Matrix Variate and Kernel Density Methods for Applications in Telematics

MATRIX VARIATE AND KERNEL DENSITY METHODS FOR APPLICATIONS IN TELEMATICS

BY

NIKOLA POČUČA, B.Sc.

A THESIS

SUBMITTED TO THE DEPARTMENT OF MATHEMATICS & STATISTICS

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

Master of Science (2019) McMaster University

(Mathematics & Statistics) Hamilton, Ontario, Canada

TITLE:	Matrix Variate and Kernel Density Methods for Applications in Telematics
AUTHOR:	Nikola Počuča B.Sc., (Mathematics and Statistics) McMaster University, Hamilton, Canada
SUPERVISOR:	Dr. Paul D. McNicholas
NUMBER OF PAGES:	viii, 50

To my parents, Frosina and Zoran.

Abstract

In the last few years, telemetric data arising from embedded vehicle sensors brung an overwhelming abundance of information to companies. There is no indication that this will be abated in future. This information concerning driving behaviour brings an opportunity to carry out analysis. The merging of telemetric data and informatics gives rise to a sub-field of data science known as telematics. This work encompasses matrix variate and kernel density methods for the purposes of analysing telemetric data. These methods expand the current literature by alleviating the issues that arise with high-dimensional data.

Acknowledgements

First and foremost, I would like to thank my supervisor, Dr. Paul McNicholas. His guidance, support, and his nurturing personality throughout the years facilitated my growth as a researcher. Secondly, I would also like to show my appreciation to Dr. Petar Jevtić who was an unofficial co-supervisor throughout this endeavour. His role in this research was absolutely essential and I could not ask for a better collaborator.

In addition, I would like to acknowledge the funds provided by the Natural Sciences and Engineering Research Council of Canada (NSERC) through the Discovery Grant, E.W.R. Steacie Memorial Fellowship and Canada Research Chairs program, the Dr. Sri Gopal Mohanty Graduate Scholarship in Statistics, and the Department of Mathematics and Statistics.

I would like to acknowledge Michael P.B. Gallaugher for his time and patience. Without him, this research could not have been completed so efficiently.

Finally, I would like to thank Dr. Ben Bolker, and Dr. Fred Hoppe for their support this year both in my classes, and in my research. Thank you for making the defence process enjoyable.

Contents

List of Figures

Acronyms

AECM alternating expectation-conditional maximization. 23–26

BIC Bayesian information criterion. 25, 26, 31–35, 37

CA class agreement. vii, 27, 29, 30, 32, 33

CC Canada Cartage. 27, 28, 37

DE density estimation. viii, 4, 6–8

EM expectation maximization. 23

KDE kernel density estimation. vii, viii, 5–10, 16, 37, 39, 43

MBI Mixtures of Bilinear Factor Analyzers. vii, 23, 24, 26, 27, 29, 30, 32–35, 37

MISE mean integrated squared error. 8, 9

MSE mean squared error. vi, 7–9, 40, 41, 43

NDE non-parametric density estimation. vi, 4–7, 9

PPCA probabilistic principal component analysis. 12, 38

VA velocity acceleration. viii, 3, 15–18, 21–23, 26–31, 34–38

Chapter 1

Introduction

Consider one of the most fundamental definitions of data science in literature. Quoting Hayashi (1998): *Data science is not only a synthetic concept to unify statistics, data analysis and their related methods but also comprises its results. Data Science intends to analyze and understand actual phenomena with"data".*

Naturally, the science begins with the foundation that all analysis stems from data. McNicholas (2019) further elaborates

...if one may wish to define data science, the key must always be data. For a piece of work to be considered data science, we require only that data are at its heart.

That being said, in this work, telemetric data is the "heart" of all statistical methodology. Using this methodology, insights into intrinsic patterns and unique phenomena are revealed that would otherwise be hidden in the data. In literature, the science of telemetric data is referred to as telematics (Zhao, 2002). A blending of telecommunications and informatics, telematics is defined as "any analysis on devices that send and receive data across distances". By this definition, telematics is regarded as a sub-field of data science encompassing all statistical methods specifically targeted

at telemetric data. Due to the nature of telemetric data, problems arise because of its enormous quantity, and high-dimensional characteristics. As a result, methods developed for telematics must take into account these characteristics and deal with them accordingly. This work focuses on using powerful statistical tools in the domain of kernel density estimation and matrix variate distributions for the purposes of dealing with both the quantity and dimensionality of data.

Fundamentally, this work provides three extensions to the current literature in telematics. First, a non-parametric extension to the current construction of what is referred to as a velocity-acceleration (VA) heat map in the actuarial literature. Second, a method is proposed for detecting outliers in telemetric data based on the driver's personal driving behaviour. Finally, a matrix variate mixture model approach for clustering VA heat maps into one of several groups is introduced. All methods are focused on handling both the dimension, and size of telemetric data for applicability in industry settings.

Chapter 2

Background

The following sections outline the historical literature on telemetry, the introduction of statistical methodology, and applications pertaining to telemetric data. Clustering is discussed and elaborated with natural extensions to mixture models and matrix variate distributions. Finally, data arising from telemetric systems is elucidated through a series of applications in both insurance and logistics.

2.1 Density Estimation

In general terms, density estimation (DE) is the problem of estimating a probability density function p, using a set of given data points. Given very few assumptions on p, a non-parametric paradigm is adopted. By definition, non-parametric density estimation (NDE) seeks to estimate p with as few assumptions as possible (Wasserman, 2006). The choice of NDE stems from the nature of the data itself. For applications discussed in further sections, NDE is used as tool for dealing with the uncertainty of driving data. The choice of NDE is a result of this uncertainty for the underlying

distribution of driving data.

2.1.1 One-Dimensional NDE

The simplest method of NDE is the histogram. Although the origin of histograms is unclear, Ioannidis (2003) provides an abridged history of the use, and speculates on its origins. Suppose one observes realizations of some random variable X as $x_1, \dots x_n$. To reconstruct the underlying probability density that characterizes X, suppose that $X \in R \subseteq \mathbb{R}$. The probability density function $p(x)$ is said to be positive on its support R. In addition, assume that $p(x)$ is smooth and

$$\left| \frac{d}{dx} p(x) \right| \leq L, \forall x \in R,$$

where L is a finite positive constant (establishing an upper bound). The histogram partitions the set R into M equal bins where

$$R = \bigcup_{m=1}^{M} R_m, \quad |R_m| = |R_{m'}|, \quad R_m \cap R_{m'} = \emptyset, \quad \forall (m \neq m^{'}) \in 1, \dots, M.$$

To clarify, $|\cdot|$ denotes the cardinality operator constituting the length of the line segment for each subset. By this construction, for any given point $x \in R_m$, the density estimator of the histogram $\hat{p}$ for p is given by

$$\hat{p}(x) = \frac{M}{n|R|} \sum_{i=1}^{n} \mathbb{1}(x_i \in R_m), \quad \forall R_m \in R.$$

Here, $\mathbb{1}(x_i \in R_m)$ is the indicator function counting the number of observations within R_m. Intuitively, this estimator assigns equal weight to every point within the

bin R_m (Wasserman, 2006). However, this approach has several limitations which are discussed in later sections. Consider a superior method known as kernel density estimation (KDE). The KDE approach uses a smooth function to approximate the true density. In theory, the KDE method converges faster to the true density and therefore, is a natural extension to the previously introduced histograms (Wasserman, 2006). Given a non-negative kernel function K, a positive number h (bandwidth), the kernel density estimator $\hat{p}_h$ is written as

$$\hat{p}_h(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - x_i}{h}\right), \quad x \in R.$$

KDE is widely discussed in literature where importance is placed on the bandwidth h as it plays a major role in determining the shape of $\hat{p}_h$ (Gramacki, 2019). Specifically, the problem of estimating the density p using KDE is reduced down to a balance between bias and variance (Wasserman, 2006). The bandwidth is sometimes referred as a smoothness parameter; this terminology is adopted going forward.

2.1.2 Two-Dimensional NDE

A natural extension to the preceding methods is to consider a two-dimensional DE problem. Observe several observations of a bivariate random variable $\boldsymbol{\mathcal{X}}$ as $\boldsymbol{x} = \{\boldsymbol{x}_1 := (x_{11}, x_{12}), ..., \boldsymbol{x}_n\}$. Using these observations, reconstruct the underlying bivariate probability density that characterizes $\boldsymbol{\mathcal{X}}$ as follows. Suppose that $\boldsymbol{\mathcal{X}} \in R \subseteq \mathbb{R}^2$, then the probability density function $p(\boldsymbol{x})$ is non-zero only within R. Assume that $p(\boldsymbol{x})$ is smooth and bounded $\forall \boldsymbol{x} \in R$. By analogy with the univariate case, a bivariate

histogram is said to partition the set R into M equal rectangular bins where

$$R = \bigcup_{m=1}^{M} R_m, \quad |R_m| = |R_{m'}|, \quad R_m \cap R_{m'} = \emptyset, \quad \forall (m \neq m^{'}) \in 1, \ldots, M.$$

For the purposes of definition, $|\cdot|$ denotes the cardinality operator constituting the area of the rectangle for each subset. By this construction, for any given point $\boldsymbol{x} \in R_m$, the density estimator of the histogram for p is given by

$$\hat{p}(\boldsymbol{x}) = \frac{M}{n|R|} \sum_{i=1}^{n} \mathbb{1}(\boldsymbol{x}_i \in R_m), \quad \forall R_m \in R$$

To clarify, $\mathbb{1}(\boldsymbol{x}_i \in R_m)$ is the indicator function that counts the number of observations within R_m. Replicating the KDE approach for the two-dimensional case, let K be a bivariate function, and $\boldsymbol{H}$ be a symmetric, positive definite, non-random, 2×2 matrix. The bi-variate kernel density estimator $\hat{p}_{\boldsymbol{H}}(\boldsymbol{x})$ is formulated as

$$\hat{p}_{\boldsymbol{H}}(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} \det(\boldsymbol{H})^{-\frac{1}{2}} K\left(\boldsymbol{H}^{-\frac{1}{2}}(\boldsymbol{x} - \boldsymbol{x}_i)\right), \quad \boldsymbol{x} \in R.$$

Here $\boldsymbol{H}$, is defined as a smoothing matrix. The entries of $\boldsymbol{H}$ denote the bandwidth in each direction. This matrix constitutes how smooth the estimated density estimate would be for a given direction. Figure 2.1 displays a visualization of both methods. Both methods seek to estimate the same density. However, the KDE method is more robust and is shown to outperform the histogram in later sections.

