Driving risk evaluation based on telematics data

Guangyuan Gao* Mario V. Wüthrich† Hanfang Yang*

November 21, 2018

Abstract

Telematics car driving data describes drivers' driving characteristics. This paper studies the predictive power of telematics data for claims frequency prediction. We first extract covariates from telematics car driving data using K-mediods clustering and principal components analysis. These telematics covariates are then used as explanatory variables for claims frequency modeling, in which we analyze their predictive power. Moreover, we use these telematics covariates to challenge the classical covariates usually used.

Keywords: Telematics data; Driving habit; Driving style; v-a heatmap; Acceleration pattern; K-mediods algorithm; Principal components analysis; Generalized additive model; Generalized linear model; Variable selection; Collinearity; Poisson regression; Deviance statistics; Claims frequency modeling; Car insurance pricing.

1 Introduction

Telematics car driving data can be collected by drivers' mobile phones or by black box devices installed in cars. While the telematics data from black box devices directly describes drivers' driving habits and driving styles, the telematics data from drivers' mobile phones needs to be pre-processed to detect driving phases. Our telematics data is collected from black box devices installed by a Chinese insurance company. Drivers are encouraged to install such devices in return for a premium discount. However, telematics data information is not yet allowed to be used in car insurance tariffication under the current regulation in China, and most relevant insurance products such as usage-based insurance (UBI) and pay-as-you-drive (PAYD) products are forbidden by the Chinese insurance regulatory authority. Nevertheless, almost all the car insurance companies and many technology companies in China have started to accumulate telematics data and evaluate driving risk. Except for privacy concerns, telematics car driving data will certainly benefit both insured and insurers in terms of optimizing premiums and promoting safe driving.

Even without the support of evidence from telematics data and accident data, it is often observed that frequent drivers are more exposed to risk than less frequent ones, and aggressive drivers tend to be more likely involved in accidents than careful ones. With the development of telematics technology (the integration of telecommunication and informatics), researchers from both the

*Center for Applied Statistics and School of Statistics, Renmin University of China, 100872 Beijing, China.
†ETH Zurich, RiskLab, Department of Mathematics, 8092 Zurich, Switzerland.

actuarial field and non-actuarial fields have started to investigate these observations in depth and to quantify the corresponding driving risk. Ayuso et al. [1] study the risk exposure to total driving distances. Paefen et al. [8] and Verbelen et al. [11] partition the total driving distances by road type and time slots. Verbelen et al. [11] further use these compositional predictors and classical risk factors in claims frequency models. Weidner et al. [12, 13] investigate driving styles using the discrete Fourier transform.

We observe that most accidents occur at low speeds, for instance, in parking spots and during traffic congestion. From this intuition, Gao et al. [3] study the interaction between speed and acceleration rate, represented by so called v-a heatmaps, in the low speed bucket $[5, 20]$km/h. The v-a heatmaps basically describe the speed-acceleration pattern as a two-dimensional functional illustrated by level curves. Gao et al. [3] have shown that the telematics covariates from low speeds have a significant predictive power for claims frequency prediction. Nevertheless, the chosen speed bucket $[5, 20]$km/h in Gao et al. [3] is rather empirical, and it is desirable to analyze different speed buckets in a more systematic way. In this paper we study how a driver behaves at different speeds, and whether driving styles at low speeds are more related to accidents.

In this paper, we implement K-mediods clustering and principal components analysis to decompose the v-a heatmap functional in different speed buckets. We refer to Kaufman and Rousseeuw [5] and Hastie et al. [4] for the K-mediods clustering and its advantages/disadvantages compared with the K-means clustering; we refer to Pearson [9] and Hastie et al. [4] for the principal components analysis. We use our telematics covariates in generalized additive models (Wood [14]) to describe the corresponding claims frequencies. Generalized additive models can address non-linear effects of telematics covariates. We then implement a smooth component selection approach proposed by Marra and Wood [6], which effectively turns generalized additive models into generalized linear models (McCullagh and Nelder [7]). Analysis of variance (ANOVA) is used to compare the in-sample predictive power of telematics covariates; see Fisher [2]. Cross-validation of Poisson deviance statistics is used to evaluate the out-of-sample predictive power of telematics covariates; see Hastie et al. [4] and Wüthrich and Buser [15]. Moreover, we use these results to challenge classical covariates like driver's age, gender, car's age, region, etc.

The paper is structured as follows. Section 2 constructs the v-a heatmaps in different speed buckets. Moreover, it compares v-a heatmaps in different speed buckets using K-mediods clustering and principal components analysis. Section 3 studies acceleration patterns. Section 4 establishes claims frequency models using both classical covariates and the covariates extracted from telematics data in Sections 2 and 3. Section 5 concludes the paper with several findings.

2 Telematics v-a heatmaps

Our telematics data collects the car's status second by second. That is, every second we receive the GPS location and the current speed and acceleration in all directions from the black box devices installed in the cars. Note that the GPS data is sometimes subject to the GPS drifting problem. The telematics data must therefore be cleaned for outliers and missing values in advance. In our analysis we consider the telematics data of three months of driving experience from 01/05/2016 to 31/07/2016. This amount of data is sufficient for getting stable results, for details see Appendix A. An assumption made here is that a driver's driving characteristics

remain the same during his/her policy period.

2.1 Construction of v-a heatmaps in different speed buckets

We focus on speed v and longitudinal acceleration a in the v-a rectangle denoted by $R = (0, 80]\text{km/h} \times [-2, 2]\text{m/s}^2$. Note that we extend the previously analyzed speed interval $[5, 20]$km/h from Gao et al. [3] to $(0, 80]$km/h because we want to study the driving styles both at low and high speeds; we cap the acceleration rate a within $[-2, 2]\text{m/s}^2$ because we do not have sufficiently many observations outside of this interval. We partition the speed interval $(0, 80]$km/h into four speed buckets $(0, 20]$km/h, $(20, 40]$km/h, $(40, 60]$km/h, and $(60, 80]$km/h, see rectangles 1-4 in Figure 1. This partition of the speed interval $(0, 80]$km/h to four sub-speed buckets

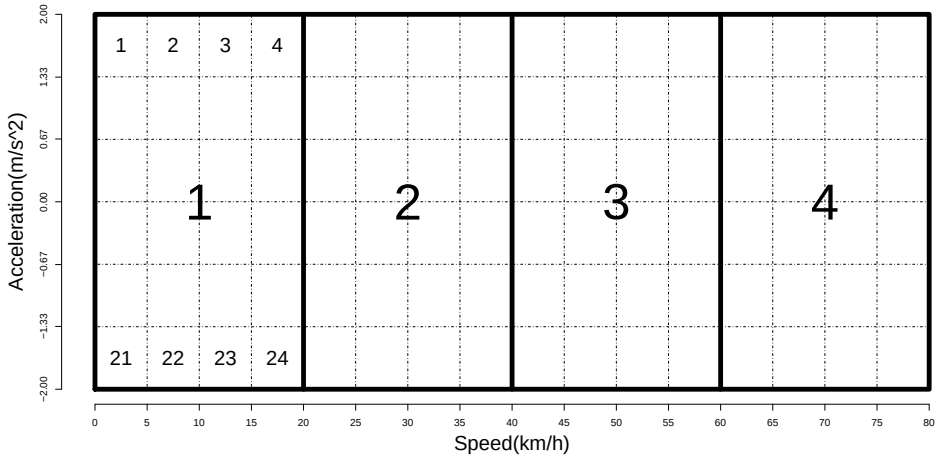

Figure 1: The partition of $R = (0, 80] \times [-2, 2]$.

