Hochschule des Bundes
für öffentliche
Verwaltung

Schriftenreihe 64

Brühl 2023

Hartmut Jakob Stenz

Topics in the statistical analysis of rare events and high-dimensional data sets

Bild: freepik.com

Hartmut Jakob Stenz

Topics in the statistical analysis of rare events and high-dimensional data sets

Brühl/Rheinland 2023

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

ISBN 978-3-96744-006-5

Impressum:
Hochschule des Bundes
für öffentliche Verwaltung
Willy-Brandt-Str. 1
50321 Brühl

www.hsbund.de

Für meine Familie

Vorwort

Es ist mir ein großes Anliegen, den Menschen zu danken, welche mich in den zurückliegenden Jahren begleitet haben und mich immer unterstützten:
Mein erster Dank gilt dabei meinem Doktorvater Prof. Dr. Jörg Breitung, der durch seine Ideen und Vorschläge wesentlichen Anteil am Gelingen meines Promotionsvorhabens hat. Auch dem Zweitbetreuer dieser Arbeit, Herrn Prof. Dr. Rainer Dyckerhoff, will ich nicht nur für die gemeinsame Zeit des Forschens, sondern auch des Lehrens von Herzen danken. Beide Beutreuer standen mir immer mit Rat und Tat zur Seite und es war eine große Bereicherung für mich, mit Ihnen arbeiten zu dürfen. Auch den übrigen Kolleg*innen vom Institut will ich danken für die vielen Stunden des Austausches und der gemeinsamen Arbeit. Die Zeit mit ihnen wird mir stets in bester Erinnerung bleiben und ich darf mich sehr glücklich schätzen, in so einem großartigen Umfeld meinen beruflichen Einstieg vollzogen zu haben.
Neben den genannten Personen ist es mir auch ein besonderes Anliegen meiner langjährigen Freundin Sarah Gansen zu danken, die mir bei der Überwindung meiner Schwächen bezüglich der englischen Sprache unermessliche Dienste erwiesen hat. Auch weitere Freunde und Familienmitglieder, die mich in besonderer Weise unterstützt haben, will ich in den Dank miteinschließen.
Abschließend will ich im Besonderem auch meiner Frau Johanna danken, die mich in den Jahren unserer noch jungen Ehe durch viele schöne, aber auch manche schweren Stunden begleitet hat. Ohne Ihren Rückhalt hätte ich es niemals geschafft, die hier vorliegende Arbeit zu vollbringen.

Sinzig, den 25. September 2022 Hartmut Jakob Stenz

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This thesis is a general work in the field of data science that cannot be assigned to any specific application context. Rather, contributions are made to several topics. Common to all contributions in this work is the focus on the statistical analysis of data sets by supervised learning. In the second chapter of their book *The Elements of Statistical Learning*, Hastie et al. (2001) define this term as follows:

"[...] there is a set of variables that might be denoted as inputs, which are measured or preset. They have some influence on one or more outputs. [...] the goal is to use the inputs to predict the values of the outputs. This exercise is called supervised learning."

A closer look at this very general problem reveals several challenges: On one hand, it must be clarified how exactly the influence of the input on the output should be modeled. In modern data analysis, classical statistical approaches compete with machine learning methods. On the other hand, possible difficulties at the level of the input and the output itself have to be addressed: In situations of high-dimensional data sets with a large number of possible input variables, a decision must be made as to which variables have a relevant influence on the target variable. What is more, when a binary output one class is observed highly rarely compared to the other class, serious problems for the data analysis can occur. This work wants to take a stand on these general challenges and develop possible solutions.
In total, the present work includes three independent essays: The second chapter is a co-authored paper with Rainer Dyckerhoff, while chapters three and four are works with sole authorship. The respective chapters are briefly introduced below:
The second chapter corresponds to the paper *Depth-based support vector classifiers to detect data nests of rare events* by Dyckerhoff and Stenz (2021) and designs a hybrid classification method: Instead of carrying out a classification directly on a data set with a binary target variable, the data is transferred to a *DD*-Plot (depth-versus-depth Plot) in a first step. For each data point, this plot depicts the data depth for both classes. A high-dimensional data set is thus transformed into a two-dimensional data set on which support vector machines (svm) as a machine learning technique are used in a second step. In the event that the target variable is only rarely observed and is also structurally present as a data nest, it

can be analytically proven that the separation of the data in the *DD*-Plot increases with an increase in the dimension in the original data set. Simulation studies substantiate these findings. The idea of combining the *DD*-Plot with the svm came from Rainer Dyckerhoff. He also developed the general structure of the article. In addition, he contributed the proof of the first part of Corollary 1. I carried out the remaining work on the project.

The third chapter provides a generalization of the approach of supervised factor models. These models are primarily used to forecast macroeconomic variables and are based on the idea that a large number of predictors can be combined into a smaller number of factors. The target variable is then regressed on these factors. Supervised factor models also include the target variable in the factor estimation and assume that all factors are relevant to the target variable. The present work drops this assumption and divides the set of factors into two groups: One factor that summarizes all effects that are relevant for the prognosis of the target variable and the other factors that only serve to explain the predictors. The advantage of such a division lies in overcoming the two-stage nature of conventional approaches: Instead of first estimating the factors and then regressing the target variable on these factors and other external variables, the factor estimation and regression of the target variable can take place simultaneously. The procedure underlying this estimation is derived as well as its algorithmic implementation. A simulation study serves to demonstrate the strength of this approach when compared to traditional approaches.

While the first chapters do not include empirical applications, the fourth and final chapter is a purely empirical project: A data mining analysis is performed on medical billing data with the aim of determining the possibility of algorithmic pre-screening for birth defects in newborn children. The analysis uses variants of the random forest as well as logistic regression. Some variants take up concepts from chapter two but others apply additional approaches. The focus of this empirical work is on the rare event problem and on the question of whether the decision-making calculus of a human decision-maker can be replicated algorithmically.

Chapter 2

Depth-based support vector machines to detect data nests of rare events

2.1 Introduction

As mentioned in the title, this chapter's approach to classification is based on data depth: Given a sample or "cloud" X of data points $x_1, \ldots, x_N$ in $\mathbb{R}^d$, a data depth is a function $\mathcal{D}(z|X) : \mathbb{R}^d \to [0,1]$ that describes how "deep" z lies in X: Values near to 0 mean that z is far away from the center of X and values close to 1 signify that z is located in or next to the center of X. The relevant literature consists of many research contributions relating to data depths. The different approaches range from the development of individual depths such as the "Mahalanobis depth" (see Mahalanobis (1936)) or the "halfspace depth" (see Donoho and Gasko (1992)) over the formalization and systematization of certain depth characteristics (see e.g. Dyckerhoff (2004)) to using data depths for classification purposes.
Concerning the latter approach, the work of Li et al. (2012), which developed the *DD*-classification as a new classification method based on data depths, is worth mentioning: They analyze the quantity $(\mathcal{D}(z_k|X), \mathcal{D}(z_k|Y)) \in [0,1]^2$ $k = 1, \ldots, N$, the so called *DD*-Plot, for given training points $z_1, \ldots, z_N$. The aim is to construct a decision rule based on the *DD*-Plot instead of using the initial data points for reference. After determining the depths $x^* = \mathcal{D}(z^*|X)$ and $y^* = \mathcal{D}(z^*|Y)$, a new data point z^* can thus be classified by applying this decision rule to (x^*, y^*). In recent years, scientists took a number of different approaches as to how a decision rule based on the *DD*-Plot can be constructed: While Li et al. (2012) applied polynomial dividing lines, Mozharovskyi (2014) among others uses a non-parametric approach in his *DD*-α-procedure.
This project aims to take up the approach of Kim et al. (2018). It is based on the application of support vector machines (svm) in the *DD*-Plot meaning that the different points of the *DD*-Plot are separated with the help of separating hyperplanes employing kernels. In doing so, one expects that the use of svm will allow for the construction of a decision rule that will produce precise results even in the event of overlapping cases. For this project, the main interest and focus is on the detection of rare events: Such data structures play a significant role in churn prediction analyses (see Reuß and Zwiesler (2006)) and credit

default or fraud management, for example. Therefore, this application serves as a potential but not sole motivation to analyze the central issue of how to predict the likelihood of a customer's churn or default given that the training set contains merely a very limited number of these customers in comparison to "normal" customers. It is very well possible that such churn or default customers deviate on average only very little from others but that they are nevertheless clustered in certain segments. As a consequence, there is a difference in dispersion rather than in location of the set X (normal customers) and Y (churn/default customers). As X is additionally much bigger than Y, this data structure can be described as a "data nest" of Y located in X.
From an analytical point of view, this project strives to demonstrate that under the assumption of an elliptical distribution, the separability in the *DD*-Plot will increase with increasing data dimension if the characteristics of data nests are fulfilled. It may seem intuitive to assume that access to more information will allow for a more precise classification. However, the observation that the structure of the data itself seems to be the reason that makes this classification possible deserves further investigation: Due to applying a data depth transformation and transferring the structure into the *DD*-Plot, the number of possible dimensions is drastically reduced from d to 2. What is more, this transformation is irreversible. Yet, the question whether important information is actually lost or rather compressed when carrying out this process remains. Ultimately, using the *DD*-Plot results in a radical simplification, thus in a trade-off between a loss of information and a reduction of complexity. The latter aspect in particular is of great significance in a big data context.
This chapter is structured as followed: After this Section serving as an introduction, Section 2 will discuss the basic concepts and ideas. In Section 3, the analytical results will be outlined and Section 4 will present the simulation studies before the conclusion and summary of the results will be given in the fifth and last Section.

2.2 Basic ideas

2.2.1 Data depth

There is a range of approaches that make the concept of data depth accessible and many of them can be found in the literature. For introductory purposes, the first part touched upon the question which insights into specific data sets data depths can offer. But there is also the possibility of an axiomatic illustration of data depth functions, which was first carried out by Liu (1990) and also Zuo and Serfling (2000). Following this approach the data depths functions can be defined either on samples X in $\mathbb{R}^d$ or upon the underlying probability distributions of these samples (see Dyckerhoff (2004)). Since it is sufficient to look at specific data sets for the objective of this paper, the first approach will be employed:

(i) **Affine invariant:** $\mathcal{D}(Az+b|AX+b) = \mathcal{D}(z|X)$ for all $b \in \mathbb{R}^d$ and regular $A \in \mathbb{R}^{d\times d}$.

(ii) **Null at infinity:** $\lim_{||z||\to\infty} \mathcal{D}(z|X) = 0$.

(iii) **Monotone on rays:** For z^* with $\mathcal{D}(z^*|X) = \max_{z\in\mathbb{R}^d} \mathcal{D}(z|X)$ and any $\alpha > 0$ and r out of the unit sphere S^{d-1}, $\mathcal{D}(z^* + \alpha r|X)$ decreases in α in a weak sense.

(iv) **Upper semi continuous:** The upper level sets $\mathcal{D}_\alpha(X) := \{z \in \mathbb{R}^d : \mathcal{D}(z|X) \geq \alpha\}$ are closed for all α.

Axiom (i) will play a crucial role for our analytic view: It implies that the data depth values of individual points do not depend on the choice of a specific coordinate system, so changing that system has no influence whatsoever on the depth relations between individual points. Axioms (ii) and (iii) refer to the initial idea that data depths are intended to reflect the centrality of points in a set: If a point's distance from the center of the set is significant, its depth should tend towards zero (axiom (ii)). Correspondingly, the data depth of points is decreasing in a weak sense when those points are moving away from a point with maximum depth (axiom (iii)). As opposed to that, axiom (iv) is of a mere technical nature.
Taking into account the numerous different existing data depths, it needs to be said that, for the vast majority of them, an exact depth value in high dimensions is very difficult or even impossible to calculate with simple computing capacity. In this case, "high" means $d > 5$ (see Mozharovskyi (2014)), meaning that one would actually have to completely refrain from using data depths in the context of big data problems. The Mahalanobis depth, which goes back to the work of Mahalanobis (1936), is a noteworthy exception: This data depth function, referred to as $\mathcal{D}_M(z|X)$ from now on, can be computed in $\mathcal{O}(n)$ for every dimension and is based on the well-known Mahalanobis distance:

$$\mathcal{D}_M(z|X) = \left(1 + (z-\mu_X)^\top \Sigma_X^{-1} (z-\mu_X)\right)^{-1} \tag{2.2.1}$$

Here μ_X denotes the empirical mean and Σ_X the empirical covariance matrix of X. The Mahalanobis depth is a so-called *convex depth* as it fulfills not only axioms (i) to (iv) but also an additional fifth axiom (see Dyckerhoff (2004)):

(v) **Quasiconcavity:** The upper level sets $\mathcal{D}_\alpha(X)$ are convex for all α.

Figure 2.2.1 is an example of visualization of the Mahalanobis depth: In the x_1-x_2-plane, set X containing 150 realizations of a random variable $\mathcal{X} \sim \mathcal{N}(\mu_\mathcal{X}, \Sigma_\mathcal{X})$ with $\mu_\mathcal{X} = (0,0)^\top$ and $\Sigma_\mathcal{X} = \left(\begin{smallmatrix} 3 & 1 \\ 1 & 2 \end{smallmatrix}\right)$ are depicted in red. The depth function $\mathcal{D}_M(z|X)$, which was calculated on the basis of the empirically determined mean value μ_X and the empirically determined covariance matrix Σ_X, rises above the x_1-x_2-plane in gray. The highest point of the function is located above the center of X and from there, it drops fast in all directions. The resulting

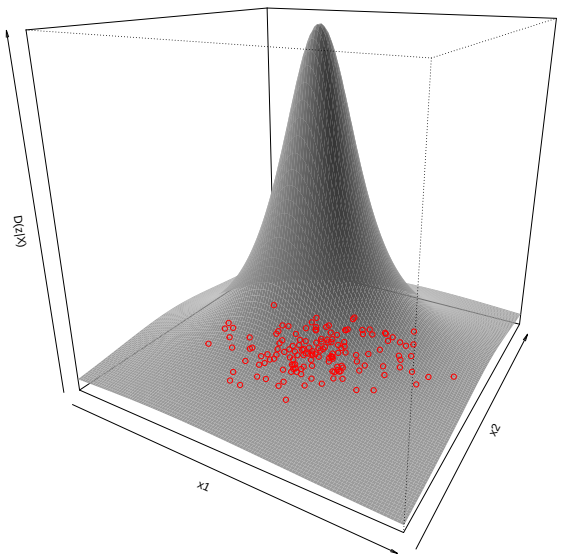


Figure 2.2.1: Mahalanobis depth (grey) on a set in $\mathbb{R}^2$ (red).

funnel structure of the function has elliptical contour sets. With the help of this example, two major weaknesses of the Mahalanobis depth can be pointed out: First of all, $\mathcal{D}_M(z|X)$ would take exactly the same course on another, non-identical set Y, which would show the same empirical first and second moments as X. This is due to the fact that the function depends solely on the first two empirical moments. On the other hand, due to its elliptical contour sets, this depth function is well-suited to depict in particular those data which are based on an elliptical distribution, such as the normal distribution. Among others, these two mentioned weak points of the Mahalanobis depth are the reasons why researchers have refrained from using this depth (see, for example, Mosler (2013)) in recent times. However, we restrict ourselves to the use of this depth function for the time being.

2.2.2 Depth based classification

Subsequent to this, we want to understand how to use data depths for the classification of (binary) data. With regard to this issue, Vencálek (2017) provides a detailed depiction of the research results in the field of depth-based classification of the past 20 years: His paper uses the term "advanced depth-based classifiers" for those methods that resort to the *DD*-Plot, an abbreviation for "depth-versus-depth"-Plot (see Li et al. (2012)), thus replacing the actual data set. How such a *DD*-Plot can be applied to binary classification problems has already briefly been mentioned in the introduction and will now be reviewed again in a formal definition:

Definition 1. *Let $z_1, \ldots, z_N \in \mathbb{R}^d$ be a sample of data points with known outcomes $w_k \in \{-1, 1\}$. Then we define classes*

$$X = \left(z_i \in \mathbb{R}^d | w_i = -1, i \in \{1, \ldots, N\}\right)$$
$$Y = \left(z_j \in \mathbb{R}^d | w_j = 1, j \in \{1, \ldots, N\}\right)$$

to get the the two-dimensional DD-Plot

$$V = \left(v_k = (x_k, y_k) \in \mathbb{R}^2 | x_k = \mathcal{D}\left(z_k | X\right), y_k = \mathcal{D}\left(z_k | Y\right), k \in \{1, \ldots, N\}\right).$$

Of course $\{0, 1\}$ instead of $\{-1, 1\}$ for the range of w_k is suitable, but later on we will see why this type of coding is especially useful.

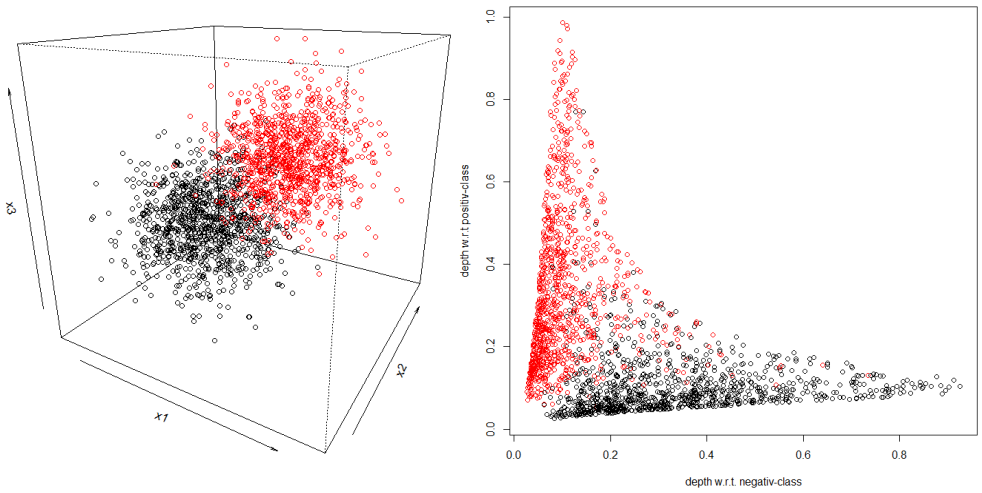


Figure 2.2.2: X (black) and Y (red) in $\mathbb{R}^3$ (left) and their *DD*-Plot (right).

A *DD*-Plot is exemplarily shown in Figure 2.2.2: In the left panel, two sets are depicted in $\mathbb{R}^3$, the set X (black, "negative-class") and Y (red, "positive-class"). Both sets follow a normal distribution with different center points but equal covariance matrix. Accordingly, the two data sets barely overlap. The *DD*-Plot corresponding to these sets can be found in the right panel: In this plot, there is hardly any superimposition of the data points either. The majority of those points are arranged along the two axes. Points which belong to the set X (black) are arranged along the x-axis ($\mathcal{D}\left(z|X\right)$) of the *DD*-Plot, while the arrangement of the points belonging to Y (red) is diametrically opposed, namely along the y-axis ($\mathcal{D}\left(z|Y\right)$). The *DD*-Plot thus condenses the given information from the original sets to determine X and Y's location, meaning that all available information is reduced to the question of where those points are located in the coordinate system and in relation to the total amount of all points. If a point is located deeper in X than in Y, it is bound to have a high x-value and

a low y-value in the *DD*-Plot, and vice versa. It is important to note that the resulting image in the *DD*-Plot attempts to map as precisely as possible the locational relationship, and nothing else but this locational relationship, for each point on the basis of the given set of constellations. The fact that this mapping depends on the choice of the data depth used and that it is accompanied by a reduction of the dimension from d to 2, is obvious.
In order to be able to use a *DD*-Plot for the classification, three steps are needed:

- **Step 1:** Calculate on given training sets a *DD*-Plot.
- **Step 2:** Construct a decision rule based on this Plot.
- **Step 3:** Classification of a new data point: Calculate its depth values and use the constructed decision rule on these values.

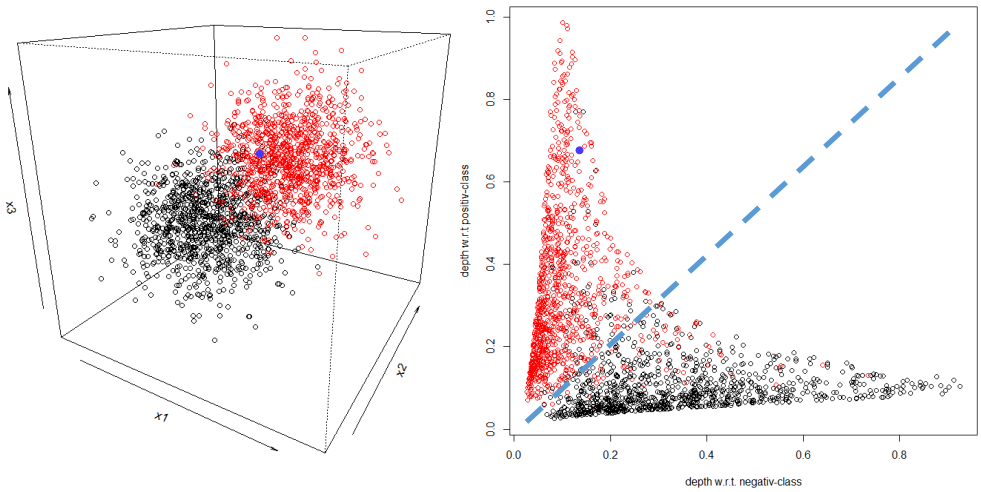


Figure 2.2.3: Same sets and *DD*-Plot like in Figure 2.2.2. The decision rule (light blue dotted) is just the line $x = y$ and the point of interest (dark blue) is classified to Y by this rule.

Aside from the choice of the data depth, which is crucial for the appearance of the *DD*-Plot, the real difficulty of this method lies in the choice of the decision rule: A very simple decision rule could be, for example, that a point in the *DD*-Plot is always assigned to that set in which its depth is maximal. Such a rule would graphically correspond to the main diagonal in the *DD*-Plot, as shown in Figure 2.2.3, a copy of Figure 2.2.2, in the right panel (in light blue). If a point, such as the point highlighted in dark blue in the right panel, was located above the main diagonal, it would accordingly be assigned to the red set Y. This decision rule may be ideal for the given sets. Applying the same rule to other set constellations,

however, does by no means guarantee that comparably good classification results will be achieved. In a nutshell, it can thus be said that for a given set or data scenario, a new search for the best decision must be carried out in each case.
How a decision rule for the respective *DD*-Plot can be efficiently constructed is the subject of a large number of publications of the past five years. In Vencálek (2017), the *DD*-α-procedure of Mozharovskyi (2014) and Lange et al. (2014) is named as one of the best methods currently available to construct a decision rule on the *DD*-Plot quickly and efficiently: The idea of this method is based on the additional construction of a "feature" space by using the coordinates of the *DD*-Plot to form features for different $k, l > 0$ of the form $\mathcal{D}(z|X)^k \mathcal{D}(z|Y)^l$. In this feature-space, the separating hyperplane, which produces the lowest classification error, then needs to be found. In order to be able to do so, said features have to be calculated explicitly as they are required to determine the hyperplane.
In contrast, the approach taken by Kim et al. (2018), that uses depth-based classifications to predict bankruptcy, relies on support vector machines. As will become evident in the next Section, features built on the *DD*-Plot are also used but only in implicit instead of explicit form. This method will be called *Depth based support vector classifiers* or *Depth based support vector machines*, in short *DD*-svm (cf. Kim et al. (2018)).

2.2.3 Support Vector Classifiers

Support vector machines, svm from hereon, are among those classification methods which are described in great detail in the literature (see, for example, Hastie et al. (2001)). Vapnik (1995) first introduced svm in his theory of statistical learning, and in the years that followed, this method was used in various areas of statistical research and practice (see Smola and Schölkopf (2004)). The core idea of svm is the construction of a separating hyperplane on a training data sample $T = ((v_1, w_1), \ldots, (v_N, w_N))$. In this sample, $w_1, ..., w_N \in \{-1, 1\}$ are the observed outcomes of the points $v_1, ..., v_N \in V \subset \mathbb{R}^d$, mentioned at the beginning of Section 2.2. These outcomes are now estimated by the expression $\hat{f}$. It uses the (for the time being linear) hyperplane $\langle \beta, v \rangle + \beta_0$ in $\mathbb{R}^d$ with $\beta \in \mathbb{R}^d$ and $\beta_0 \in \mathbb{R}$:

$$\hat{f}(v) = \text{sign}(\langle \beta, v \rangle + \beta_0) \tag{2.2.2}$$

The use of function sign$(\cdot)$ corresponds to the coding of w_k in $\{-1, 1\}$. First of all, let us suppose that the samples $X^T = (v_i | w_i = -1)$ and $Y^T = (v_j | w_j = 1)$ are linearly separable in $\mathbb{R}^d$. In that case, there would be at least one combination of β and β_0 such that for $k = 1, \ldots, N$, the following applies:

$$w_k(\langle \beta, v_k \rangle + \beta_0) \geq 1 \Rightarrow w_k \hat{f}(v_k) = 1 \tag{2.2.3}$$

In order to find the "best" combination of β and β_0 among these possible combinations, there exist several approaches. One possible approach is the minimization of $||\beta||$ (or as in

this example $\frac{1}{2}||\beta||^2$). Combined with condition 2.2.3, this results in a convex optimization problem:

$$\min_{\beta,\beta_0} \frac{1}{2}||\beta||^2 \text{ w.r.t. } w_k\left(\langle\beta, v_k\rangle + \beta_0\right) \geq 1 \text{ for } k = 1, \ldots, N \tag{2.2.4}$$

If X^T and Y^T overlap, which is almost always the case, there are two possibilities that can be combined with each other: On the one hand, the optimization problem 2.2.4 is complemented by slack variables $\xi_1, \ldots, \xi_N$:

$$\min_{\beta,\beta_0} \frac{1}{2}||\beta||^2 + C\sum_{k=1}^{N}\xi_k \text{ w.r.t. } \left\{\begin{matrix} w_k\left(\langle\beta, v_k\rangle + \beta_0\right) \geq 1 - \xi_k \\ \xi_k \geq 0, C \geq 0 \end{matrix}\right\} \text{ for } k = 1, \ldots, N \tag{2.2.5}$$

The parameter C is referred to as the cost parameter: The larger this parameter is, the more important it is to ensure that as few points v_k as possible lie on the "wrong" side of the hyperplane (if this was possible for all v_k, as in the case of linear separability of the sets X^T and Y^T, $C = \infty$ could be set). The solutions $\left(\hat{\beta}, \hat{\beta}_0\right)$ for 2.2.5 can now be determined by using Lagrangian multipliers (see, for example, Gill et al. (1981)):

$$\hat{\beta}_0 = \alpha_0, \hat{\beta} = \sum_{k=1}^{N}\alpha_k w_k v_k \Rightarrow \hat{f}(v) = \text{sign}\left(\sum_{k=1}^{N}\alpha_k w_k \langle v_k, v\rangle + \alpha_0\right) \tag{2.2.6}$$

On the other hand, it is also possible to transfer the space V into a higher-dimensional feature space by means of a map $\phi : V \to \mathbb{R}^D$ $(D > d)$ in order to be able to solve the optimization problem from 2.2.5. Hastie et al. (2001) for example show that such a solution, similar to 2.2.6, depends solely on the inner products of individual vectors:

$$\hat{f}(v) = \text{sign}\left(\sum_{k=1}^{N}\hat{\alpha}_k w_k \langle\phi(v_k), \phi(v)\rangle + \hat{\alpha}_0\right) \tag{2.2.7}$$

What is more, in order to determine 2.2.7, one exclusively needs information about inner products of vectors and not about the vectors themselves. Put differently, this means that instead of $\phi(v^*)$ and $\phi(v)$, only $\langle\phi(v^*), \phi(v)\rangle$ plays a role in the calculations. Accordingly, the individual features do not have to be calculated explicitly but the inner product in the feature space can be implicitly determined by using a kernel function $\mathcal{K}(v^*, v) = \langle\phi(v^*), \phi(v)\rangle$ to arrive at the following solution:

$$\hat{f}(v) = \text{sign}\left(\sum_{k=1}^{N}\hat{\alpha}_k w_k \mathcal{K}(v_k, v) + \hat{\alpha}_0\right) \tag{2.2.8}$$

Those v_k with $\hat{\alpha}_k \neq 0$ are called support vectors. The solution parameters $\{\hat{\alpha}_0, \ldots, \hat{\alpha}_N\}$ depend on the choice of the kernel function, but also on the height of the cost parameter C. An often used kernel function (see Hsu et al. (2003)) is the radial basis function (RFB):

$$\mathcal{K}(v^*, v) = \exp\left(-\gamma||v^* - v||^2\right) \tag{2.2.9}$$

Using 2.2.9, the solution parameters $\{\hat{\alpha}_0, \ldots, \hat{\alpha}_N\}$ in 2.2.8 thus correspond to a parameter pair (C, γ). On a given training sample, a specific validation method via this parameter pair can be used by carrying out a grid search in order to determine the best solution for the specific classification problem. A possible validation method could, for example, be the cross validation method (see Hastie et al. (2001)).
The idea of svm and the classification by means of a *DD*-Plot can now be combined to form the hybrid classification method *DD*-svm, comprising the following steps:

- **Step 1:** Given $T = ((z_1, w_1), \ldots, (z_N, w_N))$ with $z_k \in \mathbb{R}^d$ and $w_k \in \{-1, 1\}$, calculate with $X = \left(z_i \in \mathbb{R}^d | w_i = -1\right)$ and $Y = \left(z_j \in \mathbb{R}^d | w_j = 1\right)$ the *DD*-Plot $V = (v_k = (x_k, y_k) \in \mathbb{R}^2 | x_k = \mathcal{D}(z_k|X), y_k = \mathcal{D}(z_k|Y), k \in \{1, \ldots, N\})$.
- **Step 2:** Using a reasonable validation method, do a grid search on (C, γ) to find optimal $\{\hat{\alpha}_k\}_{k=0}^N$ on the training sample T for a decision rule $\hat{f}(v) = \text{sign}\left(\sum_{k=1}^N \hat{\alpha}_k w_k \exp\left(-\gamma ||v_k - v||^2\right) + \hat{\alpha}_0\right)$.
- **Step 3:** To classify new data z^*, calculate $v^* = (\mathcal{D}(z^*|X), \mathcal{D}(z^*|Y))$, then one can classify z^* with $w^* = \hat{f}(v^*) \in \{-1, 1\}$.

2.2.4 Data nests of rare events

When using a hybrid method, it is crucial to clarify why a combination of two different methods leads to better results than the use of one method by itself: Applied to the present analysis, this means that one has to analyze why *DD*-svm should provide better results than just using svm without data depth transformation. Accordingly, the aim of this paper is to provide extrinsic motivation for the use of *DD*-svm by pointing to a concrete data structure in which it can be assumed that a hybrid method of data depths and svm makes sense:

Definition 2. *Let $X = (x_1, \ldots, x_n)$ be an iid sample of the random vector $\mathcal{X}$ with moments $(\mu_\mathcal{X}, \Sigma_\mathcal{X})$ and $Y = (y_1, \ldots, y_m)$ an iid sample of the random vector $\mathcal{Y}$ with moments $(\mu_\mathcal{Y}, \Sigma_\mathcal{Y})$, both located in $\mathbb{R}^d$. Then Y is a **data nest of rare events** with respect to X if both of the following conditions are fulfilled:*

(1) $m << n$.

(2) $\Sigma_\mathcal{X} - \Sigma_\mathcal{Y}$ is positive semi definite.

For the last condition we write $\Sigma_\mathcal{X} \geq \Sigma_\mathcal{Y}$. This means that $\mathcal{Y}$ has less dispersion than $\mathcal{X}$.

Figure 2.2.4 serves to illustrate this definition: The left panel shows two samples in $\mathbb{R}^3$, which are based on normal distributions and satisfy both conditions of Definition 2: Of the

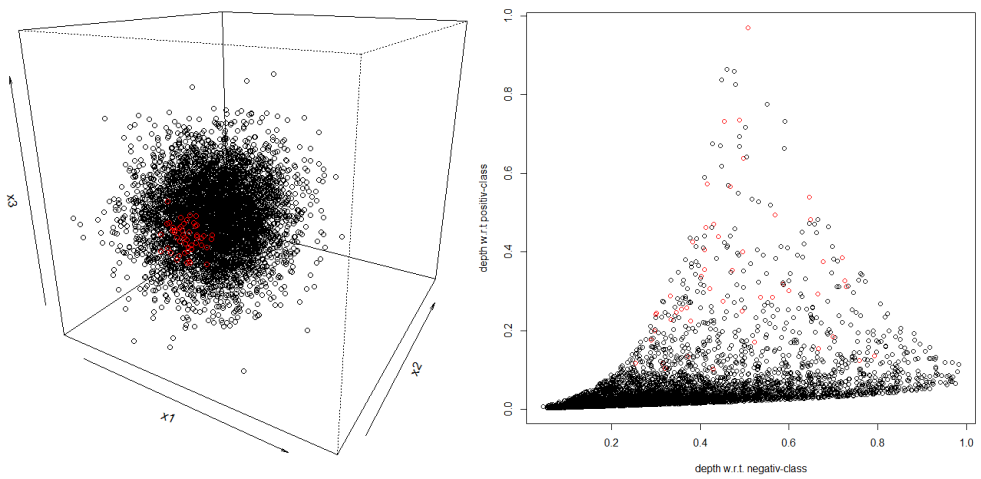


Figure 2.2.4: A data nest (red) in $\mathbb{R}^3$ (left) and its *DD*-Plot (right).

5000 data points that make up X (black) and Y (red), only 50 are attributable to sample Y. Accordingly, Y accounts for only 1% of the total data points. Condition 1 is thus satisfied. In addition, $\Sigma_{\mathcal{X}} = \mathcal{I}_d$, but for $\Sigma_{\mathcal{Y}} = Q^\top \text{diag}[\varphi_1 \ldots, \varphi_d]Q$ with a orthogonal matrix Q the eigenvalues $\varphi_1, \ldots, \varphi_d$ are randomly drawn from the interval $[0.05, 0.15]$. Consequently, because $\varphi_k < 1$ for all k, $\Sigma_{\mathcal{X}} - \Sigma_{\mathcal{Y}} = Q^\top \text{diag}[1-\varphi_1 \ldots, 1-\varphi_d]Q$ is positive definite, therefore condition 2 is also satisfied. The corresponding *DD*-Plot shown on the right panel will be discussed later in this Section.

In the literature, a term similar to "data nest of rare events" can be found: Reuß and Zwiesler (2006) speak about so-called "churn nests" in the framework of churn prediction analyses. This term is used to describe the phenomenon that cancellation customers in relation to the set of all customers are frequently located in certain segments, as has already been briefly mentioned at the beginning of this paper. The reasons for this can be manifold: Customers who have not been affiliated with a given company for an extended amount of time are more likely to resign than those who have long since been a part of the company's customer base, for example (see Kahlenberg (2005)). This higher degree of similarity of cancellation customers with each other could mathematically be described with a lower variance within the cancellation customer segment compared to the normal customer segment. Moreover, in a "healthy" company, the number of loyal customers, i.e. those who stay with the company, by far exceeds the number of customers leaving the company in a given period. It can therefore be assumed that there are significantly more normal customers than cancellation customers in the entire customer segment. As a result, a rare event problem exists and a "churn nest" corresponds exactly to Definition 2.

Why data nests of rare events can be considered data structures that motivate the use of a hybrid method such as *DD*-svm will become evident in the following discussion. For this purpose, the detection of "churn nests" needs to be put in the big data context. For each customer, there is a variety of information and thus different variables available, which is

why high-dimensional data structures are present in this case. Both the findings in the literature (see, for example, Hastie et al. (2001)) and the following simulation studies make it clear that it is difficult to apply svm in the high-dimensional case: Although accurate results can still be achieved, this can only be done at the cost of excessive calculation times. This problem can only be solved by variable selection (see Hsu et al. (2003)). However, such a selection would in itself produce new challenges. Ultimately, the classification problem is thus rendered increasingly complex and more difficult to solve.

In contrast, this problem does not occur when using *DD*-svm on high-dimensional data nest structures: In this case, a special property of the data-depth transformation even leads to a simplification of the classification problem. To understand this, the *DD*-Plot from Figure 2.2.4 needs to initially be considered. It has been generated on a three-dimensional data nest. As one might expect, most points from X accumulate along the x-axis and have a very low y-value. This reflects that hardly any points from X are deep in the data nest Y. Conversely, there are definitely points from Y with high x-values since the data nest is located near the center of X. Their y-values, however, are higher than the y-values of the X points. This is true even if, in absolute terms, more points from X than points from Y have a high y-value. In this regard, it is important to keep in mind the small share of Y of all points together.

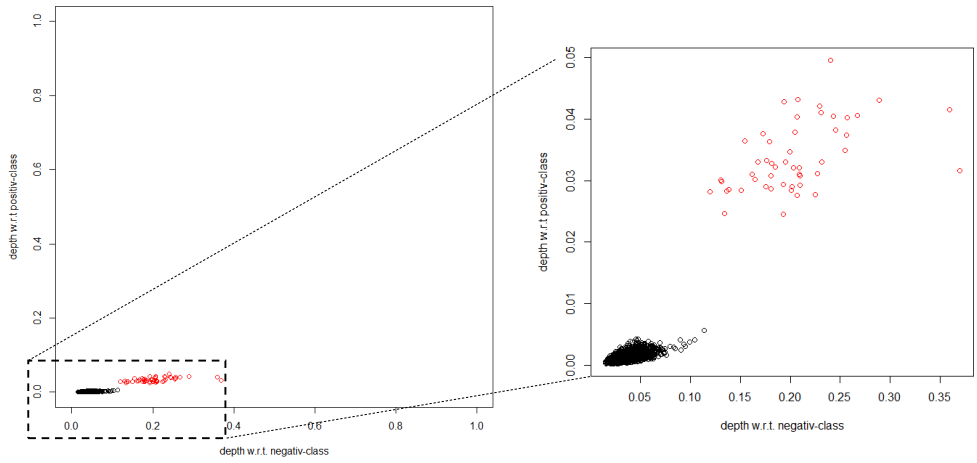


Figure 2.2.5: A *DD*-Plot of a data nest in $\mathbb{R}^{30}$ (left) with a zoom into the interesting region next to the origin (right).

