ERASMUS UNIVERSITY ROTTERDAM

TOTTA data lab

MODEL BASED SEGMENTATION WITH CONTINUOUS VARIABLE SELECTION

Marnix Koops [1]

A thesis submitted in partial fulfillment for the degree of
M.Sc. Econometrics and Management Science
Department of Econometrics
Erasmus School of Economics
Erasmus University Rotterdam

Supervisor: dr. A.J. Koning
Co-reader: dr. A. Alfons

ABSTRACT We develop a model based segmentation approach that accommodates and exploits heterogeneous data. A finite mixture regression model is extended with variable selection abilities through likelihood penalization. This approach merges simultaneous estimation of a finite mixture model based on the EM algorithm with continuous variable selection into a single feasible procedure. The result is a flexible and powerful modeling algorithm that is able to deal with todays complexity and high-dimensionality of datasets. The model combines the value of mixture modeling and continuous feature selection resulting in a synergy of their advantages. The flexibility allows for finding groups of related observations while selecting the optimal subset of variables within these groups independently. First, the model is applied on a heterogeneous population of individuals. We succeed in identifying four segments containing customers with varying desirable characteristics and behavior making them valuable for the company. Two segments with less desirable properties are revealed. The results provide a foundation for a more efficient targeted marketing approach in comparison to treating the population as a whole. Second, we use a simulation to study performance and to display the advantages of this approach. The results indicate that extending a finite mixture model with variable selection abilities yields a powerful tool. Good performance is observed in terms of selecting the correct subset of variables to include while accurately estimating the effects of these variables. The model excels in high-dimensional settings where a relatively large amount of variables are of interest.

KEYWORDS finite mixture model; variable selection; penalization; elastic net; model based clustering; segmentation

1 INTRODUCTION

Currently, data is accumulated at unprecedented rates. Access to detailed information of customers and their behavior is not an exception. As a result, the concept of customer relationship management has evolved into an important part of business strategies. This allows companies to optimize the way they analyse and target their customers. Acquiring and retaining high valued customers should be a top priority for any firm. It is often the case that retaining is more attractive than acquiring new customers in financial terms. The relationship with a customer can be seen as a capital asset which requires appropriate management. Increasing attention is given to an individual specific approach to develop and maintain long-term relationships that are beneficial for both the company and the customer. This view differs from a more traditional approach, which looks at

[1] email: marnixkoops@gmail.com, student ID: 432409

single transactions and emphasizes short-term profit. Ultimately, the goal of marketing is understanding customer needs and to provide appropriate services but not all customers are identical. Comprehension of the customer-base is essential to support choices in differentiation across customers. This research provides data-driven support to justify decisions for a targeted marketing approach.

We employ the data of medium-sized health insurance company in the Netherlands but the principles in this work hold for any company with sufficient data regarding their customer-base. Proper relationship management is of great importance and targeted marketing can be a valuable tool to improve business. This is especially true in the field of health insurance. If a customer is satisfied with the services offered by their provider, it is not unusual that he or she remains a loyal customer for a long period of time. One may argue that every customer is of potential worth and should receive the best treatment possible in any case. However, like in almost any market or population, a vast number of heterogeneity is present. For one thing, a numerous number of customer characteristics are involved. Secondly, the behavior of these respective customers can vary significantly. For example in terms of claim frequency and monetary amount. Thirdly, the health insurance system itself is also a cause of variation in individuals. This is due to the possibility of many product combinations with multiple options and coverages.

The Dutch health insurance system has been relatively stable over the last decade. The basic insurance covers common medical care and can be extended by countless additional modules and options such as extra coverage for dental treatments or foreign countries. This allows for many unique combinations of packages and modules. The system can be summarized as follows. Every Dutch citizen is required by law to have a basic level health insurance. Correspondingly, companies are not allowed to refuse anyone requesting a basic level insurance as of January 2006. However, the obligation to insure anyone does not hold for additional packages and modules that extend insurance beyond the basic level.

A combination of the three above-mentioned sources of heterogeneity implicate that customers are not all identical. Meaning they can be structurally different. Accordingly, they should also not be viewed or managed equally. Customers are diverse which implies it is not efficient to regard the entire population as an aggregate (Allenby and Rossi, 1998). To clarify, customers may have different service needs and wishes or could be more attractive than others from the viewpoint of the company. For instance, it can be argued that a loyal individual with a positive financial balance is a more desirable customer than an individual with a negative monetary value. In addition, targeted marketing is a costly and sometimes complex venture while an individual can switch health insurance provider hassle-free at the end of each year. Therefore, spreading resources and attention evenly across the customer-base is not an efficient strategy. It seems desirable and smart to differentiate the resources allocated to specific customers accordingly. Especially considering the current information environment where a large number of data on customer characteristics and behavior is available on an individual level. In order to support or justify these differentiations, we first need to gain understanding in our customers and their behavior. For that reason, the main focus of this research is capturing customer heterogeneity by modeling the structure of our population. The heterogeneity in our data is addressed by means of a a segmentation approach. Consequently, a segmented structure can be used to adequately create distinction between customers and a corresponding targeted marketing approach. Comprehension of the customer base can provide support to perform targeted marketing actions on specific segments. A more individual specific approach will most likely improve services and lead to higher customer satisfaction levels and loyalty. Another opportunity is using this information to develop strategic marketing campaigns to obtain an optimal future customer base. This can be achieved by shifting the focus towards attracting new customers that fit into an identified desirable profile. To gain insight and achieve distinction in our population we will exploit the differences and similarities in our customer data by means of a segmentation.

A segmentation addresses heterogeneity by assigning observations into groups. The goal is to find a solution where observations are relatively similar within a group but different across groups. A segmentation can be interpreted as a conceptual model of how one wishes to view a market (Wedel

and Kamakura, 2012). Insight into an underlying structure allows to differentiate between customers and possibly identify groups with higher value or desirability for the company. Numerous customer valuation approaches exist but the majority of them predominantly focus on financial transactions. In case of a non-contractual setting the lifetime or churn rate of a customer can also be taken into account, an example is the Customer Lifetime Value (CLV) model (Berger and Nasr, 1998). However, in this research multiple other characteristics which are not easily expressed in a monetary value are of interest. Moreover, we are dealing with a contractual setting in which active customers do not need to be identified.

To summarize, a segmentation of the customer base can identify subgroups present in the data. Market segmentation is an essential part of both marketing theory and application (Wedel and Kamakura, 2012). This information gives support to adjusts resource allocation across customers accordingly. Insight in the customers allows for targeted marketing actions and paves the way for an optimal customer relationship management approach. Some interesting possibilities are

- Targeted marketing programs to focus on retaining customers in desirable segments
- Invest more in relationships with valued existing customers
- Reveal desirable customer profiles or types to develop strategic marketing campaigns for new customer acquisition

The central research question is formulated as

- Can we utilize the heterogeneity in our data to reveal and identify distinguishable customer groups?

The rest of this paper is structured in the following manner: Section 2 shortly introduces the data that is used and provides some summary statistics. In Section 3 the employed methodology is described. First, we review theory and cover the fundamentals of modeling heterogeneity. Secondly, we introduce a method to perform simultaneous estimation with continuous feature selection in a single algorithm. Consequently, the results are interpreted and discussed in section 4. Next, we study performance of the developed modeling approach by means of a simulation study in Section 5. Lastly, Section 6 concludes with the main findings of this research.

2 DATA

This section serves to introduce the data and shortly covers the preparation steps to allow for modeling. In this research we employ the data of a health insurance company. Table 1 gives an overview and summary statistics of the variables in the dataset.

2.1 DATA PREPARATION

In order to collect the data full access is granted to the server of the company containing a detailed SQL database consisting of over 30 tables. The majority of these tables contain dozens of variables and well over millions of observations. The variables are mixed meaning both numerical covariates as categorical factors are present. Numerous tables provide information on an individual-specific level regarding roughly 500,000 customers. Naturally, not all the available data is relevant or useful for our research goal. The potentially meaningful data ranges from demographic details such as age, gender and location to detailed information on the composition of insurance packages and modules.

Furthermore, numerous events and behaviors are extensively logged in the database such as singular transactions for prescriptions, hospital procedures like surgeries and other claims made by customers. These event databases are very extensive and can be used to construct important variables on an individual specific level such as claim amounts and frequencies. For instance, a single claim is often represented by multiple observations in the the table logging multiple steps and various information of the process. As a result, it is no exception these tables contain over tens of millions of observations. Hence, careful processing and aggregation is needed to correctly summarize the information in these tables and explicitly assign events and behavior to specific individual customers.

The structure of the SQL database is well organized, allowing for information from different tables to be linked or matched. Again, care is needed in this process as the tables have varying levels of detail, for example in terms of time-frames. Hence, not all data can be linked as given and correct preprocessing such as aggregation is required. Many of the final variables included in the dataset were not present in the database as is but required feature engineering to be constructed based on the available information. For example, the number of people on a single

insurance policy, *N_ON_POLICY* is not a variable currently present in the database but can be extracted from the data.

Alternative to the SQL database, several other data sources were available within the company. For instance, logs regarding the details of customer complaints and corresponding processing of these complaints. Another example is behavior on the website of the company such as log-in frequencies and requests. This information is collected with less consistency and lower standards. Further investigation of these alternative sources concludes that this data is currently of insufficient quality to include on an individual-specific level in this research.

In short; extensive data preparation steps were performed to collect, clean, analyse, pre-process and join relevant information from all available sources within the company. The preparation consisted of manipulation and merging with SQL queries and further processing in R (R Development Core Team, 2008). The resulting dataset has a structure consisting of observations belonging to individual customers.

Table 1 Overview and summary statistics of the variables included in the dataset

Variable	Description	Type	Min	Mean	Median	Max	St. Dev.
RELATION_NR	unique customer id	Categoric					
BALANCE	balance of individual in euro	Numeric	-1,014,638	1744	2710	23102	5648
AGE	age of customer	Numeric	1	41.11	46	105	22.70
SEX	gender of customer	Categoric			M		
N_YEARS	number of years insured	Numeric	0	7.97	11	11	3.66
MAIN	indicator of main insurance	Categoric	0		1	1	
ADDITIONAL	additional insurances	Categoric	0		1	2	
MODULE	extra modules, such as tooth	Categoric	0		0	1	
FOREIGN	indicator of foreign coverage	Categoric	0		0	1	
BRAND	brand of insurance package	Categoric			Brand 1		
TAKER	indicator of insurance taker	Categoric	0	0.49	0	1	
N_ON_POLICY	number of people on the policy	Numeric	1	2.77	3	13	1.45
VOL_EXCESS	voluntary deductible excess	Numeric	0	62	0	900	159
N_ON_COLLECT	number of people on collectivity	Numeric	1	8033	987	35933	9207
IND_COLLECT	indicator of individual collectivity	Numeric	0		1	1	
$REGION_{GGZ}$	mental care settlement region	Categoric	0		5	10	2.98
$REGION_{VV}$	nurse and care settlement region	Categoric	0		0	5	1.93
PROVISION	payment provision amount	Numeric	0	29.61	22.56	1278	36.4
PAYMENT_TERM	payment term in months	Numeric	1	3.8	1.00	12	4.62
N_CLAIMS	number of claims made	Numeric	0	17.01	15.00	118	9.55
N_LENIANCES	numbers of leniences received	Numeric	0	0	0	9	0.05
N_MONTHS	number of different months with claims	Numeric	0	8.98	9.00	12	2.38
N_CATEGORIES	number of different care categories	Numeric	0	3.89	4.00	13	1.76
N_CLAIM_MAX	max number of claims in one category	Numeric	0	8.59	8.00	36	3.06
N_NEGLECT	number of payment neglects	Numeric	0	0.09	0	27.00	0.77

Fields without interpretation are blank

3 METHODOLOGY

The following section introduces and describes the methods used in this work. First, the t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm is discussed which we apply for a quick exploratory data analysis. Second, the response variable is defined and its transformation is explained. Third, we cover theory concerned with modeling heterogeneity. For this purpose we introduce finite mixture modeling and discuss the estimation of these models. Fourth, we review approaches to determine the number of groups present in the data and options to perform model selection. Next, a selection of interesting developments in the area of feature selection is reviewed. Lastly, we introduce a model that combines mixture estimation and variable selection. The goal is to overcome common encountered difficulties regarding mixture modeling and feature selection. This procedure merges simultaneous estimation and continuous variable selection into a single powerful and flexible algorithm.

3.1 EXPLORATORY DATA ANALYSIS

Our interest lies in modeling the structural differences in our customers. Ideally, the heterogeneity in the dataset can be exploited to reveal a grouping or clustering structure. Consequently, this structure can be used to segment the population and assign customers to specific groups. As numerous variables are involved the patterns to be discovered can be complex. As a result, the structure in a multi-dimensional dataset is a rather abstract notion and difficult to grasp. To improve comprehension of high-dimensional datasets a multiple of techniques are available. One of the most popular multivariate statistical techniques is Principal Component Analysis (PCA) invented by Pearson (1901). PCA is also known as eigenvalue decomposition in the field of linear algebra. The modern application was formalized by Hotelling (1936). This procedure aims to reduce dimensionality by describing the dominant pattern of multiple dependent variables with a new set of orthogonal variables (Abdi and Williams, 2010). An attractive consequence is that multi-dimensional data can be visualized in a two-dimensional space based on the vectors with the highest eigenvectors. These two dimensions describe the largest amount of variation in the data.

Another dimensionality reduction technique is Multi-Dimensional Scaling (MDS). MDS aims to represent the dissimilarity between pairs of observations as distances in a low-dimensional space (Groenen et al., 2005). Similar observations are represented by a smaller distance while dissimilar observations are represented by a larger distance. This distance is commonly referred to as proximity. Hence, t-SNE and MDS are somewhat related. However, MDS is based on a dissimilarity matrix of the data in contrast to the original data itself such as in t-SNE and PCA. The goal is to achieve a representation of the data that depicts the similarity of observations by their proximity as good as possible. Both PCA and MDS focus on a preserving the global structure of the data by a faithful representation of the distanced between relatively separated points. Moreover, PCA and MDS are both restricted to linear relationships between the observations.

This limitation is overcome by a more recent technique by Maaten and Hinton called t-Distributed Stochastic Neighbor Embedding (t-SNE). Another option would be nonlinear PCA. T-SNE can often reduce dimensionality more effectively in a non-linear manner (Maaten and Hinton, 2008). The power of this algorithm is creating a two- or three-dimensional map from hundreds of thousands of variables to reveal a global pattern while retaining the local structure of the data. It has been shown that t-SNE yields better results in terms of visual interpretation on many different data sets compared to other popular non-parametric visualization techniques such as Sammon mapping, Isomapping and Locally Linear Embedding (Maaten and Hinton, 2008).

As t-SNE is less well known than PCA and MDS we cover the theory behind this technique. T-SNE is well suited to explore the structure in high-dimensional datasets. The main goal is to visualize a complex structure with different scales by a single faithful representation in lower-dimensional space. It is not a clustering algorithm as input features get lost in the process. Hence, it is mainly a tool for visualization and data exploration. However, the output can be used as input for classification or clustering algorithms. In this case, t-SNE could function as a preliminary data transformation comparable to an application with Principal Components Analysis. The core principle can be summarized as assigning each datapoint to a location in a two- or three-dimensional map (Maaten and Hinton, 2008).

The technique is based on Stochastic Neighbor Embedding (SNE) by Hinton and Roweis (2003) but t-SNE alleviates two issues in SNE. For one thing, a problem known as the crowding problem. Secondly, it overcomes the difficulty of optimizing the cost function in SNE. The crowding problem can be explained as the inability to simultaneously accommodate both nearby and moderately nearby datapoints in a faithful representation in the available area of a two-dimensional map. In this case nearby refers to observations that contain similar information. Meaning if observations that are close to observation i are accurately mapped, the moderately far away points from i are drawn together in the map. Hence, this crowds observations together and prevents forming of separated clusters. T-SNE alleviates both issues. Firstly, t-SNE uses a symmetric cost function which is easier to optimize (Cook et al., 2007). Secondly, the similarity of datapoints in low-dimensional space is computed with a Student-t distribution instead of a Gaussian distribution (Maaten and Hinton, 2008).

The first step is equal in SNE and t-SNE. It consists of converting high-dimensional Euclidean distances into probabilities p_{ij} that represent pairwise similarity of observations x_i and x_k in high-dimensional space p_{ij} given by

$$p_{ij} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq l} \exp\left(-\|x_k - x_l\|^2 / 2\sigma_i^2\right)} \tag{1}$$

where σ_i is the variance of the Gaussian located in datapoint x_i. In general, the optimal value of σ_i differs per datapoint as the density of the data varies. The probability p_{ii} is zero as we are looking at pairwise distance and for similar observations x_i and x_j the value of p_{ij} is high whereas more differing points result in a low p_{ij}. Now, instead of using a Gaussian distribution in the low-dimensional map, t-SNE employs a student t-distribution with one degree of freedom, also known as a Cauchy distribution. This means that the low-dimensional similarity of datapoint in t-SNE is represented by

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|y_k - y_l\|^2\right)^{-1}} \tag{2}$$

while SNE uses

$$u_{ij} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k \neq l} \exp\left(-\|y_k - y_l\|^2\right)}. \tag{3}$$

This difference in distribution for the low-dimensional probability is motivated by the fact that a t-distribution with a single degree of freedom has much heavier tails than a Gaussian distribution. As a result, this counters the crowding problem since moderately close datapoints can now be represented with a larger distance in the low-dimensional space compared to the SNE solution. Hence, moderately dissimilar datapoints are less clustered together, allowing for gaps to form between clusters of points. A more in depth discussion of the crowding problem is given in the work of Maaten and Hinton (2008). The choice of a t-distribution is motivated by the fact that it equals an infinite mixture of Gaussians with different variances. Hence, the two distributions are closely related. Another advantage is seen when comparing Equation 2 and 3. The evaluation of q_{ij} does not involve exponential terms in contrast to u_{ij}. As a result, t-SNE is computationally easier to solve then SNE.

If the transformation of the similarity between x_i and x_j in high-dimensional space to y_i and y_j in low-dimensional space is correct we have $p_{ij} = q_{ij}$. Hence, the algorithm minimizes the difference between p_{ij} and q_{ij}. This is achieved by employing the Kullback-Leibler (KL) divergence which is further discussed in Section 3.6 as the KL divergence is used again at a later point in this research. The KL divergence can be used as a measure of dissimilarity of two distributions. The sum of the Kullback-Leibler divergences over all datapoints i is minimized by the following cost function

$$C = \sum_i KL(P_i \| Q_i) = \sum_i \sum_i p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{4}$$

where P_i is the distribution of conditional probabilities over all datapoints given observation x_i and Q_i is the distribution of conditional probabilities over all other map points given observation y_i. The KL measure is further elaborated on in Equation 30. The second difference between t-SNE and other techniques such as SNE is the fact that it uses a symmetric cost function. Meaning that $p_{ij} = p_{ji}$ and $q_{ij} = q_{ji}$ for $\forall i, j$. The symmetric cost function results in a simpler gradient form which is faster to compute. This overcomes the difficulty of optimizing the cost function in SNE.

Next, to determine the optimal value for σ_i we look at the corresponding probability distribution P_i over

all other data points. This distribution has an entropy which increases with σ_i. A binary search is performed to select the sigma that results in a fixed perplexity which is defined as

$$Perplexity\,(P_i) = 2^{H(P_i)} \tag{5}$$

where $H\,(P_i)$ represents the Shannon entropy of P_i given by

$$H\,(P_i) = -\sum_j p(j|i) \cdot \log_2 p(j|i). \tag{6}$$

The perplexity can be interpreted as a measure of information just like the Shannon entropy. Perplexity controls the effective number of neighboring observations. A straightforward explanation from Maaten and Hinton is that a fair die with k sides has a perplexity of k. The value is similar to the k nearest neighbors variable used in other algorithms. The actual value is determined by the user and several options can be used to test performance. An usual range is 5 to 50 where a more complex dataset requires a higher perplexity value (Maaten and Hinton, 2008).

An important note is that t-SNE does not provide an interpretation of relative cluster sizes in terms of standard deviation. The algorithm adapts to expand dense groups while shrinking sparse clusters resulting in an equalized visualization of the spread (Wattenberg et al., 2016). This also holds for the interpretation of distances between separated clusters. This is an important difference with the interpretation of other dimensionality reduction techniques such as PCA and MDS. As mentioned, PCA and MDS focus on retaining a global structure by faithfully representing the distances between relatively separated observations in the data. In contrast, t-SNE focuses on the local structure by preserving the distances between similar observations in the data. The main advantage of this difference is that t-SNE manages to yield a more faithful representation in terms of visualization when applied on curved manifolds in contrast to linear techniques (Maaten and Hinton, 2008).

3.2 RESPONSE VARIABLE

Two main factors which are naturally expressed in a monetary value are present in the data. Namely, the premium that a customer pays for his or her health insurance and the amount of money that the customer claims from the provider. These two covariates are used to form the basis of a response variable. In addition, a third factor needs to be taken into account which is based on a nationwide health insurance regulation. This regulation is shortly explained below. All Dutch health insurance companies have two channels of income, the first one is already mentioned, which is the premium payed by individual customers. The second one is through a contribution from the 'Zorgverzekeringsfonds' which is health insurance fund that controls and divides governmental contributions to all Dutch health insurance providers. This flexible contribution is calculated depended on the customer base of a provider and can be subdivided into two parts. The contribution can be positive or negative for an individual as it balances out nation-wide. The first part is the number of customers having basic-level insurance. Insuring more people results in a higher contribution. The second part is more complex as it is a result of individual conditions. It can be generalized as follows. If a provider insures customers that are more likely to have high expenses, a higher contribution from the fund is given to this respective provider for insuring this individual. The regulation can be interpreted as a type of risk settlement for insuring this individual. A result of this construction is that people in need of expensive health care or medication are somewhat protected. As mentioned, health insurance companies are prohibited to deny customers requesting basic-level insurance but this does not hold for any additional coverage. Hence, the settlement or contribution prevents that no provider wants to offer any additional health insurance coverages to higher risk individuals.

