Claims Frequency Modeling Using Telematics Car Driving Data

Guangyuan Gao* Shengwang Meng† Mario V. Wüthrich‡

Abstract

We investigate the predictive power of covariates extracted from telematics car driving data using the speed-acceleration heatmaps of Gao and Wüthrich (2017). These telematics covariates include K-means classification, principal components, and bottleneck activations from a bottleneck neural network. In the conducted case study it turns out that the first principal component and the bottleneck activations give a better out-of-sample prediction for claims frequencies than other traditional pricing factors such as driver's age. Based on these numerical examples we recommend the use of these telematics covariates for car insurance pricing.

Keywords: Telematics data; K-means algorithm; Principal components analysis; Bottleneck neural network; Autoencoder; Generalized additive model; v-a heatmap; Pattern recognition; Kullback-Leibler divergence; Claims frequency modeling; Car insurance pricing.

1 Introduction

With insurers collecting more and more telematics car driving data, there is the basic demand of covariate extraction from this data and testing the predictive power of these covariates in car insurance pricing models. Telematics car driving data is a type of high frequency data, typically having a record per second. Wüthrich (2017) proposed to use speed-acceleration (v-a) heatmaps to visualize this high frequency data. He applied the K-means algorithm to classify these v-a heatmaps. The classes induced may be used as categorical covariates in car insurance pricing models. Gao and Wüthrich (2017) proposed two different dimension reduction techniques: principal components analysis (PCA) and bottleneck neural networks. Both techniques may serve to represent the high dimensional v-a heatmaps by two continuous covariates. Note that the bottleneck neural network can be seen as a non-linear PCA (Kramer, 1991), and it is a special case of an autoencoder. Gao and Wüthrich (2017) have shown that the bottleneck neural network compression can better approximate the true v-a heatmaps than the PCA in terms of the Kullback-Leibler divergence; for these unsupervised learning methods and others we refer to Hastie et al. (2009) and Bishop (2006). Recent studies on the speed-acceleration pattern also include Weidner et al. (2016a,b), where the authors use a Fourier analysis decomposition. Another branch of research on telematics car driving data focuses on selecting a good exposure measure, including Ayuso et al. (2016) and Verbelen et al. (2018). These papers extract exposure

*Center for Applied Statistics and School of Statistics, Renmin University of China, 100872 Beijing, China. Corresponding email:guangyuan.gao@ruc.edu.cn. The authors are listed in alphabetical order.

†Center for Applied Statistics and School of Statistics, Renmin University of China, 100872 Beijing, China.

‡ETH Zurich, RiskLab, Department of Mathematics, 8092 Zurich, Switzerland

measures from telematics car driving data used for usage-based insurance (UBI) and pay-as-you-drive (PAYD) insurance products.
For claims count modeling we use a Poisson regression model with the expected claims frequency being described by a generalized additive model (GAM). GAMs have been introduced by Hastie and Tibshirani (1990), and they were further developed by Wood (2006). GAMs implement natural cubic splines to address non-linear effects of covariates which cannot be captured by generalized linear models (GLMs). The smoothness of these natural cubic splines and the non-linear effects, respectively, is controlled by a smoothing parameter whose value is selected by cross-validation. The smoothing parameter induces a penalty term in the likelihood function (ridge regression), and the parameters are estimated by the penalized iteratively re-weighted least squares method. Note that we frequently use cross-validation to compare models and select variable. Cross-validation evaluates the prediction error on a test sample, different from the training sample used for the parameter estimation. Hastie et al. (2009) and Wüthrich and Buser (2017) give a detailed discussion on cross-validation.

The paper is structured as follows. Section 2 describes the data being used. Section 3 investigates the predictive power of covariates extracted from telematics car driving data in a Poisson GAM for claims frequency modeling. Section 4 concludes with several findings. In the appendix we study the minimal sample size needed to receive stable v-a heatmaps; these results are important because of the vast amount of telematics car driving data available.

2 Data description

Our data contains: (a) policy information about the car and the driver, (b) insurance claims data, and (c) telematics car driving data. We describe (a)-(c) in more detail below. The data is available from 01/01/2014 to 29/06/2017. The telematics car driving data is roughly 1 GB per day, which amounts totally in 1.2 TB of data over the whole observation period.

(a) The policy information includes information about the car and the driver, such as driver's age, driver's gender, price of car, car brand, number of seats and car's age. Note that driver's age and gender refer to the most frequent driver on a particular car, and that the main driver of a car may change over time (this may also happen due to insurance premium optimization from a policyholder's perspective). In our analysis, we use the driver information that has been valid the longest on a given policy. We remark that we exclude policies where the car changes.

(b) The claims data contains the number of reported claims from 01/01/2014 to 29/06/2017 of insurance policies with effective exposures supported in the time interval from 01/01/2014 to 31/05/2017. The evaluation date is chosen as 29/06/2017 since a preliminary analysis has shown that more than 99% of all claims are reported with less than one month of reporting delay. For this reason, we do not expect a material influence on the claims frequency of claims with a reporting delay of more than one month. The policies with expiry dates after 31/05/2017 are partially exposed, and the exposures for these policies are adjusted pro-rata temporis.

(c) For computational reasons, we only consider the telematics car driving data from 01/05/2016

to 31/07/2016. This data is sufficient for constructing stable v-a heatmaps, see Appendix A. The telematics car driving data contains both GPS speed and vehicle sensor speed. We use vehicle sensor speed since it is more reliable than GPS speed which is sensitive to the GPS signal strength. For instance, there is not any signal while driving through a tunnel. The speed v range chosen for our analysis is $[5, 20]$km/h, and the acceleration rate a is capped within $[-2, 2]$m/s^2 (Weidner et al., 2016a).

After data cleaning for wrong, missing and unstable data we have been able to consider $n = 1,478$ drivers with a total exposure of 3,332 years (years-at-risk) and an observed claims frequency of 23.56%. We use the claims information on these $n = 1,478$ drivers to test the predictive power of the extracted covariates from the telematics car driving data. Summary statistics of the driver's age and the car's age of this data is provided in Appendix B.

3 Poisson generalized additive models for claims frequencies

3.1 Covariates and v-a heatmaps for claims frequency modeling

We first describe how we extract covariate information from telematics car driving data. Therefore, we follow Wüthrich (2017) to construct speed-acceleration (v-a) heatmaps. The vehicle sensor speed only takes values in integers. Therefore, we partition the v-a rectangle $R = [5, 20] \times [-2, 2]$ by dividing the v-axis (speed) into 16 intervals and the a-axis (acceleration) into 20 intervals. The resulting sub-rectangles are denoted by $(R_j)_{j=1:J}$ with $J = 320$, these are illustrated in Figure 1 (top-left).

For each car driver $i = 1, \ldots, n$, we denote the relative amount of time spent in sub-rectangle $R_j \subset R$ by $x_{i,j} \geq 0$. The induced empirical discrete distribution of driver i on the v-a rectangle R is denoted by $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,J})'$, which lies in the $(J-1)$-unit simplex $\mathcal{X} \subset \mathbb{R}^J$, i.e. has normalization $\sum_{j=1}^J x_{i,j} = 1$. Thus, every car driver $i = 1, \ldots, n$ is characterized by a discrete distribution $\boldsymbol{x}_i \in \mathcal{X}$; and $\mathcal{X}$ represents all possible car driver's discrete distributions on $\bigcup_{j=1}^J R_j = R = [5, 20] \times [-2, 2]$. We denote by $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)' \in \mathbb{R}^{n \times J}$ the $n \times J$ design matrix that contains $\boldsymbol{x}_i$ of all $n = 1,478$ car drivers. In Figure 1, we plot the resulting v-a heatmaps $\boldsymbol{x}_i \in \mathcal{X}$ of the selected car drivers $i = 72, 608$ and 718. Driver 72 tends to accelerate and brake less frequently than the other two drivers, while driver 718 tends to accelerate and brake most frequently among the three drivers.

Directly using the design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times J}$ as covariates in the claims frequency regression model would lead to over-parametrization (and over-fitting). In the following, we consider three dimension reduction techniques: the K-means algorithm, the PCA and the bottleneck neural network method. These were discussed in Gao and Wüthrich (2017). We study these three methods in the following subsections.

Before doing so, we describe the other (classical) covariates. We have performed a preliminary analysis using the available covariates of driver's age, driver's gender, price of car, car brand, number of seats and car's age. This analysis has shown that driver's age and car's age have the best out-of-sample performance for claims frequency prediction on our data. The other covariates did not provide much predictive power, probably because the number of available policies is comparably small. In addition, we have observed that the effect of car's age on the claims frequency is roughly log-linear, and that the effect of driver's age is more sophisticated.

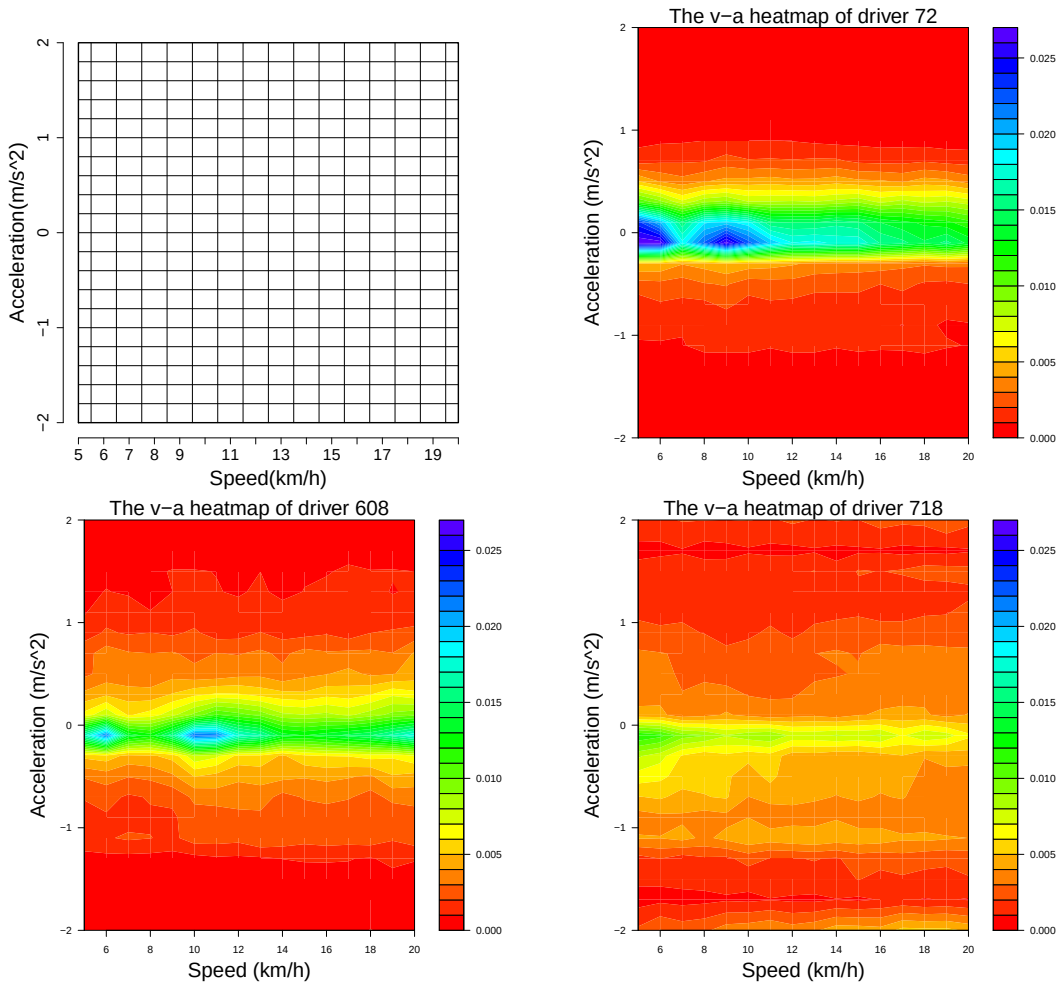

FIGURE 1: Top-left: the considered partition of sub-rectangles $(R_j)_{j=1:J}$ of $R = [5, 20] \times [-2, 2]$; the other plots illustrate the v-a heatmaps $\boldsymbol{x}_i \in \mathcal{X}$ of the selected car drivers $i = 72, 608$ and 718.

