TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

D I P L O M A R B E I T

Datenanalyse unvollständiger Beobachtungen, Multiple Imputation angewandt auf Verkehrsdaten

Ausgeführt am Institut für
Wirtschaftsmathematik
der Technischen Universität Wien

unter der Anleitung von
Ao.Univ.Prof. Wolfgang Scherrer

durch
Maximilian Leodolter

22. Juli 2013

Inhaltsverzeichnis

Abstract

Im Rahmen dieser Diplomarbeit wird das Problem unvollständiger Beobachtungen in der Datenanalyse besprochen. Hierfür wird zuallererst das Ereignis "Fehlende Daten" und dessen Eigenschaften näher erörtert, worauf aufbauend verschiedene Lösungsansätze vorgestellt werden. Abgesehen von simplen Techniken wie dem Streichen unvollkommener Beobachtungen, liegt das Hauptaugenmerk auf dem Expectation Maximization Algorithmus und Multipler Imputation. Im Detail wird Multiple Imputation by Chained Equations auf Basis der Methode Predicitve Mean Matching präsentiert. In Folge an die Diskussion der verschiedenen Prozeduren werden die Verfahren an simulierten Datensätzen getestet. Es wird evaluiert, aus welcher Technik die besten Schätzer für den Mittelwert und die Kovarianzmatrix hervorgeht. Die Ergebnisse zeigen, dass unter bestimmten Voraussetzungen auch einfache Methoden zum Ziel führen.

Abschließend steht ein lückenhafter Datensatz mit Informationen zu einer Flotte von Taxis bereit, welcher mithilfe vorgestellter Verfahren analysiert wird. Einerseits mit der simplen Herangehensweise unvollständige Beobachtungen nicht zu berücksichtigen, andererseits mit Multipler Imputation. Hierfür wird insbesondere das R-Paket *mice* bemüht. Anhand dieses praktischen Beispiels soll die Anwendung von Multipler Imputation sowie dessen Vorteile gegenüber einfacher Methoden demonstriert werden. Zu diesem Zweck werden typische Verfahren der Datenanalyse auf den Datensatz angewandt, nachdem das Problem "Fehlende Daten" behandelt wurde. Die Resultate zeigen, dass es für dieses Fallbeispiel vorzuziehen ist Multiple Imputation anzuwenden, anstatt unvollständige Beobachtungen zu löschen.

Nomenklatur

Δ_θ	Abweichung des geschätzten vom tatsächlichen Parameter, gemittelt über alle Simulationen ... 35
$\Delta_\theta^{XY(N,\lambda,\rho)}$	Abweichung des geschätzten vom tatsächlichen Parameter, gemittelt über alle Simulationen mit den Matrixparametern (N, λ, ρ), berechnet mittels der Missing-Data-Methode XY ... 35
$\Delta_{i,\theta}$	Abweichung des geschätzten vom tatsächlichen Parameter, für die i-te Simulation ... 35
ι	Einsvektor $\iota = (1 \dots 1)^T \in \mathbb{R}^n$ ... 17
Λ	das Verhältnis der Zwischen-Imputation-Varianz zur totalen Varianz eines Schätzers aus Multipler Imputation ... 30
λ	Anteil der fehlenden Daten von X ... 32
μ_j^t	Schätzer für den Mittelwert der t-ten Iteration des EM Algorithmus ... 16
Ω_θ	Parameterraum ... 3
ρ	Designparameter für die Simulation von X ... 32
$\overline{\Sigma}$	Matrix der Stichprobenkovarianzen mit den Einträgen s_{ij} ... 17
σ_{jk}^t	Schätzer für den Mittelwert der t-ten Iteration des EM Algorithmus ... 16
c_{jki}^t	Schätzer für die spezifische Kovarianz der t-ten Iteration des EM Algorithmus 15
G	Erweiterte Kovarianzmatrix ... 17
g_{ij}	Eintrag in der Zeile i und Spalte j von G ... 17
$I^{[i]}$	die Indexmenge der beobachteten Variablen für die i-te Zeile ... 9
$J^{[ij]}$	die Indexmenge der Zeilen, welche zur Regression der j-ten Variable auf jene aus $I^{[i]}$ herangezogen wurden beobachteten Variablen für die i-te Zeile ... 9
K	Anzahl der Spalten der Datenmatrix X ... 1
M	Anzahl der Imputationen für Multiple Imputation ... 20
N	Anzahl der Zeilen der Datenmatrix X ... 1
$N^{(cc)}$	Anzahl der vollständigen Zeilen von X, $N^{(cc)} \leq N$ ... 5
$N^{(jk)}$	N minus der Anzahl an Zeilen in welchen die Werte der Spalten j und/oder k fehlen ... 8
O	Matrix $\in \mathbb{R}^{N \times K}$, Einträge sind 1 und 0 ... 2
o_{ij}	Eintrag von O, 1 wenn $x_{ij} \in X_{mis}$, 0 sonst ... 2
$q(X_j, \lambda)$	Das λ-Quantil des Vektors X_j ... 33

$\tilde{s}_{jk}^{(jk)}$	Korrigierter Kovarianzschätzer der Spalten, bzw. Variablen j und k, nach Streichen der Zeilen mit fehlenden Werten in den Spalten j und/oder k ... 8
$s(X)$, s^t	eine von X abhängige Statistik, bzw. deren t-ter Iterationswert ... 13
s_{ij}	Stichprobenkovarianz der Spalten i und j ... 17
$s_{jk}^{(jk)}$	Stichprobenkovarianz der Spalten, bzw. Variablen j und k, nach Streichen der Zeilen mit fehlenden Werten in den Spalten j und/oder k ... 8
$SWP(.,.)$	Sweep-Operator ... 17
T	Totale Varianz eines Schätzers aus Multipler Imputation, bestehend aus Schwankungen innerhalb der Imputationen $\overline{V}$ und zwischen Imputationen $\overline{B}$ ... 27
u	Regressionsfehler ... 17
$\mathring{X}$	Entscheidungsmatrix nach welcher ein Wert in der Zufallsmatrix X gelöscht wird oder nicht ... 33
$\mathring{x}_i j$	Eintrag der Entscheidungsmatrix $\mathring{X}$ ... 33
$\mathring{X}_j$	j-te Spalte der Entscheidungsmatrix $\mathring{X}$ ... 33
$\overline{x}$	Vektor der Stichprobenmittelwerte mit den Einträgen $\overline{x}_i$ ... 17
$\overline{x}_i$	$\in \mathbb{R}$ Stichprobenmittel der Spalten i der Matrix X ... 17
$\overline{x}_j^{(j)}$	Stichprobenmittelwert der Spalte, bzw. Variable j, nach Streichen der Zeilen mit fehlenden Werten in der Spalte j ... 8
$\overline{x}_{l,j}^{(j)}$	Mittelwert aller $x_{l,ij}$ ... 8
X	Datenmatrix $\in \mathbb{R}^{N \times K}$ ... 1
X_j	$\in \mathbb{R}^{N \times 1}$ die j-te Spalte bzw. Variable von X ... 1
$X_{-j_{obs}}$, $X_{-j_{mis}}$	Quadranten der Matrix X abhängig von $X_{j_{obs}}$ und $X_{j_{mis}}$ ... 25
X_{cc}	Reduktion der Datenmatrix auf die vollständigen Beobachtungen $\in \mathbb{R}^{N^{(cc)} \times K}$ 5
x_{ij}^t	Schätzer für x_{ij} der t-ten Iteration des DA Algorithmus ... 21
x_{ij}^t	Schätzer für x_{ij} der t-ten Iteration des EM Algorithmus ... 15
$X_{j_{obs}}$, $X_{j_{mis}}$	beobachteter und nicht beobachteter Teil von X_j ... 25
$x_{l,ij}$	der i-te beobachtete Wert, der Variable l, in der j-ten Klasse ... 8
X_{mis}	nicht beobachtete Werte der Datenmatrix $X = (X_{obs}, X_{mis})$ ... 1
X_{obs}	beobachtete Werte der Datenmatrix $X = (X_{obs}, X_{mis})$ ... 1

1. Motivation und Einleitung

Die Problematik "Fehlende Daten" existiert schon seit Beginn der Datenerhebung. Eine von vielen Ursachen kann das Nicht-Beantworten einer unangenehmen Frage in einem Fragebogen sein, oder aber das bewusste Nicht-Beobachten bestimmter Variablen für gewisse Probanden, weil die Datenerhebung zu kostenintensiv wäre. Daher wird diese Variable nicht so oft erhoben wie die übrigen im selben Datensatz. Das grundlegende Problem mit fehlenden Daten ist der Verlust an Informationen. Nachdem die Aussagekraft jeder Datenanalyse erheblich von der Stichprobengröße N abhängt, würde ein simples Streichen der unvollständigen Beobachtungen zuallererst N verringern. Darüber hinaus ist es aber nicht nur entscheidend wie viele Daten fehlen, sondern auch nach welchem Prinzip. Folgt das Ereignis "Fehlende Daten" einem bestimmten Modus, so kann das zu gravierenden Verzerrungen von Statistiken führen. Um diese Problemstellung zu bewältigen gibt es zahlreiche Methoden, wobei nicht jede zum gewünschten Ergebnis führt. Würde man zum Beispiel die nicht beobachteten Werte einer numerischen Variable durch den Mittelwert der beobachteten Werte ersetzten, so käme es zu zwei direkten Auswirkungen. Erstens würde man am Mittelwert nichts ändern, er bliebe der selbe. Zweitens würde die Stichprobenvarianz kleiner werden, was sämtliche davon abhängige Statistiken verfälscht. Ein Streichen der unvollständigen Beobachtungen könnte hier zuverlässigere Resultate liefern.
In den Kapiteln 1 und 2 wird auf die allgemeine Problemstellung näher eingegangen und einfache Missing-Data-Strategien werden vorgestellt. Weiterentwickelte Fertigkeiten werden in den Kapitel 3 und 4 präsentiert, insbesondere der Expectation Maximization Algorithmus und Multiple Imputation mithilfe von Chained Equations, wobei die Imputationen durch das Verfahren Predictive Mean Matching erzeugt werden. In Kapitel 5 werden Ergebnisse aus empirischen Untersuchungen dargelegt.

1.1. Einführung

X sei die Matrix, deren Inhalt analysiert werden soll. Sie beinhält N viele Beobachtungen (Zeilen) von K vielen Variablen (Spalten), also $N \times K$ Werte. Angenommen die vorliegende Datenmatrix $X = (X_1 \dots X_K) \in \mathbb{R}^{N \times K}$ ist nicht vollständig beobachtet, das heißt es fehlen in einer oder mehreren Spalten Datenpunkte. Man spricht auch vom Ereignis "Fehlende Daten". X wird dann in den beobachteten (X_{obs}) und den nicht beobachteten Part (X_{mis}) unterteilt: $X = (X_{obs}, X_{mis})$. Wenn von der Imputation fehlender Werte gesprochen wird,

sind also ausschließlich Elemente aus X_{mis} gemeint. Bevor mit der Diskussion der Missing-Data-Methoden begonnen werden kann, sollen folgende 2 Fragen beantwortet werden:

1) Nach welchem Mechanismus fehlen die Daten?
   Es wird zwischen drei Mechanismen, nach welchen Daten fehlen, unterschieden. Wie erfolgreich eine Methode das Missing-Data Problem löst, kann vom zugrundeliegenden Missing-Data-Mechanismus abhängen.

MCAR: (missing completley at random) Das Fehlen der Daten ist unabhängig von den nicht beobachteten Werten und den übrigen Variablen im Datensatz, kann aber von nicht im Datensatz inbegriffenen Variablen abhängen.
Bsp: Ein Datensatz umfasst die Variablen Einkommen, Geschlecht, Wohnbezirk und Alter. Während das Einkommen nicht immer angegeben wurde, sind die restlichen Daten vollständig aufgezeichnet. Unterstellt man MCAR, so nimmt man an, dass das Ereignis [Datenwert fehlt] unabhängig von dessen Wert und jenen der anderen Variablen ist.

MAR: (missing at random) Das Fehlen der Daten ist unabhängig von den nicht beobachteten Werten, kann aber von einer der übrigen Variablen im Datensatz abhängen.
Bsp.: Männer geben seltener ihr Einkommen bekannt als Frauen, allerdings unabhängig von der Höhe des Einkommens.

NMAR: (not missing at random) Das Fehlen der Daten hängt von den nicht beobachteten Werten ab.
Bsp.: Die Angaben über das Einkommen nehmen bei sehr niedrigen und/oder hohen Werten ab. Sehr reiche/arme Menschen weigern sich zu antworten.

2) Nach welchem Muster fehlen die Daten?
   Das Muster zu kennen kann die weiteren Schritte erheblich beeinflussen, da die Eignung von Algorithmen für Missing-Data Probleme von der Struktur nach welcher die Daten fehlen möglicherweise abhängen. In Abbildung 1.1 sind die wichtigsten Muster dargestellt. Fehlende Datenpunkten werden mit einem * gekennzeichnet.

Für Untersuchungen in den Abschnitten 3 und 4 werden bezüglich des Mechanismus, nach welchem die Daten fehlen, Annahmen getroffen. Sei O eine Matrix der selben Dimension wie die Datenmatrix und befüllt nach folgender Definition:

$$o_{ij} = \begin{cases} 0 & \textit{wenn } x_{ij} \textit{ beobachtet (observed) ist,} \\ 1 & \textit{wenn } x_{ij} \textit{ nicht beobachtet ist} \end{cases} \tag{1.1}$$

O wird gegebenenfalls als Zufallsvariable betrachtet, deren Verteilung vom Parameter ψ abhängt. Die Verteilung der Datenmatrix X hängt vom Parameter θ ab. Der Mechanismus, nach welchem die Daten fehlen, kann für

- Liklihood-Untersuchungen ignoriert werden, falls

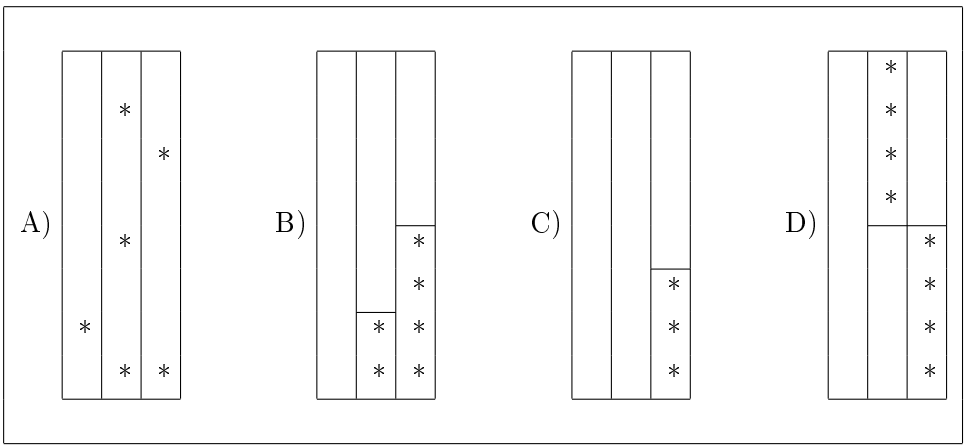

Abbildung 1.1.: A) allgemein; B) monoton; C) univariat; D) invers

a) MAR gilt und

b) die Parameterräume Ω_θ ($\ni \theta$) und Ω_ψ ($\ni \psi$) disjunkt sind.

- Bayes-Untersuchungen ignoriert werden, falls

  a) MAR gilt und

  b) die Parameter θ und ψ a priori unabhängig sind.

1.2. Klassifizierung der Missing-Data Methoden

Zur Datenanalyse einer unvollständig beobachteten Stichprobe existieren verschiedene Ansätze. Bezüglich der zugrundeliegenden Konzepte lassen sich diese wie folgt zuordnen.

Complete-Case-Methoden

Die Datenanalyse wird auf die vollständigen Beobachtungen beschränkt. Hierfür werden Beobachtungen mit fehlenden Aufzeichnungen gelöscht, sodass der neue Datensatz den ursprünglichen ersetzt. Diese Methode ist bei einem geringen Anteil (ungefähr 5 Prozent) fehlender Daten noch tolerierbar. Es kann aber auch hier, abhängig vom Missing-Data-Mechanismus, insbesondere bei einem größeren Prozentsatz zu starken Verzerrungen der Schätzer führen. Hingegen ist bei MCAR und genügend großem N mit erwartungstreuen Schätzern zu rechnen, da der bereinigte Datensatz bloß als Stichprobe des ursprünglichen zu sehen ist. Näher erläutert werden zwei dieser Methoden in den Kapitel 2.1 und 2.2.

Im Gegensatz zu den Complete-Case-Methoden soll mithilfe der nachfolgenden Alternativen der Informationsgewinn aus den unvollständigen Beobachtungen maximiert werden.

Imputations-Methoden

Anstatt unvollständige Beobachtungen zu streichen, werden die fehlenden Werte imputiert. Das heißt, es werden geschätzte Werte eingesetzt. Mögliche Imputationen sind beispielsweise die Mittelwerte (geschätzt aus den beobachteten Teilen der Variablen), was der Regression auf eine Konstante gleich kommt. Alternativ kann ebenso auf die übrigen Variablen regressiert werden. Hat man zusätzlich zu dem Datensatz X weitere Variablen zu Verfügung, welche für die Datenanalyse ad hoc nicht von Interesse sind, so kann es hilfreich sein das Regressionsmodell für das Imputieren um diese Variablen zu erweitern. In Kapitel 2.3 werden diese Ansätze eingehend diskutiert.
Konträr zu diesen Methoden stehen Hot-Deck Verfahren. Imputationen sind nun keine berechneten Schätzungen, sondern beobachtete Werte, welche anstatt der nicht beobachteten eingesetzt werden. Das Auswahlverfahren, welches die Imputationen bestimmt, beschreibt den Charakter des Hot-Deck Verfahrens. Dieses zu wählen hängt nicht zuletzt vom Datensatz ab. Sei X die Auswertung einer Volksbefragung, und seien die Angaben zum Einkommen nicht durchgängig gegeben. So kann beispielsweise das Einkommen des Nachbarn als Imputation dienen, da die Wohnsituation der beiden Probanden ähnlich ist. Setzt man den Gedanken dieser Herangehensweise fort, so können die Antworten zu den übrigen Fragen mit anderen Probanden verglichen werden, und das Einkommen jenes imputiert werden, dessen Angaben am besten mit der unvollständigen Beobachtung übereinstimmen. Es wird also ein Abstandsbegriff für die Distanz zwischen zwei Probanden definiert. Im Gegensatz zu Hot-Deck steht Cold-Deck. Dabei fungiert ein in der Vergangenheit beobachteter Wert als Imputation für den gegenwärtig fehlenden. Sei in dem Beispiel von oben das Einkommen in der Erhebung des Jahres 2013 nicht angegeben worden, im Jahr 2012 jedoch schon, so imputiert man die damalige Antwort. In Kapitel 4.2.1 wird die Methodik Predictive Mean Matching (PMM) vorgestellt, welche grundlegende Eigenschaften eines Hot-Deck Verfahrens annimmt.

Modell-basierende Methoden

Ausgehend von den beobachteten Daten trifft man Annahmen bezüglich eines Modells, welches die Datenstruktur möglichst treffend beschreibt. Aus diesem Modell lassen sich Imputationen, als Realisationen von Zufallsvariablen, ziehen und Modellparameter schätzen. Der Expectation Maximization (EM) Algorithmus und Multiple Imputation werden in den Kapitel 3 und 4 als Vertreter dieser Klasse vorgestellt.

Weitere Methoden, die sich der Problemstellung fehlender Daten widmen, werden zum Beispiel von Little und Rubin [8], und Allison [1] vorgestellt.

2. Simple Methoden

2.1. Einfaches Streichen

Das Ereignis fehlender Daten wird mithilfe dieser Methoden vollkommen außer Acht gelassen. Nach dem Löschen der jeweiligen Beobachtungen verfährt man als hätte man einen neuen Datensatz ohne fehlende Daten. Sei $X = (X_{obs}, X_{mis})$ der gesamte Datensatz, bestehend aus den beobachteten und fehlenden Werten. Streicht man nun die Zeilen in welchen ein Wert nicht vorhanden ist, so bleibt das Subset X_{cc} (complete cases) über. Unter der Voraussetzung MCAR ist X_{cc} eine zufällig gezogene Stichprobe aus X, womit die Ergebnisse der Analyse von X_{cc} auf ganz X zutreffen, sofern die Stichprobengröße von X_{cc} die Aussagekraft gewährleistet. Im Gegensatz zu den Mengen X_{obs} und X_{mis} ist X_{cc} eine Matrix $\in \mathbb{R}^{N^{(cc)} \times K}$, wobei $N^{(cc)} \leq N$ die Anzahl der vollständigen Zeilen ist. Sollten die Daten nicht MCAR sondern nur MAR sein, so kann das Streichen der Zeilen zu erheblichen Verzerrungen kommen. Ein Beispiel: Hängt die Wahrscheinlichkeit, dass Werte der Variable Ausbildung fehlen, von der Variable Arbeitsverhältnis ab, so würde eine Regression von Arbeitsverhältnis auf Ausbildung ein verzerrtes Ergebnis liefern.
Allison schreibt in seinem Buch Missing Data [1], dass Einfaches Streichen (ES) unverzerrte Regressionsschätzer (sofern die üblichen Bedingungen für lineare Regression erfüllt sind) liefern kann, falls die Wahrscheinlichkeit für das Fehlen von Daten einer erklärenden Variable (also eine unabhängige Variable im Regressionsmodell) von einer anderen erklärenden Variable abhängt, aber nicht von der zu erklärenden Variable. Insbesondere kann ES bessere Resultate erzielen als Maximum Likelihood Methoden oder Multiple Imputation, falls die Fehl-Wahrscheinlichkeit einer unabhängigen Variable nur von den eigenen (nicht beobachteten) Werten abhängt. Dazu kann es kommen, wenn im Zuge einer Volksbefragung Menschen mit niedrigem Einkommen dieses seltener angeben, weil es ihnen unangenehm ist diese Auskunft zu geben. Der Wert existiert also, ist aber nicht beobachtet. Das Fehlen der Daten ist somit von den nicht beobachteten Werten abhängig.

2.2. Doppeltes Streichen

Hierbei werden nicht wie bei ES alle Zeilen mit unvollständigen Beobachtungen gelöscht. Sondern es wird für die Berechnung von statistischen Größen von einzelnen Variablen differenziert. Nur jene Zeilen mit unvollständigen Daten in den Spalten, welche man für die

Berechnung heranzieht, werden gestrichen.
Soll beispielsweise ein lineares OLS-Regressionsmodell mit einer unvollständigen Datenmatrix geschätzt werden, so kann die Berechnung auf Formeln reduziert werden, welche ausschließlich die Mittelwerte und Kovarianzmatrix benötigen. Diese Eigenschaft macht sich Doppeltes Streichen (DS) zu nutze. Zur Mittelwertberechnung jeder einzelnen Variable X_j werden jeweils nur jene Zeilen gestrichen, welche fehlende Werte in der Variable X_j aufweisen. Zur Ermittlung der Kovarianz zweier Variablen X_j und X_k werden nur jene Zeilen gelöscht, in welchen Werte in diesen beiden Variablen fehlen. Somit wird zumindest mehr Information verwertet als bei ES. Unter MCAR liefert DS konsistente Schätzer, aber unter MAR können diese stark verzerrt sein. Es ist ein Trugschluss, dass DS unter MCAR ES vorzuziehen ist. Studien von Glasser [4], Haitovsky [5] und Kim und Curry [6] belegen, dass bei generell hoher Korrelation zwischen den Variablen ES effizientere (bedeutet, dass die Schätzer geringere Standardfehler haben) Schätzer liefert. Bei niedriger Korrelation ist DS vorzuziehen.
Ein sich durch diese Methode ergebendes Problem ist, dass es keine einheitliche Stichprobengröße N gibt. Zur Berechnung von Standardfehler wird diese jedoch benötigt. Bei der Implementierung von DS wird daher mit verschiedenen Definitionen von N vorgegangen.

- $N :=$ Anzahl der beobachteten Werte jener Variable mit dem größten Anteil an fehlenden Werten
- $N :=$ Anzahl der beobachteten Werte jener Variable mit dem kleinsten Anteil an fehlenden Werten

Keine der zwei simplen Ansätze liefert konsistente Schätzer der Standardfehler. Laut Allison [1] ist es jedoch möglich diese zu erhalten, allerdings wurden die dafür nötigen komplexen Formeln noch nicht implementiert.
Das zweite auftretende Hindernis von DS, speziell bei kleinen Stichproben, ist die nicht unbedingte positive Definitheit der Kovarianzmatrix. Dadurch ist ein Regressionsmodell, welches auf den Daten basiert, nicht mehr lösbar. Aus diesen zwei Gründen wird DS nicht als allgemeiner Favorit gegenüber ES genannt.

2.3. Einfache Imputation

Vorangegangene Methoden verwenden lediglich Informationen aus vorhandenen Daten. In diesem Kapitel wird erstmals darauf eingegangen, diese auch aus den unvollständigen Zeilen zu ziehen. Erzielt wird dies mithilfe von Imputationen für die unvollständigen Variablen, welche von den selbigen und anderen Variablen abhängig sein können. Imputationen sind aus einer prädiktiven Verteilung zu ziehen, welche entweder explizit oder implizit modelliert wird.

ex) Die Verteilung basiert auf expliziten Annahmen über ein statistisches Modell.

ex.i) *Mean imputation:* Für die fehlenden Werte werden Konstanten (Bsp.: Mittelwert, Median, Modus) eingesetzt(eventuell abhängig von Klasseneinteilungen).

ex.ii) *Regression imputation:* Durch das Regressieren der unvollständigen Variable auf andere vollständige und/oder unvollständige Variablen werden die fehlenden Werte ersetzt.

ex.iii) *Stochastic regression imputation:* zusätzlich zur vorigen Variante, werden den imputierten Werten Zufallskomponenten hinzugefügt. In normalen linearen Regressionsmodellen werden diese Zufallsvariablen aus einer Normalverteilung $N(0, \sigma)$, wobei σ die Fehlervarianz aus der linearen Regression ist, gezogen.

im) Im Vordergrund steht ein Algorithmus welcher ein Modell impliziert.

im.i) *Hot deck imputation:* Diese Methode wird sehr häufig bei der Auswertung von Daten aus Fragebögen mit fehlenden Antworten angewandt. Dabei ersetzt man die fehlenden Werte mit jenen von befragten Individuen, welche bei anderen Fragen ähnlich oder gleich geantwortet haben. Es werden also mehrere Personen in Gruppen zusammengefasst.

im.ii) *Substitution:* Sind für eine Befragung die Antworten eines Subjektes nicht verfügbar, so ersetzt man sie durch die eines anderen, welches planmäßig nicht an der Umfrage teilnehmen hätte sollen. Bsp.: Im gesuchten Haushalt ist niemand anzutreffen, also substituiert man dessen Antworten zu einer Befragung durch jene der Nachbarn.

im.iii) *Cold deck imputation:* Fehlende Werte werden aus vergangenen Beobachtungen imputiert. Bsp.: Hat ein Haushalt zu einem vergangen Zeitpunkt auf eine Frage geantwortet, welche er im aktuellen Fragebogen nicht ausgefüllt hat, so ersetzt man den fehlenden Wert durch den damalig angegebenen.

