AALTO UNIVERSITY
SCHOOL OF SCIENCE
Department of Information and Computer Science
Degree Programme of Bioinformation Technology

Tommi Vatanen

Missing Value Imputation Using Subspace Methods with Applications on Survey Data

Master's thesis

Espoo, April 23, 2012

Supervisor: Prof. Samuel Kaski
Instructors: Tapani Raiko, D.Sc.(Tech.) and Krista Lagus, D.Sc.(Tech.)

Aalto University
School of Science

AALTO UNIVERSITY
SCHOOL OF SCIENCE
Department of Information and Computer Science
Degree Programme of Bioinformation Technology

ABSTRACT OF
MASTER'S THESIS

Author	Tommi Vatanen	Date	April 23, 2012
		Pages	vii + 78

Title of thesis	Missing Value Imputation Using Subspace Methods with Applications on Survey Data

Professorship	Information and Computer Science	Code	T-61

Supervisor	Prof. Samuel Kaski
Instructor	Tapani Raiko, D.Sc.(Tech.) and Krista Lagus, D.Sc.(Tech.)

In survey practice as well as in many other data analysis tasks, missing values are a common encounter. In this thesis, the missing value imputation task is studied using three subspace methods, principal component analysis (PCA), the Self-Organizing Map (SOM) and the Generative Topographic Mapping (GTM). The application area of interest is survey imputation, where imputation is conventionally conducted using, e.g., hot deck methods or multiple imputation by chained equations (MICE). Similarities and differences between imputation in survey practice and recommendation systems are discussed, as well.

The formalism behind missing value imputation is described together with general mechanisms giving rise to missing data. A detailed review of the aforementioned subspace methods in presence of missing data is given in order to motivate the novelties and new implementations contributed. The contributions of this thesis include (i) a novel way of treating missing data in the SOM algorithm, which is shown to improve properties of the model, (ii) a fine-tuned GTM, where the number of radial basis functions is increased during learning and the initialization is made using the SOM, and (iii) a novel regularization for the GTM for binary data.

Experimental comparisons of existing and proposed methods are made using the wine data set and Likert-scale data from two wellbeing-related surveys. The variational Bayesian PCA is shown to be superior in the single imputation task. It also enables automatic relevance determination, i.e., automatic selection of the number of principal components needed. Finally, multiple imputation (MI) using the subspace methods and MICE is demonstrated. It is shown, that with survey data with less than 2 % missing data, all MI methods provide very similar population le vel results.

Keywords	Missing value imputation, Missing-at-random, Principal component analysis, Self-Organizing Map, Generative Topographic Mapping

Aalto-yliopisto
Perustieteiden korkeakoulu

AALTO-YLIOPISTO Perustieteiden korkeakoulu Bioinformaatioteknologian tutkinto-ohjelma	DIPLOMITYÖN TIIVISTELMÄ
Tekijä: Tommi Vatanen	Päiväys: 23. huhtikuuta 2012 Sivumäärä: vii + 78
Työn nimi: Puuttuvien Arvojen Korvaaminen Aliavaruusmenetelmillä	
Professuuri: Tietojenkäsittelytiede	Koodi: T-61
Työn valvoja: Prof. Samuel Kaski	
Työn ohjaaja: Tekn.tri. Tapani Raiko, Tekn.tri. Krista Lagus	

Puuttuvat arvot ovat yleisiä niin kyselyaineistoissa kuin muissakin tilastollisesti analysoitavissa aineistoissa. Tässä opinnäytetyössä tutkitaan puuttuvien arvojen korvaamista käyttäen kolmea aliavaruusmenetelmää, pääkomponenttianalyysiä (PCA), itseorganisoivaa karttaa (SOM) ja generatiivista topografista kuvausta (GTM). Sovellusalueena ovat kyselyaineistot, joiden puuttuvia arvoja korvataan perinteisesti esimerkiksi käyttäen niinsanottuja hot-deck -menetelmiä tai moninkertaista ketjutettua korvaamista (multiple imputation by chained equations, MICE). Opinnäytteessä myös tarkastellaan kyselyaineistojen korvaamisen ja suositusjärjestelmien välisistä eroavaisuuksista ja samankaltaisuuksista menetelmätasolla.

Edellä mainitut aliavaruusmenetelmät on esitelty yksityiskohtaisesti motivoiden sekä uusia muutoksia, että niiden käyttöä puuttuvien arvojen korvaamisessa. Työssä esitettyjä kontribuutioita ovat (i) uusi tapa käsitellä puuttuvie arvoja SOM-algoritmissa, minkä näytetään parantavan algoritmin ominaisuuksia, (ii) niinsanottu “fine-tuned GTM”, jossa käytettävien kantafunktioiden määrää kasvattamalla voidaan oppia parempia malleja, sekä (iii) uudella tavalla regularisoitu GTM-malli binaariselle aineistolle.

Kokeellisessa osuudessa vertaillaan ehdotettuja malleja sekä käyttäen tunnettua viiniaineistoa että kahta Likert-asteikkoista hyvinvointikyselyaineistoa. Variaatioaproksimoitu bayesilainen PCA osoittautuu parhaaksi tehtäessä yksittäisiä puuttuvien arvojen korvauksia. Se tekee myös automaattista mallinvalintaa, jolloin erillistä validointia mallin kompleksisuuden valitsemiseksi ei tarvita.

Lopuksi näytetään moninkertaista puuttuvien arvojen korvaamista (MI) käyttäen aliavaruusmenetelmiä sekä MICE-menetelmää. Menetelmät tuottavat hyvin samanlaisia tuloksia kyselyaineistolla, jossa on alle 2 % puuttuvia arvoja.

Avainsanat

Preface

This thesis was done in the Department of Information and Computer Science in Aalto university school of science and was funded by VirtualCoach research project. I would like to thank Krista Lagus, the project leader, for instructorship and for financial support. Deepest thanks to the whole VirtualCoach research and partner team for the support and encouragement.

Many thanks to Prof. Samuel Kaski for supervision and valuable comments. Instructor Tapani Raiko has provided the most valuable support, advice and comments; thank you very much. Timo Honkela has provided me the most valuable support and advice during my whole stay on the department. Warm thanks for your support and for trusting me enough to hire me after my first year of studies.

I thank Hilkka Mehtätalo, Maarit Kuoppala and others related to the nursing survey for providing a valuable and interesting data set. I also thank Harri Sintonen ja Pasi Aronen regarding their cooperation on 15D instrument data. Many thanks for your open-minded attitude for interdisciplinary collaboration and for providing me the 15D instrument data.

Thanks to Olli Kotiranta, Arttu Modig, Ilari Nieminen, Antti Heikkilä, Timo Romppanen, Juulia Suvilehto and others for providing valuable comments on my text and thoughts along the thesis project. Finally, I would like to thanks my parents, brother and friends for their valuable support and encouragement throughout my studies the thesis project.

Espoo, April 23, 2012
Tommi Vatanen

Contents

Glossary

Abbreviations

altSOM	The alternating SOM
EM	Expectation-Maximization (algorithm)
ECM	Expectation Conditional Maximization (algorithm)
GEM	Generalized EM (algorithm)
GTM	Generative Topographic Mapping
HRQoL	Health-related quality of life
impSOM	The imputation SOM
MAP	Maximum a posteriori
MAR	Missing at random
MCAR	Missing completely at random
MCMC	Markov Chain Monte Carlo
MI	Multiple Imputation
MICE	Multiple Imputation by Chained Equations
ML	Maximum likelihood
NMAR	Not missing at random
PCA	Principal Component Analysis
PPCA	Probabilistic PCA
RBF	Radial Basis Function
RMS	Root mean square (error)
SOM	Self-Organizing Map
SRMI	Sequential Regression Multiple Imputation
VBPCA	Variational Bayesian PCA

Symbols and operators

x	Scalar value (except in Chapter 2)
$\boldsymbol{x}$	Vector with elements x_i
$\boldsymbol{x}_{\text{obs}}$	Observed elements of vector $\boldsymbol{x}$
$\boldsymbol{x}_{\text{mis}}$	Missing elements of vector $\boldsymbol{x}$
$\boldsymbol{x}^{\text{T}}$	Transpose of $\boldsymbol{x}$
$\hat{x}$	Estimate of x
$\mathbb{E}(x)$	Expected value of x
$\boldsymbol{X}$	Matrix with elements X_{ij}
$\boldsymbol{X}_{\text{obs}}$	Observed elements of matrix $\boldsymbol{X}$
$\boldsymbol{X}_{\text{mis}}$	Missing elements of matrix $\boldsymbol{X}$
θ	Model parameter(s)
$\mathcal{L}(\cdot)$	Log-likelihood function
$\langle\mathcal{L}(\cdot)\rangle$	Expected log-likelihood
$\mathcal{N}(\mu, \Sigma)$	Normal distribution with parameters μ and Σ
$p(x)$	Probability density of x
$p(x\|y)$	Probability density of x given y
$C(\cdot)$	Cost function
$\varepsilon(\cdot)$	Error function
$q(\theta)$	Approximation of the posterior probabilities in variational learning
$\text{KL}(q\|\|p)$	Kullback-Leibler divergence between q and p
$\text{var}(\cdot)$	Variance operator
$\langle\boldsymbol{W}\rangle$	Variational approximation of $\boldsymbol{W}$
$\text{tr}(\boldsymbol{X})$	Trace of matrix $\boldsymbol{X}$
$\boldsymbol{m}(t)$	Value of $\boldsymbol{m}$ at time t
$\boldsymbol{m}_i$	i^{th} model vector of the SOM or the GTM
$\boldsymbol{m}_{c(\boldsymbol{x}_n)}$	Best-matching unit of the data vector $\boldsymbol{x}_n$
h_{ci}	Neighborhood function of the SOM for units c and i
$\exp\{\}$	Exponent function
$\|\boldsymbol{x}\|$	l^2 norm of vector $\boldsymbol{x}$
$\phi(\cdot)$	Radial basis function (RBF)
$\boldsymbol{\phi}(\cdot)$	Vector of RBFs (for GTM)
$\boldsymbol{\Phi}$	Matrix of the latent points transformed with RBF network
n, N	Index of data vector and number of samples
i, K	Index of map unit and number of map units (for SOM and GTM)
k, D	Index of dimension, and dimensionality of the data (for SOM and GTM)
j, M	Index of RBFs and number of RBFs (for GTM)
α	Regularization parameter

Chapter 1

Introduction

Missing data are a common problem in many fields of human endeavor ranging from social sciences to economics and from political research to entertainment industry. In fields where conducting surveys or polls is commonplace, missing data occurs when people refuse to answer to specific questions or some people cannot be contacted. In the movie business, predicting customer preferences is literally a million dollar quest. The Netflix Prize (see, e.g., Koren, 2009) was an open competition to devise the best recommendation system to predict user ratings for films based on previous ratings.

Substituting missing values with predictions is called *missing value imputation* (Rubin, 1976, 1987). It is a task as old as evidence-based science itself with more complications than one may realize at first glance. First of all, one can quite rarely know what fundamentally gives rise to the missing values. Second, models for the missing data cannot be compared against any correct test data. Even if one tries to collect the missing data afterwards, the procedure itself might interfere with the results. Last but not least, even if one had the most perfect model or expert knowledge, substituting the missing values with the expected values or best guess might harshly bias the statistics of the whole data. However, there are many sophisticated methods trying to tackle the obstacles mentioned.

Wellbeing informatics is a field of science where the goal is to use the methods and know-how of computer scientists in order to facilitate wellbeing of people in large scale. In the wellbeing context, think of an employer who wants to evaluate stress and wellbeing of her employees. She conducts a survey which results in a simple score for each anonymous employee. Her ultimate goal is to determine the mean and the variance of the score. With these simple statistics she can, for example, monitor the development of the atmosphere of the working environment of her company annually. Unfortunately, some proportion of her employees refuse to answer the questionnaire. The simplest thing the employer can do, is to base her evaluation on the results at hand, that is, the observed scores. However, this may bias the results in many ways. If the nonrespondents are, for example, stressed people, the results of the analysis omitting nonrespondents are over-optimistic.

As a second example, think of an experiment where one wants to measure if a specific intervention affects physical fitness of people. The experimenter measures the fitness using a series of tests before and after the intervention and aims to conduct analysis of variance between these two sets of measurements. However, there are

some people omitting the test after the intervention. In this kind of situation it is questionable to conduct the analysis and make inferences based only on the observed data. The underlying reason for skipping the second test may be, for example, failure to improve one's fitness. In such a case, an analysis based on the observed value over-estimates the differences between the two groups.

In this thesis, missing value imputation is conducted using so called subspace methods. The name stems from the underlying assumption that high-dimensional data—that is, data with many variables—"lives" on some lower-dimensional manifold of the original data space. As a consequence, the data can be represented with fewer variables without losing much information. One way of understanding subspace methods is to view them as data compression methods, which naturally aim to preserve as much information as possible.

Depending on the framework, subspace methods also provide means to perform missing value imputation. For example, principal component analysis (PCA) has been used extensively on recommendation systems to predict movie ratings (see, e.g., Ilin and Raiko, 2010; Kozma et al., 2009; Yu et al., 2009). However, this collaborative filtering problem is quite different from survey imputation since in recommendation systems usually the most of the data are missing. Furthermore, subspace methods rarely belong to the repertoire of the majority of statisticians and other scienctist doing survey imputation (see, e.g., Su et al., 2011; de Leeuw and Zeileis, 2011). Thus, there is a clear gap between these two areas of research which are actually solving the same problem.

This thesis aims to map the uncharted area between survey imputation and recommendation systems. Even though a single thesis cannot join the two disciplines, I hope that some survey practitioners may, in the future, consult machine learning researchers for subspace methods in their survey imputation. Also, recommendation systems may benefit and get new ideas from the methods used in survey imputation. In more detail, the aim of this thesis is threefold: 1) describe the formalism behind the missing value imputation, 2) give an overview of widely-used missing value imputation techniques, and 3) investigate the use of subspace methods for the missing value imputation. The underlying research question is very practical: what is the best way to conduct the missing value imputation on survey data.

In this thesis, missing value imputation is studied using four data sets. For each data set, missing at random data is assumed, that is, the missingness mechanism giving raise to missing data is ignored. Nonignorable models, where missingness mechanisms are known or learned from the data, do not belong to the scope of this thesis. The first two data sets provide more theoretical view on single imputation, whereas two survey data sets demonstrate the practicalities and obstacles in survey imputation. Experiments with the nursing survey data (Mehtätalo and Lagus, 2011) are meant to motivate multiple imputation over single imputation in order to obtain reliable estimates of the sample level statistics. 15D instrument (Sintonen, 2001) survey data, measuring the *Health-Related Quality of Life* (HRQoL) of respondents, is used for demonstrating multiple imputation (MI) and pooling of the results in statistical testing.

This thesis was written in a research project called VirtualCoach (VirtualCoach; Lagus, 2011a,b). In the context of the this project, various researchers and experts

have collaboratively designed wellbeing related surveys including topics such as social isolation (Lagus and Saari, 2011), stress and relaxation (Lagus, Styrman and Izzatdust, 2011, unpublished), and nursing (Mehtätalo and Lagus 2011; see also Mehtätalo 2012; Lagus 2012). Some of these surveys were then implemented in the context of a questionnaire prototype designed within the project (Klapuri et al., 2011) and used for data collection by the researchers in appropriate user communities.

This thesis is organized as follows. In Chapter 2, I introduce various mechanisms giving rise to missing data and explain the premises leading to different imputation frameworks, namely single imputation and multiple imputation. Chapter 3 gives an overview of common missing value imputation techniques whereas Chapter 4 lays out the methodological details of the subspace methods concerned, namely Principal Component Analysis (PCA), Self-Organizing Map (SOM) and Generative Topographic Map (GTM). In Chapter 5, the missing value imputation is conducted for four different data sets moving from relatively simple single imputation tasks to more complicated multiple imputation. Chapter 6 consists of discussion and conclusions.

Chapter 2

Missing-Data Problem

Missing data may arise from numerous different reasons. In surveys, missing data is typically consequence of non-response; some questions might be irrelevant for some respondents, respondents may end up interrupting the survey or results of multiple different surveys with different questions may be analyzed together. In other contexts, missing data may be caused, for example, by equipment failure or data corruption.

In any setting with missing data, one should always consider the process which generated the missing values. This process is referred as *missing-data mechanism*. The missing-data mechanisms were first acknowledged and formalized by Donald Rubin in Rubin (1976) and later refined in Rubin (1987). Four types of missingness mechanisms, moving from the simplest to the most general, are:

1. *Missingness completely at random.* When the probability of missingness is completely independent of observed data and any latent variables, the data is said to be *missing completely at random* (MCAR). Sometimes it is also said to be *observed at random*. Survey data is rarely MCAR but if respondents roll a die to decide whether they answer to a specific question, the resulting data is MCAR. Data corruption may produce MCAR data, as well.

2. *Missingness at random.* A relaxed assumption, compared to MCAR, is one where the probability of missingness depends only on observed information. It is said that data is *missing at random* (MAR). In survey data, this means that response probability on a specific question depends on other fully observed variables. For example, old people might be less desirous to answer questions about their sexual activity. However, if the data contains age of all respondents, this data may still be regarded as MAR (if it is reasonable to assume that the response of the sexual activity question itself does not affect to probability of missingness. See Section 5.4.)

3. *Missingness that depends on unobserved predictors.* When missingness depends on some unobserved information the missing data is no longer MAR. This is also the first case where data is not missing at random (NMAR). For example, suppose that depressed people are less likely to report their annual income and a survey does not cover mental state in any way. Now, the depression or the mental state of respondents is predictive of income and the mental state is unobserved. Hence, annual income is NMAR.

4. *Missingness that depends on the missing value itself.* Finally, the most complicated scenario occurs when the probability of missingness depends on the missing value itself. This kind of missingness is also called *censoring*. Censoring occurs, for example, if out-of-range values are marked as unobserved. As another example, one might assume that people with higher income are less likely to report their earnings.

In the literature, the NMAR mechanism is often said to consist of the latter two types (3. and 4.) together. Understanding the missing-data mechanisms and considering which class the missing data mechanism at hand falls into, is of paramount importance when working with data containing missing values. Generally, one cannot be sure—or in particular, prove—whether data is MAR or not. There may always be unobserved predictors which are—by definition—unknown to the observer. In practice, one is advised to include as many predictors as possible in a model, which is more likely to make MAR assumption reasonable (Gelman and Hill, 2007).

Rubin (1987) proposed treating missing-data indicators as random variables and assigning them a distribution. Hence, the missingness can be seen as a consequence of a random process that can be characterized by a missing-data model. In this chapter, the denotation that has come into common use in modern statistics literature of missing data is used. However, the notation and terminology differs slightly in different textbooks.

Let $y = (y_{\text{obs}}, y_{\text{mis}})$ represent the complete data where y_{obs} denotes the observed data and y_{mis} denotes the missing values. The notation is general—y may be a vector of univariate measurements or an $n \times d$ matrix of n observations of d dimensional multivariate responses—in order to keep equations uncluttered. Let the inclusion indicator I represent a data structure of the same size as y with each element of I equal to 1 indicating that the corresponding component of y is observed and 0 that it is missing. Usually I is completely observed. Now the joint distribution can be written as

$$p(y, I|\theta, \phi) = p(y|\theta)p(I|y, \phi), \tag{2.1}$$

where θ and ϕ are model parameters. More specifically, θ represents the parameters of the complete-data model and ϕ together with y govern the missing-data mechanism. In this general setting, the distribution of the observed data is obtained by integrating over the distribution of y_{mis}:

$$p(y_{\text{obs}}, I|\theta, \phi) = \int p(y_{\text{obs}}, y_{\text{mis}}|\theta)p(I|y_{\text{obs}}, y_{\text{mis}}, \phi)dy_{\text{mis}}. \tag{2.2}$$

Now we are ready to see how this general equation can be written down under different missingness mechanisms.

2.1 Missing-at-Random Assumption

When data is *missing completely at random*, data missingness is completely independent of y:

$$p(I|y, \phi) = p(I|\phi). \tag{2.3}$$

Under MCAR, the joint distribution (2.2) can be written as

$$\begin{aligned} p(y_{\text{obs}}, I|\theta, \phi) &= p(I|\phi) \int p(y_{\text{obs}}, y_{\text{mis}}|\theta) dy_{\text{mis}} \\ &= p(I|\phi) p(y_{\text{obs}}|\theta). \end{aligned} \tag{2.4}$$

This implies that one can totally ignore the presence of missing values and base inferences on the observed data. However, it is relatively rare in practical problems for MCAR to be plausible.

Under MAR assumption, Equation (2.2) can be factored as follows:

$$p(y_{\text{obs}}, I|\theta, \phi) = p(I|y_{\text{obs}}, \phi) p(y_{\text{obs}}|\theta). \tag{2.5}$$

This means that for maximum likelihood (ML) techniques which maximize the likelihood of the parameters, $\mathcal{L}(\theta, \phi|y_{\text{obs}}, I) \propto p(I|y_{\text{obs}}, \phi) p(y_{\text{obs}}|\theta)$, it is sufficient to maximize

$$\mathcal{L}(\theta|y_{\text{obs}}) \propto p(y_{\text{obs}}|\theta), \tag{2.6}$$

provided that one is only interested in the model parameters θ. In other words, the missing-data mechanism can be ignored for purposes of estimating θ.

For Bayesian methods, the posterior probability of the model parameters is

$$p(\theta, \phi|y_{\text{obs}}, I) \propto p(y_{\text{obs}}, I|\theta, \phi) p(\theta, \phi). \tag{2.7}$$

If one further assumes, that the parameters governing the missing data mechanism, ϕ, and the parameters of the data distribution, θ, are independent in the prior distribution, i.e., $p(\theta, \phi) = p(\theta)p(\phi)$, the missing data mechanism can be ignored in the Bayesian framework, as well.

In the previous situations, the maximum likelihood and Bayesian settings, the missing-data mechanism is said to be *ignorable*, that is, the missing-data mechanism can be ignored. The results above also imply that NMAR data cannot be handled by Bayesian or likelihood-based methods unless a model of the missing data mechanism is also learned or known.

When the missing data is *nonignorable*, that is, the missing-data mechanism cannot be ignored, missing value imputation becomes an involved task. There are some situations, where the missing data mechanism is nonignorable but known, for example, censoring where values above a known threshold are missing. However, analysis of NMAR data requires careful data-specific process. In this thesis, only ignorable models are used. Nonignorable missing data models could be a thesis topic of its own. Theory behind nonignorable models can be found in Rubin (1987) and Little and Rubin (2002).

2.2 Single Imputation

Single imputation is the most straight-forward and at the same time the most dangerous way of dealing with missing values. In single imputation, one filled-in data set is created by replacing each missing value with one predicted value. This approach

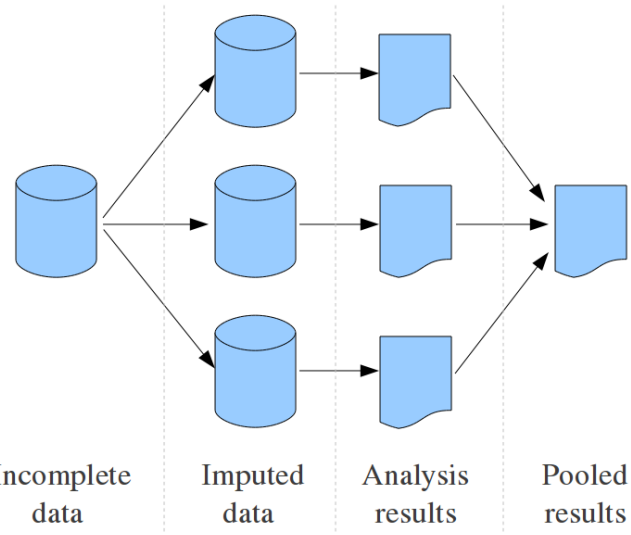


Figure 2.1: Schematic illustration of multiple imputation paradigm. Multiple imputed data sets, which simulate the uncertainty of the imputed values, can be analyzed using standard complete-data methods. The missing-data uncertainty is incorporated in the pooled results which are obtained by combining the multiple analysis results.

does not consider the uncertainty of the imputed values, hence any statistical analysis performed for the complete data may be biased. However, in some cases single imputation may be efficient and well-justified approach. For example, if one is interested in the data on the observation level, the primary interest may be the most probable value of each missing observations. This is usually the case in the collaborative filtering task, where the imputed values can be, for example, movie ratings.

2.3 Multiple Imputation

To overcome the difficulties raised in single imputation, Rubin (1987) formulated multiple imputation (MI) paradigm. In MI, each missing value is replaced with multiple imputed values, creating several simulated complete data sets. Due to the high computational complexity of many MI techniques, typical number of simulated draws is between 3 and 10. Each filled-in data set is then analyzed by standard methods and the results are combined in order to obtain pooled estimates and confidence intervals that incorporate missing-data uncertainty. Figure 2.1 illustrated the MI paradigm.

Multiply imputed data sets can be used to make inferences on any scalar quantities of the complete data $y = (y_{\text{obs}}, y_{\text{mis}})$, such as a mean or regression coefficient. Let Q denote any such quantity and suppose one obtained $m > 1$ simulated draws of the missing data y_{mis}. The overall estimate of Q is simply the average

$$\bar{Q} = \frac{1}{m} \sum_{i=1}^{m} \hat{Q}_i, \tag{2.8}$$

where $\hat{Q}_i$ is the estimate of Q evaluated using the i^{th} imputed data set. In order to

estimate the standard deviation of $\bar{Q}$, one has to take into account both the between-imputation variance $B = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{Q}_i - \bar{Q})^2$ and the within-imputation variance $\bar{U} = \frac{1}{m}\sum_{i=1}^{m} \mathrm{var}(y_{\mathrm{obs}}, y_{\mathrm{mis},i})$. The estimated total variance is

$$\mathrm{var}(\bar{Q}) = \frac{1}{1+m}B + \bar{U}. \tag{2.9}$$

Equations (2.8) and (2.8) are derived using Bayesian framework in Rubin (1987). Concise reviews of the MI paradigm can be found in Rubin (1996), Schafer (1999) and He (2010).

Chapter 3

Missing-Data Imputation

Missing data imputation methods can be divided roughly into three groups: a) single imputation methods, b) multiple imputation methods, and c) likelihood-based methods. This division is in no means general or well-established but it is apt for the purposes of this thesis. Furthermore, many methods can be used for both single and multiple imputation and the models acquired using the likelihood-based methods can be used to do single or multiple imputation, as well.

The structure of this chapter roughly follows the division above. The naive methods in 3.1 exploit heuristics or make crude simplifications. The mean imputation is probably the most simple single imputation method and it is used as a baseline measure in the experiments. *Multiple imputation using chained equations (MICE)* described in Section 3.2 is the most common choice for MI within survey administrators. This chapter is concluded by describing the expectation-maximization (EM) algorithm for missing data as an example of likelihood-based methods.

