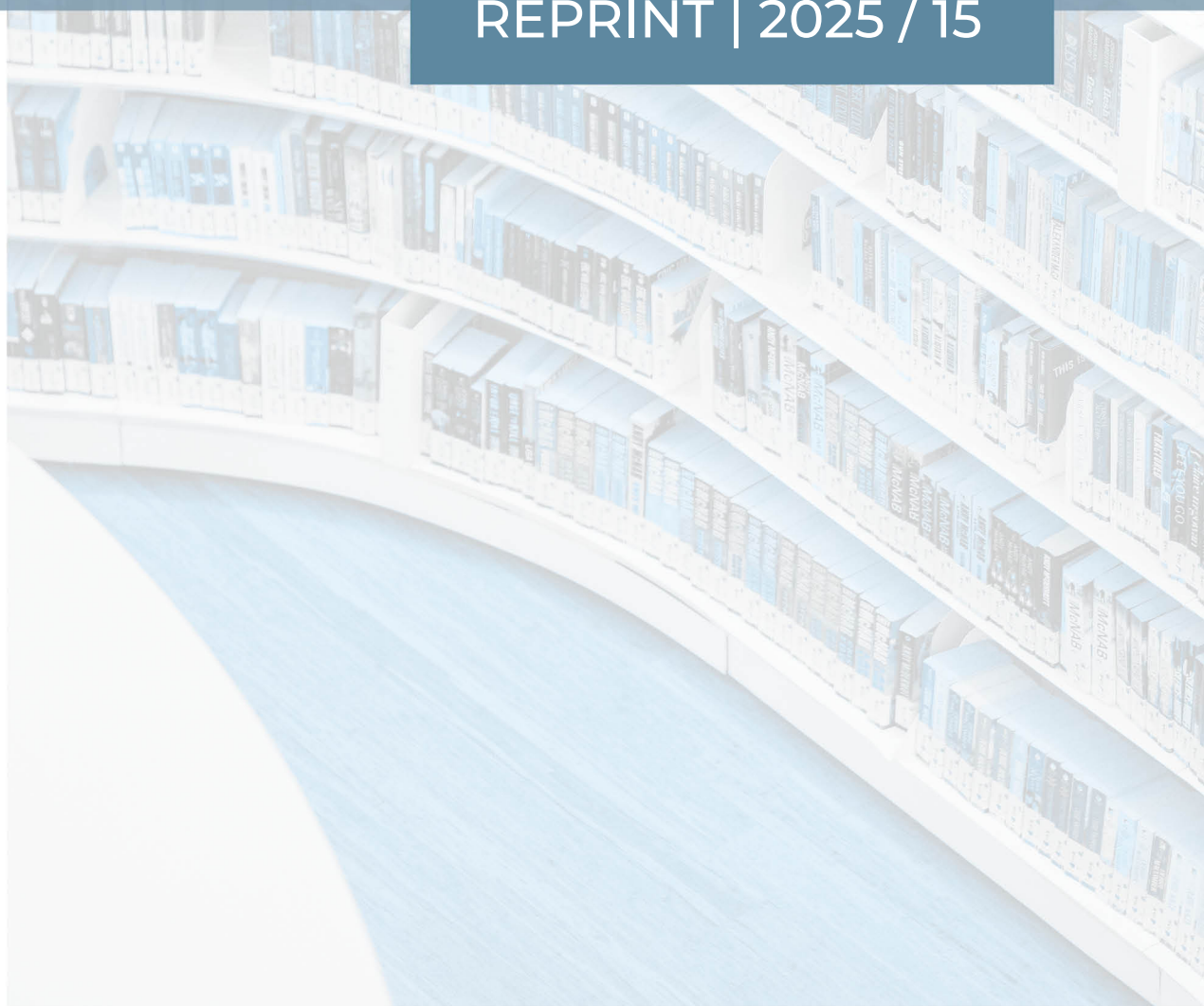
COMPARISON OF PREDICTORS' PERFORMANCE IN INSURANCE PRICING: TESTING FOR BREGMAN DOMINANCE BASED ON MURPHY DIAGRAMS

Michel Denuit, Julien Trufin, Thomas Verdebout

REPRINT | 2025 / 15



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

European Actuarial Journal (2025) 15:493–504
https://doi.org/10.1007/s13385-025-00429-4

LETTER

Comparison of predictors' performance in insurance pricing: testing for Bregman dominance based on Murphy diagrams

Michel Denuit[1] · **Julien Trufin**[2] · **Thomas Verdebout**[3]

Received: 16 December 2024 / Revised: 11 April 2025 / Accepted: 30 June 2025 /
Published online: 4 August 2025

Abstract
Ehm et al. (2016) defined forecast dominance, or Bregman dominance as dominance for every Bregman loss function. This letter explores Bregman dominance to compare competing candidate pure premiums. An actuarial interpretation is given for the expectation of extremal consistent loss functions characterizing Bregman dominance, showing their relevance in the context of insurance pricing. An effective testing procedure for Bregman dominance is proposed based on Murphy diagrams and its performance is evaluated through a simulation study. Compared to the existing literature, the proposed testing procedure is easier because it exploits the independence and homogeneity assumptions made in insurance studies and focuses on the conditional mean. An application to a Swiss motor insurance data set demonstrates the potential of the proposed procedure.