Im folgenden Abschnitt werden einige Varianten exemplarisch vorgestellt um auf deren Nachteile hinzuweisen, welche mittels der Methoden aus den Kapitel 3 und 4 behoben werden können.

Unbedingte Mean Imputation von einer prädiktiven Verteilung

Wendet man die Methode aus ex.i) auf eine Matrix X an, in welcher x_{ij} der Eintrag in der i-ten Zeile und j-ten Spalte ist, sodass die j-te Spalte unvollständig ist, so ist ein einfacher Schätzer für die fehlenden Werte der Mittelwert der vorhandenen Werte. Der Mittelwert der imputierten Variable ist damit gleich jenem aus den beobachteten Werten $\overline{x}_j^{(j)}$. Dabei bedeutet das hochgestellte (j), dass zur Berechnung der Statistik (in diesem Fall der Mittelwert) alle Zeilen von X gestrichen werden, für welche in der j-ten Spalte Werte fehlen. Wenn die

Stichprobenvarianz der beobachteten Werte $s_{jj}^{(j)}$, und $N^{(j)}$ die Anzahl der vorhandenen Werte in der Spalte j ist, dann gilt für den Varianzschätzer $\tilde{s}_{jj}^{(j)}$ der imputierten Variable:

$$\tilde{s}_{jj}^{(j)} = s_{jj}^{(j)} \frac{(N^{(j)} - 1)}{(N - 1)}$$

Unter der Annahme MCAR ist $s_{jj}^{(j)}$ ein konsistenter Schätzer. Durch das Imputieren der Stichprobenmittel wird die Stichprobenvarianz der imputierten Variable um den Term $(N^{(j)} - 1)/(N - 1)$ unterschätzt. Die Stichprobenkovarianz wird analog berechnet

$$\tilde{s}_{jk}^{(jk)} = s_{jk}^{(jk)} \frac{(N^{(jk)} - 1)}{(N - 1)}$$

Wieder gilt, dass $s_{jk}^{(jk)}$, unter der Annahme MCAR, ein konsistenter Schätzer ist. $N^{(jk)}$ ist die Anzahl der Fälle in welchen beide Variablen beobachtet sind, und $s_{jk}^{(jk)}$ ist die Stichprobenkovarianz berechnet aus diesen $N^{(jk)}$ vielen Paaren, wobei nicht $\overline{x}_j^{(jk)}$ und $\overline{x}_k^{(jk)}$, sondern $\overline{x}_j^{(j)}$ und $\overline{x}_j^{(k)}$ für die Berechnung von $s_{jk}^{(jk)}$ herangezogen werden.
Jedoch ist die Kovarianzmatrix im Allgemeinen nicht positiv definit, speziell wenn die Variablen stark korrelieren.

Bedingte Mean Imputation von einer prädiktiven Verteilung innerhalb von Zellen

Speziell beim Auswerten von Daten aus Fragebögen werden oftmals Klassen/Zellen gebildet in welchen die Mitglieder ähnliche/gleiche Antworten gegeben haben. Für eine unvollständig beobachtete Variable werden die fehlenden Werte ersetzt, durch die Mittel jener Beobachtungen, welche in den gleichen Zellen sind. Bei J vielen Klassen und N_j Mitglieder in der Klasse j, wovon r_j geantwortet haben, ergibt sich ein Mittelwertschätzer der unvollständigen Variable X_l durch:

$$\overline{x}_l^w = \frac{1}{N} \sum_{j=1}^{J} N_j \cdot \overline{x}_{l,j}^{(j)} = \frac{1}{N} \sum_{j=1}^{J} \left(r_j \cdot \overline{x}_{l,j}^{(j)} + \sum_{i=r_j+1}^{N_j} \overline{x}_{l,j}^{(j)} \right) = \frac{1}{N} \sum_{j=1}^{J} \left(\underbrace{\sum_{i=1}^{r_j} x_{l,ij}}_{beob.Werte} + \underbrace{\sum_{i=r_j+1}^{N_j} \overline{x}_{l,j}^{(j)}}_{fehl.Werte} \right)$$

was einem gewichteten Mittel der Zellenmittel gleich kommt. $x_{l,ij}$ ist einer der r_j-vielen beobachteten Werte der Variable l in der j-ten Klasse. $\overline{x}_{l,j}^{(j)}$ ist das arithmetische Mittel all dieser r_j-vielen. Hier steht das hochgestellte (j) für die vorhandenen Daten der j-ten Klasse.

Bedingte Regression Imputation von einer prädiktiven Verteilung

Nehme man den einfachen Fall eines univariaten Musters an, sodass X_j vollständig beobachtet ist für $j = 1 \dots K-1$ und unvollständig für $j = K$. Weiters sei x_{iK} für $i = 1 \dots r$ präsent. Es wird nun X_K auf die übrigen Spalten X_j für $j = 1 \dots K-1$ regressiert. Womit sich Schätzer für die fehlenden Werte x_{iK} darstellen lassen als:

$$\widehat{x}_{iK} = \beta_0 + \sum_{j=1}^{K-1} \beta_j x_{ij} \ , \qquad i = r+1, \dots, N \tag{2.1}$$

β_0 und β_j sind die OLS Schätzer des Regressions-Modells $X_K^{(K)} = \beta_0 + \beta_1 X_1^{(K)} + \dots + \beta_{K-1} X_{K-1}^{(K)} + u$, wobei $X_j^{(K)}$ der Vektor X_j, reduziert um die in der K-ten Spalte nicht beobachteten Werte (die Zeilen $r+1 \dots N$), ist. u ist der Fehlerterm des Regressions-Modells. Eine Weiterentwicklung davon für allgemeine Muster ist die Regression für jeden fehlenden Datenpunkt zu wiederholen. Um den fehlenden Datenpunkt x_{ij} zu imputieren geht man folgendermaßen vor:

1) Das Bestimmen der Variablen (Spalten), welche für die i-te Stichprobe beobachtet sind.

$$I^{[i]} := \{l \in \{1 \dots K\} | o_{il} = 0\} \tag{2.2}$$

2) Das Bestimmen der Beobachtungen (Zeilen), für welche sowohl X_j, als auch $X_l \ \forall l \in I^{[i]}$ beobachtet sind.

$$J^{[ij]} := \{m \in \{1 \dots N\} | o_{ml} = 0 \ \ \forall l \in I^{[i]} \cup \{j\} \ \} \tag{2.3}$$

3) Regressieren der j-ten Variable auf all jene, deren Indices in $I^{[i]}$ vorkommen, unter Verwendung der Zeilen, deren Indices in $J^{[ij]}$ vorkommen.

Anhand eines Beispiels sollen diese Definitionen verdeutlicht werden. Sei $X \in \mathbb{R}^{8\times 4}$:

X=

		$*_{13}$	
	$*_{22}$		$*_{24}$
			$*_{34}$
			$*_{44}$
	$*_{52}$		

	$I^{[i]}$	$J^{[ij]}$
$*_{13}$	{1,2,4}	{6,7,8}
$*_{22}$	{1,3}	{3,4,6,7,8}
$*_{24}$	{1,3}	{5,6,7,8}
$*_{34} = *_{44}$	{1,2,3}	{6,7,8}
$*_{52}$	{1,3,4}	{6,7,8}

Aus diesen Überlegungen lassen sich Regressionsmodelle ableiten, mithilfe derer die fehlenden Datenpunkte geschätzt werden. Mit entsprechender Vorbehandlung, respektive Umordnung

der Daten kann man sich Regressionsschritte ersparen, falls die Mengen $I^{[i]}$ und $J^{[ij]}$ für mehrere Punkte übereinstimmen, sowie im Beispiel $*_{44} = *_{34}$. Dadurch kann der Rechenaufwand reduziert werden.
Hier wird davon ausgegangen, dass jede Spalte, unabhängig von dem Prozentsatz an vorhandenen Daten in dieser, als unabhängige Variable zur Verfügung steht. Einzig ob in der jeweiligen Zeile die Werte der anderen Variablen beobachtet sind wird berücksichtigt.

Stochastic Regression Imputation aus einer prädiktiven Verteilung

Motiviert weil obige Methoden die Standardfehler systematisch unterschätzen, werden den Imputationen Zufallskomponenten hinzugefügt. Es sei an die Gleichung (2.1) erinnert. Nun wird eine Zufallsvariabel z_{ij} addiert.

$$\widehat{x}_{ij} = \beta_0^{(ij)} + \sum_{l \in J^{(ij)}} \beta_l^{(ij)} x_{il}^{(ij)} + z_{ij} \qquad z_{ij} \sim \mathcal{N}(0, \tilde{\sigma}^{(ij)}) \tag{2.4}$$

Wobei $\tilde{\sigma}^{(ij)}$ die Fehlervarianz der Regression von X_j auf $X^{[ij]}$ ist.

Zusammengefasst nennen Little und Rubin vier aufeinander aufbauende Methoden für einen bivariaten Fall, wenn X_2 unter dem Mechanismus MCAR unvollständig beobachtet ist:

1. Umean (Unbedingte Mean Imputation): fehlende Werte werden durch den Mittelwert $\overline{x}_2^{(2)}$ der präsenten Werte ersetzt.

2. Udraw (Unbedingte Mean Imputation mit Zufallskomponente): zu der Imputation $\overline{x}_2^{(2)}$ wird eine Zufallsvariabel $z \sim \mathcal{N}(0, s_{22}^{(2)})$ addiert, wobei $s_{22}^{(2)}$ die Stichprobenvarianz von den vorhandenen Werten der Variable X_2 ist.

3. Cmean (Bedingte Mean Imputation): bzw. Bedingte Regression Imputation wie in Gleichung (2.1) für $K = 2$.

4. Cdraw (Bedingte Mean Imputation mit Zufallskomponente): bzw. Stochastic Regression Imputation wie in Gleichung (2.4) für $K = 2$.

Dabei ist die Vierte zu bevorzugen, da nur diese konsistente Schätzer für den Mittelwert μ_2, die Varaianz σ_{22}, und die Regressionskoeffizienten (X_2 auf X_1, und X_1 auf X_2) liefert. Allerdings wird durch das Hinzufügen der Zufallskomponente ein Verlust der Effizienz (speziell bei großen Samples ist die Varianz des Schätzers $\hat{\mu}_2^{Cdraw}$ größer als jene des $\hat{\mu}_2^{Cmean}$) verursacht. Außerdem sind die Standardfehler der Cdraw-Schätzer aus dem imputierten Datensatz immer noch zu klein sind. Das Hinzufügen dieser Zufallskomponente reicht also noch nicht aus, um die Variabilität der Daten wahrheitsgetreu nachzustellen. Beide Probleme werden laut Little und Rubin [8] durch Multiple Imputation (Kapitel (4)) behoben.

Abschließend soll in diesem Kapitel noch auf Sonderheiten des multivariaten Falls eingegangen werden, insbesondere für Hot-Deck Verfahren.

Multivariater Fall für Hot-Deck Verfahren

In der Regel steht man vor dem multivariaten Problem. In der Matrix X sind mehr als eine Variable unvollständig beobachtet, und das Fehlen der Daten folgt keinem speziellen Muster. Somit können innerhalb einer Beobachtung mehrere Datenpunkte fehlen. Es wird hier zwischen zwei Ansätzen unterschieden, welche im Allgemeinen nicht zum selben Ergebnis führen. Entweder das Imputationsverfahren wird univariat Variable für Variable, oder multivariat durchgeführt. Nachfolgend werden diese zwei Optionen für den Fall m-vieler fehlender Werte innerhalb einer Beobachtung näher erläutert.

1) Die Berechnungen werden für jede Variable unabhängig voneinander durchgeführt, sodass für jede Variable ein eigenes Auswahlverfahren bestimmt ist. Zum Beispiel könnten verschiedene Distanzfunktionen definiert sein. Im Allgemeinen führt dies zu Imputationen für die m-vielen fehlenden Werte einer Beobachtung aus mehr als einer Beobachtung. Soll dies verhindert werden (sonst kann es zu ungewöhnlichen Kombinationen der Werte innerhalb einer Beobachtung führen), so ist der multivariate Ansatz zu wählen.

2) Fehlende Datenpunkte einer Zeile werden nicht aus mehreren Zeilen, sondern aus ein und der selben, imputiert.

   a) Für jede Kombination fehlender Variablen ist ein eigenes Auswahlverfahren bestimmt. Bsp.: Fehlen in der Zeile i die Werte der Variablen X_1 und X_2 wird also anders vorgegangen, als in der Zeile j, in welcher X_1 und X_3 fehlen.

   b) Es ist ein Auswahlverfahren bestimmt, welches für jede Kombination fehlender Variablen Anwendung findet. In obigem Beispiel würde für die Zeilen i und j die selbe Methode eingesetzt werden.

   c) Eine Mischform aus den beiden erstgenannten: Sodass für alle Kombinationen, außer wenigen speziellen das selbe Verfahren angewandt wird. In obigem Beispiel würde für die Zeilen i und j die selbe Methode eingesetzt werden. Allerdings wird für jene Beobachtungen, in welchen die Variablen 2 und 3 fehlen, ein zweites Verfahren bestimmt.

Predictive Mean Matching (Kapitel 4.2.1) verfährt nach Ansatz 1).

3. Expectation Maximization

Steht man vor dem Problem, dass die zu analysierenden Daten nicht vollständig sind, so gibt es, wie schon in obigen Kapitel erwähnt, mehrere Varianten damit umzugehen. Die hier vorgestellte Alternative behandelt das Thema allerdings über den Umweg der Parameter eines Modells. Es werden also die fehlenden Daten mithilfe von geschätzten Modellparametern imputiert. Die Schätzer für die Modellparameter erlangt man wiederum mithilfe der Daten. Somit entsteht ein sich wiederholender Prozess, welcher die Parameterschätzer gegen deren Maximum Likelihood Schätzer konvergieren lassen soll.
Der EM Algorithmus ist ein iteratives Verfahren, welches zur Berechnung von Maximum Likelihood Schätzern herangezogen werden kann. Jede Iteration besteht aus zwei Schritten. Im Expectation-Step wird der Erwartungswert der log-Likelihood der Daten als Funktion des aktuellen Parameterwertes von θ berechnet (siehe Gleichung (3.5)). Im Maximization-Step wird diese Funktion über den Parameter θ maximiert. Jenes θ welches die Funktion maximiert ergibt den neuen Iterationswert (siehe Gleichung (3.6)). Um Konvergenz zu erreichen werden die Schritte solange hintereinander wiederholt, bis die Differenz der geschätzten Parameter zweier aufeinanderfolgender Iterationen gering genug ist.

3.1. Allgemeine Herleitung

Sei $X = (X_{obs}, X_{mis})$, und die Dichtefunktion von X zum Parameter θ faktorisiert als

$$f(X|\theta) = f(X_{obs}, X_{mis}|\theta) = f(X_{obs}|\theta) \cdot f(X_{mis}|X_{obs}, \theta). \tag{3.1}$$

Die Loglikelihood Funktion ergibt sich dann folgendermaßen

$$l(\theta|X) = l(\theta|X_{obs}, X_{mis}) = l(\theta|X_{obs}) + ln(f(X_{mis}|X_{obs}, \theta)). \tag{3.2}$$

Um θ zu schätzen ist nun die Zielfunktion

$$l(\theta|X_{obs}) = l(\theta|X) - ln(f(X_{mis}|X_{obs}, \theta)) \tag{3.3}$$

nach θ, bei gegebenen X_{obs}, zu maximieren. Bildet man auf beiden Seiten der Gleichung (3.3) den Erwartungswert mit der Dichtefunktion von X_{mis}, bei gegebenem X_{obs} und dem

aktuellen Wert von θ, θ^t, so folgt

$$\begin{aligned}
&\int l(\theta|X_{obs}) \cdot f(X_{mis}|X_{obs},\theta^t) dX_{mis} = \\
&\underbrace{\int l(\theta|X) \cdot f(X_{mis}|X_{obs},\theta^t) dX_{mis}}_{=:Q(\theta|\theta^t)} - \underbrace{\int ln(f(X_{mis}|X_{obs},\theta)) \cdot f(X_{mis}|X_{obs},\theta^t) dX_{mis}}_{=:H(\theta|\theta^t)} \\
&l(\theta|X_{obs}) \cdot \underbrace{\int f(X_{mis}|X_{obs},\theta^t) dX_{mis}}_{=1} = Q(\theta|\theta^t) - H(\theta|\theta^t) \\
&l(\theta|X_{obs}) = Q(\theta|\theta^t) - H(\theta|\theta^t)
\end{aligned} \tag{3.4}$$

Der E-Step ist nun mathematisch formulierbar. Sei θ^t der aktuelle Schätzwert für θ, dann findet der E-Step die erwartete Complete-data-Likelihood, wenn θ gleich θ^t wäre.

$$Q(\theta|\theta^t) = \int l(\theta|X) \cdot f(X_{mis}|X_{obs}, \theta = \theta^t)\, dX_{mis} \tag{3.5}$$

Der M-Step wählt θ^{t+1} sodass

$$Q(\theta^{t+1}|\theta^t) \geqslant Q(\theta|\theta^t), \qquad \forall\, \theta \tag{3.6}$$

Die Differenz zweier aufeinander folgender Likelihoods lässt sich anschreiben als

$$l(\theta^{t+1}|X_{obs}) - l(\theta^t|X_{obs}) = \big(Q(\theta^{t+1}|\theta^t) - Q(\theta^t|\theta^t)\big) - \underbrace{\big(H(\theta^{t+1}|\theta^t) - H(\theta^t|\theta^t)\big)}_{<0} \tag{3.7}$$

wobei der letzte Ausdruck wegen Jensen's Ungleichung kleiner Null ist. Damit ergibt sich, wegen (3.5) und (3.6), dass

$$l(\theta^{t+1}|X_{obs}) \geqslant l(\theta^t|X_{obs}). \tag{3.8}$$

Erzielt wird dies durch jeden allgemeinen EM Algorithmus. Es gilt Gleichheit in (3.8) genau dann, wenn $Q(\theta^{t+1}|\theta^t) = Q(\theta|\theta^t)$.

3.2. EM für Exponential Familien

Sollte die Verteilung des kompletten Datensatzes X der Familie der Exponentialverteilungen angehören, so ist die Anwendung des EM Algorithmus relativ einfach. Die Dichtefunktion von X zu θ lässt sich dann schreiben als

$$f(X|\theta) = b(X) \cdot exp\left(\frac{s(X)\theta}{a(\theta)}\right), \tag{3.9}$$

wobei θ ein $(d\times 1)$ Parametervektor ist, $s(X)$ ist ein $(1\times d)$ Vektor suffizienter Complete-Data-Statistiken, und a und b sind Funktionen von X und θ. Die Funktion $Q(\theta|\theta^t)$ aus Gleichung (3.5) wird dann wie folgt geschrieben

$$\begin{aligned} Q(\theta|\theta^t) &= \int \underbrace{\left(\frac{s(X)\theta}{a(\theta)} + ln(b(X))\right)}_{l(\theta|X)} \cdot f(X_{mis}|X_{obs},\theta^t)\, dX_{mis} \\ &= E(s(X)|X_{obs},\theta^t) \cdot \frac{\theta}{a(\theta)} + E\left(ln(b(X)|X_{obs},\theta^t\right) \end{aligned} \tag{3.10}$$

Wobei der Term $E\left(ln(b(X)|X_{obs},\theta^t\right)$ unabhängig von θ ist, und daher bei der Maximierung über θ vernachlässigt werden kann.
Für den $(t+1)$-ten Iterationsschritt besteht der EM Algorithmus nun aus

$$\begin{aligned} E-Step: &\qquad s^{t+1} = E(s(X)|X_{obs},\theta^t) \\ M-Step: &\qquad \theta^{t+1} = argmax\left(\frac{s^{t+1}\cdot\theta}{a(\theta)}\right) \end{aligned} \tag{3.11}$$

Zuerst werden die Statistiken s neu geschätzt. Danach wird θ^{t+1} als Lösung der Likelihoodgleichungen evaluiert.
Im Anschluss wird der EM Algorithmus für den Fall multivariat normalverteilter Daten vorgestellt, wobei dies ein Spezialfall der Exponentialfamilie ist.

3.3. EM multivariater normaler Fall

Im Detail wird in diesem Abschnitt, nach der Erläuterung von Little und Rubin [8], auf die Maximum Likelihood Schätzer des Mittelwerts und der Kovarianzmatrix eines multivariaten normal verteilten Datensatzes X eingegangen, welcher ohne bestimmtes Muster unvollständig beobachtet wurde. Dabei wird angenommen, dass der Missing-Data-Mechanismus zu vernachlässigen ist. Abschließend wird noch auf den Zusammenhang zu Allison [1] (Seite 18-21) eingegangen.
Sei $X = (X_{obs}, X_{mis})$ eine K-variate normalverteilte Stichprobe der Größe N, sodass jede Zeile unabhängig identisch verteilt ist $\mathbf{x}_i \sim \mathcal{N}(\mu,\Sigma)$. Der Mittelwert des Datensatzes ist also $\mu = (\mu_1,\mu_2,\dots\mu_K)$ und die Kovarianzmatrix $\Sigma = (\sigma_{jk}\ \ j,k = 1\dots K)$. X_{obs} ist nun folgendermaßen zu verstehen: $X_{obs} = (x_{obs,1}, x_{obs,2},\dots x_{obs,N})$, dabei besteht jedes $x_{obs,i} \in \mathbb{R}^{1\times K_{i_{obs}}}$ aus den beobachteten Variablen der i-ten Zeile. $K_{i_{obs}} \leq K$ beschreibt die Anzahl der beobachteten Werte der i-ten Zeile. Analog wird X_{mis} untergliedert. Der Parameter $\mu_{obs,i}$ besteht aus den Mittelwerten jener Variablen, welche in der i-ten Zeile beobachtet sind. $\Sigma_{obs,i}$ ist analog definiert. Die log-Likelihood Funktion basierend auf X_{obs} lässt sich nun wie folgt definieren

$$l(\mu,\Sigma|X_{obs}) = const - \frac{1}{2}\sum_{i=1}^{N} ln|\Sigma_{obs,i}| - \frac{1}{2}\sum_{i=1}^{N}(x_{obs,i}-\mu_{obs,i})^T\Sigma_{obs,i}^{-1}(x_{obs,i}-\mu_{obs,i}) \tag{3.12}$$

Der EM Algorithmus für (3.12) wird laut Little und Rubin und Kapitel 3.2 mithilfe der Statistiken

$$s = \left(\underbrace{\sum_{i=1}^{N} x_{ij}, \quad j = 1 \dots K}_{s_I} \qquad ; \qquad \underbrace{\sum_{i=1}^{N} x_{ij} x_{ik}, \quad j,k = 1 \dots K}_{s_{II}} \right) \tag{3.13}$$

implementiert. Wir wollen den linken Teil als s_I und den rechten als s_{II} bezeichnen.

E-Step

Diese Statistiken werden im $(t+1)$-ten Iterationsschritt innerhalb des E-Step berechnet (vgl. (3.11)). $\theta^t = (\mu^t, \Sigma^t)$ ist bekannt:

$$\begin{aligned} s_{I_j}^{t+1} &= E\left(\sum_{i=1}^{N} x_{ij} | X_{obs}, \theta^t \right) = \sum_{i=1}^{N} x_{ij}^t, \qquad j = 1 \dots K \\ s_{II_{jk}}^{t+1} &= E\left(\sum_{i=1}^{N} x_{ij} x_{ik} | X_{obs}, \theta^t \right) = \sum_{i=1}^{N} x_{ij}^t x_{ik}^t + c_{jki}^t, \qquad j,k = 1 \dots K \end{aligned} \tag{3.14}$$

wobei die Ausdrücke x_{ij}^t und c_{jki}^t näherer Erläuterung bedürfen

$$x_{ij}^t = \begin{cases} x_{ij} & \textit{wenn } x_{ij} \textit{ beobachtet ist}, \\ E(x_{ij} | x_{obs,i}, \theta^t) & \textit{wenn } x_{ij} \textit{ nicht beobachtet ist} \end{cases} \tag{3.15}$$

$$c_{jki}^t = \begin{cases} 0 & \textit{wenn } x_{ij} \textit{ oder } x_{ik} \textit{ beobachtet ist}, \\ Cov(x_{ij}, x_{ik} | x_{obs,i}, \theta^t) & \textit{wenn } x_{ij} \textit{ und } x_{ik} \textit{ nicht beobachtet sind} \end{cases} \tag{3.16}$$

Die bedingten Erwartungswerte und Kovarianzen aus (3.15) und (3.16) erhält man durch die Anwendung des Sweep-Operators (Kapitel 3.4) auf die erweiterte Kovarianzmatrix erzeugt durch die aktuellen Parameter $\theta^t = (\mu^t, \Sigma^t)$, sodass X_{mis} regressiert wird auf X_{obs}. $E(x_{ij} | x_{obs,i}, \theta^t)$ ist also das Ergebnis der Regression der Variable, für welche der Wert in der i-ten Zeile fehlt, auf jene Variablen, für welche der Wert in der i-ten Zeile nicht fehlt. Des weiteren ist $Cov(x_{ij}, x_{ik} | x_{obs,i}, \theta^t)$ die Kovarianz der Residuen aus den zwei Regressionen X_j und X_k auf $X_{obs,i}$, durchgeführt mithilfe der aktuellen Parameter θ^t.
Die fehlenden Werte wurden durch neue Schätzungen ersetzt und die Statistiken s_I und s_{II} berechnet, womit der E-Step abgeschlossen ist.

M-Step

Um die neuen Werte des Parameters θ, $\theta^{t+1} = (\mu^{t+1}, \Sigma^{t+1})$, zu berechnen, werden s_I^{t+1} und s_{II}^{t+1} benötigt.

$$\begin{aligned}
\mu_j^{t+1} &= \frac{1}{N}\sum_{i=1}^{N} x_{ij}^t, \qquad j = 1 \dots K; \\
\sigma_{jk}^{t+1} &= \frac{1}{N} E\left(\sum_{i=1}^{N} x_{ij}^t x_{ik}^t | X_{obs}\right) - \mu_j^{t+1}\mu_k^{t+1} \\
&= \frac{1}{N}\sum_{i=1}^{N}\left((x_{ij}^t - \mu_j^{t+1}) \cdot (x_{ik}^t - \mu_k^{t+1}) + c_{jki}^t\right), \qquad j,k = 1 \dots K;
\end{aligned} \tag{3.17}$$

Die Fehlerkorrekturtherme c_{jki} werden addiert um die Unterschätzung der Kovarianzen zu kompensieren. Damit ist der M-Step abgeschlossen. Abbruchkriterien können beispielsweise folgendermaßen definiert werden. Einerseits durch eine maximale Anzahl an Iterationen, andererseits mittels einem Konvergenzkriterium.

