= \prod_{k=1}^{K}p\left(\beta_k|\eta\right)\prod_{d=1}^{D}p\left(\theta_d|\alpha\right)\prod_{n=1}^{N_d}p\left(x_{d,n}|z_{d,n},\beta_{1:K}\right)p\left(z_{d,n}|\theta_d\right) \\ &= c\prod_{k=1}^{K}\left(\prod_{v=1}^{V}\beta_{k,v}^{\eta_v-1}\right)\prod_{d=1}^{D}\left(\prod_{k=1}^{K}\theta_{d,k}^{\alpha_k-1}\right)\prod_{n=1}^{N_d}\prod_{k=1}^{K}\prod_{v=1}^{V}\left(\beta_{k,v}\theta_{d,k}\right)^{a}\,, \end{aligned} \tag{12}$$

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 $a = \mathbb{I}_{\{x_{d,n}=v,z_{d,n}=k\}}$ is an indicator function.

The denominator of the conditional Equation 11 is intractable due to the coupling of the latent variables θ and β in the summation over topics in the marginal likelihood:

$$\int_\beta\int_\theta\sum_z p\left(x,z,\theta,\beta|\alpha,\eta\right) = c\int_\beta\prod_{v=1}^{V}\beta_{k,v}^{\eta_v-1}\int_\theta\prod_{k=1}^{K}\theta_{d,k}^{\alpha_k-1}\prod_{n=1}^{N_d}\prod_{k=1}^{K}\prod_{v=1}^{V}\left(\beta_{k,v}\theta_{d,k}\right)^{a}\,. \tag{13}$$

There are two families of techniques to derive a (tractable) approximate posterior distribution: the Markov Chain Monte Carlo (MCMC) sampling approaches and the variational Bayes (VB) optimisation approaches (Wang 2011; Carpenter 2010; Darling 2011). Although the choice of a simplified parametric distribution introduce 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, Bach, and Blei 2010). To efficiently fit models to larger collections of documents, Hoffman, Bach, and Blei (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.

2.2 Variational posterior and ELBO

A variational posterior distribution $q(z, \theta, \beta|\phi, \gamma, \lambda)$, illustrated in Figure 2, is introduced to approximate the unknown distribution $p(z, \theta, \beta|x, \alpha, \eta)$. Given 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:

$$q(z, \theta, \beta|\gamma, \lambda) = \prod_{k=1}^{K} q(\beta_k|\lambda_k) \prod_{d=1}^{D} q(\theta_d|\gamma_d) \prod_{n=1}^{N_d} q\left(z_{d,n}|\phi_{d,n}\right) , \tag{14}$$

parameterised by three variational parameters:

- $\overrightarrow{\lambda_k}$ of length V for topic $k = 1, \ldots, K$;
- non-negative (and does not sum to 1) $\overrightarrow{\gamma_d}$ of length K for policy $d = 1, \ldots, D$;
- non-negative ϕ of size $D \times V \times K$ with $\sum_{k=1}^{K} \phi_{dvk} = 1, \forall d, v$.

where β_k and θ_d are sampled using Dirichlet probability distributions:

$$p(\theta_d|\gamma_d) = \mathcal{D}(\theta_d|\gamma_d) = \frac{\Gamma\left(\sum_{k=1}^{K} \gamma_{d,k}\right)}{\prod_{k=1}^{K} \Gamma\left(\gamma_{d,k}\right)} \prod_{k=1}^{K} \theta_{d,k}^{\gamma_{d,k}-1} , \tag{15}$$

$$q(\beta_k|\lambda_k) = \mathcal{D}(\beta_k|\lambda_k) = \frac{\Gamma\left(\sum_{v=1}^{V} \lambda_{k,v}\right)}{\prod_{v=1}^{V} \Gamma\left(\lambda_{k,v}\right)} \prod_{v=1}^{V} \beta_{k,v}^{\lambda_{k,v}-1} , \tag{16}$$

and $z_{d,n}$ is sampled using a Multinomial probability distribution:

$$q\left(z_{d,n} = k\right) = \mathcal{M}_K\left(\overrightarrow{\phi}_{d,n}\right) = \phi_{dx_{d,n}k} . \tag{17}$$

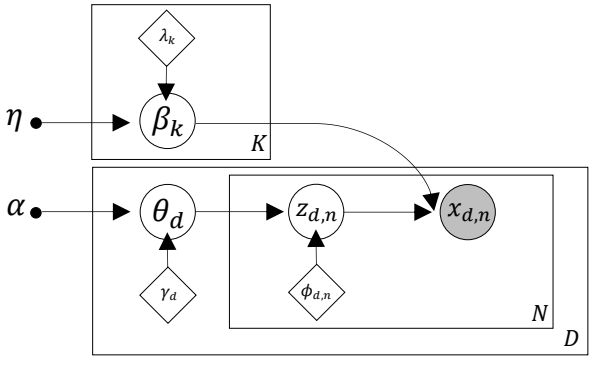

Figure 2. Representation of the variational distribution to approximate the posterior in the smoothed LDA model. We can only observe x (filled with grey). The variational parameters are framed with lozenge-shapes.

The goal of VB is to optimise the Kullback-Leibler (KL) divergence between the variational approximation q and the true posterior distribution p:

$$\min_{q} \left\{ KL\left(q(z, \theta, \beta|\phi, \gamma, \lambda) \;||\; p(z, \theta, \beta|x, \alpha, \eta)\right) = \mathbb{E}_q\left[\log \frac{q(z, \theta, \beta|\phi, \gamma, \lambda)}{p(z, \theta, \beta|x, \alpha, \eta)}\right] \right\} . \tag{18}$$

Minimising this KL divergence is equivalent to maximising the lower bound $\mathcal{L}(\gamma, \phi; \alpha, \beta)$ on the log-likelihood. This tractable evidence lower bound $\mathcal{L}$ is derived on the marginal log-likelihood $\ln p(x|\alpha, \eta)$ using variational inference and Jensen's inequality. The marginal log-likelihood is defined as:

$$\ln p(x|\alpha, \eta) = \ln \int \sum_z p(x, z, \theta, \beta|\alpha, \eta)\, d\theta\,, \tag{19}$$

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 Equation 14 and obtain the lower bound from Jensen's inequality:

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

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

$$\mathcal{L}(\phi, \gamma, \lambda; \alpha, \eta) = -KL(q(z, \theta, \beta|\phi, \gamma, \lambda)\,||p(z, \theta, \beta|x, \alpha, \eta)) + \ln p(x \mid \alpha, \eta) \tag{20}$$

The variational form of the marginal likelihood in Equation 19 is then equal to:

$$\ln p(x \mid \alpha, \eta) = \mathcal{L}(\phi, \gamma, \lambda; \alpha, \eta) + KL(q(z, \theta, \beta|\phi, \gamma, \lambda)\,||p(z, \theta, \beta|x, \alpha, \eta))\,. \tag{21}$$

To facilitate calculations (and coding), the tractable lower bound $\mathcal{L}(\phi, \gamma, \lambda; \alpha, \eta)$ in Equation 2.2 is reorganised using the complete model in Equation 6 and the variational distribution in Equation 14:

$$\begin{aligned}
\mathcal{L}(x; \phi, \gamma, \lambda; \alpha, \eta) = &\mathbb{E}_q [\ln p(\beta|\eta)] + \mathbb{E}_q [\ln p(\theta|\alpha)] + \mathbb{E}_q [\ln p(x|z, \beta)] + \mathbb{E}_q [\ln p(z|\theta)] \\
&- \mathbb{E}_q [\ln q(\beta|\lambda)] - \mathbb{E}_q [\ln q(\theta|\gamma)] - \mathbb{E}_q [\ln q(z|\phi)]\,.
\end{aligned} \tag{22}$$