Figure 2.2.5 shows a *DD*-Plot which was created on a thirty-dimensional data nest structure: The first important observation is that all points are concentrated in the lower left part of the *DD*-Plot. Compared to the low-dimensional case, the data points have "migrated" to

the origin. Even more importantly though, the points from the data nest migrate "more slowly", meaning that their y- as well as their x-values in the *DD*-Plot decrease more slowly with increasing data dimension than the corresponding values of the remaining points from X. As a result, a linearly separable data set is obtained in the *DD*-Plot, as shown in the plot itself (left) or in the zoom (right) of Figure 2.2.5. svm can thus be applied very easily and without much computational effort on this *DD*-Plot for the purpose of classification. Consequently, using a hybrid approach makes sense (see e.g. Kim et al. (2018)).
The aim of the analytic approach in the next Section is to mathematically comprehend this property of the data depth transformation, which contributes decisively to the simplification of the classification problem. In fact, it is not easy to understand why the data points migrate with different velocities in the *DD*-Plot. Especially the fact that this implies that points from the data nest Y are not only deeper in their own set but even deeper in the set X than points from X itself, needs clarification.

2.3 Analytical results

In order to be able to make analytical assessments, different restrictions must be made with regard to the distribution of the data sets. This Section solely focuses on the case of elliptical distributed data sets. It has already been mentioned in the beginning that the Mahalanobis depth $\mathcal{D}_M(z|\cdot)$ can best be applied to such data sets. Following this line of thought, the term "elliptical distribution" will be formally defined at first. This definition, following Theorem 2.1 in Fang et al. (1990), will also include the definition of spherical distributions:

Definition 3. *Let $\mathcal{X}$ and $\mathcal{Y}$ be random vectors in $\mathbb{R}^d$.*

*(1) $\mathcal{X}$ is said to have a **spherical distribution** if its characteristic function $\psi_{\mathcal{X}}(t)$ has the following form:*

$$\psi_{\mathcal{X}}(x) = \Gamma\left(x^\top x\right).$$

*The scalar function $\Gamma(\cdot)$ is called **characteristic generator** of the spherical distribution and therefore we write $\mathcal{X} \sim S_d(\Gamma)$.*

*(2) $\mathcal{Y}$ is said to have an **elliptical distribution** with parameters $\mu \in \mathbb{R}^d$ and $\Omega \in \mathbb{R}^{d\times d}$, where $rank(\Omega) = k$, if it has the same distribution as*

$$\mu + \mathcal{A}^\top \mathcal{Z}$$

where $\mathcal{Z} \sim S_k(\Gamma)$ and for $\mathcal{A} \in \mathbb{R}^{k\times d}$ holds: $\mathcal{A}^\top\mathcal{A} = \Omega$. For the case $rank(\Omega) = d$ $\mathcal{A} = \Omega^{\frac{1}{2}}$ is set. In all cases we write $\mathcal{Y} \sim EC_d(\mu, \Omega, \Gamma)$.

The following propositions along with their proofs can also be found in Fang et al. (1990) as a corollary of Theorem 2.2 in section 2.3:

Proposition 1. *Let $\mathcal{X} \sim S_d(\Gamma)$ and $\mathcal{Y} \sim EC_d(\mu, \Omega, \Gamma)$ with $rank(\Omega) = d$ defined on the same characteristic generator $\Gamma(\cdot)$. Then for a randomly drawn x from $\mathcal{X}$ and y from $\mathcal{Y}$ with corresponding $z \sim S_d(\Gamma)$ we have the* ***stochastic representation***

$$x = ||x|| \frac{x}{||x||} \stackrel{d}{=} r_d u^{(d)} \text{ and } y = \mu + ||z|| \, \Omega^{\frac{1}{2}} \frac{z}{||z||} \stackrel{d}{=} \mu + r_d \Omega^{\frac{1}{2}} u^{(d)}$$

with $\frac{x}{||x||} \stackrel{d}{=} u^{(d)} \stackrel{d}{=} \frac{z}{||z||}$ be uniformly distributed on the unit sphere S^{d-1} and independent of r.v. $||x|| \stackrel{d}{=} r_d \stackrel{d}{=} ||z|| \geq 0$. r_d is called ***generating variate*** *and identified by $\Gamma(\cdot)$ and d.*

$x \stackrel{\mathrm{d}}{=} y$ denotes here, that both r.v. are identical distributed. Elliptical can thus be traced back to a combined distribution of the uniformly distributed random variable $u^{(d)}$ and the generating variate $r_d \geq 0$. The existence of $\mathbb{E}(r_d^2)$ has consequences for the existence of the first two moments of the distribution:

Proposition 2. *Let $\mathcal{X} \sim S_d(\Gamma)$ and $\mathcal{Y} \sim EC_d(\mu, \Omega, \Gamma)$ with $rank(\Omega) = d$ and generating variate r_d. If $\mathbb{E}(r_d^2) < \infty$ then the following holds:*

$$\mu_{\mathcal{X}} = 0, \Sigma_{\mathcal{X}} = \frac{\mathbb{E}(r_d^2)}{d} \mathcal{I}_d \text{ and } \mu_{\mathcal{Y}} = \mu, \Sigma_{\mathcal{Y}} = \frac{\mathbb{E}(r_d^2)}{d} \Omega.$$

Also the relationship between generating variates of different dimensions can be clarified:

Proposition 3. *Let r_{d_1} and r_{d_2} be two generating variates of different dimensions, but related to the same characteristic generator $\Gamma(\cdot)$. If $d_1 < d_2$, then it holds:*

$$r_{d_1} \stackrel{d}{=} r_{d_2} b$$

with $b^2 \sim Beta\left(\frac{d_1}{2}, \frac{d_2 - d_1}{2}\right)$ independent of r_{d_2}. Therefore, if $\mathbb{E}\left(r_{d_2}^2\right) < \infty$ also $\mathbb{E}\left(r_{d_1}^2\right) < \infty$, and we get:

$$\mathbb{E}\left(r_{d_1}^2\right) = \mathbb{E}\left(r_{d_2}^2 b^2\right) = \mathbb{E}\left(r_{d_2}^2\right) \frac{d_1}{d_2}.$$

We will later use this proposition for comparing elliptical distributions of different dimensions. For now, however, a simple lemma will be considered:

Lemma 4. *Let $\mathcal{X}^* \sim EC_d(\mu_{\mathcal{X}^*}, \Omega_{\mathcal{X}^*}, \Gamma)$ with $rank(\Omega_{\mathcal{X}^*}) = d$ and $\mathcal{Y}^* \sim EC_d(\mu_{\mathcal{Y}^*}, \Omega_{\mathcal{Y}^*}, \Gamma)$ also with $rank(\Omega_{\mathcal{Y}^*}) = d$. Furthermore, for the generating variate $r_d \geq 0$ holds $\mathbb{E}(r_d^2) < \infty$. For an iid sample $X^* = (x_1^*, \ldots, x_n^*)$ from $\mathcal{X}^*$ and an iid sample $Y^* = (y_1^*, \ldots, y_m^*)$ from $\mathcal{Y}^*$ the two samples $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_m)$ are defined by $x_i = \Omega_{\mathcal{X}^*}^{-\frac{1}{2}}(x_i^* - \mu_{\mathcal{X}^*})$ for all $1 \leq i \leq n$ and $y_j = \Omega_{\mathcal{X}^*}^{-\frac{1}{2}}\left(y_j^* - \mu_{\mathcal{X}^*}\right)$ for all $1 \leq j \leq m$. For these samples holds:*

(1) X is an iid sample from the r.v. $\mathcal{X} \sim S_d(\Gamma)$ and Y is an iid sample from the r.v. $\mathcal{Y} \sim EC_d(\mu, \Omega, \Gamma)$ with $\mu = \Omega_{\mathcal{X}^*}^{-\frac{1}{2}}(\mu_{\mathcal{Y}^*} - \mu_{\mathcal{X}^*})$ and $\Omega = \Omega_{\mathcal{X}^*}^{-\frac{1}{2}} \Omega_{\mathcal{Y}^*} \Omega_{\mathcal{X}^*}^{-\frac{1}{2}}$.

(2) The samples (X^*, Y^*) *and* (X, Y) *have the same DD-Plot, so for any* $z^* \in \mathbb{R}^d$ *and* $z = \Omega_{\mathcal{X}^*}^{-\frac{1}{2}} (z^* - \mu_{\mathcal{X}^*})$ *holds:* $\mathcal{D}(z^*|X^*) = \mathcal{D}(z|X)$ *and* $\mathcal{D}(z^*|Y^*) = \mathcal{D}(z|Y)$.

(3) $\Sigma_{\mathcal{X}^*} \geq \Sigma_{\mathcal{Y}^*} \Leftrightarrow \Sigma_{\mathcal{X}} \geq \Sigma_{\mathcal{Y}} \Leftrightarrow 0 < \varphi_k \leq 1$ *for all eigenvalues* $\varphi_1, \ldots, \varphi_d$ *of* Ω.

The proof of Lemma 4 can be found in the appendix. Thanks to this lemma, w.l.o.g. it suffices to assume a spherically distributed data set X and an elliptically distributed data set Y in the following analytical considerations as well as in the later simulation studies.

Lemma 5. *Let* $X = (x_1, \ldots, x_n)$ *be an iid sample from r.v.* $\mathcal{X}$ *and* $Y = (y_1 \ldots, y_m)$ *an iid sample from r.v.* $\mathcal{Y}$. *Futhermore let* $D_M(z|\mathcal{X}) = \left(1 + (z - \mu_{\mathcal{X}})^\top \Sigma_{\mathcal{X}}^{-1} (z - \mu_{\mathcal{X}})\right)^{-1}$ *and* $D_M(z|\mathcal{Y}) = \left(1 + (z - \mu_{\mathcal{Y}})^\top \Sigma_{\mathcal{Y}}^{-1} (z - \mu_{\mathcal{Y}})\right)^{-1}$ *be the Mahalanobis depth functions on the theoretical moments instead the empirical moments. Then, the following holds:*

$$\mathbb{P}\big[\lim_{n\to\infty} \sup_{z\in\mathbb{R}^d} |\mathcal{D}_M(z|X) - \mathcal{D}_M(z|\mathcal{X})| = 0\big] = 1 = \mathbb{P}\big[\lim_{m\to\infty} \sup_{z\in\mathbb{R}^d} |\mathcal{D}_M(z|Y) - \mathcal{D}_M(z|\mathcal{Y})| = 0\big].$$

Lemma 5 can be found with its proof in Dyckerhoff (2016) as example 4.1 of corollary 4.1. Although in practice we compute all depths w.r.t. the given samples ($\mathcal{D}_M(z|X)$ and $\mathcal{D}_M(z|Y)$), in the following discussion we will focus on depth w.r.t. the underlying distribution ($\mathcal{D}_M(z|\mathcal{X})$ and $\mathcal{D}_M(z|\mathcal{Y})$). This is sufficient, because the empirical depth values converge uniformly to the theoretical depth values:

Theorem 6. *Let* $\mathcal{X} \sim S_d(\Gamma)$ *and* $\mathcal{Y} \sim EC_d(\mu, \Omega, \Gamma)$ *with following properties:*

- $\mathcal{X}$ *and* $\mathcal{Y}$ *have the same characteristic generator* $\Gamma(\cdot)$ *and rank*$(\Omega) = d$.
- $\mathbb{E}(r_d^2) < \infty$ *for the generating variate* $r_d \geq 0$ *related to* $\Gamma(\cdot)$ *and* $\Sigma_{\mathcal{X}} \geq \Sigma_{\mathcal{Y}}$.

Then $\Omega = Q^\top D Q$ *with orthogonal* $Q \in \mathbb{R}^{d\times d}$ *and* $D = diag[\varphi_1, \ldots, \varphi_d]$, *where* $0 < \varphi_k \leq 1$ *for all* $1 \leq k \leq d$, *and for randomly drawn* x *from* $\mathcal{X}$ *and* y *from* $\mathcal{Y}$ *holds:*

- $\mathcal{D}_M(x|\mathcal{X}) = \left(1 + ||x||^2 \left(\mathbb{E}(r_1^2)\right)^{-1}\right)^{-1}$.
- $\mathcal{D}_M(x|\mathcal{Y}) = \left(1 + \left(||x - \mu||^2 + \left|\left|\sqrt{(\mathcal{I}_d - D) D^{-1}} Q (x - \mu)\right|\right|^2\right) \left(\mathbb{E}(r_1^2)\right)^{-1}\right)^{-1}$.
- $\mathcal{D}_M(y|\mathcal{X}) = \left(1 + ||y||^2 \left(\mathbb{E}(r_1^2)\right)^{-1}\right)^{-1}$.
- $\mathcal{D}_M(y|\mathcal{Y}) = \left(1 + \left|\left|\sqrt{D^{-1}} Q (y - \mu)\right|\right|^2 \left(\mathbb{E}(r_1^2)\right)^{-1}\right)^{-1}$.
- $\mathcal{D}_M(x|\mathcal{X}) \stackrel{d}{=} \mathcal{D}_M(y|\mathcal{Y})$.

Here $r_1 \geq 0$ *is the generating variate related to the same characteristic generator* $\Gamma(\cdot)$.

Thanks to Theorem 6, whose proof can be found in the appendix, conclusions can now be drawn regarding the "migration" of the points of a *DD*-Plot of data nests by focus on points randomly drawn from $\mathcal{X}$ and $\mathcal{Y}$. The corresponding proof of these corollaries can also be found in the appendix:

Corollary 1. *Let $\mathcal{X}$ and $\mathcal{Y}$ be r.v. in $\mathbb{R}^d$ satisfying all properties of* ***Theorem*** *6 and $\{\varphi_k\}_{k=1}^d$ the eigenvalues of corresponding matrix Ω. If the cdf $F_{r_d^2}(\cdot)$ is continuous and strictly monotone, then the following holds for a randomly drawn x from $\mathcal{X}$ as well as randomly drawn y from $\mathcal{Y}$ for increasing d:*

$$\underset{d\to\infty}{plim}\, \mathcal{D}_M(x|\mathcal{X}) = 0 = \underset{d\to\infty}{plim}\, \mathcal{D}_M(y|\mathcal{Y}) \;\; and \;\; \underset{d\to\infty}{plim}\, \mathcal{D}_M(x|\mathcal{Y}) = 0 = \underset{d\to\infty}{plim}\, \mathcal{D}_M(y|\mathcal{X}).$$

The migration of the points of the *DD*-Plot to the origin therefore directly depends on the dimension of the original data or, looking at the distribution itself, on the dimension of $\mathcal{X}$ and $\mathcal{Y}$. The following corollaries will focus on the different "speeds" of said migration:

Corollary 2. *Let $\mathcal{X}$ and $\mathcal{Y}$ be r.v. in $\mathbb{R}^d$ satisfying all properties of* ***Theorem*** *6 and $\{\varphi_k\}_{k=1}^d$ the eigenvalues of corresponding matrix Ω. If the cdf $F_{r_d^2}(\cdot)$ is continuous and strictly monotone, then the following holds for a randomly drawn x from $\mathcal{X}$ and randomly drawn y from $\mathcal{Y}$*

$$\left(1 + \frac{(||x|| + ||\mu||)^2}{\varphi_d^- \mathbb{E}(r_1^2)}\right)^{-1} \leq \mathcal{D}_M(x|\mathcal{Y}) \leq \left(1 + \frac{(||x|| - ||\mu||)^2}{\varphi_d^+ \mathbb{E}(r_1^2)}\right)^{-1}$$

with $\varphi_d^+ = max\{\varphi_k\}_{k=1}^d$ and $\varphi_d^- = min\{\varphi_k\}_{k=1}^d$. Moreover, for any $0 < \alpha < 1$ holds:

$$\frac{\mathbb{E}(r_1^2)}{\mathbb{E}(r_1^2) + \frac{||\mu||^2}{p_d}} \geq \alpha \Rightarrow \mathbb{P}[\mathcal{D}_M(x|\mathcal{Y}) < \alpha] \geq \mathbb{P}[\mathcal{D}_M(y|\mathcal{Y}) < \alpha]$$

with $p_d = \min\{\varphi_d^+, \left(1 - \sqrt{\varphi_d^+}\right)^2\}$.

If $\frac{\mathbb{E}(r_1^2)}{\mathbb{E}(r_1^2) + \frac{||\mu||^2}{p_d}}$ is close to 1, most α will satisfy the inequatlity of the corollary. This inequality tells us that the probability of a low depth-value for x w.r.t. $\mathcal{Y}$ is higher than for y w.r.t. $\mathcal{Y}$: So in the y-direction of the *DD*-Plot points from $\mathcal{X}$ go faster to 0 than points from $\mathcal{Y}$.
In which cases will $\frac{\mathbb{E}(r_1^2)}{\mathbb{E}(r_1^2) + \frac{||\mu||^2}{p_d}}$ be close to 1? First, if $||\mu||$ is very small, which mean that we have a huge overlapping of $\mathcal{X}$ and $\mathcal{Y}$. Second, if $p_d = \min\{\varphi_d^+, \left(1 - \sqrt{\varphi_d^+}\right)^2\}$ is as large as possible. This is the case if the following holds:

$$\varphi_d^+ = \left(1 - \sqrt{\varphi_d^+}\right)^2 \quad \Leftrightarrow \quad 2\sqrt{\varphi_d^+} = 1 \quad \Leftrightarrow \quad \varphi_d^+ = \frac{1}{4} = 0.25$$

So, all eigenvalues should be bounded by 0.25 and therefore the data nest characteristic of $\mathcal{Y}$ is strong. If we have such a situation, then also in the x-direction points from $\mathcal{X}$ will go faster to 0 than points from $\mathcal{Y}$ in the *DD*-Plot as is shown in the following corollary:

Corollary 3. *Let $\mathcal{X}$ and $\mathcal{Y}$ be r.v. in $\mathbb{R}^d$ satisfying all properties of* ***Theorem*** *6 and $\{\varphi_k\}_{k=1}^d$ the eigenvalues of corresponding matrix Ω. If the cdf $F_{r_d^2}(\cdot)$ is continuous and strictly monotone, then the following holds for randomly drawn x from $\mathcal{X}$ and randomly drawn y from $\mathcal{Y}$ with corresponding $z \overset{d}{=} x \sim S_d(\Gamma)$:*

$$\left(1 + \frac{\left(\sqrt{\varphi_d^+}\,||z|| + ||\mu||\right)^2}{\mathbb{E}\left(r_1^2\right)}\right)^{-1} \leq \mathcal{D}_M\left(y|\mathcal{X}\right) \leq \left(1 + \frac{\left(\sqrt{\varphi_d^-}\,||z|| - ||\mu||\right)^2}{\mathbb{E}\left(r_1^2\right)}\right)^{-1}$$

with $\varphi_d^+ = max\{\varphi_k\}_{k=1}^d$ and $\varphi_d^- = min\{\varphi_k\}_{k=1}^d$. Moreover, for any $0 < \alpha < 1$ holds:

$$\frac{\mathbb{E}\left(r_1^2\right)}{\mathbb{E}\left(r_1^2\right) + \frac{||\mu||^2}{\left(1-\sqrt{\varphi_d^+}\right)^2}} \geq \alpha \Rightarrow \mathbb{P}\left[\mathcal{D}_M\left(x|\mathcal{X}\right) < \alpha\right] \geq \mathbb{P}\left[\mathcal{D}_M\left(y|\mathcal{X}\right) < \alpha\right].$$

We get the same inequality as in corollary 2 by setting $p_d = \left(1 - \sqrt{\varphi_d^+}\right)^2$. Overall, for $\mathcal{Y}$ with strong data nest characteristic in an overlapping setting, we get a different "migration speed" of the points in the *DD*-Plot. This is even more interesting because situations of overlapping data nests result in a non tivial classification problem.
In corollary 2 and corollary 3 it would be nice to show some more properties in cases of a different behavior of $||\mu||$, but for this we need more information about the structure of $F_{r_d^2}(\cdot)$, which can be easily seen by looking at the proof in the appendix.

2.4 Simulation study

2.4.1 Data generating process

To compare the hybrid method *DD*-svm with simple svm by means of a simulation study, both models should be trained on the same data nest $\left(X^T, Y^T\right)$, i.e. the training set, in order to measure their respective performance on a new data nest $\left(X^V, Y^V\right)$, i.e. the validation set. Both sets should be identical distributed. Hereafter, X is meant to be understood as synonymous to sets X^T and X^V and correspondingly, Y to sets Y^T and Y^V. When generating the data, the analytical results should be taken into account: X is spherically and Y elliptically distributed, which is sufficient according to Lemma 4. Since data dimension d and center distance $||\mu||$ affect the points in the *DD*-Plot, 100 different data nests are generated according to following description in a Monte-Carlo-procedure:

DGP I

- **Step 1:** Choose dimension $d \in \{5, 10, \ldots, 30\}$ and $\nu \in \{0, 0.5, 1\}$.
- **Step 2:** Generate 100 iid samples $X^T = \left(x_1^T, \ldots, x_n^T\right)$ and $X^V = \left(x_1^V, \ldots, x_n^V\right)$ with $n = 9900$ points $x_1^l, \ldots, x_n^l \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \mathcal{I}_d)$ for $l \in \{T, V\}$ by Monte-Carlo-procedure.
- **Step 3:** Also by Monte-Carlo-procedure, generate 100 random orthogonal matrices $Q \in \mathbb{R}^{d \times d}$ and diagonal-matrices $D = \text{diag}[\varphi_1, \ldots, \varphi_d]$ with $\varphi_1, \ldots, \varphi_d \overset{\text{iid}}{\sim} U(0, 1)$ to get iid samples $Y^T = \left(y_1^T, \ldots, y_m^T\right)$ and $Y^V = \left(y_1^V, \ldots, y_m^V\right)$ with $m = 100$ points $y_1^l, \ldots, y_m^l \overset{\text{iid}}{\sim} \mathcal{N}\left(\mu, Q^\top D Q\right)$ where $\mu = \left(\sqrt{\frac{\nu}{d}}, \ldots, \sqrt{\frac{\nu}{d}}\right)^\top \in \mathbb{R}^d$ for $l \in \{T, V\}$.

In this context, $\mathbf{0}$ denotes the zero vector $\mathbb{R}^d$. For all data nest scenarios, Y makes up 1% of the total $100 + 9900 = 10000$ data points, meaning that the first condition for data nests is fulfilled. The second condition is also satisfied because of Lemma 4, since with every new data nest Y, the eigenvalues of the covariance matrix are uniformly drawn from the interval $[0, 1]$. In addition, the individual covariance matrices differ in that the orthogonal matrix Q is also randomly generated for each data nest Y. Furthermore, it will be tested how *DD*-svm behaves in relation to svm in those cases in which, judging by the structure, a data nest is present even though there are no elliptically distributed data sets. For this, data sets were generated according to the following description:

DGP II

- **Step 1:** Choose dimension $d \in \{5, 10, \ldots, 30\}$ and $\tau \in \{0, 0.075, 0.15\}$.
- **Step 2:** Generate 100 iid samples $X^T = \left(x_1^T, \ldots, x_n^T\right)$ and $X^V = \left(x_1^V, \ldots, x_n^V\right)$ with $n = 9900$ points $x_1^l, \ldots, x_n^l \overset{\text{iid}}{\sim} \text{Beta}_d(\mathbf{3}, \mathbf{3})$, for $l \in \{T, V\}$ by Monte-Carlo-procedure.
- **Step 3:** With $T = 28\left(0.25 - \frac{\tau^2}{d}\right) - 1$, generate also by Monte-Carlo-procedure 100 vectors $\mathbf{t} \in \mathbb{R}^d$ with $\mathrm{t}_1, \ldots, \mathrm{t}_d \overset{\text{iid}}{\sim} U(T, 5T)$. With these generate $\mathbf{p} = \mathbf{t}\left(0.5 + \frac{\tau}{\sqrt{d}}\right)$ and $\mathbf{q} = \mathbf{t}\left(0.5 - \frac{\tau}{\sqrt{d}}\right)$. $Y^T = \left(y_1^T, \ldots, y_m^T\right)$ and $Y^V = \left(y_1^V, \ldots, y_m^V\right)$ with $m = 100$ points are then generated iid by $y_1^l, \ldots, y_m^l \overset{\text{iid}}{\sim} \text{Beta}_d(\mathbf{p}, \mathbf{q})$ for $l \in \{T, V\}$.

In this context, $\mathbf{3}$ denotes vector $(3, \ldots, 3) \in \mathbb{R}^d$ and $\text{Beta}_d(\mathbf{p}, \mathbf{q})$ stands for:

$$\mathbf{z} \sim \text{Beta}_d(\mathbf{p}, \mathbf{q}) \Leftrightarrow \mathrm{z}_k \sim \text{Beta}(\mathrm{p}_k, \mathrm{q}_k),\ \mathrm{z}_1, \ldots, \mathrm{z}_d \text{ are independent.}$$

The variables of all dimensions are thus beta-distributed and independent of each other. While all variables follow the same beta distribution with respect to X, the parameters of the beta distributions for Y do not only differ between the different data nests but also between the different variables. For the value of T in step 3 and to illustrate that data nests

are actually generated by **DGP II**, a corresponding proof can be found in the appendix. On each of these respective 100 samples, which accordingly differ with regard to the dimension of the data d, the distance of the center points μ as well as the underlying distribution type, the respective model is first trained on the training set by performing a 5-fold cross validation via a grid of the parameters (C, γ). The combination of parameters with the highest true positive rate (TPR) is regarded as ideal in this context. This is due to the fact that the focus lies on the prognosis of the rare events. Once the respective optimal parameters have been found on the respective training quantities, the performance of the models, which have been trained in this manner, is measured on the corresponding validation quantities and the average of all 100 scenarios is calculated for each parameter variation. The following quantities are collected in the process:

- The average true positive rate (TPR) in percentage [%]
- The average true negative rate (TNR) in percentage [%]
- The average accuracy (ACC) in percentage [%]
- The average computational time (t) in seconds [s]

With this approach, attention should be paid to the grid on which the training takes place: Following to Hsu et al. (2003), $C_I = (2^{-5}, 2^{-2}, \ldots, 2^{15})$ and for $\gamma_I = (2^{-15}, 2^{-12}, \ldots, 2^3)$ are chosen for svm and *DD*-svm. For *DD*-svm, the grid $C_{II} = (2^{-6}, 2^{-4}, \ldots, 2^{16})$ and $\gamma_{II} = (2^{-16}, 2^{-14}, \ldots, 2^6)$, which is much wider, is additionally tried out. The reason for taking a wider grid is the fact that *DD*-svm have a much lower computational time as we will see. Therefore more combinations of C and γ can be tried out.
Also, a light model variant of *DD*-svm is tested: In order to slow down the "migration speed" as a whole, the entire *DD*-Plot should be logarithmized before applying svm to it. This approach is intended to counteract a too rapid concentration at the origin, which could negatively influence the training of the svm. This issue will be the focus of the next Section. By taking logarithm, values close to 0 will decrease to infinity but in doing so, the separability of the point may also increase. This model variant, also referred to as log-*DD*-svm, is also trained on the two grid types. Therefore, five different models are tested:

- svm on grid (C_I, γ_I), referred as svm
- *DD*-svm on grid (C_I, γ_I), referred as *DD*-svm-I
- *DD*-svm on grid (C_{II}, γ_{II}), referred as *DD*-svm-II
- log-*DD*-svm on grid (C_I, γ_I), referred as log-*DD*-svm-I
- log-*DD*-svm on grid (C_{II}, γ_{II}), referred as log-*DD*-svm-II

2.4.2 Results

The means of the collected indicators are listed in tables: Tables 2.1 and 2.2 refer to data nests that were generated in accordance with **DPG I**. Accordingly, the nests are normally distributed. By contrast, Tables 2.3 and 2.4 build on data nests generated pursuant to **DGP II** and therefore, they consist of independent variables that are beta-distributed. In addition, the results from Tables 2.2 and 2.4 are graphically shown in Figures 2.4.6 and 2.4.7. Three central observations are worth mentioning: First, the models do not differ significantly with regard to ACC and TNR. Comparing the values in Tables 2.1 and 2.3, all values are close to 100%, regardless of the underlying distribution, dimension, or distance of the centers. This was to be expected because rare events are present which means that TNR and ACC, if no sampling methods are used, naturally assume high values. This, in turn, weakens the significance of these values (see chp. 4 for more discussion). In addition, the focus is on the rare events, which is why the TNR is of no relevance in the context of the research question. Consequently, these indicators should not matter in the assessment. Secondly, the TPR only corresponds to some extent to the analytical considerations in the hybrid models: With increasing dimensions, the TPR does certainly increase, however, the rate of this increase decreases more and more. In some cases, one can even observe a decrease of the TPR in high dimensions (see Figure 2.4.6 and 2.4.7). The reason for this could be the concentration towards the origin and the associated obstruction of the training of the models. This is supported by the fact that in logarithmised models, this effect of the weakening of the TPR in high dimensions is less pronounced. With a larger center distance, the increase but also the weakening of this increase of the TPR is stronger in high dimensions. Again, this suggests that too fast a concentration obstructs the training of models. Finally, a higher $||\mu||$ increases the probability that low depth values occur in at least one direction of the *DD*-Plot. Overall, it should be noted that svm has the highest TPR values, followed by log-*DD*-svm and *DD*-svm. Differences in the hybrid models with respect to the different grids are negligible. In the beta-distributed data, the distance of the TPR values of the svm to the other models is very strong, while the normally distributed data show little difference, at least in low dimensions or with high center distances. This coincides with the fact that the Mahalanobis depth is best applied to elliptical distributions (see end of Section 2.2.1). Thirdly, it can be observed that the behavior of the average computational time is very different: In all cases, the calculation time for svm increases almost linearly with the dimension of the data. In the hybrid models, however, it drops exponentially. With an increase of the distance between the centers, the computation time drops drastically even in small dimensions in the hybrid models, whereas no change is observable in svm. The hybrid models that are trained on the finer grid require more computation time than the other hybrid models, which can easily be attributed to the greater complexity of the grid search.

DGP I		svm		*DD*-svm-I		log-*DD*-svm-I		*DD*-svm-II		log-*DD*-svm-II	
τ	d	ACC	TNR	ACC	TNR	ACC	TNR	ACC	TNR	ACC	TNR
0	5	97.60	98.52	98.84	99.81	98.98	99.97	98.40	99.35	98.05	98.99
	10	97.75	98.55	98.84	99.71	98.94	99.83	98.65	99.51	98.51	99.35
	15	98.29	98.98	98.99	99.72	99.06	99.80	98.94	99.67	98.85	99.57
	20	98.73	99.32	99.18	99.81	99.23	99.84	99.18	99.80	99.16	99.76
	25	98.89	99.42	99.34	99.89	99.35	99.90	99.34	99.89	99.34	99.88
	30	99.07	99.56	99.42	99.95	99.45	99.95	99.41	99.95	99.45	99.95
0.5	5	98.08	98.89	98.75	99.63	98.90	99.79	98.50	99.35	98.39	99.23
	10	98.99	99.45	99.28	99.78	99.29	99.77	99.26	99.76	99.26	99.75
	15	99.46	99.74	99.62	99.90	99.63	99.91	99.62	99.90	99.63	99.91
	20	99.66	99.83	99.74	99.95	99.76	99.96	99.74	99.95	99.75	99.95
	25	99.79	99.90	99.78	99.99	99.81	99.99	99.78	99.99	99.81	99.99
	30	99.86	99.93	99.78	100.00	99.84	100.00	99.78	100.00	99.84	100.00
1	5	98.64	99.28	98.96	99.64	99.03	99.72	98.86	99.54	98.90	99.56
	10	99.56	99.79	99.67	99.88	99.68	99.89	99.66	99.88	99.67	99.89
	15	99.82	99.91	99.86	99.97	99.87	99.97	99.86	99.97	99.87	99.97
	20	99.92	99.96	99.92	99.99	99.94	99.99	99.92	99.99	99.94	99.99
	25	99.97	99.99	99.92	100.00	99.96	100.00	99.92	100.00	99.96	100.00
	30	99.98	99.99	99.91	100.00	99.97	100.00	99.91	100.00	99.97	100.00

Table 2.1: The average ACC in [%] and TNR in [%] of all models for increasing dimension d of data generated by **DGP I**.

DGP I		svm		*DD*-svm-I		log-*DD*-svm-I		*DD*-svm-II		log-*DD*-svm-II	
τ	d	TPR	time	TPR	time	TPR	time	TPR	time	TPR	time
0	5	6.51	469.02	2.87	730.28	1.31	482.67	4.12	2524.66	5.54	2891.62
	10	18.70	736.01	13.06	378.11	11.19	266.42	13.41	1187.93	15.20	1085.88
	15	30.42	1309.62	26.48	175.82	26.22	139.59	26.51	564.17	28.06	553.98
	20	40.20	1636.12	37.57	116.80	38.54	90.39	38.09	420.15	39.49	418.61
	25	46.27	1715.80	44.93	101.75	45.66	74.86	44.73	378.32	46.16	395.62
	30	50.51	2159.81	46.44	99.35	50.17	72.03	46.41	362.53	50.41	395.09
0.5	5	17.83	242.88	11.83	405.93	10.69	308.05	13.91	1250.35	15.27	1263.24
	10	53.05	849.63	49.96	108.77	51.08	98.00	49.97	429.82	50.73	436.11
	15	72.07	1306.94	71.87	93.60	72.08	73.18	71.41	376.93	71.85	387.85
	20	82.42	1699.75	78.50	90.32	79.69	64.31	78.47	356.29	79.86	384.74
	25	89.25	2085.00	79.47	91.92	82.46	71.35	79.46	342.67	82.58	407.99
	30	93.50	2355.95	78.62	97.00	84.54	72.46	78.43	342.60	84.44	392.04
1	5	35.38	215.97	31.24	190.05	31.31	152.66	31.88	581.33	32.93	793.18
	10	77.00	890.88	78.11	92.25	78.24	70.45	78.24	348.51	78.22	382.83
	15	91.25	1264.32	89.27	83.70	90.33	61.82	89.77	332.47	90.32	374.58
	20	95.92	1808.83	92.63	84.42	94.57	67.43	92.70	310.28	94.65	376.81
	25	98.31	2153.41	92.27	88.20	96.16	66.18	92.39	309.65	96.22	389.60
	30	99.03	2485.25	90.61	90.85	97.31	68.05	90.86	318.47	97.38	382.09

Table 2.2: The average TPR in [%] and computational time in [s] of all models for increasing dimension d of data generated by **DGP I**.

DGP II		svm		*DD*-svm-I		log-*DD*-svm-I		*DD*-svm-II		log-*DD*-svm-II	
ν	d	ACC	TNR	ACC	TNR	ACC	TNR	ACC	TNR	ACC	TNR
0	5	97.44	98.36	98.83	99.82	98.97	99.97	98.35	99.32	98.13	99.08
	10	97.71	98.54	98.80	99.72	98.90	99.83	98.68	99.59	98.54	99.43
	15	98.44	99.11	98.91	99.72	98.97	99.77	98.90	99.71	98.81	99.60
	20	98.87	99.40	99.14	99.82	99.15	99.82	99.14	99.81	99.13	99.80
	25	99.15	99.58	99.28	99.89	99.30	99.90	99.28	99.89	99.30	99.89
	30	99.36	99.72	99.36	99.94	99.39	99.95	99.36	99.94	99.40	99.95
0.075	5	97.56	98.47	98.72	99.69	98.89	99.87	98.41	99.36	98.23	99.17
	10	98.06	98.82	98.81	99.68	98.87	99.74	98.60	99.46	98.61	99.45
	15	98.87	99.42	99.09	99.77	99.10	99.77	99.08	99.76	99.04	99.71
	20	99.20	99.59	99.30	99.85	99.31	99.85	99.30	99.85	99.30	99.84
	25	99.46	99.74	99.43	99.92	99.43	99.92	99.43	99.91	99.42	99.91
	30	99.63	99.84	99.46	99.97	99.47	99.97	99.46	99.97	99.47	99.97
0.15	5	98.04	98.87	98.72	99.61	98.83	99.75	98.48	99.35	98.33	99.19
	10	99.08	99.55	99.23	99.78	99.24	99.78	99.20	99.76	99.22	99.77
	15	99.56	99.78	99.59	99.89	99.60	99.89	99.59	99.89	99.60	99.89
	20	99.76	99.88	99.73	99.95	99.74	99.96	99.73	99.95	99.74	99.95
	25	99.87	99.93	99.77	99.99	99.79	99.99	99.76	99.99	99.79	99.99
	30	99.93	99.97	99.75	100.00	99.80	100.00	99.75	100.00	99.79	100.00

Table 2.3: The average ACC in [%] and TNR in [%] of all models for increasing dimension d of data generated by **DGP II**.

DGP II		svm		*DD*-svm-I		log-*DD*-svm-I		*DD*-svm-II		log-*DD*-svm-II	
ν	d	TPR	time	TPR	time	TPR	time	TPR	time	TPR	time
0	5	5.82	450.72	1.74	785.27	0.46	503.47	2.98	2632.38	3.85	2920.96
	10	16.02	728.67	7.89	428.77	7.36	311.78	8.53	1308.14	10.03	1209.62
	15	31.22	1305.72	19.14	177.39	19.68	152.83	18.83	545.42	20.20	547.51
	20	46.54	1600.47	32.50	116.23	32.77	92.27	32.46	402.06	32.70	404.84
	25	56.65	1666.52	39.05	102.61	40.46	76.65	39.12	366.49	40.60	387.21
	30	63.83	2088.66	41.33	99.26	44.07	71.33	41.63	352.90	44.21	382.76
0.075	5	7.95	374.55	3.16	745.64	1.54	479.98	4.43	2314.56	5.32	2105.98
	10	22.64	758.68	13.09	250.78	13.16	208.20	13.04	890.77	15.25	799.57
	15	43.78	1323.40	31.97	127.96	32.90	105.88	32.27	451.75	32.78	428.75
	20	60.50	1574.88	45.00	104.17	46.01	77.73	45.02	393.98	45.63	401.98
	25	71.18	1813.61	50.88	98.38	51.00	75.85	51.12	372.40	51.10	411.29
	30	79.41	2299.25	48.93	102.06	50.11	74.36	48.86	371.57	49.90	392.16
0.15	5	16.03	226.82	9.92	411.15	8.56	294.10	11.86	1217.72	13.65	1481.23
	10	52.86	880.48	43.99	126.09	45.12	97.95	44.09	434.68	45.25	428.93
	15	77.56	1229.76	69.98	97.01	70.89	71.71	69.93	384.15	70.74	398.55
	20	88.02	1713.99	77.91	91.12	78.27	71.32	78.13	350.30	78.62	391.97
	25	93.56	2121.35	77.80	93.09	80.05	68.51	77.64	346.25	80.10	403.19
	30	96.17	2457.86	75.30	94.85	80.07	69.90	75.04	351.36	79.76	395.06

Table 2.4: The average TPR in [%] and computational time in [s] of all models for increasing dimension d of data generated by **DGP II**.