Algorithmus zusammengefasst

1. Initialwerte $\theta^0 = (\mu^0, \Sigma^0)$ bestimmen; bspw. mithilfe von ES 2.1, oder DS 2.2

E. 2. berechne (3.15) und (3.16); bspw. mithilfe des Sweep-Operators 3.4

3. berechne (3.14);

M. 4. berechne (3.17);

5. Gehe zu 2 bis das Abbruchkriterium erfüllt ist

Allison [1] beschreibt den Algorithmus für den Spezialfall einer Matrix mit 4 Spalten. Fehlt ein Wert der Spalte j, so wird er mithilfe der Regression dieser Spalte auf die restlichen imputiert (vgl. den Abschnitt über Bedingte Regression Imputation von einer prädiktiven Verteilung in Kapitel 2.3). μ wird als Mittelwert der vorhandenen und regressierten Werte berechnet. Für die Berechnung der Kovarianzen zwischen den zwei Spalten X_1 und X_2 (wobei jetzt angenommen wird, dass X_{i1} und X_{i2} fehlen) substituiert er $x_{i1}x_{i2}$ mit $x_{i1}x_{i2} + s_{12,34}$, wenn x_{i1} (und x_{i2}) durch die Regression X_1(bzw. X_2) auf X_3 und X_4 entsteht. $s_{12,34}$ ist die Kovarianz der Fehler der beiden Regressionen X_1 auf X_3 und X_4, und X_2 auf X_3 und X_4. Die Berechnung der Varianzen ist als Spezialfall analog abzuleiten, es werden die Fehlervarianzen der Regression der einen Variablen auf die anderen, zur Verfügung stehenden, addiert.

3.4. Sweep-Operator

Lineare Regression kann numerisch auf verschiedenen Wegen gelöst werden. Einer davon baut auf dem Sweep-Operator auf, welchen Little und Rubin in deren Buch [8] ab Seite 148 vorstellen. Notwendig ist hierfür der Vektor der Mittelwerte und die Kovarianzmatrix der zu analysierenden Daten $X \in \mathbb{R}^{N \times K}$. Es ist also nicht erforderlich die Daten selbst zur Verfügung zu haben, um die Regressionskoeffiezienten und deren Varianzen zu berechnen. Diesen Vorteil nutzen Algorithmen, welche mit unvollständigen Datenmatrizen hantieren, aus. Außerdem können die Parameter auf diese Weise beliebig adjustiert werden. Beispielsweise kann man statt der üblichen Kovarianzmatrix eine korrigierte verwenden (siehe Kapitel 3.3 und 4).
Der Sweep-Operator $SWP(.,.)$ bildet eine quadratische Matrix $G \in \mathbb{R}^{K+1 \times K+1}$ wieder auf eine quadratische Matrix $H = SWP(p, G)$ der selben Dimensionen ab. Für den hier erforderlichen Zweck macht es nur Sinn, wenn die Matrix G symmetrisch ist, dann ist auch H symmetrisch. Wegen der speziellen Gestalt von G wird sie von 0 bis K indiziert und nicht von 1 bis $K+1$, sodass die j-te Spalte zur j-ten Variable gehört. Das erste Argument $p \in \{1 \dots K\}$ steht für die Zeile, bzw. Spalte welche gesweept wird. Die Abbildung ist definiert durch folgende Rechenoperationen:

$$SWP(p,G) : \begin{cases} \mathbb{R}^{(K+1)\times(K+1)} & \longrightarrow \mathbb{R}^{(K+1)\times(K+1)} \\ (g_{ij})_{i,j=0\dots K} & \longmapsto (SWP(p,G)_{ij})_{i,j=0\dots K} \end{cases} \tag{3.18}$$

wobei $SWP(p,G)_{ij}$ definiert ist durch:

$$(SWP(p,G)_{ij})_{i,j=0\dots K} = h_{ij} := \begin{cases} -1/g_{p,p}, & \dots \quad i = j = p \\ g_{jp}/g_{p,p} =: h_{ji}, & \dots \quad j \neq p = i \text{ oder } j = p \neq i \\ g_{ij} - g_{ip}g_{pj}/g_{pp}, & \dots \quad j \neq i \neq p \end{cases} \tag{3.19}$$

Aus dieser Definition lässt sich ableiten, dass der Sweep-Operator kommutativ ist:

$$SWP(p, SWP(r,G)) = SWP(r, SWP(p,G)) = SWP((r,p),G) = SWP((p,r),G) \tag{3.20}$$

Für $X \in \mathbb{R}^{N \times K}$ wähle man $\tilde{G} \in \mathbb{R}^{K+1 \times K+1}$

$$\tilde{G} := \begin{bmatrix} 1 & \overline{x} \\ \overline{x}^T & \frac{1}{n} X^T X \end{bmatrix} \tag{3.21}$$

Wendet man nun $SWP(0,.)$ auf $\tilde{G}$ an, so ergeben sich die Regressionen X_i auf den Einsvektor $\iota = (1 \dots 1)^T \in \mathbb{R}^N$. Die Regressionskoeffizienten sind hier die Mittelwerte. Also folgt, dass die Varianz der Regressionsresiduen (u_i) für die Variable X_i, $Var(u_i) = N^{-1}\Sigma_j (x_{ij} - \overline{x}_i)^2 = s_{ii}$ ist. Die Matrix $\overline{\Sigma}$ ist also die Kovarianzmatrix der Regressionsfehler.

$$SWP(0,\tilde{G}) = \begin{bmatrix} -1 & \overline{x} \\ \overline{x}^T & \overline{\Sigma} \end{bmatrix} = \begin{bmatrix} -1 & \overline{x} \\ \overline{x}^T & \frac{1}{N}X^TX - \overline{x}^T\overline{x} \end{bmatrix} = \begin{bmatrix} -1 & \overline{x}_1 & \dots & \overline{x}_K \\ \overline{x}_1 & s_{11} & \dots & s_{1K} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{x}_K & s_{K1} & \dots & s_{KK} \end{bmatrix} \tag{3.22}$$

In Gleichung (3.22) wird die Spalte 0 gesweept, und damit die restlichen Variablen darauf regressiert. Die Variable 0 (bzw. der Intercept) wird dadurch zur erklärenden Variable. Das heißt, aus der Matrix $SWP(0,\tilde{G})$ sind alle Informationen zu k-vielen Regressionen direkt herauszulesen bzw. ableitbar. Wiederholt man diesen Vorgang für die Spalte 1, so werden die Variablen X_i, $i \in \{2 \dots K\}$, auf den Intercept (Spalte 0) und die Variable X_1 (Spalte 1) regressiert.

$$SWP((0,1),\tilde{G}) =$$

$$= \begin{bmatrix} -(1+\overline{x}_1^2/s_{11}) & \overline{x}_1/s_{11} & \overline{x}_2 - (s_{12}/s_{11})\overline{x}_1 & \dots & \overline{x}_k - (s_{1K}/s_{11})\overline{x}_1 \\ & -1/s_{11} & s_{12}/s_{11} & \dots & s_{1K}/s_{11} \\ \vdots & & s_{22} - s_{12}^2/s_{11} & \dots & s_{2K} - s_{1K}s_{12}/s_{11} \\ & & \ddots & & \vdots \\ \overline{x}_K - (s_{1K}/s_{11})\overline{x}_1 & & \dots & & s_{KK} - s_{1K}^2/s_{11} \end{bmatrix} =$$

$$= \begin{bmatrix} -A & B \\ B^T & C \end{bmatrix}, \qquad A \in \mathbb{R}^{2\times 2},\ B \in \mathbb{R}^{2\times K-1},\ C \in \mathbb{R}^{K-1\times K-1} \tag{3.23}$$

Für die Interpretation der Segmente A bis C sei daran erinnert welche Information aus den Zeilen und Spalten der Matrix $SWP(0,\tilde{G})$ herauszulesen ist. Sowie die erste Zeile in Gleichung (3.22), abgesehen von -1, die Koeffizienten für die Regressionen beinhaltet, so steckt diese Information für (3.23) in B. In der ersten Zeile von B stehen die $(K-1)$ vielen Intercepts zu den jeweiligen Regressionen, während die zweite Zeile die Koeffizienten der unabhängigen Variable X_1 wiedergibt. Konkret ist das Element $b_{1,j}$ der Intercept und $b_{2,j}$ der Koeffizient für die Regression von X_j ($j \in \{2 \dots K\}$) auf X_1 (und die Spalte 0). Wie auch in Gleichung (3.22) ist C die Kovarianzmatrix der Residuen der Regressionen. Außerdem ist die Varianz der Schätzer in B zu berechnen aus den Matrizen A und C.
Im allgemeinen Regressionsmodell $y = \tilde{X}\beta + u$ ist die Kovarianzmatrix des Schätzers $\hat{\beta}$ definiert durch $Cov(\hat{\beta}) = \sigma(\tilde{X}^T\tilde{X})^{-1}$. Für die einzelnen Modelle in Gleichung (3.23) ist $\tilde{X} = (\iota, X_1)$. Weil $N \cdot (\tilde{X}^T\tilde{X})^{-1} = A$, und die Schätzer für σ_i in der Diagonale der Matrix C

stehen, gilt

$$Cov(b_{.,j}) = \frac{1}{N} A \cdot c_{j,j}\,, \qquad j \in \{1 \dots K-1\}, \tag{3.24}$$

wobei $b_{.,j}$ die j-te Spalte von B ist. Zusammenfassend sind alle Informationen zu den $(K-1)$-vielen Regressionen in (3.23) aus der Matrix $SWP((0,1), \tilde{G})$ direkt herauszulesen bzw. ableitbar.

Um eine Variable auf l-viele ($l \in \{1 \dots K-1\}$) andere zu regressieren, wendet man den Sweep-Operator l-mal auf die Matrix $G := SWP(0, \tilde{G})$ an. Um die Interpretation des Ergebnisses des Sweep-Operators zu erleichtern, sei vorausgesetzt, dass die abhängigen Variablen den Spalten $l+2, \dots, K$ der Matrix G zugehören. Dadurch ist gewährleistet, dass $SWP((1 \dots l), G)$ in 4 rechteckige Segmente unterteilt werden kann, wie in Gleichung (3.23).

$$SWP((1, \dots, l), G) = \begin{bmatrix} -D & E \\ E^T & F \end{bmatrix}, \tag{3.25}$$

$$D \in \mathbb{R}^{(l+1)\times(l+1)},\ E \in \mathbb{R}^{(l+1)\times(K-l)},\ F \in \mathbb{R}^{(K-l)\times(K-l)}$$

Die Segmente D bis F haben die selben Rollen in der Bestimmung der Koeffizienten und Varianzen wie jene in Gleichung (3.23), nur dass sich die Dimensionen verändert haben. Die Matrix G nennt man die erweiterte Kovarianzmatrix.
Nach diesem Prinzip ist im Fall K=2 die Regression X_2 auf X_1 wie folgt durchzuführen:

$$G := \begin{bmatrix} -1 & \overline{x} \\ \overline{x}^T & \overline{\Sigma} \end{bmatrix} = \begin{bmatrix} -1 & \overline{x}_1 & \overline{x}_2 \\ \overline{x}_1 & s_{11} & s_{12} \\ \overline{x}_2 & s_{12} & s_{22} \end{bmatrix} \quad \Longrightarrow \tag{3.26}$$

$$SWP(1, G) = \begin{bmatrix} -A & B \\ B^T & C \end{bmatrix} = \left[\begin{array}{cc|c} -(1+\overline{x}_1^2/s_{11}) & \overline{x}_1/s_{11} & \overline{x}_2 - (s_{12}/s_{11})\overline{x}_1 \\ \overline{x}_1/s_{11} & -1/s_{11} & s_{12}/s_{11} \\ \hline \overline{x}_2 - (s_{12}/s_{11})\overline{x}_1 & s_{12}/s_{11} & s_{12} - s_{12}^2/s_{11} \end{array}\right]$$

Es gilt $A \in \mathbb{R}^{2\times2}$, $B \in \mathbb{R}^{2\times1}$ und $C \in \mathbb{R}$. Es ergibt sich, dass $h_{1,3} = b_1$ gleich dem Intercept der Regression X_2 auf X_1 ist, und $h_{2,3} = b_2$ gleich dem Slope ist. C ist die Varianz der Residuen der Regression.

4. Multiple Imputation

Bei Multipler Imputation (MI) werden mehrere neue Werte für nicht beobachtete Datenpunkte simuliert, oder sind von Simulationen abhängig. Little und Rubin [8] erwähnen, dass prinzipiell jede Methode des Einfachen Imputierens mit Zufallskomponente, durch mehrmaliges Wiederholen und unabhängige Simulationen für Multiple Imputation herangezogen werden kann. MI ist zwar bezüglich des Rechenaufwands benachteiligt, liefert jedoch konsistente Standardfehler.
Imputation, wie schon im Kapitel 2.3 vorgestellt, beschreibt in diesem Zusammenhang das Einsetzen eines Wertes anstatt eines fehlenden. Bei Multipler Imputation wird dieser Vorgang mehrmals wiederholt. Sei $M > 1$ die Anzahl der Wiederholungen, dann entstehen M viele imputierte Datensätze, jeweils unabhängig voneinander, aus einem unvollständigen gegebenen Datensatz. Ebenso ergeben sich M viele Schätzer für einen Parameter θ, welche sich aus den vervollständigten Datensätzen ableiten lassen. Sofern die a posteriori Verteilung von θ (berechnet aus dem vervollständigten Datensatz) multivariat normal ist, kann eine relativ kleine Anzahl an Wiederholungen ($1 < M \leq 10$) schon genügen, falls der Anteil an fehlenden Daten nicht zu groß ist (Little und Rubin geben in deren Buch [8] keine weiteren Angaben welcher Anteil noch als akzeptabel zu werten ist). Zusammenfassend können die Eckpunkte von Multipler Imputation folgendermaßen aufgelistet werden:

1) Auswahl einer Methode zur Daten Imputation (beispielsweise Data Augmentation oder Predictive Mean Matching)

2) Bestimmen von Startwerten für θ

3) Die Methode starten

4) Die Methode solange durchführen bis ein Abbruchkriterium gilt

5) Zurück zu 2) solange die Anzahl der imputierten Datensätze kleiner M ist

6) Die Ergebnisse aus den M vollständigen Datensätze zusammenfassen

Einer der Unterschiede zwischen EM und MI sticht sofort ins Auge. Der Output des EM-Algorithmus besteht aus genau einem vollständigen Datensatz, welcher mit den üblichen Methoden analysiert werden kann. Hingegen retourniert Multiple Imputation, je nach Eingabeparameter, M-viele Datensätze, deren gemeinsame, weiterführende Datenanalyse spezifische Methoden bedürfen (siehe dazu Kapitel 4.3). Es ist auch möglich die M Imputationen

parallel, statt hintereinander auszuführen, also innerhalb einer Iteration M-viele Imputatioen zu erzeugen. Dadurch kann die Rechenzeit verkürzt werden, allerdings ist es auch von der Imputationsmethode abhängig, ob dies möglich ist.
Im Rahmen dieser Diplomarbeit werden zwei Algorithmen vorgestellt, mithilfe derer Imputationen berechnet werden können; Eine modellbasierte (Data Augmentation, DA) und eine "Hybrid" -Methode (Predictive Mean Matching, PMM), welche Eigenschaften aus Modellbasierten Methoden und Hot Deck vereint. Der Hauptunterschied zwischen DA und PMM liegt in der Verteilung aus welcher die Imputationen gezogen werden. Für das Verfahren DA wird diese den Annahmen entsprechend modelliert, wohingegen in PMM auf die empirische Verteilung zurückgegriffen wird. Es sei hier darauf hingewiesen, dass kein Verfahren für jede beliebige Problemstellung anwendbar ist.
Für PMM wird der Iterationsprozess (die Punkte 2 bis 5) durch eine Ableitung des Gibbs Samplers, die Methode der Chained Equations, durchgeführt, welche Kapitel 4.2 vorstellt.

4.1. Data Augmentation

Data Augmentation (DA) beschreibt eine weitere Methode um fehlende Daten zu imputieren. Sie ist dem EM-Algorithmus sehr ähnlich. Im wesentlichen handelt es sich um einen Iterationsprozess, in welchem jede Iteration aus einem I-Schritt und einem P-Schritt bestehen. Multiple Imputation kann durch eine M-malige, voneinander unabhängige, Wiederholung von DA durchgeführt werden.
In den Kapitel 2 bis 3 wurden schon konkrete Varianten vorgestellt um Imputationen auszuführen. In Gleichung (2.4) wurde eine Korrektur der linearen Regression mithilfe eines Varianzterms vorgenommen. Im Kapitel 3 wurde dies, unter gewissen Voraussetzungen, mithilfe von mehrmaliger linearer Regression bewerkstelligt.

DA: I-Step

Der Imputations-Step des DA-Algorithmus ähnelt dem E-Step des EM-Algorithmus. Geht man davon aus, dass die Daten X ($\in \mathbb{R}^{N\times K}$) multivariat normalverteilt sind, das Fehlen der Daten keinem speziellen Muster genügen muss (siehe Abbildung 1.1) und der Missing-Data-Mechanismus ignoriert werden kann, so handelt es sich im Grunde genommen um eine Weiterführung der Methode aus Kapitel 3 und Kapitel 2.3 beziehungsweise der Gleichung (2.4). Die fehlenden Werte werden nicht auf die vorhandenen der selben Stichprobe regressiert, sondern auf Basis dieser Regression simuliert. Der imputierte Wert der i-ten Zeile und j-ten Spalte, der t-ten Iteration, ergibt sich somit als gezogene Zufallsvariable aus

$$x_{ij}^{t} \sim \mathcal{N}\left(\left(\beta_{i,0}^{t,(ij)} + \sum_{k\in I^{[i]}} \beta_{i,k}^{t,(ij)}\, x_{ik}\right),\ \sigma_{jj}^{t,(ij)}\right) \tag{4.1}$$

Dabei gilt, dass $k \in \{1 \dots K\} \backslash \{j\}$ in der Indexmenge $I^{[i]}$ enthalten ist, falls $o_{ik} = 0$ (vgl. (2.2)). Des Weiteren bezeichnen $\beta_{i,0}^{t,(ij)}$ bzw. $\beta_{i,k}^{t,(ij)}$ die Regressionskoeffizienten bzw. den dazugehörigen Intercept der Regression X_j auf die Variablen mit Spaltenindices $k \in I^{[i]}$ und $\sigma_{jj}^{t,(ij)}$ die Varianz der Residuen aus der selben Regression, der t-ten Iteration.
Die Regressionen werden, wie schon beim EM-Algorithmus in Kapitel 3, mithilfe der Parameter $\theta^t = (\mu 1t, \Sigma^t)$ und dem Sweepoperator (Kapitel 3.4) ausgeführt.

DA: P-Step

Nachdem man X_{mis}^t mithilfe der Parameter θ^t imputiert hat, wird $X^t = (X_{obs}, X_{mis}^t)$ behandelt als wäre es eine vollständig beobachtete Datenmatrix, und man erhält aus der $(t+1)$-ten Simulation von $\theta = (\mu, \Sigma)$ den neuen Wert des Parameters, θ^{t+1}.
Zuerst erzeugt man eine Simulation für Σ aus der Inv-Wishart Verteilung.

$$\Sigma^{t+1} \sim Inv - Wishart\left(S^t\ ,\ N-1\right) \tag{4.2}$$

wobei $S^t = Cov(X^t)$ die Stichprobenvarianz von X^t ist. Dann wird der Erwartungswert für X, μ, simuliert

$$\mu^{t+1} \sim \mathcal{N}\left(\overline{X}^t\ ,\ \frac{1}{N}\Sigma^{t+1}\right) \tag{4.3}$$

$\overline{X}^t \in \mathbb{R}^{K\times 1}$ ist das Stichprobenmittel der vollständigen Datenmatrix X^t.

Die $(t+1)$-te Iteration des Parameters $\theta = (\Sigma, \mu)$ ist somit berechnet, und der P-Step beendet. Der darauffolgende I-Step imputiert die fehlenden Datenpunkte mithilfe der soeben berechneten Schätzer.

4.2. Multiple Imputation mithilfe von Chained Equations

Das Ziel von MICE ist es, die bedingte multivariate Verteilung von X_{mis} gegeben X_{obs} zu approximieren, aus welcher die Imputationen gezogen werden. Nach der Annahme, dass die multivariate Verteilung von X ausschließlich von θ abhängt, können die fehlenden Daten imputiert werden, wenn θ bzw. dessen Verteilung bestimmt ist. MICE ermittelt die implizite multivariate a posteriori Verteilung durch iteratives Ziehen aus bedingten univariaten Verteilungen. Von besonderem Vorteil ist diese Variante, wenn das Berechnen der gemeinsamen Verteilung wesentlich komplizierter ist, als das Ziehen aus vielen bedingten Verteilungen. MICE bedient sich hier dem Prinzip eines Gibbs Samplers.

Gibbs Sampler

Rubin und Little [8] stellen den Gibbs Sampler für den allgemeinen Fall vor. Diese Methodik ermöglicht das Ziehen eines Zufallsvektors aus der gemeinsamen q-variaten Verteilung,

über den Umweg der bedingten univariaten Verteilungen. Seien $Y_1 \dots Y_q$ Zufallsvariablen, mit der gemeinsamen Verteilung $P(y_1 \dots y_q)$, so zieht der Gibbs Sampler stattdessen aus den bedingten Verteilungen $P(y_i|y_1 \dots y_{i-1}, y_{i+1} \dots y_q)$. Seien Startwerte $y^0 = (y_1^0 \dots y_q^0)$ mithilfe einfacher Schätzverfahren bestimmt, so wird die $(t+1)$-te Iteration der Ziehungen folgendermaßen berechnet:

$$\begin{aligned} y_1^{t+1} &\sim P(y_1|y_2^t \dots y_q^t) \\ y_2^{t+1} &\sim P(y_2|y_1^{t+1}, y_3^t \dots y_q^t) \\ y_3^{t+1} &\sim P(y_3|y_1^{t+1}, y_2^{t+1}, y_4^t \dots y_q^t) \\ &\vdots \\ y_q^{t+1} &\sim P(y_q|y_1^{t+1}, y_2^{t+1}, \dots y_{q-1}^{t+1}) \end{aligned} \tag{4.4}$$

Für jedes Ziehen werden die aktuellst verfügbaren Werte der restlichen Variablen verwendet um darauf zu bedingen. Allerdings wird auf den eigenen Wert der vorherigen Iteration nur indirekt Bezug genommen, und zwar über den Zusammenhang zu den anderen Variablen.
Little und Rubin [8] geben an, dass unter allgemeinen Bedingungen gezeigt werden kann, dass die Folge der Ziehungen $y^t = (y_1^t, \dots y_q^t)$ in Verteilung gegen einen Vektor $y^* = (y_1^*, \dots y_q^*)$ konvergiert, sodass gilt $y^* \sim P(y_1 \dots y_q)$.

Gibbs Sampler für Missing-data Probleme

Angenommen X sei eine Stichprobe aus der multivariaten Verteilung $P(X|\theta)$, welche durch den unbekannten Vektor θ vollständig beschrieben wird. Der MICE Algorithmus [1] approximiert die multivariate posteriori Verteilung von θ durch iteratives Ziehen aus den bedingten Verteilungen

$$\begin{aligned} &P(X_1|X_{-1}, \theta_1) \\ &\vdots \\ &P(X_K|X_{-K}, \theta_K), \end{aligned} \tag{4.5}$$

wobei diese Verteilungen durch die Parameter θ_j bestimmt werden. X_{-j} entspricht X ohne der j-ten Variable.
Der Gibbs Sampler wird hier angewandt indem sowohl die fehlenden Daten $X_{mis} = (X_{1_{mis}} \dots X_{K_{mis}})$ als auch die Parameter $\theta = (\theta_1 \dots \theta_K)$ die Rolle der Zufallsvariablen $y_1 \dots y_q$ einnehmen. Im Vergleich zu Gleichung (4.4) sind die Verteilungen zusätzlich auf die beobachteten Werte $X_{obs} = (X_{1_{obs}} \dots X_{K_{obs}})$ bedingt.
Wieder seien die Startwerte $\theta^0 = (\theta_1^0 \dots \theta_K^0)$ und $X_{1_{mis}}^0 \dots X_{K_{mis}}^0$ durch simple Statistiken bzw. Imputationen gegeben. Der $(t+1)$-te Iterationsschritt wird somit, wie in Gleichung

[1] van Buuren und Groothuis-Oudshhoorn [12], sowie van Buuren, Boshuizen und Knook [13]

(4.4), folgendermaßen berechnet:

$$\begin{aligned}
\theta_1^{t+1} &\sim P(\theta_1|X_{1_{obs}}, X_2^t \dots X_K^t) \\
X_{1_{mis}}^{t+1} &\sim P(X_{1_{mis}}|X_{1_{obs}}, X_2^t \dots X_K^t, \theta_1^{t+1}) \\
\theta_2^{t+1} &\sim P(\theta_2|X_1^{t+1}, X_{2_{obs}}, X_3^t \dots X_K^t) \\
X_{2_{mis}}^{t+1} &\sim P(X_{2_{mis}}|X_1^{t+1}, X_{2_{obs}}, X_3^t \dots X_K^t, \theta_2^{t+1}) \\
&\vdots \\
\theta_K^{t+1} &\sim P(\theta_K|X_1^{t+1}, \dots X_{K-1}^{t+1}, X_{K_{obs}}) \\
X_{K_{mis}}^{t+1} &\sim P(X_{K_{mis}}|X_1^{t+1} \dots X_{K-1}^{t+1}, X_{K_{obs}}, \theta_K^{t+1})
\end{aligned} \tag{4.6}$$

Wie auch im allgemeinen Fall, wird die zufällige Komponente $X_{j_{mis}}^t$ der Variable $X_j^t = (X_{j_{obs}}, X_{j_{mis}}^t)$ nicht in die Ziehung der $(t+1)$-ten Iteration miteinbezogen, im Sinne der Bedingung. Nur die Beobachtungen werden berücksichtigt. Die Gleichungen (4.4) und (4.6) versinnbildlichen die Namensgebung, "Verkettete Gleichungen", des MICE-Algorithmus. Wie die Verteilungen konkret aussehen, hängt von der Modellierung und dem gewählten Imputationsverfahren ab.

4.2.1. Predictive Mean Matching

Hierbei handelt es sich um eine Kombination mehrerer Verfahren, um fehlende Werte zu imputieren. Einerseits, wie der Name schon verrät, spielt predictive mean eine Rolle (vgl. den Abschnitt über Bedingte Regression Imputation von einer prädiktiven Verteilung in Kapitel 2.3), andererseits nimmt die Methode Eingenschaften eines Hot Deck-Verfahrens 2.3 an. Predictive Mean Matching [2] (PMM) unterscheidet sich von EM und DA in erster Linie dadurch, dass die Imputationen nicht neu berechnete Werte sind, sondern beobachtete. Darauf beruht die Verwandtschaft zu Hot Deck. Der Vorteil hierin liegt, dass es zu keinen unmöglichen oder unrealistischen Imputationen kommen kann (z.B.: ein Wert ungleich 0 und 1 für eine binäre Variable, negatives Einkommen, überdimensionale Körpergröße, negative Herzfrequenz, etc.). Um zu bestimmen, welcher Wert imputiert wird, bedient man sich einer Distanzfunktion $d(.,.)$. Sei der i-te Wert der Variable j nicht beobachtet, so imputiert man den k-ten Wert der Variable j, wenn die Distanz zwischen den Beobachtungen i und k die kleinste ist. Für die Berechnung der Distanzfunktion greift man auf die bedingte Regression unter einer prädiktiven Verteilung zurück (predictive mean).