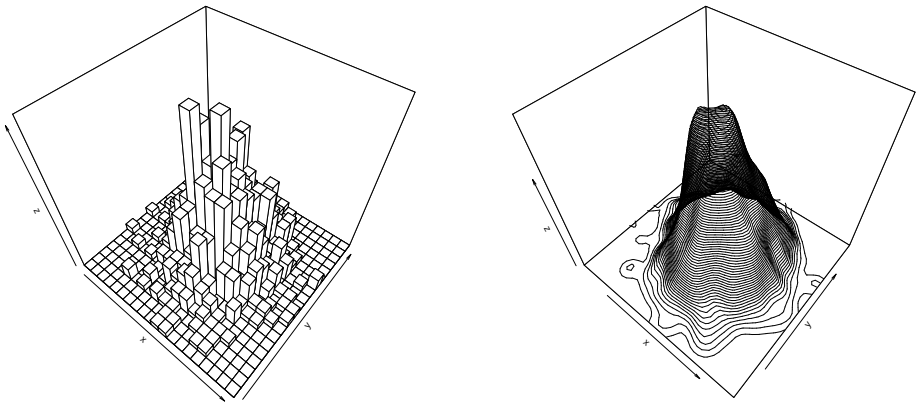

Figure 2.1: Visualization of DE methods applied to data generated from a standard bivariate Gaussian. Bivariate histogram (left) and bivariate KDE with Gaussian kernel (right) example showcasing the difference in approaches.

2.1.3 Performance

The measure of performance for DE requires a specification of error. Provided a distance measure of $\hat{p}$ to its target density p, one can assess performance. According to standard text on NDE (Gramacki, 2019), the most commonly used local error criteria is the mean squared error (MSE) . For the purposes of simplicity, consider only the univariate case. The MSE for a density estimator is given as

$$\begin{aligned}
\operatorname{MSE}(\hat{p}(x)) &= \mathbb{E}\left[(\hat{p}(x) - p(x))^2\right] \\
&= \mathbb{V}\operatorname{ar}(\hat{p}(x)) + (\mathbb{E}[\hat{p}(x)] - p(x))^2 \\
&= \mathbb{V}\operatorname{ar}(\hat{p}(x)) + \operatorname{Bias}^2(\hat{p}(x)).
\end{aligned}$$

The above formula is decomposed into an estimator's variance and its squared bias. This criterion is a measure of local error at a fixed point x. However, it is desirable to consider the global error of an estimator. The mean integrated squared error is the integral of the MSE over the domain defined as

$$\text{MISE}(\hat{p}(x)) = \int_{\mathbb{R}} \mathbb{V}\text{ar}\left(\hat{p}(x)\right) dx + \int_{\mathbb{R}} \text{Bias}^2(\hat{p}(x)) dx.$$

Often an estimator is parametrized by a bin width (in the case of histograms), or a bandwidth (for KDE). As a consequence, the best estimator in each respective approach is one that minimizes the MISE. In general, given some parameter $\gamma \in \mathbb{R}^+$ for an estimator $\hat{p}_\gamma(x)$, the MISE is minimized as

$$\gamma_{opt} = \underset{\gamma \in \mathbb{R}^+}{\text{argmin}} \left(\text{MISE}(\hat{p}_\gamma(x))\right).$$

As a comparison, the optimal parameters of both the histogram and KDE approach are provided for the MSE. Subsequently, the optimal parameter for MISE of a kernel density estimator is established. Finally, given an optimal MISE of a multidimensional KDE, performance is discussed.

2.1.4 Minimizing MSE

The KDE method can be shown to converge to the true density via MSE faster than using a histogram. This proof is left as an exercise in many books and is provided in Section A.1 (Chen, 2017). As the KDE is shown to be superior locally, the following section discusses its performance in higher dimensions.

2.1.5 Performance in Higher Dimensions

For the purposes of measuring performance for NDE in higher dimensions, the extension to multivariate KDE is generalized with the following expression. The MSE of a KDE is given by Gramacki (2019) as

$$\text{MSE}(\hat{p}_h(\boldsymbol{x})) = \mathcal{O}(h^4) + \mathcal{O}\left(\frac{1}{nh^d}\right),$$

where d is the dimension. Within this equation, the two terms correspond to the bias and variance, respectively. Note that the bias remains constant in order 4 for all higher dimensions. However, the variance is of order h^{-d} for $h < 1$. As a consequence, this prohibits reliable estimation in higher dimensions. However, if n increases sufficiently quickly this issue is reduced. KDE suffers from what is known as the "curse of dimensionality" (Bellman, 1966). The issue in higher dimensions concerns the vanishing of gradients. This is a long studied phenomenon where techniques have been developed to improve KDE (Di Marzio and Lafratta, 1999). It is noted that the optimal bandwidth selection in higher dimensions has convergence in $\mathcal{O}\left(n^{-\frac{1}{4+d}}\right)$. As a consequence, the KDE method is only efficient for $d \leq 2$.

2.2 Clustering

2.2.1 Brief Historical Context

McNicholas (2016) provides the earliest mention of mixture model-based clustering which can be traced back to Tiedeman (1955). Formally, a cluster is described in the context of a type (Tiedeman, 1955). Let G be the number of groups within a

population. Each observation belonging to the gth group is generated by a density function; the function being a Gaussian distribution as a speical case (Tiedeman, 1955). Upon removing the identity of the group to which each observation belongs, the result is a mixture of an unknown density. The problem of reconstructing the original G densities of their types is what is known as clustering (Tiedeman, 1955). Following Tiedeman's work, Wolfe (1965) defined a cluster using two different defintions. One where a cluster is a mode of a distribution, and the second definition where similarity between observations is of focus. However, similarity is often arbitrary to define (McNicholas, 2016). McNicholas (2016) specifies the definition of a cluster in the context of a mixture model and specifies: Suppose that a cluster is a uni-modal component within an appropriate finite mixture model. Here, appropriate is defined in the sense that the finite mixture model is one which has the flexibility and parametrization that is necessary to fit the data.

2.2.2 Finite Mixture Models

The framework of model-based clustering considers the underlying assumption that a finite mixture model embodies a representation of heterogeneous data. Consider a random variable $\boldsymbol{\mathcal{X}}$, from a G-component finite mixture model with probability density function of the form

$$f(\boldsymbol{x}|\boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f_g(\boldsymbol{x}|\boldsymbol{\theta_g}), \tag{2.1}$$

where $\boldsymbol{\vartheta} = \{\pi_1, \ldots, \pi_G, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G\}$, $\boldsymbol{x}$ is a realization of $\boldsymbol{\mathcal{X}}$, π_g is a mixing proportion where $\pi_g > 0$, $\sum_{g=1}^{G} \pi_g = 1$, and f_g is a probability density function parametrized by $\boldsymbol{\theta}_g$. The distribution within each cluster is usually taken to be identical, as a result

the density in (2.1) is simplified as $f_g(\boldsymbol{x}|\boldsymbol{\theta}_g) = f(\boldsymbol{x}|\boldsymbol{\theta}_g) \quad \forall g$.

2.2.3 Factor Analyzers

For the purposes of clustering high dimensional data, issues arise due to the curse of dimensionality (Di Marzio and Lafratta, 1999). A standard approach is to reduce the number of dimensions by considering a series of underlying factors of a lower dimension (Spearman *et al.*, 1950). Let $\boldsymbol{\mathcal{X}}_i$ represent an r dimensional random vector, with $\boldsymbol{x}_i$ as a realization. The factor analyzers model for $\boldsymbol{\mathcal{X}}_1, \ldots, \boldsymbol{\mathcal{X}}_N$, is given by

$$\boldsymbol{\mathcal{X}}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{U}_i + \boldsymbol{\epsilon}_i,$$

where $\boldsymbol{\mu}$ is a mean location vector, $\boldsymbol{\Lambda}$ is a $r \times s$ matrix of factor loadings, with $s < r$, $\boldsymbol{U}_i \sim \mathcal{N}_r(\boldsymbol{0}, \boldsymbol{I})$ denoting the latent factors, and $\boldsymbol{\epsilon}_i \sim \mathcal{N}_r(\boldsymbol{0}, \boldsymbol{\Psi})$ where $\boldsymbol{\Psi} = \text{diag}(\psi_1, \psi_2, ..., \psi_r)$. Here, $\mathcal{N}_r$ denotes the r-dimensional multivariate normal distribution. Furthermore, the latent factors $\boldsymbol{U}_i$ and noise $\boldsymbol{\epsilon}_i$ are independent of each other. It is noted that the probabilistic principal component analysis (PPCA) is a special case of the factor analysis model with a specific isotropic constraint on $\boldsymbol{\Psi}$ (Tipping and Bishop, 1999). The PPCA approach has been recently used in actuarial literature for the purposes of modelling frequency of claims (Gao *et al.*, 2019). The factor analyzers model is a flexible extension of PPCA. The factor analyzers model is considered to be the best choice for dealing with telemetric data as it is highly efficient in reducing dimensionality (Inui *et al.*, 2009).

2.2.4 Matrix Variate Normal Distribution

Suppose a matrix is considered to be an observation sampled from a distribution. Naturally, an appropriately sized matrix variate distribution should be considered to model randomness. Consider the matrix variate normal distribution (Gupta and Nagar, 1999). The domain of the matrix variate normal is the space of all real valued matrices. This assumption is less restrictive compared to other matrix variate distributions of the same type such as the Wishart (Gupta and Nagar, 1999). Let $\mathscr{X}$ be a random variable with an $r \times c$ matrix $\boldsymbol{X}$ as a realization. As a consequence, $\mathscr{X}$ is distributed according to a matrix variate distribution. The random matrix $\boldsymbol{X}(r \times c)$ is said to have a matrix variate normal distribution with mean matrix $\boldsymbol{M}(r \times c)$ and covariance matrix $\boldsymbol{\Psi} \otimes \boldsymbol{\Sigma}$. Each matrix is appropriately sized as $\boldsymbol{\Sigma}(r \times r)$, $\boldsymbol{\Psi}(c \times c)$, where $\text{vec}(\mathscr{X}) \sim \mathcal{N}_{rc}(\text{vec}(\boldsymbol{M}'), \boldsymbol{\Psi} \otimes \boldsymbol{\Sigma})$. Here, $\otimes$ denotes to the Kronecker product and vec denotes the vectorization of a matrix. Given this specification, the density is formulated as

$$\varphi_{r,c}(\boldsymbol{X}; \boldsymbol{M}, \boldsymbol{\Psi} \otimes \boldsymbol{\Sigma}) = \frac{\exp\left\{ -\frac{1}{2}\text{tr}\big(\boldsymbol{\Psi}^{-1}(\boldsymbol{X} - \boldsymbol{M})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{M})\big) \right\}}{(2\pi)^{\frac{rc}{2}} \det(\boldsymbol{\Psi})^{\frac{r}{2}} \det(\boldsymbol{\Sigma})^{\frac{c}{2}}}.$$

The matrix variate normal distribution is equivalent to a vectorization of a multivariate normal. Note that the covariance matrices of row and column are non-unique as they are defined through a Kronecker product (Dutilleul, 1999). As a result, both densities are parametrized by the product and not individual co-variance matrices (Gupta and Nagar, 1999). The main benefits for using a matrix variate representation are two-fold. The first is the reduction in the number of parameters used. Gallaugher and McNicholas (2018a) shows that for both matrices collectively there is

a reduction of $\frac{1}{2}\left[(r-s)^2-(r+s)+(c-v)^2-(c+v)\right]$ for $s<r$, $v<c$. As a result, this adds a second benefit as there is an increase in the speed for estimating model parameters in high dimensional settings.