The latter is addressed by considering a natural cubic spline in a GAM approach. We plot the years-at-risk (volumes) per driver's age and per car's age in Figure 9 of Appendix B. The two covariates are not collinear as shown in Figure 10 of Appendix B. Note that a complete list of models studied is summarized in Table 4, below.

3.2 The K-means algorithm

The K-means algorithm (Hastie et al., 2009) is applied to classify the $n = 1,478$ car drivers into K groups. We use exactly the same set up as in Wüthrich (2017), that is, we use the L^2-norm as dissimilarity measure, and we minimize the within-cluster dissimilarity. Denote the resulting classifier by $\mathcal{C}_K : \mathcal{X} \to \{1, \ldots, K\}$. The group of car driver i is then given by $\mathcal{C}_K(\boldsymbol{x}_i) \in \{1, \ldots, K\}$. The within-cluster dissimilarity is decreasing with the number of labels (clusters) K being increasing, see Figure 2 (left). The groups are used as a categorical covariate in a Poisson GAM for claims frequency modeling as follows: for $i = 1, \ldots, n$ the claims counts

Y_i of car drivers i are independent and fulfill

$$\begin{aligned} Y_i &\overset{\text{ind.}}{\sim} \text{Poisson}(\lambda_i e_i) \qquad \text{with} \\ \lambda_i &= \exp\left\{\beta_0 + s_1(\text{age driver}_i) + \beta_2 \cdot \text{age car}_i + \gamma_{\mathcal{C}_K(\boldsymbol{x}_i)}\right\}, \end{aligned} \tag{3.1}$$

where $e_i > 0$ is the total exposure of driver i, λ_i is the expected claims frequency of driver i, described by an intercept β_0, s_1 is a smoothing spline on the driver's age covariate, β_2 is the regression coefficient for the car's age covariate, and $(\gamma_k)_{k=1:K}$ is the parameter for the k-th group of the K-means classification. As smoothing spline we choose a natural cubic spline with knots at the unique values of its argument, i.e., s_1 has 50 knots (unique drivers' ages). The smoothness is controlled by a non-negative smoothing parameter, whose optimal value is determined as 0.0930. Wood (2006) gives a detailed discussion on the GAM. We denote the set of all parameters of model (3.1) by $\theta = \{\beta_0, \beta_1, \beta_2, \gamma\}$, where β_1 is the parameter vector of the smoothing spline s_1. The optimal value of clusters K is determined by cross-validation as follows. Denote the total data set by Ω and a subset containing $I < n$ drivers by $\mathcal{T} \subset \Omega$. The prediction error of model (3.1) with estimated parameter $\hat{\theta}$ on the subset $\mathcal{T}$ can be measured by the (average) Poisson deviance statistics (Wüthrich and Buser, 2017) as

$$D(\mathcal{T}, \hat{\theta}) = \frac{2}{I} \sum_{i \in \mathcal{T}} Y_i \left[\frac{\hat{\lambda}_i e_i}{Y_i} - 1 - \log\left(\frac{\hat{\lambda}_i e_i}{Y_i} \right) \right],$$

where the i-th term on the right-hand side is set equal to $2\hat{\lambda}_i e_i$ if $Y_i = 0$; by $\hat{\lambda}_i = \lambda_i(\hat{\theta})$ we denote the estimated claims frequency of driver i. If the data $\mathcal{T}$ is used to estimate $\hat{\theta}$, the above quantity is called in-sample deviance statistics. If $\mathcal{T}$ is not used to estimate $\hat{\theta}$, the above quantity is called out-of-sample deviance statistics. The out-of-sample deviance statistics is preferred since it is sensitive to over-fitting. Moreover, we prefer the Poisson deviance statistics as loss function because it is the natural choice for claims frequency modeling in car insurance, and because it is more robust towards outliers than the weighted square loss function in a low frequency situation, see Section 2.6.5 in Wüthrich and Buser (2017). Typically, 10-fold cross-validation is applied for estimating the out-of-sample deviance statistics. Therefore, we partition Ω into 10 roughly equally-sized parts, denoted by $\Omega_1, \ldots, \Omega_{10}$. The 10-fold cross-validation estimate of the out-of-sample deviance statistics can be obtained as

$$\widehat{\text{CV}}^{10} = \frac{1}{10} \sum_{l=1}^{10} D(\Omega_l, \hat{\theta}_{-\Omega_l}), \tag{3.2}$$

where $\hat{\theta}_{-\Omega_l}$ is the estimated parameter using all data except Ω_l. Note that we fix the smoothing parameter at 0.0930 to stablize the smoothing term of driver's age during the cross-validation, i.e., the strength of the penalization is kept fixed during the cross-validation to get more robust results. We randomly partition the data for 50 times, and calculate the cross-validation estimate of the out-of-sample deviance statistics for each partition. Note that we use the same partitioning for the cross-validations in all examples that follow below.

The resulting empirical estimates of the 10-fold cross-validations are shown in the box plots on the right-hand side of Figure 2. For comparison purposes, we provide the corresponding 5-fold cross-validation results on the left-hand side of Figure 11 (in the appendix). We note that we do

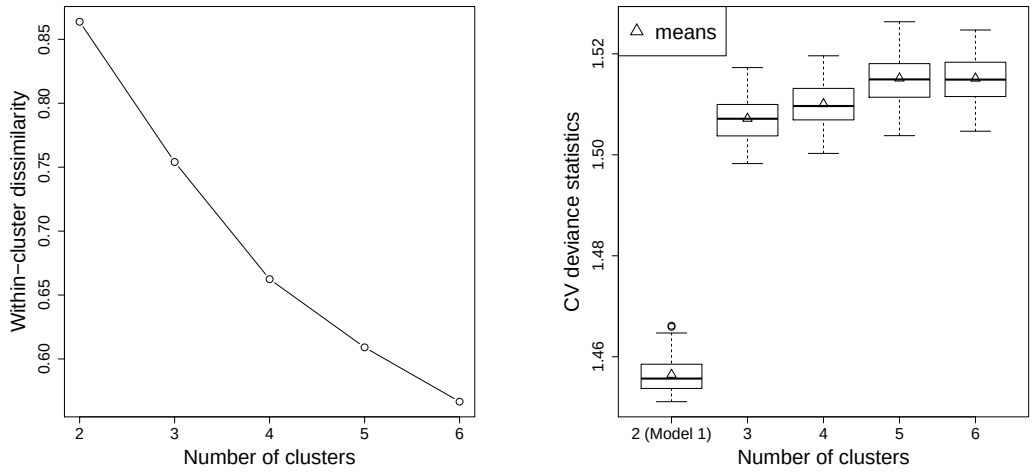

FIGURE 2: Left: within-cluster dissimilarity for $K = 2, \ldots, 6$ categorical classes; right: box plots of 10-fold cross-validation estimates of deviance statistics for the different numbers of clusters $K = 2, \ldots, 6$; below, we choose $K = 2$ for Model 1, see also Table 4 below.

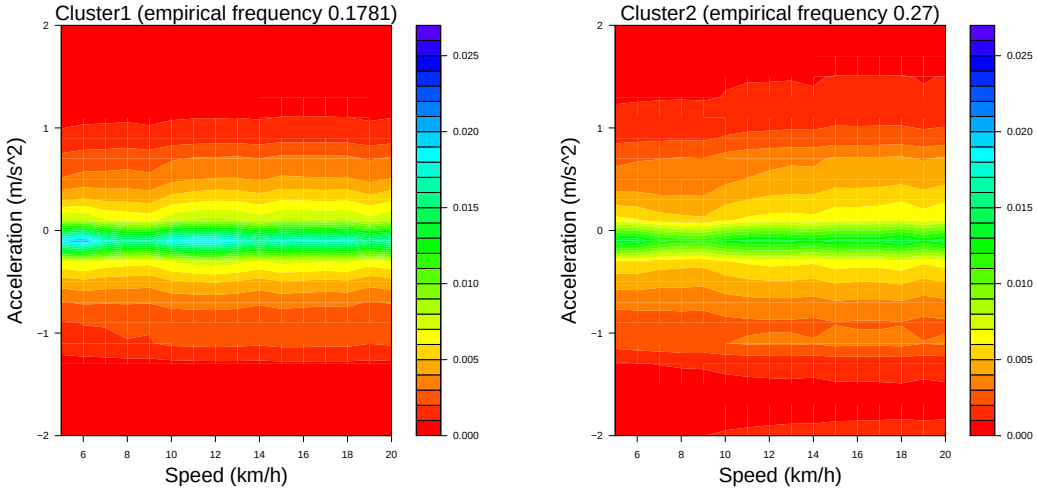

FIGURE 3: Average v-a heatmaps for the 2-means clusters ($K = 2$).

not observe any material differences between 10-fold and 5-fold cross-validation. From the graph on the right-hand side of Figure 2 we conclude that the optimal number of clusters is $K = 2$ (w.r.t. the means of the cross-validation deviances). For comparison purposes we also consider the case $K = 3$, below. We call the model with two clusters $K = 2$ as *Model 1* (see also Table 4, below). Another observation is that the cross-validation estimates are slightly positively skewed. This comes from the fact that the distribution of deviance statistics is similar to a chi-squared distribution (which has positive skewness). The empirical claims frequencies for the 2 clusters are 17.81% and 27.00%, respectively, with total years-at-risk given by 1,247 and 2,085, respectively. The average v-a heatmap for each cluster is shown in Figure 3. The v-a heatmap of cluster 1 is more concentrated at the zero acceleration axis. Drivers in cluster 1 tend to accelerate and brake less frequently (or strongly) than those in cluster 2. The estimated parameters of Model 1 are shown in Table 1. Note that cluster 1 is treated as a reference group, whose parameter is

fixed as $\gamma_1 = 0$. According to individual z-tests, the heatmap clustering is the most important covariate to explain the variation in the claims frequencies. The positive value of $\hat{\gamma}_2 = 0.4015$ is consistent with the empirical claims frequencies of the two clusters.