is done because we would like to study driving styles rather than driving habits. We consider normalized quantities on these sub-speed buckets which exactly achieves to make drivers with different driving habits (city drivers vs. highway drivers) more comparable in driving styles.
For each speed bucket $m = 1, \ldots, 4$, we further divide the v-axis (speed) into 4 intervals and the a-axis (acceleration) into 6 intervals, which results in 24 sub-rectangles $(R_{m,j})_{j=1:24}$ in each speed bucket m (see the numbers in speed bucket 1 in Figure 1). For each driver i, we denote the amount of time spent in $R_{m,j}$ by $t_{i,m,j}$. Given a speed bucket m, for each driver i we calculate the relative amount of time spent in $R_{m,j}$ as

$$z_{i,m,j} = \frac{t_{i,m,j}}{t_{i,m}} \geq 0, \tag{2.1}$$

where $t_{i,m} = \sum_{j=1}^{24} t_{i,m,j}$ is the total amount of time spent in speed bucket m by driver i. Equation (2.1) induces an empirical discrete distribution $\boldsymbol{z}_{i,m} = (z_{i,m,1}, \ldots, z_{i,m,24})'$ on speed bucket m, which lies in the $(24-1)$-unit simplex $\mathcal{Z} \subset \mathbb{R}_+^{24}$, i.e., has normalization

$$\sum_{j=1}^{24} z_{i,m,j} = 1. \tag{2.2}$$

The driving style of every car driver i is described by a J-vector $\boldsymbol{x}_i = (\boldsymbol{z}'_{i,1}, \ldots, \boldsymbol{z}'_{i,4})' \in \mathbb{R}^J$ containing the four discrete distributions on rectangles $m = 1, \ldots, 4$. This can be illustrated by

four v-a heatmaps. Note that the dimension of $\boldsymbol{x}_i$ is $J = 24 \times 4 = 96$. Also note that we have the following relationship between elements in $\boldsymbol{z}_{i,m}$ and elements in $\boldsymbol{x}_i$:

$$z_{i,m,j} = x_{i,(m-1)\times 24+j}, \tag{2.3}$$

for $i = 1, \ldots, 973,\ m = 1, \ldots, 4,\ j = 1, \ldots, 24$. We draw the four v-a heatmaps jointly for driver 155 and driver 820 in Figure 2. It shows that the width of the level sets on the a-axis of driver

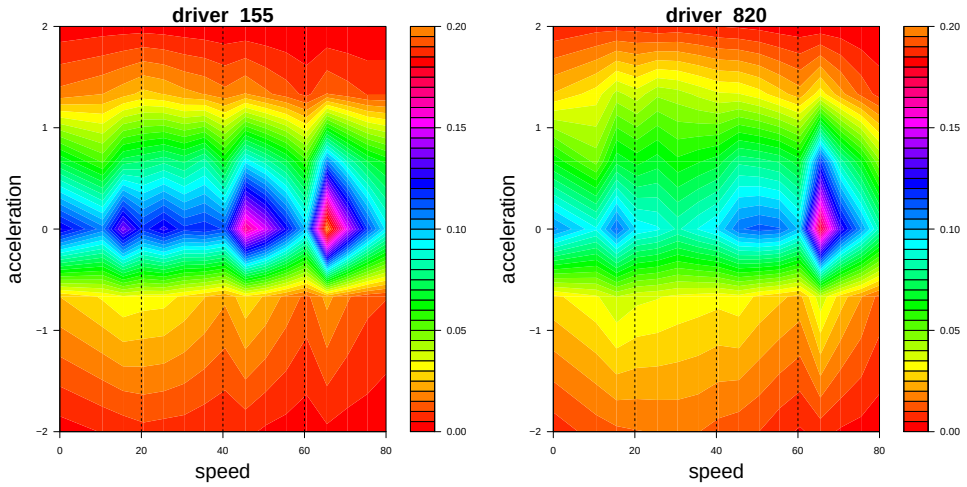

Figure 2: v-a heatmaps of drivers 155 and 820.

820 is wider than the one of driver 155 in all speed buckets, indicating a stronger acceleration and braking of driver 820.

Driving habit of driver i is described by the relative amount of time spent in each speed bucket m:

$$h_{i,m} = \frac{t_{i,m}}{t_i}, \qquad \text{for } m = 1, \ldots, 4, \tag{2.4}$$

where $t_i = \sum_{m=1}^{4} t_{i,m}$ is the total amount of time spent in the entire speed interval $(0, 80]$km/h by driver i. Equation (2.4) induces an empirical discrete distribution $\boldsymbol{h}_i = (h_{i,1}, \ldots, h_{i,4})'$ on R, which lies in the $(4-1)$-unit simplex $\mathcal{H} \subset \mathbb{R}_+^4$, and has normalization $\sum_{m=1}^{4} h_{i,m} = 1$. Suppose that a city driver i and a highway driver i' had the same driving style, we would have $h_{i,1} > h_{i',1}, h_{i,4} < h_{i',4}$, but $\boldsymbol{x}_i = \boldsymbol{x}_{i'}$. In this paper, we only consider the effects of driving style $\boldsymbol{x}_i$ on claims frequencies.

The volume of telematics data is increasing with the length of the observation period, and the above procedure can compress this increasing telematics data to a J-vector. Appendix A gives the minimum volumes required to get stable v-a heatmaps in the four speed buckets, respectively. There are $n = 973$ cars in our data meeting this minimum driving time requirements for the four speed buckets simultaneously. We stack the vectors $\boldsymbol{x}_i, i = 1, \ldots, n$, to form the $n \times J$ design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times J}$ describing these n car drivers. In the following we aim at applying the K-mediods clustering and the principal components analysis to reduce the dimension of $\boldsymbol{X}$ and extract the risk factors, still capturing explanatory power for claims frequency prediction.

2.2 K-mediods clustering

Clustering analysis aims at grouping the drivers with similar heatmaps into the same cluster, and it effectively reduces the dimension J of the design matrix $\boldsymbol{X}$ to the chosen number of clusters K. We apply the K-mediods clustering to $\boldsymbol{X}$. Compared to the K-means clustering, the K-mediods clustering is more robust against outliers and can be applied to any distance function such as Euclidean distance, Manhattan distance, Canberra distance, etc., at the expense of computational time (Hastie et al. [4]). Here, we consider the Euclidean distance.
Denote by $\mathcal{N} = \{1, \ldots, n\}$ the driver labels. Denote by $\mathcal{K} = \{1, \ldots, K\}$ the K cluster labels. Denote by $C = \{c_{K,1}, \ldots, c_{K,K}\} \subset \mathcal{N}$ the increasing ordered K mediod driver labels (i.e., $c_{K,1} < c_{K,2} < \ldots < c_{K,k}$). A classification structure is introduced by partitioning the set $\mathcal{N}$ into K disjoint clusters $\mathcal{N}_1, \ldots, \mathcal{N}_K$ satisfying

$$\bigcup_{k=1}^{K} \mathcal{N}_k = \mathcal{N} \qquad \text{and} \qquad \mathcal{N}_k \cap \mathcal{N}_{k'} = \emptyset \text{ for all } k \neq k'. \tag{2.5}$$

These K clusters define a classifier $\mathcal{C}$ on the set $\mathcal{N}$, given by

$$\mathcal{C} : \mathcal{N} \to \mathcal{K}, \qquad i \mapsto \mathcal{C}(i) = \sum_{k=1}^{K} k \mathbb{1}_{\{i \in \mathcal{N}_k\}}. \tag{2.6}$$

The within-cluster distance of the k-th cluster is defined as

$$W_k(\mathcal{C}) = \sum_{i \in \mathcal{N}_k} d(\boldsymbol{x}_i, \boldsymbol{x}_{c_{K,k}}), \tag{2.7}$$

where we use the Euclidean distance $d(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) = ||\boldsymbol{x}_i - \boldsymbol{x}_{i'}||_2$ for $\boldsymbol{x}_i, \boldsymbol{x}_{i'} \in \mathbb{R}^J$, and the mediod driver $c_{K,k}$ of cluster k satisfies

$$c_{K,k} = \arg\min_{j \in \mathcal{N}_k} \sum_{i \in \mathcal{N}_k} d(\boldsymbol{x}_i, \boldsymbol{x}_j).$$

Our goal is to find a classifier $\mathcal{C}$ that minimizes the total within-cluster distance given by

$$W(\mathcal{C}) = \sum_{k=1}^{K} W_k(\mathcal{C}). \tag{2.8}$$

The partitioning around mediods (PAM) algorithm (Kaufman and Rousseeuw [5]) can provide a K-mediod clustering. The PAM algorithm contains the following steps:

1. Randomly select K drivers $C^{(0)} = \{c^{(0)}_{K,1}, \ldots, c^{(0)}_{K,K}\}$ as mediods. Allocate each driver $i \in \mathcal{N}$ to its nearest mediod. This defines an initial classifier $\mathcal{C}^{(0)}$:

$$i \mapsto \mathcal{C}^{(0)}(i) = \arg\min_{k \in \mathcal{K}} d(\boldsymbol{x}_i, \boldsymbol{x}_{c^{(0)}_{K,k}}).$$

2. Calculate the total within-cluster distance $W(\mathcal{C}^{(0)})$ according to (2.7) and (2.8).

3. Repeat the following steps for $l \geq 1$:

(a) For each pair of a mediod in $C^{(l-1)}$ and a non-mediod driver in $\mathcal{N} \setminus C^{(l-1)}$, swap the mediod with the non-mediod driver, allocate each drivers to its nearest mediod, and calculate the total within-cluster distance.

(b) Choose the swap which leads to the smallest total within-cluster distance. This defines a proposed set of mediods $C^{(*)}$ and the corresponding classifier $\mathcal{C}^{(*)}$.