Figure 2.4.6: The average TPR in [%] (left panel) and computational time in [s] (right panel) of all models for increasing dimension d of data generated by **DGP I** with $\nu = 0$ (1^{st} line), $\nu = 0.5$ (2^{nd} line) and $\nu = 1$ (3^{rd} line).

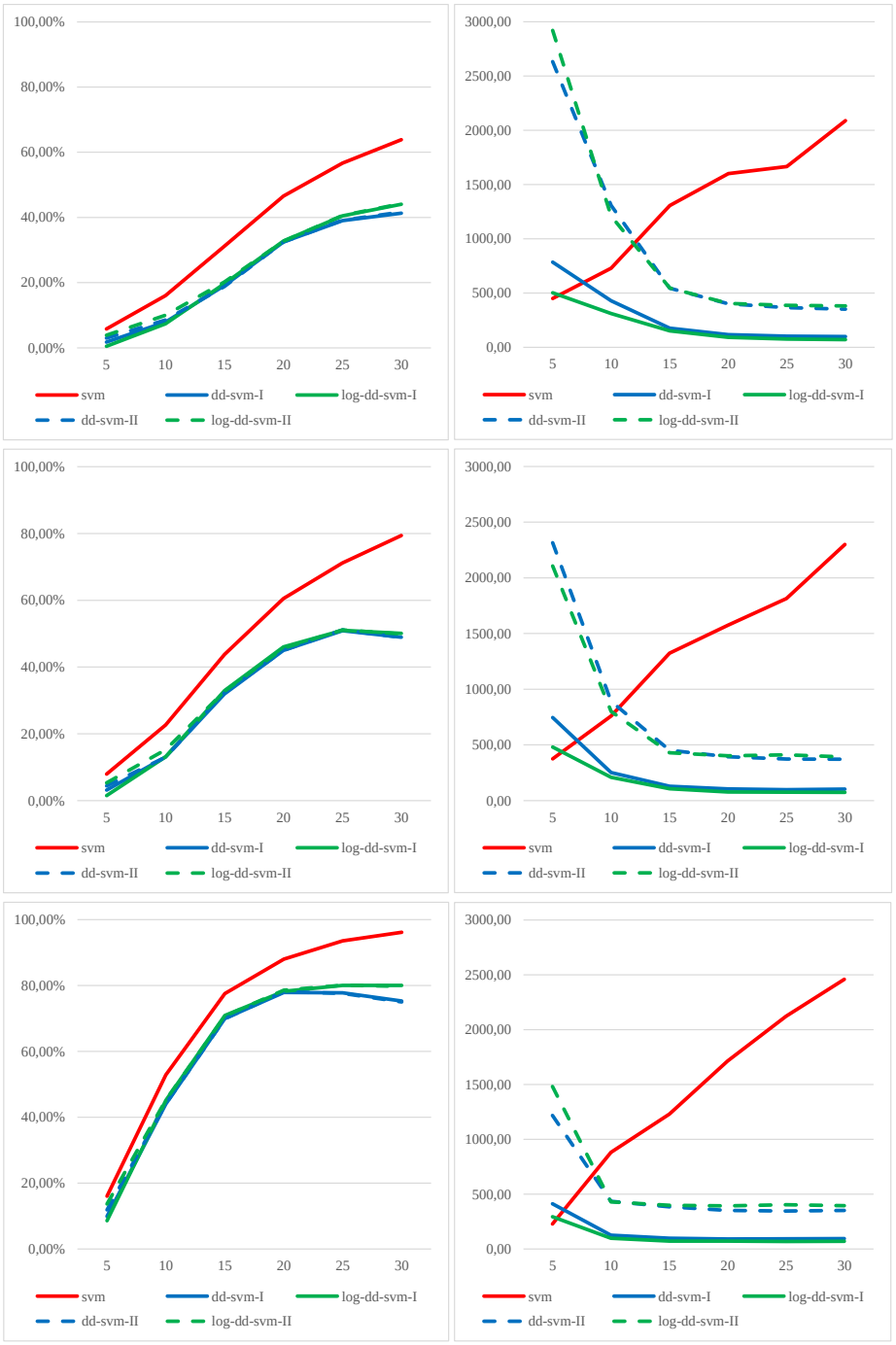


Figure 2.4.7: The average TPR in [%] (left panel) and computational time in [s] (right panel) of all models for increasing dimension d of data generated by **DGP II** with $\tau = 0$ (1^{st} line), $\tau = 0.075$ (2^{nd} line) and $\tau = 0.15$ (3^{rd} line).

Overall, it can thus be said that in comparison to svm, the hybrid model approach generates slightly weaker TPR values in the case of an elliptical distribution and significantly weaker ones in the case of a non-elliptical distribution. However, the enormous computational efficiency with which the results are generated predominates. Therefore, in the case of a "moderate" dimension, i.e. between $d = 10$ and $d = 20$, hybrid methods should be used when dealing with elliptically distributed data nests as these work fast and efficiently and produce comparably good results as simple svm.

2.5 Conclusion

Hybrid methods such as the *DD*-svm presented in this chapter are very popular in the context big data applications (see e.g. Min et al. (2006)). Combining data depths and svm as first proposed by Kim et al. (2018) is advantageous, especially when data sets are present in nest structures. Such data nests appear to be plausible, for example, in the case of cancellation analyses. Based on the individual customer characteristics, one can assume that the degree of homogeneity between cancellation customers is higher than that between the other customers (see Kahlenberg (2005)). The fact that the cancellation customers consequently accumulate in certain segments and thus form nests (see Reuß and Zwiesler (2006)) increases – according to the data depth transformation – the separability of the data, which is why svm can be used more efficiently on the *DD*-Plot than on the original data. However, this advantage, which makes use of the nest structure of the data, can be lost in too large dimensions. As the data dimension grows, the points of the *DD*-Plot concentrate at the origin, making the use of svm more difficult. Therefore, with application in practice, it will be essential to pre-select the data characteristics to be considered in order to not only avoid "too high" dimensions but above all to guarantee distinctly strong nest structures. The *DD*-svm presented here is not only advantageous when comparing it to svm but also in comparison to conventional methods which are applied in the case of rare events: While common methods are based on a sampling approach, meaning that they manipulate the data either by under- or oversampling, *DD*-svm exploits the presence of rare events in the form of data nests and renounces sampling entirely. As a result, a manipulation of the data does not take place; all available information is used and processed. It would, therefore, be of interest to analyze how a comparison between *DD*-svm and other methods using various sampling approaches would turn out (see chp. 4).

Chapter 3

Estimating factor models with generalized supervision

3.1 Introduction

Factor models are attractive for analysing high-dimensional data sets. The reason for this is that the use of factor models can make a decisive contribution to dimension reduction by replacing the many observable predictors of a model with a few non-observable diffusion indices, which we will simply refer to as factors in this chapter. Employing factor models in forecasting is typically performed in two steps. First, a small number of factors is extracted from a large set of predictors. In the second step, a dynamic regression represents the link between the target variable and the factors. Among the numerous methods of estimation, principal component analysis (PCA) is particularly popular. However, if the intended use of the model is the forecasting of a certain target variable, the PCA-based approach reveals weaknesses. The reason for this is the need for a two-stage approach: This approach only promises to be efficient enough on condition that all factors estimated from the data are also relevant for the target variable though (see, e.g., Stock and Watson (2006)). When only part of the estimated factors are of relevance to the forecast, it is of utmost importance, when estimating the factors, to consider not only the available data on possible predictors but also, and even more importantly, the target variable itself (see e.g. Bai and Ng (2008)). Factor models achieving precisely that are called supervised factor models.
In this chapter, we propose a general framework for a supervised factor model that is based on a particular rotation of the factor space that results in an optimal forecast. The main idea is to decompose the set of common factors into factors relevant to forecasting and the remaining redundant factors from the predictor space. We show that the set of relevant factors can always be combined to a single factor and, therefore, prediction based on that single factor can be shown to be optimal. This allows us to reduce the forecast exercise to finding the single relevant factor within a large set of predictors. An important advantage of our general version of a supervised factor model is that it also allows us to conveniently augment the factor model by additional variables that are not included in the factor space (e.g., the lags of the target variable).

In the following section, we offer a precise derivation of this model and illustrate how it relates to conventional approaches, in particular to the PCA approach. In doing so, the strengths of this model approach will become obvious. After that, the estimation of the model by iteratively reweighted sequential least squares (IRSLS) and the algorithmic implementation of such an estimation process will be discussed in the third section. The fourth section contains a simulation study that we use to compare our method of modeling to related approaches. By means of Monte-Carlo-experiments, we demonstrate the benefits of supervised forecasting based on a large set of predictors. The chapter concludes with a discussion of the results in the last section.

3.2 Factor models with generalized supervision

All factor models are based on the assumption that at every point in time $t = 1, \dots, T$, the common component of a variable x_{it} is generated from r factors $f_t = (f_{1t}, \dots, f_{rt})^\top \in \mathbb{R}^r$ and loading vector $\lambda_i = (\lambda_{i1}, \dots, \lambda_{ir})^\top \in \mathbb{R}^r$:

$$x_{it} = \lambda_i^\top f_t + u_{it} \tag{3.2.1}$$

The idiosyncratic component u_{it} as an error term is independent and identically distributed with $\mathbb{E}(u_{it}) = 0$ and $\mathbb{E}(u_{it} f_t) = 0$. The extent of the influence of a single factor f_{kt} on a variable x_{it} depends on the loading λ_{ik} determining how much variation is passed on from a factor to the respective variable. (3.2.1) can be represented as a matrix equation for $X = (x_1, \dots, x_n) \in \mathbb{R}^{T \times n}$ with $x_i = (x_{i1}, \dots, x_{iT}) \in \mathbb{R}^T$ through the help of the matrices $F = (f_1, \dots, f_T)^\top \in \mathbb{R}^{T \times r}$, $\Lambda = (\lambda_1, \dots, \lambda_n)^\top \in \mathbb{R}^{n \times r}$ and $U = (u_1, \dots, u_n) \in \mathbb{R}^{T \times n}$ with $u_i = (u_{i1}, \dots, u_{iT})^\top \in \mathbb{R}^T$:

$$X = F\Lambda^\top + U \tag{3.2.2}$$

Through insertion of an arbitrary regular matrix $Q \in \mathbb{R}^{r \times r}$ in (3.2.2), the common component can be rewritten in such a way that $F\Lambda^\top = FQ^{-1}Q\Lambda^\top = FQ^{-1}(\Lambda Q^\top)^\top = F'\Lambda'^\top$ applies. Consequently, the same data matrix can also be generated by the factor matrix $F' = FQ^{-1}$ and the loading matrix $\Lambda' = \Lambda Q^\top$. Therefore, futher restictions are needed to identify F and Λ. Bai and Li (2012) provide a good overview of some possible restrictions and their consequences for the estimation of factors. If $F^\top F$ is diagonal and $\Lambda^\top \Lambda = \mathcal{I}_r$ are chosen as restrictions, one can obtain the PC estimator $\hat{\Lambda}$ for the loading matrix by applying OLS to (3.2.1) or (3.2.2): The columns of $\hat{\Lambda}$ are equal to the first r eigenvectors of $\frac{1}{T}\sum_{t=1}^{T}(x_t - \overline{x})(x_t - \overline{x})^\top$, the covariance matrix of X, and $\hat{F} = X\hat{\Lambda}^\top$. For the selection of r, i.e., the number of factors, one has to use certain criteria (see Bai and Ng (2002) for some examples). Given the factors, they can be used together with additional variables

$Z = (z_1, \ldots, z_T)^\top \in \mathbb{R}^{m \times T}$ with $z_t = (z_{1t}, \ldots, z_{mt})^\top \in \mathbb{R}^m$ to forecast y_{t+h}:

$$y_{t+h} = \alpha^\top z_t + \beta^\top f_t + \varepsilon_{t+h} \tag{3.2.3}$$

$$y = Z\alpha + F\beta + \varepsilon \tag{3.2.4}$$

In (3.2.3), ε_{t+h} is, again as an error term, independent and identically distributed with $\mathbb{E}(\varepsilon_{t+h}) = 0$ and $\mathbb{E}(\varepsilon_{t+h} z_t) = 0$ such as $\mathbb{E}(\varepsilon_{t+h} f_t) = 0$. h corresponds to the forecast horizon for $y = (y_{1+h}, \ldots, y_{T+h}) \in \mathbb{R}^T$. The associated parameter vectors $\alpha = (\alpha_1, \ldots, \alpha_m)^\top \in \mathbb{R}^m$ and $\beta = (\beta_1, \ldots, \beta_r)^\top \in \mathbb{R}^r$ can only be estimated from (3.2.3) or (3.2.4) if the corresponding factors are given. However, these factors first have to be estimated. The two-step approach of first estimating the factors from the data matrix X using PCA and then regressing y on these factors along with Z does have one major disadvantage though: It is not guaranteed that all factors relevant for X are also relevant to y. More particularly, it is possible that some factors are able to explain the data in X very well but have nothing to contribute to the explanation of y. It would, therefore, be more appropriate to only take those factors for forecasting into account which explain y.

In order to overcome this disadvantage, it is advisable to implement a form of supervision when estimating the factors. This means that the factor estimation is carried out under simultaneous consideration of the target variable y. de Jong and Kiers (1992) put forward one approach of how such a supervised factor model can be designed: In the Principal Covariate Regression (PCovR), the values for F, Λ and β are estimated using a so-called criterion function. This function summarizes (3.2.2) or (3.2.4) without considering additional variables Z:

$$C_\theta(F, \Lambda, \beta) = \theta \frac{(y - F\beta)^\top (y - F\beta)}{||y||^2} + (1 - \theta) tr \left\{ \frac{(X - F\Lambda^\top)^\top (X - F\Lambda^\top)}{||X||^2} \right\} \tag{3.2.5}$$

The selection of the supervision parameter θ reflects the trade-off between the simpler regression of y on the factors ($\theta \to 1$) and PCA ($\theta \to 0$). However, in the extreme case ($\theta = 1$), this results in an inability to estimate the factors themselves because no information about the data X is included in their estimation. Conversely, a mere PC estimate ($\theta = 0$) would disregard supervision from y. Once θ has been selected, the minimization of (3.2.5) subject to the normalization restriction $T^{-1} F^\top F = \mathcal{I}_r$ results in the calculation of the eigenvalues of the following matrix:

$$\theta \frac{(X^\top X)^{-1} (y^\top X)^\top y^\top X}{||y||^2} + (1 - \theta) \frac{X^\top X}{T||X||^2} \tag{3.2.6}$$

Let C_r be the matrix constructed from the first r eigenvectors of this matrix. In that case, the supervised factor estimation holds that $\hat{F} = XC_r$. Also $\hat{\Lambda}$ and $\hat{\beta}$ can be estimated by a regression of X and y on $\hat{F}$ (see de Jong and Kiers (1992) for details). To calculate a solution in (3.2.6), θ has to be selected. Umbach (2020) proposed a criterion for this selection.

However, it is inherently based on the assumption that all factors from X are also relevant to y. This is also the reason why this criterion is a basis for decision-making not only for the selection of θ but also for the selection of the number of factors r (see Umbach (2020) for details). Moreover, the influence of additional variables Z on y is disregarded in PCovR since the part $Z\alpha$ of the predictive regression in (3.2.4) cannot be implemented in the criterion function (3.2.5) without taking into account that α is also unknown. Therefore, $C_\theta(F, \Lambda, \beta)$ becomes $C_\theta(F, \Lambda, \beta, \alpha)$ and the optimization problem is no longer solvable by the methods proposed in de Jong and Kiers (1992). Alternative approaches based on singular value decomposition (SVD, see Heij et al. (2007) for details) also failed in calculating a solution in this situation. The only way of integrating Z is again a two-stage approach: After the estimation of the factors $\hat{F}$, the parameters $\hat{\alpha}$ can be estimated like $\hat{\beta}$ by a regression of y on additional variables Z and $\hat{F}$. This leads to the question how these additional variables could be integrated in a supervised factor model in a more elegant way.

We propose a supervised factor model that aims to completely eliminate both disadvantages: On the one hand, the supervision will be designed in such a way that, when estimating the factors, a distinction is made on the basis of whether a factor from X is of relevance to y or not. On the other hand, the additional variables contained in Z will also be included directly in the factor estimation without losing the one-stage construction. To achieve the first goal, factor matrix F is set to matrix $F = \begin{pmatrix} f & G \end{pmatrix}$ with the first vector of factors $f = (f_1, \ldots, f_T)^\top \in \mathbb{R}^T$ representing all time-varying effects that are relevant for y and simultaneously affect X. The remainder matrix $G = \begin{pmatrix} g_1 & \ldots & g_s \end{pmatrix} \in \mathbb{R}^{T \times s}$ with factor-vectors $g_k = (g_{k1}, \ldots, g_{kT})^\top \in \mathbb{R}^T$ summarizes the other $s = r - 1$ factors that are only relevant for X. Accordingly, only the factor f is used for the predictive regression of y:

$$X = \begin{pmatrix} f & G \end{pmatrix} \begin{pmatrix} \varphi & \Phi \end{pmatrix}^\top + U \tag{3.2.7}$$

$$y = Z\alpha + f\gamma + \varepsilon = \begin{pmatrix} Z & f \end{pmatrix} \begin{pmatrix} \alpha \\ \gamma \end{pmatrix} + \varepsilon \tag{3.2.8}$$

Here, the loadings Λ are set to $\Lambda = \begin{pmatrix} \varphi & \Phi \end{pmatrix}$ as well. While $\varphi = (\varphi_1, \ldots, \varphi_n)^\top \in \mathbb{R}^n$ contains those loadings that correspond to factor f, $\Phi = \begin{pmatrix} \phi_1 & \ldots & \phi_s \end{pmatrix} \in \mathbb{R}^{n \times s}$ with $\phi_k = (\phi_{k1}, \ldots, \phi_{kn})^\top \in \mathbb{R}^n$ summarizes all other loadings of all the rest of factors G. This representation or the use of only one factor and corresponding one-dimensional parameter $\gamma \in \mathbb{R}$ in the regression of y is possible because (3.2.7) and (3.2.8) are generated by a simple rotation: As already mentioned, in the representation (3.2.2) a regular matrix $Q \in \mathbb{R}^{r \times r}$ can always be inserted. It is easy to show that a simple rotation matrix $R \in \mathbb{R}^{r \times r}$ exists for which $FR = \begin{pmatrix} f & G \end{pmatrix}$ and $\Lambda R = \begin{pmatrix} \varphi & \Phi \end{pmatrix}$ applies. The proof of existence of such a rotation matrix R can be found in this chapter's appendix. When estimating factor f, the variables of Z should be taken into account simultaneously to avoid the second disadvantage. Such simultaneous estimation can be derived when putting (3.2.7) and (3.2.8) in one system:

$$\begin{pmatrix} X & y \end{pmatrix} = \begin{pmatrix} Z & f & G \end{pmatrix} \underbrace{\begin{pmatrix} 0 & \alpha \\ \varphi^\top & \gamma \\ \Phi^\top & 0 \end{pmatrix}}_{:=\Psi} + \begin{pmatrix} U & \varepsilon \end{pmatrix} \tag{3.2.9}$$

The key to the supervision of the factor model is the design of the matrix $\Psi \in \mathbb{R}^{(m+1+s)\times(n+1)}$ in (3.2.9): By positioning the zeros in a block-wise fashion, it ensures that both the equations with respect to X and the equations with respect to y take into account the vector f of factors. In contrast, the additional variables Z are only included in the y-equations, just as the other factors G are only considered in the X-equations. While Z is thus only fitted to y and G only to X, f is fitted to both, y and X. This constitutes an essential difference to the PCovR approach: The supervision is not induced by an information criterion from "outside" which assumes a trade-off between fitting to y and fitting to X for all factors. Rather, all time-varying effects that affect y and X in the same way are summarized in the factor f at the model level, while the other factors are freed from the mentioned trade-off between fitting to y and X. In addition, the fitting of factor f to y is "relieved" by adding further variables Z as part of the variance in y can also be explained by Z and does not necessarily have to be explained by factors from X alone. In this sense, such a model-based supervision "from within" generalizes the concept of supervised factor models since all factors estimated from the predictor space no longer necessarily have to be fitted to X and y at the same time. Instead, the factors that have to be fitted to both X and y are combined into one factor. Also, additional factors that only need to explain X can be included in the model. The scope of supervision is thus limited to what is necessary while the basic structure of a simple unsupervised factor model is retained to the extent possible.

3.3 The estimation process

3.3.1 Iteratively reweighted sequential least squares (IRSLS)

Breitung and Eickmeier (2016) use the sequential least squares (SLS) method in their paper: They are concerned with the simultaneous estimation of factors that reflect different kinds of influences on a target variable. They differentiate between so-called "global" factors that exert an influence in every region of the target variable and the factors designated as "regional" which only affect the target variable in certain regions. Correspondingly, the associated loading matrix is built up in blocks, similar to Ψ in (3.2.9). Instead of estimating both factor types consecutively (e.g., first global, then regional), starting values are selected for each of the factors which are then used to estimate a loading matrix with OLS. In a second step, that loading matrix is taken as the starting point and new factor values are determined by means of an OLS estimation. With these new factor values, new loadings

are now determined again, etc., until the residual sum of squares of the OLS estimation no longer improves and the algorithm has thus converged (cf. Breitung and Eickmeier (2014)). Considering the model here, the loadings φ and Φ as well as the parameter vector α and parameter γ can be estimated using (3.2.7) and (3.2.8), when the factors f and G are given. On the other hand, if all loadings and parameters and, consequently, Ψ is given, $Z\alpha$ is known and (3.2.9) can be rewritten using $\tilde{y} := y - Z\alpha$:

$$\begin{pmatrix} X & \tilde{y} \end{pmatrix} = \begin{pmatrix} f & G \end{pmatrix} \begin{pmatrix} \varphi^\top & \gamma \\ \Phi^\top & 0 \end{pmatrix} + \begin{pmatrix} U & \varepsilon \end{pmatrix} \tag{3.3.1}$$

Setting $\tilde{y}_{t+h} := y_{t+h} - \alpha^\top z_t$ one can rewrite (3.3.1) in long version:

$$\begin{pmatrix} x_{11} & \cdots & x_{n1} & \tilde{y}_{1+h} \\ \vdots & \ddots & \vdots & \vdots \\ x_{T1} & \cdots & x_{nT} & \tilde{y}_{T+h} \end{pmatrix} = \begin{pmatrix} f_1 & g_{11} & \cdots & g_{s1} \\ \vdots & \vdots & \ddots & \vdots \\ f_T & g_{1T} & \cdots & g_{sT} \end{pmatrix} \begin{pmatrix} \varphi_1 & \cdots & \varphi_n & \gamma \\ \phi_{11} & \cdots & \phi_{1n} & 0 \\ \vdots & \ddots & \vdots & \vdots \\ \phi_{s1} & \cdots & \phi_{sn} & 0 \end{pmatrix} + \begin{pmatrix} u_{11} & \cdots & u_{n1} & \varepsilon_{1+h} \\ \vdots & \ddots & \vdots & \vdots \\ u_{1T} & \cdots & u_{nT} & \varepsilon_{T+h} \end{pmatrix}$$

In this SLS approach, it should be noted that both loadings and factors are fitted to a target variable in each step. The target variable's error term is assumed to display variance homogeneity, i.e., the variance of the error term is identical in each region. As Breitung and Eickmeier note, this is done for purposes of simplification, but can be dropped without losing the consistency of the estimator. The authors refer to Wang (2010) who, in his iterative PC approach, allows different variances in the respective error terms in the different regions of the target variable. However, compared to Breitung and Eickmeier, he iterates his estimation in a different way: First, Wang chooses suitable starting values for the global factors and the associated loadings. Then, he subtracts the product of the associated global factors and loadings from the target variable for each region. He uses PCA to estimate the respective regional factors and loadings from the resulting residual term. In a second step, he subtracts the product of these regional factors and loadings from the target variable in all regions. This generates new values for the global factors and associated loadings from this residual term using PCA estimation. With these new values, Wang starts another iteration, repeating this process until the residual sum no longer improves at the global level. The lack of variance homogeneity is thus compensated by an iterative change between global and regional estimation. This chapter's model allows σ_U and σ_ε to differ. However, it is not possible to compensate for this difference in variance by iterative PC estimations as in Wang (2010): Even though the factor f can be interpreted "globally" due to the supervision with regard to X and y, the factor f itself may only consist of a linear combination w of the variables $x_1, \ldots, x_n$, just like the other factors G are also only linear combinations Ω of X:

$$\begin{pmatrix} f & G \end{pmatrix} = X \begin{pmatrix} w & \Omega \end{pmatrix} \tag{3.3.2}$$

However, this precludes a PC estimation of f because this would include y, which, in turn, would render the use of f for purpose of estimation for y_{t+h} impermissible. Instead, we will make use of Breitung and Eickmeier's SLS approach and insert (3.3.2) in (3.3.1):

$$\begin{pmatrix} X & \tilde{y} \end{pmatrix} = X \begin{pmatrix} w & \Omega \end{pmatrix} \begin{pmatrix} \varphi^\top & \gamma \\ \Phi^\top & 0 \end{pmatrix} + \begin{pmatrix} U & \varepsilon \end{pmatrix} \tag{3.3.3}$$

$$\Leftrightarrow \begin{pmatrix} x_1 \\ \vdots \\ x_n \\ \tilde{y} \end{pmatrix} = \left(\begin{pmatrix} \varphi & \Phi \\ \gamma & 0 \end{pmatrix} \otimes X \right) \begin{pmatrix} w \\ \omega_1 \\ \vdots \\ \omega_s \end{pmatrix} + \begin{pmatrix} u_1 \\ \vdots \\ u_n \\ \varepsilon \end{pmatrix} \tag{3.3.4}$$

The proof of equivalence between (3.3.3) and (3.3.4) can be found in the appendix. In order to achieve the auxiliary assumption of variance homogeneity of the error terms as seen in Breitung and Eickmeier (2014), it is sufficient to substitute X by $X^* = \rho X$ with $\rho = \frac{\sigma_\varepsilon}{\sigma_U}$ on both sides of (3.3.4) and to estimate (3.3.4) with SLS. According to lemma 7, such a substitution is equivalent to an estimation with weighted sequential least squares (WSLS):

Lemma 7. *Let X in (3.3.4) be substituted by $X^* := \rho X$ with $\rho \neq 0$ on both sides. If $\hat{\varphi}^*$, $\hat{\Phi}^*$, $\hat{\gamma}^*$, $\hat{w}^*$ and $\hat{\Omega}^*$ are the OLS-estimates of (3.3.4), it holds:*

(I) If $\hat{f} = X\hat{w}$ and $\hat{G} = X\hat{\Omega}$ are given, the OLS-estimates $\hat{\varphi}^*$, $\hat{\Phi}^*$ and $\hat{\gamma}^*$ are:

$$\begin{pmatrix} \hat{\varphi}^* & \hat{\Phi}^* \end{pmatrix}^\top = \left(\begin{pmatrix} \hat{f} & \hat{G} \end{pmatrix}^\top \begin{pmatrix} \hat{f} & \hat{G} \end{pmatrix} \right)^{-1} \begin{pmatrix} \hat{f} & \hat{G} \end{pmatrix}^\top X = \begin{pmatrix} \hat{\varphi} & \hat{\Phi} \end{pmatrix}^\top$$

$$\hat{\gamma}^* = \rho^{-1}\hat{\gamma}$$

(II) If $\hat{\varphi}$, $\hat{\Phi}$ and $\hat{\gamma}$ are given, the OLS-estimates $\hat{w}^*$ and $\hat{\Omega}^*$ are:

$$\begin{pmatrix} \hat{w}^* \\ \hat{\omega}_1^* \\ \vdots \\ \hat{\omega}_s^* \end{pmatrix} = \left(\left(\begin{pmatrix} \hat{\varphi} & \hat{\Phi} \\ \hat{\gamma} & 0 \end{pmatrix} \otimes X \right)^\top W_\rho \left(\begin{pmatrix} \hat{\varphi} & \hat{\Phi} \\ \hat{\gamma} & 0 \end{pmatrix} \otimes X \right) \right)^{-1} \left(\begin{pmatrix} \hat{\varphi} & \hat{\Phi} \\ \hat{\gamma} & 0 \end{pmatrix} \otimes X \right)^\top W_\rho \begin{pmatrix} x_1 \\ \vdots \\ x_n \\ y - Z\hat{\alpha} \end{pmatrix}$$

Here, $\hat{\varphi}$, $\hat{\Phi}$, $\hat{\gamma}$, $\hat{w}$ and $\hat{\Omega}$ are the OLS-estimate of (3.3.4), if X is not substituted and the matrix $W_\rho := \mathcal{I}_{(n+1)T} \begin{pmatrix} \underbrace{\rho^2 \dots \rho^2}_{=nT} & \underbrace{1 \dots 1}_{=T} \end{pmatrix}^\top \in \mathbb{R}^{(n+1)T \times (n+1)T}$ is the weight-matrix of a WLS-estimation using X.

The proof can be found in the appendix. σ_ε or σ_U must first be estimated. This estimation can also be done iteratively: In a first run, the model is assumed to exhibit variance homogeneity ($\sigma_\varepsilon = \sigma_U$), which is equivalent to $\rho := 1$. After arriving at a solution using SLS, the variances s_y^2 and s_X^2 of the respective residuals have to be determined. The residual variances generate an estimator for ρ and for X^* and another run can be started. After WSLS generates a solution for this run, the residual variances are determined again and ρ and X^*

is re-estimated before another run is initiated. Accordingly, the corresponding weights are determined before each WSLS run. This process is repeated until the solution of the respective runs no longer changes due to a given tolerance rate. Such an iterative re-weighting corresponds to the conventional iteratively reweighted least squares (IRLS) approach except that a SLS (instead of a LS) method is used in the optimization. To emphasize this difference, this estimation method is referred to as iteratively reweighted sequential least squares (IRSLS). This term also highlights that the method runs through two iterative or sequential loops: An "inner" loop in which the SLS-estimation of the linear combinations and other parameters is determined for a given ρ, and an "outer" loop that readjusts ρ based on the respective SLS solutions. The following subsection describes in detail how such an IRSLS estimation can be implemented.

3.3.2 The algorithmic implementation

Implementation of IRSLS

(I) Start with PCA-solution $\begin{pmatrix} \hat{w}_0^* & \hat{\Omega}_0^* \end{pmatrix}$ and set $\hat{\rho}_1 := 1$.

(II) For $k = 1, 2, \ldots$ set $X^* := \hat{\rho}_k X$ and do for $i = 0, 1, 2, \ldots$:

(i) Calculate $\begin{pmatrix} \hat{f}_{i+1}^* & \hat{G}_{i+1}^* \end{pmatrix} = X^* \begin{pmatrix} \hat{w}_i^* & \hat{\Omega}_i^* \end{pmatrix}$

(ii) Calculate $M_{i+1} := \left(\begin{pmatrix} \hat{\varphi}_{i+1}^* & \hat{\Phi}_{i+1}^* \\ \hat{\gamma}_{i+1}^* & 0 \end{pmatrix} \otimes X^* \right)$ with

$$\begin{pmatrix} \hat{\alpha}_{i+1} \\ \hat{\gamma}_{i+1}^* \end{pmatrix} = \left(\begin{pmatrix} Z & \hat{f}_{i+1}^* \end{pmatrix}^\top \begin{pmatrix} Z & \hat{f}_{i+1}^* \end{pmatrix} \right)^{-1} \begin{pmatrix} Z & \hat{f}_{i+1}^* \end{pmatrix}^\top y \qquad \text{and}$$

$$\begin{pmatrix} \hat{\varphi}_{i+1}^* & \hat{\Phi}_{i+1}^* \end{pmatrix}^\top = \left(\begin{pmatrix} \hat{f}_{i+1}^* & \hat{G}_{i+1}^* \end{pmatrix}^\top \begin{pmatrix} \hat{f}_{i+1}^* & \hat{G}_{i+1}^* \end{pmatrix} \right)^{-1} \begin{pmatrix} \hat{f}_{i+1}^* & \hat{G}_{i+1}^* \end{pmatrix}^\top X^*$$

(iii) Calculate $\begin{pmatrix} \hat{w}_{i+1}^* & \hat{\Omega}_{i+1}^* \end{pmatrix}$ using the values from (ii):

$$\begin{pmatrix} \hat{w}_{i+1}^* \\ \hat{\omega}_{1_{i+1}}^* \\ \vdots \\ \hat{\omega}_{s_{i+1}}^* \end{pmatrix} = \left(M_{i+1}^\top M_{i+1} \right)^{-1} M_{i+1}^\top \begin{pmatrix} x_1^* \\ \vdots \\ x_n^* \\ y - Z\hat{\alpha}_{i+1} \end{pmatrix}$$

(iv) If $\begin{pmatrix} \hat{w}_{i+1}^* & \hat{\Omega}_{i+1}^* \end{pmatrix}$ is different from $\begin{pmatrix} \hat{w}_i^* & \hat{\Omega}_i^* \end{pmatrix}$ due to a tolerance rate, go to (i).

(III) Calculate the residual variance $\hat{s}_X^2$ from (3.2.7) and $\hat{s}_y^2$ from (3.2.8) and set $\hat{\rho}_{k+1} := \frac{\hat{s}_y}{\hat{s}_X}$.

(IV) If $\hat{\rho}_{k+1}$ is different from $\hat{\rho}_k$ due to a tolerance rate, go to (II).

(V) Set the iterative solution as the final solution.

Some comments on this implementation: Lemma 7 suggests two alternatives for implement-

ing the IRSLS algorithm, either an OLS estimate with X^* or a GLS estimate with X and the weight matrix W_ρ. Due to the fact that the weight matrix has a very high dimension ($(n+1)T \times (n+1)T$) and storing it temporarily would, therefore, require a lot of storage space, it is numerically more stable to do OLS with X^*. In contrast, $\begin{pmatrix} \hat{\varphi} & \hat{\Phi} \\ \hat{\gamma} & 0 \end{pmatrix} \otimes X^*$ is of lower dimension ($(s+1)T \times (s+1)T$) as long as $s << n$ holds. Likewise, as a termination criterion, the minimization of the residual sum of squares did not turn out to be as numerically stable as might have been expected: Due to the adjustment of the weighting via ρ, the residual square sums are no longer strictly decreasing as opposed to the SLS algorithm. On the other hand, as is common with the IRLS algorithm, the intermediate results approach the optimal solution in each iteration step. Therefore, when the intermediate results no longer differ due to a certain tolerance rate, the algorithm can be suspended.

3.4 Simulations study

3.4.1 Data generating process

It has already been mentioned that the generalized supervised factor model as proposed here only takes those factors from X which are useful to forecast the target variable into consideration. Consequently, one advantage of this method should always become apparent when these relevant factors for y are "hidden" in X in the sense that these factors are only a subset of all factors that explain X. This simulation study reconstructs such situations by adding additional factors when generating X. These factors are not included in the generation of y and are, consequently, superfluous for the target variable. The other advantage of the proposed approach is the ability of integrating additional variables directly into the model, wherever PCA and PCovR needs a two-step estimation to do so, namely some regression of y by these variables and the estimated factors. Consequently, the data generating processes (DGPs) will be extended by a different type of additional variables:

(I) Variables Z, which are independent of X

(II) Lags of y

Especially lag variables are of importance for the usage in a macro-economic context. A short remark on these processes: By drawing factor p, the autocorrelation of the individual variables/factors becomes randomly more or less pronounced. However, the selection of the variance in the error terms ϵ_t occurs in such a way that the variance adds up to a total of 1 for all t (see Umbach (2020)). By introducing factor η, the signal-to-noise-ratio of $\tilde{X}$ varies, whereas $\sigma_u^2 = 1$ remains the same. In all scenarios of **DGP I**, $m = 3$ and $k = 3$ weaning that values y_t are generated from three variables z_t and three factors f_t from x_t.

DGP I

- **Step 1:** Choose the number L of irrelevant factors, number k of relevant factors and number m of additional variables. Choose also parameter vectors $\alpha \in \mathbb{R}^m$ and $\beta \in \mathbb{R}^k$, the dimensions n and T and the signal-to-noise-ratio η.

- **Step 2:** Generate matrix $Q = (q_{ij})_{i,j=1,\dots,n} \in \mathbb{R}^{n\times n}$ with $q_{ij} \sim \mathcal{U}[0,1]$ to compute loadings $\Lambda \in \mathbb{R}^{n\times k}$ and $\Phi \in \mathbb{R}^{n\times L}$ as the first $k+L$ eigenvectors of $Q^\top Q$. Draw anew each time $p \sim \mathcal{U}[0,1]$ to generate $z_1, \dots, z_{T+1} \in \mathbb{R}^m$ and factors $f_1, \dots, f_{T+1} \in \mathbb{R}^k$ and $g_1, \dots, g_{T+1} \in \mathbb{R}^L$ as simple AR(1)-processes:

  (i) $z_{j_1} \sim \mathcal{N}(0,1) \Rightarrow z_{j_t} = p z_{j_{t-1}} + \epsilon_t$ with $\epsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, 1-p^2)$ for $j = 1, \dots, m$

  (ii) $f_{j_1} \sim \mathcal{N}(0,1) \Rightarrow f_{j_t} = p f_{j_{t-1}} + \epsilon_t$ with $\epsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, 1-p^2)$ for $j = 1, \dots, k$

  (iii) $g_{j_1} \sim \mathcal{N}(0,1) \Rightarrow g_{j_t} = p g_{j_{t-1}} + \epsilon_t$ with $\epsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, 1-p^2)$ for $j = 1, \dots, L$

- **Step 3:** For $t = 1, \dots, T+1$ draw iid $u_t \in \mathbb{R}^n$ from a multivariate normal distribution with $\mu = \mathbf{0} \in \mathbb{R}^n$ and $\Sigma = \mathcal{I}_n$ and $\varepsilon_t \sim \mathcal{N}(0,1)$ to compute $y \in \mathbb{R}^t$ and data matrix $\tilde{X} = \begin{pmatrix} \tilde{x}_1 & \dots & \tilde{x}_n \end{pmatrix} \in \mathbb{R}^{(T+1)\times n}$ with

$$\tilde{x}_t = \eta \left(\Lambda f_t + \Phi g_t \right) + u_t$$
$$y_t = \alpha^\top z_t + \beta^\top f_t + \varepsilon_t$$

- **Step 4:** Standardize $\tilde{X}$ to get X.