Table 1: Estimated parameters of Model 1 (see also Table 4).

Parameters	Estimated	Std. Error	z value	$\Pr(>\|z\|)$
β_0	-1.8609	0.0814	-22.8507	0.0000
β_2	0.0504	0.0171	2.9426	0.0033
γ_2	0.4015	0.0803	4.9966	0.0000
Smooth term	edf	Ref. df	Chi. Sq	$\Pr(>\|z\|)$
$s_1(age)$	6.0730	7.1606	26.2681	0.0005

Concerning the three drivers in Figure 1: drivers 72 and 608 are in cluster 1 and driver 718 in cluster 2. Thus, the 2-means algorithm cannot distinguish drivers 72 and 608, though their v-a heatmaps are quite different, see Figure 1. A more appropriate dimension reduction technique should generate continuous variables to represent the corresponding v-a heatmaps. This is exactly what the PCA and the bottleneck neural network achieve to do (Gao and Wüthrich, 2017).

3.3 Principal component analysis

Denote the normalized design matrix of $\boldsymbol{X}$ by $\boldsymbol{X}^0$ (all column means are set to zero and variances are normalized to one). There exists an $n \times J$ orthogonal matrix $\boldsymbol{U}$, a $J \times J$ orthogonal matrix $\boldsymbol{V}$ and a $J \times J$ diagonal matrix $\boldsymbol{\Lambda} = \operatorname{diag}(g_1, \ldots, g_J)$ with singular values $g_1 \geq \ldots \geq g_J \geq 0$, such that we have the following singular value decomposition (SVD)

$$\boldsymbol{X}^0 = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}'.$$

PCA is a linear method that explores the J-dimensional covariate space for the direction of the biggest variance in $\boldsymbol{X}^0$. The first column of $\boldsymbol{V}$ is the direction of the biggest variance in the J-dimensional covariate space. The second column of $\boldsymbol{V}$ is the direction of the second largest variance, perpendicular to the first direction. The columns of $\boldsymbol{P} = \boldsymbol{U}\boldsymbol{\Lambda}$ are the principal components, which are then used as continuous covariates in the claims frequency model as follows:

$$\begin{aligned} &Y_i \overset{\text{ind.}}{\sim} \text{Poisson}(\lambda_i e_i) \qquad \text{with} \\ &\lambda_i = \exp\left\{\beta_0 + s_1(\text{age driver}_i) + \beta_2 \cdot \text{age car}_i + s_2(P_{i,m})\right\}, \end{aligned}$$

where s_2 is a smoothing spline on $P_{i,m}$ which is the value (covariate) of driver i in the m-th principal component. The situation of considering several principal components simultaneously in the regression model is discussed later. We use cross-validation to evaluate which principal components having the smallest out-of-sample deviance statistics; note that the same partitioning and the same smoothing parameter for s_1 as in Section 3.2 are used. We also fix the smoothing parameter of the second smoothing spline s_2 at its optimal value based on all samples during the cross-validation analysis. In Figure 4, the cumulatively explained variance proportions of the principal components and the cross-validation estimate of the deviance statistics are plotted. The first principal component has the best out-of-sample prediction (the same conclusion is drawn

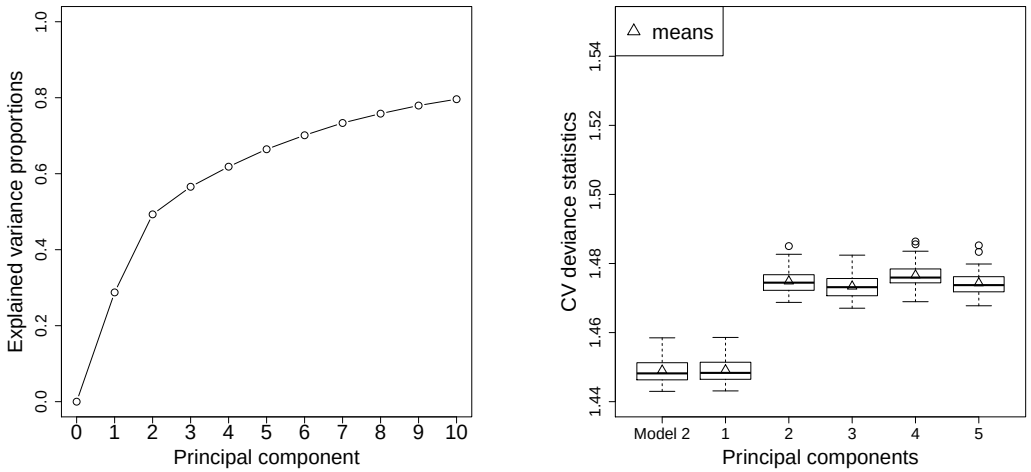

Figure 4: Left: cumulatively explained variance proportions of the first 10 principal components; right: 10-fold cross-validation estimates of the deviance statistics for different principal components (we refer to Table 4 for Model 2).

using the 5-fold cross-validation estimates as shown in the middle of Figure 11 in the appendix). We presume that it quantifies the width of the contours in the acceleration direction; see Figures 1 and 3. It turns out that the estimated effective degrees of freedom of $\hat{s}_2(P_1)$ in the GAM is close to 1. This suggests to change s_2 to the identity function for the first principal component. We call this model as *Model 2* (which is also summarized in Table 4). The estimated coefficients of Model 2 are listed in Table 2, where β_3 is the coefficient of the (log-linearly considered) first principal component. According to individual z-tests, the first principal component is the most important covariate to explain the variation in the claims frequencies. The negative sign of $\hat{\beta}_3$ indicates that the first principal component can be interpreted as a safe driving index. The cross-validation estimate of the deviance statistics for Model 2 is added to Figure 4. Comparing the second plots in Figures 2 and 4, we find that Model 2 outperforms Model 1, since Model 2 has a lower average cross-validation value (we also refer to Table 5). Further analysis shows that once the first principal component is included in the model the other principal components are no longer needed.

Table 2: Estimated parameters of Model 2 (see also Table 4).

Parameters	Estimated	Std. Error	z value	$\Pr(>\|z\|)$
β_0	-1.6066	0.0597	-26.9281	0.0000
β_2	0.0458	0.0172	2.6633	0.0077
β_3	-0.0228	0.0037	-6.1573	0.0000
Smooth term	edf	Ref. df	Chi. Sq	$\Pr(>\|z\|)$
$s_1(age)$	6.0695	7.1575	25.8762	0.0006

The first two principal components for the three drivers in Figure 1 are drawn in the top-left of Figure 6. Though driver 72 (square) and driver 608 (triangle) are in cluster 1 (red), they have quite different first principal components. From this we conclude that the first principal component can distinguish different v-a heatmaps more subtly than the 2-means clusters. Driver 718 (circle) is in cluster 2 (blue).

3.4 Bottleneck neural network approach

The last data compression method that we consider is a bottleneck neural network. A bottleneck neural network is an autoencoder which consists of an encoder $\varphi : \mathcal{X} \to \mathcal{Z}$, where $\mathcal{Z}$ is low-dimensional, and of a decoder $\psi : \mathcal{Z} \to \mathcal{X}$. The goal of this autoencoder is to choose these functions φ and ψ such that the output $\pi(\boldsymbol{x}) = \psi \circ \varphi(\boldsymbol{x})$ is close to the input $\boldsymbol{x}$. The value $\varphi(\boldsymbol{x}) \in \mathcal{Z}$ is then used as a low-dimensional representation for $\boldsymbol{x} \in \mathcal{X}$. To achieve this we use a neural network with 3 hidden layers having $(p, q, p) = (7, 2, 7)$ hidden neurons (Gao and Wüthrich, 2017). The encoder will map $\boldsymbol{x} \in \mathcal{X}$ to $\boldsymbol{z}^{(2)} = \boldsymbol{z}^{(2)}(\boldsymbol{x}) \in \mathcal{Z} = [-1, 1]^2$, the bottleneck of the neural network having $q = 2$ neurons, see (3.4), below. Moreover, we choose the neural network symmetric w.r.t. the bottleneck because this has major advantages in calibration of the corresponding encoding functions, see Kramer (1991). The first hidden layer is given by

$$z_l^{(1)}(\boldsymbol{x}) = \tanh\left(w_{l,0}^{(1)} + \sum_{j=1}^{J} w_{l,j}^{(1)} x_j\right), \qquad \text{for } l = 1, \ldots, p, \tag{3.3}$$

and the second hidden layer by

$$z_l^{(2)}(\boldsymbol{x}) = \tanh\left(w_{l,0}^{(2)} + \sum_{j=1}^{p} w_{l,j}^{(2)} z_j^{(1)}(\boldsymbol{x})\right), \qquad \text{for } l = 1, \ldots, q. \tag{3.4}$$

This provides the encoder $\varphi(\boldsymbol{x}) = \boldsymbol{z}^{(2)}(\boldsymbol{x}) = (z_1^{(2)}(\boldsymbol{x}), z_2^{(2)}(\boldsymbol{x}))' \in [-1, 1]^2$ for bottleneck $q = 2$. The third hidden layer of the neural network is given by

$$z_l^{(3)}(\boldsymbol{x}) = \tanh\left(w_{l,0}^{(3)} + \sum_{j=1}^{q} w_{l,j}^{(3)} z_j^{(2)}(\boldsymbol{x})\right), \qquad \text{for } l = 1, \ldots, p, \tag{3.5}$$

and this is then used in the regression equations

$$\mu_j(\boldsymbol{x}) = \mu_j(\boldsymbol{x}; \boldsymbol{\alpha}^{(j)}) = \alpha_0^{(j)} + \sum_{l=1}^{p} \alpha_l^{(j)} z_l^{(3)}(\boldsymbol{x}), \qquad \text{for } j = 1, \ldots, J. \tag{3.6}$$

Functions (3.5)-(3.6) provide the decoder defined by the following multinomial logistic probabilities $\pi(\cdot) = (\pi_j(\cdot))_{j=1:J}$ with

$$\pi_j(\boldsymbol{x}) = \frac{\exp\{\mu_j(\boldsymbol{x})\}}{\sum_{j'=1}^{J} \exp\{\mu_{j'}(\boldsymbol{x})\}}, \qquad \text{for all } \boldsymbol{x} \in \mathcal{X}.$$