3.1 Naive Methods

3.1.1 Mean Imputation

Mean imputation substitutes every missing value with the mean of the observations. If data is MCAR, the average of the observed values is the true expectation of the missing values. However, even in this case the filled-in data set underestimates the variance of the complete data by a factor of $(n_{\text{mis}} - 1)/(n - 1)$, where n_{mis} is the number of missing values. Many heuristic improvements to mean imputation have been proposed, for example, using within class means for data with clear cluster structure. (Little and Rubin, 2002)

3.1.2 Regression Imputation

Regression imputation replaces missing values with predictions made by a regression model of the missing value on variables observed for the data vector. The regression model may be rudimentary linear regression, $y_{\text{mis}} = \sum_k \beta_k y_{\text{obs}}$, or any generalized linear model, such as logistic or probit regression. Moreover, in many applications it is possible to build better regression models by including interactions and nonlinear terms but this kind of analysis requires expertise of its own. (Gelman and Hill, 2007)

If there are missing values in more than one variable, some workaround has to be considered. One possibility is to arrange the variables in monotonically increasing order with respect to the number of missing values and first impute the variable with least missing values using the fully observed variables as predictors. When the imputation proceeds to variables with more missing values, filled-in variables can be used as predictors. This procedure is then repeated until all the variables are imputed. More general approach, known as MICE, is described in the next section.

3.1.3 Hot Deck Imputation

If there is abundance of complete data, hot-deck imputation where one substitutes missing values according to data vectors with similar observed values is an attractive alternative. Hot deck imputation has been widely used in survey practice and it may involve very elaborate heuristics for selecting units for best match. However, there is very little theoretical results on the properties of hot deck heuristics. For discussion of hot deck applications, see Marker et al. (2002).

3.2 Multiple Imputation by Chained Equations

Multiple imputation by chained equations (MICE), also known as sequential regression multiple imputation (SRMI), is an imputation framework, where each variable with missing data is characterized by a separate conditional linear model (Buuren et al., 1999; Raghunathan et al., 2001). For each model, all variables apart from the predicted variable itself can be used as predictors. The models are used to impute one variable at the time, and imputed values are used as predictors in other models. The procedure is continued until the model parameters or imputed data distribution reach convergence. Recently, Su et al. (2011) have shown how the standard convergence measures used to evaluate Markov Chain Monte Carlo (MCMC) convergence can be used in MICE. As in regression imputation, common guidelines of regression modeling ought to be followed when characterizing the conditional models. Thus, conducting careful imputation using MICE may be time consuming, but the results are usually good. Compared to the joint modeling (see next section), it is usually easier to accommodate complex data features in univariate regression models allowing more flexible models.

There exists many good implementations of MICE for the standard statistical software such as SAS, S-plus and R. In this thesis, mice 2.9 for R (van Buuren and Groothuis-Oudshoorn, 2011) was used. However, since the regression modeling is not in the focus of this thesis, only rudimentary linear models lacking interaction terms, nonlinearities or transformed predictors were used.

3.3 Likelihood-Based Methods

The underlying idea in the likelihood-based methods is to approximate the distribution of the complete data y using a parametric probability density $p(y|\theta)$. In order to compensate for the missing data, one has to integrate over the distribution of missing values, as in Equation (2.2). Under the MAR assumption, factorization (2.5) can be exploited, hence one can maximize the likelihood $\mathcal{L}(\theta|y_{\text{obs}})$. The most difficult part

in this approach is usually the definition of the joint model $p(y|\theta)$. A downside of the likelihood-based methods is, that they are usually applicable only with very simple data which can be modeled with the standard parametric probability distributions. If the data is complex and contains many variables of a different type, there is usually no means to approximate the distribution of the data with any known parametric probability density model.

3.3.1 Expectation-Maximization Algorithm

The expectation-maximization (EM) algorithm is an iterative method for solving the ML estimates of parameters in probabilistic models with unobserved latent variables or missing values among the observed data (Dempster et al., 1977). The algorithm proceeds in two steps. The *expectation step* (E-step) evaluates the posterior probabilities of the unobserved data. The subsequent *maximization step* (M-step) updates the model parameters using the posterior distribution of the missing data evaluated in the E-step. The EM algorithm is summarized in Algorithm 3.1.

Algorithm 3.1 The EM Algorithm for data with missing values

Given a joint distribution $p(y_{\text{obs}}, y_{\text{mis}}|\theta)$ over the observed values y_{obs} and the missing values y_{mis}, and the model parameters θ, the goal is to maximize the likelihood function $p(y_{\text{obs}}|\theta)$ with respect to θ.

1. Choose an initial setting for parameters θ.
2. E-step, evaluate $p(y_{\text{mis}}|y_{\text{obs}}, \theta^{\text{old}})$
3. M-step, evaluate θ^{new} given by

   $$\theta^{\text{new}} = \arg\max_{\theta} \mathcal{Q}(\theta, \theta^{\text{old}})$$

   where

   $$\mathcal{Q}(\theta, \theta^{\text{old}}) = \int p(y_{\text{mis}}|y_{\text{obs}}, \theta^{\text{old}}) \ln p(y_{\text{obs}}, y_{\text{mis}}|\theta) dy_{\text{mis}}$$

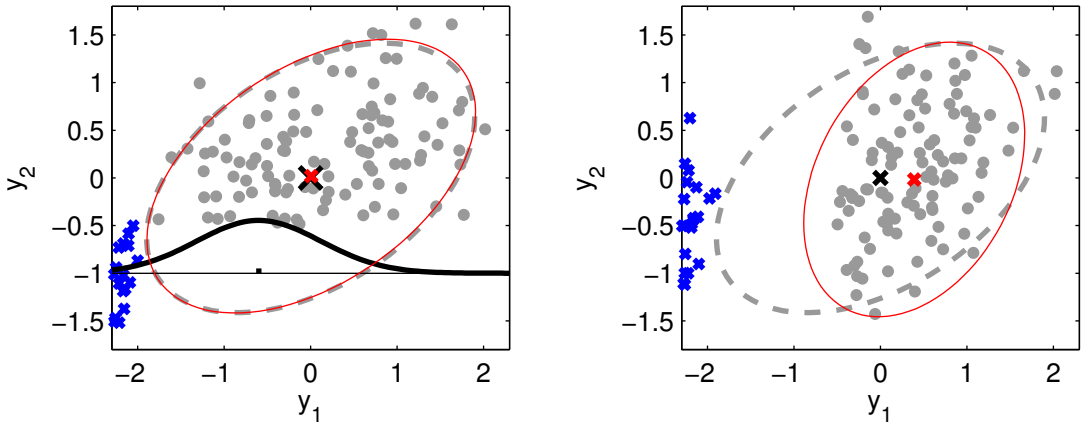
4. Check for convergence of either the log likelihood or the parameter values. If the convergence criterion is not satisfied, then let

   $$\theta^{\text{old}} \leftarrow \theta^{\text{new}}$$

   and return to Step 2.

3.3.2 Example: Multivariate Normal Data

Let us now examine an example with bivariate normal data with missing values resulting from two simple missingness mechanisms. In the first scenario depicted in Figure 3.1(a) the data is MAR such that y_1 is missing iff $y_2 < -0.5$. Moreover, data is arranged such that the values of y_1 are missing for $i = (r+1), \ldots, n$. The



(a) MAR data

(b) NMAR data

Figure 3.1: An example of estimation of data mean μ and covariance Σ using a bivariate Gaussian data with some values y_2 missing. In both figures the gray dots depict the fully-observed values and the jittered blue crosses depict the data having only y_2 observed. Gray dashed ellipse and black cross represents the covariance and mean of the distribution used to generate the data. (a) The data is MAR: y_1 is missing iff $y_2 < -0.5$. The Red cross and ellipse show the mean and covariance of the complete data estimated using the ECM algorithm. The black curve shows a conditional distribution $p(y_1|y_2 = -1)$. (b) The data is NMAR: y_1 is missing iff $y_1 < -0.5$. The ECM algorithm fails to estimate the mean and the covariance of the complete data.

log-likelihood ignoring the missing-data mechanism is

$$\mathcal{L}(\mu, \Sigma|y_{\text{obs}}) = -\frac{1}{2}\ln|\Sigma| - \frac{1}{2}\sum_{i=1}^{r}(y_1 - \mu)\Sigma^{-1}(y_i - \mu)^T - \frac{1}{2}(n-r)\ln\sigma_{11} - \frac{1}{2}\sum_{i=r+1}^{n}\frac{(y_{i1} - \mu_1)^2}{\sigma_{11}}. \quad (3.1)$$

Careful examination of the equation reveals that there is no analytical solution for the maximum of (3.1). However, it can be be maximized incrementally, for example, by using the expectation conditional maximization (ECM) algorithm (Meng and Rubin, 1993). The red cross and ellipse in Figure 3.1(a) show the estimated mean and covariance of the data using the ECM algorithm. It can be seen that the estimates are very close to the true values which were used the generate the data shown with the dashed gray ellipse and the black cross. A part of the data with missing y_1 is shown on the y_2-axis with blue crosses and some jitter added to ease the readability. Finally, the solid black curve shows a conditional distribution $p(y_1|y_1 = -0.5)$ which can be used to make inferences on data with missing y_1.

Figure 3.1(b) shows a resembling data, but now y_1 is missing iff $y_1 < -0.5$, that is, the missingness depends on the missing variable itself, hence the data is NMAR. As can be seen from the figure, the similar approach fails to estimate the complete-data distribution depicted by the dashed gray ellipse.

Chapter 4

Subspace Methods

Subspace methods refer to a collection of methods where the underlying assumption is that the data "lives" on a lower-dimensional manifold or surface embedded in the higher dimensional, original vector space. Thus, by representing the data on this manifold one can efficiently reduce the dimensionality of the data. This, in turn, can help tackling the *curse of the dimensionality* (Bellman, 1961).

In general, the objective of dimensionality reduction is to find a mapping from the original d-dimensional space to a k-dimensional subspace where $k < d$. Based on the properties of this mapping, subspace methods can be divided into linear and nonlinear methods. In linear methods, the lower-dimensional manifold is restricted to be a linear subspace. Nonlinear methods extend the set of possible surfaces to contain nonlinear manifolds with application specific restrictions.

In this thesis, three subspace methods, namely principal component analysis (PCA), the Self-Organizing Map (SOM) and the Generative Topographic Mapping (GTM) are used. PCA is an old and well-known dimensionality reduction technique which has been extended into a statistical model (Tipping and Bishop, 1999a). Also, the GTM is inherently statistical, meaning that it seeks to learn a probability distribution which resembles the distribution of the data. The SOM is rather an engineering solution and is not anchored to the statistical framework. This chapter lays out the methodological basis of the methods above and their application in presence of missing data. Moreover, possible extensions are proposed.

4.1 Principal Component Analysis

Principal component analysis (PCA) (Jolliffe, 2002) is a technique which can be used to compress data with high dimensionality by using a lower dimensional presentation computed in such a way that a minimum amount of information is lost. As such, PCA is an example of dimensionality reduction techniques where the task is to find a mapping from the original d-dimensional space to a k-dimensional subspace where $k < d$. In addition to dimensionality reduction and data compression, PCA is widely used for other applications such as feature extraction, data visualization, image processing, pattern recognition and time-series prediction.

The most common formulation of PCA, the maximum variance formulation, defines PCA as an orthogonal projection of the data onto a lower dimensional linear

space, *principal subspace*, in which the variance of the data is maximized, that is, the maximum amount of information is preserved (Hotelling, 1933). It can be shown that the optimal projection into a k-dimensional subspace is such that we choose k eigenvectors $\{\boldsymbol{w}_j\}$, $j = 1, \ldots, k$, of the data covariance matrix $\boldsymbol{S} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \bar{\boldsymbol{x}})(\boldsymbol{x}_n - \bar{\boldsymbol{x}})^{\mathrm{T}}$ corresponding to the k largest eigenvalues $\lambda_1, \ldots, \lambda_k$. Now, a linear transformation of a data vector $\boldsymbol{x}_n$ onto the principal subspace defined by the k eigenvectors is simply the product

$$\boldsymbol{z}_n = \boldsymbol{W}^{\mathrm{T}}(\boldsymbol{x}_n - \boldsymbol{\mu}), \tag{4.1}$$

where $\boldsymbol{z}_n$ are called the z-scores for the data vector, the k columns of $\boldsymbol{W}$ are the k leading eigenvectors of $\boldsymbol{S}$ and $\boldsymbol{\mu}$ is the bias term, that is, the mean of the data.

A complementary property of PCA is that it finds a k-dimensional linear representation of data such that the squared error of the reconstructed data

$$\hat{\boldsymbol{x}}_n = \boldsymbol{W}\boldsymbol{z}_n + \boldsymbol{\mu}. \tag{4.2}$$

is minimized. This formulation is also related to the original discussions of Pearson (1901).

4.1.1 Probabilistic PCA

One significant limitation of the conventional PCA is that it does not define a probability distribution. The probabilistic PCA (PPCA) (Tipping and Bishop, 1999a) offers a cure by introducing a generative latent variable model shown in Figure 4.1 (Figure is discussed shortly). Figure 4.1 corresponds to mathematical formulation of PPCA,

$$\boldsymbol{x}_n = \boldsymbol{W}\boldsymbol{z}_n + \boldsymbol{\mu} + \varepsilon_n, \tag{4.3}$$

which extends (4.2) by adding a noise term ε_n. The original authors (Tipping and Bishop, 1999a) showed that the ML solution of the model (4.3) extracts the principal components of the data.

The graphical representation in Figure 4.1 is also known as a plate diagram. In the Figure, the nodes represent random variables and the connecting edges represent their relationships. Shading of a node means that the corresponding variable is being observed. Unshaded variables are unobserved or hidden. Deterministic parameters are shown explicitly by small solid nodes. The plate (the box labeled N) implicates that there are N observation of variables z and x. Koller and Friedman (2009) provide a detailed description of graphical models and their applications.

The common choice of isotropic Gaussian noise model $\boldsymbol{\varepsilon}_n \propto \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ leads to the probability distribution

$$p(\boldsymbol{x}|\boldsymbol{z}) \propto \mathcal{N}(\boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu}, \sigma^2\boldsymbol{I}). \tag{4.4}$$

The marginal distribution of $\boldsymbol{x}$ is likewise Gaussian

$$p(\boldsymbol{x}) \propto \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{C}), \tag{4.5}$$

where $\boldsymbol{C} = \boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} + \sigma^2\boldsymbol{I}$.

The probabilistic formulation offers a number of benefits including well-founded regularization, model comparison and extensions, such as mixtures of principal component analyzers (Tipping and Bishop, 1999b).

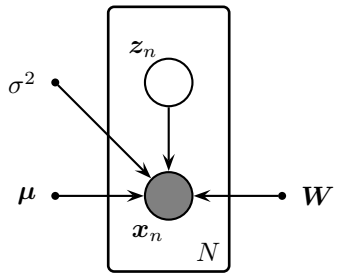


Figure 4.1: The plate diagram of PPCA.

4.1.2 Variational Bayesian PCA

Bayesian PCA treats the model parameters $\boldsymbol{W}, \boldsymbol{\mu}$ as random variables and introduces additional hyperparameters governing the distributions of the model parameters. This kind of approach allows controlling the effective dimensionality of the latent space corresponding to the number of retained principal components. In other words, this way one can avoid discrete model selection and automatically determine the appropriate dimensionality for the latent space as a natural part of the process called Bayesian inference. This kind of model selection is also called *automatic relevance determination* (unpublished work by MacKay, 1995; Neal, 1996) first introduced in the context of neural networks.

The goal above is achieved using a prior $p(\boldsymbol{W}|\boldsymbol{\alpha})$ over the matrix $\boldsymbol{W}$, governed by a k-dimensional vector of hyperparameters $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_k\}$. In his original work, Bishop (1999) treated α as a random variable following a Gamma distribution governed by further hyperparameters. In this thesis, Gaussian distribution of the form

$$p(\boldsymbol{W}) \propto \prod_{c=1}^{k} \mathcal{N}(0, \alpha_c) \tag{4.6}$$

is used. Above, each hyperparameter α_c controls a single principal component, that is, a column of the matrix $\boldsymbol{W}$. Prior over the mean vector $\boldsymbol{\mu}$ is given by

$$p(\boldsymbol{\mu}) \propto \mathcal{N}(\boldsymbol{\beta}_\mu, \beta_{\sigma^2}\boldsymbol{I}). \tag{4.7}$$

Figure 4.2 shows the plate diagram of full Bayesian PCA with hyperparameters $\boldsymbol{\alpha}, \boldsymbol{\beta}_\mu$ and β_{σ^2}.

In order to use the model above, one must be able to compute integral

$$p(\boldsymbol{x}|\boldsymbol{z}) = \iint p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{W}, \sigma^2) p(\boldsymbol{\mu}|\boldsymbol{\beta}_\mu, \beta_{\sigma^2}) p(\boldsymbol{W}|\boldsymbol{\alpha})\, d\boldsymbol{W}\, d\boldsymbol{\mu} \tag{4.8}$$

which is analytically intractable. The problem can be approached, for example, using MCMC sampling or different approximation techniques.

Variational approximation offers one way to approximate such intractable distributions. The term *variational methods* refers to a large collection of optimization techniques which have been developed for finding the extremum of an integral depending on an unknown function and its derivatives (Jaakkola, 2000). In this thesis,

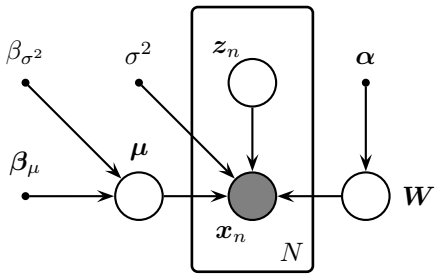


Figure 4.2: The plate diagram of Bayesian PCA.

variational learning is harnessed to approximate the posterior distribution in the E-step of the EM-algorithm with a simpler pdf

$$p(\theta|\boldsymbol{X},\gamma) \approx q(\theta), \tag{4.9}$$

where $\theta = (\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\mu})$ are the unobserved random model parameters and $\gamma = (\boldsymbol{\alpha}, \boldsymbol{\beta}_\mu, \beta_{\sigma^2})$ are the deterministic (hyper)parameters. The E-step in the variational approach updates the approximation $q(\theta)$ by minimizing the cost function

$$C(q(\theta),\gamma) = \int q(\theta) \log \frac{q(\theta)}{p(\boldsymbol{X},\theta|\gamma)} d\theta = \int q(\theta) \log \frac{q(\theta)}{p(\theta|\boldsymbol{X},\gamma)} d\theta - \log p(\boldsymbol{X}|\gamma), \tag{4.10}$$

where the first term is the Kullback-Leibler divergence $\mathrm{KL}(q||p)$ between the approximation and the true posterior. Since the second term is constant with respect to $q(\theta)$, in the E-step the goal is to find $q(\theta)$ that minimizes the $\mathrm{KL}(q||p)$, which is a (non-symmetric) measure of difference between two probability distributions. In the subsequent M-step, this approximation is used to the compute ML estimate of γ using $p(\boldsymbol{X}|\gamma)$, which can be seen as minimizing the cost function (4.10) with respect to γ (Neal and Hinton, 1999).

In this thesis, an implementation of variational Bayesian PCA (VBPCA) by Ilin and Raiko (2010) is used. The complete update equations as well as a broad review of different PCA variants can be found in the corresponding article (Ilin and Raiko, 2010). For comprehensive discussion on variational learning in Bayesian framework, see, for example, Bishop (2007).

4.1.3 PCA with Missing Data

In the conventional PCA, there is no obvious way of dealing with missing values. A technique later referred to as the *imputation algorithm* (see, e.g., Jolliffe, 2002; Ilin and Raiko, 2010) is described here. Similar ideas are also exploited in the context of the SOM in Section 4.2.3.

The imputation algorithm is an iterative procedure where one alternates between imputing the missing values in the data, $\boldsymbol{X}_{\mathrm{mis}}$ and applies the standard PCA to the complete data matrix. Initial values of the missing elements $\boldsymbol{X}_{\mathrm{mis}}$ can be set to, for example, row-wise means of $\boldsymbol{X}$. Also, the bias term $\boldsymbol{\mu}$ has to be updated on every iteration. The resulting algorithm is summarized in Algorithm 4.1.

Algorithm 4.1 The imputation algorithm for PCA.

Given an incomplete data $\boldsymbol{X}$ with observed elements $\boldsymbol{X}_{\text{obs}}$ and missing elements $\boldsymbol{X}_{\text{mis}}$.

1. Initialize the missing elements $\boldsymbol{X}_{\text{mis}}$, e.g., using the row-wise means of $\boldsymbol{X}_{\text{obs}}$

   $$\boldsymbol{X}_{\text{mis}} \leftarrow \texttt{mean}(\boldsymbol{X}_{\text{obs}})$$

   resulting in imputed complete data $\boldsymbol{X}_{\text{imp}}$.

2. Update the bias term $\boldsymbol{\mu}$ using the imputed complete data:

   $$\boldsymbol{\mu} \leftarrow \texttt{mean}(\boldsymbol{X}_{\text{imp}})$$

3. Solve k principal components $\boldsymbol{W}$ by using any known complete data technique.

4. Update the missing elements

   $$\boldsymbol{X}_{\text{mis}} \leftarrow \boldsymbol{W}\boldsymbol{W}^{\text{T}}(\boldsymbol{X}_{\text{imp}} - \boldsymbol{\mu}) + \boldsymbol{\mu},$$

5. Check for convergence of either $\boldsymbol{X}_{\text{imp}}$ or $\boldsymbol{Z} = \boldsymbol{W}^{\text{T}}(\boldsymbol{X}_{\text{imp}} - \boldsymbol{\mu})$. If the convergence criterion is not satisfied return to Step 2.

4.1.4 VBPCA with Missing Data

The PPCA offers a probabilistic model in which the missing values can be handled by integrating them out. The original authors illustrated an application with missing values in (Tipping and Bishop, 1999a). In VBPCA, the treatment of the missing data becomes more involved. The complete equations for the VBPCA with missing data can be found in (Ilin and Raiko, 2010). The reconstruction of the missing values in single imputation is obtained using (4.2), where $\boldsymbol{W}$, $\boldsymbol{z}_n$ and $\boldsymbol{\mu}$ are replaced with the respective variational approximations $\langle \boldsymbol{W} \rangle$, $\langle \boldsymbol{z}_n \rangle$ and $\langle \boldsymbol{\mu} \rangle$. In multiple imputation, the missing values are drawn from

$$\mathcal{N}(\hat{x}_{nk}, \widehat{\text{var}}(x_{nk})), \tag{4.11}$$

where $\widehat{\text{var}}(x_{nk})$ is the estimated uncertainty of the corresponding missing value

$$\widehat{\text{var}}(x_{nk}) = \widehat{\text{var}}(\mu_k) + \langle \boldsymbol{w}_k \rangle^{\text{T}} \Sigma_{\boldsymbol{z}_n} \langle \boldsymbol{w}_k \rangle + \langle \boldsymbol{z}_n \rangle^{\text{T}} \Sigma_{\boldsymbol{w}_k} \langle \boldsymbol{z}_n \rangle + \text{tr}(\Sigma_{\boldsymbol{z}_n} \Sigma_{\boldsymbol{w}_k}). \tag{4.12}$$

Above, all sources of uncertainty are taken into account; $\widehat{\text{var}}(\mu_k)$ represents the uncertainty of the bias term, and $\Sigma_{\boldsymbol{z}_n}$ and $\Sigma_{\boldsymbol{w}_k}$ represent the uncertainty of the parameters $\boldsymbol{z}_n$ and $\boldsymbol{w}_k$, respectively.

4.2 Self-Organizing Map

The *Self-Organizing Map* (SOM) (Kohonen, 2001) is an unsupervised artificial neural network which aims to discover some underlying structure in the data. The SOM is said to be topology preserving, that is, it has an explicit neighborhood function that preserves neighborhood relations of the neurons. One intuitive view of the SOM is, that it extends classical vector quantization (Gersho and Gray, 1991) by defining the neighborhood relations of the codebook vectors. However, this simple extension yields many useful properties. As a results, the theory and applications of the SOM have been a topic of active research for about three decades.

Neurons of the SOM are usually called map units or prototypes since they can be seen to be representative samples of the data (cf. codebook vectors in vector quantization). Each map unit is associated with a reference vector $\boldsymbol{m}_i$ and each data vector is mapped to a map unit whose reference vector is most similar to the data vector itself. The reference vectors $\boldsymbol{m}_i$ are usually—either emergently or explicitly—weighted local averages of the data associated with the given map unit in the original data space.

The SOM is useful for making low dimensional, usually two-dimensional, representations and visualizations of high-dimensional data. It provides a topology-preserving mapping from the original data space to the map units. If the map units are arranged to form a two-dimensional lattice this provides means to visualize the data on a plane. A SOM-type mapping has also been adapted to arbitrary data for which the mutual pairwise distances are defined (Kohonen and Somervuo, 2002).

4.2.1 SOM algorithm

Let us define some notation for the mathematical description of the SOM algorithm. In this work, the SOM with two-dimensional array of units is used, hence the SOM defines a mapping from the input data space onto a two-dimensional plane. Every map unit i has a parametric reference vector (model vector) $\boldsymbol{m}_i \in \mathbb{R}^d$, where d is the dimensionality of the data. Here, the reference vector $\boldsymbol{m}_i$ is used to refer to the units and their model vectors interchangeably. Many types of lattices can be used for the array of units but in this work a hexagonal grid is used.

Let $\boldsymbol{X} \in \mathbb{R}^{N \times d}$ be training data matrix with N samples of dimensionality d. Each data vector $\boldsymbol{x}_n$ may be compared with all the reference vectors $\boldsymbol{m}_i$ in any metric. Here, as in many other practical applications, Euclidean distance is used and the unit with the smallest Euclidean distance

$$\boldsymbol{m}_{c(\boldsymbol{x}_n)} = \arg\min_i \|\boldsymbol{x}_n - \boldsymbol{m}_i\| \tag{4.13}$$

is referred to as the *best-matching unit* of the data vector $\boldsymbol{x}_n$.

The learning starts by initializing the reference vectors $\boldsymbol{m}_i(t = 0)$, where $t = 0$ refers to a discrete-time variable representing the time scale of the training. The initialization can be done, for example, randomly or spreading the reference vectors on a plane defined by two first principal components of the data. The latter approach is used in this work. The actual learning process updates the reference vectors using

equation

$$\boldsymbol{m}_i(t+1) = \boldsymbol{m}_i(t) + h_{ci}(t)\left(\boldsymbol{x}(t) - \boldsymbol{m}_i(t)\right), \tag{4.14}$$

where $h_{ci}(t)$ refers to a neighborhood function defined over the lattice of map units. Here, widely used Gaussian neighborhood function is used:

$$h_{ci} = \alpha(t) \cdot \exp\left\{-\frac{\|r_c - r_i\|^2}{2\sigma^2(t)}\right\}, \tag{4.15}$$

where $\|r_c - r_i\|$ is the distance between the best-matching unit r_c and unit i in the array, $0 < \alpha(t) < 1$ is scalar-valued learning-rate factor and $\sigma(t)$ is the width of the neighborhood kernel. For convergence it is necessary that $h_{ci} \to 0$ when $t \to \infty$ and usually both $\alpha(t)$ and $\sigma(t)$ are decreasing monotonically in time.

For convenience, let us also define the update rule for so called *batch map* as it is easier to develop further with missing values using this notation. The basic idea is that while updating the reference vectors, all data (or a batch of the data) is taken into account at once and weighted accordingly. The batch update rule is

$$\boldsymbol{m}_i = \frac{\sum_n h_{ni}\boldsymbol{x}_n}{\sum_j h_{ni}}, \tag{4.16}$$

where and the index n runs over the data vectors whose best-matching units satisfy $h_{ni} > 0$, that is, all data points up to the range of the neighborhood function are taken into account.

