1. Introduction

Missing data are ubiquitous problem in every field of research. The handling of the missing values in an inadequate manner may lead to biased results and inference based on them are, in turn, misleading. The standard statistical techniques for analyzing data require complete cases without any missing observations. The deletion of the cases with missing information to obtain complete data will not only cause the loss of important information but can also affect inferences. In this dissertation, different imputation techniques using nearest neighbors are developed to address the missing data issues in high-dimensional as well as low dimensional data structures.

Due to the advancement and rapid growth in technology, the collection of high-dimensional data is no longer a tedious task. Regardless of considerable advances in technology over the last few decades, one has to face challenges when dealing with the high-dimensional data, specially regarding the analyses, interpretation and integration. One of the major problems, that an analyst has to face in high-dimensional data, is the occurrence of missing values. In particular, the situation becomes worse, when the distributional forms of the variables are different or these variables have been measured at different measurement scales (e.g. binary, multi-categorical, continuous, etc.). Whatever the reason, missing data may occur in all areas of applied research.

Nearest neighbors is a well known technique that has been successfully used in classification, clustering and imputation of data. A key issue in classical nearest neighbor imputation is the selection of the suitable number of nearest neighbors k. Another important topic, in particular in higher dimensions, is the selection of the relevant dimensions, since the computation of distances may suffer from the curse of dimensionality yielding poor performance in high-dimensional settings. These problems are handled by a novel approach to compute the weights on nearest neighbors.

In the first part of the thesis, some improved nearest neighbors imputation methods are developed that have better better performance than the classical NN imputation. In particular, a localized approach to missing data imputation that uses a weighted average of nearest neighbors based on L_q distances is developed. For the high-dimensional case, a new distance function that explicitly uses the correlation among variables is proposed. In contrast to classical approaches, a main advantage is that the proposed method automatically selects the relevant variables that contribute to the distance. The results from simulation studies as well as real studies show that the weighted distance procedure can successfully handle missing values for high dimensional data structures. In addition, extensions to the

binary, multi-categorical and mixed type data are also developed. It is shown in simulations, that the proposed imputation method works efficiently even when the number of samples is smaller than the number of variables. The method typically outperforms the considered competitors.

Imputation is not the ultimate goal of any analysis, an estimate with smaller imputation error may not perform well in the downstream analysis. In particular, treating the imputed data just as the complete is not recommended. This is the starting point for the second part of the thesis. Several data analytic techniques for inference from an imputed dataset in which missing values have been replaced by the proposed nearest neighbors imputation method have been proposed. A novel approach that combines the nearest neighbors imputation with bootstrap resampling estimation is suggested to obtain valid bootstrap inferences in a regression model. More specifically, imputing the bootstrap samples in the exact same way as the original data was imputed produces correct bootstrap estimates. The classification accuracy of the imputed data using different imputation methods is also compared.

Single imputation methods provide a single value as an estimate of the missing value, and thus do not account for the uncertainty of imputation. In contrast, multiple imputation takes this uncertainty into acount by providing more than one plausible values corresponding to each missing value in the data. Multiple imputation is a preferred choice of data analysts, due to its flexibility and since it can applied in a wide variety of missing data scenarios. In the presence of high-dimensional data ($p \gg n$), the missing values might be a more serious issue as the existing softwares/packages may fail. A non-parametric approach for multiple imputation in combination with the nearest neighbors is suggested. In particular, an algorithm that combines the bootstrap resampling with the nearest neighbors is given. The other algorithm uses a sequential procedure for nearest neighbors imputation of the missing values. The method successfully imputes missing values also in high-dimensional settings, in which existing software tend to fail. Using a variety of simulated data with MCAR and MAR missing patterns, the proposed algorithm is compared to existing methods. The performance is evaluated by using mean squared imputation errors and inference results obtained for the imputed data. Various measures are used to compare methods, including mean squared errors of estimated regression coefficients, their standard errors, confidence intervals and their coverage probabilities. The simulation results, for both cases $n < p$ and $n > p$, show that the sequential imputation using weighted nearest neighbors can be successfully applied to a wide range of data settings and outperforms or is close to the best when compared to existing methods.

Guidelines through the Thesis

This thesis consists of 10 chapters and 4 appendices. Due to the close interdependence, some paragraphs contain a certain degree of overlap with regard to content. These overlaps

are consciously retained to enhance comprehensibility and allow for a separate reading of the single chapters.

Chapter 2 discusses the general concepts regarding missing values. It gives an introduction to the issues that arise in the presence of missing data as well as the current methodology and literature to handle these problems.

