Random-Forest-Regression bei fehlenden Werten in den Einflussgrößen

Bachelor-Thesis

im Studiengang Statistik

Anna Rieger

Gauting, Juli 2008

Betreuer: Prof. Dr. T. Hothorn

Institut für Statistik
Ludwig-Maximilians-Universität München

Inhaltsverzeichnis

Abbildungsverzeichnis

1. Einführung

1.1. Einleitung

In statistischen Analysen versucht man, Beziehungen und ihre Stärke zwischen verschiedenen möglichen Einflussgrößen („Prädiktoren“) und einer (oder mehreren) davon abhängigen Variable(n), dem Response, herauszufinden. Dieses Verfahren nennt man allgemein „Regression“. Damit möchte man die vorliegenden Daten (die Menge der Variablen) entweder beschreiben oder andere Daten mit ihrer Hilfe vorhersagen. Um solche Regressionsprobleme zu lösen, stehen viele verschiedene Methoden zur Verfügung. Der Zusammenhang kann linear modelliert werden (mit einem als normal verteilt angenommenen Response) oder verallgemeinert werden („generalisierte lineare Modelle“; mit anderen Verteilungen, z. B. binomial verteilt).
Neben diesen Lösungen stehen so genannte „nichtparametrische Verfahren“ zur Verfügung, welche keine Verteilungsannahme an den Response stellen. In der vorliegenden Arbeit wurde die Vorgehensweise der *Classification and Regression Trees* (CART) verwendet, welche verteilungsfrei ist. Sie stellt die Daten in einer auch für Laien gut nachvollziehbaren Baumform dar (siehe z. B. Abbildung 2.1.4). Die Daten werden durch den Algorithmus der CART, der von Breiman u. a. [1984] entwickelt wurde, nach und nach in jeweils zwei Gruppen („binär“) gespalten.

Das erweiterte Verfahren der *Random Forests* von Breiman [2001] erstellt aus jeweils einer Teilmenge der ursprünglichen Daten einen Baum und mittelt die Ergebnisse. Da sich im Laufe der Zeit einige Nachteile dieser Methode herausgestellt haben, aber die Idee an sich gut ist, haben Hothorn u. a. [2006] die Methode *Conditional Tree Forests* entwickelt, die die Nachteile der Random Forests auslöscht. Sie gründet auf einer veränderten Berechnung der einzelnen Bäume, indem die Verteilung des Response – bedingt[1] auf die Prädiktoren, d. h. abhängig von ihnen – geschätzt wird.

Falls in den Daten einzelne Werte fehlen, ergeben sich mehr oder weniger große Probleme. Viele Verfahren können mit lückenhaften Datensätzen nicht umgehen. Die Lösung dafür stellen verschiedene Methoden dar, um die fehlenden Werte zu ersetzen. Andere Verfahren können selbst die fehlenden Werte mehr oder weniger gut handhaben. In dieser Arbeit soll untersucht werden, ob das Verfahren der Conditional Tree Forests damit gut umgehen kann. Außerdem wird verglichen, ob mit Hilfe von Imputation, d. h. der oben erwähnten Ersetzung der fehlenden Werte, bessere

[1] englisch: conditional

Ergebnisse erzielt werden können. Um den Rahmen nicht zu sprengen, wurde nur auf eine Imputationsmethode zurückgegriffen: den *knn*-Impute.

Es wurden Datensätze simuliert und anschließend auf verschiedene Arten fehlende Werte eingestreut. Dadurch sind die „richtigen“ Werte bekannt und man kann berechnen, wie groß das Risiko einer Fehleinschätzung des Response durch den Conditional Tree Forest ist. Dieses Risiko wurde an Hand von Lern- und Test-Datensätzen berechnet.

Die Simulation und die Berechnung der Ergebnisse erfolgte mit dem statistischen Programmpaket R [R Development Core Team, 2008] (siehe hierzu A).

1.2. Übersicht

In Kapitel 2 wird zunächst eine theoretische Einführung in die Methodik der Klassifikations- und Regressionsbäume[2] gegeben. Daran anschließend findet sich die Theorie zu den Random Forests und den bedingten Varianten der Bäume und *Forests*[3].

In Kapitel 3 werden die verschiedenen Mechanismen zur Einstreuung von fehlenden Werten vorgestellt. Das theoretische Vorgehen der Imputationsmethode *knn*-Impute wird ebenfalls hier erläutert.

In Kapitel 4 wird das verwendete Simulationsdesign erklärt und in Kapitel 5 werden die Ergebnisse dargestellt und ein Fazit aus den gewonnenen Erkenntnissen gezogen.

[2] englisch: Classification and Regression Trees (CART)
[3] englisch für „Wald“

2. Trees und Forests

2.1. Klassifikations- und Regressionsbäume

Die Methodik der *Classification and Regression Trees* (CART) wurde von Breiman u. a. [1984] entwickelt und basiert darauf, eine Menge von möglichen Prädiktoren (Einflussgrößen) $\boldsymbol{X}$ in binäre, disjunkte Teilmengen zu teilen, d. h. zwei Teilmengen $\boldsymbol{X}_1$ und $\boldsymbol{X}_2$ zu finden, so dass $\boldsymbol{X}_1 \cup \boldsymbol{X}_2 = \boldsymbol{X}$ und $\boldsymbol{X}_1 \cap \boldsymbol{X}_2 = \emptyset$. Diese Untermengen sollten in sich möglichst homogen und untereinander möglichst heterogen sein (bezüglich des Response $\boldsymbol{y}$) [Fahrmeir u. a., 1996, S. 425]. Ist der Response kategorial, spricht man von einem „Klassifikationsbaum“. Liegt dagegen ein stetiger Response vor, handelt es sich um einen „Regressionsbaum“.

Diese Partitionierung wiederholt sich rekursiv, indem man mit den entstandenen Teilmengen ebenso verfährt wie mit der ursprünglichen Menge: Man versucht, sie ebenfalls in zwei disjunkte Teilmengen zu zerlegen.
Durch diese Prozedur erhält man einen Baum, wie z. B. in Abbildung 2.1.4. Die zu teilenden Mengen werden auch „(Mutter-)Knoten“ genannt, die resultierenden Teilmengen „Tochter-Knoten“.

2.1.1. Trennpunkte

Jede Teilung eines Knotens (Split) hängt von einer einzigen Einflussgröße ab. Für eine ordinal skalierte Einflussgröße $\boldsymbol{x}$ ergibt sich eine Trennung durch $x_i \leq s; i = 1, \ldots, n; s \in \mathbb{R}$. Kategoriale Einflussgrößen $\boldsymbol{x}$ kann man trennen durch $x_i \in K$ mit einer Teilmenge K aller vorhandenen Kategorien. Somit ergeben sich n mögliche Trennpunkte bei ordinalen und $2^{L-1} - 1$ mögliche Trennpunkte bei kategorialen Einflussgrößen mit L Kategorien [Breiman u. a., 1984, S. 30]. Zusätzlich sind Linearkombinationen und Boole'sche Kombinationen denkbar.

Eine Trennung des Knotens erfolgt dadurch, dass diejenige Beobachtungen, die „wahr“ auf den Trennpunkt antworten, in den linken Tochterknoten wandern und diejenigen, für die der Trennpunkt „falsch“ auswertet, in den rechten.

2.1.2. Trennregel

In jedem Schritt des CART-Algorithmus wird dabei diejenige Einflussgröße gesucht, welche den Knoten am besten trennen kann. Dazu wird eine „Unreinheit“ des Knotens definiert: Sie sollte am größten sein, wenn alle Kategorien gleich stark vertreten sind und am kleinsten, wenn nur noch eine Klasse enthalten ist [Breiman u. a., 1984, S. 24]. Zur Teilung eines Knotens (Split) wird dann die Abnahme der „Unreinheit“ maximiert. Falls sich keine signifikante Abnahme ergibt, wird der bestehende Knoten ein Endknoten [Breiman u. a., 1984, S. 26].
Eine bekannte Unreinheitsmessung für Klassifikationsprobleme stellt der Gini-Index dar. Allerdings hat sich gezeigt, dass dieses Unreinheitskriterium stark verzerrt ist [Strobl u. a., 2007, S. 4]. Es bevorzugt bei stetigen Einflussgrößen solche mit vielen fehlenden Werten und bei kategorialen Einflussgrößen solche mit vielen Kategorien [Svedjar, 2007, S. 6].

Einen Baum zu berechnen ist sehr zeitaufwändig, da in jedem Knoten für jede Einflussgröße der beste Split berechnet werden muss. Anschließend wird der beste der besten Splits ausgewählt. Die Güte eines Splits wird, wie oben erwähnt, über die Unreinheit gemessen.

2.1.3. Stoppregel

Die Entstehung eines Baumes endet, wenn z. B. die Abnahme der Unreinheit kleiner als ein vorher festgelegter Schwellenwert ist [Breiman u. a., 1984, S. 33], die Knoten weniger als eine gewisse Anzahl an Beobachtungen enthalten [Svedjar, 2007, S. 9] oder man benutzt die Kosten-Komplexitäts-Beschneidung, die den Baum zuerst sehr komplex werden lässt und ihn dann einer Beschneidung (*Pruning*) unterzieht [Fahrmeir u. a., 1996, S. 432] [Breiman u. a., 1984, S. 37].

Der Algorithmus ist zum Beispiel im R-Paket `rpart` implementiert.

2.1.4. Beispiel

Folgendes Beispiel für einen Klassifikationsbaum stammt von Fahrmeir u. a. [1996, S. 426]:
Es handelt sich dabei um die Einteilung von Bankkunden als kreditwürdig ($y = 1$) oder nicht ($y = 2$). Die Einflussgrößen sind z. B. die gewünschte Laufzeit und die gewünschte Höhe des Kredits, ob der Kunde bereits ein laufendes Konto (ja: 1, nein: 2) hat und ob er über eine gute Zahlungsmoral verfügt. Der erstellte Baum findet sich in Abbildung 2.1.4. Daraus ist ersichtlich, dass sich ein schlechter Kunde ($y = 2$) unter anderem dadurch auszeichnet, dass er kein laufendes Konto besitzt. Ein guter Kunde hat bereits ein laufendes Konto und z. B. entweder eine lange Kreditlaufzeit von über 22.5 Monaten oder er verfügt über eine gute Zahlungsmoral (Moral ≥ 1.5).

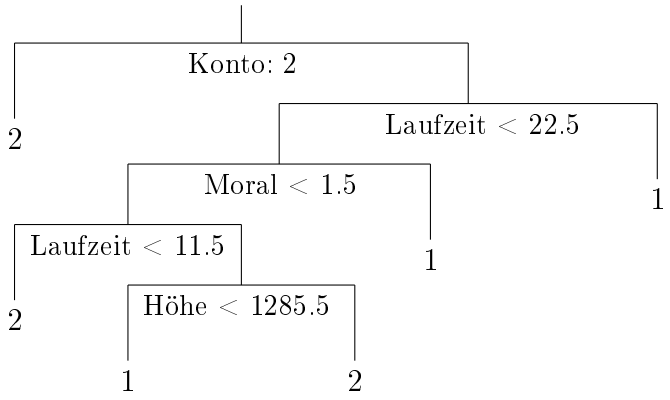

Abbildung 2.1.: Klassifikationsbaum für das Konto-Beispiel

2.2. Fehlende Werte und Surrogat-Variablen

Zwei Arten von fehlenden Werten können auftreten: Zum Einen kann der Lern-Datensatz fehlende Werte enthalten, was problematisch ist, denn mit den Daten aus dem Lern-Datensatz wird der Baum erstellt. Zum Anderen können im Test-Datensatz Werte fehlen, was die richtige Klassifikation bzw. Vorhersage erschwert. Diese Probleme kann man durch Verwenden von so genannten „Surrogat-Variablen“ minimieren.
Die Idee der Surrogat-Variablen besteht darin, eine „Ähnlichkeit“ zwischen zwei Trennpunkten (Splits) zu definieren. Falls in einer Einflussgröße $\boldsymbol{x}_i$ Werte fehlen, so wird aus allen anderen Einflussgrößen derjenige Split s_t gesucht, der dem Trennpunkt s_i in $\boldsymbol{x}_i$ am ähnlichsten ist [Breiman u. a., 1984, S. 40].

Im Basis-Algorithmus von Breiman u. a. [1984] ist die Ähnlichkeit wie folgt definiert: s^* ist der beste Trennpunkt an einem gegebenen Knoten. Ein beliebiger Trennpunkt s_t in $\boldsymbol{x}_i$ wird dann als Surrogat-Split in $\boldsymbol{x}_i$ bezeichnet, wenn die Wahrscheinlichkeit, dass der beliebige Split s_t den besten Split s^* richtig vorhersagt, maximal ist. Die Vorhersage-Wahrscheinlichkeit stellt dabei die Ähnlichkeit dar. Wenn also der Trennpunkt s_t im Vergleich zu anderen Trennpunkten am meisten Daten dem gleichen Tochter-Knoten zuordnet wie der beste Split s^*, ist der Trennpunkt s_t der Surrogat-Split von s^* [Breiman u. a., 1984, S. 140f.]. Dies ist z. B. häufig der Fall, wenn zwischen den beiden Einflussgrößen, zu denen die Trennpunkte s_t und s^* gehören, eine hohe Korrelation vorliegt.

Wenn im Lern-Datensatz fehlende Werte vorkommen, wird derjenige Surrogat-Split verwendet, der die Abnahme an Unreinheit maximiert. Falls im Test-Datensatz Werte fehlen und der beste Trennpunkt dadurch nicht definiert ist, wird ebenfalls der jeweils beste Surrogat-Split verwendet. Dieses Vorgehen ist analog zum Ersetzen eines fehlendes Wertes durch Regression auf die Einflussgröße ohne fehlende Werte

mit der höchsten Korrelation zur Einflussgröße mit dem fehlenden Wert [Breiman u. a., 1984, S. 142]. Dadurch sind hoch korrelierte Einflussgrößen gut geeignet für die Verwendung als Surrogat-Einflussgrößen. Die Methode der Surrogat-Einflussgrößen ist allerdings robuster als die Regressionsvariante.

Eine äquivalente Vorgehensweise wird von Hothorn u. a. [2006, S. 658] beschrieben: Der beste Trennpunkt s^* in der Einflussgröße mit den fehlenden Werten kann auch als Menge A^* aufgefasst werden. Die Daten werden danach geteilt, ob sie z. B. unterhalb dieses Trennpunktes liegen oder nicht ($A^* = \{x_i | x_i \leq s^*, i = 1, \ldots, n\}$). Bei kategorialen Einflussgrößen bietet sich als Trennmöglichkeit die Frage an, ob die gegebene Kategorie in einer Menge von Kategorien K enthalten ist ($A^* = \{x_i | x_i \in K, i = 1, \ldots, n\}$). Man ersetzt dann die ursprüngliche Response-Variable durch die binäre Variable $I(x_i \in A^*)$, wobei I die Indikator-Funktion ist, und wiederholt die Suche nach dem besten Split mit diesem Ersatz-Response.

2.3. Random Forests

Random Forests bestehen aus einer Anzahl von Klassifikations- oder Regressionsbäumen, bei denen jeder Baum von den Werten eines Zufallsvektors abhängt. Diese Zufallsvektoren sind untereinander unabhängig und haben die gleiche Verteilung für alle Bäume im Random Forest[1] [Breiman, 2001, S. 5]. Den zu vorhersagenden Daten wird dann jeweils das Mittel über die Bäume (Regression) oder die meist zugeteilte Kategorie (Klassifikation) zugewiesen.
In jedem Baum an jedem Knoten werden also in diesem Algorithmus zufällig gewählte Einflussgrößen oder eine Kombination von Einflussgrößen verwendet. Im Falle der zufällig gewählten Einflussgröße ergeben sich Fehlerraten ähnlich zum bekannten Boosting-Verfahren Adaboost und es zeigte sich, dass Random Forests nicht sehr anfällig sind, *wie viele* Einflussgrößen ausgewählt werden. Schon eine einzige zufällig gewählte Einflussgröße kann zu einer guten Genauigkeit in der Vorhersage führen [Breiman, 2001, S. 14]. Falls (lineare) Kombinationen von Einflussgrößen verwendet werden, erzielt man noch bessere Ergebnisse [Breiman, 2001, S. 14]. Die Kombinationen erhält man, indem man einige Einflussgrößen zufällig zieht und diese dann addiert werden mit auf $[-1, 1]$ gleichverteilten Koeffizienten. Wenn es sich um kategoriale Einflussgrößen mit L Kategorien handelt, werden die Werte zuvor in $L - 1$ Dummy-Variablen umkodiert.

Breiman [2001] hat gezeigt, dass Random Forests die Daten nicht überanpassen, je mehr Bäume erstellt werden [Breiman, 2001, Theorem 1.2, S. 7], Im Gegenteil, man erhält eine obere Grenze für den Generalisierungsfehler. Dieser hängt von der Stärke und Genauigkeit der einzelnen Klassifikationsmethoden ab (in diesem Fall also des CART) und von der Abhängigkeit zwischen ihnen. Man versucht deshalb, die mittlere Korrelation zwischen den einzelnen Bäumen so gering wie möglich zu

[1] englisch für „Wald“

halten, während die Vorhersagekraft gleichbleiben soll.
Dieses Vorgehen führt zu vielen Vorteilen eines Random Forests: Er ist bezüglich der Vorhersage mindestens genauso gut wie Adaboost, robust gegenüber Rauschen und Ausreißern, schneller als Boosting, und liefert gleichzeitig interne Schätzungen für den Fehler, die Stärke, die Korrelation usw. [Breiman, 2001, S. 10]. Random Forests können außerdem gut mit schwachen Einflussgrößen umgehen, solange deren Korrelation gering ist [Breiman, 2001, S. 23].

2.3.1. Klassifikation

Nachdem viele Bäume mit immer dem gleichen Lern-Datensatz, also ein *Forest*, erstellt wurden, wird dem Response der neuen Beobachtungen (z. B. aus dem Test-Datensatz) diejenige Klasse zugewiesen, welche die meisten Bäume zuteilen. Der Response ist also kategorial.

2.3.2. Regression

Hier ist der Response numerisch und der Fehler wird durch den mittleren quadratischen Fehler (englisch: *Mean Squared Error* (MSE)) berechnet. Äquivalent zum Klassifikations-*Forest* hängt der Fehler von der Korrelation und dem Fehler der einzelnen Bäume ab [Breiman, 2001, Theorem 11.2, S. 26]. Wie Breiman [2001] zeigte, wird der Fehler des *Forests* umso geringer, je mehr Einflussgrößen benutzt werden, aber die mittlere Korrelation zwischen den einzelnen Bäumen wird höher. Jedoch sollte diese ja möglichst klein sein (siehe oben). Außerdem wird eine relativ große Anzahl an Einflussgrößen benötigt, um den Fehler des *Forests* zu verringern. Als Vorhersage für neue Beobachtungen wird das Mittel über alle Bäume des *Forests* verwendet.

Random Forests sind also ein effektives Instrument zur Vorhersage, welche die Daten laut Breiman [2001] nicht überanpassen. Wenn man die „richtige Art von Zufälligkeit“ (für die Zufallsvektoren) verwendet, ergeben sie sehr genaue Klassifizierungs- und Regressionsverfahren, deren Ergebnisse mit Boosting vergleichbar sind. Vor allem zur Klassifikation sind sie sehr gut geeignet, etwas weniger gut für die Regression. Implementationen finden sich z. B. im R-Paket `randomForest`.

2.4. Conditional Tree Forests

Allerdings wurde bald herausgefunden, dass die Bäume nach CART sehr wohl die Daten überanpassen und außerdem einen Selektions-Bias haben, der bevorzugt Einflussgrößen mit vielen fehlenden Werten oder vielen möglichen Trennpunkten auswählt [Hothorn u. a., 2006, u. a. S. 651], da sie die Unreinheit der Knoten mit dem

Gini-Index messen. Außerdem sind sie nicht zuverlässig in Situationen, in denen die möglichen Prädiktoren in verschiedenen Skalen gemessen werden [Strobl u. a., 2007, S. 2]. Pruning-Verfahren umgehen die Überanpassung, aber der Selektions-Bias ist trotzdem vorhanden. Das R-Paket `party` von Hothorn u. a. [2006] löst auch das Bias-Problem durch bedingte Inferenz, welche auf alle Arten von Regressionsproblemen anwendbar ist, z. B. können die Methoden mit nominalem, numerischem, ordinalem oder sogar zensiertem und multivariatem Response umgehen. Zudem können die *Einflussgrößen* im Gegensatz zu Bäumen nach CART auf beliebigen Skalen gemessen werden. Dazu wird die rekursive binäre Partitionierung der CART mit Permutationstests kombiniert. Außerdem werden multiple Testverfahren (Hypothesentests) verwendet, um zu überprüfen, ob zwischen den Einflussgrößen und dem Response kein signifikanter Zusammenhang besteht. Die Vorhersagekraft ist so gut wie diejenige eines Baumes, der durch Pruning entstanden ist (optimal beschnittener Baum).

Der Algorithmus für einen *Conditional Inference Tree* geht von der bedingten Verteilung des Response bei gegebenen Einflussgrößen aus. Das Regressionsmodell wird auf einem Lern-Datensatz $\mathcal{L}_n$ (mit n Beobachtungen) angepasst, der eventuell fehlende Werte enthält. Jeder Knoten eines Baumes wird dann durch nichtnegative ganzzahlige Fall-Gewichte $\boldsymbol{w} = (w_1, \ldots, w_n)$ repräsentiert [Hothorn u. a., 2006, S. 654], deren Einträge w_i ohne Beschränkung der Allgemeinheit auf 0 oder 1 festgelegt werden [Hothorn u. a., 2006, S. 655]. Auf Grund dieser Fall-Gewichte $\boldsymbol{w}$ wird die globale Null-Hypothese der Unabhängigkeit getestet und gestoppt, falls diese Hypothese nicht abgelehnt werden kann, d. h. falls keine der Einflussgrößen $\boldsymbol{x}_j$ mit dem Response $\boldsymbol{y}$ korreliert ist. Ansonsten wird diejenige Einflussgröße ausgewählt, deren Korrelation mit dem Response am größten ist. Anschließend wird ein Trennpunkt in dieser Einflussgröße gesucht. Dabei entstehen zwei weitere Fall-Gewichte $\boldsymbol{w}_{\text{links}}$ und $\boldsymbol{w}_{\text{rechts}}$, eines für den linken und eines für den rechten Tochterknoten. Diese beiden Schritte, also zuerst die Prüfung der Null-Hypothese und anschließende Variablenselektion durch Zusammenhangsmessung und dann die Suche nach dem Split, werden rekursiv wiederholt. Dadurch, dass die Variablenselektion und die Suche nach dem Trennpunkt voneinander gelöst sind, entsteht eine Baumstruktur, die keinen Selektions-Bias aufweist [Hothorn u. a., 2006, S. 655].

Variablenselektion Im Zuge der Prüfung der globalen Null-Hypothese werden p Teil-Null-Hypothesen getestet, welche prüfen, ob die j-te Variable ($j = 1, \ldots, p$) tatsächlich Einfluss auf $\boldsymbol{y}$ hat. Der Zusammenhang zwischen der j-ten Variable und $\boldsymbol{y}$ wird durch geeignete Teststatistiken oder p-Werte gemessen.
Als Default im R-Paket `party` werden lineare Teststatistiken in Vektorform verwendet, deren bedingter Erwartungswert und die bedingte Varianz berechnet werden kann. In die Berechnung dieser Momente fließen die Permutationen ein (für genaue Formeln etc. siehe Hothorn u. a. [2006, S. 655ff.]). Weil ein Permutationstest verwendet wird, ist es möglich, Variablen auf unterschiedlichem Niveau zu benutzen.
Durch die Möglichkeit der Berechnung des Erwartungswertes und der Varianz kön-

nen die Teststatistiken standardisiert werden. Anschließend werden sie auf verschiedene Weise in eine univariate Form gebracht: Entweder durch die Auswahl des maximalen Werts (Default im Paket `party`) oder durch eine quadratische Form, die jedoch den Nachteil hat, dass die Teststatistiken für die einzelnen Variablen nicht direkt miteinander vergleichbar sind.
Falls der p-Wert zur Beurteilung herangezogen wird, werden die p-Werte der bedingten Unabhängigkeitstests adjustiert, z. B. durch einfache Bonferroni-Adjustierung, und dann diejenige Variable $\boldsymbol{x}_j$ mit dem kleinsten p-Wert gewählt [Hothorn u. a., 2006, S. 657].
Das Sicherheitsniveau α der Tests liegt für Conditional Trees bei 0.05. Falls – wie in der vorliegenden Arbeit – Conditional Tree *Forests* berechnet werden, ist das Sicherheitsniveau variabel. Per Default liegt er bei 0.1.

Suche nach dem Trennpunkt Die Güte eines Splits wird hier nicht mehr durch eine Unreinheit berechnet, sondern durch Spezialfälle der oben verwendeten linearen Teststatistiken. Sie entsprechen im Wesentlichen einem Zwei-Stichproben-Test.
Falls in einer Variable eine Beobachtung fehlt, wird ihr zugehöriges Gewicht w_i auf Null gesetzt. Wenn dann in dieser Variable ein Split gewählt werden soll, verwendet man Surrogat-Splits.

Hothorn u. a. [2006, S. 663ff.] haben gezeigt, dass Conditional Inference Trees keinen Bias haben, die Daten nicht überanpassen und eine Vorhersagekraft äquivalent zu optimal beschnittenen Bäumen besitzen. Conditional Tree Forests übertreffen die Random Forests von Breiman [2001] hinsichtlich der Vorhersagekraft und der richtigen Struktur der Bäume.

Die Wahrscheinlichkeit, einen falschen Trennpunkt zu finden, ist außerdem durch α begrenzt [Hothorn u. a., 2006, S. 670]. Ein weiterer Vorteil besteht in der beliebigen Messskala der Variablen und des Responses. Zudem sind sie zur Erklärung *und* zur Vorhersage bestens geeignet. Überdies liefern sie unverzerrte Messungen zur Relevanz der Variablen [Strobl u. a., 2007, S. 5]. Ein Nachteil der Conditional Tree Forests ist die höhere Laufzeit gegenüber den klassischen Random Forests [Strobl u. a., 2007, S. 19].

3. Methoden für fehlende Werte und Imputation

3.1. Methoden für fehlende Werte

Diese Mechanismen und deren zugehörigen R-Funktionen sind von Svedjar [2007] übernommen worden. Genauere Angaben zur Verwendung der Funktionen finden sich im Anhang (siehe D).

3.1.1. Vollständig zufälliges Fehlen

Die wohl einfachste Methode, fehlende Werte in einen Datensatz zu streuen, besteht darin, mittels einen Zufallszahlengenerators Stellen auszuwählen und diese zu streichen. Es ergibt sich also ein vollständig zufälliges Fehlen, auf englisch „missing completely at random“ (MCAR). Die R-Funktion `deleteMCAR` zieht aus den n Beobachtungen so viele Stellen wie gewünscht (ohne Zurücklegen) und ersetzt diese mit `NA`.

3.1.2. Zufälliges Fehlen

Etwas komplizierter stellen sich Methoden dar, die nicht mehr *vollständig* zufällig fehlen. Die vollständige Zufälligkeit geht verloren, da die Stelle, an welcher der Wert gestrichen wird, von einer Beurteilungsvariable abhängt. Ob ein Wert beobachtet wird oder nicht, ist von den anderen beobachteten Daten beeinflusst. Somit ergibt sich „nur“ noch zufälliges Fehlen, auf englisch „missing at random“ (MAR).

Bildung von Rängen

Die R-Funktion `deleteMAR1` erstellt hierzu einen Hilfsvektor mit den Rängen der Beurteilungsvariable. Die Wahrscheinlichkeit, gestrichen zu werden, ergibt sich dann aus dem jeweiligen zugehörigen Rang geteilt durch die Summe aller Ränge. Die

Ränge werden dabei so vergeben, dass der kleinste Wert in der Beurteilungsvariable den Rang 1 hat, der zweitkleinste Rang 2 usw. Somit hat auch der kleinste Wert die geringste Wahrscheinlichkeit, gestrichen zu werden. Mit dem resultierenden Wahrscheinlichkeitsvektor werden wieder zufällig aus den n Beobachtungen so viele Stellen wie gewünscht (ohne Zurücklegen) gezogen und mit `NA` ersetzt. Aber die Wahrscheinlichkeit, gezogen zu werden, ist nun nicht mehr für alle Beobachtungen gleich groß.

Bildung von zwei Risikogruppen

Die R-Funktion `deleteMAR2` unterteilt den gesamten Datensatz in zwei Untergruppen. Dazu dient die Beurteilungsvariable, denn die eine Gruppe bilden diejenigen Beobachtungen, deren Werte in der Beurteilungsvariable größer als bzw. gleich groß wie der Median in der Beurteilungsvariable sind. Die restlichen Beobachtungen mit Werten unter dem Median bilden eine andere Gruppe. Nun wird wieder ein Wahrscheinlichkeitsvektor benötigt. Dieser hat die Masse 1, welche auf die beiden Gruppen im Verhältnis 1:9 aufgeteilt wird. Wenn nun eine Beobachtung in der Gruppe über dem Median ist, ist ihre Streichwahrscheinlichkeit um den Faktor 9 größer als die Streichwahrscheinlichkeit in der Gruppe unter dem Median. Die einzelne Beobachtung in der jeweiligen Gruppe erhält die Streichwahrscheinlichkeit 0.1 bzw. 0.9 geteilt durch die Anzahl der Beobachtungen in dieser Gruppe [Svedjar, 2008]. Anschließend wird wieder unter Zuhilfenahme dieses Wahrscheinlichkeitsvektors zufällig aus den n Beobachtungen so viele Stellen wie gewünscht (ohne Zurücklegen) gezogen und mit `NA` ersetzt. Auch hier ist die Wahrscheinlichkeit, gezogen zu werden, nicht mehr für alle Beobachtungen gleich groß.

Rechtsseitige Trunkierung

Die R-Funktion `deleteMAR3` ist in der Handhabung wesentlich einfacher. Sie streicht einfach so viele Beobachtungen mit den jeweils größten Werten in der Beurteilungsvariable, bis der gewünschte Anteil an fehlenden Werten erreicht ist. Dabei ist allerdings zu beachten, dass in der Beurteilungsvariable selbst keine fehlenden Werte vorkommen dürfen.

Symmetrische Trunkierung

Die R-Funktion `deleteMAR4` handelt ähnlich wie `deleteMAR3`. Der einzige Unterschied besteht darin, dass sie nicht nur die Beobachtungen mit den größten Werten in der Beurteilungsvariable streicht, sondern den Anteil an gewünschten fehlenden Werten halbiert und einen Teil in den größten Werten entfernt und den anderen Teil in den Beobachtungen mit den kleinsten Werten in der Beurteilungsvariable eliminiert.

Für die beiden Trunkierungsmethoden `deleteMAR3` und `deleteMAR4` bleibt allerdings festzuhalten, dass die Werte nun nicht mehr zufällig gelöscht werden.

3.2. Imputationsmethode

Die hier verwendete Imputationsmethode nennt sich *knn*-Impute. Die Abkürzung *knn* steht dabei für *k nearest neighbours*, d. h. die „k nächsten Nachbarn“. Wenn in einer Beobachtung $\boldsymbol{x}_i$ an der j-ten Stelle ein Wert fehlend ist, werden die k ähnlichsten Beobachtungen gesucht, die an der j-ten Stelle keine fehlenden Werte haben. Die Ähnlichkeit wird dabei über die Euklidische Norm definiert [Hastie u. a.]. Die Euklidische Norm ist laut Troyanskaya u. a. [2001, S. 521] in diesem Fall die genaueste Distanzmessung. (Noch besser wäre laut Troyanskaya u. a. [2001] eine vorherige log-Transformierung der Daten, da dies die Effekte von Ausreißern minimiert. Darauf wurde jedoch hier verzichtet.)

Sind die k ähnlichsten Beobachtungen gefunden, wird der fehlende Wert in $\boldsymbol{x}_i$ ersetzt, indem der Mittelwert der Werte an der Stelle j aus diesen ähnlichsten Beobachtungen eingesetzt wird. Die Beiträge der einzelnen Beobachtungen werden dabei mit ihrer Ähnlichkeit zur Variable mit dem fehlenden Wert gewichtet. Genauer ersetzt also ein gewichteter Mittelwert den fehlenden Wert.
Falls einer oder mehrere der k nächsten Nachbarn an anderen Stellen als j fehlende Werte haben, wird die Distanz gemittelt über die nicht-fehlenden Werte.

Bei Troyanskaya u. a. [2001] erscheint der *knn*-Impute als das genaueste Imputationsverfahren. Die Methode scheint zudem robust gegenüber einem steigenden Anteil an fehlenden Werten zu sein [Troyanskaya u. a., 2001, S. 522]. Dennoch macht ein geringerer Anteil an fehlenden Daten die Imputation präziser. Zusätzlich ist der *knn*-Impute relativ widerstandsfähig gegenüber der genauen Anzahl von gesuchten nächsten Nachbarn, solange diese etwa zwischen Zehn und 20 liegt [Troyanskaya u. a., 2001, S. 524]. Allerdings sollte die Datenmatrix, auf welche die Methode angewendet wird, mindestens vier Spalten umfassen.

Die R-Funktion `impute.knn` bezieht sich auf den Artikel von Troyanskaya u. a. [2001]. Sie ist im Paket `impute` von Hastie u. a. zu finden. Die Standard-Einstellungen von `impute.knn` wurden nicht verändert. Somit liegt die Anzahl an gesuchten nächsten Nachbarn bei $k = 10$.

Der Vorteil dieser Imputationsmethode liegt darin, dass sie z. B. gegenüber dem simplen Zeilenmittelwert Informationen aus der Korrelation zwischen den einzelnen Daten zieht [Troyanskaya u. a., 2001, S. 521] und somit die Ersetzung der fehlenden Werte genauer ist. Trotzdem sind die Verteilungsannahmen etc. minimiert [Troyanskaya u. a., 2001, S. 521]. Deswegen ist dieses Verfahren für viele Skalen anwendbar. Der *knn*-Impute ist recht schnell, aber dabei genau und wenig empfindlich beim Parameter k.

4. Simulationsdesign

Wie in der Einleitung bereits erwähnt, sollen nun die beiden Methoden zum Umgang mit fehlenden Werten an Hand einer Simulationsstudie verglichen werden. Die erste Methode ist der Random Forest selbst, der mit fehlenden Werten umgehen kann. Er verwendet dazu, wie auch schon weiter oben beschrieben wurde, Surrogat-Variablen. In einem zweiten Verfahren wird erst dann ein *Forest* erstellt, nachdem die fehlenden Werte imputiert worden sind. D. h. sie werden an Hand des Verfahrens *knn*-Impute geschätzt (siehe Abschnitt 3.2). Die Daten werden auf zwei verschiedene Arten erstellt.

4.1. Datengenerierende Prozesse

Durch zwei unterschiedliche datengenerierende Prozesse erhält man eine Sorte von Datensatz, in dem der Response $\boldsymbol{y}$ kategorial ist, also ein Klassifikationsbaum erstellt werden soll. Im zweiten Fall ist der Response $\boldsymbol{y}$ stetig, sodass ein Regressionsbaum vonnöten ist.

Dazu wurden zwei Funktionen erstellt, die beide beliebig viele (`niter`) Datensätze mit je gleich vielen (`n`) Beobachtungen erzeugen. Genauere Angaben zur Verwendung der beiden Funktionen finden sich im Anhang (siehe C).