2.2.5 Mixture of Bilinear Factor Analyzers

Due to the issues of high dimensionality, an analogous extension of the factor analyzers model for matrix variate data is introduced. The mixtures of matrix variate bilinear factor analyzers model (MBI) is a powerful approach for dealing with both high-dimensional data and the presence of a mixture of populations (Gallaugher and McNicholas, 2018a). Suppose latent factors of size $s<r$, $v<c$, for a matrix variate random variable constitute the data with probability π_g of occurring as

$$\boldsymbol{X}_i = \boldsymbol{M}_g + \boldsymbol{A}_g\boldsymbol{W}_{ig}\boldsymbol{B}_g' + \boldsymbol{A}_g\mathcal{E}_{ig}^B + \mathcal{E}_{ig}^A\boldsymbol{B}_g' + \mathcal{E}_{ig},$$

where $\boldsymbol{M}_g(r,c)$ is the mean matrix, $\boldsymbol{W}_{ig}(s,v) \sim \mathcal{N}_{s\times v}(\boldsymbol{0}, \boldsymbol{I}_s, \boldsymbol{I}_v)$ is a matrix random variate of latent factors, $\boldsymbol{A}_g(r\times s)$ are column factor loadings, and $\boldsymbol{B}_g(c\times v)$ are row factor loadings, respectively. Finally, the noise is distributed according to

$$\mathcal{E}_{ig}^A \sim \mathcal{N}_{r,v}(\boldsymbol{0}, \boldsymbol{U}_g, \boldsymbol{I}_v),$$

$$\mathcal{E}_{ig}^B \sim \mathcal{N}_{s,c}(\boldsymbol{0}, \boldsymbol{I}_s, \boldsymbol{V}_g),$$

$$\mathcal{E}_{ig} \sim \mathcal{N}_{r,c}(\boldsymbol{0}, \boldsymbol{U}_g, \boldsymbol{V}_g).$$

For applications in telematics, the MBI model is used to cluster matrix variate objects pertaining to the heterogeneous population of drivers.

2.3 Telemetric Data

Telemetry is defined as an automated communications process in which measurements are collected at remote, usually inaccessible areas. These measurements are then transmitted to receiving devices. All measurements are referred synonymously as telemetries, or telemetric data. Henceforth, any analysis based on telemetric data is referred to as telematics. The applications in this work focus specifically on car telematics.

2.3.1 Historical Background

Mayo-Wells (1963) provides a complete historical background of the use of telemetry. In summary, the first industrial use of telemetric data originates from a patent in a circuit design that enabled sending of synchronized rotation information over a distance (Michalke, 1901). A decade later, the design was further expanded by Commonwealth Edison in 1912 for monitoring electrical loads of power grids (De Dutta and Prasad, 2019). Years later, due to the completion of the Panama Canal, telemetry systems were utilized extensively for the monitoring of locks and water levels. Progression of military technology throughout the 1930's utilized radio based wireless telemetry. As a consequence, this created a need for statistical analysis of telemetric data. For example, the analysis of V2-rockets required the processing of telemetric data from a variety of on-board devices monitoring temperature, pressure, and kinematics. These methods were later refined during the decade of space exploration in the sixties. Space exploration initiated the use of telemetric systems for monitoring satellite trajectories as well on-board safety systems. Near the end of the 20th century, embedded devices

provided the platform for user-facing applications of telematics. Non-military applications in motor sports like Formula 1 require telemetry systems in next generation vehicles. The Advanced Telemetry Linked Acquisition System (ATLAS) developed by McLaren Applied Technologies allows for the storage and collection of telemetry from many on-board sensors (Azzoni *et al.*, 1998). The type of data collected includes speed, g-forces, steering angles, and engine temperatures. As an example for everyday use, Tong and Hung (2010) analyzed telemetric data for the purposes of providing a speed time profile of driving cycles. The objective of a driving cycle is to measure vehicle performance and driving characteristics.

2.3.2 Telemetric Data in Insurance

For insurance purposes these driving cycles provide key insights into behaviour of drivers under a policy. Recent literature describes covariate selection expanding this methodology. Specifically, the literature introduced a matrix variate object for analyzing a driver's telemetric data. This object is referred to as a VA heat map in the actuarial literature (Wüthrich, 2017). From the perspective of insurance, the VA heat map conveys several key pieces of information. The risk of the driver can be categorized by their respective heat maps as interperation is fairly straightforward. As a result, clustering these heat maps provides a classification of risk for each driver into one of several groups. Velocity and speed are used interchangeably throughout the literature; this habit is adopted going forward. Recently, Wüthrich (2017) develops a K-means approach for clustering the VA heat maps for thousands of drivers. Within this framework, GPS data is gathered over a series of individual trips for a specific driver. Let $(\delta_{xt}, \delta_{yt})_t$ denote locational data at time t in meters. Velocity in m/s is

calculated on the time interval $(t-1, t]$ for $t \geq 1$, given by

$$v_t = \frac{\sqrt{(\delta_{xt} - \delta_{x(t-1)})^2 + (\delta_{yt} - \delta_{y(t-1)})^2}}{(t-(t-1))}, \quad a_t = \frac{v_t - v_{t-1}}{t-(t-1)}.$$

Wüthrich (2017) elaborates that this calculation for a_t is determined by the "average" speed over the time interval and not reflective of the true instantaneous acceleration. Proceeding on, a_t is considered the average acceleration for all intents and purposes. The author constructs a matrix variate object refereed to as a VA heat-map. Construct a rectangle R where v_t and a_t are considered to be values of a two-dimensional coordinate system. Next, consider a partitioning of R into M equally sized rectangles as

$$R = \bigcup_{m=1}^{M} R_m, \quad R_m \cap R_{m'} = \emptyset, \quad \forall m \neq m'.$$

Finally, consider some probability distribution of $F \in \mathcal{P}(R)$ having probability weight

$$x_m = \int_{R_m} dF \geq 0, \quad m = 1, \ldots, M, \quad \text{satisfying } \sum_{m=1}^{M} x_m = 1.$$

The resulting object is a matrix of probability weights spanning $x_m, \forall m$. This object is referred by the author as a VA heat map, where the column and row denotes the probability weight of velocity and acceleration for entry x_m, respectively. This construction is similar to that of a histogram or a KDE with a uniform kernel selection with the exception of having fixed bandwidth parameters.

Chapter 3

Methodology

3.1 Telemetric Data

Expanding the work of Wüthrich (2017), construction of the VA heat map is performed as follows. Begin with a series of telemetric data indexed by time t, and collected from driver i. Let δ_{it} be composed of a two-dimensional vector containing locational GPS coordinates defined as

$$\delta_{it} = (\delta_{xit}, \delta_{yit}), \quad \mathbf{\Delta}_i = \{\delta_{it}\}_{t=1}^{T_i}.$$

Here, T_i is the last time index received from driver i, and $\mathbf{\Delta}_i$ is considered to be the collection of positions sorted by time t. Given $\mathbf{\Delta}_i$, calculate the average velocity and

acceleration over a specified time interval as follows

$$v_{it} = \frac{\sqrt{(\delta_{xit} - \delta_{xi(t-l)})^2 + (\delta_{yit} - \delta_{yi(t-l)})^2}}{t - (t-l)} = \frac{\sqrt{(\delta_{xit} - \delta_{xi(t-l)})^2 + (\delta_{yit} - \delta_{yi(t-l)})^2}}{l},$$

$$a_{it} = \frac{v_{it} - v_{i(t-1)}}{t - (t-l)} = \frac{v_{it} - v_{i(t-1)}}{l}, \qquad \boldsymbol{v}_i = \{v_{it}\}_{t=1}^{T_i-1}, \quad \& \quad \boldsymbol{a}_i = \{a_{it}\}_{t=1}^{T_i-2}.$$

Here, l is considered to be the latency or time delay of the GPS device. Latency in this context is defined as the incremental delay between GPS readings for a specific time index. Wüthrich (2017) considers the latency of devices to be $l = 1$. However, in practice, the latency can be less than or greater than 1. Naturally, latency has an effect on smoothness of velocity (v) and acceleration (a) with respect to time. If the latency is fairly small, the v and a graphs are fairly smooth. On the other hand if latency is large, smoothness is reduced as calculated v and a will resemble more coarse behaviour. Coarseness is inconsistent with the true behaviour of v and a over time. In real environments, a driver's v and a is smooth and continuous so the calculated v and a should approximate true behaviour as accurately as possible. The collection of $\boldsymbol{v}_i$ and $\boldsymbol{a}_i$ are considered vectors of size $T_i - 1$ and $T_i - 2$ respectively. In practice, telemetric data is gathered in segments known as driving cycles (Tong and Hung, 2010). As a result, the calculation of v_{it} from one segment to another should be omitted. Furthermore, the last entry $v_{i(T_i-1)}$ should be omitted as well to allow $\boldsymbol{v}_i$ and $\boldsymbol{a}_i$ to have the same length defined as $T_i^\star$. This process is refereed to as "standardization". From actuarial literature it is known that driving behaviour can interpreted through a matrix variate object known as a VA heat map. The VA heat map is considered to be a joint probability density for v and a on a sample

space. Consider this object to be an approximation of the density function $p(\boldsymbol{w}_i)$ for a bivariate random variable $\boldsymbol{W}_i := (V_i, A_i)$, where $\boldsymbol{w}_i := (v_i, a_i)$ is a realization for $\boldsymbol{W}_i$. Note that telemetric data are also considered to be realizations of $\boldsymbol{W}_i$. However, these events are also indexed with time t as $\boldsymbol{w}_{it}$. The notation for $\boldsymbol{w}_{it}$ includes t as a subscript to denote that it is indeed a discrete measurement collected from a device. In contrast, $\boldsymbol{w}_i$ omits t as a subscript denoting that it is not measurement in a telemetric sense. In summary, the random variable $\boldsymbol{W}_i$ signifies the true continuous random behaviour on the sample space of possible speed and accelerations.