This bottleneck neural network has network parameter $\boldsymbol{\phi} = (\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}, \boldsymbol{w}^{(3)}, \boldsymbol{\alpha})$ for given hyperparameters p and q. The goal is to calibrate this bottleneck neural network such that the dissimilarity between the input $\boldsymbol{x}_i \in \mathcal{X}$ and the output $\pi(\boldsymbol{x}_i) \in \mathcal{X}$ of all observed v-a heatmaps $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subset \mathcal{X}$ is small. Note that both, the inputs and the outputs, live on the same $(J-1)$-unit simplex which contains all discrete J-dimensional probability measures. A natural choice in statistics to measure the dissimilarity between probability measures is the Kullback-Leibler (KL) divergence. In our case, for given network parameter $\boldsymbol{\phi}$, the average KL divergence is given

by

$$\mathcal{L}_{\mathrm{KL}}\left(\boldsymbol{\phi}, (\boldsymbol{x}_i)_{i=1:n}\right) = \frac{1}{n}\sum_{i=1}^{n} d_{\mathrm{KL}}\left(\boldsymbol{x}_i \| \pi(\boldsymbol{x}_i)\right), \tag{3.7}$$

with KL divergence between $\boldsymbol{x}_i$ and $\pi(\boldsymbol{x}_i)$ defined as

$$d_{\mathrm{KL}}\left(\boldsymbol{x}_i \| \pi(\boldsymbol{x}_i)\right) = -\sum_{j=1}^{J} x_{i,j} \log \frac{\pi_j(\boldsymbol{x}_i)}{x_{i,j}}, \tag{3.8}$$

where the j-th term of KL divergence is set equal to zero if $x_{i,j} = 0$.
A successful calibration for given hyperparameters p and q of this neural network requires pre-calibration of $\boldsymbol{\phi}$. The network architecture is chosen such that the calibration can be done in 3 successive steps, for details we refer to Kramer (1991) and Gao and Wüthrich (2017); we emphasize that the symmetry (p, q, p) is important for performing this pre-calibration, and for a 2-dimensional autoencoder ($q = 2$) the only free hyperparameter is p. In Figure 5, we show the last step of the calibration giving the convergence of the (in-sample) average KL divergence (3.7) with the iterations in the gradient descent method, see also Gao and Wüthrich (2017). From this reference, we also conclude that the architecture $(p, q, p) = (7, 2, 7)$ provides a good autoencoder for our v-a heatmaps; note that this is also justified by Figure 7, below.

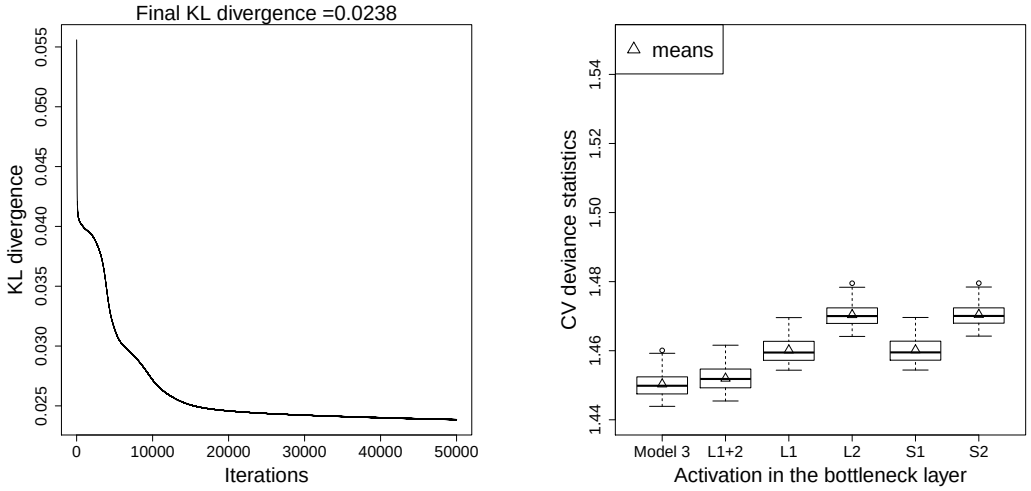

FIGURE 5: Left: convergence of the average KL divergence of the neural network calibration using the gradient descent method; right: 10-fold cross-validation estimate of the out-of-sample deviance statistics with bottleneck activation covariates, Model 3 has the transformed bottleneck activation z_0 in log-linear form, L1 has the first activation $z_1^{(2)}$ in log-linear form, S1 has the first activation $z_1^{(2)}$ in a smoothing spline form, and so for L2, S2 and L1+2 (please also refer to Table 4 for Model 3).

We consider the following model:

$$\begin{aligned} &Y_i \overset{\text{ind.}}{\sim} \text{Poisson}(\lambda_i e_i) \qquad \text{with} \\ &\lambda_i = \exp\left[\beta_0 + s_1(\text{age driver}_i) + \beta_2 \cdot \text{age car}_i + s_2\left(z_l^{(2)}(\boldsymbol{x}_i)\right)\right], \end{aligned}$$

where $z_l^{(2)}(\boldsymbol{x}_i)$ is the l-th activation at the bottleneck for driver i, for $l = 1, \ldots, q$ and $q = 2$ bot-

tleneck neurons. The model containing both bottleneck activations simultaneously is discussed later. From the 10-fold cross-validation estimates of the out-of-sample deviance statistics shown in Figure 5, we can see that the first activation $z_1^{(2)}$ (labeled by S1) has a better out-of-sample prediction than the second activation $z_2^{(2)}$ (labeled by S2). Note that the same partitioning and the same smoothing parameter for s_1 as in Section 3.2 are used. We also fix the smoothing parameter of the second smoothing splines$_2$ at its optimal value based on all samples during the cross-validation. It turns out that the estimated effective degrees of freedom of $\hat{s}_2(z_1^{(2)})$ and $\hat{s}_2(z_2^{(2)})$ are both close to 1. Therefore, we replace the smoothing spline with the identity function, and we re-evaluate the cross-validation estimates of the deviance statistics, labeled by L1 and L2 in Figure 5. Again we see that the first activation has a better out-of-sample prediction than the second one. Further investigation indicates that both activations are simultaneously needed in the model and their effects are in log-linear form with the estimated coefficients as 3.52 and -1.78. The corresponding cross-validation estimate of the deviance statistics is labeled by L1+2 in Figure 5. To facilitate the practical use and interpretation, we create a new covariate as follows

$$z_0(\boldsymbol{x}_i) = z_1^{(2)}(\boldsymbol{x}_i) - 0.5 z_2^{(2)}(\boldsymbol{x}_i), \tag{3.9}$$

and we assume the claims frequency is given by

$$\lambda_i = \exp\left[\beta_0 + s_1(\text{age driver}_i) + \beta_2 \text{age car}_i + \beta_3 z_0(\boldsymbol{x}_i)\right]. \tag{3.10}$$

Choice (3.9) is motivated by $3.52/-1.78 \approx -2$. We call this model as *Model 3* (which is also summarized in Table 4). The associated cross-validation estimate of the deviance statistics is added to Figure 5. We see that Model 3 and L1+2 are comparable (out-of-sample), the latter model having one additional parameter. Note that in the 10-fold cross validations, we use the same partitioning and the same smoothing parameter for s_1 as in Section 3.2. We have a similar observation for the 5-fold cross-validation as shown on the right-hand side of Figure 11, in the appendix. The estimated coefficients of Model 3 are listed in Table 3. From individual z-tests we see that the transformed bottleneck activation is the most import covariate to explain the variation in the claims frequencies.

Table 3: Estimated parameters of Model 3 (see also Table 4).

Parameters	Estimated	Std. Error	z value	Pr($>\|z\|$)
β_0	-1.6322	0.0603	-27.0645	0.0000
β_2	0.0496	0.0173	2.8755	0.0040
β_3	3.5353	0.5840	6.0535	0.0000
Smooth term	edf	Ref. df	Chi. Sq	Pr($>\|z\|$)
$s_1(age)$	5.9135	7.0109	24.7391	0.0009

The bottleneck activations for all drivers are drawn in the bottom-left of Figure 6. The two activations have a positive relationship. The red and blue colors in that plot reflect the corresponding clusters from the 2-means algorithm. The non-vertical boundary between the clusters indicates that the second activation is also sensitive to the 2-means clusters. This explains why the second activation is needed when already considering the first activation in the model. This is in contrast to the PCA results which are shown in the top-left of Figure 6.

3.5 Model comparisons

To analyze the predictive power of each covariate, we introduce 7 additional models that consider different covariate combinations. The full list of models is given in Table 4. Models 1-3 (having telematics covariates) are compared to Model 0 (based solely on traditional covariates). Models A.0 and A.3 replace the smoothing spline for driver's age by categorical age classes; the chosen age classes (including exposures) are listed in Table 6 of Appendix B. Models B.0 and B.3 use the two traditional pricing factors of seat count and price of car. Finally, Models C.2 and C.3 are based on telematics covariates only.

Table 4: The 10 models considered and their regression functions.

Models	Regression functions
Model 0	$\log \lambda = \beta_0 + s_1(\text{age driver}) + \beta_2 \cdot \text{age car}$
Model 1	$\log \lambda = \beta_0 + s_1(\text{age driver}) + \beta_2 \cdot \text{age car} + \gamma_{\mathcal{C}_2}$
Model 2	$\log \lambda = \beta_0 + s_1(\text{age driver}) + \beta_2 \cdot \text{age car} + \beta_3 \cdot P_1$
Model 3	$\log \lambda = \beta_0 + s_1(\text{age driver}) + \beta_2 \cdot \text{age car} + \beta_3 \cdot z_0$
Model A.0	$\log \lambda = \beta_0 + \alpha_{\text{ age class}} + \beta_2 \cdot \text{age car}$
Model A.3	$\log \lambda = \beta_0 + \alpha_{\text{ age class}} + \beta_2 \cdot \text{age car} + \beta_3 \cdot z_0$
Model B.0	$\log \lambda = \beta_0 + \beta_1 \cdot \text{seat count} + \beta_2 \cdot \text{price car}$
Model B.3	$\log \lambda = \beta_0 + \beta_1 \cdot \text{seat count} + \beta_2 \cdot \text{price car} + \beta_3 \cdot z_0$
Model C.2	$\log \lambda = \beta_0 + \beta_1 \cdot P_1$
Model C.3	$\log \lambda = \beta_0 + \beta_1 \cdot z_0$

Table 5: Cross-validation estimate of the out-of-sample deviance statistics for the models in Table 4.

