Ludwig-Maximilians-Universität München
Institut für Statistik

Der Einfluss von mtry auf Random Forests

Masterarbeit
zur Erlangung des akademischen Grades
Master of Science (M.Sc.)

Autorin:
Myriam Hatz

Betreuer:
Prof. Dr. Anne-Laure Boulesteix
Philipp Probst

München, 30. Januar 2018

Abstract

Die Korrelation der Entscheidungsbäume eines Random Forests wird unter anderem durch den Hyperparameter *mtry* beeinflusst. Dieser bestimmt die Anzahl an Variablen, die innerhalb eines Baumes als Splitkandidaten berücksichtigt werden. Oft finden für diesen Parameter bekannte Defaultwerte wie $mtry = \left\lfloor \sqrt{p} \right\rfloor$ für kategorialen Response und $mtry = \lfloor p/3 \rfloor$ für metrischen Response Anwendung. Nur wenige Untersuchungen beschäftigen sich über diese Defaultwerte hinaus mit dem Hyperparamter *mtry*. Es besteht allerdings die Vermutung, dass die genannten Defaultwerte nicht in jeder Datensituation die beste Wahl darstellen. Daher ist es Ziel dieser Arbeit, anhand einer Simulationsstudie den Einfluss von *mtry* auf Random Forests zu untersuchen. Die Analysen konzentrieren sich dabei vor allem auf die Anzahl an relevanten Kovariablen innerhalb eines Datensatzes und auf verschiedene Korrelationsstrukturen zwischen den Kovariablen. In Bezug darauf ist von Interesse, wie sich diese Datenstrukturen auf die Modellperformance eines Random Forests auswirken.

Die Ergebnisse dieser Simulationen zeigen, dass in Situationen mit sehr wenigen bzw. vielen ähnlich relevanten Kovariablen innerhalb eines Datensatzes, das optimale *mtry* von den Defaultwerten abweicht. Aber auch wenn einige der Kovariablen korreliert sind, konnte für manche Szenarien abhängig von der Stärke der Korrelation ein Einfluss auf das optimale *mtry* festgestellt werden. Bei Auftreten dieser speziellen Datenstrukturen ist es demnach ratsam, der Wahl von *mtry* besondere Aufmerksamkeit zu schenken, um einen Random Forest mit optimaler Modellperformance zu erhalten.

Abschließend konnte anhand zweier Anwendungsbeispiele gezeigt werden, wie die *mtry* Wahl durch die Messung der Relevanz von Kovariablen umgesetzt werden kann.

Inhaltsverzeichnis

Abbildungsverzeichnis

Abbildung Seite

Tabellenverzeichnis

1 Einleitung

Die von Breiman (2001) entwickelten Random Forests stellen eine beliebte nichtparametrische Klassifikations- bzw. Regressionsmethode dar, insbesondere da sie auch bei komplexen Interaktionen oder hochkorrelierten Kovariablen gute Prädiktionen liefern können. Ein Random Forest besteht aus vielen dekorrelierten Entscheidungsbäumen. Diese Eigenschaft wird unter anderem dadurch beeinflusst, dass für die einzelnen Splits eines Baumes nicht alle Variablen als Splitkandidaten berücksichtigt werden, sondern nur eine zufällige Auswahl an Variablen. Damit sinkt die paarweise Korrelation der Entscheidungsbäume. Um einen Random Forest mit optimaler Prädiktionsgüte zu erhalten, müssen verschiede Hyperparameter vorab sorgfältig vom Benutzer festgelegt werden. Einer der wichtigsten ist dabei die angesprochene Anzahl an Variablen, die als Splitkandidaten berücksichtigt werden, auch *mtry* genannt. (Hastie et al., 2009, S.587-588)

Eine Untersuchung von Bernard et al. (2009) bestätigte, dass der gebräuchliche Defaultwert $mtry = \left\lfloor \sqrt{p} \right\rfloor$ für die Klassifikation im Allgemeinen gute Ergebnisse liefert. Ebenso zeigte sich dabei allerdings auch, dass in Ausnahmefällen, zum Beispiel bei nur sehr wenigen relevanten Kovariablen, *mtry* deutlich höher als der empfohlene Defaultwert gewählt werden muss, um einen Random Forest mit bester Prädiktionsgüte zu erhalten. Daher ist es Ziel dieser Arbeit anhand einer umfangreichen Simulationsstudie den Einfluss des Hyperparamters *mtry* auf Random Forests näher zu untersuchen.
Mit den bereits bekannten Ergebnissen von Bernard et al. (2009) liegt es nahe, im ersten Schritt die Anzahl an relevanten Kovariablen zu variieren und dabei die optimalen *mtry* Werte für die verschiedenen Szenarien zu bestimmen. Als Erweiterung zu den Untersuchungen von Bernard et al. (2009) werden sowohl Klassifikations- als auch Regresssionsdatensätze betrachtet. Des Weiteren werden aber auch diverse Korrelationsstrukturen der Kovariablen definiert, da sich für einzelne Korrelationsstrukturen bereits ein beachtlicher Einfluss auf die Variablenwichtigkeiten eines Random Forest gezeigt hat (Gregorutti et al., 2016; Strobl et al., 2008), womit sich auch Auswirkungen auf das optimale *mtry* vermuten lassen.