To recap, three variables are used to construct a monetary balance. First, the premium a customer pays for his insurance. Second, the number a customer claims and third, a settlement for every individual. Now that we have defined the three parts we can formulate a response variable for each individual customer as

$$y_i = \sum_{v=1}^{V} P_{vi} - C_i + S_i \text{ for } i = 1, \ldots, N \tag{7}$$

where y_i represents the balance for individual i, P_{vi}

is the premium payed for insurance package v by individual i, C_i are the claims made by individual i and S_i is the settlement received for individual i. S_i can be positive as well as negative.

The resulting continuous response variable vector y contains balances for each individual i which can take a positive, negative or sometimes zero value. Further inspection reveals the response variable is heavy-tailed. A normal, or Gaussian, distribution is often preferred as this distribution is assumed in many statistical tests and applications. Several transformations exist to increase normality such as the well known Box-Cox transformation (Box and Cox, 1964). The Box-Cox power family is given by

$$\psi^{\mathrm{BC}}(y,\lambda) = \begin{cases} \log(y) & \lambda = 0 \\ \frac{y^{\lambda}-1}{\lambda} & \lambda \neq 0. \end{cases} \tag{8}$$

where y is the value to be transformed and λ is a transformation parameter.
This transformation is only valid for positive values of y. Our response variable is defined on the entire real line. As consequence, a Box-Cox transformation as given in Equation 8 is not defined due to the presence of zero and negative values. Naturally, the same holds for the simpler log-transformation.

Yeo and Johnson have introduced a new family of power transformations that share the desirable characteristics of the Box-Cox transformation without imposing restrictions on y (Yeo and Johnson, 2000). Zero and negative values can also be accommodated. This transformation aims to reduce excess skewness and kurtosis. In order to achieve more normality we apply this transformation on the response variable. This also decreases the large difference between the values of the response variable and the regression variables. The Yeo-Johnson power transformation family is defined as

$$\psi^{\mathrm{YJ}}(y,\lambda) = \begin{cases} \log(y+1) & \lambda = 0,\ y \geq 0 \\ \frac{(y+1)^{\lambda}-1}{\lambda} & \lambda \neq 0,\ y \geq 0 \\ -\log(-y+1) & \lambda = 2,\ y < 0 \\ \frac{-\left[(-y+1)^{2-\lambda}-1\right]}{2-\lambda} & \lambda \neq 2,\ y < 0. \end{cases} \tag{9}$$

If the value to be transformed is strictly positive, the Yeo-Johnson transformation is equal to the Box-Cox transformation of $y+1$. For strictly negative values the transformation is equal to the Box-Cox transformation of $-y+1$ with transformation parameter $2-\lambda$.

In our case we have both positive and negative values. As result, the transformation is a combination of these two (Weisberg, 2001). In addition the response variable is scaled by division with the standard error of y. We refrain from centering by subtracting the mean in this transformation. This preserves the sign of y which is desirable since the sign holds meaning in this case. To be more exact, positive balances remain positive and negative balances remain negative after transforming. Table 2 reports the skewness and kurtosis of the response variable before and after applying the transformation.

Table 2 Result of transformation on y

Measure	Original	Transformed
Min	-1014639	-29.42
Max	23102	17.10
Mean	1744	1.30
Median	2710	1.460
Skewness	2,690	0.163
Kurtosis	-27.39	10.36

3.3 FINITE MIXTURE MODELS

We will now discuss possibilities to model heterogeneous data. Heterogeneity is often considered by grouping similar observations into groups. When dealing with data of individuals this can be seen as a segmentation of the customer-base. This concept emerged in the late 1950s. In the early days, segmentations were often based on simple and common characteristics such as gender or age. Although the idea of segmentation appears simple, it is one of the most researched topics in marketing science in terms of scientific development and methodology (Wedel and Kamakura, 2012). A segmentation can be achieved by means of a finite mixture model, which is simply put a combination of several distributions. The first influential analysis based on a mixture model originates from 1894 where the biometrician Pearson fitted a two component mixture of normal densities (Pearson, 1894). Since then major advances have been made to accommodate the need for methods that can handle large and complex datasets.

Meanwhile, a surge in popularity of machine learning approaches also increased the application of cluster analysis techniques. These clustering methods used for segmentation are often heuristic in nature. A prime disadvantage of such methods is the lack of a sufficient statistical basis. These algorithms are in general based on some arbitrary measure of distance to determine the similarity of observations (Tuma and Decker, 2013). The specific choice of distance measure significantly impacts the results of the analysis. This is especially true when categorical variables are included in the analysis. In this case a preliminary transformation of the data is required to allow application, such as Gower's distance (Gower, 1966). Inference based on these heuristic approaches have lead to much discussion in terms of validity.

Finite mixture models alleviate some of the common issues associated with heuristic methods. They provide a model based approach for segmentation (Wedel and Kamakura, 2012). In order to exploit differences in the customers we require a flexible model combined with an inference method to interpret the results (Allenby and Rossi, 1998). Finite mixture models have been expanded in the 1990s with practices composed of linear regression models and generalized linear models (Wedel and DeSarbo, 1995). The practical application, potential and theoretical attention of mixture models has grown considerably since 1995 (McLachlan and Peel, 2004). This growth can be explained by the immense flexibility to model unknown distributions in a convenient manner and secondly by advances in computational power. In addition, finite mixture models are particularly useful to capture and describe some type of grouping structure present within a complex dataset. These models have seen utility in various fields such as astronomy, biology, genetics, medicine, economics, engineering and marketing (McLachlan and Peel, 2004). Mixture models can also be combined with machine learning algorithms. An interesting present-day application is the speech of Siri on Apple devices. The technology behind Siri's voice is called a deep mixture density network (MDN) which combines deep neural networks with Gaussian mixture models (Apple, 2017). In short, finite mixtures can be seen as a more elegant approach compared to heuristic methods and have obtained an important position in modern market segmentation applications (Wedel and Kamakura, 2012; McLachlan and Peel, 2004).

Whether the data is simple or complex, the principle of segmentation is similar. The fundamental idea is that a single distribution or model fails to sufficiently describe a collection of data due to the presence of heterogeneity. A finite mixture model is based on a mixture of multiple parametric distributions to describe the underlying structure of some data. In our case, we assume the entire population of customers contains unidentified subgroups. This heterogeneity is called latent, meaning it is unobserved. The groups within the population can be interpreted as a finite number of latent classes also referred to as segments or components (Muthén and Shedden, 1999). Failure to recognize the presence of subpopulations and account for heterogeneity results in misleading or incorrect inference. Finite mixture models provide an effective method to consort population heterogeneity and provide a flexible and powerful way to model univariate or multivariate data. Specifying the parametric distribution of the latent structure in the data is not required to perform estimation. This is a highly attractive feature as it prevents bias in parameter estimation as a result from potential misspecification. An interesting fact is that normal mixture models can be used to test the performance of estimators with their ability to capture deviation

from normality (McLachlan and Peel, 2004). Normal mixtures have helped in the development of robust estimators. For example the contaminated normal distribution proposed by Tukey where the density of a point is interpreted as a mixture of two normal distributions with different variances (Tukey, 1960). A more general incomplete contamination form is considered in the work of M-estimators by Huber et al. (1964). Finite mixtures are often labeled as a semi-parametric approach. Jordan and Xu describes them as an interesting niche between parametric and non-parametric. A parametric formulation of the mixture is determined whereas the number of components is allowed to vary which can be interpreted as non-parametric (Jordan and Xu, 1995). This description can be used to explain why a mixture model possesses the flexible properties of non-parametric approaches while retaining attractive analytical advantages of parametric approaches (McLachlan and Basford, 1988).

Finite mixture models can model the joint distribution of multiple variables, in contrast to non-parametric algorithms such as K-means or K-nearest neighbors. Although non-parametric methods are often fast and require no assumptions on the distribution of the data, there are some drawbacks associated with these methods. One cause of discussion is the fact that similarity between observations is based on a chosen distance measure. A finite mixture is based on a statistical model which requires to choose distribution. Yet, a result is that mixture models offer more extensive inference and interpretation possibilities. Uncertainty in the classification can be taken into account in contrast to non-parametric methods which result in hard grouping or classification. This means observations are assigned to components as if no certainty is involved in this membership. Often, this is a rough assumption as group memberships are in reality not fully certain. Moreover, the uncertainty in grouping may even be meaningful for interpretation of the cluster results. Furthermore, mixture models have the capability to handle groups with different sizes, correlation structures and overlapping of segments in contrast to many other techniques. On the contrary, non-parametric clustering techniques prefer groups of equal size and are not suited to handle overlapping segments due to hard classification. If an observation shares properties of multiple subgroups, this membership information is lost by hard clustering.

In this research we are interested in relating the response variable y with a set of explanatory features. DeSarbo and Cron introduced a methodology for cluster-wise linear regressions giving rise to finite mixture regression modeling (1988). Finite mixture regression models provides a flexible method to simultaneously estimate both group membership and separate regression functions to explain the response variable within each segment (Wedel and Kamakura, 2012). It has been proven that any continuous distribution can be estimate arbitrarily well by a finite mixture of Gaussian distributions (McLachlan and Peel, 2004; Lindsay, 1995). Consequently, a Gaussian or normal mixture regression constitutes the foundation of our model.

The density function of a general S-component finite mixture model can be formulated as

$$f(y|x,\Theta) = \sum_{s=1}^{S} \pi_s \cdot f(y|x,\theta_s), \tag{10}$$

where y is a vector of response variables as defined in Equation 7, x is a vector of regression variables given in Table 1, π_s is the prior probability of belonging to component s, each θ_s is a vector with component specific parameters for density f, and $\Theta = \{\theta_1, \ldots, \theta_p\}$ is a vector containing all parameters to specify the mixture. The prior probability π_s is also referred to as the mixing coefficient. The restrictions on the parameters are as follows. π_s is a probability, thus satisfying the follow conditions

$$\begin{aligned} &\sum_{s=1}^{S} \pi_s = 1, \\ &0 < \pi_s \leq 1 \; \forall s = 1, \ldots, S. \end{aligned} \tag{11}$$

For the component specific parameter vectors we have

$$\theta_s \neq \theta_k \; \forall s \neq k \text{ with } s,k \in \{1,\ldots,S\}. \tag{12}$$

Next, the group membership is the conditional probability of an observation belonging to segment s. This is also referred to as the posterior probability. We can compute this probability using Bayes' theorem as

$$z_{is} = \mathbb{P}(s|y_i,x_i,\Theta) = \frac{\pi_s \cdot f(y_i|x_i,\theta_s)}{\sum_{k=1}^{S} \pi_k \cdot f(y_i|x_i,\theta_k)}. \tag{13}$$

The corresponding log-likelihood function of the S-component mixture model is computed as

$$\begin{aligned}\mathcal{L}(\Theta) = \log f(y|x,\Theta) &= \log \prod_{i=1}^{N} f(y_i|x_i,\Theta) \\ &= \sum_{i=1}^{N} \log \sum_{s=1}^{S} \pi_s \cdot f(y_i|x_i,\theta_s)\end{aligned} \tag{14}$$

with corresponding maximum likelihood (ML) estimate

$$\begin{aligned}\widehat{\Theta}_{ML} &= \arg\max_{\Theta} \mathcal{L}(\Theta) \\ &= \arg\max_{\Theta} \left[\log f(y|x,\Theta)\right] \\ &= \arg\max_{\Theta} \left[\sum_{i=1}^{N} \log \sum_{s=1}^{S} \pi_s \cdot f(y_i|x_i,\theta_s)\right].\end{aligned} \tag{15}$$

In this work we use a finite mixture regression model with Gaussian distributed components such that

$$f(y_i|x_i,\Theta) = \sum_{s=1}^{S} \pi_s \cdot \frac{1}{\sqrt{2\pi}\sigma_s} \exp\left(\frac{(y_i - x_i^{\mathrm{T}}\beta_s)^2}{2\sigma_s^2}\right) \tag{16}$$

where every component s has an independent vector of regression coefficients β_s and variance σ_s^2.

3.4 ESTIMATION

As the parameters of the mixture in Equation are unknown they need to be estimated from the data. Estimation options include method of moments, maximum likelihood (ML) and Bayesian approaches (McLachlan and Peel, 2004). ML estimation can be done with numerical methods such as Newton-Raphson's algorithm. However, the likelihood function as given in Equation 15 can be difficult to solve and generally contains multiple local maxima. Numerical optimization methods often do not perform smoothly. Alternatively, a Bayesian approach based on Markov Chain Monte Carlo (MCMC) sampling can be used to estimate the parameters (Diebolt and Robert, 1994). The likelihood function can also be solved with the Expectation-Maximization (EM) algorithm by Dempster et al. (1977). The EM algorithm is an iterative hill-climbing procedure to estimate the parameters that maximize the log-likelihood function. It is a prevalent approach for problems associated with incomplete data caused by missing variables or unobserved heterogeneity (Dempster et al., 1977). Usefulness of the EM algorithm in finite mixture models is reported by McLachlan and Basford among many others (1988).

Solving Equation 15 to obtain the maximum likelihood estimates is a difficult problem. This problem can be approached by assuming that we are dealing with incomplete observations that originate from non-observed complete data. In other words, we assume that our observations originate from a finite number of groups. However, the group membership variable is not part of the available data. In order to estimate the parameters in the mixture we augment our incomplete data with a group membership variable Z yielding the complete data. This approach allows to define a complete data log-likelihood function as

$$\begin{aligned}\mathcal{L}_c(\Theta) &= \log f(y,Z|x,\Theta) \\ &= \sum_{i=1}^{N}\sum_{s=1}^{S} z_{is} \cdot \log\left[\pi_s \cdot f(y_i|x_i,\theta_s)\right].\end{aligned} \tag{17}$$

where the vector $Z = \{z_i,\ldots,z_N\}$ contains labels indicating group membership for every observation i. The complete likelihood function is also referred to as the classification likelihood in some cases.

Next, the EM algorithm is used to estimate the parameters by treating z_{is} as missing data. The algorithm can be subdivided into two steps. The Expectation-step and the Maximization-step. Every iteration provides updated parameter estimates $\widehat{\Theta}$. The procedure is stopped if a predefined convergence criterion is met. The E-step computes the expectation of the complete data log-likelihood conditional on the data and the current estimates $\widehat{\Theta}^{(t)}$ as

$$\mathbb{E}\left[\mathcal{L}_c(\Theta)\right] = \mathbb{E}\left[\log f(y,Z|x,\widehat{\Theta}^{(t)})\right]. \tag{18}$$

In this step the group memberships, also called posterior probabilities, are calculated based on the current parameter values using Equation 13 such that

$$z_{is} = \frac{\pi_s^{(t)} \cdot f(y_i|x_i^{(t)},\theta_s^{(t)})}{\sum_{k=1}^{S} \pi_k^{(t)} \cdot f(y_i|x_i^{(t)},\theta_k^{(t)})}. \tag{19}$$

Consequently, the M-step maximizes the expected value seen in Equation 18 with respect to Θ

$$\begin{aligned}\widehat{\Theta}^{(t+1)} &= \arg\max_{\Theta} \mathbb{E}\left[\mathcal{L}_c(\widehat{\Theta}^{(t)})\right] \\ &= \arg\max_{\Theta} \mathbb{E}\left[\log f(y,Z|x,\widehat{\Theta}^{(t)})\right] \\ &= \arg\max_{\Theta} \sum_{i=1}^{N}\sum_{s=1}^{S} z_{is} \cdot \log\left[\pi_s \cdot f(y_i|x_i,\theta_s^{(t)})\right].\end{aligned} \tag{20}$$

The estimation procedure described above can be summarized as follows. First, we formulate our problem as a missing data setup. Second, we iteratively estimate the parameters with the EM algorithm.

Data Setup

- Observed data: the observations as available (y_i, x_i)
- Missing data: the group membership information of each observation z_{is}
- Complete data: the observations supplemented with the group memberships

Following this setup allows the likelihood function to be maximized with the following algorithm.

Algorithm 1 EM Algorithm for a Finite Mixture Regression

1. Determine a set of initial parameter estimates Θ^{ini} that define the mixture to start the algorithm.
2. E-step: Estimate the posterior probabilities based on the current set of parameter estimates

$$z_{is} = \frac{\pi_s \cdot f(y_i|x_i, \theta_s)}{\sum_{k=1}^{S} \pi_k \cdot f(y_i|x_i, \theta_k)}. \tag{21}$$

   Derive the prior class probabilities as

$$\pi_s = \frac{1}{N} \sum_{i=1}^{N} z_{is}. \tag{22}$$

3. M-step: Update the parameter estimates using the current posterior probabilities

$$\arg\max_{\Theta} \sum_{i=1}^{N} \sum_{s=1}^{S} z_{is} \cdot \log\left[\pi_s \cdot f(y_i|x_i, \theta_s)\right]. \tag{23}$$

4. Evaluate the complete log-likelihood function

$$\mathcal{L}_c(\Theta) = \sum_{i=1}^{N} \sum_{s=1}^{S} z_{is} \cdot \log\left[\pi_s \cdot f(y_i|x_i, \theta_s)\right]. \tag{24}$$

5. Repeat steps 2 to 4 until a defined convergence criterion is met.

A potential issue of finite mixture models is identifiability. For consistent estimation of the parameters identifiability is a necessary condition (Hennig, 2000). In some cases different sets of parameter estimates can describe the same density function. The model is identifiable if one unique set of parameters is able to define the distribution. In terms of the model as introduced in Equation 10 we need that for any two parameters Θ and Θ^*

$$\begin{aligned} f(y|x, \Theta) &= f(y|x^*, \Theta^*) \\ \sum_{s=1}^{S} \pi_s \cdot f(y|x, \theta_s) &= \sum_{s=1}^{S^*} \pi_s^* \cdot f(y^*|x^*, \theta_s^*) \end{aligned} \tag{25}$$

implies $\Theta = \Theta^*$ and $S = S^*$. It has been proven that given some mild conditions many finite mixture models are identifiable, including Guassian types (Titterington et al., 1985).

3.5 COMPONENT SELECTION

A fundamental challenge in model selection is determining the number of components S used in the mixture. This problem is also referred to as order selection. In practice, the number is usually unknown beforehand and needs to be extracted from the data itself. Care is needed in selecting the number of components, too many groups may lead to over-fitting while too few may result in failure to capture the underlying structure of the data (Huang et al., 2013). Conventional tests based on the likelihood ratio do not apply as comparing nested models is not possible due to unknown S. For example the χ^2-statistic is not valid due to violation of regularity conditions (Titterington et al., 1985). Still, a wide range of options is available to perform model selection.

The following strategy is employed to determine the number of groups in our data. We fit the mixture model in a step-wise manner with an increasing number of components S. In addition, we consider the prior probability π_s to control the minimum number of observations in a group. A restriction on the prior allows for deletion of small components in the estimation process. In case the size of a group falls below the threshold the component can be removed from the model. This restriction conveniently counters overfitting while simultaneously avoiding problems in estimation. Components with little observations can lead to numerical instabilities in the EM algorithm.

Multivariate Gaussian mixture models are especially prone for this latter problem due to the estimation of full variance-covariance matrices for each component. A minimum sample size of 30 observations per component is shown to be sufficient (Garver et al., 2008) Consequently, the resulting fit of the models with varying group sizes are compared. Information criteria can be used to decide the optimal number of segments needed to describe the data. Many traditional information criteria can be generally formulated as

$$-2\mathcal{L}(\widehat{\Theta}) + \lambda \|\widehat{\Theta}\|_0 = -2\mathcal{L}(\widehat{\Theta}) + \lambda \sum_{j=1}^{p} |\widehat{\Theta}|^0 \quad (26)$$

where $\mathcal{L}(\widehat{\Theta})$ is the log-likelihood function, $\|\widehat{\Theta}\|_0$ represents the ℓ_0 "norm" which equals the number of non-zero variables in $\widehat{\Theta}$ and λ is a constant tuning parameter. λ controls the overall strength of the penalty and has restriction $\lambda \geq 0$. Strictly speaking the ℓ_0 "norm" is not an actual norm. In order for a function f to be a norm we need that $f(\alpha x) = |\alpha| f(x)$. Yet, this relation is not satisfied by the ℓ_0 "norm" since $||\alpha X||_0 \neq |\alpha| \, ||X||_0$.

Equation 26 can be used to derive some well known model selection criteria. The Akaike Information Criterion (AIC) is obtained by setting $\lambda = 2$. A modified AIC with an increased penalty on the number of variables called AIC-3 is the result of setting $\lambda = 3$ (Akaike, 1998). The Bayesian Information Criterion (BIC) by Schwartz is obtained by setting $\lambda = \log(N)$ (Schwarz et al., 1978). When the log-likelihood function $\mathcal{L}(\widehat{\Theta})$ is replaced by the regular likelihood function in Equation 26 the result is penalized least squares (Fan and Lv, 2010).

These criteria are all based on the likelihood of a model combined with a penalty for model complexity. However there are some subtle differences, BIC yields a higher penalty for complex models compared to AIC as we often have $\log(N) > 2$. Generally stated, AIC favors more complex models that might over-fit while BIC is more prone to select models that under-fit. Leroux et al. find both AIC and BIC do not underestimate the true number of components in a mixture, which is further covered in the simulation study later in this report 1992. Additionally, multiple simulation studies conclude AIC-3 performs well as a criterion in the general context of many model specifications and configurations including finite mixture regression models (Tuma and Decker, 2013). Alternative measures based on the classification likelihood function also exist such as the normalized entropy criterion (NEC) (Celeux and Soromenho, 1996) and the integrated classification likelihood criterion (ICL) (Biernacki et al., 2000).

An alternative to the deterministic methods discussed above is a stochastic approach such as Markov chain Monte Carlo (MCMC). We will not consider this approach as the computational load of MCMC is often too heavy for many applications such as pattern recognition (Figueiredo and Jain, 2002). In this work we report BIC, AIC, AIC-3 and ICL metrics to compare and evaluate models. However, our ultimate goal is to obtain a parsimonious and actionable segmentation while capturing and describing the structure sufficiently well. For this reason emphasis is put on the BIC value which in general favors a more parsimonious solution then the other used metrics.

In order to determine the number of components required to model the data we estimate models with a varying number of groups S. The selection is done in two stages to decrease computational intensity. First, we obtain results for 5 up to 30 components with a step-size of 5 after which the models are compared based on the introduced measures. This results in a rough indication of the number of components needed. In the second stage the information given by the first stage is used. We narrow the search grid and decrease the step-size to one to find the optimal number of groups. The estimation of every model specification is repeated 10 times to ensure stability, this holds for both stages. Following this approach yields a large number of models. The optimal solution of each repetition is kept as solution for that respective specification.