We now expand the expectations above to get the complete objective function (ELBO) as functions of the variational parameters. In Hoffman, Bach, and Blei (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:

$$\begin{aligned}
\mathcal{L} = &\sum_{d=1}^{D} \left\{ \sum_{v=1}^{V} n_{d,v} \sum_{k=1}^{K} \phi_{d,v,k} \left(\mathbb{E}_q \left[\ln \theta_{d,k}\right] + \mathbb{E}_q \left[\ln \beta_{k,v}\right] - \ln \phi_{d,v,k}\right) \right\} \\
&+ \sum_{d=1}^{D} \left\{ \sum_{k=1}^{K} a\left(\gamma_{d,k}\right) + \left(\alpha - \gamma_{d,k}\right) \mathbb{E}_q \left[\ln \theta_{d,k}\right] \right\} \\
&+ \sum_{d=1}^{D} \left\{ \sum_{k=1}^{K} \left(\sum_{v=1}^{V} b\left(\lambda_{k,v}\right) + \left(\eta - \lambda_v\right) \mathbb{E}_q \left[\ln \beta_{k,v}\right] \right) / D \right\} \\
&+ \sum_{d=1}^{D} \left\{ \left(\ln \Gamma(K\alpha) - K \ln \Gamma(\alpha)\right) + K \left(\ln \Gamma(V\eta) - V \ln \Gamma(\eta)\right) / D \right\}\,,
\end{aligned} \tag{23}$$

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

How 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 constraint maximisation since $\sum_k \phi_{dvk} = 1\,\forall d, v$. We form the Lagrangian by isolating the terms which contain ϕ_{dvk} and by adding the appropriate Lagrange multipliers:

$$\begin{aligned}
\mathcal{L}_{[\phi_{dvk}]} &= \phi_{dvk}\left(\mathbb{E}_q\left[\ln\beta_{k,v}\right] + \mathbb{E}_q\left[\ln\theta_{d,k}\right] - \ln\phi_{dvk}\right) + \lambda_k^{Lagrange}\left(\sum_{k=1}^{K}\phi_{dvk} - 1\right) \\
\frac{\partial\mathcal{L}}{\partial\phi_{dvk}} &= \mathbb{E}_q\left[\ln\beta_{k,v}\right] + \mathbb{E}_q\left[\ln\theta_{d,k}\right] - \left(\ln\phi_{d,v,k} + 1\right) + \lambda_k^{Lagrange}\,.
\end{aligned} \tag{24}$$

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

$$\phi_{dvk} \propto \exp\left(\mathbb{E}_q\left[\ln\beta_{k,v}\right] + \mathbb{E}_q\left[\ln\theta_{d,k}\right]\right)\,. \tag{25}$$

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

$$\begin{aligned}
\mathcal{L}_{[\gamma_{dk}]} &= \left(\sum_{v=1}^{V} n_{d,v}\phi_{d,v,k} + \left(\alpha - \gamma_{d,k}\right)\right)\left(\Psi\left(\gamma_{dk}\right) - \Psi\left(\sum_{k=1}^{K}\gamma_{dk}\right)\right) + a\left(\gamma_{dk}\right) + \text{cst} \\
\frac{\partial\mathcal{L}}{\partial\gamma_{dk}} &= \left(\Psi'\left(\gamma_{dk}\right) - \Psi'\left(\sum_{k=1}^{K}\gamma_{dk}\right)\right)\left(\sum_{v=1}^{V} n_{d,v}\phi_{d,v,k} + \left(\alpha - \gamma_{d,k}\right)\right)\,.
\end{aligned} \tag{26}$$

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

$$\gamma_{d,k} = \alpha + \sum_{v=1}^{V} n_{d,v}\phi_{d,v,k}\,. \tag{27}$$

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}:

$$\begin{aligned}
\mathcal{L}_{[\lambda_{kv}]} &= \left(\sum_{d=1}^{D} n_{d,v}\phi_{d,v,k} + \left(\eta - \lambda_{kv}\right)\right)\left(\Psi\left(\lambda_{kv}\right) - \Psi\left(\sum_{v=1}^{V}\lambda_{kv}\right)\right) + b\left(\lambda_{k,v}\right) + \text{cst} \\
\frac{\partial\mathcal{L}}{\partial\lambda_{kv}} &= \left(\sum_{d=1}^{D} n_{d,v}\phi_{d,v,k} + \left(\eta - \lambda_{kv}\right)\right)\left(\Psi'\left(\lambda_{kv}\right) - \Psi'\left(\sum_{v=1}^{V}\lambda_{kv}\right)\right)\,.
\end{aligned} \tag{28}$$

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

$$\lambda_{kv} = \eta + \sum_{d=1}^{D} n_{d,v}\phi_{d,v,k}\,. \tag{29}$$

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 $\phi_{d,v,k}$ 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 Equations 29, 27, and 25), the online variational Bayes EM steps (Hoffman, Bach, and Blei 2010) for solving LDA are detailed in the following Algorithm 1.

Algorithm 1: Online variational Bayes for LDA (Hoffman, Bach, and Blei 2010).

Define iteration weight $\rho_t \triangleq (\tau_0 + t)^{-\kappa}$ where $\tau_0 > 1$ is the learning offset, t is the iteration number of the EM step, and $\kappa \in (0.5, 1.0]$ is the learning decay.
Define convergence condition $K^{-1} \sum_{k=1}^{K} |\text{change in } \gamma_{tk}| < 1e-5$.
Randomly initialise variational parameter $\boldsymbol{\lambda}$.
while *not converged* **do**
    `// E-step`
    Initialize γ_{tk} to an arbitrary constant (e.g., $\gamma_{tk} = 1$);
    repeat
        Compute $\phi_{tvk} \propto \exp\left(\mathbb{E}_q[\ln \beta_{kv}] + \mathbb{E}_q[\ln \theta_{tk}]\right)$;
        Set $\gamma_{tk} = \alpha + \sum_{v=1}^{V} n_{tv} \phi_{tvk}$;
    until *convergence*;
    `// M-step`
    Compute $\tilde{\lambda}_{kv} = \eta + D n_{tv} \phi_{tvk}$;
    Set $\boldsymbol{\lambda} = (1 - \rho_t)\boldsymbol{\lambda} + \rho_t \tilde{\boldsymbol{\lambda}}$;
end
Output : β, θ
Compute $\mathcal{L}$ for diagnostics.

3. Results and Discussion

We use two distinct insurance portfolios to evaluate the validity and consistency of the LDA model across various types of covariates (i.e., (positive) numerical, binary, Multinomial, and ordinal) with a variety of marginal distributions (e.g., multi-modal, long tail) and different assumptions regarding the target variable's distribution (Poisson or Binomial). The first dataset is a loan default portfolio that contains both structured and unstructured data. Structured data include numeric and categorical data, whereas unstructured data refer to the claims text descriptions. The three types of data are preprocessed and cleaned according to their respective characteristics. Table 1 provides a description of the data in the loan portfolio which consists of 25 917 policies. Each policy is described by six categorical variables, two binary variables, four continuous variables, and a loan description provided by the borrower. The target variable takes two possible values (Charged Off or Fully Paid) and is assumed to follow a Binomial distribution. In the following analyses, each continuous variable is discretized into four distinct groups based on its 25th, 50th, and 75th percentile.