Im Folgenden befasst sich Kapitel 2 näher mit der statistischen Methodik, die dieser Arbeit zugrunde liegt. Dazu zählt neben verschiedenen Modellgütemaßen und den sogenannten CART-Entscheidungsbäumen auch der im Fokus stehende Random Forest. Die Unterkapitel 2.3.2 und 2.3.3 geben dabei einen Überblick über die bisherige Forschung zu den angesprochenen Einflussfaktoren des Parameters *mtry*.
Das Design der Simulationsstudie, mit den entsprechenden Kovariableneinflüssen und Kovarianzstrukturen wird zusammen mit den Ergebnissen in Kapitel 3 dargestellt.
Ziel dieser Arbeit ist es außerdem eine Empfehlung zur *mtry* Wahl auszusprechen, die auf Basis der Datenstruktur bestimmt werden kann. Dafür beschreibt Kapitel 4 Möglichkeiten zur Messung der Korrelation und Relevanz von Kovariablen und überprüft, ob sich mit diesem Ansatz das optimale *mtry* zweier Beispieldatensätze bestimmen lässt.
Abschließend werden die Ergebnisse dieser Arbeit in Kapitel 5 zusammengefasst.

Alle Analysen in dieser Arbeit wurden mit der statistischen Software **R** durchgeführt (R Core Team, 2015, Version 3.2.3). Zur Erstellung der Abbildungen wurde dabei das **R**-Package `ggplot2` (Wickham, 2009, Version 2.2.1) genutzt.

2 Statistische Methodik

Das folgende Kapitel behandelt die Methodik, die dieser Arbeit zugrunde liegt. Für die Auswertungen sind unter anderem die vorgestellten Modellgütemaße von Bedeutung, welche auch die Performance eines Random Forests messen können. Neben Klassifikations- und Regressionsbäumen, welche die Basis eines Random Forest bilden, werden ebenso Eigenschaften eines Random Forests beschrieben, die unter anderem Grundlage für die durchgeführte Simulationsstudie waren.

2.1 Modellgütemaße

Modellgütemaße werden verwendet, um die Anpassung eines statistischen Modells an vorliegende Daten zu quantifizieren. Dabei liegt der Fokus zumeist auf der Prädiktionsfähigkeit des Modells. Abhängig vom Response und der gewünschten Struktur, die durch die Modellierung dargestellt werden soll, können eine Vielzahl verschiedener Maße betrachtet werden.
Im Weiteren werden vier Maße vorgestellt, die zwei unterschiedliche Strukturen für Regressions- und Klassifikationsmethoden abbilden. Die Bestimmung der Modellgüte kann damit zum einen basierend auf den Residuen und zum anderen basierend auf den Rängen der Beobachtungen erfolgen. Rosset et al. (2006) vergleichen diese beiden Ansätze für Regressionsmethoden. Residuenbasierte Maße haben demnach den Vorteil, dass sie eine Likelihood-Interpretation ermöglichen und oft die „wahren“ Kosten des Prädiktionsfehlers darstellen. Allerdings können in manchen Situationen auch rangbasierte Maße gewisse Vorteile mit sich bringen, beispielsweise, wenn anstelle der konkret geschätzten Werte eines Modells eher im Vordergrund steht, ob das Modell die Beobachtungen anhand der Prädiktionen in ihrer korrekten Reihenfolge anordnet. Außerdem sind diese Maße einfach zu interpretieren und im Vergleich zu residuenbasierten deutlich robuster gegenüber Ausreißern im Response oder auch in den Kovariablen.

2.1.1 Regression

Vorzugsweise wird die Modellgüte anhand eines Testdatensatzes ermittelt, welcher nicht zur Modellierung berücksichtigt wurde. Dieser besteht aus N Beobachtungen $(\mathbf{x}_i, y_i)$,

$i = 1, \ldots, N$, mit den Kovariablenausprägungen $\mathbf{x}_i$ und einem Response y_i, welcher im Fall einer Regressionsmethode metrisch skaliert ist.