4.1.1. Prozess für die Klassifikation

Der kategoriale datengenerierende Prozess wurde von Svedjar [2007] übernommen. Die Funktion `dgp1` (im Original `create`) erzeugt dabei einen binären Faktor als Response mit den Ausprägungen $y = 1$ und $y = 2$. Dies basiert auf fünf multivariat normalverteilten Zufallsvariablen $\boldsymbol{x}_j, j = 1, \ldots, 5$ mit dem Erwartungswert Null und beliebiger Varianz. Die Varianz kann durch den Eingabe-Parameter `sigma` verändert werden. Der Einfluss der einzelnen Kovariablen kann mittels dem Eingabe-Parameter `coef` beliebig variiert werden. Per Default liegt dieser bei `coef` $= (1, 2, 3, 4, 5)'$. Der Response berechnet sich durch Modellierung von Logits, d. h.

$$P(\boldsymbol{y} = 2|\boldsymbol{X}) = \pi(\boldsymbol{X}) \quad = \quad \frac{\exp(\boldsymbol{X}'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{X}'\boldsymbol{\beta})}, \tag{4.1}$$

wobei `coef` dem Koeffizienten-Vektor β entspricht und $\boldsymbol{X}$ die Matrix der Einflussgrößen $\boldsymbol{x}_j, j = 1, \ldots, 5$ ist. Durch diesen Wahrscheinlichkeitsvektor wird dem Response $\boldsymbol{y}$ zufällig die Klasse 1 oder 2 zugeteilt [Svedjar, 2007, S. 25].

4.1.2. Prozess für die Regression

Der Prozess, der den stetigen Response berechnet, wurde schon in mehreren Artikeln verwendet und stammt von Friedman [1991]. Im Original-Modell „Friedman1“ [Friedman, 1991, Formel (61)] werden genau zehn gleichverteilte Variablen gebraucht, von denen nur die ersten fünf einen Einfluss haben:

$$\boldsymbol{y} = 10 \cdot \sin(\pi \boldsymbol{x}_1 \boldsymbol{x}_2) + 20 \cdot (\boldsymbol{x}_3 - 0.5)^2 + 10 \cdot \boldsymbol{x}_4 + 5 \cdot \boldsymbol{x}_5$$

In der Funktion `dgp2` ist es möglich, die Anzahl an Variablen ohne Einfluss beliebig groß zu machen, sie ist allerdings per Default auf ebenfalls fünf gesetzt. Diese Variablen ohne Einfluss sind untereinander unabhängig, die Varianz und damit die Korrelation der fünf Einflussgrößen kann man mittels `sigma` verändern.

Die gleichverteilten Zufallsvariablen konnten wie folgt simuliert werden: Zuerst wurde die entsprechende Anzahl an multivariat normalverteilten Zufallsvariablen mit Hilfe der R-Funktion `rmvnorm` aus dem Paket `mvtnorm` gezogen. Dabei wurden bereits die später gewünschten Korrelationen verwendet. Anschließend hat man die Verteilungsfunktion der Standardnormalverteilung auf diese Zahlen angewandt, woraus auf $[0, 1]$ gleichverteilte Zufallszahlen resultieren. Es gilt also:

$$\begin{aligned} X &\sim F(x) \\ \Rightarrow F(X) &\sim U(0, 1) \end{aligned}$$

Diese Methode ist auch unter dem Namen „Inversionsverfahren“ bekannt.
Zur Sicherheit wurde an dieser Stelle ein Kolmogorow-Smirnow-Test gerechnet, damit die Zufallszahlen tatsächlich gleichverteilt sind. Wurde dieser Test abgelehnt, mussten erneut multivariat standardnormalverteilte Zufallszahlen gezogen werden.

4.2. Korrelationsmatrizen

Es wurden je datengenerierendem Prozess drei verschiedene Kovarianzmatrizen benutzt. Alle drei haben auf der Diagonale Einsen, sodass die Kovarianz- gleich der Korrelationsmatrix ist. Denn die Kovarianz zweier Zufallsvariablen bzw. Einflussgrößen $\boldsymbol{x}_{j_1}$ und $\boldsymbol{x}_{j_2}$ berechnet sich durch

$$V(\boldsymbol{x}_{j_1}, \boldsymbol{x}_{j_2}) = E\left[(\boldsymbol{x}_{j_1} - E(\boldsymbol{x}_{j_1}))(\boldsymbol{x}_{j_2} - E(\boldsymbol{x}_{j_2}))'\right]$$

und die Korrelation ergibt sich aus

$$\varrho(\boldsymbol{x}_{j_1}, \boldsymbol{x}_{j_2}) = \frac{V(\boldsymbol{x}_{j_1}, \boldsymbol{x}_{j_2})}{\sqrt{V(\boldsymbol{x}_{j_1}) V(\boldsymbol{x}_{j_2})}}.$$

Wenn nun die Varianzen der Zufallsvariablen $V(\boldsymbol{x}_{j_1}) = 1$ und $V(\boldsymbol{x}_{j_2}) = 1$ sind, so folgt

$$\varrho(\boldsymbol{x}_{j_1}, \boldsymbol{x}_{j_2}) \quad = \quad V(\boldsymbol{x}_{j_1}, \boldsymbol{x}_{j_2}).$$

Die erste Matrix besteht aus gleichmäßig hohen Korrelationen zwischen den einzelnen Einflussgrößen von 0.9:

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0.9 & 0.9 & 0.9 & 0.9 \\ 0.9 & 1 & 0.9 & 0.9 & 0.9 \\ 0.9 & 0.9 & 1 & 0.9 & 0.9 \\ 0.9 & 0.9 & 0.9 & 1 & 0.9 \\ 0.9 & 0.9 & 0.9 & 0.9 & 1 \end{pmatrix}$$

Die zweite Matrix beinhaltet ebenfalls hohe Korrelationen, allerdings sind die Variablen in zwei Blöcke unterteilt, sodass $\boldsymbol{x}_1$ bis $\boldsymbol{x}_3$ eine Gruppe bilden sowie $\boldsymbol{x}_4$ und $\boldsymbol{x}_5$ eine zweite:

$$\boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & 0.9 & 0.9 & 0 & 0 \\ 0.9 & 1 & 0.9 & 0 & 0 \\ 0.9 & 0.9 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0.9 \\ 0 & 0 & 0 & 0.9 & 1 \end{pmatrix}$$

Die dritte Matrix hat wieder gleichmäßige Korrelationen zwischen den einzelnen Variablen, aber diesmal niedrige Werte von 0.1:

$$\boldsymbol{\Sigma}_3 = \begin{pmatrix} 1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 & 1 \end{pmatrix}$$

Der Einfluss der Variablen im kategorialen Prozess wurde beim Default belassen. Beim Regressionsprozess ist er durch das Friedman-Modell vorgegeben.

4.3. Methoden für fehlende Werte

Diese Mechanismen wurden bereits in Kapitel 3 erläutert.

Der Anteil an fehlenden Werten blieb in jeder Simulation gleich. In der ersten Einflussgröße $\boldsymbol{x}_1$ fehlen sowohl im jeweiligen Lern- als auch im Test-Datensatz 20% der Daten, in der dritten Einflussgröße $\boldsymbol{x}_3$ 10% und in der vierten Einflussgröße $\boldsymbol{x}_4$ wiederum 20% der Daten. Der Anteil 20% erschien sinnvoll, da mit dieser Aufteilung maximal 50% aller Beobachtungen unvollständig sein können und außerdem bei Troyanskaya u. a. [2001] ebenfalls mit dieser Obergrenze gearbeitet wurde.

Für `deleteMCAR` benötigt man bezüglich der fehlenden Werte keine weiteren Angaben, für die `MAR`-Funktionen muss noch die Beurteilungsvariable bestimmt werden.

Für `deleteMAR1` und `deleteMAR2` hat man sich hier für $\boldsymbol{x}_2$ als Beurteilungsvariable von $\boldsymbol{x}_1$ entschieden, da diese beiden Variablen in $\boldsymbol{\Sigma}_2$ einem Korrelationsblock angehören. Dadurch kann das Verhalten bei hoch korrelierten Variablen zusätzlich zur Korrelationsmatrix $\boldsymbol{\Sigma}_1$ untersucht werden. Bei $\boldsymbol{x}_3$ fiel die Wahl auf $\boldsymbol{x}_4$, um die Zuordnungsfähigkeiten von `cforest` bei nicht-korrelierten Variablen zu testen. Denn genau dies ist der Fall beim Korrelationsdesign von $\boldsymbol{\Sigma}_2$. Für $\boldsymbol{x}_4$ ist die Beurteilungsvariable $\boldsymbol{x}_5$, aus dem gleichen Grund wie bei $\boldsymbol{x}_1$. Außerdem ist dann in jedem Block in $\boldsymbol{\Sigma}_2$ jeweils eine Variable `NA`-behaftet und eine ist Beurteilungsvariable. Der „gemischte“ Fall mit der unkorrelierten Beurteilungsvariable muss die Blockgrenzen überschreiten, um eben die gewünschte Unkorreliertheit zu erreichen.

Bei `deleteMAR3` und `deleteMAR4` werden komplett beobachtete Beurteilungsvariablen benötigt, d. h. in der jeweiligen Beurteilungsvariable dürfen keine Werte fehlen. Aus diesem Grund werden in der Einflussgröße $\boldsymbol{x}_3$ ebenfalls mit Hilfe der Variable $\boldsymbol{x}_5$ Werte gestrichen. Denn auch zwischen diesen beiden Variablen besteht in der nicht-gleichmäßigen Matrix $\boldsymbol{\Sigma}_2$ keine Korrelation und in $\boldsymbol{x}_4$ werden ja ebenfalls Beobachtungen eliminiert. Dadurch ist $\boldsymbol{x}_4$ nicht vollständig beobachtet und aus diesem Grund nicht als Beurteilungsvariable bei `deleteMAR3` und `deleteMAR4` geeignet. Aber durch die äquivalente Unkorreliertheit von $\boldsymbol{x}_3$ mit $\boldsymbol{x}_5$ ist der Effekt gleichbedeutend mit der Verwendung der Beurteilungsvariable $\boldsymbol{x}_4$ bei `deleteMAR1` und `deleteMAR2`.

4.4. Imputationsmethode

Die Theorie zur verwendeten Imputationsmethode *knn*-Impute findet sich in Kapitel 3.2.

4.5. Berechnung der Random Forests

Für die Berechnung der Conditional Tree Forests wurde die Funktion `cforest` aus dem bereits mehrfach erwähnten R-Paket `party` verwendet. Die angewandten Einstellungen dieser Funktion sind im Folgenden noch einmal zusammengefasst:

Als univariate Teststatistik wurde der maximal erzielte Wert der linearen Teststatistiken verwendet, und kein p-Wert. Als Sicherheitsniveau α wurde die Standard-Einstellung 0.1 beibehalten. Die Anzahl an Bäumen je erstelltem *Forest* (`ntree`) lag bei 50, die maximale Anzahl an Surrogat-Splits bei 3. Damit ein Knoten als möglicher „Split-Kandidat“ betrachtet wird, musste die Summe an Gewichten mindestens 30 betragen.

Da die Funktion noch nicht ausgereift ist und aus diesem Grund in manchen Fällen Fehler warf, wurde dann das zugehörige Gütemaß (siehe unten, Formeln (4.2) und (4.3)) mit `NA` abgespeichert. Genauere Angaben zur Verwendung der Funktion finden sich im Anhang (siehe E).

4.6. Simulation

Gesamt ergaben sich also alle möglichen Kombinationen aus den

zwei datengenerierenden Prozessen

- `dgp1`: kategorialer Response; fünf normalverteilte Variablen mit steigendem Einfluss
- `dgp2`: stetiger Response; zehn gleichverteilte Variablen, davon fünf mit Einfluss

den drei Kovarianz-/Korrelationsmatrizen

- gleichmäßig hohe Korrelation von 0.9
- hohe Korrelation von 0.9 in zwei Gruppen
- gleichmäßig niedrige Korrelation von 0.1

den fünf Methoden, um fehlende Werte einzustreuen

- `deleteMAR1`: Bildung von Rängen, Entfernung mit Wahrscheinlichkeit „Rang durch Summe der Ränge“
- `deleteMAR2`: Bildung von zwei Risikogruppen, Entfernung mit hohem (über dem Median der Beurteilungsvariable) oder niedrigem Risiko (unter dem Median der Beurteilungsvariable)
- `deleteMAR3`: rechtsseitige Trunkierung, Entfernung der größten Werte
- `deleteMAR4`: symmetrische Trunkierung, Entfernung der größten und der kleinsten Werte
- `deleteMCAR`: vollständig zufälliges Fehlen, Entfernung per Zufallsauswahl

und alles mit

- mit oder
- ohne Imputation, d. h. dann wurden Surrogat-Variablen verwendet.

Für jedes Szenario wurden die drei Möglichkeiten

- fehlende Werte in den Lern-Datensätzen,
- fehlende Werte im Test-Datensatz und
- fehlende Werte sowohl in den Lern-Datensätzen als auch im Test-Datensatz

absimuliert. Für sämtliche Simulationen galt dabei ein Seed von 856329.

Es wurden zwei Test-Datensätze erstellt – für jede datengenerierende Funktion einen. Sie beinhalteten jeweils 5000 Beobachtungen. Die Lern-Datensätze wurden für jede Simulation neu erzeugt und bestanden aus 100 Beobachtungen. Pro Simulationsdurchlauf wurden 200 Lern-Datensätze generiert.

Zum Vergleich wurden noch für jede Korrelationsmatrix $\mathbf{\Sigma}_k, k = 1, \ldots, 3$ ein Goldstandard komplett ohne fehlende Werte berechnet und für jede Matrix $\mathbf{\Sigma}_k$ und jede `delete`-Funktion ein vollständig beobachteter, reduzierter Fall „`-NA`“. In diesem Fall wurden in den Lern-Datensätzen fehlende Werte eingestreut und anschließend die Beobachtungen mit fehlenden Werten gestrichen, so dass der Lern-Datensatz nur noch vollständige Beobachtungen enthielt, aber weniger als 100.

Als Güte-Maß wurde bei den kategorialen Datensätzen die mittlere Binomial-Log-Likelihood

$$ll = \frac{1}{n} \sum_{i=1}^{n} y_i \cdot \log(\widehat{p}_i) + (1 - y_i) \cdot \log(1 - \widehat{p}_i) \tag{4.2}$$

verwendet, wobei $\widehat{p}_i$ die an Hand der Lern-Datensätze geschätzte Wahrscheinlichkeit für $y_i = 2$ im Test-Datensatz darstellt. Die Binomial-Log-Likelihood kann nur negative Werte annehmen bzw. maximal Null werden.

Bei den Datensätzen mit stetigem Response wurde die mittlere quadratische Abweichung (englisch: *Mean Squared Error* (MSE))

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \tag{4.3}$$

als Vergleichsmaß herangezogen. Der Fehler wurde dabei an Hand des Test-Datensatzes berechnet, indem man auf Grund der Lern-Daten die Response-Werte der Test-Daten vorhersagen ließ und diese vom jeweiligen wahren Wert abgezogen hat.

Zwei-Stichproben-t-Tests Damit der Unterschied in der Fehlerverteilung zwischen den Conditional Tree Forests mit imputierten Daten und den *Forests* mit Surrogat-Variablen (ohne Imputation) belegt werden konnte, wurden Zwei-Stichproben-t-Tests gerechnet. Die eine Stichprobe ist dabei der imputierte Datensatz und die andere Stichprobe ist der Datensatz mit den fehlenden Werten. Dabei wurden jeweils die Lern-Datensätze mit fehlenden Werten und Surrogat-Variablen mit den Lern-Datensätzen mit fehlenden Werten und Imputation verglichen. Ebenso wurde mit den Test-Datensätzen verfahren und mit den Fällen, bei denen in beiden Datensätzen Werte fehlten.
Der Zwei-Stichproben-t-Test vergleicht den Mittelwert. Die Null-Hypothese lautet hier $\mu_{\boldsymbol{Y}} = \mu_{\boldsymbol{Z}}$, wobei $\boldsymbol{Y}$ und $\boldsymbol{Z}$ die zu vergleichenden Datensätze sind. Die Alternativ-Hypothese lautet entsprechend $\mu_{\boldsymbol{Y}} \neq \mu_{\boldsymbol{Z}}$. Wenn nun ein p-Wert $< 0.05 = \alpha$ resultiert, so wird die Null-Hypothese abgelehnt und der Datensatz $\boldsymbol{Y}$ hat keinen gleich großen Mittelwert wie der Datensatz $\boldsymbol{Z}$. Das bedeutet, dass die Methode, die

in Datensatz $\boldsymbol{Y}$ angewendet wurde, sich signifikant von derjenigen des Datensatzes $\boldsymbol{Z}$ unterscheidet. Wenn ein Test nicht abgelehnt wurde, galten die beiden Methoden als im Mittelwert gleich gut.
Genauso wurde beim Vergleich der einzelnen Situationen mit dem Goldstandard und dem Fall „`-NA`“ verfahren. Der Goldstandard steht dann an Stelle des Datensatzes $\boldsymbol{Z}$ und die jeweilige Methode wird mit ihm verglichen. Analog funktioniert es bei Vergleichen mit dem reduzierten Fall „`-NA`“.
Beim Zwei-Stichproben-t-Test ist eine Normalverteilung Voraussetzung. Diese wird über den Zentralen Grenzwertsatz abgedeckt. Außerdem sind die Risiken einer Fehleinschätzung recht symmetrisch verteilt, wie man in den Grafiken im Kapitel 5 erkennen kann. Ein separater Test zur Überprüfung der Normalverteilung wurde nicht vorgeschaltet.
Die Zwei-Stichproben-t-Tests wurden mit der R-Funktion `t.test` ausgeführt. Die Ergebnisse aus den t-Tests werden im Ergebnis-Teil 5 nur teilweise genannt. Sämtliche p-Werte finden sich in Anhang H.

5. Ergebnisse

Die Verteilung des jeweiligen Güte-Maßes („risk“) wurde mittels Boxplots dargestellt. Die Grafiken sind aufgeteilt nach „Benchmark“[1], „RF[2] + Surrogat“ (ohne Imputation) und „impute.knn + RF“ (mit Imputation). Es existiert für jede Art von Datensatz (Klassifikation, Regression), jede Korrelation $\mathbf{\Sigma}_k, k = 1, \ldots, 3$ und jede `delete`-Funktion eine Grafik.
Die Abteilung „Benchmark“ unterteilt sich in den Goldstandard und den reduzierten Fall „`-NA`“, die Imputationskategorien in die drei verschiedenen Szenarien an fehlenden Werten[3] (in den Lern-Datensätzen („lernMV“), im Test-Datensatz („testMV“), sowohl in den Lern-Datensätzen als auch im Test-Datensatz („lerntestMV“)).
Die Breite der Boxplots entspricht der Anzahl an Beobachtungen darin. Wenn eine Box also relativ dünn ist, sind viele Beobachtungen mit `NA` kodiert.

Außerdem wurden die Differenzen der Güte-Maße zum Güte-Maß des Goldstandards berechnet und auch diese mit der gleichen Aufteilung als Boxplots dargestellt. Das Ideal wäre eine Box mit Median um Null oder einer Tendenz ins Positive, da dann der Goldstandard besser ist als der jeweils andere Fall.

Zum besseren optischen Vergleich wurde die nur Werte ≤ 0 annehmende Binomial-Log-Likelihood negativ gezeichnet, sodass die Orientierung derjenigen des MSE entspricht. Für eine bessere Gegenüberstellung innerhalb der Datensatz-Arten ist die Skalierung für alle Grafiken zur Klassifikation und für alle Grafiken zur Regression identisch. Falls Datenpunkte nicht sichtbar sind, stehen sie in den Tabellen im Anhang G.

Zusammenfassungen der in den Simulationen erzielten Gütemaße finden sich im Anhang (siehe F).

5.1. Simulation zur Klassifikation

Für einen Klassifikationsbaum bzw. -*Forest* ist ein kategorialer Response nötig, der mit Hilfe der Funktion `dgp1` erzeugt wurde. Der binäre Responsevektor $\boldsymbol{y}$ wurde durch Logits (siehe (4.1)) modelliert.

[1] englisch für: „Bezugswert“
[2] Abkürzung für Random Forest
[3] englisch: missing values; deshalb in den Grafiken mit „MV“ abgekürzt

Für beide Datensatz-Arten wurden alle drei Varianten der Korrelationsmatrix $\mathbf{\Sigma}_k$ verwendet.

Für jede Art von Korrelation wurde wiederum das Verhalten der verschiedenen Mechanismen für fehlende Werte untersucht.

5.1.1. Gleichmäßige, hohe Korrelationen

MAR 1

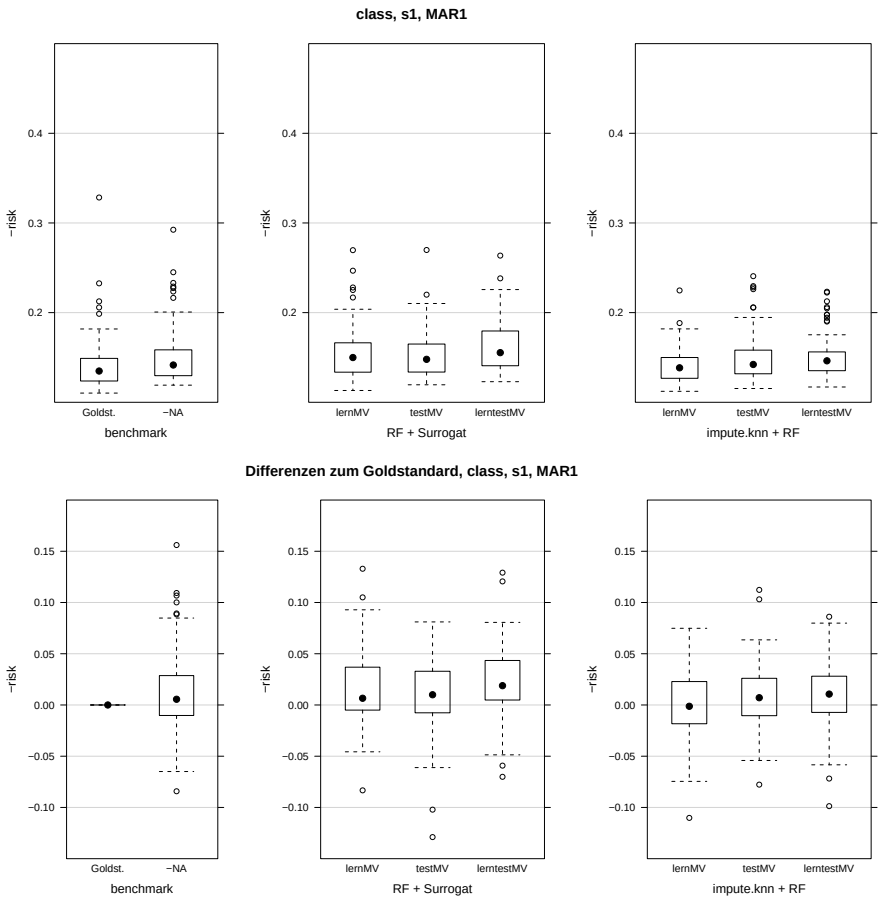

Abbildung 5.1.: Verteilung der negativen Binomial-Log-Likelihood sowie der Differenzen zum Goldstandard bei gleichmäßig hohen Korrelationen und fehlenden Werten nach MAR1 (Bildung von Rängen) [s1 entspricht Σ_1]

Da man auf Grund der guten Ergebnisse bei Troyanskaya u. a. [2001] davon ausgeht, dass die *knn*-Imputation korrekt arbeitet, ist es nicht weiter verwunderlich, dass bei erfolgter Imputation („impute.knn + RF“) die Maximalwerte der Binomial-

Log-Likelihood sowie die oberen Zäune der Boxplots in allen drei Varianten der fehlenden Werte unterhalb des reduzierten Falls liegen und die Boxen generell nicht sehr viel höher liegen. Ein signifikanter Unterschied wird allerdings nur bei den Lern-Datensätzen erzielt.
Zur Erinnerung: Der reduzierte Fall hat in den Lern-Datensätzen fehlende Werte und diese unvollständigen Beobachtungen wurden anschließend gestrichen.
Ohne Imputation, d. h. bei Verwendung von Surrogat-Variablen, ist das Bild etwas schlechter: Die Boxen und auch die oberen Zäune reichen etwas weiter nach oben, was auf ein schlechteres Ergebnis hindeutet. In allen drei Fällen ist der Mittelwert signifikant von dem des Goldstandards verschieden. Ein signifikanter Unterschied im Mittelwert zwischen Imputation und Surrogat-Variablen ergibt sich für die eliminierten Werte in den Lern-Datensätzen und im kombinierten Fall.

Wenn man die Differenzen zum Goldstandard betrachtet, ist die Tendenz besser zu erkennen: Mit Imputation werden keine schlechteren Ergebnisse als im reduzierten Fall ohne `NA` erzielt und ohne Imputation ergeben sich etwas größere Fehler. Man erkennt aber auch, dass die Mediane sämtlicher Boxen fast nicht von der Null abweichen und 0.02 nicht übersteigen. Das bedeutet, dass der Conditional Tree Forest den datengenerierenden Prozess in allen Fällen gut erkennt.

MAR 2

Bei fehlenden Werten durch Bildung zweier Risikogruppen fällt als erstes der schlechte Fall von Lern- und Testdaten mit `NA` und ohne Imputation („RF + Surrogat“) auf. Wegen der hohen Korrelation ist auch das vermeintlich „schlechte“ Ergebnis – insgesamt gesehen – nicht schlecht, jedoch nicht so gut wie die anderen Fälle in dieser Simulation. Der Median ist nur ca. 0.009 über dem des reduzierten Falls, dennoch resultierte ein signifikanter Unterschied im Mittelwert sowohl zum reduzierten Fall „`-NA`“ als auch zum Imputationsfall und zum Goldstandard. Ein weiterer signifikanter t-Test beim Vergleich zwischen Imputation und Surrogat-Variablen resultiert bei den Lern-Datensätzen.
Die Lern-Datensätze mit Surrogat-Variablen liegen im Median ebenfalls ca. 0.006 über dem reduzierten Fall. Alle anderen Mediane befinden sich zwischen Goldstandard und dem reduzierten Fall ohne `NA`, wobei `cforest` mit den Lern-Datensätzen mit Imputation im Median sogar noch ca. 0.003 besser als mit dem Goldstandard umgehen kann.

Bei den Differenzen zum Goldstandard zeigt sich in etwa die gleiche Situation: Der Fall der mit `NA` besetzten Lern- und Test-Datensätzen mit Verwendung von Surrogat-Variablen sticht heraus. Im genannten Fall (fast doppelt so weit von der Null entfernt), in den Lern-Datensätzen ohne Imputation und im Test-Datensatz mit Imputation (beide etwa gleich hoch wie der reduzierte Fall) ergeben sich größere Fehler als im reduzierten Fall und deshalb ein höherer Boxplot der Differenzen des Güte-Maßes.

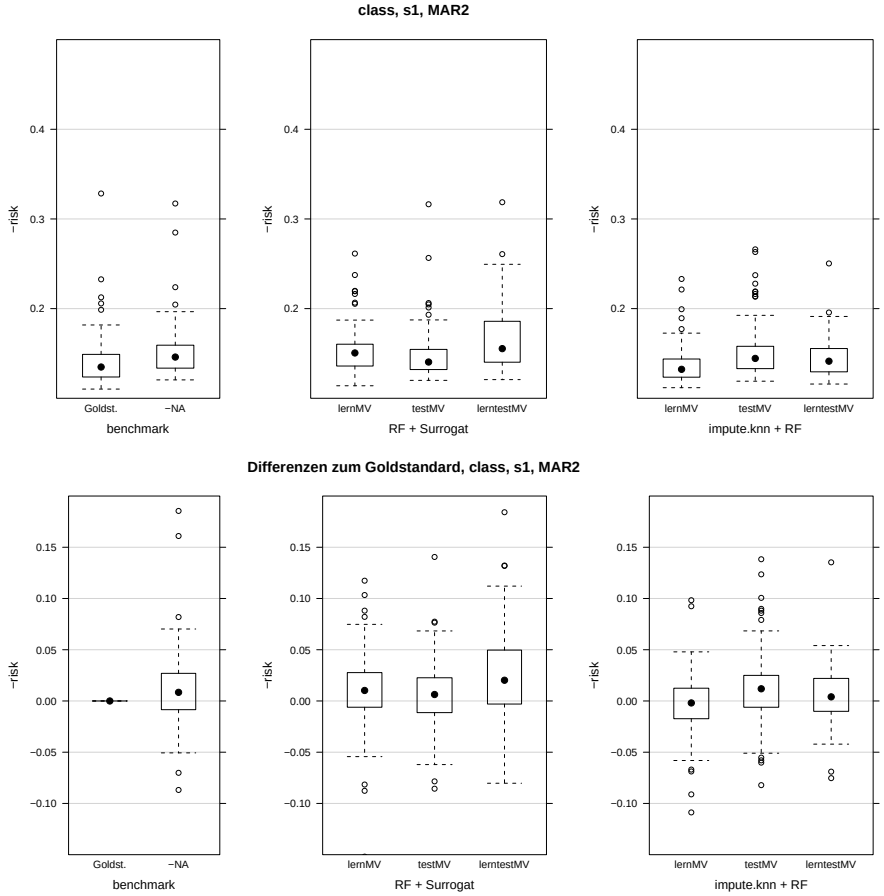

Abbildung 5.2.: Verteilung der negativen Binomial-Log-Likelihood sowie der Differenzen zum Goldstandard bei gleichmäßig hohen Korrelationen und fehlenden Werten nach MAR2 (Bildung von zwei Risikogruppen) [s1 entspricht Σ_1]

MAR 3

Während der Fehler beim Test-Datensatz mit fehlenden Werten und erfolgter Imputation bei MAR2 nur leicht erhöht ist, ragt er bei MAR3 weit heraus. Wegen der aus Vergleichsgründen immer gleichen Skalierung sind zwei Ausreißer nicht zu sehen. Der am weitesten von Null entfernte Punkt liegt bei fast 0.9. Der Median liegt um ca. 0.1 höher als der Median des reduzierten Falls ohne `NA`.

Weiterhin auffallend ist, dass sich Goldstandard und der reduzierte Fall kaum unterscheiden (bis auf wenige Ausreißer). Zudem wurden ohne Imputation bessere Ergebnisse erzielt als mit Imputation, was die signifikanten t-Tests (siehe Anhang H) für die Vergleiche mit und ohne Imputation sowie für den Vergleich Imputationsfall – Goldstandard bestätigen. Es werden sogar signifikante Abweichungen zum reduzierten Fall „`-NA`“ beobachtet. Dies könnte daran liegen, dass die fehlenden Werte

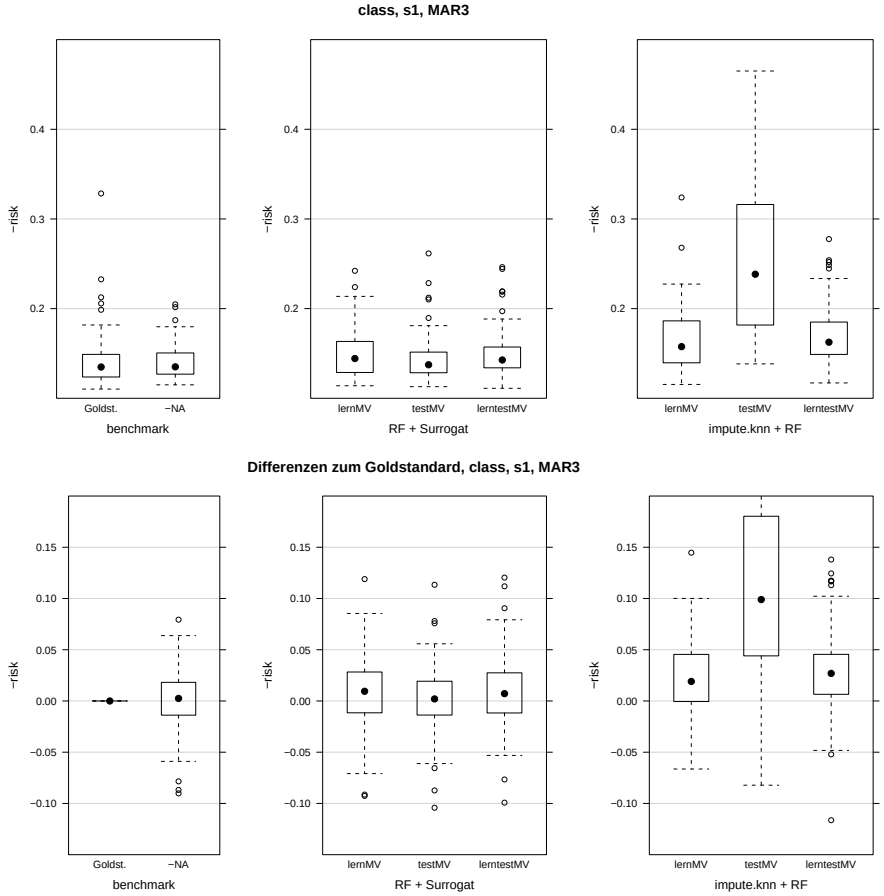

Abbildung 5.3.: Verteilung der negativen Binomial-Log-Likelihood sowie der Differenzen zum Goldstandard bei gleichmäßig hohen Korrelationen und fehlenden Werten nach MAR3 (rechtsseitige Trunkierung) [s1 entspricht Σ_1]

nicht über die Daten verteilt sind, sondern generell die höchsten Werte gestrichen wurden und deshalb trotz hoher Korrelation keine sinnvollen $k = 10$ nächsten Nachbarn gefunden werden können, da zweimal $20 > k$ und einmal $10 = k$ Daten mit `NA` kodiert werden.

Bei den Differenzen zum Goldstandard liegt der Median des herausragenden Falles sogar um das fast 40-fache höher als derjenige des Falls „`-NA`“. Dieser liegt eng um Null verteilt, d. h. der Conditional Tree Forest kann auch in einem reduzierten Datensatz die Einflussgrößen gut erkennen. Der Maximalwert des mit `NA` besetzten Test-Datensatzes mit Imputation liegt bei ca. 0.7. Die Boxen ohne Imputation sind auch bei den Differenzen besser um Null gelagert als die Boxplots mit Imputation.