This thesis can be divided into three main parts which are dedicated to single imputation (Chapter 3 to 6), impact of imputation on inference and classification (Chapter 7 and 8) and multiple imputation (Chapter 9). In order to keep the single chapters self-contained, every chapter contains separate introductions to the relevant topics and a separate conclusion. Therefore, every chapter can also be read separately.

Chapter 3 and Chapter 4 deal with missing values problem where the covariates are metric in nature.

Chapter 5 extends the imputation method for binary and multi-categorical data.

Chapter 6 is dedicated to mixed-type data.

Chapter 7 is based on using bootstrap sampling to reach valid inferences from imputed data.

Chapter 8 investigates the impact of imputation on the classification of data.

Chapter 9 presents multiple imputation algorithms.

Throughout the thesis, the scalars are represented by lowercase letters (e.g. x), column vectors are represented by boldfaced lowercase letters (e.g. $\mathbf{x}$), the row vector as the transpose of the column vector (e.g. $\mathbf{x}^T$), and the matrices are represent by boldfaced uppercase letters (e.g. $\mathbf{X}$).

In the following we present the summaries of the individual chapters.

In **Chapter 3** we present improved versions of the nearest neighbor imputation method. First, a weighted nearest neighbor imputation method based on L_q distances is proposed. It is demonstrated that the method tends to have a smaller imputation error than other nearest neighbor estimates. We then consider weighted- neighbor imputation methods that use distances for selected covariates. The careful selection of distances that carry information about the missing values yields an imputation tool that can outperform competing nearest neighbor methods . This approach performs well, especially when the number of predictors is large. The methods are evaluated in simulation studies and with several real data sets from different fields.

Chapter 4 focuses on high-dimensional data setting in the real world situations and applies the methods developed in Chapter 3 to the real data. High-dimensional data like gene expression and RNA-sequences often contain missing values. The subsequent analysis and results based on these incomplete data can suffer strongly from the presence of these missing values. Several approaches to imputation of missing values in gene expression data have

been developed but the task is difficult due to the high-dimensionality (number of genes) of the data. In this chapter, an imputation procedure is proposed that uses weighted nearest neighbors. Instead of using nearest neighbors defined by a distance that includes all genes the distance is computed for genes that are apt to contribute to the accuracy of imputed values. The method aims at avoiding the curse of dimensionality, which typically occurs if local methods as nearest neighbors are applied in high-dimensional settings. The proposed weighted nearest neighbors algorithm is compared to existing missing value imputation techniques like mean imputation, KNNimpute and the recently proposed imputation by random forests. We use RNA-sequence and microarray data from studies on human cancer to compare the performance of the methods. The results from simulations as well as real studies show that the weighted distance procedure can successfully handle missing values for high-dimensional data structures where the number of predictors is larger than the number of samples. The method in chapter 4 typically outperforms the considered competitors.

In **Chapter 5** the weighted nearest neighbors imputation method is extended for the binary and categorical variables. While various imputation methods are available for metrically scaled variables, methods for categorical data are scarce. An imputation method that has been shown to work well for high-dimensional metrically scaled variables is the imputation by nearest neighbor methods. One has to use specific distances or similarity measures, which are typically based on contingency tables for categorical data. The Euclidean or variants of the Minkowski distance give an equal importance to all the variables in the data matrix when computing the distance. But for a larger number of variables, the equal weighting ignores the complex structure of correlation/association among these variables. In Chapter 5, we propose a weighted nearest neighbors approach based on dummy variables to impute missing values in categorical variables. As demonstrated in the chapter, better distance measures are obtained by utilizing the association between variables. More specific, we propose a weighted distance that explicitly takes the association among covariates into account. Strongly associated covariates are given higher weights forcing them to contribute more strongly to the computation of the distances than weakly associated covariates. The performance of different imputation methods is compared in terms of the proportion of falsely imputed values. Simulation results show that the weighting of attributes yields smaller imputation errors than existing approaches. A variety of real data sets is used to support the results obtained by simulations.

Chapter 6 is dedicated to mixed type of data. One has to deal with combination of continuous and nominal variables in many real world applications, therefore the methods to impute mixed data become more important. Since the multiple imputation techniques fail to impute high-dimensional missing data, the nonparametric single imputation methods are gaining more popularity. We propose an improved version of the popular nonparametric nearest neighbors method which uses information only on potentially relevant neighbors to impute missing values. More specifically, We introduce a distance function that is more appropriate for mixed data. It is an extension of Tutz and Ramzan (2015) and uses information on association among variables. A particular advantage of the proposed method is that while imputing the missing values, it simultaneously takes into account the similarities between samples and the relationships between covariates. The performance of the pro-

posed method is investigated under a variety of data settings. The results show a smaller imputation error and better performance when compared to other approaches. It is shown that the proposed imputation method works efficiently even when the number of samples is smaller than the number of variables.