Weit verbreitet zur Evaluierung der Performance eines Regressionsmodells ist folgendes additive Fehlermaß, welches mit den Prädiktionen $\hat{y}_i$ aus den Residuen $r_i = y_i - \hat{y}_i$ gebildet wird:

$$\text{Mittlerer Prädiktionsfehler } = \frac{1}{N} \sum_{i=1}^{N} L(r_i). \tag{2.1}$$

Hierbei können verschiedene Verlustfunktionen $L(r)$ zum Einsatz kommen. Zum Beispiel ergibt sich durch den quadratischen Fehler $L(r) = r^2$, der sehr oft verwendete *Mean Squared Error* (MSE), wobei ein Modell mit einem möglichst geringen MSE angestrebt wird. (Rosset et al., 2006)

Eine andere Möglichkeit der Modellevaluierung bieten rangbasierte Maße. Um diese formal einfach darzustellen wird im Weiteren vorausgesetzt, dass weder im Response noch in den Kovariablen des Testdatensatzes Bindungen auftreten. Zusätzlich werden die Prädiktionen des Modells in absteigender Reihenfolge angenommen, also:

$$\hat{y}_1 > \hat{y}_2 > \cdots > \hat{y}_N.$$

Ebenso wird auch der beobachtete Response $\mathbf{y}$ im Testdatensatz absteigend angeordnet, womit sich der ursprüngliche Rang einer Beobachtung i ergibt zu:

$$s_i = \Big|\{h \leq N : y_i \leq y_h\}\Big|. \tag{2.2}$$

Mithilfe dieser Ränge und der Indikatorfunktion I kann die Anzahl an vertauschten Beobachtungspaaren im Modell definiert werden als

$$T = \sum_{i<h} I(s_i > s_h). \tag{2.3}$$

Dieses Maß wird anschließend transformiert, sodass sich *Kendall's* τ,

$$\tau = 1 - \frac{4T}{N(N-1)}, \tag{2.4}$$

ergibt. Die Transformation stellt sicher, dass τ innerhalb des Wertebereichs $[-1, 1]$ liegt. Denn in einem optimalen Modell ist der beobachtete Response in der gleichen Rangfolge angeordnet wie die Prädiktionen, das heißt es gilt $T = 0$, womit τ den Wert 1 annimmt. Im Gegensatz dazu können aber auch alle möglichen Rangfolgen des beobachteten Responses vertauscht sein, was einer invertierten Rangfolge entspricht. Dabei gilt für die Anzahl an vertauschten Beobachtungspaaren $T = N(N-1)/2$, womit τ wiederum den Wert -1 annimmt. (Rosset et al., 2006)

2.1.2 Klassifikation

Auch für Klassifikationsverfahren sollten Testdatensätze zur Modellgütebestimmung herangezogen werden. Diese haben die gleiche Definition wie schon bei der Regression, allerdings mit kategorialem Response. Im Weiteren werden zwei Maße speziell für binäre Ausprägungen des Responses vorgestellt.

Ähnlich zum MSE für die Regression, misst der Brier Score den mittleren quadratischen Fehler einer Klassifikation. Brier (1950) entwickelte diesen Score ursprünglich um meteorologische mehrkategoriale Vorhersagen auszuwerten, er kann allerdings auch für jedes andere Klassifikationsproblem herangezogen werden. Für einen binären Response ist der Brier Score gegeben durch

$$Brier\ Score = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - y_i)^2, \tag{2.5}$$

wobei y_i der beobachtete Response ist und $\hat{p}_i$ die vom Modell vorhergesagte Wahrscheinlichkeit, dass für die i-te Beobachtung $y_i = 1$ gilt. Wie schon beim MSE induziert auch hier ein *Brier Score* von geringerem Wert eine bessere Prädiktionsgüte für ein Modell. (Roulston, 2007)

Das AUC ist ein weiteres Gütemaß für Klassifikationsverfahren und entspricht der Fläche unter der ROC-Kurve. Diese sogenannten *Receiver Operating Characteristics* Kurven ermöglichen eine visuelle Darstellung der Modellperformance. Hierfür wird zunächst für jede Beobachtung die beobachtete Responseausprägung mit der Prädiktion aus dem Modell verglichen. Dabei können vier verschiedene Fälle auftreten, die in einer sogenannten Konfusionsmatrix zusammengefasst werden. Tabelle 2.1 stellt solch eine Konfusionsmatrix allgemein dar.