Keywords Insurance pricing · Model comparison · Bregman dominance · Murphy diagram · Statistical test

1 Setting

Given a response Y (claim number or claim amount) and a set of features $X_1, \ldots, X_p$ gathered in the vector $\boldsymbol{X}$, actuarial pricing aims to evaluate the pure premium $\mu(\boldsymbol{X}) = \mathrm{E}[Y|\boldsymbol{X}]$ as accurately as possible. To this end, the function $\boldsymbol{x} \mapsto \mu(\boldsymbol{x}) = \mathrm{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$ is approximated by a (working, or actual) pure premium $\boldsymbol{x} \mapsto \pi(\boldsymbol{x})$ based on an additive or piece-wise constant score. Once fitted on the training data set using an

✉ Julien Trufin
julien.trufin@ulb.ac.be

1 Institute of Statistics, Biostatistics and Actuarial Science - ISBA, Louvain Institute of Data Analysis and Modeling - LIDAM, UCLouvain, Louvain-la-Neuve, Belgium

2 Department of Mathematics, Université Libre de Bruxelles (ULB), Bruxelles, Belgium

3 ECARES and Department of Mathematics, Université Libre de Bruxelles (ULB), Brussels, Belgium

appropriate learning procedure, this produces estimates, or predicted values $\widehat{\pi}(\boldsymbol{x})$ for $\mu(\boldsymbol{x})$.

The developments in this paper hold the training set as fixed. Precisely, $\widehat{\pi}$ is assumed to have been obtained by minimizing the average empirical loss given some training set and all formulas in the text are meant given this training set. The respective performances of competing models can then be assessed on the basis of a validation set $\{(Y_i, \boldsymbol{X}_i), \quad i = 1, 2, \ldots, n\}$, that has not been used to obtain $\widehat{\pi}$. Provided the validation set comprises a large number of independent and identically distributed random vectors $(Y_i, \boldsymbol{X}_i)$, this allows us to perform calculations with a generic random vector $(Y, \boldsymbol{X})$, independent of, and distributed as the observations contained in the training set. This assumption is retained throughout the paper.

2 Bregman dominance

Patton [9] established that Bregman loss functions are those attaining their minimum at the expected value of the response, subject to weak regularity conditions. The class of Bregman loss functions is then said to be strictly consistent for the (conditional) mean functional. We refer the interested reader to [5] for a general overview of consistency in statistics.

Since pure premiums correspond to conditional expectations, every Bregman loss function is thus eligible to fit $\widehat{\pi}$. See e.g. Table 1 in [6] as well as [8] for examples of Bregman loss functions. Interestingly, all deviance functions associated to Exponential Dispersion distributions are Bregman loss functions.

We refer the reader to [4] for a general account of model comparison in an insurance context. As pointed out in 2015 by Patton (the reference is now updated to Patton, 2020), rankings of two predictors may depend on the specific consistent loss function used for out-of-sample evaluation. That is why [2, 3, 14, 16] and [1] proposed graphical tools and hypothesis tests to assess the robustness of empirical forecast rankings. In their terminology, one predictor dominates another if it performs better with respect to every consistent loss function. This approach leads to Bregman dominance when considering pure premiums. Stated precisely, $\widehat{\pi}_2$ outperforms $\widehat{\pi}_1$ in terms of Bregman dominance if the inequality $\mathrm{E}[L(Y, \widehat{\pi}_2(\boldsymbol{X}))] \leq \mathrm{E}[L(Y, \widehat{\pi}_1(\boldsymbol{X}))]$ holds true for every Bregman loss function L. This guarantees that the better performance of $\widehat{\pi}_2$ compared to $\widehat{\pi}_1$ is robust across all Bregman loss functions.

3 Murphy diagram

Krüger and Ziegel [6] established that $\widehat{\pi}_2$ outperforms $\widehat{\pi}_1$ in terms of Bregman dominance if, and only if,

$$\mathrm{E}\big[(t - Y)\big(\mathrm{I}[\widehat{\pi}_2(\boldsymbol{X}) > t] - \mathrm{I}[\widehat{\pi}_1(\boldsymbol{X}) > t]\big)\big] \leq 0 \quad \text{for all } t \in \mathbb{R}, \tag{1}$$

where $\mathrm{I}[\cdot]$ is the indicator function (equal to 1 if the event appearing within the brackets is realized and to 0 otherwise). Note that in applications to insurance pricing, $\widehat{\pi}_k$ is

positive and finite for $k \in \{1, 2\}$ so that condition (1) is automatically fulfilled for every $t < 0$ and we can restrict our attention to finite values of $t \geq 0$. Therefore, we will assume below that the set on which (1) must hold to get dominance is a compact subset–Ω say–of the positive real line.

The aim of insurance ratemaking, or actuarial pricing is to evaluate the pure premium as accurately as possible. This means that the target is the conditional expectation $\mu(\boldsymbol{X})$ of the response Y. Thus, the goal is to estimate the unknown $\mu(\boldsymbol{X})$ and not to predict response values. This point is crucial for proper understanding of insurance ratemaking: predicting the actual losses Y is meaningless since the very nature of insurance is to compensate losses within large groups. If actuaries were able to predict the response Y accurately, insurance would even become useless because every policyholder would be charged his or her actual losses. Instead, actuaries are only interested in the true pure premium $\mu(\boldsymbol{X})$ because insurance exploits the fact that individual departures $Y - \mu(\boldsymbol{X})$ just cancel each other when they are summed over a sufficiently large portfolio. We refer the reader to [7] for more details about the very aim of insurance ratemaking, which is not to predict the actual losses Y but to create accurate estimates of $\mu(\boldsymbol{X})$, which is unobserved. This point of view differs from the literature devoted to predictive accuracy and forecast dominance.

