TESTING FOR MORE POSITIVE EXPECTATION DEPENDENCE WITH APPLICATION TO MODEL COMPARISON

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

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

Thomas Verdebout
Department of Mathematics and ECARES
Université Libre de Bruxelles (ULB)
Bruxelles, Belgium

July 13, 2021

Abstract

Modern data science tools are effective to produce predictions that strongly correlate with responses. Model comparison can therefore be based on the strength of dependence between responses and their predictions. Positive expectation dependence turns out to be attractive in that respect. The present paper proposes an effective testing procedure for this dependence concept and applies it to model selection. A simulation study is performed to evaluate the performances of the proposed testing procedure. Empirical illustrations using insurance loss data demonstrate the relevance of the approach for model selection in supervised learning. The most positively expectation dependent predictor can then be autocalibrated to obtain its balance-corrected version that appears to be optimal with respect to Bregman, or forecast dominance.

Keywords: Expectation dependence, concentration curve, Lorenz curve, autocalibration, convex order, balance correction.

1 Introduction and motivation

Several notions of dependence have been used in insurance studies, including quadrant, expectation and regression dependence. See, e.g., Denuit et al. (2005) for a detailed account of dependence structures and their links with stochastic dominance rules. Expectation dependence introduced by Wright (1987) has been shown to play a key role in many financial problems, such as asset allocation (Wright, 1987; Denuit and Eeckhoudt 2016), demand for risky asset under background risk (Li, 2011) and insurance under background risk (Hong et al., 2011; Li et al. 2016; Denuit and Mesfioui, 2017). This dependence concept can be traced back to Kowalczyk and Pleszczynska (1977) where it was termed as expectation quadrant dependence.

Positive expectation dependence expresses some form of positive relationship between two random variables Y and Z. It assesses the influence of Z on Y by specifying the impact of the information that Z is small (i.e. below some threshold z, say) on the expectation of Y. Precisely, Y is positively expectation dependent on Z if

$$\mathrm{E}[Y] \geq \mathrm{E}[Y|Z \leq z] \text{ for all } z \Leftrightarrow \mathrm{E}[Y|Z > z] \geq \mathrm{E}[Y] \text{ for all } z. \tag{1.1}$$

Negative expectation dependence is defined by reversing the sign of the inequalities appearing in (1.1). Notice that (1.1) is not symmetric in Y and Z so that expectation dependence distinguishes among two dimensions: a random variable Y of interest and the information provided by the auxiliary variable Z (as in regression problems).

In this paper, we apply expectation dependence to insurance pricing. Modern data science tools are effective to produce predictions that strongly correlate with responses. This is related to the submodularity of standard loss functions adopted in machine learning. Model comparison can therefore be based on the strength of dependence between responses and their predictions. Positive expectation dependence turns out to be attractive in that respect. Our approach builds on Denuit et al. (2019) who demonstrated that the variability of model predictions as well as the strength of their association with the response must both be taken into account to select the optimal pricing tool. These aspects are translated into mathematical terms with the help of convex order (probabilistic tool to assess the dispersion of random variables, beyond simple indicators such as standard deviations) and expectation dependence. The latter concept turns out to be closely related with concentration and Lorenz curves that are known to apply to model selection since Frees et al. (2011, 2014). According to Property 3.4 in Denuit et al. (2019), the concentration curve lies below the 45-degree line under positive expectation dependence.

The expectation of Y involved in the definition (1.1) for positive dependence can be considered as the conditional expectation of Y given $Z \leq z$ or given $Z > z$ when Y and Z are mutually independent (so that the condition does not modify the expected value of Y). Hence, it is natural to extend the concept to compare the strength of expectation dependence of Y on two random variables Z_1 and Z_2. This extension was proposed in the conclusion to Wright (1987). It appears to be relevant for the application considered in the present paper, since it allows the actuary to compare different supervised learning models to select the optimal one. In that setting, Y corresponds to the response under consideration while Z_1 and Z_2 correspond to ranks of model predictions under two competing insurance pricing tools (so that Z_1 and Z_2 are both uniformly distributed over the unit interval). Then, model

1 outperforms model 2 if the response Y is more positively expectation dependent on Z_1 than on Z_2, that is, if

$$\mathrm{E}[Y|Z_1 \leq z] \leq \mathrm{E}[Y|Z_2 \leq z] \text{ for all } z \Leftrightarrow \mathrm{E}[Y|Z_1 > z] \geq \mathrm{E}[Y|Z_2 > z] \text{ for all } z. \tag{1.2}$$

The inequalities in (1.2) express that the response is more reactive to the scores produced by model 1 compared to model 2: a small score (that is, below threshold z) makes the expected response smaller under model 1 compared to model 2 and a large score (that is, above threshold z) makes the expected response larger under model 1 compared to model 2.

In order to apply expectation dependence to model selection, we need to be able to test for this dependence concept. The problem of testing for positive expectation dependence consists in testing the null hypothesis $\mathcal{H}_0 : \mathrm{E}[Y|Z \leq z] \leq \mathrm{E}[Y]$ for all z against the alternative $\mathcal{H}_1 : \mathrm{E}[Y|Z \leq z] > \mathrm{E}[Y]$ for some z. Tests for expectation dependance have been proposed in Zhu et al. (2016), Cmiel and Ledwina (2017) and Linton et al. (2018). Guo and Li (2016) proposed a method to construct uniform confidence band for quantities defining expectation dependence, derived from Hoeffding's inequality. In this paper, we design a testing procedure for comparing the strength of expectation dependence of Y on Z_1 and on Z_2. The test is obtained by adapting Zhu et al. (2016) approach to the testing problem $\mathcal{H}_0 : \mathrm{E}[Y|Z_1 \leq z] \leq \mathrm{E}[Y|Z_2 \leq z]$ for all z against $\mathcal{H}_1 : \mathrm{E}[Y|Z_1 \leq z] > \mathrm{E}[Y|Z_2 \leq z]$ for some z for a triplet of random variables Y, Z_1 and Z_2 with Z_1 and Z_2 identically distributed.

The remainder of the paper is organized as follows. In Section 2, we explain the role of expectation dependence in model comparison. Section 3 presents the testing procedure for comparing the strength of expectation dependence between Y and Z_1 or Z_2. A simulation study is conducted to assess its performances. A case study is performed in Section 4 with a motor insurance data set to demonstrate the relevance of the proposed approach for model selection. The final Section 5 discusses the results and relates them to autocalibration discussed in Denuit et al. (2021).

2 Expectation dependence in supervised learning

2.1 Supervised learning

Consider a response Y and a set of features $X_1, \ldots, X_p$ gathered in the vector $\boldsymbol{X}$. The dependence structure inside the random vector $(Y, X_1, \ldots, X_p)$ is exploited to extract the information contained in $\boldsymbol{X}$ about Y. Often, the target is the conditional expectation $\mu(\boldsymbol{X}) = \mathrm{E}[Y|\boldsymbol{X}]$ of the response Y given the available information $\boldsymbol{X}$. This is the situation considered in the present paper. The function $\boldsymbol{x} \mapsto \mu(\boldsymbol{x}) = \mathrm{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$ is unknown and approximated by a predictor $\boldsymbol{x} \mapsto \pi(\boldsymbol{x})$ with a simpler structure. Predicting a response variable from a function of features and parameters is also referred to as supervised learning.

Models are generally calibrated so that a measure of the goodness-of-fit is optimized (deviance or log-likelihood, in most cases). Formally, the analyst selects a loss function $L(\cdot, \cdot)$ giving the error made when predicting the response. The empirical loss is then minimized on the training data set, maximizing goodness-of-fit or adherence to observed data. Cross validation is used for early-stopping rules, for instance. Model performances are then evaluated

on a separate data set. We refer the reader to Denuit et al. (2020) for a general presentation of this approach to insurance pricing.

We assume that $\pi(\boldsymbol{X})$ is a continuous random variable and we denote as

$$F_\pi(t) = \mathrm{P}[\pi(\boldsymbol{X}) \leq t], \;\; t \geq 0,$$

its distribution function and as F_π^{-1} the associated quantile function defined as the generalized inverse of F_π, i.e.

$$F_\pi^{-1}(\alpha) = \inf\{t \in \mathbb{R} | F_\pi(t) \geq \alpha\} \text{ for a probability level } \alpha.$$

2.2 Concentration curve

Following Frees et al. (2011, 2014) and Denuit et al. (2019, 2021), predictor performances are measured with the help of concentration curves. Recall that the concentration curve of the response Y with respect to the predictor $\pi(\cdot)$ based on the information contained in the vector $\boldsymbol{X}$ is defined as

$$\mathrm{CC}[Y, \pi(\boldsymbol{X}); \alpha] = \frac{\mathrm{E}\big[Y \mathrm{I}[\pi(\boldsymbol{X}) \leq F_\pi^{-1}(\alpha)]\big]}{\mathrm{E}[Y]} \text{ for a probability level } \alpha.$$

The interested reader is referred to the book by Yitzhaki and Schechtman (2013) for an exhaustive review of the properties of concentration curves. Notice that Lorenz curves (LC) correspond to concentration curves of a random variable with respect to itself.