		Beobachtete Klasse y		
		1	0	$\sum$
Vorhergesagte Klasse $\hat{y}$	1	richtig positiv (rp)	falsch positiv (fp)	$n_{\hat{1}}$
	0	falsch negativ (fn)	richtig negativ (rn)	$n_{\hat{0}}$
	$\sum$	n_1	n_0	N

Tabelle 2.1: *Konfusionsmatrix bzw. Kontingenztabelle für den Vergleich eines beobachteten binären Responses und einer Modellprädiktion.*

Tritt bei einer Beobachtung $y_i = 1$ auf und die Prädiktion stimmt mit dieser Ausprägung überein, wird die Beobachtung als *richtig positiv* klassifiziert bezeichnet. Stimmt

die Prädiktion jedoch nicht überein, so ist sie *falsch negativ* klassifiziert. Andererseits, wenn $y_i = 0$ die wahre Ausprägung ist und $\hat{y}_i = 0$ vorhergesagt wurde, ist die Beobachtung *richtig negativ* klassifiziert, bzw. bei $\hat{y}_i = 1$ *falsch positiv*. Mithilfe der Anzahl des Auftretens der einzelnen Fälle können verschiedene Kennzahlen ermittelt werden. Zwei wichtige Größen für die *ROC*-Kurve sind dabei die *richtig positiv Rate* und die $falsch\ positiv\ Rate$:

$$richtig\ positiv\ Rate = \frac{rp}{n_1}, \tag{2.6}$$

$$falsch\ positiv\ Rate = \frac{fp}{n_0}. \tag{2.7}$$

Mit n_0 und n_1 werden dabei jeweils die Anzahl an Beobachtungen mit Responseausprägung $y = 0$ bzw. $y = 1$ bezeichnet. Für probabilistische Klassifikationsmethoden, bei denen der geschätzte Response nicht konkret 0 oder 1 ist, sondern die Wahrscheinlichkeit $\hat{p}$ für eine der beiden Klassen, kann die Konfusionsmatrix für verschiedene Schwellenwerte zwischen 0 und 1 aufgestellt werden. Erst wenn die geschätzte Wahrscheinlichkeit $\hat{p}$ diesen Schwellenwert überschreitet, wird einer Beobachtung die vorhergesagte Klasse $\hat{y} = 1$ zugewiesen. Für jeden Schwellenwert gelten demnach auch andere richtig positiv bzw. falsch positiv Raten. Diese verschiedenen Raten können zweidimensional gegeneinander abgetragen werden, wodurch sich die *ROC*-Kurve ergibt. Dabei liegt die falsch positiv Rate auf der x-Achse und die richtig positiv Rate auf der y-Achse. In Abbildung 2.1 ist beispielhaft eine *ROC*-Kurve dargestellt.

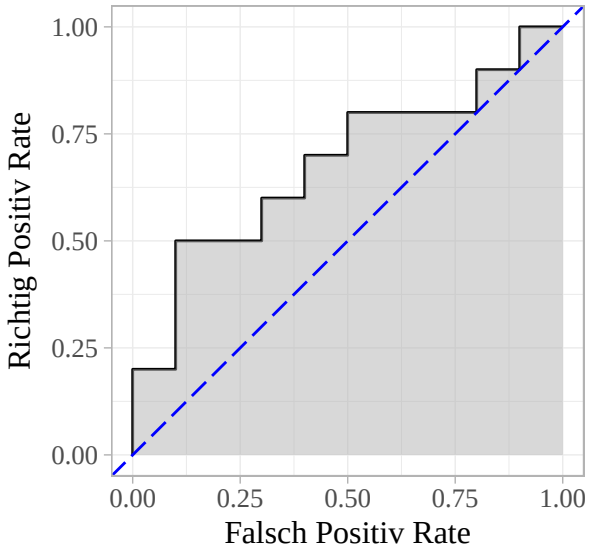


Abbildung 2.1: *ROC-Kurve eines Beispieldatensatzes (In Anlehnung an Fawcett (2006)). Die grau eingezeichnete Fläche unterhalb der Kurve entspricht dabei dem AUC. Die ROC-Kurve einer Klassifikationsmethode mit zufälligen Vorhersagen liegt auf der gestrichelten blauen Linie.*