Univariater Fall

Sei $X = (X_1, \cdots, X_K)$ und außer X_1 sind alle Variablen vollständig beobachtet. X_1 ist für $i = 1 \dots r$ vorhanden und fehlt für $r+1 \dots N$. Darüber hinaus sei X_2 gleich der konstante

[2] Little [7] sowie van Buuren, Boshuizen und Knook [13]

Einsvektor ι^T. Der Schätzer, $\tilde{x}_{i1}$, für einen fehlenden Wert x_{i1} wird für $i = r+1 \dots N$ wie folgt bestimmt:

$$\tilde{x}_{i1} := x_{k1} \tag{4.7}$$

für ein bestimmtes $k \in \{1 \dots r\}$, falls

$$\begin{aligned} &d(i,k) \leqslant d(i,l) \qquad \forall l \in \{1 \dots r\} \\ &d(i,k) := (\hat{x}_{i1} - \hat{x}_{k1})^2 \end{aligned} \tag{4.8}$$

wobei $\widehat{X}_1 = (\widehat{X}_{1_{obs}}^T, \widehat{X}_{1_{mis}}^T)^T = (\widehat{x}_{11} \dots \widehat{x}_{r1}, \widehat{x}_{r+1,1} \dots \widehat{x}_{N1})$ analog zu (2.1) bestimmt wird. Sei X untergliedert wie folgt

$$X = \left[\begin{array}{c|ccc} x_{11} & x_{12} & \dots & x_{1K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{r1} & x_{r2} & \dots & x_{rK} \\ \hline NA & x_{r+1,2} & \dots & x_{r+1,K} \\ \vdots & \vdots & \ddots & \vdots \\ NA & x_{N,2} & \dots & x_{NK} \end{array}\right] = \left[\begin{array}{c|c} X_{1_{obs}} & X_{-1_{obs}} \\ \hline X_{1_{mis}} & X_{-1_{mis}} \end{array}\right] \tag{4.9}$$

Dann berechnet sich $\widehat{X}_1$ aus

$$\begin{aligned} \widehat{X}_{1_{obs}} &= X_{-1_{obs}} \underbrace{\left(X_{-1_{obs}}^T X_{-1_{obs}}\right)^{-1} X_{-1_{obs}}^T X_{1_{obs}}}_{\hat{\beta}} \\ \widehat{X}_{1_{mis}} &= X_{-1_{mis}} \left(X_{-1_{obs}}^T X_{-1_{obs}}\right)^{-1} X_{-1_{obs}}^T X_{1_{obs}} \end{aligned} \tag{4.10}$$

Die Distanzfunktion kann auch andere Formen annehmen. Ein Beispiel hierfür ist die Mahalanobis-Distanz (siehe Little und Rubin [8]).
Um den Anforderungen für Multiple Imputation zu genügen, werden die Gleichungen (4.8) bis (4.10) adjustiert, sodass die Imputationen abhängen von einer, aus der a posteriori Verteilung der fehlenden Daten gezogenen, Zufallsvariable.

$$d(i,k) := (\check{x}_{i1} - \hat{x}_{k1})^2 \qquad \forall k \in \{1 \dots r\},\ \forall i \in \{r+1 \dots N\} \tag{4.11}$$

wobei

$$\begin{aligned} \check{X}_{1_{mis}} &= \widehat{X}_{1_{mis}} + X_{-1_{mis}} \cdot e \cdot \frac{1}{g} \\ e &\sim \mathcal{N}(0, cov(\hat{\beta})) \qquad e \in \mathbb{R}^{K-1 \times 1} \\ g &= \sqrt{\frac{u^2}{r-(K-1)}} \\ u^2 &\sim \chi^2_{r-(K-1)} \\ cov(\hat{\beta}) &= \left(X_{-1_{obs}}^T X_{-1_{obs}}\right)^{-1} \cdot \left(\widehat{X}_{1_{obs}} - X_{1_{obs}}\right)^T \left(\widehat{X}_{1_{obs}} - X_{1_{obs}}\right) \frac{1}{r-(K-1)} \end{aligned} \tag{4.12}$$

Predictive Mean Matching mit Chained Equations für den multivariaten Fall

Um Multiple Imputation mit der Methode PMM, implementiert mithilfe von MICE, im multivariaten Fall bei allgemeinem Muster fehlender Daten, durchzuführen, folge man diesem Ablauf:

1) Initialwerte für die fehlenden Datenpunkte werden aus den Randverteilungen gezogen und in die Matrix X eingesetzt. Die Randverteilungen werden aus den beobachteten Daten geschätzt.

2) $j = 1$

3) Die Imputationen der j-ten Variable $\widehat{X}_{j_{mis}}$ werden wieder gelöscht.

4) $\widehat{X}_{j_{mis}}$ wird aus der Verteilungsfunktion, bedingt auf die restlichen Variablen und $X_{j_{obs}}$ imputiert. Die neuen Imputationen werden für das Bedingen auf X_j beim Imputieren der Variablen $X_{j+1} \dots X_K$ verwendet. Siehe Gleichung (4.6) bzw. (4.11) und (4.12)

5) Falls $j < K$ zutrifft rechne $j = j + 1$ und zurück zu 3). Sonst zu 6).

6) Wiederhole 2) bis 5) solange bis Konvergenz erreicht ist. Die zuletzt imputierten Datenpunkte füllen X auf. Eine von M-vielen Imputationen ist damit beendet.

7) Wiederhole 1) bis 6) M-mal

Wie viele Iterationen bis zur Konvergenz notwendig sind ist in der Literatur nicht genau definiert. Rubin [9] diskutiert dieses Thema, dabei gilt prinzipiell, dass eine sehr geringe Anzahl an Iterationen schon reicht. Auf jeden Fall sollten die Ergebnisse nach der Anwendung von MI analysiert, und auf Konvergenz geprüft. Hinweise hierfür finden sich in van Buuren und Groothuis-Oudshoorn [14].
Diese Gibbs Sampling Methode lässt die Folge der Realisationen $\widehat{X}^t_{mis}$ unter recht allgemeinen Bedingungen in Verteilung gegen $\widehat{X}^*_{mis}$ konvergieren, wobei $\widehat{X}^*_{mis} \sim P(X_{mis}|X_{obs})$. Dennoch kann es vorkommen, dass wegen inkompatiblen Formulierungen der bedingten Verteilungen keine gemeinsame Verteilung existiert. Der Algorithmus wird hier nicht zu Konvergenz führen. Untersuchungen zu diesem Thema finden sich beispielsweise in Brand [3].

4.3. Zusammenfassung der M Imputationen

Nachdem durch mehrmaliges Wiederholen des Imputations-Algorithmus (Bsp.: PMM, DA) M viele Datensätze entstanden sind, gilt es nun diese sinnvoll zusammenzufassen. Hierfür werden die Parameter jedes Datensatzes separat geschätzt. So ergibt sich der Erwartungswert

des Parameters θ[3] als arithmetisches Mittel der M vielen Erwartungswerte der θ_m. Komplizierter ist die richtige Herangehensweise für die Evaluierung eines vernünftigen Varianzmaßes des Schätzers. Es werden zwei Terme addiert, welche der Varianz innerhalb einer, und zwischen mehreren Imputationen, entsprechen.
Im Allgemeinen wird jeder der M Datensätze mithilfe der selben Methode analysiert. Es enstehen also für einen Schätzer $\hat{\theta}$ M verschiedene $\hat{\theta}_1 \dots \hat{\theta}_M$ und M verschiedene Varianzen $V_1 \dots V_M$ der Schätzer. Der totale Schätzer entsteht aus dem Mittelwert der einzelnen.

$$\overline{\theta}_M = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m \tag{4.13}$$

Aus den Varianzen innerhalb einer Imputation

$$\overline{V} = \frac{1}{M} \sum_{m=1}^{M} V_m \tag{4.14}$$

und solcher zwischen zwei Imputationen

$$\overline{B} = \frac{1}{M-1} \sum_{m=1}^{M} (\hat{\theta}_m - \overline{\theta}_M)^2 \tag{4.15}$$

kombiniert man die totale Varianz des Schätzers $\overline{\theta}_M$ wie folgt

$$T = \overline{V} + \frac{M+1}{M} \overline{B}. \tag{4.16}$$

Little und Rubin [8] leiten die Formeln (4.13) bis (4.16) her, indem sie die

- a posteriori Verteilung des Parameters θ, bedingt auf die beobachteten Werte X_{obs} (4.17),

  und die

- a posteriori Verteilung des Parameters θ, bedingt auf den ganzen Datensatz $X = (X_{obs}, X_{mis})$ - wäre dieser vollständig beobachtet - (4.18),

  in Relation zueinander setzen.

$$p(\theta|X_{obs}) = const \cdot p(\theta) \cdot p(X_{obs}|\theta) \tag{4.17}$$

$$p(\theta|X_{obs}, X_{mis}) = const \cdot p(\theta) \cdot p(X_{obs}, X_{mis}|\theta) \tag{4.18}$$

[3] θ ist der zu schätzende Parameter (dieser entspricht nicht notwendigerweise den Parametern θ aus den Abschnitten 4.1 bzw. 4.2)

$$\begin{aligned} p(\theta|X_{obs}) &= \int p(\theta, X_{mis}|X_{obs})dX_{mis} \\ &= \int p(\theta|X_{mis}, X_{obs})p(X_{mis}|X_{obs})dX_{mis} \end{aligned} \tag{4.19}$$

Laut der Bayes-Statistik ist der Parameter θ eine Zufallsvariable, dessen a priori Verteilung $p(\theta)$ ist. Des weiteren beschreibt die Funktion $p(X_{obs}|\theta)$ die Dichte der beobachteten Werte, bedingt auf θ.
Gleichung (4.19) impliziert, dass die a posterior Verteilung von θ simuliert werden kann durch

- das m-malige Ziehen der fehlenden Werte (X^m_{mis}) aus der bedingten Verteilung $(p(X_{mis}|X_{obs}))$,
- dem Imputieren dieser und
- dem Ziehen von θ aus der Verteilung bedingt auf die vervollständigten Daten $(p(\theta|X_{obs}, X^m_{mis}))$.

In diesem Sinn approximiert Multiple Imputation das Integral in Gleichung (4.19) durch Summenbildung:

$$p(\theta|X_{obs}) \approx \frac{1}{M}\sum_{m=1}^{M} p(\theta|X_{obs}, X^m_{mis}) \tag{4.20}$$

wobei $X^m_{mis} \sim p(X_{mis}|X_{obs})$. Sofern der bedingte Erwartungswert und die bedingte Varianz der a posteriori Verteilung in (4.19) existieren, werden sie gleichermaßen approximiert:

$$\begin{aligned} E(\theta|X_{obs}) &\overset{(1)}{=} E\left[E(\theta|X_{obs}, X_{mis})|X_{obs}\right] \\ &\overset{(2)}{=} \int\int \underbrace{E(\theta|X_{obs}, X_{mis})}_{=:g} \cdot p(X_{obs}, X_{mis}|X_{obs})\, dX_{mis}\, dX_{obs} \\ &\overset{(3)}{=} \int E(\theta|X_{obs}, X_{mis}) \cdot p(X_{mis}|X_{obs})\, dX_{mis} \\ &\overset{(4)}{\approx} \frac{1}{M}\sum_{m=1}^{M} E(\theta|X_{obs}, X^m_{mis}) \\ &= \frac{1}{M}\sum_{m=1}^{M} \hat{\theta}_m = \overline{\theta} \end{aligned} \tag{4.21}$$

Wobei $\hat{\theta}_m := E(\theta|X_{obs}, X^m_{mis})$ gilt. Die Gleichung (1) in (4.21) hält wegen der Turmeigenschaft des bedingten Erwartungswertes. (2) ist die Bildung des Erwartungswertes der Funktion g, abhängig von X_{obs} und X_{mis}, mit der gemeinsamen und bedingten Dichte $p(X_{obs}, X_{mis}|X_{obs})$. (3) folgt weil $p(X_{obs}, X_{mis}|X_{obs}) = p(X_{mis}|X_{obs})$ und das Integral nach X_{obs} wegfällt, weil darauf schon bedingt wird. Schließlich ist (4) die Approximation des Integrals, wie sie schon

in Gleichung 4.20 durchgeführt wurde.
Die Herleitung für die Formel (4.16) basiert auf den analogen Argumenten:

$$Var(\theta|X_{obs}) = \underbrace{E[Var(\theta|X_{obs}, X_{mis})|X_{obs}]}_{\overline{V}} + \underbrace{Var[E(\theta|X_{obs}, X_{mis})|X_{obs}]}_{\overline{B}} \tag{4.22}$$

$$\begin{aligned} \overline{V}: \quad & E[Var(\theta|X_{obs}, X_{mis})|X_{obs}] \\ & \approx \frac{1}{M}\sum_{m=1}^{M} E(\theta^2|X_{mis}^m, X_{obs}) - (E(\theta|X_{mis}^m, X_{obs}))^2 \\ & = \frac{1}{M}\sum_{m=1}^{M} V_m = \overline{V} \end{aligned} \tag{4.23}$$

Wobei $V_m := Var(\theta|X_{obs}, X_{mis}^m)$ gilt.

$$\begin{aligned} \overline{B}: \quad & Var[\underbrace{E(\theta|X_{obs}, X_{mis})}_{:=a}|X_{obs}] = Var[a|X_{obs}] \\ & = E\left[\left(a - \underbrace{E(a|X_{obs})}_{\overline{\theta}}\right)^2 |X_{obs}\right] = E\left[\left(a-\overline{\theta}\right)^2|X_{obs}\right] \\ & = E\left[\left(E(\theta|X_{obs}, X_{mis}) - \overline{\theta}\right)^2|X_{obs}\right] \\ & \approx \frac{1}{M-1}\sum_{m=1}^{M}\left(E(\theta|X_{mis}^m, X_{obs}) - \overline{\theta}\right)^2 \\ & = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\theta}_m - \overline{\theta}\right)^2 = \overline{B} \end{aligned} \tag{4.24}$$

$$\begin{aligned} \Longrightarrow Var(\theta|X_{obs}) &\approx \frac{1}{M}\sum_{m=1}^{M} V_m + \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\theta}_m - \overline{\theta})^2 \\ &\approx \overline{V} + \frac{1+M}{M}\overline{B} \end{aligned} \tag{4.25}$$

V_m ist die Varianz von θ, berechnet mithilfe des m-ten durch Imputation vervollständigten Datensatzes (X_{obs}, X_{mis}^m). $\overline{V}$ beschreibt den Mittelwert eben dieser Schwankungen innerhalb jeder einzelnen Imputation. $\overline{B}$ steht für die Varianz zwischen den Imputationen. Zum Anschluss an Gleichung (4.16) fehlt noch der Korrekturterm $(1 + M^{-1})$, welcher für kleines M hinzugefügt wird (für den multivariaten Fall werden die Varianzen durch Kovarianzmatrizen und Quadrate durch Vektorprodukte ersetzt).
Für große Stichproben ist folgende Statistik, im skalaren Fall, t-verteilt

$$\frac{\theta - \overline{\theta}_M}{\sqrt{T}} \sim t_\nu\,, \qquad \nu = (M-1)\left(1 + \frac{M}{M+1}\frac{\overline{V}}{\overline{B}}\right)^2 \tag{4.26}$$

Daraus folgt ein 95% Intervallschätzer für den Parameter μ :

$$\left(\overline{\mu}_M - t_{\nu,0.975}\sqrt{T}\,,\, \overline{\mu}_M + t_{\nu,0.975}\sqrt{T}\right) \tag{4.27}$$

Fehlende Information wegen unvollständiger Daten

Um zu evaluieren wie ausschlaggebend das Fehlen der Daten auf die Analyse ist, stellen Rubin und Little [8] das Verhältnis der Zwischen-Imputation-Varianz zur totalen Varianz des Schätzers vor.

$$\Lambda = \frac{\frac{1+M}{M}\overline{B}}{\overline{V} + \frac{1+M}{M}\overline{B}} \tag{4.28}$$

Eine weitere Kennzahl (fraction of missing information due to nonresponse, fmi) wird in Rubin [9] präsentiert.

$$\begin{aligned} \gamma &= \frac{r + \frac{2}{\nu+3}}{r+1} \\ r &= \left(1 + \frac{1}{M}\right)\frac{\overline{B}}{\overline{V}} \end{aligned} \tag{4.29}$$

ν ist aus Gleichung (4.26) zu entnehmen. r beschreibt den prozentualen Zuwachs der Varianz durch fehlende Daten.
Hält man die Anzahl der Imputationen nicht fest, variiert also M, so notiert man die Variablen T, $\overline{B}$, $\overline{V}$, ν, Λ, r, γ mit einem zusätzlichen $_M$, also; T_M, $\overline{B}_M$, $\overline{V}_M$, ν_M, Λ_M, r_M, γ_M.

4.4. Erklärendes Beispiel

Um die Erläuterung der Methoden des Kapitels 4 abzurunden, sei hier ein Beispiel angeführt. Sei X multivariat normal verteilt und unvollständig beobachtet, und $M = 5$ gewählt. Gesucht sind Imputationen für die fehlenden Werte, sowie ein Schätzer für den Mittelwert μ und dessen Varianz. Um MI durchzuführen wird folgendermaßen vorgegangen:

- Data Augmentation M-mal unabhängig voneinander, mit den selben Startwerten für $\theta = (\mu, \Sigma)$, starten. Diese seien aus einem der Verfahren ES, DS oder auch EM berechnet.
- $\forall\ m \in \{1 \ldots 5\}$ iteriert DA solange bis Konvergenz, bzw. eine maximale Anzahl an Iterationen erreicht ist.
- Jede t-te Iteration besteht dabei aus einem I-Step und einem P-Step.
  - Im I-Step werden fehlende Daten, mithilfe der Parameter $\theta^t = (\mu^t, \Sigma^t)$ nach der Formel 4.1 imputiert.

  - Im P-Step werden die Parameter $\theta^{t+1} = (\mu^{t+1}, \Sigma^{t+1})$ mithilfe des eben imputierten Datensatzes nach den Formeln 4.2 und 4.3 erneut geschätzt.

- Nach der letzten Iteration wird der Datensatz der m-ten Imputation X^m gespeichert. Das Schätzen von Statistischen Größen eines jeden Datensatzes wird nun durchgeführt, als wäre dieser vollständig beobachtet worden. So wird der Mittelwert einer Variable (sei die Variable j festgehalten) mit der bekannten Formel berechnet, wobei die x_{ij}^m entweder tatsächlich beobachtete, oder, im letzten Iterationsschritt des DA-Algorithmus, imputierte Werte sind.

$$\hat{\mu}_m = \frac{1}{N} \sum_{i=1}^{N} x_{ij}^m \tag{4.30}$$

- Das Zusammenfassen dieser Größen basiert auf den Formeln (4.13) bis (4.16). Daraus lässt sich schließlich der Schätzer $\overline{\mu}_M$ berechnen, in welchem die Informationen der M-vielen Datensätze vereint werden.

$$\overline{\mu}_M = \frac{1}{5} \sum_{m=1}^{5} \hat{\mu}_m \tag{4.31}$$

  Die Varianz dieses Schätzers wird mit Gleichung (4.16) aus der Varianz innerhalb jeder Imputation und jener zwischen diesen wie folgt kalkuliert:

$$\begin{aligned} Var(\overline{\mu}_M) &= \frac{1}{5} \sum_{m=1}^{5} V_m + \frac{6}{5} \cdot \frac{1}{4} \sum_{m=1}^{5} (\hat{\mu}_m - \overline{\mu}_M)^2 \\ V_m &= \frac{1}{N-1} \sum_{i=1}^{N} (x_{ij}^m - \hat{\mu}_m)^2 \end{aligned} \tag{4.32}$$

5. Anwendung von R-Prozeduren

In den Kapitel 2 bis 4 wurden mehrere Techniken erläuterten, mit welchen das Missing-Data-Problem behandelt werden kann. Nachfolgend sollen einige dieser Anwendung finden. Zuerst wird anhand simulierter Matrizen der Vergleich zwischen ES, DS, EM und MI unter MCAR und MAR angestellt. Im Anschluss liegt ein unvollständig beobachteter Datensatz vor, welcher mit ES und MI behandelt wird. Darüber hinaus werden basierend auf diesen Missing-Data-Methoden folgende Analysen berechnet; Lineare Regressionsmodelle, Hauptkomponentenanalyse, Autokorrelationsfunktion und Partielle Autokorrelationsfunktion. Sowohl in Kapitel 5.1 als auch in 5.2 wird MI mittels PMM durchgeführt, wie es in Abschnitt 4.2.1 vorgestellt wurde.

5.1. Testen mit Zufallsmatrizen

Bevor es zur Gegenüberstellung der Methoden kommt, soll erklärt werden was, und wie verglichen wird. Hierfür wird in den nächsten Abschnitten erläutert wie die Zufallsmatrizen simuliert werden, wie der Missing-Data-Mechanismus implementiert ist, und womit die Güte der Resultate evaluiert wird. Erst dann wird auf die Vergleiche eingegangen.

5.1.1. Wie wird X simuliert?

In diesem Abschnitt werden Tests mit Zufallsmatrizen durchgeführt, welche stets nach dem selben Prinzip generiert, $X \sim \mathcal{N}(m, \Sigma)$ werden. Einzig die Anzahl der Zeilen N variiert zwischen 500 und 2500, sowie der Anteil an fehlenden Werten λ zwischen 0% und 50% gewählt wird, wobei der Mechanismus MCAR oder MAR gilt. Des weiteren hängt die Struktur der Zufallsmatrix von dem Parameter ρ ab. Dieser definiert die Matrix Σ folgendermaßen:

$$\Sigma = (\sigma_{ij}) \tag{5.1}$$

$$\sigma_{ij} := \rho^{|i-j|} \tag{5.2}$$

Für den Wert $\rho = 0.9$ ergibt sich also folgendes Σ:

$$\Sigma = \begin{pmatrix} 1.00 & 0.90 & 0.81 & 0.73 & 0.66 \\ 0.90 & 1.00 & 0.90 & 0.81 & 0.73 \\ 0.81 & 0.90 & 1.00 & 0.90 & 0.81 \\ 0.73 & 0.81 & 0.90 & 1.00 & 0.90 \\ 0.66 & 0.73 & 0.81 & 0.90 & 1.00 \end{pmatrix} \tag{5.3}$$

Für jede Kombination der Matrix-Parameter (N, λ, ρ) wird X 100-mal erzeugt. Dabei ist die Zahl 100 beliebig gewählt, um möglichst allgemeine Aussagen treffen zu können, und gleichzeitig den Rechenaufwand auf einem bewältigbaren Niveau zu belassen. Die Matrix-Parameter nehmen im Rahmen der Simulationen folgende Werte an:

- N: 500, 1000, 1500, 2000, 25000
- λ: 0%, 12.5%, 25%, 37.5%, 50%
- ρ: 0.9, 0.8, 0.7, 0.6, 0.5

Die Spaltenanzahl $K = 5$ von X wird nicht variiert. Auch der Mittelwert-Vektor m wird einmal zufällig gewählt und konstant gehalten. Daraus ergeben sich $(125 \cdot 100)$-verschiedene Varianten für X.

$$m = \begin{pmatrix} 100.0 & 80.0 & 20.0 & -100.0 & 50.0 \end{pmatrix} \tag{5.4}$$

Ein Parameter der das Fehlen der Daten mitbestimmt wurde schon erwähnt, λ. Im nächsten Abschnitt wird beschrieben, wie die , erst zufällig gezogenen, Datenpunkte wieder gelöscht werden, sodass sie imputiert werden können und die Güte des Testverfahren evaluiert werden kann.

5.1.2. Wie wird das Fehlen der Daten simuliert?

Für das Simulieren des Missing-Data-Mechanismus werden hier zwei Methoden vorgestellt um einerseits MCAR, und andererseits MAR zu erzeugen.

MCAR

Nachdem X aus einer Verteilung gezogen ist (und damit noch vollständig), wird eine Matrix der selben Größe definiert:

$$\begin{aligned} &\mathring{X} \in \mathbb{R}^{N \times K} \\ &\mathring{X} \sim \mathcal{N}(0,1) \;\; sodass \;\; \mathring{x}_{ij} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1) \end{aligned} \tag{5.5}$$

Der Wert x_{ij} wird durch NA ersetzt, falls der Wert $\mathring{x}_{ij}$ kleiner als das λ-Quantil der Matrix $\mathring{X}$ ist.

$$\mathring{x}_{ij} < q(\mathring{X}, \lambda) \implies x_{ij} \leftarrow NA \tag{5.6}$$

Das Fehlen der Datenpunkte ist somit vollkommen unabhängig vom Wert des Datenpunktes selbst und der anderen Variablen.

MAR

Unter MCAR bedient man sich also der Entscheidungshilfe $\mathring{X}_j$ für das Löschen der Zeilen des Vektors X_j. An die Stelle dieser Entscheidungshilfe rückt unter MAR der Vektor X_{j+2}.

$$\begin{array}{l|rcl} j = 1 \dots K-2 & x_{i(j+2)} < q(X_{(j+2)}, \lambda) & \implies & x_{ij} \leftarrow NA \\ j = K-1 & x_{i1} < q(X_1, \lambda) & \implies & x_{i(K-1)} \leftarrow NA \\ j = K & x_{i2} < q(X_2, \lambda) & \implies & x_{iK} \leftarrow NA \end{array} \tag{5.7}$$

Folglich ist das Fehlen der Datenpunkte unabhängig vom Wert der fehlenden Variable selbst, jedoch direkt abhängig vom Wert einer anderen Variable im selben Datensatz. Zu bedenken ist, dass über den Parameter ρ die Kovarianzmatrix von X bestimmt wird. Ein großer Wert von ρ impliziert eine indirekte Abhängigkeit des Fehlens der Daten vom eigenen Wert.
Bsp.: Würde X aus den Variablen $X_1 = Bruttoeinkommen$ und $X_2 = Nettoeinkommen$ von N-vielen Personen bestehen, so ist die Kovarianz, bzw. Korrelation der beiden Variablen sehr hoch. Sei nun der Nettowert von Probanden mit einem Bruttoeinkommen von über EUR 10.000 nicht angegeben, dann wird wegen der hohen Kovarianz MAR als Missing-Data-Mechanismus nicht mehr plausibel erscheinen, sondern NMAR. Das Fehlen der Variable X_2 hängt also direkt von X_1, und wegen der Beziehung der beiden Variablen indirekt von sich selbst, ab.