Table 1. Descriptions of the loan portfolio attributes.

Categorical variables	**Description**	**Modalities**	**Proportions**		
Inquiries	The number of inquiries in the past 6 months.	0	0.48%		
		1	0.27%		
		2	0.15%		
		≥ 3	0.10%		
Region	The region of the state provided by the borrower in the loan application.	East	0.51%		
		West	0.29%		
		Central	0.20%		
Purpose	A category provided by the borrower for the loan request.	Debt consolidation	0.47%		
		Other	0.19%		
		Credit card	0.13%		
		Home improvement	0.07%		
		Major purchase	0.05%		
		Small business	0.05%		
Grade	LC assigned loan subgrade.	1-10	0.56%		
		11-20	0.34%		
		21-35	0.10%		
Home ownership	The home ownership status provided by the borrower during registration.	Rent	0.48%		
		Mortgage	0.44%		
		Own	0.08%		
Verification status	Indicates if the income (source) was verified by LC.	Not verified	0.45%		
		Verified	0.32%		
		Source verified	0.23%		
Term	The number of payments on the loan.	36 months	0.76%		
		60 months	0.24%		
Public records	Number of derogatory public records.	0	0.95%		
		1, 2 or 3	0.05%		
Target variable					
Loan status	Current status of the loan.	Fully Paid	0.85%		
		Charged Off	0.15%		
Continuous variables		**min**	**max**	**mean**	**std**
Loan amount	The listed amount of the loan applied for by the borrower.	500	35 000	11 369.9	7 274.08
Interest rate	Interest rate on the loan.	5.42	24.4	11.95	3.6
Annual income	The self-reported annual income provided by the borrower during registration.	400	3 900 000	68 854	59 173
Revolving line utilization rate	The amount of credit the borrower is using relative to all available revolving credit.	0	99.9	48.49	28.22
Claim description	Loan description provided by the borrower.				

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 and Johansson 2010). Table 2 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, $20-29$, $30-39$, $\ldots$, $80-89$, ≥ 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.

Table 2. Descriptions of the motorcycle portfolio attributes (Ohlsson and Johansson 2010).

Categorical variables	**Description**	**Modalities**		**Proportions**	
Gender	Male (ma)	M		0.85%	
	Female (kvinnor)	K		0.15%	
Geographic area	Central and semi-central parts of Sweden's three largest cities.	1		0.13%	
	Suburbs plus middle-sized cities.	2		0.18%	
	Lesser towns, except those in 5 or 7.	3		0.20%	
	Small towns and countryside.	4		0.39%	
	Northen towns.	5		0.04%	
	Northen countryside.	6		0.06%	
	Gotland (Sweden's largest island).	7		0.01%	
Vehicle power	EV ratio -5	1		0.11%	
	EV ratio 6-8	2		0.08%	
	EV ratio 9-12	3		0.30%	
	EV ratio 13-15	4		0.19%	
	EV ratio 16-19	5		0.18%	
	EV ratio 20-24	6		0.13%	
	EV ratio 25-	7		0.01%	
Continuous variables		**min**	**max**	**mean**	**std**
Owner's age	Age of the policyholder.	16	92	42.54	12.95
Vehicle age	Age of the policyholder's vehicle.	0	99	12.56	9.68
Target variable					
Claim frequency	The ratio between the number of claims and the contract duration.	0	182.51	0.028	0.947
Exposure	The contract duration.	0.002	11.99	1.008	1.06
Number of claims	The number of motor claims.	0		0.9898%	
		1		0.0101%	
		2		0.0004%	

3.1 LDA soft clustering for structured data

As explained in Section 2, by considering each distinct value of the categorical (and discretized 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 $\boldsymbol{\theta}^{(D\times K)}$ as a soft clustering tool where $k = 1, \ldots, K$ are the topic clusters and $d = 1, \ldots, D$ 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, 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 $\boldsymbol{\theta}_{d^{new}}$ of these unseen (new) insurance policies (in the test set) using the optimised posterior distribution $q(\boldsymbol{\theta})$.

Table 3. Number of observations (insurance policies) in the training-testing subsets.

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

To evaluate the goodness of fit of the models, each policy is assigned to the topic for which it has the highest relative probability $\max(\boldsymbol{\theta}_d)$. 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:

$$D_{Bin}(y, \hat{y}) = 2\sum_{d=1}^{D}\left(y_d \ln\left(\frac{y_d}{\hat{y_d}}\right) + (1-y_d)\ln\left(\frac{1-y_d}{1-\hat{y_d}}\right)\right), \tag{30}$$

where y_d is the observed loan default probability of the dth policy, and $\hat{y_d}$ is the predicted loan default probability of the dth policy (which is equal to the average number of defaults of all policies assigned to the same (topic) cluster as the dth policy). The Poisson deviance is defined as:

$$D_{Poi}(y, \hat{y}) = 2\sum_{d=1}^{D} N_d\left(\frac{\boldsymbol{\nu}_d}{N_d}\hat{\boldsymbol{\lambda}_d} - \left(\ln\left(\frac{\boldsymbol{\nu}_d}{N_d}\hat{\boldsymbol{\lambda}_d}\right)+1\right)\mathbb{I}_{\{N_d\geq 1\}}\right), \tag{31}$$

where N_d and $\boldsymbol{\nu}_d$ are respectively the number of claims and the exposure of the dth policy, and $\hat{\boldsymbol{\lambda}_d}$ 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 dth 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. Figure 3 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 (of interest) 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 Figure 3) which contains more observations (see Table 3). In the following analyses, all results (including goodness of fit or deviance values) are evaluated on the test sets, unless otherwise specified.

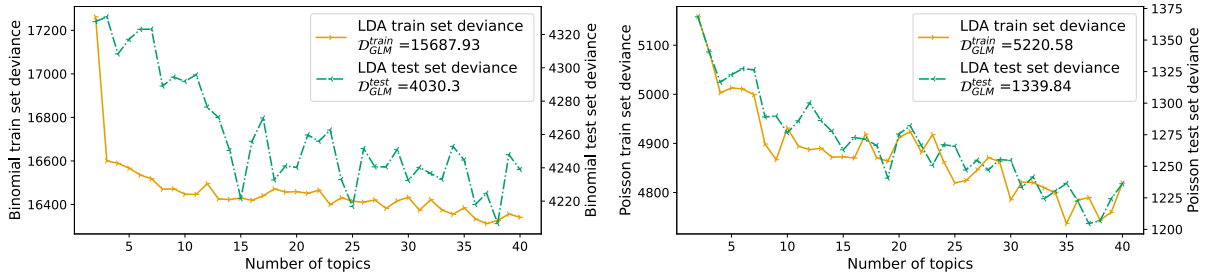

Figure 3. Evolution of the LDA Binomial (left plot) and Poisson (right plot) deviances with respect to the number of LDA topics $K \in (2, 40)$.