It turns out that

$$\mathrm{CC}[Y, \pi(\boldsymbol{X}); \alpha] = \mathrm{CC}\big[\mu(\boldsymbol{X}), \pi(\boldsymbol{X}); \alpha\big] \text{for any probability level } \alpha. \tag{2.1}$$

Formula (2.1) shows that we can equivalently replace the response Y with the pure premium $\mu(\boldsymbol{X})$ in the concentration curve. Thus, the concentration curve assesses the dependence within the pair $(\mu(\boldsymbol{X}), \pi(\boldsymbol{X}))$, that is, between the target $\mu(\boldsymbol{X})$ and its predictor $\pi(\boldsymbol{X})$. But it can also be expressed in terms of observed response Y, wich appears to be useful for estimation.

The concentration curve can be equivalently rewritten as

$$\begin{aligned} \mathrm{CC}[Y, \pi(\boldsymbol{X}); \alpha] &= \frac{\mathrm{E}\big[Y \big| \pi(\boldsymbol{X}) \leq F_\pi^{-1}(\alpha)\big]}{\mathrm{E}[Y]} \times \alpha \\ &= \frac{\mathrm{Cov}\big[Y, \mathrm{I}[\pi(\boldsymbol{X}) \leq F_\pi^{-1}(\alpha)]\big]}{\mathrm{E}[Y]} + \alpha \end{aligned}$$

for every probability level α, where $\mathrm{I}[\cdot]$ denotes the indicator function, equal to 1 if the event appearing in the brackets is realized and to 0 otherwise. We refer the reader to Yitzhaki and Schechtman (2013) for the proofs.

The concentration curve $\alpha \mapsto \mathrm{CC}[Y, \pi(\boldsymbol{X}); \alpha]$ is defined from $\pi(\boldsymbol{X}) \leq F_\pi^{-1}(\alpha)$. This means that it is enough to consider the ranking induced by the predictor, that is, we are free to replace every predictor $\pi(\boldsymbol{X})$ with the corresponding rank

$$\Pi = F_\pi\big(\pi(\boldsymbol{X})\big)$$

obeying the unit uniform distribution. The concentration curve at probability level α can be rewritten as

$$\begin{aligned}
\mathrm{CC}[Y, \pi(\boldsymbol{X}); \alpha] &= \frac{\mathrm{E}\big[Y \mathrm{I}[\Pi \leq \alpha]\big]}{\mathrm{E}[Y]} \\
&= \frac{\mathrm{E}\big[Y \big| \Pi \leq \alpha\big]}{\mathrm{E}[Y]} \times \alpha \\
&= \frac{\mathrm{Cov}\big[Y, \mathrm{I}[\Pi \leq \alpha]\big]}{\mathrm{E}[Y]} + \alpha.
\end{aligned}$$

These expressions only involve the rank Π induced by the predictor under consideration.

2.3 Model comparison

Assume now that we have two predictors π_1 and π_2 for $\mu(\boldsymbol{X})$. These predictors may differ in their functional form (π_1 instead of π_2) and/or in the information ($\boldsymbol{X}_1$ instead of $\boldsymbol{X}_2$) on which they are based. The respective distribution functions of the two predictors π_1 and π_2 are denoted as F_{π_1} and F_{π_2}. Both F_{π_1} and F_{π_2} are assumed to be continuous and strictly increasing. Define $\Pi_1 = F_{\pi_1}(\pi_1)$ and $\Pi_2 = F_{\pi_2}(\pi_2)$ that are both uniformly distributed over the unit interval $[0, 1]$.

Better predictions result in a lower concentration curve, meaning that they induce a larger average decrease of the target when they fall below their αth quantile, for every probability level α. Thus,

$$\Pi_1 \text{ outperforms } \Pi_2 \Leftrightarrow \mathrm{E}[Y|\Pi_1 \leq \alpha] \leq \mathrm{E}[Y|\Pi_2 \leq \alpha] \text{ for all probability levels } \alpha. \tag{2.2}$$

This definition is in line with (1.2): model performances are assessed by the degree of expectation dependence of the response with the predictor. In words, (2.2) means that the reduction in the expectation resulting from the knowledge that $\Pi_k \leq \alpha$ is larger for Π_1 compared to Π_2. The ranking of the concentration curves amounts to requiring that Y is more positively expectation dependent on Π_1 than on Π_2 so that (2.2) can be equivalently stated in terms of concentration curves.

3 Testing for more positive expectation dependence

The problem under investigation can be stated as follows: we have a random variable Y and two random variables Π_1 and Π_2 that are both uniformly distributed over the unit interval $[0, 1]$ and possibly correlated between each other and with Y. We observe n realizations of these random variables. Let $(Y_1, \Pi_{11}, \Pi_{21}), \ldots, (Y_n, \Pi_{1n}, \Pi_{2n})$ be the corresponding triplets, assumed to be independent copies of (Y, Π_1, Π_2). We want to test the null hypothesis $\mathcal{H}_0 : \mathrm{E}[Y|\Pi_1 \leq \alpha] \leq \mathrm{E}[Y|\Pi_2 \leq \alpha]$ for all $\alpha \in (0, 1)$ against the alternative $\mathcal{H}_1 : \mathrm{E}[Y|\Pi_1 \leq \alpha] > \mathrm{E}[Y|\Pi_2 \leq \alpha]$ for some $\alpha \in (0, 1)$. The null hypothesis thus corresponds to (2.2) and supports predictor π_1. Now since both Π_1 and Π_2 have the same distribution we readily have the null hypothesis $\mathcal{H}_0$ is equivalent to

$$\mathrm{E}[\mathrm{I}[\Pi_1 \leq \alpha]](\mathrm{E}[Y|\Pi_1 \leq \alpha] - \mathrm{E}[Y]) \leq \mathrm{E}[\mathrm{I}[\Pi_2 \leq \alpha]](\mathrm{E}[Y|\Pi_2 \leq \alpha] - \mathrm{E}[Y]) \text{ for all } \alpha \in (0, 1),$$

so that $\mathcal{H}_0$ is also equivalent to

$$\mathrm{Cov}\big[Y, \mathrm{I}[\Pi_1 \le \alpha] - \mathrm{I}[\Pi_2 \le \alpha]\big] \le 0 \text{ for all } \alpha \in (0,1).$$

Letting

$$D(\alpha) := \mathrm{Cov}\big[Y, \mathrm{I}[\Pi_1 \le \alpha] - \mathrm{I}[\Pi_2 \le \alpha]\big],$$

the most natural estimator of $D(\alpha)$ is obtained by computing an empirical covariance based on the observed sequences; that is,

$$\widehat{D}(\alpha) \;:=\; \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})(\mathrm{I}[\Pi_{1i} \le \alpha] - \mathrm{I}[\Pi_{2i} \le \alpha] - (\overline{\mathrm{I}[\Pi_1 \le \alpha]} - \overline{\mathrm{I}[\Pi_2 \le \alpha]}))$$

where $\bar{Y} := n^{-1}\sum_{i=1}^{n} Y_i$ and $\overline{\mathrm{I}[\Pi_k \le \alpha]} := n^{-1}\sum_{i=1}^{n}\mathrm{I}[\Pi_{ki} \le \alpha]$, $k = 1, 2$. We then consider a test that rejects the null hypothesis at level $\beta \in (0,1)$ when

$$T_n := \sup_{\alpha\in(0,1)} \sqrt{n}\widehat{D}(\alpha) > \xi_\beta,$$

where the critical value ξ_β can be obtained by studying the limiting behavior of the empirical process $\sqrt{n}(\widehat{D}(\alpha) - D(\alpha))$ under the null hypothesis. We have the following result.