To further explore the structure of the determined components after estimation we consider two approaches. First, the ratio of the prior probability and the number of observations having a posterior probability larger than ε. Epsilon is set to a small number larger than zero such that $\varepsilon > 0$ and is interpreted as follows. When a probability is smaller than epsilon it is considered as zero as many observations are given a very small probability of belonging to a segment. This ratio can be interpreted as a measure of how well a component is separated from the other components based on the posterior probabilities. We formulate

this ratio as

$$\text{ratio} = \frac{\text{size}}{\#\{post > \varepsilon\}} \tag{27}$$

where *size* represents the number of observations assigned to this component based on the posterior probabilities. Third, $\#\{post > \varepsilon\}$ represents the number of observations with a posterior probability of belonging to this component larger than epsilon. This measure is bounded between 0 and 1. A value of 1 means perfect separation is achieved for the respective component. In contrast, a value closer to 0 indicates a larger amount of overlap in segments. Second, we use the Kullback-Leibler (KL) divergence measure which is introduced in the next section.

3.6 KULLBACK-LEIBLER DIVERGENCE

Kullback-Leibler divergence originates from 1951 and has its roots in the field of information theory Kullback and Leibler. This concept is sometimes also referred to as information gain or relative entropy (Kullback, 1997). Simply stated, the Kullback-Leibler divergence can be used as a measure of dissimilarity of two distributions. As starting point we take a fundamental concept in information theory called entropy. Entropy aims to quantify the amount of information present in a collection of data. The entropy H for a discrete probability distribution $p(x)$ is given by

$$H = -\sum_{i=1}^{N} p(x) \cdot \log p(x). \tag{28}$$

The continuous version of H is known as differential entropy (Cover and Thomas, 2012) defined as

$$h(P) = -\int P(x) \cdot \log P(x). \tag{29}$$

A small modification to Equation 29 yields the Kullback-Leibler divergence (Kullback, 1997). For two continuous probability distributions P and Q the KL divergence from Q to P is

$$\begin{aligned} D_{KL}(P\|Q) &= \int P(x) \cdot (\log P(x) - \log Q(x)) \\ &= \int P(x) \cdot \log\left(\frac{P(x)}{Q(x)}\right). \end{aligned} \tag{30}$$

This measure is also used in the t-SNE algorithm in Section 3.1 where it's purpose is to to preserve a local high-dimensional structure between two data-points while mapping them into a lower dimensional space. A divergence of zero would indicate the distributions are equal. KL divergence is often interpreted as a distance metric out of convenience. Theoretically this is incorrect as it is does not satisfy the triangle inequality and is asymmetric. Formulated in a more exact manner, this means for two distributions P and Q we can have

$$D_{KL}(P\|Q) \neq D_{KL}(Q\|P). \tag{31}$$

In case of two Gaussian distributions P and Q, such as two components of our finite mixture model, the KL divergence can be formulated as

$$\begin{aligned} D_{KL}(P\|Q) = \frac{1}{2}\Bigg[\log\frac{|\Sigma_Q|}{|\Sigma_P|} + \text{Tr}\left[\Sigma_Q^{-1}\Sigma_P\right] - d \\ + (\mu_p - \mu_q)^T\Sigma_Q^{-1}(\mu_p - \mu_q)\Bigg] \end{aligned} \tag{32}$$

where μ denotes the mean and Σ denotes the variance of the Gaussian distribution of the respective component (Hershey and Olsen, 2007). This expression is insightful for our results as the ratio defined in Equation 27 is merely a global indication of overlap. The ratio does not provide information on which specific components are well separated or overlapping in contrast to the KL divergence measure. Therefore, we consider Kullback-Leibler divergence as a measure to explore the pairwise relationships between the components in our mixture model after estimation.

3.7 INITIALIZATION STRATEGY

Well known issues of the EM algorithm are slow convergence and high sensitivity to initial value specification $\widehat{\Theta}^{(0)}$. Different starting strategies and stopping criteria can lead to a range of parameter estimates as final solution (Seidel et al., 2000). Although convergence is ensured, the EM algorithm is greedy. Hence, the solution can be a local optimum yielding a sub-optimal maximum of the log-likelihood. Straightforward approaches are based on multiple random starts after which the best solution is kept to avoid ending in a local optimum. However, more sophisticated strategies have been proposed to overcome initialization problems which often outperform random starting (Biernacki et al., 2003). For example the split and merge EM (SMEM) algorithm designed to escape local maxima in mixture models (Ueda et al., 1999) or the deterministic annealing EM (DAEM) algorithm designed to recover from a poor initialization based on entropy measure (Ueda and Nakano, 1998). Another option is to first run a variant of the EM

algorithm such as the classification EM (CEM) or stochastic EM (SEM) (Celeux and Govaert, 1992). Both CEM and SEM have faster convergence than the EM algorithm and the optimal solution can be used to initialize the EM algorithm. CEM yields a starting solution which is comparable to a K-means type algorithm as a result of hard classification but does not provide ML estimates as it employs the complete likelihood. SEM also classifies observations into a single component but does so in a stochastic manner.

Instead of utilizing an EM variant for initialization it is also possible to perform multiple short runs of the EM algorithm itself. Again, the best solution is then used to initialize a longer run. In this case, the length of the run is controlled by a hyper-parameter in the EM algorithm. A convergence tolerance is defined to stop the estimation when the relative change in log-likelihood is small enough. Such strategies all aim to overcome slow convergence and avoid ending in a local maximum by obtaining more sensible starting positions compared to a multiple of longer runs with random starts. In addition, computational intensity can be immensely decreased. The strategy consisting of shorter EM runs followed by a longer run has been shown to yield good results on both simulated and real life data in a various situations without assuming a particular form of the mixture (Biernacki et al., 2003). Therefore, we use this approach for the initialization of our model.

3.8 VARIABLE SELECTION

Like in almost any model, feature selection is an important aspect. Feature or variable selection has been given increasing attention in statistical research. The current era of high-dimensional problems require adequate techniques to deal with a large number of variables. Therefore, it is desirable to exclude irrelevant information from the model considering the goal of a parsimonious solution. In addition to increasing the goodness of fit, variable selection has the potential to improve the interpretability of our model (James et al., 2013). First we cover traditional approaches. Second, we review some developments in the field of feature selection based on regularization. Thereafter, we describe how to merge variable selection and simultaneous estimation of parameters with the EM algorithm into a single feasible mixture modeling procedure.

As introduced in Equation 26, ℓ_0 penalization is fundamental in various model selection methods. This penalization provides a clear interpretation for subset selection while having convenient sampling properties (Barron et al., 1999). Common feature selection methods are stepwise procedures where variables are iteratively added or removed to find the best subset of features. Often applied examples are stepwise selection, forward selection and backward elimination. The resulting models are compared based on goodness-of-fit measures such as AIC or BIC. However, due to increasing data complexity and size, these stepwise procedures quickly explode to the point of computational infeasibility. Even when a mixture consists of a moderate number of components and variables, classical subset selection approaches are intensive (Khalili and Chen, 2007). In addition, these algorithms are greedy and do not provide any guarantee in finding the optimal subset of variables. Moreover, subset selection approaches are shown to be unstable and further limitations are evident (Breiman, 1995).

As consequence, recent advances have given rise to multiple new forms of penalized likelihood methods with the ability to perform feature selection. The purpose of these methods is to control the number of variables included in the model while taking parsimony and therewith computational intensity into account (Fan and Lv, 2010). Some of these developments are sparked by ultra-high dimension problems where the number of variables p is larger than the number of observations N such that $p > N$. This situation is currently no exception in various fields such as genomics, web analysis, health sciences, finance, economics and machine learning (Fan and Lv, 2010). Hence, it is no surprise that regularization techniques have obtained an important place in modern statistical research and applications.

We are not facing such a high-dimensional problem with more variables than observations. However, we do have numerous variables of which not all may be of equal importance. It is ideal to obtain a parsimonious and well interpretable model while capturing the structure of our data in a satisfactory manner. Naturally, this is very often the goal. This trade-off accounts to finding a good balance in the amount of information needed to explain the structure of the data. Hence, our goal is to estimate

variable effects while simultaneously selecting the important ones by excluding irrelevant variables from the model. This is a complicated optimization problem as we are iteratively estimating a mixture of models instead of a single model. As explained, we assume the data originates from multiple subpopulations. A key consequence stems from this assumption. Namely, the presence of subgroups implies that variables may also vary across components. In turn, this gives rise to a particular interest in selecting the optimal subset of features within each separate segment while correctly estimating the effects of these variables. The variation in features across components can surface in two ways. Firstly, through a difference in the optimal subset of variables and secondly, through a varying importance of the selected variables within a component. In order to achieve this high amount of flexibility, we need to combine estimation of our model with a continuous variable selection algorithm that has the freedom to operate independently across components.

We now introduce several forms of penalization methods from the starting point of Ordinary Least Squares (OLS). Thereafter, we formulate an approach that combines a finite mixture model with penalization. OLS minimizes the residual sum of squares (RSS) formulated as

$$\beta_{OLS} = \min_{\beta} \underbrace{\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2}_{\text{RSS}} . \tag{33}$$

In order to obtain an estimation method that can perform feature selection we extend the model with penalization. The principle of ℓ_0 penalization was introduced in Equation 26. It can be seen as part of the general families of ℓ_q penalties, also referred to as bridge functions (Frank and Friedman, 1993). This form of penalties is given by

$$\lambda \sum_{j=1}^{p} |\beta_j|^q \tag{34}$$

where $0 < q \leq 1$ in order to achieve variable selection abilities. For $q = 0$ we obtain the AIC or BIC penalty depending on λ as described in Equation 26. This function of families can be used to introduce penalization methods starting with ridge regression by Hoerl and Kennard (1970). Ridge regression has lead to more recent advances such as the lasso by Tibshirani (1996) and the elastic net by Zou and Hastie (2005). The lasso and elastic net both posses the ability to perform continuous variable selection which is further discussed in the next sections.

3.8.1 RIDGE REGRESSION

Ridge regression is the foundation of many modern penalization methods (Hoerl and Kennard, 1970). It is also known as Tikhonov regularization (Tikhonov et al., 1977) or as weight decay in neural networks in the field of machine learning (Friedman et al., 2001). Instead of ℓ_0 penalization it is based on the ℓ_2 norm (Hoerl and Kennard, 1970). This form of penalization is obtained by setting $q = 2$ in Equation 34 resulting in the following objective function

$$\beta_{RDG} = \min_{\beta} \underbrace{\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2}_{\text{RSS}} + \underbrace{\lambda \sum_{j=1}^{p} |\beta_j|^2}_{\text{Penalty}} . \tag{35}$$

Ridge regression has the property to decrease the effect of non-important variables, this is referred to as shrinking. The amount of shrinkage is controlled by the λ parameter (Friedman et al., 2001). In addition, the variance of the coefficient estimates can be significantly decreased as result of shrinking (James et al., 2013). Although the effect of a variable can be decreased with ridge regression, it cannot be nullified. In other words, ridge regression cannot perform feature selection to obtain a more parsimonious model (Zou and Hastie, 2005). Yet, shrinking to exactly zero is highly desirable when the goal is to select the most important variables in the model. A similar procedure that does possess the ability to perform feature selection is the least absolute shrinkage and selection operator (lasso) introduced by Tibshirani (1996).

3.8.2 LASSO

In contrast to ridge regression the lasso is based on ℓ_1 instead of ℓ_0 penalization. This is achieved by setting $q = 1$ in Equation 34 yielding the following objective function

$$\beta_{LAS} = \min_{\beta} \underbrace{\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)}_{\text{RSS}} + \underbrace{\lambda \sum_{j=1}^{p} |\beta_j|}_{\text{Penalty}} . \tag{36}$$

The lasso can be described as a continuous subset selection algorithm with the ability to shrink the effect

of unimportant variables similar to ridge regression (Tibshirani, 1996). The algorithm constrains the total magnitude of the coefficients resulting in the scaling of a variables effect based on importance. In contrast to ridge regression, the lasso possesses variable selection properties. This is achieved by the ability to shrink the effect of a certain variable all the way down to zero. This can be interpreted as exclusion of this respective variable from the model. A numerical advantage of the lasso is a convex penalty function. This is very convenient from a computational viewpoint.

The concept of the lasso is influenced by Breiman's non-negative garrotte (Breiman, 1995). A drawback of the non-negative garrotte is that it is not defined when a problem involves more parameters p than observations N which is not uncommon present-day. The lasso is still valid in this case but shrinkage of the non-zero coefficient causes non-ignorable bias towards zero yielding inconsistent estimates (Fan and Li, 2001). The bias can be reduced by a modification of the penalty function such that large coefficients are shrunken less (Fan et al., 2004). This idea is used in another variable selection algorithm known as the smoothly clipped absolute deviation (SCAD) (Fan et al., 2004).

Alternatively, the lasso can be extended by including data-dependent weights which is known as the adaptive lasso (Zou, 2006). Now, the strength of penalization is allowed to vary across different coefficients due to adding adaptive weights in the penalty giving the following objective function

$$\beta_{ALS} = \min_{\beta} \underbrace{\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)}_{\text{RSS}} + \underbrace{\lambda \sum_{j=1}^{p} \hat{w}_j |\beta_j|}_{\text{Penalty}} . \tag{37}$$

where $\hat{w}_j$ are the coefficient dependent weights with the power to control penalty strength per coefficient. This weighting vector is determined by

$$\hat{w}_j = \frac{1}{|\hat{\beta_j^{\text{ini}}}|^{\gamma}} \tag{38}$$

where $\hat{\beta_j^{\text{ini}}}$ are initial estimates of the coefficients which can be obtained from a consistent estimator for $\hat{\beta}_j$ such as OLS or ridge regression. In order for the adaptive lasso to be consistent $\hat{\beta_j^{\text{ini}}}$ need to be consistent. Coefficients with lower initial estimates are penalized more through the weights vector $\hat{w}_j$. It has been shown that this extension yields the oracle property (Zou, 2006; Fan and Li, 2001; Fan et al., 2004). An estimator has the oracle property if it has the ability to be consistent in both parameter estimation as well as variable selection. This is further examined in the Simulation study in section 5. On the contrary, the regular lasso does not posses the oracle property which has been shown to be associated with the bias problem (Zou, 2006). The adaptive lasso consistently estimates parameters while retaining the desirable convexity property (Friedman et al., 2001).

Recent studies have discovered that the lasso is related to the maximum margin explanation which is key in support vector machines (SVM) and boosting algorithms (AdaBoost, XGBoost) in the field of machine learning (Rosset et al., 2004). The lasso has been used to explain the success of boosting which can be interpreted as a high-dimensional lasso without explicit use of the ℓ_1 penalty (Friedman et al., 2004, 2001). However, a drawback of both lasso algorithms is the performance in presence of multicollinearity. In practice, variables can be highly correlated especially when the number of variables is relatively large. In this situation the lasso has the tendency to select merely one of these correlated variables in an arbitrary fashion while ignoring the others. Zou and Hastie have shown the lasso path to be unstable in case of multicollinearity yielding unsatisfactory results (2005). These difficulties are overcome by a more recent regularization technique called the elastic net (Zou and Hastie, 2005). For this reason, we select the elastic net as variable selection algorithm in our model.

3.8.3 ELASTIC NET

A relatively new regularization and variable selection method is the elastic net (Zou and Hastie, 2005). This method is closely related to the lasso which has proven to be a valuable asset in modern model fitting and covariate selection. Some of the limitations of the lasso are solved by combining the ℓ_1- and ℓ_2 norm into a new penalty function given by

$$\xi_{NET}(\beta_j) = \lambda \sum_{j=1}^{p} \left(\alpha |\beta_j| + \frac{(1-\alpha)}{2} |\beta_j|^2 \right) \tag{39}$$

such that the following problem is solved

$$\beta_{NET} = \min_{\beta} \underbrace{\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)}_{\text{RSS}} + \underbrace{\lambda \sum_{j=1}^{p} \left(\alpha |\beta_j| + \frac{(1-\alpha)}{2} |\beta_j|^2 \right)}_{\text{Penalty}} \tag{40}$$

where α is a parameter that determines the mix of the penalties. Setting $\alpha = 0$ results in ridge regression whereas setting $\alpha = 1$ results in the lasso. Hence, this method can be seen as a dynamic blend of ridge regression and the lasso. The elastic net possesses all the desirable properties of the lasso, it can perform automatic variable selection through continuous shrinkage while overcoming the issues regarding multicollinearity. The second term in Equation 39 causes variables with high correlation to be averaged, whereas the first term encourages a parsimonious solution and stabilizes the solution (Friedman et al., 2001). Zou and Hastie describes this method as a stretchable fishing net with the ability to retain all the big fish (2005). It has been shown that elastic net often yields better results than the lasso in simulations and real world data (Zou and Hastie, 2005). To implement the elastic net we make use of the *glmnet* algorithm developed by (Friedman et al., 2009) which is specifically designed for speed and dealing with relatively large datasets.

3.8.4 HYPERPARAMETER SELECTION

The choice of mixing parameter α depends on preference and the problem at hand. Commonly, some experimentation is done with different values using cross-validation. Sometimes α is merely used to obtain a more stable application of the lasso. This can be done by setting $\alpha = 1 - \varepsilon$ with some small value $\varepsilon > 0$. This approach increases numerical stability by negating undesirable behavior of the lasso caused by high correlations in the data. Another common choice is $\alpha = 0.5$ which gives an evenly divided mix of the ridge and lasso penalty term. As consequence, groups of correlated variables are selected to be included or excluded together. We test three values for the mixing parameter $\alpha = \{0.5, 0.7, 0.9\}$. These values tend more towards the lasso than ridge regression as we prefer a parsimonious solution. We exclude the lasso to avoid problems associated with multicollinearity. Since we are fitting a finite mixture model, an alternative is to select α per component based on the smallest error. This results in a different penalization method per group. Some components will tend towards a smaller value for α. This is not desired in our application, hence we select the same penalization method for all groups. In general, penalization towards ridge regression is good for prediction purposes but yields a less interpretable solution. This approach would be more appropriate if the main focus were to accurately predict the balance of individuals. For the sake of parsimony and further interpretation we are trying find a small subset of the most important variables per component.

For all three penalization methods introduced above, the strength of the penalty parameter λ can not be estimated directly due to identifiability problems. To solve this issue we use 10-fold cross-validation to obtain a sequence of models with different penalty strengths over the grid of α values (Golub et al., 1979). The regularization path is fitted based on a range of 100 different values of λ. The minimum value in the range is 0. This equals no penalization such that all variables are included. The maximum value for λ is set to the value for which all coefficients are zero. This means that at this value for λ all variables are excluded from the regression. The strength of the penalty in each component is estimated independently. In general, a higher penalty value leads to more severe shrinkage of parameters and a smaller selection of variables. Simultaneously, exclusion of variables increases the error. Hence, the purpose of this cross-validation is to find a balanced trade-off between error and parsimony. If the relative improvement falls below the threshold of 10^{-5} the computation is stopped. The results allow to select an appropriate value for λ. Generally one of the following options is used for λ. First, the value which minimizes the mean cross-validation error (MSE) denoted by λ_{min}. Second, the value which results in the most regularized model within one standard error of λ_{min}, denoted by λ_{1se}. Both options are supported by literature and used in applications. The restriction $\lambda_{1se} > \lambda_{min}$ holds in all cases. For our model we select λ_{1se} as this value encourages a more parsimonious solution in comparison with λ_{min}. The selection of α and λ is further discussed in the results section.

3.9 EXTENDED FINITE MIXTURE MODEL (MIXNET)

We have first discussed the fundamentals regarding the formulation and estimation of finite mixture models. Second, we introduced penalized estimation methods. The methodology is now expanded by merging these two principles into a single estimation and variable selection algorithm. This approach is inspired by Khalili and Chen who makes use of the lasso to perform variable selection in mixture models (2007). Khalili and Chen have shown that this procedure is consistent and yields equal or better performance than traditional methods such as BIC in terms of model selection whilst greatly reducing computational burden.

We now introduce a model which combines a finite mixture model with the elastic net algorithm. We refer to this model as MIXNET in short. MIXNET combines the power of statistical based finite mixture modeling with the convenience of automatic variable selection. The result is a highly feasible and relatively fast procedure in terms of computational intensity. Variable selection is achieved by shrinkage of parameters through the elastic net algorithm. As consequence, all desirable properties of the elastic net are adopted. MIXNET has the ability to deal with a large number of variables while simultaneously performing continuous selection of the relevant ones. We would like to emphasize the power of this algorithm as it possesses the ability to operate independently within components. Hence, both estimation and variable selection is done in a component specific manner. This increases both the flexibility and potential interpretability of groups in comparison to a variable selection procedure that takes the entire population into account as a whole. Moreover, in case the problem contains more variables than observations, such that $p > N$, MIXNET can still be applied in contrast to a regular likelihood approach.