Inequality (1) possesses a nice interpretation in terms of performances of the candidate pure premiums $\widehat{\pi}_1$ and $\widehat{\pi}_2$ as it can be rewritten as

$$\mathrm{E}\big[(\mu(\boldsymbol{X}) - t)\mathrm{I}[\widehat{\pi}_1(\boldsymbol{X}) > t]\big] \leq \mathrm{E}\big[(\mu(\boldsymbol{X}) - t)\mathrm{I}[\widehat{\pi}_2(\boldsymbol{X}) > t]\big] \quad \text{for all } t \in \Omega. \quad (2)$$

Inequality (2), which is equivalent to (1), compares $\widehat{\pi}_1$ and $\widehat{\pi}_2$ with the help of the true pure premium $\mu(\boldsymbol{X})$. Now, (2) explains why the actuary should consider $\widehat{\pi}_2$ as being superior to $\widehat{\pi}_1$ when (1) holds true. Inequality (2) requires that whatever the threshold t, the average excess of true pure premium $\mu(\boldsymbol{X})$ over threshold t when the candidate premium exceeds this threshold is larger for $\widehat{\pi}_2$ compared to $\widehat{\pi}_1$. When this holds for every t, it means that $\widehat{\pi}_2$ better captures the behavior of the true pure premium above any selected threshold t than $\widehat{\pi}_1$. The same interpretation holds with shortfalls since (1) can be rewritten as

$$\mathrm{E}\big[(t - \mu(\boldsymbol{X}))\mathrm{I}[\widehat{\pi}_1(\boldsymbol{X}) \leq t]\big] \leq \mathrm{E}\big[(t - \mu(\boldsymbol{X}))\mathrm{I}[\widehat{\pi}_2(\boldsymbol{X}) \leq t]\big] \quad \text{for all } t \in \Omega.$$

Of course, since the true pure premium $\mu(\cdot)$ is unknown in applications, the inequalities with $\mu(\boldsymbol{X})$ remain conceptual and only serve to provide actuaries with an intuitive meaning for (1) and its relevance in insurance studies. In practice, the inequalities with the observed response Y replacing the true pure premium $\mu(\boldsymbol{X})$ can be estimated from the available data and can be used to assess the performances of the candidate pure premiums $\widehat{\pi}_1$ and $\widehat{\pi}_2$ under consideration.

Subtracting $\mathrm{E}[(\mu(\boldsymbol{X}) - t)_+]$ to both sides of inequality (2), we recover the extemal, or elementary scores identified by Ehm et al. ([2]), but where Y is replaced with $\mu(\boldsymbol{X})$. However, the interpretation given by these authors in the context of forecasting does not translate to insurance ratemaking.

Denoting as

$$L_t(\widehat{\pi}_k(\boldsymbol{X}), Y) = \frac{1}{2}(t - Y)\mathrm{I}[\widehat{\pi}_k(\boldsymbol{X}) > t] \tag{3}$$

the extremal consistent loss function of [2] for the mean indexed by the parameter $t \in \Omega$ (up to a term that does not depend on $\widehat{\pi}_k(\boldsymbol{X})$), $k = 1, 2$, we see that (1) is equivalent to $\mathrm{E}[L_t(\widehat{\pi}_2(\boldsymbol{X}), Y)] \leq \mathrm{E}[L_t(\widehat{\pi}_1(\boldsymbol{X}), Y)]$ for all $t \in \Omega$.

Ehm et al. [2] used the extremal consistent loss functions (3) to graphically compare performances of two predictors. Those graphs are known as Murphy diagrams; see, e.g., [4] for applications to insurance. Coming back to the interpretation of (1) in terms of performances of candidate pure premiums, we see that Murphy diagrams actually assess the accuracy of the pricing structure. The actuary must then favor candidate premiums leading to smaller expected loss (3) because this means that the corresponding premium better capture the behavior of the true, unknown pure premium.

While the Murphy diagram is a useful tool, it only provides graphical evidence of the performance differences but gives no formal statistical justification. [3, 6] and [14] already developed statistical tests for Bregman dominance based on the extremal consistent loss functions of [2]. However, these authors consider forecasts based on time series whereas the present paper considers predictions made from independent data as commonly found in insurance. This allows us to propose a simple testing procedure of Kolmogorov-Smirnov type.

Every Bregman loss function can be recovered as a weighted average of the extremal consistent loss functions (3). Let us illustrate this result with the classical mean squared error (MSE):

$$\begin{aligned}\mathrm{E}\left[\int_0^\infty L_t(\widehat{\pi}_k(\boldsymbol{X}), Y)\mathrm{d}t\right] &= \frac{1}{2}\mathrm{E}\left[\int_0^{\widehat{\pi}_k(\boldsymbol{X})}(t - Y)\mathrm{d}t\right] \\ &= \frac{1}{4}\mathrm{E}\big[(\widehat{\pi}_k(\boldsymbol{X}) - Y)^2 - Y^2\big]\end{aligned} \tag{4}$$

which reduces to the MSE since the second term does not involve the predictor $\widehat{\pi}_k(\boldsymbol{X})$. This shows that MSE is an aggregation of the finer information provided by Murphy diagram.