5.1.3. Wie wird die Performance einer Methode gemessen?

Es wird ein simpler Ansatz gewählt, um zu ermitteln welche Missing-Data-Methode sich eignet um die Schätzer der Mittelwerte und Kovarianzmatrizen zu berechnen, bzw. ob eine Methode einer anderen für diesen Zweck vorzuziehen ist.
Anhand der Abweichung der geschätzten Parameter $\hat{\theta} = (\hat{\mu}, \hat{\Sigma})$ von deren tatsächlichen Werten $\theta = (\mu, \Sigma)$ wird evaluiert, wie gut sich die jeweilige Methode eignet. Die tatsächlichen Werte der Parameter sind bekannt, weil die Matrix X mithilfe dieser simuliert wird. Die Abweichung wird gemessen mit der 2-Norm $\|.\|_2$, die für eine Matrix $A \in \mathbb{R}^{K \times K}$ definiert ist wie folgt:

$$\|A\|_2 = \sqrt{\sum_{i=1}^{K} \sum_{j=1}^{K} (a_{ij})^2} \tag{5.8}$$

Also ergeben sich zwei Maßzahlen für die Performance einer Missing-Data-Methode:

$$\Delta_{i,\mu} := \|\mu - \hat{\mu}_i\|_2 \qquad i \in \{1 \dots 100\} \tag{5.9}$$

$$\Delta_{i,\Sigma} := \|\Sigma - \hat{\Sigma}_i\|_2 \qquad i \in \{1 \dots 100\} \tag{5.10}$$

$$\Delta_{\mu} := \frac{1}{100} \sum_{i=1}^{100} \Delta_{i,\mu} \tag{5.11}$$

$$\Delta_{\Sigma} := \frac{1}{100} \sum_{i=1}^{100} \Delta_{i,\Sigma} \tag{5.12}$$

Diese Differenzen werden für jede Methode (MCAR, MAR) und jede Kombination von (N, λ, ρ) 100 mal berechnet und über die entsprechenden Matrix-Parameter gemittelt gemittelt. Das Ergebnis daraus wird im entsprechenden Ergebnis-Array (bspw.: Tabelle A.1) an der Stelle (N, λ, ρ) eingetragen. So ergeben sich Ergebnis-Tabellen, anhand derer die Performance der Methode abgelesen werden kann, sowie mit anderen Methoden verglichen. Die Resultate werden bis zur zweiten Nachkommastelle gerundet angegeben, und insofern analysiert.

5.1.4. Vergleich der Methoden ES und DS

Sowohl für den Missing-Data-Mechanismus MCAR, als auch MAR werden die Berechnungen durchgeführt.

- Jeder Eintrag der Tabellen A.1 bis A.5 entspricht $\Delta_{\mu}^{ES(N,\lambda,\rho)}$ bzw. $\Delta_{\mu}^{DS(N,\lambda,\rho)}$. Wobei $\Delta_{\mu}^{ES(N,\lambda,\rho)}$ als Mittel aus 100 Abweichungen (jeweils mit der 2-Norm gemessen) des geschätzten Parameters μ von seinem tatsächlichen Wert gebildet wird. μ wird aus der, mit den Parametern (N, λ, ρ) erzeugten, Matrix X geschätzt, nachdem ES angewandt wurde. Die Daten fehlen nach MCAR. $\Delta_{\mu}^{DS(N,\lambda,\rho)}$ ist analog zu verstehen.
- Jeder Eintrag der Tabellen A.6 bis A.10 entspricht $\Delta_{\Sigma}^{ES(N,\lambda,\rho)}$ bzw. $\Delta_{\Sigma}^{DS(N,\lambda,\rho)}$, bei MCAR. Die Definition von $\Delta_{\Sigma}^{ES(N,\lambda,\rho)}$ ist analog wie jene von $\Delta_{\mu}^{ES(N,\lambda,\rho)}$ zu verstehen.
- Jeder Eintrag der Tabellen A.11 bis A.15 entspricht $\Delta_{\mu}^{ES(N,\lambda,\rho)}$ bzw. $\Delta_{\mu}^{DS(N,\lambda,\rho)}$, bei MAR.
- Jeder Eintrag der Tabellen A.16 bis A.20 entspricht $\Delta_{\Sigma}^{ES(N,\lambda,\rho)}$ bzw. $\Delta_{\Sigma}^{DS(N,\lambda,\rho)}$, bei MAR

Ergebnisse zu MCAR (Tabellen A.1 bis A.10):

a) Sowohl $\Delta_{\mu}^{DS(N,\lambda,\rho)}$, als auch $\Delta_{\Sigma}^{ES(N,\lambda,\rho)}$ sind nahezu unabhängig von ρ.

b) Je größer N und kleiner λ, desto besser schätzen die beiden Methoden.

c) DS ist ES vorzuziehen, für das Schätzen von μ, insbesondere bei hohen Werten von λ und N.

d) DS ist ES vorzuziehen, für das Schätzen von Σ, insbesondere bei hohen Werten von λ und N.

Ergebnisse zu MAR (Tabellen A.11 bis A.20):

a) Konträr zu MCAR-a) sind sowohl $\Delta_{\mu}^{DS(N,\lambda,\rho)}$, als auch $\Delta_{\Sigma}^{ES(N,\lambda,\rho)}$ von ρ abhängig. Je kleiner ρ desto besser schätzen die beiden Verfahren. Dies resultiert aus der Konstruktion von X und der Missing-Data-Mechanismus.

b) Die Schätzfehler sind für MAR deutlich größer als unter MCAR

c) Der in MCAR-b) beschriebene Trend für N ist hier nicht zu erkennen. Die Schätzfehler scheinen sogar bei großem λ mit N zu wachsen.

d) Der in MCAR-b) beschriebene Trend für λ ist hier ebenfalls zu erkennen. Als je kleiner λ ist, desto besser schätzen die Methoden.

e) Auch unter MAR ist, für das Schätzen von μ und Σ, DS der Methode ES vorzuziehen.

5.1.5. Vergleich der Methoden EM und MI

Die Berechnungen werden in R durchgeführt. Der EM Algorithmus wird aus dem Paket *norm* [2] verwendet und MI aus *mice* [11]. Die Eingabeparameter für die Funktion *mice* sind die Imputationsmethode PMM, mit fünf Iterationen und fünf Imputationen.

- Jeder Eintrag der Tabellen A.21 bis A.25 entspricht $\Delta_{\mu}^{MI(N,\lambda,\rho)}$ bzw. $\Delta_{\mu}^{EM(N,\lambda,\rho)}$, bei MCAR. Die Definition von $\Delta_{\mu}^{MI(N,\lambda,\rho)}$ ist analog wie jene von $\Delta_{\mu}^{ES(N,\lambda,\rho)}$ zu verstehen, die Daten werden also mit MI anstatt ES vorbehandelt. Aus dem R-Paket *norm* wird der schon implementierte EM-Algorithmus für die Berechnungen herangezogen, sowie *mice* für MI.

- Jeder Eintrag der Tabellen A.26 bis A.30 entspricht $\Delta_{\Sigma}^{MI(N,\lambda,\rho)}$ bzw. $\Delta_{\Sigma}^{EM(N,\lambda,\rho)}$, bei MCAR.

- Jeder Eintrag der Tabellen A.31 bis A.35 entspricht der durchschnittlichen Durchlaufzeit des jeweiligen Algorithmus, bei MCAR.

- Wie in Kapitel 5.1.4 wird auch für die Methoden EM und MI zwischen MCAR und MAR unterschieden. Die Ergebnisse für MAR sind in den Tabellen A.40 bis A.50 eingetragen.

Ergebnisse zu MCAR (Tabellen A.21 bis A.35):

a) μ wird für $\lambda \neq 0$ annähernd so gut geschätzt wie für $\lambda = 0$. Das heißt die Schätzfehler der beiden Methoden bei $\lambda \neq 0$ sind nur geringfügig größer als jene der gewöhnlichen Stichprobenschätzer bei $\lambda = 0$.

b) Eine Abhängigkeit von ρ lässt sich nur in einem geringen Maß feststellen. Für $\rho = 0.5$ sind die Deltas um bis zu 0.09 größer als jene für $\rho = 0.9$.

c) Je größer N und kleiner λ, desto besser schätzen die beiden Methoden.

d) Den Tabellen A.31 bis A.35 ist zu entnehmen, dass MI eine deutlich höhere Durchlaufzeit bedarf.

e) Die Schätzverfahren EM und MI eignen sich für den Fall MCAR deutlich besser als ES um die Parameter zu schätzen. Jedoch ist nur EM der Methode DS in jedem Fall vorzuziehen, MI liefert lediglich bei kleinem N bessere Ergebnisse als DS.[1]

$$\Delta_\theta^{EM(N,\lambda,\rho)} \lesssim \Delta_\theta^{MI(N,\lambda,\rho)} < \Delta_\theta^{ES(N,\lambda,\rho)} \tag{5.13}$$

Ergebnisse zu MAR (Tabellen A.36 bis A.50):

a) sowohl für $\theta = \mu$ als auch $\theta = \Sigma$ lässt sich aus den Tabellen A.36 bis A.45 ableseen, dass $\forall\ N$, λ, ρ gilt:

$$\Delta_\theta^{EM(N,\lambda,\rho)} \lesssim \Delta_\theta^{MI(N,\lambda,\rho)} \tag{5.14}$$

b) Die Abhängigkeit vom Faktor ρ ist im Fall MAR deutlich zu sehen. Je kleiner ρ, desto geringer sind die Kovarianzen der Variablen. Also hängen die Spalten der Matrix X weniger voneinander ab, woraus der in Abschnitt 5.1.2 beschriebene indirekte Effekt schwächer ausfällt.

c) Die Durchlaufzeiten sind für MI deutlich größer als jene für EM. Zwischen MCAR und MAR besteht nur ein geringer Unterschied. Die Algorithmen rechnen jedoch für MCAR etwas schneller.

d) Die Schätzverfahren EM und MI eignen sich für den Fall MAR deutlich besser als ES und DS um die Parameter zu schätzen.

$$\Delta_\theta^{EM(N,\lambda,\rho)} \lesssim \Delta_\theta^{MI(N,\lambda,\rho)} < \Delta_\theta^{DS(N,\lambda,\rho)} \lesssim \Delta_\theta^{ES(N,\lambda,\rho)} \tag{5.15}$$

[1] $a \lesssim b$ bedeutet, dass a in den meisten Fällen kleiner ist als b

Die in diesem Abschnitt angestellten Vergleiche bevorzugen den EM Algorithmus klar, da die simulierten Daten X wie auf dessen Modellannahmen zugeschnitten sind. Dennoch ergab die Diskussion, dass unter bestimmten Umständen die Unterschiede der Techniken verschwindend klein sein können. Außerdem stellte sich heraus, dass EM und MI den einfachen Methoden ES und DS insbesondere im Fall MAR zu bevorzugen sind.

5.2. Unvollständig beobachtete Verkehrsdaten

In diesem Abschnitt soll anhand eines praktischen Beispiels vorgeführt werden, wie Multiple Imputation eingesetzt wird, und inwiefern diese Methode den, durch das Fehlen der Daten bedingten, Informationsverlust beim Schätzen von Statistiken kompensieren kann.
Es wird ein Datensatz [10] analysiert, welcher dankenswerterweise vom AIT Austrian Institute of Technology zur Verfügung gestellt wurde. Das AIT ist Österreichs größte außeruniversitäre Forschungseinrichtung und ist unter den europäischen Forschungseinrichtungen der Spezialist für die zentralen Infrastrukturthemen der Zukunft.[2]
Nachfolgend wird die Problemstellung erläutert und die Grundstruktur des Datensatzes angegeben. Anschließend werden verschiedene Modelle formuliert, mit welchen das Problem der fehlenden Daten behandelt wird. Im Detail wird die Methode ES und MI durch PMM betrachtet. Im Anschluss werden die Resultate der Imputationen überprüft. Abschließend wird ein Vergleich zwischen ES und MI für mehrere typische Problemstellung der Datenanalyse angestellt.

Der Datensatz $\tilde{Z} = (\tilde{Y}, \tilde{X})$ speichert Informationen einer Flotte von Taxifahrern. Jedes Fahrzeug dieser Flotte sendet in regelmäßigen Abständen seine Position und Geschwindigkeit an die Zentrale. Viermal pro Stunde werden diese Daten, betreffend die letzten 15 Minuten, zusammengefasst. Die Datenerhebung findet in ganz Wien statt, wobei sich die hier präsentierte Analyse nur auf einen Teilabschnitt des Straßennetzes bezieht. Dafür wird jede Straße in Abschnitte unterteilt, sogenannte Links. Die Segmentierung ist so geregelt, dass innerhalb eines Links keine Abbiegemöglichkeit vorkommt. Das bedeutet, kann man auf einer beliebigen Strecke 2 mal abbiegen, so besteht diese Strecke aus zumindest drei Links. Mittels dieser Erhebung entstehen je Link l und Zeitintervall t zwei Werte:

- $\tilde{y}_{tl} \in \tilde{Y} \ldots$ ist die Anzahl an Taxis, die im Zeitraum t den Link l befahren haben
- $\tilde{x}_{tl} \in \tilde{X} \ldots$ ist die Durchschnittliche Geschwindigkeit der Taxis, die im Zeitraum t den Link l befahren haben

Im Zuge dieser Diplomarbeit werden sechs Links näher analysiert, sie sollen im folgenden als (A, B, C, D, E, F) benannt werden. Aneinandergereiht stellen diese einen nahezu geraden Straßenabschnitt von ungefähr 650 Meter dar, zu sehen in Abbildung 5.1.

[2] AIT Austrian Institute of Technology: `http://www.ait.ac.at/`

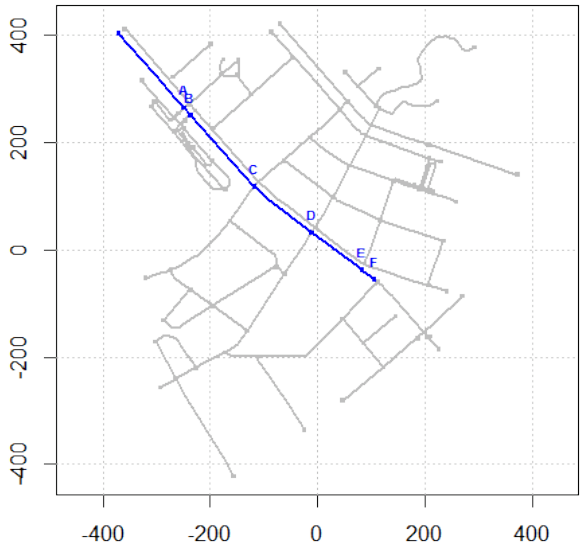

Abbildung 5.1.: Ein Ausschnitt des Wiener Straßennetzes: Die blau gefärbten Links werden analysiert.

Aus den zur Verfügung stehenden Daten wird ein Zeitabschnitt von 2 Monaten analysiert, was in etwa einer Stichprobengröße von $N = 5800 \approx 60 \cdot 96$ entspricht. In Tabelle 5.1 werden die den Links entsprechenden Datensätze $\tilde{X}$ und $\tilde{Y}$ zusammengefasst. Bewegt sich im Zeitraum $t \in \{1, \ldots, 5800\}$ kein Fahrer durch den Link $l \in \{A, B, C, D, E, F\}$, so gilt $\tilde{y}_{t,l} = 0$ bzw. $\tilde{x}_{t,l} = NA$. Für $\tilde{Y}$ entsteht also sehr wohl ein gültiger Wert, nämlich 0, jedoch kann aus fehlenden Beobachtungen für $\tilde{X}$ kein Wert interpretiert werden, also NA.
Im ursprünglichen Datensatz, bezogen auf den hier berücksichtigten Zeitabschnitt und die Links A bis F, fehlen die Werte für 105 Beobachtungen für alle sechs Links gleichzeitig. Diese Zeilen werden entfernt, man erhält den schon erwähnten Datensatz $\tilde{X}$. 60 dieser 105 Zeitpunkte entsprechen den Viertelstunden jedes Tages zwischen 23:45 und 00:00 des nächsten Tages. Weitere 26 treten zwischen 6:30 und 6:45 auf. Die übrigen 19 verteilen sich ohne zu erkennendes Muster. Dieses Auftreten an fehlenden Daten hat keinen Einfluss auf die MAR-Annahme, da man die Zeit als zusätzliche Variable betrachten könnte, womit das Fehlen von eben dieser abhängen würde, also MAR entspräche.
In Tabelle 5.2 wird angegeben nach welchem Muster Daten von $\tilde{X}$ fehlen. Es gibt demnach 5743 vollständige Beobachtungen. Für 4 Zeitfenster sind ausschließlich im Link B keine Taxifahrer gewesen. Analog kann man die restlichen Zeilen interpretieren. Folglich sind in keinem der Zeitfenstern gar keine Fahrer auf dieser Strecke zusehen gewesen (wegen der erwähnten

Vorbehandlung der Daten). Insgesamt fehlen 151 $\tilde{x}_{t,l}$, was sich ebenfalls aus der Summe der NA's in Tabelle 5.1 ergibt.

	A	B	C	D	E	F	A	B	C	D	E	F
Min.	5.1	4.9	5.4	4.8	5.5	5.6	0	0	0	0	0	0
1st Qu.	24.4	26.0	25.0	24.6	24.8	24.1	6	5	6	6	6	5
Median	29.2	30.4	29.1	28.4	28.4	27.8	10	8	11	10	10	9
Mean	28.5	30.0	28.8	28.1	28.0	27.4	9	7	10	9	9	8
3rd Qu.	33.3	34.3	32.8	31.8	31.5	31.1	14	11	15	13	13	12
Max.	65.8	65.8	65.8	68.5	68.5	68.5	45	34	51	39	38	35
NA's	27.0	39.0	25.0	26.0	17.0	17.0						

Tabelle 5.1.: Summary des Datensatzes: Auf der linken Seite die durchschnittlichen Geschwindigkeiten der Links $\tilde{X}$, auf der rechten Seite die Anzahl der Beobachtungen der Links $\tilde{Y}$

In den Abbildungen 5.2 und 5.3 wird nochmals auf die Datenstruktur eingegangen. Für jede Viertelstunde des Tages wird aus den ungefähr 60 Tagen die Durchschnittsgeschwindigkeit ermittelt, bzw. die Fehlwahrscheinlichkeit (die Anzahl der fehlenden Datenpunkte je Link und Tageszeit mal $^{96}/_{N}$). So lässt sich erkennen, dass die Durchschnittsgeschwindigkeiten während der Nacht um bis zu 10 km/h höher liegen als tagsüber. Darüber hinaus fehlen in den Morgenstunden (speziell für die Links A bis C) die meisten Aufzeichnungen, wohingegen um die Mittagszeit wenige Beobachtungen unvollständig sind.

Vorkommen	E	F	C	D	A	B	#NA
5743	1	1	1	1	1	1	0
4	1	1	1	1	1	0	1
5	1	1	1	0	1	1	1
1	1	0	1	1	1	1	1
7	1	1	1	1	0	0	2
3	1	1	0	1	1	0	2
4	0	0	1	1	1	1	2
12	1	1	0	1	0	0	3
1	1	1	1	0	0	0	3
2	1	1	0	0	1	0	3
8	0	0	1	0	1	1	3
5	1	1	0	0	0	0	4
1	0	0	1	0	1	0	4
1	0	1	0	0	0	0	5
1	0	0	1	0	0	0	5
2	0	0	0	0	1	0	5
	17	17	25	26	27	39	151

Tabelle 5.2.: Muster der fehlenden Daten von $\tilde{X}$

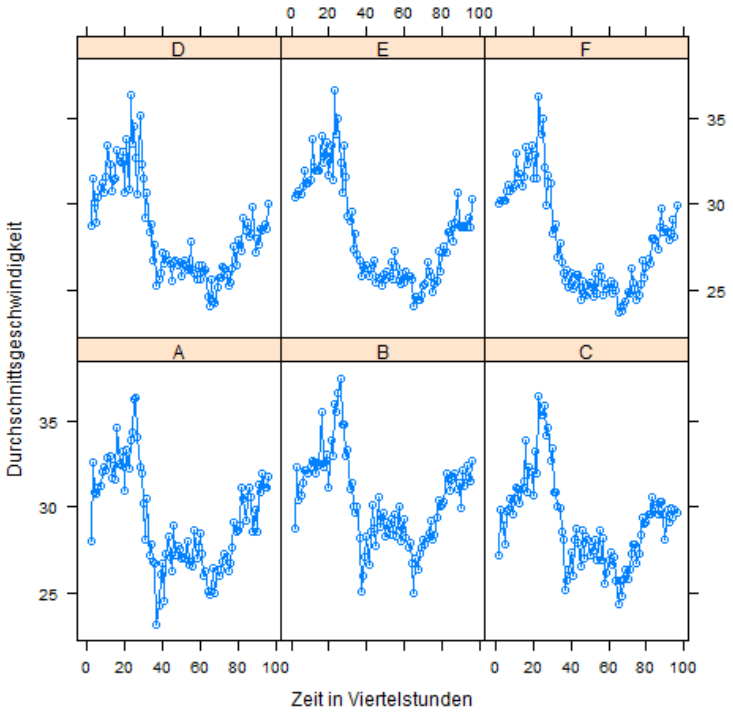

Abbildung 5.2.: Plot der Durchschnittsgeschwindigkeit für jede Viertelstunde des Tages

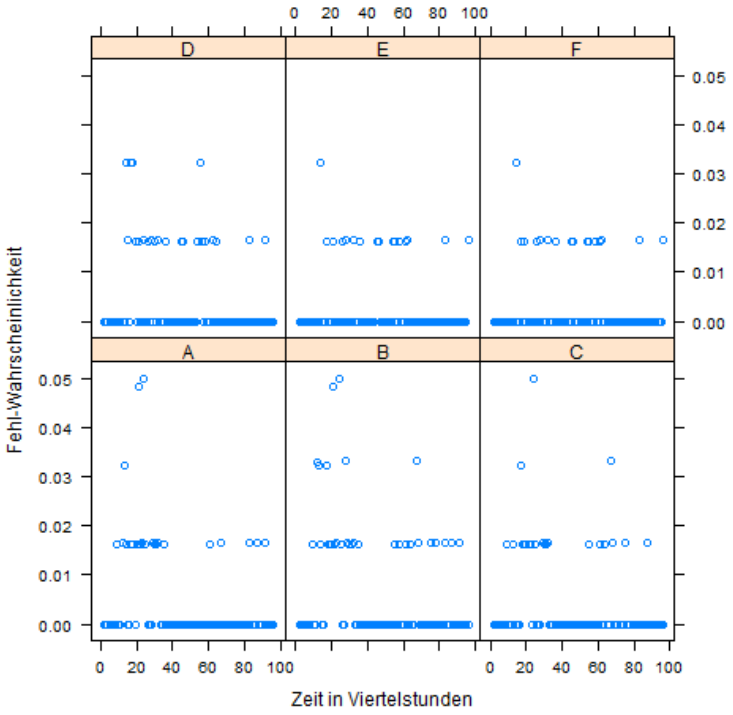

Abbildung 5.3.: Plot der Fehlwahrscheinlichkeit für jede Viertelstunde des Tages

5.2.1. MI angewandt auf die Verkehrsdaten

In Abschnitt 5.1 werden die Methoden EM und MI (mit PMM) auf normalverteilt simulierte Daten angewandt, sodass Schätzer für Mittelwerte und Kovarianzmatritzen berechnet werden können. Da der Expectation Maximization Algorithmus genau für diesen Fall konstruiert ist, liefert er bessere Resultate als MI. Jedoch ist der hier vorliegende Datensatz nicht normalverteilt, wie aus Abbildung 5.7 hervorgeht, und MI (mit PMM) ist verteilungs-unabhängig konzipiert. Darüber hinaus sollen die Auswirkungen des Auftretens unvollständiger Beobachtungen auf zusätzliche Statistiken untersucht werden, was mithilfe der in *mice* [11] implementierten Prozeduren (beispielsweise für lineare Regression) möglich ist.
Daher wird der beschriebenen Datensatz $\tilde{X}$ mit Multipler Imputation analysiert. Hierfür wird das R-Paket *mice* angewandt. Die Imputationen werden nach der Methode PMM 4.2.1 mithilfe von MICE 4.2 erzeugt. Anschließend wird die Plausibilität der Imputationen überprüft, wobei vernünftige Imputationen Werte annehmen, welche tatsächlich hätten beobachtet werden können. Zu diesem Zweck werden vor der Anwendung von MI zusätzliche 30% der Datenpunkte, nach MCAR, gelöscht. Die Imputationsfehler dieser künstlich erzeugten NA's können somit berechnet, und zwischen verschiedenen Modellen verglichen werden. Sei $\tilde{X}$ die originale Datenmatrix, dann erhält man X nach dem zusätzlichen Löschen der Daten aus $\tilde{X}$. Nachdem es sich beim Datensatz $\tilde{X}$ um Zeitreihen handelt, stecken Informationen über einen Wert $\tilde{x}_{tl}$ nicht nur in der t-ten Zeile, sondern auch in der l-ten Spalte. Für X gilt dasselbe. Soll heißen, um einen fehlenden Datenpunkt zu imputieren hat man in diesem Fall mehr Informationen auf die man zurückgreifen kann. Daher werden zusätzliche Modelle formuliert, welche die Vergangenheit und Zukunft jedes Datenpunktes berücksichtigen, und verglichen. Hierfür wird die Matrix X erweitert um die verzögerten und zukünftigen Beobachtungen. L ist der Lag-Operator, sodass

$$\begin{aligned} L^{j}(x_t) &= x_{t-j} \\ L^{-j}(x_t) &= x_{t+j} \end{aligned} \tag{5.16}$$

Will man also die Imputation eines fehlenden Wertes, zusätzlich zu den Werten der anderen Links zur selben Zeit, auf die Aufzeichnungen der vorigen Viertelstunde bedingen, so erweitert man die Matrix X um $L^1(X)$. Man schreibt auch nur $L(X)$. Um neben den vergangenen Werten auch die zukünftigen zu berücksichtigen, fügt man noch $L^{-1}(X)$ hinzu. Darüber hinaus kann die Datenmatrix noch um weitere Lags ausgedehnt werden. Will man das Zeitintervall vor 24 Stunden in die Datenmatrix aufnehmen, so wählt man den Lag 96.
In Folge werden mehrere Modelle formuliert und getestet. Die Ergebnisse beziehen sich jedoch immer nur auf die relevanten Links in X, also nicht auf deren Lags. Es werden neue, erweiterte, Datenmatrizen definiert, welche mit *mice* imputiert werden. Die Analyse über die

Modellname	zugrundeliegende Datenmatrix	Statistische Methoden werden angwandt auf
M.orig	$\tilde{X}_{cc}$	die originale Datenmatrix $\tilde{X}$, nach ES. Kontrollfall.
M.cc	X_{cc}	X (= die Datenmatrix $\tilde{X}$ nach zusätzlichem Löschen von 30%), nach ES.
M.0	X	die imputierten Datensätze von X, nach MI von X
M.1	$X^{(1)}$	die imputierten Datensätze von X, nach MI von $X^{(1)}$
M.2	$X^{(1,2)}$	die imputierten Datensätze von X, nach MI von $X^{(1,2)}$
M.96	$X^{(1,96)}$	die imputierten Datensätze von X, nach MI von $X^{(1,96)}$

Tabelle 5.3.: Verschiedene Methoden um mit dem Missing-Data-Problem umzugehen.

