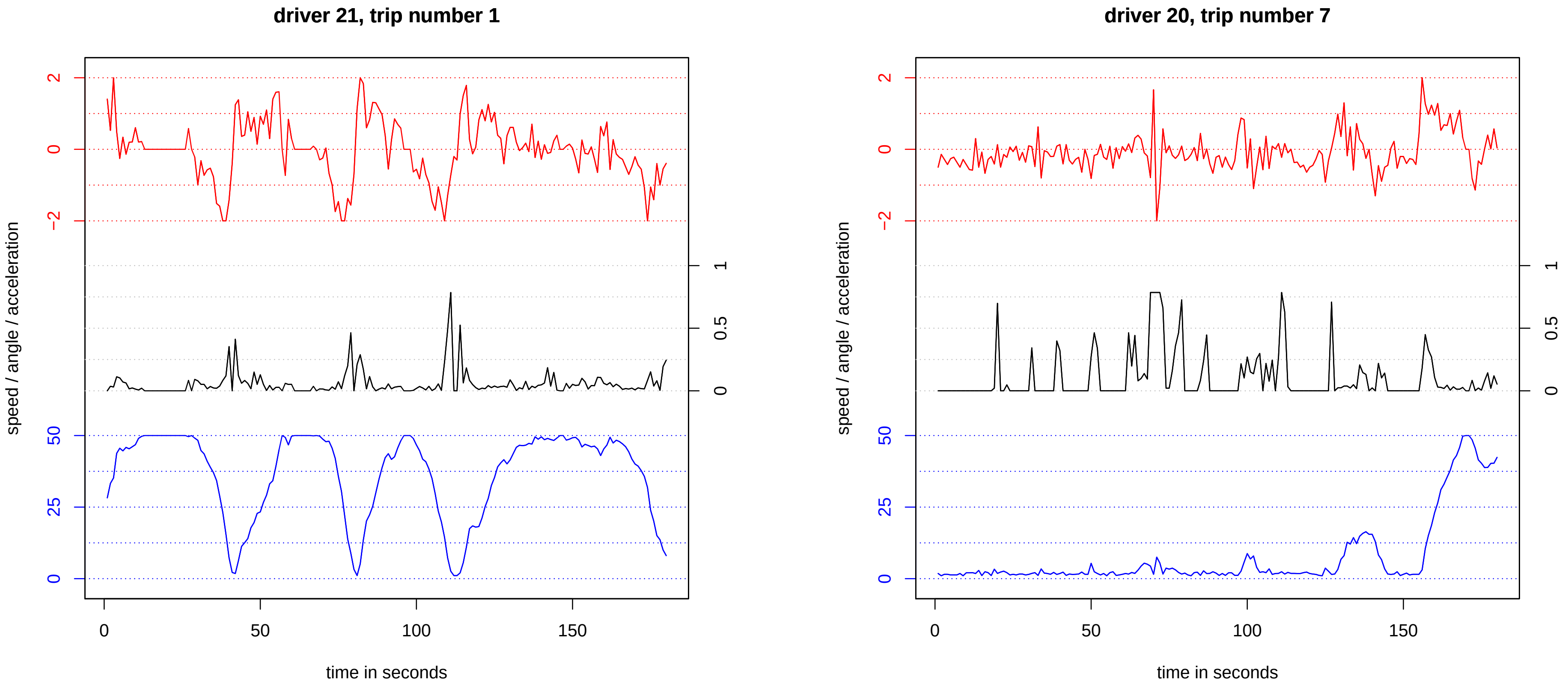
Claims Frequency Modeling using Telematics Car Driving Data

Mario V. Wüthrich
RiskLab, ETH Zurich

joint work with
Guangyuan Gao & Shengwang Meng
Renmin University, Beijing

Insurance Data Science Conference
London, July 16, 2018

Telematics Car Driving Data



acceleration / change in direction / speed

Available Car Driving Data

▷ Find structure (driving styles) in features

$$\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subset \mathcal{X},$$

of n insurance policies in a given feature space $\mathcal{X}$.

- **Data**: 12'076 drivers with
  - ⋆ classical features like age, gender, type of car, prize of car, etc.,
  - ⋆ telematics data of all trips including GPS location (sec by sec), time stamp, speed, acceleration (in all directions), engine revolutions per minute,
  - ⋆ claims data,

  from 2014-2017 (1GB per day, 1.5TB in total).