Liegt die *ROC*-Kurve auf der Diagonalen spricht dies für ein Modell, das absolut zufällig eine Klasse vorhersagt. In der Regel befindet sich die *ROC*-Kurve also oberhalb dieser Diagonale. Perfekte Performance ist dadurch gekennzeichnet, dass der Kurvenverlauf vertikal ausgehend vom Punkt $(0, 0)$ zu $(0, 1)$ verläuft und anschließend horizontal bei $(1, 1)$ endet. Die Performance lässt sich nun auch mithilfe einer Maßzahl ausdrücken, indem die Fläche unterhalb der *ROC*-Kurve betrachtet wird. Dieses Performancemaß wird als *AUC* (*Area Under the Curve*) bezeichnet und kann durch Integrieren der Funktion, welche die *ROC*-Kurve beschreibt, ermittelt werden. Theoretisch kann das *AUC* innerhalb des Wertebereichs $[0, 1]$ liegen, da jedoch die *ROC*-Kurve für gewöhnlich oberhalb der angesprochenen Diagonale liegt, sollte keine realistische Klassifikationsmethode ein *AUC* kleiner als 0.5 aufweisen. Es gilt, je größer das *AUC*, desto größer auch die Performance des Modells. Eine interessante Eigenschaft des *AUC* ist außerdem, dass diese Fläche der Wahrscheinlichkeit entspricht, dass zwei zufällig gezogene Beobachtungen $i1$ und $h0$ mit $y_{i1} = 1$ und $y_{h0} = 0$ von der Klassifikationsmethode korrekt geordnet werden, das heißt es gilt $AUC = P(\hat{p}_{i1} > \hat{p}_{h0})$. (Fawcett, 2006)
Hanley und McNeil (1982) zeigen, dass das empirische *AUC* daher äquivalent zur Wilcoxon Teststatistik W ist und sich auch folgendermaßen berechnen lässt:

$$AUC = W = \frac{1}{n_1 \cdot n_0} \sum_{i=1}^{n_1} \sum_{h=1}^{n_0} S(\hat{p}_{i1}, \hat{p}_{h0}), \tag{2.8}$$

$$\text{mit } S(\hat{p}_{i1}, \hat{p}_{h0}) = \begin{cases} 1, & \text{falls } \hat{p}_{i1} > \hat{p}_{h0} \\ 0.5, & \text{falls } \hat{p}_{i1} = \hat{p}_{h0} \\ 0, & \text{falls } \hat{p}_{i1} < \hat{p}_{h0} \end{cases}. \tag{2.9}$$

Dabei bezeichnen n_0 und n_1 die jeweilige Anzahl an Beobachtungen, für die $y = 0$ bzw. $y = 1$ gilt, und $\hat{p}_{gk}$ ($g \in \{h, i\}$, $k \in \{0, 1\}$) sind die entsprechenden Modellprädiktionen einer Beobachtung g mit wahrer Klasse $y = k$.
Das *AUC* kann damit also ähnlich zu *Kendall's* τ aus Gleichung (2.4) auch als rangbasiertes Maß interpretiert werden.

2.2 CART - Classification and Regression Trees

CART (*Classification and Regression Trees*) bezeichnet eine Methode, mit der baumbasierte Klassifikation und Regression durchgeführt werden kann. Dabei wird der Variablenraum rekursiv in verschiedene Partitionen eingeteilt.
Rechts in Abbildung 2.2 ist beispielhaft die Partitionierung eines Datensatzes mit zwei Variablen $\mathbf{X}_1$ und $\mathbf{X}_2$ dargestellt. Die Unterräume $R_1, \ldots, R_5$ sind durch wiederholtes Teilen der Einzelräume in zwei Gruppen entstanden. Dabei wird für jeden Split eine Variable und

eine entsprechende Variablenausprägung t gewählt, sodass ein vorab bestimmtes Kriterium optimiert wird. Ebenso wird vorab ein Stopkriterium definiert, welches den Zeitpunkt festlegt, ab dem keine weiteren Splits mehr durchgeführt werden.

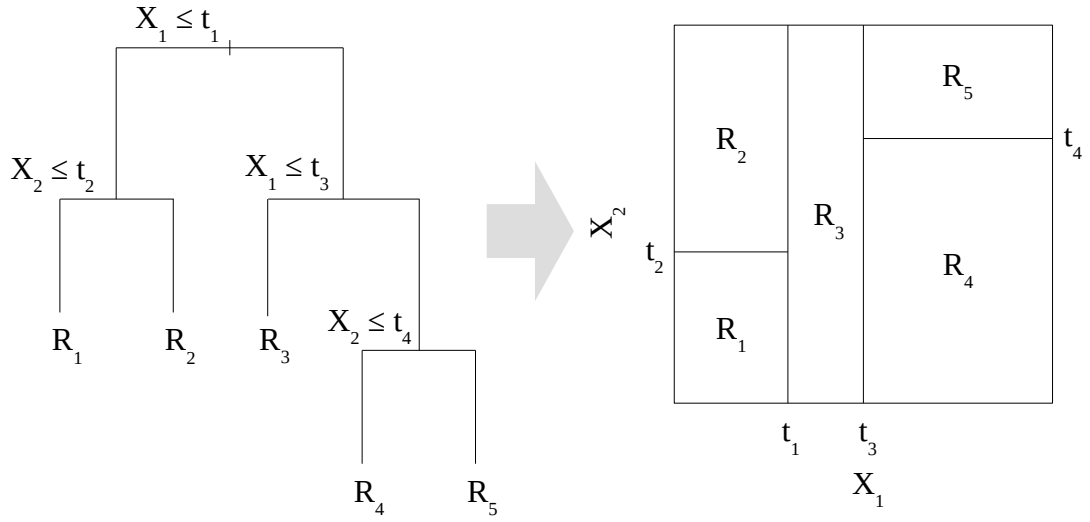


Abbildung 2.2: *CART Beispiel für einen zweidimensionalen Variablenraum, der rekursiv binär aufgeteilt wurde. Links der zugehörige Entscheidungsbaum, der die rechte Partition bildet (In Anlehnung an Hastie et al. (2009, S. 306)).*