Models	Mean	Std.	1st quantile	3rd quantile
Model 0	1.4734	0.0035	1.4707	1.4755
Model 1	**1.4565**	0.0038	1.4537	1.4585
Model 2	**1.4490**	0.0038	1.4463	1.4513
Model 3	**1.4503**	0.0037	1.4475	1.4524
Model A.0	1.4823	0.0028	1.4804	1.4839
Model A.3	**1.4579**	0.0030	1.4557	1.4598
Model B.0	1.4872	0.0013	1.4862	1.4882
Model B.3	**1.4607**	0.0016	1.4595	1.4623
Model C.2	**1.4556**	0.0017	1.4545	1.4566
Model C.3	**1.4572**	0.0016	1.4563	1.4579

The 10-fold cross-validation estimates of the out-of-sample deviance statistics are listed in Table 5 (using the same partitioning as in Sections 3.2-3.4). Comparing Models 0-3 we see that including the telematics covariates leads to a clearly lower out-of-sample prediction error in terms of quantile range than Model 0 which is solely based on traditional covariate information. Comparing Models 1-3 we see that including the continuous telematics covariates P_1 and z_0 leads to a lower out-of-sample prediction error than Model 1 which uses the categorical telematics covariates $\mathcal{C}_2$. Comparing Models 0, 3, A.0, A.3, we see that the smoothing term of driver's age leads to a slight lower out-of-sample prediction error than categorical driver's age classes. Comparing Models 0, 3, B.0, B.3, we see that using the driver's age and car's age lead to a lower out-of-sample prediction error than using seat counts and price of car. Comparing Models 2,

3, C.2, C.3, we see that the claims frequency models using telematics covariates can be slightly improved by adding the driver's age and car's age covariates.
To compare the effects of different covariates on the claims frequency, we choose Model 3 for illustration. The years-at-risk per transformed bottleneck activation z_0 are plotted in Figure 9 of the appendix. The scatter plot matrix for the three covariates is shown in Figure 10, which does not indicate an issue of collinearity. Finally, we make marginal predictions for each covariate value and compare them to the observed claims frequency in Figure 12 of the appendix. For instance, to make a marginal prediction for driver's age 50, we first find the average car's age and the average z_0 of all 50 years old drivers in the portfolio, and then make the prediction. We see that the effects of driver's age and z_0 are of a similar magnitude, while the effect of car's age is rather mild.
From this we conclude that the v-a heatmaps are very relevant information for car frequency prediction. In our analysis, the corresponding compressed v-a heatmaps (PCA or bottleneck activations) have a better out-of-sample prediction than the classical covariate driver's age. The only reservation to be made is that this analysis has been done of a comparably small portfolio.

3.6 Further observations

We explore further Models 1 and 2. It turns out that once the first principal component is considered in the model (Model 2), the clusters of Model 1 are no longer needed. This is because the first principal component is highly related to the selected clusters. The 2-means clusters are shown in the left column of Figure 6 (blue and red color). We see that the separation between the two clusters is almost a vertical line w.r.t. the first two principal components of the PCA (top-left graph of Figure 6). Therefore, the first principal component is able to explain the clustering of the 2-means algorithm. A similar argument applies to the bottom-right plot of Figure 6, from which we see that the transformed bottleneck activation is highly related with the first principal component. So we do not need both covariates in the model.
The 3-means clusters are shown in the middle column of Figure 6 (green, cyan and black color). The bottom-middle plot indicates that the bottleneck activations can capture the difference among the 3-means clusters, while the top-middle plot indicates that the first two principal components cannot capture the difference among the 3-means clusters. In fact, the first and third principal components of the PCA can capture the difference among the 3-means clusters as shown in the top-right plot of Figure 6.
Finally, we investigate the KL divergence of the three methods. Denote the average v-a heatmap of cluster k by the J-dimensional vector $\bar{\boldsymbol{x}}_{|k} \in \mathcal{X}$ with the following components

$$\bar{x}_{j|k} = \frac{1}{\sum_{i=1}^{n} \mathbb{1}_k(\mathcal{C}_K(\boldsymbol{x}_i))} \sum_{i=1}^{n} \mathbb{1}_k(\mathcal{C}_K(\boldsymbol{x}_i)) x_{i,j}.$$

Note that $\bar{\boldsymbol{x}}_{|k}, k = 1, \ldots, K$, is an approximation of the heatmaps for the drivers in cluster k. Next we turn to the PCA of Section 3.3. Denote $\boldsymbol{\Lambda}_1 = \operatorname{diag}(g_1, 0, \ldots, 0) \in \mathbb{R}^{J \times J}$, $\boldsymbol{\Lambda}_2 = \operatorname{diag}(g_1, g_2, 0, \ldots, 0) \in \mathbb{R}^{J \times J}$, and so on for $\boldsymbol{\Lambda}_q, q = 1, \ldots, J$. The best q-dimensional approximation with respect to the Frobenius norm of $\boldsymbol{X}^0$ is given by the first q principal components as follows

$$\hat{\boldsymbol{X}}_q^0 = \boldsymbol{U} \boldsymbol{\Lambda}_q \boldsymbol{V}'.$$

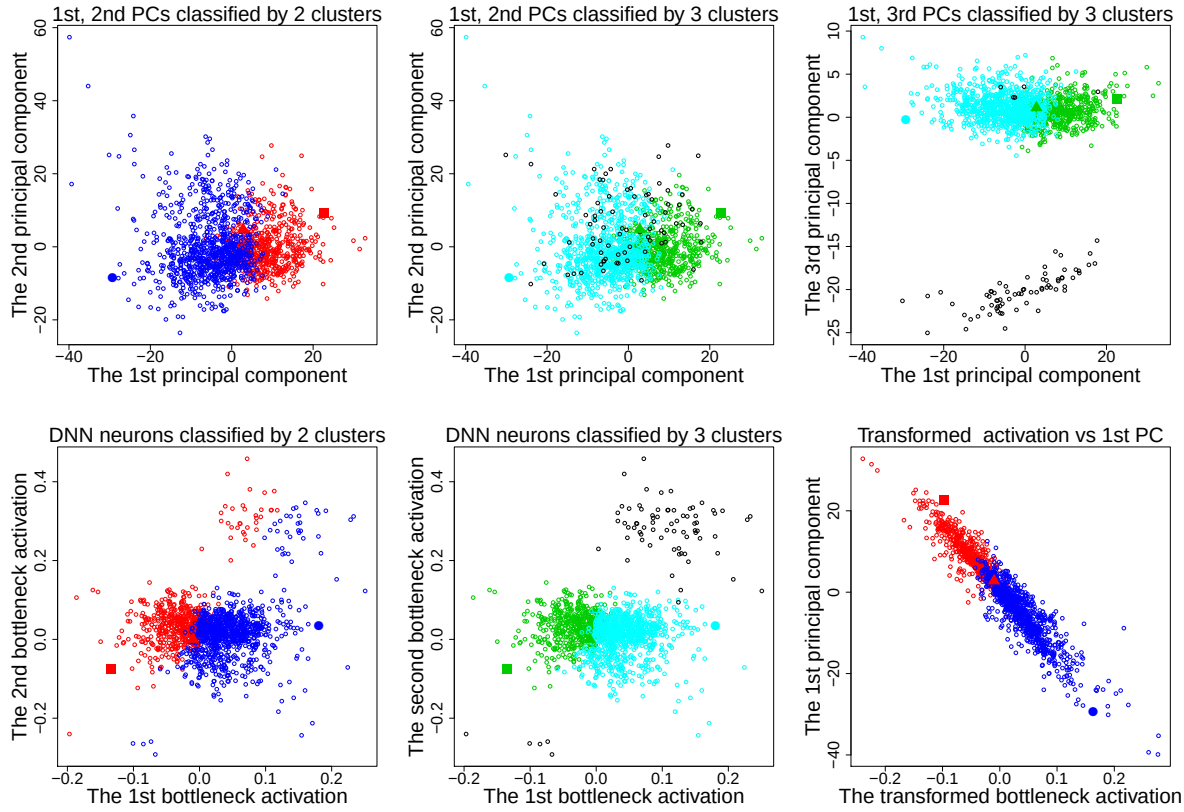

FIGURE 6: Red indicates cluster 1 and blue indicates cluster 2 of the 2-means clustering; green indicates cluster 1, cyan indicates cluster 2 and black indicates cluster 3 of the 3-means clustering. Square, triangle and circle symbols indicate drivers 72, 608 and 718, respectively. Each point corresponds to one of the $n = 1,478$ drivers' heatmaps.
Top-left: the first two principal components representations of heatmaps, classified by the 2-means clusters; top-middle: the first two principal components representations of heatmaps, classified by the 3-means clusters; top-right: the first and third principal components representations of heatmaps, classified by the 3-means clusters.
Bottom-left: the bottleneck activations representations of heatmaps, classified by the 2-means clusters; bottom-middle: the bottleneck activations representations of heatmaps, classified by the 3-means clusters; bottom-right: the relationship between the transformed bottleneck activation z_0 and the first principal component.

Denote the i-th row of the back-transformed (for column means and variances) $\hat{\boldsymbol{X}}_q$ by $\hat{\boldsymbol{x}}_i^q$, which is an approximation of the v-a heatmap $\boldsymbol{x}_i$ for driver i. If the minimum element in $\hat{\boldsymbol{x}}_i^q$ is non-positive denoted by $\hat{x}_{i,j'}^q$, we do the transformation $(\hat{\boldsymbol{x}}_i^q - (\hat{x}_{i,j'}^q - 0.001)\boldsymbol{e})/\sum_{j=1}^{J}(\hat{x}_{i,j}^q - \hat{x}_{i,j'}^q + 0.001)$, such that each element is positive, and where $\boldsymbol{e} = (1,\ldots,1)' \in \mathbb{R}^J$. For the bottleneck neural network, the approximation of v-a heatmap for driver i is given by $\pi(\boldsymbol{x}_i)$.
These approximations of the v-a heatmaps are compared to the original v-a heatmaps $\boldsymbol{x}_i$ using the KL divergence $d_{\mathrm{KL}}(\boldsymbol{x}_i||\hat{\boldsymbol{x}}_i)$ for $i = 1,\ldots,n$. Note that we set for the K-means algorithm approximation, $\hat{\boldsymbol{x}}_i = \bar{\boldsymbol{x}}_{|\mathcal{C}_K(\boldsymbol{x}_i)}$, for the principal component approximation, $\hat{\boldsymbol{x}}_i = \hat{\boldsymbol{x}}_i^q$ and for the bottleneck neural network approximation, $\hat{\boldsymbol{x}}_i = \pi(\boldsymbol{x}_i)$, respectively. The KL divergence $d_{\mathrm{KL}}(\boldsymbol{x}_i||\hat{\boldsymbol{x}}_i)$ is a measure of approximation error. In Appendix A, we consider another KL divergence (A.9) which is a measure of finite sample size error.
We compute the KL divergence of each driver for the three approximation methods as shown in