- **Step 5:** Repeat **Step 2** to **Step 4** N-times to get Monte-Carlo-Simulations $\left(X^{(1)}, y^{(1)}\right), \dots, \left(X^{(N)}, y^{(N)}\right)$

The corresponding parameter vectors $\alpha = (1.2, 0.8, -0.4)^\top$ and $\beta = (-0.7, 1.3, -0.9)^\top$ are fixed as well whereas the number L of irrelevant factors g_t varies: The higher this number, the more irrelevant factors for y_t are used to generate x_t. In **DGP II** the first lag of y_t is used instead of z_t with $\delta = 0.8$. Considering the variances of the error-terms in both DGPs, $\sigma_\varepsilon^2 = 1$ is set to be constant, but different values for the variances of the error-terms in X are generated by standardizing $\tilde{X}$. Only three factors are ever relevant for the value of the y-variable. As a consequence, the first three factors, which correspond to the highest eigenvalues of the loading matrix, were always included in the estimation for all models. Since the loading matrix was generated randomly, it may well be the case that precisely those factors that are relevant for y are not selected due to this selection criterion. However, this often occurs in practice if only those factors are taken into account that are, as measured by the associated eigenvalues, of the greatest relevance for X. Our study will therefore also examine how the models react to such a misspecification.

DGP II

- **Step 1:** Choose the number L of irrelevant factors, number k of relevant factors and number m of additional variables. Choose also parameter vectors $\delta \in \mathbb{R}$ and $\beta \in \mathbb{R}^k$, the dimensions n and T and the signal-to-noise-ratio η.

- **Step 2:** Generate matrix $Q = (q_{ij})_{i,j=1,\dots,n} \in \mathbb{R}^{n \times n}$ with $q_{ij} \sim \mathcal{U}[0,1]$ to compute loadings $\Lambda \in \mathbb{R}^{n \times k}$ and $\Phi \in \mathbb{R}^{n \times L}$ as the first $k+L$ eigenvectors of $Q^\top Q$. Draw anew each time $p \sim \mathcal{U}[0,1]$ to generate factors $f_1, \dots, f_{T+1} \in \mathbb{R}^k$ and $g_1, \dots, g_{T+1} \in \mathbb{R}^L$ as simple AR(1)-processes:

  (i) $f_{j1} \sim \mathcal{N}(0,1) \Rightarrow f_{jt} = p f_{jt-1} + \epsilon_t$ with $\epsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, 1-p^2)$ for $j = 1, \dots, k$

  (ii) $g_{j1} \sim \mathcal{N}(0,1) \Rightarrow g_{jt} = p g_{jt-1} + \epsilon_t$ with $\epsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, 1-p^2)$ for $j = 1, \dots, L$

- **Step 3:** For $t = 1, \dots, T+1$ draw iid $u_t \in \mathbb{R}^n$ from a multivariate normal distribution with $\mu = \mathbf{0} \in \mathbb{R}^n$ and $\Sigma = \mathcal{I}_n$ and $\varepsilon_t \sim \mathcal{N}(0,1)$ to compute $y \in \mathbb{R}^t$ and data matrix $\tilde{X} = \begin{pmatrix} \tilde{x}_1 & \dots & \tilde{x}_n \end{pmatrix} \in \mathbb{R}^{(T+1) \times n}$ with

$$\tilde{x}_t = \eta \left(\Lambda f_t + \Phi g_t \right) + u_t$$
$$y_t = \delta y_{t-1} + \beta^\top f_t + \varepsilon_t$$

  For initial value draw $y_0 \sim \mathcal{N}(0,1)$.

- **Step 4:** Standardize $\tilde{X}$ to get X.

- **Step 5:** Repeat **Step 2** to **Step 4** N-times to get Monte-Carlo-Simulations $\left(X^{(1)}, y^{(1)}\right), \dots, \left(X^{(N)}, y^{(N)}\right)$

For each number $L = 0, 1, \dots, 9$ and $\eta = 0.25, 0.50, 0.75, 1$, Monte-Carlo-simulations were carried out with $N = 1,000$ runs. In each run, a sample of $T = 200$ observations and $n = 50$ different variables in X are generated as well as a sample of the target variable y. To measure the forecasting accuracy, x_{T+1} and y_{T+1} are generated in each case and each run. Therefore, the one-step-ahead mean squared error (MSE) or out-of-sample-MSE can be evaluated on the basis of $1,000$ different runs. The in-sample-R^2 is evaluated in each run and averaged in the end as well. While this is of less interest when focusing on forecasting, it may be helpful to interpret the results.

DGP I		Average in-sample-R^2						
		IRSLS	SLS	PCA	PCovR with $\theta =$			
η	L				0.2	0.4	0.6	0.8
0.25	0	0.405	0.482	0.385	0.502	0.505	0.505	0.505
	1	0.414	0.489	0.390	0.504	0.506	0.506	0.506
	2	0.420	0.494	0.393	0.505	0.507	0.507	0.507
	3	0.424	0.493	0.397	0.503	0.505	0.505	0.505
	4	0.432	0.499	0.403	0.509	0.510	0.510	0.510
	5	0.432	0.501	0.404	0.508	0.509	0.509	0.509
	6	0.440	0.507	0.411	0.510	0.511	0.511	0.511
	7	0.448	0.507	0.419	0.512	0.513	0.513	0.513
	8	0.446	0.509	0.417	0.510	0.511	0.511	0.511
	9	0.452	0.513	0.422	0.513	0.514	0.514	0.514
0.50	0	0.462	0.550	0.403	0.551	0.552	0.551	0.551
	1	0.473	0.548	0.411	0.552	0.553	0.553	0.552
	2	0.480	0.555	0.417	0.553	0.554	0.554	0.554
	3	0.483	0.555	0.425	0.553	0.554	0.553	0.553
	4	0.491	0.557	0.432	0.558	0.558	0.558	0.558
	5	0.491	0.557	0.434	0.556	0.557	0.557	0.556
	6	0.501	0.561	0.444	0.559	0.560	0.559	0.559
	7	0.502	0.563	0.449	0.560	0.560	0.560	0.560
	8	0.504	0.565	0.453	0.560	0.560	0.560	0.560
	9	0.507	0.566	0.460	0.563	0.563	0.563	0.563
0.75	0	0.564	0.617	0.457	0.608	0.607	0.606	0.605
	1	0.565	0.616	0.469	0.610	0.608	0.607	0.607
	2	0.567	0.617	0.475	0.610	0.609	0.609	0.608
	3	0.569	0.617	0.484	0.611	0.610	0.609	0.608
	4	0.574	0.623	0.493	0.615	0.614	0.613	0.613
	5	0.570	0.620	0.495	0.613	0.612	0.612	0.611
	6	0.574	0.617	0.506	0.616	0.615	0.615	0.614
	7	0.576	0.620	0.511	0.616	0.616	0.615	0.615
	8	0.577	0.621	0.516	0.617	0.617	0.616	0.616
	9	0.580	0.624	0.521	0.620	0.619	0.619	0.618
1.00	0	0.642	0.672	0.551	0.660	0.656	0.654	0.653
	1	0.643	0.669	0.557	0.661	0.658	0.656	0.655
	2	0.642	0.674	0.560	0.662	0.659	0.657	0.656
	3	0.639	0.671	0.565	0.662	0.660	0.658	0.657
	4	0.642	0.674	0.573	0.666	0.663	0.662	0.661
	5	0.633	0.670	0.572	0.664	0.662	0.661	0.660
	6	0.638	0.668	0.580	0.667	0.665	0.663	0.663
	7	0.638	0.668	0.583	0.667	0.665	0.664	0.663
	8	0.639	0.672	0.588	0.668	0.666	0.665	0.665
	9	0.640	0.672	0.591	0.668	0.666	0.665	0.665

Table 3.1: Average in-sample-R^2 of $N = 1,000$ draws given η and L

DGP II		Average in-sample-R^2						
		IRSLS	SLS	PCA	PCovR with $\theta =$			
η	L				0.2	0.4	0.6	0.8
0.25	0	0.842	0.858	0.837	0.850	0.850	0.850	0.850
	1	0.841	0.854	0.835	0.847	0.847	0.847	0.847
	2	0.842	0.856	0.836	0.848	0.848	0.848	0.848
	3	0.847	0.860	0.840	0.851	0.851	0.851	0.851
	4	0.846	0.858	0.839	0.850	0.850	0.850	0.850
	5	0.846	0.858	0.840	0.850	0.850	0.850	0.850
	6	0.851	0.863	0.843	0.853	0.854	0.853	0.853
	7	0.855	0.866	0.848	0.857	0.857	0.857	0.857
	8	0.854	0.865	0.847	0.856	0.857	0.857	0.857
	9	0.851	0.862	0.844	0.853	0.854	0.854	0.854
0.50	0	0.856	0.875	0.842	0.860	0.860	0.860	0.860
	1	0.854	0.871	0.840	0.858	0.857	0.857	0.857
	2	0.856	0.873	0.841	0.858	0.858	0.858	0.858
	3	0.861	0.877	0.846	0.862	0.862	0.862	0.862
	4	0.860	0.875	0.846	0.861	0.861	0.861	0.861
	5	0.861	0.873	0.847	0.861	0.861	0.861	0.861
	6	0.864	0.877	0.851	0.864	0.864	0.864	0.864
	7	0.868	0.881	0.855	0.868	0.868	0.868	0.868
	8	0.867	0.880	0.856	0.868	0.868	0.868	0.867
	9	0.865	0.877	0.853	0.865	0.865	0.865	0.865
0.75	0	0.880	0.892	0.854	0.875	0.874	0.873	0.873
	1	0.877	0.890	0.853	0.872	0.871	0.871	0.871
	2	0.878	0.889	0.856	0.873	0.872	0.872	0.872
	3	0.882	0.893	0.860	0.877	0.876	0.876	0.875
	4	0.880	0.891	0.860	0.876	0.875	0.875	0.875
	5	0.879	0.890	0.861	0.876	0.875	0.875	0.875
	6	0.883	0.892	0.865	0.879	0.879	0.878	0.878
	7	0.885	0.895	0.870	0.883	0.883	0.882	0.882
	8	0.885	0.894	0.870	0.882	0.882	0.882	0.881
	9	0.883	0.891	0.868	0.880	0.880	0.880	0.880
1.00	0	0.900	0.907	0.877	0.890	0.888	0.887	0.886
	1	0.897	0.904	0.875	0.888	0.886	0.885	0.885
	2	0.898	0.904	0.877	0.889	0.887	0.886	0.886
	3	0.900	0.906	0.880	0.892	0.890	0.890	0.889
	4	0.900	0.906	0.882	0.893	0.891	0.891	0.891
	5	0.897	0.903	0.881	0.891	0.890	0.889	0.889
	6	0.898	0.906	0.884	0.894	0.893	0.892	0.892
	7	0.901	0.907	0.888	0.897	0.896	0.895	0.895
	8	0.901	0.907	0.888	0.897	0.896	0.895	0.895
	9	0.898	0.905	0.886	0.895	0.894	0.894	0.894

Table 3.2: Average in-sample-R^2 of $N = 1,000$ draws given η and L

DGP I		Out-of-sample-MSE						
		IRSLS	SLS	PCA	PCovR with $\theta =$			
η	L				0.2	0.4	0.6	0.8
	1	0.684	0.740	0.691	0.813	0.828	0.833	0.835
	2	0.665	0.724	0.669	0.818	0.840	0.847	0.850
	3	0.655	0.707	0.672	0.776	0.794	0.800	0.802
0.25	4	0.674	0.762	0.692	0.837	0.854	0.860	0.862
	5	0.703	0.781	0.710	0.854	0.869	0.874	0.876
	6	0.746	0.826	0.754	0.877	0.894	0.900	0.903
	7	0.680	0.745	0.690	0.800	0.814	0.819	0.821
	8	0.640	0.717	0.646	0.768	0.783	0.788	0.790
	9	0.687	0.752	0.706	0.798	0.808	0.811	0.813
	0	0.626	0.636	0.638	0.695	0.715	0.721	0.724
	1	0.657	0.687	0.669	0.752	0.769	0.774	0.777
	2	0.633	0.669	0.645	0.756	0.779	0.787	0.791
	3	0.622	0.640	0.642	0.710	0.730	0.736	0.739
0.50	4	0.644	0.695	0.665	0.771	0.790	0.796	0.799
	5	0.670	0.711	0.681	0.781	0.798	0.803	0.806
	6	0.706	0.722	0.714	0.789	0.807	0.813	0.816
	7	0.693	0.732	0.723	0.772	0.784	0.788	0.790
	8	0.612	0.665	0.618	0.696	0.710	0.715	0.718
	9	0.642	0.663	0.660	0.719	0.731	0.736	0.738
	0	0.549	0.540	0.590	0.601	0.623	0.631	0.635
	1	0.587	0.585	0.610	0.666	0.687	0.694	0.698
	2	0.561	0.585	0.570	0.671	0.698	0.707	0.711
	3	0.552	0.550	0.578	0.625	0.647	0.654	0.658
0.75	4	0.582	0.607	0.599	0.683	0.703	0.710	0.713
	5	0.603	0.606	0.615	0.686	0.705	0.712	0.716
	6	0.626	0.624	0.635	0.678	0.698	0.704	0.708
	7	0.632	0.643	0.649	0.670	0.683	0.688	0.690
	8	0.553	0.568	0.560	0.610	0.626	0.631	0.634
	9	0.584	0.580	0.583	0.629	0.644	0.649	0.651
	0	0.469	0.457	0.485	0.513	0.539	0.548	0.553
	1	0.494	0.497	0.511	0.584	0.611	0.621	0.626
	2	0.487	0.499	0.479	0.584	0.614	0.625	0.631
	3	0.472	0.472	0.482	0.545	0.570	0.579	0.584
1.00	4	0.513	0.524	0.508	0.594	0.616	0.624	0.628
	5	0.519	0.505	0.522	0.596	0.619	0.628	0.632
	6	0.521	0.505	0.501	0.566	0.585	0.592	0.596
	7	0.554	0.547	0.552	0.584	0.600	0.605	0.608
	8	0.517	0.495	0.505	0.549	0.566	0.572	0.575
	9	0.524	0.526	0.502	0.573	0.590	0.596	0.599

Table 3.3: Out-of-sample-MSE based on $N = 1,000$ draws given η and L

DGP II		Out-of-sample-MSE						
		IRSLS	SLS	PCA	PCovR with $\theta =$			
η	L				0.2	0.4	0.6	0.8
0.25	0	0.184	0.187	0.185	0.202	0.204	0.205	0.205
	1	0.179	0.187	0.179	0.201	0.203	0.203	0.204
	2	0.187	0.197	0.188	0.210	0.212	0.213	0.213
	3	0.180	0.195	0.182	0.201	0.203	0.204	0.204
	4	0.195	0.211	0.197	0.229	0.231	0.232	0.232
	5	0.178	0.190	0.178	0.197	0.199	0.200	0.200
	6	0.169	0.185	0.171	0.188	0.190	0.190	0.191
	7	0.181	0.193	0.185	0.205	0.206	0.207	0.207
	8	0.180	0.204	0.186	0.198	0.200	0.201	0.201
	9	0.160	0.173	0.167	0.183	0.183	0.184	0.184
0.50	0	0.177	0.180	0.182	0.195	0.198	0.199	0.199
	1	0.173	0.177	0.175	0.193	0.196	0.197	0.197
	2	0.178	0.180	0.181	0.201	0.204	0.205	0.206
	3	0.175	0.183	0.178	0.193	0.196	0.197	0.197
	4	0.190	0.199	0.194	0.222	0.225	0.226	0.226
	5	0.170	0.177	0.170	0.188	0.191	0.191	0.192
	6	0.162	0.173	0.164	0.176	0.178	0.179	0.179
	7	0.175	0.183	0.179	0.194	0.196	0.196	0.196
	8	0.174	0.186	0.181	0.194	0.195	0.196	0.196
	9	0.189	0.202	0.191	0.222	0.225	0.227	0.227
0.75	0	0.161	0.161	0.166	0.180	0.184	0.186	0.187
	1	0.156	0.155	0.164	0.179	0.183	0.184	0.185
	2	0.161	0.157	0.162	0.183	0.188	0.189	0.190
	3	0.161	0.161	0.165	0.179	0.183	0.184	0.184
	4	0.175	0.178	0.181	0.204	0.208	0.209	0.210
	5	0.156	0.153	0.156	0.172	0.175	0.176	0.177
	6	0.147	0.148	0.148	0.159	0.161	0.162	0.162
	7	0.163	0.169	0.161	0.176	0.178	0.178	0.179
	8	0.163	0.168	0.168	0.179	0.182	0.182	0.183
	9	0.175	0.176	0.174	0.204	0.208	0.209	0.210
1.00	0	0.143	0.140	0.140	0.163	0.170	0.172	0.173
	1	0.138	0.133	0.139	0.161	0.167	0.169	0.170
	2	0.139	0.139	0.137	0.164	0.170	0.172	0.173
	3	0.145	0.147	0.143	0.161	0.166	0.168	0.169
	4	0.139	0.139	0.134	0.157	0.162	0.163	0.164
	5	0.141	0.140	0.139	0.154	0.158	0.159	0.160
	6	0.131	0.125	0.126	0.140	0.143	0.145	0.145
	7	0.143	0.144	0.146	0.160	0.163	0.164	0.165
	8	0.146	0.150	0.147	0.160	0.163	0.164	0.165
	9	0.157	0.150	0.150	0.181	0.185	0.187	0.188

Table 3.4: Out-of-sample-MSE based on $N = 1,000$ draws given η and L

3.4.2 Results

The average in-sample-R^2 values are summarized in Tables 3.1 and 3.2. The value of R^2 clearly increases on average when the signal-to-noise ratio increases. Increasing the number of additional factors has the same effect as this also makes the signal-to-noise ratio larger. This was to be expected since a stronger signal in X allows for a better estimation of the factors. Between the models, the PCovR models and the simple SLS method perform better than IRSLS and PCA with regard to the R^2.
This changes when considering the results of the out-of-sample-MSE, which are summarized in Tables 3.3 and 3.4. Here, IRSLS, SLS and also PCA perform significantly better than PCovR. If, in addition, the factor η is small (0.25 and 0.50), the IRSLS model proves to be the strongest model compared to not only SLS but also PCA. However, this prevalence is not as pronounced for high values of η (0.75 and 1.00) as there are some cases where PCA in particular produces better results. It can thus be concluded that a high signal-to-noise ratio favors the classic PCA approach whereas the IRSLS approach works better than other estimation approaches with a low signal-to-noise ratio.

3.5 Conclusion

The present chapter examines a factor model with a form of supervision that can be understood as a generalization of supervised factor models: All factors that are relevant for the target variable are combined into one factor and estimated simultaneously using X and y and additional variables z. The two-stage approach from classic factor analysis is thus avoided. In addition to the derivation of the model, the IRSLS approach, which is to be understood as a combination of the classic IRLS estimation and the SLS estimation, presents an estimation method for this model and also the algorithmic implementation. A simulation study also shows situations in which this model approach generates better out-of-sample results than classic methods. In particular, the results of the simulation study illustrate that the generalized form of supervision outperforms the PCovR approach and is, therefore, preferable.

Chapter 4

Algorithmic pre-screening for birth defects using medical invoice data

4.1 Introduction

The provision of midwifery care is an integral part of a society's healthcare system. Despite the ongoing medical progress and the new technical possibilities of diagnostics, in rare cases, severe human error can occur in this field that causes the newborn child to suffer health damages and permanent impairments from the birth process. Such damages, which cannot be attributed to genetic or physiological malformations in the womb, but to misconduct on the part of the person providing obstetrics, are called birth defects in the narrower sense. In a welfare state, it goes without saying that those affected are adequately cared for or compensated in such cases and that, in addition to damage prevention measures, provisions are also made to settle possible damages. The different payers in the healthcare system pursue this mandate. While better healthcare reduces the frequency of damage in the field of obstetrics, the simultaneous increase in life expectancy of the children affected and the associated longer duration of expenditures for therapy, loss of earnings and compensation for pain and suffering, has brought about a substantial rise in costs when damage does occur. For example, the Association of German Insurers (GDV) states that for the years 2003 to 2020 alone, the average claims expenditure for severe birth defects more than doubled from EUR 1.5 million to EUR 3.7 million per damaged child[1]. Detecting birth defects and correctly classifying them is, therefore, not only in the interest of those involved as well as the general public so that further preventive measures can be developed, but also affects the economic interests of all payers in the healthcare system.

The aim of this work is to discuss the possibility of an algorithmic pre-screening for birth defects using medical invoice data. So far, a preliminary check in the sense of a preselection based on these data is carried out manually by a human decision-maker. However, there are some works in the literature that address the possibility of an algorithmic replication of human decisions, above all, but not only, in economic business processes. In the field of

[1] https://www.gdv.de/de/themen/news/behandlungsfehler-warum-die-schicksalsschlaege-immer-hoehere-kosten-verursachen-84998

machine learning in particular, there are numerous papers that even propagate an improved decision-making process through the use of algorithmic structures, or at least suggest that such structures can map the human decision-making calculus sufficiently well and can thus contribute to gains in efficiency. One example is the work of Harding and Vasconcelos (2022). It provides evidence that bank managers' decisions on credit risk assessment issues can be replicated using machine learning methods. The present work follows a very similar approach but pays special attention to two aspects: On one hand, the special structure of the invoice data analyzed here entails some challenges and peculiarities. On the other hand, as explained in the introduction, birth defects are rare events. This aspect is of central importance to the modeling. Before discussing these aspects, we focus contextualizing the term 'data mining' in the existing literature and describing it in more detail with the help of a process model in the second section. The subsequent two sections are based on the phases of this process model and refer back to the two aspects already mentioned: In particular, after the specification of the research question in the third section, strategies and modifications of algorithms, which can favor the prognosis of rare events, are shown in the fourth section. Furthermore, the various possible algorithms are applied to the correspondingly prepared data and a performance comparison is carried out. The fifth and last section closes with a brief summary and assessment of the results.

4.2 Data Mining as generic cyclical model

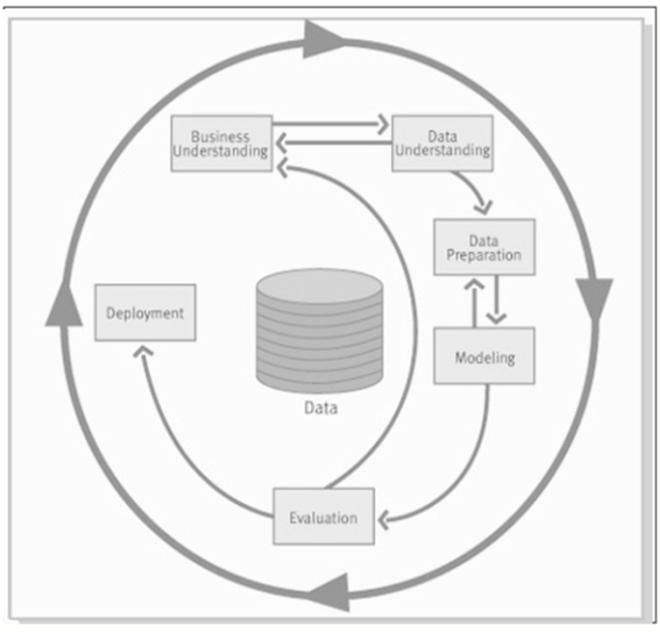


Figure 4.2.1: CRISP-DM, figure taken from Chapman et al. (2000).

Data mining processes are the subject of numerous scientific works. Due to the growing importance of data, they have also increasingly come into focus of companies' economic interests. This is not surprising provided that data mining processes are meant to describe the algorithmic "mining" of information from data. In a second step, this information serves as the basis for economic decision-making. With the model of the cross-industry standard process for data mining (abbreviated 'CRISP-DM'), a standardized method for data mining in an economic context was developed in the late 1990s. It is shown in Figure 4.2.1. This draft model is the starting point for many further developments in the field of data mining. However, due to the fact that it is a highly idealized model, which one has to deviate from in practice for various reasons, some modifications of CRISP-DM exist in the literature. In this work, we refer to the modification of Reuß and Zwiesler (2006) who embed the data mining process in a generic cyclical model and apply it to a industry case study. They convert the various work steps of CRISIP-DM into four working phases:

(I) Specification of the research question
   (a) Specification of the goals and framework conditions
   (b) Analysis of the data basis

(II) Execution of the data mining analysis
   (a) Pre-processing of the data
   (b) Model-based Analysis of the data
   (c) Evaluation of the results

(III) Conversion of the findings into actionable steps (measures)
   (a) Derivation of appropriate actionable steps
   (b) Implementation of appropriate actionable steps

(IV) Measurement and evaluation of the results

Phase (III) includes interventions in the operative businesses of the different payers in the healthcare system and phase (IV) concerns the measurement and evaluation of these interventions. Therefore, this work can only discuss phases (I) and (II) because they can run independent of the different business processes as well as different types of payers in the healthcare system. In addition, phase (I) deals solely with the data basis and its interaction with the underlying business question while phase (II) addresses further aspects such as model selection and the associated rare event problem.
These first two phases of the circular model by Reuß and Zwiesler (2006) are adapted to the problem at hand and combined to form a separate process model:

(1) Defining the research question
    (a) Defining the research question based on the business process
    (b) Defining the research question based on the database
(2) Data mining analysis
    (a) Data preparation
    (b) Modeling
    (c) Evaluation

The second phase of this modification is more strongly oriented towards CRISP-DM again. In contrast, the first phase takes up the suggestions of Reuß and Zwiesler (2006) (i) to narrow down the question, (ii) to include the definition of the goals of a data mining analysis as a separate step in the process and (iii) to link it to the initial analysis of the database. Since the underlying business process in the detection of birth defects and the use of invoice data from the medical field are associated with some challenges and restrictions, such an independent consideration is worthwhile here as well.

4.3 Defining the research question

4.3.1 Defining the research question based on the business process

The economic importance of pre-screening for birth defects has already been explained in the introduction. This section discusses the concrete implementation of such a screening in the business process in more detail. Said implementation is shown schematically in Figure 4.3.2. In a first step, if an employee of a payer in the healthcare system discovers anomalies in the invoice data of a newborn or is persuaded to do so by external sources (e.g., by a call from the parents), she or he notes that the newborn child in question should be examined further for possible birth defects (*Notice*). The subsequent steps, namely contacting the parents (*Inform*) and initiating an official examination for birth defects (*Investigate*) do not necessarily have to follow the first step. They are also not documented in the data on which this work is based. This results in the decisive limitation for the objective of algorithmic pre-screening for birth defects: Such a screening cannot be carried out to determine whether birth defects really exist or not. Instead, it determines whether or not a human decision-maker would recommend further testing for birth defects in a given case. Due to this limitation resulting from the database, the actual goal is to replicate human decision-making in a given case as precisely as possible rather than improving it.

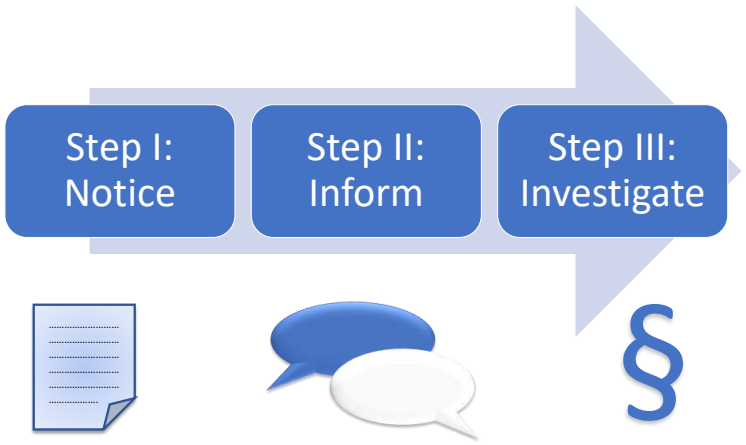


Figure 4.3.2: Detection of birth defects

4.3.2 Defining the research question based on the database

We analyze invoices of over 30,000 children born between October 2019 and December 2020. During this period, there were fewer than 750 children who received a note and had billing data at all, but more than 100 children with a note but no bills. The assumption here is that the human decision-maker did not base his decision on invoice data but on external sources. However, the data analyzed in this work does not conclusively prove this. Limiting the analysis to an informative database, i.e., from the 30,000 to the 25,000 children for whom any billing data are available, demonstrates that the existence of a note regarding birth defects is a highly rare occurrence: Less than 3% of the children were issued such a note. Said billing data includes the following information per bill for each child:

- Amount to pay [EUR]
- Days in Hospital [0,1,2...]
- Treatment [Different categories]
- Diagnoses [ICD-10-Codes]
- Quality of these diagnoses [sure/unsure]
- Months between birth and treatment [0,1,2...]
- Months between birth and payment [0,1,2...]

Only the payment amount and the temporal variables are scaled numerically; the diagnoses, types of treatment and the variable that reflects the quality of the diagnosis are categorical

variables. While an invoice is always assigned to only one type of treatment, it can have multiple diagnosis codes. Regarding the temporal variables, the number of months between the child's birth and the payment of the bill contains a key piece of information: It specifies the point in time - relative to the birth of the respective child - at which a payer in the healthcare system finds out about the treatment for which it has to pay. This is more informative than the moment of treatment itself as solely the facts known to the employee of the respective payer in the healthcare system when making the decision determine the outcome. Therefore, the number of months between payment and birth rather than the number of months between treatment and birth is considered for the purposes of this analysis. Just like the billing information, the birth defect notes also document the number of months between the note and the birth for each child. Figure 4.3.3 shows the number of entries made until the respective month after the birth of the child in question relative to all entries: It is noteworthy that two-thirds of all entries are made within the first quarter after the birth of the respective child. Almost 95% of all entries are made within the first six months after the birth of the respective child. In other words, in the event of a child receiving a birth defect note, the probability that this occurs within the first six months of the child's life is extremely high. This would justify a simple cross-sectional as opposed to a panel data structure: For each child, the billing data of all bills that were incurred in the first six months after birth could be summarized. Based on these aggregated values, a decision could then be made as to whether or not a note should be made regarding birth defects. Taking this in account, the number of invoices which have to be analyzed is still over 140,000.

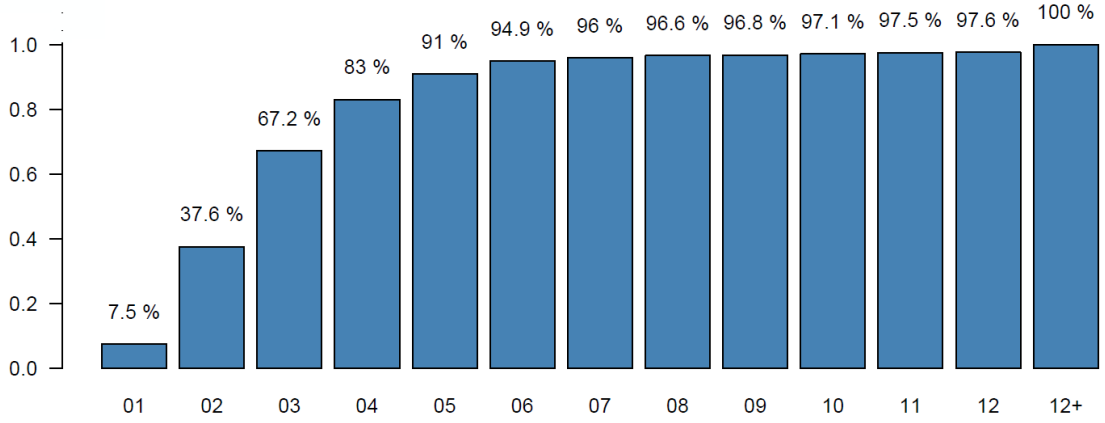


Figure 4.3.3: Ratio of detections in a given month after birth relative to all detections.

Before this invoice data information is processed further in the course of the subsequent data mining analysis (phase (2)), something fundamental about this database must be noted: Concluding from such invoice data whether there may be birth defects assumes that the

bills provide information on the child's health. This assumption can be viewed critically: On one hand, it must be clarified whether the number of days spent in the hospital, the respective treatment group and, in particular, the diagnoses that are made really adequately reflect the child's medical condition. On the other hand, and more generally speaking, the quality of the data, i.e., whether the specified values really correspond to the actual values, has to be determined. At least with regard to the diagnoses, arguments that substantiate the basic assumption of a connection between the children's billing and health data can be found for both sides. There already is some literature that successfully draws conclusions regarding a diagnosed person's medical condition through analyzing diagnoses in the form of ICD-10-Codes (e.g., Baumel et al. (2018), Sailer et al. (2015) or Burkul et al. (2020)): Burkul et al. (2020), for example, performed a decision tree-based analysis of ICD-10-Codes to further investigate the relationship between mortality and the medical equipment used that poses a risk of infection. Their goal is to use data mining analysis to make more precise predictions as to which devices pose a higher risk of infection, measured with regard to the disease the patient has been diagnosed with. They also use machine learning methods and reconstruct the patient's medical condition from the ICD-10-Codes. With regard to the data quality, the information about the degree of certainty with which a diagnosis was made is another piece of meaningful information. It can decisively relativize or correct the informative value of the diagnosis. Nonetheless, the question whether the computational data can really measure what it is intended to remains a valid point of discussion, even though it should be noted (once more) that the goal has been limited to the replication, and not improvement, of human decision-making. Therefore, the data are meaningful as long as they are able to accurately replicate the human decision-making process, regardless of how precise the human decision in relation to issuing birth defect notes actually is.

4.4 Data mining analysis

4.4.1 Data preparation

While the analysis of the data in the first working phase only served to define the research question more precisely, the pre-processing of the data in the second working phase focuses on the selection of the specific data mining method: The method to be chosen has to be adjusted to the question as well as to the existing database (cf. Reuß and Zwiesler (2006)). As already mentioned, the diagnoses documented on the invoices in the form of ICD-10 codes are particularly suitable for assessing the state of health of the child in question. However, the ICD-10 codes are categorical variables with a large number of characteristics and also feature incorrect entries that arose from the digitization of the handwritten diagnoses by doctors. This, in turn, must be taken into account when choosing the method: Some algorithms are certainly capable of processing categorical variables. One could deal with this problem by using dummy variables, an approach that is very common in these cases

(cf. Hastie et al. (2001)). However, the number of diagnoses differs from one invoice to the next. The difference in diagnoses documented could be informative with regard to the child's medical condition: A healthy child is likely to have fewer diagnoses documented than a sick child. The invoice data should, therefore, be transformed in such a way that it reflects this information. Coding using dummy variables can only achieve this to a limited extent. Instead, a different transformation was carried out: To begin with, since the first digit of the code corresponds to a specific disease category (all main disease categories according to ICD-10 can be found in the appendix), the individual ICD-10-codes were summarized in main diagnoses. After summarizing these main diagnoses, the number and the amounts of all invoices for each child that were known to the respective payer within the first six months after the birth of that child and carried the respective main diagnosis were added up. Furthermore, the average quality per main diagnosis were determined by averaging the value of a dummy variable 1/0 coding "sure"/"unsure" from all corresponding invoices. Instead of a simple dummy variable, the result generated was the total medical cost associated with a main diagnosis per child within the first six months after birth. Seven different main diagnoses occur on the invoices with a relative frequency of $> 3\%$:

- **A/B**: Certain infectious and parasitic diseases
- **J**: Diseases of the respiratory system
- **M**: Diseases of the musculoskeletal system and connective tissue
- **P**: Certain conditions originating in the perinatal period
- **Q**: Congenital malformations, deformations and chromosomal abnormalities
- **R**: Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
- **Z**: Factors influencing health status and contact with health services

The remaining diagnoses were summarized under "**Rare**" (see Figure 4.4.4). In doing so, we generate 24 numerical variables (7 main diagnoses and **Rare** with 3 variables each) from the database per child. However, in both groups of children (with/without note, coded by 1/0), there are still children with no registered diagnoses at all (see Table 4.1) but other cases of an accumulation of invoices (and therefore registered diagnoses). As a consequence, the total number of invoices is still higher than the total number of children (cf. 4.4.4).
Of these main diagnoses, category **P** is of particular interest as this code summarizes all pathological conditions that record damage to the newborn that was caused during birth.

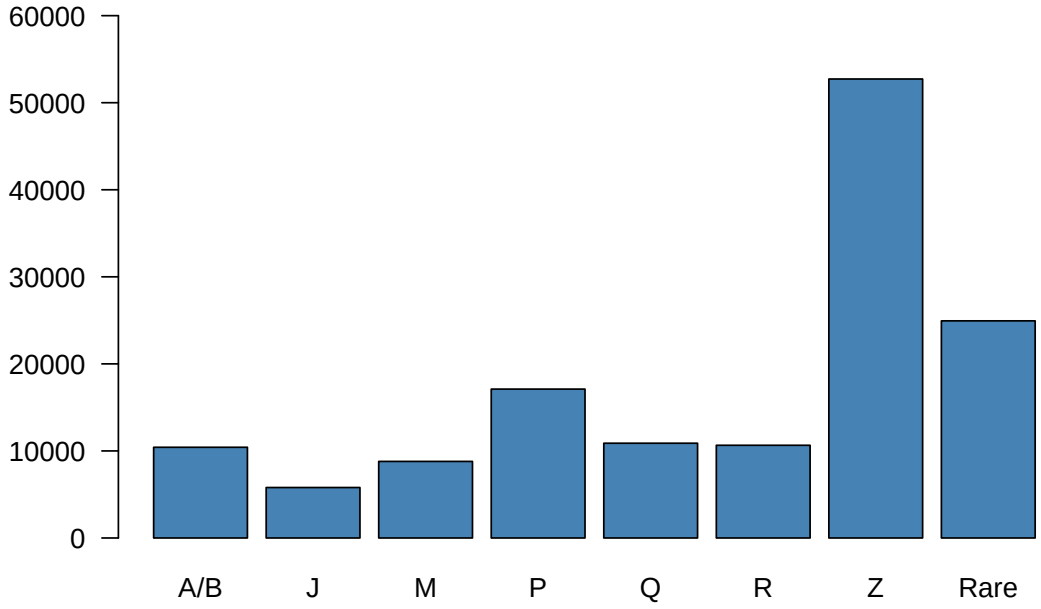


Figure 4.4.4: Number of invoices per main diagnoses.

main diagnose	with note [1]	without note [0]
A/B	25.23%	18.03%
J	15.51%	11.18%
M	11.34%	11.99%
P	73.03%	14.99%
Q	30.32%	17.84%
R	32.18%	19.61%
Z	76.04%	55.66%
Rare	55.79%	33.48%
None	19.79%	35.33%

Table 4.1: Percentage of children (grouped by notes) with at least one invoice on which the corresponding main diagnose is registered.