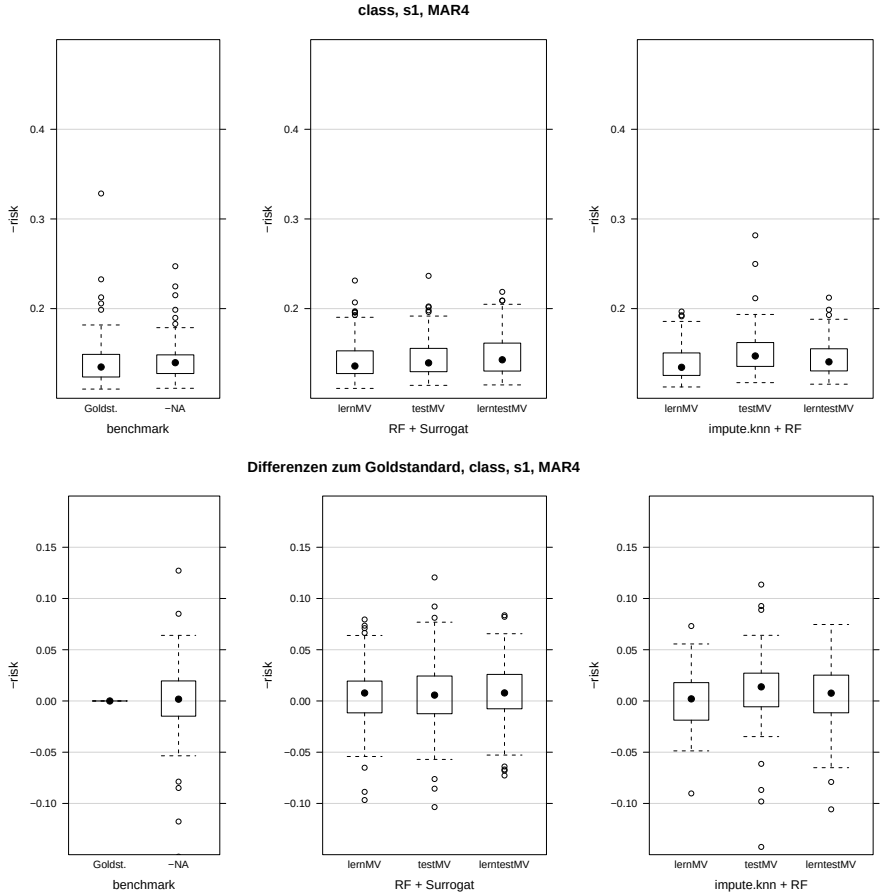

Abbildung 5.4.: Verteilung der negativen Binomial-Log-Likelihood sowie der Differenzen zum Goldstandard bei gleichmäßig hohen Korrelationen und fehlenden Werten nach MAR4 (symmetrische Trunkierung) [s1 entspricht Σ_1]

MAR 4

Ein sehr gutes Ergebnis lässt sich dafür bei MAR4 erzielen. Die symmetrische Trunkierung scheint die Imputation nicht zu stören. Die Mediane liegen höchstens ca. 0.008 über dem Median des reduzierten Falls. Dieser liegt selbst sehr nahe beim Goldstandard, die anderen Boxplots liegen leicht erhöht. Der Fehler beim Test-Datensatz mit fehlenden Werten und Imputation ist wiederholt höher, ebenso der Fall mit fehlenden Werten in Lern- und Test-Datensatz mit Surrogat-Variablen. Ein signifikanter Unterschied zu Goldstandard und „`-NA`“ ergibt sich nur beim Test-Datensatz mit Imputation. Der Vergleich zwischen Surrogat-Variablen und Imputation ergibt kein signifikantes Ergebnis.

Bei den Differenzen zum Goldstandard zeigt sich das bis jetzt beste Ergebnis, obwohl alle Mediane betragsmäßig größer sind als der des reduzierten Falls. Außerdem ist

die Streuung der Boxplots mit und ohne Imputation größer als die Streuung des reduzierten Falls. Doch verglichen mit den vorher gehenden Simulationen liegt diese Methode bzw. deren Fehler bis jetzt am nächsten an der Null.

MCAR

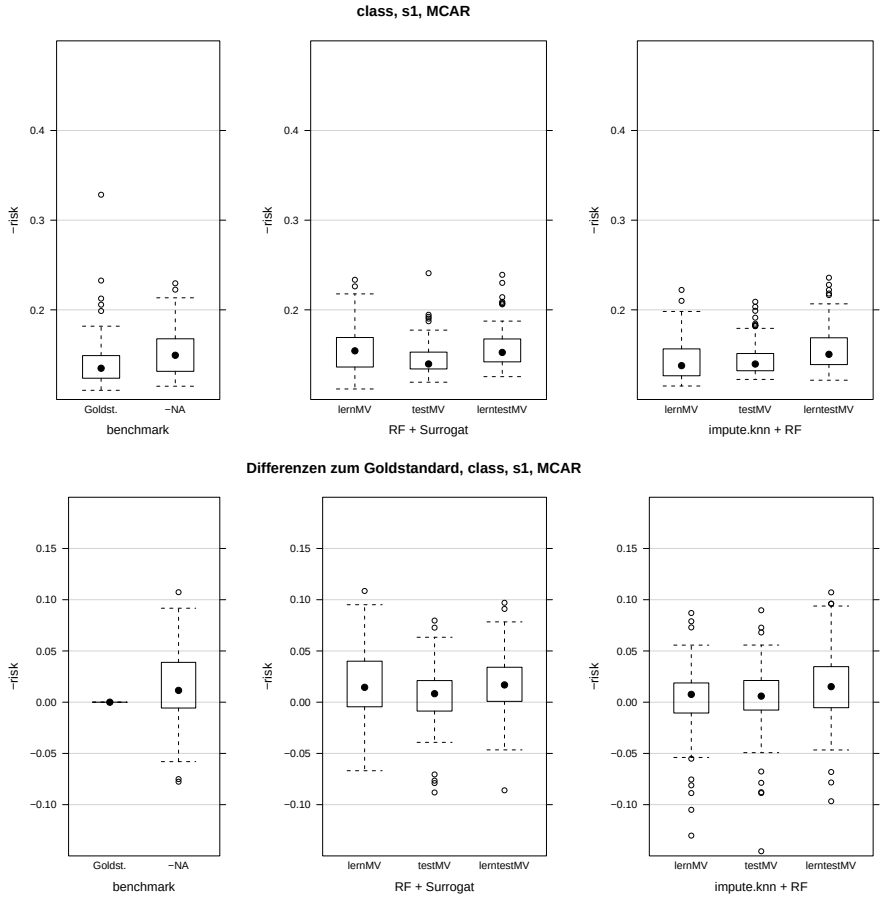

Abbildung 5.5.: Verteilung der negativen Binomial-Log-Likelihood sowie der Differenzen zum Goldstandard bei gleichmäßig hohen Korrelationen und fehlenden Werten nach MCAR (vollständig zufällig) [s1 entspricht Σ_1]

Bei vollständig zufälliger Streichung von Beobachtungen sind die Lern-Datensätze ohne Imputation und der Fall mit fehlenden Werten in beiden Datensätzen (mit Imputation) gesamt etwas schlechter als der reduzierte Fall ohne `NA`. Zusätzlich ist noch der Fall mit fehlenden Werten in beiden Datensätzen (ohne Imputation) im Median schlechter als der reduzierte Fall. Allerdings resultiert in keinem der drei genannten Szenarien eine signifikante Abweichung im Mittelwert vom reduzierten Fall.

Die anderen drei Fälle sind signifikant verschieden vom Fall „-NA“ und ergeben eine kleinere Box, also einen kleineren Interquartilsabstand als der reduzierte Fall und damit eine kleinere Streuung. Bis auf die beiden Test-Datensätze ist der Interquartilsabstand des Goldstandards am geringsten, d. h. die Höhe der Box. Für die Methoden mit den fehlenden Werten in den Test-Datensätzen ist auch kein signifikanter Unterschied zum Goldstandard zu verzeichnen. Die Fehlerverteilung der Lern-Datensätze unterscheidet sich beim Vergleich zwischen Surrogat-Variablen und Imputation signifikant im Mittelwert.

In den Differenzen zum Goldstandard wird der Median des reduzierten Falls von den drei oben genannten Fällen ebenfalls überschritten. Die Mediane des Test-Datensatzes ohne und mit Imputation und des Lern-Datensatzes mit Imputation sind niedriger als 0.01, also ziemlich nahe an Null. Die Streuung der Differenzen ist allerdings kleiner als beim reduzierten Fall ohne `NA`.

5.1.2. Blockweise hohe Korrelationen

In den folgenden Simulationen mit der Korrelationsmatrix $\mathbf{\Sigma}_2$, in der $\boldsymbol{x}_1$ bis $\boldsymbol{x}_3$ die eine Gruppe bilden und $\boldsymbol{x}_4$ und $\boldsymbol{x}_5$ die andere, ist das Risiko einer Fehleinschätzung generell höher. Dies hängt damit zusammen, dass durch die blockweise Anordnung der Korrelationen die Bildung von Surrogat-Variablen und auch die Berechnung der k nächsten Nachbarn nicht so gut funktioniert wie bei (mehr oder weniger) gleichmäßiger hoher Korrelation.

MAR 1

In dieser Situation liegen die Mediane aller Fälle zwischen dem des Goldstandard und dem des reduzierten Falls. Außerdem liegt die Box des Falls „-NA“ hier am höchsten, was zu signifikanten Abweichungen sämtlicher Szenarien von diesem reduzierten Fall führt. Dieser hat auch den geringsten Interquartilsabstand, gefolgt von fehlenden Werten in den Lern-Datensätzen mit Imputation und fehlenden Werten in Lern- und Test-Datensätzen ohne Imputation. Zudem ist der Fall der Lern-Datensätze mit `NA` und mit Imputation nach dem Goldstandard am besten. Beide Situationen mit den fehlenden Werten in den Lern-Datensätzen ergeben keinen signifikanten Unterschied zum Goldstandard. Der Vergleich zwischen Surrogat-Variablen und Imputation bei den Lern-Datensätzen ist dagegen signifikant im Mittelwert.

Dies zeigt sich auch bei den Differenzen zum Goldstandard, da dort der Median der Lern-Datensätze mit Imputation sehr nahe Null liegt (ca. 0.001). Allerdings ist die Streuung relativ groß. Die Streuungen der Differenzen sind annähernd derjenigen des reduzierten Falls ähnlich. Dessen oberer Zaun bzw. die gesamte Box liegt wieder am höchsten, was positiv ist. Die Mediane liegen alle unterhalb von ca. 0.019.

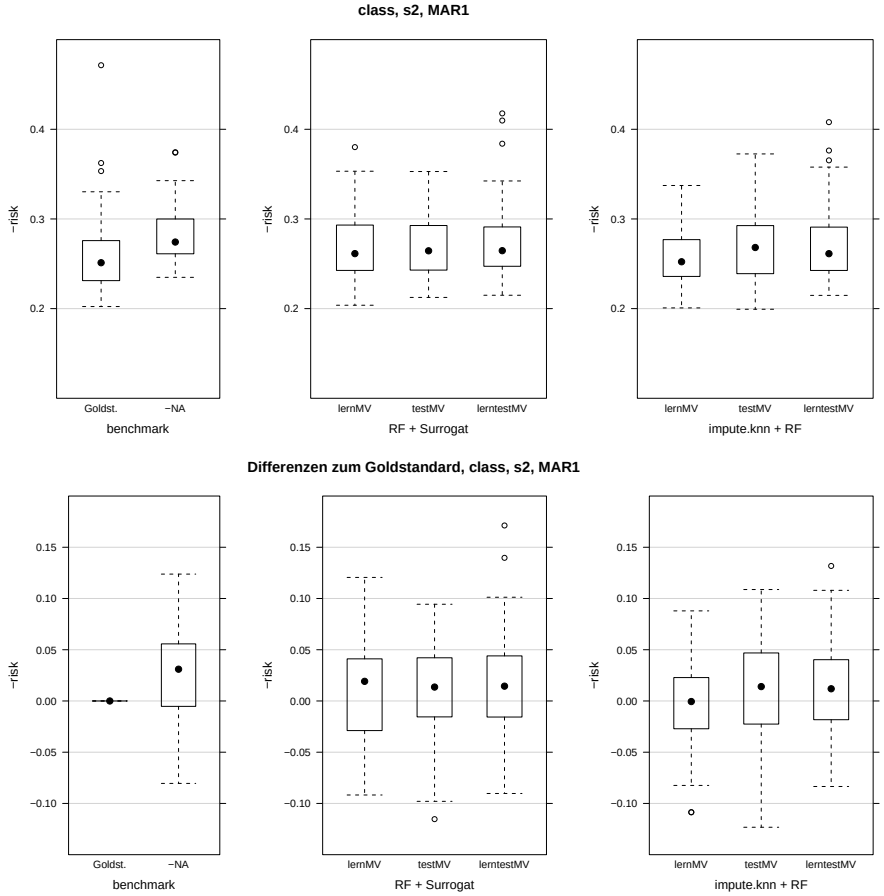

Abbildung 5.6.: Verteilung der negativen Binomial-Log-Likelihood sowie der Differenzen zum Goldstandard bei blockweise hohen Korrelationen und fehlenden Werten nach MAR1 (Bildung von Rängen) [s2 entspricht Σ_2]

MAR 2

Während bei der hohen Korrelation durch $\mathbf{\Sigma}_1$ der Fall mit fehlenden Werten in den Lern-Datensätzen und im Test-Datensatz ohne Imputation relativ schlecht war, ergibt sich hier eine ziemlich gute Verteilung des Güte-Maßes. Sämtliche Szenarien sind signifikant verschieden vom reduzierten Fall ohne `-NA` und nur für den kombinierten Fall mit Imputation ergibt sich ein Unterschied zum Goldstandard. Die Box des vorher kritischen kombinierten Falles ohne Imputation liegt ungefähr auf der Höhe des Goldstandards. Die Beträge aller Mediane befinden sich zwischen dem Goldstandard und dem reduzierten Fall, ebenso die Quartile. Außerdem ist der Lern-Datensatz mit fehlenden Werten und Imputation sehr nahe am Goldstandard.
Beim Vergleich zwischen Surrogat-Variablen und Imputation ergibt sich in keinem Szenario ein signifikanter t-Test.

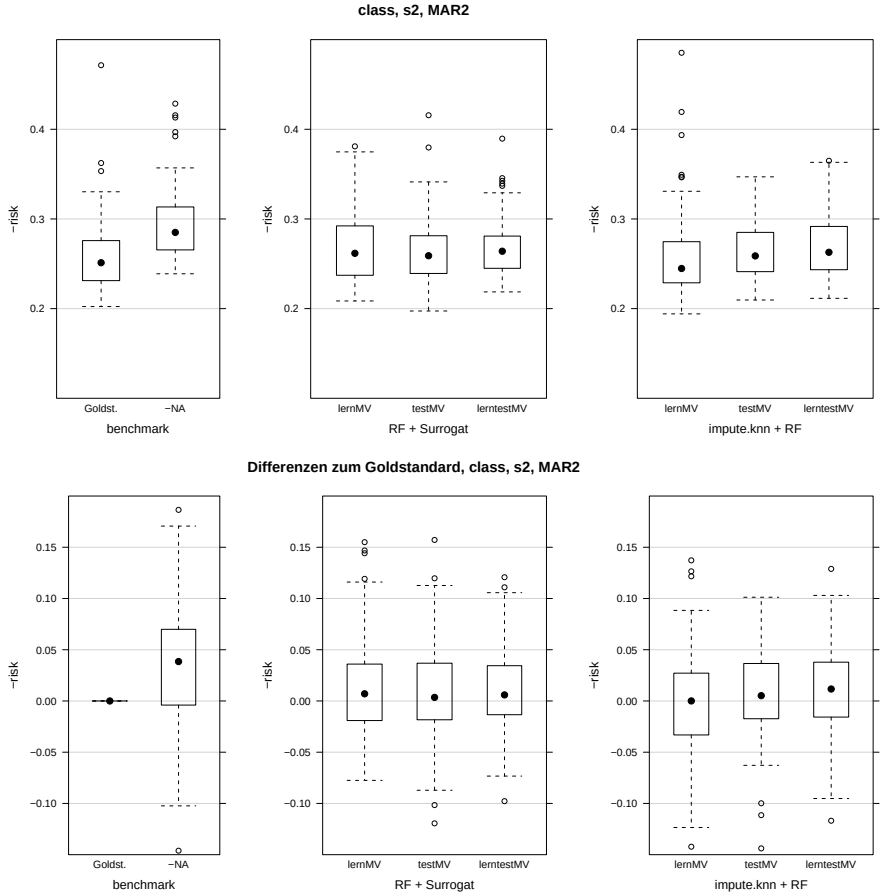

Abbildung 5.7.: Verteilung der negativen Binomial-Log-Likelihood sowie der Differenzen zum Goldstandard bei blockweise hohen Korrelationen und fehlenden Werten nach MAR2 (Bildung von zwei Risikogruppen) [s2 entspricht Σ_2]

Die Mediane der Differenzen zum Goldstandard sind – abgesehen vom reduzierten Fall – nahe der Null (maximale Abweichung: ca. 0.014). Der Median der Lern-Datensätze mit Imputation ist im Prinzip Null. Dadurch, dass der Fall „`-NA`“ mehr oder weniger dem nach oben verschobenen Goldstandard entspricht, ist die Streuung seiner Differenzen relativ groß. Sie übertreffen die anderen Fälle bezüglich der Höhe des Fehlers, der Höhe der Box und der Länge der Zäune. Wenn man die Differenzen betrachtet, ergibt sich für diese Simulation ein sehr gutes Bild.

MAR 3

Wie auch schon bei $\mathbf{\Sigma}_1$ ist die Fehler-Verteilung des Test-Datensatzes mit fehlenden Werten und mit Imputation nach oben verzerrt. Sie liegt leicht unter dem reduzierten Fall, was bedeutet, dass sie zwar nicht gut, aber etwas besser als dieser ist. Der

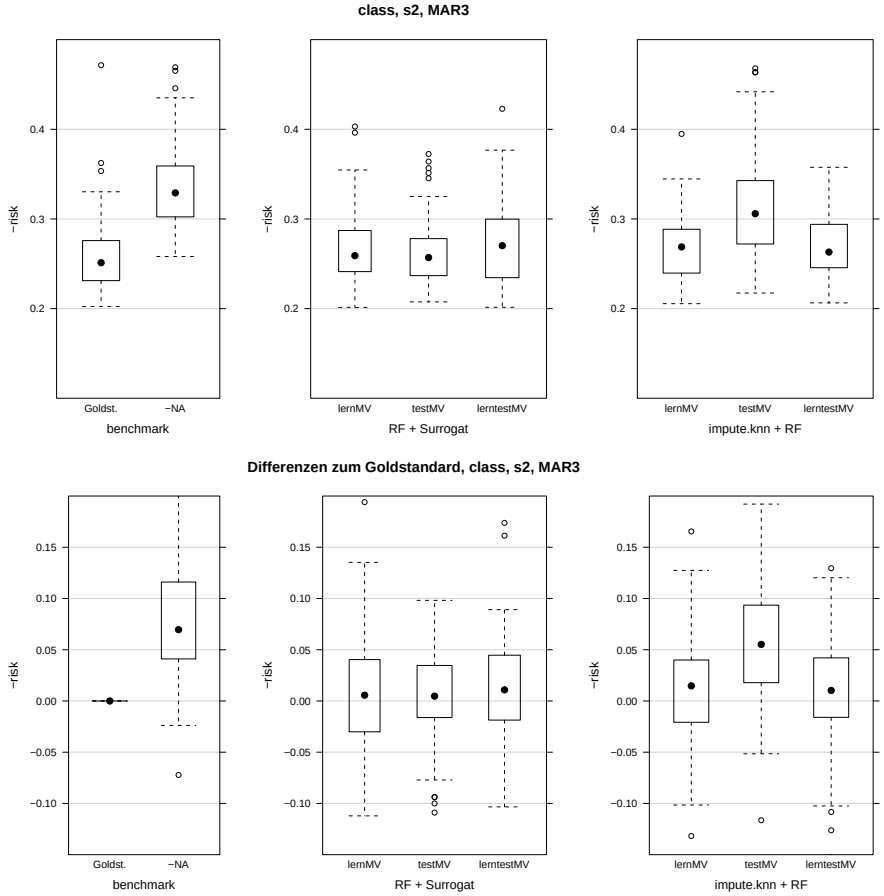

Abbildung 5.8.: Verteilung der negativen Binomial-Log-Likelihood sowie der Differenzen zum Goldstandard bei blockweise hohen Korrelationen und fehlenden Werten nach MAR3 (rechtsseitige Trunkierung) [s2 entspricht Σ_2]

Unterschied ist signifikant, sowohl zum reduzierten Fall „`-NA`“ als auch zum Goldstandard.
Das Minimum des Goldstandards wird von den anderen Fällen zweimal unterschritten: bei den Lern-Datensätzen ohne Imputation (Vergleich im Mittelwert nicht signifikant) und bei den Lern- und Test-Datensätzen ohne Imputation (im Mittelwert signifikant verschieden). Die Mediane liegen allerdings auch hier alle zwischen dem des Goldstandards und dem des reduzierten Falls ohne `NA`.
Beim Vergleich zwischen Surrogat-Variablen und Imputation ergibt sich bei den Test-Datensätzen ein signifikanter Unterschied.

In diesem Szenario ergeben sich zum ersten Mal mehr als fünf ungewollte fehlende Werte durch die Funktion `cforest`. Zum Einen fehlen in den Lern-Datensätzen ohne Imputation 22 Beobachtungen (von 100), zum Anderen konnten in den Lern- und Test-Datensätzen ohne Imputation 18 Beobachtungen nicht ausgewertet werden.

Betrachtet man wieder die Differenzen zum Goldstandard, wird für die Fälle ohne Imputation, d. h. bei Verwendung von Surrogat-Variablen, ein gutes Ergebnis erzielt. Die Mediane liegen weniger als 0.011 von der Null entfernt und die Interquartilsabstände liegen weit unter dem des Falls „`-NA`“.
Die Fälle mit Imputation bringen eine weiter nach oben verschobene Fehlerverteilung mit größerer Streuung hervor: Die Mediane weichen um mindestens 0.010 von der Null ab und die Zäune sind länger als bei den Beobachtungen ohne Imputation.

Zu diesem Ergebnis kam man ja schon bei $\mathbf{\Sigma}_1$, da immer die Beobachtungen mit den größten Werten gestrichen werden und somit die Berechnung der Surrogat-Variablen und der nächsten Nachbarn noch zusätzlich zur blockweisen Korrelation behindert werden.

MAR 4

Für fehlende Werte nach MAR4 liegen sämtliche Boxen unterhalb des reduzierten Falls, somit auch die Mediane. Die Mittelwerte weichen ebenfalls signifikant ab. Der Fall der Lern-Datensätze mit Imputation ist dem Goldstandard sehr ähnlich (kein signifikanter Unterschied). Für die Lern- und Testdatensätze mit gleichzeitig fehlenden Werten resultiert ein etwas erhöhte Streuung bzw. längere Zäune als für den Rest. Außerdem ergibt sich für diesen Fall und die Test-Datensätze ein signifikanter Unterschied zwischen Imputationsverfahren und Surrogat-Variablen.

In den Differenzen wird der Median des reduzierten Falls nicht überstiegen, obwohl er selbst nicht sehr hoch liegt. Dies bedeutet, zusammen mit der fehlenden Überschreitung, dass der Conditional Tree Forest die relevanten Variablen und ihren Einfluss gut erkennen kann.
Die Mediane der Fälle ohne Imputation liegen nicht weiter als ca. 0.006 von der Null entfernt. Auffällig ist, dass die Lern-Datensätze mit fehlenden Werten in beiden *Forest*-Varianten einen Median sehr nahe Null haben.

MCAR

Für diese Simulation ergibt sich nur eine geringe Schwankung. Der Fall „`-NA`“ wird nur von den Lern- und Test-Datensätzen mit `NA` und ohne Imputation leicht überschritten (in Bezug auf Median und Quartile). Der schon mehrmals sehr gute Lern-Datensatz mit Imputation ist auch wieder etwas besser als der Goldstandard. Beide Überschreitungen der Bezugsmarken sind aber nicht signifikant. Der Vergleich zwischen den Mittelwerten der Fehlerverteilungen bei Verwendung von Surrogat-Variablen und bei Imputation ist dagegen signifikant für die Lern-Datensätze und die kombinierten Fälle.
Eventuell erwähnenswert ist noch die Tatsache, dass die Fälle mit Imputation Ausreißer in geringeren Höhen haben.

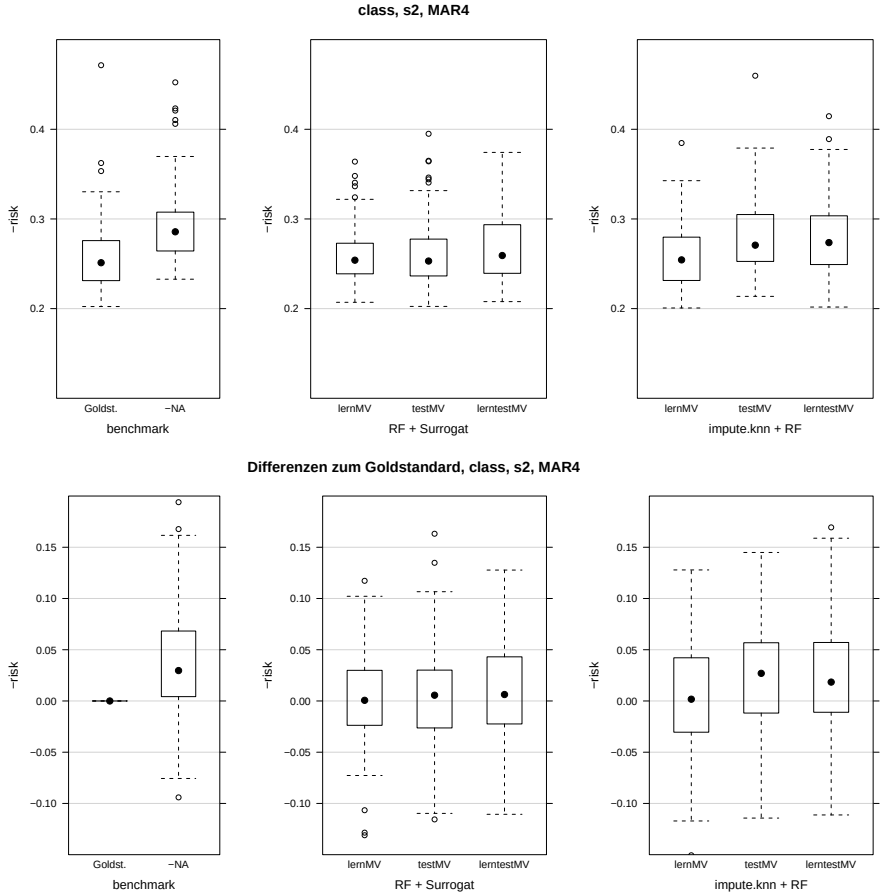

Abbildung 5.9.: Verteilung der negativen Binomial-Log-Likelihood sowie der Differenzen zum Goldstandard bei blockweise hohen Korrelationen und fehlenden Werten nach MAR4 (symmetrische Trunkierung) [s2 entspricht Σ_2]

Demzufolge sind auch die Differenzen zum Goldstandard bei erfolgter Imputation besser, d. h. näher an der Null. Sie weichen betragsmäßig um maximal 0.011 von der Null ab, während ohne Imputation dies die Mindestmarke darstellt. Die Boxplots der Differenzen weisen außerdem eine relativ große Streuung auf, obwohl die eigentlichen Werte kaum schwanken. Dennoch lässt sich auch hier sagen, dass der Conditional Tree Forest an sich den datengenerierenden Prozess gut lernt, wie man an der doch recht niedrigen Abweichung der Mediane von der Null erkennen kann.

5.1.3. Gleichmäßige, niedrige Korrelationen

In den folgenden Simulationen mit gleichmäßigen, aber niedrigen Korrelationen von 0.1 ist das Risiko einer Fehleinschätzung noch höher. Die Ursache dafür sind nun die niedrigen Korrelationen, auf Grund derer die Bildung von Surrogat-Variablen und

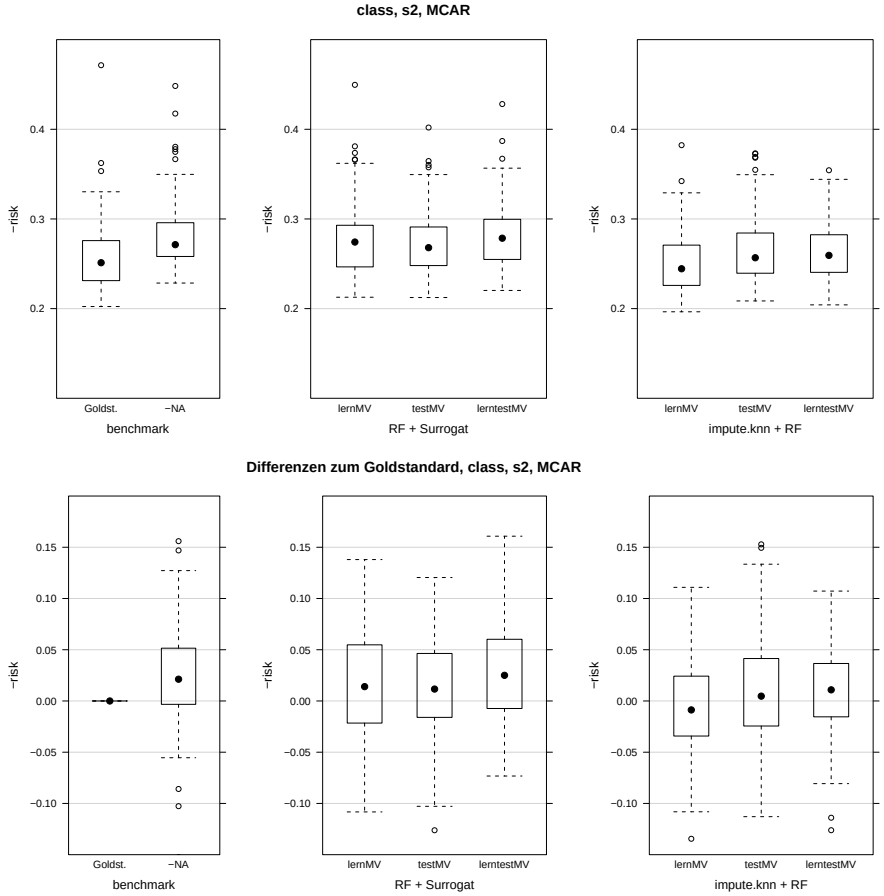

Abbildung 5.10.: Verteilung der negativen Binomial-Log-Likelihood sowie der Differenzen zum Goldstandard bei blockweise hohen Korrelationen und fehlenden Werten nach MCAR (vollständig zufällig) [s2 entspricht Σ_2]

die Suche nach den k nächsten Nachbarn ebenfalls nicht erfolgreich ist.

MAR 1

In diesem Szenario lässt sich ein steigendes Risiko der Fehleinschätzung feststellen. Von den Lern-Datensätzen mit fehlenden Werten wandern die Boxplots über die Test-Datensätze zum kombinierten Fall nach oben. Ohne Imputation verändert sich dabei die Streuung kaum bzw. wird eher geringer. Mit Imputation liegt eine größere Streuung vor, die sich ebenfalls wie die „Höhenlage“ steigert. Aus diesen Gründen wird der reduzierte Fall dreimal bezüglich des Medians überstiegen: vom kombinierten Fall mit Surrogat-Variablen und mit Imputation und vom Test-Datensatz mit Imputation. Im Mittelwert ist eine signifikante Abweichung bei allen Fällen, außer dem Test-Datensatz mit Surrogat-Variablen, festzustellen. Es fällt wiederum auf,

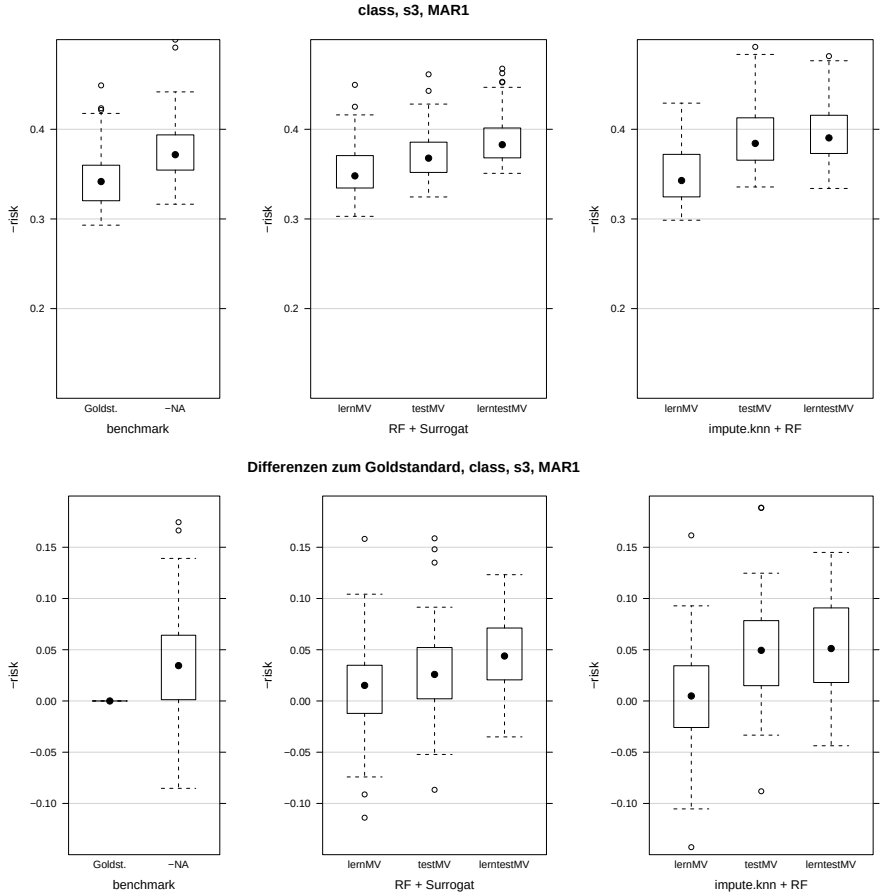

Abbildung 5.11.: Verteilung der negativen Binomial-Log-Likelihood sowie der Differenzen zum Goldstandard bei gleichmäßig niedrigen Korrelationen und fehlenden Werten nach MAR1 (Bildung von Rängen) [s3 entspricht Σ_3]

dass die Fälle mit erfolgter Imputation größere Fehler ergeben, teilweise eben sogar schlechtere als der reduzierte Fall. Der Vergleich mit den Fehlern bei Verwendung von Surrogat-Variablen ist deswegen für die Test-Datensätze und die kombinierten Fälle signifikant.

Ein relativ schlechtes Bild zeigt sich auch bei Betrachtung der Differenzen zum Goldstandard. Die Mediane liegen bis auf die Lern-Datensätze mit Imputation und dem Test-Datensatz ohne Imputation *über* dem des reduzierten Falls ohne `NA`. Generell haben die Boxen eine Tendenz ins Positive (was bedeutet, dass der jeweilige Fall schlechter als der Goldstandard ist), allerdings sind die Boxen nicht sehr symmetrisch um Null gelagert, was zu einem erstrebenswerten Median um Null führen würde. Allein die Lern-Datensätze mit Imputation führen in die Nähe eines guten Ergebnisses. Deren Mittelwert ist in den ursprünglichen Fehlerdaten nicht signifikant verschieden vom Mittelwert des Goldstandards.