(c) If $W(\mathcal{C}^{(*)}) - W(\mathcal{C}^{(l-1)}) < 0$, accept the proposed clustering and let $C^{(l)} = C^{(*)}, \mathcal{C}^{(l)} = \mathcal{C}^{(*)}, W(\mathcal{C}^{(l)}) = W(\mathcal{C}^{(*)})$. If $W(\mathcal{C}^{(*)}) - W(\mathcal{C}^{(l-1)}) \geq 0$, terminate the algorithm and accept $\mathcal{C}^{(l-1)}$ as the final clustering.

Note that the PAM algorithm finds a local minimum in the sense that no single switch of a mediod and a non-mediod will decrease the total within-cluster distance (2.8). Hence, different initial sets of mediods $C^{(0)}$ will lead to different clustering. Reynolds et al. [10] propose a build phase which looks for a good initial set of mediods following certain rules. The R function `pam` implements this build phase by default, and always returns the same clustering for a particular data set. Compared with K-means clustering, K-mediods clustering is more robust against outliers and can be applied to any distance function at the expense of computational time. For our data set containing $n = 973$ car drivers, it takes less than one second to get the result of 2-mediods clustering. For a large portfolio containing millions of car drivers, K-means clustering might be a better choice.

The cluster $\mathcal{C}(i) \in \mathcal{K}$ of each driver i could be used in claims frequency models as a categorical covariate with K levels. However, for two drivers in the same cluster we cannot distinguish them by their clusters. Here, we consider the distance between a driver and each mediod driver as a continuous covariate. The distance covariates can distinguish two drivers in the same cluster, i.e., the distance covariates provide more information than the cluster covariates.

We fix $K = 2$ for the visualization of our results. Denote by $c_{2,1}$ and $c_{2,2}$ ($c_{2,1} < c_{2,2}$) the two mediods obtained from the PAM algorithm with the build phrase for our $n = 973$ car drivers. We have $c_{2,1} = 155$ and $c_{2,2} = 820$, whose v-a heatmaps are shown in Figure 2. Given a speed bucket m, for each driver i the distances from the two mediods $c_{2,1}, c_{2,2}$ are defined as

$$d^m_{i,1|K=2} = ||\boldsymbol{z}_{i,m} - \boldsymbol{z}_{c_{2,1},m}||_2 \qquad \text{and} \qquad d^m_{i,2|K=2} = ||\boldsymbol{z}_{i,m} - \boldsymbol{z}_{c_{2,2},m}||_2, \tag{2.9}$$

where $\boldsymbol{z}_{i,m}$ is the v-a heatmap on speed bucket m for driver i with normalization (2.2). Hence, we have eight distances for each driver i in total, i.e., two distances per speed bucket. The pair plot of the eight Euclidean distances is drawn in Figure 3. It shows that there is a strong linear relationship among the distances from the same mediod, i.e., $(d^m_{i,1|K=2})_{m=1:4}$ (the yellow plots in Figure 3) and $(d^m_{i,2|K=2})_{m=1:4}$ (the red plots in Figure 3), respectively. This implies that a driver tends to have similar v-a heatmaps in different speed buckets. Given a speed bucket m, the n points $(d^m_{i,1|K=2}, d^m_{i,2|K=2})_{i=1:n}$ lie in a rectangle that lies at a 45 degree angle with the mediods in the corner; see the green plots in Figure 3. For later purposes, we denote the overall distances between driver i and the two mediods by

$$d_{i,1|K=2} = d(\boldsymbol{x}_i, \boldsymbol{x}_{c_{2,1}}) = ||\boldsymbol{x}_i - \boldsymbol{x}_{c_{2,1}}||_2 = \sqrt{\sum_{m=1}^{4} \left(d^m_{i,1|K=2}\right)^2},$$

$$d_{i,2|K=2} = d(\boldsymbol{x}_i, \boldsymbol{x}_{c_{2,2}}) = ||\boldsymbol{x}_i - \boldsymbol{x}_{c_{2,2}}||_2 = \sqrt{\sum_{m=1}^{4} \left(d^m_{i,2|K=2}\right)^2}.$$

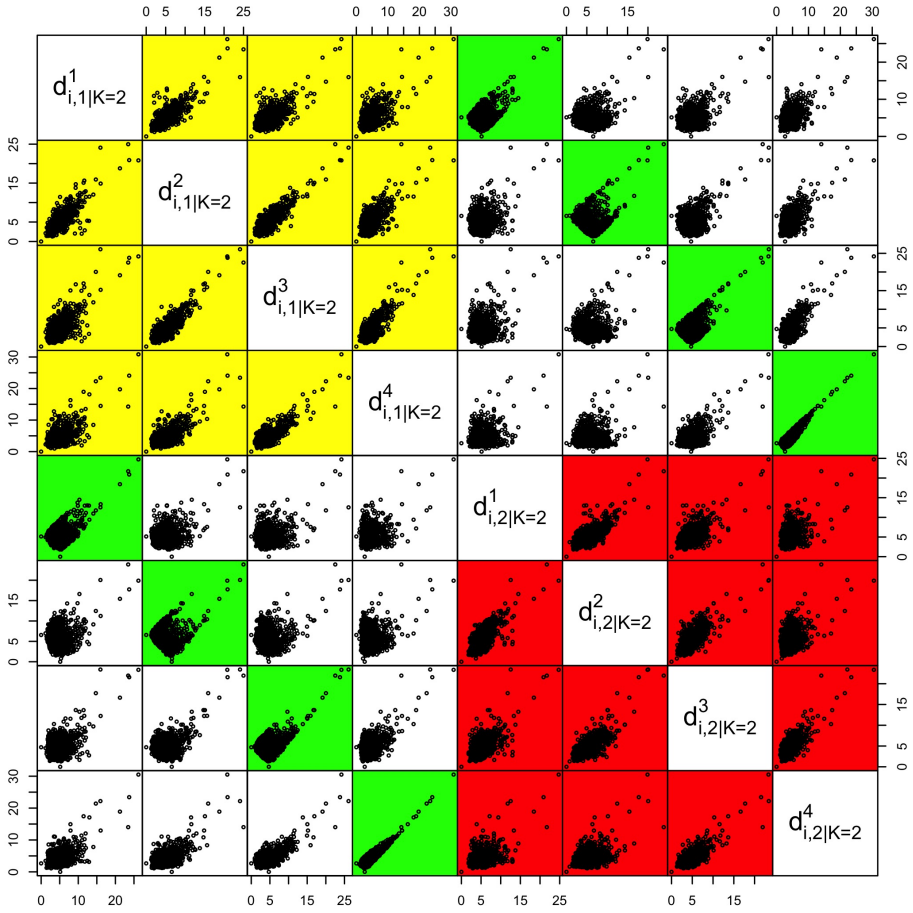

Figure 3: The pair plot of the 8 Euclidean distances in the 2-mediods clustering analysis, where $d^m_{i,k|K=2}$ denotes the distance between driver i and mediod k in speed bucket m.

2.3 Principal components analysis

We apply principal components analysis to directly analyze the design matrix $\boldsymbol{X} \in \mathbb{R}^{n\times J}$. Denote the normalized design matrix by $\boldsymbol{X}^0$ (all column means are set to zero and variances are normalized to one). Denote the J covariates in $\boldsymbol{X}^0$ by $(X^0_j)_{j=1:J}$, i.e., these are the columns of $\boldsymbol{X}^0$. Consider a direction $\boldsymbol{v}_1 = (v_{1,1}, \ldots, v_{J,1})'$ of the J-dimensional covariate space with normalization constraint

$$\sum_{j=1}^{J} v^2_{j,1} = 1. \tag{2.10}$$

The first principal component of the covariates $(X^0_j)_{j=1:J}$ is their projected value onto the direction $\boldsymbol{v}_1$

$$P_1 = v_{1,1}X^0_1 + \cdots + v_{J,1}X^0_J,$$

which has the largest variance. The elements $v_{1,1}, \ldots, v_{J,1}$ are the loadings of the first principal components, and the vector $\boldsymbol{v}_1$ is the first principal component loading vector also called the first right-singular vector of the corresponding matrix $\boldsymbol{V}$ (which we are going to introduce below).