Accordingly, these could be birth defects in the narrower sense. Diagnosis **Z**, similar to diagnosis **R**, is a quite unspecific term for all pathological conditions that lead to a strain on or utilization of the health care system without the origin of the disease being clear. It is precisely this diagnosis that is made most frequently in both groups (cf. 4.4.4). However, a clear difference becomes apparent here as well: While slightly more than every other child without a note has been diagnosed with **Z**, the diagnosis occurred in more than 75% of the

children with a note. In addition, half of children with potential birth defects are diagnosed with a rare diagnosis, compared to only one-third of children without birth defects. Also, the proportion of children without any registered diagnosis in the group of children without a note is almost twice as large as in those who received one. Accordingly, more diagnoses, especially those falling in the **P**, **Z** and "**Rare**" categories, are documented for children with possible birth defects than for children without a corresponding note. Examining the 24 diagnosis-specific variables exploratively confirms this finding: For each main diagnosis X, Figures 4.4.6 and 4.4.5 below and Figures A.3.1 to A.3.6 in the appendix show the box plots of the total invoice amount $Sum.X$, the number $N.X$ of associated invoices and the average measured quality $n.X$ of the respective diagnosis. All boxplots were created separately for both groups (1 = blue, 0 = white) based on the number of children who had at least one invoice with the respective main diagnosis (i.e., those children whose share is registered in Table 4.1). For the sake of clarity, outliers, i.e., values that exceed 1.5 times the upper quartile, were excluded. However, the proportion of these outliers is negligible at less than 0.5% per diagnosis. Whereas for diagnosis **M** no differences between the two groups are observable, the boxplots of the invoice amounts for diagnose **P** is clearly in the higher numerical range for group 1 while the boxplots are almost completely at 0 for group 0.

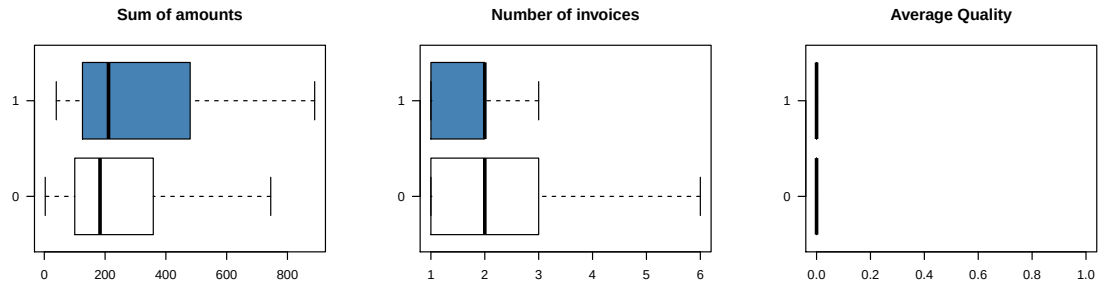


Figure 4.4.5: Boxplots of main diagnose **M**.

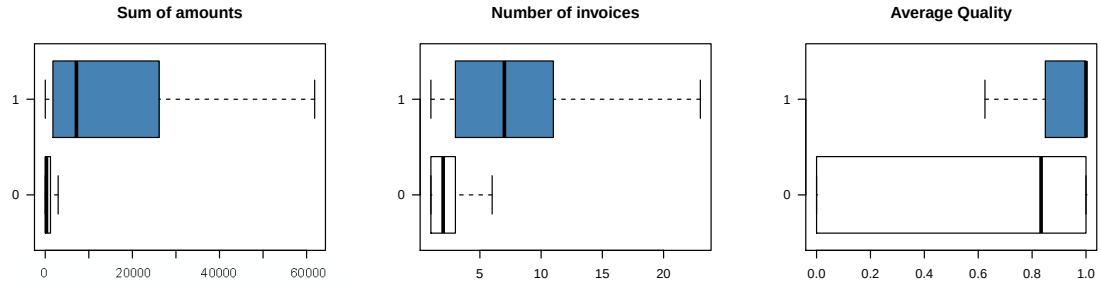


Figure 4.4.6: Boxplots of main diagnose **P**.

The same observation can be made for the boxplots on the average diagnosis quality. What is more, the number of bills displays the same finding. Not only have children with a birth defect been diagnosed more often than children without one, the associated invoice amounts and the number of invoices per main diagnosis are also higher on average. This is also true for the average quality of the diagnosis, i.e., the doctors are more certain of their diagnosis when treating children with a note than when they treat those without one. What is shown here for diagnose **P** can also be observed for diagnoses **Z** and **"Rare"** (see appendix).
The same was done with the types of treatment. Due to the manageable number of characteristics, a total of nine categories, a pre-selection was not necessary:

- **OT**: Outpatient treatment
- **MR**: Medical remedies
- **HP**: Homeopathic practitioner
- **MA**: Medical aids
- **RS**: Rehabilitation stay
- **MD**: Medications
- **CS**: Care services
- **IT**: Inpatient treatment
- **DT**: Dental treatment

treatment	with note [1]	without note [0]
OT	66.44%	68.57%
MR	12.27%	5.12%
HP	20.37%	20.73%
MA	14.70%	6.93%
RS	0%	0%
MD	68.63%	67.60%
CS	3.82%	0.28%
IT	75.58%	18.58%
DT	0.93%	1.03%
None	16.55%	23.52%

Table 4.2: Percentage of children (grouped by notes) with at least one invoice on which the corresponding treatment is registered.

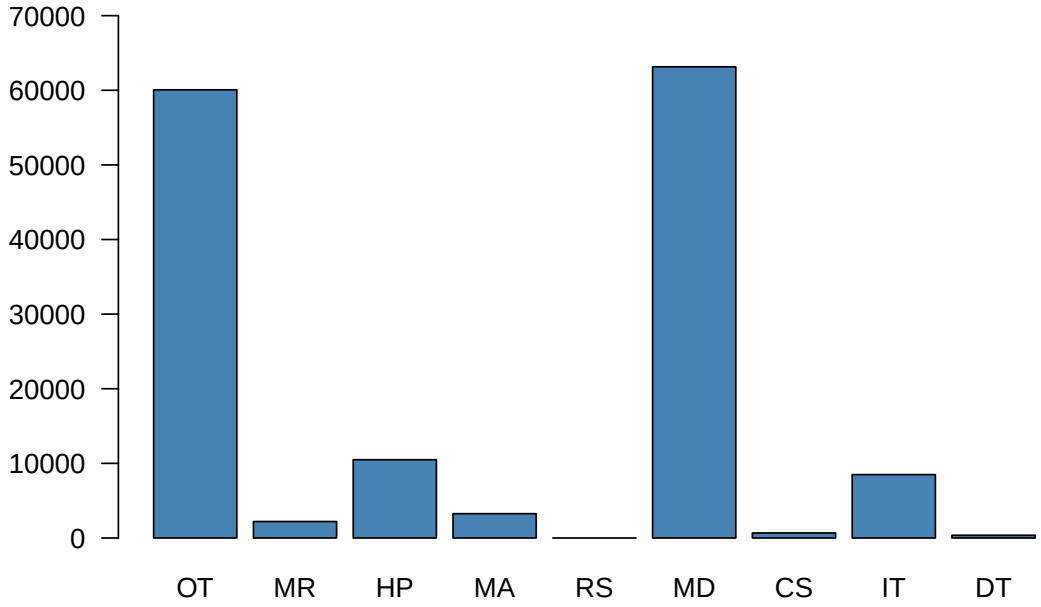


Figure 4.4.7: Number of invoices per treatment group.

With regard to these different types of treatment, differences can also be seen on an exploratory level. However, they are less stark than those that occur with the diagnoses made: Just like with the diagnoses, the total invoice amount and the number of invoices for the first six months after birth of a given child were determined. None of the children had a rehabilitation stay (**RS**), therefore, this treatment will be dropped. Also, most of the treatments seem to have no differences between both groups. However, looking at Table 4.2, which, like 4.1, shows the proportion of children per group for whom there is at least one invoice with the respectively associated type of treatment, it should be noted that differences between the groups can be observed with regard to the medical remedies (**MR**) and the medical aids (**MA**) as well as the inpatient treatments (**IT**). In those instances, the differences are very apparent: Relatively speaking, children with a note need a medical remedy or medical aid twice as often as those without one. With inpatient treatment, the difference is even more significant: While in group 0, less than one in five children has an invoice for an inpatient treatment, one was recorded for three out of four children with a note. This trend continues in the exploratory analysis when looking at the boxplots in Figures A.3.7 to A.3.13 in the appendix and in Figure 4.4.8 below:

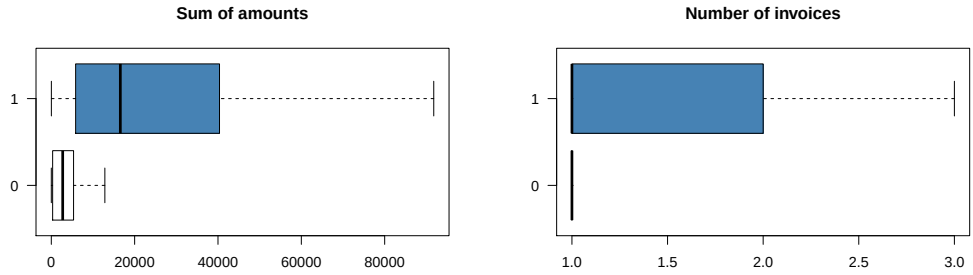


Figure 4.4.8: Boxplots of inpatient treatment (**IT**).

The total costs for inpatient stays are on average significantly higher, as is the total number of bills. Children with possible birth defects are, therefore, treated more often in a hospital and their stay is more expensive. The variable **Days in Hospital** whose boxplot can be found in Figure 4.4.9 supports this finding: While children without a note spend almost no day in the hospital within the first six months of their life, 50% of the children with a note undergo almost 20 days of inpatient treatment.

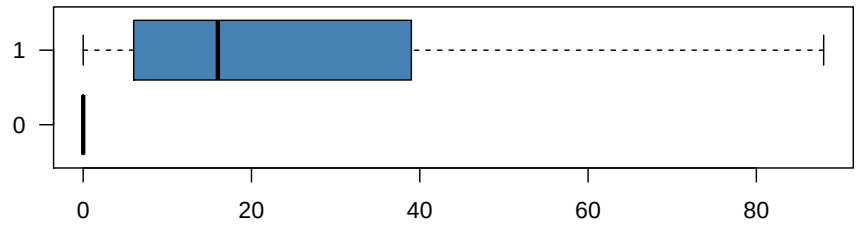


Figure 4.4.9: Boxplot of the **Days in Hospital**.

The exploratory analysis thus leads to the conclusion that the transformation of the database or its transfer into 41 numerical variables (24 from the diagnoses, 8 treatments with 2 variables each and the variable **Days in Hospital**) provides a meaningful basis for further model-based analysis. Before looking at this model-based analysis, it is important to point to the correlation analysis of the diagnoses and types of treatment: Within the variables that relate to the diagnoses or types of treatment, only low values can be observed (see Figures A.3.14 and A.3.15 in appendix). Only the total invoice amount of the diagnoses **P**, **Z** and **Q** and "**Rare**" show a stronger correlation. Between the diagnoses and the types of treatment, there are sometimes strong correlations between the variables of the diagnoses **P**, **Q**, **R**, **Z** and "**Rare**" (rows of Table 4.3) to the total invoice amount and number of inpatient

treatments (**IT**) and **Days in Hospital** (columns of Table 4.3). The findings also allow for a reconstruction of the children's medical condition from the overall consideration of the invoice data if a transformation with regard to the main diagnoses and types of treatment takes place.

correlation	Sum.IT	N.IT	Days in Hospital
Sum.P	0.770	0.260	0.665
N.P	0.684	0.549	0.739
n.P	0.418	0.647	0.527
Sum.Q	0.537	0.223	0.384
N.Q	0.263	0.253	0.228
n.Q	0.371	0.363	0.360
Sum.R	0.462	0.202	0.348
N.R	0.112	0.194	0.127
n.R	0.263	0.356	0.271
Sum.Z	0.737	0.289	0.646
N.Z	0.232	0.319	0.264
n.Z	0.397	0.650	0.479
Sum.Rare	0.579	0.222	0.435
N.Rare	0.273	0.370	0.274
n.Rare	0.308	0.460	0.323

Table 4.3: Correlation analysis of diagnoses **P**, **Q**, **R**, **Z** and "**Rare**" (rows) to the total invoice amount, number of inpatient treatments and **Days in Hospital** (columns).

4.4.2 Modeling

The exploratory analysis of the data allows for the conclusion that some variables, particularly those associated with the diagnosis **P**, can explain the target variable. This justifies the use of regression-based models. The advantage of these models over alternative models is the possibility of evaluating the relevance of individual explanatory variables for the target variable and, thus, deriving a possible factual connection from the model analysis. Especially in this application context that aims to replicate the human decision algorithmically, the use of algorithms, which could also be interpreted, would be easier to justify. In her fundamental article, Rudin (2019) explains that it makes more sense to use models that can be interpreted directly, particularly when applied in the health sector, instead of using black-box models such as artificial neural networks and striving to explain their results after use. She further elaborates that a lack of interpretability of the models is not inevitably compensated by better performance. The present work, therefore, deals, on the one hand, with the application of different variations of logistic regression and uses, on the other hand, variants of the random forest model, an algorithm from the field of machine learning, which is also able to map the importance of the individual explanatory variables, as a benchmark.

The different variations of these models also result from the necessity to address the rare event problem: Many models used for binary classification are based on the fact that both classes are observed with the same frequency or that the costs resulting from an incorrect classification are the same. Neither is true in our case: On one hand, birth defects are noted much less often. On the other hand, the scenario in which a child is mistakenly considered healthy so that no examination for birth defects is carried out results in significantly higher costs compared to the alternative in which a healthy child is examined for birth defects unnecessarily. As a consequence, this work faces a classic rare event problem.
Krawczyk (2016) distinguishes between three approaches that can address such problems:

- Data-level methods
- Algorithm-level methods
- Hybrid methods

While sampling methods are mainly used at the data-level to artificially guarantee the balance of the data, the methods used at the algorithm-level aim to make the underlying model cost-sensitive to the underrepresented class (see Krawczyk (2016) for more details). These two approaches, considered individually, eliminate the inequality with regard to the observation of both classes or the inequality with regard to the costs of an incorrect classification. In order to do justice to both aspects of the rare event problem, the third approach of hybrid methods, in which sampling methods are combined with modifications of the algorithms, might be an alternative.
Logistic regression is one of the most widely used classification methods in econometrics and is also used as a benchmark model in many works examining the performance of machine learning methods. The basis of logistic regression is the assumption of a probabilistic model:

$$F(\beta, x_i) := \mathbb{P}(y_i = 1|x_i) = \frac{1}{1 + \exp(-\tilde{x}_i^\top \beta)} \text{ for } \beta \in \mathbb{R}^{m+1} \text{ with } \tilde{x}_i = (1, x_i) \in \mathbb{R}^{m+1} \quad (4.4.1)$$

Based on this model, the associated log-likelihood is numerically maximized:

$$\mathcal{L}(\beta, X, y) = \Sigma_{i=1}^{N} \left(y_i \log \left(F(\beta, x_i) \right) + (1 - y_i) \log \left(1 - F(\beta, x_i) \right) \right) \quad (4.4.2)$$

Two problems arise from this approach against the background of the rare event problem: First of all, only a few summands in Formula 4.4.2 are observed at all with the value $y_i = 1$ (= child with a note), which raises the question of whether there are enough cases at all to adjust the model accordingly. To solve this problem, it is advisable to use sampling methods (data-level method). Secondly, Formula 4.4.2 suggests that a misclassification causes the same costs in both classes. In their work, Günnemann and Pfeffer (2017) use the latter issue as a starting point to proceed to present various cost-sensitive approaches on how the log-likelihood can be adjusted in the case of rare events: The aim of their work is to use this adjustment to better adapt the logistic regression model to the classification of

the rare class (algorithm-level method). However, this only succeeds if the costs generated by misclassifying the rare events are known or can be reasonably estimated (see Günnemann and Pfeffer (2017) for details). This is not possible in our case as no corresponding data on children who erroneously did not receive a birth defect note exists. Instead, we would like to further develop another approach that takes up the idea of adjusting the log-likelihood: Shen et al. (2020) supplement the cost-sensitive log-likelihood with the AUC value of the associated ROC curve. In order to find the optimal parameterization β that minimizes the costs and maximizes the AUC value, Shen et al. (2020) employ a particle swarm optimization algorithm. Although our approach is very similar, we do not adjust the log-likelihood with regard to possible cost parameters since we could not estimate them from the data. Hence, we only add two penalty terms to the log-likelihood: The AUC penalty term $\rho \log\left(\text{AUC}(\beta, X, y)\right)$ and a LASSO penalty term $\lambda\|\beta\|_1$ (cf. Hastie et al. (2001)) to account for possible multicollinearity in β:

$$\mathcal{H}(\beta, X, y, \rho, \lambda) = \mathcal{L}(\beta, X, y) + \rho \log\left(\text{AUC}(\beta, X, y)\right) - \lambda\|\beta\|_1 \tag{4.4.3}$$

As in Shen et al. (2020), a swarm particle optimization algorithm is used to search for the optimal parameterization β for certain values of ρ and λ. It is implemented as follows:

Particle swarm algorithm

(I) Draw particles $B^0 = \{\beta_1^0, \ldots, \beta_P^0\}$ by taking P times vector β and setting randomly 90% of the entries of β to 0 (**=less chance for correlation in** X).

(II) Draw directions $V = \{v_1^0, \ldots, v_P^0\}$ with $v_k^0 \overset{\text{iid}}{\sim} \mathcal{N}\left(\mathbf{0}, \mathcal{I}_{m+1}\right)$ with $\mathbf{0} \in \mathbb{R}^{m+1}$.

(III) Set $\{\beta_1^*, \ldots, \beta_P^*\} = B^0$ (**individual optimum**) and compute β^*, whereas $\mathcal{H}(\beta^*, X, y, \rho, \lambda) \geq \mathcal{H}(\beta_k^*, X, y, \rho, \lambda)\ \forall\ k = 1, \ldots, P$ (**global optimum**).

(IV) For $k = 1, \ldots, P$ do:

- (i) $\beta_k^{i+1} = \beta_k^i + v_k^i$.
- (ii) If $\mathcal{H}(\beta_k^{i+1}, X, y, \rho, \lambda) > \mathcal{H}(\beta_k^*, X, y, \rho, \lambda)$: $\beta_k^* = \beta_k^{i+1}$ (**individual update**).
- (iii) If $\mathcal{H}(\beta_k^{i+1}, X, y, \rho, \lambda) > \mathcal{H}(\beta^*, X, y, \rho, \lambda)$: $\beta^* = \beta_k^{i+1}$ (**global update**).
- (iv) $v_k^{i+1} = r_1 v_k^i + r_2(\beta_k^* - \beta_k^{i+1}) + r_3(\beta^* - \beta_k^{i+1})$ with $r_q \sim \mathcal{U}[0, \frac{1}{3}]$ for $q = 1, 2, 3$.

(V) Repeat (IV) until β^* does not change due to a tolerance rate and set $\hat{\beta} = \beta^*$.

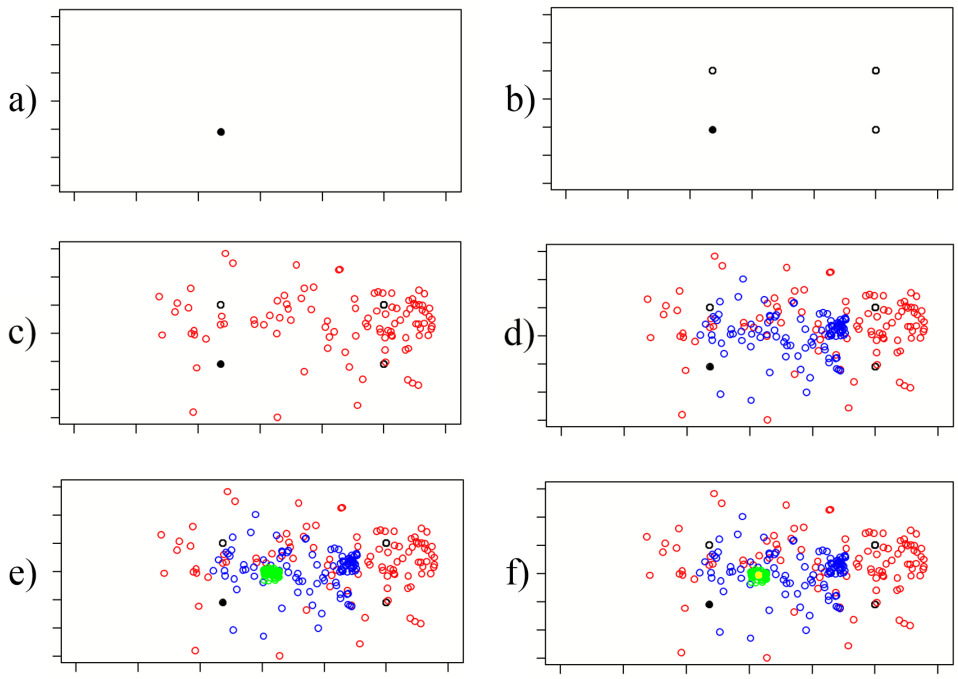


Figure 4.4.10: Visualization of a particle swarm optimization in $\mathbb{R}^2$.

Illustration 4.4.10 shows how the algorithm works in the case of two variables: After determining the log-likelihood solution (step I, see (a)), the particles are generated by setting different entries of β to the value 0 (step II and III, see (b)). After that, the individual particles are updated step by step, with each particle moving in the direction of the global and the individual optimum while taking into account a certain inertia (steps IV.i to IV.iv, see (c) to (e)). Finally, all particles reach the point at which the individual and global optimum no longer differ (step V, see (f)). The precise location of this solution depends on the hyper-parameters ρ and λ. A grid search is used to determine which values of ρ and λ lead to the best result (more on this in the next section). With regard to the logistic regression, two model variants ensue: The first one is the simple logistic regression that uses oversampling to address the rare event problem (data-level method). We decide to use oversampling according to the findings of Marqués et al. (2013): In their work, they provide evidence that oversampling of the data generates better results than undersampling because no relevant information from the database is left out. Our second model is the AUC-corrected logistic regression in which the model is adjusted to the classification of rare events by supplementing corresponding penalty terms in the log-likelihood (algorithm-level method). Aside from these model variants, we will also consider a third approach: Instead of the AUC with regard to the ROC curve, the AUC with regard to the precision-recall curve (short: PR curve) is used in (4.4.3). We expect this to account for the rare event problem to an even larger extent as there are some articles in the literature that recommend the use of the PR curve instead of the ROC curve in the case of rare events (e.g. Davis

and Goadrich (2006), Sofaer et al. (2019) and Ozenne et al. (2015), more on this in the next section). In order to distinguish both model variants from each other, we will speak of a ROC-AUC-corrected logistic regression versus a PR-AUC-corrected logistic regression. However, besides using a different AUC, the implementation remains the same.
Random forest from the field of machine learning will serve as a benchmark for these models: This algorithm is often contrasted with logistic regression in application papers, such as Couronné et al. (2018). The authors of this paper justify the comparison of both models with the fact that logistic regression has been widely used in classical statistics. This model allows an interpretation of the coefficients and thus the influence of the individual variables on the target variable. In contrast, random forest to some extend is a sort of a black box offering only limited possibilities for interpretation. However, in many cases random forest performs better than logistic regression (see Couronné et al. (2018)). For a brief illustration of the algorithm developed by Breiman (2001), see Figure 4.4.11:
In step (I), N sets of data are created by bootstrapping. In step (II), an unpruned CART decision tree is created on these data sets. In this decision tree, only a certain number of randomly selected features (instead of the whole set) generate a decision rule for each split. A new data point passes through all the decision trees in step (III). The respective trees then assign it to one of the two classes. The relative number of trees assigning the data point to a particular class is set as the probability that this point belongs to the corresponding class (see Breiman (2001) for details). Random forest and logistic regression both have a probability as output which provides another argument for the comparability of both models. In addition, random forest can be easily adapted to the rare event problem as the developer of random forest himself explains in his 2004 paper (see Chen et al. (2004) for details): As a first possibility, a weight adjustment can be made in step (II) during tree creation which results in weighing the misclassification error of the rare class more heavily than that of the common class (weighted random forest). In contrast, there is also the possibility of pulling the bootstrap samples in step (I) separately from both classes in order to create balanced data sets overall on which the decision trees are generated (balanced random forest). While the weighted random forest belongs to the group of algorithm-level methods, the balanced random forest is a data-level method. However, the weighted random forest approach cannot be pursued further for the same reasons that have already been discussed in the framework of the cost-sensitive logistic regression: As long as the costs of misclassifying a child with birth defects are not documented, these costs cannot be calculated from the available data. For this reason, the balanced random forest approach is being pursued further.
Another modification of the method is to be tried out on the data as well: In accordance with the hybrid classification approach of the first chapter of this work, the balanced random forest is applied to a *DD*-Plot of the data by referring back to the Mahalanobis depth. When creating this *DD*-Plot, the inclusion of those variables that are present in a data nest structure (see chp.1 for details) is also made a priority.

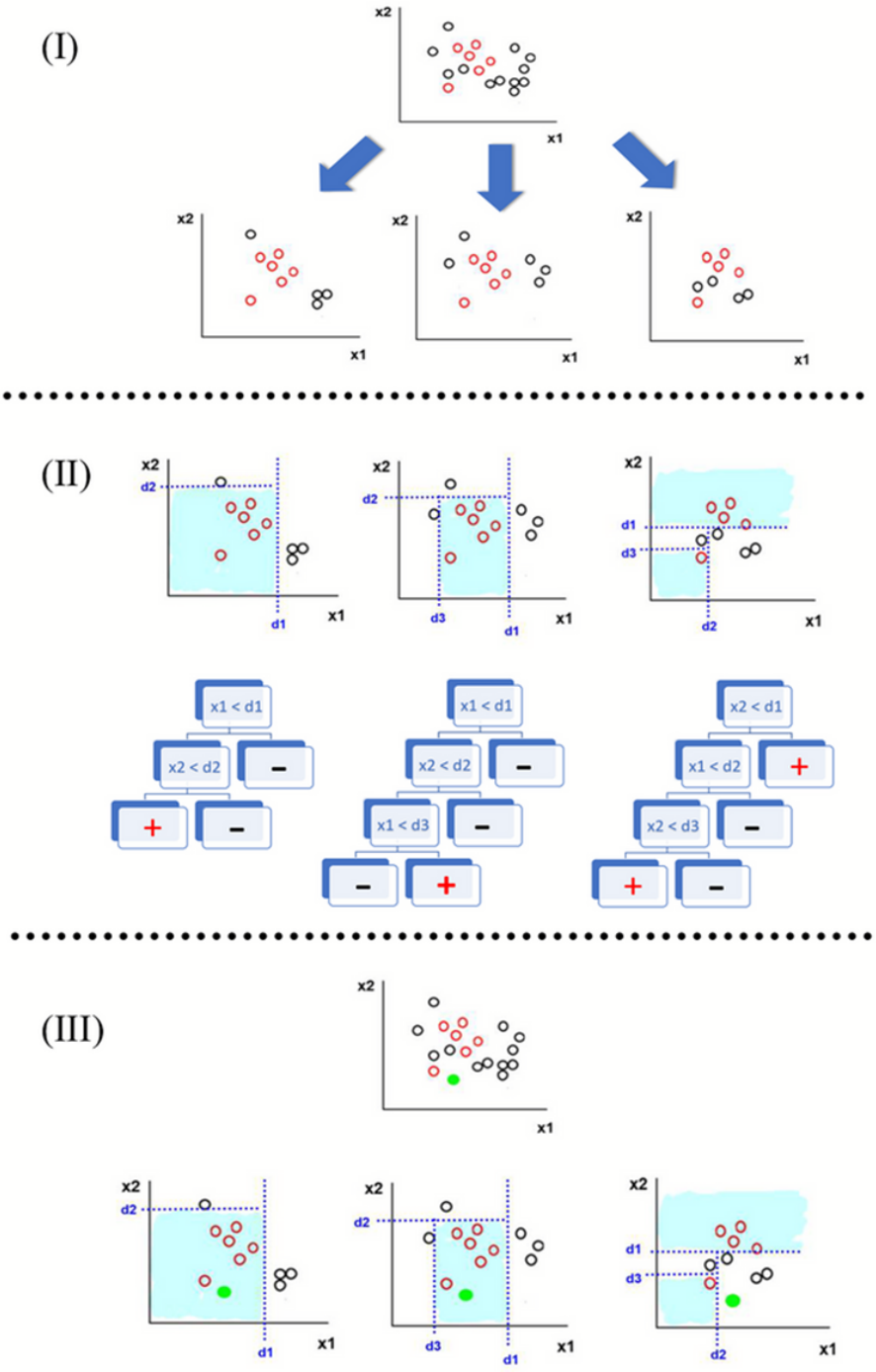


Figure 4.4.11: Steps of random forest in $\mathbb{R}^2$.

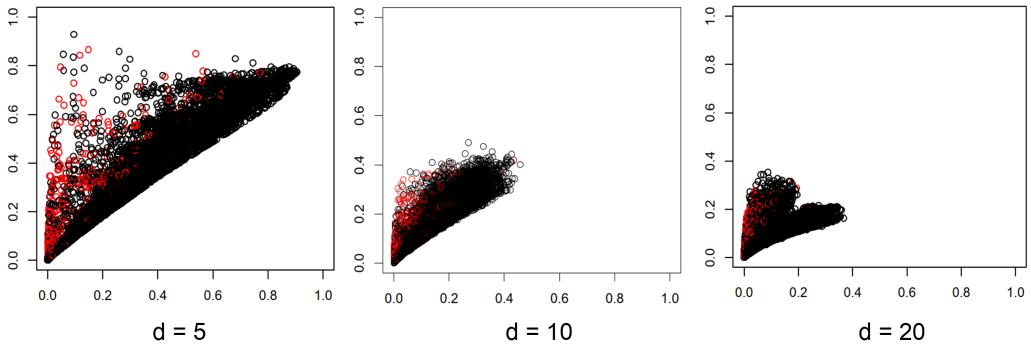


Figure 4.4.12: *DD*-Plots for three different combination of 5, 10 and 20 variables.

Figure 4.4.12 shows three different *DD*-Plots based on different combinations of variables: As is to be expected from the theoretical explanations in the first chapter, the *DD*-Plot collapses as the dimension increases. Whether the separation of the data increases depends on how strong the data nest structure is. Ideally, a *DD*-transformation of the data can therefore support the classification of rare events. However, this only applies if the plot has not yet collapsed to such an extent that separation is no longer possible. A data transformation (data-level method) in combination with a balanced random forest (also data-level method) thus produces a fifth model which can be attributed to the data-level methods with regard to the rare events problem as well.

4.4.3 Evaluation

The rare event problem must also be taken into account when measuring the performance of the models as not all measures are suitable in this case. In addition, it is important to consider the background of the data analysis in this context: The aim is to identify those children who would be selected to be screened for birth defects by a human decision-maker. Since all models generate probabilities as output, it would be possible to measure performance using a confusion matrix, as shown in Figure 4.4.13. The entries in the matrix would be evaluated for a specific threshold value π, but weighting would have to be carried out due to the rare event problem. However, since such weights would be difficult to estimate, another measure is preferable: The area under the curve (AUC). This approach works both with the ROC as well as the PR curve and can be determined independently of a specific threshold. Such a threshold is not suitable for the present application since the goal is to carry out a possible birth defect screening. Accordingly, it is sufficient to make a list of the N children to be examined, with children at the top of this list for whom the algorithm assumes with a high probability that a human decision-maker would carry out a birth defect test. This list would then be processed starting from the top over the first $n < N$ children depending on the test capacity. Consequently, it is not necessary to determine a

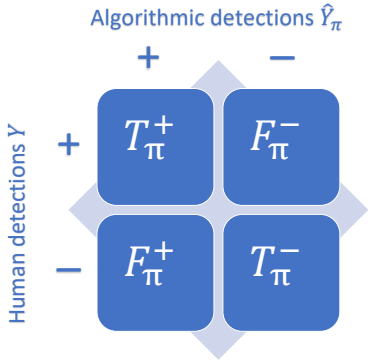


Figure 4.4.13: Confusion-matrix with True-Positives (T_π^+), False-Positives (F_π^+), False-Negatives (F_π^-) and True-Negatives (T_π^-) depended on threshold π.

specific threshold but rather to check whether children with a higher probability do in fact receive more birth defect notes. This is precisely what the AUC measure does. As Davis and Goadrich (2006) elaborate in their paper, the ROC and PR curve are closely related. But not every model that maximizes the ROC curve will also maximize the PR curve (see Davis and Goadrich (2006) for details). In addition, the PR curve is better suited for rare events than the ROC curve which is otherwise widely used in areas of machine learning. Ozenne et al. (2015) highlight this in their work: They use a simulation study to show that in the case of rare events, the ROC-AUC overestimates the performance of the models. In contrast, the PR-AUC estimates the ability of the models to correctly forecast rare events more realistically. Sofaer et al. (2019) provide additional empirical evidence that the PR curve is better suited to measure the performance of a model than the ROC curve in the case of rare events: Their work deals with statistical models that estimate the distribution area of certain animal species in the United States based on data relating to their sighting and additional geographic data. They find that a performance measurement using ROC-AUC overestimates the distribution area precisely when it comes to rare animal species. In contrast, a performance comparison using PR-AUC provides more accurate results in those cases. In the case of rare events, a consideration of the PR curve in addition to the ROC curve can therefore be quite advantageous following these research contributions. Another advantage of the PR curve is that it is easy to interpret. In Figure 4.4.14 we have a PR curve which was calculated on the training set using oversampled logistic regression: For all thresholds, starting at 1 (violet) and continuing to 0 (red), the combinations of precision and recall, also known as sensitivity or TPR (true positive rate, see chp. 1), are determined:

$$\text{Recall}(\pi) = \frac{T_\pi^+}{T_\pi^+ + F_\pi^-} \qquad \text{Precision}(\pi) = \frac{T_\pi^+}{T_\pi^+ + F_\pi^+}$$

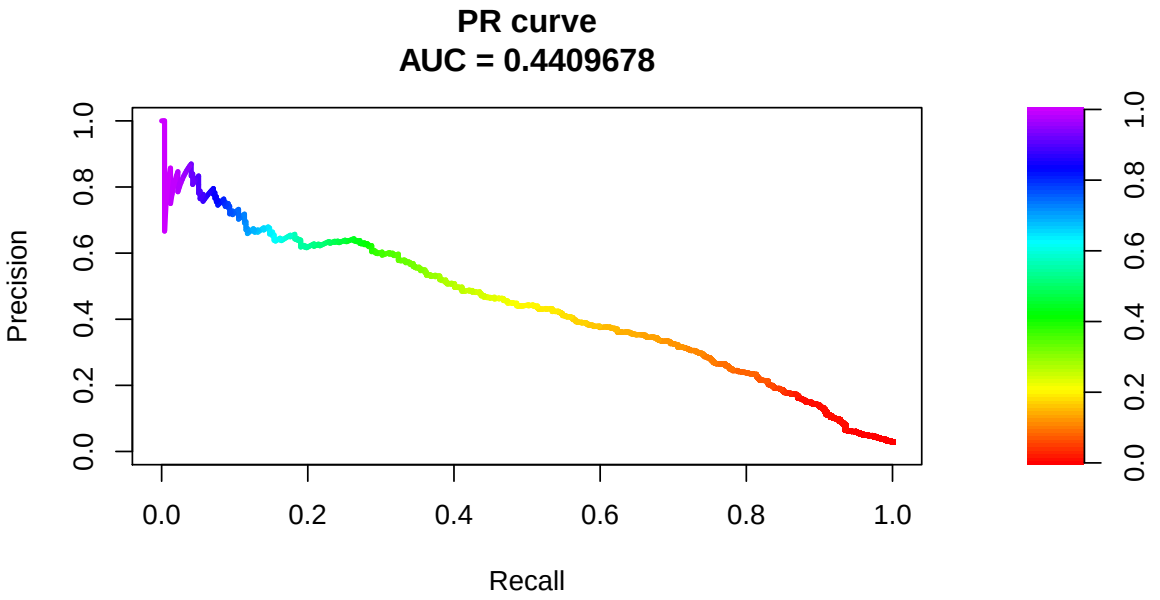


Figure 4.4.14: PR curve on the training set according to oversampled logistic regression.

Looking at a specific point on the PR curve, the number of children given a note by the model can be deduced: The number of flagged children with a note is divided by the precision value (y-axis) at that point. The resulting value encompasses all children that the model flagged. The recall value (x-axis) then shows which percentage of the children who received a note were identified by the model. The trajectory of a good PR curve ideally achieves high precision and recall values at the same time. Such a trajectory appears when the AUC value is close to 1. PR-AUC is therefore used when evaluating the results. For reason of completeness, ROC-AUC will be used though. On top of that, precision and recall values will be calculated for a certain n: Both values will show how the corresponding algorithm will perform when only the first n children with the highest output are taken into account. Said results were measured on a test set containing 20% ($\approx 5,000$ / ≈ 150 ones) randomly selected cases from the total data set. The remaining 80% ($\approx 20,000$ / ≈ 600 ones) served as a training set on which the models were tuned. Different tuning-approaches were applied:

- **Oversampled logistic regression**: The model was fitted on a training set that was offset by oversampling ($\approx 25,000$ / $\approx 5,000$ ones).
- **AUC-corrected logistic regression (PR and ROC)**: On the training set that has not been sampled, a 3-fold-cross-validation was carried out on a grid for various combinations of ρ and λ and a selection of the parameterization which led to the highest averaged AUC value (see Figure 4.4.15) occurred ($\rho = 100,000$ and $\lambda = 1,000$ for PR-AUC-corrected, $\rho = 100$ and $\lambda = 100$ for ROC-AUC-corrected).