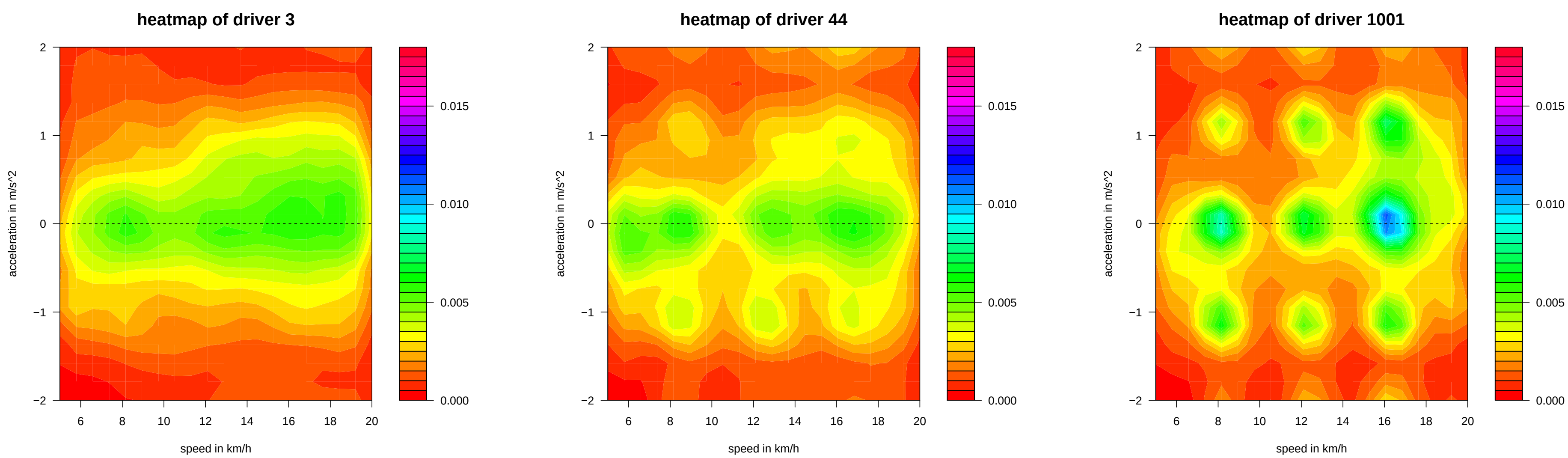
Two Different Approaches for Driving Styles

- Score individual trips.

- Build summary statistics per driver (law of large numbers) and score those.

Normalized v-a Heatmaps

- Calculate v-a heatmap of all trips in speed bucket $[5, 20)$km/h for all n drivers.
- These heatmaps measure the amount of time spent in a (v, a) location.
- Normalization gives (discrete) probability distributions $\boldsymbol{x}_i$ for drivers $i = 1, \ldots, n$.



v-a heatmaps of drivers $i = 3, 44, 1001$ in speed bucket $[5, 20)$km/h.

Autoencoders for Data Compression

- **Encoder:**
$$\varphi : \mathcal{X} \to \mathcal{Z},$$
where $\mathcal{Z}$ is low-dimensional.