The second principal component P_2 is the projected value of $(X_j^0)_{j=1:J}$ onto the direction $\boldsymbol{v}_2$ perpendicular to $\boldsymbol{v}_1$, which has the second largest variance; and so on.
Given the data $\boldsymbol{X}^0$ the first principal component loading vector can be derived by maximizing the sample variance of P_1:

$$\arg \max_{v_{1,1},\ldots,v_{J,1}} \left[\frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{J} v_{j,1} x_{i,j}^0 \right)^2 \right], \tag{2.11}$$

under constraint (2.10). The above optimization problems can be solved via singular value decomposition as follows:

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

where $\boldsymbol{U}$ is an $n \times J$ orthogonal matrix, $\boldsymbol{V}$ is a $J \times J$ orthogonal matrix and $\boldsymbol{\Lambda} = \text{diag}(g_1, \ldots, g_J)$ is a $J \times J$ diagonal matrix with singular values $g_1 \geq \ldots \geq g_J \geq 0$. The w-th column of the rotation matrix $\boldsymbol{V}$ is the w-th principal component loading vector (or right-singular vector) $\boldsymbol{v}_w = (v_{1,w}, \ldots, v_{J,w})', w = 1, \ldots, J$. The w-th principal component is the projected value of $(X_j^0)_{j=1:J}$ onto the direction $\boldsymbol{v}_w$, that is,

$$P_w = \sum_{j=1}^{J} v_{j,w} X_j^0. \tag{2.12}$$

In Figure 4 we show the first and second loading vectors $\boldsymbol{v}_1, \boldsymbol{v}_2$ in its corresponding sub-rectangle. We note that the loadings for the same acceleration interval are close no matter which speed in-

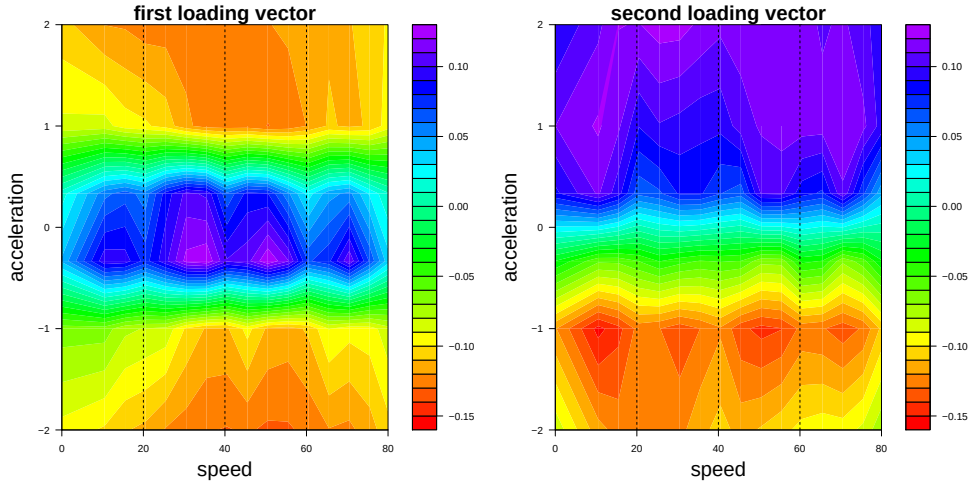

Figure 4: The first and second loading vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$.

terval we consider. The loadings mainly depend on the acceleration interval rather than the speed interval. The first principal component loadings describe abrupt braking in $[-2, -2/3)\text{m/s}^2$ and strong acceleration in $(2/3, 2]\text{m/s}^2$. Thus, the first principal component reflects the relative frequency of smooth acceleration/braking, i.e., the degree of concentration on the zero acceleration rate, see Figure 4 (left). The signs of the second principal component loadings switch at the acceleration rate zero, which indicates that the second principal component illustrates the difference in absolute value between acceleration and braking, see Figure 4 (right).

We define the w-th principal component for speed bucket m of driver i as

$$p_{i,w}^{m} = \sum_{j=1}^{24} v_{(m-1)\times 24+j,w} x_{i,(m-1)\times 24+j}^{0}, \quad w = 1, \ldots, J \quad \text{and} \quad m = 1, \ldots, 4, \tag{2.13}$$

where the element $x_{i,(m-1)\times 24+j}^{0}$ in matrix $\boldsymbol{X}^{0}$ corresponds to the j-th element $z_{i,m,j}$ in vector $\boldsymbol{z}_{i,m}$, see equation (2.3). We construct the pair plot of the first two principal components for all speed buckets in Figure 5. The yellow plots illustrate that there is strong collinearity between

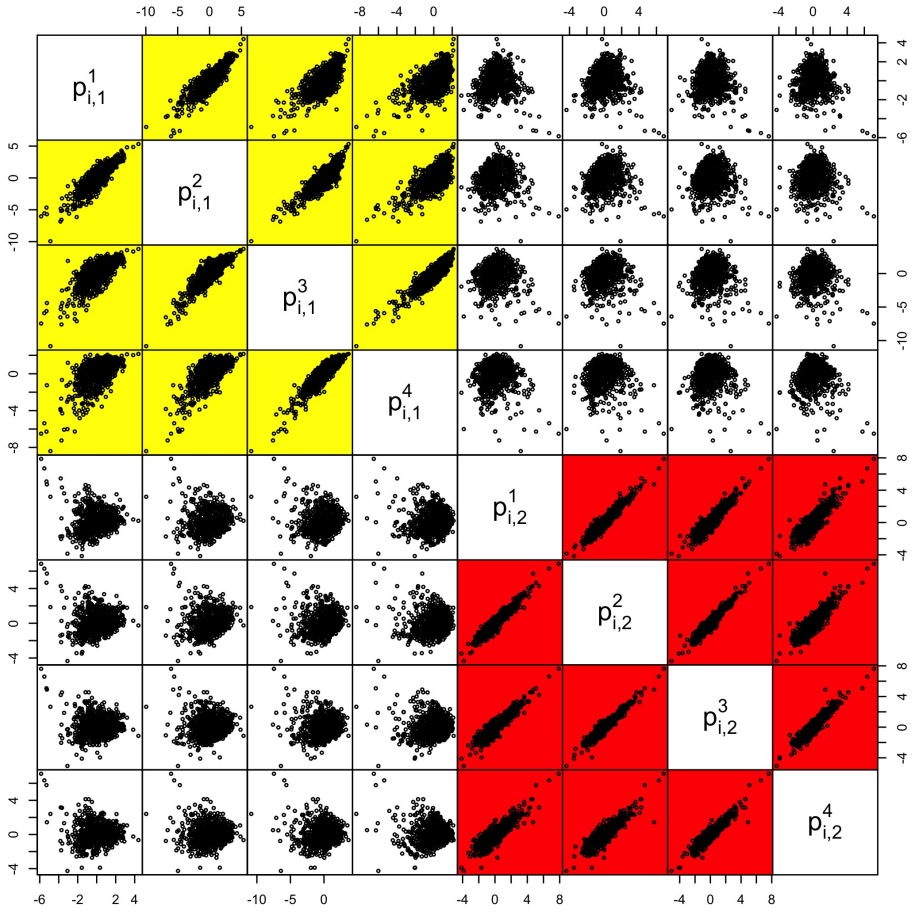

Figure 5: The pair plot of the first two principal components in the four speed buckets, where $p_{i,w}^{m}$ denotes the w-th principal component in speed bucket m of driver i .

the first principal components of all speed buckets, where collinearity slightly decreases for an increasing distance in speed buckets. The red plots show that collinearity seems bigger for the second principal component than for the first principal component. These collinearities indicate that a driver behaves similarly in all speed buckets. Furthermore, the first and second principal components do not seem to have any collinearity. We conclude that the K-mediods clustering and the principal components analysis come to a similar result for $K = 2$. For later purpose,

following (4.3), we denote the w-th overall principal component of driver i as

$$p_{i,w} = \sum_{j=1}^{J} v_{j,w} x_{i,j}^0.$$

3 Acceleration patterns

The symmetry shown in Figure 4 implies that the marginal acceleration patterns account for the main difference among v-a heatmaps and it might be sufficient to only consider the acceleration patterns. For this reason, we compress the two dimensional heatmap $\boldsymbol{z}_{i,m}$ in each speed bucket m of every driver i across the speed to obtain the marginal distribution of acceleration rates $\boldsymbol{b}_{i,m} = (b_{i,m,1}, \ldots, b_{i,m,6})'$ with elements

$$b_{i,m,u} = \sum_{v=1}^{4} z_{i,m,(u-1)\times 4+v} \quad \text{for } u = 1, \ldots, 6, \tag{3.1}$$

see Figure 1. Note that $\boldsymbol{b}_{i,m}$ has normalization

$$\sum_{u=1}^{6} b_{i,m,u} = 1. \tag{3.2}$$

Every car driver i is then characterized by a 24-vector $\boldsymbol{a}_i = (\boldsymbol{b}_{i,1}', \ldots, \boldsymbol{b}_{i,4}')' \in \mathbb{R}^{24}$ which contains four marginal acceleration distributions of dimension 6. Note that we have the following relationship between elements in $\boldsymbol{b}_{i,m}$ and elements in $\boldsymbol{a}_i$:

$$b_{i,m,u} = a_{i,(m-1)\times 6+u}, \tag{3.3}$$

for $i = 1, \ldots, 973,\ m = 1, \ldots, 4,\ u = 1, \ldots, 6$.
We denote by $\boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n)' \in \mathbb{R}^{n\times 24}$ the design matrix of these acceleration patterns. In analogy to Sections 2.2 and 2.3, we apply the K-mediods clustering and the principal components analysis to design matrix $\boldsymbol{A}$. Denote by $(r_{i,k|K})_{k=1:K}$ the overall distances between driver i and mediod k in the K-mediods clustering, by $(q_{i,w})_{w=1:24}$ the overall w-th principal components of driver i in the principal components analysis, and by $r_{i,k|K}^m, q_{i,w}^m$ the corresponding values in speed bucket $m = 1, \ldots, 4$. Denote by $o_{K,k} \in \mathcal{N}$ the driver label of mediod k in K-mediods clustering. We show the relationship between heatmaps covariates d, p and acceleration patterns covariates r, q in Figure 6. The first two plots of Figure 6 show that there is a strong linear relationship between $r_{i,1|K=2}, d_{i,2|K=2}$ and $r_{i,2|K=2}, d_{i,1|K=2}$, respectively. Note that we get two mediod drivers $o_{2,1} = 285$ and $o_{2,2} = 440$, which are close to drivers $c_{2,2}$ and $c_{2,1}$, respectively; see the first two principal components representations of all drivers in Figure 8. We have a similar observation for the overall principal components $p_{i,w}$ and $q_{i,w}$ for $w = 1, 2$ in the last two plots of Figure 6. This leads us to the conclusion that the covariates extracted from the acceleration patterns contain almost the same information as the covariates extracted from the v-a heatmaps, and therefore we can restrict our studies to acceleration patterns on the corresponding 4 speed intervals $(0, 20]$km/h, $(20, 40]$km/h, $(40, 60]$km/h, and $(60, 80]$km/h.
One advantage of using acceleration patterns rather than heatmaps is that we can reduce the minimum driving time required, see Appendix A. On the one hand, this enables us to use less

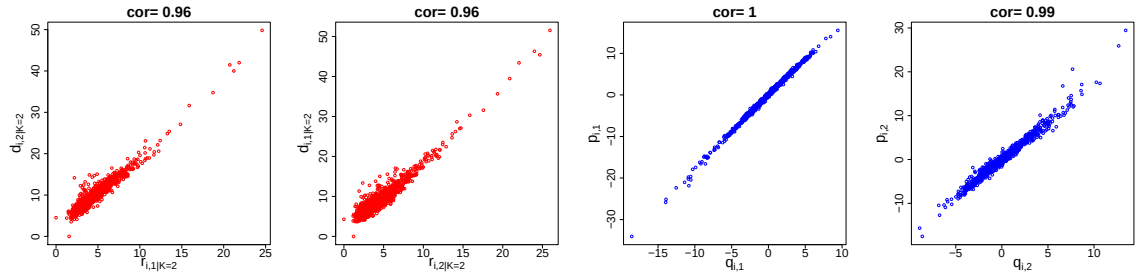

Figure 6: The relationship between the heatmaps and the acceleration patterns in terms of the 2-mediods overall distances and the first two overall principal components.

telematics data of the $n = 973$ car drivers, and it significantly shortens the computational time. On the other hand, given the three months' telematics data we could analyze a larger portfolio containing 1011 car drivers, which meet the minimum driving time requirements for a stable acceleration pattern in the four speed buckets, simultaneously.

4 Claims frequency modeling

Our claims data contains the number of reported claims from 01/01/2014 to 29/06/2017 of compulsory third party liability policies with effective exposures supported in the time interval from 01/01/2014 to 31/05/2017. The evaluation date is chosen as 31/05/2017 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 expiration dates after 31/05/2017 are partially exposed, and the exposures for these policies are adjusted pro-rata temporis. For the $n = 973$ cars, the average exposure per car driver is 2.24 years (also called years-at-risk), measured by the effective policy duration (in years) to the evaluation date 31/05/2017. The average claims frequency is 0.24 per year per car driver, which is consistent with the market benchmark in China.

In this section, we establish Poisson regression models for the number of reported claims of the $n = 973$ cars under both a generalized additive model framework and a generalized linear model framework. We mainly investigate three aspects: (a) comparing the predictive performance of the K-mediods covariates r with the principal components covariates q; (b) comparing the predictive performance of different speed buckets using $q^m, m = 1, \ldots, 4$; (c) comparing the predictive performance of v-a heatmaps covariates d, p with acceleration patterns covariates r, q. As discussed in Section 2, we assume that a driver's driving style does not change during the whole policy period.

4.1 The benchmark model

We begin with a full model containing the classical actuarial covariates including driver's age, gender, car's age, region and average driving time per month. We estimate average driving time per month using the three months' telematics data from 01/05/2016 to 31/07/2016. We show the distribution of the years-at-risk for the continuous covariates driver's age, car's age and average driving hours per month in Figure 7, and for the categorical covariates gender and

regions in Table 1. Figure 7 shows that we do not have many exposures at small ages, and

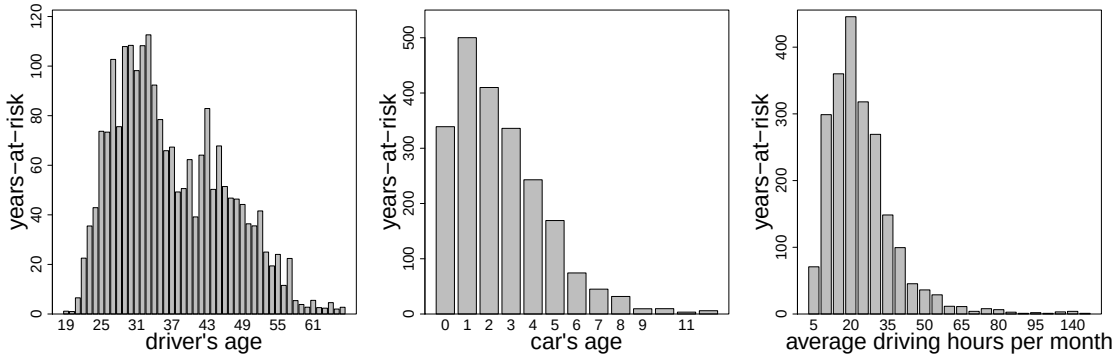

Figure 7: Distribution of years-at-risk for the continuous covariates driver's age, car's age and average driving hours per month.

the distribution for average driving time per month is positively skewed. Table 1 shows that most car drivers considered are in Zhejiang and Hebei provinces and there are many more male drivers than female drivers in all regions.

Table 1: Distribution of years-at-risk across the categorical covariates gender and regions.

		Region				Total
		Zhejiang	Hebei	Shanghai	Others	
Gender	Male	776	642	186	31	1635
	Female	317	170	28	27	542
Total		1093	812	214	58	2177

We specify the full model as follows:

$$\begin{aligned} Y_i &\overset{\text{ind.}}{\sim} \text{Poisson}(\lambda_i e_i) \qquad \text{with} \\ \log \lambda_i =& \beta_0 + s(\text{driver\_age}_i) + \alpha_{\text{gender}_i} + s(\text{car\_age}_i) + \gamma_{\text{region}_i} + s(\text{duration}_i), \end{aligned} \tag{4.1}$$

where $e_i > 0$ is the total exposure in years (years-at-risk) of driver i measured by the effective policy duration to the evaluation date 31/05/2017, $\lambda_i > 0$ is the expected claims frequency of driver i, described by an intercept β_0, and s are thin plate splines addressing potential non-linear effects of covariates on $\log \lambda_i$. All the parameters can be estimated in a generalized additive model framework (Wood [14]). We apply the smooth component selection technique from Marra and Wood [6]: (1) remove smooth terms with effective degrees-of-freedom of less than 0.1; (2) treat smooth terms with effective degrees-of-freedom in $[0.1, 1.3]$ as linear terms; (3) discretize smooth terms with effective degrees-of-freedom of larger than 1.3 into categorical variables. This smooth component selection always leads to a generalized linear model. It turns out that the effective degrees-of-freedom of driver's age, car's age and duration are 0.005, 0.587 and 0.901, respectively.