3.2 Velocity-Acceleration Heat Map

Consider the use of KDE for the problem of estimating an unknown joint probability density $p(\boldsymbol{w}_i)$ on the space $R \subseteq \mathbb{R}^2$ for driver i. Let K be a bivariate function defined on the space $\mathbb{R}^2$. Furthermore, let $\boldsymbol{H}$ be a constant, positive definite, symmetric matrix defined as the smoothness parameter. The KDE of $p(\boldsymbol{w}_i)$ is written as

$$\hat{p}_{\boldsymbol{H}}(\boldsymbol{w}_i) = \frac{1}{T_i^\star} \sum_{t=1}^{T_i^\star} \det(\boldsymbol{H})^{-\frac{1}{2}} K\left(\boldsymbol{H}^{-\frac{1}{2}}(\boldsymbol{w}_i - \boldsymbol{w}_{it})\right), \quad \boldsymbol{x} \in R.$$

Due to popularity and radial symmetry, the standard bivariate Gaussian distribution is selected as the kernel function K. As a result, K is formulated as

$$K\left(\boldsymbol{H}^{-\frac{1}{2}}(\boldsymbol{w}_i - \boldsymbol{w}_{it})\right) = (2\pi)^{-1} \exp\left\{-\frac{1}{2}(\boldsymbol{w}_i - \boldsymbol{w}_{it})^{\mathrm{T}} \boldsymbol{H}^{-1}(\boldsymbol{w}_i - \boldsymbol{w}_{it})\right\}.$$

This selection of K allows the kernel estimator to be the weighted sum of normal densities centered at telemetry points $\boldsymbol{w}_{it}$. In summary, the true density $p(\boldsymbol{w}_i)$ is

approximated using the kernel estimator

$$\hat{p}_{\boldsymbol{H}}(\boldsymbol{w}_i) = \frac{1}{T_i^\star} \sum_{t=1}^{T_i^\star} \varphi_2(\boldsymbol{w}_i, \boldsymbol{w}_{it}, \boldsymbol{H}).$$

The selection of the smoothness matrix $\boldsymbol{H}$ affects both the shape and orientation of the kernels on the two-dimensional space. The smoothing matrix as a consequence induces an orientation. Here, orientation is regarded as a basic difference between a bivariate and univariate KDE since orientation is not defined for the univariate case. As mentioned before, the bandwidth is a key parameter for optimizing performance of KDE. For the purposes of simplicity, the normal scale selector (Chacón *et al.*, 2011) is selected as

$$\boldsymbol{H}_{NS} = \left(\frac{1}{T_i^\star}\right)^{\frac{1}{3}} \hat{\boldsymbol{\Sigma}}^\star,$$

where $\hat{\boldsymbol{\Sigma}}^\star$ is the sample covariance matrix. The use of this method for estimating the probability density function is superior than that of a histogram or an uniform kernel. Unlike in the actuarial literature, the use of a multivariate histogram is avoided as it requires to specify the size of bins, the origin, and the orientation (Silverman, 1998). Furthermore, the discontinuous nature of uniform based estimators contradict driving behaviour as previously discussed. The current use of these estimators by Wüthrich (2017) has a large number of bins (size 40000) and is purposely avoided in this work. Problems arise when telemetric data are poor or the latency is too large. The use of a uniform kernel or histograms fail to capture the density in settings such as this. In contrast, the KDE method with a Gaussian kernel converges faster to the true density. Furthermore, the KDE method captures the continuous nature of data even when latency is large. For implementation see the `MASS` package regarding the use of

two-dimensional KDE (Venables and Ripley, 2002).

With all methods introduced, the construction of a VA heat map is conceived as follows. Consider a partitioning of R into M equally sized rectangles as

$$R = \bigcup_{m=1}^{M} R_m, \quad R_m \cap R_{m'} = \emptyset, \quad \forall m \neq m'.$$

Secondly, construct the matrix variate object $\boldsymbol{X}_i$ pertaining to the collection of probability weights as

$$x_{im} = \int_{R_m} p_{\boldsymbol{H}_{NS}}(\boldsymbol{w}_i) d\boldsymbol{w}_i > 0, \quad m = 1, \ldots, M,$$

$$\text{satisfying } \sum_{m=1}^{M} x_{im} = 1, \quad \boldsymbol{X}_i = \{x_{im}\}_{m=1}^{M}.$$

This matrix variate object is defined as the VA heat map for driver i. Note that this extension differs from the original construction by Wüthrich (2017) in one key distinction; the use of KDE to estimate the joint density on R. This imposes a continuous non-zero probability on R such that it covers the entire space of possible driving situations that may occur. In addition, the choice of a smooth kernel allows the matrix $\boldsymbol{X}_i$ to have non-zero entries which more accurately resemble a driver's behaviour. Lastly, the use of a smooth kernel allows for the construction of the matrix variate object to be viable even when $l > 1$.

3.3 Detection of Deviant Events

Consider the problem of detecting which events fall outside of the normal behaviour of driving. The introduction of a new method for the detection of deviant events defined

as the α-level deviation test is formulated as follows. Let α be the significance level of the test. Given this level, the test will capture the $100\alpha\%$ of events that deviate most from the original style. Driving style is a term used in actuarial literature for the driver's behaviour exhibited on the vehicle over time. Let $\boldsymbol{\omega}_i = (\boldsymbol{v}_i, \boldsymbol{a}_i)$ be a collection of standardized telemetries (size $T_i^\star$) from which a VA heat map $\boldsymbol{X}_i$ is estimated. Next, let $\mathcal{M}$ be a function which maps a telemetry event $\boldsymbol{w}_{it} \in \boldsymbol{\omega}_i$ to its corresponding probability weight within the map defined as

$$\mathcal{M}(\boldsymbol{w}_{it}; \boldsymbol{X}_i) = \begin{cases} x_{im} & \boldsymbol{w_{it}} \in R_m \\ 0 & \text{otherwise.} \end{cases}$$

Furthermore, let $\hat{p}_{it} = \mathcal{M}(\boldsymbol{w}_{it}; \boldsymbol{X}_i)$ which corresponds to the estimated probability over the region R_m that the telemetric event $\boldsymbol{w}_{it}$ belongs to. The rationale is that the rarest telemetry events belong to the regions with the smallest probability. The problem is then reduced down to using DE to estimate a density function of probability weights $g(p)$. Few assumptions are imposed, and the non-parametric method of a one-dimensional KDE is utilized. Let the kernel density estimator for $g(p)$ be

$$\hat{g}_h(p) = \frac{1}{T_i^\star} \sum_{t=1}^{T_i^\star} \frac{1}{h} K\left(\frac{p - \hat{p}_{it}}{h}\right),$$

where $p \in \mathbb{R}$, and $\hat{p}_{it}$ are the estimated probabilities of each telemetric event. K is once again chosen as the Gaussian kernel, and the bandwidth $h > 0$ is selected via Silverman's rule of thumb (Gramacki, 2019). The goal is to identify the $100\alpha\%$ of

rarest events with the given density $\hat{g}_h(p)$. By integrating $\hat{g}_h(p)$ as

$$\int_0^{p^\star} \hat{g}_h(p)dp = \alpha,$$

and numerically solving for $p^\star$. The $100\alpha\%$ of rarest events are captured as

$$\boldsymbol{w}_{it}^{\alpha} := \begin{cases} \boldsymbol{w}_{it}, & \hat{p}_{it} < p^\star \\ \emptyset, & \text{otherwise,} \end{cases} \qquad \boldsymbol{\omega}_i^{\alpha} = \bigcup_{t=1}^{T_i^\star} \{\boldsymbol{w}_{it}^{\alpha}\}.$$

The greatest advantage of this method for detecting deviant events $\boldsymbol{\omega}_i^{\alpha}$ is in its flexibility. Each driver has their own particular style which is encapsulated in the VA heat map. Using the heat maps in such a way results in detecting when each driver is deviating from their own dominantly established behaviour. This method can also be expanded to compare two different drivers. For example, events from one driver could be considered normal with respect to their own driving style. However, said events may be considered deviant with respect to another driver's style. For industry use, these types of analysis are important when comparing drivers for training purposes.

3.4 Estimation and Performance for MBI

The estimation procedure for MBI is based on local maximum likelihood estimation. A common approach for estimating finite mixture models is with the expectation maximization (EM) algorithm (Dempster *et al.*, 1977). Many extensions to the EM algorithm have been proposed (McLachlan and Krishnan, 2008). For the purposes of

dealing with latent factor models, McNicholas and Murphy (2008) uses the alternating expectation–conditional maximization algorithm (AECM; Meng and Van Dyk, 1997). Estimation of parameters pertaining to the MBI model is performed as follows. Consider a latent variable Z_{ig} denoting membership of observation i belonging to group g as,

$$Z_{ig} = \begin{cases} 1, & \boldsymbol{X}_i \text{ belongs to group } g \\ 0, & \text{otherwise.} \end{cases}$$

For example, the component membership for observation 1 is given as $\boldsymbol{z}_1 := (z_{11}, \ldots, z_{1G})$. Suppose observation i is in group g. The formulation of factor analysers in the matrix variate case has the density

$$\boldsymbol{X}_i | z_{ig} = 1 \sim \mathcal{N}_{r,c}(\boldsymbol{X}_i; \boldsymbol{M}_g, \boldsymbol{U}_g + \boldsymbol{A}_g \boldsymbol{A}_g^{'}, \boldsymbol{V}_g + \boldsymbol{B}_g \boldsymbol{B}_g^{'}).$$