If we chose to train our LDA model with $K = 15$ topics on a subset (80%) of the loan portfolio (in Table 1), we get a Binomial deviance equal to 4 221.52 (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 discretized and categorical covariates are one-hot encoded) gives a lower (and better) deviance equal to 4 030.3. The granularity of the GLM (4 809 distinct $\hat{y_d}$) is however much higher than LDA clustering (15 distinct $\hat{y_d}$). The choice of setting the number of topics K to 15 was determined through visual analysis of the evolution of the deviance (green line) in the left plot of Figure 3. Moreover, this number is sufficiently large to capture the data underlying structure, while still being small enough to allow for a concise analysis of the composition of the different clusters.

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 4 that the LDA approach gives some insights into the prominent risk factors contributing to higher or lower claim frequencies. This table 4 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. For ease of analysis, 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 4 as none of their respective categories (modalities) has a consistent association with higher or lower loan default rates (see Figure 13 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 Figure 14 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. We observe in Table 4 and in the left plot of Figure 4 that the two topic clusters with the highest number of defaults have the

Table 4. 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.

Topic	% of policies	term	sub_grade	int_rate	revol_util	Frequency (%)
1	0.77	36 months	1-10	6.86	37.0	0.05
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	['11-20', '1-10']	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	['11-20', '1-10']	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	1.6	60 months	['21-35', '11-20']	15.22	61.08	0.253
15	3.74	60 months	21-35	15.33	72.8	0.2835

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 Figure 4 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 4. 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 4.

With regard to numeric covariates, we observe in the left plot of Figure 5 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 Figure 5 for the revolving line utilisation rate.

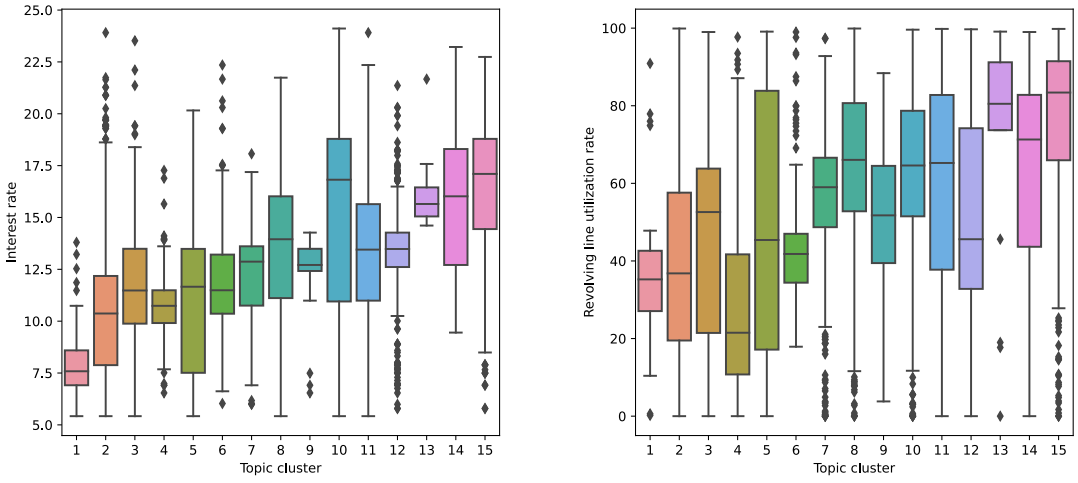

Figure 5. Boxplot of the numeric variables `int_rate` and `revol_util` by topic cluster. The topic clusters are ordered by increasing (average) number of loan defaults, as in Table 4.

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 2). To strike a balance between capturing the underlying data structure and maintaining a manageable number of clusters for analysis, and after visually analysing the evolution of the deviance on the test set (green line) in the right plot of Figure 3, 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 discretized 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 5, further highlights LDA's ability to model the data underlying structure and give insights into the prominent risk factors contributing to higher or lower claim frequencies. This Table 5 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 `VehiculeAge`) in each topic cluster. For ease of analysis, the topic clusters (as well as the corresponding figures) are ordered by increasing (average) motor claim frequencies (see `Frequency` column).

Table 5. Most common modalities and average numeric values with LDA soft clustering (motor portfolio). The table is ordered by increasing (average) motor claim frequencies.

Topic	% of policies	Zone	Class	Gender	OwnersAge	VehiculeAge	Frequency (%)
1	7.53	4	1	M	48.17	32.19	0.13
2	14.44	4	['3', '4']	M	46.2	14.5	0.29
3	6.01	['4', '1']	3	K	45.0	6.62	0.52
4	15.61	4	['6', '3']	M	56.46	14.0	0.61
5	8.35	['4', '2']	4	M	46.5	13.0	0.79
6	8.01	['4', '1']	['3', '4']	M	32.8	15.0	1.2
7	13.42	['2', '3']	3	M	50.79	2.0	1.38
8	7.71	['3', '6']	5	M	47.18	15.0	1.54
9	10.05	['4', '3']	['6', '3']	M	35.69	6.0	2.03
10	8.87	['4', '2']	['5', '4']	M	26.75	8.0	4.89

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 Figure 6). 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 Figure 6).

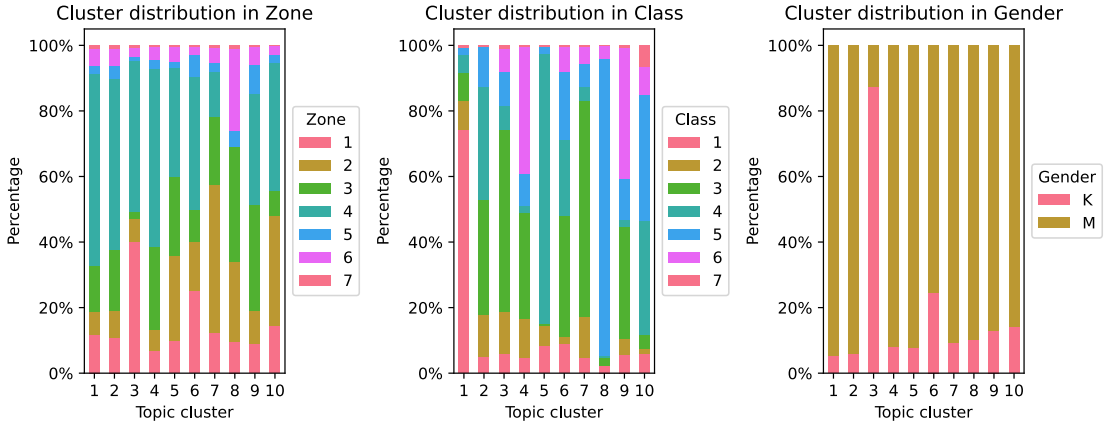

Figure 6. Topic cluster distribution in the categorical variables `Zone`, `Class` and `Gender`. The topic clusters are ordered by increasing (average) motor claim frequencies, as in Table 5.

The left plot of Figure 7 also highlights the differences in age distribution across topic clusters. It is interesting to note that the topic cluster associated with 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 obverse in the right plot of Figure 7 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 Figure 6). Collection vehicles might be driven less frequently, reducing the likelihood of accidents.