MAR 2

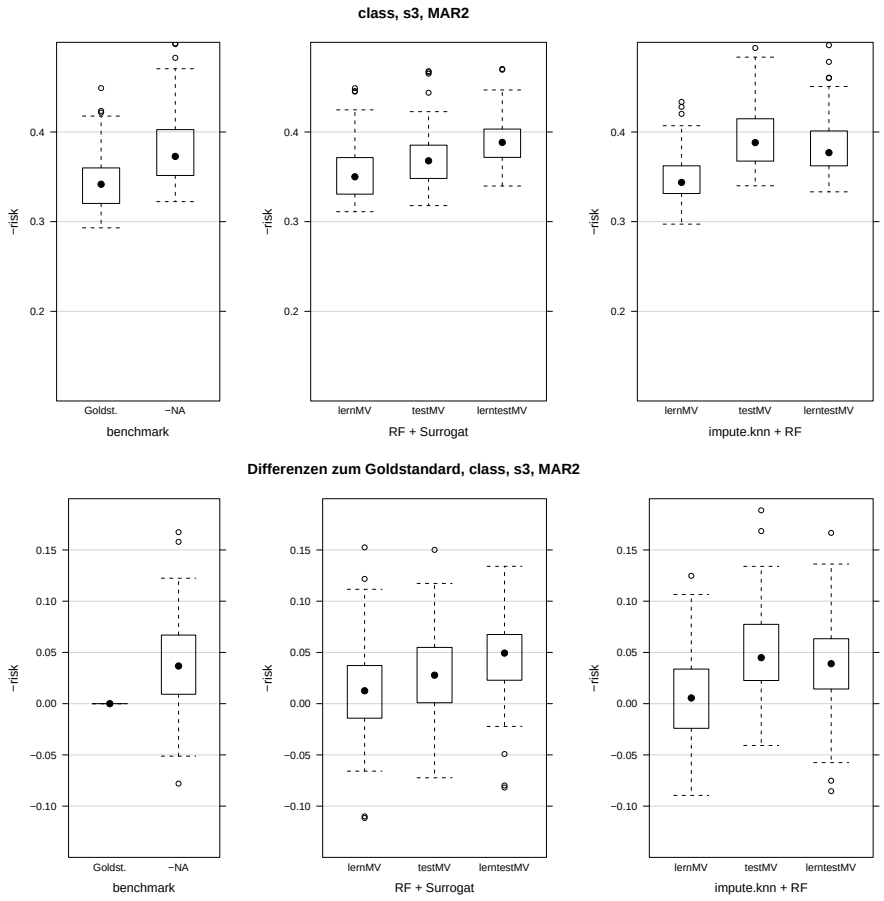

Abbildung 5.12.: Verteilung der negativen Binomial-Log-Likelihood sowie der Differenzen zum Goldstandard bei gleichmäßig niedrigen Korrelationen und fehlenden Werten nach MAR2 (Bildung von zwei Risikogruppen) [s3 entspricht Σ_3]

Der reduzierte Fall wird hier vom Test-Datensatz mit Imputation und dem kombinierten Fall ohne Imputation überschritten, was die Box angeht. Zudem ist der Median vom kombinierten Fall mit Imputation höher als der des reduzierten Falls. Die Fehler der kombinierten Fälle sind als einzige Situationen nicht signifikant im Mittelwert verschieden vom Fehler des Fall „`-NA`“. Wie immer recht gute Boxplots ergeben die Lern-Datensätze mit fehlenden Werten, wobei bei Verwendung von Surrogat-Variablen der t-Test einen signifikanten Unterschied im Mittelwert zum Goldstandard zeigt. Bei fehlenden Werten in den Test-Datensätzen ist der Mittelwert bei einem *Forest* mit Surrogat-Variablen signifikant verschieden vom Mittelwert bei Imputation.

Auch der Median der Differenzen vom Fall „`-NA`“ ist nicht der höchste: der kombinierte Fall ohne und mit Imputation sowie der Test-Datensatz mit Imputation sind etwas höher. Die beiden Fälle der Lern-Datensätze mit und ohne Imputation sind relativ nahe an der Null (ca. 0.012 und ca. 0.006). Aber diese Differenzen streuen stark, was wohl daran liegt, dass die ursprünglichen Risiken im Gegensatz zum Goldstandard mehr oder weniger verschoben sind und nicht bis zu seinem unteren Ende streuen.

MAR 3

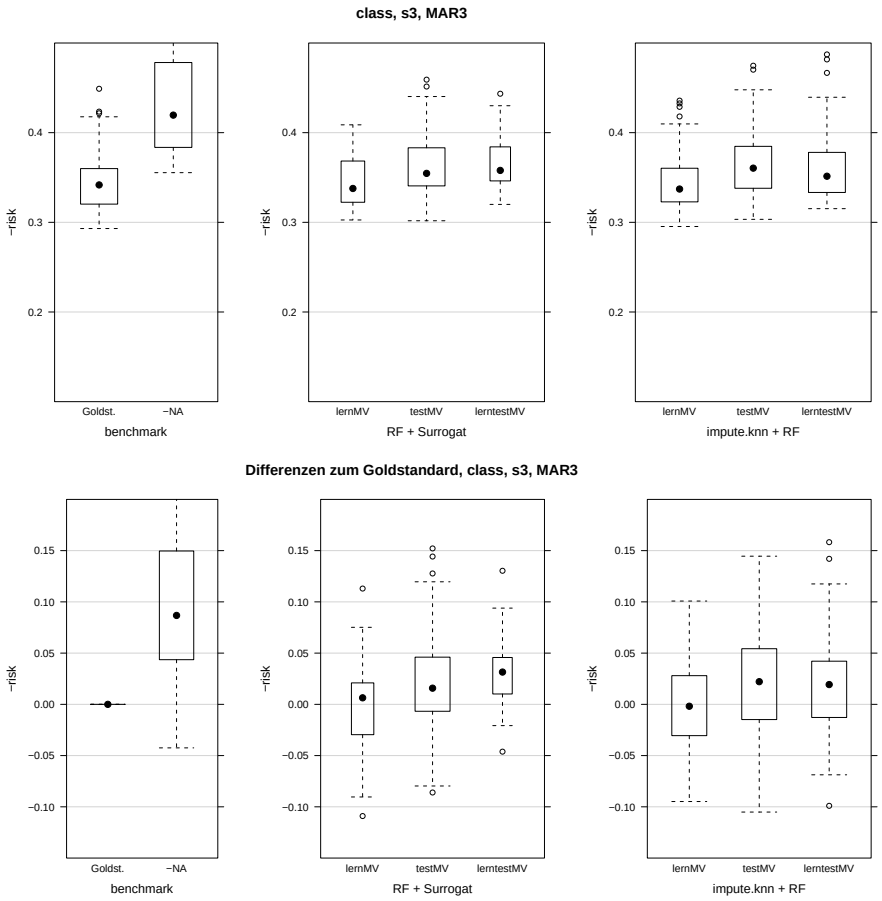

Abbildung 5.13.: Verteilung der negativen Binomial-Log-Likelihood sowie der Differenzen zum Goldstandard bei gleichmäßig niedrigen Korrelationen und fehlenden Werten nach MAR3 (rechtsseitige Trunkierung) [s3 entspricht Σ_3]

Auch bei niedriger Korrelation ergeben sich bei MAR3 unfreiwillige fehlende Beobachtungen durch die Funktion `cforest`. In den Lern-Datensätzen ohne Imputation wurden 59 Beobachtungen mit `NA` kodiert und in den Lern- und Test-Datensätzen

ohne Imputation fehlen sogar 69 Werte. Dies ist deutlich mehr als die Hälfte!

Dafür schneidet diesmal der sonst bei MAR3 so schlechte Test-Datensatz mit Imputation besser ab, er liegt sogar niedriger als der reduzierte Fall ohne `NA` und unterscheidet sich im Mittelwert wie die anderen Fälle auch signifikant von „`-NA`“. Dieser ergibt sich hier als die schlechteste Situation – so, wie es sein soll. Die anderen Boxen haben sogar das obere Quartil in Höhe des unteren Quartils des reduzierten Falls. Die Lern-Datensätze schneiden wiederholt gut ab, übertreffen im Median sogar ein wenig den Goldstandard und ergeben keine signifikante Abweichung im Mittelwert. Ebenfalls nicht signifikant ist in allen Fällen der Test zum Vergleich zwischen Surrogat-Variablen und Imputation.

Die Boxplots der Differenzen zum Goldstandard liegen niedriger als noch bei MAR2. Die Mediane der Lern-Datensätze sind sehr nahe an der Null (ca. 0.006 bzw. -0.002), der Median des reduzierten Falls wird immer unterschritten (um mindestens 0.056). Hinsichtlich der Streuung unterscheiden sich die Fälle nicht sehr.

MAR 4

Die symmetrische Trunkierung führt zu leicht besseren Ergebnissen des Conditional Tree Forests als die rechtsseitige. Der Fall „`-NA`“ ist näher am Goldstandard, aber dennoch liegen die meisten Mediane dazwischen. Eine Ausnahme bilden die Lern-Datensätze und der kombinierte Fall mit Imputation, diese Mediane sind sogar noch unter dem Goldstandard (aber kein signifikanter Unterschied im Mittelwert). Ansonsten ähneln sie dem Goldstandard stark.
Im Vergleich zwischen Surrogat-Variablen und *knn*-Imputation zeigen sich signifikante Unterschiede bei den Test-Datensätzen und bei den kombinierten Fällen.

Das gute Ergebnis spiegeln auch die Differenzen zum Goldstandard wider: Die Mediane der Imputationsfälle sind maximal ca. 0.021 von der Null entfernt. Dies beträgt etwas mehr als die Hälfte des reduzierten Falls.

MCAR

Bei vollständig zufälligem Fehlen rutschen der reduzierte Fall ohne `NA` und der Goldstandard noch näher zusammen. Deswegen ist der kombinierte Fall ohne Imputation und der Test-Datensatz mit fehlenden Werten und mit Imputation schlechter als der reduzierte Fall (beide allerdings nicht signifikant im Mittelwert). Zudem kann man in den Fällen mit Surrogat-Variablen die Steigerung des Risikos einer Fehleinschätzung bei kleiner werdenden Zäunen feststellen. Die Verwendung von Surrogat-Variablen ist bei den Lern-Datensätzen und bei den Test-Datensätzen signifikant verschieden von der Imputation. Für die imputierten Lern-Datensätze ergibt sich keine im Mittelwert schlechtere Fehlerverteilung als der Goldstandard.

Weniger gute Ergebnisse zeigen allerdings die Differenzen zum Goldstandard. Die

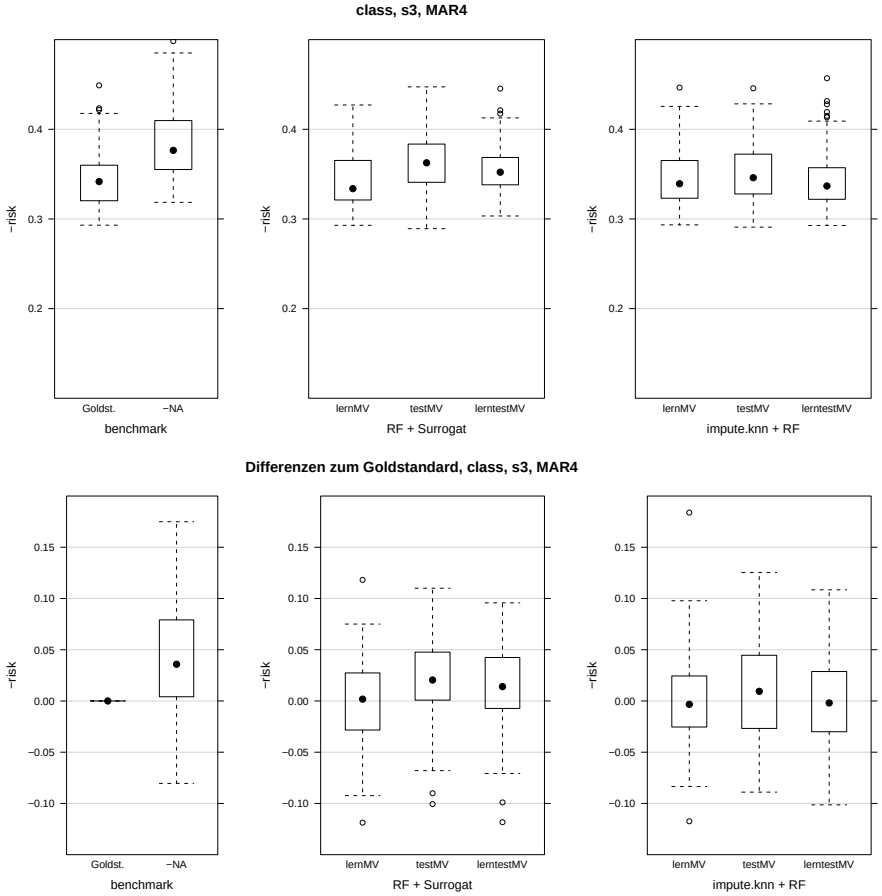

Abbildung 5.14.: Verteilung der negativen Binomial-Log-Likelihood sowie der Differenzen zum Goldstandard bei gleichmäßig niedrigen Korrelationen und fehlenden Werten nach MAR4 (symmetrische Trunkierung) [s3 entspricht Σ_3]

Boxen und Mediane gehören mit zu den höchsten der bis jetzt betrachteten. So befinden sich die Mediane mindestens 0.011 von der Null-Linie entfernt. Diese Marke war oftmals bereits das Maximum! Lediglich die Boxen der Lern-Datensätze beinhalten die Null überhaupt. Alle anderen Boxen liegen darüber und zusätzlich überrunden deren Mediane den „`-NA`“-Median.
Eventuell kann es daran liegen, dass der Abstand zwischen dem reduzierten Fall und Goldstandard doch recht gering ist und somit auch kleine Verschiebungen nach oben stärker ins Gewicht fallen.

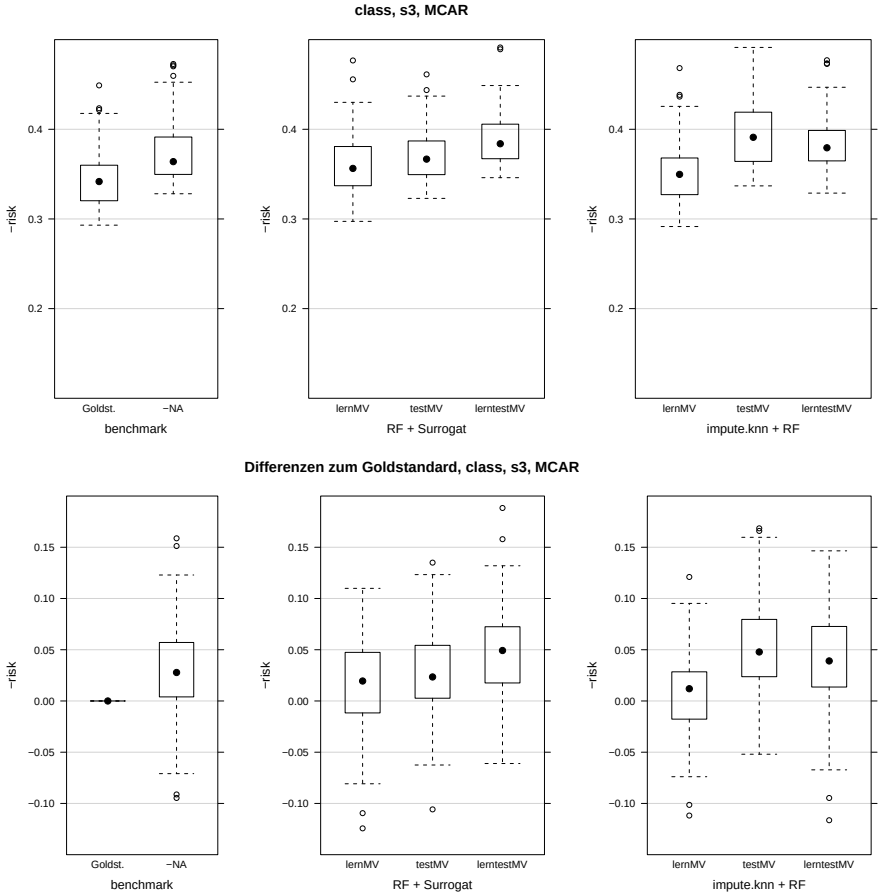

Abbildung 5.15.: Verteilung der negativen Binomial-Log-Likelihood sowie der Differenzen zum Goldstandard bei gleichmäßig niedrigen Korrelationen und fehlenden Werten nach MCAR (vollständig zufällig) [s3 entspricht Σ_3]

5.2. Simulation zur Regression

Mit dem hier vorliegenden stetigen Response soll die Leistung der Funktion `cforest` im Bereich der Regressionsbäume überprüft werden. Dafür wurde die Funktion `dgp2` verwendet, welche wiederum das Regressionsmodell „Friedman1“ benutzt (siehe (4.1.2)).

Das Vorgehen entspricht demjenigen für Klassifikationsbäume: Alle drei Korrelationsmatrizen $\mathbf{\Sigma}_k$ wurden betrachtet sowie die fünf verschiedenen Methoden, um fehlende Werte einzustreuen.

5.2.1. Gleichmäßige, hohe Korrelationen

MAR 1

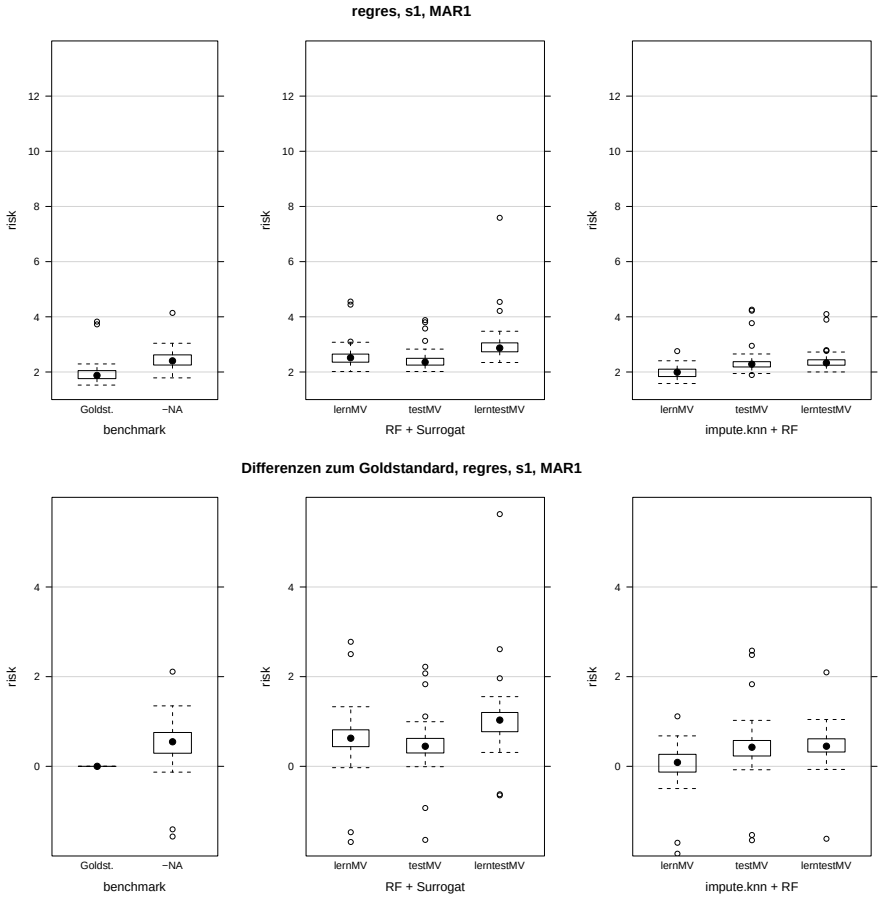

Abbildung 5.16.: Verteilung der mittleren quadratischne Abweichung sowie der Differenzen zum Goldstandard bei gleichmäßig hohen Korrelationen und fehlenden Werten nach MAR1 (Bildung von Rängen) [s1 entspricht Σ_1]

In dieser ersten Simulation zur Regression reicht nur ein Fall an den Goldstandard heran. Sämtliche Boxen sind höher gelagert als dieser, bis auf die Lern-Datensätze mit Imputation. Die beiden Mediane unterscheiden sich nur um ca. 0.116. Der Mittelwertsvergleich der Fehlerverteilung der imputierten Lern-Datensätze mit derjenigen des Goldstandards fällt als einziger nicht signifikant aus. Die Lern-Datensätze ohne Imputation dagegen sind im Median sogar höher als der reduzierte Fall und unterscheiden sich im Mittelwert sogar signifikant von diesem.
Der Test-Datensatz mit fehlenden Werten und ohne Imputation ist in etwa auf der Höhe des reduzierten Falls. Der kombinierte Fall ohne Imputation liegt darüber. Dagegen sind die drei Fälle mit Imputation besser als der Fall „`-NA`“.

Der Vergleich zwischen den Mittelwerten bei Imputation und bei Surrogat-Variablen ist für die Lern-Datensätze und den kombinierten Fall signifikant.

Bei den Differenzen zum Goldstandard sind die drei Fälle mit Imputation ebenfalls besser als diejenigen ohne. Während die Boxplots für die Fälle ohne Imputation sogar mit ihren Zäunen nur knapp die Null beinhalten, liegt der Median der Lern-Datensätze mit Imputation nahe bei der Null. Die beiden anderen Boxplots mit Imputation entsprechen in etwa dem besten ohne.
Der Median des kombinierten Falles ohne Imputation, der auch bei den ursprünglichen Risiken über dem reduzierten Fall liegt, beträgt bei den Differenzen beinahe das Doppelte des Medians dieses Falls.

MAR 2

Das Risiko einer Fehleinschätzung bei fehlenden Werten nach MAR2, also einer Bildung von zwei Streich-Risiko-Gruppen, liegt insgesamt etwas unter dem von MAR1, aber in der Tendenz ist es in etwa gleich. Die Lern-Datensätze mit Imputation sind etwa so gut wie der Goldstandard, allerdings ist diesmal der Median nicht besser. Der zugehörige Mittelwert ist wieder als einziger nicht vom Goldstandard verschieden. Die Lern-Datensätze, deren *Forests* mit Surrogat-Variablen berechnet wurden, unterscheiden sich als einzige nicht signifikant vom reduzierten Fall „`-NA`“. Der kombinierte Fall ohne Imputation ist aber im Median schlechter als der reduzierte Fall und die drei Fälle mit Imputation sind besser. Wenn man deren Fehler mit dem der Fälle mit Surrogat-Variablen vergleicht, ergibt sich nur bei den Test-Datensätzen *keine* signifikante Abweichung.

Die Situationen mit Imputation zeigen sich auch bei den Differenzen erfolgreicher als die Fälle ohne Imputation. Der Median der guten Lern-Datensätze mit Imputation liegt sehr nahe bei Null und die beiden anderen Mediane liegen unter dem des reduzierten Falls. Die Null wird von den Fällen mit Imputation mindestens von den Zäunen der Boxplots eingeschlossen.
Der schlechte kombinierte Fall ohne Imputation liegt auch bis auf eine Beobachtung oberhalb der Null und schließt somit diese nicht einmal mit den Zäunen ein. Sein Median beträgt diesmal sogar mehr als das Doppelte desjenigen des reduzierten Falls.

Hier sind auch wieder fehlende Beobachtungen zu erwähnen: In den Lern-Datensätzen ohne Imputation sind neun `NA`'s vorhanden und im kombinierten Fall ohne Imputation fehlen sieben Werte. Wären diese vorhanden, würden sie dennoch das relativ schlechte Ergebnis ihrer Kategorie vermutlich nicht verbessern.

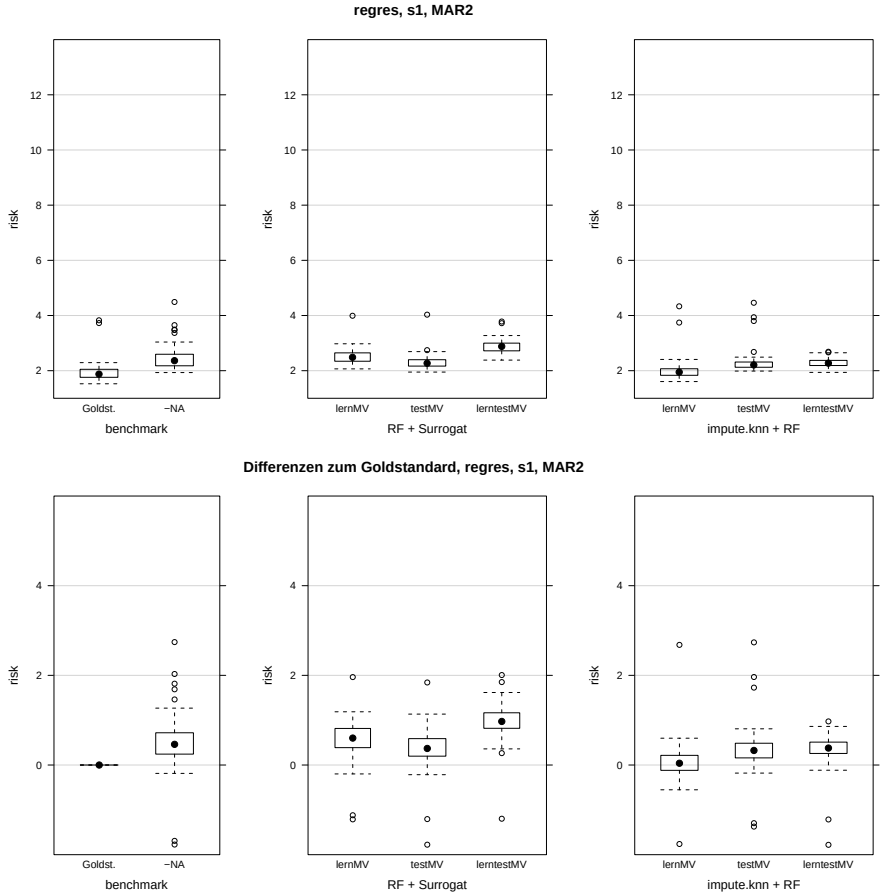

Abbildung 5.17.: Verteilung der mittleren quadratischen Abweichung sowie der Differenzen zum Goldstandard bei gleichmäßig hohen Korrelationen und fehlenden Werten nach MAR2 (Bildung von zwei Risikogruppen) [s1 entspricht Σ_1]

MAR 3

Hier ist es schwer, vernünftige Aussagen über die Fälle ohne Imputation zu machen, da durch die Funktion `cforest` leider sowohl in den Lern-Datensätzen ohne Imputation als auch im kombinierten Fall ohne Imputation jeweils 94 Beobachtungen fehlen. Dies bedeutet, dass lediglich sechs Beobachtungen vorhanden sind und dies ist zu wenig, um aussagekräftig zu sein.

Die Konstellationen mit Imputation sind bis auf die Lern-Datensätze deutlich schlechter als der reduzierte Fall. Sie haben ihr Minimum etwa auf der Höhe des Medians vom reduzierten Fall ohne `NA`. Die Lern-Datensätze sind in der Verteilung dem Goldstandard ähnlich, abgesehen vom Median. Dieser liegt ca. 0.252 höher. Der Test-Datensatz ohne Imputation liegt auch in etwa auf derselben Höhe, mit einem

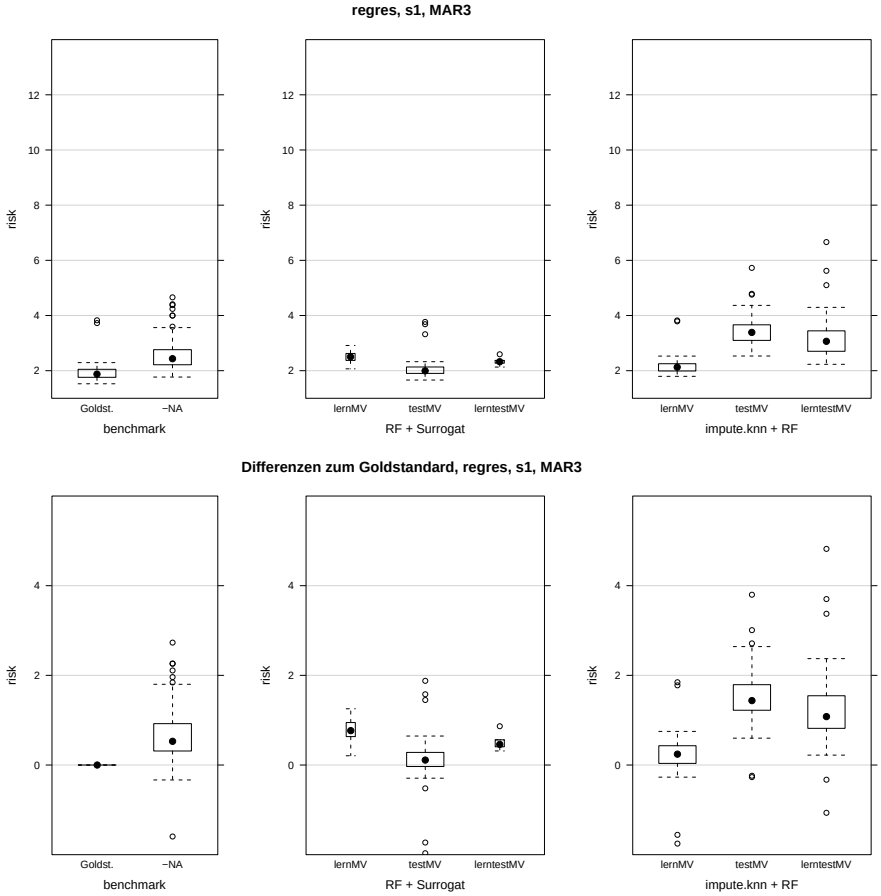

Abbildung 5.18.: Verteilung der mittleren quadratischen Abweichung sowie der Differenzen zum Goldstandard bei gleichmäßig hohen Korrelationen und fehlenden Werten nach MAR3 (rechtsseitige Trunkierung) [s1 entspricht Σ_1]

Median ca. 0.118 über dem Goldstandard. Dennoch resultiert für alle drei Fälle ein signifikanter Unterschied zum Goldstandard, aber auch zum reduzierten Fall „`-NA`“. Für den allein aussagekräftigen Mittelwertsvergleich zwischen dem Fehler bei Imputation und dem Fehler bei Surrogat-Variablen bei den Test-Datensätzen ergibt sich eine signifikante Abweichung.

Die beiden letztgenannten Situationen sind in den Differenzen zum Goldstandard an der Null gelegen mit einem Median von 0.110 bzw. 0.243. Die schlechten Fälle mit Imputation, also der Test-Datensatz und die kombinierte Zusammenstellung, liegen im Median dreifach bzw. doppelt so hoch wie der Fall „`-NA`“. Sie beinhalten die Null nicht einmal mit ihren Zäunen.

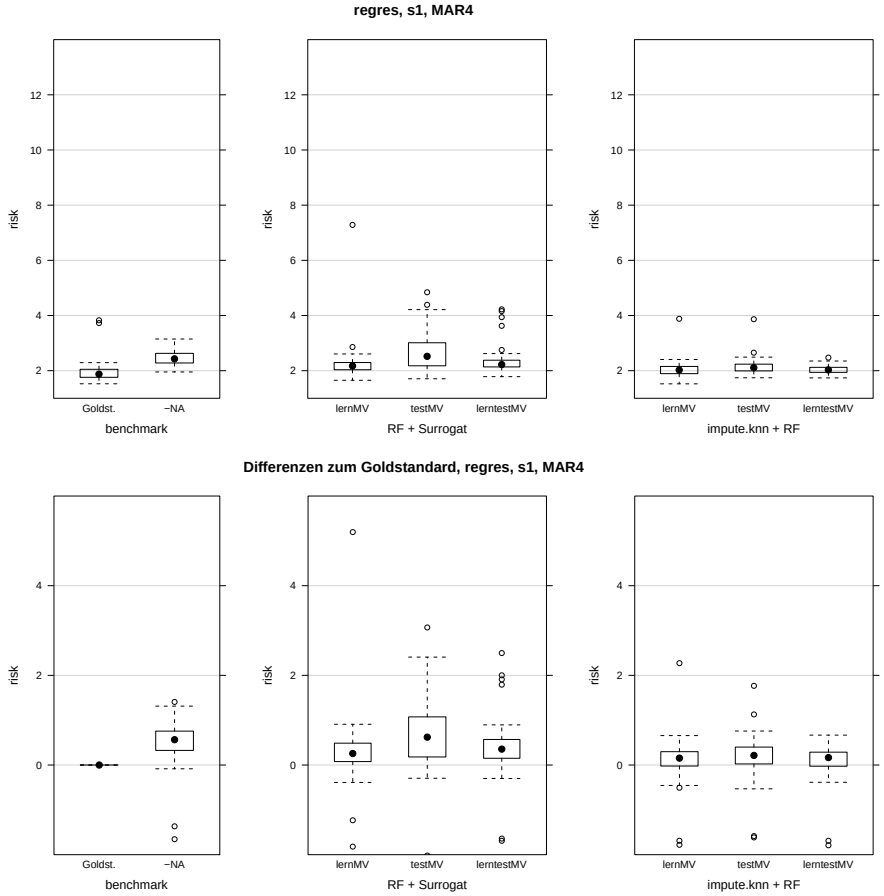

Abbildung 5.19.: Verteilung der mittleren quadratischen Abweichung sowie der Differenzen zum Goldstandard bei gleichmäßig hohen Korrelationen und fehlenden Werten nach MAR4 (symmetrische Trunkierung) [s1 entspricht Σ_1]

MAR 4

Die symmetrische Trunkierung ist wie von der Klassifikation gewohnt auch bei der Regression besser geeignet für `cforest` als die rechtsseitige Trunkierung. Die Boxplots befinden sich relativ nahe am Goldstandard, die Mediane allesamt zwischen Goldstandard und reduziertem Fall ohne `NA`. Eine Ausnahme bildet der Test-Datensatz ohne Imputation, dieser überschreitet den reduzierten Fall im Median und ebenso im dritten Quartil. Die Interquartilsabstände sind relativ ähnlich, bis auf der des Test-Datensatzes ohne Imputation. Dieser ist etwa doppelt so groß wie die anderen. Sämtliche Mittelwertsvergleiche ergeben bei dieser Methode einen signifikanten Unterschied.

Bei solch ähnlichen Ergebnissen ist es wichtig, sich die Differenzen zum Goldstandard anzuschauen. Doch auch hier zeigt sich ein relativ gutes Bild. Die Null wird

mindestens mit den Zäunen eingeschlossen. Abgesehen vom Test-Datensatz ohne Imputation liegen die Mediane maximal 0.355 von der Null entfernt. Der Median des Falls „-NA“ liegt bei 0.565, der des Test-Datensatzes bei 0.621 und damit annähernd am reduzierten Fall. Die Streuungen sind relativ gering und die dritten (oberen) Quartile liegen unterhalb des reduzierten Falls.