When written in this formulation, the complete data-likelihood is taken to be

$$L(\boldsymbol{X}; \boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{g=1}^{G} [\pi_g \varphi_{r,c}(\boldsymbol{X}_i; \boldsymbol{M}_g, \boldsymbol{U}_g + \boldsymbol{A}_g \boldsymbol{A}_g^{'}, \boldsymbol{V}_g^{'} + \boldsymbol{B}_g \boldsymbol{B}_g^{'})]^{z_{ig}},$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1 := (\pi_1, \boldsymbol{z}_1, \boldsymbol{M}_1, \boldsymbol{U}_1, \boldsymbol{V}_1, \boldsymbol{A}_1, \boldsymbol{B}_1), \ldots, \boldsymbol{\theta}_G)$. Estimation based on the AECM algorithm is complicated, having multiple intermediate steps and algebraic expressions. To maintain clarity, the following is a summarized version where several intermediate steps $\hat{\boldsymbol{S}}_g^A$ and $\hat{\boldsymbol{S}}_g^B$ are omited. For specifics, see Gallaugher and McNicholas (2018a). The AECM algorithm consists of three stages. Within the first stage the complete-data is taken to be the observed matrices $\boldsymbol{X}_1, ..., \boldsymbol{X}_N$, and the

component memberships $\boldsymbol{z} = (\boldsymbol{z}_1, ..., \boldsymbol{z}_N)$. The E-step for this stage is given by

$$\hat{z}_{ig}^{(t)} = \frac{\pi_g \varphi_{r,c}(\boldsymbol{X}_i; \boldsymbol{\theta}_g)}{\sum_{h=1}^{G} \pi_h \varphi_{r,c}(\boldsymbol{X}_i; \boldsymbol{\theta}_h)}.$$

In the conditional maximization step, the updates for π_g and $\boldsymbol{M}_g$, at some iteration t is given by

$$\hat{\boldsymbol{M}}_g^{(t)} = \frac{\sum_{i=1}^{N} \hat{z}_{ig}^{(t)} \boldsymbol{X}_i,}{\sum_{i=1}^{N} \hat{z}_{ig}^{(t)}} \quad \text{and} \quad \hat{\pi}_g^{(t)} = \frac{\sum_{i=1}^{N} \hat{z}_{ig}^{(t)}}{N}.$$

In the second stage, the complete-data is taken to be the observed $\boldsymbol{X}_1, ..., \boldsymbol{X}_N$, the component memberships $\boldsymbol{z}$, and the $r \times s$ latent matrices for column factors. In addition, $N_g^{(t)} = \sum_{i=1}^{N} \hat{z}_{ig}^{(t)}$. The expectation step for this stage is given in (Gallaugher and McNicholas, 2018a, 3.2) yields some intermediate terms for $\hat{\boldsymbol{S}}_g^B$. In the conditional maximization step the parameter update for $\boldsymbol{U}$ is taken to be

$$\hat{\boldsymbol{U}}_g^{(t)} = \frac{1}{N_g^{(t)} c} \text{diag}\{\hat{\boldsymbol{S}}_g^B\}.$$

In the third stage, the complete-data is taken to be the observed $\boldsymbol{X}_1, ..., \boldsymbol{X}_N$, the component memberships $\boldsymbol{z}$, and the $c \times v$ latent matrices for row factors. The expectation step for this stage is given in (Gallaugher and McNicholas, 2018a, 3.2) yields some intermediate terms for $\hat{\boldsymbol{S}}_g^A$. In the conditional maximization step the parameter update for $\boldsymbol{V}$ is given as

$$\hat{\boldsymbol{V}}_g^{(t)} = \frac{1}{N_g^{(t)} r} \text{diag}\{\hat{\boldsymbol{S}}_g^A\}.$$

Convergence of the AECM algorithm is based on the Aitken acceleration criterion (Aitken, 1926) defined as

$$a^{\star(t)} = \frac{l^{(t+1)} - l^{(t)}}{l^{(t)} - l^{(t-1)}},$$

where $l^{(t)}$ is the observed log likelihood at iteration t. Let

$$l_{\infty}^{(t+1)} = l^{(t)} + \frac{l^{(t+1)} - l^{(t)}}{1 - a^{\star(t)}}$$

be the observed estimate after many iterations at $t+1$. Termination of the algorithm occurs when $l_{\infty}^{(t+1)} - l_{\infty}^{(t)} \in (0, \varepsilon)$ for some pre-specified ε (McNicholas, 2010). Model selection is based on the Bayesian information criterion (BIC; Schwarz *et al.*, 1978). The BIC is a criterion to assess the performance of the model fit, while penalizing for the number of parameters used. For interpretability, the BIC is used for assessing model performance (larger is better). Let ρ be the number of parameters used. The BIC is then formulated as $\text{BIC} = 2l(\boldsymbol{\theta}) - \rho \log N$. With all methods relating to MBI introduced, the clustering problem is formulated as follows. Assuming there exists a heterogeneous population of drivers of up to G types. Let $\boldsymbol{X}_i$ be a VA heat map of driver i. Formally, $\mathscr{X} \sim \mathcal{N}_{r,c}^{G}(\boldsymbol{\theta})$ with probability density function

$$f(\boldsymbol{X}_i; \boldsymbol{\theta}) = \sum_{g=1}^{G} \pi_g \varphi_{r,c}(\boldsymbol{X}_i; \boldsymbol{M}_g, \boldsymbol{U}_g + \boldsymbol{A}_g \boldsymbol{A}_g^{'}, \boldsymbol{V}_g^{'} + \boldsymbol{B}_g \boldsymbol{B}_g^{'}),$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1 = (\pi_1, \boldsymbol{M}_1, \boldsymbol{U}_1, \boldsymbol{V}_1, \boldsymbol{A}_1, \boldsymbol{B}_1), \ldots, \boldsymbol{\theta}_G)$. Estimation of the model is performed by using the AECM algorithm as mentioned previously. Once the model has been estimated, classification of drivers into one of G types is done in accordance with selecting the maximum a posteriori of component memberships. The classification $\mathcal{C}_i$ for driver i is given by

$$\mathcal{C}_i = \begin{cases} g, & \hat{z}_{ig} = \max\limits_{\forall g}\{\hat{z}_{i1}, \ldots \hat{z}_{iG}\}, \\ 0, & \text{otherwise.} \end{cases}$$

Chapter 4

Analyses

This chapter is an amalgamation of analyses on a series of telemetric datasets. Each section contains an in-depth discussion of results for each type of analysis. All methods were implemented using the Julia language within the TeleMap.jl package (Počuča, 2019). The first section is devoted to the Cartage Canada dataset which utilizes the VA heat map and α-level deviation test to identify outlying events. The second section contains class agreement (CA) comparisons between the MBI and K-means model for the ETH Zurich synthetic dataset (Wüthrich, 2017). The third section is a cluster analysis of VA heat maps from the National Renewable Energy Laboratory, and is a practical application of clustering real data.

4.1 Cartage Canada Data

The Cartage Canada dataset is a collection of telemetries for $N = 26$ truck drivers. Canada Cartage (CC) is a logistics company that provided the data as part of a research agreement with McMaster. Beginning with the construction of the VA heat

maps for each driver, consider the splitting up of the driving styles into parking, city, and highway driving. The splitting is performed in accordance with Wüthrich (2017), and is considered the standard approach. Highway driving is classified as speeds $[90, V_{\max})$ km/h. Here, $V_{\max}$ is some upper limit of velocity that a vehicle can endure. The visualization of the VA heat map for highway driving is shown in Figure 4.1. Only highway driving will be the focus for deviation analysis. Deviation analysis refers to the method for identifying $\alpha-$level deviation events for a particular driver and is adopted moving forward. An example of this analysis is visualized in Figure 4.2. Given an α, telemetric events are identified as being deviant if they exceed a certain threshold that is determined via KDE. In Figure 4.2a, $\alpha = 0.05$. Those events considered deviant of this level are superimposed on the VA heat map. There are a total of 67 events considered to be deviant for this particular driver. Furthermore, Figure 4.2b refers to those events considered deviant with level $\alpha = 0.01$. An interpretation of these events show that all of them exhibit hard breaking, speeding, or high acceleration in combination with high speeds. In addition, these events are time stamped which can then be investigated further to analyse when these events occur. This allows time of day to be a possible factor in when deviant events occur. Due to the anonymity of the dataset, GPS coordinates are not recorded with the telemetric events. However, in practice, GPS is included and can be taken into account where exactly these deviant events occur. In summary, deviation analysis provides a tool for safety investigators to analyse where and when deviation events occur. For the shareholders of CC, this allows decision makers to plan for the necessary changes in their routes and training to mitigate risk.

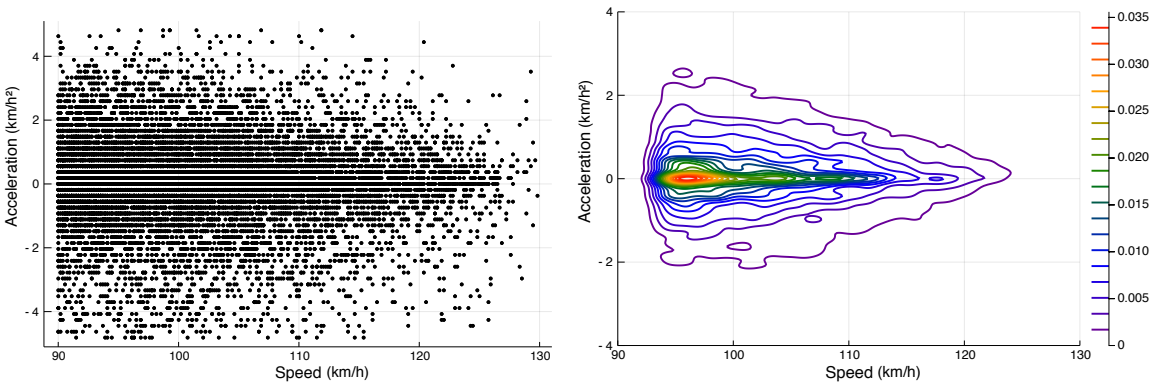

Figure 4.1: A visual representation of the construction of the highway VA for driver id 89675 (speeds are in km/h).

4.2 ETH Synthetic Dataset

The publically available ETH Synthetic dataset is generated from a simulation machine provided by Dr. Mario Wüthrich at ETH Zurich. The machine is a neural network that has been trained on a real VA heat map dataset for parking speeds. The machine has the ability to generate heat maps from up to four labelled groups for parking speeds only. Figure 4.3 shows the average VA heat map from all four groups generated from 1000 observations in each group. It is noted that these labels are the result of the analysis done in Wüthrich (2017) and, therefore, do not necessarily correspond to true labels. This section is split into two types of analysis. The first is a type of simulation study where the labels provided by the machine are used to assess CA between the MBI model and K-means. In the second analysis, the labels provided by the machine are disregarded, and the analysis is performed in a clustering setting.