Proposition 3.1. *Provided that* $\mathrm{E}[Y_i^2] < \infty$*, we have that when* $D(\alpha) = 0$ *(at the boundary between the null hypothesis and the alternative), the empirical process* $\sqrt{n}\widehat{D}(\alpha)$ *converges weakly to a Gaussian process with mean zero and covariance function*

$$\Sigma(\alpha_1, \alpha_2) := \mathrm{E}\big[(Y - \mathrm{E}[Y])^2(\mathrm{I}[\Pi_1 \le \alpha_1] - \mathrm{I}[\Pi_2 \le \alpha_1])(\mathrm{I}[\Pi_1 \le \alpha_2] - \mathrm{I}[\Pi_2 \le \alpha_2])\big]. \tag{3.1}$$

The proof of Proposition 3.1 is given in the appendix. Note that (3.1) is equal to 0 when α_1 or α_2 is equal to 0 or 1. This is in line with the fact that for $\alpha = 0$ or $\alpha = 1$, $\sqrt{n}\widehat{D}(\alpha)$ has variance zero.

Now, it directly follows from Proposition 3.1 and the continuous mapping theorem that a Kolmogorov-Smirnov type test can be obtained by rejecting the null hypothesis $\mathcal{H}_0$ at the asymptotic level $\beta \in (0,1)$ when

$$T_n = \sup_{\alpha\in(0,1)} \sqrt{n}\widehat{D}(\alpha) > c_\beta, \tag{3.2}$$

where the critical value c_β can in principle be obtained from Proposition 3.1. Although Proposition 3.1 shows that T_n in (3.2) is a very natural test statistic for the problem considered, the computation of the critical value c_β is delicate since it requires the computation of $\Sigma(\alpha_1, \alpha_2)$ which is not realistic in practice (in particular, we do not know the distribution of Y). Therefore, we suggest a Monte-Carlo procedure to perform the test that follows along the same lines as in Zhu et al. (2016):

1. Generate M independent samples $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_M$, where the mth sample $\boldsymbol{U}_m = (U_{m1}, \ldots, U_{mn})$ contains n independent standard Gaussian random variables.

2. Compute

$$\widehat{p} := \frac{1}{M}\sum_{m=1}^{M} \mathrm{I}\left[\max_{0\le\alpha\le1} \Delta\left(\alpha, \widehat{D}(\alpha), \boldsymbol{U}_m\right) > \max_{0\le\alpha\le1} \sqrt{n}\widehat{D}(\alpha)\right],$$

where

$$\Delta\left(\alpha, \widehat{D}(\alpha), \boldsymbol{U}_m\right) := \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left((Y_i - \bar{Y})(\mathrm{I}[\Pi_{1i} \le \alpha] - \mathrm{I}[\Pi_{2i} \le \alpha]) - \widehat{D}(\alpha)\right) U_{mi}.$$

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

We conclude the section with a simulation exercise whose objective is to show that the test performed using the steps 1-3 above reaches the correct nominal β-level constraint and enjoys power under the alternative. In this simulation exercise, we generated $R = 1000$ independent samples of the form

$$(Y_1^\ell, \Pi_{11}, \Pi_{21}), \ldots, (Y_n^\ell, \Pi_{1n}, \Pi_{2n}), \quad \ell = 0, 1, 2, 3,$$

via the following schemes:

(i) the Π_{1j}'s and the Π_{2j}'s are mutually independent and uniformly distributed over $[0, 1]$. The random variables $Y_1^\ell, \ldots, Y_n^\ell$ are obtained as

$$Y_j^\ell = 2\ell\Pi_{2j} + V_j,$$

where V_j is uniform on $[-1, 1]$ and independent from Π_{2j} (and Π_{1j}). Clearly, $\mathrm{E}[Y_j^\ell] = \ell = \mathrm{E}[Y_j^\ell|\Pi_{1j} \le \alpha]$ for all α while $\mathrm{E}[Y_j^0|\Pi_{2j} \le \alpha] = 0 = \mathrm{E}[Y_j^0|\Pi_{1j} \le \alpha]$ and $\mathrm{E}[Y_j^\ell|\Pi_{2j} \le \alpha] \le \ell$ for all α and $\ell = 1, 2, 3$. The value $\ell = 0$ therefore coincides with the (boundary) of the null hypothesis while the values $\ell = 1, 2, 3$, provide data generating processes that are increasingly under the alternative.

(i) the Π_{1j}'s are i.i.d. uniform over $[0, 1]$, while the Π_{2j}'s are defined as

$$\Pi_{2j} = \Pi_{1j}\mathrm{I}[\Pi_{1j} \le \frac{1}{2}] + (\frac{3}{2} - \Pi_{1j})\mathrm{I}[\Pi_{1j} > \frac{1}{2}].$$

It is easy to check that the Π_{2j}'s are uniform over $[0, 1]$ and correlated with Π_{1j}'s. The random variables $Y_1^\ell, \ldots, Y_n^\ell$ are obtained as

$$Y_j^\ell = 2\ell\Pi_{2j} + V_j,$$

where V_j is uniform on $[-1, 1]$ and independent from Π_{2j} (and Π_{1j}). As in the first scheme above, the value $\ell = 0$ therefore coincides with the (boundary) of the null hypothesis while the values $\ell = 1, 2, 3$, provide data generating processes that are increasingly under the alternative.

(iii) the Π_{1j}'s and the Π_{2j}'s are mutually independent and uniformly distributed over $[0,1]$. The random variables $Y_1^\ell, \ldots, Y_n^\ell$ are obtained as

$$Y_j^\ell = \ell \Pi_{1j} + 2\ell \Pi_{2j} + V_j,$$

where V_j is uniform on $[-1,1]$ and independent from Π_{2j} (and Π_{1j}). Clearly, $\mathrm{E}[Y_j^0 | \Pi_{2j} \leq \alpha] = 0 = \mathrm{E}[Y_j^0 | \Pi_{1j} \leq \alpha]$ while $\mathrm{E}[Y_j^\ell | \Pi_{2j} \leq \alpha] \leq \mathrm{E}[Y_j^\ell | \Pi_{1j} \leq \alpha]$ for all α and $\ell = 1,2,3$. The value $\ell = 0$ therefore coincides with the (boundary) of the null hypothesis while the values $\ell = 1,2,3$, provide data generating processes that are increasingly under the alternative.

Figures 3.1, 3.2 and 3.3 display the empirical rejection frequencies of our test performed using steps 1-3 above with M=500 at the nominal level $\beta = .05$ for various sample sizes. Note that the maxima over $(0,1)$ required in step 2 of the Algorithm have been obtained by taking maxima over grids of the form $(0, \frac{1}{100}, \frac{2}{100}, \ldots, 1)$. Inspection of the various Figures clearly reveals that the proposed testing procedure is valid in all the sampling schemes; it reaches the correct nominal level constraint in the various scenarii involving different dependance structures between the random variables. It furthermore clearly shows power.

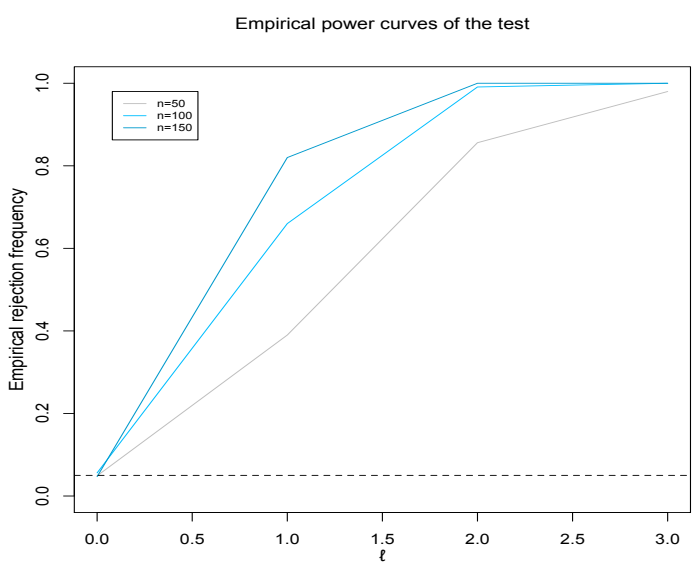

Figure 3.1: Empirical power curves of the test in the sampling scheme (i) for various sample sizes.