Güte der Imputationen bezieht sich im Anschluss ausschließlich auf X selbst.

$$\begin{aligned} X^{(1)} &:= \big(X, L(X), L^{-1}(X)\big) && \in \mathbb{R}^{N\times 3K} \\ X^{(1,2)} &:= \big(X, L(X), L^{-1}(X), L^{2}(X), L^{-2}(X)\big) && \in \mathbb{R}^{N\times 5K} \\ X^{(1,96)} &:= \big(X, L(X), L^{-1}(X), L^{96}(X), L^{-96}(X)\big) && \in \mathbb{R}^{N\times 5K} \end{aligned} \tag{5.17}$$

Würden die dadurch hinzugefügten Spalten auf die selbe Art imputiert werden, so bliebe die Beziehung zwischen den Variablen X_j und $L(X_j)$ im Allgemeinen nicht bestehen. Daher wird passive Imputation (siehe van Buuren und Oudshoorn [12]) angewandt. Die Beziehung zwischen den entsprechenden Spalten wird der R-Funktion mithilfe eines Lagoperators übergeben. Die Lags im Modell dienen also ausschließlich dem Bestimmen der Imputationen für die originalen Variablen.
Um aus dem unvollständigen Datensatz Statistische Größen zu berechnen, werden verschiedene Herangehensweisen verglichen. Tabelle 5.3 fasst diese zusammen. Das Modell M.orig dient zur Kontrolle. Vom originalen Datensatz $\tilde{X}$ fehlt nur ein kleiner Prozentsatz der Daten (siehe Tabelle 5.2). Nach ES werden Schätzverfahren darauf angewandt. Für die restlichen Modelle werden aus der Matrix $\tilde{X}$ erst 30% der Beobachtungen nach MCAR gelöscht. Man erhält X. Im Modell M.cc wird genauso verfahren wie im Modell M.orig, abgesehen davon, dass man von X ausgeht. Die Modelle M.0 bis M.96 wenden MI auf X, bzw. auf um Lags erweiterte Matrizen an, und beschränken die Anwendung der Statistiken auf die Imputierten Datensätze von X.

Fehlen die Daten von $\tilde{X}$ nach MAR?

Bevor mit dem Imputieren der fehlenden und gelöschten Datenpunkten begonnen wird, soll überprüft werden, ob der Mechanismus nach welchem die Daten fehlen MAR genügt.
Hierfür wird einerseits der Mittelwert der Geschwindigkeiten für alle Tageszeiten (96 Viertelstunden) evaluiert, andererseits die Wahrscheinlichkeit, dass ein Wert zu einer Tageszeit fehlt, also die Summe aller fehlenden Werte je Tageszeit durch die Anzahl an Beobachtungen je Tageszeit (also die selben Berechnungen die für die Abbildungen 5.3 und 5.2 angestellt wurden). Diese zwei Größen werden in Abbildung 5.4 gegeneinander geplottet. Es ist zu erkennen, dass das Fehlen von Datenpunkten nahezu unbeeinflusst ist von der mittleren Geschwindigkeit. In Tabelle 5.4 werden die Ergebnisse aus den linearen Regressionsmodellen, in welchen für jeden Link die Fehl-Wahrscheinlichkeit auf die durchschnittliche Geschwindigkeit regressiert wird, wiedergegeben. Zwar sind die geschätzten Koeffizienten für die ersten drei Links signifikant, jedoch liegen alle nahe bei Null. Darüber hinaus sind die R-squared's in Tabelle 5.5 ebenfalls sehr klein. Die Annahme, dass die Daten nach MAR fehlen, erscheint also vernünftig.

	Estimate	Std.Err	t-value	p-value
A: interc	-0.037	0.009	-3.916	0.000
A: coef	0.001	0.000	4.433	0.000
B: interc	-0.042	0.013	-3.132	0.002
B: coef	0.002	0.000	3.649	0.000
C: interc	-0.023	0.010	-2.305	0.023
C: coef	0.001	0.000	2.737	0.007
D: interc	-0.019	0.009	-2.101	0.038
D: coef	0.001	0.000	2.614	0.010
E: interc	-0.002	0.007	-0.371	0.711
E: coef	0.000	0.000	0.812	0.419
F: interc	-0.001	0.006	-0.198	0.843
F: coef	0.000	0.000	0.658	0.512

Tabelle 5.4.: Ergebnisse der linearen Regressionen: Die Fehl-Wahrscheinlichkeit regressiert auf die Durchschnittsgeschwindigkeit

	A	B	C	D	E	F
R-squared	0.174	0.125	0.075	0.068	0.007	0.005

Tabelle 5.5.: R-squared der linearen Regressionen: Die Fehl-Wahrscheinlichkeit regressiert auf die Durchschnittsgeschwindigkeit

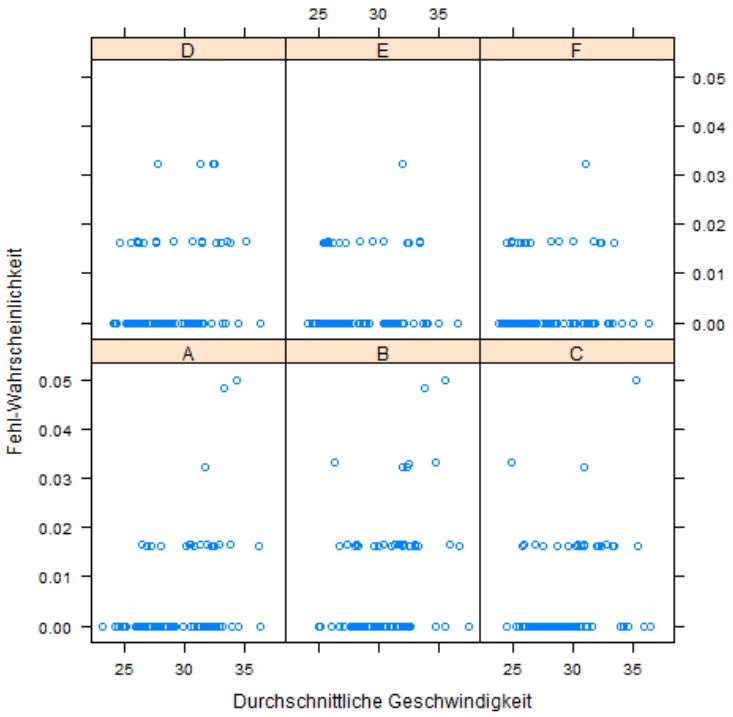

Abbildung 5.4.: Überprüfung der Annahme, dass die Daten nach MAR fehlen. Das Fehlen der Daten soll unabhängig von der durchschnittlichen Geschwindigkeit sein.

Ist θ a posteriori normal verteilt?

Zu Beginn des Kapitels 4 wird erwähnt, dass, sofern die mittels imputierten Datensatz ermittelte a posteriori Verteilung von θ eine multivariate Normalverteilung ist, schon wenige Imputationen genügen. Um die empirische Verteilungsfunktion von θ zu schätzen sei M groß gewählt, zum Beispiel 500. Exemplarisch werden für den Datensatz aus dem Modell M.0, also X, die Mittelwerte und Varianzen nach jeder Imputation herangezogen. Für jede Imputation wird zwanzig mal iteriert. So entstehen M viele Werte für μ und Σ. Es wird getestet ob die empirische Verteilung dieser einer Normalverteilung entspricht. In den Tabellen 5.6 und 5.7 sind die Ergebnisse von Shapiro Wilk Tests eingetragen. Nur für den Mittelwert des Links C und die Varianzen des Links A wird die Null Hypothese H_0, dass die Grundgesamtheit der Stichprobe normalverteilt ist, zu einem Signifikanzniveau von 5% verworfen. In den übrigen Fällen wird also von einer Normalverteilung ausgegangen. In Abbildung 5.5 werden die entsprechenden QQ-Plots dargestellt. Die empirischen Quantile der Stichproben werden gegen jene einer Normalverteilung dargestellt. Zusammengefasst ist diese Annahme zwar nicht vollständig belegt, jedoch genügen die Ergebnisse um fortzufahren und die Anzahl der Imputationen M klein zu belassen.

	A	B	C	D	E	F
W-Statistic	0.9977	0.9967	0.9937	0.9982	0.9965	0.9978
p-value	0.7425	0.4037	0.0370	0.8895	0.3418	0.7509

Tabelle 5.6.: Shapiro Wilk Testergebnisse für die Mittelwerte

	A	B	C	D	E	F
W-Statistic	0.9939	0.9971	0.9982	0.9966	0.9981	0.9979
p-value	0.0403	0.5373	0.8736	0.3654	0.8476	0.7855

Tabelle 5.7.: Shapiro Wilk Testergebnisse für die Varianzen

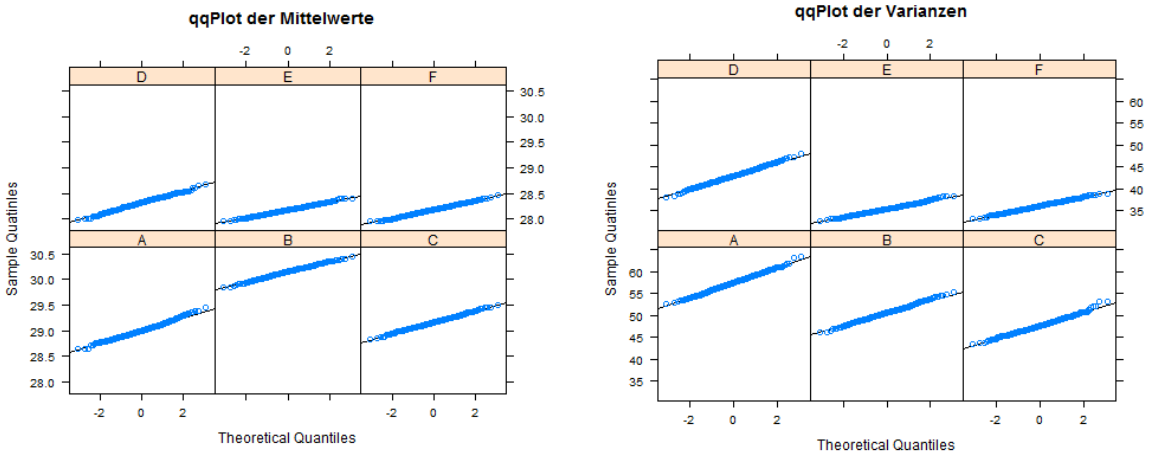

Abbildung 5.5.: Quantil-Quantil-Plots der empirischen Verteilungen für die Mittelwerte und Varianzen

Mittelwert und Kovarianz

Zuallererst werden die Mittelwertschätzer der einzelnen Variablen berechnet. Die Ergebnisse sind zusammengefasst in Tabelle 5.8. In den letzten drei Spalten sind die 2-Normen

- den Vektoren der Schätzer für die Mittelwerte $\|\hat{\mu}_\alpha\|_2$,
- den Differenzen dieser zum Kontroll-Modell M.orig $\|\hat{\mu}_{M.orig} - \hat{\mu}_\alpha\|_2$, und
- den Differenzen in Relation zur Norm aus M.orig $\|\hat{\mu}_{M.orig} - \hat{\mu}_\alpha\|_2 / \|\hat{\mu}_{M.orig}$

eingetragen, wobei α für eins der Modelle steht. Nachdem die 30% nach MCAR gelöscht wurden, überrascht es nicht, dass der Mittelwertschätzer des Modells M.cc jenen aus M.orig sehr gut approximiert, jedoch nicht so gut wie jene aus den MI-Modellen.
Tabelle 5.9 gibt die Resultate bezüglich der Schätzungen der Kovarianzmatrix wider. Die Bedeutung der Spalten ist analog wie für die Mittelwertschätzer zu verstehen. MI liefert hier deutlich bessere Schätzer als ES, hingegen ist zwischen den MI-Modellen wenig Unterschied. Aus M.96 geht die kleinste Differenz zum Kontrollmodell bzgl. der Kovarianz hervor.

	A	B	C	D	E	F	Norm	NDiff	NDiff%
M.orig	29.15	30.34	29.09	28.40	28.34	27.82	70.712	0.000	0.000
M.cc	29.18	30.42	29.12	28.29	28.28	27.74	70.671	0.174	0.246
M.0	29.10	30.36	29.11	28.37	28.37	27.87	70.728	0.087	0.123
M.1	29.10	30.37	29.11	28.36	28.37	27.89	70.736	0.106	0.150
M.2	29.11	30.39	29.12	28.39	28.40	27.88	70.773	0.111	0.157
M.96	29.12	30.39	29.13	28.37	28.36	27.86	70.749	0.089	0.126

Tabelle 5.8.: Mittelwertschätzer der verschiedenen Modelle

	Norm	NDiff	NDiff%
M.orig	198.003	0.000	0.000
M.cc	182.369	17.384	8.780
M.0	200.507	3.478	1.757
M.1	200.819	3.721	1.879
M.2	200.994	3.706	1.872
M.96	200.232	3.310	1.672

Tabelle 5.9.: Normen der Kovarianz-Matrix-Schätzfehler der verschiedenen Modelle

Imputationsfehler

Für die Modelle M.0, M.1, M.2 und M.96 können die Imputationsfehler (Differenz aus beobachtetem und imputiertem Wert) der zusätzlich gelöschten 30% berechnet werden. In Abbildung 5.6 werden diese aus den verschiedenen Modellen in Boxplots und Histogrammen dargestellt. Um die Unterschiede zwischen den Modellen besser erkennen zu können, sei auf die Tabelle 5.10 verwiesen, in welcher die Imputationsfehler (linke Seite) zusammengefasst werden, sowie deren Quadrate (rechte Seite). Die durch die Statistiken beschriebene Struktur der Fehler weist ebenfalls auf sehr ähnliche Modelle hin, jedoch wird M.96 wegen der Quartile und des Maximums der quadrierten Fehler präferiert.
Daher, und wegen den Ergebnissen aus den Schätzern für den Mittelwert und die Kovarianzmatrix, wird im weiteren Verlauf dieser Arbeit auf das Modell M.96 näher eingegangen.

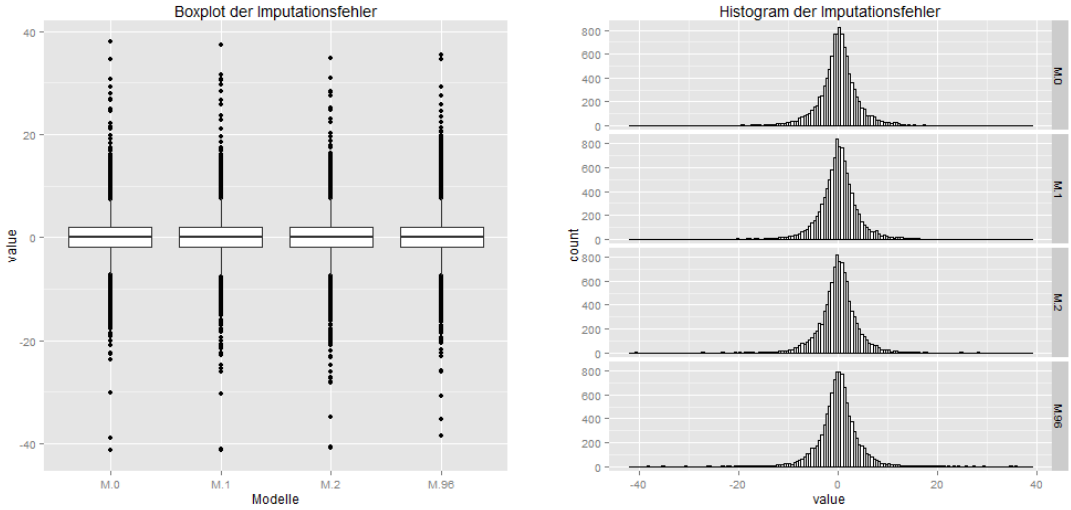

Abbildung 5.6.: Die Imputationsfehler nach dem Löschen von 30% der Beobachtungen

	M.0	M.1	M.2	M.96	M.0^2	M.1^2	M.2^2	M.96^2
Min.	-41.34	-41.37	-40.75	-38.45	0.00	0.00	0.00	0.00
1st Qu.	-1.76	-1.85	-1.82	-1.83	0.64	0.67	0.68	0.66
Median	0.06	0.07	0.01	0.04	16.65	16.12	16.20	15.84
Mean	0.10	0.09	0.07	0.06	3.39	3.61	3.47	3.50
3rd Qu.	1.91	1.94	1.90	1.90	13.73	13.39	13.76	13.40
Max.	38.06	37.34	34.84	35.50	1709.34	1711.33	1660.72	1478.69

Tabelle 5.10.: Summary der Imputationsfehler und der quadrierten Imputationsfehler

Konvergenz und Güte der Imputationen

Um fundierte Analysen aus den imputierten Datensätzen zu erhalten, bedarf es davor einer Überprüfung der Imputationen. Wie sich diese auf die Schätzer der Mittelwerte und Kovarianzen auswirken wurde schon eingehend in Kapitel 5.1 besprochen und für dieses Fallbeispiel in den Tabellen 5.8 und 5.9 ausgewertet.
Das Anwenden von MI soll darüber hinaus die Grundstruktur der Daten nicht verändern. Für die Überprüfung betrachte man Abbildung 5.7. Die Dichtefunktion der beobachteten Werte (blau) und jene der Imputationen (rot) ist für jeden Link dargestellt. Wobei die roten Graphen sich ausschließlich auf die Imputationen bezieht, und nicht auf die imputierten Datensätze.
In Abbildung 5.8 werden die Mittelwerte und Standardabweichungen der imputierten Datensätze nach jeder Iteration für jede der M Imputationen abgebildet. Bei wenigen Graphen kann ein Einpendeln beobachtet werden, sodass sich erst nach den ersten 10 bis 20 Iterationen kein Trend mehr abzeichnet. Besonders gut zu beobachten ist dies beispielsweise für den Schätzer der Standardabweichungen der Variablen E und F. Für die übrigen Statistiken reichen schon wenige Iterationen um Konvergenz zu erreichen, also keinem Trend mehr zu folgen.
Somit ist sowohl die Konvergenz der Iterationen als auch die Gültigkeit der Imputationen überprüft, da diese die empirisch geschätzte Verteilung der Daten nicht maßgeblich beeinflussen.

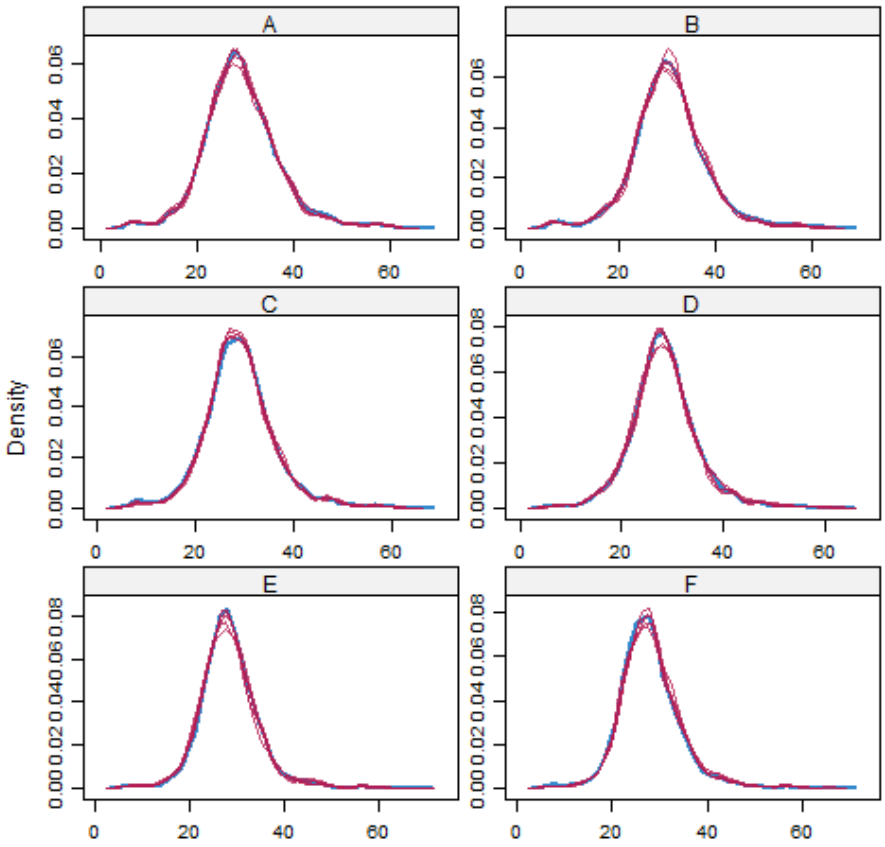

Abbildung 5.7.: Dichteplots der fünf Imputationen (rot) und beobachteten Daten (blau), für die sechs Links der Matrix Xv, nach zusätzlichem Löschen von 30%

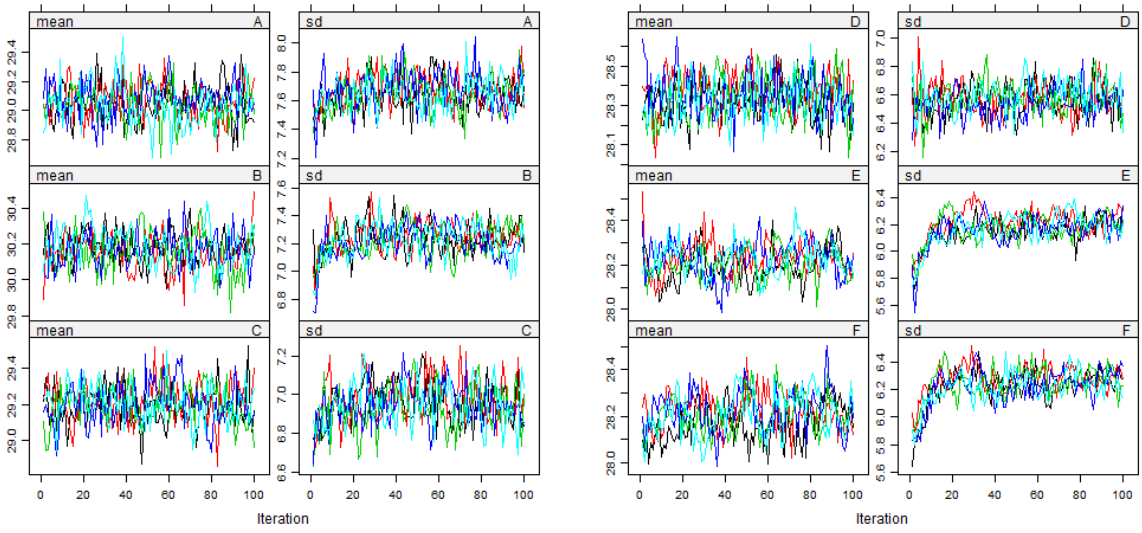

Abbildung 5.8.: Die Mittelwerte und Standardabweichungen der imputierten Datensätze nach den jeweiligen Iterationen, für alle fünf Imputationen.

Lineare Regression

Als Beispiel einer typischen Datenanalyse sei ein lineares Regressionsmodell geschätzt. Es wird der Link D, welcher von den restlichen Links eingeschlossen ist, auf diese regressiert. Hierfür seien wieder drei verschiedene Herangehensweisen verglichen, M.orig, M.cc und M.96. Die Tabellen 5.11 bis 5.13 enthalten die Ergebnisse. Während laut den Ergebnissen aus M.orig der Link D von allen anderen Links linear abhängt, signalisieren die p-Werte aus M.cc und M.96 Differenziertes. Außerdem sind die Standardfehler beider Modelle (bei M.cc mehr als M.96) deutlich größer als jene aus M.orig. Lambda gibt den Anteil der durch fehlende Daten verursachten Varianz an der totalen Varianz wieder (vgl. Gleichung (4.28)). Wegen dem hohen Anteil an fehlenden Daten kann auch MI die ursprünglichen Ergebnisse von M.orig nicht erreichen, dennoch ist dieser Ansatz jenem von M.cc vorzuziehen.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.172	0.241	4.867	0.000
A	0.030	0.011	2.686	0.007
B	-0.059	0.015	-3.824	0.000
C	0.310	0.013	23.217	0.000
E	0.955	0.022	43.021	0.000
F	-0.285	0.021	-13.769	0.000

Tabelle 5.11.: Lineare Regression mit M.orig

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.311	0.689	1.901	0.058
A	0.011	0.028	0.378	0.705
B	-0.015	0.040	-0.366	0.714
C	0.241	0.036	6.678	0.000
E	1.056	0.059	17.942	0.000
F	-0.352	0.056	-6.264	0.000

Tabelle 5.12.: Lineare Regression mit M.cc

	est	se	t	df	Pr(>|t|)	lo 95	hi 95	nmis	fmi	lambda
(Intercept)	1.347	0.343	3.928	15.825	0.001	0.619	2.075		0.554	0.501
A	0.015	0.022	0.655	7.181	0.533	-0.038	0.067	1820	0.795	0.745
B	-0.042	0.029	-1.453	8.011	0.184	-0.108	0.024	1700	0.759	0.705
C	0.297	0.029	10.125	6.406	0.000	0.227	0.368	1721	0.833	0.788
E	0.965	0.048	20.257	6.543	0.000	0.851	1.080	1721	0.826	0.780
F	-0.294	0.051	-5.745	5.729	0.001	-0.420	-0.167	1755	0.871	0.833

Tabelle 5.13.: Lineare Regression mit M.96

Hauptkomponenten-Analyse

Das Durchführen einer Hauptkomponenten-Analyse (Principal Component Analysis, PCA) soll dabei helfen einen multivariaten Datensatz durch einige (wenige) Linearkombinationen der Variablen aus diesem Datensatz, besser zu strukturieren und zu verstehen. Die folgenden Ergebnisse stammen aus einer zentrierten und skalierten PCA.
In den Tabellen 5.14 bis 5.16 sind in den ersten Zeilen die Standardabweichungen der jeweiligen Komponenten eingetragen. Diese ergeben sich aus den Eigenwerten der Kovarianzmatrix der normalisierten Datenmatrix. Die Zeilen darunter geben an, welchen Anteil die jeweilige Komponente an der gesamten Varianz trägt, beziehungsweise kumuliert. Zwischen den drei Modellen ist nahezu kein Unterschied zu erkennen.
In den Tabellen 5.17 bis 5.19 werden die Rotationsmatrizen abgebildet, welche das neue Koordinatensystem aufspannen. Dies sind die Eigenvektoren zu den oben erwähnten Eigenwerten. Die Modelle M.cc und M.96 approximieren die Vektoren aus M.orig, bis auf die Vorzeichen, recht gut. Betrachtet man wieder die Normen der Differenzen in Tabelle 5.20, so sind diese im Verhältnis zur Norm der Rotationsmatrix aus M.orig bei 12 und 5%. Wegen der Vorzeichen werden die Differenzen der Absolutbeträge der Matrizen berechnet. M.96 eignet sich also besser um die Rotationsmatrix zu schätzen. Die Normen der Rotationsmatrizen aus den verschiedenen Modellen sind ident, weil dies orthonormale Matrizen sind, also $2.449 \approx \sqrt{6} = \|R_\alpha\|_2$.
Die Abbildungen 5.9 bis 5.11 visualisieren die Ergebnisse der PCA für die drei Modelle. Auch hier ist, bis auf die Vorzeichen der Eigenvektoren, kaum ein Unterschied zu erkennen.