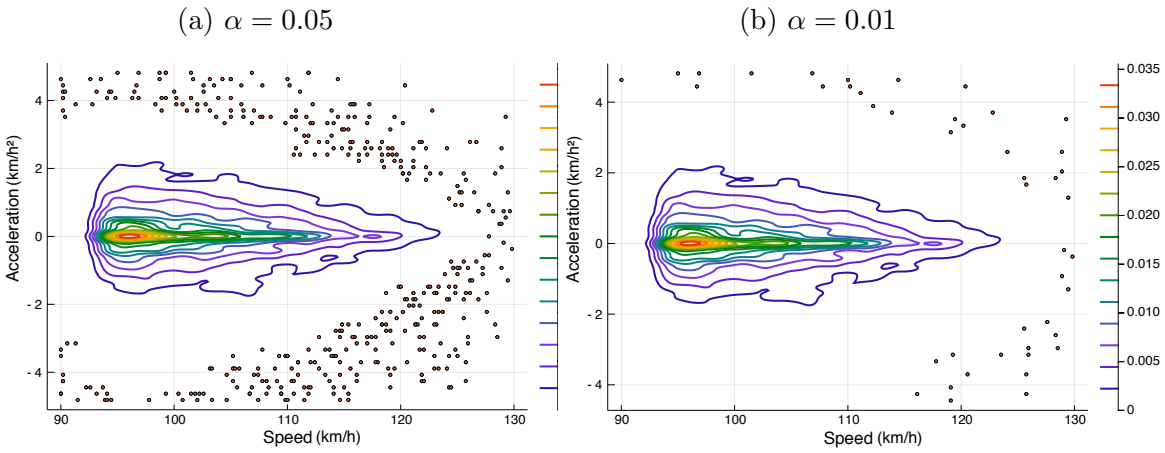

Figure 4.2: α-level deviation events overlayed on a VA heat map for driver 89675 (speeds are in km/h).

4.2.1 Class Agreement Comparison

This section outlines analysis comparing CAs between the ETH dataset and the MBI model. Machine labels are taken into account to assess comparison between MBI and K-means. Within this setting, 1000 observations were generated for groups $G = 1, \ldots, 4$ for a total of $N = 4000$. All possible combinations of groups were considered to compare the MBI and the standard K-means approach. The settings for possible combinations of $2, \ldots, 4$ groups are outlined in column one of Tables 4.1 and 4.2. For example, setting (1,2,3) is a dataset with a combination of groups $1, 2, 3$ containing 1000 observations sampled from each group (total $N = 3000$). Due to the numerical instability for estimating the MBI model, the standardization of matrix variate data is necessary (Gallaugher and McNicholas, 2018a). The traditional approach to standardize data is to simply scale the data (each entry is subtracted by the sample mean and divided by the sample standard deviation within each matrix) and multiply by a constant number. However, consider an additional method for

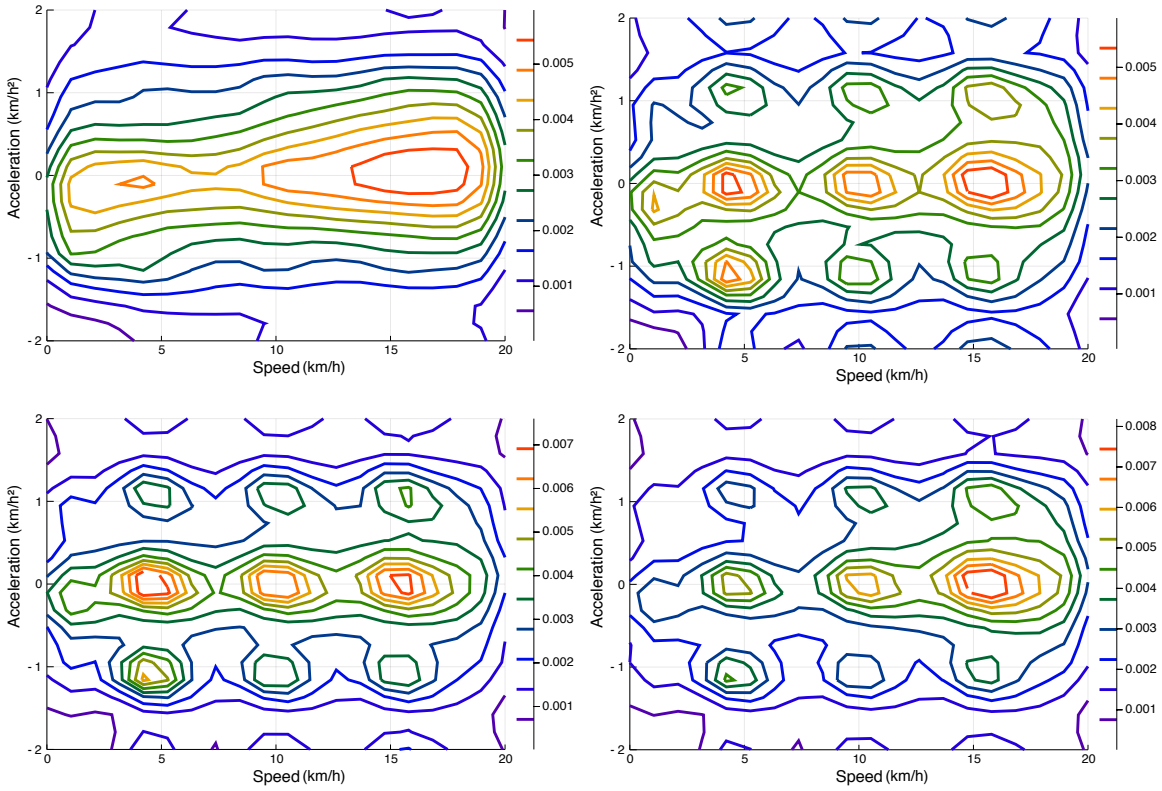

Figure 4.3: Average parking VA heat maps from machine labelled groups $1, \ldots 4$ (speed in km/h).

standardization. Since each heat map contains matrix entries within $[0, 1)$; the logit function which maps from $[0, 1) \to (-\infty, \infty) =: \mathbb{R}$ is appropriate (Ashton, 1972). This allows each entry of the observation to be in the domain of the matrix variate normal distribution. Neither of the transformations have much impact on the data itself as they are one-to-one functions. These standardizations are used to stabilize the estimation procedure of the model. The logit method applied to this data has better model BIC values compared to the scaled method. The values are compared in Tables 4.1 and 4.2, respectively. The number of iterations it takes to estimate parameters within the EM algorithm is less for the logit method when compared to the scaling

method. It is noted once again, that the labels are not necessarily reflective of true labels and are more used as a guide. These labels are provided by the maintainer of the dataset from his own analysis (Wüthrich, 2017). The best results are shown in bold for each setting. The MBI model has greater CA than K-means for several of the settings. Table 4.1 shows greater CA of the MBI under most of the mixture settings for scaled data but not under the logit standardized data as reported in Table 4.2. The best MBI for each setting is selected via BIC, which is reported in the second column of Tables 4.1 and 4.2.

Table 4.1: Model summary for scaled data between MBI and K-Means on CA and ARI.

Mixture	MBI			K-Means	
Setting	BIC Max	CA (%)	ARI	CA (%)	ARI
(1,2)	-8,321,290	**91.15**	**0.677**	90.25	0.649
(1,3)	-8,109,690	**99.65**	**0.986**	98.30	0.933
(1,4)	-8,040,320	**98.10**	**0.925**	94.95	0.808
(2,3)	-8,158,130	84.75	0.482	**86.45**	**0.531**
(2,4)	-8,140,840	**90.60**	**0.659**	89.75	0.631
(3,4)	-7,866,430	**91.00**	**0.672**	89.40	0.620
(1,2,3)	-12,279,600	80.00	0.517	**84.53**	**0.591**
(1,2,4)	-12,359,300	**87.23**	**0.654**	85.40	0.607
(1,3,4)	-12,038,600	**92.40**	**0.790**	85.43	0.615
(2,3,4)	-12,112,400	**80.70**	**0.505**	80.60	0.503
(1,2,3,4)	-16,080,100	78.67	0.525	**79.15**	**0.522**

In summary, all tables show how analyses differ under both methods for various standardizations. Logit standardization has an impact in the CA ARI, and BIC for the MBI model. In addition, the K-means approach suffers greatly under this standardization. The original analysis by Wüthrich (2017) are labels resultant from using K-means. As this simulation study suggests, the CA is difficult to regain K-means labels for matrix variate data even when using K-means. This leads to the

Table 4.2: Model summary for logit standardized data between MBI and K-Means on CA and ARI.

Mixture	MBI			K-Means	
Setting	BIC Max	CA (%)	ARI	CA (%)	ARI
(1,2)	-304,192	**94.10**	**0.777**	89.10	0.611
(1,3)	-524,070	**97.95**	**0.919**	84.72	0.481
(1,4)	-412,970	**97.22**	**0.891**	87.15	0.552
(2,3)	-616,595	82.95	0.434	**84.93**	**0.487**
(2,4)	-691,203	84.80	0.484	**88.42**	**0.600**
(3,4)	-640,034	84.80	0.484	**88.45**	**0.590**
(1,2,3)	-348,144	81.71	0.561	**82.93**	**0.550**
(1,2,4)	-721,927	79.52	0.531	**84.17**	**0.575**
(1,3,4)	-748,776	**90.63**	**0.748**	57.27	0.307
(2,3,4)	-724,082	**78.57**	**0.460**	75.81	0.415
(1,2,3,4)	-648,698	**60.82**	**0.376**	58.13	0.330

conclusion that the class membership according to Wüthrich (2017) is not stable. In some cases, the MBI model diverges from the original approach. However, despite the ambiguity of the labels, the analysis shows the MBI model can reliably recover some of the original analysis done by Wüthrich (2017).

4.2.2 Clustering Setting

Within the clustering setting, machine labels are disregarded and the entire dataset is taken to have an unknown number of groups. The MBI models are estimated with a combination of the following schemes: $G = 1, \ldots, 7$, $q = 1, \ldots, 12$, $r = 1, \ldots, 12$. The best model was selected according to BIC with a value of $-632,509$, and $q = 9, r = 11$ as the number of row and column factors, respectively. The estimated mixing proportions are reported to be $\boldsymbol{\pi} = (0.314, 0.686)$. Table 4.3 shows that the majority of groups E-G2, E-G3 and E-G4 are placed within group M-G2 by the MBI model. Furthermore, group M-G1 of the MBI model places the majority of observations in

E-G1 from the ETH analysis, and very few observations in other groups. This result is further confirmed when viewing the average heat maps of each group. Figure 4.4 shows the average heat maps from the estimated groups. Group M-G1's heat map appears to be similar to E-G1 from the given labels in Figure 4.3. The remaining group M-G2 in Figure 4.4 has a heatmap which resembles a combination of all other E-G2, E-G3 and E-G4 from Figure 4.3. This analysis implies that the labels given by the original author may not be four separate groups. The model selection via the BIC criterion implies that under MBI, $G = 2$ gives the best result. Both the MBI and the K-means approach do not validate with absolute certainty the presence of underlying populations. All this considered, this analysis should not be taken as absolute evidence that there only exists exactly two groups.