4 Case study

4.1 Data set

We consider the motor third-party liability insurance portfolio used in Denuit et al. (2020). It relates to an insurance company operating in the EU that has been observed during one

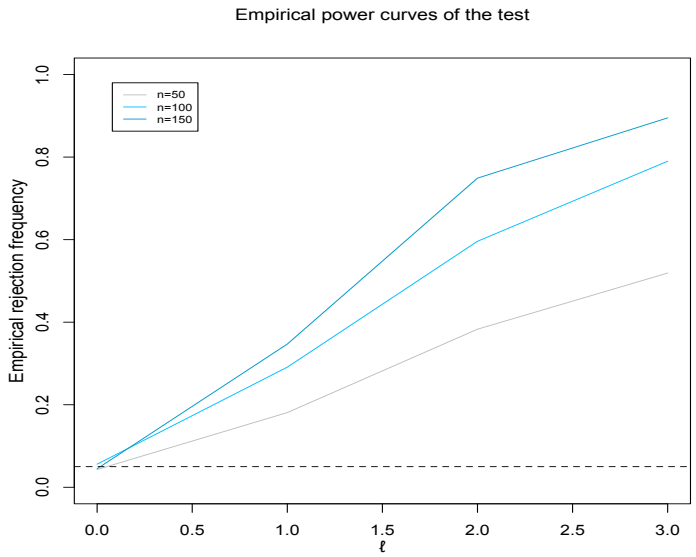

Figure 3.2: Empirical power curves of the test in the sampling scheme (ii) for various sample sizes.

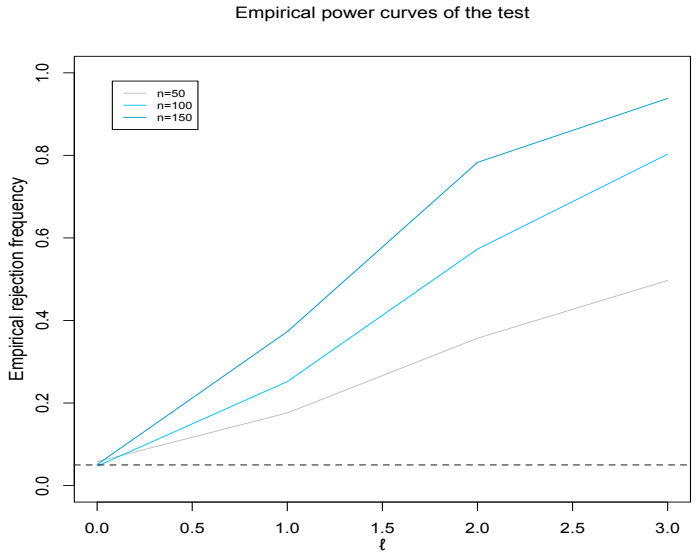

Figure 3.3: Empirical power curves of the test in the sampling scheme (iii) for various sample sizes.

Number of claims	Exposure-to-risk
0	126 499.7
1	15 160.4
2	1424.9
3	145.4
4	14.3
5	1.4
≥ 6	0

Table 4.1: Descriptive statistics for the number of claims.

year. The portfolio comprises 160 944 insurance policies. For each policy i, the data set contains the numbers of claims Y_i filed by policyholder i, the corresponding exposure-to-risk $e_i \leq 1$ (expressed in policy-year), and the following eight features $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{i8})$:

- X_{i1} = AgePh: policyholder's age;
- X_{i2} = AgeCar: age of the car;
- X_{i3} = Fuel: fuel of the car, with two categories (gas or diesel);
- X_{i4} = Split: splitting of the premium, with four categories (annually, semi-annually, quarterly or monthly);
- X_{i5} = Cover: extent of the coverage, with three categories (from compulsory third-party liability cover to comprehensive);
- X_{i6} = Gender: policyholder's gender, with two categories (female or male);
- X_{i7} = Use: use of the car, with two categories (private or professional);
- X_{i8} = PowerCat: the engine's power, with five categories.

Figure 4.1 displays the exposure-to-risk by category/value for each of the eight features and Table 4.1 shows the observed numbers of claims with corresponding exposures-to-risk.

We partition the data set into a training set $\mathcal{D}$ and a validation set $\overline{\mathcal{D}}$. The training set $\mathcal{D}$ is composed of 80% of the observations taken at random from the entire data set and the validation set $\overline{\mathcal{D}}$ is made of the 20% remaining observations. We refer the reader to Denuit et al. (2020) for more details about the data set, the notions of training and validation sets as well as for the regression techniques used throughout this section.

4.2 Models under consideration

The observations are assumed to be independent. Given $\boldsymbol{X} = \boldsymbol{x}$ and the exposure-to-risk e, the response Y is assumed to be Poisson distributed with mean $e\mu(\boldsymbol{x})$. So, $\mu(\boldsymbol{x}_i)$ represents the expected annual claim frequency for policyholder i. We aim to estimate the unknown function $\boldsymbol{x} \mapsto \mu(\boldsymbol{x})$.

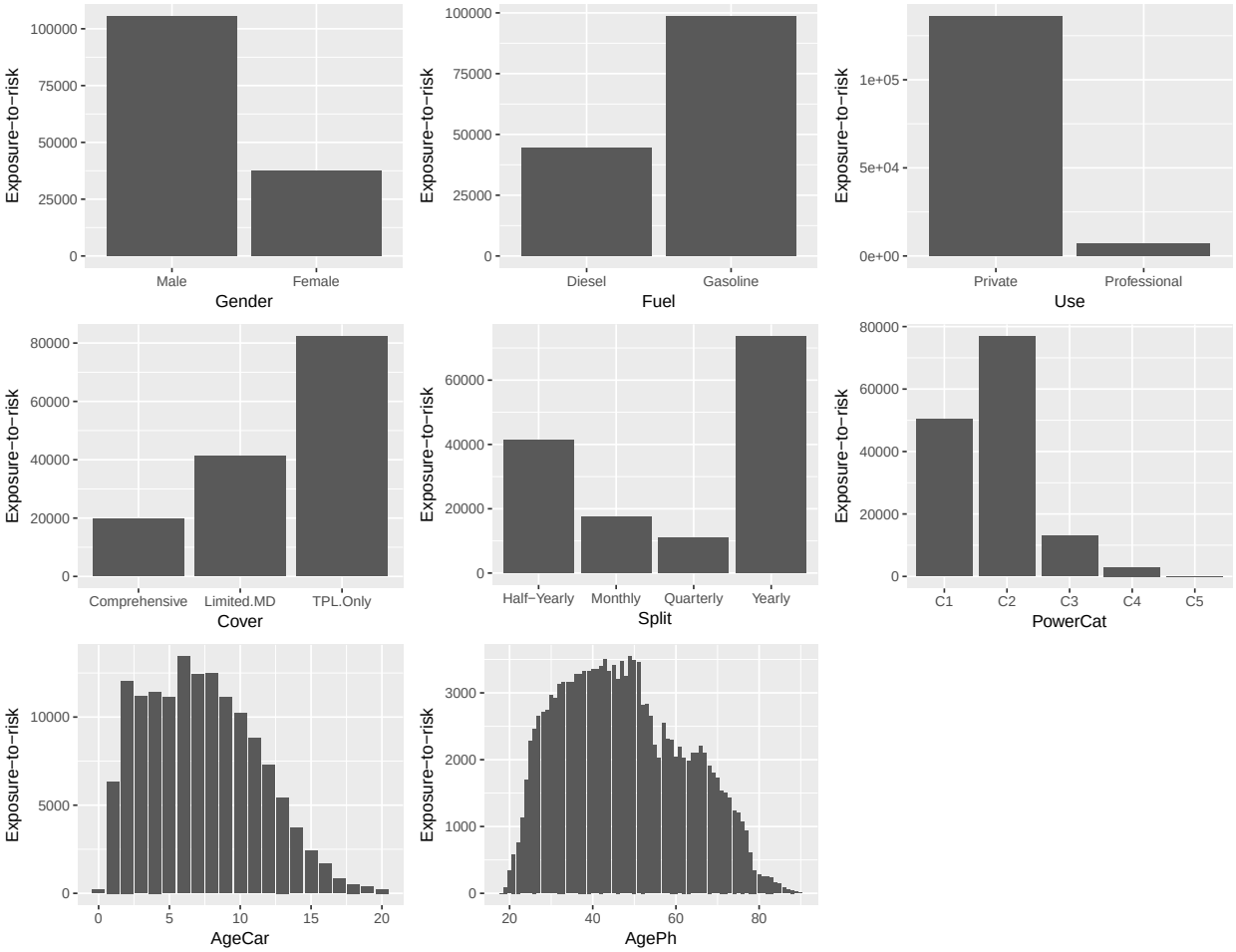

Figure 4.1: Levels/values of the features and corresponding exposures-to-risk.