We now cover the mathematical formulation of this model. To obtain the ability to perform feature selection through shrinkage we take the log-likelihood function as given in Equation 14 and extend it with a penalty term such that we have a penalized log-likelihood function defined as

$$\widetilde{\mathcal{L}}(\Theta) = \mathcal{L}(\Theta) - \text{Penalty}(\Theta). \tag{41}$$

To obtain the MIXNET model we employ the elastic net penalty as given in Equation 39

$$\xi_{NET}(\Theta) = \sum_{s=1}^{S} \lambda_s \sum_{j=1}^{p} \left(\alpha|\beta_{sj}| + \frac{(1-\alpha)}{2}|\beta_{sj}|^2 \right) \tag{42}$$

resulting in a penalized log-likelihood function

$$\begin{aligned} \widetilde{\mathcal{L}}(\Theta) = & \underbrace{\sum_{i=1}^{N} \log \sum_{s=1}^{S} \pi_s \cdot f(y_i|\theta_s)}_{\text{Log-Likelihood}} \\ & \underbrace{- \sum_{s=1}^{S} \lambda_s \sum_{j=1}^{p} \left(\alpha|\beta_j| + \frac{(1-\alpha)}{2}|\beta_j|^2 \right)}_{\text{Penalty}}. \end{aligned} \tag{43}$$

The corresponding maximum likelihood (ML) is then computed by

$$\begin{aligned} \widehat{\Theta}_{ML} &= \arg\max_{\Theta} \widetilde{\mathcal{L}}(\Theta) \\ &= \arg\max_{\Theta} \left[\log f(y|\Theta) - \xi_{NET}(\Theta) \right] \\ &= \arg\max_{\Theta} \left[\sum_{i=1}^{N} \log \sum_{s=1}^{S} \pi_s \cdot f(y_i|\theta_s) \right. \\ &\qquad \left. - \sum_{s=1}^{S} \lambda_s \sum_{j=1}^{p} \left(\alpha|\beta_j| + \frac{(1-\alpha)}{2}|\beta_j|^2 \right) \right]. \end{aligned} \tag{44}$$

Lastly, the complete data log-likelihood function is defined as

$$\begin{aligned} \widetilde{\mathcal{L}}_c(\Theta) &= \log f(y, Z|\Theta) - \xi_{NET}(\Theta) \\ &= \sum_{i=1}^{N} \sum_{s=1}^{S} z_{is} \cdot \log \left[\pi_s \cdot f(y_i|\theta_s) \right] - \xi_{NET}(\Theta). \end{aligned} \tag{45}$$

To obtain estimates of the parameters $\widehat{\Theta}$ the EM algorithm as described in Algorithm 1 is used. The algorithm can be subdivided into two separate steps, the Expectation-step and the Maximization-step. Every iteration provides new parameter estimates $\widehat{\Theta}$. The E-step computes the expectation of the complete data log-likelihood conditional on y and the current estimate $\widehat{\Theta}^{(t)}$. The E-step is given by

$$\mathbb{E}\left[\widetilde{\mathcal{L}}_c(\Theta) \right] = \mathbb{E}\left[\log f(y, Z|\widehat{\Theta}^{(t)}) - \xi_{NET}(\widehat{\Theta}^{(t)}) \right] \tag{46}$$

Consequently, the M-step maximizes the expected

value in Equation 46 with respect to Θ such that

$$\begin{aligned}\widehat{\Theta}^{(t+1)} &= \arg\max_{\Theta} \mathbb{E}\left[\widetilde{\mathcal{L}}_c(\widehat{\Theta}^{(t)})\right] \\ &= \arg\max_{\Theta} \mathbb{E}\left[\log f(y, Z|\widehat{\Theta}^{(t)}) - \xi_{NET}(\widehat{\Theta}^{(t)})\right]\end{aligned} \tag{47}$$

yielding updated parameter estimates $\widehat{\Theta}^{(t+1)}$. The two steps are repeated until convergence is met resulting in a final solution. The log-likelihood function can be extended with the different penalties introduced above in a similar manner. For instance, to obtain a log-likelihood function with the adaptive lasso penalty.

4 RESULTS

We now report the results obtained by following the estimation and simultaneous variable selection procedure referred to as MIXNET. The results are organized in the following manner. We start with a exploratory data visualization. Second, the component selection procedure is reported. Third, we discuss the grouping structure. Next, the coefficient estimates are reported and interpreted. Lastly, we conclude with segment-level results by discussing the most important properties of the components. Our goal is to achieve a clear and concise interpretation of the segments which can be used to improve business.

4.1 EXPLORATORY DATA ANALYSIS

In order to visually explore our dataset and potentially reveal some structure we apply both PCA and the t-SNE algorithm. Consequently, the results are mapped into two-dimensional space. For optimal results in the PCA we first center and scale the data such that each variable has a mean equal to zero and variance equal to one. Figure 1 displays the results of plotting the first two principal components of the PCA. The first component describes 16% of the variability in the data while the second describes 9%. Together the first two principal components capture 25% of the total variation in the data. Figure 2 shows the data mapped into two-dimensional space by the t-SNE algorithm. In both figures the observations are colored by age. We find PCA manages to find some structure in the data based on the first two principal components. Clearly, younger customers are in the bottom part of the point cloud whereas older individuals are seen in the top part. Yet, no clear separated grouping structure is found in the data by plotting the first two dimensions.

The t-SNE solution reveals somewhat separated groups of observations in comparison with PCA. Still, many datapoints overlap, especially in the center part of the figure. Datapoints that are close together represent similar observations while two observations in separated point clouds indicate that the observations are dissimilar. Most noticeable are the darker colored clusters corresponding to younger individuals. Further inspection reveals that many point clouds have a lighter colored edge corresponding to older individuals.

There is a large amount of overlap in the two-dimensional visualization of both techniques. The solutions do not point towards a clear presence of groups that are easy to separate based on the relations in our data. This finding can be supported by the fact that our data contains relatively many observations and variables. Of the included variables there are many that do not show a large amount of variation across the observations. For instance, the majority of individuals possess a main insurance. There are merely 2972 observations in the population without a main insurance. When two observations share the same value on a certain variable they are already somewhat similar. An ideal solution before applying a finite mixture model would reveal a clearly separated grouping structure of the observations. For instance, three dense and separated clusters of datapoints would implicate observations are easy to group and the groups easy to separate based on the high-dimensional patterns of the data. This finding could support the appropriateness of fitting a three-component mixture model.

Figure 1 Visualization of the data structure of the first two principal components of PCA.

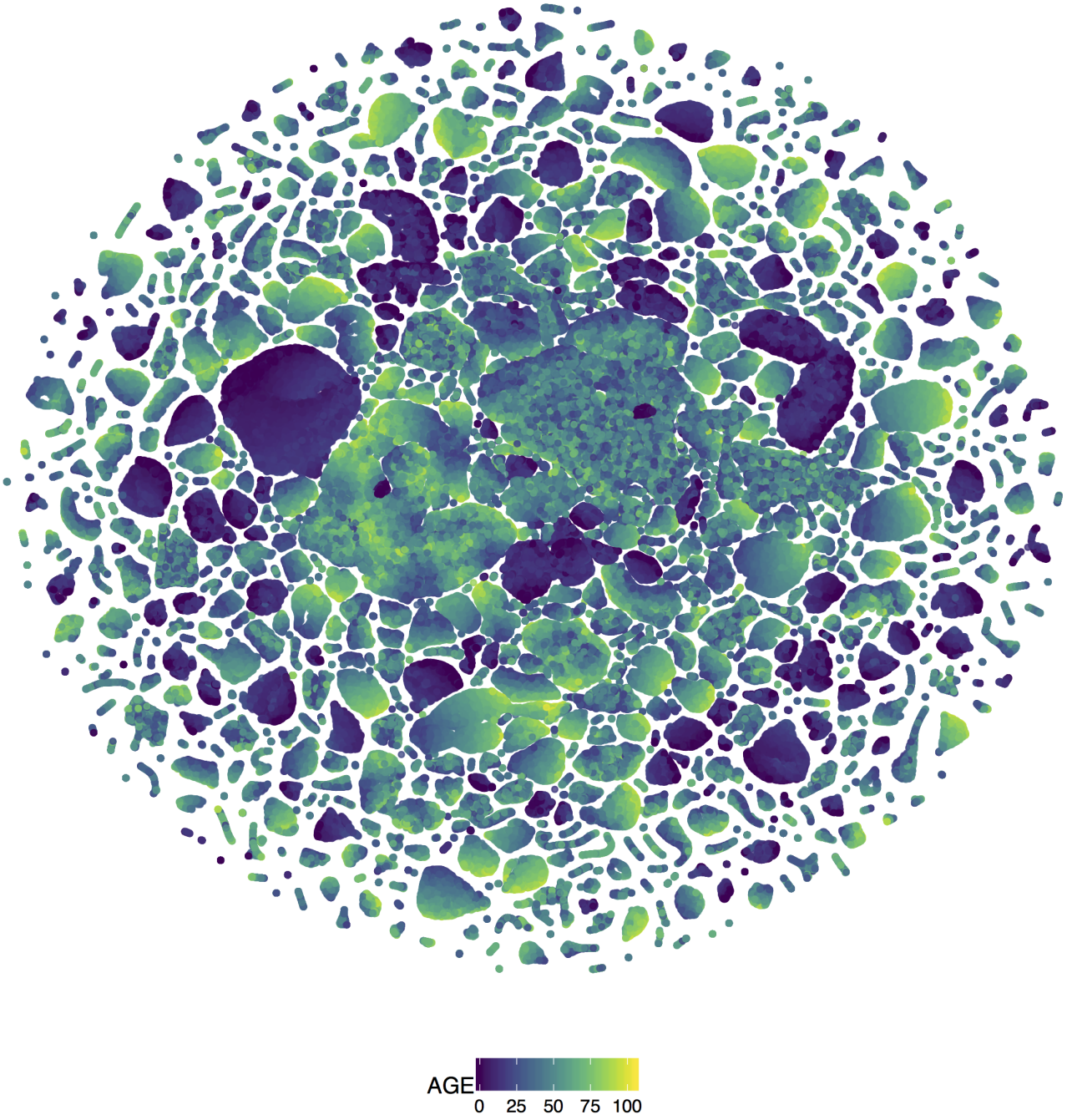

Figure 2 Visualization of the data structure mapped to 2-dimensional space with the t-SNE algorithm.

4.2 COMPONENT SELECTION

This section summarizes the results of our selection procedure to determine the optimal number of components S needed to model the structure of our data. To control for computation time we impose two separate stopping criteria on the EM algorithm. First, a tolerance threshold of 0.01 as described in the initialization strategy. This means convergence is met when the relative improvement in the log-likelihood drops below 1%. Second, we set a maximum number of iterations. If either of these conditions is met, the EM algorithm is forced to stop and the current estimate is taken as solution. In addition, we examine the size and prior probabilities π_s of the components in the solution. If the size or prior probability of a component is relatively low compared to the others we consider the segment as not substantial and remove it from the solution. Table 3 presents the best solution of the 10 repetitions for each specification component size of the first stage. According to the used criteria, the optimal solution is obtained with roughly 10 components.

Next, we perform the second stage of our component selection procedure based on the information provided by the first stage. Now we estimate our model with 6 up to 14 components in a step-size of 1. Table 4 reports the solutions of every step. Based on this comparison we conclude that a model with 12 components performs the best in terms of describing the structure of the data based on our diagnostic measures. However, the solution with 12 segments contains a component with merely 2157 observations and corresponding low prior probability of 0.004. The size of these group is not substantial enough to target from a business point of view, in addition a smaller model is preferred. Hence, we decrease the number of components to 11 and repeat the entire estimation procedure. Again, the solution contains a small component with only 2852 observations. Therefore, we set $S = 10$ and obtain a final solution where the smallest group contains 4158 observations. We accept this solution in order to not deviate too much from the optimal number of 12 components which may lead to inability of capturing structural differences in the data. The results of this solution are reported in Table 5.

Table 3 Component selection stage 1

S	df	log-Lik	AIC	AIC-3	BIC	ICL
5	194	-241778	483943	484136	485989	920496
10•	352	-232969	466642	468641	470375•	1251132
15	544	-232556	466200	466744	471969	1426217
20	717	-231448	464329	465045	471921	1579343
25	932	-241636	485135	486066	495006	1854422
30	1114	-243742	489711	490824	501512	2013732

Best solution highlighted in gray
• Decisive measure

Table 4 Component selection stage 2

S	df	log-Lik	AIC	AIC-3	BIC	ICL
7	155	-237636	475581	475736	477225	1115935
8	174	-236717	473781	473955	475626	1129006
9	191	-236518	473419	473610	475444	1210953
10	213	-236021	472468	472681	474726	1244810
11	232	-234981	470426	470658	472886	1344398
12•	256	-233102	466717	466973	469431•	1324802
13	272	-234090	468724	468996	471608	1340071
14	290	-233798	468176	468466	471251	1400102

Best solution highlighted in gray
• Decisive measure

4.3 CLUSTERING STRUCTURE

Now that we have determined the number of components we employ the initialization strategy consisting of short EM runs followed by a full estimation run to obtain a final solution. Consequently, we look at the obtained grouping structure as reported in Table 5. The following metrics are used to describe the structure. First, the *size* of a group represents the number of observations assigned to this component based on the posterior probabilities. Second, *prior* refers to the prior probability π_s of of an observation i belonging to group s. Third, $\#\{post > \varepsilon\}$ represents the number of observations with a posterior probability of belonging to this component larger than epsilon. We have set $\varepsilon = 0.05$, meaning posterior probabilities of belonging to a certain component smaller than 5% are interpreted as zero. Lastly, the *ratio* as defined in Equation 27 is given which is an indication of how well a segment is separated from the others. The grouping structure obtained by the elastic model is reported in Table 5. The components are sorted in ascending order by their prior probability π_s. Figure 3 displays a comparison of the size of each component. We find component 1 is the smallest containing roughly 0.5% of the sample whereas segment 10 is by far the largest containing 38% of the population based on posterior probabilities.

Table 5 Grouping structure of MIXNET result

Component	Prior	Size	$\#\{\text{post} > \varepsilon\}$	Ratio
1	0.009	4158	13568	0.307
2	0.067	27433	324985	0.084
3	0.083	15199	478109	0.032
4	0.088	54472	435699	0.125
5	0.094	83160	432770	0.192
6	0.094	61723	425107	0.145
7	0.111	143899	345995	0.416
8	0.118	89336	461861	0.193
9	0.135	77158	517791	0.149
10	0.202	344931	673340	0.512

Components are ordered by prior probability π_s

The results in Table 5 indicate a large amount of overlap between the components. This finding is in agreement with the results of the exploratory data

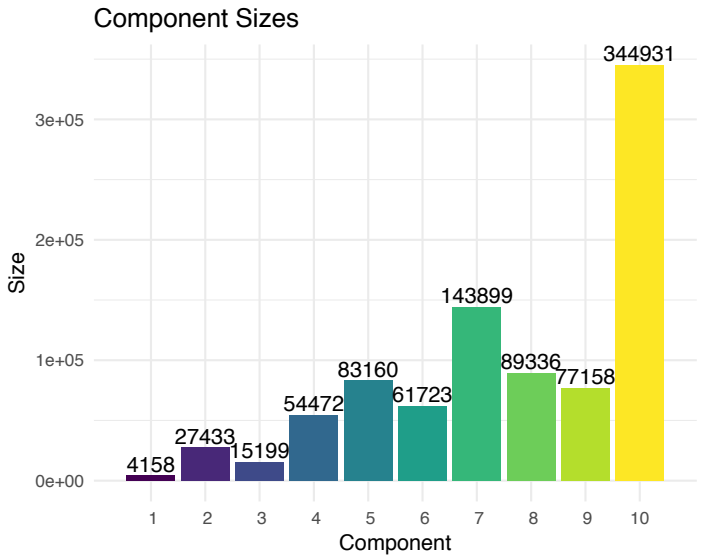

Figure 3 Comparison of component sizes.

visualization. Cluster 10 is the most separated followed by segment 7 and 1. Segment 3 is the least separated from other clusters. However, the *ratio* in Table 5 does not provide information regarding which specific pairs of components overlap. For the segment interpretation we would like to examine if a certain group is part of a larger group of customers or if the group can be seen as a distinct market segment. To obtain more detailed information regarding the overlap and separation in our clustering solution we use the Kullback-Leibler divergence measure as introduced in Section 3.6. This measure allows for calculating pairwise dissimilarities of the distributions of the clusters. Table 6 reports the pairwise Kullback-Leibler divergence measures. The values are divided by 100,000 and rounded to whole numbers to increase convenience of interpretation. Naturally, the diagonal is zero as the distance of a components distribution to its own distribution is zero. As discussed, KL divergence is not symmetric. This is evident when comparing the upper and lower triangle of Table 6. High values are observed in the row of component 1. This indicates that the density of *1* has the highest deviation from the other groups. The maximum value is observed for the pairwise distance from component *1* to *7*.

4.4 HYPERPARAMETER ESTIMATES

We wish to obtain a clear description for the different groups focusing on the most distinct and important differences. Hence, we prefer a relatively small subset of the most important variables per component. For

the mixing parameter of the penalties we considered three values, $\alpha = [0.5, 0.7, 0.9]$. We exclude $\alpha = 1$ to avoid arbitrary selection within groups of correlated variables and computational issues. For each α value in the grid cross-validation is done to determine the optimal penalty strength λ per component. The resulting final solutions per α are reported in Table 7. The results indicate that even when setting α to a relatively high value for all components, many of the variables are still included. We find $\alpha = 0.9$ yields the lowest BIC score and the most parsimonious solution. Hence, this value of α is selected. Figure 4 displays the result of 10-fold cross validation for λ with $\alpha = 0.9$. Two values of lambda are highlighted in the plots with a dashed vertical line. The first line corresponds with the value that minimizes the mean cross-validated error, λ_{min}. The second line corresponds with the stronger penalized model within one standard deviation, λ_{1se}.

We select λ_{1se} as our penalty value which yields the most parsimonious solution of the two lambda values. Figure 4 shows that different penalty strengths are selected within components. The values displayed on the top vertical axes above each plot represent the number of non-zero variables at that respective value of λ. The number of selected variables varies across the clusters as a results of differing penalty strengths. For instance, in component 1 we obtain $\lambda_{1se} \approx \log(2) \approx 0.3$ whereas in component 3 we find $\lambda_{1se} \approx \log(5) \approx 0.7$. We find the least amount of penalization is done in component 1 resulting in a selection of 21 variables while component 3 is most penalized leading to a selection of merely 9 variables. The shapes of the lambda estimates look comparable, however differences can be seen when looking at the range of the axes. For one thing, the mean squared error ranges from roughly 10 to 13 in component 1 whereas most other components are below a value of 1. Furthermore, component 7 has a noticeably smaller error ranging from 0.02 to 0.08.

Table 6 Pairwise Kullback-Leibler divergence of components

	Component									
	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
1	.	299	1100	2679	2367	1255	2438	1336	117	794
2	15	.	81	194	113	44	253	226	37	769
3	20	23	.	24	29	332	156	46	20	13
4	24	26	10	.	14	27	117	39	20	13
5	24	18	14	15	.	9	150	74	26	20
6	21	12	29	58	16	.	224	116	30	36
7	24	36	75	131	149	118	.	113	23	44
8	21	56	440	80	134	111	21	.	11	18
9	11	100	163	358	425	283	363	81	.	96
10	19	27	18	31	55	55	126	28	16	.

Values are divided by 100,000 and rounded to whole numbers

Table 7 Comparison of α values

α	**S**	**df**	**log-Lik**	**AIC**	**AIC-3**	**BIC**	**ICL**
0.5	10	170	-677075	1354491	1354661	1356482	3157545
0.7	10	171	-677830	1356003	1356174	1358005	3155693
0.9	10	158	-674553	1349422	1349580	1351272•	3043364

Best solution highlighted in gray

• Decisive measure

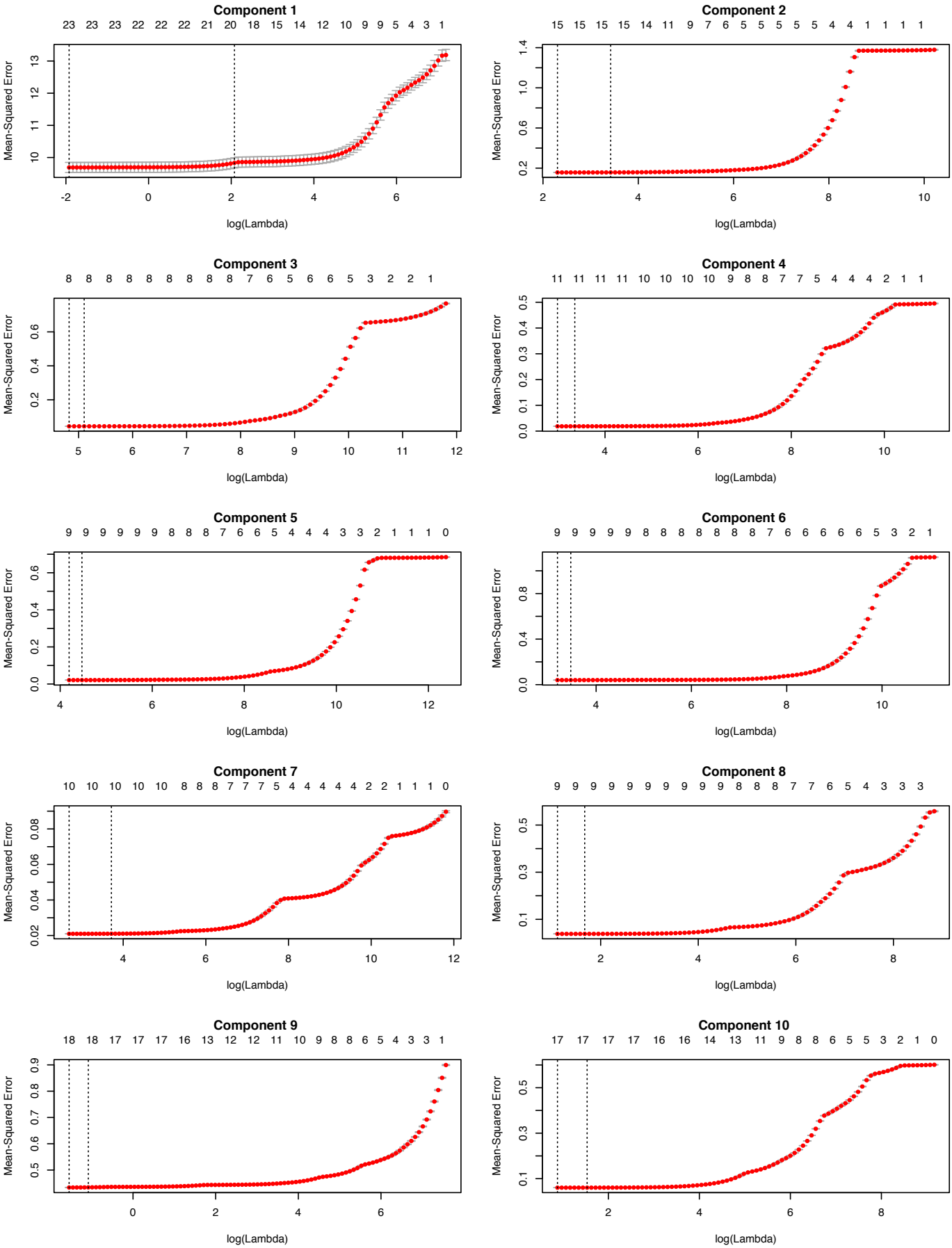

Figure 4 Results of cross-validation for λ in each component. Dashed lines indicate λ_{min} and λ_{1se}. Values on the top vertical axes represent the number of non-zero variables.