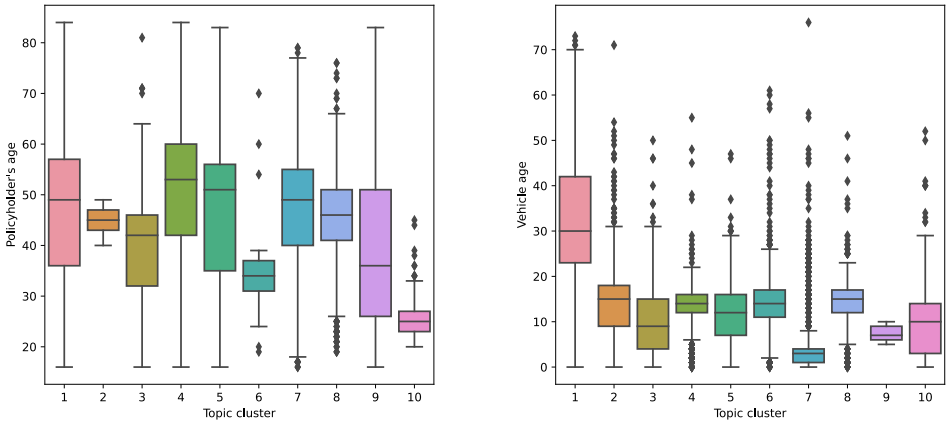

Figure 7. Boxplot of the numeric variables `OwnersAge` and `VehiculeAge` by topic cluster. The topic clusters are ordered by increasing (average) motor claim frequencies, as in Table 5.

The following Table 6 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 1 (loan portfolio) and Table 2 (motor portfolio).

Table 6. Summary of the LDA goodness of fit results on the test sets.

	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
GLM deviance	4 030.3	1 339.84

3.2 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 2023). In this Burt framework, categorical data are mapped to numeric vectors as follow:

$$\mathbf{B}^{\mathbf{W}} = \frac{1}{q}\mathbf{CBC}\,, \tag{32}$$

where q is equal to the number of categorical (and discretized numeric) variables, $\mathbf{B}$ is equal to the product of the disjunctive matrix $\mathbf{D}$ and its transpose, and the diagonal weight matrix $\mathbf{C} = diag\left(n_{11}^{-1/2} \dots n_{mm}^{-1/2}\right)$ contains on its diagonal the number of policies characterised by the ith modality, denoted by n_{ii}. For more details, we refer to Jamotton, Hainaut, and Hames (2023). For a survey of others state-of-the-art mixed data clustering algorithms, see Ahmad and Khan (2019). The following Figure 8 shows the evolution of the Binomial and

Poisson deviances with respect to the number of fuzzy K-means clusters. The decrease in deviance appear more smooth with the Burt approach than with LDA for the loan portfolio (see left plot of Figure 3). If we train our Burt adapted fuzzy K-means on the same subset

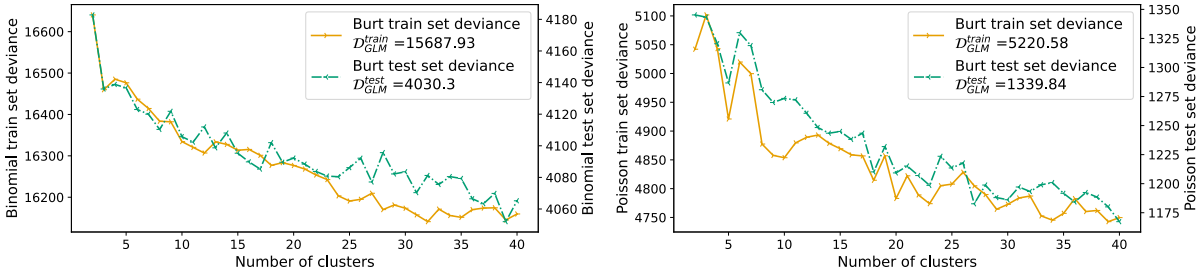

Figure 8. Evolution of the Binomial (left plot) and Poisson (right plot) deviances with respect to the number of K-means clusters $K \in (2, 40)$ in our Burt adapted fuzzy K-means.

(80%) of the loan insurance portfolio (in Table 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 4 095.31 (with 15 distinct $\hat{y_d}$). In comparison, the GLM deviance was equal to 4 030.3 (with 4 809 distinct $\hat{y_d}$) and the LDA deviance was equal to 4 221.52 (with 15 distinct $\hat{y_d}$). The following Table 7 reports some of the most common modalities and average numeric values in each topic cluster.

Table 7. 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.

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	8.2	36 months	1-10	7.53	19.3	0.0847
4	4.98	36 months	1-10	9.16	36.56	0.093
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	0.158
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

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 Figure 9 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 7 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 Figure 9.

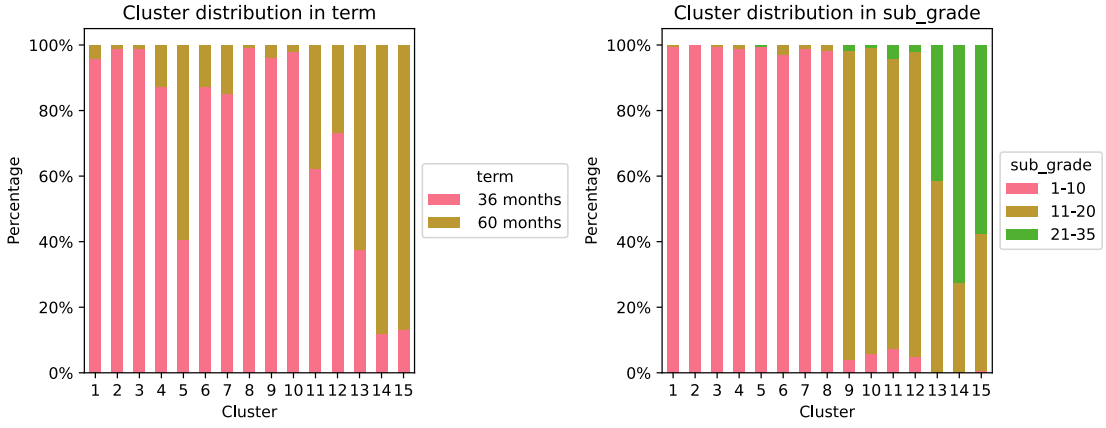

Figure 9. Cluster distribution in the categorical variables `term` and `sub_grade`. The clusters are ordered by increasing (average) number of loan defaults, as in Table 4.

In Figure 10, 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 Figure 5.

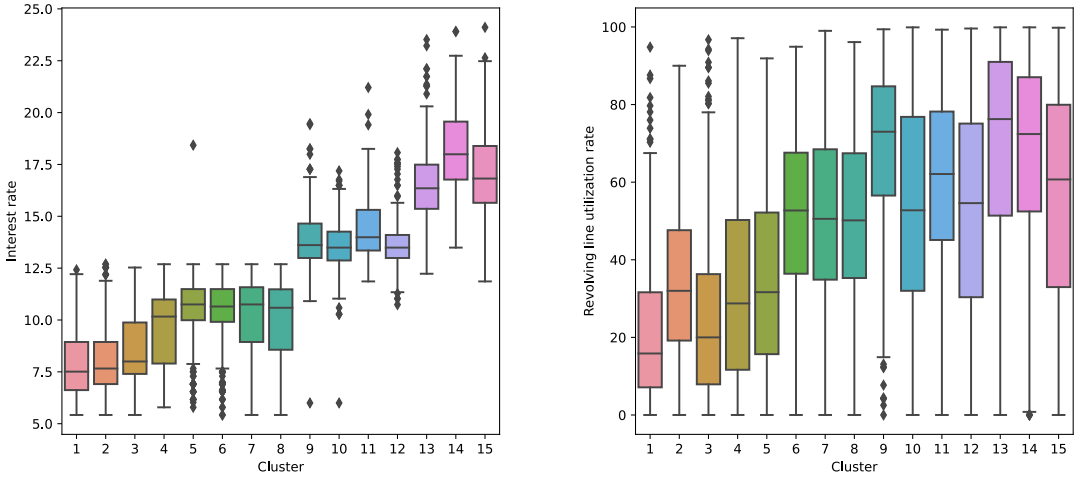

Figure 10. Boxplot of the numeric variables `int_rate` and `revol_util` by cluster. The clusters are ordered by increasing (average) number of loan defaults, as in Table 4.