MCAR

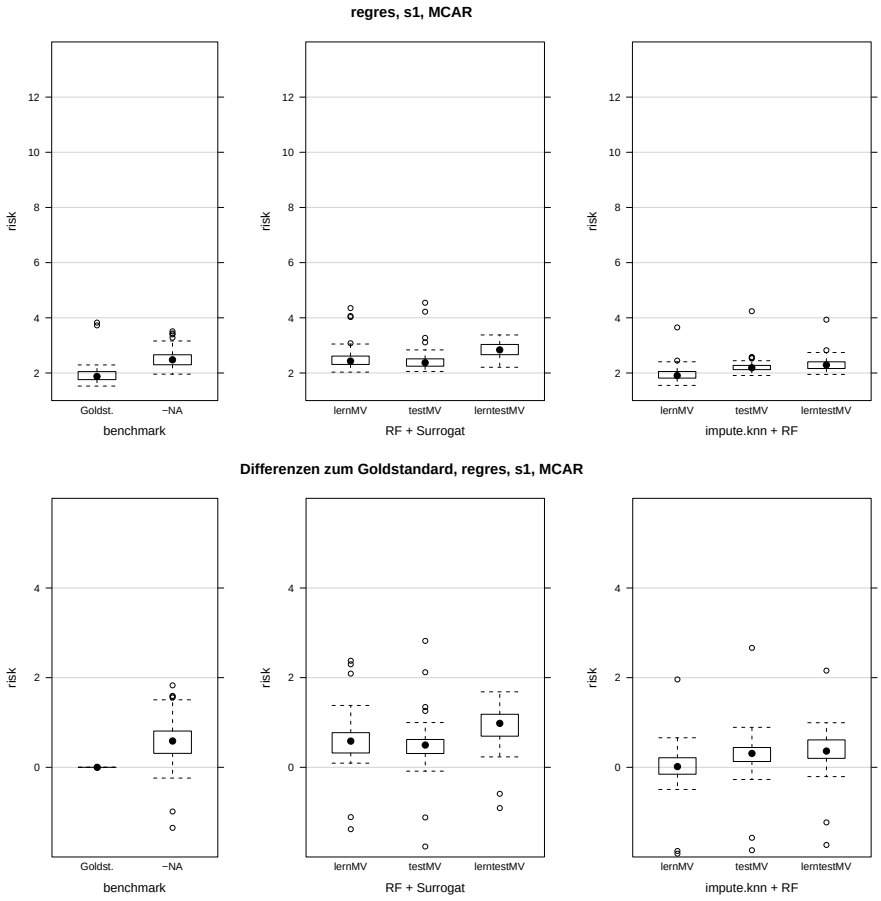

Abbildung 5.20.: Verteilung der mittleren quadratischen Abweichung sowie der Differenzen zum Goldstandard bei gleichmäßig hohen Korrelationen und fehlenden Werten nach MCAR (vollständig zufällig) [s1 entspricht Σ_1]

Bei vollständig zufällig fehlenden Werten sind etwas deutlichere Unterschiede zum bemerken. Die Anordnungen mit Imputation weisen eine leichte Steigerung auf (Test-Datensatz und kombinierter Fall sind fast gleich). Ohne Imputation ist ebenfalls der kombinierte Fall der schlechteste. Dennoch unterscheiden sich die beiden kombinierten Fälle im Mittelwert signifikant vom reduzierten Fall. Bis auf den Test-

Datensatz ohne Imputation liegen alle Mediane zwar unterhalb des reduzierten Falls, aber näher daran als am Goldstandard. Für die Mittelwerte ergibt sich deswegen ein Unterschied zum Goldstandard. Nur bei den Lern-Datensätzen mit Imputation ist er nicht signifikant. Vom reduzierten Fall weichen alle Situationen ab, bis auf die Lern-Datensätze ohne Imputation und die Test-Datensätze ohne Imputation. Die Interquartilsabstände sind recht übereinstimmend. Der Vergleich zwischen Conditional Tree Forests mit Surrogat-Variablen und *Forests* mit vorheriger Imputation sind alle Mittelwerte signifikant voneinander verschieden.

Die Boxplots der Differenzen zum Goldstandard sind für Fälle ohne Imputation etwas weiter von der Null entfernt als für Fälle mit Imputation. Die Lern-Datensätze und der kombinierte Fall ohne Imputation schließen die Null nicht mit den Zäunen ein wie der Test-Datensatz ohne Imputation. Die Steigerung in den ursprünglichen Risiko-Daten mit Imputation lässt sich ebenso in den Differenzen erkennen, allerdings ist der Median des Test-Datensatzes und des kombinierten Falls fast gleich (Unterschied: 0.053). Sehr schön: Bei den Lern-Datensätzen mit Imputation ist der Median fast Null (0.016).

5.2.2. Blockweise hohe Korrelationen

Wie in den Simulationen zur Klassifikation ist eine allgemeine Verschlechterung auf Grund des Korrelationsdesigns zu verzeichnen.

MAR 1

Die Ergebnisse dieser Simulation liegen zwischen Goldstandard und reduziertem Fall. In beiden Imputationssituationen ist der kombinierte Fall der schlechteste. Die Boxplots weisen (ohne Ausreißer betrachtet) eine geringe Streuung auf. Der Test-Datensatz mit fehlenden Werten hat neben den Lern-Datensätzen mit Imputation ein Ergebnis, das dem des Goldstandards am nächsten ist. Die Lern-Datensätze fallen dabei wiederholt mit guten Ergebnissen auf. Sämtliche Mittelwertsvergleiche sind signifikant.

Auch die Differenzen zum Goldstandard steigen nicht über den reduzierten Fall. Die Null wird nur vom kombinierten Fall ohne Imputation nicht mehr von den Zäunen eingeschlossen. Dessen Median liegt bei 0.893, der Median der anderen Situationen bei maximal der Hälfte des Falls „`-NA`“ (Median: 1.21, die anderen sind ≤ 0.614). Die Streuungen sind ziemlich einheitlich.

MAR 2

Auch diese Simulation verläuft im Hinblick auf die Lage der Risiko-Verteilungen zufrieden stellend. Trotz der nur blockweisen Korrelationen liegen die Mediane und

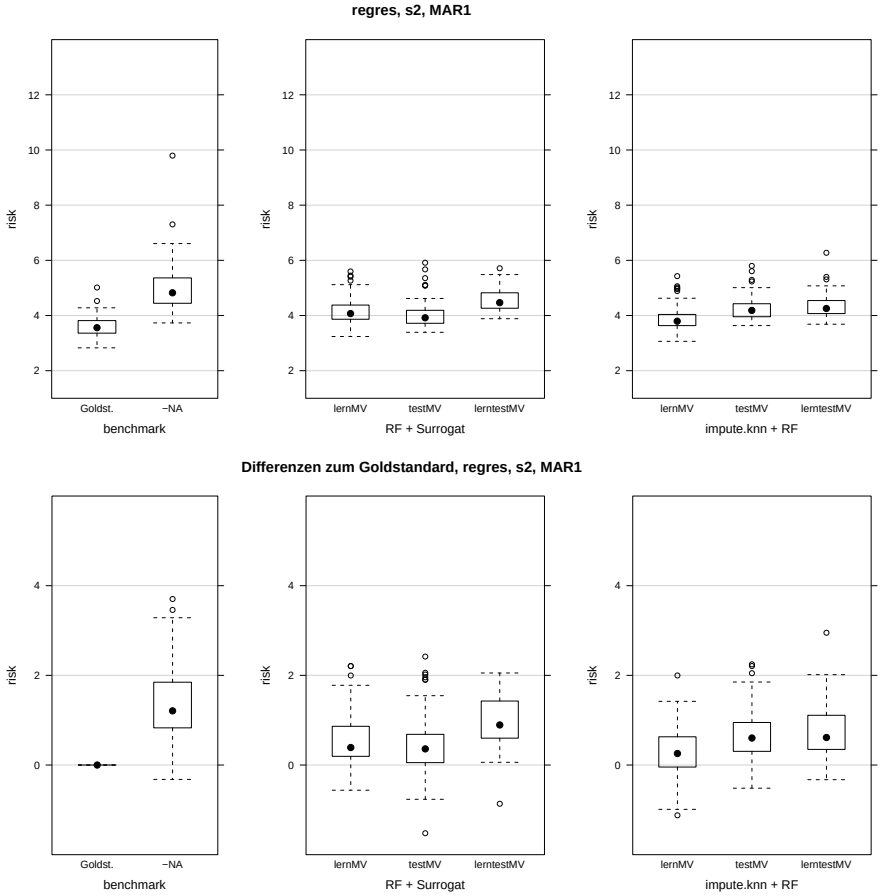

Abbildung 5.21.: Verteilung der mittleren quadratischen Abweichung sowie der Differenzen zum Goldstandard bei blockweise hohen Korrelationen und fehlenden Werten nach MAR1 (Bildung von Rängen) [s2 entspricht Σ_2]

die Boxen in der Mitte von Goldstandard und reduziertem Fall. Dadurch lassen sich kaum Schwankungen feststellen. Dennoch sind alle Mittelwertsvergleiche signifikant. Der Fall mit fehlenden Werten in den Lern- und Test-Datensätzen ohne Imputation ist leicht nach oben verschoben.

Sechs der *Forests* konnten bei den Lern-Datensätzen mit fehlenden Werten (ohne Imputation) nicht berechnet werden.

Eine ähnliche Situation wie in den ursprünglichen Daten spiegelt das Ergebnis der Differenzen zum Goldstandard wider. Der etwas schlechtere kombinierte Fall ohne Imputation schließt knapp die Null nicht mit den Zäunen ein. Dennoch liegt der Median ca. 0.451 und damit ein Drittel unter demjenigen des reduzierten Falls ohne `NA`. Die anderen Mediane sind maximal ca. 0.551 von der Null entfernt. Auch hier sind die Streuungen etwa gleich groß.

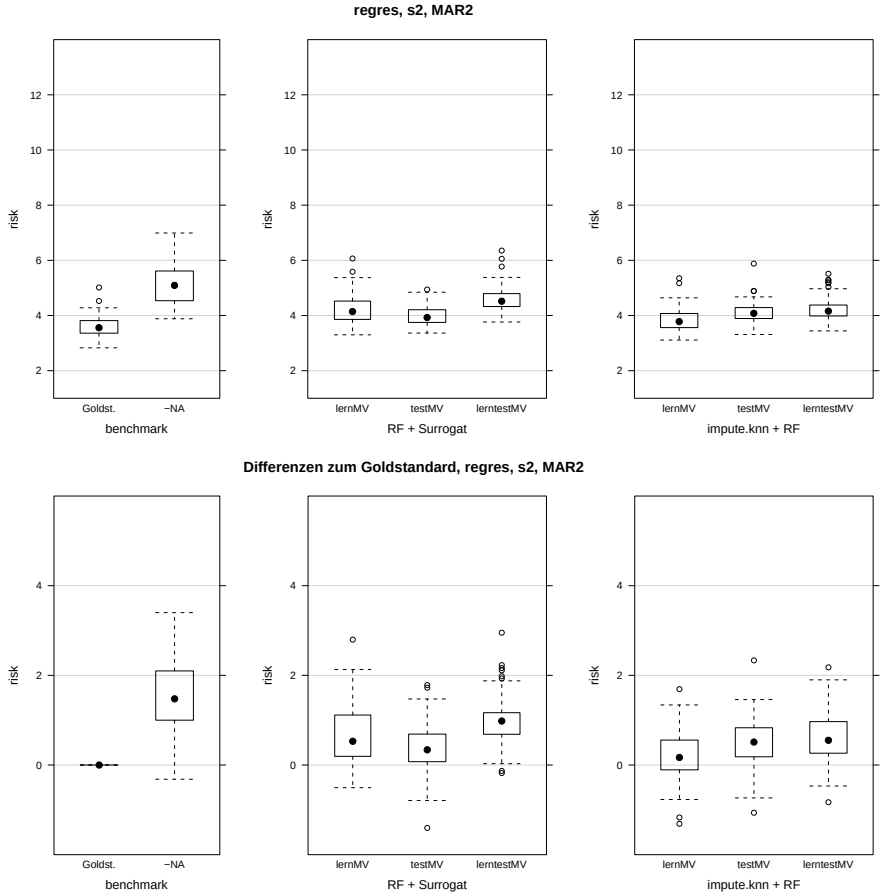

Abbildung 5.22.: Verteilung der mittleren quadratischen Abweichung sowie der Differenzen zum Goldstandard bei blockweise hohen Korrelationen und fehlenden Werten nach MAR2 (Bildung von zwei Risikogruppen) [s2 entspricht Σ_2]

MAR 3

Bei fehlenden Werten nach MAR3 – diese Methode hat schon mehrfach zu schlechten Ergebnissen der Conditional Tree Forests geführt – ist die Verteilung des MSE erstaunlich gut. Die Mediane liegen eher am Goldstandard als am reduzierten Fall und die Boxen reichen nicht bis zur unteren Grenze der „`-NA`“-Box hinauf. Der Test-Datensatz mit fehlenden Werten und mit Imputation verlagert sich bei dieser Simulation ebenfalls nach oben, dennoch nicht so hoch, wie man es aus bisherigen Ergebnissen erwarten würde. Alle sechs Fälle unterscheiden sich im Mittelwert signifikant vom Fall „`-NA`“, aber auch vom Goldstandard. Der Mittelwertsvergleich zwischen den Methoden mit Imputation und denen mit Surrogat-Variablen ist nur für die Test-Datensätze signifikant.

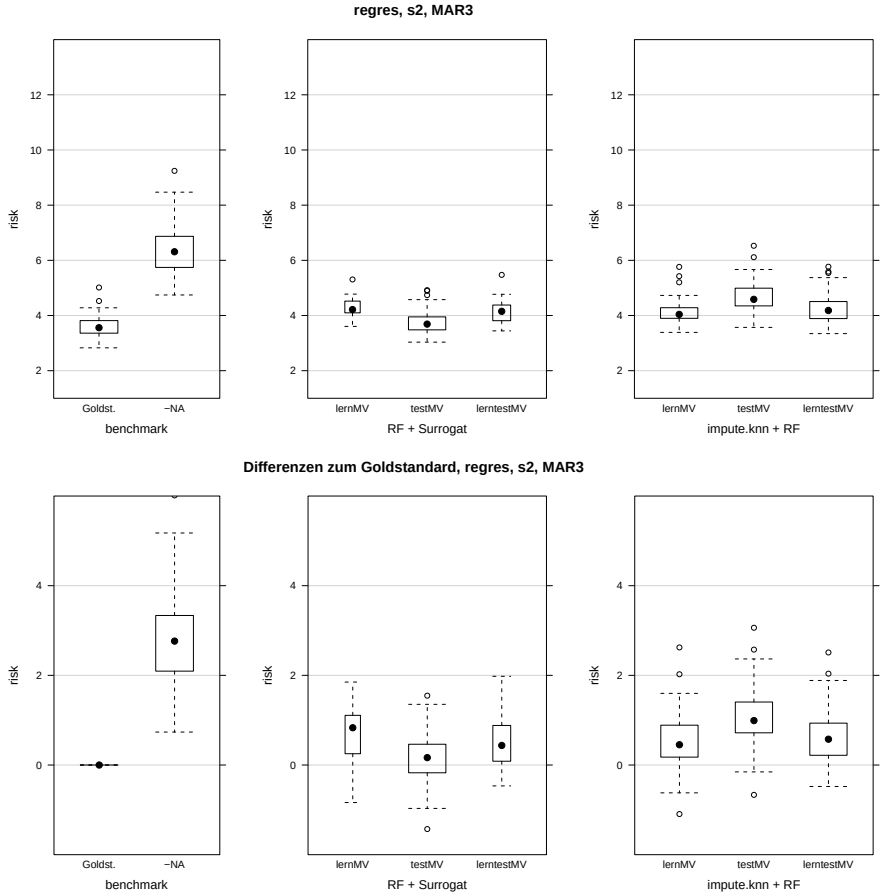

Abbildung 5.23.: Verteilung der mittleren quadratischen Abweichung sowie der Differenzen zum Goldstandard bei blockweise hohen Korrelationen und fehlenden Werten nach MAR3 (rechtsseitige Trunkierung) [s2 entspricht Σ_2]

Die Fälle ohne Imputation lassen sich kaum vergleichen, da in den Lern-Datensätzen 83 Beobachtungen fehlen und im kombinierten Fall 77 *Forests* mit `NA` kodiert werden mussten.

In den Differenzen zum Goldstandard wird sogar die Null-Linie von den Zäunen des Test-Datensatzes mit erfolgter Imputation eingeschlossen. Überhaupt ergibt sich hier eine relativ geringe Abweichung der Mediane von der Null (maximal 0.991, d. h. ca. 1/3 des reduzierten Falls). Wiederholt fällt die in etwa gleiche Stärke der Streuung auf.

MAR 4

Trotz des blockweisen Designs der Korrelationen ergeben sich auch in diesem Szenario Fehler-Verteilungen unterhalb des reduzierten Falls ohne `NA`. Mit MAR4 können

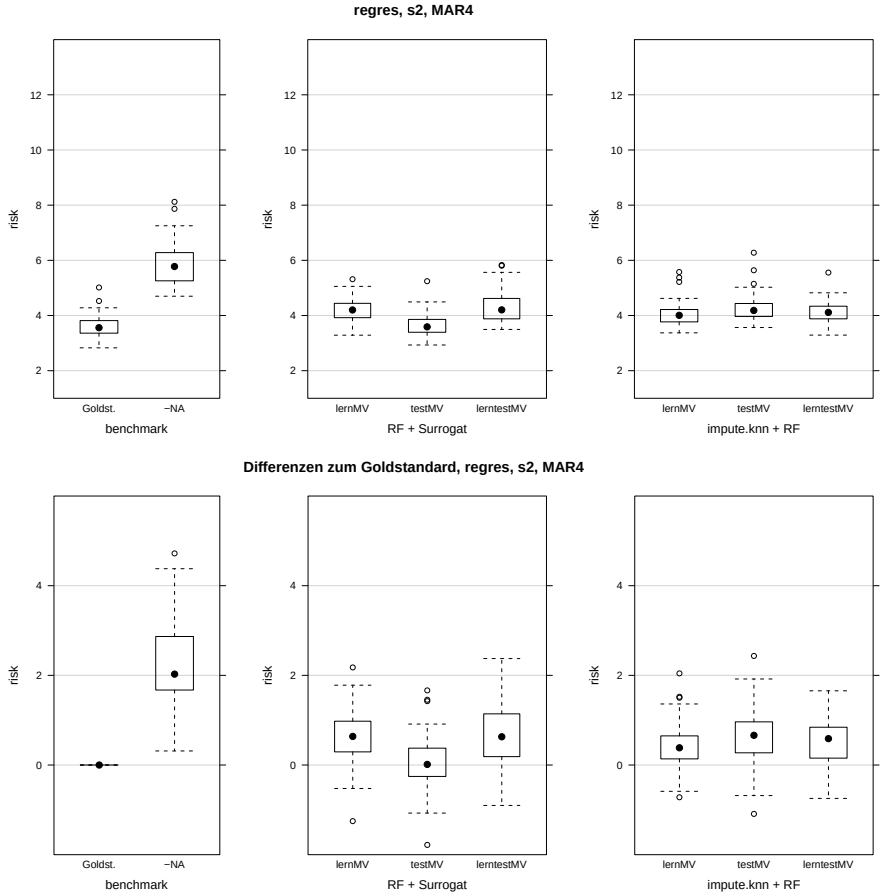

Abbildung 5.24.: Verteilung der mittleren quadratischen Abweichung sowie der Differenzen zum Goldstandard bei blockweise hohen Korrelationen und fehlenden Werten nach MAR4 (symmetrische Trunkierung) [s2 entspricht Σ_2]

bei fehlenden Werten im Test-Datensatz ohne Imputation sogar bessere Ergebnisse erzielt werden als bei fehlenden Werten in den Lern-Datensätzen mit Imputation: Der Median liegt näher am Goldstandard (0.030 darüber). Der Unterschied im Mittelwert ist ebenfalls nicht signifikant. Alle anderen Mittelwertsvergleiche sind signifikant, auch diejenigen zum Vergleich der Ergebnisse mit und ohne Imputation.
Die Fälle mit Imputation sind sich sehr ähnlich, die Mediane unterscheiden sich nur um maximal 0.176 (ohne Imputation: maximal 0.615). Das sehr gute Ergebnis des Test-Datensatzes ohne Imputation gibt auch der kombinierte Fall wider: Er ist nur minimal schlechter als wenn nur in den Lern-Datensätzen Werte fehlen würden.

Die Mediane der Differenzen zum Goldstandard betragen ein gutes Viertel des reduzierten Falls. Obwohl die Mediane also ein gutes Ergebnis erwarten lassen, wird die Null nur beim Test-Datensatz ohne Imputation von der Box eingeschlossen. In den anderen Fällen liegen die Fehlerverteilungen höher, so dass die Null-Linie nur

noch innerhalb der Zäune liegt. Außerdem lässt sich festhalten, dass die Differenzen dieser Simulation nicht so gleichmäßig gestreut sind wie in den vorangegangenen Methoden für fehlende Werte.

MCAR

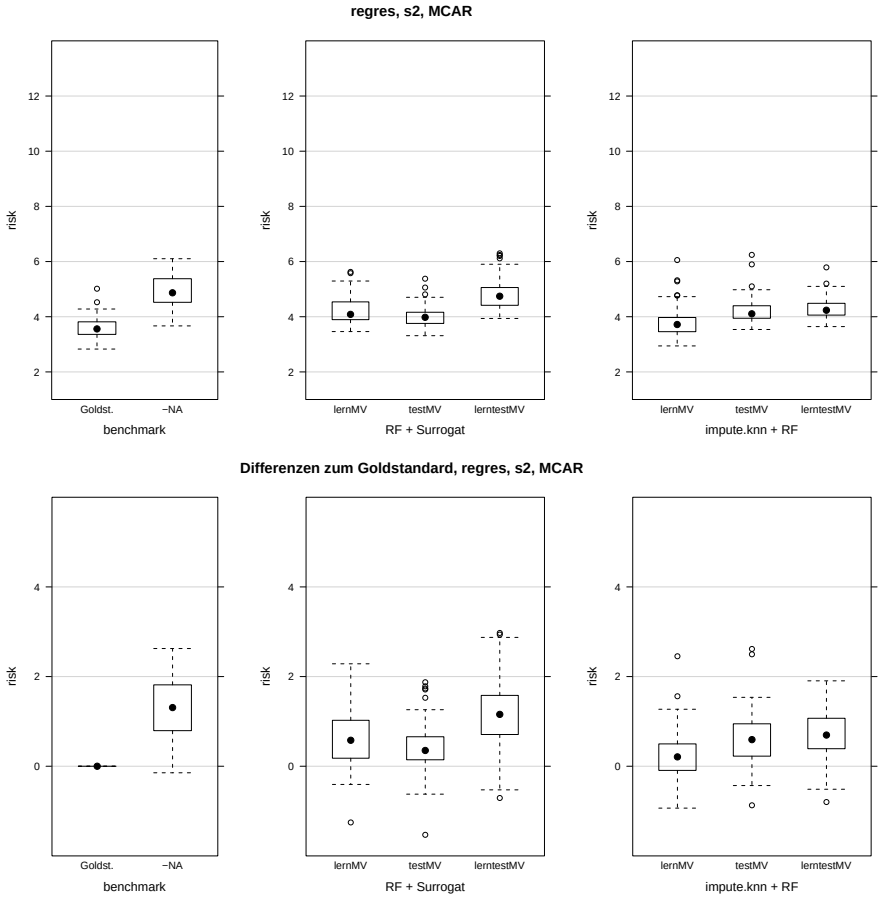

Abbildung 5.25.: Verteilung der mittleren quadratischen Abweichung sowie der Differenzen zum Goldstandard bei blockweise hohen Korrelationen und fehlenden Werten nach MCAR (vollständig zufällig) [s2 entspricht Σ_2]

Werden einige Werte vollständig zufällig eliminiert, so erhält man für blockweise hohe Korrelationen im Vergleich mit den Bezugswerten Goldstandard und reduzierter Fall akzeptable Ergebnisse. Erstaunlich dabei ist, dass der Test-Datensatz ohne Imputation bei gleichmäßig hohen Korrelationen im Gegensatz zu hier *oberhalb* des Falls „`-NA`“ liegt. Im vorliegenden Fall entspricht die kombinierte Zusammenstellung ohne Imputation etwa dem Fall „`-NA`“, ist aber im Median leicht besser. Der Mittelwertsvergleich mit dem reduzierten Fall ist nur bei diesem Fall nicht signifikant.

Die Lern-Datensätze mit Imputation liegen dem Goldstandard am nächsten (ca. 0.1 nach oben verschoben), eine signifikante Abweichung ist trotzdem vorhanden. Allgemein befinden sich allerdings die unteren Quartile über dem Goldstandard-Median. Vergleicht man die Mittelwerte der Methode „Surrogat-Variablen“ mit der Methode der Imputation, so ist in allen Fällen ein signifikanter Unterschied festzustellen.

Betrachtet man die Differenzen zum Goldstandard, ergibt sich auch hier beim kombinierten Fall ohne Imputation eine Ähnlichkeit zum reduzierten Fall. Nur die Lern-Datensätze mit fehlenden Werten und mit Imputation schließen die Null mit der Box ein. Erfreulicherweise beinhalten somit alle Anordnungen mindestens mit den Zäunen die Null-Linie. Die zu vergleichenden Situationen sind außerdem näher an der Null als der reduzierte Fall ohne `NA`. Wie meistens sind auch hier die Fälle mit Imputation etwas besser als diejenige ohne.

5.2.3. Gleichmäßige, niedrige Korrelationen

Ein weiterer genereller Sprung der Risiken einer Fehleinschätzung nach oben wird auch hier beobachtet. Diese Simulationen sind weiter gestreut und höher angesiedelt als die vorhergehenden. Vermutlich hängt es auch hier mit dem Design der Korrelationsmatrix zusammen.

MAR 1

Auf den ersten Blick fällt auf, dass die Risiken in ihren Beträgen steigen. Die Lern-Datensätze mit fehlenden Daten sind die niedrigsten Boxplots und die kombinierten Fälle ergeben die größten Fehler. Die Mediane springen um jeweils ca. 1 nach oben. Dabei sind die Lern-Datensätze ohne Imputation bereits auf Höhe des reduzierten Falls (ca. 0.08 darunter, kein signifikanter Unterschied), lediglich die Lern-Datensätze mit Imputation sind besser (ca. 1.16 darunter, signifikante Abweichung im Mittelwert von Goldstandard und „`-NA`“). Dieser Fall ist also der einzige, der sich zwischen Goldstandard und dem reduzierten Fall befindet. Es spricht nicht unbedingt dafür, dass `cforest` mit dieser Situation gut umgehen kann.
Der Unterschied zwischen Imputation und Verwendung von Surrogat-Variablen ist bei den Lern-Datensätzen und beim kombinierten Fall im Mittelwert signifikant.

Die Differenzen zum Goldstandard liegen weit im Positiven. Zum Teil liegt sogar das Minimum über der Null. In den Differenzen liefern die Lern-Datensätze ohne Imputation, welche im Median der ursprünglichen Daten noch unterhalb des Falls „`-NA`“ liegen, etwa die gleichen Fehlerbeträge wie dieser (im Median ca. 0.06 darüber). Der Median des in den ursprünglichen Daten besten Falls (Lern-Datensätze mit Imputation) liegt bei ca. einem Fünftel des höchsten Medians (kombinierter Fall ohne Imputation). Dies verdeutlicht die starke Variabilität des Ergebnisses.

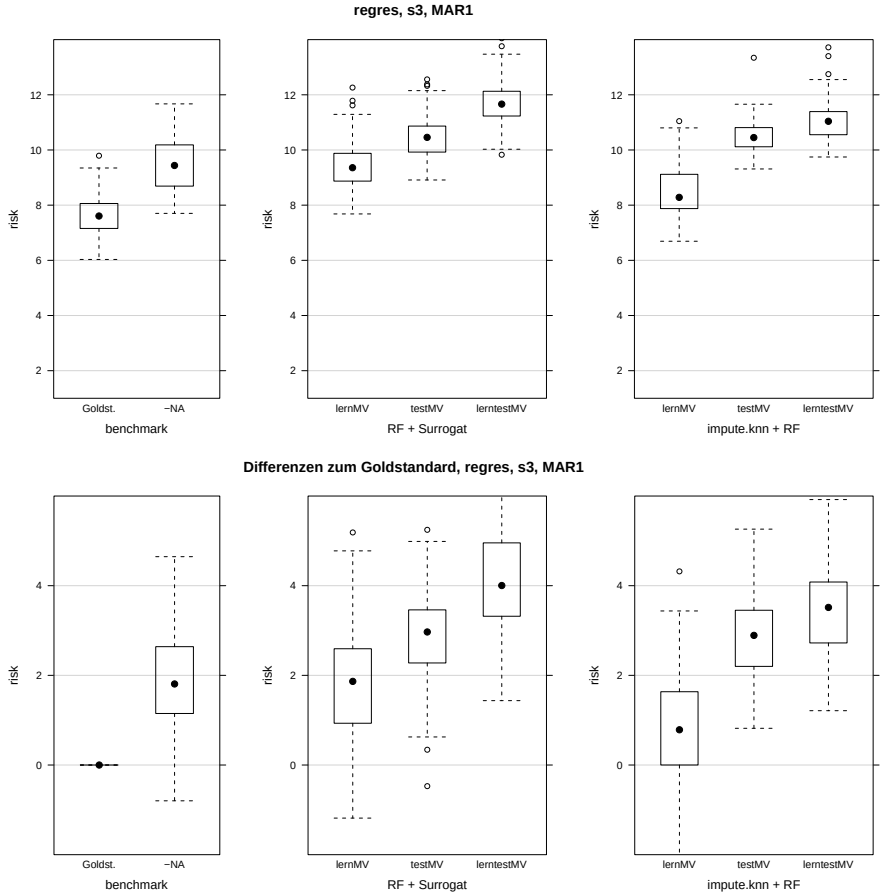

Abbildung 5.26.: Verteilung der mittleren quadratischen Abweichung sowie der Differenzen zum Goldstandard bei gleichmäßig niedrigen Korrelationen und fehlenden Werten nach MAR1 (Bildung von Rängen) [s3 entspricht Σ_3]

MAR 2

Bei fehlenden Werten, die durch die Methode MAR2 eingestreut werden, ergibt sich kein wesentlicher Unterschied zu fehlenden Werten nach MAR1. Der reduzierte Fall ohne `NA` liegt etwas höher, was das Gesamtergebnis allerdings nicht verbessert. Nach wie vor liegen vier von sechs Situationen im Median darüber, eine nur knapp darunter. Die Lern-Datensätze mit Imputation liegen ebenso wie bei MAR1 näher am Goldstandard als am reduzierten Fall. Dennoch weichen sie im Mittelwert wie alle anderen Situationen auch sowohl vom Goldstandard als auch vom reduzierten Fall „`-NA`“ signifikant ab. Es ergibt sich kein deutlicher Unterschied im Mittelwert zwischen dem Fall mit und dem Fall ohne Imputation bei den Test-Datensätzen.

Auch die Differenzen zum Goldstandard sind außer bei den Lern-Datensätzen mit fehlenden Werten nicht besser als die des reduzierten Falls. Bei den beiden Test-

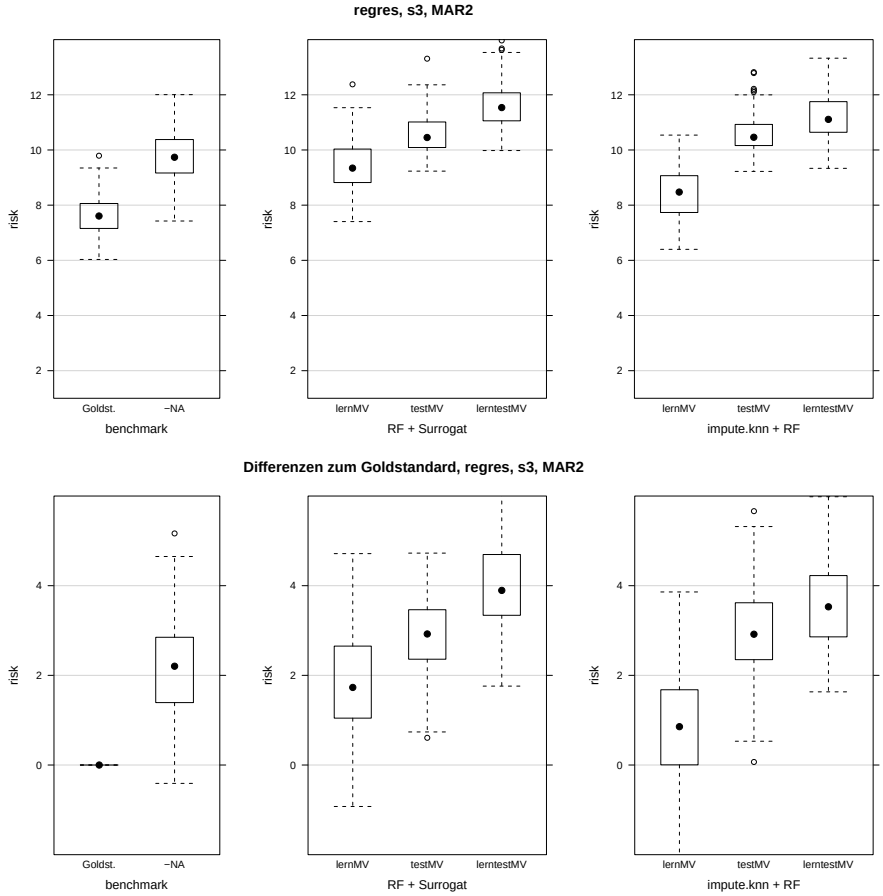

Abbildung 5.27.: Verteilung der mittleren quadratischen Abweichung sowie der Differenzen zum Goldstandard bei gleichmäßig niedrigen Korrelationen und fehlenden Werten nach MAR2 (Bildung von zwei Risikogruppen) [s3 entspricht Σ_3]

Datensätzen mit und ohne Imputation unterscheiden sich die Boxplots kaum bis auf die Tatsache, dass derjenige mit Imputation noch weiter hinauf reicht. Bei den kombinierten Fällen zeigt sich eine kleine Besserung für den Fall mit Imputation im Vergleich zum Fall ohne Imputation. Der Median der Lern-Datensätze ohne Imputation verringert sich ungefähr um die Hälfte, falls Imputation erfolgt.

MAR 3

Wie schon mehrmals bei MAR3-Anwendung, fehlen auch hier wieder in den Lern-Datensätzen und im kombinierten Fall – jeweils ohne Imputation – Beobachtungen bzw. konnten die Conditional Tree Forests nicht berechnet werden. In den Lern-Datensätzen trifft dies 76-mal, im kombinierten Fall 84-mal zu. Somit stehen nur

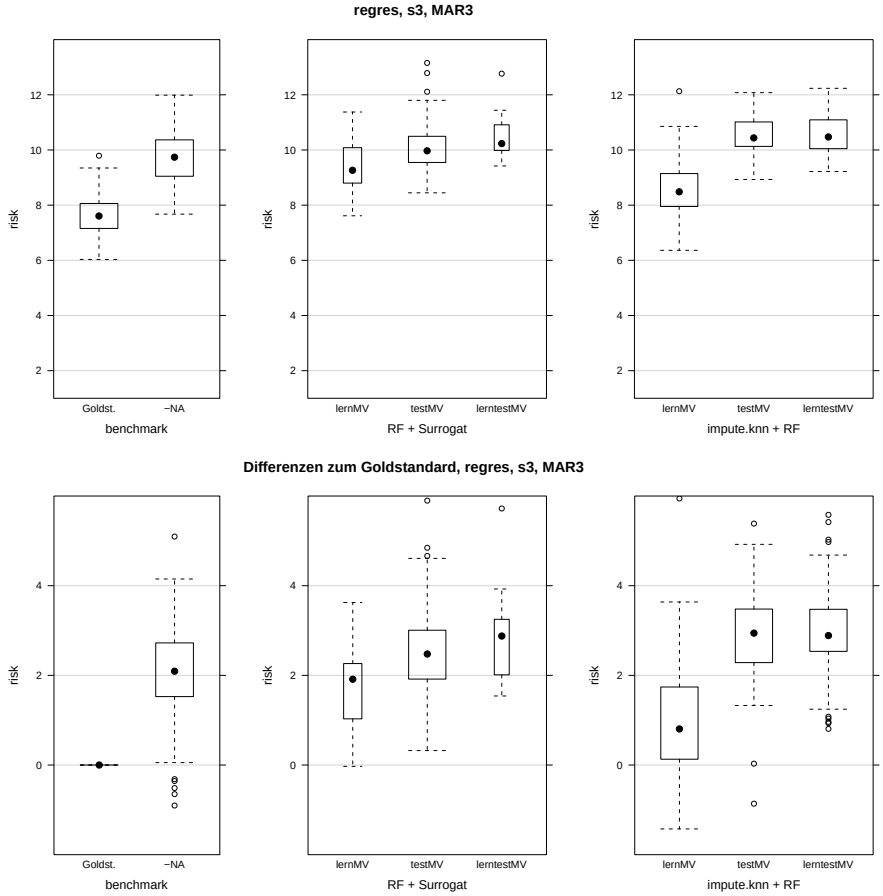

Abbildung 5.28.: Verteilung der mittleren quadratischen Abweichung sowie der Differenzen zum Goldstandard bei gleichmäßig niedrigen Korrelationen und fehlenden Werten nach MAR3 (rechtsseitige Trunkierung) [s3 entspricht Σ_3]

24 bzw. 16 Risiko-Werte zur Verfügung und damit ist eine aussagekräftige Analyse nicht gegeben.