4.2.2 Quality and Size of SOM

Selecting the size of the array of map units in the SOM is a subtle task since SOM can be used for different purposes. The question of the size can be approached from the point of view of different quality measures. Two most commonly used error measures are the *quantization error* and the *topological error*. The former measures the mean of the reconstruction errors $\|\boldsymbol{x} - \boldsymbol{m}_c\|$ when each data point used in learning is replaced by its best-matching unit, while the latter measures the proportion of data points for which the two nearest map units are not neighbors in the array topology. As the number of map units increases, quantization error decreases and topological error tends to increase. Hence, there is no straightforward way of choosing the number of map units based on the measures above.

Kaski and Lagus (1996) proposed combining the errors above by computing the sum of the quantization error and the distance from the best-matching unit to the second-best-matching unit of each data vector along the shortest path following the neighborhood relations. This measure is defined with rigorous mathematical notation in (Kaski and Lagus, 1996). In this work, this kind of error is referred to as the *combined error*.

4.2.3 SOM with Missing Values

SOM has been used for missing value imputation with many kinds of data, such as survey data (Fessant and Midenet, 2002; Wang, 2003), socio-economic data (Cottrell

and Letrémy, 2007; Gaubert et al., 1996), industrial data (Rustum and Adeloye, 2007; Sorjamaa et al., 2009; Merlin et al., 2010) and climate data (Sorjamaa, 2010). All the work above deal with the missing values as proposed by Cottrell and Letrémy (2007). They compute the best-matching units for the data vectors with missing values

$$\boldsymbol{m}_{c(\boldsymbol{x}_n)} = \arg\min_i \|\boldsymbol{x}_n - \boldsymbol{m}_i\| = \sum_{k \text{ s.t. } \boldsymbol{I}_{nk}=1} (\boldsymbol{x}_{nk} - \boldsymbol{m}_{ik})^2, \tag{4.17}$$

that is, the distance is computed using only the components present in vector $\boldsymbol{x}_n$. The missing values are ignored also while updating the reference vectors. This approach is implemented in the widely used MATLAB toolbox, SOM Toolbox (Vesanto et al., 2000; Alhoniemi et al., 2005). After the training, missing values can be filled according to the best-matching units of corresponding data vectors. Cottrell and Letrémy (2007) also discuss posteriori estimation of missing values but that does not affect the convergence properties of the SOM algorithm.

Incomplete Data on Training

Cottrell and Letrémy (2007) identify two options for using the incomplete data with the SOM. First, one may define distances as (4.17) and use only the components present in each data vector $\boldsymbol{x}_n$ when updating the weights $\boldsymbol{m}_i$. Second, if there is sufficient data, the mere full data vectors can be used for training the SOM. After the training, the best-matching units of the resulting SOM can be used to impute the sparse data vectors. Later in this thesis, the approaches above are referred as *sparse* and *full*, since their training data consists of sparse and full data, respectively.

Novel Ways of Using Incomplete Data

Inspired by the imputation algorithm (see Section 4.1.3 above) and handling of the missing values with GTM (see Section 4.3.3 below), this thesis proposes two novel ways of treating incomplete data during the SOM training. The first method, named the *alternating SOM (altSOM)*, imputes the incomplete data with new values from the corresponding best-matching units on every epoch of the batch training and computes the best-matching unit for data according to (4.13). The update rule for reference vectors $\boldsymbol{m}_i$ is revised slightly, which allows units which have data with missing values in their neighborhood adapt more easily according to the neighboring units. This revision is done by introducing weights $\boldsymbol{w}_n$ in the update rule

$$\boldsymbol{m}_i = \frac{\sum_n h_{ni} \boldsymbol{w}_n \cdot \boldsymbol{x}_n}{\sum_n h_{ni} \boldsymbol{w}_n}, \tag{4.18}$$

where product and division are taken componentwise, $w_{nk} = 1 \,\forall\, x_{nk}$ not missing and $w_{nk} = w \leq 1 \,\forall\, x_{nk}$ missing. In other words, while updating the reference vectors $\boldsymbol{m}_i$, the missing values have weight $w \leq 1$. Consequences of this revision are demonstrated in Section 5.1.1. When $w = 0$, this equals the common treatment of missing values described above. The weight w for missing values in (4.18) is another free parameter in the training phase which can be learned from the data. It might be reasonable to

alter this parameter during the learning but this is out of the scope of this thesis and is left for the future research.

The second approach, named the *imputation SOM (impSOM)*, stems from the way missing values are treated while using the GTM with an isotropic noise model (see Section 4.3.3). The distances between data points and reference vectors are evaluated according to (4.17), that is, only observed components are used for calculating distances. While updating the reference vectors, instead of ignoring the missing values their "expected values"

$$\hat{\boldsymbol{x}}_{ni,\text{mis}} = \mathbb{E}\left[\boldsymbol{x}_{n,\text{mis}} | \boldsymbol{m}_i\right] = \boldsymbol{m}_i \tag{4.19}$$

are used. Above, expectation is used in an informal sense, since the SOM is not a statistical model. This results in an update rule, where the reference vectors are updated according to (4.16) such that for each unobserved component of $\boldsymbol{x}_n$ the current value $\boldsymbol{m}_i$ is used.

4.2.4 Binary Data

The are many improvements for the SOM for processing binary data (see, e.g., Lebbah et al., 2007, 2008). In this work, the standard SOM is used in order to process binary data. It has been shown that one can achieve reasonable results by applying the SOM on this kind of data (see, e.g., Kohonen, 2001, page 162). The resulting codebook vectors may be interpreted as parameters of Bernoulli distributions and used in sampling to obtain MI.

4.3 Generative Topographic Mapping

The Generative Topographic Mapping (GTM) (Bishop et al., 1998; Bishop and Williams, 1998) is a nonlinear latent variable model which was proposed as a probabilistic alternative to the SOM and as such it overcomes some limitations of the SOM. Loosely speaking, it extends the SOM in similar manner as Gaussian mixture model extends k-means clustering. This is achieved by working in a probabilistic framework where points have posterior probabilities given a map unit, to use the SOM terminology. Instead of one best-matching unit, each data vector contributes to many reference vectors directly.

The GTM can be seen consisting of three parts: 1) discrete set of point in usually one or two-dimensional latent space, 2) nonlinear mapping, usually radial basis function (RBF) network, between the latent space and the data space, and 3) a Gaussian noise model in the data space such that the resulting model is a constrained mixture of Gaussians. These three parts are illustrated schematically in Figure 4.3 from left to right: the nine dots in a square represent map units in two-dimensional latent space spanned by $\{\boldsymbol{u}_1, \boldsymbol{u}_2\}$. These points are mapped to three-dimensional data space on the right using a nonlinear mapping ($\boldsymbol{y}(\boldsymbol{u}; \boldsymbol{w})$ in the figure). The shaded spheres represent the noise model of the map units in the data space spanned by $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$.

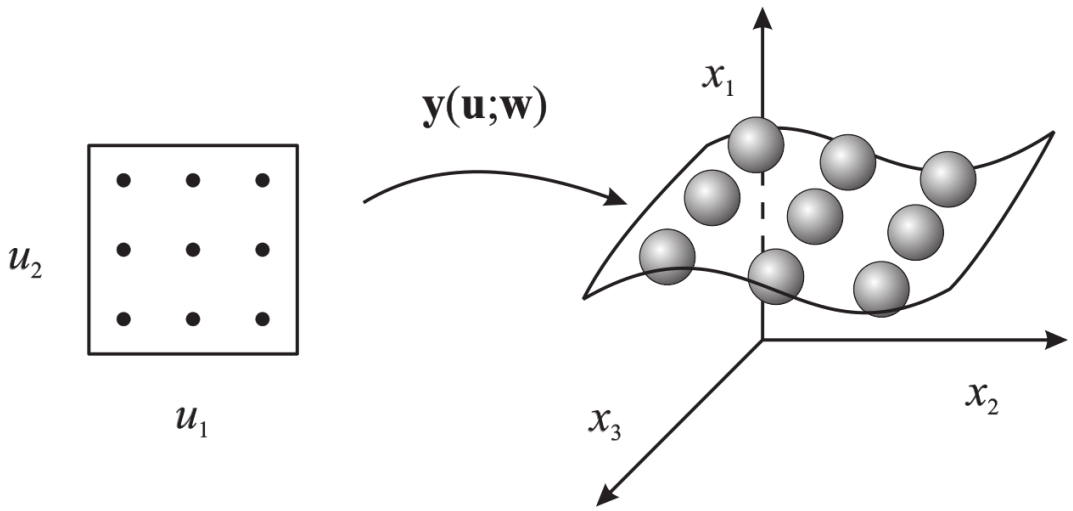


Figure 4.3: Schematic illustration of the GTM: discrete set of points in two-dimensional latent space on the left-hand side are mapped to three-dimensional data space using a nonlinear mapping $y(u; w)$. The speheres on the right-hand side represent the isotropic Gaussian distributions which comprise the probability distribution in the data space. (Bishop and Williams, 1998)

4.3.1 Latent-Variable Model

In this section, notation consistent with the article by the original authors Bishop and Williams (1998) is used. The goal of the GTM is to approximate the density $p(\boldsymbol{x})$ of data in a d-dimensional data space in terms of q latent variables $\boldsymbol{u} = (u_1, \ldots, u_q)$. The GTM uses a regular array of K nodes in the latent space, $\{\boldsymbol{u}_i\}$, $i = 1, \ldots, K$, analogous to the map units of the SOM. The latent points are mapped from the latent space into the data space using a nonlinear function $\mathbf{y}(\boldsymbol{u}, \boldsymbol{W})$, where $\boldsymbol{W}$ represents the parameters of the mapping. The most interesting situation is such that q equals one or two allowing SOM-style visualization of the data. However, any dimensionality $q < d$ might be of interest. Figure 4.3 illustrates the case $q = 2$ and $d = 3$.

The non-linear mapping is achieved by using a set of M fixed radial basis functions $\boldsymbol{\phi}(\boldsymbol{u}_i) = \{\phi_j(\boldsymbol{u}_i)\}$, where $\phi_j(\boldsymbol{u}_i) = \exp\{-\|\boldsymbol{c}_j - \boldsymbol{u}_i\|/\sigma^2\}$, σ is the width parameter of the RBFs, $\{\boldsymbol{c}_j\}$ are the RBF centers and $j = 1, \ldots, M$. The number of RBFs, M, is a free parameter which has to be chosen by the experimenter. Later, techniques for avoiding this kind of discrete model selection are discussed. In this thesis, square and uniform lattices of 4, 9, 16 and 25 RBFs are used. The radius of the RBFs is chosen according to

$$\sigma = \frac{d_{\max}}{\sqrt{N}}, \tag{4.20}$$

where $d_{\max}$ is the maximum distance between two RBF centers. This is a common advice in the neural network text books (see, e.g., Haykin, 2008).

In more general setting, the nonlinear mapping may consist of any non-linear functions such as Gaussian or sigmoidal functions. Each map unit $\boldsymbol{u}_i$ in the latent

space is mapped to a corresponding point $\boldsymbol{y}_i$ in the data space given by

$$\boldsymbol{y}_i = \boldsymbol{W}\boldsymbol{\phi}(\boldsymbol{u}_i), \tag{4.21}$$

where $\boldsymbol{W}$ is a $D \times M$ matrix of weight parameters. Using the SOM terminology, the node locations in latent space, $\boldsymbol{u}_i$, define a corresponding set of reference vectors

$$\boldsymbol{m}_i = \boldsymbol{W}\boldsymbol{\phi}(\boldsymbol{u}_i), \quad i = 1, \ldots, K, \tag{4.22}$$

in the data space. In this work, each reference vector $\boldsymbol{m}_i$ serves as a center of an isotropic Gaussian distribution in the data space

$$p(\boldsymbol{x}|\boldsymbol{m}_i) = \left(\frac{\beta}{2\pi}\right)^{D/2} \exp\left\{-\frac{\beta}{2}\|\boldsymbol{m}_i - \boldsymbol{x}\|^2\right\}, \tag{4.23}$$

where β is the precision or inverse variance. The Gaussian distribution above also represents a noise model accounting for the fact that the data will not be confined precisely to the lower-dimensional manifold in the data space. More general noise models has been proposed, as well (see, e.g., Bishop and Williams, 1998).

The probability density function of the GTM is obtained by summing over the Gaussian components yielding

$$p(\boldsymbol{x}|\boldsymbol{W}, \beta) = \sum_{i=1}^{K} P(\boldsymbol{m}_i) p(\boldsymbol{x}|\boldsymbol{m}_i) = \sum_{i=1}^{K} \frac{1}{K} \left(\frac{\beta}{2\pi}\right)^{D/2} \exp\left\{-\frac{\beta}{2}\|\boldsymbol{m}_i - \boldsymbol{x}\|^2\right\}, \tag{4.24}$$

where K is the total number grid points in the latent space, or map units in the SOM terminology, and the prior probabilities $P(\boldsymbol{m}_i)$ are given equal probabilities $1/K$. Figure 4.3 illustrates a GTM with nine map units schematically. Each map unit corresponds to an isotropic Gaussian in the data space illustrated by the spheres in the figure.

The GTM represents a parametric probability density model, with parameters $\boldsymbol{W}$ and β, and it can be fitted to a data set $\{\boldsymbol{x}_n\}$, where $n = 1, \ldots, N$, by maximum likelihood. The log likelihood function of the GTM is given by

$$\mathcal{L}(\boldsymbol{W}, \beta) = \sum_{n=1}^{N} \ln p(\boldsymbol{x}_n|\boldsymbol{W}, \beta), \tag{4.25}$$

where $p(\boldsymbol{x}_n|\boldsymbol{W}, \beta)$ is given by (4.24) and independently, identically distributed (iid) data is assumed. The log likelihood can be maximized using standard non-linear optimization techniques or alternatively using the EM algorithm. Figure 4.4 shows the plate diagram of the GTM. Note that $\boldsymbol{W}$ and $\boldsymbol{\Phi}$ are deterministic, hence matrix $\boldsymbol{M}$, which represents the reference vectors in the data space, is determined by the matrix product $\boldsymbol{W}\boldsymbol{\Phi}$ and is not directly optimized.

4.3.2 The EM Algorithm

The EM algorithm is an important general technique for finding maximum-likelihood estimates in incomplete-data problems. In order to transform the GTM into a latent-variable model, let us introduce a K-dimensional binary random variable $\boldsymbol{z}$ using a

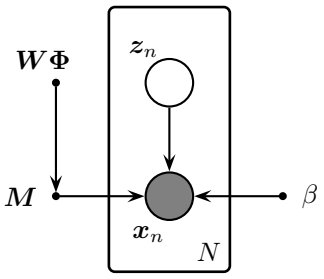


Figure 4.4: The plate diagram of the GTM. $\boldsymbol{x}_n$ denotes the observations, latent variables $\boldsymbol{z}_n$ govern which map unit data vectors "belong to",$\boldsymbol{W\Phi}$ is the nonlinear mapping, $\boldsymbol{M}$ is a matrix of reference vectors $\boldsymbol{m}_i$ and β is precision of the noise model.

1-of-K representation such that a particular element of z_i is equal to 1 and all other elements of z are equal to 0. In other words, $p(z_{ni} = 1|\boldsymbol{x}_n, \boldsymbol{W}, \beta) = R_{ni}$ is the probability that data vector $\boldsymbol{x}_n$ was generated by map unit i.

In the E-step the current values of model parameters $\boldsymbol{W}$ and β are used to evaluate the posterior probabilities, or responsibilities, which each map unit $\boldsymbol{m}_i$ takes for every data point $\boldsymbol{x}_n$, which is given by

$$R_{ni} = p(z_{ni} = 1|\boldsymbol{x}_n) = \frac{p(\boldsymbol{x}_n|\boldsymbol{m}_i)}{\sum_j p(\boldsymbol{x}_n|\boldsymbol{m}_j)}. \tag{4.26}$$

The prior probabilities $p(\boldsymbol{m}_i) = 1/K$ cancel out between numerator and denominator. In the subsequent M-step the responsibilities R_{ni} are used to re-estimate the model parameters $\boldsymbol{W}$ and β. The log likelihood function for the expected complete data is

$$\langle \mathcal{L}_{\text{comp}}(\boldsymbol{W}, \beta) \rangle = \sum_{n=1}^{N} \sum_{i=1}^{K} R_{ni} \ln p(\boldsymbol{x}_n|\boldsymbol{m}_i, \boldsymbol{W}, \beta). \tag{4.27}$$

Maximizing this with respect to $\boldsymbol{W}$ yields

$$\sum_{n=1}^{N} \sum_{i=1}^{K} R_{ni} \left\{ \boldsymbol{W}\boldsymbol{\phi}(\boldsymbol{u}_i) - \boldsymbol{x}_n \right\} \boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{u}_i) = 0. \tag{4.28}$$

This can be written in matrix notation and solved as follows:

$$\boldsymbol{W}^{\mathrm{T}} = (\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{X}, \tag{4.29}$$

where $\boldsymbol{\Phi}$ is a $K \times M$ matrix with elements $\Phi_{ij} = \phi_j(\boldsymbol{u}_i)$, $\boldsymbol{X}$ is an $N \times D$ data matrix with elements x_{nk}, $\boldsymbol{R}$ is a $K \times N$ matrix with elements R_{ni}, and $\boldsymbol{G}$ is a $K \times K$ diagonal matrix with elements $G_{ii} = \sum_n R_{ni}$. The precision parameter is also updated in the M-step by evaluating

$$\frac{1}{\beta} = \frac{1}{ND} \sum_{n=1}^{N} \sum_{i=1}^{K} R_{ni} \left\| \boldsymbol{W}\boldsymbol{\phi}(\boldsymbol{u}_i) - \boldsymbol{x}_n \right\|^2, \tag{4.30}$$

which is obtained by maximizing (4.27) with respect to β.

4.3.3 The GTM with Missing Values

As Bishop et al. (1998) point out, the GTM offers a robust framework for dealing with missing values. If missing values are MAR, the likelihood function is obtained by integrating out the unobserved values

$$p(\boldsymbol{X}_{\text{obs}}|\boldsymbol{W},\beta) = \int p(\boldsymbol{X}_{\text{obs}}|\boldsymbol{X}_{\text{mis}},\boldsymbol{W},\beta)d\boldsymbol{X}_{\text{mis}}. \tag{4.31}$$

This integration can performed analytically for the GTM with an isotropic noise model.

If only the unit responsibilities $\boldsymbol{R}$ are unknown, that is, the values of $\boldsymbol{Z}$ are unobserved, the E-step is reduced to estimating $\mathbb{E}(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{W},\beta)$. We are interested in the case where both $\boldsymbol{Z}$ and $\boldsymbol{X}_{\text{mis}}$ are missing. The complete-data likelihood is

$$\begin{aligned}\mathcal{L}_{\text{comp}}(\boldsymbol{W},\beta|\boldsymbol{X}_{\text{obs}},\boldsymbol{X}_{\text{ mis}},\boldsymbol{R}) = \sum_{n=1}^{N}\sum_{i=1}^{K} R_{ni}\Bigg\{ &\frac{D}{2}\ln\beta - \frac{D}{2}\ln 2\pi \\ &- \frac{\beta}{2}\|\boldsymbol{W}\phi(\boldsymbol{u}_i) - \boldsymbol{x}_{n,\text{mis}}\|^2 \\ &- \frac{\beta}{2}\|\boldsymbol{W}\phi(\boldsymbol{u}_i) - \boldsymbol{x}_{n,\text{obs}}\|^2\Bigg\}. \end{aligned} \tag{4.32}$$

The sufficient statistics for the parameters $\boldsymbol{W}$ and β include three unknown terms $\mathbb{E}[\boldsymbol{R}|\boldsymbol{X}_{\text{obs}},\boldsymbol{W}^k,\beta^k]$, $\mathbb{E}[\boldsymbol{R}\boldsymbol{X}_{\text{mis}}|\boldsymbol{X}_{\text{obs}},\boldsymbol{W}^k,\beta^k]$ and $\mathbb{E}[\boldsymbol{R}\boldsymbol{X}_{\text{mis}}\boldsymbol{X}_{\text{mis}}^{\text{T}}|\boldsymbol{X}_{\text{obs}},\boldsymbol{W}^k,\beta^k]$. The use of isotropic Gaussian for the noise model significantly simplifies evaluating the statistics above. For more involved case in presence of covariances, the derivation of the expectations for the Gaussian mixture model can be found in (Ghahramani and Jordan, 1994).

Let us define

$$\hat{\boldsymbol{x}}_{ni,\text{mis}} = \mathbb{E}\left[\boldsymbol{x}_{n,\text{mis}}|z_{ni}=1,\boldsymbol{x}_{i,\text{obs}},\boldsymbol{W}^k,\beta^k\right] = \boldsymbol{m}_i, \tag{4.33}$$

which is the least-squares linear regression between $\boldsymbol{x}_{n,\text{obs}}$ and $\boldsymbol{x}_{n,\text{mis}}$ predicted by the map unit $\boldsymbol{m}_i$ on k^{th} iteration. The expectation $\mathbb{E}[z_{ni}|\boldsymbol{x}_{n,\text{obs}},\boldsymbol{W}^k,\beta^k] = R_{ni}$ as defined in (4.26) measured only on the observed dimensions of $\boldsymbol{x}_n$. Similarly,

$$\mathbb{E}[z_{ni}\boldsymbol{x}_{n,\text{mis}}|\boldsymbol{x}_{n,\text{obs}},\boldsymbol{W}^k,\beta^k] = R_{ni}\hat{\boldsymbol{x}}_{ni,\text{mis}}. \tag{4.34}$$

and

$$\mathbb{E}[z_{ni}\boldsymbol{x}_{n,\text{mis}}\boldsymbol{x}_{n,\text{mis}}^{\text{T}}|\boldsymbol{x}_{n,\text{obs}},\boldsymbol{W}^k,\beta^k] = R_{ni}(\beta^{-1,k} + \hat{\boldsymbol{x}}_{ni,\text{mis}}\hat{\boldsymbol{x}}_{ni,\text{mis}}^{\text{T}}). \tag{4.35}$$

These expectations are substituted into Equations (4.29) and (4.30) to re-estimate the weights and the precision. As a consequence, the matrix product $\boldsymbol{R}\boldsymbol{X}$ has to be evaluated separately for each map unit $\boldsymbol{m}_i$. In the update equation of β, the squared-norm term for the missing data is given by $\|\boldsymbol{W}\boldsymbol{\phi}(\boldsymbol{u}_i) - \boldsymbol{x}_n\|^2 = \sigma^{2,\text{old}} = 1/\beta^{\text{old}}$. The resulting algorithm is summarized in Algorithm 4.2.

After the training, there are at least two possibilities to perform single imputation using the GTM. One may use the expected values $\mathbb{E}(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}},\boldsymbol{W},\beta)$ or impute using the maximum-a-posteriori (MAP) estimates $p_{\text{MAP}}(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}},\boldsymbol{W},\beta)$ which takes the missing values from the most similar map unit. Additionally, multiple imputation can be conducted by sampling the posterior probability $p(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}},\boldsymbol{W},\beta)$.

Algorithm 4.2 The EM Algorithm for the GTM for data with missing values

Given a joint distribution $p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{W}, \beta)$ over the observed values $\boldsymbol{X}_{\text{obs}}$, the missing values $\boldsymbol{X}_{\text{mis}}$ and unobserved unit responsibilities $\boldsymbol{R}$, governed by parameters $\boldsymbol{W}$ and β, the goal is to maximize the likelihood function $p(\boldsymbol{X}_{\text{obs}}|\boldsymbol{W}, \beta)$ with respect to $\boldsymbol{W}$ and β.

1. Initialize the parameters $\boldsymbol{W}$ and β using, e.g., PCA.
2. **E-step**, evaluate

   $$\boldsymbol{R} = p(\boldsymbol{Z}|\boldsymbol{X}_{\text{obs}}, \boldsymbol{W}^{\text{old}}, \beta^{\text{old}}).$$

3. **M-step**, evaluate $\boldsymbol{W}^{\text{new}}$ and β^{new} given by

   $$\boldsymbol{W}^{\text{T}} = (\boldsymbol{\Phi}^{\text{T}}\boldsymbol{G}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\text{T}}\boldsymbol{R}\boldsymbol{X} \text{ and}$$

   $$\frac{1}{\beta} = \frac{1}{ND}\sum_{n=1}^{N}\sum_{i=1}^{K} R_{ni}\left\|\boldsymbol{W}\boldsymbol{\phi}(\boldsymbol{u}_i) - \boldsymbol{x}_n\right\|^2,$$

   where $\hat{\boldsymbol{X}}_{i,\text{mis}} = \boldsymbol{W}\boldsymbol{\phi}(\boldsymbol{u}_i)$ are used as expected values for the missing data.
4. Check for convergence of either the log likelihood or the parameter values. If the convergence criterion is not satisfied, then let

   $$\boldsymbol{W}^{\text{old}} \leftarrow \boldsymbol{W}^{\text{new}} \text{ and } \beta^{\text{old}} \leftarrow \beta^{\text{new}}$$

   and return to step 2.

4.3.4 Binary Data

Bishop and Williams (1998) have also formulated the GTM for discrete data. In this thesis, the GTM for binary data, $x_k \in \{0, 1\}$, is called the *Bernoulli GTM*. Components of $\boldsymbol{x}$ are assumed conditionally independent, given the map unit $\boldsymbol{m}_i$. The conditional distribution of observation $\boldsymbol{x}$ given a map unit $\boldsymbol{m}_i$ is given by a product of Bernoulli distributions

$$p(\boldsymbol{x}|\boldsymbol{m}_i) = \prod_{k=1}^{D} m_{ik}^{x_k}(1 - m_{ik})^{1-x_k}, \tag{4.36}$$

where the conditional means are given by $m_{ik} = \sigma(\boldsymbol{w}_k^{\text{T}}\boldsymbol{\phi}(\boldsymbol{u}_i))$, $\sigma(x) = (1+\exp(-x))^{-1}$ is the logistic sigmoid function, and $\boldsymbol{w}_k$ is the k^{th} column of $\boldsymbol{W}$.

The parameters $\boldsymbol{W}$ can again be estimated using the EM algorithm. The E-step updates the posterior probabilities R_{ni} using (4.26). The M-step requires nonlinear optimization. In this thesis, a novel error function to be minimized in the M-step is proposed, given by

$$\varepsilon(\boldsymbol{W}) = \frac{1}{2}\|\sigma(\boldsymbol{W}\boldsymbol{\Phi}) - \boldsymbol{M}_{\text{ML}}\|^2 + \frac{1}{2}\alpha\|\boldsymbol{w}\|^2, \tag{4.37}$$

where $\boldsymbol{M}_{\text{ML}}$ is the ML estimate for the cluster centers in the data space, given the posterior probabilities R_{ni}, and $\frac{1}{2}\alpha\|\boldsymbol{w}\|^2$ is a regularization term (see Section 4.3.6).

Any known optimization method can be used to minimize (4.37). Bishop and Williams (1998) propose using the iterative re-weighted least squares (IRLS) or generalized EM (GEM) algorithms. They also note that it is sufficient and computationally more efficient to perform only partial optimization in the M-step. Girolami (2001, 2002) has developed alternative, and more specific methods and demonstrated their performance with multiple data sets. In this work, MATLAB Optimization toolbox `fminunc` function, which exploits a subspace trust-region method (Coleman and Li, 1996), is used in order to minimize (4.37).