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	2.04	1.07	0.57	0.48	0.31	0.23
Proportion of Variance	0.69	0.19	0.05	0.04	0.02	0.01
Cumulative Proportion	0.69	0.88	0.94	0.97	0.99	1.00

Tabelle 5.14.: PCA mit M.orig

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	2.02	1.09	0.56	0.51	0.34	0.24
Proportion of Variance	0.68	0.20	0.05	0.04	0.02	0.01
Cumulative Proportion	0.68	0.88	0.93	0.97	0.99	1.00

Tabelle 5.15.: PCA M.cc

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	2.04	1.07	0.57	0.48	0.32	0.23
Proportion of Variance	0.69	0.19	0.05	0.04	0.02	0.01
Cumulative Proportion	0.69	0.88	0.94	0.98	0.99	1.00

Tabelle 5.16.: PCA mit M.96

	PC1	PC2	PC3	PC4	PC5	PC6
A	0.38	0.46	0.48	0.53	0.36	0.04
B	0.41	0.45	0.04	-0.19	-0.77	-0.07
C	0.43	0.29	-0.40	-0.55	0.51	0.06
D	0.42	-0.25	-0.64	0.56	-0.07	-0.20
E	0.42	-0.46	0.17	-0.06	-0.09	0.76
F	0.39	-0.48	0.41	-0.25	0.07	-0.61

Tabelle 5.17.: PCA mit M.orig, Rotationsmatrix

	PC1	PC2	PC3	PC4	PC5	PC6
A	-0.38	0.45	0.60	0.46	-0.29	-0.01
B	-0.41	0.44	-0.07	-0.21	0.76	0.07
C	-0.42	0.32	-0.43	-0.45	-0.57	-0.10
D	-0.42	-0.26	-0.53	0.63	0.03	0.26
E	-0.42	-0.46	0.15	-0.03	0.11	-0.76
F	-0.39	-0.47	0.37	-0.38	-0.06	0.58

Tabelle 5.18.: PCA M.cc, Rotationsmatrix

	PC1	PC2	PC3	PC4	PC5	PC6
A	-0.38	0.45	-0.46	0.56	0.35	-0.02
B	-0.41	0.45	-0.01	-0.17	-0.77	0.05
C	-0.43	0.30	0.33	-0.59	0.52	-0.05
D	-0.42	-0.26	0.68	0.51	-0.03	0.19
E	-0.42	-0.46	-0.18	-0.04	-0.08	-0.76
F	-0.39	-0.48	-0.43	-0.23	0.03	0.62

Tabelle 5.19.: PCA mit M.96, Rotationsmatrix

	Norm	NDiff	NDiff%
M.orig	2.449	0.000	0.000
M.cc	2.449	0.294	11.988
M.96	2.449	0.125	5.122

Tabelle 5.20.: Normen der Rotationsmatrizen aus den jeweiligen Modellen, sowie die Normen (absolut und relativ) der Differenzen der Rotationsmatrizen zu jenen aus M.orig. Die Differenzen werden von den Absolutbeträgen berechnet.

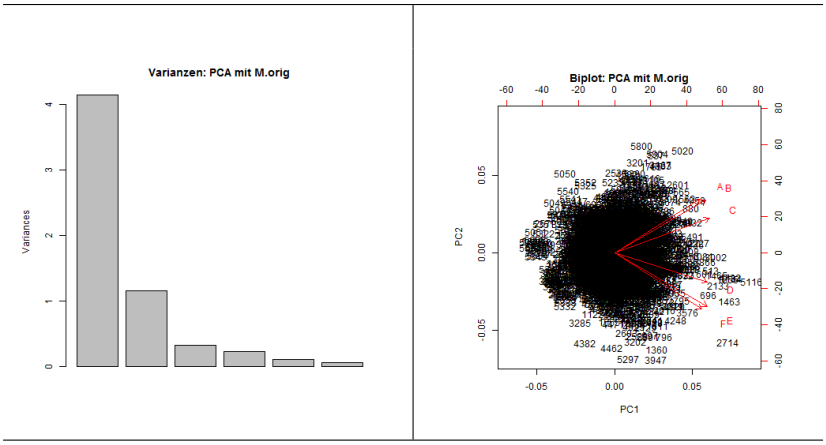

Abbildung 5.9.: PCA der Durchschnittsgeschwindigkeiten, M.orig

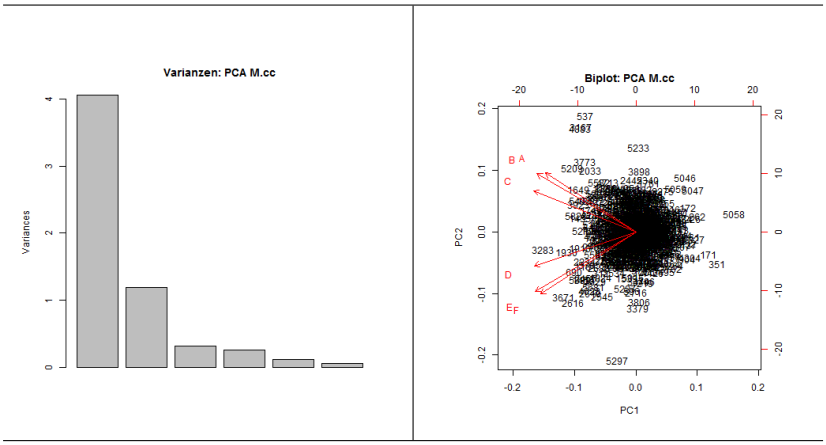

Abbildung 5.10.: PCA der Durchschnittsgeschwindigkeiten, M.cc

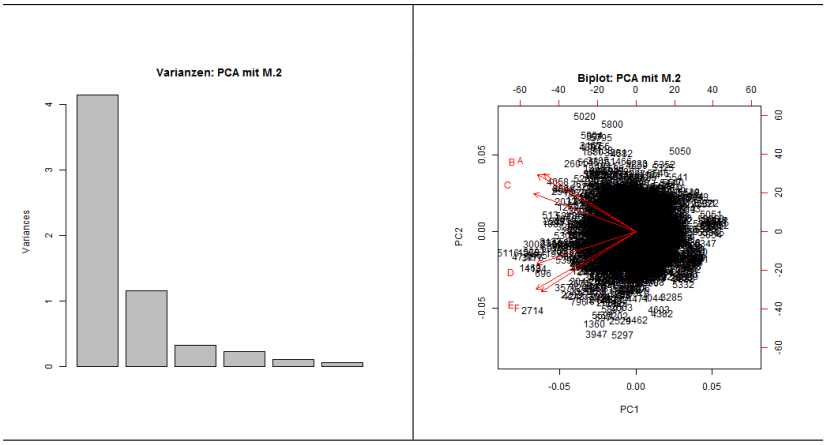

Abbildung 5.11.: PCA der Durchschnittsgeschwindigkeiten, M.2

Autokovarianzfunktion

Nachdem die Variablen in X Zeitreihen sind, wird in diesem Abschnitt analysiert wie sich das Fehlen von Daten auf die Autokovarianzfunktion (ACF) und die Partielle Autokovarianzfunktion (PACF) auswirkt.
Wieder werden die drei Modelle M.orig, M.cc und M.96 verglichen, allerdings mit Abänderungen. Die Berechnung der ACF entspricht der Berechnung der Kovarianzfunktion eines Vektors und dessen Lag. Eben dafür wurde in Kapitel 2.2 schon die Anwendung von DS vorgestellt, sowie der Vorteil gegenüber ES in Kapitel 5.2 demonstriert. Somit macht es nur Sinn, in den Modellen M.orig und M.cc anstatt ES das Verfahren DS anzuwenden[3]. Sonst führte ein Löschen aller unvollständiger Zeilen zur Zerstörung der Struktur der Zeitreihe.
Tabelle 5.21 beinhält die Resultate. Für jeden Link werden die ACF und PACF bis zu einem Lag von 100 Viertelstunden, also mehr als einem Tag, berechnet. Wie sehr diese von jener aus M.orig abweichen wird anhand der Norm der Differenz angegeben (die Zeilen NDiff in der Tabelle 5.21). Diese Abweichungen werden in Relation zur Norm $\|ACF_{M.orig}\|_2$ gesetzt und in den Zeilen NDiff ACF % und NDiff PACF % eingetragen. Daraus ergibt sich, dass für die Berechnung der ACF das Modell M.96 für fünf der sechs Links bessere Resultate erzielt, und, dass die PACF aus M.96 jene aus M.orig wesentlich genauer approximiert als jene aus M.cc, und zwar für alle sechs Links.
In Abbildung 5.12 werden für den Link A die ACF und PACF aus den drei Modellen visualisiert. Unterschiede der Plots sind beispielsweise bei genauer Betrachtung der ersten zehn Lags der PACF zu erkennen. Das Modell M.96 kann die Struktur aus M.orig besser rekonstruieren als M.cc.
Hätte man in den Modellen M.orig und M.cc ES anstatt DS belassen, so wäre die ACF für M.orig nicht stark verändert, weil der Prozentsatz an fehlenden Daten in $\tilde{X}$ gering genug ist. Für M.cc hätte dies jedoch fatale Folgen genommen, zu sehen in Abbildung 5.13.

[3] In R wählt man für die Funktion acf(...) den Parameter *na.action* = *na.pass* anstatt *na.action* = *na.exclude*.

Modell	Norm Typ	A	B	C	D	E	F
M.orig	ACF	1.644	1.679	1.725	1.789	2.164	2.162
M.cc	NDiff ACF	0.110	0.117	0.126	0.144	0.131	0.134
M.96	NDiff ACF	0.123	0.071	0.078	0.063	0.079	0.118
M.cc	NDiff ACF %	6.685	6.943	7.329	8.036	6.072	6.176
M.96	NDiff ACF %	7.508	4.222	4.499	3.539	3.672	5.439
M.orig	PACF	0.520	0.530	0.558	0.556	0.643	0.639
M.cc	NDiff PACF	0.152	0.176	0.182	0.200	0.173	0.173
M.96	NDiff PACF	0.094	0.075	0.080	0.067	0.069	0.070
M.cc	NDiff PACF %	29.172	33.178	32.555	35.969	26.901	27.045
M.96	NDiff PACF %	18.161	14.252	14.286	11.994	10.764	11.014

Tabelle 5.21.: Normen der ACF aus den zwei Modellen M.cc und M.96; Die Absolutwerte und die Verhältnisse zur Norm des Modells M.orig.

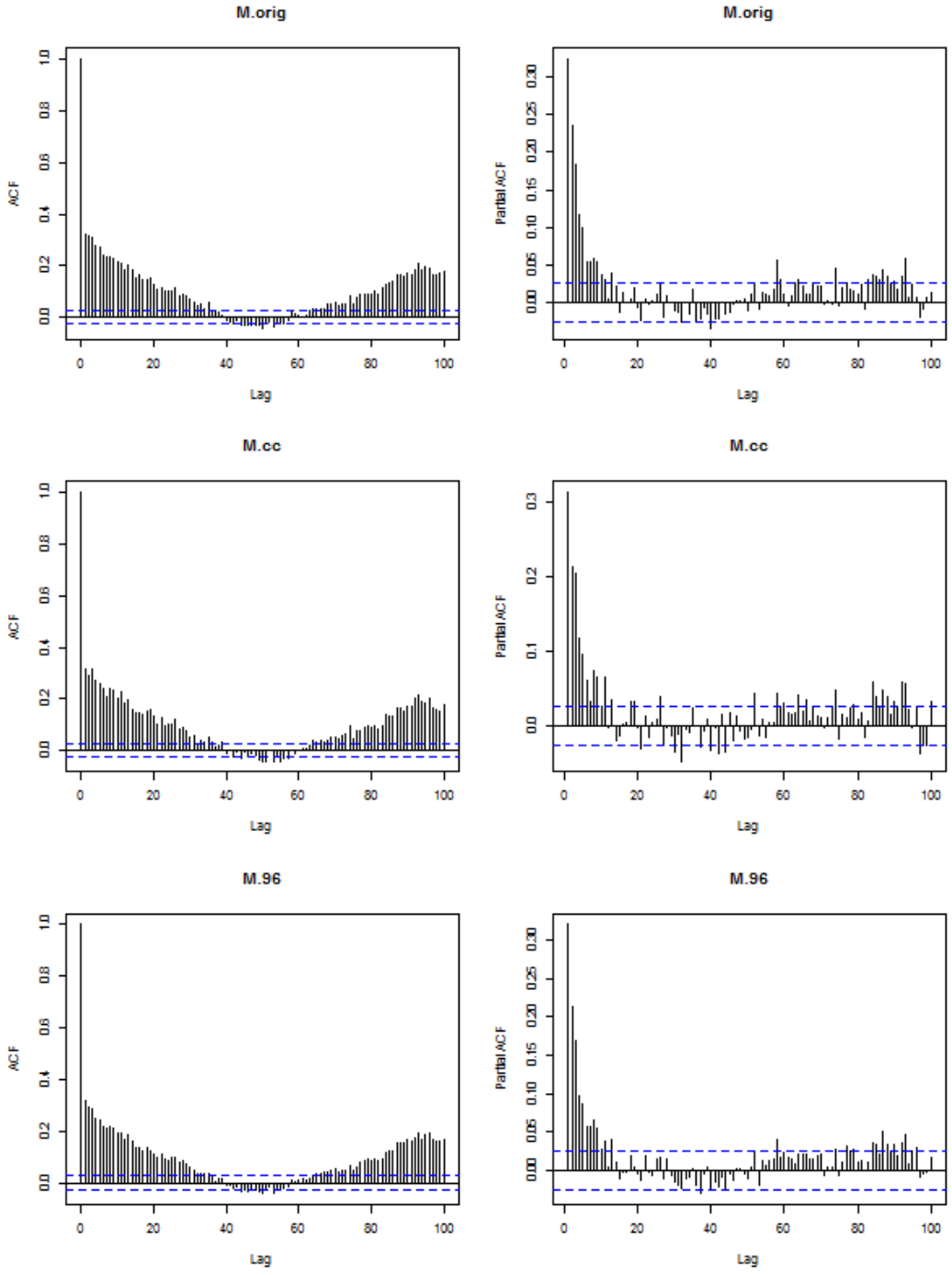

Abbildung 5.12.: ACF und PACF für den Link A aus den drei Modellen M.orig, M.cc und M.2

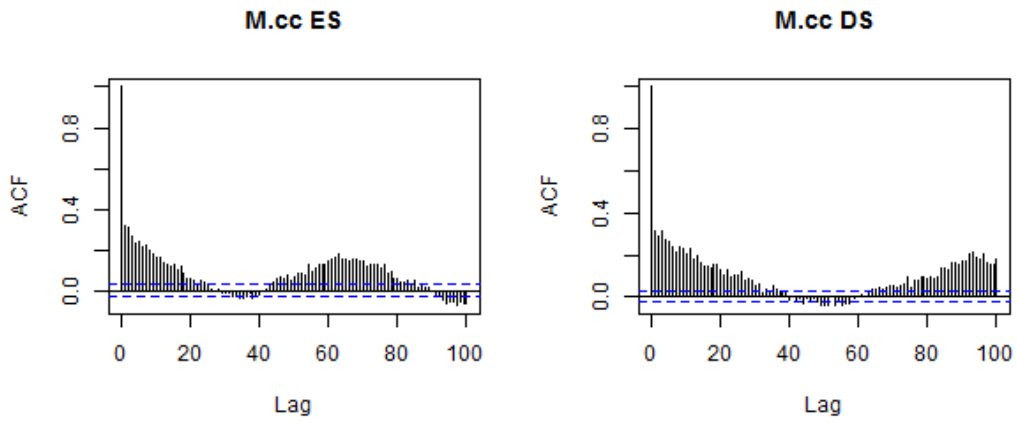

Abbildung 5.13.: ACF für den Link A aus dem Modell M.cc; berechnet mithilfe von ES und DS

6. Conclusion

Die Erläuterung des Problems "Fehlende Daten" in Kapitel 1 diente der Hervorhebung der Komplexität des Themas, sowie dass jeder Datensatz singulär zu bearbeiten ist, also im Allgemeinen für zwei verschiedene Missing-Data Probleme nicht gleich verfahren werden kann. Anschließend wurden in Kapitel 2 erste Methoden vorgestellt, mit welchen die Problemstellung behandelt wurde. Es wurde nicht nur präsentiert wie man die Datenanalyse auf beobachtete Daten beschränkt, sondern schon darauf eingegangen, mit Imputationsmethoden den aus dem Datensatz gewonnenen Informationsgehalt zu erhöhen.

In Kapitel 3 und 4 sind weiterentwickelte Methoden der zuvor diskutierten Ansätze präsentiert worden. Der Expectation Maximization Algorithmus sowie Data Augmentation und Multiple Imputation wurden eingehend besprochen. Darüber hinaus wurde auf den Gibbs Sampler eingegangen, welcher in Multiple Imputation by Chained Equations Anwendung findet. Neben modellbasierten Berechnungen wurde auch das Hot Deck Verfahren, Predictive Mean Matching, dargebracht.

Einige der zuvor diskutierten Techniken finden in Kapitel 5.1 Anwendung. Aus den Berechnungen, basierend auf je 12500 simulierten Zufallsmatrizen für MCAR und MAR, wurde geschlossen, dass für die Schätzung des Mittelwertes und der Kovarianzmatrix DS der Methode ES vorzuziehen ist. Dies gilt unabhängig vom Missing-Data-Mechanismus, der sich in Abhängigkeit von der simulierten Kovarianz der Datenmatrix von MAR zu NMAR wandelt. Allerdings ist zu beachten, dass die mit DS geschätzte Kovarianzmatrix nicht positiv definit sein muss. Auch EM und MI wurde auf die selben Simulationen angewandt, woraus sich die Erkenntnisse ergeben, dass diese ES vorzuziehen sind. Dieser Vergleich zu DS ist allerdings abhängig vom Missing-Data-Mechanismus. Je weniger dieser den Bedingungen von MCAR genügt, desto eher ist MI dem Verfahren DS vorzuziehen. Die Schätzer aus dem Verfahren EM approximieren die tatsächlichen Parameter am besten, was nicht überraschen soll, da die Datenmatrizen normalverteilt simuliert wurden, und somit den Modellannahmen des EM Algorithmus genügen.
In Kapitel 5.2 wurden Verkehrsdaten einer Flotte von Taxis vorgestellt. Dieser unvollständig beobachtete Datensatz wurde mithilfe von Multipler Imputation bearbeitet. Mehrere Modellansätze wurden diskutiert und verglichen, darunter auch das Imputieren mithilfe von vergangenen Beobachtungen. Um die Güte der Imputationen evaluieren zu können wurden

zusätzliche Datenpunkte nach MCAR gelöscht. Daraus war zu erkennen, dass das Verfahren einerseits konvergente Iterationen, andererseits konsistente Imputationen lieferte, und somit die Datenanalysen basierend auf dem imputierten Datensatz jener aus dem, um die unvollständigen Beobachtungen reduzierten, Datensatz vorzuziehen ist.

Appendices

A. Tabellen zum Vergleich der Zufallsmatrizen, $\triangle_\theta$

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.08	0.12	0.18	0.29	0.50	0.08	0.09	0.10	0.12	0.13
1000	0.06	0.09	0.12	0.19	0.32	0.06	0.07	0.07	0.08	0.09
1500	0.05	0.07	0.11	0.18	0.27	0.05	0.05	0.06	0.06	0.07
2000	0.04	0.05	0.09	0.15	0.23	0.04	0.04	0.05	0.06	0.06
2500	0.04	0.05	0.08	0.12	0.19	0.04	0.04	0.05	0.05	0.06

Tabelle A.1.: $\Delta_\mu^{ES(N,\lambda,\rho)}$ und $\Delta_\mu^{DS(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus ES (links) und DS (rechts), MCAR, rho = 0.9

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.09	0.13	0.19	0.30	0.52	0.09	0.10	0.11	0.12	0.13
1000	0.06	0.09	0.13	0.20	0.33	0.06	0.07	0.07	0.08	0.09
1500	0.05	0.07	0.11	0.18	0.28	0.05	0.05	0.06	0.06	0.07
2000	0.04	0.06	0.09	0.16	0.24	0.04	0.04	0.05	0.06	0.06
2500	0.04	0.05	0.08	0.12	0.19	0.04	0.04	0.05	0.05	0.06

Tabelle A.2.: $\Delta_\mu^{ES(N,\lambda,\rho)}$ und $\Delta_\mu^{DS(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus ES (links) und DS (rechts), MCAR, rho = 0.8

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.09	0.13	0.19	0.30	0.53	0.09	0.10	0.11	0.12	0.14
1000	0.06	0.09	0.13	0.20	0.34	0.06	0.07	0.07	0.08	0.09
1500	0.05	0.07	0.11	0.18	0.29	0.05	0.05	0.06	0.06	0.07
2000	0.04	0.06	0.09	0.16	0.25	0.04	0.05	0.05	0.06	0.06
2500	0.04	0.06	0.08	0.12	0.20	0.04	0.04	0.05	0.05	0.06

Tabelle A.3.: $\Delta_\mu^{ES(N,\lambda,\rho)}$ und $\Delta_\mu^{DS(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus ES (links) und DS (rechts), MCAR, rho = 0.7

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.09	0.13	0.19	0.31	0.54	0.09	0.10	0.11	0.12	0.13
1000	0.07	0.09	0.13	0.21	0.35	0.07	0.07	0.08	0.08	0.09
1500	0.05	0.07	0.11	0.18	0.29	0.05	0.05	0.06	0.07	0.07
2000	0.04	0.06	0.10	0.16	0.25	0.04	0.05	0.05	0.06	0.06
2500	0.04	0.06	0.08	0.13	0.20	0.04	0.04	0.05	0.05	0.06

Tabelle A.4.: $\Delta_\mu^{ES(N,\lambda,\rho)}$ und $\Delta_\mu^{DS(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus ES (links) und DS (rechts), MCAR, rho = 0.6

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.09	0.13	0.20	0.31	0.54	0.09	0.10	0.11	0.12	0.13
1000	0.07	0.09	0.13	0.21	0.36	0.07	0.07	0.08	0.08	0.09
1500	0.05	0.07	0.11	0.18	0.30	0.05	0.06	0.06	0.07	0.07
2000	0.04	0.06	0.10	0.16	0.26	0.04	0.05	0.05	0.06	0.06
2500	0.04	0.06	0.08	0.13	0.21	0.04	0.04	0.05	0.05	0.06

Tabelle A.5.: $\Delta_\mu^{ES(N,\lambda,\rho)}$ und $\Delta_\mu^{DS(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus ES (links) und DS (rechts), MCAR, rho = 0.5

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.26	0.37	0.51	0.76	1.48	0.26	0.30	0.36	0.44	0.54
1000	0.20	0.28	0.41	0.60	1.08	0.20	0.23	0.26	0.31	0.39
1500	0.15	0.22	0.34	0.49	0.94	0.15	0.18	0.22	0.26	0.32
2000	0.13	0.19	0.29	0.43	0.82	0.13	0.16	0.18	0.22	0.26
2500	0.12	0.18	0.25	0.40	0.67	0.12	0.14	0.17	0.20	0.24

Tabelle A.6.: $\Delta_\Sigma^{ES(N,\lambda,\rho)}$ und $\Delta_\Sigma^{DS(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus ES (links) und DS (rechts), MCAR, rho = 0.9

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.25	0.36	0.51	0.77	1.48	0.25	0.29	0.34	0.42	0.52
1000	0.19	0.27	0.40	0.60	1.07	0.19	0.22	0.25	0.30	0.37
1500	0.15	0.21	0.33	0.49	0.91	0.15	0.17	0.20	0.25	0.30
2000	0.13	0.19	0.28	0.42	0.80	0.13	0.15	0.18	0.21	0.25
2500	0.12	0.17	0.25	0.39	0.66	0.12	0.14	0.16	0.19	0.23

Tabelle A.7.: $\Delta_\Sigma^{ES(N,\lambda,\rho)}$ und $\Delta_\Sigma^{DS(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus ES (links) und DS (rechts), MCAR, rho = 0.8

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.25	0.35	0.50	0.77	1.48	0.25	0.29	0.33	0.41	0.50
1000	0.19	0.27	0.39	0.60	1.06	0.19	0.21	0.24	0.28	0.36
1500	0.15	0.21	0.32	0.48	0.89	0.15	0.16	0.20	0.24	0.29
2000	0.13	0.18	0.27	0.41	0.78	0.13	0.15	0.17	0.20	0.24
2500	0.11	0.17	0.24	0.38	0.65	0.11	0.13	0.16	0.18	0.22

Tabelle A.8.: $\Delta_\Sigma^{ES(N,\lambda,\rho)}$ und $\Delta_\Sigma^{DS(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus ES (links) und DS (rechts), MCAR, rho = 0.7

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.25	0.35	0.50	0.78	1.47	0.25	0.29	0.32	0.39	0.47
1000	0.18	0.26	0.39	0.59	1.05	0.18	0.21	0.24	0.27	0.34
1500	0.14	0.20	0.31	0.48	0.87	0.14	0.16	0.19	0.22	0.28
2000	0.13	0.18	0.27	0.41	0.77	0.13	0.15	0.17	0.20	0.24
2500	0.11	0.16	0.24	0.37	0.64	0.11	0.13	0.15	0.18	0.21

Tabelle A.9.: $\Delta_\Sigma^{ES(N,\lambda,\rho)}$ und $\Delta_\Sigma^{DS(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus ES (links) und DS (rechts), MCAR, rho = 0.6

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.25	0.34	0.50	0.78	1.46	0.25	0.28	0.32	0.38	0.46
1000	0.18	0.25	0.38	0.59	1.04	0.18	0.20	0.23	0.27	0.33
1500	0.14	0.20	0.30	0.47	0.85	0.14	0.16	0.18	0.22	0.27
2000	0.12	0.17	0.26	0.40	0.75	0.12	0.14	0.16	0.19	0.23
2500	0.11	0.16	0.23	0.36	0.63	0.11	0.13	0.15	0.17	0.21

Tabelle A.10.: $\Delta_\Sigma^{ES(N,\lambda,\rho)}$ und $\Delta_\Sigma^{DS(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus ES (links) und DS (rechts), MCAR, rho = 0.5

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.06	0.86	1.46	1.95	2.36	0.06	0.47	0.77	1.09	1.42
1000	0.03	0.81	1.38	1.85	2.33	0.03	0.42	0.73	1.08	1.42
1500	0.02	0.83	1.35	1.88	2.34	0.02	0.40	0.73	1.05	1.41
2000	0.03	0.79	1.38	1.87	2.33	0.03	0.40	0.74	1.06	1.43
2500	0.02	0.80	1.35	1.87	2.35	0.02	0.41	0.75	1.08	1.43

Tabelle A.11.: $\Delta_\mu^{ES(N,\lambda,\rho)}$ und $\Delta_\mu^{DS(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus ES (links) und DS (rechts), MAR, rho = 0.9

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.06	0.89	1.48	1.97	2.46	0.06	0.37	0.61	0.84	1.05
1000	0.03	0.84	1.47	1.89	2.37	0.03	0.31	0.55	0.82	1.09
1500	0.02	0.86	1.40	1.93	2.41	0.02	0.30	0.55	0.79	1.07
2000	0.03	0.85	1.44	1.93	2.40	0.03	0.30	0.56	0.81	1.08
2500	0.02	0.86	1.42	1.94	2.41	0.02	0.32	0.57	0.83	1.10

Tabelle A.12.: $\Delta_\mu^{ES(N,\lambda,\rho)}$ und $\Delta_\mu^{DS(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus ES (links) und DS (rechts), MAR, rho = 0.8