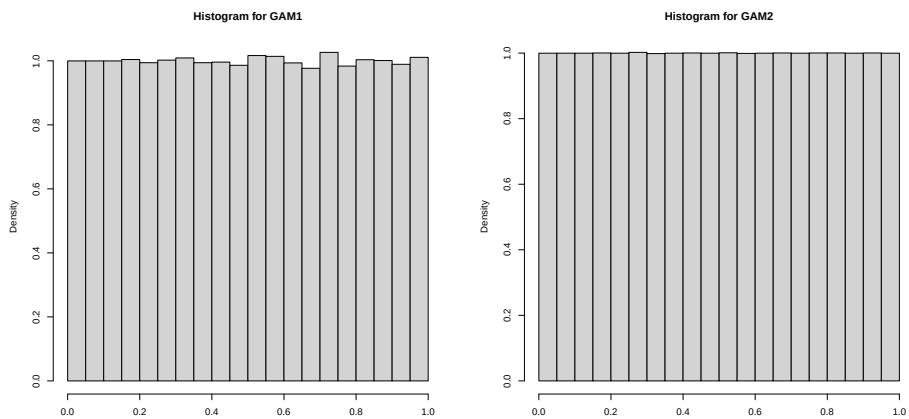

Figure 4.2: Histograms for Π^{GAM1} (left panel) and Π^{GAM2} (right panel) estimated from $\overline{\mathcal{D}}$.

To that end, we first 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:

- $\pi^{\text{GAM1}}(\boldsymbol{x})$, with only two features, namely X_1 (AgePh) and X_2 (AgeCar);
- $\pi^{\text{GAM2}}(\boldsymbol{x})$, using all 8 available features.

We expect the second model $\pi^{\text{GAM2}}(\boldsymbol{x})$ to clearly outperform the more simple one $\pi^{\text{GAM1}}(\boldsymbol{x})$ as variables such as X_4 (Split) or X_8 (PowerCat) are also known to be important in this setting (as shown for instance in Section 5.8 in Denuit et al. (2020)). In both models, the effects of the covariates AgePh and AgeCar are captured by splines and we do not consider interaction terms. We denote by Π^{GAM1} and Π^{GAM2} the ranks corresponding to $e\pi^{\text{GAM1}}(\boldsymbol{x})$ and $e\pi^{\text{GAM2}}(\boldsymbol{x})$, respectively. Figure 4.2 depicts their distribution estimated on the validation set $\overline{\mathcal{D}}$.

Then, we fit gradient boosting trees (GBT) on $\mathcal{D}$ with Poisson deviance loss and log-link function with the help of the `R` package `gbm`. We expect to get better results than with the GAMs because we allow for interaction effects between feature components. The bagging fraction γ (that is, the fraction of observations randomly selected from the training set to fit the next tree in the expansion), is set at 0.5. The shrinkage parameter τ is set at the low value 0.01. The size of the trees is controlled by the interaction depth `ID`, and we consider four values for `ID`, namely $\texttt{ID} = 1, 2, 3, 4$. The training set $\mathcal{D}$ is divided into two sets, $\mathcal{D}_1$ and $\mathcal{D}_2$, where $\mathcal{D}_1$ comprises 80% of the observations of $\mathcal{D}$. We train the GBT on $\mathcal{D}_1$. Figure 4.3 displays the in-sample (computed on $\mathcal{D}_1$) and out-of-sample (computed on $\mathcal{D}_2$) estimates of the generalization error for the GBT with $\texttt{ID} = 1, 2, 3, 4$ against the number of trees T. The out-of-sample estimates of the generalization error for $\texttt{ID} = 1, 2, 3, 4$ is minimized for $T = 1186, 1043, 633, 721$, respectively. We denote by $\pi^{\text{GBT1}}(\boldsymbol{x})$, $\pi^{\text{GBT2}}(\boldsymbol{x})$, $\pi^{\text{GBT3}}(\boldsymbol{x})$ and $\pi^{\text{GBT4}}(\boldsymbol{x})$ the GBT corresponding to $(\texttt{ID} = 1, T = 1186)$, $(\texttt{ID} = 2, T = 1043)$, $(\texttt{ID} = 3, T = 633)$ and $(\texttt{ID} = 4, T = 721)$, respectively. We turn the predictions $\pi^{\text{GBT1}}(\boldsymbol{x})$, $\pi^{\text{GBT2}}(\boldsymbol{x})$, $\pi^{\text{GBT3}}(\boldsymbol{x})$ and $\pi^{\text{GBT4}}(\boldsymbol{x})$ on the validation set into the corresponding ranks Π^{GBT1}, Π^{GBT2}, Π^{GBT3} and Π^{GBT4} by transforming them using the distribution function of the predictors estimated from the

π^{GAM1}	$549.93 \cdot 10^{-3}$
π^{GAM2}	$548.05 \cdot 10^{-3}$
π^{GBT1}	$545.06 \cdot 10^{-3}$
π^{GBT2}	$544.54 \cdot 10^{-3}$
π^{GBT3}	$544.29 \cdot 10^{-3}$
π^{GBT4}	$544.30 \cdot 10^{-3}$

Table 4.2: Out-of-sample estimates (on $\overline{\mathcal{D}}$) of the generalization error.

training set. Their distributions on $\overline{\mathcal{D}}$ are shown in Figure 4.4. Conditionally on the training set, the ranks Π^{GBT1}, Π^{GBT2}, Π^{GBT3} and Π^{GBT4} can be seen as independent realizations so that we are in a position to apply the testing procedure described in Section 3.

In Table 4.2, we compute out-of-sample estimates (on $\overline{\mathcal{D}}$) of the generalization errors for the six models under consideration. As expected, π^{GAM2} outperforms π^{GAM1}, which confirms the importance of some additional features used in π^{GAM2} as highlighted in Denuit et al. (2020). Moreover, we notice that π^{GBT1} outperforms π^{GAM2}. Finally, π^{GBT3} and π^{GBT4} have the lowest errors, which may indicate that there are three-way or even four-way interactions between the features that have not been captured in the other models.

4.3 Testing procedure

The conditional expectation $\mathrm{E}[Y|\Pi \leq \alpha]$ can be estimated on $\overline{\mathcal{D}}$ as

$$\widehat{\mathrm{E}}[Y|\Pi \leq \alpha] = \frac{\sum_{i \in \overline{\mathcal{D}}} y_i \mathrm{I}[\Pi(\boldsymbol{x}_i) \leq \alpha]}{\sum_{i \in \overline{\mathcal{D}}} \mathrm{I}[\Pi(\boldsymbol{x}_i) \leq \alpha]}.$$

In Figure 4.5, we display $\alpha \mapsto \widehat{\mathrm{E}}[Y|\Pi \leq \alpha]$ for the models under consideration.

Table 4.3 contains the values of $\widehat{p}$ computed with $M = 500$. The null hypothesis $\mathcal{H}_0 : \mathrm{E}[Y|\Pi_1 \leq \alpha] \leq \mathrm{E}[Y|\Pi_2 \leq \alpha]$ for all $\alpha \in (0,1)$ is always rejected for $\Pi_1 = \Pi^{\text{GAM1}}$ whatever Π_2. This shows that the first GAM model with only two features in inferior to all other models under consideration because it fails to produce more positively expectation dependent predictions. For $\Pi_1 = \Pi^{\text{GAM2}}$, the same observation holds at the level 0.05 except for $\Pi_2 = \Pi^{\text{GAM1}}$. This shows that GBT outperforms GAMs on this data set. The testing procedure does not identify one GBT dominating the others. This confirms the similar performances of all GBTs on $\overline{\mathcal{D}}$.