- **Balanced random forest**: There are also two hyper-parameters; the number of trees (*ntree*) and the number of variables selected randomly for a split generation (*mtry*). The grid search occurred automatically using the `R`-function `tuneRF` based on the OOB error and leads to *mtry*= 28, whereas *ntree*= 500 was chosen by default.

- ***DD*-random forest**: Since only two variables are available, i.e., *mtry*= 2, a grid search was not necessary. The number of trees from balanced random forest was simply taken over whereas the variables belonging to diagnoses **P** and treatment **IT** as well as **Days in Hospital** according to the exploratory analyses were used for the *DD*-Plot. These variable selection also produced a *DD*-Plot which was not collapsed (see Figure 4.4.16).

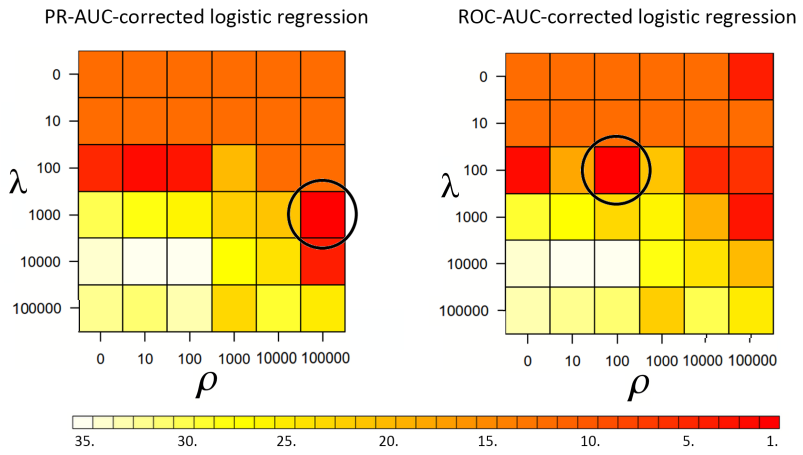


Figure 4.4.15: Rank-Heat-Plots based on a 3-fold-cross-validation of highest averaged AUC-values. The Grid is built over λ (LASSO-Term) and ρ (AUC-Term) for the PR-AUC-corrected logistic regression (left) and ROC-AUC-corrected logistic regression (right). Highest AUC-value is cycled in black.

Figure 4.4.17 summarizes the results: In addition to the individual PR curves and ROC curves, the AUC values of the models are also listed as well as the precision and recall when $n = 500$. We chose this value for n to simulate the realistic scenario in which, for capacity reasons, only 10% of the children can examined further for possible birth defects. As a result, a user of these algorithms will only look on the first 500 children of the test set with the highest outcome. The highest PR-AUC is achieved with the PR-AUC-corrected logistic regression (47.45%), which beats both oversampled logistic regression (44.72%) and ROC-AUC-corrected logistic regression (45.86%). Balanced random forest (39.86%) and

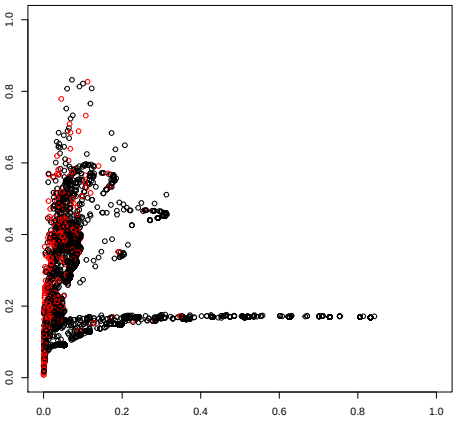
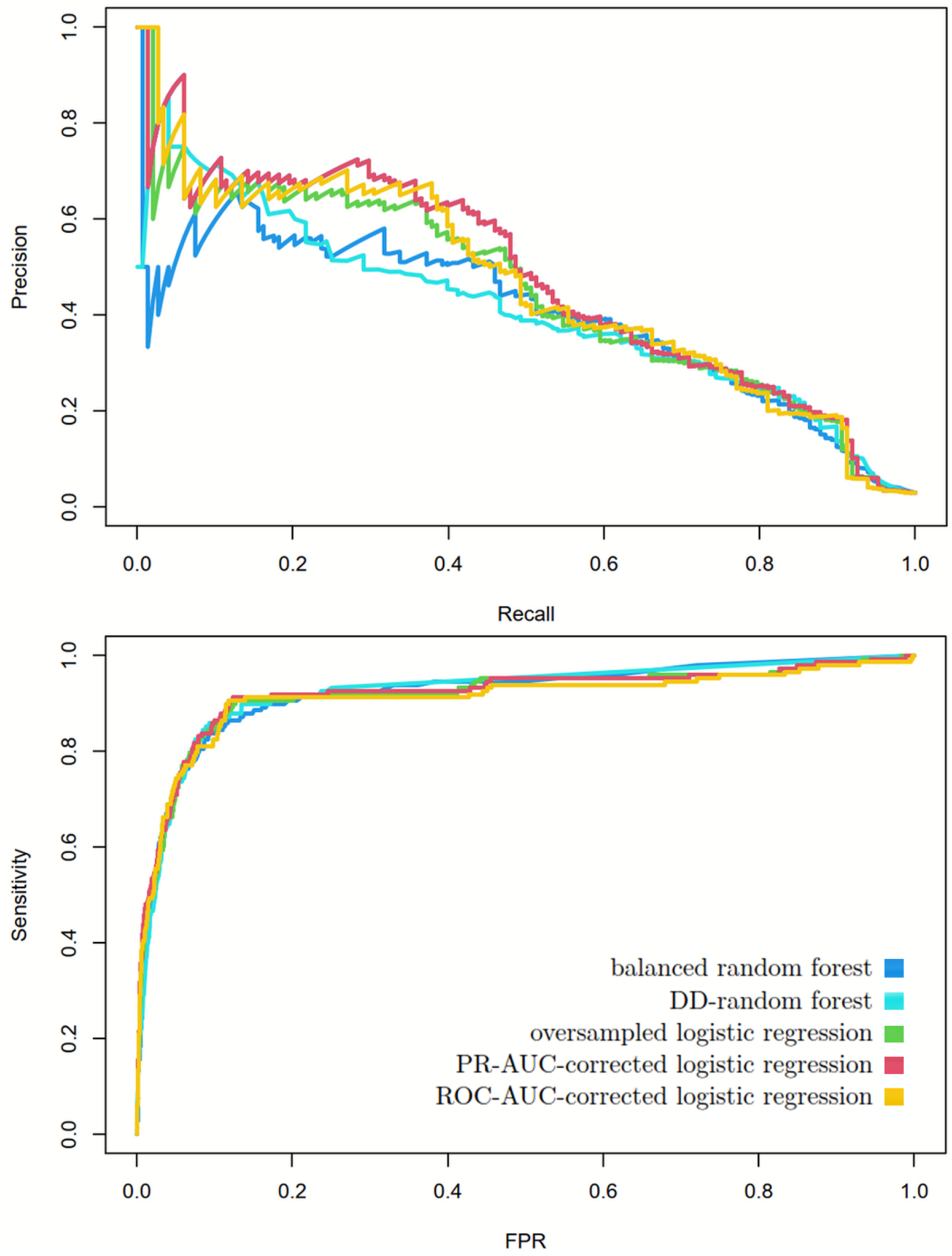


Figure 4.4.16: *DD*-Plot on the training set using all variables belonging to diagnose **P** and treatment **IT** as well as variable**Days in Hospital**.

DD-random forest (40.83%) are beaten by all logistic regression models. However, these observations do not hold when looking at the precision and recall for $n = 500$: *DD*-random forest generates the highest values (precision=24.40% / recall=82.43%), but still PR-AUC-corrected logistic regression is the second best model (precision=24.20% / recall=81.76%). ROC-AUC-corrected logistic regression (precision=23.60% / recall=79.73%) and balanced random forest (precision=23.40% / recall=79.05%) have the lowest performance with these values. When comparing the course of the ROC curves with the PR curves, it is noticeable that almost no differences between the models can be observed if the ROC curve alone is considered as a measure. The same applies to the AUC values. Such a comparison would, therefore, not be very informative. Evidence from the literature, namely that the analysis of the ROC curve loses its meaningfulness in the case of rare events, confirms this. The PR curve, which allows for a clearer distinction between the models in terms of both the course and the AUC values, proves to be significantly more meaningful. A comparison of the respective models with regard to the underlying methods that were used to address the rare event problem (data-level versus algorithm-level) leads to the following observation: When using the more meaningful PR-AUC measure for comparison, the two best models are the two AUC-corrected logistic regressions which are based on algorithm-level methods. The other models, which perform worse, are based on data-level methods. Addressing the rare event problem at the algorithm-level rather than at the data-level thus seems to generate better results. However, looking at a specific value of precision and recall for a given n, PR-AUC-corrected logistic regression still performances well but in this case, the *DD*-random forest also promises to produce good results.



Model	PR-AUC	ROC-AUC	precision	recall
balanced random forest	39.86%	92.13%	23.40%	79.05%
DD-random forest	40.83%	92.40%	24.40%	82.43%
oversampled logistic regression	44.72%	91.69%	24.00%	81.08%
PR-AUC-corrected logistic regression	47.45%	91.94%	24.20%	81.76%
ROC-AUC-corrected logistic regression	45.86%	90.99%	23.60%	79.73%

Figure 4.4.17: PR curves (top) and ROC curves (bottom) for all models. PR-AUC and ROC-AUC are also listed as well as values of precision and recall for the $n = 500$ children with the highest corresponding outcome.

x_i	β_i	p-Value	
Intercept	-3.684	$< 2 \cdot 10^{-16}$	(***)
n.P	1.131	$9.48 \cdot 10^{-11}$	(***)
N.P	0.1122	$5.47 \cdot 10^{-9}$	(***)
Sum.IT	0.00002381	0.000244	(***)
Days in Hospital	-0.01569	0.01986	(*)

Table 4.4: Most significant variables according PR-AUC-corrected logistic regression.

Despite the different model approaches and difference in performance, there are also commonalities in all models. This becomes apparent when looking at the balanced random forest (lowest PR-AUC) and the PR-AUC-corrected logistic regression (highest PR-AUC), for example: Table 4.4 shows the four variables (except for the constant) whose parameters were significant. A comparison of these variables to the first ten entries of the variable importance plot (Figure 4.4.18) of balanced random forest highlights that the variables for diagnosis **P** (**n.P**/**N.P**/**Sum.P**) as well as **Sum.IT** and **Days in Hospital** can be observed in both cases. This not only corresponds to the results of the exploratory data analysis but also allows conclusions to be drawn about human decision-making: As already mentioned, the diagnoses of group **P** point towards a complication in the birth process (see appendix). In addition, there are high costs for inpatient treatment and a longer stay at the hospital for a given child. These three variables are arguably also the crucial ones a human decision-maker considers when deciding on whether to issue a birth defect note.

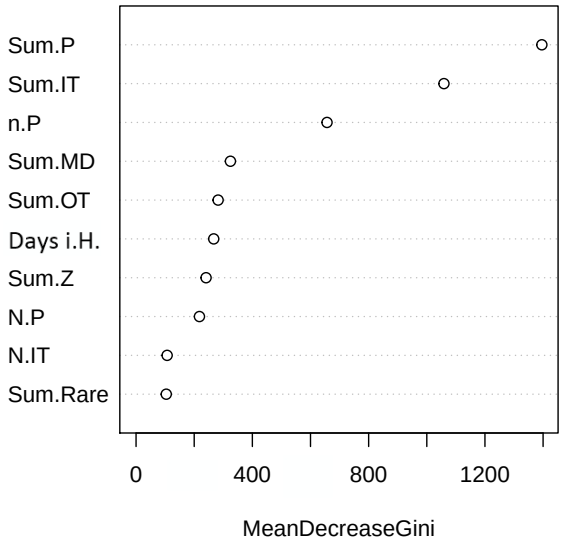


Figure 4.4.18: Variable-Importance-Plot of balanced random forest for the 10 most important variables according to the mean decrease of Gini.

The data mining analysis deliberately contrasts model variants of the logistic regression with the random forest ones. However, models are often combined in the specific application context in order to improve performance (ensemble learning, see Hastie et al. (2001)). This also applies in the present case: If the geometric mean is calculated from the outputs of the *DD*-random forest (highest precision and recall at $n = 500$) and the output of the PR-AUC-corrected logistic regression (highest PR-AUC), the resulting model has a better course of the PR curve, a higher PR-AUC value (see Figure 4.4.19) and the highest values for precision and recall at $n = 500$ (precision=24.6% / recall=83.11%). The same applies to the ROC curve compared to the other models (see Figure 4.4.20). The combined model of *DD*-random forest (data-level method) and PR-AUC-corrected logistic regression (algorithm-level method) uses a hybrid model approach to overcome the rare event problem, which could be an additional reason for the better performance.

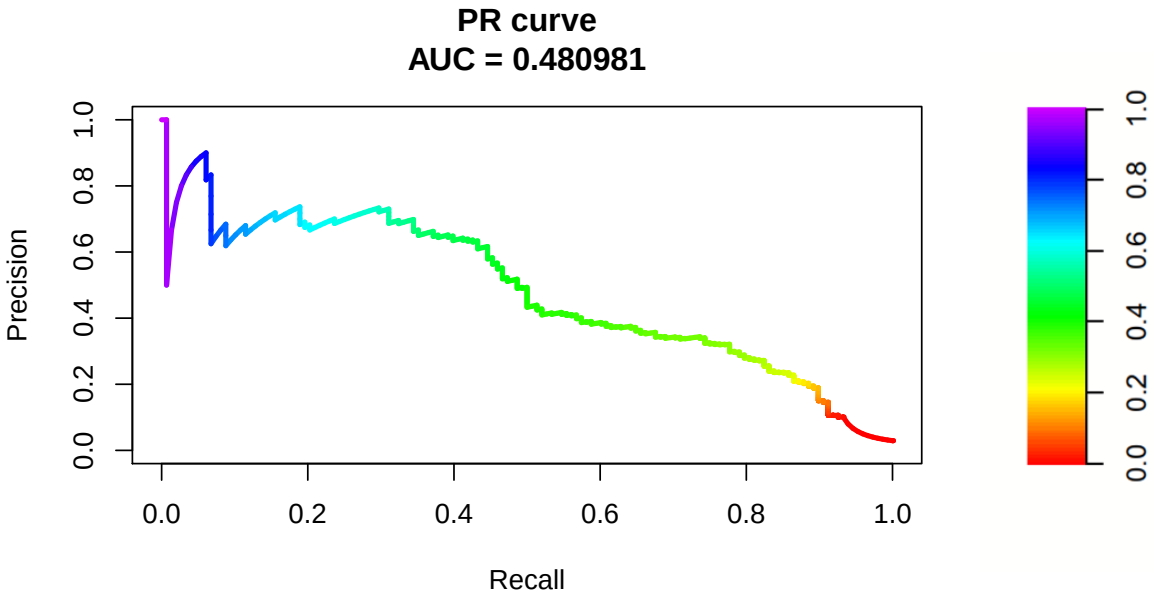


Figure 4.4.19: PR curve of the combined model.

Concerning the performance itself, it ultimately has to be concluded that even the best model cannot precisely reproduce human decisions, even if the decision-making process is based on a similar calculation. Once more, Figure 4.4.17 serves to highlight this: Following the trajectory of the three highest PR curves, they would only have a precision of 20% with a recall value of 90%. This means that if the models predicted 90% of the cases in which the

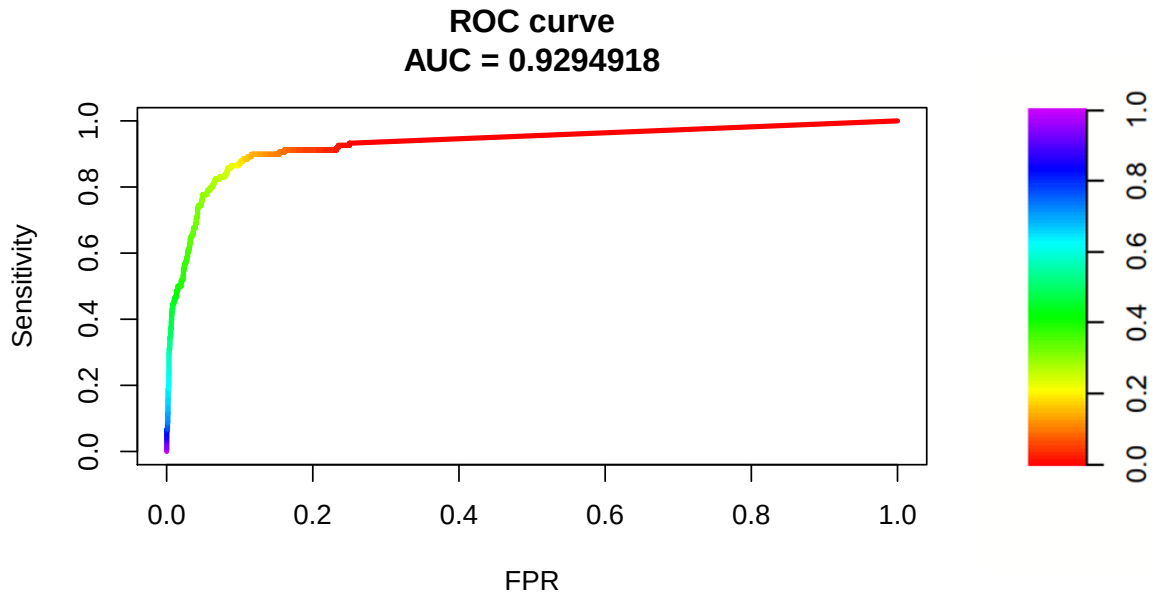


Figure 4.4.20: ROC curve of the combined model.

human decision-maker would also issue a note, they would mark five times as many children overall. Applied to this specific data set, this would mean that out of 25,000 children, the models would identify around 3,400 children to have birth defects but fail to flag more than 70 children which would be issued a birth defect note by a human decision-maker. At the same time, the models would flag 2,700 children who had not received a note from the human decision-maker. With more than 10% of the children, an algorithmic pre-screening would, therefore, trigger a false alarm. In addition, the algorithm would fail to detect 10% of the children whom a human decision-maker had issued a note to. However, this point of view could be criticized: A human decision-maker using these algorithms would not choose a point on the PR or ROC curve but a number n of children that should be investigated further. Taking this into account, the most reasonable value for interpretation is the precision and recall given $n = 500$ as these values show how well a decision-maker will perform starting at the top of the list of children with the corresponding highest outcomes. Looking from this perspective, most models receive a recall of 80% meaning that four out of five children who receive a note from the human decision-maker will also be flagged by the algorithms. And because only $n = 500$ children will be flagged, the number of false alarms will be much lower. This implementation of an algorithmic pre-screening is more realistic and suitable for the use in practices.

4.5 Conclusion

This chapter is dedicated to the question of whether an algorithmic pre-selection with regard to possible birth defects in newborns can be carried out on the basis of medical invoice data. For this purpose, we ran a data mining process in the form of a generic cycle model which corresponds to the approach of Reuß and Zwiesler (2006). In the course of this evaluation, the database was analyzed and the question was specified to allow for an algorithmic replication of the human decision-making calculation. We used a logistic regression model from the field of classical statistics and a random forest model from the field of machine learning. Both models were enriched with their own modifications. However, the PR-AUC-corrected logistic regression combined with *DD*-random forest produced the best result. Both the logistic regression and the random forest confirmed the insights gained from the exploratory data analysis, namely that the main diagnosis **P**, the costs of inpatient treatment and the number of days that a given child spent in the hospital are meaningful indicators for the human decision-maker to determine whether to issue a birth defect note. For the implementation in practice, one should generate a list of children ordered by the highest corresponding outputs. In a second step, the first n children of the list will be examined further for possible birth defects. n will be chosen by means of capacity. This implementation generates models that show, when $n = 500$, a recall around 80% and a precision between 20% to 25%. This might be caused by the fact that children without any invoices data still receive notes. In conclusion, algorithmic pre-screening based on invoice data can only be recommended as a supplementary, but not substitute, means of improving the detection of birth defects: The two model types that achieved the best results when combined with each other seem well-suited to support the human decision-making process or to interact with the human decision-maker. For example, the human decision-maker can pass the variables which he bases his decision on to the *DD*-random forest and thus consciously direct a machine translation of his decision-making process. Conversely, the PR-AUC correction of the logistic regression places a special focus on the detection of rare events which counteracts the risk that the human decision-maker overlooks critical cases. This last point should also be given special attention when criticizing the models – the data makes it impossible to determine whether the children which the models flagged but who did not receive a note from the human decision-maker, are, in fact, children who should have actually received a note based on the decision-making process. Future research in the field should, therefore, subject the cases that were flagged algorithmically to a case-by-case examination. This analysis would then allow for a final judgment as to whether algorithmic pre-screening using medical invoice data to detect birth defects is possible and advisable.

Appendix

A.1 Appendix of Chapter 2

Proof of ***Lemma*** 4:

□ For 1: With corresponding $z \sim S_d(\Gamma)$ we get for every $1 \leq i \leq n$ and $1 \leq j \leq m$:

$$x_i = \Omega_{\mathcal{X}^*}^{-\frac{1}{2}}(x_i^* - \mu_{\mathcal{X}^*}) = \Omega_{\mathcal{X}^*}^{-\frac{1}{2}}\left(\mu_{\mathcal{X}^*} + \Omega_{\mathcal{X}^*}^{\frac{1}{2}} z - \mu_{\mathcal{X}^*}\right) = \Omega_{\mathcal{X}^*}^{-\frac{1}{2}}\Omega_{\mathcal{X}^*}^{\frac{1}{2}} z = z \sim S_d(\Gamma) \text{ and}$$

$$\begin{aligned} y_j = \Omega_{\mathcal{X}^*}^{-\frac{1}{2}}\left(y_j^* - \mu_{\mathcal{X}^*}\right) &= \Omega_{\mathcal{X}^*}^{-\frac{1}{2}}\left(\mu_{\mathcal{Y}^*} + \Omega_{\mathcal{Y}^*}^{\frac{1}{2}} z - \mu_{\mathcal{X}^*}\right) = \Omega_{\mathcal{X}^*}^{-\frac{1}{2}}(\mu_{\mathcal{Y}^*} - \mu_{\mathcal{X}^*}) + \Omega_{\mathcal{X}^*}^{-\frac{1}{2}}\Omega_{\mathcal{Y}^*}^{\frac{1}{2}} z \\ &= \mu + \left(\left(\Omega_{\mathcal{Y}^*}^{\frac{1}{2}}\right)^\top \left(\Omega_{\mathcal{X}^*}^{-\frac{1}{2}}\right)^\top\right)^\top z = \mu + \left(\Omega_{\mathcal{Y}^*}^{\frac{1}{2}}\Omega_{\mathcal{X}^*}^{-\frac{1}{2}}\right)^\top z \sim EC_d(\mu, \Omega, \Gamma) \end{aligned}$$

because $\left(\Omega_{\mathcal{Y}^*}^{\frac{1}{2}}\Omega_{\mathcal{X}^*}^{-\frac{1}{2}}\right)^\top \Omega_{\mathcal{Y}^*}^{\frac{1}{2}}\Omega_{\mathcal{X}^*}^{-\frac{1}{2}} = \Omega_{\mathcal{X}^*}^{-\frac{1}{2}}\Omega_{\mathcal{Y}^*}\Omega_{\mathcal{X}^*}^{-\frac{1}{2}} = \Omega$. In 2 we just realize that $\Omega_{\mathcal{X}^*}^{-\frac{1}{2}} \in \mathbb{R}^{d\times d}$ is regular and $-\Omega_{\mathcal{X}^*}^{-\frac{1}{2}}\mu_{\mathcal{X}^*} \in \mathbb{R}^d$, so we can use Axiom (i) of the data depth axioms. For 3:

$$\begin{aligned} \Sigma_{\mathcal{X}^*} \geq \Sigma_{\mathcal{Y}^*} &\Leftrightarrow \Sigma_{\mathcal{X}^*} - \Sigma_{\mathcal{Y}^*} \text{ is pos. semi. def.} \underset{\mathbb{E}(r_d^2)<\infty}{\Leftrightarrow} \frac{\mathbb{E}(r_d^2)}{d}(\Omega_{\mathcal{X}^*} - \Omega_{\mathcal{Y}^*}) \text{ is pos. semi. def.} \\ &\Leftrightarrow \frac{\mathbb{E}(r_d^2)}{d}\Omega_{\mathcal{X}^*}^{-\frac{1}{2}}(\Omega_{\mathcal{X}^*} - \Omega_{\mathcal{Y}^*})\Omega_{\mathcal{X}^*}^{-\frac{1}{2}} \text{ is pos. semi. def.} \\ &\Leftrightarrow \frac{\mathbb{E}(r_d^2)}{d}(\mathcal{I}_d - \Omega) \text{ is pos. semi. def.} \underset{\mathbb{E}(r_d^2)<\infty}{\Leftrightarrow} \Sigma_{\mathcal{X}} - \Sigma_{\mathcal{Y}} \text{ is pos. semi. def.} \\ &\Leftrightarrow 0 < \varphi_k \leq 1 \text{ for all eigenvalues } \varphi_1, \ldots, \varphi_d \text{ of } \Omega. \quad \square \end{aligned}$$

Proof of ***Theorem*** 6:

□ Let $X = (x_1, \ldots, x_n)$ be an iid sample drawn from a r.v. $\mathcal{X} \sim S_d(\Gamma)$ and $Y = (y_1, \ldots, y_m)$ an iid sample drawn from a r.v. $\mathcal{Y} \sim EC_d(\mu, \Omega, \Gamma)$. Furthermore, for the generating variate $r_d \geq 0$ holds $\mathbb{E}(r_d^2) < \infty$ and $\operatorname{rank}(\Omega) = d$. Because of Proposition 2 we get:

$$\mu_{\mathcal{X}} = 0, \Sigma_{\mathcal{X}} = \frac{\mathbb{E}(r_d^2)}{d}\mathcal{I}_d \text{ and } \mu_{\mathcal{Y}} = \mu, \Sigma_{\mathcal{Y}} = \frac{\mathbb{E}(r_d^2)}{d}\Omega.$$

Therefore, because $\Sigma_{\mathcal{Y}}$ has to be symmetric and so do Ω, we find orthogonal $Q \in \mathbb{R}^{d\times d}$ for $\Omega = Q^\top DQ$ and $D = \operatorname{diag}[\varphi_1, \ldots, \varphi_d]$ with eigenvalues φ_k. If Y is a data nest of rare events w.r.t. X, $0 < \varphi_k \leq 1$ holds for $1 \leq k \leq d$ because of Lemma 4 (3). Moreover, the following

holds for randomly drawn x from $\mathcal{X}$ and y from $\mathcal{Y}$ because of Proposition 3:

$$\begin{aligned}
\mathcal{D}_M(x|\mathcal{X}) &= \left(1+(x-\mu_{\mathcal{X}})^\top \Sigma_{\mathcal{X}}^{-1}(x-\mu_{\mathcal{X}})\right)^{-1} = \left(1+x^\top\left(\frac{\mathbb{E}(r_d^2)}{d}\mathcal{I}_d\right)^{-1}x\right)^{-1} \\
&= \left(1+\frac{d}{\mathbb{E}(r_d^2)}||x||^2\right)^{-1} = \left(1+\frac{d}{\mathbb{E}(r_1^2)\,d}||x||^2\right)^{-1} = \left(1+\frac{||x||^2}{\mathbb{E}(r_1^2)}\right)^{-1}. \\
\mathcal{D}_M(y|\mathcal{Y}) &= \left(1+(y-\mu_{\mathcal{Y}})^\top \Sigma_{\mathcal{Y}}^{-1}(y-\mu_{\mathcal{Y}})\right)^{-1} = \left(1+(y-\mu)^\top\left(\frac{\mathbb{E}(r_d^2)}{d}\Omega\right)^{-1}(y-\mu)\right)^{-1} \\
&= \left(1+\frac{d}{\mathbb{E}(r_d^2)}\left|\left|\Omega^{-\frac{1}{2}}(y-\mu)\right|\right|^2\right)^{-1} = \left(1+\frac{d}{\mathbb{E}(r_d^2)}\left|\left|Q^\top\sqrt{D^{-1}}Q(y-\mu)\right|\right|^2\right)^{-1} \\
&= \left(1+\frac{d}{\mathbb{E}(r_1^2)\,d}\left|\left|\sqrt{D^{-1}}Q(y-\mu)\right|\right|^2\right)^{-1} = \left(1+\frac{\left|\left|\sqrt{D^{-1}}Q(y-\mu)\right|\right|^2}{\mathbb{E}(r_1^2)}\right)^{-1}. \\
\mathcal{D}_M(y|\mathcal{X}) &= \left(1+(y-\mu_{\mathcal{X}})^\top \Sigma_{\mathcal{X}}^{-1}(y-\mu_{\mathcal{X}})\right)^{-1} = \left(1+y^\top\left(\frac{\mathbb{E}(r_d^2)}{d}\mathcal{I}_d\right)^{-1}y\right)^{-1} \\
&= \left(1+\frac{d}{\mathbb{E}(r_1^2)\,d}||y||^2\right)^{-1} = \left(1+\frac{||y||^2}{\mathbb{E}(r_d^2)}\right)^{-1}. \\
\mathcal{D}_M(x|\mathcal{Y}) &= \left(1+(x-\mu_{\mathcal{Y}})^\top \Sigma_{\mathcal{Y}}^{-1}(x-\mu_{\mathcal{Y}})\right)^{-1} = \left(1+(x-\mu)^\top\left(\frac{\mathbb{E}(r_d^2)}{d}\Omega\right)^{-1}(x-\mu)\right)^{-1} \\
&= \left(1+\frac{d}{\mathbb{E}(r_1^2)\,d}(x-\mu)^\top Q^\top D^{-1}Q(x-\mu)\right)^{-1}.
\end{aligned}$$

Here $\sqrt{D^{-1}}$ stands for $\mathrm{diag}[\sqrt{\varphi_1^{-1}},\ldots,\sqrt{\varphi_d^{-1}}]$ and $\sqrt{D}$ for $\mathrm{diag}[\sqrt{\varphi_1},\ldots,\sqrt{\varphi_d}]$ and the term D^{-1} for $\mathrm{diag}[\varphi_1^{-1},\ldots,\varphi_d^{-1}]$ and therefore $\Omega^{-1} = \left(Q^\top DQ\right)^{-1} = Q^\top D^{-1}Q$ and what is even more $\Omega^{-\frac{1}{2}} = \left(\Omega^{\frac{1}{2}}\right)^{-1} = \left(Q^\top\sqrt{D}Q\right)^{-1} = Q^\top\sqrt{D^{-1}}Q$.
Moreover with $Q^\top D^{-1}Q = Q^\top\left(\mathcal{I}_d + (\mathcal{I}_d - D)\,D^{-1}\right)Q = \mathcal{I}_d + Q^\top\left((\mathcal{I}_d - D)\,D^{-1}\right)Q$ holds:

$$\begin{aligned}
\mathcal{D}_M(x|\mathcal{Y}) &= \left(1+\frac{1}{\mathbb{E}(r_1^2)}(x-\mu)^\top\left(\mathcal{I}_d + Q^\top(\mathcal{I}_d - D)\,D^{-1}Q\right)(x-\mu)\right)^{-1} \\
&= \left(1+\frac{1}{\mathbb{E}(r_1^2)}\left((x-\mu)^\top(x-\mu) + (x-\mu)^\top Q^\top(\mathcal{I}_d - D)\,D^{-1}Q(x-\mu)\right)\right)^{-1} \\
&= \left(1+\frac{1}{\mathbb{E}(r_1^2)}\left(||x-\mu||^2 + \left|\left|\sqrt{(\mathcal{I}_d - D)\,D^{-1}}Q(x-\mu)\right|\right|^2\right)\right)^{-1}.
\end{aligned}$$

Here $\sqrt{(\mathcal{I}_d - D)\,D^{-1}}$ stands for $\mathrm{diag}[\sqrt{\frac{1-\varphi_1}{\varphi_1}}, \ldots, \sqrt{\frac{1-\varphi_d}{\varphi_d}}]$. Moreover because of Proposition 2 for $z \sim S_d(\Gamma)$ corresponding to y holds:

$$\mathcal{D}_M(x|\mathcal{X}) = \left(1 + \frac{||x||^2}{\mathbb{E}(r_1^2)}\right)^{-1} \overset{\mathrm{d}}{=} \left(1 + \frac{\left|\left|r_d u^{(d)}\right|\right|^2}{\mathbb{E}(r_1^2)}\right)^{-1} \overset{\mathrm{d}}{=} \left(1 + \frac{||z||^2}{\mathbb{E}(r_1^2)}\right)^{-1},$$

$$\text{but } \left(1 + \frac{||z||^2}{\mathbb{E}(r_1^2)}\right)^{-1} = \left(1 + \frac{\left|\left|\Omega^{-\frac{1}{2}}(y-\mu)\right|\right|^2}{\mathbb{E}(r_1^2)}\right)^{-1} = \mathcal{D}_M(y|\mathcal{Y}).$$

So $\mathcal{D}_M(x|\mathcal{X})$ and $\mathcal{D}_M(y|\mathcal{Y})$ are identical distributed. □

Proof of the ***Corollary*** 1:

□ For $\mathcal{D}_M(x|\mathcal{X}) \overset{\mathrm{d}}{=} \mathcal{D}_M(y|\mathcal{Y})$ holds for any $\varepsilon > 0$:

$$\mathbb{P}\big[\mathcal{D}_M(y|\mathcal{Y}) < \varepsilon\big] = \mathbb{P}\big[\mathcal{D}_M(x|\mathcal{X}) < \varepsilon\big] = \mathbb{P}\Big[\left(1 + \frac{||x||^2}{\mathbb{E}(r_1^2)}\right)^{-1} < \varepsilon\Big] = \mathbb{P}\Big[\left(\frac{1}{\varepsilon} - 1\right)\mathbb{E}(r_1^2) < ||x||^2\Big]$$

$$= \mathbb{P}\big[\underbrace{\left(\frac{1}{\varepsilon} - 1\right)\mathbb{E}(r_1^2)}_{C_\varepsilon \geq 0} < r_d^2\big] = 1 - \mathbb{P}\big[r_d^2 \leq C_\varepsilon\big] = 1 - F_{r_d^2}(C_\varepsilon).$$

Here $F_{r_d^2}(\cdot)$ denotes the cdf of r_d^2. Accordingly to Proposition 3 holds for $d_1 < d_2$ and $c \geq 0$:

$$1 \geq F_{r_{d_1}^2}(c) = \mathbb{P}\big[r_{d_1}^2 \leq c\big] = \mathbb{P}\big[r_{d_2}^2 b^2 \leq c\big] = \int_0^1 \mathbb{P}\big[r_{d_2}^2 b^2 \leq c \mid b^2 = t\big] f_{b^2}(t)\,\mathrm{d}t$$

$$\underset{0 \leq b^2 \leq 1}{\geq} \int_0^1 \mathbb{P}\big[r_{d_2}^2 \leq c \mid b^2 = t\big] f_{b^2}(t)\,\mathrm{d}t \underset{b^2 \perp r_{d_2}^2}{=} \int_0^1 \mathbb{P}\big[r_{d_2}^2 \leq c\big] f_{b^2}(t)\,\mathrm{d}t$$

$$= \mathbb{P}\big[r_{d_2}^2 \leq c\big] \underbrace{\int_0^1 f_{b^2}(t)\,\mathrm{d}t}_{=1} = F_{r_{d_2}^2}(c) \geq 0.$$

Therefore, $F_{r_{d_1}^2}(c) \geq F_{r_{d_2}^2}(c)$ and therefore $\lim_{d\to\infty} F_{r_d^2}(c) = a \geq 0$ for all $c \geq 0$, because $F_{r_d}(c)$ is monotone decreasing in d and bounded by 0. Assume $a \neq 0$, then the following holds for any $c \geq 0$ and $0 < p \leq 1$:

$$F_{r_{d_1}^2}(cp) \underset{\text{s.a.}}{=} \int_0^1 \mathbb{P}\big[r_{d_2}^2 b^2 \leq cp \mid b^2 = t\big] f_{b^2}(t)\,\mathrm{d}t \geq \int_0^p \mathbb{P}\big[r_{d_2}^2 \frac{b^2}{p} \leq c \mid b^2 = t\big] f_{b^2}(t)\,\mathrm{d}t$$

$$\underset{0 \leq b^2 \leq p}{\geq} \int_0^p \mathbb{P}\big[r_{d_2}^2 \leq c \mid b^2 = t\big] f_{b^2}(t)\,\mathrm{d}t \underset{\text{s.a.}}{=} \underbrace{\mathbb{P}\big[r_{d_2}^2 \leq c\big]}_{=F_{r_{d_2}^2}(c) \geq a} \underbrace{\int_0^p f_{b^2}(t)\,\mathrm{d}t}_{=F_{b^2}(p)} \geq a F_{b^2}(p).$$

Because, for $b^2 \sim \mathrm{Beta}\left(\frac{d_1}{2}, \frac{d_2 - d_1}{2}\right)$ $\overset{d_2\to\infty}{\longrightarrow}$ $\mathrm{Beta}\left(\frac{d_1}{2}, \infty\right)$ holds $F_{b^2}(p)$ $\overset{d_2\to\infty}{\longrightarrow}$ 1 for any $p > 0$,

$$F_{r_d^2}(cp) \geq a F_{b^2}(p) \overset{d\to\infty}{\longrightarrow} a$$

is true for any $c \geq 0$. Therefore we get

$$0 = F_{r_d^2}(0) = \lim_{p \to 0} F_{r_d^2}(cp) \geq a F_{b^2}(p) \overset{d \to \infty}{\longrightarrow} a \Rightarrow \lim_{d \to \infty} F_{r_d^2}(c) = 0 \qquad (*)$$

for any $c \geq 0$ if $F_{r_d^2}(\cdot)$ is continuous, so a has to be 0. Therefore it holds:

$$\begin{aligned} \mathbb{P}[\mathcal{D}_M(y|\mathcal{Y}) < \varepsilon] = \mathbb{P}[\mathcal{D}_M(x|\mathcal{X}) < \varepsilon] = 1 - F_{r_d^2}(C_\varepsilon) \overset{d \to \infty}{\longrightarrow} 1 \\ \Rightarrow \operatorname*{plim}_{d \to \infty} \mathcal{D}_M(y|\mathcal{Y}) = 0 = \operatorname*{plim}_{d \to \infty} \mathcal{D}_M(x|\mathcal{X}). \end{aligned}$$