4 Testing for Bregman dominance

Consider two predictors $\widehat{\pi}_1(\boldsymbol{X})$ and $\widehat{\pi}_2(\boldsymbol{X})$ valued in the sets $\Omega_1 \subseteq \mathbb{R}$ and $\Omega_2 \subseteq \mathbb{R}$, respectively, and let $\Omega = [\min(\Omega_1 \cup \Omega_2), \max(\Omega_1 \cup \Omega_2)] \subseteq \mathbb{R}$ be a compact interval. From (1), we want to test the null hypothesis

$$\mathcal{H}_0 : \mathrm{E}\big[(t - Y)\left(\mathrm{I}[\widehat{\pi}_2(\boldsymbol{X}) > t] - \mathrm{I}[\widehat{\pi}_1(\boldsymbol{X}) > t]\right)\big] \leq 0 \quad \text{for all} \quad t \in \Omega$$

against the alternative

$$\mathcal{H}_1 : \mathrm{E}\Big[(t - Y)\left(\mathrm{I}[\widehat{\pi}_2(\boldsymbol{X}) > t] - \mathrm{I}[\widehat{\pi}_1(\boldsymbol{X}) > t]\right)\Big] > 0 \quad \text{for some } t \in \Omega.$$

Let

$$D(t) := \mathrm{E}\Big[(t - Y)\left(\mathrm{I}[\widehat{\pi}_2(\boldsymbol{X}) > t] - \mathrm{I}[\widehat{\pi}_1(\boldsymbol{X}) > t]\right)\Big], \quad t \in \Omega, \text{ and } \overline{D} := \sup_{t\in\Omega} D(t).$$

The null hypothesis $\mathcal{H}_0$ is then equivalent to $\overline{D} \leq 0$. A natural estimator $D_n(t)$ of $D(t)$ is

$$D_n(t) = \frac{1}{n}\sum_{i=1}^{n}(t - Y_i)\left(\mathrm{I}[\widehat{\pi}_2(\boldsymbol{X}_i) > t] - \mathrm{I}[\widehat{\pi}_1(\boldsymbol{X}_i) > t]\right).$$

A test statistic of Kolmogorov-Smirnov type can then be defined as

$$T_n = \sup_{t\in\Omega} \sqrt{n} D_n(t).$$

The test will reject the null hypothesis $\mathcal{H}_0$ at level $\beta \in (0, 1)$ if the value of T_n exceeds some critical value c_β. The critical values c_β can be obtained by studying the limiting behavior of the empirical process $\sqrt{n}(D_n(t) - D(t))$ under the null hypothesis. We have the following result.

Proposition 4.1 *Assume that the second moment of Y exists, i.e. $\mathrm{E}[Y^2] < \infty$. Then, when $D(t) = 0$, which is at the boundary between $\mathcal{H}_0$ and $\mathcal{H}_1$, the empirical process $\sqrt{n} D_n(t)$ converges weakly to a Gaussian process with mean zero and covariance function*

$$\begin{aligned}\Sigma(t_1, t_2) := \mathrm{E}\Big[&(t_1 - Y)\big(\mathrm{I}[\widehat{\pi}_2(\boldsymbol{X}) > t_1] - \mathrm{I}[\widehat{\pi}_1(\boldsymbol{X}) > t_1]\big) \\ &(t_2 - Y)\big(\mathrm{I}[\widehat{\pi}_2(\boldsymbol{X}) > t_2] - \mathrm{I}[\widehat{\pi}_1(\boldsymbol{X}) > t_2]\big)\Big]. \end{aligned} \tag{5}$$

The proof of Proposition 4.1 is given in Appendix A. From Proposition 4.1 and the continuous mapping theorem, it follows that a Kolmogorov-Smirnov type test can be derived by rejecting the null hypothesis $\mathcal{H}_0$ at the asymptotic level $\beta \in (0, 1)$ when

$$T_n = \sup_{t\in\Omega} \sqrt{n} D_n(t) > c_\beta, \tag{6}$$

where the critical value c_β can, in principle, be obtained from Proposition 4.1. Although Proposition 4.1 shows that T_n in (6) is a natural test statistic for the problem considered, computing the critical value c_β is challenging as it requires the computation of $\Sigma(t_1, t_2)$, which is impractical due to the unknown distribution of $(Y, \widehat{\pi}_1(\boldsymbol{X}), \widehat{\pi}_2(\boldsymbol{X}))$. Hence, we propose in Appendix B a Monte Carlo procedure to perform the test, following [15]. A simulation exercise is performed in Appendix C, demonstrating that the procedure described in Appendix B achieves the correct nominal β-level constraint and maintains power under the alternative hypothesis.

5 Case study

5.1 Data set

We consider the Swiss motor insurance data set used in Wüthrich [12] and in Wüthrich and Buser [13] consisting of 500 000 insurance policies. This is a simulated data set which as often been used in the literature. In the present setting, it has the advantage to contain the true pure premium $\mu(\boldsymbol{X})$ so that we can illustrate the intuitive meaning of Murphy diagram in terms of the assessment of the superiority of some candidate pure premium. Of course, the analysis can be conducted without the knowledge of $\mu(\boldsymbol{X})$, as it is the case with a real data set.

For each policy i ($i = 1, \ldots, 500\,000$), the data set records the numbers of claims Y_i filed by the policyholder i, the corresponding exposure-to-risk $e_i \leq 1$ and the features $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{i8})$ where X_{i1} is the age of the driver (`age`), X_{i2} is the age of the car (`ac`), X_{i3} is the power of the car (`power`), X_{i4} is the fuel type of the car (`gas`), X_{i5} is the vehicle brand (`brand`), X_{i6} is the area code of the living place of the driver (`area`), X_{i7} is the population density at the living place of the driver (`dens`), and X_{i8} is the Swiss canton of the license plate of the car (`ct`).