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.06	0.90	1.49	1.95	2.41	0.06	0.29	0.48	0.62	0.76
1000	0.03	0.85	1.44	1.87	2.35	0.03	0.23	0.41	0.59	0.80
1500	0.02	0.86	1.40	1.92	2.41	0.02	0.21	0.40	0.58	0.78
2000	0.03	0.88	1.43	1.92	2.36	0.03	0.21	0.41	0.59	0.80
2500	0.02	0.86	1.42	1.92	2.41	0.02	0.23	0.42	0.61	0.81

Tabelle A.13.: $\Delta_\mu^{ES(N,\lambda,\rho)}$ und $\Delta_\mu^{DS(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus ES (links) und DS (rechts), MAR, rho = 0.7

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.06	0.87	1.43	1.95	2.38	0.06	0.21	0.36	0.44	0.60
1000	0.03	0.82	1.41	1.81	2.30	0.03	0.14	0.28	0.41	0.53
1500	0.02	0.82	1.34	1.90	2.41	0.02	0.15	0.29	0.40	0.55
2000	0.03	0.82	1.38	1.84	2.32	0.03	0.15	0.28	0.44	0.57
2500	0.02	0.83	1.35	1.90	2.37	0.02	0.17	0.30	0.44	0.58

Tabelle A.14.: $\Delta_\mu^{ES(N,\lambda,\rho)}$ und $\Delta_\mu^{DS(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus ES (links) und DS (rechts), MAR, rho = 0.6

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.06	0.82	1.37	1.94	2.28	0.06	0.17	0.27	0.30	0.43
1000	0.03	0.77	1.32	1.78	2.28	0.03	0.09	0.18	0.27	0.34
1500	0.02	0.77	1.30	1.87	2.31	0.02	0.09	0.20	0.29	0.38
2000	0.03	0.79	1.34	1.78	2.24	0.03	0.09	0.18	0.30	0.39
2500	0.02	0.77	1.30	1.83	2.34	0.02	0.11	0.20	0.30	0.39

Tabelle A.15.: $\Delta_\mu^{ES(N,\lambda,\rho)}$ und $\Delta_\mu^{DS(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus ES (links) und DS (rechts), MAR, rho = 0.5

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.16	1.73	2.21	2.61	2.77	0.16	1.28	1.64	1.96	2.22
1000	0.31	1.66	2.22	2.63	2.87	0.31	1.19	1.65	2.02	2.29
1500	0.24	1.69	2.18	2.59	2.92	0.24	1.16	1.65	1.96	2.27
2000	0.33	1.62	2.23	2.62	2.91	0.33	1.15	1.64	1.97	2.28
2500	0.33	1.62	2.19	2.61	2.91	0.33	1.15	1.64	1.99	2.29

Tabelle A.16.: $\Delta_\Sigma^{ES(N,\lambda,\rho)}$ und $\Delta_\Sigma^{DS(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus ES (links) und DS (rechts), MAR, rho = 0.9

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.20	1.49	1.90	2.21	2.36	0.20	0.97	1.20	1.39	1.55
1000	0.28	1.44	1.98	2.24	2.48	0.28	0.84	1.17	1.42	1.66
1500	0.23	1.47	1.88	2.24	2.50	0.23	0.82	1.19	1.41	1.63
2000	0.30	1.42	1.96	2.29	2.50	0.30	0.79	1.16	1.39	1.59
2500	0.28	1.44	1.93	2.27	2.50	0.28	0.82	1.15	1.41	1.66

Tabelle A.17.: $\Delta_\Sigma^{ES(N,\lambda,\rho)}$ und $\Delta_\Sigma^{DS(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus ES (links) und DS (rechts), MAR, rho = 0.8

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.23	1.27	1.62	1.86	2.01	0.23	0.69	0.85	0.97	1.11
1000	0.26	1.25	1.70	1.95	2.11	0.26	0.60	0.84	0.98	1.13
1500	0.22	1.25	1.61	1.96	2.15	0.22	0.57	0.82	0.99	1.15
2000	0.28	1.25	1.69	1.98	2.18	0.28	0.53	0.80	0.94	1.10
2500	0.25	1.24	1.66	1.96	2.19	0.25	0.54	0.79	0.98	1.15

Tabelle A.18.: $\Delta_\Sigma^{ES(N,\lambda,\rho)}$ und $\Delta_\Sigma^{DS(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus ES (links) und DS (rechts), MAR, rho = 0.7

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.25	1.04	1.36	1.64	1.82	0.25	0.47	0.62	0.65	0.78
1000	0.25	1.05	1.48	1.73	1.90	0.25	0.39	0.56	0.71	0.77
1500	0.21	1.06	1.39	1.70	1.95	0.21	0.41	0.56	0.68	0.79
2000	0.26	1.05	1.47	1.73	1.87	0.26	0.42	0.55	0.68	0.76
2500	0.22	1.08	1.42	1.71	1.93	0.22	0.37	0.51	0.64	0.75

Tabelle A.19.: $\Delta_\Sigma^{ES(N,\lambda,\rho)}$ und $\Delta_\Sigma^{DS(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus ES (links) und DS (rechts), MAR, rho = 0.6

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.26	0.90	1.24	1.47	1.60	0.26	0.36	0.42	0.42	0.52
1000	0.24	0.93	1.34	1.55	1.71	0.24	0.30	0.43	0.48	0.54
1500	0.21	0.91	1.22	1.53	1.69	0.21	0.31	0.40	0.49	0.57
2000	0.24	0.92	1.32	1.53	1.64	0.24	0.30	0.37	0.45	0.53
2500	0.20	0.93	1.25	1.53	1.75	0.20	0.23	0.30	0.41	0.46

Tabelle A.20.: $\Delta_\Sigma^{ES(N,\lambda,\rho)}$ und $\Delta_\Sigma^{DS(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus ES (links) und DS (rechts), MAR, rho = 0.5

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.08	0.09	0.09	0.10	0.10		0.09	0.09	0.10	0.11
1000	0.06	0.06	0.07	0.07	0.07		0.06	0.07	0.07	0.08
1500	0.05	0.05	0.05	0.05	0.06		0.05	0.05	0.05	0.06
2000	0.04	0.04	0.04	0.04	0.05		0.04	0.04	0.05	0.05
2500	0.04	0.04	0.04	0.04	0.04		0.04	0.04	0.04	0.04

Tabelle A.21.: $\Delta_\mu^{EM(N,\lambda,\rho)}$ und $\Delta_\mu^{MI(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus EM (links) und MI (rechts), MCAR, rho = 0.9

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.09	0.09	0.09	0.10	0.11		0.09	0.10	0.11	0.12
1000	0.06	0.07	0.07	0.07	0.08		0.07	0.07	0.07	0.08
1500	0.05	0.05	0.05	0.06	0.06		0.05	0.06	0.06	0.06
2000	0.04	0.04	0.04	0.05	0.05		0.04	0.04	0.05	0.05
2500	0.04	0.04	0.04	0.04	0.05		0.04	0.04	0.04	0.05

Tabelle A.22.: $\Delta_\mu^{EM(N,\lambda,\rho)}$ und $\Delta_\mu^{MI(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus EM (links) und MI (rechts), MCAR, rho = 0.8

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.09	0.09	0.10	0.11	0.12		0.09	0.10	0.11	0.12
1000	0.06	0.07	0.07	0.07	0.08		0.07	0.07	0.08	0.09
1500	0.05	0.05	0.06	0.06	0.06		0.05	0.06	0.06	0.07
2000	0.04	0.04	0.05	0.05	0.06		0.04	0.05	0.05	0.06
2500	0.04	0.04	0.04	0.04	0.05		0.04	0.04	0.05	0.05

Tabelle A.23.: $\Delta_\mu^{EM(N,\lambda,\rho)}$ und $\Delta_\mu^{MI(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus EM (links) und MI (rechts), MCAR, rho = 0.7

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.09	0.09	0.10	0.11	0.12		0.09	0.10	0.11	0.13
1000	0.07	0.07	0.07	0.08	0.09		0.07	0.07	0.08	0.10
1500	0.05	0.05	0.06	0.06	0.07		0.05	0.06	0.06	0.07
2000	0.04	0.04	0.05	0.05	0.06		0.05	0.05	0.05	0.06
2500	0.04	0.04	0.04	0.05	0.05		0.04	0.04	0.05	0.06

Tabelle A.24.: $\Delta_\mu^{EM(N,\lambda,\rho)}$ und $\Delta_\mu^{MI(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus EM (links) und MI (rechts), MCAR, rho = 0.6

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.09	0.09	0.10	0.11	0.13		0.10	0.11	0.12	0.13
1000	0.07	0.07	0.07	0.08	0.09		0.07	0.08	0.08	0.09
1500	0.05	0.05	0.06	0.07	0.07		0.05	0.06	0.07	0.07
2000	0.04	0.05	0.05	0.06	0.06		0.05	0.05	0.06	0.07
2500	0.04	0.04	0.04	0.05	0.05		0.04	0.04	0.05	0.06

Tabelle A.25.: $\Delta_\mu^{EM(N,\lambda,\rho)}$ und $\Delta_\mu^{MI(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus EM (links) und MI (rechts), MCAR, rho = 0.5

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.26	0.26	0.26	0.28	0.30		0.27	0.30	0.35	0.45
1000	0.20	0.20	0.20	0.21	0.23		0.21	0.24	0.30	0.38
1500	0.15	0.15	0.16	0.16	0.18		0.16	0.19	0.25	0.35
2000	0.13	0.13	0.13	0.14	0.15		0.15	0.18	0.24	0.34
2500	0.12	0.12	0.12	0.13	0.14		0.13	0.17	0.24	0.34

Tabelle A.26.: $\Delta_\Sigma^{EM(N,\lambda,\rho)}$ und $\Delta_\Sigma^{MI(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus EM (links) und MI (rechts), MCAR, rho = 0.9

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.25	0.26	0.27	0.30	0.33		0.27	0.32	0.38	0.49
1000	0.19	0.20	0.20	0.22	0.25		0.21	0.25	0.32	0.41
1500	0.15	0.15	0.16	0.17	0.20		0.16	0.21	0.28	0.38
2000	0.13	0.13	0.14	0.15	0.17		0.15	0.20	0.27	0.37
2500	0.12	0.12	0.13	0.13	0.15		0.14	0.19	0.26	0.36

Tabelle A.27.: $\Delta_\Sigma^{EM(N,\lambda,\rho)}$ und $\Delta_\Sigma^{MI(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus EM (links) und MI (rechts), MCAR, rho = 0.8

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.25	0.26	0.28	0.32	0.37		0.27	0.32	0.40	0.51
1000	0.19	0.19	0.21	0.23	0.26		0.21	0.26	0.33	0.44
1500	0.15	0.16	0.17	0.18	0.21		0.17	0.22	0.29	0.40
2000	0.13	0.14	0.14	0.16	0.18		0.15	0.20	0.28	0.38
2500	0.11	0.12	0.13	0.14	0.17		0.14	0.20	0.28	0.38

Tabelle A.28.: $\Delta_\Sigma^{EM(N,\lambda,\rho)}$ und $\Delta_\Sigma^{MI(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus EM (links) und MI (rechts), MCAR, rho = 0.7

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.25	0.26	0.29	0.33	0.39		0.28	0.33	0.41	0.52
1000	0.18	0.19	0.21	0.23	0.28		0.20	0.26	0.34	0.44
1500	0.14	0.16	0.17	0.19	0.22		0.17	0.22	0.30	0.40
2000	0.13	0.14	0.15	0.16	0.19		0.16	0.21	0.28	0.39
2500	0.11	0.12	0.13	0.15	0.17		0.14	0.20	0.28	0.38

Tabelle A.29.: $\Delta_\Sigma^{EM(N,\lambda,\rho)}$ und $\Delta_\Sigma^{MI(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus EM (links) und MI (rechts), MCAR, rho = 0.6

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.25	0.27	0.29	0.33	0.39		0.29	0.34	0.43	0.53
1000	0.18	0.19	0.21	0.24	0.29		0.21	0.26	0.34	0.44
1500	0.14	0.15	0.16	0.19	0.22		0.16	0.22	0.30	0.40
2000	0.12	0.14	0.15	0.17	0.20		0.15	0.21	0.29	0.39
2500	0.11	0.12	0.14	0.15	0.18		0.14	0.20	0.28	0.38

Tabelle A.30.: $\Delta_\Sigma^{EM(N,\lambda,\rho)}$ und $\Delta_\Sigma^{MI(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus EM (links) und MI (rechts), MCAR, rho = 0.5

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.01	0.01	0.01	0.01	0.02		2.58	3.76	4.78	5.58
1000	0.01	0.01	0.01	0.02	0.03		6.22	9.85	12.60	14.41
1500	0.01	0.01	0.02	0.02	0.03		11.54	18.80	23.96	27.01
2000	0.01	0.02	0.02	0.03	0.04		18.72	30.81	39.19	43.78
2500	0.02	0.03	0.03	0.04	0.05		27.70	45.84	58.10	64.41

Tabelle A.31.: Vergleich der Prozessdurchlaufzeiten der Methoden EM (links) und MI (rechts), MCAR, rho = 0.9

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.00	0.01	0.01	0.01	0.02		2.59	3.75	4.78	5.60
1000	0.01	0.01	0.01	0.02	0.03		6.21	9.85	12.61	14.41
1500	0.01	0.02	0.02	0.02	0.04		11.54	18.83	23.94	27.00
2000	0.01	0.02	0.02	0.03	0.04		18.69	30.76	39.13	43.73
2500	0.02	0.02	0.03	0.04	0.05		27.71	45.83	58.06	64.40

Tabelle A.32.: Vergleich der Prozessdurchlaufzeiten der Methoden EM (links) und MI (rechts), MCAR, rho = 0.8

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.01	0.01	0.01	0.01	0.02		2.59	3.77	4.78	5.59
1000	0.01	0.01	0.01	0.02	0.03		6.20	9.85	12.59	14.39
1500	0.01	0.01	0.02	0.03	0.04		11.56	18.83	23.97	27.03
2000	0.02	0.02	0.02	0.03	0.04		18.70	30.85	39.21	43.77
2500	0.02	0.02	0.02	0.04	0.05		27.73	45.82	58.16	64.43

Tabelle A.33.: Vergleich der Prozessdurchlaufzeiten der Methoden EM (links) und MI (rechts), MCAR, rho = 0.7

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.01	0.01	0.01	0.01	0.02		2.58	3.76	4.77	5.59
1000	0.01	0.01	0.01	0.02	0.03		6.22	9.88	12.61	14.42
1500	0.01	0.01	0.02	0.02	0.03		11.57	18.84	23.98	27.07
2000	0.01	0.02	0.02	0.03	0.04		18.71	30.78	39.17	43.78
2500	0.02	0.02	0.02	0.04	0.06		27.70	45.86	58.10	64.44

Tabelle A.34.: Vergleich der Prozessdurchlaufzeiten der Methoden EM (links) und MI (rechts), MCAR, rho = 0.6

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.01	0.01	0.01	0.01	0.02		2.59	3.76	4.80	5.59
1000	0.01	0.01	0.01	0.02	0.03		6.22	9.85	12.60	14.40
1500	0.01	0.01	0.02	0.03	0.03		11.56	18.79	23.97	27.01
2000	0.01	0.02	0.02	0.03	0.04		18.70	30.77	39.22	43.79
2500	0.02	0.02	0.03	0.05	0.06		27.70	45.87	58.09	64.34

Tabelle A.35.: Vergleich der Prozessdurchlaufzeiten der Methoden EM (links) und MI (rechts), MCAR, rho = 0.5

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.08	0.17	0.38	0.63	0.91		0.19	0.41	0.65	0.94
1000	0.06	0.17	0.39	0.64	0.92		0.19	0.41	0.66	0.95
1500	0.05	0.17	0.39	0.64	0.92		0.18	0.40	0.66	0.94
2000	0.04	0.17	0.39	0.63	0.91		0.19	0.41	0.65	0.93
2500	0.04	0.18	0.39	0.64	0.92		0.19	0.41	0.66	0.94

Tabelle A.36.: $\Delta_\mu^{EM(N,\lambda,\rho)}$ und $\Delta_\mu^{MI(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus EM (links) und MI (rechts), MAR, rho = 0.9

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.09	0.11	0.21	0.37	0.57		0.11	0.23	0.39	0.60
1000	0.06	0.10	0.21	0.38	0.58		0.10	0.22	0.40	0.60
1500	0.05	0.09	0.21	0.38	0.58		0.09	0.22	0.39	0.60
2000	0.04	0.09	0.21	0.37	0.57		0.09	0.22	0.39	0.59
2500	0.04	0.09	0.22	0.38	0.58		0.10	0.22	0.39	0.60

Tabelle A.37.: $\Delta_\mu^{EM(N,\lambda,\rho)}$ und $\Delta_\mu^{MI(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus EM (links) und MI (rechts), MAR, rho = 0.8

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.09	0.10	0.14	0.23	0.36		0.10	0.15	0.25	0.39
1000	0.06	0.08	0.13	0.22	0.36		0.08	0.14	0.24	0.38
1500	0.05	0.06	0.12	0.22	0.36		0.07	0.13	0.24	0.38
2000	0.04	0.06	0.12	0.22	0.35		0.07	0.13	0.24	0.37
2500	0.04	0.06	0.12	0.22	0.36		0.07	0.14	0.24	0.38

Tabelle A.38.: $\Delta_\mu^{EM(N,\lambda,\rho)}$ und $\Delta_\mu^{MI(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus EM (links) und MI (rechts), MAR, rho = 0.7

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.09	0.10	0.12	0.16	0.23		0.10	0.13	0.18	0.26
1000	0.07	0.07	0.09	0.14	0.22		0.08	0.11	0.17	0.25
1500	0.05	0.06	0.08	0.13	0.21		0.07	0.10	0.16	0.24
2000	0.04	0.05	0.07	0.13	0.20		0.06	0.10	0.15	0.24
2500	0.04	0.05	0.08	0.13	0.21		0.06	0.10	0.16	0.25

Tabelle A.39.: $\Delta_\mu^{EM(N,\lambda,\rho)}$ und $\Delta_\mu^{MI(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus EM (links) und MI (rechts), MAR, rho = 0.6

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.09	0.10	0.11	0.14	0.18		0.10	0.13	0.16	0.21
1000	0.07	0.07	0.09	0.10	0.15		0.08	0.11	0.14	0.18
1500	0.05	0.06	0.07	0.09	0.14		0.07	0.10	0.13	0.18
2000	0.04	0.05	0.06	0.08	0.13		0.06	0.09	0.12	0.17
2500	0.04	0.05	0.06	0.08	0.13		0.06	0.09	0.12	0.17

Tabelle A.40.: $\Delta_\mu^{EM(N,\lambda,\rho)}$ und $\Delta_\mu^{MI(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus EM (links) und MI (rechts), MAR, rho = 0.5

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.26	0.56	0.97	1.32	1.62		0.70	1.17	1.57	1.91
1000	0.20	0.54	0.96	1.33	1.63		0.68	1.16	1.57	1.91
1500	0.15	0.54	0.96	1.33	1.63		0.67	1.15	1.56	1.89
2000	0.13	0.55	0.98	1.32	1.62		0.68	1.17	1.56	1.88
2500	0.12	0.56	0.97	1.32	1.62		0.68	1.16	1.55	1.88

Tabelle A.41.: $\Delta_\Sigma^{EM(N,\lambda,\rho)}$ und $\Delta_\Sigma^{MI(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus EM (links) und MI (rechts), MAR, rho = 0.9

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.25	0.34	0.51	0.70	0.89		0.40	0.62	0.87	1.12
1000	0.19	0.28	0.47	0.69	0.89		0.34	0.59	0.86	1.12
1500	0.15	0.26	0.46	0.68	0.88		0.31	0.58	0.85	1.09
2000	0.13	0.26	0.48	0.68	0.88		0.31	0.59	0.85	1.09
2500	0.12	0.25	0.47	0.68	0.87		0.31	0.58	0.85	1.09

Tabelle A.42.: $\Delta_\Sigma^{EM(N,\lambda,\rho)}$ und $\Delta_\Sigma^{MI(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus EM (links) und MI (rechts), MAR, rho = 0.8

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.25	0.29	0.35	0.45	0.56		0.33	0.45	0.60	0.76
1000	0.19	0.22	0.29	0.39	0.52		0.26	0.39	0.55	0.72
1500	0.15	0.18	0.26	0.37	0.50		0.23	0.37	0.53	0.70
2000	0.13	0.17	0.25	0.36	0.49		0.22	0.36	0.53	0.69
2500	0.11	0.16	0.25	0.36	0.48		0.21	0.36	0.54	0.71

Tabelle A.43.: $\Delta_\Sigma^{EM(N,\lambda,\rho)}$ und $\Delta_\Sigma^{MI(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus EM (links) und MI (rechts), MAR, rho = 0.7

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.25	0.29	0.33	0.40	0.50		0.33	0.42	0.54	0.66
1000	0.18	0.21	0.25	0.32	0.43		0.25	0.35	0.47	0.61
1500	0.14	0.17	0.21	0.29	0.40		0.22	0.32	0.45	0.59
2000	0.13	0.15	0.19	0.26	0.38		0.20	0.31	0.44	0.58
2500	0.11	0.13	0.18	0.26	0.38		0.19	0.31	0.44	0.59

Tabelle A.44.: $\Delta_\Sigma^{EM(N,\lambda,\rho)}$ und $\Delta_\Sigma^{MI(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus EM (links) und MI (rechts), MAR, rho = 0.6

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.25	0.28	0.33	0.41	0.55		0.32	0.42	0.53	0.68
1000	0.18	0.21	0.25	0.33	0.47		0.25	0.36	0.47	0.62
1500	0.14	0.17	0.21	0.30	0.44		0.22	0.33	0.46	0.61
2000	0.12	0.14	0.19	0.27	0.41		0.20	0.31	0.44	0.58
2500	0.11	0.13	0.17	0.26	0.40		0.19	0.29	0.43	0.57

Tabelle A.45.: $\Delta_\Sigma^{EM(N,\lambda,\rho)}$ und $\Delta_\Sigma^{MI(N,\lambda,\rho)}$: Vergleich der Parameterschätzfehler aus EM (links) und MI (rechts), MAR, rho = 0.5

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.01	0.01	0.01	0.01	0.02		2.55	3.79	4.78	5.65
1000	0.01	0.01	0.01	0.02	0.02		6.23	9.89	12.68	14.47
1500	0.01	0.02	0.02	0.02	0.03		11.65	18.87	24.12	27.11
2000	0.01	0.02	0.02	0.03	0.04		18.71	30.97	39.21	44.03
2500	0.02	0.02	0.03	0.04	0.05		27.82	45.90	58.30	64.50

Tabelle A.46.: Vergleich der Prozessdurchlaufzeiten der Methoden EM (links) und MI (rechts), MAR, rho = 0.9

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.00	0.01	0.01	0.01	0.02		2.56	3.77	4.79	5.65
1000	0.01	0.01	0.02	0.02	0.02		6.23	9.85	12.69	14.46
1500	0.01	0.01	0.02	0.02	0.03		11.66	18.87	24.09	27.05
2000	0.01	0.02	0.02	0.03	0.04		18.68	31.00	39.25	43.96
2500	0.02	0.03	0.03	0.04	0.05		27.77	45.92	58.35	64.57

Tabelle A.47.: Vergleich der Prozessdurchlaufzeiten der Methoden EM (links) und MI (rechts), MAR, rho = 0.8

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.01	0.01	0.01	0.01	0.02		2.55	3.79	4.78	5.65
1000	0.01	0.01	0.02	0.02	0.02		6.22	9.88	12.65	14.46
1500	0.01	0.02	0.02	0.02	0.03		11.68	18.86	24.06	27.05
2000	0.01	0.02	0.02	0.03	0.04		18.71	30.96	39.16	43.99
2500	0.02	0.03	0.03	0.04	0.05		27.75	45.96	58.28	64.58

Tabelle A.48.: Vergleich der Prozessdurchlaufzeiten der Methoden EM (links) und MI (rechts), MAR, rho = 0.7

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.01	0.01	0.01	0.02	0.02		2.56	3.79	4.78	5.64
1000	0.01	0.01	0.01	0.02	0.03		6.24	9.88	12.65	14.47
1500	0.01	0.02	0.02	0.03	0.03		11.66	18.87	24.04	27.06
2000	0.01	0.02	0.03	0.03	0.04		18.72	31.00	39.23	44.06
2500	0.02	0.02	0.03	0.04	0.05		27.76	45.97	58.31	64.56

Tabelle A.49.: Vergleich der Prozessdurchlaufzeiten der Methoden EM (links) und MI (rechts), MAR, rho = 0.6

	0	0.125	0.25	0.375	0.5	0	0.125	0.25	0.375	0.5
500	0.01	0.01	0.01	0.01	0.02		2.56	3.78	4.81	5.64
1000	0.01	0.01	0.02	0.02	0.03		6.24	9.88	12.64	14.49
1500	0.01	0.02	0.02	0.03	0.03		11.68	18.90	24.05	27.09
2000	0.01	0.02	0.03	0.03	0.05		18.71	30.97	39.18	44.02
2500	0.02	0.02	0.03	0.04	0.05		27.81	45.91	58.32	64.60

Tabelle A.50.: Vergleich der Prozessdurchlaufzeiten der Methoden EM (links) und MI (rechts), MAR, rho = 0.5

Literaturverzeichnis

[1] Allison P. D.: Missing Data, *Sage University Papers Series on Quantitative Applications in the Social Sciences. 07-136. Thousand Oaks, CA: Sage.*, 2001

[2] Alvaro A. N., Original by Joseph L. Schafer: Analysis of multivariate normal datasets with missing values, *Package "norm"*, 2013

[3] Brand J. P. L.: Developement, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets, *Academic Thesis, Erasmus University, Rotterdam*, 1998

[4] Glasser M.: Linear regression analysis wirh missing observations among the independent variables., *Journal of the American Statistical Association*, 59, 834-844, 1964

[5] Haitovsky Y.: Missing data in regression analysis, *Journal of the Royal Statistical Society, Serie B*, 30, 67-82, 1986

[6] Kim J. O., Curry J.: The treatment of missing data in multivariate analysis, *Sociological Methods & Research*, 6, 215-240, 1977

[7] Little Roderick J. A.: Missing-Data Adjustments in Large Surveys, *Journal of Business & Economic Statistics*, 2000

[8] Little Roderick J. A., Rubin Donald B.: Statistical Analysis with Missing Data (2nd edn), *Wiley: Hoboken, New Jersey,* 2002

[9] Rubin D. B.: Multiple Imputation for Nonresponse in Surveys, *Wiley: New York,* 1987

[10] Toplak W., Koller H., Dragaschnig M., Bauer D., Asamer J.: Novel Road Classifications for Large Scale Traffic Networks, *in Proceedings of the 13th International IEEE Conference on Intelligent Transportation Systems, Madeira*, 19.09.2010 - 22.09.2010, Paper-Nr. 450

[11] van Buuren S., Oudshoorn C.G.M.: Multivariate Imputation by Chained Equations, *MICE V1.0 User's manual*, 2000

[12] van Buuren S., Oudshoorn C.G.M.: Multivariate Imputation by Chained Equations, *Journal of Statistical Software*, 2011

[13] van Buuren S., Boshuizen H.C., Knook D.L.: Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis, *Statistics in Medicine*, 1999

[14] White I. R., Royston P.: Imputing missing covariate values for the Cox model, *Statistics in Medicine*, 2009