However, note that the size of the different clusters is more or less the same (see column titled `% of policies` in Table 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 3.1, we also train our Burt adapted fuzzy K-means model on a subset (80%) of the motor portfolio (in Table 2), with the same number of fuzzy clusters $K = 10$ as the number of LDA topic clusters. It results in a Poisson deviance equal to 1 273.37 (with 10 distinct $\hat{y_d}$), which is quite close to the one obtained with the LDA approach (1 276.51 with 10 distinct $\hat{y_d}$) and still better than the one obtained with the GLM (1 339.84 with more than 1 269 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 8), we observe in the LDA approach in Table 5 an average claim frequency of 0.13%, 0.29%, 0.52%, and 0.61% in the first four clusters.

Table 8. Dominant modalities and average numeric values of the continuous v variables (from the motor database) with K-means Burt adapted soft clustering. The table is ordered by increasing (average) motor claim frequencies.

Cluster	% of policies	Zone	Class	Gender	OwnersAge	VehiculeAge	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	46.2	7.5	0.46
4	13.78	4	['5', '3', '4']	M	47.0	14.0	0.49
5	7.87	3	['5', '3']	M	44.94	15.0	0.57
6	6.62	['4', '2']	['3', '4']	M	53.5	6.0	1.01
7	15.77	['2', '1']	['3', '5']	M	47.4	15.5	1.16
8	7.12	['4', '3', '2']	3	M	46.24	2.67	1.2
9	9.91	['4', '3', '2']	['4', '5']	M	36.5	2.0	3.35
10	7.45	['4', '3']	['6', '5']	M	25.5	7.0	4.8

Nevertheless, the dominant categories and average numeric values obtained across the hardened fuzzy clusters in Table 8 are highly similar to our LDA topic clusters analysis in Table 5. The evolution of the numeric variables with respect to the clusters average claim frequency in Figure 11, as well as the cluster distribution in the categorical variables in Figure 12 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 Figure 12) with the lowest power class (see middle plot of Figure 11). Whereas, the cluster with the highest claim frequency exclusively consists of young (less experiences) drivers (around 25), as illustrated in the left plot of Figure 12 and has the largest proportions of higher power vehicles (see middle plot of Figure 11). The cluster distribution in the `Gender` variable (see right plot of the following Figure 11) 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.

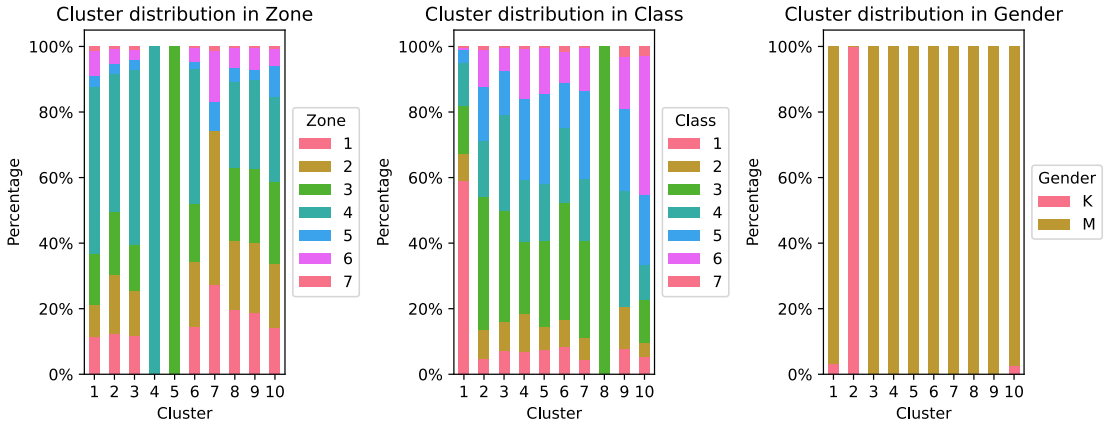

Figure 11. Cluster distribution in the categorical variables `Zone`, `Class` and `Gender`. The clusters are ordered by increasing (average) motor claim frequencies, as in Table 8.

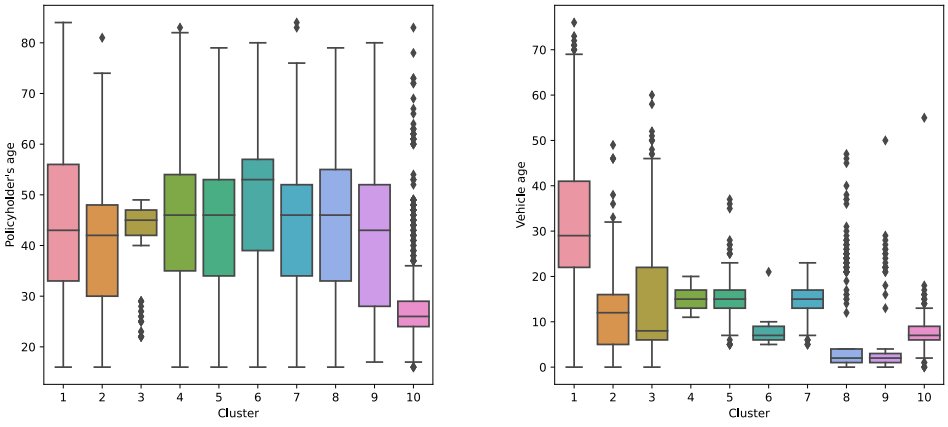

Figure 12. Boxplot of the numeric variables `OwnersAge` and `VehiculeAge` by cluster. The clusters are ordered by increasing (average) motor claim frequencies, as in Table 8.

The following Table 9 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.

Table 9. Summary of the goodness of fit results on the test sets.

	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	1273.37
GLM deviance	4 030.3	1 339.84

4. Conclusions

In this article, 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 discretized 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 discretized 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.

For future research, the flexibility of LDA in handling varying document lengths could be an interesting prospect to the analysis of heterogeneous insurance datasets containing missing values.

Funding Statement This research was supported by the project ASTeRISK under research grant ID 40007517, from the Excellence of Science (EoS) program call by FWO and FNRS.

Competing Interests No potential conflict of interest was reported by the authors.

Data availability statement The loan data set used to support the findings of this study is available on the link: https://www.kaggle.com/datasets/imsparsh/lending-club-loan-dataset-2007-2011/data. The motor data set used to support the findings of this study is available on the companion website of the book "Non-Life Insurance Pricing with Generalized Linear Models" by Ohlsson and Johansson (2010). Link: https://staff.math.su.se/esbj/GLMbook/case.html

References

Ahmad, Amir, and Shehroz S Khan. 2019. Survey of state-of-the-art mixed data clustering algorithms. *Ieee Access* 7:31883–31902.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3 (Jan): 993–1022.