4.3.5 Initialization

According to Kiviluoto and Oja (1998) it seems that the GTM requires a careful initialization in order to self-organize. Bishop and Williams (1998) propose using PCA to initialize the parameters $\boldsymbol{W}$ and β. They determine $\boldsymbol{W}$ by minimizing the sum-of-squares error between the projections of the latent points into the data space by the GTM and the corresponding projections obtained from PCA. The value of β^{-1} is initialized to be the larger of either the $q+1$ eigenvalue from PCA, representing the largest variance of the data perpendicular to the PCA subspace, or the square of half of the grid spacing of the PCA-projected latent points in the data space. In this thesis, a novel way of initializing the GTM using the reference vectors of a SOM trained using the same data set, is studied.

4.3.6 Improvements

Model selection using the GTM involves selecting the size of the latent variable grid and the RBF grid as well as the width parameter of the RBFs σ. The model selection regarding the RBF network roughly corresponds to the selection of the width of the neighborhood function in the SOM, which in turn control the "stiffness" of the SOM. During the learning phase of the SOM, the stiffness of the map is usually altered in order to allow better self-organization and to speed up the learning. More precisely, the training is started using a rigid grid and by narrowing the radius of the neighborhood function the map is "loosened" during the training.

In this thesis, this engineering approach which benefits SOM, is investigated with the GTM. Altering the number of RBFs and controlling the magnitude of the elements of $\boldsymbol{W}$, that is, regularization, are studied. In Section 5.3.3, the number of RBFs is increased as the training proceeds. This approach is called the *fine-tuned GTM*, and it effectively makes the mapping more elastic during the training. With the same data set and the Bernoulli GTM, a simple quadratic regularization is used. The quadratic regularization adds a term

$$\frac{1}{2}\alpha\|\boldsymbol{w}\|^2 \tag{4.38}$$

to the error measure, which in the case of the original GTM is the log likelihood (4.25). In the regularization term (4.38), $\boldsymbol{w}$ is a column vector consisting of the concatenation of the successive columns of $\boldsymbol{W}$ and α is a constant regularization parameter which has to be selected according to the data at hand. Any other known regularization technique may be used as well.

To overcome the constraints due to the finite number of RBFs used, the Gaussian Process formulation of the GTM was introduced by Bishop and Williams (1998). Loosely speaking this means that one assumes that the covariance between two map units $\boldsymbol{m}_i$ and $\boldsymbol{m}_j$ depends on the positions of their respective nodes $\boldsymbol{u}_i$ and $\boldsymbol{u}_j$ through, for example,

$$cov_{i,j} = v \cdot \exp\left\{-\frac{\|\boldsymbol{u}_i - \boldsymbol{u}_j\|^2}{2\lambda^2}\right\}, \tag{4.39}$$

where v and λ are constants. However, the GTM using the Gaussian Process formulation is not studied in this thesis.

Chapter 5

Experiments and Results

In this chapter, the missing data imputation methods described in the previous chapters are studied using four different data sets:

1. *Artificial data* is a simple, three-dimensional data set generated using a function $f(x_1, x_2) = \sin(x_1) + \cos(x_2) = y$. This data set allows visualizations and careful inspection of the properties of different imputation methods.

2. *Wine data set* is a widely used data set with 13 variables from the UCI machine learning repository (Frank and Asuncion, 2010). MCAR data is generated artificially and properties of different imputation methods are further studied.

3. *Nursing survey data* was collected in VirtualCoach research project. The survey was targeted to mothers who had experienced one or more breastfeeding periods. The data consists of responses to 36 Likert-scale questions from 1101 respondents with less than 1 % missing data.

4. *15D instrument data* was collected in unidentified clinical trials. 15D instrument (Sintonen, 2001) is a survey tool for measuring the *Health-Related Quality of Life* (HRQoL) of respondents. Since the analysis of 15D data frequently consists of statistical tests, it is an apt data set for comparing multiple imputation methods.

In all the experiments, MAR data is assumed. This allows ignoring the missingness mechanisms as explained in Section 2.1. For the first two data sets this is well-justified, since the missing data is generated randomly. For the real-life survey data sets this assumption is more dubious. However, using nonignorable missingness models is far more complicated and deserves a study of its own. In any case, assuming MAR is the only reasonable choice in most real-life data analysis tasks.

Experiments in the chapter were done mostly in MATLAB environment. The PCA experiments were conducted using *Matlab package for PCA for datasets with missing values* (Ilin and Raiko, 2010, 2008). For the traditional SOM, the SOM toolbox (Vesanto et al., 2000; Alhoniemi et al., 2005) was used. The alternating SOM and the imputation SOM together with the function for the combined error were implemented for this thesis. For the rudimentary GTM, The NETLAB toolbox (Nabney, 2002) was used. The treatment for missing values, imputation and the Bernoulli GTM were implemented for this work. R package *mice: Multivariate Imputation by Chained*

Equations (van Buuren and Groothuis-Oudshoorn, 2011) was used to carry out MICE for the last two data sets.

5.1 Artificial Data

The first data set studied consists of artificial data generated using a function $f(x_1, x_2) = \sin(x_1) + \cos(x_2) = y$. The surface $y = f(x_1, x_2)$ is shown in Figure 5.1(a). After adding Gaussian noise with standard deviation $\sigma = 0.1$, the data follows the Gaussian distribution

$$y \sim \mathcal{N}(f(x_1, x_2), \sigma). \tag{5.1}$$

Figure 5.1(b) shows a sample ($N = 1000$) of the artificial data, where x_1 and x_2 were sampled uniformly in the range $[-3, 3]$. The data lies on a nonlinear, two-dimensional manifold which makes it a suitable test problem for the subspace methods.

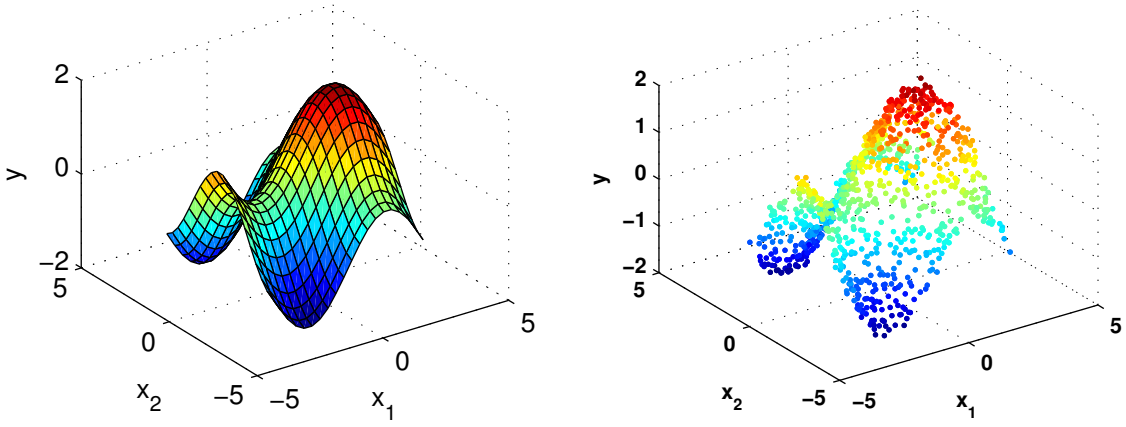
Missing values were generated by removing y for all $(x_1, x_2) \in (-1.5, 1.5) \times (-1.5, 1.5)$. This produced a hole in the data surface as depicted in Figures 5.2(a) and 5.2(b).

The problem setting is a bit naive, but it enables a study of the properties of the SOM and the GTM combined with visual inspection of the results. Furthermore, it is a scenario where the most simple approaches, such as hot deck, fail since there is a broad area in the data space with no observed values y. Resembling missing data patterns may also appear in real life experiments. For example in climate research, Tangayika Lake surface temperature data is used widely (see, e.g., Tierney et al., 2010; Sorjamaa et al., 2010). The measurements are made using satellites, hence cloud coverage may inflict wide, continuous areas of missing data.

5.1.1 Imputation with SOM

The goal of the artificial data experiments with the SOM was to compare the SOM update rule proposed by Cottrell and Letrémy (2007, see Section 4.2.3) with the novel imputation techniques introduced in Section 4.2.3. The experiments were carried out using (1) the sparse data (where some values of y were missing) with distance metric (4.17), (2) only the full data (the data vectors with missing y were removed from the training data) with traditional SOM algorithm, (3) the imputation SOM, which treats the missing data as described on page 21, and (4) the alternating SOM, where missing data is imputed on each training epoch according to the best matching units of the data vectors and reference vectors are updated according to (4.18). For the alternating SOM, the weight parameter $w = 0.05$ for missing values in (4.18) is selected such that the beneficial properties of the revision are emphasized.

Figure 5.3 shows the combined error (see page 19) and root mean square (RMS) imputation error with respect to the number of map units used. The RMS imputation error was computed over imputations of 50 randomly generated data sets. The plot of the combined error in Figure 5.3(a) shows that even though the differences are small one can obtain maps with the lowest combined error by using the imputation SOM, whereas using only the full data vectors for training, results in maps with significantly worse quality. However, measured with the RMS imputation error the imputation



(a) $f(x_1, x_2) = \sin(x_1) + \cos(x_2) = y$ (b) A random sample

Figure 5.1: (a) The function used to generate artificial data, $f(x_1, x_2) = \sin(x_1) + \cos(x_2) = y$. (b) A random sample $N = 1000$ of y in (5.1) with Gaussian noise $\mathcal{N}(0, 0.1)$.

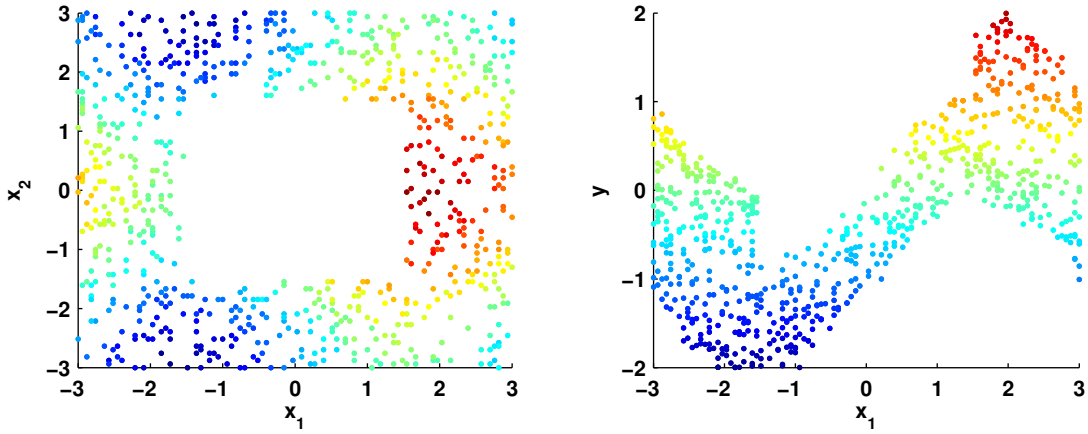
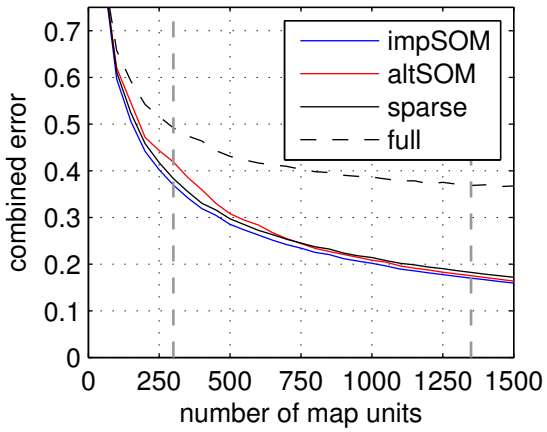


Figure 5.2: Artificial data set with Gaussian noise and value of y removed for all $(x_1, x_2) \in (-1.5, 1.5) \times (-1.5, 1.5)$.

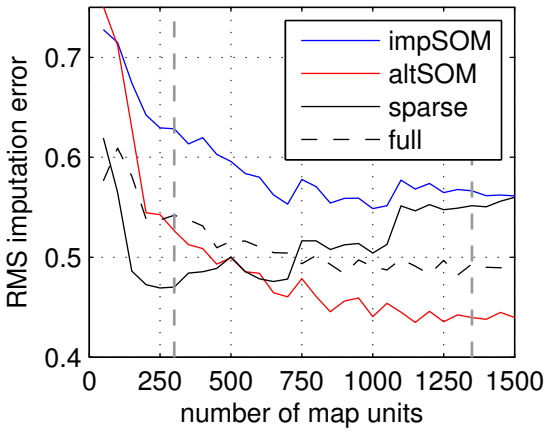
SOM obtains worst results as can be seen in Figure 5.3(b). The RMS plot also shows that the traditional SOM trained using the sparse data obtains good imputation results with relatively small maps (small number of map units) but the alternating SOM outperforms the traditional method as the map size increases. Figure 5.4 shows the corresponding quantization and topological errors. The differences between the imputation and traditional SOMs are minimal.

Selecting the number of map units for SOM is a subtle task. If the purpose of using the SOM is missing value imputation, one does not have the RMS plot of Figure 5.3(b) available without first performing some kind of validation. If the data contains "holes"—concentrated areas with plenty of missing values—as is the case here, a suitable selection of validation data might be difficult. Moreover, the error plots in Figures 5.3(a) and 5.4 do not give any straightforward way of determining a suitable map size. These difficulties in mind, the number of map units was deliberately increased above the number of data points. This allowed experimenting the hypothesis that the excess map units interpolate the data space allowing more precise imputation (see, e.g., Sorjamaa, 2010, this property of the SOM was discussed specifically in the dissertation).

The SOM arrays shown in the original data space in Figures 5.5–5.12 provide another view on the properties of SOM algorithms at hand. Figures show maps of two sizes, 294 and 1350 map units, shown with vertical dashed lines in Figure 5.3. The coloring shows the difference between the map plane and the actual plane $y = f(x_1, x_2)$. Inspection of the figures reveals that even though the RMS imputation error of the imputation SOM is worse compared to other methods, it provides the smoothest interpolation of the area $(x_1, x_2) \in (-1.5, 1.5) \times (-1.5, 1.5)$, that is, the hole in data. Also, the alternating SOM preserves the topology better compared to the tradition SOM when the map size is increased as can be seen in Figures 5.3(b) and 5.9. Figure 5.8 shows that when only full data is used in training, the topology is distorted in the area of the data space where the missing data lies. This distortion is harshly amplified when the number of map units is increased to 1350. In that case, the both scenarios of using the traditional SOM fail to preserve smooth topology, as can be seen in Figures 5.11 and 5.12, whereas the imputation SOM and the alternating SOM better interpolate the area of the sparse data.

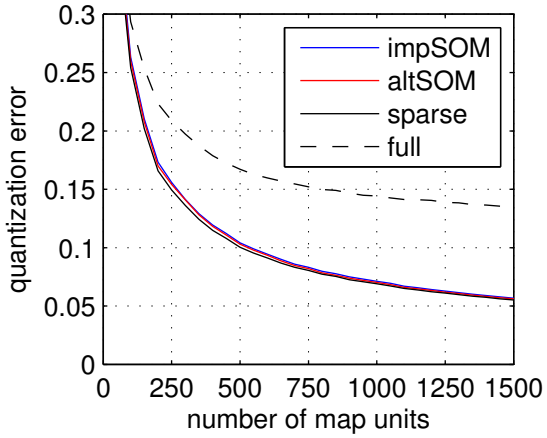


(a) Combined error

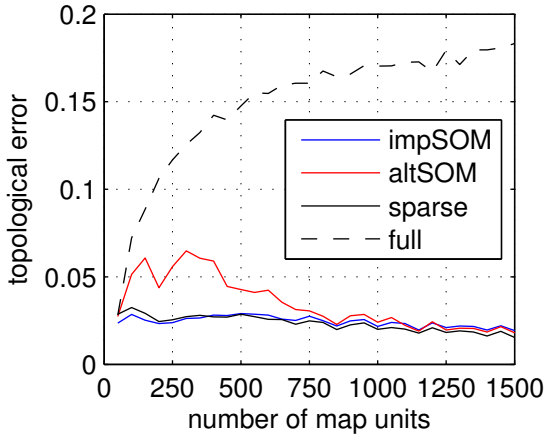


(b) Imputation error

Figure 5.3: The results of the experiments with artificial data and different SOM imputation methods. (a) The combined error with respect to the map size using the imputation SOM (impSOM), the alternating SOM (altSOM), the traditional SOM using the sparse data in training (sparse) and the traditional SOM using only full data in training (full). (b) The mean imputation error for the methods above. The vertical dashed lines show the map sizes visualized in Figures 5.5–5.12 below.



(a) Quantization error



(b) Topological error

Figure 5.4: The results of the experiments with artificial data and different SOM imputation methods. (a) The quantization error with respect to the map size with imputation SOM (impSOM), the alternating SOM (altSOM), the traditional SOM using the sparse data in training (sparse) and the traditional SOM using only full data in training (full). (b) The topological error for the methods above.

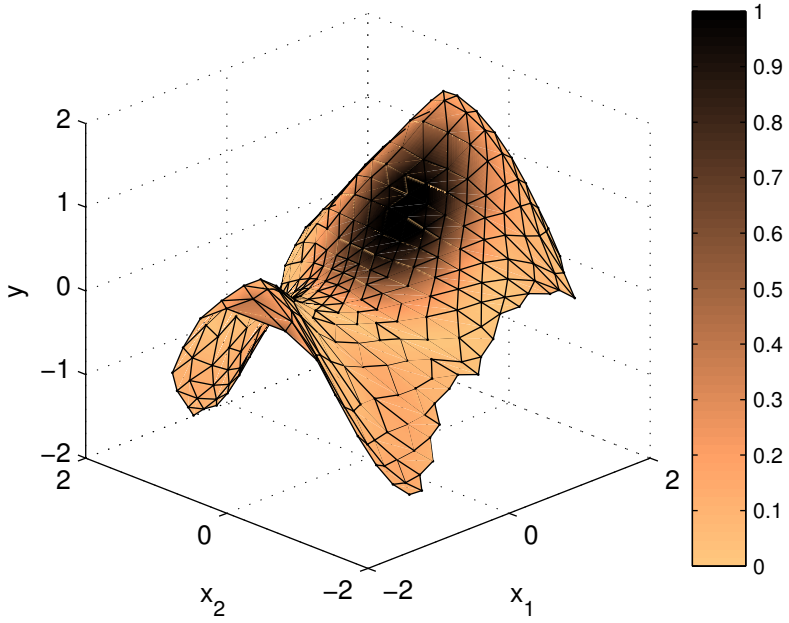


Figure 5.5: A SOM with 294 (21×14) map units trained using the artificial data and the **imputation SOM** algorithm. Coloring depicts the difference between the map units and the surface $y = f(x_1, x_2)$. RMS imputation error is 0.580.

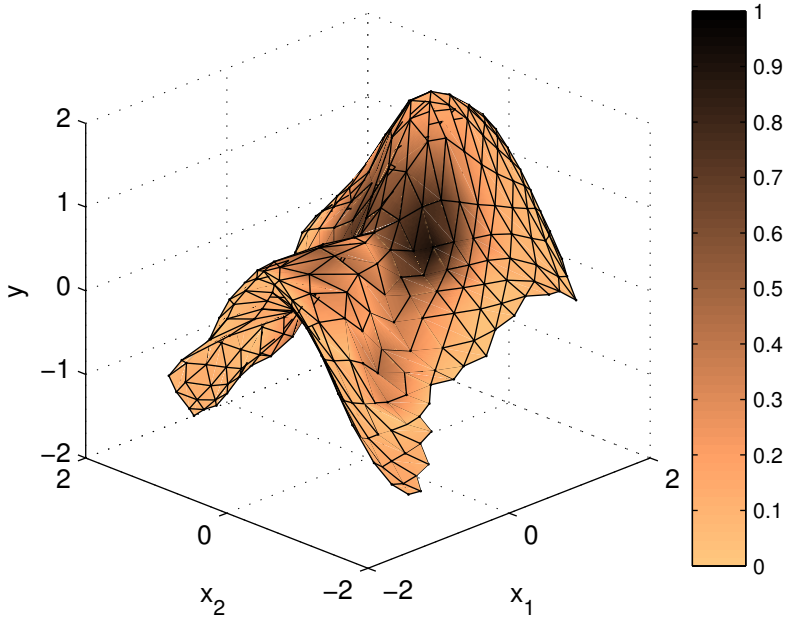


Figure 5.6: A SOM with 294 (21×14) map units trained using the artificial data and the **alternating SOM** algorithm. Coloring depicts the difference between the map units and the surface $y = f(x_1, x_2)$. RMS imputation error is 0.486.

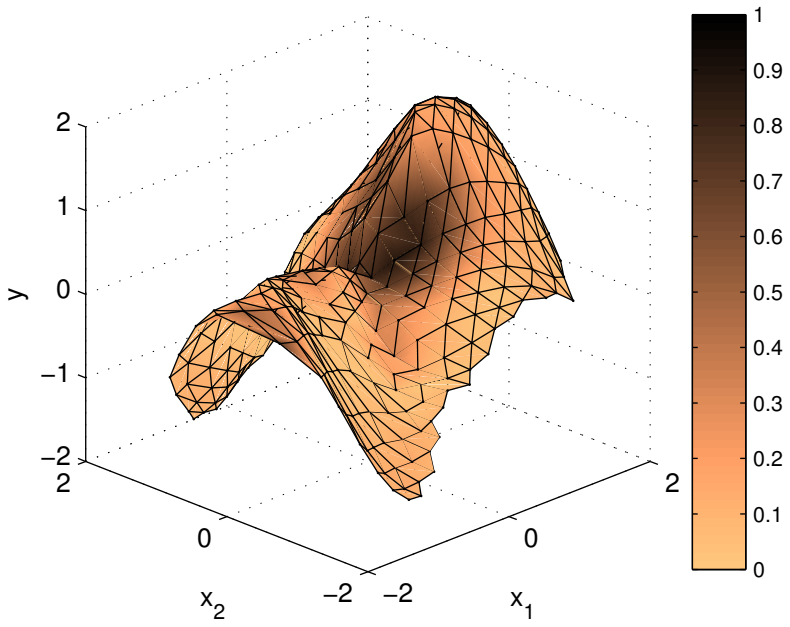


Figure 5.7: A SOM with 294 (21×14) map units trained using the **sparse artificial data and the traditional SOM** algorithm. Coloring depicts the difference between the map units and the surface $y = f(x_1, x_2)$. RMS imputation error is 0.439.

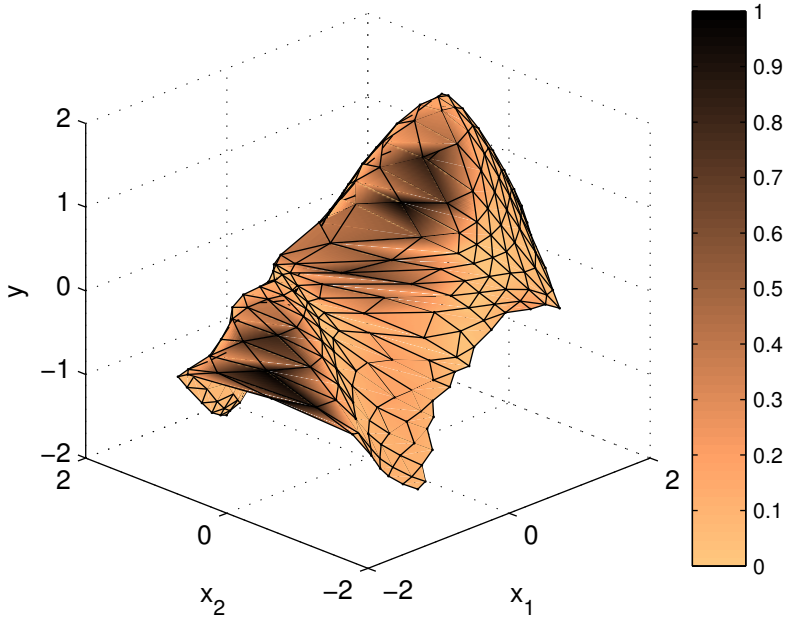


Figure 5.8: A SOM with 294 (21×14) map units trained using only the **full vectors of the artificial data and the traditional SOM** algorithm. Coloring depicts the difference between the map units and the surface $y = f(x_1, x_2)$. RMS imputation error is 0.515.

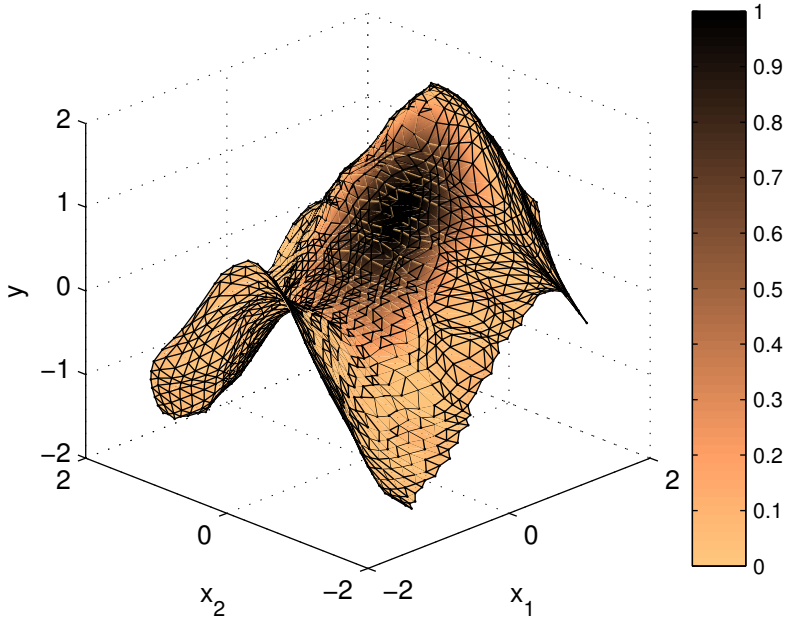


Figure 5.9: A SOM with 1350 (45×30) map units trained using the artificial data and the **imputation SOM** algorithm. Coloring depicts the difference between the map units and the surface $y = f(x_1, x_2)$. RMS imputation error is 0.554.