- **Decoder:**
$$\psi : \mathcal{Z} \to \mathcal{X}.$$

- **Goal:** Choose functions φ and ψ such that

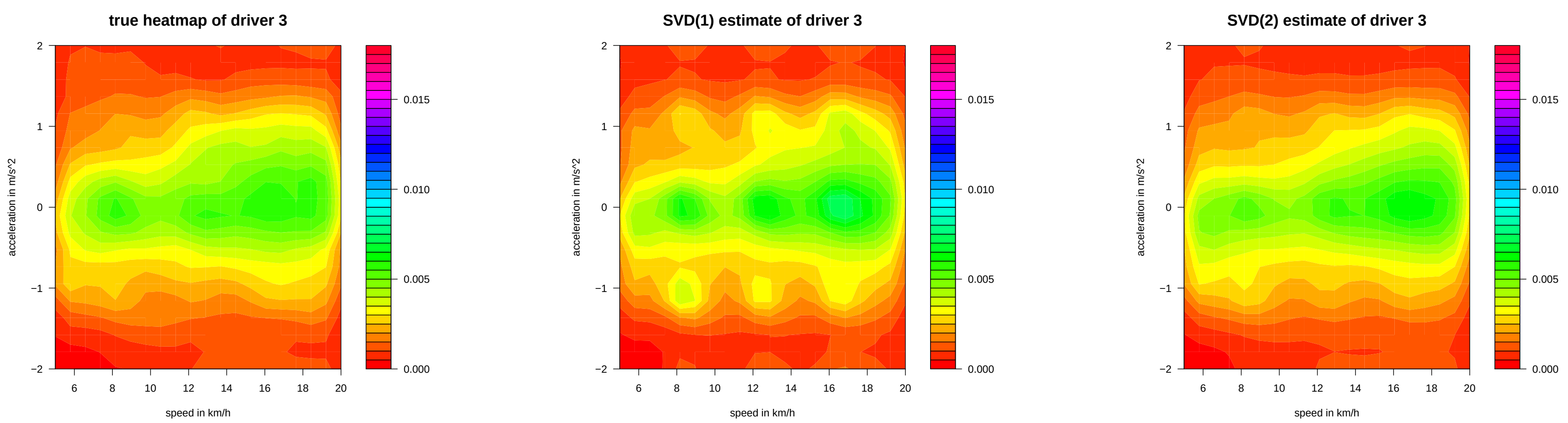
output $\pi(\boldsymbol{x}) = \psi \circ \varphi(\boldsymbol{x})$ is close to input $\boldsymbol{x}$.

$\triangleright$ $\varphi(\boldsymbol{x}) \in \mathcal{Z}$ is used as low-dimensional representation for $\boldsymbol{x} \in \mathcal{X}$.

Principal Component Analysis (PCA)

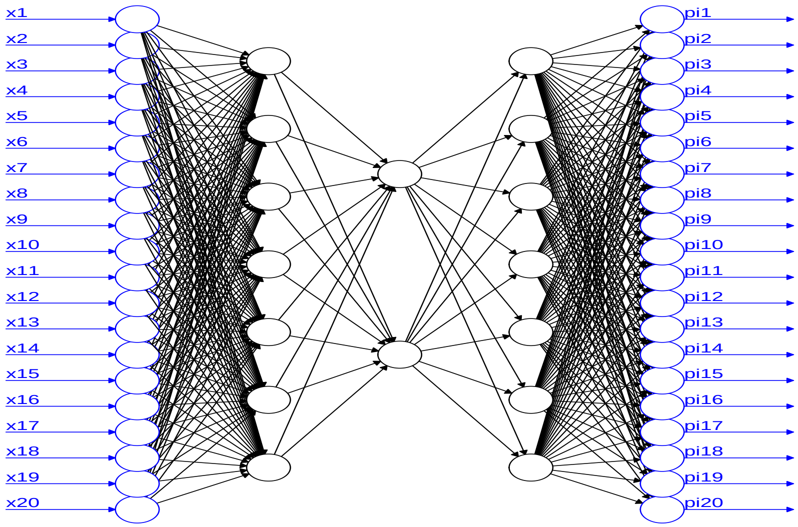
- Consider the design matrix $\boldsymbol{X} = (\boldsymbol{x}_1', \dots, \boldsymbol{x}_n')' \in \mathbb{R}^{n \times d}$ of rank $d \le n$.

- Singular value decomposition (SVD) provides (an) optimal approximation $\boldsymbol{X}_q$ of design matrix $\boldsymbol{X}$ of (smaller) rank $q \le d$ (in Frobenius norm).



SVD result of driver $i = 3$ for ranks $q = 1, 2$ (true heatmap on the left).