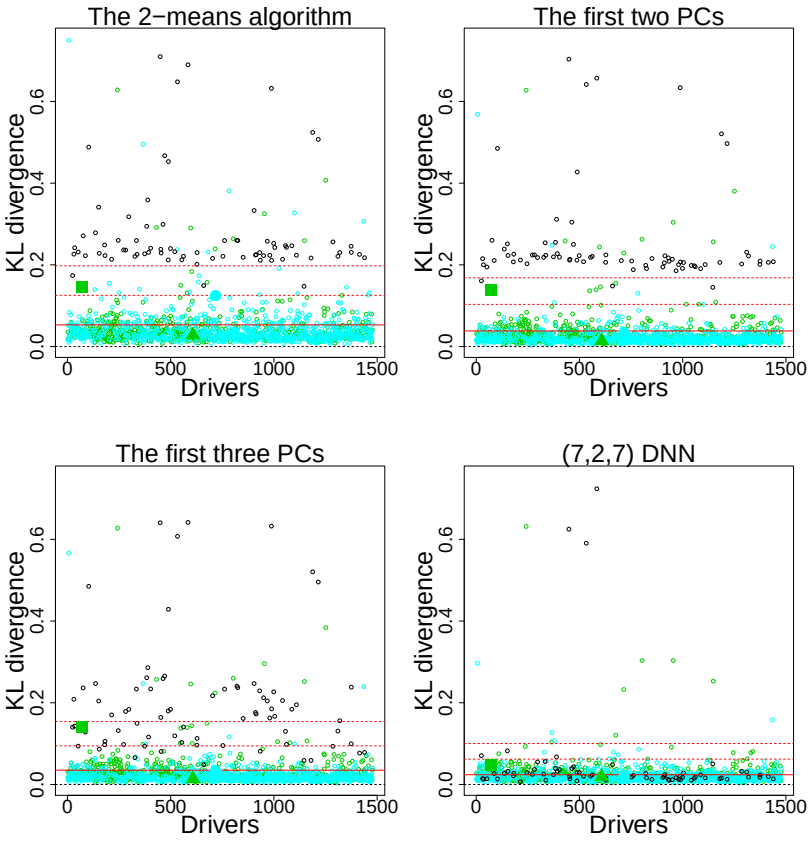

FIGURE 7: Green indicates cluster 1, cyan indicates cluster 2 and black indicates cluster 3 of the 3-means clusters. Square, triangle and circle symbols indicate drivers 72, 608 and 718, respectively. The four plots show the KL divergence for each driver individually using the four approximation methods: 2-means clusters, the first two principal components, the first three principal components and the (7,2,7) bottleneck neural network. The solid line indicates the average KL divergence and the dotted lines indicate 1 and 2 standard deviations.

Figure 7. The first three plots (1st row and left 2nd row) indicate that the 2-means algorithm and the first three principal components cannot approximate the heatmaps of cluster 3 (in black) as accurately as cluster 1 and 2. In contrast, the bottleneck neural network can approximate almost all heatmaps well. Unfortunately, for this data set, this marginal advantage of the neural network does not provide obvious benefits in the claims frequency modeling, comparing Models C.2 and C.3 in Table 5. Model 2 and Model 3 perform almost equally well. This is due to the fact that in Model 3 we consider the transformed bottleneck activations $z_0 = z_1^{(2)} - 0.5 z_2^{(2)}$ in the regression model, see (3.9)-(3.10). These transformed bottleneck activations are constant on straight lines $z_1^{(2)} \mapsto z_2^{(2)} = c + 2z_1^{(2)}$. This indicates that the black dots in the bottom-middle plot of Figure 6 are divided into two groups under Model 3 which have the same structure as the red-blue colored dots in the bottom-left plot of Figure 6.

4 Conclusions

The most important finding is that the telematics covariates extracted from speed-acceleration heatmaps improves out-of-sample prediction in claims frequency modeling compared to tradi-

tional actuarial pricing covariates. These telematics covariates include the first principal component and the transformed bottleneck activation from the bottleneck neural network (autoencoder). These two telematics covariates have a linear relationship as shown in the bottom-right of Figure 6. The two covariates can be related to the claims frequency in a simple log-linear form, which increases the chance of their practical application in actuarial pricing models. The first principal component is interpreted as a safe driving index at low speeds $[5, 20]$km/h, and a similar interpretation applies to the transformed bottleneck activation. Though the details of telematics car driving data cleaning are not shown in this paper, we emphasize that the telematics car driving data cleaning is crucial (and time consuming) to obtain a meaningful result.
We have investigated claims frequency models, the low speed interval and the longitudinal acceleration rate. The logic underlying is that most accidents occur at low speeds. For claims severity models, high speed intervals and lateral acceleration rates are also of interest and should be investigated next.

Acknowledgments

Guangyuan Gao and Shengwang Meng acknowledge the financial support from National Social Science Fund of China (Grant No. 16ZDA052) and MOE National Key Research Bases for Humanities and Social Sciences (Grant No. 16JJD910001).

References

Ayuso, M., Guillen, M., and Pérez-Marín, A. M. (2016). Telematics and gender discrimination: some usage-based evidence on whether men's risk of accidents differs from women's. *Risks* **4,** article 10.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York.

Gao, G. and Wüthrich, M. V. (2017). Feature extraction from telematics car driving heatmaps. *SSRN ID: 3070069* .

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman & Hall, New York.

Hastie, T. J., Tibshirani, R. J., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, second edition.

Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal* **37,** 233–243.

Verbelen, R., Antonio, K., and Claeskens, G. (2018). Unraveling the predictive power of telematics data in car insurance pricing. *Journal of the Royal Statistical Society: Series C (Applied Statistics), forthcoming* .

Weidner, W., Transchel, F. W. G., and Weidner, R. (2016a). Classification of scale-sensitive telematic observables for riskindividual pricing. *European Actuarial Journal* **6,** 1–22.

Weidner, W., Transchel, F. W. G., and Weidner, R. (2016b). Telematic driving profile classification in car insurance pricing. *Annals of Actuarial Science* **11,** 213–236.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R.* Chapman & Hall, New York.

Wüthrich, M. V. (2017). Covariate selection from telematics car driving data. *European Actuarial Journal* **7,** 89–108.

Wüthrich, M. V. and Buser, C. (2017). Data analytics for non-life insurance pricing. *SSRN ID: 2870308* .

A Stability of v-a heatmaps

In our study we have restricted the vast amount of telematics car driving data of 1.2 TB to 3 months of observations, i.e. from 01/05/2016 to 31/07/2016. In this appendix we investigate whether these 3 months of observations are sufficient to receive stable v-a heatmaps in the speed interval $[5, 20]$km/h. We therefore fix a driver i and we split his/her telematics car driving data into the different days $d = 1, \ldots, D$ of observations. This splitted data is considered for each day individually and we assume that the non-driving days have already been dropped. For notational convenience, we drop index i of the chosen driver in the derivations in this appendix.
Denote by T^d the total time (in seconds) which driver i spents in speed bucket $[5, 20]$km/h on day $d = 1, \ldots, D$, and denote by $\boldsymbol{x}^d \in \mathcal{X}$ the resulting v-a heatmap of driver i on that day d.
We make the following model assumption: the random vectors $(T^d, \boldsymbol{x}^d)$ are i.i.d. for $d \geq 1$. Note that we ignore the effects of weather, personal emotions, etc., which may affect several days in a row and, hence, potentially violate the i.i.d. assumption. The i.i.d. assumption allows us to apply the law of large numbers which implies that

$$\lim_{D\to\infty} \frac{1}{D} \sum_{d=1}^{D} T^d = \mathbb{E}[T^1] \qquad \text{and} \qquad \lim_{D\to\infty} \frac{1}{D} \sum_{d=1}^{D} T^d \boldsymbol{x}^d = \mathbb{E}[T^1 \boldsymbol{x}^1], \qquad \mathbb{P}\text{-a.s.} \tag{A.1}$$

Note that all considered random variables are bounded, $\mathbb{P}$-a.s., and henceforth all moments exist. Merging the two limits above provides, $\mathbb{P}$-a.s.,

$$\boldsymbol{x} = \boldsymbol{x}_i \overset{\text{def.}}{=} \lim_{D\to\infty} \frac{\sum_{d=1}^{D} T^d \boldsymbol{x}^d}{\sum_{d=1}^{D} T^d} = \lim_{D\to\infty} \frac{\frac{1}{D}\sum_{d=1}^{D} T^d \boldsymbol{x}^d}{\frac{1}{D}\sum_{d=1}^{D} T^d} = \frac{\mathbb{E}[T^1 \boldsymbol{x}^1]}{\mathbb{E}[T^1]}. \tag{A.2}$$

Note that we drop subscript i in the v-a heatmaps $\boldsymbol{x} = \boldsymbol{x}_i$ in the following. Limit (A.2) defines the true v-a heatmap of driver i in speed interval $[5, 20]$km/h under our model assumptions. The aim of the subsequent analysis is to study the speed of convergence in (A.2), in order to receive the minimal D where the following expression becomes stable in D

$$\widehat{\boldsymbol{x}}^{(D)} = \widehat{\boldsymbol{x}}_i^{(D)} = \frac{\sum_{d=1}^{D} T^d \boldsymbol{x}^d}{\sum_{d=1}^{D} T^d}, \tag{A.3}$$

i.e. what is the minimal D such that $\boldsymbol{x} \approx \widehat{\boldsymbol{x}}^{(D)}$ where the accuracy is going to be measured in the KL divergence. Remark that this minimal D may depend on the chosen driver i.