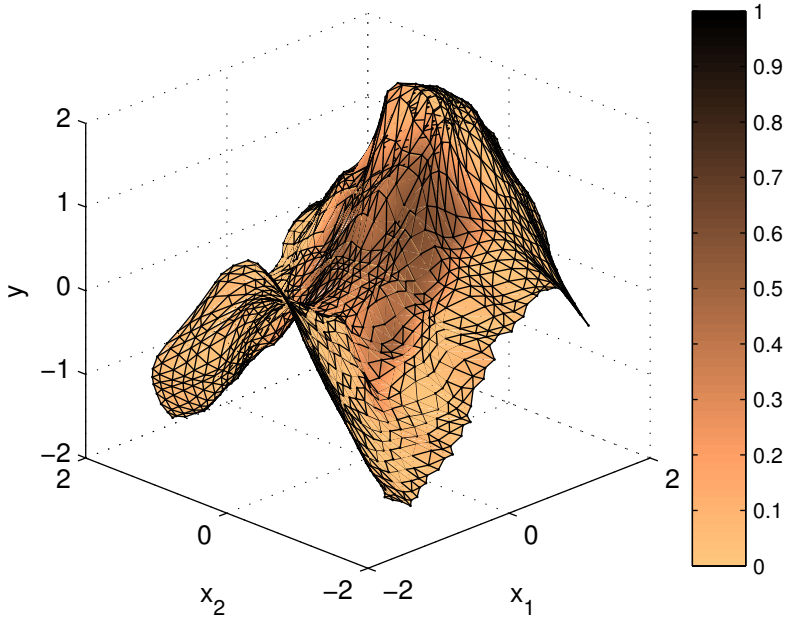


Figure 5.10: A SOM with 1350 (45×30) map units trained using the artificial data and the **alternating SOM** algorithm. Coloring depicts the difference between the map units and the surface $y = f(x_1, x_2)$. RMS imputation error is 0.385.

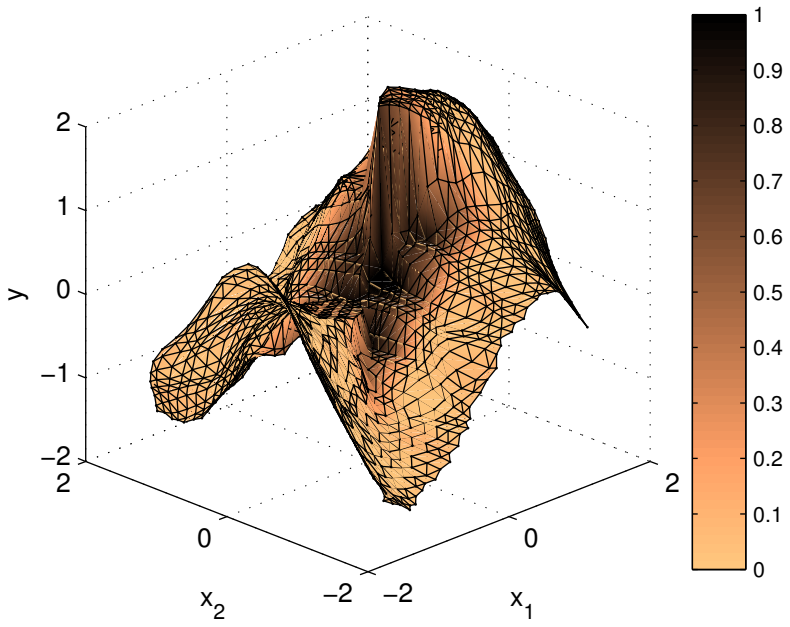


Figure 5.11: A SOM with 1350 (45×30) map units trained using the **sparse artificial data and the traditional SOM** algorithm. Coloring depicts the difference between the map units and the surface $y = f(x_1, x_2)$. RMS imputation error is 0.512.

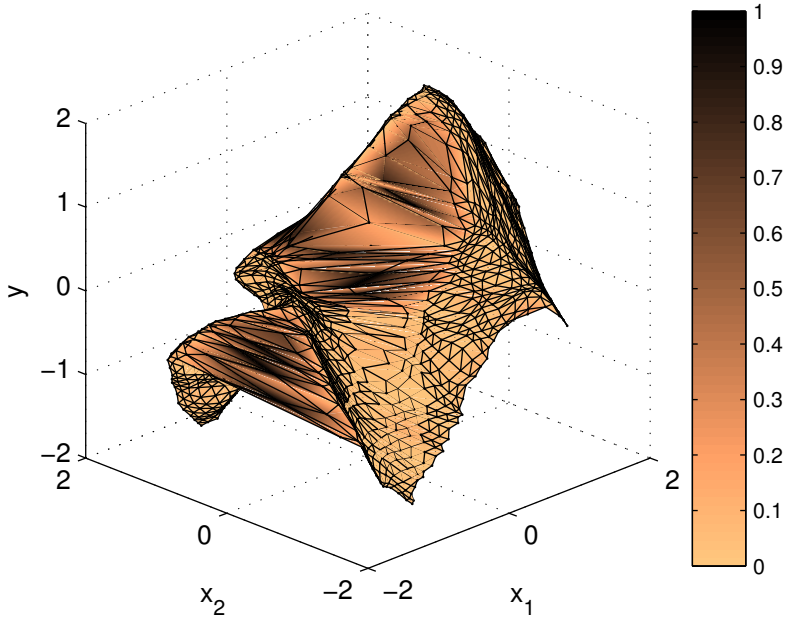


Figure 5.12: A SOM with 1350 (45×30) map units trained using only the **full vectors of the artificial data and the traditional SOM** algorithm. Coloring depicts the difference between the map units and the surface $y = f(x_1, x_2)$. RMS imputation error is 0.487.

5.1.2 Imputation with GTM

The primary goal of the experiments with the GTM was to study the properties of the GTM imputation methods, namely expectation and MAP imputation explained on page 25. Second, the effects of initializing the GTM using the reference vectors of a SOM trained using the same data set are studied. A suitable convergence criterion for the GTM with artificial data was found out to be such that the iteration is considered to be converged, when the increase in the log-likelihood (4.25) is less than 10^{-2}. A GTM with $M = 16$ (4×4) RBFs and σ selected using (4.20) was used. Imputation was conducted using both the expectation and the MAP estimates for the missing values. Figure 5.13 shows the plot of the RMS imputation error with respect to the number of map units. The RMS imputation error was computed over imputations of 50 randomly generated data sets. The GTM provides significantly smaller RMS imputation error compared to the SOM. Furthermore, imputing with expected values provides better results compared to using MAP estimates, with all map sizes used.

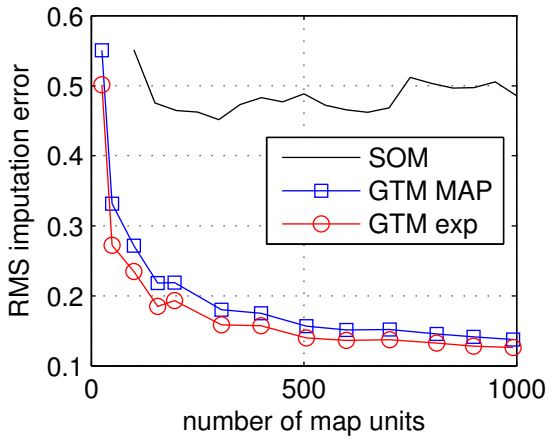


Figure 5.13: The results of the experiments with artificial data and the GTM imputation methods. The plot shows the RMS imputation error with respect to the number of map units, K, when imputing using the expected values (GTM exp) and according to the best-matching map unit (GTM MAP). The results of the traditional SOM trained with sparse data is shown for comparison.

Figures 5.14 and 5.15 show two resulting GTM models with $K = 100$ and $K = 992$ units, respectively. The coloring shows the difference between the map plane and the actual plane $y = f(x_1, x_2)$ on the same scale with the SOM Figures 5.7–5.12. It is also notable that the GTM plane expands slightly beyond the original data domain $(x_1, x_2) \in [-3, 3] \times [-3, 3]$. Thus, it seems that the GTM does not suffer from the border shrinkage effects typical to the SOM (see Kohonen, 2001, page 140).

The experiments confirmed the hypothesis that initializing the GTM with SOM reference vectors can speed up the convergence. Figure 5.16 shows that the SOM initialization can cut off 20–30 % of the number of iterations needed until convergence. The experiments were not suitable for comparing the CPU time consumed by SOM and GTM since the GTM algorithm used in this thesis was not implemented com-

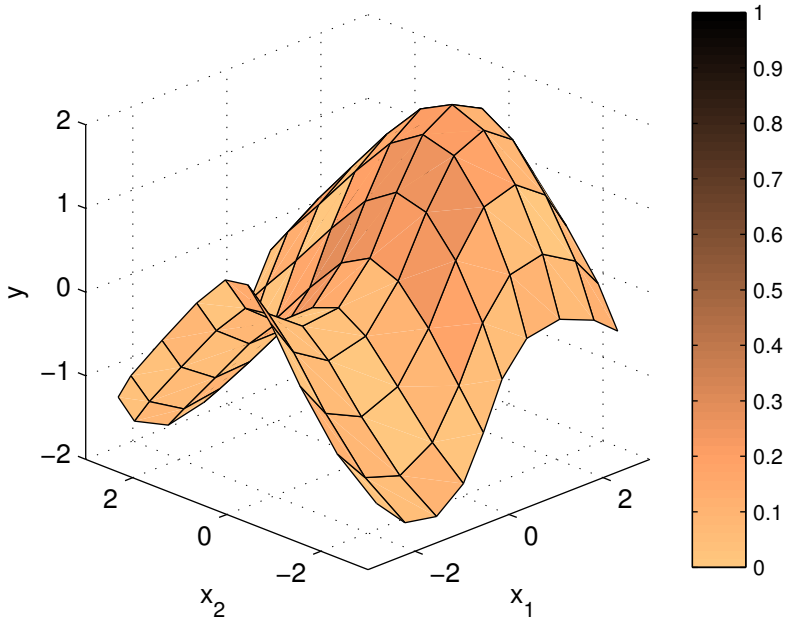


Figure 5.14: GTM with 100 (10×10) map units trained using the artificial data. Coloring depicts the difference between the map units and the surface $y = f(x_1, x_2)$. RMS imputation error is 0.231.

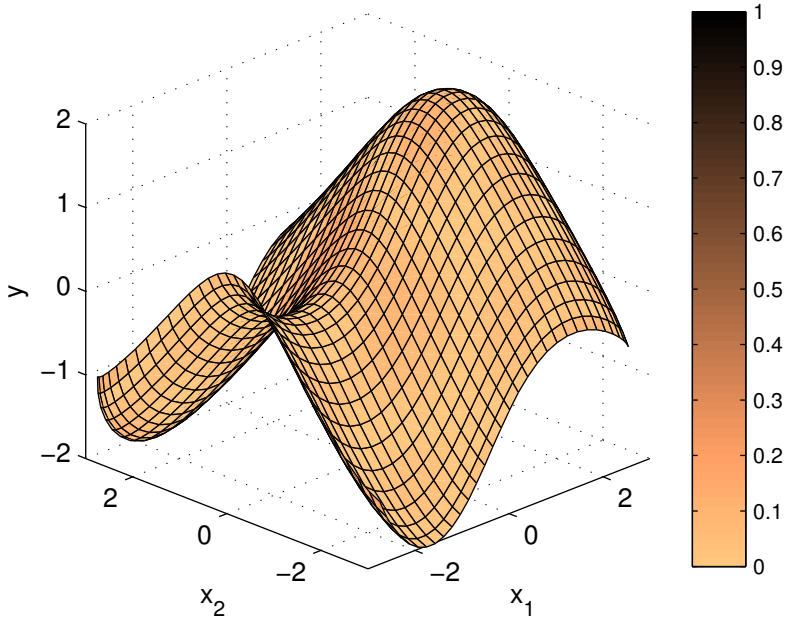


Figure 5.15: GTM with 992 (31×32) map units trained using the artificial data. Coloring depicts the difference between the map units and the surface $y = f(x_1, x_2)$. RMS imputation error is 0.139.

putational efficiency in mind. However, for the both algorithms above, the dominant computational cost arises from the evaluation of the Euclidean distances between data points and reference vectors (Bishop et al., 1998). This was verified using the MATLAB Profiler which measures where a program spends computational time.

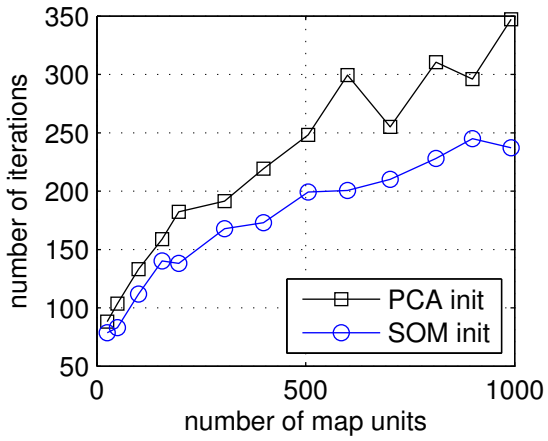
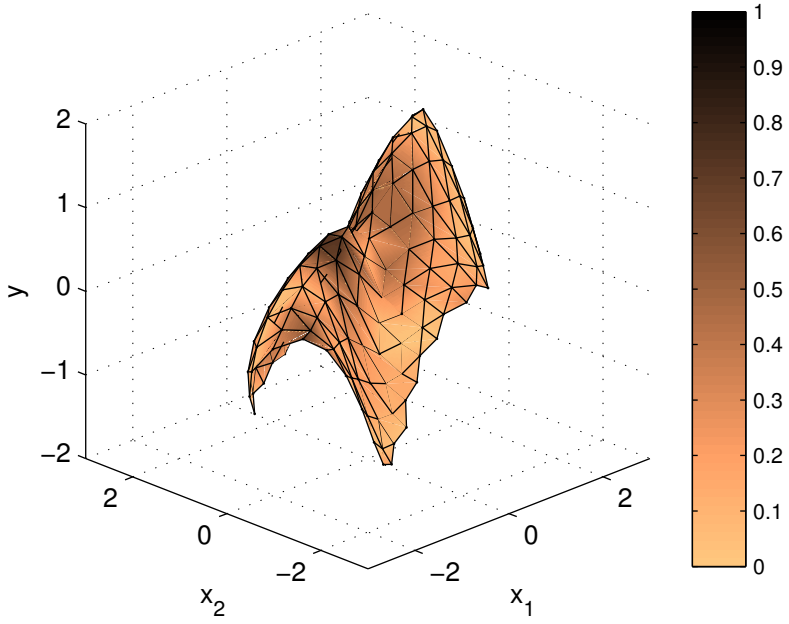
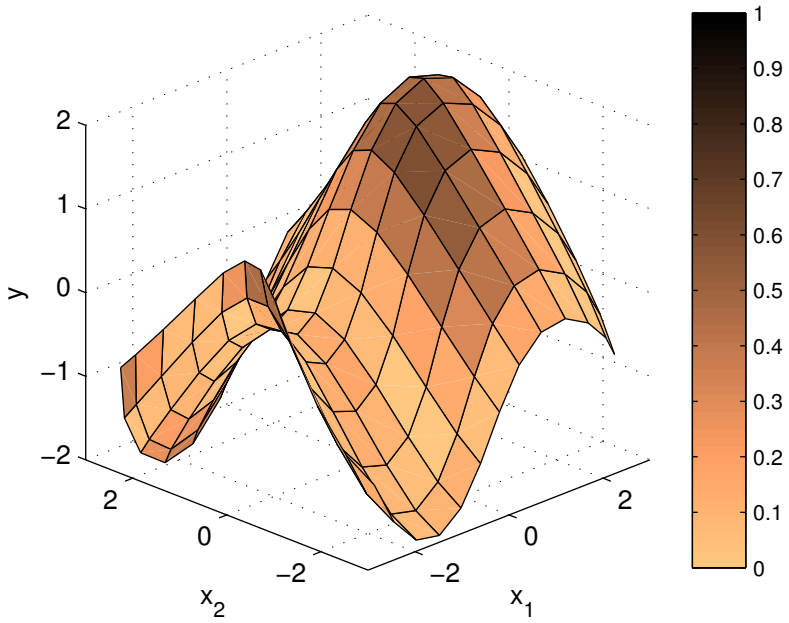


Figure 5.16: The number of iterations with respect to the number of map units until convergence of the GTM algorithm.

Further experiments included adding increased noise to the artificial data and comparing the mappings obtained using the SOM and the GTM. The array grid size was chosen according to the SOM that was able to acquire the smoothest topology (this was evaluated visually, not validated as should be done in rigorous experiments). The traditional SOM trained using the sparse data was used. Figure 5.17 shows the resulting models with RMS imputation errors 0.666 for the SOM and 0.605 for the GTM. Even though there is no profound difference between the RMS imputation errors obtained by the methods, the resulting mappings have quite distinctive properties. While the SOM suffer from the border shrinkage effect, as discussed above, the GTM tends to expand beyond the training data domain.



(a) SOM



(b) GTM

Figure 5.17: (a) SOM and (b) GTM models with 150 (10×15) map units trained using the artificial data set with increased noise having standard deviation $\sigma = 0.5$. The RMS imputation errors are 0.666 and 0.605, respectively.

5.2 Wine Data set

The wine data set consists of chemical properties of 178 wines divided in three types of wines. It is available at UCI Machine Learning Repository (Frank and Asuncion, 2010) and has been used on many research papers.[1] The variables in the data set are alcohol percentage, malic acid content, ash, alcalinity of ash, magnesium, total phenols, flavanoids, nonflavanoid phenols, proanthocyanins, color intensity, hue, OD280/OD315 of diluted wines and proline. For the purposes of this thesis, understanding the meaning of the variables is not relevant.

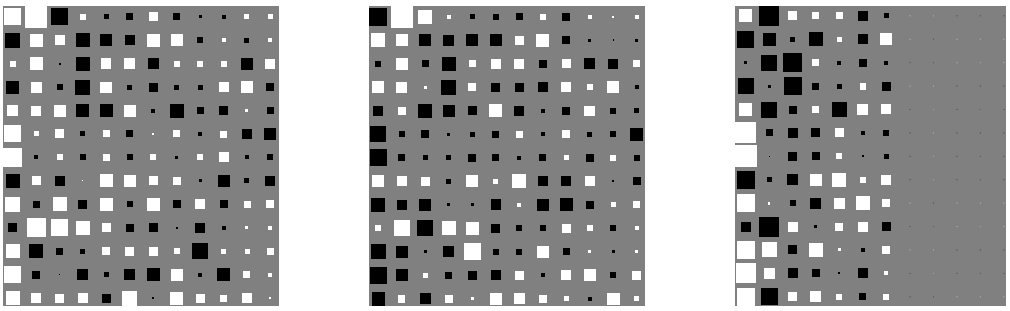
The data was normalized such that each variable had zero mean and unit variance. Artificial imputation data was created by randomly hiding values from the data according to five different missingness proportions: 0.01, 0.05, 0.1, 0.3 and 0.5. In survey imputation missingness proportion is usually low, hence emphasis was given to such data. In the resulting data sets, the data is MCAR, that is, missing completely at random. In this kind of setting the normalization before creating the test data does not give rise to any additional biases. All the results below are computed using the normalized data. The abbreviations used in this section are put together in Table 5.1.

5.2.1 Model selection with PCA

The wine data set with missing values was imputed using PCA imputation algorithm, the maximum likelihood PPCA and the variational Bayesian PCA. In order to illustrate the capacity of the VBPCA model to automatically determine the appropriate number of principal components, algorithms were run with the original dimensionality of the data $c = 13$. Figure 5.18 shows Hinton diagrams of the resulting matrices $\boldsymbol{W}$ and $\langle \boldsymbol{W} \rangle$ when the missingness proportion was 0.1. In Hinton diagram, each element of the matrix is depicted as a square whose area is proportional to the magnitude of that element and white squares correspond to positive and black squares to negative values. The algorithms sort the eigenvectors from left to right according to decreasing eigenvalue. The diagrams show that all principal components obtained by the imputation algorithm and the ML PPCA have non-zero entries which means that they are only rotating and scaling the data. Conversely, the VBPCA is able to suppress the five last principal components which means that it is able to estimate the underlying dimensionality of the data. Thus, while using VBPCA one can usually avoid model selection and the number of required principal components is determined automatically.

For further experiments, the number of principal components used for imputation algorithm PCA and ML PPCA was set to $c = 2$. Selecting more components seemed to lead to over-fitting and worse imputation results. Comparison of the RMS imputation error obtained using different PCA methods is discussed shortly together with other subspace imputation methods.

[1] See `http://archive.ics.uci.edu/ml/datasets/Wine` for a list of papers citing the wine data set.



(a) Imputation algorithm (b) PPCA (c) VBPCA

Figure 5.18: Hinton diagrams of (a) $\boldsymbol{W}$ using imputation algorithm PCA, (b) $\boldsymbol{W}$ using maximum likelihood PPCA and (c) $\langle \boldsymbol{W} \rangle$ using VBPCA on the wine data set with 10 % missing data. In diagrams each element of the matrix is depicted as a square (white for positive and black for negative values) whose area is proportional to the magnitude of that element. The variational Bayesian PCA is able to suppress five components from the matrix.

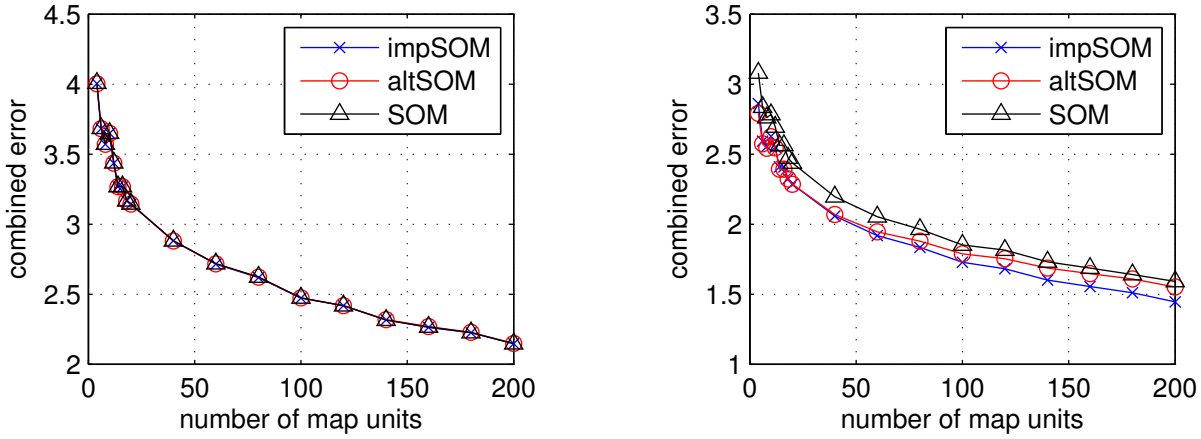
5.2.2 Model Selection with SOM and GTM

The purpose of the experiments with the wine data was to justify the choice of methods used with the proceeding data sets. In other words, the goal was to compare the methods at their best. Hence, the grid sizes for the SOM and the GTM were chosen according to their lowest RMS imputation error. Rigorous validation and the model selection is demonstrated with other data sets. The alternating SOM with weight parameter $w = 1$ was used.

Figure 5.19 shows the combined error for the different SOM imputation methods used. When only 1 % of the data is missing, there is no difference between the SOM imputation methods. However, Figure 5.19(b) shows that when 50 % of the data is missing, the imputation SOM gives the lowest combined error. The different between the imputation SOM and the alternating SOM becomes visible when the grid size is increased.

Figure 5.20 shows the behavior of the RMS imputation error with the different SOM imputation techniques. In Figure 5.20(a) with 10 % missing data there are no significant differences between the SOM methods and all methods obtain best results when map size equals 60 units. In Figure 5.20(b) with 50 % missing data, the imputation is most robust with the imputation SOM and the alternating SOM methods. All the SOM methods perform best when the grid size is in the proximity of 20 map units. Thus, this map size is selected for further comparison with other methods below. The map size with different missingness proportions was chosen in similar fashion resulting in map sizes 200 (1 %), 60 (5 %), 60 (10 %), 40 (30 %), 20 (50 %) map units.

Figure 5.21 shows the behavior of the RMS imputation error with the different GTM imputation techniques. The GTM with $M = 9$ RBFs was used. In Figure 5.21(a) with 10 % missing data, all the GTM imputation techniques improve as the map size is increased up to 63 map units. The GTMs initialized with SOM



(a) 1 % of data missing

(b) 50 % of data missing

Figure 5.19: The combined error with respect to number of map units with wine data and (a) 1 % of data missing, (b) 50 % of data missing. In (a), the results are practically identical, whereas in (b), the lowest combined error is obtained using the imputation SOM.

reference vectors obtain the best imputation performance with this map size whereas larger maps initialized using PCA are required to obtain similar imputation results. In Figure 5.21(b) with 50 % missing data, the behavior is very different. The best imputation results are obtained using the smallest reasonable map size, three map units. In this plot, there is still some evidence supporting the initialization with SOM reference vectors; maps with 9–15 units obtain slightly better imputation results when initialized with the SOM. The map sizes of GTMs with different missingness ratios were chosen in similar fashion resulting in map sizes 180 (1 %), 99 (5 %), 99 (10 %), 4 (30 %), 3 (50 %) map units. It is notable, that the GTM with only 3 map units is able to provide the lowest RMS imputation error when 50 % of data is missing. The advantage of using three map units will become clear when the results are examined in the light of the underlying cluster structure of the data below.

Table 5.2 and Figure 5.22 summarize the results of the experiments with the wine data set. One hundred randomly generated data sets with each missingness ratio were imputed using all the methods. Figure 5.22 shows the box plots of the results. Each box contains the results between 25th and 75th percentiles and the whiskers show the range of the results. Results further than 1.5 times the size of the box away from the box are considered as outliers and marked with red plus symbol. Red and black horizontal lines shows the median the mean of the results, respectively. Three best methods for each missingness proportion are enumerated on top of each sub-figure. Table 5.2 lists the mean result for each method and missingness ratio. The top three results are bold face. When the GTM with 3 map units is used, SOM initialization was not feasible, since only rectangular grid sizes are implemented in the SOM toolbox.

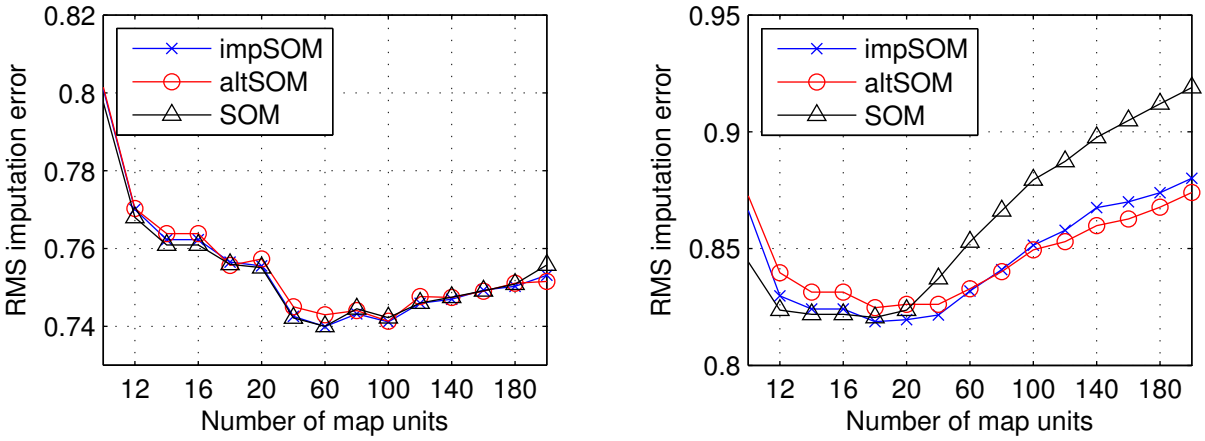
The best overall results are obtained using the VBPCA. It falls slightly second behind the GTM only in the case of 50 % missing data. The comparison between the

Table 5.1: Acronyms for different imputation methods used in Section 5.2.