This leads to our benchmark model as follows:

$$\begin{aligned} Y_i &\overset{\text{ind.}}{\sim} \text{Poisson}(\lambda_i e_i) \qquad \text{with} \\ \log \lambda_i =& \beta_0 + \alpha_{\text{gender}_i} + \beta_1 \text{car\_age}_i + \gamma_{\text{region}_i} + \beta_2 \text{duration}_i, \end{aligned} \tag{4.2}$$

with intercept β_0 and regression parameters $\alpha_{\text{gender}}, \beta_1, \gamma_{\text{region}}$ and β_2 to be estimated in a generalized linear framework. The estimated coefficients are shown in Table 2. Note that

Table 2: Estimated coefficients in the benchmark model (4.2).

parameters	estimate	standard error	z value	$\Pr(>\|z\|)$
$\hat{\beta}_0$	-1.67	1.63×10^{-1}	-10.26	0.00
$\hat{\alpha}_{\text{female}}$	1.69×10^{-1}	9.90×10^{-2}	1.71	0.09
$\hat{\beta}_1$	3.31×10^{-2}	2.04×10^{-2}	1.62	0.10
$\hat{\gamma}_{\text{Others}}$	-9.06×10^{-2}	2.92×10^{-1}	-0.31	0.76
$\hat{\gamma}_{\text{Hebei}}$	-4.26×10^{-1}	1.58×10^{-1}	-2.70	0.01
$\hat{\gamma}_{\text{Zhejiang}}$	3.79×10^{-2}	1.46×10^{-1}	0.26	0.80
$\hat{\beta}_2$	1.38×10^{-4}	4.17×10^{-5}	3.30	0.00

male drivers in Shanghai are treated as reference class. Though car's age passes our smooth component selection, it is not significant in a z-test on 5% level given all other covariates in the model. Female drivers tend to have a higher claims frequency than male drivers. While drivers in Hebei province have a lower claims frequency than those in Shanghai, drivers in Zhejiang province and other regions do not have a significant different claims frequency from those in Shanghai.

We implement a sequential deviance test as shown in Table 3. It indicates that driving region is

Table 3: Analysis of deviance table of the benchmark model (4.2).

	Df	Explained deviance	Residual Df.	Residual deviance	Pr(>Chi)
Null			972	1061.6	
Gender	1	4.25	971	1057.4	0.04
Car's age	1	5.00	970	1052.4	0.03
Region	3	19.23	967	1033.1	0.00
Duration	1	9.45	966	1023.7	0.00

the most important risk factor among all the traditional risk factors for our portfolio. Indeed, the traffic conditions are quite different among the four regions considered. We observe contradicting test results for car's age from Table 2 and Table 3, because the z-test is a marginal test assuming all other covariates are already in the model while the sequential deviance test indicates that car's age is significant given only the gender in the model. Note that both the z-test and the sequential deviance test are in-sample tests which might be distorted by over-fitting. We will conduct an out-of-sample test via cross-validation in Section 4.5. Finally, we remark that driver's age is usually an important risk factor in car insurance. However, since our portfolio is small and we do not have enough exposures for small ages, the model cannot detect the effect of driver's age at young ages on the claims frequency very well, and therefore this covariate drops out of the model.

4.2 Claims frequency modeling with K-mediods distances r

We first investigate the relationship between the K-mediods distances r from the acceleration patterns for $K = 2, 3, 4$. We draw the mediod drivers $(o_{K,k})_{k=1:K}$ for $K = 2, 3, 4$ in Figure 8, where each driver i is represented by its first two principal components $q_{i,1}, q_{i,2}$ and different colors and sizes are used to distinguish different K's. We note that the mediods $o_{2,1}, o_{3,2}, o_{4,4}$ are

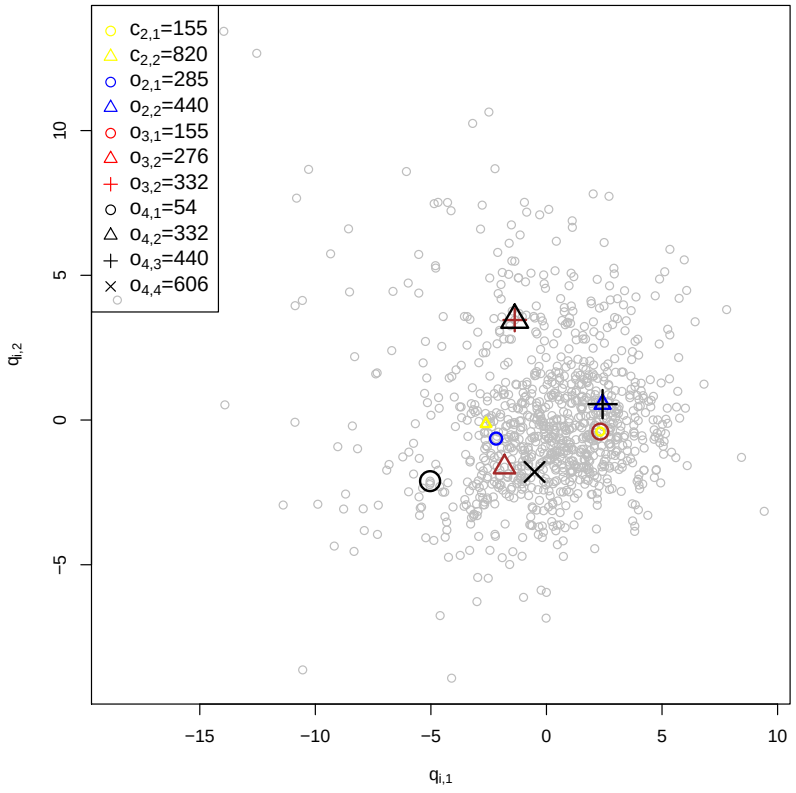

Figure 8: The first two principal components of the acceleration patterns for all $n = 973$ drivers. Yellow indicates the two mediods from v-a heatmaps. Blue indicates the two mediods ($K = 2$) from acceleration patterns. Red indicates the three mediods ($K = 3$) from acceleration patterns. Black indicates the four mediods ($K = 4$) from acceleration patterns. Note that the mediods are naturally ordered according to their driver labels.

close; $o_{2,2}$ is the same driver as $o_{4,3}$ and it is close to $o_{3,1}$; $o_{3,3}$ is the same driver as $o_{4,2}$; $o_{4,1}$ is far away from the other mediods. We conclude that 3-mediods clustering and 4-mediods clustering recognize the 2-mediods clustering structure, and they also discover a new cluster around driver 332. Furthermore, 4-mediods clustering discover another new cluster around driver 54. However, not all these clusters are related to claims frequencies, which can be seen from the following claims frequencies modeling. We draw the pair plot of $(r_{i,k|K})_{k=1:K}$ for $K = 2, 3, 4$ in Figure 9. The linearity observed in the yellow plots is consistent with our observation in Figure 8. The colored dots lie in a rectangle that lies at a 45 degree angle with the mediods in the corner; also see the green plots in Figure 3.

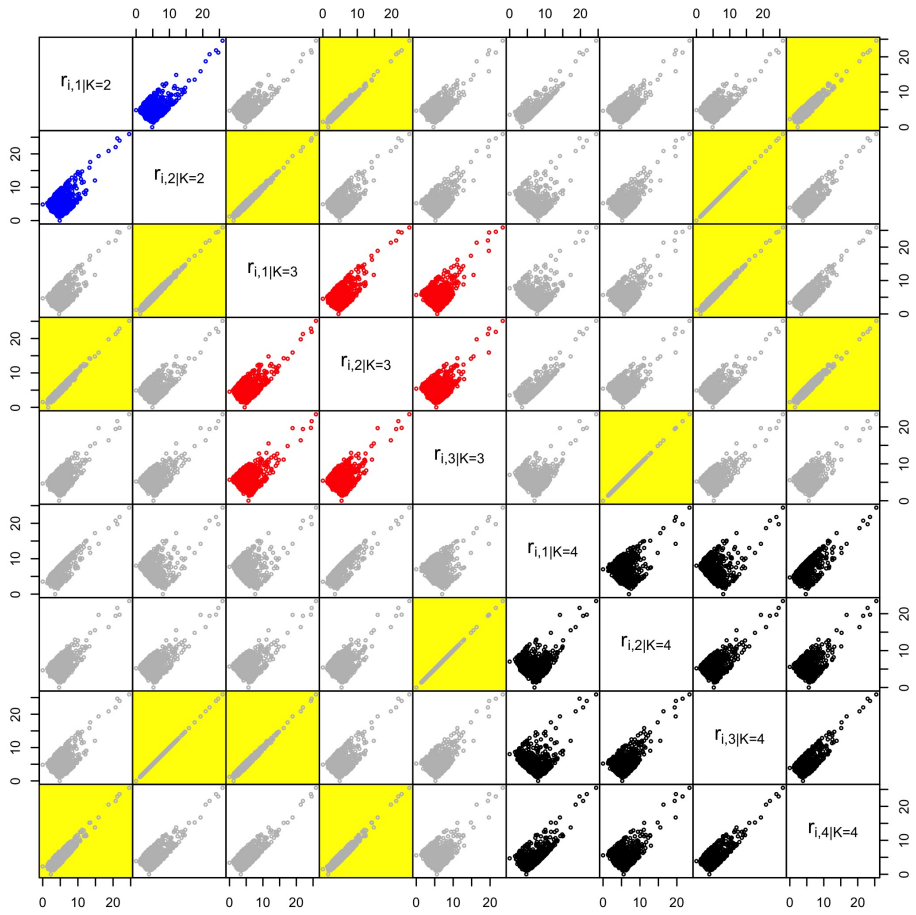

Figure 9: The pair plot of the overall distances for $K = 2, 3, 4$.