Trotz der bis jetzt als problematisch erschienenen Methode MAR3 sind in dieser Simulation die Unterschiede in den Boxplots nicht so stark. Der nicht aussagekräftige kombinierte Fall ohne Imputation, der bis jetzt bei $\mathbf{\Sigma}_3$ und Regression am schlechtesten war, liegt im Median nur ca. 0.49 über dem Median des Falles „`-NA`“. Der Median des kombinierten Falls mit Imputation liegt ca. 0.73 darüber (bei MAR1 und MAR2: mehr als 1 darüber), der Mittelwert weicht signifikant ab. Der kombinierte Fall und der Test-Datensatz, jeweils mit Imputation, weichen kaum voneinander ab. Lediglich die Lern-Datensätze mit Imputation sind einigermaßen in der Nähe des Goldstandards, der Mittelwert ist allerdings signifikant verschieden. Bei den aussagekräftigen Test-Datensätzen ist der Unterschied im Mittelwert zwischen Surrogat-Variablen

und Imputation signifikant.

In den Differenzen zum Goldstandard zeigt sich allerdings, dass auch vom Boxplot dieses Falls die Null nicht von der Box eingeschlossen wird: Das untere Quartil liegt bereits im Positiven. Die Gleichheit des Test-Datensatzes und des kombinierten Falls (mit Imputation) ist auch bei den Differenzen zu beobachten. Die Mediane liegen dabei etwa anderthalb mal so hoch wie der des reduzierten Falls.

MAR 4

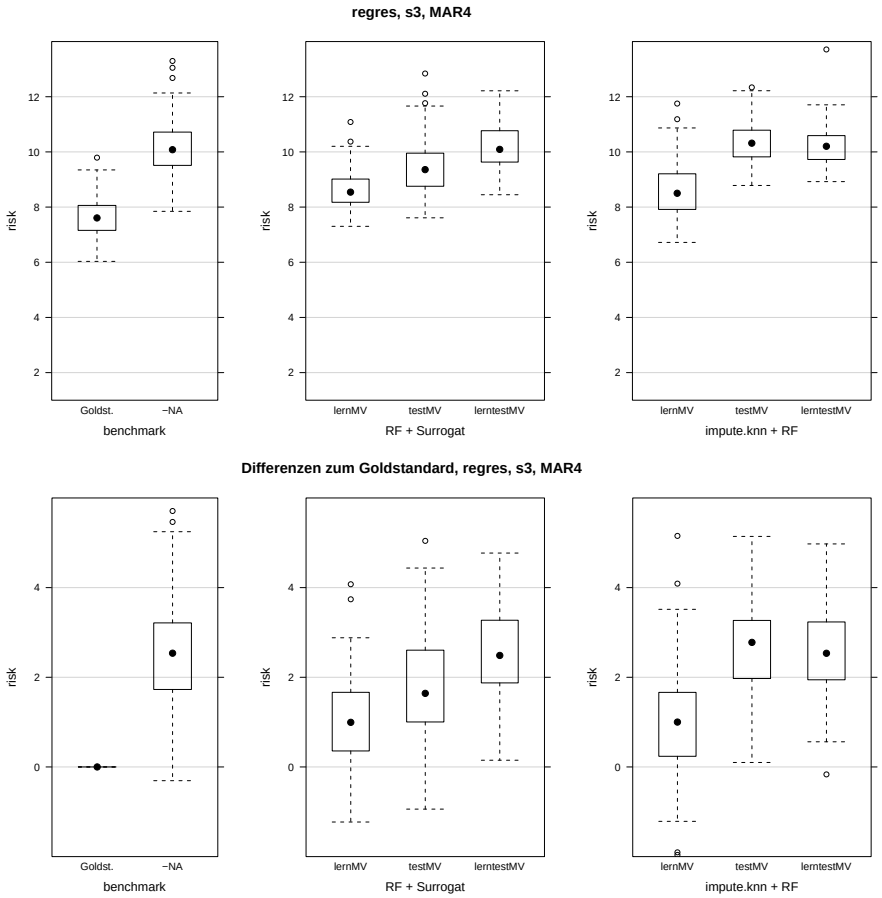

Abbildung 5.29.: Verteilung der mittleren quadratischen Abweichung sowie der Differenzen zum Goldstandard bei gleichmäßig niedrigen Korrelationen und fehlenden Werten nach MAR4 (symmetrische Trunkierung) [s3 entspricht Σ_3]

Es ist eine leichte Verbesserung festzustellen. In dieser Simulation sind wieder einige Mediane zwischen dem Fall „`-NA`“ und dem Goldstandard. Derjenige vom kombinierten Fall ohne Imputation ist 0.01 schlechter als der reduzierte Fall ohne `NA` (im

Mittelwert nicht signifikant verschieden). Die Mediane des Test-Datensatzes und des kombinierten Falls mit Imputation liegen 0.24 bzw. 0.12 über demjenigen des reduzierten Falls (kein signifikanter Unterschied im Mittelwert). Auf die Höhe des Betrags (10.08) betrachtet sind dies sehr geringe Abweichungen. Die Lern-Datensätze mit fehlenden Werten und der kombinierte Fall unterscheiden sich kaum. Speziell die Lern-Datensätze liegen beide näher am Goldstandard als am Fall „`-NA`“. Die Abweichungen vom Goldstandard sind trotzdem alle signifikant.

Die Zäune der Differenzen reichen zwar weiter nach unten als z. B. bei MAR3, dennoch fassen immer noch nicht alle die Null ein. Außer bei den beiden Anordnungen mit Lern-Datensätzen ist dies nun bei fehlenden Werten im Test-Datensatz ohne Imputation der Fall. Der Test-Datensatz mit Imputation überschreitet dagegen als einziger im Median den reduzierten Fall; die kombinierten Fälle liegen etwa gleich auf mit diesem. Die beiden Boxplots für die Lern-Datensätze unterscheiden sich nicht sehr.

MCAR

Bei vollständig zufällig fehlenden Werten ist das Risiko einer Fehleinschätzung wieder etwas erhöht. Außerdem ist hier auch wieder die Steigerung zu beobachten. Zwischen Goldstandard und „`-NA`“ liegen die Lern-Datensätze. Die anderen Mediane liegen ca. 0.46 und mehr über dem des reduzierten Falls ohne `NA`. Die Fälle mit Imputation liegen tiefer als die zugehörigen Fälle ohne Imputation und `cforest` erzielt damit bessere Ergebnisse. Dies bestätigen die t-Tests, die nur für die Test-Datensätze keinen signifikante Unterschied ergeben. Alle anderen Mittelwertsvergleiche, auch zum Goldstandard und zum reduzierten Fall, sind signifikant.

Fehlende Werte in den Lern-Datensätzen sind bei den Differenzen zum Goldstandard ebenfalls die besten Situationen. Jedoch wird dort die Null wiederholt nur von den Zäunen eingeschlossen. Die Fälle mit Imputation ergeben auch in den Differenzen kleinere Fehler als ihre Gegenstücke ohne Imputation. Der Median des kombinierten Falles ohne Imputation (schlechtester Median) ist knapp doppelt so hoch wie derjenige des reduzierten Falls. Der beste Median (fehlende Werte in den Lern-Datensätzen, mit Imputation) beträgt ungefähr die Hälfte des reduzierten Falls.
Die Interquartilsabstände sind in dieser Simulation ziemlich gleich. Der kombinierte Fall mit Imputation hat eine etwas kleineren Boxlänge.

5.3. Fazit

Die beiden verschiedenen Arten von Datensätzen (aus `dgp1` und `dgp2`) lassen sich auf Grund der verschiedenen Risiko-Berechnungen nicht miteinander vergleichen.

Bei den verschiedenen Korrelationen hat sich die Vermutung bestätigt, dass Daten

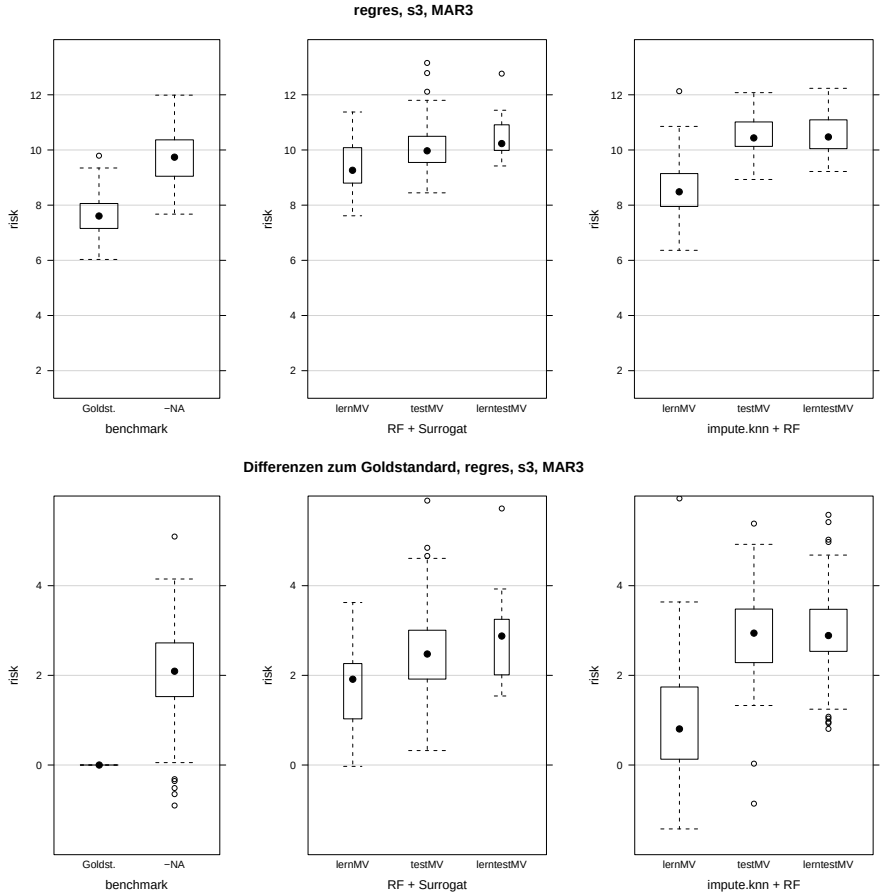

Abbildung 5.30.: Verteilung der mittleren quadratischen Abweichung sowie der Differenzen zum Goldstandard bei gleichmäßig niedrigen Korrelationen und fehlenden Werten nach MCAR (vollständig zufällig) [s3 entspricht Σ_3]

mit hohen Korrelationen besser geeignet sind. Sie sind für die sinnvolle Berechnung von Surrogat-Variablen und damit für gute Ergebnisse des Conditional Tree Forests unerlässlich. Außerdem werden sie zur Berechnung der k nächsten Nachbarn benötigt. Bei der Regression hat sich allerdings ergeben, dass für die blockweise hohen Korrelationen keine wesentlich schlechteren Differenzen zum Goldstandard resultierten als für die gleichmäßig hohen Korrelationen. Zudem fiel auf, dass bei gleichmäßig niedrigen Korrelationen oftmals eine Steigerung zu verzeichnen war, sodass die Fälle mit fehlenden Werten in Lern- und Test-Datensätzen schlechtere Ergebnisse brachten als die Lern-Datensätze mit fehlenden Werten.

Häufig ergab der Fall „`-NA`“ aus verschiedenen Gründen nicht die größten Fehlerbeträge. Bessere Situationen als der Goldstandard kamen so gut wie nicht vor. Wenn der Median tatsächlich näher an der Null lag, dann war sein Betrag nicht sehr viel geringer als der des Goldstandard-Medians.

Bezüglich der verschiedenen Mechanismen für fehlende Werte konnte festgestellt werden, dass die rechtsseitige Trunkierung (MAR3) der Funktion `cforest` Probleme bereitet, da die Werte nur einseitig gestrichen werden und somit die Berechnung von Surrogat-Variablen erschwert wird. Weniger schwierig war dies bei symmetrischer Trunkierung (MAR4) und den zufälligen Methoden MAR1 und MAR2 sowie MCAR. Für MCAR wurden allgemein innerhalb der Simulationen recht gleichmäßige Verteilungen der einzelnen Risiken beobachtet.

Weiterhin fiel auf, dass die Lern-Datensätze mit fehlenden Werten meist gute Ergebnisse erzielen konnten. Vor allem mit erfolgter Imputation reichten ihre Gütemaße häufig an den Goldstandard heran. Die Test-Datensätze mit `NA`s waren ohne Imputation meist besser als die Lern-Datensätze ohne Imputation – zumindest bei hohen Korrelationen ($\boldsymbol{\Sigma}_2$ und $\boldsymbol{\Sigma}_3$).

Die Fälle mit Imputation waren größtenteils besser als die jeweiligen Fälle ohne Imputation. Dies bestätigt das gute Ergebnis, das die Methode *knn*-Impute bei Troyanskaya u. a. [2001] erzielt hat.

Generell war festzustellen, dass die Risiken einer Fehleinschätzung ziemlich symmetrisch verteilt waren. Das arithmetische Mittel wich selten stark vom Median ab (vgl. Tabellen im Anhang F). Aus diesem Grunde war die Verwendung des Zwei-Stichproben-t-Tests nicht problematisch.

Als sehr gute Methoden wurden unter den Klassifikationen diejenige mit gleichmäßigen hohen Korrelationen und symmetrischer Trunkierung (5.1.1), mit blockweise hohen Korrelationen und Bildung von zwei Risikogruppen (5.1.3) und bei gleichmäßig niedrigen Korrelationen wieder diejenige mit symmetrischer Trunkierung ermittelt (5.1.3). Unter den Simulationen zur Regression wurden die beiden Methoden mit hohen Korrelationen und symmetrischer Trunkierung als gut beurteilt (5.2.1 und 5.2.2).

6. Zusammenfassung und Ausblick

In dieser Arbeit sollte das Verhalten der R-Funktion `cforest` bei verschiedenen Situationen mit fehlenden Werten untersucht werden.

Die Funktion `cforest` berechnet einen Conditional Tree Forest nach der Theorie von Hothorn u. a. [2006]. Dieses Verfahren bedingt dabei auf die Prädiktoren und schätzt die Verteilung des Response $\boldsymbol{y}$. Dadurch werden die Nachteile des ursprünglichen Random Forests von Breiman [2001] umgangen, wie z. B. Überanpassung der Daten und Selektions-Bias.

Um herauszufinden, ob `cforest` mit fehlenden Werten umgehen kann, wurden `NA`-behaftete Datensätze und imputierte Datensätze in die Funktion eingespeist und die Ergebnisse verglichen. Dabei wurden bezüglich der fehlenden Werte drei Situationen absimuliert: fehlende Werte in den 100 Lern-Datensätzen, fehlende Werte im Test-Datensatz und beides kombiniert – jeweils eben mit und ohne Imputation. Ohne vorheriger Imputation benutzt `cforest` Surrogat-Variablen, um die fehlenden Werte vorherzusagen.

Es wurden Simulationen zur Klassifikation und zur Regression erstellt, welche allerdings nicht direkt miteinander vergleichbar sind. Denn das Risiko einer Fehleinschätzung wurde auf Grund der vorliegenden Datenstruktur verschieden berechnet. Zusätzlich wurden Zwei-Stichproben-t-Tests durchgeführt, um die Unterschiede zwischen den einzelnen Methoden belegen zu können.

Aus den Ergebnissen ist ersichtlich, dass Daten mit hohen Korrelationen besser geeignet sind. Dies hilft der Funktion `cforest` bei der Berechnung von sinnvollen Surrogat-Variablen. Außerdem ist es dann für die Imputationsfunktion einfacher, geeignete k nächste Nachbarn zu finden.
Das Design der Korrelationen fällt anscheinend bei Regressionsbäumen weniger ins Gewicht als bei Klassifikationsbäumen.

Zur besseren Vergleichbarkeit der einzelnen Situationen innerhalb einer Simulation wurde ein Goldstandard, d. h. ein *Forest* mit einem Datensatz ohne fehlende Werte, und ein reduzierter Fall „`-NA`“ berechnet. Bei diesem Fall wurden in den Lern-Datensätzen Werte eliminiert und die Beobachtungen mit diesen fehlenden Werten wurden zur Berechnung des *Forests* einfach gestrichen. Jedoch ergaben sich oft Situationen, die aus verschiedenen Gründen schlechter waren als der reduzierte

Fall. Deutlich bessere Situationen als der Goldstandard wurden nicht beobachtet.

Allgemein scheinen einseitig fehlende Werte der Funktion `cforest` Probleme zu bereiten, wie bei den Simulationen mit fehlenden Werten nach MAR3 zu sehen war. Symmetrisch oder zufällig fehlende Werte erschweren die Berechnung von Surrogat-Variablen nicht so sehr und konnten dadurch zu besseren Ergebnissen führen.

Weiterhin fiel auf, dass fehlende Werte in den Lern-Datensätzen wenig problematisch sind. in diesen Situationen konnte `cforest` meist niedrige Risiken einer Fehleinschätzung erzielen. Vor allem mit vorheriger Imputation waren sie häufig dem Goldstandard ähnlich. Ohne Imputation waren die Test-Datensätze mit `NA`s meist besser als die Lern-Datensätze ohne Imputation – zumindest bei hohen Korrelationen ($\mathbf{\Sigma}_2$ und $\mathbf{\Sigma}_3$).

Das gute Ergebnis von Troyanskaya u. a. [2001] für die Imputationsmethode *knn*-Impute konnte bestätigt werden: Die Fälle mit Imputation waren meist besser als die jeweiligen Fälle ohne Imputation.

Es könnte interessant sein, ob dieses Ergebnis für weniger gute bzw. andere Imputationsmethoden zu halten ist. Ebenso wäre ein Überprüfen der Leistungen des Conditional Tree Forests mit realen Datensätzen möglich. Zudem wäre eventuell ein Vergleich mit den Regressionsmethoden des generalisierten linearen Modells und der logistischen Regression aufschlussreich.

Literaturverzeichnis

[Breiman 2001] Breiman, Leo: Random Forests. In: *Machine Learning* (2001), Nr. 45, S. 5–32

[Breiman u. a. 1984] Breiman, Leo ; Friedman, Jerome ; Stone, Charles J. ; Olshen, R.A.: *Classification And Regression Trees*. Chapman & Hall, 1984

[Fahrmeir u. a. 1996] Fahrmeir, Ludwig ; Hamerle, Alfred ; Tutz, Gerhard: *Multivariate statistische Verfahren*. 2., überarbeitete Auflage. de Gruyter, 1996

[Friedman 1991] Friedman, Jerome H.: Multivariate adaptive regression splines. In: *The Annals of Statistics* 19 (1991), Nr. 1, S. 1–67

[Hastie u. a.] Hastie, Trevor ; Tibshirani, Robert ; Narasimhan, Balasubramanian ; Chu, Gilbert: *impute: Imputation for microarray data*. – R package version 1.0-5

[Hothorn u. a. 2006] Hothorn, Torsten ; Hornik, Kurt ; Zeileis, Achim: Unbiased Recursive Partitioning: A Conditional Inference Framework. In: *Journal of Computational and Graphical Statistics* 15 (2006), Nr. 3, S. 651–674

[R Development Core Team 2008] R Development Core Team: *R: A Language and Environment for Statistical Computing*. Wien, Österreich: R Foundation for Statistical Computing (Veranst.), 2008. – URL http://www.R-project.org. – ISBN 3-900051-07-0

[Strobl u. a. 2007] Strobl, Carolin ; Boulesteix, Anne-Laure ; Zeileis, Achim ; Hothorn, Torsten: Bias in random forest variable importance measures: Illustrations, sources and a solution. In: *BMC Bioinformatics* 8 (2007), Nr. 25

[Svedjar 2007] Svedjar, Viola: *Variablenselektion in Klassifikationsbäumen unter spezieller Berücksichtigung von fehlenden Werten*, Ludwig-Maximilians-Universität München, Diplomarbeit, 2007

[Svedjar 2008] Svedjar, Viola: *persönliche Kommunikation*. 2008

[Troyanskaya u. a. 2001] Troyanskaya, Olga ; Cantor, Michael ; Sherlock, Gavin ; Brown, Pat ; Hastie, Trevor ; Tibshirani, Robert ; Botstein, David ; Altman, Russ B.: Missing value estimation methods for DNA microarrays. In: *Bioinformatics* 17 (2001), Nr. 6, S. 520–525

A. Technische Daten

Die Simulationen wurden mit R-Version 2.7.0 erstellt. Die Version des Pakets `party` hat die Nummer 0.9-97, die Versionsnummer des Pakets `impute` lautet 1.0-5. Das Paket `mvtnorm` wurde in der Version 0.9-0 verwendet.

B. elektronischer Anhang

Der elektronische Anhang enthält alle R-Dateien, die zur Simulation benötigt wurden. Dies umfasst die im Folgenden vorgestellten Funktionen zu den datengenerierenden Prozessen („funktionenZumDGP.R“), zur Erzeugung von fehlenden Werten („funktionenZuMV.R“) und zur Berechnung der Random Forests („funktionenZuRF.R“). Des Weiteren sind die Dateien zum Durchlauf der Simulationen („Simulationen.R“) und zur Erstellung der Grafiken („funktionenFürDiePlots.R“) enthalten. Außerdem finden sich dort R-Dateien zur Ausgabe der Ausreißer („ausgabeAusreißer.R“) bzw. der nicht sichtbaren Datenpunkte in den Grafiken, zur Ausgabe der Zusammenfassungen („ausgabeSummary.R“) und zur Berechnung der t-Tests („t-test.R“). In der Datei „Berechnungen.R“ sind die eigentlichen R-Läufe zu finden.

Die Datei „res.Rda“ enthält die während der Simulation gewonnenen Ergebnisse.

Diese Arbeit selbst findet sich im PDF-Format ebenfalls im elektronischen Anhang.

C. Funktionen zu den datengenerierenden Prozessen

C.1. dgp1

aufbauend auf der Funktion `create` von Svedjar [2007]

Argumente:

`niter`	gewünschte Anzahl der Datensätze, Default: 5
`n`	gewünschte Anzahl an Beobachtungen je Datensatz, Default: 100
`sigma`	Kovarianzmatrix, Default: Einheitsmatrix
`coef`	Koeffizienten-Vektor β, Default: $(1, 2, 3, 4, 5)'$

Wert:

`datsimul`	Liste mit `niter` Datensätzen (`data.frames`)

```
> dgp1 <- function(niter = 5, n = 100, sigma = diag(rep(1, 5)),
+     coef = 1:5) {
+     stopifnot(require("mvtnorm"))
+     datsimul <- list()
+     for (i in 1:niter) {
+         X <- rmvnorm(n, mean = rep(0, 5), sigma = sigma)
+         pi <- as.vector(exp(X %*% coef)/(1 + exp(X %*% coef)))
+         y <- rbinom(n, 1, pi) + 1
+         dat <- as.data.frame(cbind(y, X))
+         dat$y <- as.factor(y)
+         names(dat) <- c("y", paste("x", 1:5, sep = ""))
+         datsimul[[i]] <- dat
+     }
+     return(datsimul)
+ }
```

C.2. dgp2

Argumente:

`niter`	gewünschte Anzahl der Datensätze, Default: 5
`n`	gewünschte Anzahl an Beobachtungen je Datensatz, Default: 100
`sigma`	Kovarianzmatrix, Default: Einheitsmatrix
`u`	gewünschte Anzahl an Variablen ohne Einfluss, Default: 5

Wert:

`datsimul`	Liste mit `niter` Datensätzen (`data.frames`)

```
> dgp2 <- function(niter = 5, n = 100, sigma = diag(rep(1, 5)),
+     u = 5) {
+     stopifnot(require("mvtnorm"))
+     datsimul <- list()
+     for (i in 1:niter) {
+         p <- rep(0.01, 5)
+         while (all(p <= 0.05)) {
+             X <- rmvnorm(n, mean = rep(0, 5), sigma = sigma)
+             U <- apply(X, 2, pnorm)
+             for (j in 1:5) {
+                 p[j] <- ks.test(U[, j], y = punif,
+                     exact = FALSE)$p.value
+             }
+         }
+         y <- 10 * sin(pi * U[, 1] * U[, 2]) +
+             20 * (U[, 3] - 0.5)^2 + 10 * U[, 4] + 5 * U[, 5]
+         p <- rep(0.01, 5)
+         while (all(p <= 0.05)) {
+             W <- rmvnorm(n, mean = rep(0, u),
+                 sigma = diag(rep(1, u)))
+             V <- apply(W, 2, pnorm)
+             for (j in 1:5) {
+                 p[j] <- ks.test(V[, j], y = punif)$p.value
+             }
+         }
+         dat <- as.data.frame(cbind(y, U, V))
+         names(dat) <- c("y", paste("x", 1:5, sep = ""),
+             paste("u", 1:u, sep = ""))
+         datsimul[[i]] <- dat
+     }
+     return(datsimul)
+ }
```

D. Funktionen zur Erzeugung von fehlenden Werten

Diese Funktionen sind von Svedjar [2007] übernommen worden.

D.1. deleteMAR1

Argumente:

`datsimul`	Liste von gleich großen Datensätzen, z. B. aus einer `dgp`-Funktion
`mv`	„missing values“-Matrix mit Spalten 1-3 (pro Variable eine Zeile) Spalte 1: Anteil fehlender Werte in der Variable Spalte 2: Variable, in der die Werte gestrichen werden sollen Spalte 3: Variable, die als Beurteilungskriterium dient

Wert:

`datsimul`	Liste von Datensätzen mit zufällig fehlenden Werten nach MAR1

```
> deleteMAR1 <- function(datsimul,
+     mv = matrix(data = c(0.2, 1, 2,
+                          0.2, 4, 5,
+                          0.1, 3, 4), byrow = TRUE, ncol = 3), ...) {
+     niter <- length(datsimul)
+     n <- nrow(datsimul[[1]])
+     mv[, 1] <- mv[, 1] * n
+     for (i in 1:niter) {
+         x <- datsimul[[i]][, -1]
+         for (j in 1:nrow(mv)) {
+             z <- rank(x[, mv[j, 3]])
+             p <- z/sum(1:n)
+             x[sample(n, mv[j, 1], prob = p), mv[j, 2]] <- NA
+         }
+         datsimul[[i]][, -1] <- x
+     }
+     datsimul
+ }
```

D.2. deleteMAR2

Argumente:

datsimul	Liste von gleich großen Datensätzen, z. B. aus einer dgp-Funktion
mv	„missing values“-Matrix mit Spalten 1-3 (pro Variable eine Zeile) Spalte 1: Anteil fehlender Werte in der Variable Spalte 2: Variable, in der die Werte gestrichen werden sollen Spalte 3: Variable, die als Beurteilungskriterium dient

Wert:

datsimul	Liste von Datensätzen mit zufällig fehlenden Werten nach MAR2

```
> deleteMAR2 <- function(datsimul,
+     mv = matrix(data = c(0.2, 1, 2,
+                          0.2, 4, 5,
+                          0.1, 3, 4), byrow = TRUE, ncol = 3), ...) {
+     niter <- length(datsimul)
+     n <- nrow(datsimul[[1]])
+     mv[, 1] <- mv[, 1] * n
+     for (i in 1:niter) {
+         x <- datsimul[[i]][, -1]
+         for (j in 1:nrow(mv)) {
+             z <- rep(0, n)
+             z[x[, mv[j, 3]] >= median(x[, mv[j, 3]])] <- 1
+             S <- sum(z)
+             p <- rep(0.1/(n - S), n)
+             p[z == 1] <- 0.9/(n - S)
+             x[sample(n, mv[j, 1], prob = p), mv[j, 2]] <- NA
+         }
+         datsimul[[i]][, -1] <- x
+     }
+     datsimul
+ }
```

D.3. deleteMAR3

Argumente:

datsimul	Liste von gleich großen Datensätzen, z. B. aus einer dgp-Funktion
mv	„missing values“-Matrix mit Spalten 1-3 (pro Variable eine Zeile) Spalte 1: Anteil fehlender Werte in der Variable Spalte 2: Variable, in der die Werte gestrichen werden sollen Spalte 3: Variable, die als Beurteilungskriterium dient

Wert:

datsimul	Liste von Datensätzen mit zufällig fehlenden Werten nach MAR3

```
> deleteMAR3 <- function(datsimul,
+     mv = matrix(data = c(0.2, 1, 2,
+                          0.2, 4, 5,
+                          0.1, 3, 5), byrow = TRUE, ncol = 3), ...) {
+     niter <- length(datsimul)
+     n <- nrow(datsimul[[1]])
+     mv[, 1] <- mv[, 1] * n
+     for (i in 1:niter) {
+         x <- datsimul[[i]][, -1]
+         for (j in 1:nrow(mv)) {
+             a <- quantile(x[, mv[j, 3]],
+                 probs = (1 - (mv[j, 1]/n)))
+             x[, mv[j, 2]][x[, mv[j, 3]] >= a] <- NA
+         }
+         datsimul[[i]][, -1] <- x
+     }
+     datsimul
+ }
```

D.4. deleteMAR4

Argumente:

datsimul	Liste von gleich großen Datensätzen, z. B. aus einer dgp-Funktion
mv	„missing values“-Matrix mit Spalten 1-3 (pro Variable eine Zeile) Spalte 1: Anteil fehlender Werte in der Variable Spalte 2: Variable, in der die Werte gestrichen werden sollen Spalte 3: Variable, die als Beurteilungskriterium dient

Wert:

datsimul	Liste von Datensätzen mit zufällig fehlenden Werten nach MAR4

```
> deleteMAR4 <- function(datsimul,
+     mv = matrix(data = c(0.2, 1, 2,
+                          0.2, 4, 5,
+                          0.1, 3, 5), byrow = TRUE, ncol = 3), ...) {
+     niter <- length(datsimul)
+     n <- nrow(datsimul[[1]])
+     mv[, 1] <- mv[, 1] * n
+     for (i in 1:niter) {
+         x <- datsimul[[i]][, -1]
+         for (j in 1:nrow(mv)) {
+             a <- quantile(x[, mv[j, 3]],
+                 probs = (1 - (0.5 * mv[j, 1]/n)))
+             x[, mv[j, 2]][x[, mv[j, 3]] >= a] <- NA
```

```
+             b <- quantile(x[, mv[j, 3]],
+                 probs = (0.5 * mv[j, 1]/n))
+             x[, mv[j, 2]][x[, mv[j, 3]] <= b] <- NA
+         }
+         datsimul[[i]][, -1] <- x
+     }
+     datsimul
+ }
```

D.5. deleteMCAR

Argumente:

datsimul	Liste von gleich großen Datensätzen, z. B. aus einer dgp-Funktion
mv	„missing values“-Vektor mit den Anteilen fehlender Werte (pro Variable eine Zeile)

Wert:

datsimul	Liste von Datensätzen mit vollständig zufällig fehlenden Werten

```
> deleteMCAR <- function(datsimul, mv = c(0.2, 0, 0.1, 0.2, 0), ...) {
+     niter <- length(datsimul)
+     nvar <- length(mv)
+     n <- nrow(datsimul[[1]])
+     mv <- mv * n
+     for (i in 1:niter) {
+         x <- datsimul[[i]][, -1]
+         for (j in 1:nvar) {
+             if (mv[j] > 0) {
+                 x[sample(n, mv[j]), j] <- NA
+             }
+         }
+         datsimul[[i]][, -1] <- x
+     }
+     datsimul
+ }
```

E. Funktionen zur Berechnung der Random Forests

Diese Hilfsfunktion ist nötig, da $\log(0) = -\infty$ ist. `x` ist ein Vektor.

```
> mylog <- function(x) {
+     x[x < 1e-05] <- 1e-05
+     log(x)
+ }
```

E.1. RF1

Argumente:	
`mvfun`	eine der `delete`-Funktionen; nur nötig, falls auch fehlende Werte eingestreut werden
`test`	kompletter Test-Datensatz als Liste mit einem Element
`dgpfun.niter`	gewünschte Anzahl der zu erzeugenden Lerndatensätze, Default: 500
`lernMV, testMV`	logischer Wert: TRUE, falls im Lern- bzw. Testdatensatz fehlende Werte eingestreut werden sollen, Default: FALSE
`imp`	logischer Wert: TRUE, falls *knn*-Imputation angewendet werden sollen, Default: FALSE
`na.omit`	logischer Wert: TRUE, falls fehlende Werte gestrichen werden sollen, Default: FALSE
Wert:	
`loglik`	Vektor mit der mittleren Binomial-Log-Likelihood pro erzeugtem Lern-Datensatz

```
> RF1 <- function(mvfun = NULL, test, dgpfun.niter = 500,
+     lernMV = FALSE, testMV = FALSE, imp = FALSE, na.omit = FALSE,
+     ...) {
+     dat <- dgp1(niter = dgpfun.niter, n = 200, ...)
+     if (lernMV) {
+         dat <- mvfun(dat, ...)
+     }
+     if (testMV) {
```

```
+         test <- mvfun(test, ...)
+     }
+     test <- test[[1]]
+     if (imp & testMV) {
+         stopifnot(require("impute"))
+         test[, -1] <-
+             as.data.frame(impute.knn(as.matrix(test[, -1])))
+     }
+     stopifnot(require("party"))
+     loglik <- vector("numeric")
+     niter <- length(dat)
+     for (i in 1:niter) {
+         loglik[i] <- NA
+         if (imp & lernMV) {
+             dat[[i]][-1] <-
+                 as.data.frame(impute.knn(as.matrix(dat[[i]][-1])))
+         }
+         if (na.omit)
+             dat[[i]] <- dat[[i]][complete.cases(dat[[i]]), ]
+         rf <- try(cforest(y ~ ., data = dat[[i]],
+             control = cforest_control(maxsurrogate = 3,
+             ntree = 50, minsplit = 30)))
+         if (inherits(rf, "try-error"))
+             next()
+         print(i)
+         p <- treeresponse(rf, newdata = test)
+         P <- matrix(unlist(p), byrow = TRUE, ncol = 2)
+         p <- P[, 2]
+         yvec <- as.numeric(test$y) - 1
+         loglik[i] <- mean(yvec * mylog(p) +
+             (1 - yvec) * mylog(1 - p))
+     }
+     return(loglik)
+ }
```

E.2. RF2

Argumente:

`mvfun`	eine der `delete`-Funktionen; nur nötig, falls auch fehlende Werte eingestreut werden
`test`	kompletter Test-Datensatz als Liste mit einem Element
`dgpfun.niter`	gewünschte Anzahl der zu erzeugenden Lerndatensätze, Default: 500
`lernMV, testMV`	logischer Wert: TRUE, falls im Lern- bzw. Testdatensatz fehlende Werte eingestreut werden sollen, Default: FALSE
`imp`	logischer Wert: TRUE, falls *knn*-Imputation angewendet werden sollen, Default: FALSE
`na.omit`	logischer Wert: TRUE, falls fehlende Werte gestrichen werden sollen, Default: FALSE

Wert:

`MSE`	Vektor mit der mittleren quadratischen Abweichung pro erzeugtem Lern-Datensatz

```
> RF2 <- function(mvfun = NULL, test, dgpfun.niter = 500,
+     lernMV = FALSE, testMV = FALSE, imp = FALSE, na.omit = FALSE,
+     ...) {
+     dat <- dgp2(niter = dgpfun.niter, n = 200, ...)
+     if (lernMV) {
+         dat <- mvfun(dat, ...)
+     }
+     if (testMV) {
+         test <- mvfun(test, ...)
+     }
+     test <- test[[1]]
+     if (imp & testMV) {
+         stopifnot(require("impute"))
+         test[, -1] <-
+             as.data.frame(impute.knn(as.matrix(test[, -1])))
+     }
+     stopifnot(require("party"))
+     MSE <- vector("numeric")
+     niter <- length(dat)
+     for (i in 1:niter) {
+         MSE[i] <- NA
+         if (imp & lernMV) {
+             dat[[i]][-1] <-
+                 as.data.frame(impute.knn(as.matrix(dat[[i]][-1])))
+         }
+         if (na.omit)
```

```
+             dat[[i]] <- dat[[i]][complete.cases(dat[[i]]), ]
+         rf <- try(cforest(y ~ ., data = dat[[i]],
+             control = cforest_control(maxsurrogate = 3, ntree = 50,
+             minsplit = 30)))
+         if (inherits(rf, "try-error"))
+             next()
+         print(i)
+         f <- predict(rf, newdata = test)
+         MSE[i] <- mean((test$y - f)^2)
+     }
+     return(MSE)
+ }
```

F. Tabellen zum Fehler der Simulationen

„Srgt“ steht für diejenigen Fälle, in denen Surrogat-Variablen verwendet wurden. „Imp“ dagegen bezeichnet die Fälle mit erfolgter Imputation.