5 Discussion

In the present paper, we have shown that model comparison can be based on the strength of dependence between responses and their predictions. Positive expectation dependence is adopted to that end and an effective testing procedure is proposed for model selection. Numerical illustrations with both simulated and real data demonstrate the relevance of the approach. It is worth mentioning that model comparisons are based here on a formal test. This is in contrast with the classical cross-validation approach that consists in estimating

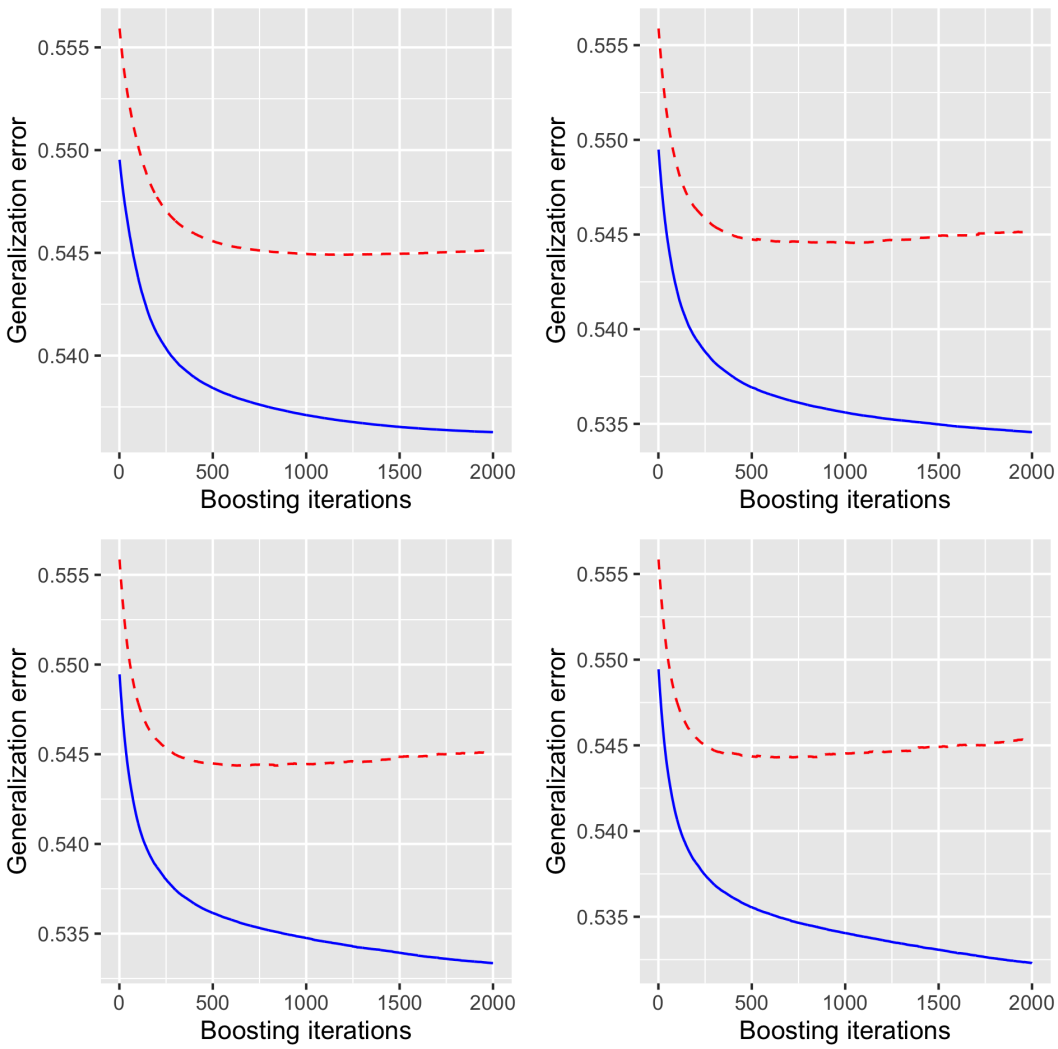

Figure 4.3: In-sample (solid blue line) and out-of-sample (dotted red line) estimates of the generalization error for `ID` = 1 (top-left), `ID` = 2 (top-right), `ID` = 3 (bottom-left) and `ID` = 4 (bottom-right).

		Π_2					
		Π^{GAM1}	Π^{GAM2}	Π^{GBT1}	Π^{GBT2}	Π^{GBT3}	Π^{GBT4}
Π_1	Π^{GAM1}	/	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
	Π^{GAM2}	0.998	/	**0.049**	**0.008**	**0.022**	**0.010**
	Π^{GBT1}	1.000	0.710	/	0.420	0.256	0.232
	Π^{GBT2}	0.998	0.856	0.990	/	0.902	0.250
	Π^{GBT3}	1.000	0.616	1.000	0.792	/	0.230
	Π^{GBT4}	0.998	0.806	1.000	0.910	0.964	/

Table 4.3: Values of $\widehat{p}$ for $M = 500$. The null hypothesis $\mathcal{H}_0 : \mathrm{E}[Y|\Pi_1 \leq \alpha] \leq \mathrm{E}[Y|\Pi_2 \leq \alpha]$ for all $\alpha \in (0, 1)$ is rejected at the level 0.05 when $\widehat{p} < 0.05$ (cases printed in bold in the table).

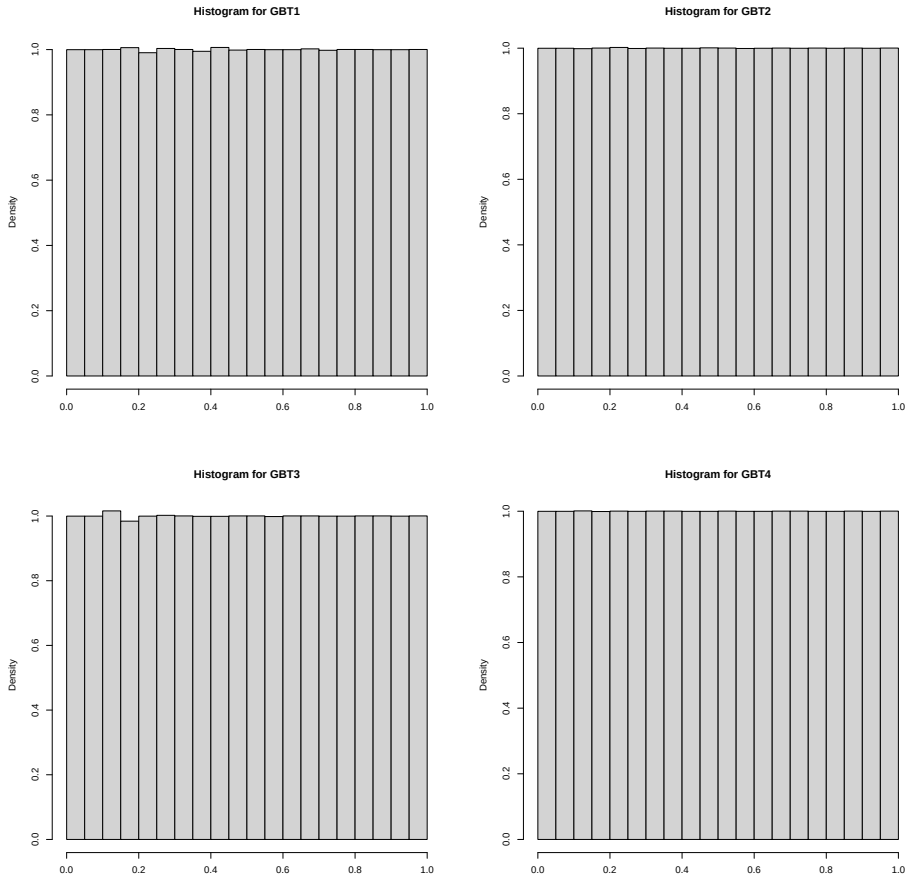

Figure 4.4: Distribution functions for Π^{GBT1} (top-left), Π^{GBT2} (top-right), Π^{GBT3} (bottom-left) and Π^{GBT4} (bottom-right) estimated on $\overline{\mathcal{D}}$.