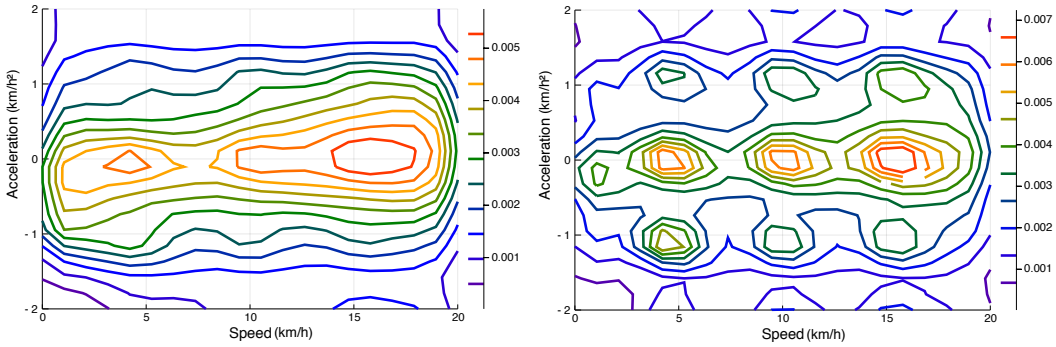

Figure 4.4: Average parking VA heat maps for groups 1 and 2 (speed in km/h).

Table 4.3: Classification agreement between ETH labels (E-G#) and the best MBI model (M-G#).

	E-G1	E-G2	E-G3	E-G4
M-G1	968	276	4	8
M-G2	32	724	996	992

4.3 ARC Study Dataset

The ARC study open source dataset contains a series of vehicle performance measurements over a period of 24 hours on $1,651$ vehicles in the United States (Transportation Secure Data Center, 2015). The study records a large variety of information pertaining to vehicle efficiency in driving cycles at every second. However, only the speed and acceleration records for every driver are taken into account. Due to computational limitations, only highway driving will be considered for the analysis as the dataset is extremely large. Highway driving, according to the authors of the dataset, are vehicle speeds consisting of anywhere between $[55, V_{\max})$ mph. After taking this into account, only $1,523$ drivers exhibited highway driving over the 24 hour period. Each driver's speed and acceleration data was processed and reduced down to their respective VA heatmaps. Each VA heatmap is a 24×24 matrix to which has been logit standardized for numerical stability as previously aforementioned. The MBI model is estimated according to $G = 1, \ldots, 4$ groups, and $q = r = 1, \ldots, 12$ latent factors for row and column respectively. According to the BIC, the best MBI model consists of $G = 3, q = 10, r = 11$, $\boldsymbol{\pi} = (0.18, 0.23, 0.59)$, with a BIC of $-3,023,861$.

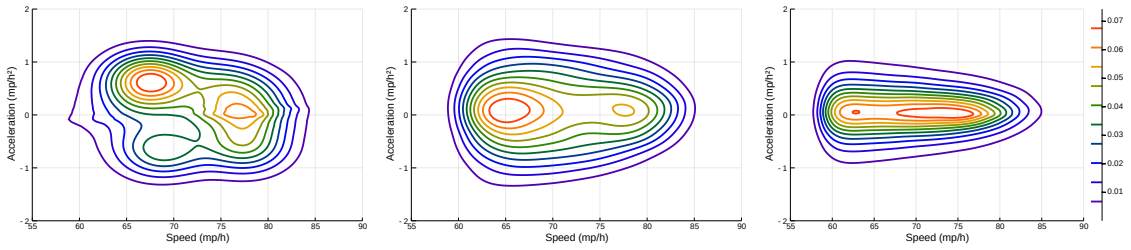

Figure 4.5: Average VA heat maps for Groups $1, 2,$ and 3 respectively.

Figure 4.5 shows the mean VA heat map for each group. The first group shows the widest heat map along acceleration indicating that observations in this group

have aggressive highway driving habits. In addition, the first group has asymmetric accelerations, indicating that these drivers have a preference for higher acceleration. The second group shows a smoother VA heat map average which resembles several of the known heat maps in actuarial literature (Wüthrich, 2017). These drivers have a preference for lower speeds but still contain the same variability with acceleration as the last group. The third group shows the safest average VA heat map with the smallest variability of acceleration. In summary, the ARC study dataset has been classified into three distinct groups with meaningful interpretations pertaining to risk. From an insurer's point of view, a premium discount model can be created around classifying drivers into one of the three groups.

Chapter 5

Conclusions and Future Work

This work expands on the current literature of telematics by using models and methods adapted to the problem at hand. The first extension is the use of KDE. The KDE method was proven to converge faster to the true density, and deal with the discreteness of telemetry by using smooth kernels. The second extension to telematics allowed for the detection of deviant events. CC was searching for a way to detect when their drivers were deviating from normal behaviour. The third extension to telematics is the clustering of driving types. Previous work on clustering VA heat maps involved the use of ambiguous dissimilarity functions (Wüthrich, 2017). The MBI approach uses BIC to select for both the model and the number of groups. Section 4.3 shows the segmentation of drivers into one of three types of driving behaviour using the MBI approach. All methods utilized in this work embody two characteristics. The first is the ease in the interpretation of results. Methods such as MBI, and deviation detection have natural explainations pertaining to driving behaviour. The second characteristic is the reduction of dimensionality. The VA heat map allows for the comparison of any sized collection of telemetries. For example, if two drivers have

telemetries collected over two different lengths of time, then both of them can be compared via VA heat maps (provided that the sample of telemetric data is sufficiently large).

Future work entails three possible avenues. The first is to reliably capture the same behaviour of VA heat maps but with increased lag. Interpolation of $\boldsymbol{v}_i$ can be used to estimate speeds at various time points with a reduced size sample size of telemetries. The second is to model frequency and severity of auto-mobile claims using these VA heat maps. VA heat maps can be used as covariates in a generalized linear model for modelling the frequency and severity of claims. Current literature uses the PPCA approach which can be expanded upon (Gao *et al.*, 2019). Finally, the assumption of normality is fairly strong for classifying matrix variate data. Skewed distributions may lead to better classification performance with matrix variate data and should be considered in the future (e.g. Gallaugher and McNicholas, 2018b).

Appendix A

Comparison of Estimators

A.1 Comparing KDE and Histogram

Lemma A.1 The kernel density estimator converges faster to the true density than a histogram.

Proof. Let $\hat{p}_M(x)$ be a histogram density estimator parametrized by bin length M. The expectation of this estimator is written as

$$\begin{aligned}
\mathbb{E}\left[\hat{p}_M(x)\right] &= MP(x_i \in R_m) \\
&= M \int_{\frac{m-1}{M}}^{\frac{m}{M}} p(u)du = M\left(F\left(\frac{m}{M}\right) - F\left(\frac{m-1}{M}\right)\right) \\
&= \frac{\left(F\left(\frac{m}{M}\right) - F\left(\frac{m-1}{M}\right)\right)}{1/M} = \frac{\left(F\left(\frac{m}{M}\right) - F\left(\frac{m-1}{M}\right)\right)}{\frac{m}{M} - \frac{m-1}{M}} \\
&= p(x^*), \quad x \in R_m.
\end{aligned}$$

By the mean value theorem,

$$\frac{p(x^*) - p(x+\delta)}{x^* - x} = p^{'}(x^{**}) \text{ for some } x^{**} \in \left[\frac{m-1}{M}, \frac{m}{M}\right].$$

Therefore, the bias of $\hat{p}_M(x)$ is calculated as

$$\begin{aligned}
\text{Bias}\left(\hat{p}_M(x)\right) &= \mathbb{E}\left[\hat{p}_M(x)\right] - p(x) \\
&= p(x*) - p(x) \\
&= p^{'}(x^{**})(x^* - x) \\
&\leq |p^{'}(x^{**})||x^* - x| \\
&\leq \frac{L}{M}.
\end{aligned}$$

The last line is derived from the fact that each line segment is bounded by $\frac{1}{M}$. By definition, the density is also bounded by L and thus the bias is bounded by $\frac{L}{M}$. M is defined to be the number of bins used. Therefore, as M increases, the bias decreases. The variance of the histogram density estimator is calculated as

$$\begin{aligned}
\mathbb{V}\text{ar}\left(\hat{p}_M(x)\right) &= M^2 \mathbb{V}\text{ar}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(\boldsymbol{x}_i \in R_m)\right) \\
&= M^2 \frac{P(x_i \in R_m)(1 - P(x_i \in R_m))}{n} \\
&= M^2 \frac{\frac{p(x^*)}{M}\left(1 - \frac{p(x^*)}{M}\right)}{n} \\
&= M\frac{p(x^*)}{n} + \frac{p^2(x^*)}{n}.
\end{aligned}$$

Therefore, the upper bound of the MSE is established as

$$\text{MSE}\left(\hat{p}_M(x)\right) \leq \frac{L^2}{M^2} + M\frac{p(x^*)}{n} + \frac{p^2(x^*)}{n}$$

The minimization of MSE is straight forward yielding

$$M_{opt} = \left(\frac{nL^2}{p(x^*)}\right)^{\frac{1}{3}}.$$

Now, consider a kernel density estimator $\hat{p}_h(x_0)$ for some fixed point x_0 and bandwidth h. The bias is calculated as

$$\begin{aligned}
\mathbb{E}\left[\hat{p}_h(x_0)\right] - p(x_0) =& \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x_i - x_0}{h}\right)\right] - p(x_0) \\
=& \frac{1}{h}\mathbb{E}\left[K\left(\frac{x_i - x_0}{h}\right)\right] - p(x_0) \\
=& \frac{1}{h}\int_R K\left(\frac{x - x_0}{h}\right)p(x)dx - p(x_0).
\end{aligned}$$

Performing a change of variable $y = \frac{x-x_o}{h}, \quad dy = dx/h,$

$$\frac{1}{h}\int_R K\left(\frac{x - x_0}{h}\right)p(x)dx - p(x_0) = \int_R K(y)p(x_o + hy)dy - p(x_o).$$