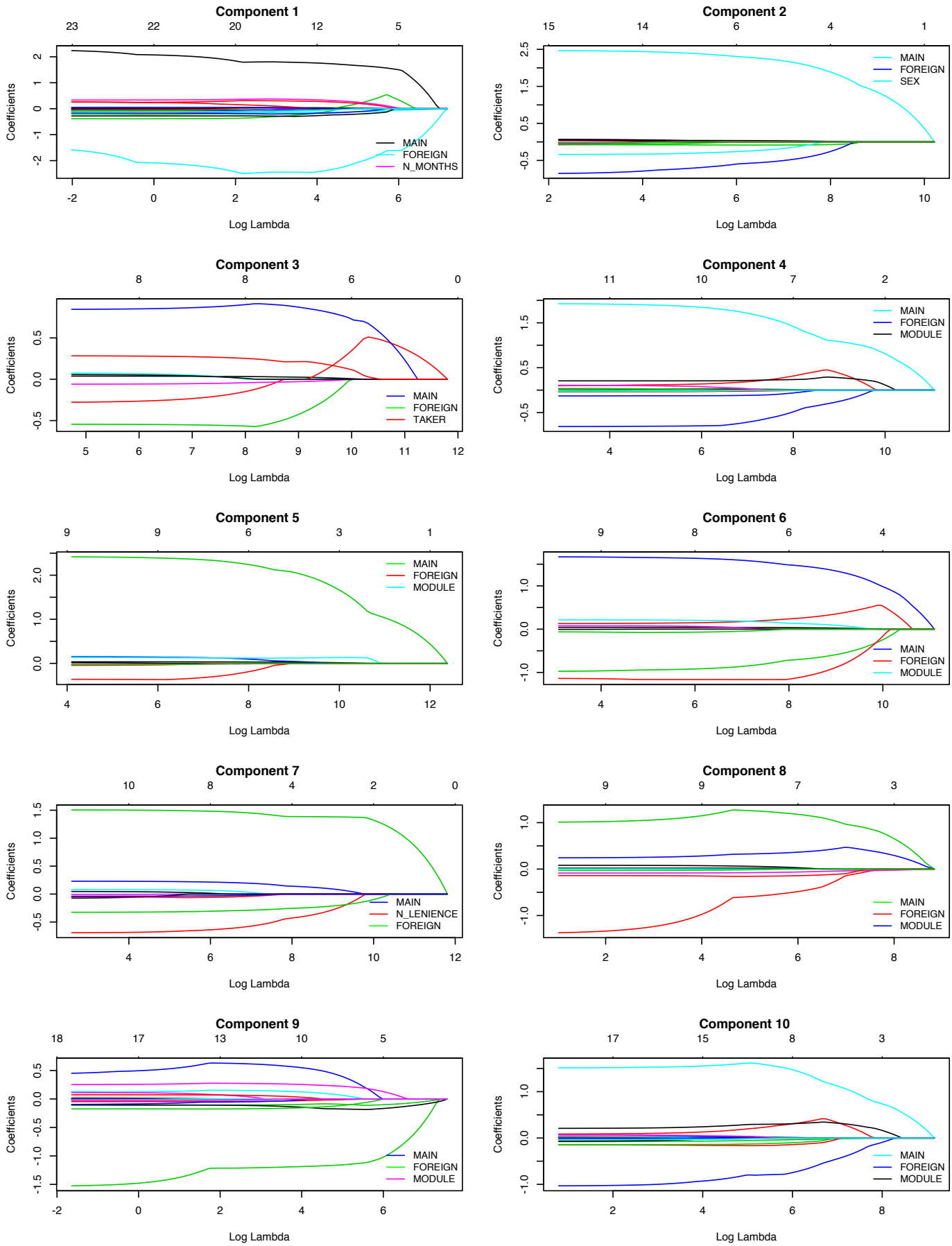

Figure 5 Coefficient paths of each variable. Values on the top vertical axes represent the number of non-zero variables.

Table 8 Coefficient estimates within each component

	Component										
Variable	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	**Population**
INTERCEPT	-1.56	-1.57	-0.56	-1.49	-2.33	-1.92	0.41	-0.28	1.59	-0.75	-0.65
AGE	0.05	0.04	0.04	0.03	0.04	0.04	.	0.02	.	0.02	0.03
N_YEARS	.	.	.	.	.	.	.	.	-0.01	.	.
TAKER	-0.2	0.04	-0.28	0.1	.	0.13	0.04	.	.	0.09	0.02
N_ON_POLICY	-0.07	-0.08	.	-0.01	.	.	.	.	0.02	-0.04	-0.01
FOREIGN_IND	-2.46	-0.82	-0.54	-0.81	-0.36	-0.97	-0.69	-1.35	-1.52	-1.03	-0.96
PAYMENT_TERM	.	.	.	.	.	.	.	.	.	.	.
N_ON_COLLECTIVITY	.	.	.	.	.	.	.	.	.	.	.
MAIN_IND	1.82	2.45	0.85	1.92	2.41	1.67	1.5	1.02	0.46	1.52	1.34
ADDITIONAL	0.31	0.06	0.07	0.11	0.15	0.21	0.23	.	0.13	-0.02	0.09
MODULE	-0.02	0.06	.	0.21	0.14	0.08	0.08	0.24	0.25	0.21	0.13
PROVISION	.	.	.	.	.	.	.	.	.	0.01	.
N_CLAIMS	-0.16	.	-0.06	-0.04	-0.02	.	-0.01	-0.09	-0.05	-0.08	-0.07
N_MONTHS	0.36	.	.	.	.	.	.	.	-0.11	0.04	0.02
VOLUNT_EXCESS	.	.	.	.	.	.	.	.	.	.	.
N_NEGLECT	-0.01	.	.	.	.	.	.	.	-0.01	.	.
N_LENIENCE	-0.37	-0.03	.	.	.	-1.14	-0.06	.	0.11	.	.
N_HEALTH_CATS	-0.19	-0.06	.	.	.	-0.06	-0.05	.	-0.11	0.04	-0.01
MAX_N_CATS	0.03	.	0.06	0.04	0.01	.	.	0.08	0.07	0.06	0.05
COMPARISON_IND	-0.11	-0.03	.	.	.	.	.	.	.	-0.07	.
SEX	-0.28	-0.33	0.28	.	.	.	-0.32	.	-0.17	-0.15	-0.08
$BRAND_2$	0.15	-0.03	.	-0.13	-0.04	.	.	-0.14	-0.09	-0.15	-0.02
$BRAND_3$	-0.11	0.04	.	.	.	.	.	.	-0.02	.	.
$REGION_{GGZ}$	0.04	-0.01	.	.	.	.	.	.	.	-0.01	.
$REGION_{VV}$	-0.11	-0.02	.	-0.03	-0.03	.	-0.04	-0.03	-0.02	-0.03	-0.02
# of Variables	21	16	9	12	10	10	11	10	19	18	15

Estimates are rounded to two decimal places, excluded features are marked with a.

4.5 COEFFICIENT ESTIMATES

The plots in Figure 5 display the coefficient path versus the penalty strength parameter λ in each segment. This figure visualizes the shrinkage behavior of the variable selection procedure. Each curve corresponds to a single variable, factor levels are count as separate variables. The values displayed on the top vertical axes above each plot represent the number of non-zero variables at that respective value of λ. In addition, the legend in each plot reports the three variables with the largest absolute influence within that component. The results clearly show different paths due to independent estimation and variable selection within the segments. A higher penalty results in more severe shrinkage of the parameter estimates, which can be observed by tracing the coefficient paths from left to right in any given component. Consequently, a stronger λ means stricter selection resulting in the exclusion of more variables within the component. The coefficient estimates within each component are reported in Table 8. The excluded variables which are shrunken down to exactly zero are denoted with a dot.

Standard errors are not reported as they can be unmeaningful or misleading in regularized regressions. Shrinkage can significantly reduce the variance of the estimators which is achieved by introducing a bias. As a result, the introduced bias can form a substantial part of the mean squared error. Several methods have been proposed to obtain reliable standard errors in a penalized setting. For instance, the sandwich formula by Fan and Li which estimates the covariance of the estimates and an extension for the adaptive lasso by Zou and Hastie (Fan and Li, 2001; Zou and Hastie, 2005). However, both methods yield a variance equal to zero for estimates that are shrunken towards zero. Alternatively, a bootstrapping approach can be applied to obtain standard errors (Tibshirani, 1996). However, bootstrapping can be very intensive computationally. The bootstrapping approach has been thoroughly studied and performance is argued in this case. For instance, Knight and Fu report estimation problems caused by a bias in bridge estimators, including the lasso, when the true parameter values are shrunken to or just above zero (2000). Further issues are discussed by Leeb and Pötscher who show difficulties in estimating precision of shrinkage estimators and Beran who prove inconsistencies in estimation (Leeb and Pötscher, 2006; Beran, 1982). Thus, obtaining valid standard error estimates with a bootstrap approach proves to be problematic in practice. In short, the estimates are inconsistent for the variables that are shrunken towards or set to zero (Kyung et al., 2010). If our goal were to find an optimal model to predict the balance of an individual then standard errors would be of bigger concern. In this case, an approach such as a Bayesian lasso would be more appropriate as this would allow to produce valid standard errors (Kyung et al., 2010). However, this is not our purpose here since the main focus is not prediction of just the financial balance but describing the structure of the data and finding the most important relations.

We find the following relations provided by the coefficient estimates in Table 8. All coefficient are interpreted as average within the component and relative across components. Moreover, the discussed effects are under the assumption of keeping all other variables fixed (ceteris paribus). Numbers refer to components to improve readability. For instance, *9* refers to component *9*.

First, we look at demographics and general characteristics of customers. For most components being a male is negatively related to balance with exception of *3* where the effect is positive. Age appears to have a relatively small positive effect which seems counter-intuitive. We expect to find a negative effect for age. In general, older individuals are less healthy compared to younger people and require more health services yielding a less positive balance. This result is possibly explained by the settlement which is taken into account in our response variable. The settlement ensures a higher compensation for older individuals compared to younger customers. In addition, the number of packages is positively related to age. This means older customers have a higher insurance coverage in general, resulting in a better negation of health care costs. Next, the number of people on a policy do not show a meaningful relation with financial value. This variable can be used as proxy to distinguish families with children and individually insured customers or couples. The number of years a customer is insured has no effect on the monetary balance of that person, except in component 9 where it is negative. The number of people on a collectivity is also not important. Mental health care region is in general not influential while the nursing and care

region shows a negative effect. This region is an ordered variable from 1 to 10 with 10 being the most expensive, hence this result is as expected.

Second, we interpret insurance package related variables. Having a main (basic coverage) insurance package has a strong positive effect in all cases, as opposed to being insured without main insurance package. These individuals have concluded main insurance at another provider as they are obligated to by law. These customers decide to purchase additional coverage at an alternative provider. An explanation is that a comparable coverage at the provider of their main insurance is less attractive in terms of features or costs. This implicates that customers who have a main insurance at a competitor are less attractive for the company from a monetary point of view. Next, we find having additional packages has a positive impact, except in *8* where it has no effect and segment 10 where the effect negative. In addition, a more expensive additional insurance yields a positive effect. Furthermore, having extra modules such as a dental coverage has a positive impact in general. In each segment we find having foreign insurance coverage has a negative effect in comparison with having no foreign coverage, this is most notable in *1*. Having an insurance of brand two has a negative impact in general, except in *1*. Lastly, information regarding use of a insurance comparison site is not useful. Individuals that use such a site could be labeled as more price sensitive, as they take effort to find the best suited or least expensive insurance provider but this is not reflected in the financial balance

Third, we discuss monetary variables related to insurance packages. The amount of voluntary excess is insignificant in all components which is unexpected. Intuitively, the chosen amount of voluntary excess would have explanatory power regarding the level of health care services required by an individual. Interestingly, this voluntary excess appears to be unrelated with the resulting balance of a customer. This result is most likely explained by the fact that choosing a higher amount of voluntary deductible excess is compensated by a lower priced insurance. Therefore, we expect customers with a lower amount of excess to require more health care services which is then compensated by a higher premium for their package. Paying a provisional service fee to a third party, the term of payment and the number of payment neglects do not have any effect.

Lastly, we look at the behavior of customers regarding claims. As expected, the number of claims has a negative effect on the monetary balance. However, the number of different months containing claims does not appear to be of importance in general, except in *1* where the effect is positive. The number of leniences provided to customers on their claims is negatively related with balance in segment 1 and to a lesser extent in *6*. Lenience is provided by the company in certain situations where the individual is not (completely) insured for the treatment or service he is claiming. The number of different health care categories in which claims are done has a negative effect in general but not in all groups. Lastly, we find a positive effect in most segments for the number of claims in the health category containing the maximum number of claims (MAX_N_CATS). If the number of claims is high in a single specific health care category this could implicate a more structural or expected requirement of health services allowing for insuring against these costs. This is as opposed to for instance an unpredictable accident resulting in high cost treatment.

To conclude this section we recap the most important findings. Results indicate that age does not have a negative relation with our response variable. This is likely explained by the settlement we have taken into account in our response variable. Second, not having a main insurance package has a relatively large negative effect on the monetary balance of an individual. The same result holds for having insurance packages with foreign coverage. On the contrary, holding more packages such as additional insurances and modules like dental coverage is positively related with ones balance. These customers have insurances with better coverage negating the impact of costly health care services on their financial balance. Next, the number of claims is negatively correlated with balance as expected. In contrast, the number of different months in which claims are done is not important. We find segment *1* is an exception in many of the general relations revealed by the model. Correspondingly, we find the largest number of variables is selected in this group. This finding is in agreement with the values observed in the pairwise KL divergence of component 1 to the others in Table 6.

Table 9 Component wise means and standard deviations

	Component										
Variable	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	**Population**
BALANCE	-20727.22 (55022.68)	4950.27 (1508.76)	1943.7 (3645.78)	2262.17 (1724.96)	2617.39 (1584.69)	3287.43 (2250.78)	3089.98 (467.98)	728.68 (4335.05)	-3890.7 (9262.19)	2144.78 (2038.15)	1744.15 (5648.42)
AGE	54 (29)	58 (22)	49 (24)	40 (23)	43 (24)	52 (26)	40 (15)	43 (22)	50 (21)	35 (22)	41 (23)
SEX	0.55 (0.5)	0.46 (0.5)	0.54 (0.5)	0.43 (0.5)	0.49 (0.5)	0.59 (0.49)	0.46 (0.5)	0.45 (0.5)	0.5 (0.5)	0.56 (0.5)	0.51 (0.50)
N_YEARS	8.51 (3.61)	9.09 (3.1)	8.34 (3.46)	7.98 (3.59)	8.27 (3.58)	8.86 (3.29)	8.03 (3.7)	7.99 (3.71)	8.21 (3.7)	7.55 (3.71)	7.97 (3.66)
MAIN	1 (0.02)	1 (0.01)	0.99 (0.1)	1 (0.07)	1 (0.04)	1 (0.04)	0.99 (0.08)	1 (0.05)	1 (0.04)	1 (0.06)	1.00 (0.06)
FOREIGN_IND	0.01 (0.1)	0 (0.06)	0.02 (0.16)	0.02 (0.13)	0.01 (0.1)	0.01 (0.08)	0.01 (0.08)	0.01 (0.1)	0.01 (0.09)	0.01 (0.07)	0.01 (0.09)
$BRAND_1$	0.78 (0.41)	0.78 (0.41)	0.76 (0.42)	0.79 (0.41)	0.8 (0.4)	0.77 (0.42)	0.77 (0.42)	0.78 (0.42)	0.78 (0.41)	0.75 (0.43)	0.77 (0.42)
$BRAND_2$	0.1 (0.29)	0.11 (0.32)	0.14 (0.34)	0.1 (0.3)	0.1 (0.3)	0.12 (0.32)	0.11 (0.31)	0.12 (0.32)	0.1 (0.3)	0.12 (0.33)	0.11 (0.32)
$BRAND_3$	0.12 (0.33)	0.1 (0.31)	0.1 (0.3)	0.1 (0.31)	0.1 (0.3)	0.11 (0.31)	0.12 (0.33)	0.11 (0.31)	0.12 (0.32)	0.13 (0.33)	0.12 (0.32)
TAKER	0.61 (0.49)	0.66 (0.47)	0.53 (0.5)	0.49 (0.5)	0.47 (0.5)	0.57 (0.5)	0.52 (0.5)	0.52 (0.5)	0.61 (0.49)	0.42 (0.49)	0.49 (0.50)
N_ON_POLICY	2.19 (1.28)	2.14 (1.2)	2.49 (1.41)	2.71 (1.41)	2.73 (1.39)	2.45 (1.31)	2.8 (1.5)	2.62 (1.4)	2.37 (1.34)	3.03 (1.48)	2.77 (1.45)
VOLUNT_EXCESS	22 (96)	23 (100)	47 (142)	63 (160)	55 (151)	28 (108)	86 (183)	76 (175)	42 (133)	64 (162)	62 (159)
PAYMENT_TERM	3.91 (4.64)	3.7 (4.53)	3.77 (4.61)	3.8 (4.63)	3.78 (4.6)	3.91 (4.65)	3.72 (4.58)	3.92 (4.69)	3.9 (4.67)	3.77 (4.61)	3.80 (4.62)
N_ON_COLLECT	5099 (9232)	4842 (8679)	5013 (8737)	5107 (8980)	4929 (8706)	4942 (8821)	5463 (9505)	4933 (8968)	4938 (9188)	5510 (9414)	5249.13 (9207)
IND_COLLECT	0.26 (0.44)	0.23 (0.42)	0.21 (0.41)	0.23 (0.42)	0.21 (0.41)	0.21 (0.41)	0.19 (0.39)	0.2 (0.4)	0.23 (0.42)	0.16 (0.37)	0.19 (0.39)
$REGION_{GGZ}$	5.26 (2.96)	5.35 (2.9)	5.56 (2.99)	5.23 (2.98)	5.5 (2.94)	5.62 (2.88)	5.24 (3.04)	5.36 (2.95)	5.44 (2.97)	5.59 (2.98)	5.46 (2.98)
$REGION_{VV}$	1.7 (1.94)	1.62 (1.93)	1.91 (2.03)	1.58 (1.91)	1.59 (1.92)	1.67 (1.94)	1.57 (1.91)	1.66 (1.91)	1.7 (1.92)	1.64 (1.94)	1.63 (1.93)
PROVISION	39 (65)	36 (37)	35 (43)	29 (40)	25 (38)	28 (36)	34 (33)	35 (44)	41 (42)	25 (31)	30 (36)
N_CLAIMS	23 (14)	20 (8)	23 (15)	17 (10)	17 (9)	18 (8)	14 (7)	20 (11)	22 (11)	15 (9)	17.01 (9.55)
N_LENIENCE	0 (0.06)	0 (0.07)	0.01 (0.12)	0 (0.06)	0 (0.04)	0 (0.04)	0 (0.04)	0 (0.05)	0 (0.06)	0 (0.04)	0.00 (0.05)
RETND_EXCESS	299 (180)	285 (148)	236 (186)	141 (183)	147 (180)	193 (180)	144 (171)	200 (208)	338 (186)	117 (171)	166 (190)
N_MONTHS	9.84 (2.94)	10.01 (1.91)	9.66 (2.57)	9 (2.36)	9 (2.36)	9.41 (2.21)	8.4 (2.37)	9.62 (2.23)	10.07 (2.04)	8.6 (2.38)	8.98 (2.38)
N_CATEGORIES	4.49 (1.96)	4.34 (1.42)	4.79 (2.17)	3.88 (1.87)	3.84 (1.69)	3.95 (1.48)	3.47 (1.64)	4.43 (1.87)	4.88 (1.69)	3.62 (1.7)	3.89 (1.76)
MAX_N_CATS	9.56 (4.06)	9.31 (2.93)	9.45 (3.67)	8.72 (3.12)	8.58 (3)	8.81 (2.98)	8.13 (2.92)	9.05 (3.09)	9.48 (3.16)	8.3 (2.99)	8.59 (3.06)
ADDITIONAL	1.09 (0.66)	1.08 (0.63)	1.18 (0.66)	1 (0.67)	1.06 (0.61)	1.13 (0.6)	0.89 (0.67)	1.08 (0.64)	1.08 (0.66)	0.98 (0.62)	1.01 (0.64)
MODULE	0.43 (0.49)	0.44 (0.5)	0.42 (0.49)	0.44 (0.5)	0.37 (0.48)	0.34 (0.47)	0.47 (0.5)	0.46 (0.5)	0.5 (0.5)	0.36 (0.48)	0.41 (0.49)
N_NEGLECT	0.11 (0.86)	0.12 (0.87)	0.11 (0.88)	0.08 (0.72)	0.08 (0.73)	0.09 (0.76)	0.1 (0.75)	0.11 (0.83)	0.14 (0.94)	0.08 (0.72)	0.09 (0.77)

Standard deviations are denoted between brackets.