This data set is unique because the values $e_i\mu(\boldsymbol{X}_i)$ of the true model are also provided. In fact, the number of claims Y_i has been generated from a Poisson distribution with mean $e_i\mu(\boldsymbol{X}_i)$. We partition the dataset into a training set $\mathcal{D}$ and a validation set $\overline{\mathcal{D}}$. The training set $\mathcal{D}$ comprises 80% of the observations from the entire dataset, while the validation set $\overline{\mathcal{D}}$ contains the remaining 20% of observations.

5.2 Models under consideration

The observations $(Y_i, \boldsymbol{X}_i)$ are assumed to be independent. Given $\boldsymbol{X}_i = \boldsymbol{x}_i$ and the exposure-to-risk e_i, the response Y_i is assumed to follow a Poisson distribution with mean $e_i\mu(\boldsymbol{x}_i)$. Thus, $\mu(\boldsymbol{x}_i)$ represents the expected annual claim frequency for policyholder i. Our objective is to estimate the function $\boldsymbol{x} \mapsto \mu(\boldsymbol{x})$ using the observations $(Y_i, \boldsymbol{X}_i)$ in the training set $\mathcal{D}$.

We fit generalized additive models (GAMs) on $\mathcal{D}$ with Poisson deviance loss and log-link function using the `R` package `gam`. More precisely, we fit two GAMs producing the following predictors:

- $\widehat{\pi}^{\mathrm{GAM1}}$, using one feature, namely X_1 (`age`);
- $\widehat{\pi}^{\mathrm{GAM2}}$, using two features, namely X_1 (`age`) and X_6 (`area`).

In both models, the effect of the covariate `age` is captured by splines, and we do not consider interaction terms.

We also train gradient boosting trees (GBTs) on $\mathcal{D}$ with Poisson deviance loss and log-link function with the help of the `R` package `gbm`. The bagging fraction γ is set at 0.5. The shrinkage parameter κ is set at the low value 0.01. The size of the trees is controlled by the interaction depth `ID`. We fit two GBTs on 80% of the observations in $\mathcal{D}$ and the optimal values for the number of trees T are determined by minimizing the out-of-sample Poisson deviance loss computed on the remaining 20% of observations in $\mathcal{D}$, producing the following estimators:

Table 1 Out-of-sample estimates (on $\overline{\mathcal{D}}$) of the generalization error

μ	$279.23 \cdot 10^{-3}$
$\widehat{\pi}^{\text{GAM1}}$	$313.26 \cdot 10^{-3}$
$\widehat{\pi}^{\text{GAM2}}$	$312.74 \cdot 10^{-3}$
$\widehat{\pi}^{\text{GBT1}}$	$291.74 \cdot 10^{-3}$
$\widehat{\pi}^{\text{GBT2}}$	$279.75 \cdot 10^{-3}$

Table 2 Values of $\widehat{p}$ for $B = 500$. The null hypothesis $\mathcal{H}_0$: $\mathrm{E}\big[(t - Y)\,(\mathrm{I}[\widehat{\pi}_2(\boldsymbol{X}) > t] - \mathrm{I}[\widehat{\pi}_1(\boldsymbol{X}) > t])\big] \leq 0$ for all $t \in \Omega$ is rejected at the level 0.05 when $\widehat{p} < 0.05$

		$\widehat{\pi}_2$				
		μ	$\widehat{\pi}^{\text{GAM1}}$	$\widehat{\pi}^{\text{GAM2}}$	$\widehat{\pi}^{\text{GBT1}}$	$\widehat{\pi}^{\text{GBT2}}$
$\widehat{\pi}_1$	μ	/	0.000	0.000	0.000	0.272
	$\widehat{\pi}^{\text{GAM1}}$	1.000	/	0.000	0.908	0.960
	$\widehat{\pi}^{\text{GAM2}}$	1.000	0.000	/	0.924	0.962
	$\widehat{\pi}^{\text{GBT1}}$	1.000	0.000	0.000	/	0.976
	$\widehat{\pi}^{\text{GBT2}}$	0.980	0.000	0.000	0.000	/

- $\widehat{\pi}^{\text{GBT1}}$, using features X_1 (`age`) and X_6 (`area`) and `ID` $= 2$;
- $\widehat{\pi}^{\text{GBT2}}$, using all 8 available features and `ID` $= 2$.

We expect $\widehat{\pi}^{\text{GBT1}}$ to produce better results than with the GAMs because we allow for interaction effects between feature components `age` and `area`. Moreover, $\widehat{\pi}^{\text{GBT2}}$ is expected to perform better than the others since it uses all features, unlike the other models considered.

In Table 1, we compute out-of-sample estimates (on $\overline{\mathcal{D}}$) of the generalization errors (with the Poisson loss function) for the four models under consideration as well as for the true model. As expected, we notice that $\widehat{\pi}^{\text{GAM2}}$ slightly outperforms $\widehat{\pi}^{\text{GAM1}}$ and that $\widehat{\pi}^{\text{GBT1}}$ outperforms $\widehat{\pi}^{\text{GAM2}}$. Moreover, $\widehat{\pi}^{\text{GBT2}}$ has the lowest error among the four models considered. It is interesting to notice that the error of $\widehat{\pi}^{\text{GBT2}}$ is very closed to the one of the true model μ, so that $\widehat{\pi}^{\text{GBT2}}$ seems to be a good candidate to approximate μ.