F.1. Simulation zur Klassifikation

Sämtliche Tabellen sind auf vier Nachkommastellen gerundet.

F.1.1. Gleichmäßige, hohe Korrelationen

MAR 1

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
Goldstandard	-0.3284	-0.1486	-0.1349	-0.1418	-0.1238	-0.1103	
-NA	-0.2925	-0.1583	-0.1416	-0.1515	-0.1297	-0.1191	
Srgt – lernMV	-0.2697	-0.1656	-0.1499	-0.1544	-0.1336	-0.1132	2
Srgt – testMV	-0.2699	-0.1649	-0.1479	-0.1522	-0.1339	-0.1195	
Srgt – lerntestMV	-0.2636	-0.1795	-0.1554	-0.1617	-0.1408	-0.1230	1
Imp – lernMV	-0.2247	-0.1499	-0.1385	-0.1407	-0.1268	-0.1122	
Imp – testMV	-0.2407	-0.1576	-0.1422	-0.1490	-0.1319	-0.1154	
Imp – lerntestMV	-0.2233	-0.1561	-0.1463	-0.1504	-0.1355	-0.1171	

Differenzen zum Goldstandard

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
-NA	-0.1561	-0.0285	-0.0056	-0.0097	0.0102	0.1949	
Srgt – lernMV	-0.1329	-0.0364	-0.0066	-0.0126	0.0048	0.1874	2
Srgt – testMV	-0.0810	-0.0324	-0.0100	-0.0104	0.0076	0.1289	
Srgt – lerntestMV	-0.1291	-0.0434	-0.0189	-0.0200	-0.0048	0.1883	1
Imp – lernMV	-0.0748	-0.0228	0.0013	0.0011	0.0179	0.1676	
Imp – testMV	-0.1123	-0.0260	-0.0071	-0.0072	0.0104	0.1813	
Imp – lerntestMV	-0.0861	-0.0280	-0.0106	-0.0087	0.0062	0.1708	

MAR 2

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
Goldstandard	-0.3284	-0.1486	-0.1349	-0.1418	-0.1238	-0.1103	
-NA	-0.3172	-0.1591	-0.1460	-0.1510	-0.1337	-0.1205	
Srgt – lernMV	-0.2614	-0.1601	-0.1506	-0.1523	-0.1360	-0.1140	2
Srgt – testMV	-0.3164	-0.1544	-0.1405	-0.1489	-0.1321	-0.1200	
Srgt – lerntestMV	-0.3187	-0.1859	-0.1554	-0.1665	-0.1402	-0.1208	5
Imp – lernMV	-0.2331	-0.1438	-0.1323	-0.1366	-0.1236	-0.1118	
Imp – testMV	-0.2661	-0.1578	-0.1445	-0.1528	-0.1332	-0.119	
Imp – lerntestMV	-0.2504	-0.1555	-0.1414	-0.1447	-0.1295	-0.1159	

Differenzen zum Goldstandard

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
-NA	-0.1855	-0.0266	-0.0084	-0.0092	0.0080	0.1822	
Srgt – lernMV	-0.1174	-0.0275	-0.0103	-0.0104	0.0060	0.1519	2
Srgt – testMV	-0.2061	-0.0226	-0.0063	-0.0071	0.0113	0.1750	
Srgt – lerntestMV	-0.1842	-0.0496	-0.0202	-0.0245	0.0030	0.1821	5
Imp – lernMV	-0.0983	-0.0120	0.0019	0.0052	0.0173	0.1812	
Imp – testMV	-0.1383	-0.0244	-0.0119	-0.0110	0.0058	0.1987	
Imp – lerntestMV	-0.1353	-0.0219	-0.0040	-0.0030	0.0100	0.1905	

MAR 3

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
Goldstandard	-0.3284	-0.1486	-0.1349	-0.1418	-0.1238	-0.1103	
-NA	-0.2048	-0.1504	-0.1351	-0.1403	-0.1270	-0.1150	
Srgt – lernMV	-0.2422	-0.1634	-0.1444	-0.1490	-0.1288	-0.1140	1
Srgt – testMV	-0.2616	-0.1515	-0.1374	-0.1431	-0.1285	-0.1129	
Srgt – lerntestMV	-0.2463	-0.1572	-0.1427	-0.1489	-0.1340	-0.1111	1
Imp – lernMV	-0.324	-0.1860	-0.1576	-0.1656	-0.1396	-0.1154	
Imp – testMV	-0.8878	-0.3160	-0.2384	-0.2670	-0.1819	-0.1384	
Imp – lerntestMV	-0.2776	-0.1849	-0.1626	-0.1715	-0.1490	-0.1171	

Differenzen zum Goldstandard

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
-NA	-0.0794	-0.0179	-0.0025	0.0015	0.0138	0.2040	
Srgt – lernMV	-0.1189	-0.0283	-0.0094	-0.0071	0.0116	0.2010	1
Srgt – testMV	-0.1134	-0.0188	-0.0020	-0.0013	0.0134	0.1603	
Srgt – lerntestMV	-0.1204	-0.0274	-0.0072	-0.0069	0.0117	0.1827	1
Imp – lernMV	-0.2079	-0.0449	-0.0191	-0.0238	0.0002	0.1542	
Imp – testMV	-0.7060	-0.1778	-0.0990	-0.1252	-0.0450	0.0822	
Imp – lerntestMV	-0.1380	-0.0454	-0.0269	-0.0297	-0.0067	0.1163	

MAR 4

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
Goldstandard	-0.3284	-0.1486	-0.1349	-0.1418	-0.1238	-0.1103	
-NA	-0.2473	-0.1482	-0.1397	-0.1432	-0.1276	-0.1111	
Srgt – lernMV	-0.2313	-0.1529	-0.1360	-0.1437	-0.1276	-0.1110	1
Srgt – testMV	-0.2366	-0.1551	-0.1395	-0.1461	-0.1297	-0.1144	
Srgt – lerntestMV	-0.2187	-0.1614	-0.1429	-0.1481	-0.1304	-0.1149	
Imp – lernMV	-0.1968	-0.1505	-0.1346	-0.1404	-0.1257	-0.1126	
Imp – testMV	-0.2817	-0.1621	-0.1472	-0.1521	-0.1356	-0.1174	
Imp – lerntestMV	-0.2123	-0.1552	-0.1406	-0.1454	-0.1306	-0.1157	

Differenzen zum Goldstandard

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
-NA	-0.1272	-0.0193	-0.0018	-0.0014	0.0147	0.1519	
Srgt – lernMV	-0.0795	-0.0194	-0.0078	-0.0019	0.0116	0.2083	1
Srgt – testMV	-0.1206	-0.0243	-0.0057	-0.0044	0.0124	0.2011	
Srgt – lerntestMV	-0.0836	-0.0257	-0.0079	-0.0063	0.0074	0.1867	
Imp – lernMV	-0.0731	-0.0177	-0.0021	0.0014	0.0187	0.1982	
Imp – testMV	-0.1136	-0.0271	-0.0138	-0.0103	0.0054	0.1424	
Imp – lerntestMV	-0.0746	-0.0250	-0.0077	-0.0036	0.0115	0.1749	

MCAR

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
Goldstandard	-0.3284	-0.1486	-0.1349	-0.1418	-0.1238	-0.1103	
-NA	-0.2296	-0.1675	-0.1493	-0.1522	-0.1314	-0.1148	
Srgt – lernMV	-0.2335	-0.1690	-0.1543	-0.1557	-0.1365	-0.1117	
Srgt – testMV	-0.2409	-0.1528	-0.1397	-0.1457	-0.1341	-0.1193	
Srgt – lerntestMV	-0.2391	-0.1673	-0.1525	-0.1576	-0.1421	-0.1255	
Imp – lernMV	-0.2222	-0.1564	-0.1378	-0.1434	-0.1265	-0.1151	
Imp – testMV	-0.2090	-0.1511	-0.1396	-0.1450	-0.1323	-0.1224	
Imp – lerntestMV	-0.2359	-0.1688	-0.1504	-0.1560	-0.1390	-0.1215	

Differenzen zum Goldstandard

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
-NA	-0.1073	-0.0388	-0.0115	-0.0104	0.0056	0.2008	
Srgt – lernMV	-0.1085	-0.0398	-0.0145	-0.0139	0.0045	0.1946	
Srgt – testMV	-0.0797	-0.0210	-0.0083	-0.0040	0.0086	0.1916	
Srgt – lerntestMV	-0.0970	-0.0340	-0.0169	-0.0158	-0.0012	0.1792	
Imp – lernMV	-0.0870	-0.0186	-0.0076	-0.0017	0.0102	0.1302	
Imp – testMV	-0.0897	-0.0211	-0.0059	-0.0032	0.0075	0.1455	
Imp – lerntestMV	-0.1072	-0.0344	-0.0152	-0.0142	0.0053	0.1708	

F.1.2. Blockweise hohe Korrelationen

MAR 1

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
Goldstandard	-0.4714	-0.2758	-0.2512	-0.2585	-0.2315	-0.2023	
-NA	-0.3744	-0.2995	-0.2743	-0.2835	-0.2616	-0.2349	
Srgt – lernMV	-0.3801	-0.2927	-0.2614	-0.2689	-0.2427	-0.2038	
Srgt – testMV	-0.3529	-0.2921	-0.2645	-0.2683	-0.2431	-0.2125	
Srgt – lerntestMV	-0.4177	-0.2910	-0.2646	-0.2729	-0.2475	-0.2149	
Imp – lernMV	-0.3373	-0.2767	-0.2523	-0.2559	-0.2361	-0.2008	
Imp – testMV	-0.3725	-0.2924	-0.2682	-0.2680	-0.2395	-0.1993	
Imp – lerntestMV	-0.4080	-0.2898	-0.2613	-0.2714	-0.2429	-0.2148	

Differenzen zum Goldstandard

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
-NA	-0.1238	-0.0554	-0.0310	-0.0250	0.0050	0.2116	
Srgt – lernMV	-0.1206	-0.0410	-0.0192	-0.0104	0.0283	0.1802	
Srgt – testMV	-0.0944	-0.0419	-0.0136	-0.0098	0.0153	0.1605	
Srgt – lerntestMV	-0.1713	-0.0435	-0.0144	-0.0144	0.0152	0.1877	
Imp – lernMV	-0.0879	-0.0227	0.0006	0.0026	0.0270	0.2028	
Imp – testMV	-0.1087	-0.0463	-0.0140	-0.0095	0.0224	0.2022	
Imp – lerntestMV	-0.1318	-0.0402	-0.0119	-0.0129	0.0178	0.1887	

MAR 2

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
Goldstandard	-0.4714	-0.2758	-0.2512	-0.2585	-0.2315	-0.2023	
-NA	-0.4286	-0.3131	-0.2850	-0.2936	-0.2655	-0.2388	
Srgt – lernMV	-0.3810	-0.2922	-0.2617	-0.2682	-0.2373	-0.2085	
Srgt – testMV	-0.4157	-0.2811	-0.2590	-0.2641	-0.2395	-0.1973	
Srgt – lerntestMV	-0.3896	-0.2809	-0.2640	-0.2671	-0.2450	-0.2186	3
Imp – lernMV	-0.4853	-0.2740	-0.2448	-0.2575	-0.2289	-0.1941	
Imp – testMV	-0.3470	-0.2849	-0.2588	-0.2653	-0.2415	-0.2096	
Imp – lerntestMV	-0.3649	-0.2916	-0.2628	-0.2698	-0.2442	-0.2115	

Differenzen zum Goldstandard

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
-NA	-0.1865	-0.0691	-0.0385	-0.0350	0.0037	0.1461	
Srgt – lernMV	-0.155	-0.0360	-0.0070	-0.0096	0.0188	0.2303	
Srgt – testMV	-0.1572	-0.0367	-0.0035	-0.0056	0.0180	0.1698	
Srgt – lerntestMV	-0.1208	-0.0344	-0.0059	-0.0076	0.0134	0.2106	3
Imp – lernMV	-0.2536	-0.0270	0.0000	0.0010	0.0328	0.1727	
Imp – testMV	-0.1012	-0.0359	-0.0052	-0.0067	0.0170	0.1439	
Imp – lerntestMV	-0.1289	-0.0376	-0.0117	-0.0113	0.0148	0.1767	

MAR 3

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
Goldstandard	-0.4714	-0.2758	-0.2512	-0.2585	-0.2315	-0.2023	
-NA	-0.5131	-0.3591	-0.3291	-0.3375	-0.3025	-0.2581	
Srgt – lernMV	-0.4031	-0.2861	-0.2591	-0.2681	-0.2416	-0.2014	22
Srgt – testMV	-0.3724	-0.2778	-0.2569	-0.2614	-0.2369	-0.2075	
Srgt – lerntestMV	-0.4228	-0.2992	-0.2702	-0.2709	-0.2354	-0.2015	18
Imp – lernMV	-0.3949	-0.2881	-0.2688	-0.2684	-0.2397	-0.2055	
Imp – testMV	-0.5550	-0.3421	-0.3058	-0.3168	-0.2722	-0.2174	
Imp – lerntestMV	-0.3576	-0.2935	-0.2631	-0.2698	-0.2459	-0.2064	

Differenzen zum Goldstandard

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
-NA	-0.2949	-0.1161	-0.0696	-0.0789	-0.0414	0.1863	
Srgt – lernMV	-0.1941	-0.0400	-0.0056	-0.0081	0.0299	0.2592	22
Srgt – testMV	-0.0982	-0.0343	-0.0047	-0.0028	0.0150	0.2088	
Srgt – lerntestMV	-0.1738	-0.0442	-0.0109	-0.0127	0.0177	0.1033	18
Imp – lernMV	-0.1655	-0.0394	-0.0147	-0.0099	0.0205	0.2113	
Imp – testMV	-0.3129	-0.0934	-0.0552	-0.0583	-0.0183	0.2088	
Imp – lerntestMV	-0.1296	-0.0419	-0.0103	-0.0113	0.0151	0.1526	

MAR 4

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
Goldstandard	-0.4714	-0.2758	-0.2512	-0.2585	-0.2315	-0.2023	
-NA	-0.4523	-0.3073	-0.2857	-0.2949	-0.2647	-0.2328	
Srgt – lernMV	-0.3640	-0.2729	-0.2540	-0.2601	-0.2390	-0.2070	
Srgt – testMV	-0.3950	-0.2768	-0.2531	-0.2616	-0.2368	-0.2025	
Srgt – lerntestMV	-0.3742	-0.2930	-0.2592	-0.2673	-0.2396	-0.2078	
Imp – lernMV	-0.3847	-0.2793	-0.2544	-0.2580	-0.2314	-0.2007	
Imp – testMV	-0.4597	-0.3031	-0.2708	-0.2824	-0.2526	-0.2136	
Imp – lerntestMV	-0.4146	-0.3034	-0.2736	-0.2798	-0.2492	-0.2017	

Differenzen zum Goldstandard

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
-NA	-0.2189	-0.0682	-0.0297	-0.0363	-0.0046	0.2001	
Srgt – lernMV	-0.1173	-0.0291	-0.0006	-0.0016	0.0237	0.1310	
Srgt – testMV	-0.1632	-0.0294	-0.0056	-0.0031	0.0263	0.1629	
Srgt – lerntestMV	-0.1277	-0.0430	-0.0063	-0.0088	0.0224	0.1989	
Imp – lernMV	-0.1279	-0.0413	-0.0017	0.0005	0.0302	0.2440	
Imp – testMV	-0.2574	-0.0566	-0.0270	-0.0238	0.0109	0.2188	
Imp – lerntestMV	-0.1694	-0.0568	-0.0184	-0.0213	0.0110	0.2447	

MCAR

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
Goldstandard	-0.4714	-0.2758	-0.2512	-0.2585	-0.2315	-0.2023	
-NA	-0.4483	-0.2948	-0.2713	-0.2819	-0.2586	-0.2285	
Srgt – lernMV	-0.4496	-0.2930	-0.2743	-0.2768	-0.2465	-0.2127	1
Srgt – testMV	-0.4020	-0.2908	-0.2680	-0.2713	-0.2482	-0.2124	
Srgt – lerntestMV	-0.4282	-0.2993	-0.2785	-0.2815	-0.2550	-0.2203	
Imp – lernMV	-0.3823	-0.2705	-0.2444	-0.2512	-0.2260	-0.1965	
Imp – testMV	-0.3732	-0.2841	-0.2567	-0.2649	-0.2398	-0.2086	
Imp – lerntestMV	-0.3543	-0.2823	-0.2594	-0.2643	-0.2406	-0.2042	

Differenzen zum Goldstandard

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
-NA	-0.2460	-0.0515	-0.0212	-0.0234	0.0030	0.1524	
Srgt – lernMV	-0.2375	-0.0548	-0.0140	-0.0180	0.0216	0.1984	1
Srgt – testMV	-0.1205	-0.0460	-0.0116	-0.0127	0.0159	0.2405	
Srgt – lerntestMV	-0.1608	-0.0599	-0.0250	-0.0230	0.0063	0.2254	
Imp – lernMV	-0.1109	-0.0242	0.0087	0.0074	0.0333	0.2279	
Imp – testMV	-0.1529	-0.0411	-0.0047	-0.0063	0.0238	0.2327	
Imp – lerntestMV	-0.1072	-0.0364	-0.0109	-0.0058	0.0154	0.1788	

F.1.3. Gleichmäßige, niedrige Korrelationen

MAR 1

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
Goldstandard	-0.4490	-0.3596	-0.3417	-0.3431	-0.3204	-0.2931	
-NA	-0.5106	-0.3934	-0.3717	-0.3778	-0.3548	-0.3164	
Srgt – lernMV	-0.5135	-0.3705	-0.3480	-0.3569	-0.3347	-0.3029	
Srgt – testMV	-0.4613	-0.3855	-0.3678	-0.3713	-0.3523	-0.3245	
Srgt – lerntestMV	-0.4676	-0.4014	-0.3828	-0.3883	-0.3682	-0.3509	
Imp – lernMV	-0.5395	-0.3720	-0.3428	-0.3496	-0.3248	-0.2986	
Imp – testMV	-0.5085	-0.4120	-0.3843	-0.3923	-0.3660	-0.3357	
Imp – lerntestMV	-0.4816	-0.4151	-0.3904	-0.3980	-0.3731	-0.3340	

Differenzen zum Goldstandard

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
-NA	-0.1745	-0.0638	-0.0345	-0.0348	-0.0015	0.0853	
Srgt – lernMV	-0.1582	-0.0347	-0.0152	-0.0138	0.0114	0.1139	
Srgt – testMV	-0.1587	-0.0520	-0.0259	-0.0282	-0.0022	0.0867	
Srgt – lerntestMV	-0.1232	-0.0712	-0.0439	-0.0452	-0.0206	0.0350	
Imp – lernMV	-0.1616	-0.0342	-0.0049	-0.0065	0.0256	0.1427	
Imp – testMV	-0.1887	-0.0781	-0.0494	-0.0492	-0.0150	0.0881	
Imp – lerntestMV	-0.1450	-0.0907	-0.0512	-0.0550	-0.0187	0.0437	

MAR 2

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
Goldstandard	-0.4490	-0.3596	-0.3417	-0.3431	-0.3204	-0.2931	
-NA	-0.5714	-0.4016	-0.3728	-0.3845	-0.3515	-0.3223	
Srgt – lernMV	-0.4488	-0.3710	-0.3500	-0.3551	-0.3309	-0.3111	2
Srgt – testMV	-0.4678	-0.3849	-0.3679	-0.3719	-0.3485	-0.3179	
Srgt – lerntestMV	-0.4703	-0.4032	-0.3884	-0.3904	-0.3717	-0.3397	1
Imp – lernMV	-0.4336	-0.3620	-0.3438	-0.3488	-0.3314	-0.2972	
Imp – testMV	-0.5702	-0.4147	-0.3882	-0.3975	-0.3678	-0.3400	
Imp – lerntestMV	-0.5793	-0.4012	-0.3769	-0.3839	-0.3624	-0.3332	

Differenzen zum Goldstandard

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
-NA	-0.2395	-0.0663	-0.0368	-0.0415	-0.0100	0.0780	
Srgt – lernMV	-0.1525	-0.0364	-0.0126	-0.0125	0.0142	0.1115	2
Srgt – testMV	-0.1501	-0.0548	-0.0278	-0.0289	-0.0017	0.0723	
Srgt – lerntestMV	-0.1341	-0.0675	-0.0494	-0.0469	-0.0229	0.0819	1
Imp – lernMV	-0.1248	-0.0335	-0.0056	-0.0057	0.0240	0.0895	
Imp – testMV	-0.2097	-0.0774	-0.0449	-0.0544	-0.0227	0.0408	
Imp – lerntestMV	-0.2767	-0.0628	-0.0390	-0.0408	-0.0145	0.0854	

MAR 3

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
Goldstandard	-0.4490	-0.3596	-0.3417	-0.3431	-0.3204	-0.2931	
-NA	-0.8132	-0.4772	-0.4196	-0.4446	-0.3840	-0.3554	
Srgt – lernMV	-0.4087	-0.3684	-0.3378	-0.3459	-0.3224	-0.3026	59
Srgt – testMV	-0.5520	-0.3819	-0.3546	-0.3647	-0.3406	-0.3018	
Srgt – lerntestMV	-0.4434	-0.3842	-0.3579	-0.3660	-0.3462	-0.3200	69
Imp – lernMV	-0.4358	-0.3600	-0.3371	-0.3438	-0.3229	-0.2953	
Imp – testMV	-0.4747	-0.3840	-0.3604	-0.3637	-0.3384	-0.3033	
Imp – lerntestMV	-0.4873	-0.3780	-0.3514	-0.3601	-0.3336	-0.3152	

Differenzen zum Goldstandard

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
-NA	-0.4584	-0.1478	-0.0868	-0.1015	-0.0438	0.0425	
Srgt – lernMV	-0.1130	-0.0209	-0.0064	-0.0015	0.0296	0.1090	59
Srgt – testMV	-0.2284	-0.0460	-0.0158	-0.0217	0.0060	0.0860	
Srgt – lerntestMV	-0.1304	-0.0457	-0.0315	-0.0296	-0.0101	0.0461	69
Imp – lernMV	-0.1008	-0.0275	0.0018	-0.0008	0.0303	0.0948	
Imp – testMV	-0.1446	-0.0543	-0.0221	-0.0206	0.0148	0.1051	
Imp – lerntestMV	-0.1583	-0.0401	-0.0193	-0.0170	0.0121	0.0990	

MAR 4

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
Goldstandard	-0.4490	-0.3596	-0.3417	-0.3431	-0.3204	-0.2931	
-NA	-0.5994	-0.4095	-0.3765	-0.3874	-0.3553	-0.3185	
Srgt – lernMV	-0.4271	-0.3652	-0.3338	-0.3438	-0.3214	-0.2929	2
Srgt – testMV	-0.4473	-0.3832	-0.3626	-0.3642	-0.3411	-0.2892	
Srgt – lerntestMV	-0.4454	-0.3684	-0.3522	-0.3558	-0.3382	-0.3033	2
Imp – lernMV	-0.6658	-0.3647	-0.3393	-0.3474	-0.3231	-0.2934	
Imp – testMV	-0.4458	-0.3707	-0.3460	-0.3536	-0.3283	-0.2909	
Imp – lerntestMV	-0.4569	-0.3570	-0.3368	-0.3434	-0.3220	-0.2927	

Differenzen zum Goldstandard

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
-NA	-0.2760	-0.0788	-0.0358	-0.0443	-0.0044	0.0804	
Srgt – lernMV	-0.1182	-0.0273	-0.0018	-0.0004	0.0278	0.1187	2
Srgt – testMV	-0.1100	-0.0474	-0.0205	-0.0211	-0.0011	0.1006	
Srgt – lerntestMV	-0.0958	-0.0419	-0.0140	-0.0131	0.0071	0.1183	2
Imp – lernMV	-0.3410	-0.0244	0.0033	-0.0043	0.0254	0.1174	
Imp – testMV	-0.1253	-0.0444	-0.0094	-0.0105	0.0261	0.0890	
Imp – lerntestMV	-0.1085	-0.0281	0.0019	-0.0003	0.0300	0.1014	

MCAR

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
Goldstandard	-0.4490	-0.3596	-0.3417	-0.3431	-0.3204	-0.2931	
-NA	-0.4728	-0.3911	-0.3639	-0.3736	-0.3497	-0.3281	
Srgt – lernMV	-0.4768	-0.3807	-0.3564	-0.3610	-0.3371	-0.2973	
Srgt – testMV	-0.4612	-0.3866	-0.3667	-0.3722	-0.3498	-0.3229	
Srgt – lerntestMV	-0.5166	-0.4048	-0.3839	-0.3904	-0.3673	-0.3460	
Imp – lernMV	-0.4682	-0.3679	-0.3497	-0.3504	-0.3274	-0.2915	
Imp – testMV	-0.5583	-0.4180	-0.3911	-0.3977	-0.3642	-0.3369	
Imp – lerntestMV	-0.4770	-0.3987	-0.3794	-0.3835	-0.3648	-0.3288	

Differenzen zum Goldstandard

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
-NA	-0.1587	-0.0570	-0.0278	-0.0305	-0.0043	0.0946	
Srgt – lernMV	-0.1099	-0.0473	-0.0195	-0.0180	0.0112	0.1243	
Srgt – testMV	-0.1350	-0.0536	-0.0234	-0.0291	-0.0033	0.1058	
Srgt – lerntestMV	-0.2125	-0.0723	-0.0493	-0.0474	-0.0181	0.0610	
Imp – lernMV	-0.1210	-0.0284	-0.0120	-0.0073	0.0177	0.1118	
Imp – testMV	-0.2514	-0.0792	-0.0478	-0.0546	-0.0241	0.0520	
Imp – lerntestMV	-0.1465	-0.0722	-0.0391	-0.0404	-0.0139	0.1164	

F.2. Simulation zur Regression

Sämtliche Tabellen sind auf vier gültige Ziffern bzw. vier Nachkommastellen gerundet – je nachdem, was kürzer ist.

F.2.1. Gleichmäßige, hohe Korrelationen

MAR 1

Fall	Min.	1. Qu.	Median	Mean	3. Qu.	Max.	NA
Goldstandard	1.525	1.763	1.877	1.931	2.049	3.827	
-NA	1.787	2.257	2.404	2.451	2.612	4.142	
Srgt – lernMV	2.017	2.357	2.518	2.557	2.643	4.553	4
Srgt – testMV	2.018	2.253	2.357	2.422	2.495	3.880	
Srgt – lerntestMV	2.344	2.732	2.871	2.963	3.056	7.589	5
Imp – lernMV	1.582	1.835	1.993	1.981	2.101	2.756	
Imp – testMV	1.892	2.183	2.285	2.342	2.371	4.260	
Imp – lerntestMV	2.002	2.247	2.330	2.387	2.441	4.100	

Differenzen zum Goldstandard

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
-NA	-1.566	0.2937	0.5474	0.5202	0.7496	2.112	
Srgt – lernMV	-1.688	0.4389	0.6273	0.6235	0.8084	2.777	4
Srgt – testMV	-1.641	0.2989	0.4474	0.4910	0.6216	2.221	
Srgt – lerntestMV	-0.6472	0.7720	1.031	1.0280	1.201	5.626	5
Imp – lernMV	-1.949	-0.1232	0.0870	0.0501	0.2615	1.115	
Imp – testMV	-1.649	0.2319	0.4261	0.4115	0.5754	2.580	
Imp – lerntestMV	-1.616	0.3193	0.4486	0.4568	0.6116	2.096	

MAR 2

Fall	Min.	1. Qu.	Median	Mean	3. Qu.	Max.	NA
Goldstandard	1.525	1.763	1.877	1.931	2.049	3.827	
-NA	1.932	2.183	2.364	2.443	2.593	4.491	
Srgt – lernMV	2.066	2.343	2.485	2.505	2.646	3.992	9
Srgt – testMV	1.950	2.167	2.271	2.295	2.397	4.031	
Srgt – lerntestMV	2.383	2.720	2.881	2.878	2.999	3.783	7
Imp – lernMV	1.606	1.837	1.948	1.981	2.068	4.334	
Imp – testMV	1.986	2.125	2.209	2.277	2.313	4.462	
Imp – lerntestMV	1.938	2.189	2.273	2.284	2.372	2.688	

Differenzen zum Goldstandard

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
-NA	-1.774	0.2490	0.4615	0.5122	0.7166	2.742	
Srgt – lernMV	-1.212	0.3875	0.6011	0.5778	0.8156	1.961	9
Srgt – testMV	-1.777	0.2015	0.3696	0.3642	0.5875	1.841	
Srgt – lerntestMV	-1.198	0.8190	0.9720	0.9618	1.165	2.007	7
Imp – lernMV	-1.762	-0.1148	0.0399	0.0502	0.2130	2.680	
Imp – testMV	-1.374	0.1606	0.3250	0.3467	0.4857	2.736	
Imp – lerntestMV	-1.780	0.2582	0.3788	0.3536	0.5095	0.9729	

MAR 3

Fall	Min.	1. Qu.	Median	Mean	3. Qu.	Max.	NA
Goldstandard	1.525	1.763	1.877	1.931	2.049	3.827	
-NA	1.770	2.217	2.435	2.589	2.761	4.659	
Srgt – lernMV	2.067	2.391	2.502	2.497	2.609	2.914	94
Srgt – testMV	1.661	1.902	1.995	2.053	2.133	3.768	
Srgt – lerntestMV	2.132	2.285	2.325	2.337	2.360	2.595	94
Imp – lernMV	1.795	1.991	2.129	2.158	2.251	3.822	
Imp – testMV	2.532	3.104	3.388	3.448	3.661	5.727	
Imp – lerntestMV	2.231	2.708	3.065	3.175	3.443	6.662	

Differenzen zum Goldstandard

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
-NA	-1.595	0.3182	0.5288	0.6587	0.9137	2.730	
Srgt – lernMV	0.2066	0.6480	0.7668	0.7626	0.9213	1.254	94
Srgt – testMV	-1.967	-0.0313	0.1107	0.1220	0.2803	1.878	
Srgt – lerntestMV	0.3135	0.4164	0.4601	0.5111	0.5395	0.8652	94
Imp – lernMV	-1.753	0.0379	0.2426	0.2272	0.4254	1.847	
Imp – testMV	-0.2693	1.224	1.437	1.517	1.789	3.799	
Imp – lerntestMV	-1.066	0.8197	1.080	1.244	1.538	4.821	

MAR 4

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
Goldstandard	1.525	1.763	1.877	1.931	2.049	3.827	
-NA	1.954	2.282	2.430	2.461	2.626	3.148	
Srgt – lernMV	1.652	2.034	2.173	2.216	2.296	7.285	3
Srgt – testMV	1.707	2.179	2.520	2.627	3.010	4.843	
Srgt – lerntestMV	1.784	2.137	2.223	2.319	2.380	4.224	5
Imp – lernMV	1.523	1.893	2.025	2.043	2.154	3.882	
Imp – testMV	1.744	1.993	2.110	2.135	2.235	3.866	
Imp – lerntestMV	1.738	1.940	2.028	2.034	2.120	2.474	

Differenzen zum Goldstandard

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
-NA	-1.654	0.3352	0.5649	0.5307	0.7552	1.407	
Srgt – lernMV	-1.820	0.0768	0.2567	0.2835	0.4873	5.196	3
Srgt – testMV	-2.019	0.1799	0.6211	0.6959	1.064	3.068	
Srgt – lerntestMV	-1.690	0.1503	0.3551	0.3909	0.5696	2.497	5
Imp – lernMV	-1.779	-0.0185	0.1541	0.1126	0.2972	2.272	
Imp – testMV	-1.615	0.0246	0.2148	0.2042	0.3993	1.767	
Imp – lerntestMV	-1.791	-0.0251	0.1662	0.1035	0.2859	0.6669	

MCAR

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
Goldstandard	1.525	1.763	1.877	1.931	2.049	3.827	
-NA	1.957	2.300	2.478	2.508	2.658	3.511	
Srgt – lernMV	2.031	2.310	2.433	2.502	2.612	4.352	
Srgt – testMV	2.056	2.247	2.373	2.426	2.511	4.546	
Srgt – lerntestMV	2.207	2.667	2.838	2.850	3.034	3.379	
Imp – lernMV	1.550	1.816	1.907	1.946	2.053	3.651	
Imp – testMV	1.909	2.122	2.189	2.217	2.274	4.240	
Imp – lerntestMV	1.948	2.166	2.292	2.308	2.402	3.932	

Differenzen zum Goldstandard

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
-NA	-1.350	0.3142	0.5868	0.5770	0.8076	1.828	
Srgt – lernMV	-1.380	0.3200	0.5848	0.5715	0.7723	2.377	
Srgt – testMV	-1.767	0.3096	0.4972	0.4951	0.6176	2.82	
Srgt – lerntestMV	-0.9089	0.6949	0.9808	0.9194	1.182	1.684	
Imp – lernMV	-1.926	-0.1528	0.0164	0.0154	0.2121	1.96	
Imp – testMV	-1.849	0.1299	0.3097	0.2860	0.4401	2.664	
Imp – lerntestMV	-1.731	0.2007	0.3629	0.3773	0.6093	2.157	