This shows the one part of corollary 1. To see the other part we derive a upper bound $U_{y|\mathcal{X}}$ for $\mathcal{D}_M(y|\mathcal{X})$. Let therefore $z \sim S_d(\Gamma)$ be the corresponding r.v. to a randomly drawn y from $\mathcal{Y}$ and eigenvalues $\varphi_1, \ldots, \varphi_d$ of Ω with $\varphi_d^+ = \max\{\varphi_k\}_{k=1}^d$ and $\varphi_d^- = \min\{\varphi_k\}_{k=1}^d$:

$$\begin{aligned} \mathcal{D}_M(y|\mathcal{X}) &= \left(1 + \frac{||y||^2}{\mathbb{E}(r_1^2)}\right)^{-1} = \left(1 + \frac{\left|\left|\Omega^{\frac{1}{2}} z - (-\mu)\right|\right|^2}{\mathbb{E}(r_1^2)}\right)^{-1} \\ &\leq \left(1 + \frac{\left|\left|\left|Q^\top \sqrt{D} Q z\right|\right| - ||-\mu||\right|^2}{\mathbb{E}(r_1^2)}\right)^{-1} = \left(1 + \frac{\left(\left|\left|\sqrt{D} Q z\right|\right| - ||\mu||\right)^2}{\mathbb{E}(r_1^2)}\right)^{-1} \\ &\leq \left(1 + \frac{\left(\sqrt{\varphi_d^-}\,||z|| - ||\mu||\right)^2}{\mathbb{E}(r_1^2)}\right)^{-1} = U_{y|\mathcal{X}}. \end{aligned}$$

$$\begin{aligned} &\Rightarrow \mathbb{P}[\mathcal{D}_M(y|\mathcal{X}) < \varepsilon] \geq \mathbb{P}[U_{y|\mathcal{X}} < \varepsilon] = \mathbb{P}\left[\left(1 + \frac{\left(\sqrt{\varphi_d^-}\,||z|| - ||\mu||\right)^2}{\mathbb{E}(r_1^2)}\right)^{-1} < \varepsilon\right] \\ &= \mathbb{P}[\underbrace{\left(\frac{1}{\varepsilon} - 1\right)\mathbb{E}\left(r_1^2\right)}_{C_\varepsilon} < \left(\sqrt{\varphi_d^-}\,||z|| - ||\mu||\right)^2] \\ &= \mathbb{P}[\frac{C_\varepsilon}{\varphi_d^-} < \left(||z|| - \frac{||\mu||}{\sqrt{\varphi_d^-}}\right)^2] = \mathbb{P}[\frac{C_\varepsilon}{\varphi_d^-} < \left(r_d - \frac{||\mu||}{\sqrt{\varphi_d^-}}\right)^2] = \mathbb{P}[\sqrt{\frac{C_\varepsilon}{\varphi_d^-}} < \left|r_d - \frac{||\mu||}{\sqrt{\varphi_d^-}}\right|] \\ &= \mathbb{P}[\frac{||\mu|| + \sqrt{C_\varepsilon}}{\sqrt{\varphi_d^-}} < r_d] + \mathbb{I}_{(\sqrt{C_\varepsilon} < ||\mu||)} \mathbb{P}[r_d < \frac{||\mu|| - \sqrt{C_\varepsilon}}{\sqrt{\varphi_d^-}}] \\ &= 1 - \mathbb{P}[r_d^2 < \frac{\left(||\mu|| + \sqrt{C_\varepsilon}\right)^2}{\varphi_d^-}] + \mathbb{I}_{(\sqrt{C_\varepsilon} < ||\mu||)} \mathbb{P}[r_d^2 < \frac{\left(||\mu|| - \sqrt{C_\varepsilon}\right)^2}{\varphi_d^-}] \\ &= 1 - F_{r_d^2}\left(\frac{\left(||\mu|| + \sqrt{C_\varepsilon}\right)^2}{\varphi_d^-}\right) + \mathbb{I}_{(\sqrt{C_\varepsilon} < ||\mu||)} F_{r_d^2}\left(\frac{\left(||\mu|| - \sqrt{C_\varepsilon}\right)^2}{\varphi_d^-}\right). \end{aligned}$$

Now we derive a upper bound $U_{x|\mathcal{Y}}$ for $\mathcal{D}_M(x|\mathcal{Y})$:

$$\mathcal{D}_M(x|\mathcal{Y}) = \left(1 + \frac{1}{\mathbb{E}(r_1^2)}\left(||x-\mu||^2 + \left|\left|\sqrt{(\mathcal{I}_d - D)D^{-1}}Q(x-\mu)\right|\right|^2\right)\right)^{-1}$$
$$\leq \left(1 + \frac{1}{\mathbb{E}(r_1^2)}\left(||x-\mu||^2 + \left|\left|\sqrt{\left(1-\varphi_d^+\right)\varphi_d^{+-1}}\mathcal{I}_d Q(x-\mu)\right|\right|^2\right)\right)^{-1}$$
$$\leq \left(1 + \frac{1}{\mathbb{E}(r_1^2)}\left(1 + \left(1-\varphi_d^+\right)\varphi_d^{+-1}\right)(||x|| - ||\mu||)^2\right)^{-1} = \left(1 + \frac{(||x|| - ||\mu||)^2}{\varphi_d^+ \mathbb{E}(r_1^2)}\right)^{-1} = U_{x|\mathcal{Y}}.$$

$$\Rightarrow \mathbb{P}[\mathcal{D}_M(x|\mathcal{Y}) < \varepsilon] \geq \mathbb{P}[U_{x|\mathcal{Y}} < \varepsilon] = \mathbb{P}[\left(1 + \frac{(||x|| - ||\mu||)^2}{\varphi_d^+ \mathbb{E}(r_1^2)}\right)^{-1} < \varepsilon]$$
$$= \mathbb{P}[\sqrt{C_\varepsilon \varphi_d^+} < |r_d - ||\mu|||]$$
$$= \mathbb{P}[||\mu|| + \sqrt{C_\varepsilon \varphi_d^+} < r_d] + \mathbb{I}_{\left(\sqrt{C_\varepsilon \varphi_d^+} < ||\mu||\right)}\mathbb{P}[r_d < ||\mu|| - \sqrt{C_\varepsilon \varphi_d^+}]$$
$$= 1 - \mathbb{P}[r_d^2 < \left(||\mu|| + \sqrt{C_\varepsilon \varphi_d^+}\right)^2] + \mathbb{I}_{\left(\sqrt{C_\varepsilon \varphi_d^+} < ||\mu||\right)}\mathbb{P}[r_d^2 < \left(||\mu|| - \sqrt{C_\varepsilon \varphi_d^+}\right)^2]$$
$$= 1 - F_{r_d^2}\left(\left(||\mu|| + \sqrt{C_\varepsilon \varphi_d^+}\right)^2\right) + \mathbb{I}_{\left(\sqrt{C_\varepsilon \varphi_d^+} < ||\mu||\right)}F_{r_d^2}\left(\left(||\mu|| - \sqrt{C_\varepsilon \varphi_d^+}\right)^2\right).$$

Let now $0 < \varphi_\infty^- = \inf_{d\in\mathbb{N}} \varphi_d^- \leq 1$ and $0 < \varphi_\infty^+ = \sup_{d\in\mathbb{N}} \varphi_d^+ \leq 1$ for $d \longrightarrow \infty$:

- $\lim_{d\to\infty} ||\mu|| > \sqrt{C_\varepsilon} \geq \sqrt{C_\varepsilon \varphi_\infty^+}$

$$\Rightarrow \mathbb{I}_{\left(\sqrt{C_\varepsilon \varphi_d^+} < ||\mu||\right)} \overset{d\to\infty}{\longrightarrow} \mathbb{I}_{\left(\sqrt{C_\varepsilon \varphi_\infty^+} < ||\mu||\right)} \geq \mathbb{I}_{\left(\sqrt{C_\varepsilon} < ||\mu||\right)} \overset{d\to\infty}{\longrightarrow} 1$$

$$\Rightarrow \lim_{d\to\infty} F_{r_d^2}\left(\frac{\left(||\mu|| + \sqrt{C_\epsilon}\right)^2}{\varphi_d^-}\right) = \lim_{d\to\infty} F_{r_d^2}\left(\frac{\left(||\mu|| - \sqrt{C_\epsilon}\right)^2}{\varphi_d^-}\right)$$
$$= \lim_{d\to\infty} F_{r_d^2}\left(\left(||\mu|| + \sqrt{C_\varepsilon \varphi_d^+}\right)^2\right) = \lim_{d\to\infty} F_{r_d^2}\left(\left(||\mu|| - \sqrt{C_\varepsilon \varphi_d^+}\right)^2\right) = 0 \,(\text{see } (*)).$$

Therefore we get:

$$\mathbb{P}[\mathcal{D}_M(y|\mathcal{X}) < \varepsilon], \mathbb{P}[\mathcal{D}_M(x|\mathcal{Y}) < \varepsilon] \overset{d\to\infty}{\longrightarrow} 1 \Rightarrow \operatorname*{plim}_{d\to\infty} \mathcal{D}_M(y|\mathcal{X}) = 0 = \operatorname*{plim}_{d\to\infty} \mathcal{D}_M(x|\mathcal{Y}).$$

- $\sqrt{C_\varepsilon} \geq \lim_{d\to\infty} ||\mu|| \geq \sqrt{C_\varepsilon \varphi_\infty^+}$

$$\Rightarrow \mathbb{I}_{\left(\sqrt{C_\varepsilon \varphi_d^+} < ||\mu||\right)} \overset{d\to\infty}{\longrightarrow} \mathbb{I}_{\left(\sqrt{C_\varepsilon \varphi_\infty^+} < ||\mu||\right)} \overset{d\to\infty}{\longrightarrow} 1 \text{ and } \mathbb{I}_{\left(\sqrt{C_\varepsilon} < ||\mu||\right)} \overset{d\to\infty}{\longrightarrow} 0$$

$$\Rightarrow \lim_{d\to\infty} F_{r_d^2}\left(\frac{\left(||\mu|| + \sqrt{C_\epsilon}\right)^2}{\varphi_d^-}\right)$$
$$= \lim_{d\to\infty} F_{r_d^2}\left(\left(||\mu|| + \sqrt{C_\varepsilon\varphi_d^+}\right)^2\right) = \lim_{d\to\infty} F_{r_d^2}\left(\left(||\mu|| - \sqrt{C_\varepsilon\varphi_d^+}\right)^2\right) = 0\,(\text{see } (*))\,.$$

Therefore we get:

$$\mathbb{P}[\mathcal{D}_M(y|\mathcal{X}) < \varepsilon], \mathbb{P}[\mathcal{D}_M(x|\mathcal{Y}) < \varepsilon] \overset{d\to\infty}{\longrightarrow} 1 \Rightarrow \operatorname*{plim}_{d\to\infty} \mathcal{D}_M(y|\mathcal{X}) = 0 = \operatorname*{plim}_{d\to\infty} \mathcal{D}_M(x|\mathcal{Y})\,.$$

- $\sqrt{C_\varepsilon\varphi_\infty^+} > \lim_{d\to\infty} ||\mu||$

$$\Rightarrow \mathbb{I}_{\left(\sqrt{C_\varepsilon\varphi_d^+} < ||\mu||\right)} \overset{d\to\infty}{\longrightarrow} \mathbb{I}_{\left(\sqrt{C_\varepsilon\varphi_\infty^+} < ||\mu||\right)} \overset{d\to\infty}{\longrightarrow} 0 \text{ and } \mathbb{I}_{\left(\sqrt{C_\varepsilon} < ||\mu||\right)} \overset{d\to\infty}{\longrightarrow} 0$$

$$\Rightarrow \lim_{d\to\infty} F_{r_d^2}\left(\frac{\left(||\mu|| + \sqrt{C_\epsilon}\right)^2}{\varphi_d^-}\right) = \lim_{d\to\infty} F_{r_d^2}\left(\left(||\mu|| + \sqrt{C_\varepsilon\varphi_d^+}\right)^2\right) = 0\,(\text{see } (*))\,.$$

Overall we get for any z randomly drawn from $\mathcal{X}$ or $\mathcal{Y}$:

$$\operatorname*{plim}_{d\to\infty} \mathcal{D}_M(z|\mathcal{X}) = 0 = \operatorname*{plim}_{d\to\infty} \mathcal{D}_M(z|\mathcal{Y})\,. \quad \square$$

Proof of the ***Corollary*** 2:

$\square$ Now we derive a lower bound $L_{x|\mathcal{Y}}$ for $\mathcal{D}_M(x|\mathcal{Y})$:

$$\mathcal{D}_M(x|\mathcal{Y}) = \left(1 + \frac{1}{\mathbb{E}(r_1^2)}\left(||x-\mu||^2 + \left|\left|\sqrt{(\mathcal{I}_d - D)D^{-1}}Q(x-\mu)\right|\right|^2\right)\right)^{-1}$$
$$\geq \left(1 + \frac{1}{\mathbb{E}(r_1^2)}\left(||x-\mu||^2 + \left|\left|\sqrt{\left(1-\varphi_d^-\right)\varphi_d^{-\,-1}}\mathcal{I}_d Q(x-\mu)\right|\right|^2\right)\right)^{-1}$$
$$= \left(1 + \frac{1}{\mathbb{E}(r_1^2)}\left(1 + \left(1-\varphi_d^-\right)\varphi_d^{-\,-1}\right)||x-\mu||^2\right)^{-1} = \left(1 + \frac{||x-\mu||^2}{\varphi_d^-\mathbb{E}(r_1^2)}\right)^{-1}$$
$$\geq \left(1 + \frac{(||x|| + ||\mu||)^2}{\varphi_d^-\mathbb{E}(r_1^2)}\right)^{-1} = L_{x|\mathcal{Y}}.$$

This shows the first part of corollary 2. To see the second part we use our results from the proof of corollary 1. Let therefore be $0 < \alpha < 1$, then in any dimension holds:

$$\mathbb{P}[U_{x|\mathcal{Y}} < \alpha] \leq \mathbb{P}[\mathcal{D}_M(x|\mathcal{Y}) < \alpha] \leq \mathbb{P}[L_{x|\mathcal{Y}} < \alpha] \text{ with}$$

$$\mathbb{P}[U_{x|\mathcal{Y}} < \alpha] = 1 - F_{r_d^2}\left(\left(||\mu|| + \sqrt{C_\alpha \varphi_d^+}\right)^2\right) + \mathbb{I}_{\left(\sqrt{C_\alpha \varphi_d^+} < ||\mu||\right)} F_{r_d^2}\left(\left(||\mu|| - \sqrt{C_\alpha \varphi_d^+}\right)^2\right).$$

$$\begin{aligned}
\mathbb{P}[L_{x|\mathcal{Y}} < \alpha] &= \mathbb{P}[\left(1 + \frac{\left(||x|| + ||\mu||\right)^2}{\varphi_d^- \mathbb{E}\left(r_1^2\right)}\right)^{-1} < \alpha] \\
&= \mathbb{P}[\underbrace{\left(\frac{1}{\alpha} - 1\right)\mathbb{E}\left(r_1^2\right)}_{C_\alpha} < \frac{\left(||x|| + ||\mu||\right)^2}{\varphi_d^-}] = \mathbb{P}[\sqrt{C_\alpha \varphi_d^-} < ||x|| + ||\mu||] \\
&= 1 - \mathbb{I}_{\left(\sqrt{C_\alpha \varphi_d^-} \geq ||\mu||\right)} F_{r_d^2}\left(\left(\sqrt{C_\alpha \varphi_d^-} - ||\mu||\right)^2\right).
\end{aligned}$$

We rewrite these conditions on $||\mu||$:

$$\sqrt{C_\alpha \varphi_d^-} \geq ||\mu|| \Leftrightarrow \left(\frac{1}{\alpha} - 1\right)\mathbb{E}\left(r_1^2\right)\varphi_d^- \geq ||\mu||^2 \Leftrightarrow \frac{\mathbb{E}\left(r_1^2\right)}{\mathbb{E}\left(r_1^2\right) + \frac{||\mu||^2}{\varphi_d^-}} \geq \alpha \qquad \text{(I)}$$

$$\sqrt{C_\alpha \varphi_d^+} < ||\mu|| \Leftrightarrow \left(\frac{1}{\alpha} - 1\right)\mathbb{E}\left(r_1^2\right)\varphi_d^+ < ||\mu||^2 \Leftrightarrow \frac{\mathbb{E}\left(r_1^2\right)}{\mathbb{E}\left(r_1^2\right) + \frac{||\mu||^2}{\varphi_d^+}} < \alpha \qquad \text{(II)}$$

We look now on two different cases:

(a) $\lim_{d\to\infty} ||\mu|| = \infty$:
Then for any $0 < \alpha < 1$ we can find dimension D such that for any $d \geq D$ holds:

$$\alpha > \frac{\mathbb{E}\left(r_1^2\right)}{\mathbb{E}\left(r_1^2\right) + \frac{||\mu||^2}{\varphi_d^+}} \geq \frac{\mathbb{E}\left(r_1^2\right)}{\mathbb{E}\left(r_1^2\right) + \frac{||\mu||^2}{\varphi_d^-}}$$

so (I) is false and (II) is true for any $d \geq D$ and we get:

$$\begin{aligned}
&\mathbb{P}[\mathcal{D}_M\left(x|\mathcal{Y}\right) < \alpha] \geq 1 - \left(F_{r_d^2}\left(\left(||\mu|| + \sqrt{C_\alpha \varphi_d^+}\right)^2\right) - F_{r_d^2}\left(\left(||\mu|| - \sqrt{C_\alpha \varphi_d^+}\right)^2\right)\right) \\
\Leftrightarrow\ &\mathbb{P}[\mathcal{D}_M\left(x|\mathcal{Y}\right) < \alpha] \geq 1 - \mathbb{P}[\left(||\mu|| - \sqrt{C_\alpha \varphi_d^+}\right)^2 \leq\ r_d^2\ \leq \left(||\mu|| + \sqrt{C_\alpha \varphi_d^+}\right)^2].
\end{aligned}$$

$\lim_{d\to\infty} \mathbb{P}[\left(||\mu|| - \sqrt{C_\alpha \varphi_d^+}\right)^2 \leq\ r_d^2\ \leq \left(||\mu|| + \sqrt{C_\alpha \varphi_d^+}\right)^2] = 0$ because $\lim_{d\to\infty} ||\mu|| = \infty$, so we can find another $d^* \geq D$ such that for all $d \geq d^*$ holds:

$$\begin{aligned}
&\mathbb{P}[\left(||\mu|| - \sqrt{C_\alpha \varphi_d^+}\right)^2 \leq\ r_d^2\ \leq \left(||\mu|| + \sqrt{C_\alpha \varphi_d^+}\right)^2] < \mathbb{P}[r_d^2 \leq C_\alpha] \\
\Leftrightarrow\ &1 - \mathbb{P}[\left(||\mu|| - \sqrt{C_\alpha \varphi_d^+}\right)^2 \leq\ r_d^2\ \leq \left(||\mu|| + \sqrt{C_\alpha \varphi_d^+}\right)^2] > 1 - \mathbb{P}[r_d^2 \leq C_\alpha] \\
\Leftrightarrow\ &1 - \left(F_{r_d^2}\left(\left(||\mu|| + \sqrt{C_\alpha \varphi_d^+}\right)^2\right) - F_{r_d^2}\left(\left(||\mu|| - \sqrt{C_\alpha \varphi_d^+}\right)^2\right)\right) > 1 - F_{r_d^2}\left(C_\alpha\right) \\
&\Leftrightarrow\ \mathbb{P}[\mathcal{D}_M\left(x|\mathcal{Y}\right) < \alpha] > \mathbb{P}[\mathcal{D}_M\left(y|\mathcal{Y}\right) < \alpha].
\end{aligned}$$

(b) $\lim_{d\to\infty} ||\mu|| = 0$:
Then for any $0 < \alpha < 1$ we can find dimension D such that for any $d \geq D$ holds:

$$\alpha \leq \frac{\mathbb{E}\left(r_1^2\right)}{\mathbb{E}\left(r_1^2\right) + \frac{||\mu||^2}{\varphi_d^-}} \leq \frac{\mathbb{E}\left(r_1^2\right)}{\mathbb{E}\left(r_1^2\right) + \frac{||\mu||^2}{\varphi_d^+}}$$

so (I) is true and (II) is false for any $d \geq D$ and we get:

$$\mathbb{P}\big[\mathcal{D}_M\left(x|\mathcal{Y}\right) < \alpha\big] \geq 1 - F_{r_d^2}\left(\left(||\mu|| + \sqrt{C_\alpha \varphi_d^+}\right)^2\right)$$

$\lim_{d\to\infty} F_{r_d^2}\left(\left(||\mu|| + \sqrt{C_\alpha \varphi_d^+}\right)^2\right) = F_{r_d^2}\left(C_\alpha \varphi_\infty^+\right) \leq F_{r_d^2}\left(C_\alpha\right)$ because $\lim_{d\to\infty} ||\mu|| = 0$, so we can find another $d^* \geq D$ such that for all $d \geq d^*$ holds:

▷ If $\varphi_d^+ < 1$ for all $d \geq d^*$ and $F_{r_d^2}(\cdot)$ continuous:

$$F_{r_d^2}\left(\left(||\mu|| + \sqrt{C_\alpha \varphi_d^+}\right)^2\right) \leq F_{r_d^2}\left(C_\alpha\right)$$

$$\Rightarrow \mathbb{P}\big[\mathcal{D}_M\left(x|\mathcal{Y}\right) < \alpha\big] \geq 1 - F_{r_d^2}\left(\left(||\mu|| + \sqrt{C_\alpha \varphi_d^+}\right)^2\right) \geq 1 - F_{r_d^2}\left(C_\alpha\right) = \mathbb{P}\big[\mathcal{D}_M\left(y|\mathcal{Y}\right) < \alpha\big].$$

▷ If $\varphi_d^+ = 1$ for all $d \geq d^*$ and $F_{r_d^2}(\cdot)$ continuous:

$$F_{r_d^2}\left(\left(||\mu|| + \sqrt{C_\alpha \varphi_d^+}\right)^2\right) = F_{r_d^2}\left(\left(||\mu|| + \sqrt{C_\alpha}\right)^2\right) \geq F_{r_d^2}\left(C_\alpha\right),$$

so no conclusion is possible. □

Proof of the ***Corollary*** 3:

□ Now we derive a lower bound $L_{y|\mathcal{X}}$ for $\mathcal{D}_M\left(y|\mathcal{X}\right)$. Let $z \overset{\mathrm{d}}{=} x \sim S_d\left(\Gamma\right)$ be the corresponding r.v. to the drawn y from $\mathcal{Y}$. Then the following holds:

$$\begin{aligned}
\mathcal{D}_M\left(y|\mathcal{X}\right) &= \left(1 + \frac{||y||^2}{\mathbb{E}\left(r_1^2\right)}\right)^{-1} = \left(1 + \frac{\left|\left|\Omega^{\frac{1}{2}} z + \mu\right|\right|^2}{\mathbb{E}\left(r_1^2\right)}\right)^{-1} \\
&\geq \left(1 + \frac{\left(\left|\left|Q^\top \sqrt{D} Q z\right|\right| + ||\mu||\right)^2}{\mathbb{E}\left(r_1^2\right)}\right)^{-1} = \left(1 + \frac{\left(\left|\left|\sqrt{D} Q z\right|\right| + ||\mu||\right)^2}{\mathbb{E}\left(r_1^2\right)}\right)^{-1} \\
&\geq \left(1 + \frac{\left(\sqrt{\varphi_d^+}\,||z|| + ||\mu||\right)^2}{\mathbb{E}\left(r_1^2\right)}\right)^{-1} = L_{y|\mathcal{X}}.
\end{aligned}$$

This shows the first part of corollary 3. To see the second part we use our results from the

proof of corollary 1. Let therefore be $0 < \alpha < 1$, then in any dimension holds:

$$\mathbb{P}[U_{y|\mathcal{X}} < \alpha] \leq \mathbb{P}[\mathcal{D}_M(y|\mathcal{X}) < \alpha] \leq \mathbb{P}[L_{y|\mathcal{X}} < \alpha] \text{ with}$$

$$\mathbb{P}[U_{y|\mathcal{X}} < \alpha] = 1 - F_{r_d^2}\left(\frac{\left(||\mu|| + \sqrt{C_\alpha}\right)^2}{\varphi_d^-}\right) + \mathbb{I}_{\left(\sqrt{C_\alpha} < ||\mu||\right)} F_{r_d^2}\left(\frac{\left(||\mu|| - \sqrt{C_\alpha}\right)^2}{\varphi_d^-}\right).$$

$$\mathbb{P}[L_{y|\mathcal{X}} < \alpha] = \mathbb{P}[\left(1 + \frac{\left(\sqrt{\varphi_d^+}\,||z|| + ||\mu||\right)^2}{\mathbb{E}\left(r_1^2\right)}\right)^{-1} < \alpha]$$

$$= \mathbb{P}[\underbrace{\left(\frac{1}{\alpha} - 1\right)\mathbb{E}\left(r_1^2\right)}_{C_\alpha} < \left(\sqrt{\varphi_d^+}\,||z|| + ||\mu||\right)^2]$$

$$= \mathbb{P}[\frac{C_\alpha}{\varphi_d^+} < \left(||z|| + \frac{||\mu||}{\sqrt{\varphi_d^+}}\right)^2] = \mathbb{P}[\frac{C_\alpha}{\varphi_d^+} < \left(r_d + \frac{||\mu||}{\sqrt{\varphi_d^+}}\right)^2]$$

$$= \mathbb{P}[\sqrt{\frac{C_\alpha}{\varphi_d^+}} < r_d + \frac{||\mu||}{\sqrt{\varphi_d^+}}] = \mathbb{I}_{\left(\sqrt{C_\alpha} < ||\mu||\right)} + \mathbb{I}_{\left(\sqrt{C_\alpha} \geq ||\mu||\right)} \mathbb{P}[\frac{\left(\sqrt{C_\alpha} - ||\mu||\right)^2}{\varphi_d^+} < r_d^2]$$

$$= \mathbb{I}_{\left(\sqrt{C_\alpha} < ||\mu||\right)} + \mathbb{I}_{\left(\sqrt{C_\alpha} \geq ||\mu||\right)} \left(1 - \mathbb{P}[\frac{\left(\sqrt{C_\alpha} - ||\mu||\right)^2}{\varphi_d^+} \geq r_d^2]\right)$$

$$= 1 - \mathbb{I}_{\left(\sqrt{C_\alpha} \geq ||\mu||\right)} \mathbb{P}[r_d^2 \leq \frac{\left(\sqrt{C_\alpha} - ||\mu||\right)^2}{\varphi_d^+}] = 1 - \mathbb{I}_{\left(\sqrt{C_\alpha} \geq ||\mu||\right)} F_{r_d^2}\left(\frac{\left(\sqrt{C_\alpha} - ||\mu||\right)^2}{\varphi_d^+}\right).$$

We rewrite the conditions on $||\mu||$:

$$\sqrt{C_\alpha} \geq ||\mu|| \Leftrightarrow \left(\frac{1}{\alpha} - 1\right)\mathbb{E}\left(r_1^2\right) \geq ||\mu||^2 \Leftrightarrow \frac{\mathbb{E}\left(r_1^2\right)}{\mathbb{E}\left(r_1^2\right) + ||\mu||^2} \geq \alpha \qquad \text{(I)}$$

$$\sqrt{C_\alpha} < ||\mu|| \Leftrightarrow \left(\frac{1}{\alpha} - 1\right)\mathbb{E}\left(r_1^2\right) < ||\mu||^2 \Leftrightarrow \frac{\mathbb{E}\left(r_1^2\right)}{\mathbb{E}\left(r_1^2\right) + ||\mu||^2} < \alpha \qquad \text{(II)}$$

We look now again on two different cases:

(a) $\lim_{d\to\infty} ||\mu|| = \infty$:
Then for any $0 < \alpha < 1$ we can find dimension D such that for any $d \geq D$ holds:

$$\alpha > \frac{\mathbb{E}\left(r_1^2\right)}{\mathbb{E}\left(r_1^2\right) + ||\mu||^2}$$

so (I) is false and (II) is true for any $d \geq D$ and we get:

$$\mathbb{P}\big[\mathcal{D}_M\left(y|\mathcal{X}\right) < \alpha\big] \geq 1 - \left(F_{r_d^2}\left(\frac{\left(||\mu|| + \sqrt{C_\alpha}\right)^2}{\varphi_d^-}\right) - F_{r_d^2}\left(\frac{\left(||\mu|| - \sqrt{C_\alpha}\right)^2}{\varphi_d^-}\right)\right)$$

$$\Leftrightarrow \mathbb{P}\big[\mathcal{D}_M\left(y|\mathcal{X}\right) < \alpha\big] \geq 1 - \mathbb{P}\big[\frac{\left(||\mu|| - \sqrt{C_\alpha}\right)^2}{\varphi_d^-} \leq \ r_d^2 \ \leq \frac{\left(||\mu|| + \sqrt{C_\alpha}\right)^2}{\varphi_d^-}\big].$$

$\lim\limits_{d\to\infty} \mathbb{P}\big[\frac{\left(||\mu|| - \sqrt{C_\alpha}\right)^2}{\varphi_d^-} \leq \ r_d^2 \ \leq \frac{\left(||\mu|| + \sqrt{C_\alpha}\right)^2}{\varphi_d^-}\big] = 0$ because $\lim\limits_{d\to\infty} = \infty$, so we can find another $d^* \geq D$ such that for all $d \geq d^*$ holds:

$$\mathbb{P}\big[\frac{\left(||\mu|| - \sqrt{C_\alpha}\right)^2}{\varphi_d^-} \leq \ r_d^2 \ \leq \frac{\left(||\mu|| + \sqrt{C_\alpha}\right)^2}{\varphi_d^-}\big] < \mathbb{P}\big[r_d^2 \leq C_\alpha\big]$$

$$\Leftrightarrow 1 - \mathbb{P}\big[\frac{\left(||\mu|| - \sqrt{C_\alpha}\right)^2}{\varphi_d^-} \leq \ r_d^2 \ \leq \frac{\left(||\mu|| + \sqrt{C_\alpha}\right)^2}{\varphi_d^-}\big] > 1 - \mathbb{P}\big[r_d^2 \leq C_\alpha\big]$$

$$\Leftrightarrow 1 - \left(F_{r_d^2}\left(\frac{\left(||\mu|| + \sqrt{C_\alpha}\right)^2}{\varphi_d^-}\right) - F_{r_d^2}\left(\frac{\left(||\mu|| - \sqrt{C_\alpha}\right)^2}{\varphi_d^-}\right)\right) > 1 - F_{r_d^2}\left(C_\alpha\right)$$

$$\Leftrightarrow \mathbb{P}\big[\mathcal{D}_M\left(y|\mathcal{X}\right)\big] > \mathbb{P}\big[\mathcal{D}_M\left(x|\mathcal{X}\right)\big].$$

(b) $\lim\limits_{d\to\infty} ||\mu|| = 0$:
Then for any $0 < \alpha < 1$ we can find dimension D such that for any $d \geq D$ holds:

$$\alpha \leq \frac{\mathbb{E}\left(r_1^2\right)}{\mathbb{E}\left(r_1^2\right) + ||\mu||^2}$$

so (I) is true and (II) is false for any $d \geq D$ and we get:

$$\mathbb{P}\big[\mathcal{D}_M\left(x|\mathcal{Y}\right) < \alpha\big] \leq 1 - F_{r_d^2}\left(\frac{\left(\sqrt{C_\alpha} - ||\mu||\right)^2}{\varphi_d^+}\right)$$

$\lim\limits_{d\to\infty} F_{r_d^2}\left(\frac{\left(\sqrt{C_\alpha} - ||\mu||\right)^2}{\varphi_d^+}\right) = F_{r_d^2}\left(\frac{C_\alpha}{\varphi_d^+}\right) \geq F_{r_d^2}\left(C_\alpha\right)$ because $\lim\limits_{d\to\infty} ||\mu|| = 0$, so we can find another $d^* \geq D$ such that for all $d \geq d^*$ holds:

▷ If $\varphi_d^+ < 1$ for all $d \geq d^*$ and $F_{r_d^2}\left(\cdot\right)$ continuous: $F_{r_d^2}\left(\frac{\left(\sqrt{C_\alpha} - ||\mu||\right)^2}{\varphi_d^+}\right) \geq F_{r_d^2}\left(C_\alpha\right)$

$$\Rightarrow \mathbb{P}\big[\mathcal{D}_M\left(x|\mathcal{Y}\right) < \alpha\big] \leq 1 - F_{r_d^2}\left(\frac{\left(\sqrt{C_\alpha} - ||\mu||\right)^2}{\varphi_d^+}\right) \leq 1 - F_{r_d^2}\left(C_\alpha\right) = \mathbb{P}\big[\mathcal{D}_M\left(y|\mathcal{Y}\right) < \alpha\big].$$

▷ If $\varphi_d^+ = 1$ for all $d \geq d^*$ and $F_{r_d^2}\left(\cdot\right)$ continuous:

$$F_{r_d^2}\left(\frac{\left(\sqrt{C_\alpha} - ||\mu||\right)^2}{\varphi_d^+}\right) = F_{r_d^2}\left(\left(||\mu|| - \sqrt{C_\alpha}\right)^2\right) \leq F_{r_d^2}\left(C_\alpha\right),$$

so no conclusion is possible. □

Proof of ***DGP II***:

□ We have to proof condition 2 for data nests because $n = 9900 >> 100 = m$, so condition 1 is already satisfied. For this enough to focus on only one dimension $d = 1$, so $z \sim \text{Beta}_{d=1}(p, q) = \text{Beta}(p, q)$, because all generated variables are independent of each other. It is true that

$$Z \sim \text{Beta}(p, q) \Rightarrow \ \mu_Z = \frac{p}{p+q} \text{ and } \sigma_Z^2 = \frac{pq}{(p+q+1)(p+q)^2}$$

$$\text{and therefore we get: } X \sim \text{Beta}(3, 3) \Rightarrow \ \mu_X = \frac{1}{2} = 0.5 \text{ and } \sigma_X^2 = \frac{9}{7 \cdot 6^2} = \frac{1}{28}.$$

Now we check for $p' = t(0.5 + \tau)$ and $q' = t(0.5 - \tau)$:

$$Y \sim \text{Beta}(p', q') \Rightarrow \mu_Y = \frac{p'}{p'+q'} = \frac{t(0.5+\tau)}{t(0.5+\tau) + t(0.5-\tau)} = 0.5 + \tau = \mu_X + \tau.$$

$$\text{So we get: } \sigma_Y^2 \leq \sigma_X^2 \Leftrightarrow \frac{p'q'}{(p'+q'+1)(p'+q')^2} \leq \frac{1}{28} \Leftrightarrow \frac{t^2(0.5+\tau)(0.5-\tau)}{(t+1)t^2} \leq \frac{1}{28}$$

$$0.25 - \tau^2 \leq \frac{1+t}{28} \Leftrightarrow 28\left(0.25 - \tau^2\right) - 1 \leq t \Leftrightarrow T \leq t.$$

So also the second condition for data nest holds because $t \sim U(T, 5T)$. □

A.2 Appendix of Chapter 3

Proof of the existence of rotation matrix R:

□ (3.2.4) and (3.2.8) are equal if $F\beta = f\gamma \Leftrightarrow F\frac{1}{\gamma}\beta = f$. Scale factor f such that

$$(\frac{1}{\gamma}\beta)^\top (\frac{1}{\gamma}\beta) = 1 \Leftrightarrow \gamma = \sqrt{\beta^\top \beta}.$$

Then one can find vectors $r_1, \ldots, r_{k-1}$ to construct the orthogonal matrix $R \in \mathbb{R}^{k \times k}$:

$$R = \begin{pmatrix} \frac{\beta}{\sqrt{\beta^\top \beta}} & r_1 & \ldots & r_{k-1} \end{pmatrix} \text{ such that } R^\top R = \mathcal{I}_k.$$

Now with $\begin{pmatrix} f & G \end{pmatrix} = FR$ and $\Lambda R = \begin{pmatrix} \lambda & \Phi \end{pmatrix}$ we get

$$X = F\Lambda^\top + U = FRR^\top \Lambda^\top + U = FR(\Lambda R)^\top + U = \begin{pmatrix} f & G \end{pmatrix} \begin{pmatrix} \lambda & \Phi \end{pmatrix}^\top + U \quad \square$$

Proof of the equivalence between (3.3.3) and (3.3.4):

□ We start from (3.3.3):

$$\begin{pmatrix} X & \tilde{y} \end{pmatrix} = X \begin{pmatrix} w & \Omega \end{pmatrix} \begin{pmatrix} \varphi & \Phi \\ \gamma & 0 \end{pmatrix}^{\top} + \begin{pmatrix} U & \varepsilon \end{pmatrix}$$

$$\Leftrightarrow \begin{pmatrix} x_1 & \cdots & x_n & \tilde{y} \end{pmatrix} = X \begin{pmatrix} w & \Omega \end{pmatrix} \begin{pmatrix} \varphi_1 & \cdots & \varphi_n & \gamma \\ \phi_{11} & \cdots & \phi_{1n} & 0 \\ \vdots & \ddots & \vdots & \vdots \\ \phi_{s1} & \cdots & \phi_{sn} & 0 \end{pmatrix} + \begin{pmatrix} u_1 & \cdots & u_n & \varepsilon \end{pmatrix}$$

$$\Leftrightarrow \tilde{y} = y - Z\alpha = X \begin{pmatrix} w & \Omega \end{pmatrix} \begin{pmatrix} \gamma \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \varepsilon \wedge \forall 1 \leq i \leq n : x_i = X \begin{pmatrix} w & \Omega \end{pmatrix} \begin{pmatrix} \varphi_i \\ \phi_{1i} \\ \vdots \\ \phi_{si} \end{pmatrix} \qquad (*)$$