We start by introducing some transformations. Note that $\boldsymbol{x}^d = (x_1^d, \ldots, x_J^d)'$ lives on the $(J-1)$-unit simplex $\mathcal{X}$, therefore we have normalization

$$x_J^d = 1 - \sum_{j=1}^{J-1} x_j^d. \tag{A.4}$$

This implies that it is sufficient to study the first $J-1$ components of $\boldsymbol{x}^d$. This also motivates the definition of the $(J-1)$-dimensional (uniformly bounded) random vector

$$\boldsymbol{Z}^d = (Z_1^d, \ldots, Z_{J-1}^d)' = T^d(x_1^d, \ldots, x_{J-1}^d)'.$$

Note that $\boldsymbol{Z}^d$ are by assumption i.i.d. for $d \geq 1$. Equation (A.3) is then modified accordingly by dropping the last component to receive

$$\widehat{\boldsymbol{x}}_\circ^{(D)} = \frac{\sum_{d=1}^D \boldsymbol{Z}^d}{\sum_{d=1}^D T^d} = \frac{\mathbb{E}[\boldsymbol{Z}^1] + \left(\frac{1}{D}\sum_{d=1}^D \boldsymbol{Z}^d - \mathbb{E}[\boldsymbol{Z}^1]\right)}{\mathbb{E}[T^1] + \left(\frac{1}{D}\sum_{d=1}^D T^d - \mathbb{E}[T^1]\right)} = \frac{\mathbb{E}[\boldsymbol{Z}^1] + \varepsilon^{(D)}}{\mathbb{E}[T^1] + \delta^{(D)}},$$

where the latter defines the $(J-1)$-dimensional residual $\varepsilon^{(D)}$ in the numerator and the one-dimensional residual $\delta^{(D)}$ in the denominator, respectively. The law of large numbers (A.1) implies that these residuals converge to 0, $\mathbb{P}$-a.s., see also (A.2). Considering the first order Taylor expansion of the last ratio provides for small residuals

$$\widehat{\boldsymbol{x}}_\circ^{(D)} - \frac{\mathbb{E}[\boldsymbol{Z}^1]}{\mathbb{E}[T^1]} = \frac{\varepsilon^{(D)}}{\mathbb{E}[T^1]} - \frac{\delta^{(D)}\mathbb{E}[\boldsymbol{Z}^1]}{\mathbb{E}[T^1]^2} + o(\varepsilon^{(D)}, \delta^{(D)}).$$

Under sufficient regularity on the underlying distribution we can apply the central limit theorem which provides

$$\sqrt{D}\boldsymbol{\Sigma}^{-1/2}\left(\widehat{\boldsymbol{x}}_\circ^{(D)} - \frac{\mathbb{E}[\boldsymbol{Z}^1]}{\mathbb{E}[T^1]}\right) \implies \mathcal{N}(0, \mathbb{1}_{J-1}), \qquad \text{for } D \to \infty,$$

where the covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{(J-1)\times(J-1)}$ is given by

$$\boldsymbol{\Sigma} = \frac{1}{\mathbb{E}[T^1]^2}\left(\operatorname{Cov}\left[Z_j^1 - \frac{T^1}{\mathbb{E}[T^1]}\mathbb{E}[Z_j^1],\ Z_l^1 - \frac{T^1}{\mathbb{E}[T^1]}\mathbb{E}[Z_l^1]\right]\right)_{j,l=1,\ldots J-1}. \tag{A.5}$$

Next we translate the above central limit theorem to convergence in distribution in the KL divergence. A second order Taylor expansion of the KL divergence provides the following approximation

$$d_{\mathrm{KL}}(\boldsymbol{x}\|\widehat{\boldsymbol{x}}^{(D)}) = -\sum_{j=1}^J x_j \log\frac{\widehat{x}_j^{(D)}}{x_j} \approx \left(\widehat{\boldsymbol{x}}_\circ^{(D)} - \frac{\mathbb{E}[\boldsymbol{Z}^1]}{\mathbb{E}[T^1]}\right)' \boldsymbol{H}(\boldsymbol{x})\left(\widehat{\boldsymbol{x}}_\circ^{(D)} - \frac{\mathbb{E}[\boldsymbol{Z}^1]}{\mathbb{E}[T^1]}\right),$$

with Hessian for $\boldsymbol{x} \in \mathcal{X}$

$$\boldsymbol{H}(\boldsymbol{x}) = \left(x_j^{-1}1_{\{j=l\}} + x_J^{-1}\right)_{j,l=1,\ldots J-1} \in \mathbb{R}^{(J-1)\times(J-1)}. \tag{A.6}$$

In the calculation we also use the normalization (A.4). Again under suitable regularity conditions on the underlying distributions this provides the following convergence in distribution

$$\lim_{D\to\infty} D\; d_{\mathrm{KL}}(\boldsymbol{x}\|\widehat{\boldsymbol{x}}^{(D)}) = \boldsymbol{Y}'\boldsymbol{V}(\boldsymbol{x})\boldsymbol{Y}, \tag{A.7}$$

with $\boldsymbol{Y} \sim \mathcal{N}(0, \mathbb{1}_{J-1})$ and

$$\boldsymbol{V}(\boldsymbol{x}) = \boldsymbol{\Sigma}^{1/2}\boldsymbol{H}(\boldsymbol{x})\boldsymbol{\Sigma}^{1/2}.$$

We analyze the right-hand side of (A.7). We start by calculating the moment generating function of the random variable $\boldsymbol{Y}'\boldsymbol{V}(\boldsymbol{x})\boldsymbol{Y}$ for sufficiently small real number r. A standard calculation with Gaussian densities provides

$$M_{\boldsymbol{Y}'\boldsymbol{V}(\boldsymbol{x})\boldsymbol{Y}}(r) = \mathbb{E}\left[\exp\left\{r\boldsymbol{Y}'\boldsymbol{V}(\boldsymbol{x})\boldsymbol{Y}\right\}\right] = \left(\det\left(\mathbb{1}_{J-1} - 2r\boldsymbol{V}(\boldsymbol{x})\right)\right)^{-1/2}.$$

Denote by $\eta_1(\boldsymbol{x}) \geq \ldots \geq \eta_{J-1}(\boldsymbol{x}) \geq 0$ the $J-1$ eigenvalues of $\boldsymbol{V}(\boldsymbol{x})$. Then the moment generating function of $\boldsymbol{Y}'\boldsymbol{V}(\boldsymbol{x})\boldsymbol{Y}$ for $r < 1/(2\eta_1(\boldsymbol{x}))$ is given by

$$M_{\boldsymbol{Y}'\boldsymbol{V}(\boldsymbol{x})\boldsymbol{Y}}(r) = \prod_{j=1}^{J-1}\left(\frac{1}{1-2r\eta_j(\boldsymbol{x})}\right)^{1/2}.$$

This implies that

$$\boldsymbol{Y}'\boldsymbol{V}(\boldsymbol{x})\boldsymbol{Y} \overset{(d)}{=} \sum_{j=1}^{J-1}\eta_j(\boldsymbol{x})\chi_j^2,$$

where χ_j^2 are i.i.d. χ^2-distributed random variables with 1 degree of freedom. From this and (A.7) we conclude the following convergence in distribution (under suitable regularity conditions)

$$\lim_{D\to\infty} D\; d_{\mathrm{KL}}(\boldsymbol{x}\|\widehat{\boldsymbol{x}}^{(D)}) = \sum_{j=1}^{J-1}\eta_j(\boldsymbol{x})\chi_j^2. \tag{A.8}$$

The first two moments on the right-hand side are given by

$$\mathbb{E}\left[\boldsymbol{Y}'\boldsymbol{V}(\boldsymbol{x})\boldsymbol{Y}\right] = \sum_{j=1}^{J-1}\eta_j(\boldsymbol{x}) \qquad \text{and} \qquad \mathrm{Var}\left(\boldsymbol{Y}'\boldsymbol{V}(\boldsymbol{x})\boldsymbol{Y}\right) = 2\sum_{j=1}^{J-1}\eta_j^2(\boldsymbol{x}).$$

From this we conclude that for sufficiently large D we have the following finite sample size error approximation

$$d_{\mathrm{KL}}(\boldsymbol{x}\|\widehat{\boldsymbol{x}}^{(D)}) \;\approx\; \frac{1}{D}\sum_{j=1}^{J-1}\eta_j(\boldsymbol{x}), \tag{A.9}$$

and the uncertainty in this approximation can be quantified by the asymptotic variance

$$\frac{2}{D^2}\sum_{j=1}^{J-1}\eta_j^2(\boldsymbol{x}).$$

Numerical results

We first estimate the covariance matrix (A.5). In order to get a full $(J-1)$-rank covariance matrix we need to choose $J-1 < D = 92$ (number of days between 01/05/2016 and 31/07/2016). Therefore, we work on a coarser grid with $J = 16 \times 4 = 64$, i.e. similar to the top-left plot in Figure 1 but dividing the acceleration axis into 4 segments with the same length. For each driver i, we estimate $\mathbb{E}[T_i^1]$ and $\mathbb{E}[\boldsymbol{Z}_i^1]$ by

$$\hat{\mathbb{E}}[T_i^1] = \frac{1}{D_i}\sum_{d=1}^{D_i} T_i^d \qquad \text{and} \qquad \hat{\mathbb{E}}[\boldsymbol{Z}_i^1] = \frac{1}{D_i}\sum_{d=1}^{D_i} \boldsymbol{Z}_i^d,$$

note that we also introduce the variable $D_i \leq 92$ to indicate that driver i is not necessarily driving on all $D = 92$ days between 01/05/2016 and 31/07/2016.
The covariance matrix $\hat{\boldsymbol{\Sigma}}_i$ is estimated by the sample covariance of the D_i available $(J-1)$-dimensional observations of driver i

$$\left(Z_{i,1}^d - \frac{T_i^d}{\hat{\mathbb{E}}[T_i^1]}\hat{\mathbb{E}}[Z_{i,1}^1], \ldots, Z_{i,J-1}^d - \frac{T_i^d}{\hat{\mathbb{E}}[T_i^1]}\hat{\mathbb{E}}[Z_{i,J-1}^1] \right)', \qquad \text{for } d = 1, \ldots, D_i.$$

Then we estimate the Hessian as $\boldsymbol{H}(\hat{\boldsymbol{x}}_i^{(D_i)})$ for each driver i. This is easily obtained by applying (A.3) and (A.6). Finally, we calculate the eigenvalues of matrix $\hat{\boldsymbol{V}}(\boldsymbol{x}_i) = \hat{\boldsymbol{\Sigma}}_i^{1/2}\boldsymbol{H}(\hat{\boldsymbol{x}}_i^{(D_i)})\hat{\boldsymbol{\Sigma}}_i^{1/2}$, and hence the approximated asymptotic KL divergence and its uncertainty for each driver i. These two quantities are plotted in the first row of Figure 8. We observe that the finite sample size errors are much smaller than the approximation errors in Figure 7. Note that 390 drivers have no more than $J-1$ days' telematics data, i.e. $D_i \leq J-1 = 63$. For those drivers, $\hat{\boldsymbol{\Sigma}}^{1/2}$ cannot be obtained and we ignore those drivers in Figure 8.
Finally, we determine the minimal days D_i^* and the minimal amount in seconds T_i^* for each driver i such that the finite sample size error in (A.9) is of similar size (of 0.05) as in Figure 7. The minimal days D_i^* is obtained as $\sum_{j=1}^{J-1}\eta_j(\boldsymbol{x}_i)/0.05$, see (A.9). The minimal amount in seconds T_i^* is then given by $D_i^*\hat{\mathbb{E}}[T_i^1]$. The median, mean, 3rd quantile and 99% quantile of the minimal amount are 67, 81, 87 and 305 in minutes (ignoring the drivers with $D_i \leq 63$). So for 99% of the evaluated car drivers 300 minutes of telematics car driving data is sufficient to obtain a stable v-a heatmap. During data cleaning, we have discarded the drivers with less than 300 minutes' telematics car driving data. We plot the minimal days and the minimal seconds on the second row of Figure 8. Another observation is that none of these quantities are sensitive to the 3-means clusters, see coloring in Figure 8.