5.3 Testing for Bregman dominance

Table 2 contains the values of $\widehat{p}$ of our testing procedure computed on the validation set $\overline{\mathcal{D}}$ with $B = 500$. We observe that the null hypothesis $\mathcal{H}_0$ is always rejected when $\widehat{\pi}_1 = \mu$, as expected, except for $\widehat{\pi}_2 = \widehat{\pi}^{\text{GBT2}}$. Similarly, $\mathcal{H}_0$ is always rejected when $\widehat{\pi}_1 = \widehat{\pi}^{\text{GBT2}}$ except for $\widehat{\pi}_2 = \mu$, as expected. Hence, for $\widehat{\pi}^{\text{GBT2}}$ and μ, our test cannot determine if either of the two models outperforms the other in terms of Bregman dominance. Moreover, for $\widehat{\pi}_1 = \widehat{\pi}^{\text{GAM1}}$ (resp. $\widehat{\pi}_1 = \widehat{\pi}^{\text{GAM2}}$), we do not reject $\mathcal{H}_0$, except for $\widehat{\pi}_2 = \widehat{\pi}^{\text{GAM2}}$ (resp. $\widehat{\pi}_2 = \widehat{\pi}^{\text{GAM1}}$). Therefore, for $\widehat{\pi}^{\text{GAM1}}$ and $\widehat{\pi}^{\text{GAM2}}$, one can say that neither model outperforms the other in terms of Bregman dominance. Finally, for $\widehat{\pi}_1 = \widehat{\pi}^{\text{GBT1}}$, $\mathcal{H}_0$ is rejected for $\widehat{\pi}_2 = \widehat{\pi}^{\text{GAM1}}$ and $\widehat{\pi}_2 = \widehat{\pi}^{\text{GAM2}}$ and is not rejected for $\widehat{\pi}_2 = \widehat{\pi}^{\text{GBT2}}$ and $\widehat{\pi}_2 = \mu$, which is not particularly surprising here.

Figure 1 displays Murphy diagrams for the extremal consistent loss functions $L_t(\widehat{\pi}(\boldsymbol{X}), Y)$ defined in (3). More precisely, it shows the estimates of $\mathrm{E}\big[L_t(\widehat{\pi}(\boldsymbol{X}), Y)\big]$ on the validation set $\overline{\mathcal{D}}$ for the models $\widehat{\pi}$ under consideration as well as for the true model μ. It is interesting to notice that Murphy diagrams almost coincide for $\widehat{\pi} = \widehat{\pi}^{\mathrm{GBT2}}$ and $\widehat{\pi} = \mu$, which confirms the difficulty of our test in distinguishing between these two models. Also, one sees that Murphy diagrams intersect for $\widehat{\pi} = \widehat{\pi}^{\mathrm{GAM1}}$ and $\widehat{\pi} = \widehat{\pi}^{\mathrm{GAM2}}$, which suggests that neither model dominates the other, as confirmed by our testing procedure. Notice that the difference in MSE can be obtained (up to a constant factor) as the area between the curves in Figure 1 as shown by (4).

Appendix

A Proof of Proposition 4.1

Let

$$g_n(t) := \sqrt{n} D_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (t - Y_i) \left(\mathrm{I}[\widehat{\pi}_2(\boldsymbol{X}_i) > t] - \mathrm{I}[\widehat{\pi}_1(\boldsymbol{X}_i) > t] \right).$$

For $t_1 < t_2$, since $D(t_1) = D(t_2) = 0$ and the random vectors $\{(Y_i, \boldsymbol{X}_i), i = 1, \ldots, n\}$ are i.i.d., it is easy to see that

$$\begin{aligned} \mathrm{E}[g_n(t_1) g_n(t_2)] = \mathrm{E}\Big[& (t_1 - Y) \left(\mathrm{I}[\widehat{\pi}_2(\boldsymbol{X}) > t_1] - \mathrm{I}[\widehat{\pi}_1(\boldsymbol{X}) > t_1] \right) \\ & (t_2 - Y) \left(\mathrm{I}[\widehat{\pi}_2(\boldsymbol{X}) > t_2] - \mathrm{I}[\widehat{\pi}_1(\boldsymbol{X}) > t_2] \right) \Big]. \end{aligned}$$

Note that

$$g_n(t) = t\sqrt{n}(F_2^{(n)}(t) - F_1^{(n)}(t)) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i \left(\mathrm{I}[\widehat{\pi}_2(\boldsymbol{X}_i) > t] - \mathrm{I}[\widehat{\pi}_1(\boldsymbol{X}_i) > t] \right)$$

where

$$F_j^{(n)}(t) := n^{-1} \sum_{i=1}^{n} \mathrm{I}[\widehat{\pi}_j(\boldsymbol{X}_i) > t], \quad j = 1, 2.$$

Since $t \to \frac{1}{n} \sum_{i=1}^{n} Y_i \mathrm{I}[\widehat{\pi}_j(\boldsymbol{X}_i) > t]$ is monotone, it is a VC-class of functions and it is therefore Donsker. Moreover, since Ω is compact, there exists a $C > 0$ such that

$$|t F_j^{(n)}(t)| \leq C F_j^{(n)}(t).$$

As the enveloppe of $t \to t F_j^{(n)}(t)$ is square integrable, Theorem 19.14 of [11] directly entails that it is Donsker. Now, since the sum of Donsker classes is Donsker (see for instance [10]), the result follows.