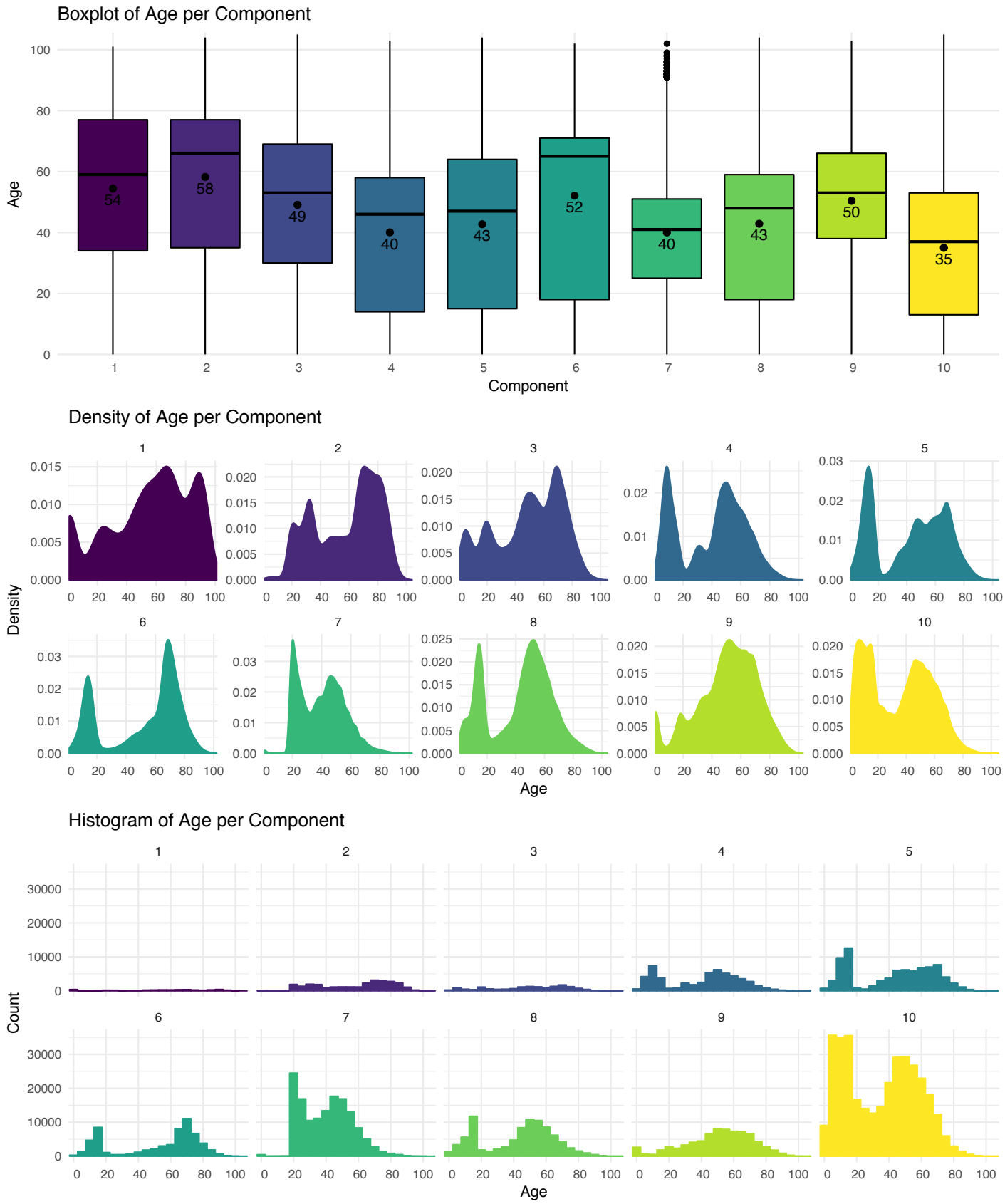

Figure 6 Comparison of age per component with a boxplot (top), density plot (center) and histogram (bottom). Black dots and numbers in the boxplot indicate the component averages, horizontal lines mark the medians.

4.6 SEGMENT INTERPETATION

We now further analyze the segments. First, segment-level results are reported and interpreted. Emphasis is put on the captured differences that could provide support for targeting specific components. The total combination of information needs to be taken into account simultaneously in order to sufficiently create a relative distinction between customers. Interpretation and judgment of a customer profile cannot be done based on a shallow combination of some customer characteristics or without taking other customers into account. This is an important consideration when interpreting or assigning value to a customer segment and exactly the power of applying a finite mixture modeling approach for market segmentation. Second, we discuss the quality of the segmentation based on three general criteria.

Table 9 reports summary statistics of the groups based on the posterior probabilities of the observations. The mean of each variable within the components is given and the standard deviation is denoted between brackets. This table includes all variables because an important distinction is taken into account here. The coefficient estimates reported in Table 8 are based on the relation with our response variable, the monetary balance of an individual. However, certain characteristic of a customer or other features can be of value regardless of the relation with the balance of this individual. This value is not necessarily reflected by means of a direct financial aspect described by our response variable. For instance, we do not find the age of a person or the size of a collectivity to have a meaningful relation with the monetary balance of a customer. It is of importance to remember the goal of this research, which is to support differentiation in groups of customers. From a business point of view, the age of a customer holds value regardless. Younger customers have the highest potential customer lifetime duration, if they are satisfied with the services of the company they can remain a loyal customer for many years. In addition, a specific segment of interesting customers are children, usually below the age of 18, who are still on their parents policy. When the time comes for them to insure their own policy the company is very interested in retaining them as a customer instead of losing them to a competitor. Furthermore, large collectivities can be of more interest than smaller ones, for example a large company that has a insurance deal with the provider for their employees. These interpretations may still hold meaning and be of use without finding a a direct correlation with the response variable. Hence, certain variables that have been excluded by the elastic net in the mixture regression model can still provide relevant information. Moreover, processing claims requires effort and time hence claims are very expensive for the company. Therefore, customers with a more extensive claiming behavior are less desirable on top of the negative effect that claims have on a customers financial balance.

We will not discuss all of the 24 variables in each of the 10 segments. Instead, emphasis is put on the results that provide the most useful information for differentiating the components. First, we look at the balance in Table 9. The results indicate that component *1* is on average the most negative. Figure 8 plots the balance of the different components. Indeed, a large number of customers with a negative balance are included in this segment. Yet, a conflicting finding is that many customers with a relatively high positive balance are also contained in *1*. Further inspection reveals the median balance is actually the highest in *1* of all components. This result can also be seen in the boxplot in Figure 8 where the median of each component is marked with a horizontal line. The dispersion of balance in *1* is clear by looking at the range of the box which contains half of the observations. This finding complicates the practical interpretation of this component as the range of balances is very wide compared to the other groups. Next, *9* also has a negative mean and median balance and does not contain many positive balances. All other segments have a positive mean and median with *2* having the highest financial balance. In addition, we find *8* also includes a large number of customers with a negative balance in comparison with the other groups. Another interesting observation is the standard deviation in *7*, which is noticeably smaller compared to other groups and the total population. The largest difference is observed when comparing the deviation from *1* to *7* which is in agreement with the highest pairwise Kullback-Leibler divergence measure found between this components in Table 6.

We continue with the demographics and general characteristics of the customers. The results reveal that component *2* contains the oldest customers

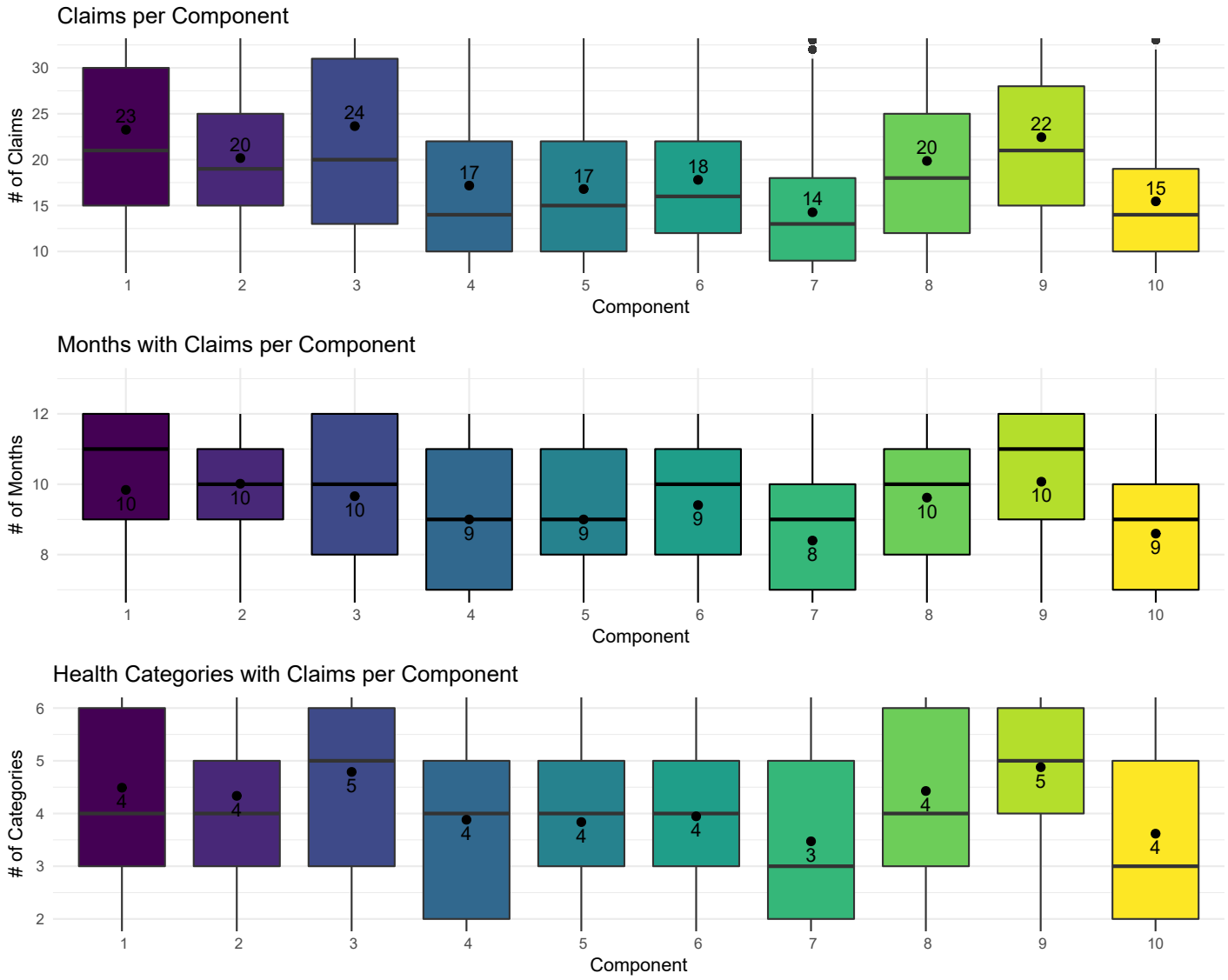

Figure 7 Comparison of claim behavior per component. Averages of number of claims (top), months with claims (middle) and health categories with claims (bottom). Black dots and numbers indicate the component averages, horizontal lines mark the medians.

whereas *10* comprises many young individuals, this can also be seen in Figure 6. A majority of families with young children and young adults are assigned to this group. The density plot of age per component shows that *1* and *2* have the most mass in the older age categories whereas many other groups show a bimodal density where younger individuals are also present. However, the bottom histogram plot shows that in terms of absolute number of customers, component 10 contains the majority of young individuals between the age 0 and 20. This is an important group of customers for the company.

We also find *10* contains fewer insurance takers than for example the older segments *1* and *2*. This can be explained by the fact that many customers below the age of 18, and frequently beyond this age, are included in the same policy as their parents. In this case one of the parents is responsible for this insurance policy. This finding is also reflected in the higher number of people on the same policy, *N_ON_POLICY* in this group compared to others. Component *1* and *2* have a relatively high number of insurance takers in combination with a relatively low number of individuals on the same policy, this indicates that *1* and *2* contain more individually insured customers compared to the population average. This finding is supported by the high value for *IND_COLLECT* in *1* and *2*, and a low value in *10*. These conclusions are intuitive,

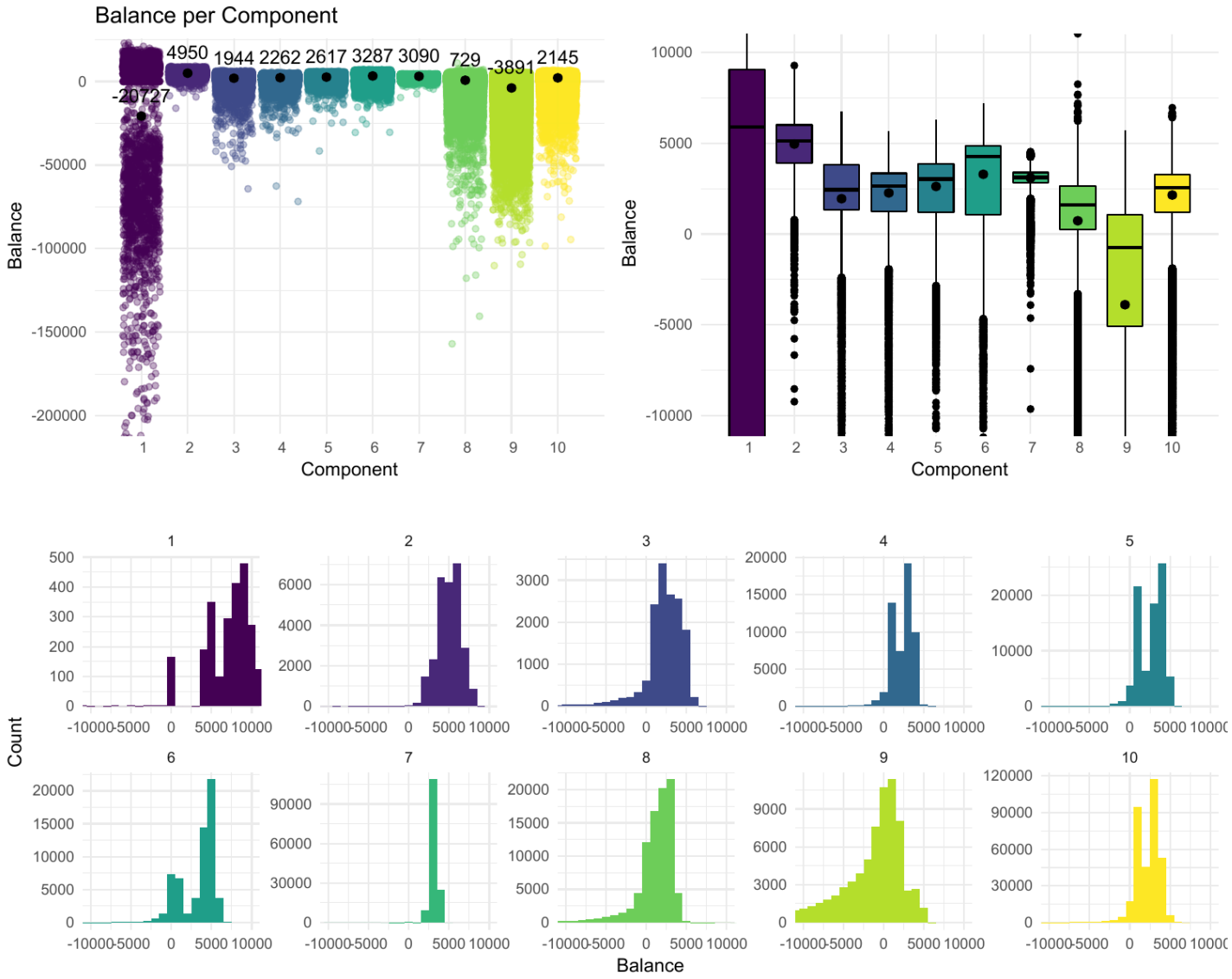

Figure 8 Comparison of balance per component with a jitter plot (top left) and boxplot (top right) and histogram (bottom). Black dots and numbers in the top row indicate the component averages, horizontal lines in the boxplot mark the medians.

older aged couples tend to be on a policy together while younger couples have their children included. This makes component *10* a very attractive market segment for targeted campaigns focusing on younger customers or families with children. Gender is approximately equally divided in general but *4* contains more females and *6* contains more males than the population average.

Third, we discuss monetary variables related to insurance packages. The amount of retained voluntary deductible excess is the highest in *1* and *9* but is also high in segment *2*. On the contrary, the retained amount is the lowest in *10* followed by *7* and *4*. In combination with the amount of voluntary deductible excess, which is set by the customer, we find that individuals in segment *1* expect to have high costs as the amount is chosen to be low. On the contrary we find the amount to be high in *10* as they likely expect to have low costs. However, this relation is not reflected in our regression results in Table 8. This is likely caused by the fact that choosing a higher amount of voluntary deductible excess is compensated by a lower priced insurance and vice versa.

Next, we interpret behavior regarding claims. The number of claims is high in both segments with a negative balance, *1* and *9* but the number is also high in component *3*. In contrast, less claims are made in *10* and *7*. Interestingly, the number of claims is higher than the population average in *2* while the monetary balance is also the highest in this segment. It is

likely these customers expect structural requirement of health care services and have chosen insurance packages accordingly compensating the costs of their treatments or medicine. Thus, meaning a high claim frequency and high costs do not necessarily constitute a less valuable customer than an individual who requires less health services. This finding underlines the importance of objectively taking the entire combination of data into account.

Taking all results into account we conclude that there are four components with desirable properties. First, component 2 which is the oldest group and seem to require structural health care as can be seen from their claim frequency of 20 times in 10 different months during a year and low voluntary deductible excess. Nevertheless, this group of customers simply has the highest average financial value yielding the most profit. This group can likely be labeled as wealthy retirees who are quite aware of their health cost requirements and have contracted appropriate insurance coverage. The premium of their insurance in turn compensates the costs. This component constitutes 3% of the customer base.

Second, component 10 which contains the majority of young customers and families. The claim frequency is the lowest in this group after segment 7. The customers in this group have a high potential to remain a long-time customer if they are happy with the services provided by company. In addition, many children that reach the age of 18 move from their parents insurance policy to their own. It is desirable to retain these individuals which can be achieved by performing targeted campaigns on segment *10*. This group is the most voluminous with 38% of the population.

Third, segment *7*. The financial value is slightly less than 6 but the standard deviation is significantly lower. Above that, the average customer in this group is 12 years younger compared to segment *6* with again the smallest range of all groups. On average these customers claim 14 times in 8 different months during a year. This is the lowest frequency of all segments. In addition, we observe the highest voluntary deductible excess which is generally seen as a proxy for not expecting the requirement of high cost health services. The standard deviation of the majority examined variables is very low compared to the other groups. Furthermore, customers in component *7* posses insurance packages with less additional coverage but more dental coverage in comparison with the population. In addition, the component forms a substantial part of the total population as it contains roughly 16 %.

Fourth, segment is *6* which has the second highest financial value of all groups. With an average age of 52 and median age of 66 it is slightly younger than segment *2* but considerably older than *7*. The average number of claims is 18 times during a year in 9 months which is just above group *7*. Furthermore, we find this group owns additional insurances with more coverage but the least amount of dental coverage compared to the average customer. *6* is also smaller than *7* as it includes 7% of the observations.

Beside the segments with valuable customers we identify two components with less desirable properties. First, segment *9* which contains many negative financial valued individuals and has the lowest median balance and lowest average balance after component *1*. These customers claim on average 22 times in 10 different months within 5 different health care categories during a year. These are the highest number of months and number of different care categories of the entire customer base. Segment *9* comprises 9%.

The second less desirable group is segment 8. The average balance is the lowest after *1* and *9* and the median balance is the lowest after *9*. Figure 8 reveals the inclusion of a large number of customers with a negative balance. Furthermore, we find a high average claim frequency of 20 within 10 months in 4 different health care categories. This group is slightly larger than *9* containing 10%.

Lastly, component *3*, *5* and *6* have not been discussed yet. These groups are in general average when comparing their properties to the population. However, noteworthy diverging results are firstly the extensive claiming behavior in group *3* which is costly for the company seen in Figure 7. Secondly, the high median age of 65 in component *6*, seen in Figure 6. Together they constitute approximately 16% of the total sample. Not all groups are easily interpreted. We find segment *1* has the most exceptions regarding the general relations in the regression. Which is also confirmed by the pairwise KL measures. This component requires 21 of the 26 variables in order to explain the variance in the response variable which is the largest amount. Component *1* is a bit more complicated to assess due to the very wide range of balances contained in this group. The majority of very negative financial valued customers are included

together with a large number of high financial value customers. Only 0.5% of the observations are in this segment and merely 2000 of these customers have a positive balance. The average balance is by far the lowest of all components due to the large number of highly negative balances whereas the median is actually the highest of all groups. This complication is not solved by fitting a mixture model with more components. For example, an 11 and 12-component solution still included a component with extreme diverging financial balances. Hence, we conclude that the data of these individuals must be highly comparable while their balance is evidently not.

Taking all results into account we conclude that there are four components with varying desirable properties. First, segment *2* which is roughly 3% of the population and has the most attractive financial balance yielding the highest profit per customer. Second, segment *7* is the most specific group due small deviations in it's properties compared to other groups. This allows for targeting a specific profile of customers fitting in this group. Conveniently, this component has a substantial size of 16% and very desirable properties such as the lowest number of claims in the least number of different health categories. Third, segment 6 which is comparable to *7* but with more deviation, slightly more claims and smaller as it concerns around 7% of the observations. Lastly and fourth, segment *10* which contains the majority of families including young customers and constitutes 38% of the sample. On the contrary, two components contain less desirable customers. Both segment *8* and *9* have a negative financial balance due to inclusion of many customers that are costly to insure for the company. In addition, *8* and *9* have the most extensive claiming behavior. Processing claims is time consuming ultimately adding to the financial resources required for insuring these customers. These components together form approximately 19% of the population. The remaining components *3*, *5* and *6* can be labeled as average based on their properties and constitute roughly 18% of the customer base.

4.6.1 SEGMENT QUALITY

In this section we try to validate the results from a more practical point of view. The aim is to evaluate our segmentation in terms of quality and applicability. Bluntly stated, does the outcome make sense and is it useful? Segment quality is a very vague term as it is highly dependent on the purpose and opinion of the interpreter. Therefore, in order to allow evaluation of the quality of the obtained segmentation we consider three rather general criteria: identifiability, substantially and actionability (Wedel and Kamakura, 2012).

- **Identifiability:** Does the segmentation reveal significant variation across the defined components?

The different components describe an adequate number of variation to interpret customers assigned to that group. This does not hold for each included variable in the data, some features do not show useful distinction across components. For example, the size of collectivities is relatively stable in each group. However, a multiple of variables that can be used to interpret the components do show meaningful variation. A good example is the age within groups. We find individuals in component *2* to be old on average while the majority of young customers are included in component *10*. Hence, a sufficient amount of variation is present in the grouping structure to properly create distinction between groups and differ targeting to specific segments as desired. However, we find a large amount of overlap between the components indicating that the groups are hard to separate. An ideal solution would consist of perfectly separated components.

- **Substantiality:** Are the segments large enough in size to allow targeting?