In **Chapter 7** we present analytic techniques for inference in datasets in which missing values have been replaced by nearest neighbors imputation methods. It is not advisable to treat the imputed data just as the complete data. To apply the existing methods for analyzing the data, for example, to estimate the variance and/or statistical inference will probably produce invalid results because these methods do not account for the uncertainty of imputations. To overcome this, we presents a bootstrap algorithm that combines the nearest neighbors imputation with bootstrap resampling estimation to obtain valid bootstrap inference in a linear regression model. The proposed procedure that provides estimated values for the missings which have not only smaller MSEs, but also provide better inference, when one variable is a response variable and the rest of the variables are predictors in a linear regression model. The confidence intervals for the regression coefficients are constructed using bootstrap sampling. Simulation results show that the suggested imputation method provides promising results and the bootstrap has the desired nominal coverage.

Chapter 8 investigates the impact of imputation on the classification of the data. Although some methods in machine learning can be tolerant to the presence of missing values, many statistical methods require a complete data matrix and so does classification methods. We compare the accuracy of different classifiers using the imputed data obtained by different imputation methods. The results show that some classification methods are robust to the presence of missing values and hence their accuracy is not affected much. While some methods really influenced by imputed values. Our study shows that the impact of imputation varies between data sets and different classifiers. Overall the proposed weighted nearest neighbors imputation provided best results in four out of six datasets considered and second best for the remaining two datasets. The random forests imputation provided the second best classification results in our experiments, whereas `MEAN` imputation was seen to yield poor performance. The results based on misclassification error and Brier score show that imputation improves the classification accuracy and in particular, beneficial for the higher amounts of missing data. When the amount of missing data is small, there is not much difference among the imputation methods.

In **Chapter 9**, multiple imputation methods are proposed based on the sequential nearest neighbors. In the presence of high-dimensional data ($p \gg n$), the missing values lead to a more serious issue as the existing softwares/packages fail respond. In this chapter we present a multiple imputation method based on nearest neighbors which sequentially imputes the missing values. The specific distances are computed using the information of correlation among the target and candidate predictors. Thus only the relevant predictors contribute to the computation of distances. The method successfully imputes missing values in high-dimensional settings. Using a variety of simulated data with MCAR and MAR missing patterns, the proposed algorithm is compared to three existing methods, Multivariate Imputation by Chained Equations (MICE), Iterative robust model-based imputation

(IRMI) and Amelia. Our extensive simulation results show that the sequential imputation using weighted nearest neighbors can be applied to a wide range of data settings considered here and outperforms or is close to the best when compared to existing methods.

The thesis concludes with a short summary and an outlook to potential aspects for future research in **Chapter 10**.

Contributing Manuscripts

Parts of this thesis were published as articles in peer reviewed journals or as pre-prints at the Cornell University Library's open access archive `arXiv.org`. Other parts were published in proceedings of scientific conferences or as technical reports at the Department of Statistics of the Ludwig-Maximilians-Universität München. All manuscripts have been written in cooperation with (supervising) coauthors. In the following, chapter by chapter all contributing manuscripts are listed together with a declaration of the personal contributions of the respective authors:

Chapter 3 is based on the published paper

- Tutz and Ramzan (2015). Improved methods for the imputation of missing data by nearest neighbor methods. *Computational Statistics & Data Analysis*, 84–99. https://doi.org/10.1016/j.csda.2015.04.009

Gerhard Tutz initiated the project and developed the imputation approach. Shahla Faisal wrote the code for the method in R (R Core Team, 2017). Shahla Faisal was responsible for the implementation of the method, of the simulation studies and the application to real data. Gerhard Tutz wrote the first version of the paper. Furthermore, Shahla Faisal developed the corresponding R package `wNNSel` (available on CRAN). The technical report (Tutz and Ramzan, 2014) contains first versions of work on the project. Preliminary work on the topics of Chapter 3 has been presented at IWSM 2014 (Ramzan et al., 2014) as a talk

- Ramzan, Tutz, and Heumann (2014). Improved methods for nearest neighbor imputation. 29th International Workshop on Statistical Modelling. Georg-August-Universität Göttingen, Germany.