Campbell, Malcolm. 1986. An integrated system for estimating the risk premium of individual car models in motor insurance. *ASTIN Bulletin: The Journal of the IAA* 16 (2): 165–183.

Carpenter, Bob. 2010. Integrating out multinomial parameters in latent dirichlet allocation and naive bayes for collapsed gibbs sampling. *Rapport Technique* 4:464.

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,* 642–647.

Dinh, Duy-Tai, Van-Nam Huynh, and Songsak Sriboonchitta. 2021. Clustering mixed numerical and categorical data with missing values. *Information Sciences* 571:418–442.

Fukuyama, Julia, Kris Sankaran, and Laura Symul. 2021. Multiscale analysis of count data through topic alignment. *arXiv preprint arXiv:2109.05541.*

Gan, Guojun, and Emiliano A Valdez. 2020. Data clustering with actuarial applications. *North American Actuarial Journal* 24 (2): 168–186.

Hancock, John T, and Taghi M Khoshgoftaar. 2020. Survey on categorical data for neural networks. *Journal of Big Data* 7 (1): 1–41.

Hoffman, Matthew, Francis Bach, and David Blei. 2010. Online learning for latent dirichlet allocation. *advances in neural information processing systems* 23.

Hofmann, Thomas. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international acm sigir conference on research and development in information retrieval,* 50–57.

Jain, Anil K, M Narasimha Murty, and Patrick J Flynn. 1999. Data clustering: a review. *ACM computing surveys (CSUR)* 31 (3): 264–323.

Jamotton, Charlotte, Donatien Hainaut, and Thomas Hames. 2023. *Insurance analytics with clustering techniques.* Technical report. Université catholique de Louvain, Institute of Statistics, Biostatistics and ...

Jose, Alex, Angus Smith Macdonald, George Tzougas, and George Streftaris. 2024. Interpretable zero-inflated neural network models for predicting admission counts. *Annals of Actuarial Science.*

Krestel, Ralf, Peter Fankhauser, and Wolfgang Nejdl. 2009. Latent dirichlet allocation for tag recommendation. In *Proceedings of the third acm conference on recommender systems,* 61–68.

Ng, Kai Wang, Man-Lai Tang, Ming Tan, and Guo-Liang Tian. 2008. Grouped dirichlet distribution: a new tool for incomplete categorical data analysis. *Journal of Multivariate Analysis* 99 (3): 490–509.

Ohlsson, Esbjörn, and Björn Johansson. 2010. *Non-life insurance pricing with generalized linear models.* Vol. 2. Springer.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Richman, Ronald. 2021. Ai in actuarial science–a review of recent advances–part 1. *Annals of Actuarial Science* 15 (2): 207–229.

Wang, Yi. 2011. Distributed gibbs sampling of latent dirichlet allocation: the gritty details.

Wang, Yibo, and Wei Xu. 2018. Leveraging deep learning with lda-based text analytics to detect automobile insurance fraud. *Decision Support Systems* 105:87–95.

Xuan, Junyu, Jie Lu, Guangquan Zhang, and Xiangfeng Luo. 2014. Release 'bag-of-words' assumption of latent dirichlet allocation. In *Foundations of intelligent systems: proceedings of the eighth international conference on intelligent systems and knowledge engineering, shenzhen, china, nov 2013 (iske 2013)*, 83–92. Springer.

Yang, M-S. 1993. A survey of fuzzy clustering. *Mathematical and Computer modelling* 18 (11): 1–16.

Yao, Ji. 2016. Clustering in general insurance pricing. *Predictive modeling applications in actuarial science*, 159–79.

Appendix 1. LDA soft clustering

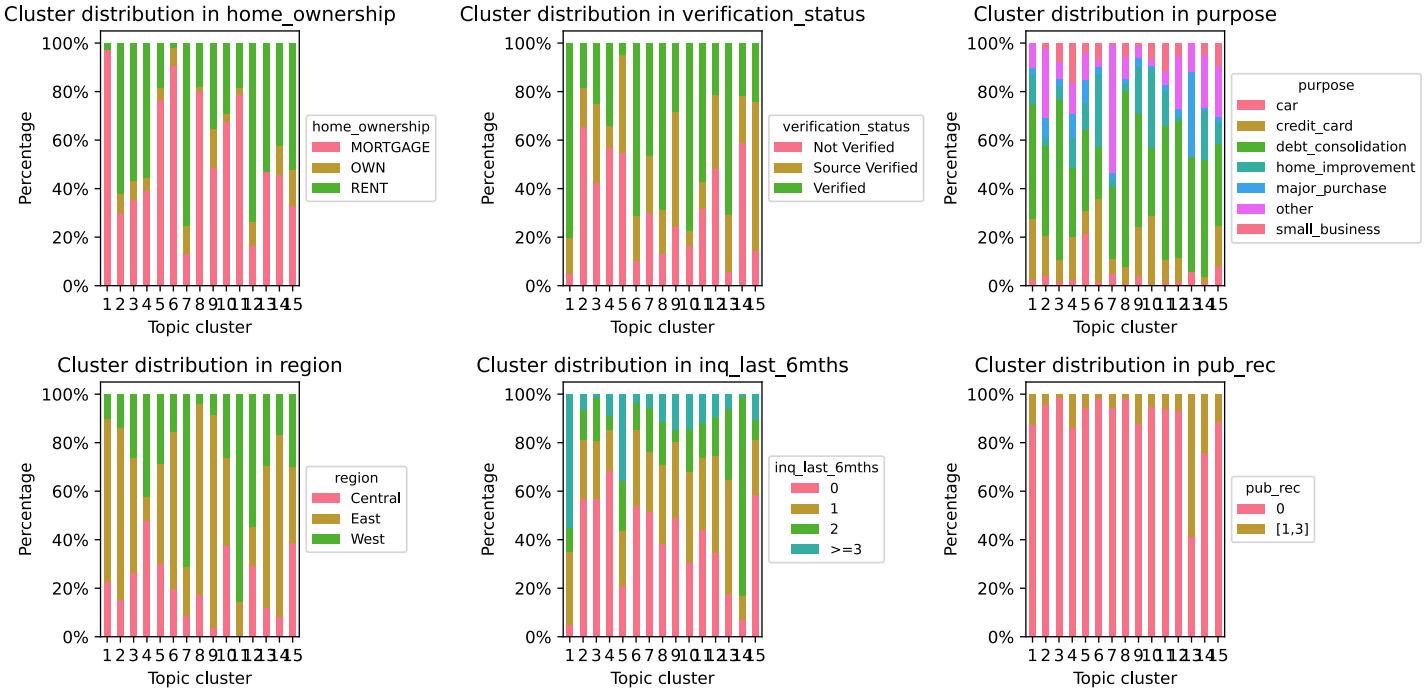

Figure 13. Topic cluster distribution (from the LDA approach) in the categorical variables `home_ownership`, `verification_status`, `purpose`, `region`, `inq_last_6mths`, and `pub_rec`. The topic clusters are ordered by increasing (average) number of loan defaults, as in Table 4.