The regression function using K-mediods covariates $(r_{i,k|K})_{k=1:K}$ is specified as:

$$\log \lambda_i = \beta_0 + \alpha_{\text{gender}_i} + \beta_1 \text{car\_age}_i + \gamma_{\text{region}_i} + \beta_2 \text{duration}_i + s(r_{i,1|K}) + \ldots + s(r_{i,K|K}),$$

where $r_{i,k|K}$ describes the Euclidean distance of driver i to mediod k in K-mediods clustering. We follow the same smooth component selection approach as in the previous section. For $K = 2$, it turns out that $r_{i,1|K=2}$ and $r_{i,2|K=2}$ should be in linear form in the model. For $K = 3$, it turns out that $r_{i,1|K=3}$ and $r_{i,2|K=3}$ should be in linear form in the model and $r_{i,3|K=3}$ can be removed. For $K = 4$, it turns out that $r_{i,3|K=4}$ and $r_{i,4|K=4}$ should be in linear form in the model and $r_{i,1|K=4}, r_{i,2|K=4}$ can be removed. We conclude that $K = 2$ is sufficient since the additional mediods found by $K = 3$ and $K = 4$ drop out of the regression model. In the following cross-validation, we only focus on $K = 2$ since $K = 3, 4$ lead to a similar result. When $K = 2$, the regression function is as follows:

$$\log \lambda_i = \beta_0 + \alpha_{\text{gender}_i} + \beta_1 \text{car\_age}_i + \gamma_{\text{region}_i} + \beta_2 \text{duration}_i + \beta_3 r_{i,1|K=2} + \beta_4 r_{i,2|K=2}, \qquad (4.3)$$

with two additional regression parameters β_3 and β_4 compared to (4.2). Following the analysis of deviance in Table 3, given all the classical covariates already in the model the 2-mediods distances covariates can explain another 24.15 of the deviance, which is even larger than 19.23

of the deviance explained by the most important classical covariate region; see also the last column of Table 4.

4.3 Claims frequency modelling with principal components q

Claims frequency modelling with the principal components is more straightforward than with K-mediods distances covariates, since we do not have a tuning parameter K in the principal components analysis. We again focus on the acceleration patterns. The regression function using principal components $(q_{i,w})_{w=1:24}$ is specified as:

$$\log \lambda_i = \beta_0 + \alpha_{\text{gender}_i} + \beta_1 \text{car\_age}_i + \gamma_{\text{region}_i} + \beta_2 \text{duration}_i + s(q_{i,1}) + \ldots + s(q_{i,24}).$$

It turns out that only the first principal component is significant with effective degree-of-freedom of 1.01. Hence, we get the following regression function

$$\log \lambda_i = \beta_0 + \alpha_{\text{gender}_i} + \beta_1 \text{car\_age}_i + \gamma_{\text{region}_i} + \beta_2 \text{duration}_i + \beta_3 q_{i,1}, \tag{4.4}$$

with one additional regression parameter β_3 compared to (4.2). The sign of $\hat{\beta}_3$ is negative and we can interpret $q_{i,1}$ as the safe driving index of driver i. This is consistent with our explanation of the first loading vector in Figure 4. Following the analysis of deviance in Table 3, given all the classical covariates already in the model the first principal components can explain another 25.99 of the deviance, and it is better than the 2-mediods distances covariates in (4.3); also see the last column of Table 4.

One of the purposes in this section is to compare the predictive power of different speed buckets. Hence, we establish 4 more regression functions using the first principal component $q_{i,1}^m$ in speed bucket m, as follows:

$$\log \lambda_i = \beta_0 + \alpha_{\text{gender}_i} + \beta_1 \text{car\_age}_i + \gamma_{\text{region}_i} + \beta_2 \text{duration}_i + \beta_3 q_{i,1}^m, \tag{4.5}$$

with $m = 1, \ldots, 4$. Following the analysis of deviance in Table 3, it turns out that $q_{i,1}^m$ can explain another 28.30, 24.61, 23.77, 14.84 of the deviance for $m = 1, \ldots, 4$, respectively, given all the classical covariates already in the model; also see the last column of Table 4. Finally, for comparison we set up a regression function only using the telematics covariates

$$\log \lambda_i = \beta_0 + \beta_2 \text{duration}_i + \beta_3 q_{i,1}^1, \tag{4.6}$$

and keeping only the average driving time per month from the classical covariates. We will compare this model with others in Section 4.5.

4.4 Claims frequency modelling with v-a heatmaps covariates d and p

Figure 6 has indicated that v-a heatmaps covariates d, p have a strong linear relationship with acceleration patterns covariates, r, q. However, it is still desirable to study the predictive power of the variations in v-a heatmaps covariates d, p not explained by acceleration patterns covariates r, q as shown in Figure 6. Following the analysis in Sections 4.2 and 4.3, we establish two regression functions:

$$\log \lambda_i = \beta_0 + \alpha_{\text{gender}_i} + \beta_1 \text{car\_age}_i + \gamma_{\text{region}_i} + \beta_2 \text{duration}_i + \beta_3 d_{i,1|K=2} + \beta_4 d_{i,2|K=2}, \tag{4.7}$$

$$\log \lambda_i = \beta_0 + \alpha_{\text{gender}_i} + \beta_1 \text{car\_age}_i + \gamma_{\text{region}_i} + \beta_2 \text{duration}_i + \beta_3 p_{i,1}, \tag{4.8}$$

where $d_{i,k|K=2}$ is the distance between driver i and mediod k in 2-mediods clustering, and $p_{i,1}$ is the first principal component of v-a heatmaps. Following the analysis of deviance in Table 3, given all the classical covariates already in the model the 2-mediods distances and the first principal components can explain another 20.46, 26.17 of the deviance, respectively; also see the last column of Table 4.

4.5 Model comparisons

We compare the predictive performance of different models based on their average Poisson deviance statistics, defined as

$$D(\mathcal{T}, \hat{\theta}_{\mathcal{S}}) = \frac{2}{I} \sum_{i \in \mathcal{T}} Y_i \left[\frac{\hat{\lambda}_i e_i}{Y_i} - 1 - \log \left(\frac{\hat{\lambda}_i e_i}{Y_i} \right) \right], \tag{4.9}$$

where $\mathcal{T}, \mathcal{S} \subset \mathcal{N}$ are two subsets of all car drivers $\mathcal{N}$ containing $I, L < n$ drivers, respectively, $\hat{\theta}_{\mathcal{S}}$ is the set on which we estimate the parameters using the drivers in $\mathcal{S}$, and we denote by $\hat{\lambda}_i = \lambda_i(\hat{\theta}_{\mathcal{S}})$ the estimated claims frequency of driver i using estimate $\hat{\theta}_{\mathcal{S}}$. Note that the i-th term on the right-hand side is set equal to $2\hat{\lambda}_i e_i$ if $Y_i = 0$. 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 (2016) [15].

If we let $\mathcal{T} = \mathcal{S}$, equation (4.9) would be the average residual deviance on the drivers in $\mathcal{S}$ which is subject to over-fitting. If we let $\mathcal{T} \cap \mathcal{S} = \emptyset$, equation (4.9) gives us an out-of-sample average Poisson deviance statistics which is sensitive to over-fitting. Typically, 10-fold cross-validation is applied for estimating the out-of-sample average Poisson deviance statistics. Therefore, we randomly partition $\mathcal{N}$ into 10 roughly equally-sized disjoint parts, denoted by $\mathcal{T}_1, \ldots, \mathcal{T}_{10}$. The 10-fold cross-validation estimate of the average Poisson deviance statistics can be obtained as

$$\widehat{D} = \frac{1}{10} \sum_{l=1}^{10} D(\mathcal{T}_l, \hat{\theta}_{-\mathcal{T}_l}), \tag{4.10}$$

where $\hat{\theta}_{-\mathcal{T}_l}$ is the estimated parameter using all drivers except $\mathcal{T}_l$. We randomly partition all drivers $\mathcal{N}$ for 50 times, and calculate the cross-validation estimate of the average Poisson deviance statistics $\hat{D}_s$ for each partition $s = 1, \ldots, 50$. We use the same partitioning for the cross-validations in models (4.2)-(4.8), and calculate the sample mean, the sample standard deviance, and the first and third quantiles of $(\hat{D}_s)_{s=1:50}$ for each model. The results are listed in Table 4.