Links in Abbildung 2.2 ist die gleiche Partitionierung in ihrer Baumstruktur dargestellt, welche eine einfache Interpretierbarkeit des gruppierten Variablenraums ermöglicht. Der vollständige Datensatz befindet sich dabei an der Spitze des Baumes und wird an den jeweiligen Ästen entlang in die Unterräume $R_1, \ldots, R_5$, die auch *Terminal Nodes* genannt werden, aufgesplittet. Um nun auf Basis solch eines Baumes einen Response $\mathbf{y}$ vorherzusagen, wird dieser in jeder Terminal Node separat als Konstante modelliert. (Hastie et al., 2009, S. 305)

Je nachdem, welcher Response mit einem Entscheidungsbaum abgebildet werden soll, unterscheidet sich diese Modellierung und auch das Splitkriterium. Daher wird im Weiteren zwischen Regressionsbäumen für metrischen Response und Klassifikationsbäumen für kategorialen Response unterschieden.

2.2.1 Regressionsbäume

Ein Regressionsbaum wird bei metrischem Response $\mathbf{y}$ angewendet. Dabei sind für N Beobachtungen sowohl der Response y_i, $i = 1, \ldots, N$, als auch die Ausprägungen von p Kovariablen $\mathbf{X}_j$, $j = 1, \ldots, p$, bekannt. Jeder Beobachtung i kann also ein Vektor $(\mathbf{x}_i, y_i)$ mit $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ zugeordnet werden.

Bei einer Partitionierung in M Terminal Nodes $R_1, \ldots, R_M$ wird der Response in jeder

dieser Node als Konstante c_m modelliert:

$$f(\mathbf{x}) = \sum_{m=1}^{M} c_m I(\mathbf{x} \in R_m). \tag{2.10}$$

Mit der Quadratsumme $\sum(y_i - f(\mathbf{x}_i))^2$ als Minimierungskriterium ergibt sich daraufhin der Mittelwert aller y_i in einer Terminal Node R_m als optimaler Schätzer für c_m, also

$$\hat{c}_m = \frac{1}{|R_m|} \sum_{i|\mathbf{x}_i \in R_m} y_i, \tag{2.11}$$

wobei $|R_m|$ die Anzahl an Beobachtungen in einer Node R_m angibt.
Um nun die erste Splitvariable j und den optimalen Splitpunkt t zu bestimmen, werden zwei Halbebenen,

$$R_1(j,t) = \{\mathbf{X}|\mathbf{X}_j \leq t\} \text{ und } R_2(j,t) = \{\mathbf{X}|\mathbf{X}_j > t\}, \tag{2.12}$$

definiert. Die Splitvariable j und der Splitpunkt t sind dann diejenigen, die folgende Bedingung erfüllen:

$$\min_{j,t} \left[\min_{c_1} \sum_{i|\mathbf{x}_i \in R_1(j,t)} (y_i - c_1)^2 + \min_{c_2} \sum_{i|\mathbf{x}_i \in R_2(j,t)} (y_i - c_2)^2 \right]. \tag{2.13}$$

Die inneren Minimierungen werden dabei für alle j und t mit Gleichung (2.11) und den in Gleichung (2.12) definierten Halbebenen gelöst, womit sich

$$\begin{aligned} \hat{c}_1 &= \frac{1}{|R_1(j,t)|} \sum_{i|\mathbf{x}_i \in R_1(j,t)} y_i \text{ und} \\ \hat{c}_2 &= \frac{1}{|R_2(j,t)|} \sum_{i|\mathbf{x}_i \in R_2(j,t)} y_i \end{aligned} \tag{2.14}$$

ergeben. Für jede Splitvariable werden dabei alle beobachteten Ausprägungswerte als Splitpunkte geprüft, um das beste Paar (j,t) in Bezug auf das Minimierungskriterium zu finden. Mit diesem Paar werden die Daten in zwei Unterräume gesplittet und in jedem dieser Unterräume wieder ein binärer Split durchgeführt. Laut Duroux und Scornet (2016) hat der Response $\mathbf{y}$ in den mit diesem Splitkriterium jeweils entstehenden Unterräumen minimale (empirische) Varianz, was auch in Gleichung (2.13) ersichtlich wird. Dieses Vorgehen wird solange rekursiv angewendet bis ein vorab definiertes Abbruchkriterium erfüllt ist. Meist wird dafür die minimale Anzahl an Beobachtungen in einer Node verwendet, welche nicht unterschritten werden soll. (Hastie et al., 2009, S. 307)
Zusätzlich gibt es auch die Möglichkeit, einen so aufgestellten Baum T_0 zu stutzen. Dafür kann zum Beispiel das sogenannte *cost-complexity pruning* angewendet werden. Durch das Entfernen von einzelnen Ästen in T_0 entstehen Unterbäume $T \subset T_0$ mit einer geringeren