Bottleneck Neural Network Autoencoder

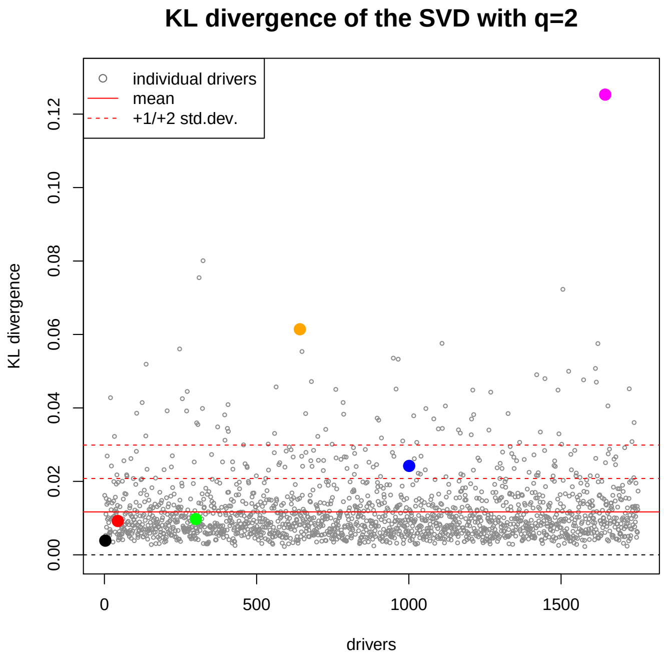


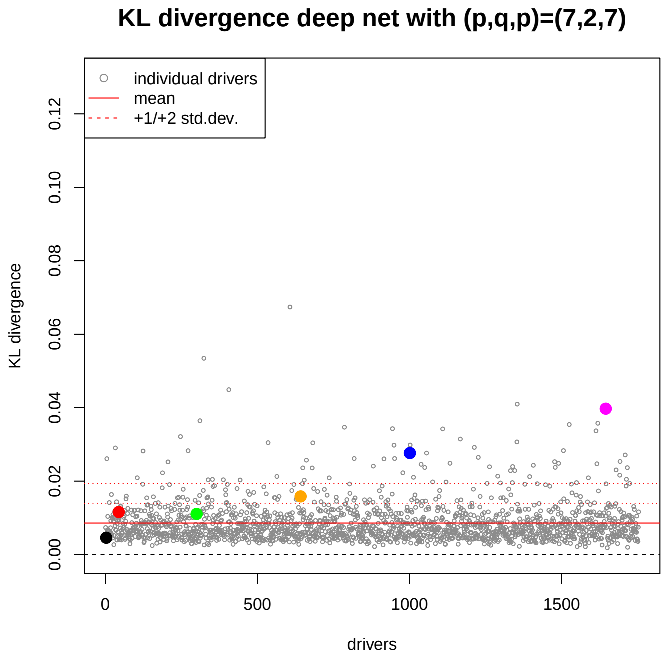
- Calibrate bottleneck neural network such that inputs $\boldsymbol{x}_i$ and outputs $\pi_i = \pi(\boldsymbol{x}_i)$ are close in Kullback-Leibler (KL) divergence

$$\mathcal{L}_{\mathrm{KL}}\left((\boldsymbol{x}_i)_i, (\pi_i)_i\right) \;=\; \frac{1}{n}\sum_{i=1}^{n} d_{\mathrm{KL}}(\boldsymbol{x}_i \| \pi_i).$$

- Signals at the bottleneck are the $\mathcal{Z}$-representations of drivers $i = 1, \ldots, n$.

SVD vs. Bottleneck Network for $q = 2$





KL divergences of SVD and the bottleneck neural network

(drivers $i = 3, 44, 300, 1001;\ 642, 1645$).

- **Predictive Power of v-a Heatmaps?**

Poisson GAM Regression Models

Assume for $i = 1, \ldots, n$

$$Y_i \overset{\text{ind.}}{\sim} \text{Poi}\left(\lambda(\boldsymbol{x}_i)v_i\right),$$

with exposures $v_i > 0$ and regression function $\lambda : \mathcal{X} \to \mathbb{R}_+$ given by

Model 0: $\log \lambda(\boldsymbol{x}) = \beta_0 + s_1(\text{age driver}) + \beta_2 \cdot \text{age car},$

Model 1: $\log \lambda(\boldsymbol{x}) = \beta_0 + s_1(\text{age driver}) + \beta_2 \cdot \text{age car} + \beta_3 \cdot \text{PCA(heatmap)},$

Model 2: $\log \lambda(\boldsymbol{x}) = \beta_0 + s_1(\text{age driver}) + \beta_2 \cdot \text{age car} + \beta_3 \cdot \text{BN(heatmap)}.$

	cross-validation out-of-sample loss	std. dev. error
Model 0 (GAM classic)	1.4806	0.0240
Model 1 (PCA)	1.4573	0.0266
Model 2 (bottleneck net)	1.4579	0.0232

Conclusions

- v-a heatmaps allow for low-dimensional representations and approximations.
- Do these heatmaps have predictive power? Preliminary analysis shows "yes"!
- We have central limit theorems and rate of convergence for v-a heatmaps.
- Other speed buckets and claim sizes?