Now we reorder $\begin{pmatrix} X & \tilde{y} \end{pmatrix}$ as in (3.3.4):

$$(*) \Leftrightarrow \begin{pmatrix} x_1 \\ \vdots \\ x_n \\ \tilde{y} \end{pmatrix} = \begin{pmatrix} X \begin{pmatrix} w & \Omega \end{pmatrix} \begin{pmatrix} \varphi_1 \\ \phi_{11} \\ \vdots \\ \phi_{s1} \end{pmatrix} \\ \vdots \\ X \begin{pmatrix} w & \Omega \end{pmatrix} \begin{pmatrix} \varphi_n \\ \phi_{1n} \\ \vdots \\ \phi_{sn} \end{pmatrix} \\ X \begin{pmatrix} w & \Omega \end{pmatrix} \begin{pmatrix} \gamma \\ 0 \\ \vdots \\ 0 \end{pmatrix} \end{pmatrix} + \begin{pmatrix} u_1 \\ \vdots \\ u_n \\ \varepsilon \end{pmatrix}$$

$$\Leftrightarrow \begin{pmatrix} x_1 \\ \vdots \\ x_n \\ \tilde{y} \end{pmatrix} = \begin{pmatrix} X\left(\varphi_1 w + \phi_{11}\omega_1 + \cdots + \phi_{s1}\omega_s\right) \\ \vdots \\ X\left(\varphi_n w + \phi_{1n}\omega_1 + \cdots + \phi_{sn}\omega_s\right) \\ X\left(\gamma w + 0\omega_1 + \cdots + 0\omega_s\right) \end{pmatrix} + \begin{pmatrix} u_1 \\ \vdots \\ u_n \\ \varepsilon \end{pmatrix}$$

$$\Leftrightarrow \begin{pmatrix} x_1 \\ \vdots \\ x_n \\ \tilde{y} \end{pmatrix} = \begin{pmatrix} \varphi_1 X & \phi_{11} X & \cdots & \phi_{s1} X \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_n X & \phi_{1n} X & \cdots & \phi_{sn} X \\ \gamma X & 0X & \cdots & 0X \end{pmatrix} \begin{pmatrix} w \\ \omega_1 \\ \vdots \\ \omega_s \end{pmatrix} + \begin{pmatrix} u_1 \\ \vdots \\ u_n \\ \varepsilon \end{pmatrix}$$

$$\Leftrightarrow \begin{pmatrix} x_1 \\ \vdots \\ x_n \\ \tilde{y} \end{pmatrix} = \left(\begin{pmatrix} \varphi & \Phi \\ \gamma & 0 \end{pmatrix} \otimes X \right) \begin{pmatrix} w \\ \omega_1 \\ \vdots \\ \omega_s \end{pmatrix} + \begin{pmatrix} u_1 \\ \vdots \\ u_n \\ \varepsilon \end{pmatrix} \quad \square$$

Proof of Lemma 7:

$\square$ Let X be substituted by $X^* := \rho X$ in (3.3.4). To show (I) we plug in X^* into (3.3.2) and use $\hat{w}$ and $\hat{\Omega}$ which are given:

$$\begin{pmatrix} \hat{f}^* & \hat{G}^* \end{pmatrix} := X^* \begin{pmatrix} \hat{w} & \hat{\Omega} \end{pmatrix} = \rho X \begin{pmatrix} \hat{w} & \hat{\Omega} \end{pmatrix} = \rho \begin{pmatrix} \hat{f} & \hat{G} \end{pmatrix}$$

Now we plug in these factors in (3.2.7) to compute the OLS-estimate of $\hat{\varphi}^*$ and $\hat{\Phi}^*$:

$$\begin{pmatrix} \hat{\varphi}^* & \hat{\Phi}^* \end{pmatrix}^\top = \left(\begin{pmatrix} \hat{f}^* & \hat{G}^* \end{pmatrix}^\top \begin{pmatrix} \hat{f}^* & \hat{G}^* \end{pmatrix} \right)^{-1} \begin{pmatrix} \hat{f}^* & \hat{G}^* \end{pmatrix} X^*$$

From there it follows:

$$\begin{aligned} \begin{pmatrix} \hat{\varphi}^* & \hat{\Phi}^* \end{pmatrix}^\top &= \left(\rho \begin{pmatrix} \hat{f} & \hat{G} \end{pmatrix}^\top \rho \begin{pmatrix} \hat{f} & \hat{G} \end{pmatrix} \right)^{-1} \rho \begin{pmatrix} \hat{f} & \hat{G} \end{pmatrix} \rho X \\ &= \left(\rho^2 \begin{pmatrix} \hat{f} & \hat{G} \end{pmatrix}^\top \begin{pmatrix} \hat{f} & \hat{G} \end{pmatrix} \right)^{-1} \rho^2 \begin{pmatrix} \hat{f} & \hat{G} \end{pmatrix} X \\ &= \rho^{-2} \left(\begin{pmatrix} \hat{f} & \hat{G} \end{pmatrix}^\top \begin{pmatrix} \hat{f} & \hat{G} \end{pmatrix} \right)^{-1} \rho^2 \begin{pmatrix} \hat{f} & \hat{G} \end{pmatrix} X \\ &= \left(\begin{pmatrix} \hat{f} & \hat{G} \end{pmatrix}^\top \begin{pmatrix} \hat{f} & \hat{G} \end{pmatrix} \right)^{-1} \begin{pmatrix} \hat{f} & \hat{G} \end{pmatrix} X = \begin{pmatrix} \hat{\varphi} & \hat{\Phi} \end{pmatrix}^\top \end{aligned}$$

So the OLS-estimate of φ and Φ does not depend of ρ. To see the second part of (I) we consider (3.2.8):

$$y = Z\alpha + f\gamma + \varepsilon = Z\alpha + \rho f \rho^{-1} \gamma + \varepsilon = Z\alpha + f^* \rho^{-1} \gamma + \varepsilon$$

Therefore, if the OLS-estimate of (3.2.8) is $\hat{\gamma}$, $\hat{\gamma}^* := \rho^{-1}\hat{\gamma}$ is the OLS-estimate of (3.2.8) where f is substituted by $f^* := \rho f$. To see (II) we use (3.3.4) substituting X by X^* and compute the OLS-estimate $\begin{pmatrix} \hat{w} & \hat{\omega}_1 & \dots & \hat{\omega}_s \end{pmatrix}^\top$:

$$\begin{pmatrix} \hat{w}^* \\ \hat{\omega}_1^* \\ \vdots \\ \hat{\omega}_s^* \end{pmatrix} = \left(\left(\begin{pmatrix} \hat{\varphi}^* & \hat{\Phi}^* \\ \hat{\gamma}^* & 0 \end{pmatrix} \otimes X^* \right)^\top \left(\begin{pmatrix} \hat{\varphi}^* & \hat{\Phi}^* \\ \hat{\gamma}^* & 0 \end{pmatrix} \otimes X^* \right) \right)^{-1} \left(\begin{pmatrix} \hat{\varphi}^* & \hat{\Phi}^* \\ \hat{\gamma}^* & 0 \end{pmatrix} \otimes X^* \right)^\top \begin{pmatrix} x_1^* \\ \vdots \\ x_n^* \\ y - Z\hat{\alpha} \end{pmatrix}$$

$\hat{\varphi}$, $\hat{\Phi}$ and $\hat{\gamma}$ are given, therefore we get:

$$\begin{aligned}
\begin{pmatrix} \hat{\varphi}^* & \hat{\Phi}^* \\ \hat{\gamma}^* & 0 \end{pmatrix} \otimes X^* &= \begin{pmatrix} \hat{\varphi} & \hat{\Phi} \\ \rho^{-1}\hat{\gamma} & 0 \end{pmatrix} \otimes (\rho X) \\
&= \left(\left(\mathcal{I}_{n+1} \begin{pmatrix} 1 & \dots & 1 & \rho^{-1} \end{pmatrix}^\top \right) \begin{pmatrix} \hat{\varphi} & \hat{\Phi} \\ \hat{\gamma} & 0 \end{pmatrix} \right) \otimes \left(\left(\mathcal{I}_T \begin{pmatrix} \rho & \dots & \rho \end{pmatrix}^\top \right) X \right) \\
&= \left(\left(\mathcal{I}_{n+1} \begin{pmatrix} 1 & \dots & 1 & \rho^{-1} \end{pmatrix}^\top \right) \otimes \left(\mathcal{I}_T \begin{pmatrix} \rho & \dots & \rho \end{pmatrix}^\top \right) \right) \left(\begin{pmatrix} \hat{\varphi} & \hat{\Phi} \\ \hat{\gamma} & 0 \end{pmatrix} \otimes X \right) \\
&= \left(\mathcal{I}_{(n+1)T} \left(\underbrace{\rho \dots \rho}_{=nT} \ \underbrace{1 \dots 1}_{=T} \right)^\top \right) \left(\begin{pmatrix} \hat{\varphi} & \hat{\Phi} \\ \hat{\gamma} & 0 \end{pmatrix} \otimes X \right) \\
\begin{pmatrix} x_1^* \\ \vdots \\ x_n^* \\ y - Z\hat{\alpha} \end{pmatrix} &= \left(\mathcal{I}_{(n+1)T} \begin{pmatrix} \rho & \dots & \rho & 1 & \dots & 1 \end{pmatrix}^\top \right) \begin{pmatrix} x_1 \\ \vdots \\ x_n \\ y - Z\hat{\alpha} \end{pmatrix}
\end{aligned}$$

Now we define:

$$W_\rho := \left(\mathcal{I}_{(n+1)T} \begin{pmatrix} \rho & \dots & \rho & 1 & \dots & 1 \end{pmatrix}^\top \right) \left(\mathcal{I}_{(n+1)T} \begin{pmatrix} \rho & \dots & \rho & 1 & \dots & 1 \end{pmatrix}^\top \right)$$

Therefore we get:

$$\begin{aligned}
\begin{pmatrix} \hat{w}^* \\ \hat{\omega}_1^* \\ \vdots \\ \hat{\omega}_s^* \end{pmatrix} &= \left(\left(\begin{pmatrix} \hat{\varphi}^* & \hat{\Phi}^* \\ \hat{\gamma}^* & 0 \end{pmatrix} \otimes X^* \right)^\top \left(\begin{pmatrix} \hat{\varphi}^* & \hat{\Phi}^* \\ \hat{\gamma}^* & 0 \end{pmatrix} \otimes X^* \right) \right)^{-1} \left(\begin{pmatrix} \hat{\varphi}^* & \hat{\Phi}^* \\ \hat{\gamma}^* & 0 \end{pmatrix} \otimes X^* \right)^\top \begin{pmatrix} x_1^* \\ \vdots \\ x_n^* \\ y - Z\hat{\alpha} \end{pmatrix} \\
&= \left(\left(\begin{pmatrix} \hat{\varphi} & \hat{\Phi} \\ \hat{\gamma} & 0 \end{pmatrix} \otimes X \right)^\top W_\rho \left(\begin{pmatrix} \hat{\varphi} & \hat{\Phi} \\ \hat{\gamma} & 0 \end{pmatrix} \otimes X \right) \right)^{-1} \left(\begin{pmatrix} \hat{\varphi} & \hat{\Phi} \\ \hat{\gamma} & 0 \end{pmatrix} \otimes X \right)^\top W_\rho \begin{pmatrix} x_1 \\ \vdots \\ x_n \\ y - Z\hat{\alpha} \end{pmatrix} \qquad \square
\end{aligned}$$

A.3 Appendix of Chapter 4

Main diseases categories according to ***ICD-10*** (Version 2019):

Taken from website of WHO (`https://icd.who.int/browse10/2019/en`):

- **A/B**: Certain infectious and parasitic diseases
- **C/D**: Neoplasms and Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism

- **E**: Endocrine, nutritional and metabolic diseases
- **F**: Mental and behavioural disorders
- **G**: Diseases of the nervous system
- **H**: Diseases of the eye and adnexa and Diseases of the ear and mastoid process
- **I**: Diseases of the circulatory system
- **J**: Diseases of the respiratory system
- **K**: Diseases of the digestive system
- **L**: Diseases of the skin and subcutaneous tissue
- **M**: Diseases of the musculoskeletal system and connective tissue
- **N**: Diseases of the genitourinary system
- **O**: Pregnancy, childbirth and the puerperium
- **P**: Certain conditions originating in the perinatal period
- **Q**: Congenital malformations, deformations and chromosomal abnormalities
- **R**: Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
- **S/T**: Injury, poisoning and certain other consequences of external causes
- **U**: Codes for special purposes
- **V/W/X/Y**: External causes of morbidity and mortality
- **Z**: Factors influencing health status and contact with health services

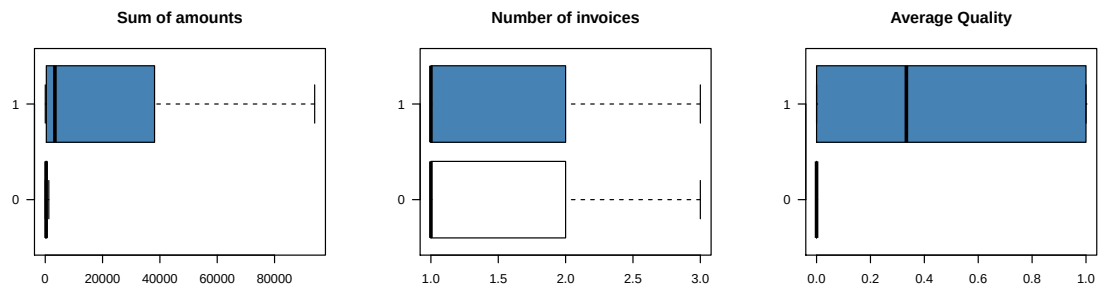


Figure A.3.1: Boxplots of main diagnose **A/B**.

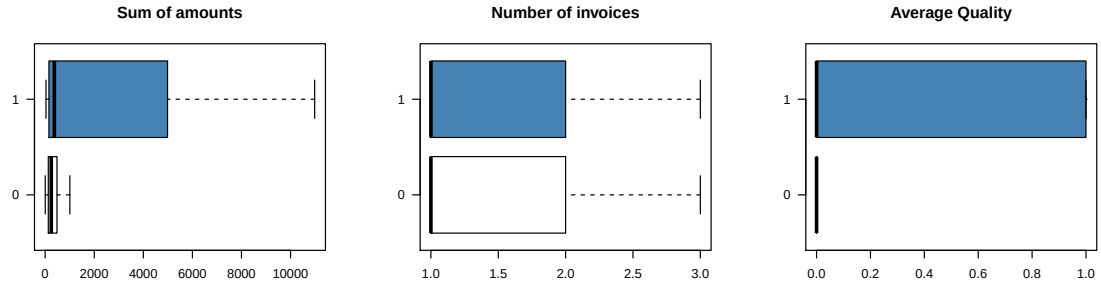


Figure A.3.2: Boxplots of main diagnose **J**.

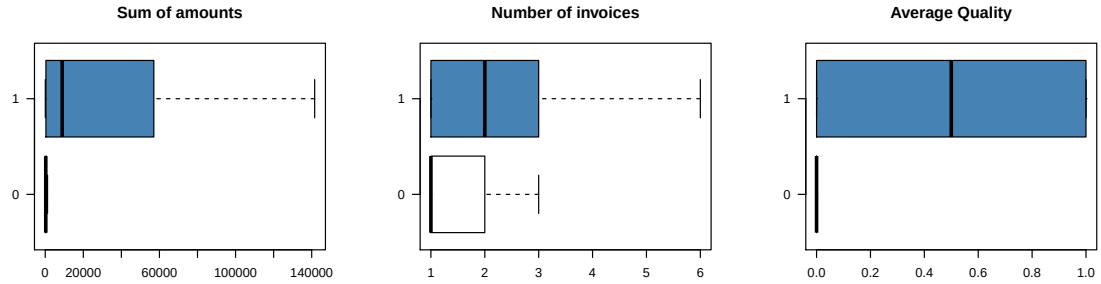


Figure A.3.3: Boxplots of main diagnose **Q**.

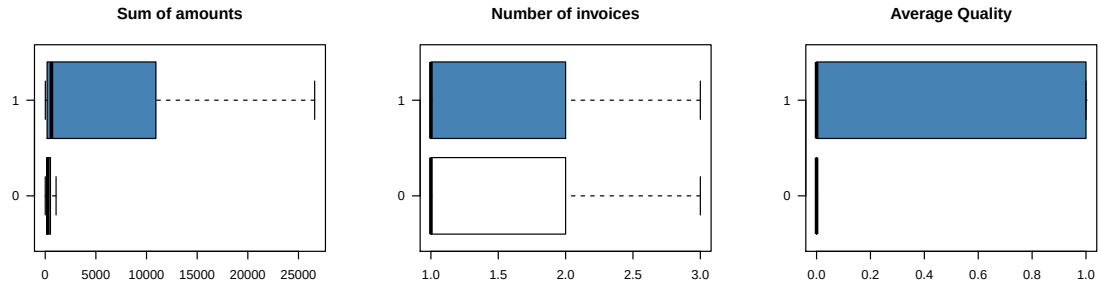


Figure A.3.4: Boxplots of main diagnose **R**.

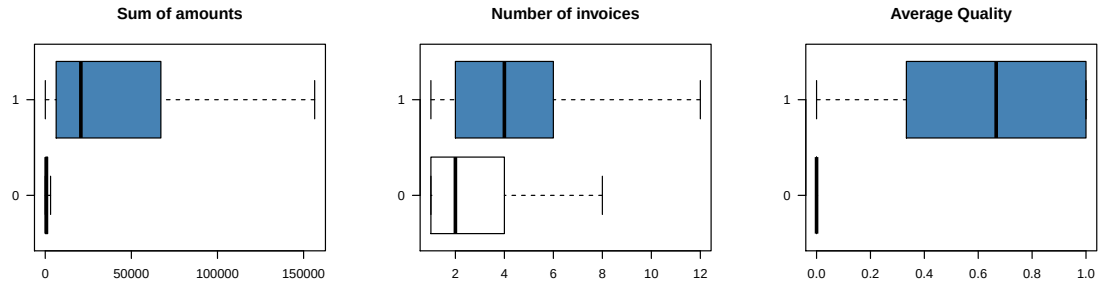


Figure A.3.5: Boxplots of main diagnose **Z**.

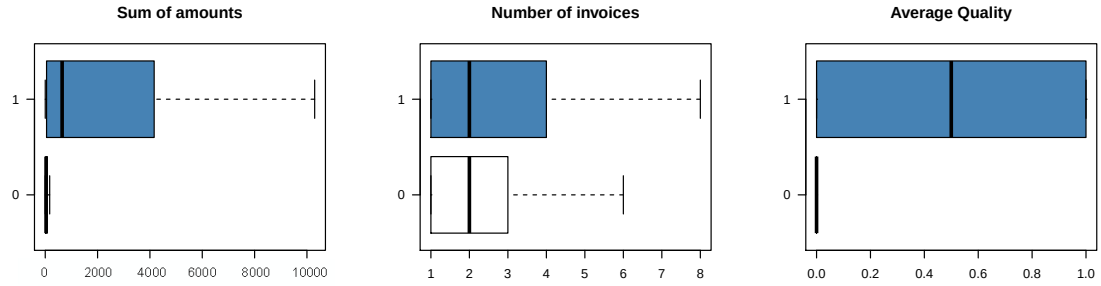


Figure A.3.6: Boxplots of main diagnose '**Rare**'.

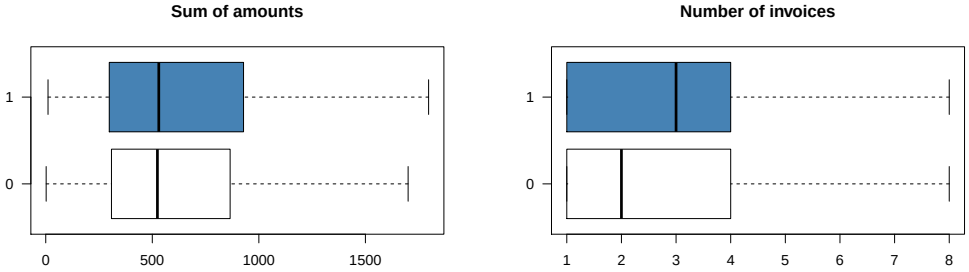


Figure A.3.7: Boxplots of outpatient treatment (**OT**).

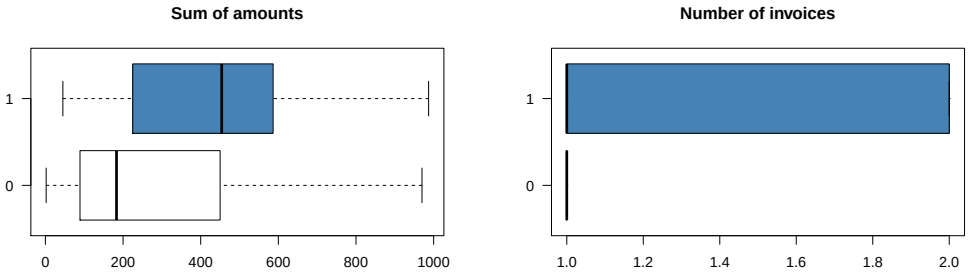


Figure A.3.8: Boxplots of medical remedies (**MR**).

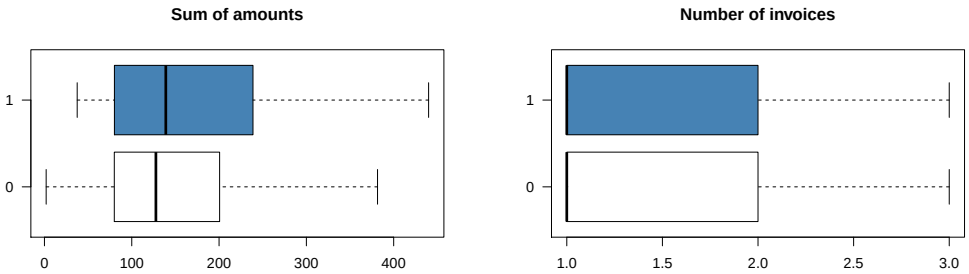


Figure A.3.9: Boxplots of homeopathic practitioner (**HP**).

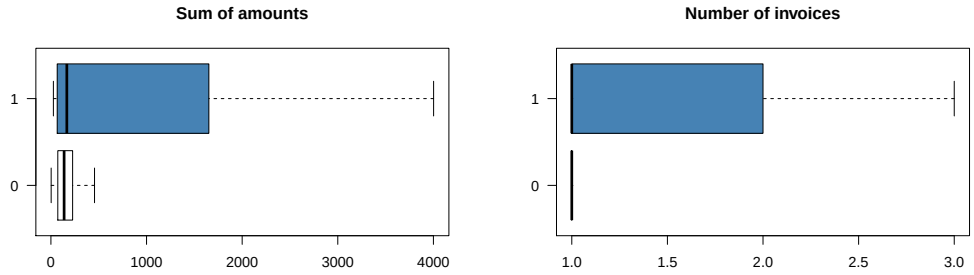


Figure A.3.10: Boxplots of medical aids (**MA**).

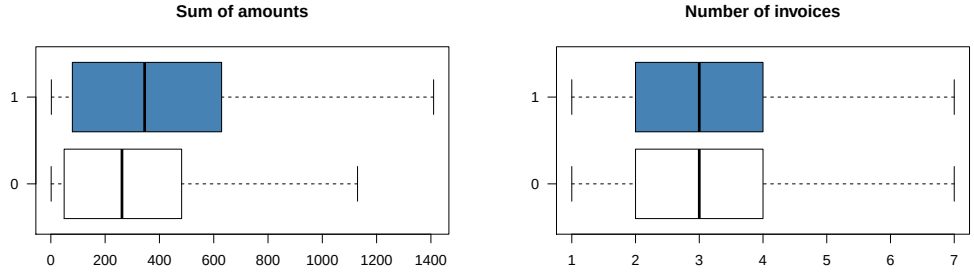


Figure A.3.11: Boxplots of medications (**MD**).

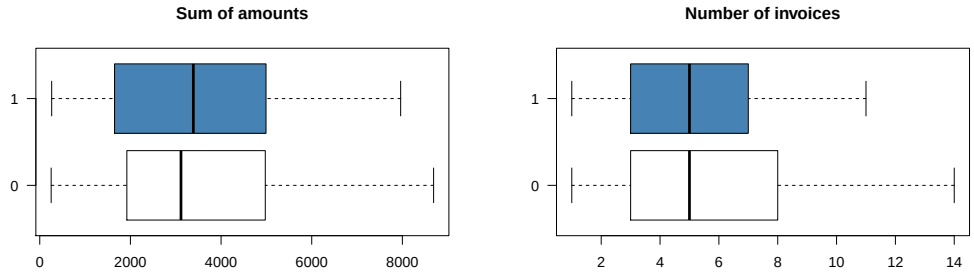


Figure A.3.12: Boxplots of care services (**CS**).

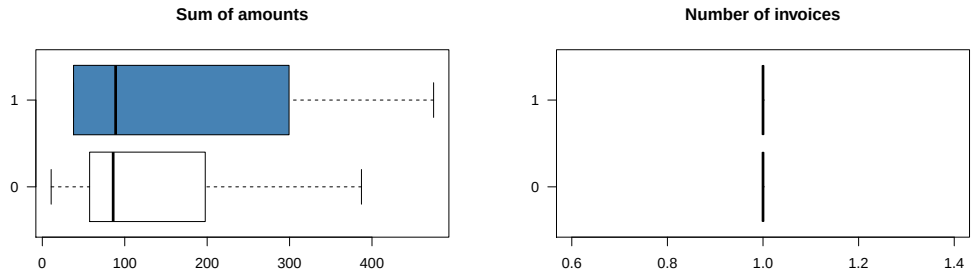


Figure A.3.13: Boxplots of main dental treatment (**DT**).

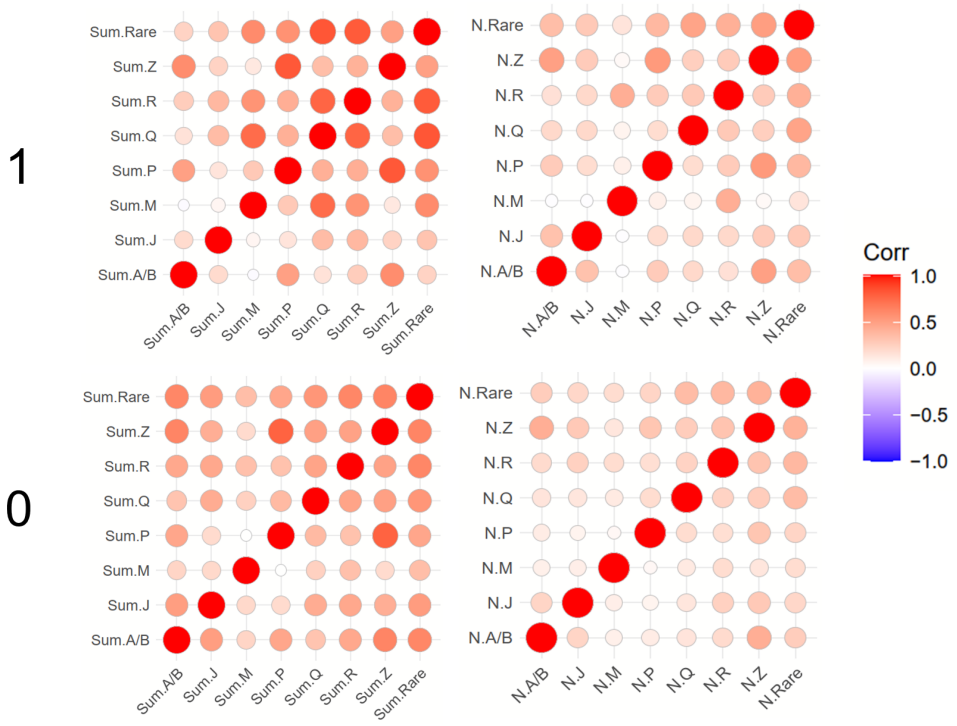


Figure A.3.14: Correlation analysis of the sum of amount (left) and number of invoices (right) for all main diagnoses, ordered by with / without notice (1/0).

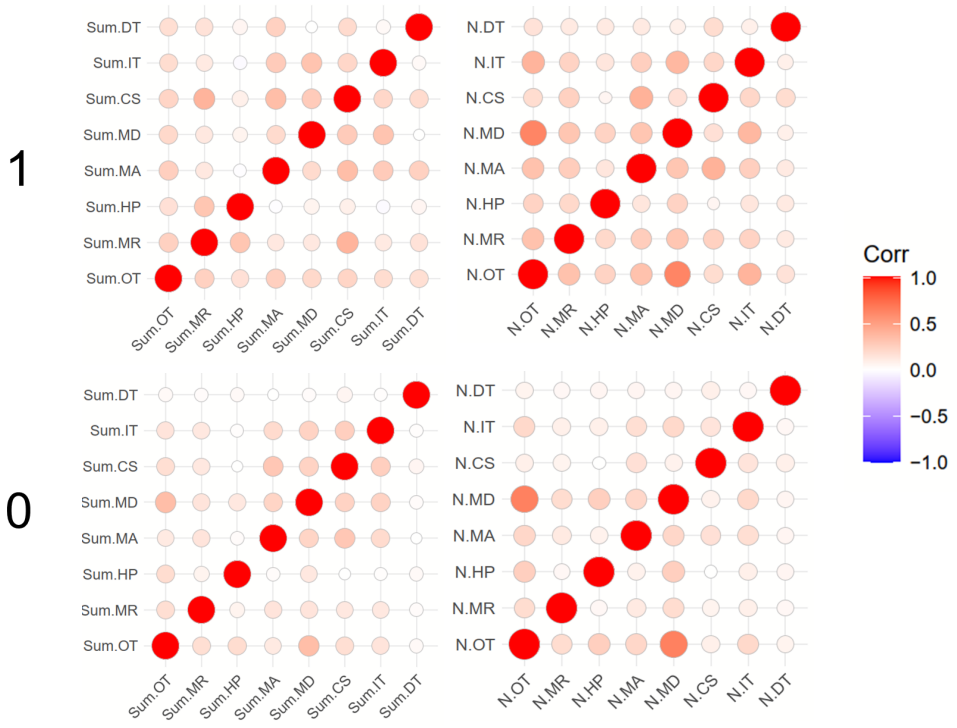


Figure A.3.15: Correlation analysis of the sum of amount (left) and number of invoices (right) for all treatment groups, ordered by with / without notice (1/0).

Bibliography

Bai, J. and Li, K. (2012). Statistical analysis of factor models of high dimension. *The Annals of Statistics*, 40(1):436 – 465.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191 – 221.

Bai, J. and Ng, S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146(2):304 – 317.

Baumel, T., Nassour-Kassis, J., Cohen, R., Elhadad, M., and Elhadad, N. (2018). Multi-label classification of patient notes: Case study on ICD code assignment. In *Proceedings of the 2018 AAAI Joint Workshop on Health Intelligence (W3PHIAI 2018)*, pages 409 – 416. AAAI press.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5 – 32.

Breitung, J. and Eickmeier, S. (2016). Analyzing international business and financial cycles using multi-level factor models. In Koopman, S. and Hillebrand, E., editors, *Advance in Econometrics*, volume 35, chapter II, pages 117 – 214. Emerald Publishing Limited, 1 edition.

Burkul, P., Umapathy, K., Asaithambi, A., and Huang, H. (2020). Data mining pipeline for performing decision tree analysis on mortality dataset with ICD-10 codes. In *SAIS 2020 Proceedings*. 28.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). Crisp-dm 1.0 step-by-step data mining guide. Technical report, The CRISP-DM consortium.

Chen, C., Liaw, A., and Breiman, L. (2004). Using random forest to learn imbalanced data. techreport 666, Department of Statistics, UC Berkley.

Couronné, R., Probst, P., and Boulesteix, A.-L. (2018). Random forest versus logistic regression: A large-scale benchmark experiment. *BMC Bioinformatics*, 19(1):1 – 14.

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 233 – 240, New York, NY, USA. ACM.

de Jong, S. and Kiers, H. A. (1992). Principal covariates regression: Part I. Theory. *Chemometrics and Intelligent Laboratory Systems*, 14(1):155 – 164. Proceedings of the 2nd Scandinavian Symposium on Chemometrics.

Donoho, D. L. and Gasko, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics*, 20(4):1803 – 1827.

Dyckerhoff, R. (2004). Data depths satisfying the projection property. *Allgemeines Statistisches Archiv*, 88(2):163 – 190.

Dyckerhoff, R. (2016). Convergence of depths and depth-trimmed regions. arXiv:1611.08721 [math.ST].

Dyckerhoff, R. and Stenz, H. J. (2021). Depth-based support vector classifiers to detect data nests of rare events. *International Journal of Computational Economics and Econometrics*, 11(2):107 – 142.

Fang, K., Kotz, S., and Ng, K. (1990). *Symmetric multivariate and related distributions*. Number 36 in Monographs on statistics and applied probability. Chapman and Hall, London, UK.

Gill, P. E., Murray, W., and Wright, M. H. (1981). *Practical optimization*. Academic Press Inc., London, UK.

Günnemann, N. and Pfeffer, J. (2017). Cost matters: A new example-dependent cost-sensitive logistic regression model. In Kim, J., Shim, K., Cao, L., Lee, J.-G., Lin, X., and Moon, Y.-S., editors, *Advances in Knowledge Discovery and Data Mining*, pages 210 – 222, Cham. Springer.

Harding, M. and Vasconcelos, G. F. R. (2022). Managers versus Machines: Do algorithms replicate human intuition in credit ratings? arXiv:2202.04218 [econ.EM].

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning*, volume 1 of *Springer Series in Statistics*. Springer, New York, NY.

Heij, C., Groenen, P. J., and van Dijk, D. (2007). Forecast comparison of principal component regression and principal covariate regression. *Computational Statistics & Data Analysis*, 51(7):3612 – 3625.

Hsu, C.-W., Chang, C.-C., and Lin, C.-J. (2003). A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University.

Kahlenberg, J. (2005). *Storno und Profitabilität in der Privathaftpflichtversicherung: eine Analyse unter Verwendung von univariaten und bivariaten verallgemeinerten linearen Modellen*. Berichte aus der Statistik. Shaker.

Kim, S., Mun, B. M., and Bae, S. J. (2018). Data depth based support vector machines for predicting corporate bankruptcy. *Applied Intelligence*, 48(3):791 – 804.

Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221 – 232.

Lange, T., Mosler, K., and Mozharovskyi, P. (2014). Fast nonparametric classification based on data depth. *Statistical Papers*, 55(1):49 – 69.

Li, J., Cuesta-Albertos, J. A., and Liu, R. Y. (2012). DD-classifier: Nonparametric classification procedure based on DD-plot. *Journal of the American Statistical Association*, 107(498):737 – 753.

Liu, R. Y. (1990). On a notion of data depth based on random simplices. *The Annals of Statistics*, 18(1):405 – 414.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2(1):49 – 55.

Marqués, A., García, V., and Sánchez, J. (2013). On the suitability of resampling techniques for the class imbalance problem in credit scoring. *Journal of the Operational Research Society*, 64(7):1060 – 1070.

Min, S.-H., Lee, J., and Han, I. (2006). Hybrid genetic algorithms and support vector machines for bankruptcy prediction. *Expert Systems with Applications*, 31(3):652 – 660.

Mosler, K. (2013). Depth statistics. In Becker, C., Fried, R., and Kuhnt, S., editors, *Robustness and complex data structures*, pages 17 – 34. Springer, 1 edition.

Mozharovskyi, P. (2014). *Contributions to depth-based classification and computation of the tukey depth*. Dr. Kovac Verlag, Hamburg.

Ozenne, B., Subtil, F., and Maucort-Boulch, D. (2015). The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *Journal of Clinical Epidemiology*, 68(8):855 – 859.

Reuß, A. and Zwiesler, H.-J. (2006). Ein generisches Kreislaufmodell zur Einbettung von Data-Mining-Analysen in die Geschäftsprozesse von Unternehmen mit einem Fallbeispiel aus der Unfallversicherung. *Zeitschrift für die gesamte Versicherungswissenschaft*, 95(1):201 – 230.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206 – 215.

Sailer, F., Pobiruchin, M., Bochum, S., Martens, U., and Schramm, W. (2015). Prediction of 5-Year survival with data mining algorithms. In Mantas, J., Hasman, A., and Househ, M. S., editors, *Studies in health technology and informatics*, volume 213, pages 75 – 78. IOS Press BV.

Shen, F., Wang, R., and Shen, Y. (2020). A cost-sensitive logistic regression credit scoring model based on multi-objective optimization approach. *Technological and Economic Development of Economy*, 26(2):405 – 429.

Smola, A. J. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3):199 – 222.

Sofaer, H. R., Hoeting, J. A., and Jarnevich, C. S. (2019). The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution*, 10(4):565 – 577.

Stock, J. H. and Watson, M. (2006). Forecasting with many predictors. In Elliott, G., Granger, C., and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 1, chapter 10, pages 515 – 554. Elsevier, 1 edition.

Umbach, S. L. (2020). Forecasting with supervised factor models. *Empirical Economics*, 58(1):169 – 191.

Vapnik, V. N. (1995). *The nature of statistical learning theory*, volume 1. Springer, New York, NY.

Vencálek, O. (2017). Depth-based classification for multivariate data. *Austrian Journal of Statistics*, 46(3-4):117 – 128.

Wang, P. (2010). Large dimensional factor models with a multi-level factor structure. *Mimeo, Hong Kong University of Science and Technology*.

Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. *The Annals of Statistics*, 28(2):461 – 482.

Autorenhinweis:

Hartmut Jakob Stenz studierte Mathematik, Volkswirtschaftslehre und Statistik an den Universitäten Tübingen, Bonn, Ulm und Milwaukee (USA). Die vorliegende Publikation schließt seine Promotion an der Wirtschafts- und Sozialwissenschaftlichen Fakultät der Universität zu Köln ab. Sie entstand am Institut für Ökonometrie und Statistik der Universität zu Köln (Tag der Promotion 19. Dezember 2022, Erstgutachter Prof. Dr. Jörg Breitung, Zweitgutachter Prof. Dr. Rainer Dyckerhoff) und beinhaltet Beiträge zur Analyse seltener Ereignisse und großer Datenmengen. Herr Stenz ist als hauptamtlich Lehrender am zentralen Lehrbereich der Hochschule des Bundes für öffentliche Verwaltung tätig und unterrichtet in Fachgebieten mit mathematischen und datenwissenschaftlichen Schwerpunkten.


115
IHRE BEHÖRDENNUMMER