Anzahl an Terminal Nodes ($\hat{=}|T|$). Nun wird das Kosten-Komplexitäts Kriterium

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha|T|, \tag{2.15}$$
$$\text{mit } Q_m(T) = \frac{1}{N_m} \sum_{i|\mathbf{x}_i \in R_m} (y_i - \hat{c}_m)^2,$$
$$N_m = |R_m|,$$
$$\hat{c}_m = \frac{1}{N_m} \sum_{i|\mathbf{x}_i \in R_m} y_i$$

minimiert. Ziel ist das Auffinden eines Unterbaums $T_\alpha \subseteq T_0$ für jedes α. Dabei stellt α einen Tuningparameter größer oder gleich Null dar, der den Tradeoff zwischen Baumgröße und der Anpassung an die Daten reguliert. Je größer dabei α, desto kleiner fallen die Bäume T_α aus. Um T_α zu ermitteln, werden schrittweise diejenigen Nodes m entfernt, die den kleinsten Anteil zur Summe $\sum_m N_m Q_m(T)$ beitragen. Aus all diesen Unterbäumen wird dann mithilfe einer Kreuzvalidierung der Wert $\hat{\alpha}$ für α geschätzt, der die Quadratsumme minimiert, womit der finale Baum $T_{\hat{\alpha}}$ ist. (Hastie et al., 2009, S. 308)

2.2.2 Klassifikationsbäume

Im Gegensatz zum metrischen Response $\mathbf{y}$ für die Regressionsbäume liegt der Response für einen Klassifikationsbaum als kategoriale Variable mit den Ausprägungen $1, \ldots, K$ vor. Die Konstruktion solch eines Baumes ähnelt stark der eines Regressionsbaumes, mit dem einzigen Unterschied im Kriterum $Q_m(T)$, das für einen Split und für das Stutzen Verwendung findet.
Auch hierbei entstehen Unterräume R_m, denen jeweils N_m Beobachtungen zugeordnet werden. Die Klassenzuordnung dieser Beobachtungen wird dabei über das Mehrheitsverhältnis bestimmt. Eine Beobachtung in Node m wird demnach der Klasse $k(m)$ zugeordnet, welche am häufigsten innerhalb dieser Node auftritt:

$$k(m) = \arg\max_k \hat{p}_{mk}, \tag{2.16}$$
$$\text{mit } \hat{p}_{mk} = \frac{1}{N_m} \sum_{i|\mathbf{x}_i \in R_m} I(y_i = k). \tag{2.17}$$

Häufig verwendete Splitkriterien eines Klassifikationsbaumes sind der Missklassifizierungsfehler oder der Gini Index. Beide Kriterien sind Maße $Q_m(T)$ für die Unreinheit einer Node. Dabei misst der Missklassifizierungsfehler den Anteil an Beobachtungen, die durch das oben beschriebene Mehrheitsverhältnis einer falschen Klasse zugeordnet werden. Der Gini Index berücksichtigt dagegen für jede Klasse k das Produkt aus Wahrscheinlichkeit und Gegenwahrscheinlichkeit für eine Klasse.

Formal lassen sich diese beiden Maße auch folgendermaßen ausdrücken:

$$\text{Missklassifizierungsfehler:} \quad Q_m(T) = \frac{1}{N_m} \sum_{i|\mathbf{x}_i \in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_{mk(m)}. \tag{2.18}$$

$$\text{Gini Index:} \quad Q_m(T) = \sum_{k \neq k'} \hat{p}_{mk}\hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}). \tag{2.19}$$

Im Klassifikationsbaum wird die Splitvariable j und ensprechende Splitausprägung t gesucht, sodass die Unreinheit in beiden neu entstehenden Unterräumen simultan minimiert wird. Dafür werden die Unreinheitsmaße $Q_m(T)$ der beiden Unterräume mit der Anzahl der jeweiligen Beobachtungen N_1 und N_2 gewichtet. (Hastie et al., 2009, S. 308–310)
Demnach muss für einen Split die Bedingung

$$\min_{j,t} \left[Q_1(T) \cdot N_1 + Q_2(T) \cdot N_2 \right] \tag{2.20}$$

erfüllt sein. Wie auch schon bei den Regressionsbäumen sind auch hier die Maße $Q_1(T)$ und $Q_2(T)$ abhängig von den zwei definierten Halbebenenen aus Gleichung (2.12) und somit auch von dem Paar (j, t).
Das Verfahren zum Stutzen eines Baumes verläuft mit dem entsprechend ausgetauschten Unreinheitsmaß $Q_m(T)$ analog zum Regressionsbaum.