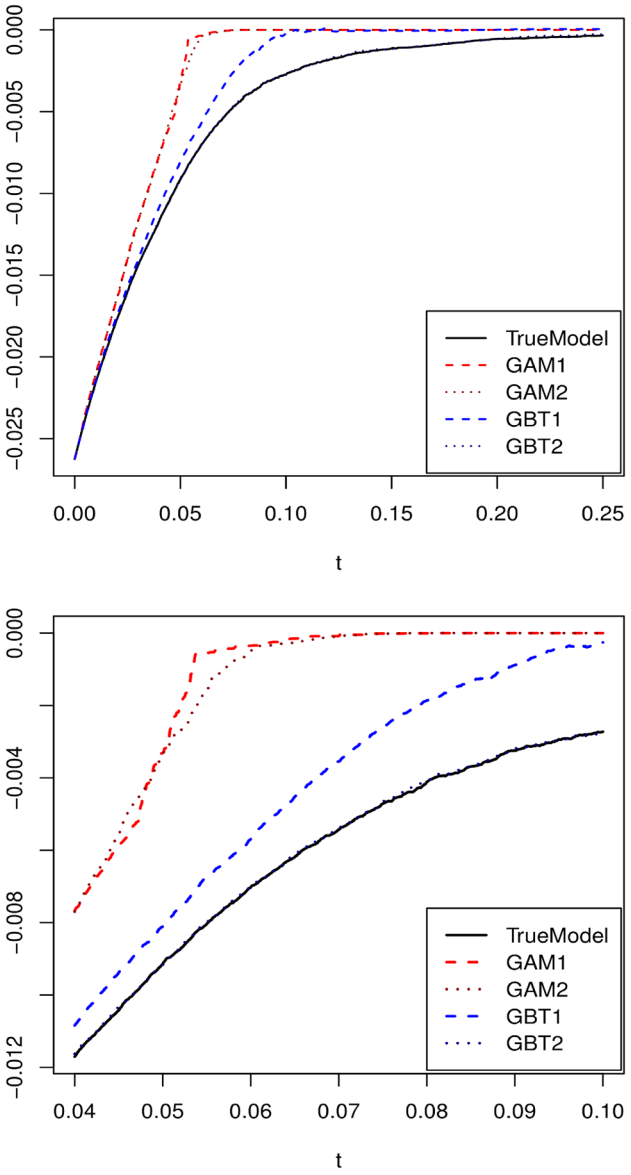


Fig. 1 Murphy diagrams for $\widehat{\pi}^{\mathrm{GAM1}}$ and $\widehat{\pi}^{\mathrm{GAM2}}$ in red, $\widehat{\pi}^{\mathrm{GBT1}}$ and $\widehat{\pi}^{\mathrm{GBT2}}$ in blue, and for μ in black

B Monte Carlo testing procedure

1. Generate B independent samples $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_B$, where the bth sample $\boldsymbol{U}_b = (U_{b1}, \ldots, U_{bn})$ contains n independent standard Gaussian random variables.

2. Compute

$$\widehat{p} := \frac{1}{B}\sum_{b=1}^{B} \mathrm{I}\left[\max_{t\in\Omega} \Delta\left(t, D_n(t), \boldsymbol{U}_b\right) > \max_{t\in\Omega} \sqrt{n} D_n(t)\right],$$

where

$$\Delta\left(t, D_n(t), \boldsymbol{U}_b\right) := \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\left(t - Y_i\right)\left(\mathrm{I}[\widehat{\pi}_2(\boldsymbol{X}_i) > t] - \mathrm{I}[\widehat{\pi}_1(\boldsymbol{X}_i) > t]\right) - D_n(t)\Big)\, U_{bi}.$$

3. Reject the null hypothesis at the level β when $\widehat{p} < \beta$.

C Simulated example

In this simulation exercise, $\mu(\boldsymbol{X})$, $\widehat{\pi}_1(\boldsymbol{X})$ and $\widehat{\pi}_2(\boldsymbol{X})$ are assumed to be uniformly distributed over $[a, b]$, with $b > a > 0$, $\mu(\boldsymbol{X})$ and $\widehat{\pi}_2(\boldsymbol{X})$ are supposed to be mutually independent and $\mu(\boldsymbol{X})$ and $\widehat{\pi}_1(\boldsymbol{X})$ are made dependent through the use of a Gaussian copula with a correlation coefficient $\rho \geq 0$. Therefore, since $\mathrm{E}\big[(t - Y)\mathrm{I}[\widehat{\pi}_k(\boldsymbol{X}) > t]\big] = \mathrm{E}\big[(t - \mu(\boldsymbol{X}))\mathrm{I}[\widehat{\pi}_k(\boldsymbol{X}) > t]\big]$ for $k = 1, 2$, we have

$$\begin{aligned}
&\mathrm{E}\big[(t - Y)\mathrm{I}[\widehat{\pi}_2(\boldsymbol{X}) > t]\big] \leq \mathrm{E}\big[(t - Y)\mathrm{I}[\widehat{\pi}_1(\boldsymbol{X}) > t]\big] \quad \text{for all} t \in [a, b] \\
&\Leftrightarrow \mathrm{E}\big[\mu(\boldsymbol{X})|\widehat{\pi}_2(\boldsymbol{X}) > t\big] \geq \mathrm{E}\big[\mu(\boldsymbol{X})|\widehat{\pi}_1(\boldsymbol{X}) > t\big] \quad \text{for all} t \in [a, b] \\
&\Leftrightarrow \mathrm{E}\big[\mu(\boldsymbol{X})\big] \geq \mathrm{E}\big[\mu(\boldsymbol{X})|\widehat{\pi}_1(\boldsymbol{X}) > t\big] \quad \text{for all} t \in [a, b].
\end{aligned}$$

When $\rho = 0$, we get $\mathrm{E}\big[\mu(\boldsymbol{X})|\widehat{\pi}_1(\boldsymbol{X}) > t\big] = \mathrm{E}\big[\mu(\boldsymbol{X})\big]$, so that this case coincides with the boundary of the null hypothesis. Increasing the value of ρ increases $\mathrm{E}\big[\mu(\boldsymbol{X})|\widehat{\pi}_1(\boldsymbol{X}) > t\big]$, so that the larger ρ, the further the data generating process is from the null hypothesis.