F.2.2. Blockweise hohe Korrelationen

MAR 1

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
Goldstandard	2.828	3.378	3.559	3.599	3.815	5.016	
-NA	3.732	4.445	4.824	4.977	5.351	9.795	
Srgt – lernMV	3.239	3.865	4.071	4.154	4.379	5.598	1
Srgt – testMV	3.391	3.718	3.918	4.031	4.183	5.914	
Srgt – lerntestMV	3.884	4.263	4.467	4.564	4.823	5.713	1
Imp – lernMV	3.064	3.642	3.794	3.899	4.036	5.428	
Imp – testMV	3.635	3.959	4.185	4.238	4.421	5.799	
Imp – lerntestMV	3.685	4.072	4.255	4.334	4.536	6.274	

Differenzen zum Goldstandard

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
-NA	-0.3212	0.8298	1.209	1.378	1.843	6.398	
Srgt – lernMV	-0.5622	0.1938	0.3901	0.5496	0.8634	2.207	1
Srgt – testMV	-1.521	0.0560	0.3600	0.4319	0.6744	2.419	
Srgt – lerntestMV	-0.8644	0.5990	0.8934	0.9742	1.427	2.052	1
Imp – lernMV	-1.122	-0.0379	0.2566	0.3001	0.6079	1.998	
Imp – testMV	-0.5170	0.3085	0.6024	0.6392	0.9469	2.246	
Imp – lerntestMV	-0.3270	0.3506	0.6135	0.7358	1.109	2.950	

MAR 2

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
Goldstandard	2.828	3.378	3.559	3.599	3.815	5.016	
-NA	3.882	4.536	5.093	5.148	5.608	6.991	
Srgt – lernMV	3.300	3.859	4.143	4.243	4.518	6.066	6
Srgt – testMV	3.364	3.752	3.927	3.989	4.197	4.940	
Srgt – lerntestMV	3.765	4.327	4.518	4.585	4.790	6.354	4
Imp – lernMV	3.110	3.564	3.780	3.814	4.072	5.352	
Imp – testMV	3.312	3.893	4.079	4.089	4.284	5.881	
Imp – lerntestMV	3.443	3.983	4.159	4.226	4.379	5.514	

Differenzen zum Goldstandard

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
-NA	-0.3170	1.021	1.476	1.550	2.081	3.399	
Srgt – lernMV	-0.5058	0.2001	0.5299	0.6389	1.110	2.797	6
Srgt – testMV	-1.403	0.0762	0.3389	0.3905	0.6883	1.783	
Srgt – lerntestMV	-0.1780	0.6957	0.9811	0.9925	1.167	2.951	4
Imp – lernMV	-1.311	-0.1070	0.1688	0.2153	0.5561	1.692	
Imp – testMV	-1.065	0.1843	0.5121	0.4905	0.8310	2.334	
Imp – lerntestMV	-0.8300	0.2646	0.5511	0.6277	0.9596	2.179	

MAR 3

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
Goldstandard	2.828	3.378	3.559	3.599	3.815	5.016	
-NA	4.744	5.751	6.312	6.388	6.856	9.243	
Srgt – lernMV	3.605	4.097	4.216	4.257	4.522	5.307	83
Srgt – testMV	3.034	3.481	3.689	3.740	3.947	4.919	
Srgt – lerntestMV	3.443	3.812	4.152	4.158	4.380	5.473	77
Imp – lernMV	3.388	3.899	4.041	4.111	4.278	5.763	
Imp – testMV	3.570	4.349	4.587	4.658	4.988	6.528	
Imp – lerntestMV	3.344	3.893	4.182	4.227	4.503	5.769	

Differenzen zum Goldstandard

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
-NA	0.7351	2.097	2.764	2.789	3.328	6.014	
Srgt – lernMV	-0.8367	0.2518	0.8317	0.6782	1.108	1.850	83
Srgt – testMV	-1.428	-0.1705	0.1652	0.1417	0.4599	1.547	
Srgt – lerntestMV	-0.4657	0.0846	0.4368	0.5108	0.8824	1.978	77
Imp – lernMV	-1.094	0.1775	0.4526	0.5124	0.8831	2.623	
Imp – testMV	-0.6658	0.7228	0.9914	1.059	1.405	3.062	
Imp – lerntestMV	-0.4787	0.2192	0.5760	0.6285	0.9336	2.511	

MAR 4

Fall	Min.	1. Qu.	Median	Mean	3. Qu.	Max.	NA
Goldstandard	2.828	3.378	3.559	3.599	3.815	5.016	
-NA	4.699	5.260	5.777	5.818	6.257	8.121	
Srgt – lernMV	3.287	3.922	4.201	4.213	4.442	5.316	5
Srgt – testMV	2.931	3.395	3.589	3.642	3.857	5.244	
Srgt – lerntestMV	3.496	3.884	4.204	4.329	4.614	5.830	6
Imp – lernMV	3.371	3.770	4.008	4.025	4.215	5.577	
Imp – testMV	3.565	3.968	4.184	4.244	4.434	6.277	
Imp – lerntestMV	3.290	3.883	4.113	4.119	4.337	5.556	

Differenzen zum Goldstandard

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
-NA	0.3136	1.675	2.027	2.219	2.852	4.720	
Srgt – lernMV	-1.252	0.2929	0.6381	0.6100	0.9772	2.177	5
Srgt – testMV	-1.782	-0.2526	0.0139	0.0429	0.3718	1.665	
Srgt – lerntestMV	-0.9017	0.1953	0.6300	0.7289	1.135	2.374	6
Imp – lernMV	-0.7195	0.1589	0.3843	0.4267	0.6393	2.044	
Imp – testMV	-1.091	0.2777	0.6630	0.6454	0.9610	2.433	
Imp – lerntestMV	-0.7449	0.1550	0.5886	0.5204	0.8424	1.655	

MCAR

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
Goldstandard	2.828	3.378	3.559	3.599	3.815	5.016	
-NA	3.668	4.528	4.868	4.895	5.376	6.103	
Srgt – lernMV	3.462	3.894	4.086	4.222	4.538	5.625	3
Srgt – testMV	3.313	3.758	3.980	3.999	4.161	5.378	
Srgt – lerntestMV	3.937	4.414	4.744	4.780	5.058	6.298	1
Imp – lernMV	2.942	3.461	3.719	3.781	3.970	6.054	
Imp – testMV	3.537	3.946	4.106	4.202	4.394	6.243	
Imp – lerntestMV	3.643	4.059	4.236	4.307	4.476	5.789	

Differenzen zum Goldstandard

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
-NA	-0.1461	0.8162	1.309	1.297	1.814	2.625	
Srgt – lernMV	-1.254	0.1788	0.5794	0.6210	1.025	2.285	3
Srgt – testMV	-1.529	0.1461	0.3518	0.4007	0.6497	1.874	
Srgt – lerntestMV	-0.7082	0.7079	1.157	1.183	1.580	2.973	1
Imp – lernMV	-0.9344	-0.0863	0.2090	0.1823	0.4983	2.453	
Imp – testMV	-0.8713	0.2297	0.5937	0.6034	0.9309	2.616	
Imp – lerntestMV	-0.7983	0.3941	0.6959	0.7082	1.069	1.905	

F.2.3. Gleichmäßige, niedrige Korrelationen

MAR 1

Fall	Min.	1. Qu.	Median	Mean	3. Qu.	Max.	NA
Goldstandard	6.032	7.158	7.607	7.634	8.050	9.792	
-NA	7.702	8.697	9.439	9.479	10.18	11.67	
Srgt – lernMV	7.682	8.883	9.358	9.471	9.872	12.26	
Srgt – testMV	8.913	9.939	10.46	10.47	10.85	12.56	
Srgt – lerntestMV	9.828	11.23	11.66	11.73	12.13	14.39	1
Imp – lernMV	6.692	7.880	8.283	8.473	9.116	11.05	
Imp – testMV	9.312	10.12	10.45	10.45	10.81	13.34	
Imp – lerntestMV	9.746	10.56	11.04	11.09	11.38	13.72	

Differenzen zum Goldstandard

Fall	Min.	1. Qu.	Median	Mean	3. Qu.	Max.	NA
-NA	-0.7985	1.160	1.808	1.845	2.638	4.647	
Srgt – lernMV	-1.184	0.9458	1.866	1.837	2.593	5.188	
Srgt – testMV	-0.4710	2.284	2.968	2.839	3.455	5.246	
Srgt – lerntestMV	1.436	3.318	4.004	4.101	4.955	7.013	1
Imp – lernMV	-2.133	0.0011	0.7863	0.8392	1.632	4.319	
Imp – testMV	0.8191	2.204	2.892	2.820	3.451	5.261	
Imp – lerntestMV	1.212	2.727	3.515	3.452	4.077	5.921	

MAR 2

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
Goldstandard	6.032	7.158	7.607	7.634	8.050	9.792	
-NA	7.425	9.169	9.737	9.831	10.34	12.01	
Srgt – lernMV	7.405	8.820	9.343	9.464	10.03	12.38	1
Srgt – testMV	9.233	10.10	10.45	10.54	11.01	13.31	
Srgt – lerntestMV	9.979	11.06	11.54	11.62	12.07	13.97	
Imp – lernMV	6.399	7.746	8.475	8.480	9.056	10.54	
Imp – testMV	9.223	10.17	10.46	10.60	10.93	12.82	
Imp – lerntestMV	9.334	10.65	11.11	11.18	11.75	13.33	

Differenzen zum Goldstandard

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
-NA	-0.4098	1.396	2.204	2.197	2.836	5.166	
Srgt – lernMV	-0.9249	1.046	1.731	1.829	2.652	4.715	1
Srgt – testMV	0.6074	2.376	2.923	2.910	3.462	6.751	
Srgt – lerntestMV	1.760	3.358	3.895	3.985	4.685	7.076	
Imp – lernMV	-2.235	0.0072	0.8554	0.8457	1.677	3.860	
Imp – testMV	0.0679	2.353	2.918	2.968	3.616	5.663	
Imp – lerntestMV	1.632	2.859	3.529	3.546	4.219	6.307	

MAR 3

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
Goldstandard	6.032	7.158	7.607	7.634	8.050	9.792	
-NA	7.674	9.053	9.739	9.697	10.36	11.99	
Srgt – lernMV	7.615	8.869	9.264	9.359	10.07	11.38	76
Srgt – testMV	8.446	9.549	9.970	10.11	10.49	13.15	
Srgt – lerntestMV	9.421	9.991	10.23	10.50	10.86	12.77	84
Imp – lernMV	6.363	7.954	8.484	8.601	9.138	12.13	
Imp – testMV	8.931	10.13	10.44	10.55	11.01	12.08	
Imp – lerntestMV	9.220	10.05	10.47	10.55	11.09	12.23	

Differenzen zum Goldstandard

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
-NA	-0.9034	1.544	2.092	2.063	2.712	5.096	
Srgt – lernMV	-0.0303	1.074	1.914	1.773	2.242	3.624	76
Srgt – testMV	0.3224	1.920	2.476	2.480	2.999	5.897	
Srgt – lerntestMV	1.540	2.045	2.877	2.854	3.189	5.722	84
Imp – lernMV	-1.426	0.1342	0.8046	0.9670	1.733	5.946	
Imp – testMV	-0.8618	2.285	2.942	2.920	3.476	5.385	
Imp – lerntestMV	0.8061	2.537	2.888	2.916	3.470	5.581	

MAR 4

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
Goldstandard	6.032	7.158	7.607	7.634	8.050	9.792	
-NA	7.846	9.516	10.08	10.16	10.72	13.30	
Srgt – lernMV	7.302	8.174	8.542	8.642	9.016	11.08	1
Srgt – testMV	7.611	8.756	9.358	9.451	9.953	12.84	
Srgt – lerntestMV	8.447	9.640	10.09	10.18	10.77	12.22	2
Imp – lernMV	6.720	7.922	8.501	8.620	9.204	11.75	
Imp – testMV	8.784	9.824	10.32	10.33	10.79	12.34	
Imp – lerntestMV	8.923	9.73	10.20	10.21	10.59	13.71	

Differenzen zum Goldstandard

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
-NA	-0.3066	1.731	2.537	2.528	3.193	6.209	
Srgt – lernMV	-1.228	0.3568	0.9951	1.021	1.664	4.075	1
Srgt – testMV	-0.9406	1.022	1.641	1.817	2.594	5.043	
Srgt – lerntestMV	0.1498	1.877	2.487	2.529	3.265	4.771	2
Imp – lernMV	-1.962	0.2400	1.001	0.9860	1.657	5.154	
Imp – testMV	0.0998	1.998	2.779	2.700	3.267	5.142	
Imp – lerntestMV	-0.1663	1.960	2.536	2.575	3.231	6.100	

MCAR

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
Goldstandard	6.032	7.158	7.607	7.634	8.05	9.792	
-NA	7.815	9.106	9.711	9.757	10.32	12.97	
Srgt – lernMV	7.035	8.802	9.339	9.447	10.01	12.27	
Srgt – testMV	8.96	9.915	10.42	10.46	10.99	12.29	
Srgt – lerntestMV	10.09	10.99	11.42	11.62	12.29	14.48	1
Imp – lernMV	7.066	7.959	8.420	8.525	9.015	10.99	
Imp – testMV	8.907	9.869	10.17	10.32	10.67	14.72	
Imp – lerntestMV	9.064	10.28	10.78	10.92	11.43	13.02	

Differenzen zum Goldstandard

Fall	Min.	1. Qu.	Median	Mittelw.	3. Qu.	Max.	NA
-NA	-0.7079	1.363	2.051	2.122	2.815	4.675	
Srgt – lernMV	-0.8043	1.025	1.721	1.812	2.639	4.567	
Srgt – testMV	0.1701	2.104	2.759	2.825	3.489	5.356	
Srgt – lerntestMV	1.265	3.290	3.933	3.989	4.750	6.315	1
Imp – lernMV	-2.226	0.2053	0.9134	0.8910	1.666	3.288	
Imp – testMV	-0.2309	1.965	2.563	2.687	3.389	8.688	
Imp – lerntestMV	1.162	2.620	3.169	3.287	3.802	6.454	

G. Tabellen zu den nicht sichtbaren Datenpunkten

G.1. Simulation zur Klassifikation

Sämtliche Tabellen sind auf vier Nachkommastellen gerundet.

G.1.1. Gleichmäßige, hohe Korrelationen

MAR 1

Differenzen zum Goldstandard

Abteilung	Fall	Datenwert
Benchmark	-NA	-0.1561

MAR 2

Differenzen zum Goldstandard

Abteilung	Fall	Datenwert
Benchmark	-NA	-0.1610
		-0.1855
Surrogatv.	testMV	-0.2061
Surrogatv.	lerntestMV	-0.1842

MAR 3

Abteilung	Fall	Datenwert
Imputation	testMV	-0.5731
		-0.8878

Differenzen zum Goldstandard

Abteilung	Fall	Datenwert
Benchmark	-NA	0.2040
Surrogatv.	lernMV	0.2010
Imputation	testMV	-0.1506

Abteilung	Fall	Datenwert
Imputation	testMV	-0.1604
		-0.1612
		-0.1628
		-0.1753
		-0.1852
		-0.1904
		-0.2186
		-0.2230
		-0.2260
		-0.2299
		-0.2306
		-0.2406
		-0.2438
		-0.2465
		-0.2490
		-0.2534
		-0.2539
		-0.2594
		-0.2615
		-0.2625
		-0.2846
		-0.2886
		-0.2914
		-0.3098
		-0.3133
		-0.3186
		-0.3229
		-0.4529
		-0.7060
		-0.2079

MAR 4

Differenzen zum Goldstandard

Abteilung	Fall	Datenwert
Surrogatv.	lernMV	0.2083
Surrogatv.	testMV	0.2011

MCAR

Differenzen zum Goldstandard

Abteilung	Fall	Datenwert
Benchmark	-NA	0.2008

G.1.2. Blockweise hohe Korrelationen

MAR 1

Differenzen zum Goldstandard

Abteilung	Fall	Datenwert
Benchmark	-NA	0.2116
Surrogatv.	lerntestMV	-0.1713
Imputation	lernMV	0.2028
Imputation	testMV	0.2022

MAR 2

Differenzen zum Goldstandard

Abteilung	Fall	Datenwert
Benchmark	-NA	-0.1548
		-0.1669
		-0.1707
		-0.1865
Surrogatv.	lernMV	0.2303
		-0.1550
Surrogatv.	testMV	-0.1572
Surrogatv.	lerntestMV	0.2106
Imputation	lernMV	-0.2536

MAR 3

Abteilung	Fall	Datenwert
Benchmark	-NA	-0.5123
		-0.5131
Imputation	testMV	-0.5550

Differenzen zum Goldstandard

Abteilung	Fall	Datenwert
Benchmark	-NA	-0.1538
		-0.1557
		-0.1698

Abteilung	Fall	Datenwert
Benchmark	-NA	-0.1725
		-0.1741
		-0.1750
		-0.1789
		-0.1955
		-0.2021
		-0.2134
		-0.2385
		-0.2735
		-0.2949
Surrogatv.	lernMV	0.2592
		-0.1941
Surrogatv.	testMV	0.2088
Surrogatv.	lerntestMV	-0.1613
		-0.1738
Imputation	testMV	0.2088
		-0.1517
		-0.1526
		-0.1713
		-0.1730
		-0.1745
		-0.1757
		-0.1819
		-0.1873
		-0.1921
		-0.2328
		-0.3129
Imputation	lernMV	0.2113
		-0.1655

MAR 4

Differenzen zum Goldstandard

Abteilung	Fall	Datenwert
Benchmark	-NA	0.2001
		-0.1510
		-0.1535
		-0.1616
		-0.1677
		-0.1940
		-0.2189
Surrogatv.	testMV	-0.1632

Abteilung	Fall	Datenwert
Imputation	lernMV	0.2440
Imputation	testMV	0.2188
		-0.2574
Imputation	lerntestMV	0.2447
		-0.1588
		-0.1694

MCAR

Differenzen zum Goldstandard

Abteilung	Fall	Datenwert
Benchmark	-NA	-0.1560
		-0.2460
Surrogatv.	lernMV	-0.2375
Surrogatv.	testMV	0.2405
Surrogatv.	lerntestMV	0.2254
		-0.1518
		-0.1608
Imputation	lernMV	0.2279
Imputation	testMV	0.2327
		-0.1529

G.1.3. Gleichmäßige, niedrige Korrelationen

MAR 1

Abteilung	Fall	Datenwert
Benchmark	-NA	-0.5106
Surrogatv.	lernMV	-0.5135
		-0.5395
Imputation	testMV	-0.5085

Differenzen zum Goldstandard

Abteilung	Fall	Datenwert
Benchmark	-NA	-0.1664
		-0.1745
Surrogatv.	lernMV	-0.1582
Surrogatv.	testMV	-0.1587
Imputation	lernMV	-0.1616
Imputation	testMV	-0.1883
		-0.1887

MAR 2

Abteilung	Fall	Datenwert
Benchmark	-NA	-0.5709
		-0.5714
Imputation	testMV	-0.5092
		-0.5702
Imputation	lerntestMV	-0.5793

Differenzen zum Goldstandard

Abteilung	Fall	Datenwert
Benchmark	-NA	-0.1579
		-0.1674
		-0.2289
		-0.2395
Surrogatv.	lernMV	-0.1525
Surrogatv.	testMV	-0.1501
Imputation	testMV	-0.1684
		-0.1887
		-0.2047
		-0.2097
Imputation	lerntestMV	-0.1667
		-0.2767

MAR 3

Abteilung	Fall	Datenwert
Benchmark	-NA	-0.5000
		-0.5039
		-0.5054
		-0.5055
		-0.5069
		-0.5103
		-0.5213
		-0.5269
		-0.5279
		-0.5312
		-0.5328
		-0.5330
		-0.5352
		-0.5804
		-0.5845
		-0.6011
		-0.6837

Abteilung	Fall	Datenwert
Benchmark	-NA	-0.7070
		-0.7785
		-0.8132
Surrogatv.	testMV	-0.5520

Differenzen zum Goldstandard

Abteilung	Fall	Datenwert
Benchmark	-NA	-0.1533
		-0.1578
		-0.1634
		-0.1660
		-0.1699
		-0.1702
		-0.1739
		-0.1740
		-0.1743
		-0.1761
		-0.1781
		-0.1798
		-0.1814
		-0.1826
		-0.1838
		-0.2010
		-0.2095
		-0.2143
		-0.2238
		-0.2311
		-0.3080
		-0.3689
		-0.3696
		-0.4521
		-0.4584
Surrogatv.	testMV	-0.1522
		-0.2284
Imputation	lerntestMV	-0.1583

MAR 4

Abteilung	Fall	Datenwert
Benchmark	-NA	-0.5359
		-0.5361
		-0.5994
Imputation	lernMV	-0.5322
		-0.6658

Differenzen zum Goldstandard

Abteilung	Fall	Datenwert
Benchmark	-NA	-0.1696
		-0.1749
		-0.2046
		-0.2156
		-0.2760
Imputation	lernMV	-0.1838
		-0.3410

MCAR

Abteilung	Fall	Datenwert
Surrogatv.	lerntestMV	-0.5166
Imputation	testMV	-0.5051
		-0.5445
		-0.5583

Differenzen zum Goldstandard

Abteilung	Fall	Datenwert
Benchmark	-NA	-0.1512
		-0.1587
Surrogatv.	lerntestMV	-0.1579
		-0.1883
		-0.2125
Imputation	testMV	-0.1528
		-0.1597
		-0.1660
		-0.1685
		-0.2163
		-0.2514

G.2. Simulation zur Regression

Sämtliche Tabellen sind auf vier gültige Ziffern gerundet.

G.2.1. Gleichmäßige, hohe Korrelationen

MAR 4

Differenzen zum Goldstandard

Abteilung	Fall	Datenwert
Surrogatv.	testMV	-2.019

G.2.2. Blockweise hohe Korrelationen

MAR 1

Differenzen zum Goldstandard

Abteilung	Fall	Datenwert
Benchmark	-NA	6.398

MAR 3

Differenzen zum Goldstandard

Abteilung	Fall	Datenwert
Benchmark	-NA	6.014

G.2.3. Gleichmäßige, niedrige Korrelationen

MAR 1

Abteilung	Fall	Datenwert
Surrogatv.	lerntestMV	14.05
		14.35
		14.39

Differenzen zum Goldstandard

Abteilung	Fall	Datenwert
Surrogatv.	lerntestMV	6.299
		6.341
		6.504
		6.756
		6.957
		7.013
Imputation	lernMV	-2.133

MAR 2

Differenzen zum Goldstandard

Abteilung	Fall	Datenwert
Surrogatv.	testMV	6.751
Surrogatv.	lerntestMV	6.102
		6.153
		7.076
Imputation	lernMV	-2.235
Imputation	lerntestMV	6.307

MAR 4

Differenzen zum Goldstandard

Abteilung	Fall	Datenwert
Benchmark	-NA	6.209
Imputation	lerntestMV	6.100

MCAR

Abteilung	Fall	Datenwert
Surrogatv.	lerntestMV	14.48
Imputation	testMV	14.72

Differenzen zum Goldstandard

Abteilung	Fall	Datenwert
Surrogatv.	lerntestMV	6.126
		6.308
		6.315
Surrogatv.	lernMV	-2.226
Imputation	testMV	8.688
Imputation	lerntestMV	6.454

H. Tabellen über die p-Werte der t-Tests

Sämtliche Tabellen sind auf vier Nachkommastellen bzw. vier gültige Ziffern gerundet – je nachdem, was kürzer ist.

Die beiden Mittelwerte unterscheiden sich signifikant, falls $p < 0.05$ ist.

H.1. Simulation zur Klassifikation

H.1.1. Gleichmäßige, hohe Korrelationen

MAR 1

Fall	lernMV	testMV	lerntestMV
Imputation – Surrogatv.	0.0001	0.3670	0.0016
Surrogatv. – Goldstandard	0.0027	0.0069	0.0000
Surrogatv. – -NA	0.5095	0.8698	0.0154
Imputation – Goldstandard	0.7505	0.0639	0.0190
Imputation – -NA	0.0037	0.5306	0.7784

MAR 2

Fall	lernMV	testMV	lerntestMV
Imputation – Surrogatv.	0.0000	0.3377	0.0000
Surrogatv. – Goldstandard	0.0080	0.0812	0.0000
Surrogatv. – -NA	0.7341	0.6038	0.0012
Imputation – Goldstandard	0.1455	0.0089	0.4068
Imputation – -NA	0.0001	0.6596	0.0820

MAR 3

Fall	lernMV	testMV	lerntestMV
Imputation – Surrogatv.	0.0001	0.0000	0.0000
Surrogatv. – Goldstandard	0.0616	0.7332	0.0716
Surrogatv. – -NA	0.0057	0.3640	0.0080
Imputation – Goldstandard	0.0000	0.0000	0.0000
Imputation – -NA	0.0000	0.0000	0.0000

MAR 4

Fall	lernMV	testMV	lerntestMV
Imputation – Surrogatv.	0.2826	0.0834	0.3736
Surrogatv. – Goldstandard	0.6137	0.2447	0.0889
Surrogatv. – -NA	0.8798	0.3697	0.1319
Imputation – Goldstandard	0.6875	0.0079	0.2991
Imputation – -NA	0.3622	0.0099	0.4580

MCAR

Fall	lernMV	testMV	lerntestMV
Imputation – Surrogatv.	0.0003	0.7761	0.6286
Surrogatv. – Goldstandard	0.0004	0.2529	0.0000
Surrogatv. – -NA	0.3370	0.0417	0.1143
Imputation – Goldstandard	0.6526	0.3550	0.0003
Imputation – -NA	0.0104	0.0238	0.2972

H.1.2. Blockweise hohe Korrelationen

MAR 1

Fall	lernMV	testMV	lerntestMV
Imputation – Surrogatv.	0.0060	0.9438	0.7846
Surrogatv. – Goldstandard	0.0522	0.0600	0.0083
Surrogatv. – -NA	0.0024	0.0011	0.0297
Imputation – Goldstandard	0.5929	0.0756	0.0207
Imputation – -NA	0.0000	0.0012	0.0165

MAR 2

Fall	lernMV	testMV	lerntestMV
Imputation – Surrogatv.	0.0798	0.8041	0.5947
Surrogatv. – Goldstandard	0.0832	0.2884	0.0971
Surrogatv. – -NA	0.0000	0.0000	0.0000
Imputation – Goldstandard	0.8674	0.1761	0.0371
Imputation – -NA	0.0000	0.0000	0.0000

MAR 3

Fall	lernMV	testMV	lerntestMV
Imputation – Surrogatv.	0.9634	0.0000	0.8550
Surrogatv. – Goldstandard	0.1100	0.5751	0.0426
Surrogatv. – -NA	0.0000	0.0000	0.0000
Imputation – Goldstandard	0.0608	0.0000	0.0328
Imputation – -NA	0.0000	0.0116	0.0000

MAR 4

Fall	lernMV	testMV	lerntestMV
Imputation – Surrogatv.	0.6595	0.0003	0.0207
Surrogatv. – Goldstandard	0.7513	0.5721	0.1014
Surrogatv. – -NA	0.0000	0.0000	0.0000
Imputation – Goldstandard	0.9208	0.0001	0.0002
Imputation – -NA	0.0000	0.0391	0.0101

MCAR

Fall	lernMV	testMV	lerntestMV
Imputation – Surrogatv.	0.0000	0.2314	0.0005
Surrogatv. – Goldstandard	0.0018	0.0197	0.0000
Surrogatv. – -NA	0.3707	0.0488	0.9412
Imputation – Goldstandard	0.1557	0.2482	0.2496
Imputation – -NA	0.0000	0.0019	0.0005

H.1.3. Gleichmäßige, niedrige Korrelationen

MAR 1

Fall	lernMV	testMV	lerntestMV
Imputation – Surrogatv.	0.1419	0.0000	0.0227
Surrogatv. – Goldstandard	0.0023	0.0000	0.0000
Surrogatv. – -NA	0.0000	0.1551	0.0238
Imputation – Goldstandard	0.1747	0.0000	0.0000
Imputation – -NA	0.0000	0.0057	0.0001

MAR 2

Fall	lernMV	testMV	lerntestMV
Imputation – Surrogatv.	0.1376	0.0000	0.1355
Surrogatv. – Goldstandard	0.0067	0.0000	0.0000
Surrogatv. – -NA	0.0000	0.0194	0.2500
Imputation – Goldstandard	0.1622	0.0000	0.0000
Imputation – -NA	0.0000	0.0322	0.9103

MAR 3

Fall	lernMV	testMV	lerntestMV
Imputation – Surrogatv.	0.6925	0.8395	0.3634
Surrogatv. – Goldstandard	0.5831	0.0000	0.0006
Surrogatv. – -NA	0.0000	0.0000	0.0000
Imputation – Goldstandard	0.8581	0.0000	0.0003
Imputation – -NA	0.0000	0.0000	0.0000

MAR 4

Fall	lernMV	testMV	lerntestMV
Imputation – Surrogatv.	0.5281	0.0249	0.0038
Surrogatv. – Goldstandard	0.8639	0.0000	0.0020
Surrogatv. – -NA	0.0000	0.0001	0.0000
Imputation – Goldstandard	0.4407	0.0233	0.9391
Imputation – -NA	0.0000	0.0000	0.0000

MCAR

Fall	lernMV	testMV	lerntestMV
Imputation – Surrogatv.	0.0199	0.0000	0.1079
Surrogatv. – Goldstandard	0.0001	0.0000	0.0000
Surrogatv. – -NA	0.0081	0.7626	0.0004
Imputation – Goldstandard	0.0985	0.0000	0.0000
Imputation – -NA	0.0000	0.0000	0.0261

H.2. Simulation zur Regression

H.2.1. Gleichmäßige, hohe Korrelationen

MAR 1

Fall	lernMV	testMV	lerntestMV
Imputation – Surrogatv.	0.0000	0.0858	0.0000
Surrogatv. – Goldstandard	0.0000	0.0000	0.0000
Surrogatv. – -NA	0.0282	0.4985	0.0000
Imputation – Goldstandard	0.1854	0.0000	0.0000
Imputation – -NA	0.0000	0.0207	0.1318

MAR 2

Fall	lernMV	testMV	lerntestMV
Imputation – Surrogatv.	0.0000	0.6779	0.0000
Surrogatv. – Goldstandard	0.0000	0.0000	0.0000
Surrogatv. – -NA	0.1953	0.0016	0.0000
Imputation – Goldstandard	0.2817	0.0000	0.0000
Imputation – -NA	0.0000	0.0018	0.0002

MAR 3

Fall	lernMV	testMV	lerntestMV
Imputation – Surrogatv.	0.0308	0.0000	0.0000
Surrogatv. – Goldstandard	0.0035	0.0072	0.0004
Surrogatv. – -NA	0.4953	0.0000	0.0083
Imputation – Goldstandard	0.0000	0.0000	0.0000
Imputation – -NA	0.0000	0.0000	0.0000

MAR 4

Fall	lernMV	testMV	lerntestMV
Imputation – Surrogatv.	0.0062	0.0000	0.0000
Surrogatv. – Goldstandard	0.0000	0.0000	0.0000
Surrogatv. – -NA	0.0001	0.0137	0.0039
Imputation – Goldstandard	0.0062	0.0000	0.0033
Imputation – -NA	0.0000	0.0000	0.0000

MCAR

Fall	lernMV	testMV	lerntestMV
Imputation – Surrogatv.	0.0000	0.0000	0.0000
Surrogatv. – Goldstandard	0.0000	0.0000	0.0000
Surrogatv. – -NA	0.9083	0.0791	0.0000
Imputation – Goldstandard	0.7038	0.0000	0.0000
Imputation – -NA	0.0000	0.0000	0.0000

H.2.2. Blockweise hohe Korrelationen

MAR 1

Fall	lernMV	testMV	lerntestMV
Imputation – Surrogatv.	0.0001	0.0006	0.0001
Surrogatv. – Goldstandard	0.0000	0.0000	0.0000
Surrogatv. – -NA	0.0000	0.0000	0.0000
Imputation – Goldstandard	0.0000	0.0000	0.0000
Imputation – -NA	0.0000	0.0000	0.0000

MAR 2

Fall	lernMV	testMV	lerntestMV
Imputation – Surrogatv.	0.0000	0.0467	0.0000
Surrogatv. – Goldstandard	0.0000	0.0000	0.0000
Surrogatv. – -NA	0.0000	0.0000	0.0000
Imputation – Goldstandard	0.0001	0.0000	0.0000
Imputation – -NA	0.0000	0.0000	0.0000

MAR 3

Fall	lernMV	testMV	lerntestMV
Imputation – Surrogatv.	0.2004	0.0000	0.5386
Surrogatv. – Goldstandard	0.0000	0.0067	0.0000
Surrogatv. – -NA	0.0000	0.0000	0.0000
Imputation – Goldstandard	0.0000	0.0000	0.0000
Imputation – -NA	0.0000	0.0000	0.0000

MAR 4

Fall	lernMV	testMV	lerntestMV
Imputation – Surrogatv.	0.0005	0.0000	0.0017
Surrogatv. – Goldstandard	0.0000	0.3985	0.0000
Surrogatv. – -NA	0.0000	0.0000	0.0000
Imputation – Goldstandard	0.0000	0.0000	0.0000
Imputation – -NA	0.0000	0.0000	0.0000

MCAR

Fall	lernMV	testMV	lerntestMV
Imputation – Surrogatv.	0.0000	0.0002	0.0000
Surrogatv. – Goldstandard	0.0000	0.0000	0.0000
Surrogatv. – -NA	0.0000	0.0000	0.1411
Imputation – Goldstandard	0.0036	0.0000	0.0000
Imputation – -NA	0.0000	0.0000	0.0000

H.2.3. Gleichmäßige, niedrige Korrelationen

MAR 1

Fall	lernMV	testMV	lerntestMV
Imputation – Surrogatv.	0.0000	0.8439	0.0000
Surrogatv. – Goldstandard	0.0000	0.0000	0.0000
Surrogatv. – -NA	0.9529	0.0000	0.0000
Imputation – Goldstandard	0.0000	0.0000	0.0000
Imputation – -NA	0.0000	0.0000	0.0000

MAR 2

Fall	lernMV	testMV	lerntestMV
Imputation – Surrogatv.	0.0000	0.5606	0.0001
Surrogatv. – Goldstandard	0.0000	0.0000	0.0000
Surrogatv. – -NA	0.0073	0.0000	0.0000
Imputation – Goldstandard	0.0000	0.0000	0.0000
Imputation – -NA	0.0000	0.0000	0.0000

MAR 3

Fall	lernMV	testMV	lerntestMV
Imputation – Surrogatv.	0.0014	0.0000	0.8335
Surrogatv. – Goldstandard	0.0000	0.0000	0.0000
Surrogatv. – -NA	0.1280	0.0005	0.0020
Imputation – Goldstandard	0.0000	0.0000	0.0000
Imputation – -NA	0.0000	0.0000	0.0000

MAR 4

Fall	lernMV	testMV	lerntestMV
Imputation – Surrogatv.	0.8543	0.0000	0.7590
Surrogatv. – Goldstandard	0.0000	0.0000	0.0000
Surrogatv. – -NA	0.0000	0.0000	0.9120
Imputation – Goldstandard	0.0000	0.0000	0.0000
Imputation – -NA	0.0000	0.1686	0.7011

MCAR

Fall	lernMV	testMV	lerntestMV
Imputation – Surrogatv.	0.0000	0.1867	0.0000
Surrogatv. – Goldstandard	0.0000	0.0000	0.0000
Surrogatv. – -NA	0.0264	0.0000	0.0000
Imputation – Goldstandard	0.0000	0.0000	0.0000
Imputation – -NA	0.0000	0.0000	0.0000

Hiermit versichere ich, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel verfasst habe.

Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Gauting, den 30. Juli 2008

Anna Rieger