B Summary statistics and figures

In Table 6 we illustrate the chosen age classes and the aggregated years-at-risk distribution. The age classes are used in Models A.0 and A.3 in Table 4. In Figure 9 (left-hand side and middle) we illustrate the portfolio of the 1,478 drivers considered. The graph on the left-hand side gives the driver's age structure and the graph in the middle gives the car's age structure. In Figure 10 we illustrate the scatter plot of these two covariates. We note that the volume below age 25 is comparably small. In Figures 9-10 we also illustrate the corresponding volumes for the

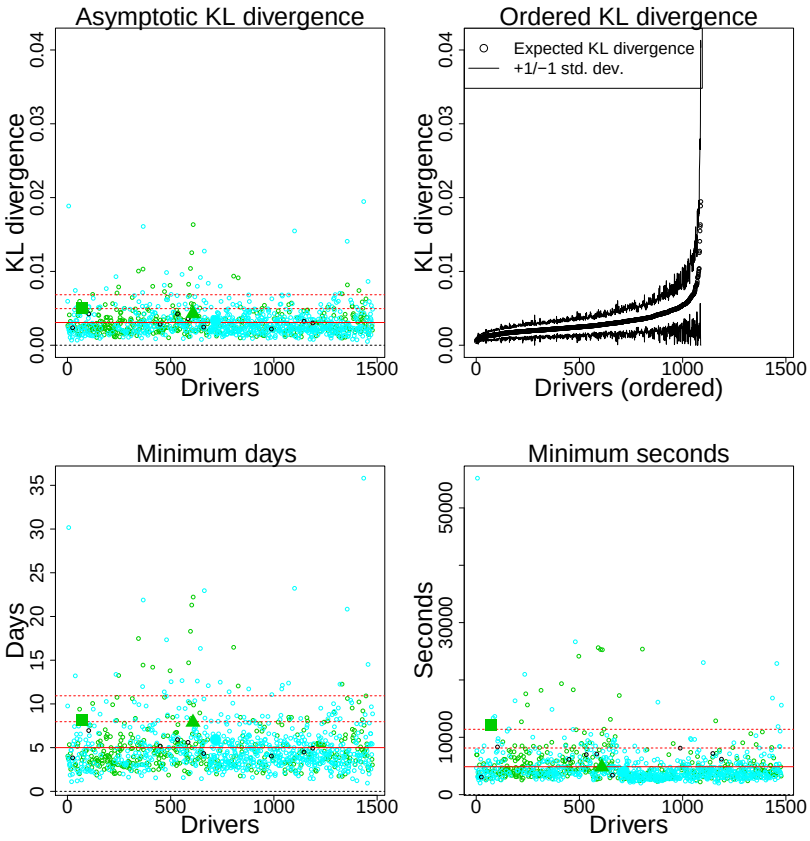

Figure 8: Green indicates cluster 1, cyan indicates cluster 2 and black indicates cluster 3 of 3-means clusters. Square, triangle and circle symbols indicate drivers 72, 608 and 718, respectively. The four plots show the asymptotic KL divergence, the 1 standard deviation confidence bound, the minimal days and the minimal seconds for each driver. The solid line indicates the average value and the dotted lines indicate 1 and 2 standard deviations.

transformed bottleneck activations z_0.

In Figure 11 we provide the 5-fold cross-validation estimates of deviance statistics, comparing to the 10-fold cross-validation results in Section 3.2-3.4. In Figure 12 we provide the marginally predicted claims frequencies for the covariates driver's age, car's age and transformed bottleneck activation. These are obtained from Model 3 in Table 4.

Table 6: The aggregated years-at-risk distribution across the chosen age classes.

Age classes	[18, 25]	[26, 30]	[31, 35]	[36, 40]	[41, 45]	[46, 50]	[51, 80]
Exposures	236	696	714	492	496	338	360

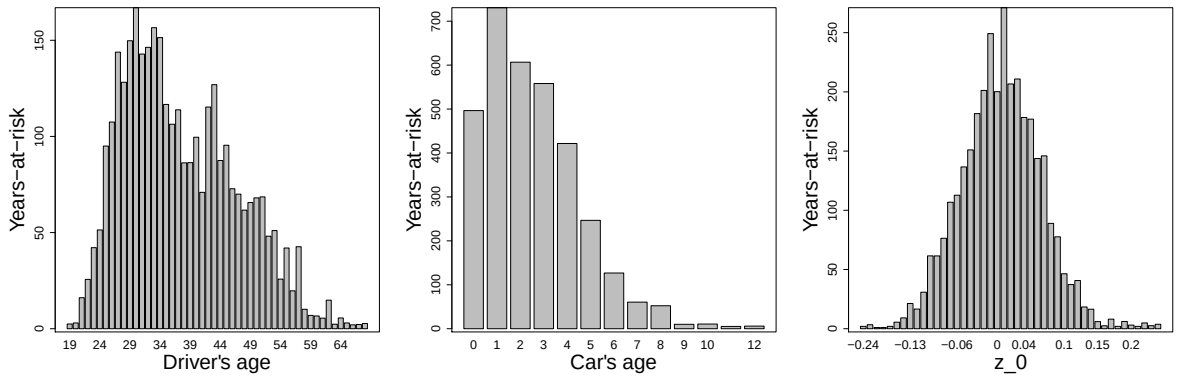

FIGURE 9: The years-at-risk (volume) per covariate label in Model 3 (driver's age, car's age and bottleneck activation z_0); the transformed bottleneck activation z_0 is rounded to 0.01.

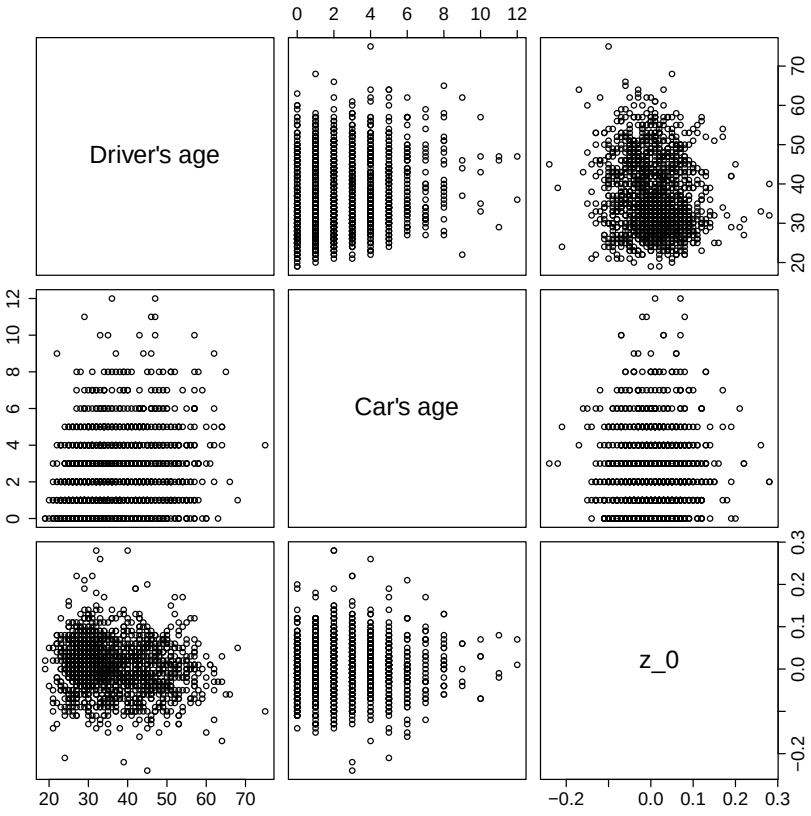

FIGURE 10: Scatter plot matrix for the covariates used in Model 3 (driver's age, car's age and bottleneck activation z_0).

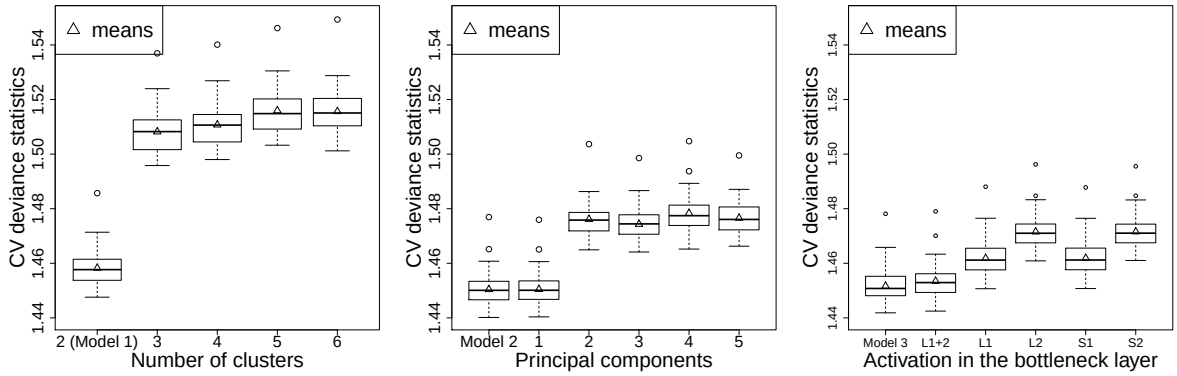

FIGURE 11: 5-fold cross-validation estimates of deviance statistics, comparing to the left graphs of Figures 2, 4, and 5 (please also refer to Table 4 for Models 1-3).

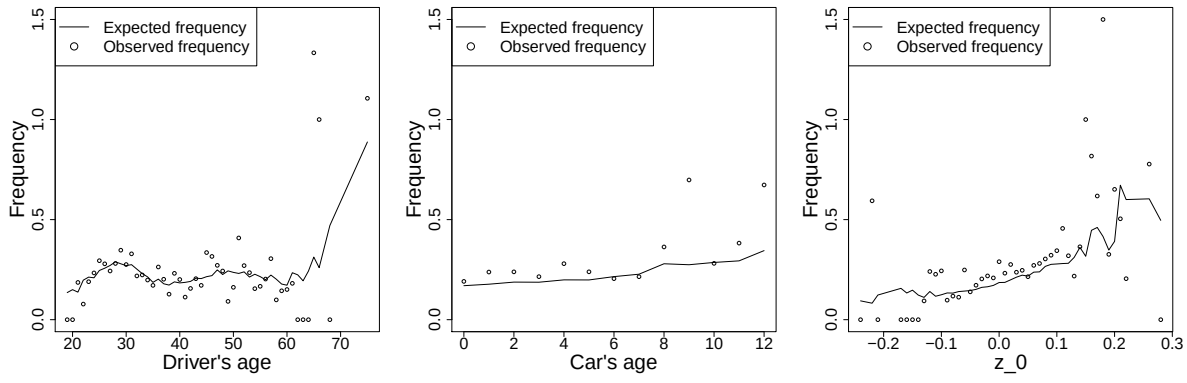

FIGURE 12: Predicted marginal claims frequency per driver's age, per car's age, and per transformed bottleneck activation z_0 from Model 3.