Component *1* contains a relative small number of observations compared to the population size. All other groups contain at least 15,000 observations. This seems to be a sufficient amount depending on the specific marketing action to be taken. In addition, some marketing campaigns may be applicable for a multiple of segments. For instance, one possibility is to offer a discount to all segments with a relatively young age and good financial value in order to increase customer loyalty. Segment *7*, *4* and *5* would all qualify for such a campaign. In this scenario, the three segments can easily be pooled to increase the number of targeted customers as desired.

- **Actionability:** Is the variation across segments interpretable and does it provide guidance?

In other words, can insights be acted upon to improve business? The grouping structure can be employed

to target a subpopulation of customers. Key drivers that show variation across the components include financial balance, age, claim behavior and package selection. All of which can be used to support a distinction in resource allocation or targeted marketing depending on the campaign of choice. These revealed structural differences in the properties of the segments can be used to select or target specific groups of customers which seem appropriate for the action to be taken.

5 SIMULATION STUDY

Lastly, we compare the performance of Gaussian mixture regression models with different penalization methods by means of a simulation study. The included modeling approaches are:

- **MIXREG**: Regular Gaussian mixture model without variable selection
- **MIXRDG**: Gaussian mixture model combined with the ridge penalty function
- **MIXNET**: Gaussian mixture model combined with the elastic net penalty function
- **MIXLAS**: Gaussian mixture model combined with the lasso penalty function
- **MIXALS**: Gaussian mixture model combined with the adaptive lasso penalty function

The penalty specifications are given in Equation 35 for ridge, Equation 39 for the elastic net, Equation 36 for the lasso and Equation 37 for the adaptive lasso. For the elastic net we consider two penalty mixing proportions, $\alpha = 0.5$ and $\alpha = 0.9$. The choice of alpha is indicated with a subscript. $\alpha = 0.9$ results in stricter regularization as it tends more towards the lasso than the ridge penalty. For completeness we list the (penalized) log-likelihood function of each tested modeling approach. In MIXREG we do not add a penalty term. For MIXALS the coefficient dependent weights $\hat{w}_{sj}$ are obtained through a preliminary ridge regression.

- **MIXREG**:

$$\mathcal{L}(\Theta) = \underbrace{\sum_{i=1}^{N} \log \sum_{s=1}^{S} \pi_s \cdot f(y_i|x_i, \theta_s)}_{\text{Log-Likelihood}} . \tag{48}$$

- **MIXRDG**:

$$\widetilde{\mathcal{L}}(\Theta) = \underbrace{\sum_{i=1}^{N} \log \sum_{s=1}^{S} \pi_s \cdot f(y_i|x_i, \theta_s)}_{\text{Log-Likelihood}} - \underbrace{\sum_{s=1}^{S} \lambda_s \sum_{j=1}^{p} |\beta_{sj}|^2}_{\text{Penalty}} . \tag{49}$$

- **MIXNET$_5$**:

$$\widetilde{\mathcal{L}}(\Theta) = \underbrace{\sum_{i=1}^{N} \log \sum_{s=1}^{S} \pi_s \cdot f(y_i|x_i, \theta_s)}_{\text{Log-Likelihood}} - \underbrace{\sum_{s=1}^{S} \lambda_s \sum_{j=1}^{p} \left(0.5|\beta_{sj}| + 0.25|\beta_{sj}|^2\right)}_{\text{Penalty}} . \tag{50}$$

- **MIXNET$_9$**:

$$\widetilde{\mathcal{L}}(\Theta) = \underbrace{\sum_{i=1}^{N} \log \sum_{s=1}^{S} \pi_s \cdot f(y_i|x_i, \theta_s)}_{\text{Log-Likelihood}} - \underbrace{\sum_{s=1}^{S} \lambda_s \sum_{j=1}^{p} \left(0.9|\beta_{sj}| + 0.05|\beta_{sj}|^2\right)}_{\text{Penalty}} . \tag{51}$$

- **MIXLAS**:

$$\widetilde{\mathcal{L}}(\Theta) = \underbrace{\sum_{i=1}^{N} \log \sum_{s=1}^{S} \pi_s \cdot f(y_i|x_i, \theta_s)}_{\text{Log-Likelihood}} - \underbrace{\sum_{s=1}^{S} \lambda_s \sum_{j=1}^{p} |\beta_{sj}|}_{\text{Penalty}} . \tag{52}$$

- **MIXALS**:

$$\widetilde{\mathcal{L}}(\Theta) = \underbrace{\sum_{i=1}^{N} \log \sum_{s=1}^{S} \pi_s \cdot f(y_i|x_i, \theta_s)}_{\text{Log-Likelihood}} - \underbrace{\sum_{s=1}^{S} \lambda_{sj} \sum_{j=1}^{p} \hat{w}_{sj} |\beta_{sj}|}_{\text{Penalty}} . \tag{53}$$

A highly desirable property of regularization algorithms is shrinking the influence of the least important variables. Or in case of algorithms with variable selection abilities such as the elastic net, we wish to select the optimal subset of variables while excluding the non-influential variables from the model by shrinking their effect down to zero. The elastic net and lasso approaches both have feature selection abilities while the regular and ridge approach yield non-zero estimates for all coefficients. Secondly, besides selecting the correct subset of variables we wish to obtain accurate estimate of the effects of the non-zero coefficients. Thirdly, another fundamental challenge in mixture modeling is determining the optimal amount of components to describe the data. The number of components is in general unknown and must be extracted from the data. Hence, the ability to select the correct amount of components is also investigated.

In short, we compare performance of the different modeling approaches with a simulation in which three aspects are examined:

- Selection of correct subset of variables
- Accuracy of non-zero coefficient estimates
- Recovery of true component amounts

We specify the following general 2-component finite mixture form to generate a response variable y

$$\pi \cdot \phi\left(y_1; x^\top \beta_1, \sigma^2\right) + (1-\pi) \cdot \phi\left(y_2; x^\top \beta_2, \sigma^2\right) \quad (54)$$

with $\sigma^2 = 1$. Three different prior probabilities are tested, $\pi_1 = \{0.15, 0.3, 0.6\}$ implying $\pi_2 = 1 - \pi_1 = \{0.85, 0.7, 0.4\}$. The covariates x are generated from a multivariate normal distribution with mean 0, variance 1 and a correlation structures ρ_{ij} such that

$$x \sim \mathcal{N}\left(\mu = 0, \sigma^2 = 1\right) \quad (55)$$

with $\rho_{ij} = cor(x_i, x_j) = 0.6^{|i-j|}$.

Next, we use Equation 54 to define two different models M_1 and M_2 to simulate a set of data. The component specifications of both models are given in Table 10. The first model, M_1, contains $p = 10$ covariates of which 5 zero coefficients in component 1 and 6 in component 2. The second model, M_2, presents a higher-dimensional and more realistic variable selection problem and includes $p = 25$ covariates. Component 1 contains 10 zero coefficients while component 2 contains 15 zero coefficients. Hence, in both models component 2 contains more zero coefficients than component 1. In each case, a sample size of $N = 100$ observations is used.

A widely used performance metric is the hit-rate. Hit-rate is simply the ratio of correct predictions to the total of observations. Indeed, this is an intuitive and effective measure when dealing with symmetric data. However, in case of an unbalanced class distribution the hit-rate may fail to provide a proper indication of performance. In order to compare the detection of true zero coefficients we consider the following metrics; precision (specificity), recall (sensitivity), and F1 score. We define precision as the ratio of correctly estimated zero coefficients (true positives) to the total estimated number of zero coefficients (true positives and false positives) such that

$$\text{Precision} = \frac{\text{TP}_0}{\text{TP}_0 + \text{FP}_0}. \quad (56)$$

Next, recall is given by the ratio of correctly estimated zero coefficients to the true number of zero coefficients defined as

$$\text{Recall} = \frac{\text{TP}_0}{\text{TP}_0 + \text{FN}_0}. \quad (57)$$

The F1 score (Van Rijsbergen, 1979) is the weighted average of precision and recall given by

$$\text{F1} = \frac{2 \cdot (\text{Recall} \times \text{Precision})}{(\text{Recall} + \text{Precision})}. \quad (58)$$

A flawless performance would results in a ratio of 1 for all three metrics. By combining the precision and recall we can take true and false positives and negatives into account simultaneously. This allows us to quickly compare the subset selection performance of each model and specification with a single metric. Each scenario is repeated for 100 iterations.

The component-wise results for the detection of zero coefficients are reported in Table 11. We exclude the MIXREG and MIXRDG in this part of the simulation as they both do not have variable selection abilities. The numbers are rounded to four decimal points. In case of an unbalanced mix of the component sizes for $\pi_1 = 0.15$, the first component contains a small amount of observations. In general, the performance drops when the amount of observations in the second component decreases. There is no universal best method. In case of the higher-dimensional variable

selection problem in M_2 MIXREG would fail to obtain a solution for $\pi_1 = 0.15$. In this scenario 25 coefficients are to be estimated based on approximately 15 observations which is not possible with the regular likelihood approach. All other models have no estimation issues in this scenario when $p > N$. This is a very attractive property of the penalized likelihood approaches.

There is no indication of a single overall superior shrinkage algorithm in this part of simulation. In general we find that MIXNET performs well when the amount of observations in component 2 is larger. The lasso based models perform better when the amount of observations in component 2 decreases. In the higher dimensional problem in model M_2 MIXALS is the best method when component 2 contains little observations. This situation is the most challenging in terms of selecting the correct subset of variables.

Table 10 Simulation model specifications M_1 and M_2.

Parameters	**Model M**$_1$ $(p = 10)$	**Model M**$_2$ $(p = 25)$
$\beta_{s=1}$	(2, -0.8, 1, 0, 0, 1.2, 0, 0, 1.2, 0)	(0, 2, -24, 1, 0, 3, 15, 22, -5, 28, 0, 0, 14, 29, 0, 0, 19, -6, 0, 21, 31, 0, 0, -19, 0)
$\beta_{s=2}$	(0, 0, 0, 1, 2, 0, 0, -1.5, 0, 1.2)	(-6, 0, 0, 15, 0, 0, 0, 8, 0, 22, 0, -3, 0, 17, 0, 0, 5, 0, 13, 0, 0, -19, 0, 0, 1)
ρ_{ij}	$0.6^{\lvert i-j\rvert}$	$0.6^{\lvert i-j\rvert}$
π_1	0.15, 0.3, 0.6	0.15, 0,3, 0.6

Table 11 Detection of zero coefficients based on 100 simulation repetitions.

	Model M$_1$ $(p = 10)$						**Model M**$_2$ $(p = 25)$					
	Component 1			Component 2			Component 1			Component 2		
Method	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
$\pi_1 = 0.15$												
MIXNET$_5$	.8202	.9400	**.8745**	.8333	1.000	**.9091**	.7271	.8473	.7809	.9487	.9832	.9655
MIXNET$_9$	.8220	.9340	.8714	.8333	1.000	**.9091**	.7327	.8513	**.7861**	.9493	.9853	.9668
MIXLAS	.8430	.7640	.7846	.8333	1.000	**.9091**	.7790	.7933	.7829	.9995	.9432	**.9697**
MIXALS	.8163	.9340	.8698	.8290	.9960	.9048	.7315	.8513	.7854	.9483	.9842	.9658
$\pi_1 = 0.3$												
MIXNET$_5$	.8188	.9760	.8898	.8246	.9920	.9004	.7886	.8820	.8307	.9292	.9268	.9256
MIXNET$_9$	.8247	.9740	.8925	.8289	1.000	.9062	.8013	.8900	**.8417**	.9253	.9347	**.9289**
MIXLAS	.8380	.9040	.8661	.8307	.9980	**.9067**	.8372	.8020	.8153	.9667	.8489	.9006
MIXALAS	.8303	.9820	**.8993**	.8333	1.000	.9091	.7884	.8740	.8265	.9178	.9053	.9096
$\pi_1 = 0.6$												
MIXNET$_5$	.8217	.9900	.8978	.8293	.982	.8980	.9073	.8826	.8904	.8343	.9067	.8675
MIXNET$_9$	.8217	.9840	.8954	.8259	.994	**.9019**	.8936	.8605	.8731	.8170	.8967	.8536
MIXLAS	.8260	.9880	**.8996**	.8266	.970	.8912	.9471	.7953	.8610	.8539	.8373	.8427
MIXALS	.8213	.9880	.8962	.8292	.986	.9003	.9038	.8868	**.8920**	.8322	.9153	**.8704**

Best results per component marked in bold.

Next, the same simulation setup is used to study the accuracy of the non-zero coefficient estimates. In order to compare the behavior of the tested models we look at several error metrics. We consider the mean absolute error (MAE), mean squared error (MSE) and root mean squared error (RMSE) of the coefficient estimates. These metrics are calculated based on the p true coefficient estimates β as reported in Table 10 and the p estimated coefficient estimates $\widehat{\beta}$ resulting from the tested models. As reported, we have $p = 10$ covariates in model M_1 and $p = 25$ covariates in model M_2. The used metrics are formulated as

$$\begin{aligned} \text{MAE} &= \frac{1}{p}\sum_{j=1}^{p}|\beta_j - \widehat{\beta}_j|, \\ \text{MSE} &= \frac{1}{p}\sum_{j=1}^{p}\left(\beta_j - \widehat{\beta}_j\right)^2, \\ \text{RMSE} &= \sqrt{\frac{1}{p}\sum_{j=1}^{p}\left(\beta_j - \widehat{\beta}_j\right)^2}. \end{aligned} \tag{59}$$

Every iteration of the simulation results in an error value. Hence, for a more convenient comparison of the models we summarize the metrics by reporting the mean over the 100 simulation repetitions. This yields a single value for each used error metric. This means we report the average of the errors over all $n = 100$ iterations such that

$$\begin{aligned} \overline{\text{MAE}} &= \frac{1}{n}\sum_{i=1}^{n}\text{MAE}_i, \\ \overline{\text{MSE}} &= \frac{1}{n}\sum_{i=1}^{n}\text{MSE}_i, \\ \overline{\text{RMSE}} &= \frac{1}{n}\sum_{i=1}^{n}\text{RMSE}_i. \end{aligned} \tag{60}$$

The component-wise results for the accuracy of non-zero coefficient estimates are reported in Table 12. The numbers are rounded to four decimal points. Figure 9 visualizes the average error and standard deviation over the iterations for model M_1 and Figure 10 for model M_2. The standard deviation of each metric over the repetitions are shown graphically with an error bar. Note that the scales of the error on the y-axis differ per component and prior. RDGMIX is clearly the least accurate method in both simulation models. MIXREG performs well in estimating the coefficients in all cases for the easier problem M_1. The penalized likelihood approaches come with the price of introducing a bias in the estimates which is reflected in this simulation. The difference between the performance of MIXREG and the penalized methods decrease when the problem becomes more difficult. That is, when the amount of observations belonging to component 2 decreases. Note that the range of the error on the y-axis differs per choice of prior π_s. This finding proves even more substantive when the variable selection problem is complicated further by increasing the amount of zero- and non-zero covariates from $p = 10$ to $p = 25$ in M_2 while still using $N = 100$ observations. In general, the deviation of the MIXREG is now considerably larger than the penalized approaches. In this case we find that performance of the lasso and elastic net models approach the MIXREG. Again, MIXALS provides the most accurate solution in the most difficult case for $\pi_1 = 0.6$. This exceptional performance compared to all other tested models is likely explained by the fact that the adaptive lasso possesses the oracle property as discussed in Section 3.8.2 (Zou, 2006).

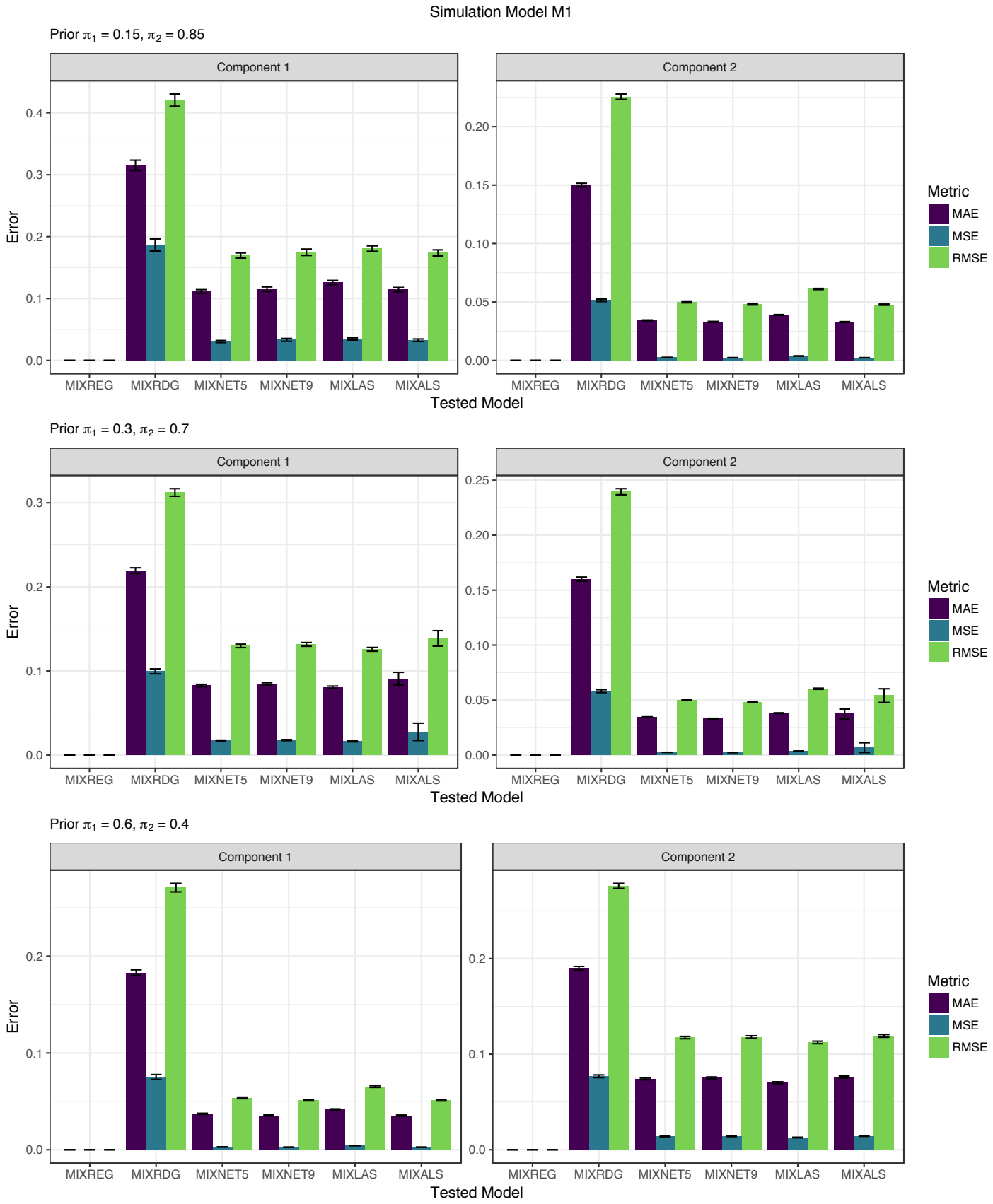

Figure 9 Comparison of the MAE, MSE and RMSE of the coefficients estimated per component by each tested mode in simulation model M_1. Black error bars indicate the standard deviation.

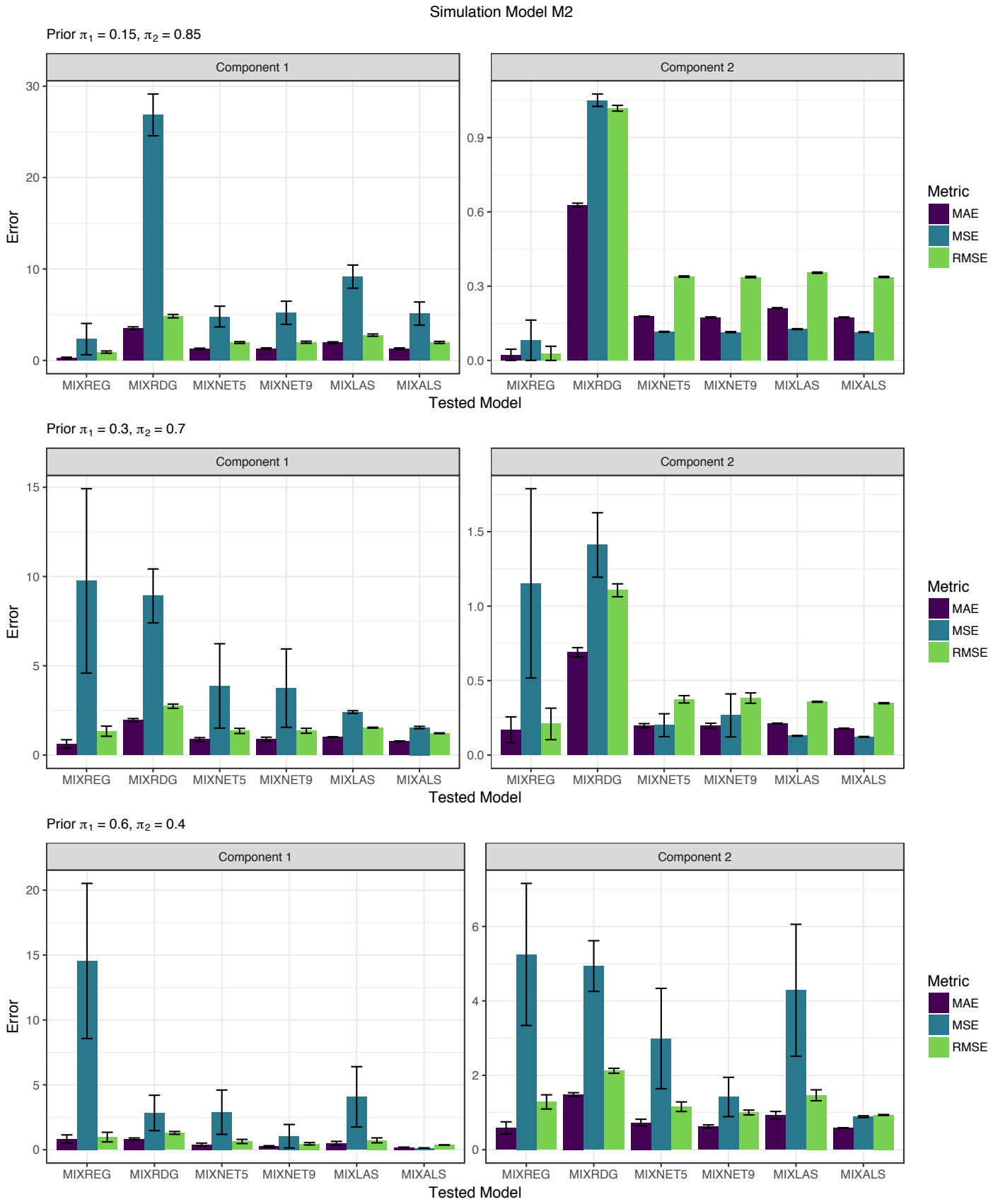

Figure 10 Comparison of the MAE, MSE and RMSE of the coefficients estimated per component by each tested model in simulation model M_2. Black error bars indicate the standard deviation.