Chapter 4 is based on the published paper

- Faisal and Tutz (2017c). Missing value imputation for gene expression data by tailored nearest neighbors. Statistical Applications in Genetics and Molecular Biology 16 (2), 95–106. DOI 10.1515/sagmb-2015-0098.

This project was initiated jointly by Shahla Faisal and Gerhard Tutz. Shahla Faisal developed the R code (R Core Team, 2017) as well as implemented and conducted the numerical experiments and the data analyses. She mainly wrote the paper and Gerhard Tutz added valuable remarks and complementary notes which improved the manuscript. Apart from some minor modifications in the notations, Chapter 4 and Faisal and Tutz (2017c) match.

Chapter 5 is based on a paper available on arXiv

- Faisal and Tutz (2017e). Nearest Neighbor Imputation for Categorical Data by Weighting of Attributes. `arXiv:1710.01011 [stat.ME]`

The project was initiated by Shahla Faisal and further developed jointly by the two authors. Shahla Faisal wrote the R code (R Core Team, 2017) as well as implemented the method and conducted the simulations and applications on real data. She mainly wrote the manuscript in close collaboration with Gerhard Tutz. Preliminary work on Chapter 5 has been presented at IWSM-2016 (Ramzan and Tutz, 2016) as a talk.

- Ramzan and Tutz (2016). Nearest neighbor imputation for categorical data by weighting of attributes. 31st International Workshop on Statistical Modelling. Rennes, France.

Chapter 6 The project was jointly developed by the two authors. Shahla Faisal was responsible for writing the code in R (R Core Team, 2017). Shahla Faisal implemented and conducted the numerical experiments and the data analyses. She also wrote the first version of the paper. Preliminary results linked to the topic of Chapter 5 have been presented at conference HDDA-IV (Faisal and Tutz, 2017a) as a talk

- Faisal and Tutz (2017a) Imputation for Missing Values in High-Dimensional Data Structures. International Workshop on Perspectives on High Dimensional Data (HDDA) VII. Guanajuato, Mexico.

and at the conference IWSM-2017 as a poster,

- Faisal and Tutz (2017b). Imputation in High-dimensional Mixed-Type data by Nearest Neighbors. 29th International Workshop on Statistical Modelling. University of Groningen, Netherlands.

The conference paper (Faisal and Tutz, 2017b) contains the preliminary work on the project. Shahla Faisal presented the poster and won the *best poster presentation* award in the conference (Faisal and Tutz, 2017b).

Chapter 7 The idea was initiated by both coauthors. Christian Heumann helped to develop the algorithm. The manuscript was mainly written by Shahla Faisal. Christian Heumann and Gerhard Tutz added their valuable comments to improve the manuscript. Currently, the manuscript is submitted for publication.

- Faisal, S. and Heumann, C. (2017). Bootstrap Inference for Weighted Nearest Neighbors Imputation.

Preliminary results linked to the topic of Chapter 7 have been presented in the conferences in the talks

- Ramzan, Heumann, and Tutz (2016a). Bootstrap Confidence Interval after Nearest Neighbors Imputation. 14th International Conference on Statistical Sciences: Statistics for Better Decision-Making and Development, Jinnah Sindh Medical University, Karachi, Pakistan. March 14-16, 2016.
- Ramzan, Heumann, and Tutz (2016b). Inference when using Nearest Neighbors methods and the Bootstrap. 22nd International Conference on Computational Statistics, Auditorium/Congress Palace Principe Felipe, Oviedo, Spain.

Chapter 8 The idea was initiated by all the coauthors. Shahla Faisal mainly wrote the manuscript. She was also responsible for conducting the simulations in `R` (R Core Team, 2017). Gerhard Tutz added valuable remarks which helped to improve the manuscript.

Chapter 9 The project was initiated and developed in close collaboration. Shahla Faisal developed the R code for the methods. She, as the first author, mainly wrote most of the manuscript and performed the presented analyses. Gerhard Tutz helped to improve the manuscript by extensive discussions. Shahla Faisal implemented the method and conducted the simulations and applications on real data. Preliminary results linked to the topic of Chapter 9 have been presented in the conference as a talk

- Faisal and Tutz (2017d). Multiple Imputation using Sequential Nearest Neighbors. The Third Annual Kliakhandler Conference on *Bayesian Inference in Statistics and Statistical Genetics*. Michigan Technological University, USA. August 16-20, 2017.