Comparing model (4.2) with (4.3), (4.4), (4.7) and (4.8), we conclude that the telematics covariates can significantly improve out-of-sample predictive power, i.e., we get a statistical decrease in mean from 1.0727 to 1.0517, 1.0478, 1.0561 and 1.0476, respectively. We also conclude that the telematics covariates reveal risk factors not explained by traditional risk factors. Comparing models (4.5) for different m, we conclude that the telematics covariates from the low speed buckets have a better out-of-sample predictive power, which, of course, is what is expected. Comparing model (4.6) with (4.2), we conclude that the telematics covariates have a better out-of-sample predictive power than the classical covariates. Thus, our telematics covariates are more important than the classical covariates for claims frequency prediction. Comparing

Table 4: 10-fold cross-validation estimates of the average Poisson deviance statistics for models (4.2)-(4.8). The last column is the additional explained deviance compared to the benchmark model (4.2).

models	Mean	Std.	1st quantile	3rd quantile	add. explained dev.
(4.2)	1.0727	0.0035	1.0700	1.0751	-
(4.3)	1.0517	0.0039	1.0491	1.0549	24.15
(4.4)	1.0478	0.0036	1.0457	1.0506	25.99
(4.5) with $m = 1$	1.0454	0.0036	1.0432	1.0485	28.30
(4.5) with $m = 2$	1.0494	0.0038	1.0469	1.0522	24.61
(4.5) with $m = 3$	1.0502	0.0037	1.0479	1.0527	23.77
(4.5) with $m = 4$	1.0594	0.0036	1.0566	1.0617	14.84
(4.6) with $m = 1$	1.0502	0.0019	1.0491	1.0517	-
(4.7)	1.0561	0.0039	1.0537	1.0592	20.46
(4.8)	1.0476	0.0036	1.0454	1.0504	26.17

models (4.7) and (4.8) with models (4.3) and (4.4), we conclude that there is no important risk information lost by only considering marginal acceleration patterns from v-a heatmaps. This is consistent with our conclusion that a driver tends to behave similarly at different speeds. Comparing model (4.3) with (4.4), we conclude that the principal components covariates have a slight better out-of-sample predictive power. Moreover, we fit an ordinary linear regression with $q_{i,1}$ as response and $r_{i,1}, r_{i,2}$ as predictors. The estimated regression functions is

$$\hat{q}_{i,1} = 1.77 + 0.84 r_{i,1} - 1.23 r_{i,2},$$

with the adjusted R-squared of 0.86. This indicates that there is collinearity between $q_{i,1}$ and $r_{i,1}, r_{i,2}$. Hence, we do not need all these covariates in the model simultaneously, i.e., they contain quite similar risk information.

Remark: In our data $\mathcal{N}$, the average years-at-risk is $\bar{e} = 2.24$ and the empirical claims frequency is $\bar{\lambda} = 0.24$. By Monte Carlo simulation we estimate the average Poisson deviance statistics as

$$\mathbb{E}\left(2Y\left[\frac{\bar{\lambda}\bar{e}}{Y} - 1 - \log\left(\frac{\bar{\lambda}\bar{e}}{Y}\right)\right]\right) \approx 1.0205,$$

with a standard error of 0.0012. We note that the average Poisson deviance statistics in Table 4 are in a reasonable range, i.e., Table 4 indicates that we do not miss any important covariate.

5 Conclusions

We have investigated the v-a heatmaps and acceleration patterns in the four speed buckets. According to the principal components analysis, the distinguishing features of v-a heatmaps are mainly due to the acceleration patterns. A driver tends to show a similar acceleration pattern in both low speed buckets and high speed buckets. The claims frequency modeling using a relatively small portfolio implies the following points:

1. The telematics covariates reveal risk factors which are not contained in traditional risk factors.

2. The principal components covariates have a slight better out-of-sample predictive power than the distances covariates from the K-mediods clustering.

3. The low speed bucket covariates have a better out-of-sample predictive power than the high speed bucket covariates.

4. There is no important risk information lost by only considering marginal acceleration patterns from v-a heatmaps.

If the computational time and the limited telematics data volume are the main restrictions, one might use acceleration patterns in low speed buckets only for claims frequency modeling.

Acknowledgment

Guangyuan Gao gratefully acknowledges financial support from the Forschungsinstitut für Mathematik (FIM) during his research stay at ETH Zurich.

References

[1] Ayuso, M., Guillen, M., 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* **4/2**, article 10.

[2] Fisher, R.A. (1925). *Statistical Methods for Research Workers.* Edinburgh: Oliver & Boyd.

[3] Gao, G., Meng, S., Wüthrich, M.V. (2018). Claims frequency modeling using telematics car driving data. *Scandinavian Actuarial Journal*, DOI: 10.1080/03461238.2018.1523068.

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

[5] Kaufman, L., Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis.* New York: Wiley.

[6] Marra, G., Wood, S.N. (2011). Practical variable selection for generalized additive models. *Computational Statistics and Data Analysis* **55**, 2372-2387.

[7] McCullagh, P., Nelder, J. (1989). *Generalized Linear Models*, second edition. New York: Chapman & Hall.

[8] Paefgen, J., Staake, T., Fleisch, E. (2014). Multivariate exposure modeling of accident risk: insights from pay-as-you-drive insurance data. *Transportation Research A: Policy and Practice* **61**, 27-40.

[9] Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* **2**, 559-572.

[10] Reynolds, A., Richards, G., de la Iglesia, B., Rayward-Smith, V. (1992). Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms* **5**, 475-504.

[11] Verbelen, R., Antonio, K., Claeskens, G. (2018). Unraveling the predictive power of telematics data in car insurance pricing. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **67**, 1275-1304.

[12] Weidner, W., Transchel, F.W.G., Weidner, R. (2016). Classification of scale-sensitive telematic observables for riskindividual pricing. *European Actuarial Journal* **6/1**, 3-24.

[13] Weidner, W., Transchel, F.W.G., Weidner, R. (2016). Telematic driving profile classification in car insurance pricing. *Annals of Actuarial Science* **11/2**, 213-236.

[14] Wood, S.N. (2017). *Generalized Additive Models: An Introduction with R*, second edition. New York: Chapman & Hall.

[15] Wüthrich, M.V., Buser, C. (2016). Data analytics for non-life insurance pricing. *SSRN Manuscript* ID 2870308. Version June 13, 2018.

A Stability of v-a heatmaps and acceleration patterns

In the 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. We follow Appendix A in Gao et al. [3] to estimate the minimum driving time in each speed bucket needed to make the finite sample error less than 0.05. For the definition of finite sample error and the choice of the cut-off value 0.05, we refer to equation (A.9) and Figure 7 in Gao et al. [3].
Table 5 shows the 90%, 95% and 99% quantiles of the minimal amount in days and in minutes for a stable v-a heatmap in a specific speed bucket. If we require the driving time in the four speed buckets being at least 160, 190, 290 and 360 minutes, respectively and simultaneously, we receive $n = 973$ cars meeting this requirement. The same calculation is applied for the acceleration patterns. Table 5 shows the 90%, 95% and 99% quantiles of the minimal amount in days and in minutes for a stable acceleration pattern in a specific speed bucket. If we require the driving time in the four speed buckets being at least 110, 140, 210, and 230 minutes, respectively and simultaneously, we receive 1011 cars meeting this requirement. Thus, in the latter approach we typically can consider more drivers.

Table 5: The 90%, 95% and 99% quantiles of the minimal amount in the four speed buckets.

Speed bucket	v-a heatmaps						Acceleration patterns					
	In days			In minutes			In days			In minutes		
	90%	95%	99%	90%	95%	99%	90%	95%	99%	90%	95%	99%
(0, 20] km/h	8	11	19	**163**	220	460	6	8	15	**111**	161	371
(20, 40] km/h	10	13	21	**194**	252	455	7	10	18	**142**	195	378
(40, 60] km/h	22	29	44	**292**	372	615	15	20	35	**214**	278	506
(60, 80] km/h	40	51	86	**366**	461	702	30	41	68	**234**	325	531