Table 12 accuracy of non-zero coefficients based on 100 simulation repetitions.

	Model $\mathbf{M}_1$ ($p = 10$)						**Model $\mathbf{M}_2$** ($p = 25$)					
	Component 1			Component 2			Component 1			Component 2		
Method	$\overline{\text{MAE}}$	$\overline{\text{MSE}}$	$\overline{\text{RMSE}}$	$\overline{\text{MAE}}$	$\overline{\text{MSE}}$	$\overline{\text{RMSE}}$	$\overline{\text{MAE}}$	$\overline{\text{MSE}}$	$\overline{\text{RMSE}}$	$\overline{\text{MAE}}$	$\overline{\text{MSE}}$	$\overline{\text{RMSE}}$
$\pi_1 = 0.15$												
MIXREG	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.2524**	**2.3295**	**0.9078**	**0.0226**	**0.0814**	**0.0285**
MIXRDG	0.3152	0.1865	0.4205	0.1501	0.0514	0.2256	3.5306	26.8569	4.8393	0.6277	1.0511	1.0186
MIXNET_5	0.1114	0.0305	0.1695	0.0342	0.0025	0.0497	1.2665	4.7945	1.9511	0.1779	0.1154	0.3387
MIXNET_9	0.1152	0.0332	0.1748	0.0330	0.0023	0.0479	1.2928	5.2074	1.9865	0.1741	0.1143	0.3367
MIXLAS	0.1259	0.0346	0.1807	0.0389	0.0038	0.0611	1.9363	9.1635	2.7603	0.2111	0.1262	0.3542
MIXALS	0.1145	0.0326	0.1737	0.0328	0.0023	0.0477	1.9363	5.1196	1.9652	0.1741	0.1141	0.3370
$\pi_1 = 0.3$												
MIXREG	**0.0000**	**0.0000**	**0.0000**	0.0000	**0.0000**	**0.0000**	0.6190	9.7534	1.3362	**0.1701**	1.1529	**0.2091**
MIXRDG	0.2193	0.0995	0.3123	0.1600	0.0581	0.2395	1.9523	8.9116	2.7358	0.6898	1.4105	1.1061
MIXNET_5	0.0827	0.0173	0.1298	0.0345	0.0025	0.0501	0.8724	3.8717	1.3509	0.1969	0.2004	0.3742
MIXNET_9	0.0845	0.0178	0.1316	0.0331	0.0023	0.0481	0.8888	3.7490	1.3590	0.1964	0.2664	0.3825
MIXLAS	0.0806	0.0163	0.1257	0.0381	0.0037	0.0603	1.0077	2.4079	1.5330	0.2125	0.1292	0.3578
MIXALAS	0.0908	0.0276	0.1388	0.0374	0.0068	0.0541	**0.7759**	**1.5442**	**1.2197**	0.1790	**0.1225**	0.3483
$\pi_1 = 0.6$												
MIXREG	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	0.8306	14.5471	0.9749	0.5863	5.2478	1.2853
MIXRDG	0.1830	0.0752	0.2708	0.1899	0.0769	0.2760	0.8307	2.8373	1.2989	1.4831	4.9359	2.1207
MIXNET_5	0.0371	0.0029	0.0535	0.0741	0.0139	0.1174	0.3936	2.8946	0.6387	0.7319	2.9880	1.1563
MIXNET_9	0.0354	0.0027	0.0512	0.0752	0.0141	0.1180	0.2518	1.0416	0.4594	0.6266	1.4171	1.0005
MIXLAS	0.0417	0.0043	0.0653	0.0702	0.0128	0.1123	0.4948	4.0784	0.7244	0.9285	4.2848	1.4632
MIXALS	0.0353	0.0027	0.0511	0.0760	0.0144	0.1191	**0.1911**	**0.1379**	**0.3691**	**0.5834**	**0.8880**	**0.9342**

Best results per component marked in bold.

Lastly, we consider the performance of the models in terms of recovering the true number of components in the mixture, this is also known as order selection. We define a S-component Gaussian mixture model where we can vary the number of components as

$$\sum_{s=1}^{S} \pi_s \cdot \phi(y_s, x^\mathsf{T}\beta_s, \sigma^2), \tag{61}$$

with $\sigma^2 = 1$. We generate random priors $\pi_s = \{\pi_1, \ldots, \pi_S\}$ by splitting the value 1 into S parts based on a binomial distribution with a restriction on π_s to ensure $\sum_{s=1}^{S} \pi_s = 1$. The p regression coefficients $\beta_s = \{\beta_{s1}, \ldots, \beta_{sp}\}$ per component are drawn from a uniform distribution such that

$$\beta_{sj} \sim \mathcal{U}(-3, 3) \;\; \forall s = 1, \ldots, S, \; \forall j = 1, \ldots p. \tag{62}$$

The covariates x are generated from a multivariate normal distribution with mean 0, variance 1 and correlation structure $\rho_{ij} = 0.6^{|i-j|}$ as described above. In order to resemble a problem where variable selection is of importance we set all regression coefficients with an absolute value smaller than 0.5 to zero. This results in a varying amount of zero-coefficients per component, generally zero to three. We use this framework to simulate a mixture with varying component amounts $S = \{2, 4, 6, 8, 10, 15\}$ and test each modeling approach. A stepwise component selection procedure is used as explained in Section 3.5 of the Methodology. In short, we fit each model starting with the following initial amount of components $\widehat{S} = \{1, 2, 5, 10, 15\}$ and select the best solution based on the BIC measure. To somewhat decrease computational intensity a limit of 100 iterations is used in the EM algorithm. If the prior probability of a component falls below the value of 0.05 it is removed from the solution after which the EM algorithm continues fitting with $\widehat{S} - 1$ components. This allows the algorithms to perform component selection. Consequently, we compare the performance of the models in terms of selecting the true amount of components present based on the data. Again 100 repetitions are performed. We report the average amount of determined components present in the mixture $\overline{\widehat{S}} = \frac{1}{n}\sum_{i=1}^{n} \widehat{S}$ to examine the solutions over the repetitions. In addition, we look at the hit-rate of the determined amount and true amount of present components, $\widehat{S} = S$. Lastly, the ratio of converged solutions over the repetitions is also reported. A ratio of 1 indicates all solutions over the $n = 100$ repetitions converged.

The results of the component selection simulation are reported in Table 13. The numbers are rounded to two decimal points. In general we observe accurate performance when the amount of components in the mixture is small. When the amount of true components in the mixture grows, all approaches prove ineffective in detecting the correct amount. The performance of all models drop in each step from $S = 6$ onwards. This is possibly due to the fact that the number of observations per component decreases when S increases. Very noticeable is the accuracy of the adaptive lasso model in this aspect of the simulation. MIXLAS yields the most accurate result when the amount of true components in the mixture increases. For $S = 8, 10$ and 15 it is the only method that manages to approach the true component amount while all other techniques yield worse results. Even the adaptive lasso is clearly outperformed by the normal lasso in this comparison. Again, we find that MIXRDG does not perform well compared to the other methods. Interestingly, MIXRDG is the only model that manages to obtain good convergence over all simulation repetitions.

Taking all three aspects into account, both the mixture modeling approaches with a penalized likelihood based on the lasso and the elastic net prove to yield good performance in comparison with a traditional mixture model. There is no universally dominant method in this simulation. The added value of the extension of feature selection abilities is most evident when a higher-dimensional variable selection problem is of interest. In this situation, the extended models clearly outperform the traditional mixture model. We find that both approaches based on the lasso yield good results, closely followed by the models based on the elastic net. In terms of selecting the correct subset of variables while also providing an accurate estimate of non-excluded variables in the mixture component, MIXALS is the optimal method. Naturally, this is a golden combination when dealing with a regression problem where heterogeneity and variable selection are of both of importance which is often the case. When determining the optimal amount of components in larger mixtures we find MIXLAS is the only model that performs well.

Table 13 Recovery of true component amounts based on 100 simulation repetitions.

Method	$\widehat{\overline{S}}$	Hit-rate	Converged	**Method**	$\widehat{\overline{S}}$	Hit-rate	Converged	**Method**	$\widehat{\overline{S}}$	Hit-rate	Converged
$S = 2$				$S = 4$				$S = 6$			
MIXREG	2.16	0.84	0.98	MIXREG	**4.00**	**1.00**	**1.00**	MIXREG	5.12	0.12	0.64
MIXRDG	**2.00**	**1.00**	**1.00**	MIXRDG	**4.00**	**1.00**	**1.00**	MIXRDG	4.68	0.02	**1.00**
MIXNET_5	2.04	0.96	**1.00**	MIXNET_5	**4.00**	**1.00**	0.40	MIXNET_5	5.16	0.16	0.34
MIXNET_9	**2.00**	**1.00**	0.99	MIXNET_9	**4.00**	**1.00**	0.46	MIXNET_9	5.26	0.30	0.36
MIXLAS	2.06	0.96	0.94	MIXLAS	4.50	0.50	0.62	MIXLAS	**6.72**	**0.42**	0.28
MIXALS	2.04	0.96	**1.00**	MIXALS	**4.00**	**1.00**	0.50	MIXALS	5.14	0.16	0.38
$S = 8$				$S = 10$				$S = 15$			
MIXREG	5.12	0.00	0.38	MIXREG	4.18	0.00	0.32	MIXREG	4.46	0.00	0.34
MIXRDG	4.66	0.00	**0.96**	MIXRDG	4.56	0.00	**1.00**	MIXRDG	4.66	0.00	**1.00**
MIXNET_5	5.20	0.00	0.49	MIXNET_5	4.64	0.00	0.50	MIXNET_5	5.06	0.00	0.57
MIXNET_5	5.26	0.00	0.58	MIXNET_9	4.62	0.00	0.58	MIXNET_9	5.02	0.00	0.52
MIXLAS	**8.48**	**0.30**	0.46	MIXLAS	**9.36**	**0.90**	0.38	MIXLAS	**9.58**	**0.06**	0.38
MIXALS	5.10	0.00	0.37	MIXALS	4.62	0.00	0.66	MIXALS	5.00	0.00	0.65

Best results per scenario marked in bold.

6 CONCLUSION

This research reports the analysis and development of a finite mixture model. Our purpose was to model and interpret the structure of a heterogeneous customer base. In this process we were also interested in exploring and revealing features that are related to the financial balance of a customer. In order to reveal the key features we extend the mixture model with a regularization algorithm that has the ability to perform variable selection. The model merges simultaneous estimation of a finite mixture model based on the EM algorithm with variable selection abilities into a single procedure. This approach provides a flexible and powerful modeling algorithm which is able to handle todays high-dimensionality and complexity of datasets.

The results of the feature selection reveal the following relations with the financial value of customers. We find that age does not have a negative relation with our response variable. This is likely explained by the settlement we have taken into account in our response variable. Second, not having a main insurance package has a relatively large negative effect on the monetary balance of an individual. The same result holds for having insurance packages with foreign coverage. On the contrary, holding more packages such as additional insurances and modules like dental coverage is positively related with ones balance. Next, the number of claims negatively influences financial balance as expected. In contrast, the frequency, or number of different months in which claims are made is in general not of importance.

Ultimately, the revealed structure is used to describe and interpret segments of distinct individuals. We obtain a 10-component solution which can be used to support differentiation of resources within the company. Four segments are identified as containing customers with desirable characteristics and behavior which making them a valuable asset for the company. The model provides structure in the heterogeneous population of customers. The results can be used to support choices regarding differentiation of segments with desirable and less desirable properties. The segmentation provides a solid foundation which allows for a more efficient business strategy in

comparison to treating the customer base as a whole. For instance, more resources can be invested in the relationship of customers in valuable components. The results provide a big step towards a more data-driven business approach within the company.

Moreover, we have performed a simulation study in which we prove the value of extending mixture models with the power of variable selection algorithms. Results show that the models that combine simultaneous fitting of the components and selecting the most important variables within each component perform well. In addition, the models provide a very convenient algorithm by combining fitting and feature selection into a single procedure while greatly alleviating the computational burden associated with traditional subset approaches.

High-dimensional problems are encountered in many different fields today. We have shown that especially in these situations the extended mixture models clearly outperform an approach with normal likelihood function in terms of selecting the correct subset of variables while accurately estimating the corresponding coefficients. Moreover, the extended mixture models are more accurate in determine the optimal amount of components present in the data. In addition, the combination of a mixture model and feature selection allows for the freedom of selecting the most important subset of variables within each component independently. Another major advantage is that the extended models are able to handle ultra-dimensional problems with more variables than observations. Hence, we conclude that the model proves to be not only a flexible, but also an accurate approach especially when dealing with many covariates.

All things considered, the results of this research are very promising. The discussed approach has a high potential for successfully dealing with regression problems on heterogeneous data. The model excels when a large number of variables are of interest and it is desirable to select the most important ones which is often the case. To conclude, combining finite mixture models with simultaneous variable selection abilities results in a highly relevant technique both for modern applications on complex datasets as well as further academic research.

6.1 CONTRIBUTION TO ACADEMICS

The need for techniques that have the ability to handle large and complex datasets is ever increasing. This is evident by looking at the surge in popularity of variable selection algorithms in current scientific research (Fan and Lv, 2010). Much attention has been given to algorithms that allow for feature selection such as the lasso and elastic net among many others techniques (Tibshirani, 1996; Zou and Hastie, 2005). Moreover, the assumption of perfectly homogeneous data is often not realistic. A single regression or model may fail to adequately capture and describe the structure of complex data.

An efficient way of dealing with heterogeneity is by means of a mixture model. Like in any model, the problem of feature selection is relevant. Moreover, traditional techniques associated with finite mixture modeling such as a best subset approach for variable selection are computationally infeasible when applied on relatively large datasets. In addition, problems often include many covariates. Ultra-high-dimensional problems where the number of parameters is larger than the number of observations are no exception today. This situation is becoming more frequent in various fields such as genomics, web analysis, health sciences, finance, economics and machine learning. Hence, efficient and flexible methods that can deal with heterogeneity and high-dimensionality are greatly relevant, both in practical applications as well as academic research. Khalili (2011) provides a broad overview of variable selection in mixture models and concludes the story is far from complete. Especially for high-dimensional problems much research is still left to be done.

We aim to contribute to this area of research by showing usefulness in both a real-world data application and a simulation study. The performance is tested by comparing behavior in terms of selecting the correct subset of variables to include in the model while accurately providing an accurate estimate of the effect of these variables. In addition, we study the issue of order selection in mixture models. We show that finite mixture models with variable selection can be a very successful approach in regression problems. Moreover, we showcase the added value over a traditional mixture model. The value is most evident when applied in high-dimensional situations where the extended models excel.

7 REFERENCES

H. Abdi and L. J. Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, 1998.

G. M. Allenby and P. E. Rossi. Marketing models of consumer heterogeneity. *Journal of Econometrics*, 89 (1):57–78, 1998.

S. T. Apple. Deep learning for siri's voice: On-device deep mixture density networks for hybrid unit selection synthesis. *Apple Machine Learning Journal*, 1 (4), 2017.

A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability theory and related fields*, 113(3):301–413, 1999.

R. Beran. Estimated sampling distributions: the bootstrap and competitors. *The Annals of Statistics*, pages 212–225, 1982.

P. D. Berger and N. I. Nasr. Customer lifetime value: Marketing models and applications. *Journal of interactive marketing*, 12(1):17–30, 1998.

C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, 22(7):719–725, 2000.

C. Biernacki, G. Celeux, and G. Govaert. Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics & Data Analysis*, 41 (3):561–575, 2003.

G. E. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 211–252, 1964.

L. Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384, 1995.

G. Celeux and G. Govaert. A classification em algorithm for clustering and two stochastic versions. *Computational statistics & Data analysis*, 14(3):315–332, 1992.

G. Celeux and G. Soromenho. An entropy criterion for assessing the number of clusters in a mixture model. *Journal of classification*, 13(2):195–212, 1996.

J. Cook, I. Sutskever, A. Mnih, and G. Hinton. Visualizing similarity data with a mixture of maps. In *Artificial Intelligence and Statistics*, pages 67–74, 2007.

T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B*, pages 1–38, 1977.

W. S. DeSarbo and W. L. Cron. A maximum likelihood methodology for clusterwise linear regression. *Journal of classification*, 5(2):249–282, 1988.

J. Diebolt and C. P. Robert. Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 363–375, 1994.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

J. Fan and J. Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101, 2010.

J. Fan, H. Peng, et al. Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961, 2004.

M. A. T. Figueiredo and A. K. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on pattern analysis and machine intelligence*, 24(3):381–396, 2002.

L. E. Frank and J. H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

J. Friedman, T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu. Discussion of boosting papers. Citeseer, 2004.

J. Friedman, T. Hastie, and R. Tibshirani. glmnet: Lasso and elastic-net regularized generalized linear models. *R package version*, 1(4), 2009.

M. S. Garver, Z. Williams, and G. S. Taylor. Employing latent class regression analysis to examine logistics theory: an application of truck driver retention. *Journal of Business Logistics*, 29(2):233–257, 2008.

G. H. Golub, M. Heath, and G. Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

J. C. Gower. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3-4):325–338, 1966.

F. Groenen, J. Patrick, and M. Velden. *Multidimensional scaling*. Wiley Online Library, 2005.

C. Hennig. Identifiablity of models for clusterwise linear regression. *Journal of Classification*, 17(2):273–296, 2000.

J. R. Hershey and P. A. Olsen. Approximating the kullback leibler divergence between gaussian mixture models. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–317. IEEE, 2007.

G. E. Hinton and S. T. Roweis. Stochastic neighbor embedding. In *Advances in neural information processing systems*, pages 857–864, 2003.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

T. Huang, H. Peng, and K. Zhang. Model selection for gaussian mixture models. *arXiv preprint arXiv:1301.3558*, 2013.

P. J. Huber et al. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1): 73–101, 1964.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, volume 7. Springer, 2013.

M. I. Jordan and L. Xu. Convergence results for the em approach to mixtures of experts architectures. *Neural networks*, 8(9):1409–1431, 1995.

A. Khalili. An overview of the new feature selection methods in finite mixture of regression models. *Journal of The Iranian Statistical Society*, 10(2):201–235, 2011.

A. Khalili and J. Chen. Variable selection in finite mixture of regression models. *Journal of the American Statistical Association*, 102(479):1025–1038, 2007.

K. Knight and W. Fu. Asymptotics for lasso-type estimators. *Annals of statistics*, pages 1356–1378, 2000.

S. Kullback. *Information theory and statistics*. Courier Corporation, 1997.

S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22 (1):79–86, 1951.

M. Kyung, J. Gill, M. Ghosh, G. Casella, et al. Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, 5(2):369–411, 2010.

H. Leeb and B. M. Pötscher. Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics*, pages 2554–2591, 2006.

B. G. Leroux et al. Consistent estimation of a mixing distribution. *The Annals of Statistics*, 20(3):1350–1360, 1992.

B. G. Lindsay. Mixture models: theory, geometry and applications. In *NSF-CBMS regional conference series in probability and statistics*, pages i–163. JSTOR, 1995.

L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov): 2579–2605, 2008.

G. McLachlan and D. Peel. *Finite mixture models*. John Wiley & Sons, 2004.

G. J. McLachlan and K. E. Basford. *Mixture models: Inference and applications to clustering*, volume 84. Marcel Dekker, 1988.

B. Muthén and K. Shedden. Finite mixture modeling with mixture outcomes using the em algorithm. *Biometrics*, 55(2):463–469, 1999.

K. Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

K. Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL http://www.R-project.org. ISBN 3-900051-07-0.

S. Rosset, J. Zhu, and T. Hastie. Boosting as a regularized path to a maximum margin classifier. *Journal of Machine Learning Research*, 5(Aug):941–973, 2004.

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

W. Seidel, K. Mosler, and M. Alker. Likelihood ratio tests based on subglobal optimization: A power comparison in exponential mixture models. *Statistical Papers*, 41(1):85–98, 2000.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

A. N. Tikhonov, V. I. Arsenin, and F. John. *Solutions of ill-posed problems*, volume 14. Winston Washington, DC, 1977.

D. M. Titterington, A. F. Smith, and U. E. Markov. *Statistical analysis of finite mixture distributions*. Wiley,, 1985.

J. W. Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, 2:448–485, 1960.

M. Tuma and R. Decker. Finite mixture models in market segmentation: a review and suggestions for best practices. *Electronic Journal of Business Research Methods*, 11(1):2–15, 2013.

N. Ueda and R. Nakano. Deterministic annealing em algorithm. *Neural networks*, 11(2):271–282, 1998.

N. Ueda, R. Nakano, Z. Ghahramani, and G. E. Hinton. Smem algorithm for mixture models. In *Advances in neural information processing systems*, pages 599–605, 1999.

C. Van Rijsbergen. Information retrieval. *Dept. of computer science, University of Glasgow*, 14, 1979.

M. Wattenberg, F. Viégas, and I. Johnson. How to use t-sne effectively. *Distill*, 1(10):e2, 2016.

M. Wedel and W. S. DeSarbo. A mixture likelihood approach for generalized linear models. *Journal of Classification*, 12(1):21–55, 1995.

M. Wedel and W. A. Kamakura. *Market segmentation: Conceptual and methodological foundations*, volume 8. Springer Science & Business Media, 2012.

S. Weisberg. Yeo johnson power transformations. *Department of Applied Statistics, University of Minnesota. Retrieved June*, 1:2003, 2001.

I. K. Yeo and R. A. Johnson. A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959, 2000.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476): 1418–1429, 2006.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67 (2):301–320, 2005.