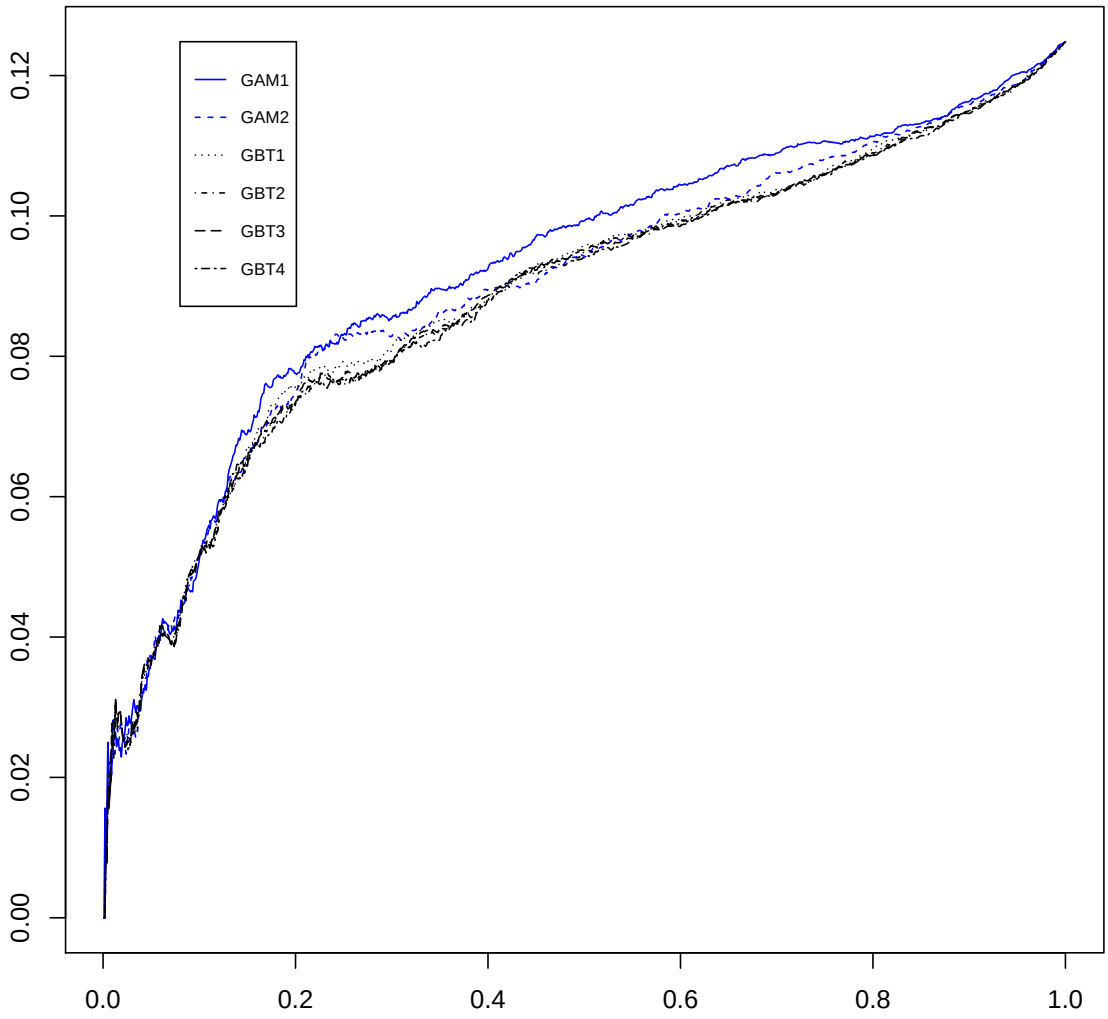

Figure 4.5: Conditional expectations $\alpha \to \widehat{\mathrm{E}}[Y|\Pi \leq \alpha]$ for Π^{GAM1} and Π^{GAM2} in blue, and for Π^{GBT1}, Π^{GBT2}, Π^{GBT3} and Π^{GBT4} in black.

the generalization errors of two different models and select the model with the smaller cross-validation error. Here, we perform the comparison on a validation set by applying a formal testing procedure for comparing the strength of expectation dependence of the response on the predictors of the models under consideration. The rejection of the null hypothesis yields to the conclusion that one model is significantly superior than the other. Cross-validation and the test for positive expectation dependence should both be used as diagnostic tools, providing the analyst with complementary comparison criteria for models under consideration (as it can be seen from Tables 4.2-4.3 in the numerical illustration performed on motor insurance data).

The approach proposed in this paper appears to be closely related to Bregman, or forecast dominance, as well as to autocalibration, as discussed next. The more Π_k is correlated to Y, the more information the corresponding predictor π_k contains. More informative predictors thus lead to greater variability of the conditional expectation $\mathrm{E}[Y|\Pi_k]$. Assessing the performances of predictors can thus also be based on the comparison of the variability of these conditional expectations. In that respect, the convex order is often used in applied probability to compare the variability inherent to probability distributions beyond standard deviations. It is therefore a natural candidate to assess the variability of conditional expectations $\mathrm{E}[Y|\Pi_k]$. Recall that a random variable Z_1 is said to be smaller than another random variable Z_2 in the convex order, henceforth denoted as $Z_1 \preceq_{\mathrm{cx}} Z_2$, if

$$\mathrm{E}[Z_1] = \mathrm{E}[Z_2] \text{ and } \mathrm{E}[(Z_1 - t)_+] \leq \mathrm{E}[(Z_2 - t)_+] \text{ for all } t \in \mathbb{R}.$$

The name convex order comes from the fact that $Z_1 \preceq_{\mathrm{cx}} Z_2 \Leftrightarrow \mathrm{E}[g(Z_1)] \leq \mathrm{E}[g(Z_2)]$ for all the convex functions g for which the expectations exist. For more details, we refer the reader e.g. to Denuit et al. (2005) or Shaked and Shanthikumar (2007).

It is easy to see that

$$Z_1 \preceq_{\mathrm{cx}} Z_2 \Rightarrow \mathrm{Var}[Z_1] \leq \mathrm{Var}[Z_2]. \tag{5.1}$$

This explains why $\preceq_{\mathrm{cx}}$ is a variability order: it only applies to random variables with the same expected value and compares the dispersion of these variables. The convex order is a more sophisticated comparison than only focusing on the variances, yet (5.1) indicates that it agrees with this approach. Henceforth, we can interpret $Z_1 \preceq_{\mathrm{cx}} Z_2$ as "Z_2 is more variable than Z_1", keeping in mind that the variability in question extends beyond the simple comparison of standard deviation.

Here, $\mathrm{E}[Y|\Pi_k]$ measures how the rank Π_k induced by the predictor π_k explains the response Y. Therefore, we consider that Π_1 is more informative than Π_2 if $\mathrm{E}[Y|\Pi_2] \preceq_{\mathrm{cx}} \mathrm{E}[Y|\Pi_1]$. This ensures that the mean square error of prediction (MSEP) is smaller with Π_1 compared to Π_2:

$$\mathrm{E}[Y|\Pi_2] \preceq_{\mathrm{cx}} \mathrm{E}[Y|\Pi_1] \Rightarrow \mathrm{E}\Big[\big(Y - \mathrm{E}[Y|\Pi_1]\big)^2\Big] \leq \mathrm{E}\Big[\big(Y - \mathrm{E}[Y|\Pi_2]\big)^2\Big],$$

that is, Y is closer to $\mathrm{E}[Y|\Pi_1]$ in the L^2-norm. The literature about auction theory says that Π_1 is more integral precise than Π_2 in such a case. See Ganuza and Penalva (2010).

In their study of dependence orderings based on generalized Lorenz curves, Muliere and Petrone (1992) established that, provided the functions $\alpha \mapsto \mathrm{E}[Y|\Pi_k = \alpha]$ are continuous and strictly increasing for $k \in \{1, 2\}$,

$$\mathrm{E}[Y|\Pi_2] \preceq_{\mathrm{cx}} \mathrm{E}[Y|\Pi_1] \Leftrightarrow \mathrm{E}[Y|\Pi_1 \geq \alpha] \geq \mathrm{E}[Y|\Pi_2 \geq \alpha] \text{ for all } \alpha. \tag{5.2}$$

Notice that the condition appearing in (5.2) corresponds to (2.2) since the identity

$$\mathrm{E}[Y] = \alpha \mathrm{E}[Y|\Pi_k \leq \alpha] + (1-\alpha)\mathrm{E}[Y|\Pi_k > \alpha]$$

holds for $k \in \{1, 2\}$ and all probability levels α. See also Denuit (2010). The test to check for the assumption $\mathrm{E}[Y|\Pi_2] \preceq_{\mathrm{cx}} \mathrm{E}[Y|\Pi_1]$ can thus also be based on the procedure proposed for testing for more positive expectation dependence.

Recall that a predictor π is said to be autocalibrated if $\pi(\boldsymbol{X}) = \mathrm{E}[Y|\pi(\boldsymbol{X})]$. We refer the reader to Kruger and Ziegel (2020) for a general presentation of this concept. By Jensen inequality, autocalibration thus ensures that $\pi(\boldsymbol{X}) \preceq_{\mathrm{cx}} Y$. Thus, autocalibration implies that the predictor is less variable than the response, in the sense of the convex order.

Bregman dominance, also called forecast dominance is defined as dominance for every Bregman loss function. Precisely, π_2 outperforms π_1 in terms of Bregman dominance if the inequality $\mathrm{E}[L(Y, \pi_2)] \leq \mathrm{E}[L(Y, \pi_1)]$ holds true for every Bregman loss function L. We refer the interested reader to Kruger and Ziegel (2020) and the references therein for an extensive presentation of this concept.