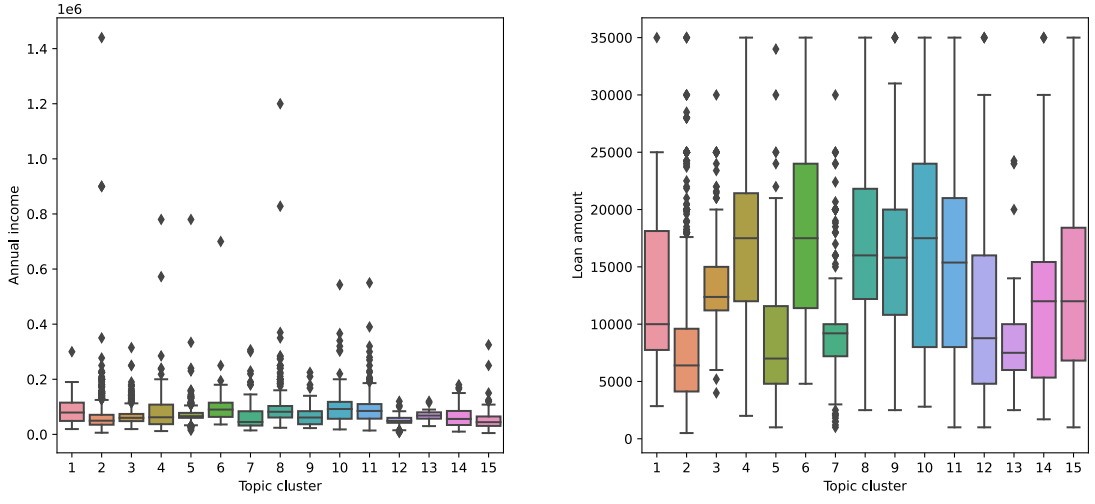

Figure 14. Boxplot of the numeric variables `annual_inc` and `loan_amnt` by topic cluster (LDA clustering approach). The topic clusters are ordered by increasing (average) number of loan defaults, as in Table 4.

Appendix 2. Burt-adapted soft clustering

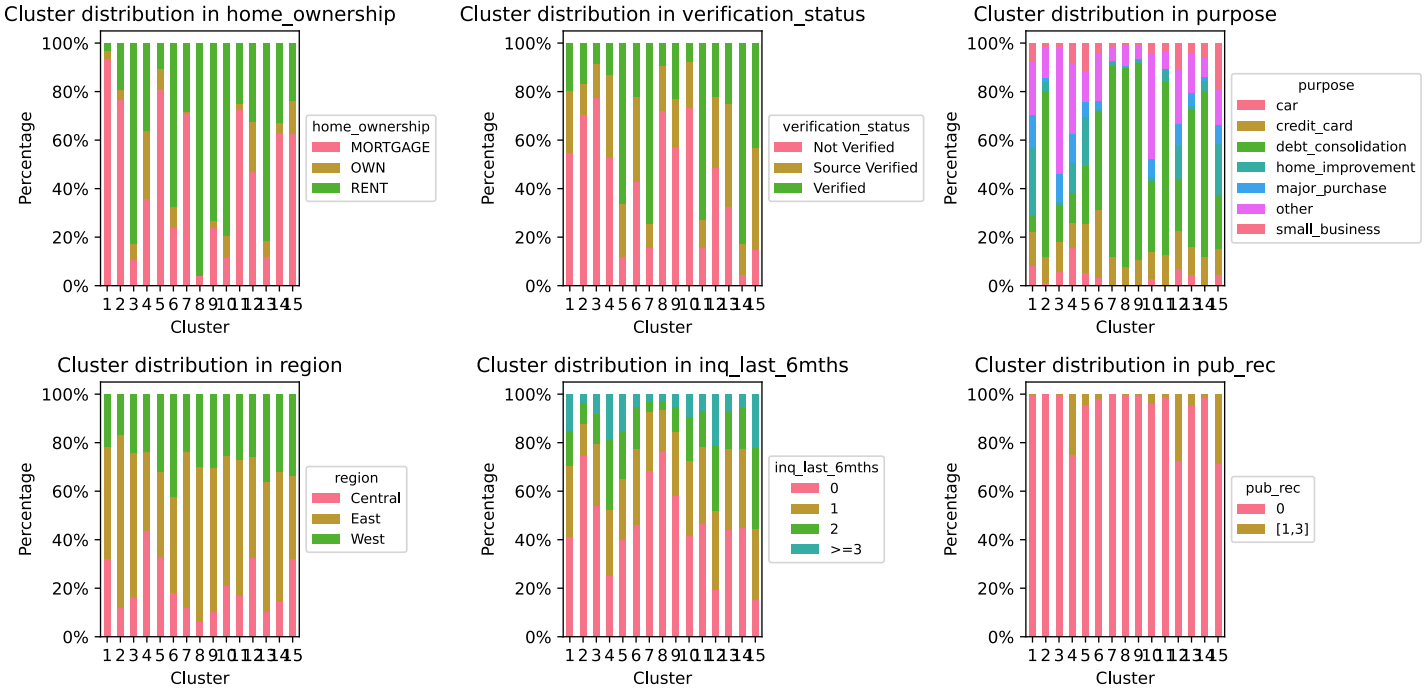

Figure 15. Cluster distribution (from the Burt approach) in the categorical variables `home_ownership`, `verification_status`, `purpose`, `region`, `inq_last_6mths`, and `pub_rec`. The clusters are ordered by increasing (average) number of loan defaults, as in Table 7.

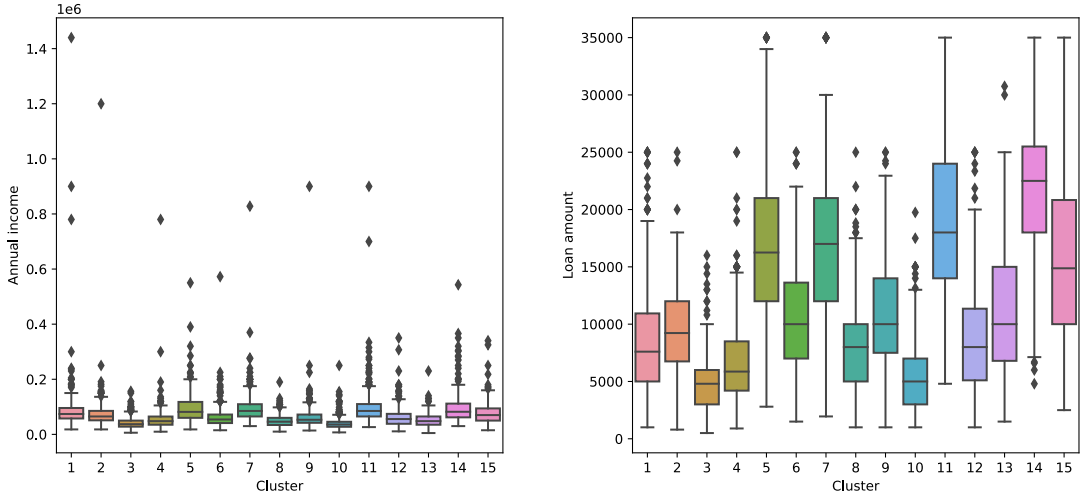

Figure 16. Boxplot of the numeric variables `annual_inc` and `loan_amnt` by cluster (Burt adapted clustering approach). The clusters are ordered by increasing (average) number of loan defaults, as in Table 7.