Acronym	Method
m	Mean imputation
PCA1	PCA imputation algorithm
PCA2	Maximum likelihood PPCA
PCA3	Variational Bayesian PCA
SOM1	traditional SOM algorithm
SOM2	alternating SOM
SOM3	imputation SOM
GTM1	expectation of GTM with SOM initialization
GTM2	expectation of GTM with PCA initialization
GTM3	MAP of GTM with SOM initialization
GTM4	MAP of GTM with PCA initialization

SOM and the GTM does not reveal large differences. The GTM is able to provide lower RMS imputation error when the missingness proportion is 1, 10 and 50 %, and with 5 and 30 % missing data the SOM gives slightly better results.

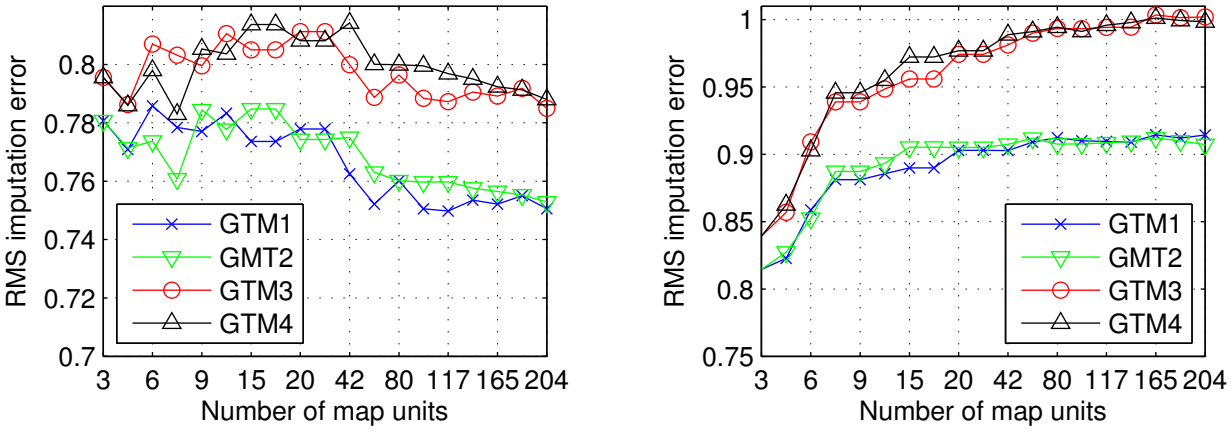
Internal comparisons within the variants for PCA, SOM and GTM reveal more explicit results. The VBPCA is superior to the other two PCA imputation techniques used. When 50 % of the data is missing, the PCA imputation algorithm and the ML PPCA provide no better results compared to naive mean imputation. There are no large differences in terms of the RMS imputation error between the different SOM imputation techniques used. The imputation SOM outperforms the tradition SOM with three missingness ratios, whereas the traditional SOM is better in one occasion, and with 10 % missing data the results are practically equal. However, it was shown in the previous section, that the imputation error does not tell the whole truth, so to speak, and for example the topologies of the maps with similar RMS imputation errors may differ considerably. The GTM using the expected values for imputation proved to be superior compared to the MAP estimation of the missing values. There is not much evidence supporting the SOM initialization in the imputation results. However, when the missingness proportion is 5 or 10 %, some improvement is gained using the SOM initialization. On the other hand, with only 1 % missing data PCA initialization gives better results.



(a) 10 % of data missing

(b) 50 % of data missing

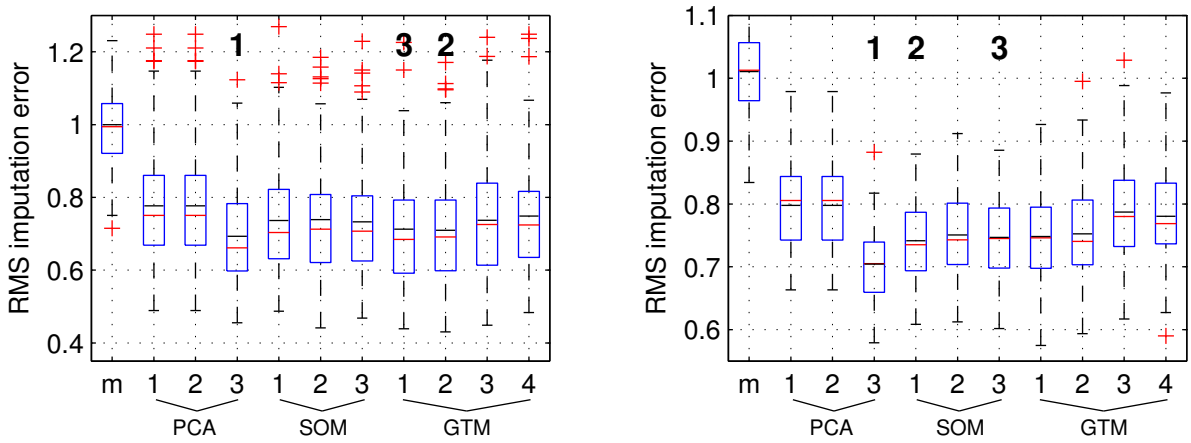
Figure 5.20: The RMS imputation error for SOM with (a) 10 % missing data and (b) 50 % missing data with different map sizes. In (a), the differences are minimal, whereas in (b), both the novel methods, the imputation SOM and the alternating SOM, are more robust when the grid size is increased. Note the nonlinear x-axis.



(a) 10 % of data missing

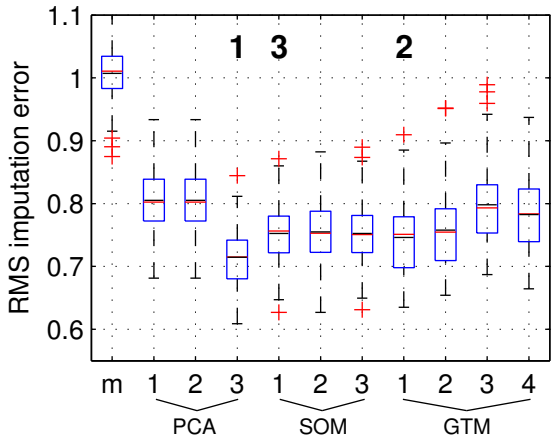
(b) 50 % of data missing

Figure 5.21: The RMS imputation error for GTM with (a) 10 % missing data and (b) 50 % missing data with different map sizes. The better results (GTM1, GTM2) are obtained using expected values for imputation. The acronyms for curves are listed in Table 5.1. Note the nonlinear x-axis.

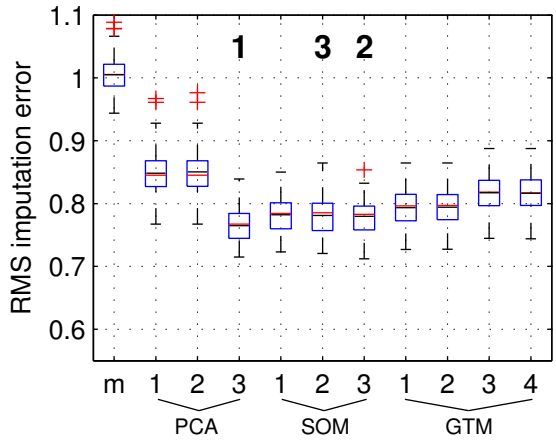


(a) 1 % missing data

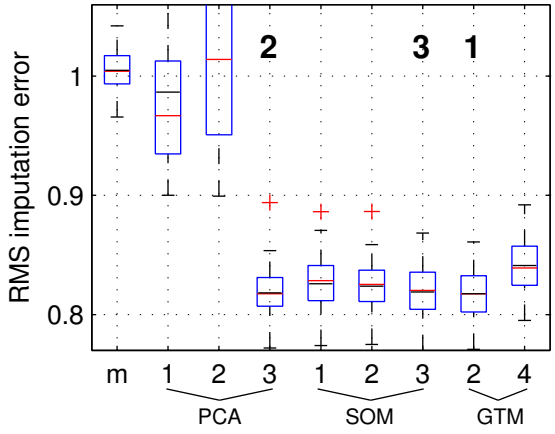
(b) 5 % missing data



(c) 10 % missing data



(d) 30 % missing data



(e) 50 % missing data

Figure 5.22: Comparison of imputation methods for PCA, the SOM and the GTM with different missingness proportions. Each subfigure (a)–(e) shows box plots of the RMS imputation errors for different methods obtained by imputing a hundred data sets with randomly generated missing data. Red and black horizontal lines shows the median the mean of the results, respectively. Three best methods for each missingness proportion are enumerated on top of each subfigure. The acronyms are listed in Table 5.1.

Table 5.2: The means of the RMS imputation errors for different imputation methods obtained by imputing a hundred data sets with randomly generated missing data and five missingness proportions using the wine data set. Three best results for each column are bold face. The acronyms are given in Table 5.1 above.

Method	1 % missing	5 % missing	10 % missing	30 % missing	50 % missing
m	1.000	1.011	1.007	1.005	1.005
PCA1	0.777	0.798	0.805	0.848	0.987
PCA2	0.777	0.798	0.805	0.850	1.778
PCA3	**0.693**	**0.705**	**0.715**	**0.765**	**0.818**
SOM1	0.737	**0.741**	**0.752**	0.782	0.826
SOM2	0.739	0.751	0.755	**0.781**	0.824
SOM3	0.733	**0.747**	0.752	**0.780**	**0.820**
GTM1	**0.713**	0.748	**0.746**	0.794	–
GTM2	**0.710**	0.752	0.758	0.794	**0.817**
GTM3	0.737	0.787	0.798	0.817	–
GTM4	0.748	0.780	0.783	0.816	0.841

An indirect and somewhat subjective way of evaluating the performance of the algorithms is to investigate the visualizations they provide while operating with missing data. Figure 5.23 shows the representation of the full wine data using the three methods. The gray-scale coloring behind SOM and GTM in Figures 5.23(b) and 5.23(c) show U-Matrix and the magnification factors of the maps, respectively. The three colors—blue, green and red—represent wines from three different wine regions, and the size of the colored markers in Figure 5.23(b) is proportional to the number of data vectors mapped to the corresponding map unit. Figure 5.23(a) shows that the three wine regions are almost[2] separable in the two-dimensional principal subspace. Hence, it is not surprising that the nonlinear methods, the SOM and the GTM, are able to produce good clustering of the data.

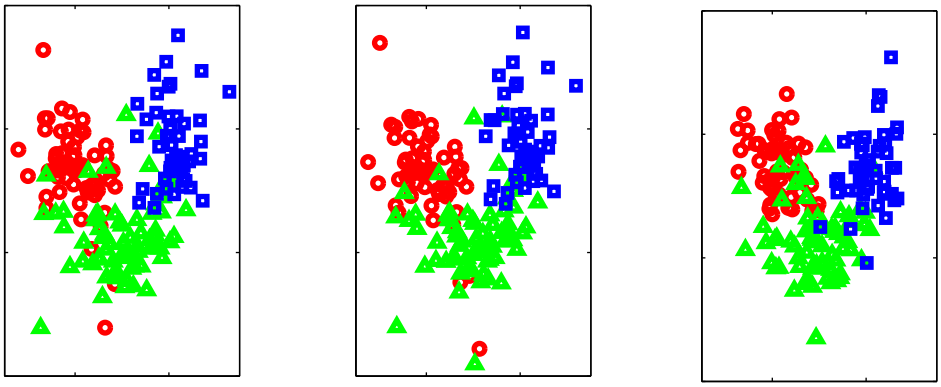
(a) PCA (b) SOM (c) GTM

Figure 5.23: Visualizations of the complete wine data set using (a) PCA, (b) the SOM and (c) the GTM. The labels of the observations belonging to three different groups, denoted by red, green and blue coloring, are used to show how well the methods are able to cluster the data.

Figure 5.24 shows visualizations of sparse wine data with 50 % of the values missing using different PCA imputation methods. The corresponding RMS imputation errors are 0.986 for the imputation algorithm, 1.008 for the ML PPCA, and 0.832 for the VBPCA. Note that the visualization using VBPCA is obtained using only two principal components, instead of all principal components used above. Comparing Figures 5.24(a) and 5.24(b) with Figure 5.23(a) reveals, that the imputation algorithm and the ML PPCA tend to disperse some data points, hence cluttering the cluster structure. From the visualizations in Figure 5.24, the one provided by the VBPCA is the most similar compared to Figure 5.23(a) indicating the robustness of the method.

Figure 5.25 shows visualizations of sparse wine data with 50 % of the values missing using different SOM imputation methods with 21 (7×3) map units. For all the maps, the RMS imputation errors are relatively equal: 0.812 for the traditional SOM and the alternating SOM, and 0.803 for the imputation SOM. Comparing the imputation

[2]Word "almost" is used here in its informal sense.



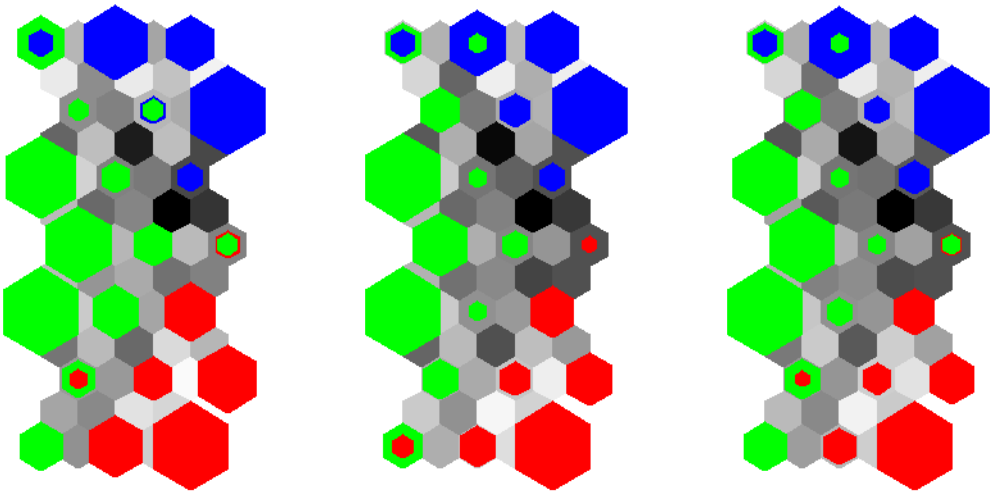
(a) impPCA (b) ML PPCA (c) VBPCA

Figure 5.24: Clustering of sparse wine data with 50 % missing data using (a) the imputation algorithm PCA, (b) the ML PPCA and (c) the VBPCA. The data is more dispersed in (a) and (b).

SOM with the other two SOMs reveals that the obtained clustering is slightly better using the imputation SOM: red and green clusters are better concentrated on the border regions of the map and there are less data points mapped on the central area of the map.

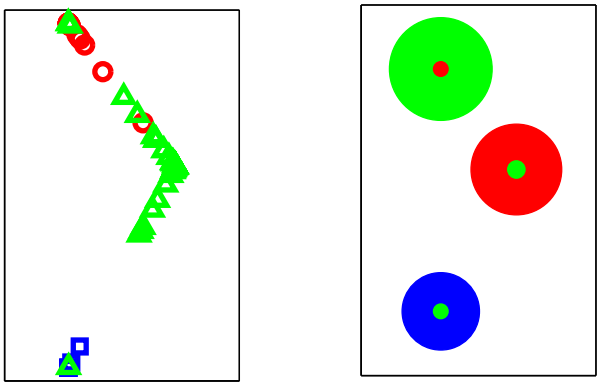
Figure 5.26 shows visualizations of sparse wine data with 50 % of the values missing using the two different GTM imputation methods. An optimal number of map units for the GTM with 50 % missingness ratio equals 3, hence the resulting visualizations differ from ones obtained using the SOM. The latent points $\boldsymbol{u}_i$ are assigned such that they form an equilateral triangle in the latent space; a configuration resembling the array of the hexagonal SOM. In the visualizations, the distances between the units are proportional to their distances in the original data space, that is, $d(\boldsymbol{u}_i, \boldsymbol{u}_j) \propto d(\boldsymbol{m}_i, \boldsymbol{m}_j)$. The resulting RMS imputation errors are 0.811 for the MAP imputation and 0.790 for the expectation imputation. In Figure 5.26(b), the size of the markers is proportional to the number of data vectors mapped to the corresponding map unit. It is notable, that the GTM is able to provide results comparable with the SOM, with only 3 map units. However, this is understandable since the data actually consists of three different clusters, wines from three distinct regions.

All in all, the experiments with the wine data set are used to motivate the choices of methods for the proceeding data sets. There are small pieces of evidence—more robust imputation with increased grid size, slightly better clustering properties and better combined error—supporting the imputation SOM over the other SOM imputations techniques. Moreover, it's novelty makes it an interesting subject of study. Regarding the GTM imputation, using the expectation of missing values proved to be the superior over the MAP estimates. This is natural, since using the MAP estimates discards information and is rarely a wise choice when dealing with multimodal distributions.



(a) SOM (b) altSOM (c) impSOM

Figure 5.25: Clustering of the wine data set with 50 % missing data using (a) the traditional SOM, (b) the alternating SOM and (c) the imputation SOM. A SOM with 21 (7×3) map units was used. The size of the colored markers is proportional to the number of data vectors mapped to the corresponding map unit.



(a) GTM1 (b) GTM3

Figure 5.26: Clustering of the wine data set with 50 % missing data using (a) the GTM and expected values and (b) the GTM and MAP estimates. A GTM with 3 map units was used. The size of the colored markers in (b) is proportional to the number of data vectors mapped to the corresponding map unit.

5.3 Nursing Survey

In the VirtualCoach research project, many wellbeing-related surveys have been implemented using a web-based system (Klapuri et al., 2011). One of them was the nursing survey, which was targeted to mothers who had experienced one or more breastfeeding periods (Mehtätalo and Lagus, 2011; Mehtätalo, 2012). The survey data consists of 65 questions, out of which 36 were answered with six-point Likert scale, from 1101 respondents. For the purposes of this thesis, the answers to the 36 questions mentioned were taken apart from the rest of the data and used to demonstrate single imputation. The questions can be seen in Appendix A. Single imputation might be useful, for example, if one is interested in the imputations on the respondent level. That is, one would like to predict how a particular respondent would have most likely answered questions, which the values are missing for. A suitable measure for evaluating the single imputation performance is the RMS imputation error. In order to show the problems arising from the single imputation and to motivate multiple imputation, which is studied further in the proceeding Section 5.4, means and standard deviations of the imputed data set are also under inspection.

Figures 5.27 and 5.28 show the histograms of the variables in the data as well as number of missing values in each variable. In total, only 0.46 % of the values in the data were missing. Observing the histograms reveals many differently shaped distributions; skewed to the left (e.g., Q2 and Q12), skewed to the right (e.g., Q4 and Q7), peaked in the middle (e.g., Q10 and Q24), peaked on the both edges (e.g., Q34 and Q35), relatively uniform (e.g., Q22 and Q32) and many combinations of the properties mentioned. Six variables with relatively different distributions were chosen to test the single imputation on: Q2, Q9, Q22, Q24, Q32 and Q33. This kind of selection is aimed to get good overall insight of the versatility of the tested methods. For example, mean imputation can be expected to work well on variabled with peaked distribution, but it works worse on variables peaked on the both edges.

For testing, 100 randomly selected values from each five variable were taken aside, that is, the other responses of the corresponding respondent were kept in the validation data set, but the test values were marked as missing. For each method requiring model selection, a 10-fold cross-validation was conducted.

The model selection can be done based upon many different criteria. Having the single imputation task in mind, one may randomly hide some known values of the imputed variables and evaluate the RMS imputation error of models with different complexity. This was also the first approach used in this thesis. Validation was done based on the same six variables mentioned above and the validation indices are chosen independently for each variable, that is, on each cross-validation fold, missing values were scattered on different rows in each validation variable. Second alternative is to use some other error criterion on validation data which is kept aside during the training phase. For probabilistic models, this equals evaluating the likelihood of a model given separate validation data. For the SOM, combined error was used instead of the likelihood of the model.

Imputation was also conducted using models which assume binary data. Without heuristic tricks, these models provide discrete imputations. Thus, also the imputations provided by the models which assume continuous data were rounded to integers

before evaluating the RMS imputation error of the test data. Evaluating the error on continuous imputations would bias the comparison between methods assuming binary and continuous data. Also, it makes sense to use the same discrete scale which would be used on single imputation.

The nursing survey data was first imputed using the mean imputation. This provided a useful base-line result which other methods are supposed to improve. The RMS imputation error using mean imputation was 1.616.

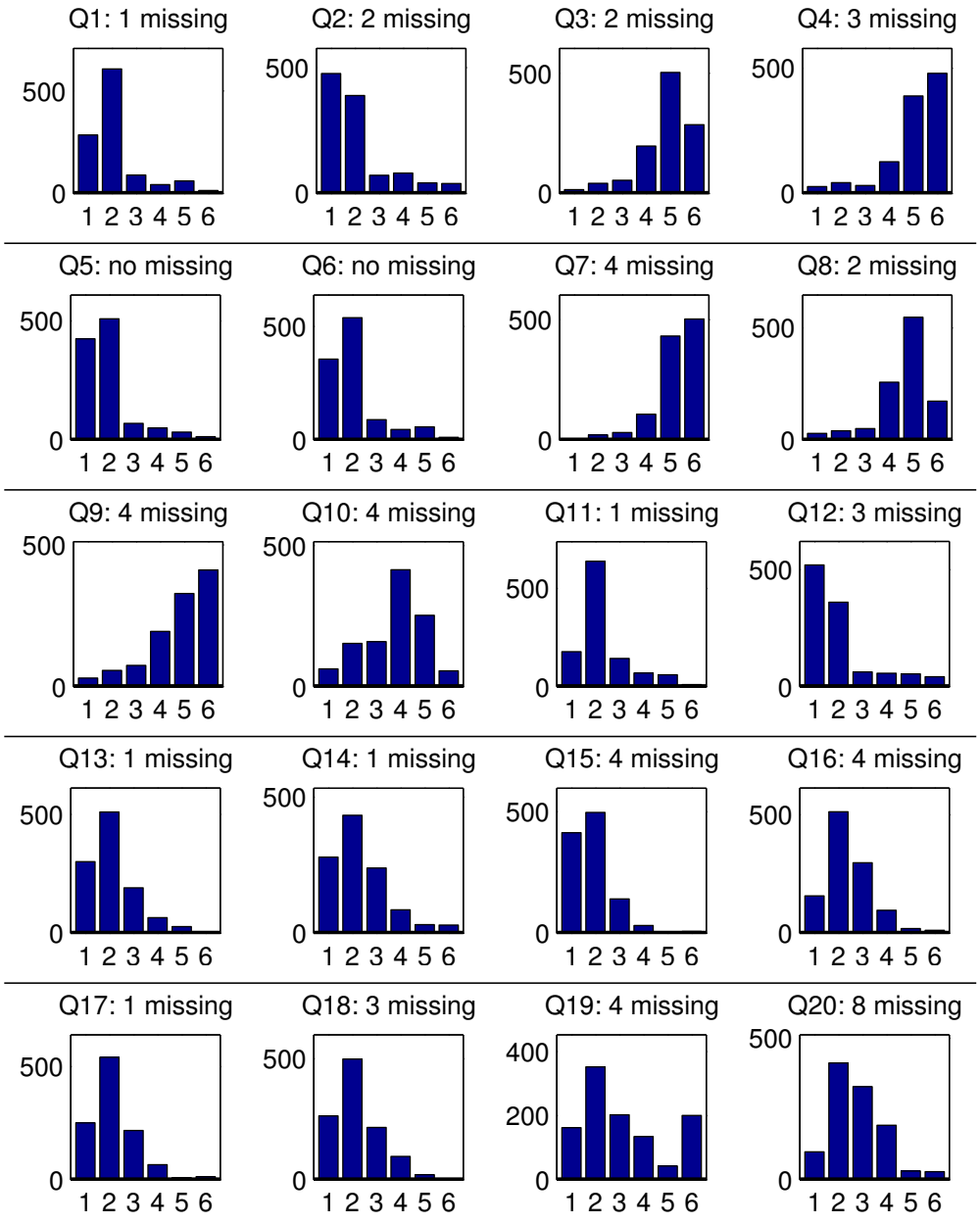


Figure 5.27: Histograms of the variables (Questions) and number of missing values in each variable in the nursing survey (1 of 2). Q2, Q9, Q22, Q24, Q32 and Q33 were used as test data.

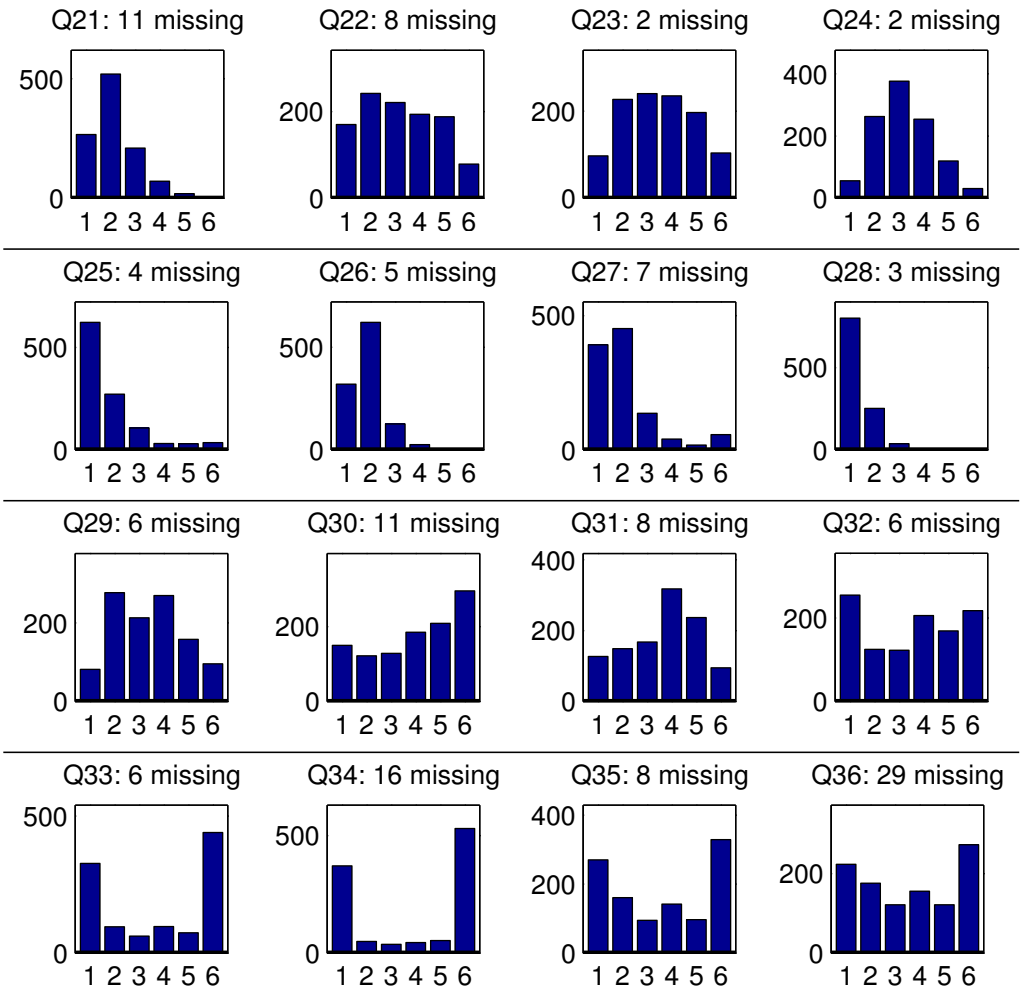


Figure 5.28: Histograms of the variables (Questions) and number of missing values in each variable in the nursing survey (2 of 2). Q2, Q9, Q22, Q24, Q32 and Q33 were used as test data.