Let π_1 and π_2 be two autocalibrated predictors. Bregman dominance reduces to the convex order for autocalibrated predictors, as pointed out by Kruger and Ziegel (2020). Precisely, π_1 outperforms π_2 in terms of Bregman dominance if, and only if,

$$\begin{aligned} & \pi_2(\boldsymbol{X}) \preceq_{\mathrm{cx}} \pi_1(\boldsymbol{X}) \\ \Leftrightarrow\ & \mathrm{LC}[\pi_1(\boldsymbol{X}); \alpha] \leq \mathrm{LC}[\pi_2(\boldsymbol{X}); \alpha] \text{ for all probability levels } \alpha \\ \Leftrightarrow\ & \mathrm{CC}[\mu(\boldsymbol{X}), \pi_1(\boldsymbol{X}); \alpha] \leq \mathrm{CC}[\mu(\boldsymbol{X}), \pi_2(\boldsymbol{X}); \alpha] \text{ for all probability levels } \alpha \end{aligned}$$

since Lorenz and concentration curves coincide for autocalibrated predictors, as shown by Denuit et al. (2021). It is interesting to notice that, for autocalibrated predictors, π_1 outperforms π_2 in terms of Bregman dominance if, and only if, π_1 is more discriminatory than π_2 in the sense defined in Denuit et al. (2019).

Denuit et al. (2021) defined the balance-corrected version π_{BC} of the predictor π as

$$\pi_{\mathrm{BC}}(\boldsymbol{X}) = \mathrm{E}[Y|\pi(\boldsymbol{X})].$$

It is shown there that if $s \mapsto \mathrm{E}[Y|\pi(\boldsymbol{X}) = s]$ is continuously increasing then the balance-corrected version π_{BC} of π satisfies the autocalibration property. The strategy proposed in this paper can be decomposed as follows:

- fit the models under consideration to the training data set $\mathcal{D}$ to get estimated distribution functions F_{π_k}.
- compute the ranks Π_k on $\overline{\mathcal{D}}$ and select the more expectation dependent $\Pi_{k^\star}$ with the response Y.
- use the selected model to produce the balance-corrected version

$$\pi_{\mathrm{BC}}(\boldsymbol{X}) = \mathrm{E}[Y|\Pi_{k^\star}].$$

The resulting π_{BC} is autocalibrated and dominates the competing models in Bregman dominance.

Models under consideration produce predictions $\pi_k(\boldsymbol{x}_i)$ on $\overline{\mathcal{D}}$. Therefore, we can apply existing tools to test for convex order based on the corresponding samples (see, e.g., Barrett and Donald, 2003). This can be an alternative to the testing procedure for more positive expectation dependence developed in the present paper.

References

- Barrett, G.F., Donald, S.G. (2003). Consistent tests for stochastic dominance. Econometrica 71, 71–104.

- Cmiel, B., Ledwina, T. (2017). Validation of positive expectation dependence. ESAIM: Probability and Statistics 21, 536-561.

- Denuit, M. (2010). Positive dependence of signals. Journal of Applied Probability 47, 893-897.

- Denuit, M., Charpentier, A., Trufin, J. (2021). Autocalibration and Tweedie-dominance for insurance pricing with machine learning. Available from https://dial.uclouvain.be.

- Denuit, M., Dhaene, J., Goovaerts, M.J., Kaas, R. (2005). Actuarial Theory for Dependent Risks: Measures, Orders and Models. Wiley, New York.

- Denuit, M., Eeckhoudt, L. (2016). Risk aversion, prudence and asset allocation: A review and some new developments. Theory and Decision 80, 227-243.

- Denuit, M., Hainaut, D., Trufin, J. (2020). Effective Statistical Learning Methods for Actuaries II: Tree-based Methods and Extensions. Springer Actuarial Lecture Notes Series.

- Denuit, M., Mesfioui, M. (2017). Preserving the Rothschild-Stiglitz type increase in risk with background risk: A characterization. Insurance: Mathematics and Economics 72, 1-5.

- Denuit, M., Sznajder, D., Trufin, J. (2019). Model selection based on Lorenz and concentration curves, Gini indices and convex order. Insurance: Mathematics and Economics 89, 128-139.

- Frees, E.W., Meyers, G., Cummings, A.D. (2011). Summarizing insurance scores using a Gini index. Journal of the American Statistical Association 106, 1085-1098.

- Frees, E.W., Meyers, G., Cummings, A.D. (2014). Insurance ratemaking and a Gini index. Journal of Risk and Insurance 81, 335-366.

- Ganuza, J.-J., Penalva, J.S. (2010). Signal orderings based on dispersion and the supply of private information in auctions. Econometrica 78, 1007-1030.

- Guo, X., Li, J. (2016). Confidence band for expectation dependence with applications. Insurance: Mathematics and Economics 68, 141-149.

- Hong, S.K., Lew, K.O., MacMinn, R., Brockett, P. (2011). Mossin's theorem given random initial wealth. Journal of Risk and Insurance 78, 309-324.

- Kowalczyk, T., Pleszczynska, E. (1977). Monotonic dependence functions of bivariate distributions. Annals of Statistics 5, 1221-1227.

- Kruger, F., Ziegel, J.F. (2020). Generic conditions for forecast dominance. Journal of Business & Economic Statistics, in press.

- Li, J. (2011). The demand for a risky asset in the presence of a background risk. Journal of Economic Theory 146, 372-391.

- Li, J., Liu, D., Wang, J. (2016). Risk aversion with two risks: A theoretical extension. Journal of Mathematical Economics 63, 100-105.

- Linton, O., Whang, Y., Yen, Y. (2018). The lower regression function and testing expectation dependence dominance hypotheses. https://doi.org/10.17863/CAM.36019

- Muliere, P., Petrone, S. (1992). Generalized Lorenz curve and monotone dependence orderings. Metron 50, 19-38.

- Shaked, M., Shanthikumar, J.G. (2007). Stochastic Orders. Springer, New York.

- Van der Vaart, A. W. (2000). Asymptotic statistics (Vol. 3). Cambridge university press. Chicago.

- Wright, R. (1987). Expectation dependence of random variables, with an application in portfolio theory. Theory and Decision 22, 111-124.

- Yitzhaki, S., Schechtman, E. (2013). The Gini Methodology: A Primer on Statistical Methodology. Springer.

- Zhu, X., Guo, X., Lin, L., Zhu, L. (2016). Testing for positive expectation dependence. Annals of the Institute of Statistical Mathematics 68, 135-153.

APPENDIX

A Proof of Proposition 3.1

The classical CLT and the Glivenko-Cantelli Lemma entail that

$$\begin{aligned}
\sqrt{n}\widehat{D}(\alpha) &:= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Y_i-\bar{Y})(\mathrm{I}[\Pi_{1i}\le\alpha]-\mathrm{I}[\Pi_{2i}\le\alpha]) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Y_i-\mathrm{E}[Y_i])(\mathrm{I}[\Pi_{1i}\le\alpha]-\mathrm{I}[\Pi_{2i}\le\alpha]) \\
&\quad +(\mathrm{E}[Y_i]-\bar{Y})\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\mathrm{I}[\Pi_{1i}\le\alpha]-\mathrm{I}[\Pi_{2i}\le\alpha]) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Y_i-\mathrm{E}[Y_i])(\mathrm{I}[\Pi_{1i}\le\alpha]-\mathrm{I}[\Pi_{2i}\le\alpha]) \\
&\quad +\sqrt{n}(\mathrm{E}[Y_i]-\bar{Y})\frac{1}{n}\sum_{i=1}^{n}(\mathrm{I}[\Pi_{1i}\le\alpha]-F_{\mathrm{unif}}(\alpha))-(\mathrm{I}[\Pi_{2i}\le\alpha]-F_{\mathrm{unif}}(\alpha)) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Y_i-\mathrm{E}[Y_i])(\mathrm{I}[\Pi_{1i}\le\alpha]-\mathrm{I}[\Pi_{2i}\le\alpha])+o_{\mathrm{P}}(1) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Y_i-\mathrm{E}[Y_i])((\mathrm{I}[\Pi_{1i}\le\alpha]-F_{\mathrm{unif}}(\alpha))-(\mathrm{I}[\Pi_{2i}\le\alpha]-F_{\mathrm{unif}}(\alpha)))+o_{\mathrm{P}}(1)
\end{aligned}$$

as $n\to\infty$, where F_{unif} stands for the unit uniform distribution function. Letting

$$M(Y,\Pi_1,\Pi_2,z):=(Y-\mathrm{E}[Y])((\mathrm{I}[\Pi_1\le z]-F_{\mathrm{unif}}(z))-(\mathrm{I}[\Pi_2\le z]-F_{\mathrm{unif}}(z))),$$

we have that since Y is square integrable, the class of functions $\{M(Y,\Pi_1,\Pi_2,z), z\in(0,1)\}$ is P- Donsker in the sense of Theorem 19.5 in van der Vaart (2000). The result follows.