Now, consider the Taylor expansion of $p(x_o + hy)$ as

$$p(x_o + hy) = p(x_o) - hyp^{'}(x_o) + \frac{1}{2}h^2y^2p^{''}(x_o) + \mathcal{O}(h^2).$$

Substituting the form of the Taylor expansion into the expectation yields,

$$\begin{aligned}
\mathbb{E}\left[\hat{p}_h(x_0)\right] - p(x_o) &= \int_R K(y)p(x_o + hy)dy - p(x_o) \\
&= \int_R K(y)\left[p(x_o) - hyp^{'}(x_o) + \frac{1}{2}h^2y^2p^{''}(x_o) + \mathcal{O}(h^2)\right]dy - p(x_o) \\
&= p(x_o)\left(\int_R K(y)dy\right) - hp^{'}(x_o)\left(\int_R K(y)ydy\right) \\
&+ \frac{1}{2}h^2p^{''}(x_o)\left(\int_R K(y)y^2dy\right) + \mathcal{O}(h^2) - p(x_o) \\
&= p(x_o) - 0 + \frac{1}{2}h^2p^{''}(x_o)\left(\int_R K(y)y^2dy\right) + \mathcal{O}(h^2) - p(x_o) \\
&= \frac{1}{2}h^2p^{''}(x_o)\left(\int_R K(y)y^2dy\right) + \mathcal{O}(h^2) \\
&= \frac{1}{2}h^2p^{''}(x_o)\mu_K + \mathcal{O}(h^2).
\end{aligned}$$

Therefore, the bias of a kernel density estimator is given by

$$\text{Bias}\left(\hat{p}_h(x)\right) = \frac{1}{2}h^2p^{''}(x_o)\mu_K + \mathcal{O}(h^2), \quad \mu_K = \int_R K(y)y^2dy.$$

Now, consider the variance of a kernel density estimator,

$$\begin{aligned}
\mathbb{V}\text{ar}(\hat{p}_h(x)) &= \mathbb{V}\text{ar}\left(\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_i - x_o}{h}\right)\right) \\
&= \frac{1}{nh^2}\mathbb{V}\text{ar}\left(K\left(\frac{x_i - x_o}{h}\right)\right) \\
&\leq \frac{1}{nh^2}\mathbb{E}\left(K^2\left(\frac{x_i - x_o}{h}\right)\right) \\
&= \frac{1}{nh^2}\int_R K^2\left(\frac{x - x_o}{h}\right)p(x)dx.
\end{aligned}$$

Performing a change of variable $y = \frac{x-x_o}{h}, \quad dy = dx/h$, and using the same Taylor

series expansion as before we have

$$\begin{aligned}
&= \frac{1}{nh^2}\int_R K^2(y)p(x_o+hy)dy \\
&= \frac{1}{nh^2}\int_R K^2(y)\left[p(x_o)+hyp^{'}(x_o)+\mathcal{O}(h)\right]dy \\
&= \frac{1}{nh^2}p(x_o)\int_R K^2(y)dy + 0 + \mathcal{O}\left(\frac{1}{nh}\right).
\end{aligned}$$

Therefore, the variance of a kernel density estimator is

$$\mathbb{V}\text{ar}(\hat{p}_h(x)) = \frac{1}{nh^2}p(x_o)\sigma_K + \mathcal{O}\left(\frac{1}{nh}\right), \quad \sigma_K = \int_R K^2(y)dy$$

Using the preceding results above, the MSE of KDE is given by

$$\begin{aligned}
\text{MSE}(\hat{p}_h(x)) &= \text{Bias}^2(\hat{p}_h(x)) + \mathbb{V}\text{ar}(\hat{p}_h(x)) \\
&= h^4|p^{''}(x_o)|^2\mu_K^2 + \frac{1}{nh^2}p(x_o)\sigma_K + \mathcal{O}\left(\frac{1}{nh}\right)
\end{aligned}$$

Supposing that $h \to 0$ and $n \to \infty$, the terms of interest are

$$h^4|p^{''}(x_o)|^2\mu_K^2 + \frac{1}{nh^2}p(x_o)\sigma_K.$$

Under these asymptotic conditions, the optimal smoothing bandwidth is

$$h_{opt}(x_o) = \left(\frac{4}{n}\frac{p(x_o)}{|p''(x_o)|^2}\frac{\sigma_K^2}{\mu_K^2}\right)^{\frac{1}{5}}$$

Comparing the optimal parameters M_{opt} and h_{opt} for the two NDE methods the convergence for MSE $\left(\hat{p}_{M_{opt}}(x)\right)$ is $\mathcal{O}\left(n^{-\frac{2}{3}}\right)$ while the convergence for MSE $\left(\hat{p}_{h_{opt}}(x)\right)$

is $\mathcal{O}\left(n^{-\frac{4}{5}}\right)$. Therefore, the KDE method converges faster to the true probability density function $p(x)$. $\square$

Bibliography

Aitken, A. (1926). A series formula for the roots of algebraic and transcendental equations. *Proceedings of the Royal Society of Edinburgh*, **45**(1), 14–22.

Ashton, W. D. (1972). *The logit transformation: with special reference to its uses in bioassay*. Griffin London.

Azzoni, P., Moro, D., and Rizzoni, G. (1998). Time-frequency signal analysis of the acoustic emission of Formula 1 engines. In *Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis (Cat. No.98TH8380)*, pages 441–444.

Bellman, R. (1966). Dynamic programming. *Science*, **153**(3731), 34–37.

Chacón, J. E., Duong, T., and Wand, M. P. (2011). Asymptotics for general multivariate kernel density derivative estimators. *Statistica Sinica*, **21**(2), 807–840.

Chen, Y.-C. (2017). A tutorial on kernel density estimation and recent advances.

De Dutta, S. and Prasad, R. (2019). Security for smart grid in 5g and beyond networks. *Wireless Personal Communications*, **106**(1), 261–273.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM-algorithm. *Journal of the Royal Statistical Society B*, **39**, 1–38.

Di Marzio, M. and Lafratta, G. (1999). Reducing dimensionality effects on kernel density estimation: The bivariate Gaussian case. In M. Vichi and O. Opitz, editors, *Classification and Data Analysis*, pages 287–294, Berlin, Heidelberg. Springer Berlin Heidelberg.

Dutilleul, P. (1999). The MLE algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation*, **64**(2), 105–123.

Gallaugher, M. P. B. and McNicholas, P. D. (2018a). A Mixture of Matrix Variate Bilinear Factor Analyzers. in 'Proceedings of the Joint Statistical Meetings', American Statistical Association, Alexandria, VA. Preprint available as arXiv:1712.08664.

Gallaugher, M. P. B. and McNicholas, P. D. (2018b). Finite mixtures of skewed matrix variate distributions. *Pattern Recognition*, **80**, 83–93.

Gao, G., Meng, S., and Wüthrich, M. V. (2019). Claims frequency modeling using telematics car driving data. *Scandinavian Actuarial Journal*, **2019**(2), 143–162.

Gramacki, A. (2019). *Nonparametric Kernel Density Estimation and Its Computational Aspects*, volume 37 of *Studies in Big Data*. Springer.

Gupta, A. and Nagar, D. (1999). *Matrix Variate Distributions*. Monographs and Surveys in Pure and Applied Mathematics. Taylor & Francis.

Hayashi, C. (1998). What is data science? fundamental concepts and a heuristic example. In C. Hayashi, K. Yajima, H.-H. Bock, N. Ohsumi, Y. Tanaka, and

Y. Baba, editors, *Data Science, Classification, and Related Methods*, pages 40–51, Tokyo. Springer Japan.

Inui, M., Kawahara, Y., Goto, K., Yairi, T., and Machida, K. (2009). Adaptive limit checking for spacecraft telemetry data using kernel principal component analysis. *Transactions of The Japan Society for Aeronautical and Space Sciences, Space Technology Japan*, **7**.

Ioannidis, Y. (2003). The history of histograms (abridged). In *Proceedings of the 29th International Conference on Very Large Data Bases - Volume 29*, VLDB '03, pages 19–30. VLDB Endowment.

Mayo-Wells, W. J. (1963). The origins of space telemetry. *Technology and Culture*, **4**(4), 499–514.

McLachlan, G. and Krishnan, T. (2008). *The EM algorithm and extensions*. Wiley series in probability and statistics. Wiley, Hoboken, NJ, 2. ed edition.

McNicholas, P. D. (2010). Model-based classification using latent gaussian mixture models. *Journal of Statistical Planning and Inference*, **140**(5), 1175 – 1181.

McNicholas, P. D. (2016). *Mixture Model-Based Classification*. Boca Raton: Chapman and Hall.

McNicholas, P. D. (2019). Data science. *FACETS*, **4**(1), 131–135.

McNicholas, P. D. and Murphy, T. B. (2008). Parsimonious Gaussian mixture models. *Statistics and Computing*, **18**(3), 285–296.

Meng, X.-L. and Van Dyk, D. (1997). The EM Algorithm—an Old Folk-song Sung to a Fast New Tune. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **59**(3), 511–567.

Michalke, C. J. (1901). Means for operating electrical machines synchronously. US684579A.

Počuča, N. (2019). *TeleMap.jl: Driving style and deviation analysis for telemetric data*. Julia package version 0.1.0, https://github.com/nikpocuca/TeleMap.jl.

Schwarz, G. *et al.* (1978). Estimating the dimension of a model. *The annals of statistics*, **6**(2), 461–464.

Silverman, B. W. (1998). *Density estimation for statistics and data analysis*. Routledge.

Spearman, C., Jones, L., *et al.* (1950). Human ability. *American Psychological Association*.

Tiedeman, D. V. (1955). On the study of types. *Symposium on pattern analysis*.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, **61**(3), 611–622.

Tong, H. Y. and Hung, W. T. (2010). A framework for developing driving cycles with on road driving data. *Transport Reviews*, **30**(5), 589–615.

Transportation Secure Data Center (2015). Arc study dataset. *National Renewable Energy Laboratory*. 29-08-2019, www.nrel.gov/tsdc.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer Texts in Statistics. Springer New York.

Wolfe, J. (1965). A computer program for the maximum likelihood analysis of types. *USNPRA Technical Bulletin*, pages 15–65.

Wüthrich, M. V. (2017). Covariate selection from telematics car driving data. *European Actuarial Journal*, **7**(1), 89–108.

Zhao, Y. (2002). Telematics: safe and fun driving. *IEEE Intelligent systems*, **17**(1), 10–14.