5.3.1 Imputation with VBPCA

Imputing the nursing survey data with the VBPCA is a straightforward task. No model selection is required, since the VBPCA is able to accomplish automatic relevance determination as explained above. Figure 5.29 shows the Hinton diagram of $\langle \boldsymbol{W} \rangle$ for the data. More than half of the 36 principal components are suppressed. The resulting RMS imputation error for the test data is 1.161.

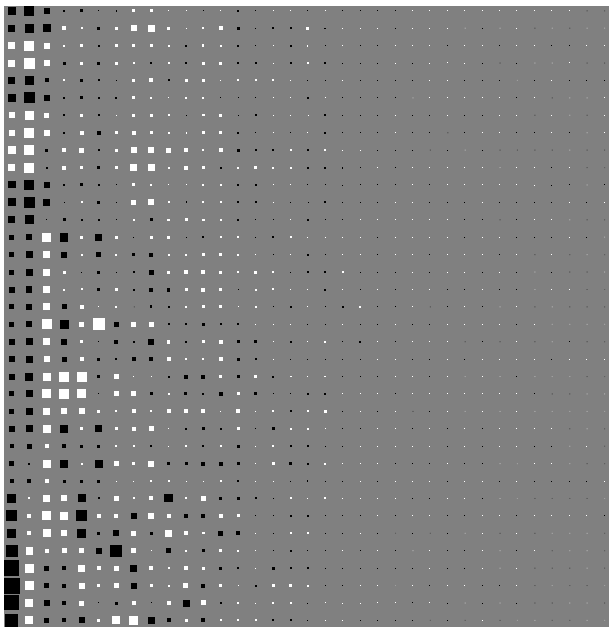


Figure 5.29: Hinton diagram of $\langle \boldsymbol{W} \rangle$ using VBPCA for the nursing survey data. More than half of the components are suppressed from the matrix.

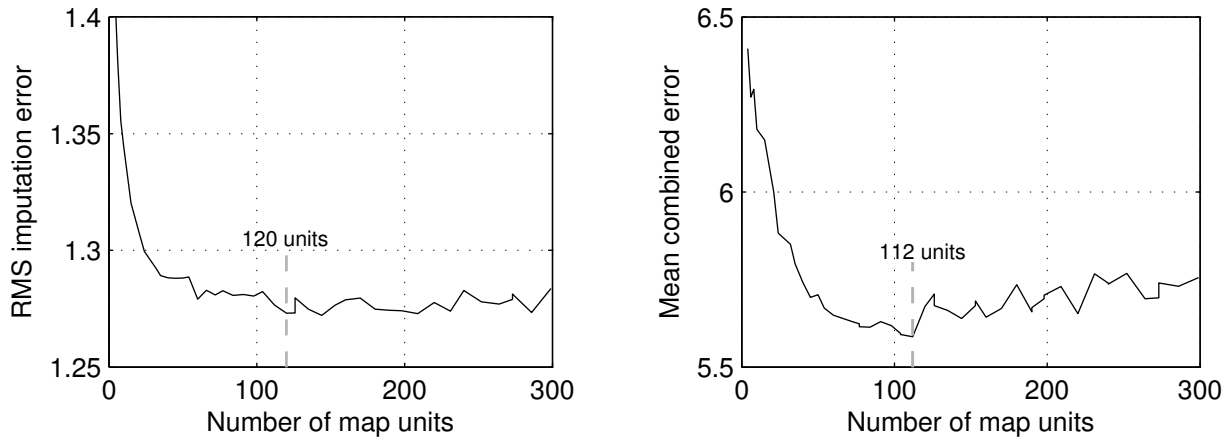
5.3.2 Imputation with SOM

The size of the imputation SOM used in imputation was selected by 10-fold cross-validation based on the RMS imputation error and the combined error. For the validation with the RMS imputation error, all respondents were used in training for all folds and randomly chosen validation values were hidden for each fold, as explained above. For the validation with combined error, nine parts out of ten were used for training the model on the last part was used to evaluate to combined error on each fold.

Figure 5.30 shows the results of the validation. The best model can be chosen by combining the results from two validation techniques. Figure 5.30(b) clearly shows that the minimum of the combined error is obtained around the model with 112 map units. This result can be expected to be slightly too small, since the map is trained with less data compared to the validation on the RMS imputation error and the final imputation task (remember that one tenth of the data was taken aside on each validation fold). Figure 5.30(a) shows that the best RMS imputation error on validation data is obtained using a model with 120 map units. There are some more

complex models that obtain infinitesimally better results, but even with only the RMS imputation error validation results at hand, the selection of a model with 120 map units can be motivated with parsimony: a simpler model is preferred over the more complex one. Combining the results from the two validation results gives evidence for selecting a SOM with 120 map units.

Imputing the test data with an imputation SOM with 120 map units gives an RMS imputation error 1.265. This is significantly better compared to the base-line error 1.616 given by the mean imputation but worse compared to the result obtained by the VBPCA. Using the traditional SOM of the same size gives RMS imputation error 1.264, hence there is no big difference between the traditional SOM and the imputation SOM. The combined errors for imputation and traditional SOM are 5.356 and 5.404, respectively, which are again in slight favor for the imputation SOM. The results are combined in Table 5.3.



(a) RMS imputation error

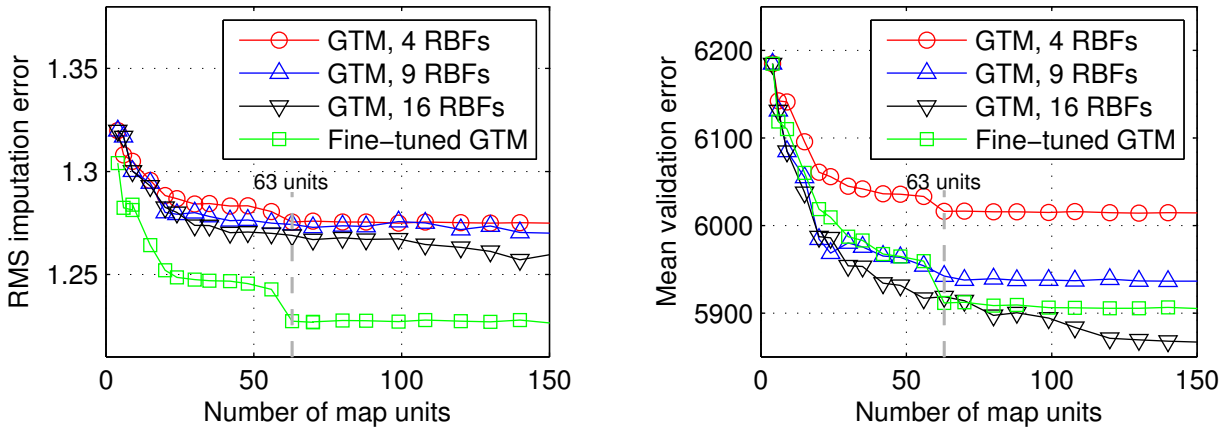
(b) Mean combined error

Figure 5.30: The model selection results of the imputation SOM using 10-fold cross-validation. Figures show (a) the mean of the RMS imputation errors and (b) the mean combined errors between the folds with respect to the number of map units. The grid size that the corresponding validation suggests is shown with vertical dashed line.

5.3.3 Imputation with GTM

The size of the GTM was selected similarly as in the case of the SOM except that instead of using the combined error, the negative log likelihood evaluated using the validation data was used as an error measure. The GTM was initialized with the reference vectors of the SOM of the same grid size. Red, blue and black curves in Figure 5.31 show the validation results for GTMs with $M = 4$ (2×2), $M = 9$ (3×3) and $M = 16$ (4×4) RBFs, respectively. Both the RMS imputation error and the validation error give evidence for choosing the map size $K = 63$ units for the first two models. For the GTM with $M = 16$ RBFs, a model with $K = 140$ map units was chosen. Imputing the test data with the three GTMs provided RMS imputation

errors 1.254 ($M = 4$), 1.248 ($M = 9$) and 1.245 ($M = 16$). When the GTM was initialized using the PCA the resulting RMS imputation errors were slightly worse.



(a) RMS imputation error

(b) Mean validation error

Figure 5.31: The model selection results of the GTM using 10-fold cross-validation. Figures show (a) the mean of the RMS imputation errors and (b) the mean validation error, the negative log likelihood evaluated using validation data, between the folds with respect to the number of map units. The green curve denotes the "fine-tuned" GTM where number of RBFs, M, is increased during the training whereas the other curves are results of the traditional GTM where number of RBFs, M, is kept constant. The grid size that the corresponding validation suggests is shown with vertical dashed line.

As suggested in Section 4.3.6, it is possible to control the stiffness of the GTM by altering the number of RBFs, M. As comparison to the three GTMs above, a fine-tuned GTM using three different RBF-network structures was used. The initial GTM was trained using $M = 4$ (2×2) RBFs, followed by maps with $M = 9$ (3×3) and $M = 16$ (4×4) RBFs. Each map was trained until convergence before M was increased.

The green curve in Figure 5.31 shows the validation results using the fine-tuned GTM. The suggested map size is 63 units. There is significant improvement in the RMS imputation error validation results. For the test data, the RMS imputation error was 1.247, which is slightly worse compared to the unmodified GTM. However, since the validation provides strong evidence supporting the fine-tuned GTM, this test result may be a statistical defect. Again, initializing the first GTM with PCA instead of the SOM gives slightly worse result 1.271. All in all, the GTM was able to provide slightly better single imputation results compared to the SOM, but the results falls far behind the results obtained using the VBPCA. See Table 5.3 for further comparison.

5.3.4 Binary Data

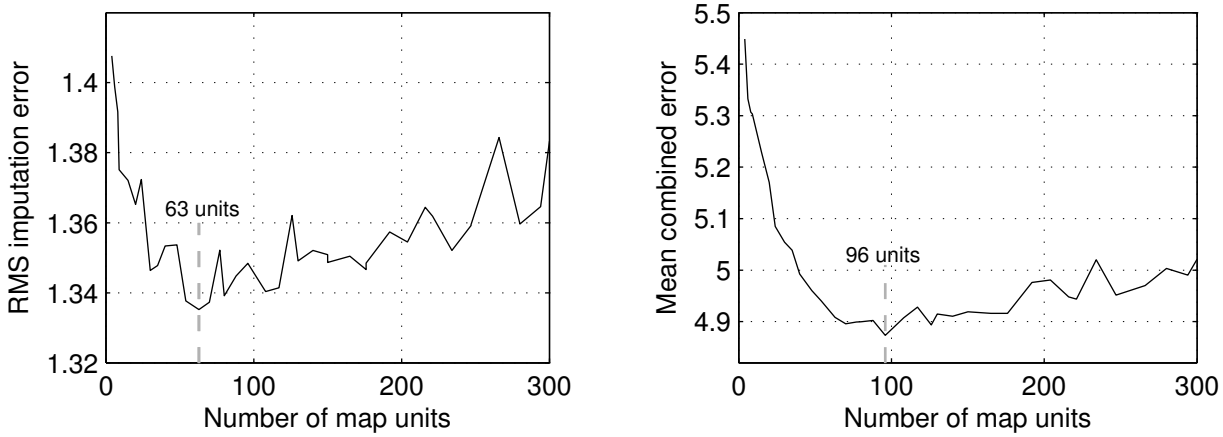
The Likert-scale data is discrete but we have so far used continuous-data methods for modeling it. Logical follow-up is to take the discrete nature of the data into

account and conduct experiments using discrete-data methods. One possibility is to use binarization, where the Likert-scale variables are encoded with few binary variables as follows:

$$
\begin{aligned}
1 &\rightarrow 00000 \\
2 &\rightarrow 10000 \\
3 &\rightarrow 11000 \\
4 &\rightarrow 11100 \\
5 &\rightarrow 11110 \\
6 &\rightarrow 11111
\end{aligned}
$$

This kind of scheme has been used, for example, in Kozma et al. (2009). After binarization, any binary-data method can be applied to the data.

The results of the SOM map size validation on the binarized, unnormalized nursing survey data are shown in Figure 5.32. This time the validation results are more ambiguous: grid sizes suggested by two different validation strategies suggest using SOMs with 63 and 96 map units. The fluctuating behavior of the RMS imputation error is probably due to the discrete nature of the data. The smaller grid size suggested by the validation on the RMS imputation error could be motivated by parsimony. However, the combined error of the validation data offers a smoother curve, thus, providing less ambiguous evidence for the model selection. Imputing the test data using the imputation SOM with 96 (12×8) map units, gives RMS imputation error 1.296. This result is slightly worse compared to the results without binarization.

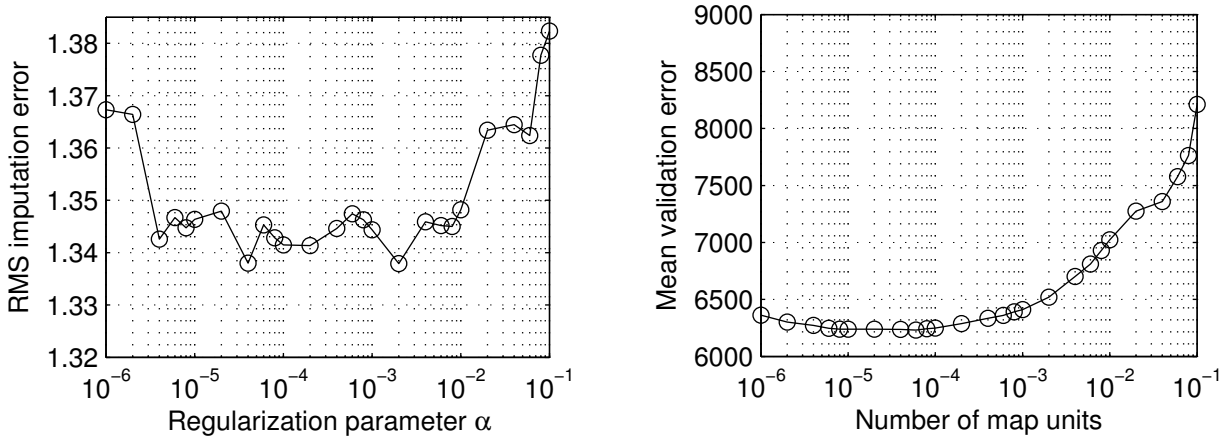


(a) RMS imputation error

(b) Mean combined error

Figure 5.32: The model selection results of the imputation SOM on the binarized nursing survey data using 10-fold cross-validation. Figures show (a) the mean of the RMS imputation errors and (b) the mean combined errors between the folds with respect to the number of map units. The grid size that the corresponding validation suggests is shown with vertical dashed line.

The validation was also done for the Bernoulli GTM, described in Section 4.3.4. This time, the number of map units was fixed to $K = 96$, and the goal was to validate with respect to the regularization parameter α in (4.37). Again, combining the evidence from both validation curves, suggests choosing the regularization parameters $\alpha = 10^{-4}$. The relatively large α between 10^{-3} and 10^{-2} can be ruled out, since the validation error is far from the optimal in this region of alphas. On the other hand, large α means more regularization and less complex model. Hence, $\alpha = 10^{-4}$ can be preferred over smaller alphas and more complex models in terms of parsimony.



(a) RMS imputation error

(b) Mean combined error

Figure 5.33: The model selection results of the Bernoulli GTM using 10-fold cross-validation. Figures show (a) the mean of the RMS imputation errors and (b) the mean combined errors between the folds with respect to the regularization parameter α.

Table 5.3: The RMS imputation errors for different methods on the nursing survey data. The best result obtained by the VBPCA is bold face.

Method	RMS imputation error
Mean imputation	1.616
VBPCA	**1.161**
SOM	1.264
impSOM	1.265
GTM	1.245
fine-tuned GTM	1.247
Binary SOM	1.296
Bernoulli GTM	1.271

5.3.5 Sample-Wide Statistics

The single-imputations data sets obtained were used to estimate the sample-wide statistics of the nursing survey data. In order to compare with MICE, single imputations conducted using mice package in R (van Buuren and Groothuis-Oudshoorn, 2011) were taken into comparison. Ten multiply imputed data sets were created, and their means were used as single imputations for the missing values. The mice package supports proportional odds logistic regression, which commonly used for ordered categorical variables such as Likert-scale, only up to five levels. Thus, regular Bayesian linear regression for continuous data was used in the regression models.

Figure 5.34 shows histograms of the imputed values together with the histogram of the test values. The most significant differences between the methods can be seen in Question 33. Only the methods using binary data are able to capture the multimodal distribution of the test data. Also, the distribution of test values in Question 2 seems to be challenging for the imputation methods. Again, the binary methods together with the VBPCA capture the distribution of the data best. Note that based on the visual inspection of the histograms, MICE does not provide the best single imputation results.

Tables 5.4 and 5.5 show how the imputed values bias the sample-wide statistics. The estimates closest to the real statistics, shown on the first row, are bold face. In Table 5.4, the means of the imputed questions are compared to the means of the full data. The few missing values in each question, shown in Figures 5.27 and 5.28, are ignored and the statistics are evaluated based on the values present in the test data. All the methods acquire good estimates for the means. It is notable, that for all but Questions 2 and 9, one can acquire better estimates of the means, compared to the mean imputation, using any other method. The mean imputation should provide an unbiased estimate of the mean, since the test data was sampled from each question randomly.

Table 5.5 shows the standard deviations of each question after the imputation. All methods underestimate the standard deviations of the imputed variables, apart from one exception. In Question 33, binary methods provide imputed data which overestimates the standard deviation. Moreover, these methods, the Binary SOM and the Bernoulli GTM, provide imputations whose standard deviations are closest to the corresponding statistics of the test data.

Even though we have now acquired some estimates of the sample-wide statistics, we have no insights how reliable these estimates are. In order to evaluate the uncertainty of these estimates, caused by the missing data, we have to utilize multiple imputation. This is demonstrated with the last data set in the proceeding section.

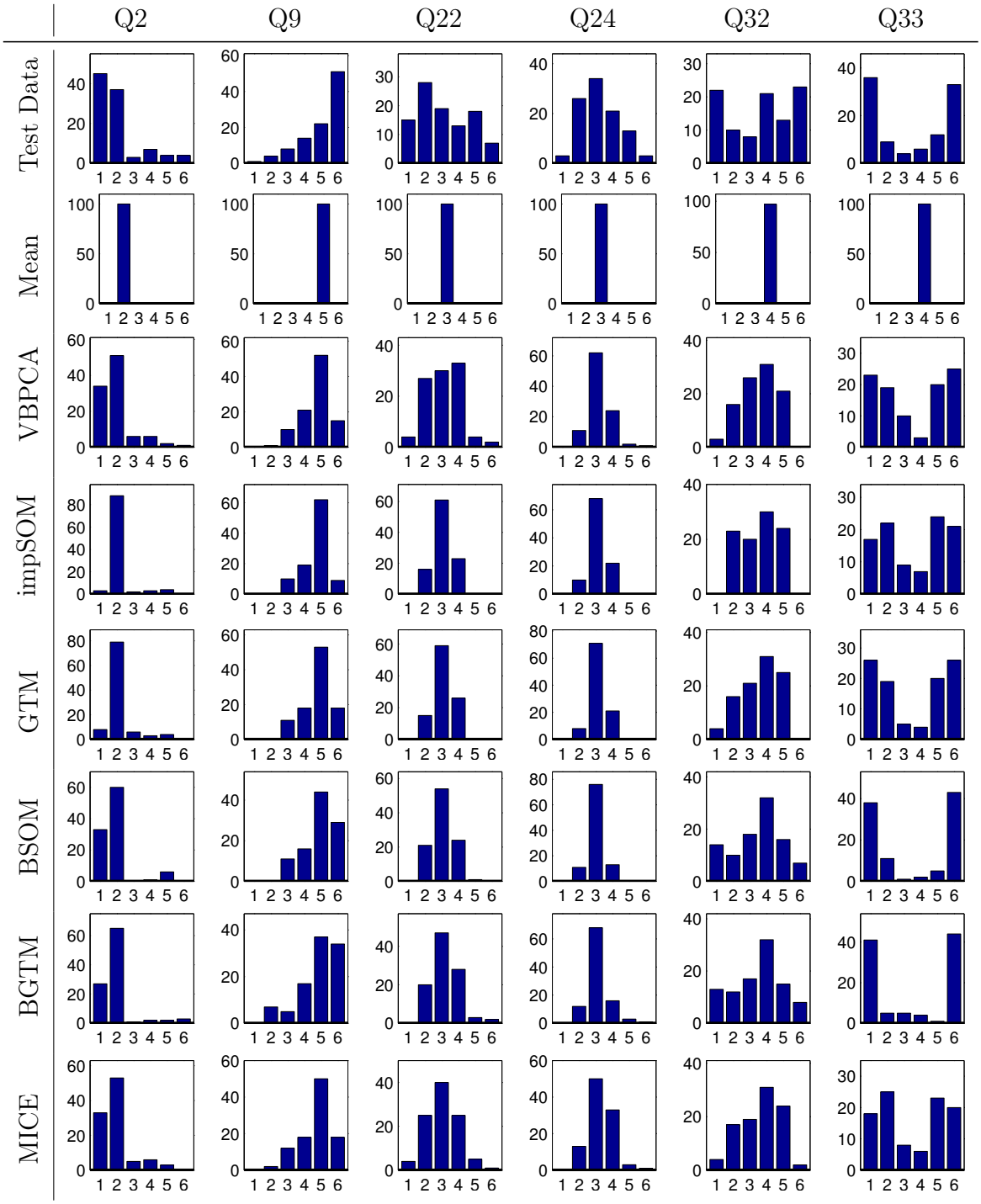


Figure 5.34: Histograms of the test data and the imputed values using different imputation methods. The binary methods are abbreviated to BSOM (binary SOM) and BGTM (Bernoulli GTM).

Table 5.4: The sample-wide means of the imputed variables after imputations. The real mean, computed using the test data, is shown on the first row.

Method	Q2	Q9	Q22	Q24	Q32	Q33
Real mean	2.035	4.796	3.206	3.193	3.514	3.744
Mean imputation	**2.035**	**4.791**	3.195	3.171	3.546	3.791
VBPCA	2.033	4.772	**3.207**	**3.194**	**3.513**	3.748
impSOM	2.051	4.764	3.201	3.182	3.507	3.756
GTM	2.050	4.771	**3.205**	3.183	3.509	**3.746**
Binary SOM	2.024	4.783	3.199	3.173	3.500	3.749
Bernoulli GTM	2.032	4.779	3.213	3.183	3.501	**3.746**
MICE	2.029	4.764	3.199	3.197	3.512	**3.746**

Table 5.5: The sample-wide standard deviations of the imputed variables after imputations. The real standard deviation, computed using the test data, is shown on the first row.

Method	Q2	Q9	Q22	Q24	Q32	Q33
Real std	1.306	1.300	1.522	1.152	1.838	2.169
Mean imputation	1.243	1.246	1.451	1.100	1.757	2.065
VBPCA	1.274	1.272	**1.485**	1.117	1.788	2.155
impSOM	1.262	1.266	1.462	1.111	1.782	2.139
GTM	1.265	1.272	1.463	1.110	1.785	2.155
Binary SOM	1.276	1.277	1.466	1.110	**1.804**	**2.181**
Bernoulli GTM	**1.280**	**1.292**	1.473	1.118	**1.804**	2.182
MICE	1.275	1.278	1.480	**1.123**	1.788	2.141

5.4 15D Instrument Data

15D Instrument is a comprehensive and self-administered survey tool for measuring the *Health-Related Quality of Life* (HRQoL) among adults (Sintonen, 2001). 15D consists of 15 questions on five-point Likert scale. Results can be used to examine the HRQoL on respondent and population level. In clinical studies, 15D is often used to test whether a known treatment or intervention has an effect on HRQoL of the patients. It can be used to compare patient groups, as well. Separate versions have been developed for adolescents (16D) and children (17D). The 15D questionnaire is shown in Appendix B.

There is a standard equation for calculating a standardised and sensitive single index score measure for HRQoL based on the 15D responses,

$$v_{\mathrm{HM2}} = \sum_{j=1}^{15} l_j(x_j) w_j(x_j), \quad v_{\mathrm{HM2}} \in [0, 1], \tag{5.2}$$

where $l_j(x_j)$ is a set of positive constants for the j^{th} dimension, representing the relative importance of the dimension at its various levels, and $w_j(x_j)$ is a function, representing the relative value of the various levels of the j^{th} dimension (Sintonen, 2001).

In this thesis, 15D survey data among four unidentified patient groups of size 408, 243, 385 and 297 were used. The survey results in the groups contained 58 (1.0 %), 9 (0.3 %), 53 (0.9 %) and 79 (1.8 %) missing values, respectively. Most of the missing values occurred in a single question about respondents sexual activity. In addition to the 15 questions, age and gender of each patient are included in the data and the imputation models. For the SOM, the Likert-scale data was binarized and the grid size was validated based on the combined error as above. The resulting grid size was 90 map units. For the GTM, grid size of 65 units was chosen and regularization parameters $\alpha = 10^{-4}$ was selected using 10-fold cross-validation. For each patient group, ten multiply imputed data sets were produced using the SOM, the VBPCA and MICE. Again, MAR data is assumed although in this case, this assumption is unlikely to hold strictly. For example, some people with problems with their sex life—the question most often left unanswered is about respondents sexual activity—may rather leave this question unanswered. However, a part of this behavior can be predicted with other variables, age and gender included. Investigating how this missingness may be modeled is and interesting research question beyond the scope of this thesis.

After the imputation, the v_{HM2}-score in (5.2) was computed for all patients and the pooled estimates of the mean v_{HM2}, depicting the group level HRQoL, for each patient group were computed. Table 5.6 summarizes the results. Ignoring the respondents with missing values (the first row) biases the results slightly. Notably, the difference between the most naive approach, the mean imputation, and MI methods is minimal. However, the mean imputation does not provide any measure of uncertainty of the patient group level mean estimates. In addition to the group level mean estimates, pooling the multiply imputed data sets provides standard deviations of the estimated means, which are shown in parenthesis next to the corresponding mean estimate in

Table 5.6: The pooled estimates of the mean v_{HM2}, depicting the group level HRQoL, and their estimated total variances in parenthesis, for the four patient groups under study. All MI methods and the mean imputation provide nearly equal estimates whereas there is slightly more bias when the patients with missing values are ignored (the first row).

Method	group 1	group 2	group 3	group 4
Ignore	0.779	0.802	0.810	0.796
Mean	0.776	0.800	0.807	0.793
MICE	0.775 (0.112)	0.800 (0.110)	0.807 (0.112)	0.793 (0.091)
VBPCA	0.775 (0.111)	0.801 (0.110)	0.807 (0.112)	0.793 (0.090)
GTM	0.775 (0.111)	0.801 (0.110)	0.807 (0.112)	0.793 (0.090)
BSOM	0.775 (0.111)	0.801 (0.110)	0.807 (0.112)	0.793 (0.090)

Table 5.6. The dominant term in the estimated total variance (2.9) is the within-imputation variance $\bar{U}$.