Software

Most computations in this thesis were done with the statistical program `R` (R Core Team, 2017). For most of the methods proposed in this thesis add-on packages for `R` were developed and some package are being developed. In particular, the following `R`-package was developed which is available on CRAN (R Core Team, 2017):

`wNNSel` provides the methods proposed in Chapter 3 and Chapter 4. (Faisal, 2017)

2. Methodological Concepts for Missing Data

In this chapter, we briefly describe missing data, their impact on downstream analysis and issues in high-dimensional data. Section 2.1 presents the basic concepts related to missing values, and the mechanisms of missing values. A brief overview of the available techniques to handle missing data is provided in Section 2.2. Section 2.3 is aimed at a brief history of the nearest neighbors methods, the classic nearest neighbors algorithm is described in Section 2.2. The next sections are dedicated to the proposed approaches regarding single imputation (Section 2.4-2.7). The remaining sections describes the concepts and algorithms related to inference based on imputed data and the multiple imputation of missing values.

2.1. Missing Values

Missing data are often a major problem in all areas of quantitative research. The term *missing* or *incomplete* refers to unavailability of an information on some characteristics of the data. Missing values values may occur due to many reasons, for example, patients may fail to respond to certain question. The subjects may withdraw or expire before completion of the studies. The respondents may not provide the information at all, for example income, age etc. It is also possible that some respondents do not provide the complete information on the queries, which is the most common reason for missing values in surveys. Sometimes the information may not be recorded or included into the database due to failure of recording mechanisms. Whatever the reason, missing values or incomplete data occur in all areas of applied research.

The major issue with missing data is that almost all standard (classic and modern) statistical techniques for analyzing the data require complete cases without any missing observation. The commonly used statistical packages are also set to the default options for dealing missing data, i.e., to discard the incomplete cases before preforming the actual analysis, despite that the case may have potential information to contribute to the overall analysis.

In real data applications, if the data contains fifteen percent or more missing values, it can greatly affect the results, some proper technique/procedure is required to handle five to fifteen percent of missing values. However, one to five percent missing values are considered to be manageable and less than one percent are usually trivial (Edgar Acuna and Rodriguez, 2004, Jiang and Yang, 2015).

Now-a-days, the collection of high-dimensional data is no more a tedious task due to the advancement and significantly rapid growth in technology. Therefore it is not uncommon to have a large number of variables measured on a few number of samples in a dataset, for example, in biomedical, epidemiological and social research. It becomes more challenging for the analysts to deal with missing values if they occur in high-dimensional data. An algorithm based on the popular random forests technique was proposed by Stekhoven and Bühlmann (2012). It is able to impute missing data in high-dimensional settings. Random forests avoid overfitting by bootstrap aggregation of multiple regression trees. Furthermore, the accuracy of predictions is enhanced by combining the predictions across all the trees (Breiman, 2001). An adaptation of this method and its comparison to parametric imputation methods was given by Shah et al. (2014). The results of their studies showed that their proposed method is more efficient and may produce confidence interval that are narrower than standard MI approaches. Four different variants of the nearest neighbors imputation were developed by Liao et al. (2014). But all of these methods do not properly account for the uncertainty of imputation, Deng et al. (2016) regarded them as *improper* in the sense of Rubin (1987).

To deal with high-dimensional data, some model based methods are also available in the literature apart from kNN and random forests. But, in general, these MI methods suffer from the curse of dimensionality and hence may not be suitable for imputing the high-dimensional data. When the number of predictors is less than the sample size, many software packages provide an easy application of MI methods, e.g., `R` packages `mice` and `Amelia`. When the number of predictors is less than but close to the sample size, they can be applied to get imputation results but may not perform well (Zhao and Long, 2016). But when the number of predictors exceeds the sample size, the available software packages crash or fail to respond. MI methods, particularly, can have problems when $n < p$, since imputation model may not be defined in this case. Some researchers have suggested to use regularized regression methods, which allow for variable selection before building the final imputation model (Long and Johnson, 2015).

2.1.1. Missing Data Mechanism

An important aspect in missing data imputation is the pattern of missing values because it determines the selection of an imputation procedure (Little and Rubin, 2002; Allison, 2001). Most imputation methods assume the data to be at least MAR, if not MCAR, and so does the NN method. Little and Rubin (2002) defined three categories of missing data, missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). If any of MCAR or MAR assumptions hold, the missing data mechanism is said to be *ignorable*. The missingness is known to be *non-ignorable* mechanism if the data has NMAR missing values. In the following we briefly describe these mechanism with explanatory example.