For this simulation exercise, we assume that Y given $\boldsymbol{X}$ is Poisson distributed with mean $\mu(\boldsymbol{X})$, which is a typical case in insurance when the response represents the number of claims made by a policyholder over one year for instance, and we suppose that $a = 0.05$ and $b = 0.5$. We consider $\rho = 0, 0.1, 0.5, 1$, and for each value of ρ, we generate $R = 1000$ independent samples $(Y_1, \widehat{\pi}_1(\boldsymbol{X}_1), \widehat{\pi}_2(\boldsymbol{X}_1)), \ldots,$ $(Y_n, \widehat{\pi}_1(\boldsymbol{X}_n), \widehat{\pi}_2(\boldsymbol{X}_n))$.

Figure 2 displays empirical rejection frequencies of the test performed using steps 1-3 of the Monte-Carlo procedure described in the previous section with $B = 500$ at the nominal level $\beta = 5\%$ for sample sizes $n = 500$, $n = 1000$ and $n = 1500$. The test satisfies the level constraint since the rejection frequencies for $\rho = 0$ are close

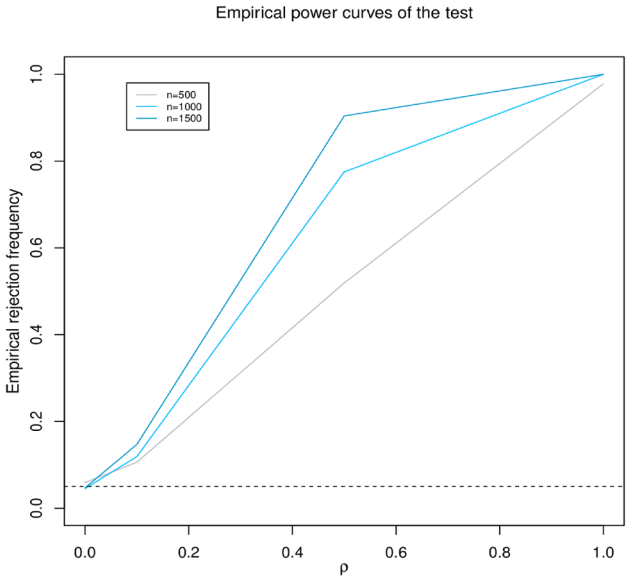


Fig. 2 Empirical power curves of the test for various sample sizes

to the nominal level 5%. Moreover, the testing procedure shows higher power as n increases.

Acknowledgements The authors would like to express their gratitude to the anonymous Referee and Handling Editor whose comments have been extremely useful to revise a previous version of the present work. Julien Trufin gratefully acknowledges the support received from the Research Chair ACTIONS under the aegis of the Risk Foundation, an initiative by BNP Paribas Cardif and the French Institute of Actuaries.

References

1. Barendse S, Patton AJ (2021) Comparing predictive accuracy in the presence of a loss function shape parameter. Journal of Business & Economic Statistics 40:1057–1069
2. Ehm W, Gneiting T, Jordan A, Kruger F (2016) Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings. Journal of the Royal Statistical Society - Series B 78:505–562
3. Ehm W, Kruger F (2018) Forecast dominance testing via sign randomization. Electronic Journal of Statistics 12:3758–3793
4. Fissler T, Lorentzen C, Mayer M (2022) Model comparison and calibration assessment: User guide for consistent scoring functions in machine learning and actuarial practice. arXiv preprint arXiv:2202.12780
5. Gneiting T (2011) Making and evaluating point forecasts. J Am Stat Assoc 106:746–762
6. Krüger F, Ziegel JF (2021) Generic conditions for forecast dominance. Journal of Business & Economic Statistics 39:972–983
7. Meyers G, Cummings AD (2009) "Goodness of Fit" vs. "Goodness of Lift". Actuarial Review 36:16–17
8. Patton AJ (2020) Comparing possibly misspecified forecasts. Journal of Business & Economic Statistics 38:796–809
9. Savage LJ (1971) Elicitation of personal probabilities and expectations. J Am Stat Assoc 66:783–810
10. van der Vaart AW (1996) New Donsker classes. Ann Probab 24:2128–2140
11. van der Vaart AW (2000) Asymptotic Statistics. Cambridge University Press

12. Wüthrich MV (2020) Bias regularization in neural network models for general insurance pricing. Eur Actuar J 10:179–202
13. Wüthrich, M.V., Buser, C. (2023). Data analytics for non-life insurance pricing. SSRN Manuscript ID 2870308, Version of June 19, 2023
14. Yen T-J, Yen Y-M (2021) Testing forecast accuracy of expectiles and quantiles with the extremal consistent loss functions. Int J Forecast 37:733–758
15. Zhu X, Guo X, Lin L, Zhu L (2016) Testing for positive expectation dependence. Ann Inst Stat Math 68:135–153
16. Ziegel JF, Krüger F, Jordan A, Fasciati F (2020) Robust forecast evaluation of expected shortfall. J Financ Economet 18:95–120