The most common statistical tests used to compare 15D scores are two-sample t-test and Wilcoxon rank sum test, which is equivalent to a Mann-Whitney U-test (Hollander and Wolfe, 1999). In this thesis, two sample t-test, which assumes that both data sets follow normal distributions with unknown but equal variance, is used. Figure 5.35 shows a histogram of v_{HM2}-score in all four patient groups. Visual inspection confirms that the data can be assumed to follow normal distribution.

In terms of comparing two patients groups, groups 1 and 2 provide an interesting comparison. First, t-tests were conducted for data where patients with missing values were ignored, and data imputed using the mean imputation. The resulting p-values were 0.0139 and 0.0054, respectively. In both cases, t-test was unable to reject the null hypothesis, that the data follows the same normal distribution, on 0.005 confidence level. Thus, the conclusion is that there is no statistically significant difference between the patient groups 1 and 2. In MI, statistical tests can be done between all pairs of imputed data sets. Inferences are then drawn by combining the results of the tests. Similar testing was conducted for all pair-wise imputed data sets in patient groups 1 and 2 resulting in 100 test results for each MI method. Using the VBPCA and the Binary SOM, 53 out of the 100 tests reject the null hypothesis on 0.005 confidence level while the mean p-value was 0.0050 for the both methods. Using MICE and the GTM, 78 and 89 out of 100 tests reject the null hypothesis, respectively, giving the most clear evidence of difference between the patient groups. The corresponding mean p-values were 0.0046 for the MICE and 0.0045 for the GTM.

In other pair-wise comparisons between the four patient groups on 0.005 confidence level, there is no such controversy between different MI techniques. According to the tests, group 1 has similar overall HRQoL with group 4, and groups 2, 3 and 4 are also similar in this sense. If the confidence level is loosened to be 0.05, controversial conclusions arise when comparing groups 1 and 4. All MI methods and mean

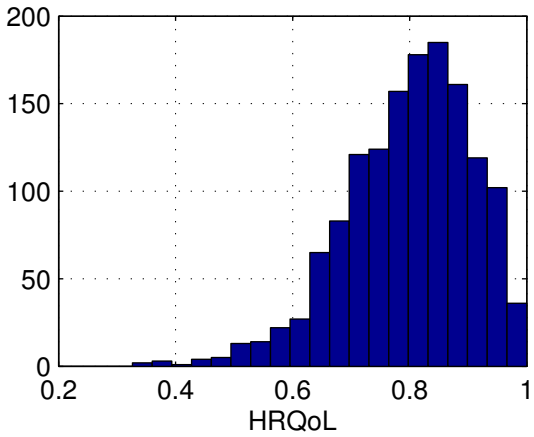


Figure 5.35: A histogram of the v_{HM2}-scores, representing the health-related quality of life (HRQoL) of the patiens, of all patient groups. The data follows approximately the normal distribution.

imputation suggest that the patients have difference in their average HRQoL while ignoring the patients with missing values declares the groups similar.

The statistical testing above was conducted in order to demonstrate difficulties one may confront when performing statistical analysis on data with missing values. In the task above, there is no "correct answer" to compare the results with. However, the results obtained using MI techniques are more reliable, since the take the uncertainty over the missing values into account. The most important finding was, that MI conducted using the subspace methods is able to provide results comparable with MICE.

Chapter 6

Discussion and Conclusions

In this thesis, the missing value imputation task was approached using three subspace methods, PCA, the SOM and the GTM. Properties of these methods in presence of missing data and applicability to multiple imputation were studied.

In all the experiments, missing-at-random data was assumed, that is, mechanisms giving rise to missing data were ignored. This is known to be a proper approach if other observed variables can be assumed to account for the missingness, that is, the missingness of y can be predicted by observed values x in the same observation. However, in many data sets, the observed data does not fully explain the missingness, hence the data is not-missing-at-random and rigorous modeling requires nonignorable models which take the missingness mechanism into account. These models did not belong to the scope of this thesis. All in all, usually nonignorable models are able to provide reasonably good imputations given that the data is not censored.

The contributions of this thesis consist of improvements in the SOM in presence of missing data as well as novel investigations and improvements in the GTM. A novel revision to the SOM algorithm, the imputation SOM, which borrows an idea from the probabilistic framework, was proposed. It was shown that the imputation SOM is more robust in terms of the combined error and the RMS imputation error compared to the traditional SOM algorithm in presence of missing data.

Self-organization, initialization and regularization of the GTM were studied. It was discovered that the GTM can benefit an initialization done using the reference vectors of the SOM. It was also shown that loosening the stiffness of the map during the training—a stunt that is usually carried out while training the SOM in order to allow better self-organization and speeding up the learning—also benefits the GTM. This suggests, that other engineering adjustments implemented in the SOM toolbox, such as hexagonal grid, may benefit the GTM in a similar manner. The imputation using the expectation over the missing values was found out to be the better choice for single imputation compared to using the MAP estimates of the missing values.

A variant of the GTM for binary data, the Bernoulli GTM, with novel regularization was implemented and applied to binarized survey data. It was shown, that binarization may be a useful transformation in order to model a discrete survey data and the methods modeling the binary data—the binary SOM and the Bernoulli GTM—were better able to capture the distribution of the missing test values in the nursing survey data.

The VBPCA was shown to be superior in single imputation task in comparison to the other methods used. Furthermore, the VBPCA can automatically select an optimal number of principal components, that is, it does not require discrete model selection. The encouraging results of the binary SOM and the Bernoulli GTM suggest that PCA adapted for binary data (see, e.g., Kozma et al., 2009) might further improve the properties of the VBPCA.

The MI paradigm offers a robust framework for assessing the uncertainty of any population level statistic after the missing value imputation. This property was demonstrated using 15D instrument survey data. The pooled estimates for population level means, as well as their estimated standard deviations, were nearly equal when comparing the results obtained using MICE and the subspace methods. However, some inconsistent results were obtained while comparing two patient groups using the two-sample t-test. Conducting t-test on multiply imputed data generated using MICE and the GTM rejected the null hypothesis—claiming a difference between the patient groups—while data generated using the SOM and the VBPCA approved the null hypothesis. Thus, there obviously are differences in the imputations provided by the methods in question.

One objective of this thesis was to investigate similarities and possible synergies between known methods for survey imputation and collaborative filtering. During the completion of the work, the reasons for the gap between these two research areas became more apparent. First, people use different software. Researchers whose main focus is in the survey rather than the statistical methods themselves, use SAS, S, and R. On the other hand, collaborative filtering is studied mainly by machine learning researchers, who use MATLAB, Python and other software for scientific computing, in their experiments. However, R is also gaining ground among machine learning researchers which might bring about a rapprochement between the two disciplines. While there are similarities in the collaborative filtering and the missing value imputation tasks, there is also at least one distinctive feature. In collaborative filtering, the amount of data and the missingness proportion are usually very high. Thus, approaches such as MICE are out of question and researchers in this area have to concentrate on the scalability of their methods.

The research question framed in Introduction was: what is the best way to conduct the missing value imputation on survey data? Obviously, and as was shown, the most simple approaches, such as ignoring the missing data or the mean imputation, are unable to provide satisfactory results in most cases. Conducting MICE is an all-round approach which usually requires knowledge on regression modeling and may be computationally expensive. However, MICE enables flexible models which can accommodate complex data features. Last but not least, the subspace methods used in this thesis proved out to be appropriate for the missing value imputation task. Especially, in additional to the excellent test results, the variational Bayesian PCA provides a solid probabilistic framework for dealing with missing values and selecting model complexity. I trust that this evidence motivates its usage in survey imputation, as well.

Bibliography

Esa Alhoniemi, Johan Himberg, Juha Parhankangas, and Juha Vesanto. SOM toolbox: `http://www.cis.hut.fi/somtoolbox/`, 2005.

Richard E. Bellman. *Adaptive control processes - A guided tour*. Princeton University Press, 1961.

Christopher M. Bishop. Variational principal components. In *In Proceedings Ninth International Conference on Artificial Neural Networks, ICANN'99*, pages 509–514, 1999.

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2007.

Christopher M. Bishop and Christopher K. I. Williams. Developments of the Generative Topographic Mapping. *Neurocomputing*, 21:203–224, 1998.

Christopher M. Bishop, Markus Svensén, and Christopher K. I. Williams. GTM: The generative topographic mapping. *Neural Computation*, 10, 1998.

S. Van Buuren, H. C. Boshuizen, and D. L. Knook. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, 18:681–694, 1999.

T. F. Coleman and Y. Li. An interior trust region approach for nonlinear minimization subject to bounds. *SIAM Journal on Optimization*, 6(2):418–445, 1996.

Marie Cottrell and Patrick Letrémy. Missing values : processing with the Kohonen algorithm, 2007.

Jan de Leeuw and Achim Zeileis. *Journal of Statistical Software, Volume 45: Multiple Imputation*. American Statistical Association, 2011.

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

Françoise Fessant and Sophie Midenet. Self-organising map for data imputation and correction in surveys. *Neural Computing and Applications*, 10(4):300–310, 2002.

A. Frank and A. Asuncion. UCI machine learning repository, `http://archive.ics.uci.edu/ml`, 2010.

Patrice Gaubert, Smaül Ibbou, and Christian Tutin. Segmented real estate markets and price mechanisms: The case of Paris. *International Journal of Urban and Regional Research*, 20(2):270–298, 1996.

Andrew Gelman and Jennifer Hill. Cambridge University Press, 2007.

Allen Gersho and Robert M. Gray. *Vector quantization and signal compression*. Kluwer Academic Publishers, Norwell, MA, USA, 1991.

Zoubin Ghahramani and Michael I. Jordan. Supervised learning from incomplete data via an EM approach. In *Advances in Neural Information Processing Systems 6*, pages 120–127. Morgan Kaufmann, 1994.

Mark Girolami. The topographic organisation and visualisation of binary data using multivariate-bernoulli latent variable models. *IEEE Transactions on Neural Networks*, 12(6):1367–1374, 2001.

Mark Girolami. Latent variable models for the topographic organisation of discrete and strictly positive data. *Neurocomputing*, 48(1-4):185–198, 2002.

Simon Haykin. *Neural Networks and Learning Machines*. Prentice Hall, 3rd edition, 2008.

Yulei L. He. Missing data analysis using multiple imputation: Getting to the heart of the matter. *Circulation-cardiovascular Quality and Outcomes*, 3(1):98–U145, 2010.

Myles Hollander and Douglas A. Wolfe. *Nonparametric Statistical Methods*. Wiley-Interscience, 2nd edition, 1999.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *J. Educ. Psych.*, 24, 1933.

Alexander Ilin and Tapani Raiko. Matlab package for pca for datasets with missing values: `http://users.ics.tkk.fi/alexilin/software/`, 2008.

Alexander Ilin and Tapani Raiko. Practical approaches to principal component analysis in the presence of missing values. *Journal of Machine Learning Research*, 99: 1957–2000, August 2010.

Tommi S. Jaakkola. Tutorial on variational approximation methods. In *In Advanced Mean Field Methods: Theory and Practice*, pages 129–159. MIT Press, 2000.

I. T. Jolliffe. *Principal Component Analysis*. Springer, 2nd edition, 2002.

Samuel Kaski and Krista Lagus. Comparing self-organizing maps. In Christoph von der Malsburg, Werner von Seelen, Jan Vorbrüggen, and Bernhard Sendhoff, editors, *Artificial Neural Networks (ICANN) 1996*, volume 1112 of *Lecture Notes in Computer Science*, pages 809–814. Springer Berlin / Heidelberg, 1996.

K. Kiviluoto and E. Oja. S-Map: A Network with a Simple Self-Organization Algorithm for Generative Topographic Mappings. In *Advances in Neural Information Processing Systems*, pages 549–555. Morgan Kaufmann Publishers, 1998.

Antti Klapuri, Larri Haaranen, and Ilari T. Nieminen. Questionnaire prototype designed at the virtualcoach project. Unpublished, 2011.

Teuvo Kohonen. *Self-Organizing Maps*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 3rd edition, 2001.

Teuvo Kohonen and Panu Somervuo. How to make large self-organizing maps for nonvectorial data. *Neural Networks*, 15:945–952, October 2002.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

Yehuda Koren. The BellKor solution to the Netflix grand prize, 2009.

L. Kozma, A. Ilin, and T. Raiko. Binary principal component analysis in the Netflix collaborative filtering task. In *Machine Learning for Signal Processing, 2009. MLSP 2009. IEEE International Workshop on*, pages 1 –6, September 2009.

Krista Lagus. Hyvinvoinnin polut ja osa-alueet: neuroverkkoanalyysejä laajoista aineistoista. Talk at the Hyvinvointitutkimuksen workshop, 2011a. Pieksämäki, Finland, March 9-10.

Krista Lagus. VirtualCoach – Mitä ja miksi. Talk at the Hyvinvoinnin polut hankeseminaari, 2011b. Helsinki, Finland, May 13.

Krista Lagus. Wellbeing informatics - VirtualCoach project. Talk at Design For All seminar, 2012. Espoo, Finland, April 18.

Krista Lagus and Juho Saari. Any hope for the socially isolated? data mining study of loneliness questionnaires using the Self-Organising Map. Talk at the Hyvinvointitutkimuksen workshop, 2011. Kuopio, Finland, September 20-21.

Mustapha Lebbah, Nicoleta Rogovschi, and Younès Bennani. BeSOM : Bernoulli on Self-Organizing Map. In *International Joint Conference on Neural Networks 2007*, pages 631 –636, Aug 2007.

Mustapha Lebbah, Younès Bennani, and Nicoleta Rogovschi. A probabilistic self-organizing map for binary data topographic clustering. *International Journal of Computational Intelligence and Applications*, 7(4):363–383, 2008.

Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. Wiley, New York, 2nd edition, 2002.

David A. Marker, David R. Judkins, and Marianne Winglee. Large-scale imputation for complex surveys, Chapter 22. In Robert M. Groves, D. Dillman, J. Eltinge, and R. Little, editors, *Survey nonresponse*. Wiley, New York, 2002.

Hilkka Mehtätalo. Imettäjän hyvinvointi ja onni. Presentation in Sosiologipäivät, 2012. Kuopio, Finland, March 29-30.

Hilkka Mehtätalo and Krista Lagus. Nursing survey questions. Unpublished, 2011. August.

Xiao-Li Meng and Donald B. Rubin. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2):267–278, 1993.

Paul Merlin, Antti Sorjamaa, Bertrand Maillet, and Amaury Lendasse. X-SOM and L-SOM: a double classification approach for missing value imputation. *Neurocomputing*, 73(7-9):1103–1108, 2010.

I. T. Nabney. *NETLAB: Algorithms for Pattern Recognition*. Advances in Pattern Recognition. Springer, 2002.

Radford M. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1996.

Radford M. Neal and Geoffrey E. Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In Michael I. Jordan, editor, *Learning in graphical models*, pages 355–368. MIT Press, Cambridge, MA, USA, 1999.

K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901.

Trivellore E. Raghunathan, James M. Lepkowski, John Hoewyk, and Peter Solenberger. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27(1):85–95, 2001.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley, 1987.

Donald B. Rubin. Multiple Imputation After 18+ Years. *Journal of the American Statistical Association*, 91(434):473–489, 1996.

Rabee Rustum and Adebayo J. Adeloye. Replacing outliers and missing values from activated sludge data using Kohonen Self-Organizing Map. *Journal of Environmental Engineering*, 133(9):909–916, 2007.

Joseph L Schafer. Multiple imputation: a primer. *Stat Methods Med Res*, 8:3–15, February 1999.

Harri Sintonen. The 15D instrument of health-related quality of life: properties and applications. *Ann Med*, 33(5):328–36, 2001.

Antti Sorjamaa. *Methodologies for Time Series Prediction and Missing Value Imputation*. PhD thesis, Aalto University School of Science and Technology, November 2010.

Antti Sorjamaa, Francesco Corona, Yoan Miche, Paul Merlin, Bertrand Maillet, Eric Séverin, and Amaury Lendasse. Sparse linear combination of SOMs for data imputation: Application to financial database. In *WSOM*, pages 290–297, 2009.

Antti Sorjamaa, Amaury Lendasse, Yves Cornet, and Eric Deleersnijder. An improved methodology for filling missing values in spatiotemporal climate data set. *Computational Geosciences*, 14(1):55–64, 2010.

Yu-Sung Su, Andrew Gelman, Jennifer Hill, and Masanao Yajima. Multiple imputation with diagnostics (mi) in R: opening windows into the black box. *Journal of Statistical Software*, 45(2):1–31, 12 2011.

J. E. Tierney, M. T. Mayes, N. Meyer, C. Johnson, P. W. Swarzenski, A. S. Cohen, and J. M. Russell. Late-twentieth-century warming in lake tanganyika unprecedented since AD 500. *Nature Geoscience*, 3(6):422–425, 2010.

Michael E. Tipping and Christopher M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, 61:611–622, 1999a.

Michael E. Tipping and Christopher M. Bishop. Mixtures of probabilistic principal component analyzers. *Neural Comput.*, 11:443–482, February 1999b.

Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67, 12 2011.

Juha Vesanto, Johan Himberg, Esa Alhoniemi, and Juha Parhankangas. Self-organizing map in MATLAB: the SOM toolbox. In *the Matlab DSP Conference*, pages 35–40, 2000.

VirtualCoach. Project web page `http://www.pathsofwellbeing.com/`.

Shouhong Wang. Application of Self-Organising Maps for data mining with incomplete data sets. *Neural Computing and Applications*, 12:42–48, 2003.

Kai Yu, Shenghuo Zhu, John Lafferty, and Yihong Gong. Fast nonparametric matrix factorization for large-scale collaborative filtering. *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval SIGIR 09*, 4(6):211–218, 2009.

Appendix A

Nursing Survey Questions

Kun ajattelet valitsemasi lapsen imetystaivalta ja imetyskokemuksia, tunsitko itsesi (1=Kaiken aikaa tai lähes kaiken aikaa, 2=Suurimman osan ajasta, 3=Puolet ajasta, 4=Osan ajasta, 5=Pienen osan ajasta, 6=En koskaan)

1. onnelliseksi
2. korvaamattomaksi
3. surulliseksi
4. pettyneeksi
5. hyväksytyksi
6. tyytyväiseksi
7. vihaiseksi
8. turhautuneeksi
9. yksinäiseksi
10. väsyneeksi
11. iloiseksi
12. ylpeäksi

Kun ajattelet valitsemasi lapsen imetysaikaa, kuinka tyytyväinen olit silloin? (1=Erittäin tyytyväinen, 2=Melko tyytyväinen, 3=Osin tyytyväinen, osin tyytymätön, 4=Melko tyytymätön, 5=Täysin tyytymätön, 6=Ei koske minua)

13. itseesi
14. parisuhteeseesi
15. perheeseesi
16. sukulaisiisi
17. ystäviisi
18. elintasoosi
19. opintoihisi/ työhösi
20. vapaa-aikaasi
21. terveyteesi

22. sairaalan imetysohjaukseen
23. neuvolan imetysohjaukseen
24. yleiseen yhteiskunnassa vallitsevaan imetysilmapiiriin
25. puolisosi tukeen imetykseen liittyen

Kuinka tyytyväinen olet tällä hetkellä? (1=Erittäin tyytyväinen, 2=Melko tyytyväinen, 3=Osin tyytyväinen, osin tyytymätön, 4=Melko tyytymätön, 5=Täysin tyytymätön, 6=Ei koske minua)

26. itseesi
27. parisuhteeseesi
28. suhteeseesi lapseen, jonka imetyksestä olet kertonut

Vertaistuki on samanlaisia tilanteita läpikäyneiden ihmisten tasavertaista kokemusten vaihtoa. Oletko halunnut ja kuinka paljon olet saanut vertaistukea imetykseen valitsemasi lapsen imetysaikana seuraavilta tahoilta? (1=En ole halunnut, 2=Paljon, 3=Melko paljon, 4=Jonkin verran, 5=Vähän, 6=En ollenkaan)

29. kaverit
30. saman ikäluokan sukulaiset (esim. sisko/serkku/käly)
31. vanhemman ikäluokan sukulaiset (esim. äiti/anoppi/täti)
32. vauvalehtien nettikeskustelut
33. Imetyksen tuki ry:n imetysryhmä
34. muu imetysryhmä
35. Imetyksen tuki ry:n nettikeskustelut
36. muu saamaasi vertaistuki

Appendix B

QUALITY OF LIFE QUESTIONNAIRE (15D©)

Please read through all the alternative responses to each question before placing a cross (x) against the alternative which best describes **your present health status**. Continue through all 15 questions in this manner, giving only **one** answer to each.

QUESTION 1. MOBILITY
1 ( ) I am able to walk normally (without difficulty) indoors, outdoors and on stairs.
2 ( ) I am able to walk without difficulty indoors, but outdoors and/or on stairs I have slight difficulties.
3 ( ) I am able to walk without help indoors (with or without an appliance), but outdoors and/or on stairs only with considerable difficulty or with help from others.
4 ( ) I am able to walk indoors only with help from others.
5 ( ) I am completely bed-ridden and unable to move about.

QUESTION 2. VISION
1 ( ) I see normally, i.e. I can read newspapers and TV text without difficulty (with or without glasses).
2 ( ) I can read papers and/or TV text with slight difficulty (with or without glasses).
3 ( ) I can read papers and/or TV text with considerable difficulty (with or without glasses).
4 ( ) I cannot read papers or TV text either with glasses or without, but I can see enough to walk about without guidance.
5 ( ) I cannot see enough to walk about without a guide, i.e. I am almost or completely blind.

QUESTION 3. HEARING
1 ( ) I can hear normally, i.e. normal speech (with or without a hearing aid).
2 ( ) I hear normal speech with a little difficulty.
3 ( ) I hear normal speech with considerable difficulty; in conversation I need voices to be louder than normal.
4 ( ) I hear even loud voices poorly; I am almost deaf.
5 ( ) I am completely deaf.

QUESTION 4. BREATHING
1 ( ) I am able to breathe normally, i.e. with no shortness of breath or other breathing difficulty.
2 ( ) I have shortness of breath during heavy work or sports, or when walking briskly on flat ground or slightly uphill.
3 ( ) I have shortness of breath when walking on flat ground at the same speed as others my age.
4 ( ) I get shortness of breath even after light activity, e.g. washing or dressing myself.
5 ( ) I have breathing difficulties almost all the time, even when resting.

QUESTION 5. SLEEPING

1 ( ) I am able to sleep normally, i.e. I have no problems with sleeping.
2 ( ) I have slight problems with sleeping, e.g. difficulty in falling asleep, or sometimes waking at night.
3 ( ) I have moderate problems with sleeping, e.g. disturbed sleep, or feeling I have not slept enough.
4 ( ) I have great problems with sleeping, e.g. having to use sleeping pills often or routinely, or usually waking at night and/or too early in the morning.
5 ( ) I suffer severe sleeplessness, e.g. sleep is almost impossible even with full use of sleeping pills, or staying awake most of the night.

QUESTION 6. EATING

1 ( ) I am able to eat normally, i.e. with no help from others.
2 ( ) I am able to eat by myself with minor difficulty (e.g. slowly, clumsily, shakily, or with special appliances).
3 ( ) I need some help from another person in eating.
4 ( ) I am unable to eat by myself at all, so I must be fed by another person.
5 ( ) I am unable to eat at all, so I am fed either by tube or intravenously.

QUESTION 7. SPEECH

1 ( ) I am able to speak normally, i.e. clearly, audibly and fluently.
2 ( ) I have slight speech difficulties, e.g. occasional fumbling for words, mumbling, or changes of pitch.
3 ( ) I can make myself understood, but my speech is e.g. disjointed, faltering, stuttering or stammering.
4 ( ) Most people have great difficulty understanding my speech.
5 ( ) I can only make myself understood by gestures.

QUESTION 8. ELIMINATION

1 ( ) My bladder and bowel work normally and without problems.
2 ( ) I have slight problems with my bladder and/or bowel function, e.g. difficulties with urination, or loose or hard bowels.
3 ( ) I have marked problems with my bladder and/or bowel function, e.g. occasional 'accidents', or severe constipation or diarrhea.
4 ( ) I have serious problems with my bladder and/or bowel function, e.g. routine 'accidents', or need of catheterization or enemas.
5 ( ) I have no control over my bladder and/or bowel function.

QUESTION 9. USUAL ACTIVITIES

1 ( ) I am able to perform my usual activities (e.g. employment, studying, housework, free-time activities) without difficulty.
2 ( ) I am able to perform my usual activities slightly less effectively or with minor difficulty.
3 ( ) I am able to perform my usual activities much less effectively, with considerable difficulty, or not completely.
4 ( ) I can only manage a small proportion of my previously usual activities.
5 ( ) I am unable to manage any of my previously usual activities.

QUESTION 10. MENTAL FUNCTION

1 ( ) I am able to think clearly and logically, and my memory functions well
2 ( ) I have slight difficulties in thinking clearly and logically, or my memory sometimes fails me.
3 ( ) I have marked difficulties in thinking clearly and logically, or my memory is somewhat impaired.
4 ( ) I have great difficulties in thinking clearly and logically, or my memory is seriously impaired.
5 ( ) I am permanently confused and disoriented in place and time.

QUESTION 11. DISCOMFORT AND SYMPTOMS

1 ( ) I have no physical discomfort or symptoms, e.g. pain, ache, nausea, itching etc.
2 ( ) I have mild physical discomfort or symptoms, e.g. pain, ache, nausea, itching etc.
3 ( ) I have marked physical discomfort or symptoms, e.g. pain, ache, nausea, itching etc.
4 ( ) I have severe physical discomfort or symptoms, e.g. pain, ache, nausea, itching etc.
5 ( ) I have unbearable physical discomfort or symptoms, e.g. pain, ache, nausea, itching etc.

QUESTION 12. DEPRESSION

1 ( ) I do not feel at all sad, melancholic or depressed.
2 ( ) I feel slightly sad, melancholic or depressed.
3 ( ) I feel moderately sad, melancholic or depressed.
4 ( ) I feel very sad, melancholic or depressed.
5 ( ) I feel extremely sad, melancholic or depressed.

QUESTION 13. DISTRESS

1 ( ) I do not feel at all anxious, stressed or nervous.
2 ( ) I feel slightly anxious, stressed or nervous.
3 ( ) I feel moderately anxious, stressed or nervous.
4 ( ) I feel very anxious, stressed or nervous.
5 ( ) I feel extremely anxious, stressed or nervous.

QUESTION 14. VITALITY

1 ( ) I feel healthy and energetic.
2 ( ) I feel slightly weary, tired or feeble.
3 ( ) I feel moderately weary, tired or feeble.
4 ( ) I feel very weary, tired or feeble, almost exhausted.
5 ( ) I feel extremely weary, tired or feeble, totally exhausted.

QUESTION 15. SEXUAL ACTIVITY

1 ( ) My state of health has no adverse effect on my sexual activity.
2 ( ) My state of health has a slight effect on my sexual activity.
3 ( ) My state of health has a considerable effect on my sexual activity.
4 ( ) My state of health makes sexual activity almost impossible.
5 ( ) My state of health makes sexual activity impossible.