2.3 Random Forests

Ein *Random Forest* ist eine Abwandlung der *Bootstrap Aggregation*, kurz auch *Bagging* genannt. Beim Bagging wird im Allgemeinen ein Modell auf verschiedene Bootstrap-Stichproben des Datensatzes angewendet und die resultierenden Schätzer gemittelt. Das führt zu einer erheblichen Varianzreduktion der Schätzfunktion.
Im Folgenden wird beschrieben, wie sich ein Random Forest auf Basis der vorab vorgestellten Entscheidungsbäume bilden lässt und welche Bedeutung dabei korrelierte Kovariablen und der Hyperparamter *mtry* besitzen.

2.3.1 Verfahren

Konkret wird bei einem Random Forest eine große Anzahl an Entscheidungsbäumen betrachtet. Diese werden auf Basis von Bootstrap-Stichproben (Ziehen mit Zurücklegen) der gleichen Größe des ursprünglichen Datensatz gebildet. Ergebnisse aus Entscheidungsbäumen können bereits bei kleineren Änderungen in den Daten unterschiedlich ausfallen. Aus diesem Grund und da besonders tiefe Bäume (zum Beispiel ungestutzte Bäume) einen relativ geringen Bias aufweisen, eignen sie sich hervorragend für das Bagging und profitieren stark von der Mittelwertbildung der Schätzer aus den einzelnen Bäumen. Eine Schwachstelle dabei ist allerdings, dass die paarweise Korrelation der

einzelnen Bäume die Varianzreduktion begrenzt. Dem wird im Random Forest dadurch entgegengewirkt, dass in jedem Split nicht alle Variablen berücksichtigt werden, sondern eine zufällige Auswahl einer bestimmten Anzahl ($mtry$) an Splitkandidaten vorgenommen wird. Das reduziert die Korrelation zwischen den Bäumen ohne die Varianz des gemittelten Schätzers stark ansteigen zu lassen. Je kleiner $mtry$ dabei gewählt wird, desto geringer fällt auch die Korrelation zwischen den Bäumen aus. (Hastie et al., 2009, S. 587–588)

Der Algorithmus für einen Random Forest wird von Hastie et al. (2009, S. 588) folgendermaßen zusammengefasst:

Algorithmus 1: Random Forest nach Hastie et al. (2009, S. 588).

Sei B die Anzahl an Bäumen in einem Random Forest.

für $b = 1$ **bis** B

(a) Ziehe eine Bootstrap-Stichprobe $\mathbf{Z}^*$ der Größe N aus den Trainingsdaten.

(b) Bilde mit dieser Stichprobe einen Baum T_b, durch rekursive Wiederholung der folgenden Schritte in jeder Terminal Node, bis n_{min}, die minimale Anzahl an Beobachtungen in einer Node erreicht ist.

   i. Wähle zufällig $mtry$ Variablen aus den p Kovariablen.

   ii. Ermittle die optimale Variable und Variablenausprägung unter den $mtry$ Variablen.

   iii. Splitte die Node in zwei Tochter-Nodes.

Gebe alle Bäume $\{T_b\}_1^B$ zurück.

Dieser Algorithmus lässt sich sowohl für einen Random Forest mit metrischem Response als auch mit kategorialem Response anwenden. Dabei werden bestimmte Defaultwerte für $mtry$ und das Abbruchkriterium n_{min} abhängig vom Response empfohlen. Bei Regressionsproblemen wird hierfür $mtry = \lfloor p/3 \rfloor$ und $n_{min} = 5$ vorgeschlagen, bei Klassifikationsproblemen $mtry = \left\lfloor \sqrt{p} \right\rfloor$ und $n_{min} = 1$. Allerdings weisen Hastie et al. (2009, S. 592) auch darauf hin, dass diese Parameter abhängig von den vorliegenden Daten sein können und als Tuningparameter behandelt werden sollten.
Unterschiede treten außerdem auf, wenn die Prädiktion für eine neue Beobachtung mit Variablenausprägungen $\mathbf{x}$ bestimmt werden soll. Im Regressionsfall mit metrischem Response wird dieser über den Mittelwert der Prädiktionen $\hat{y}_b = T_b(\mathbf{x})$ der einzelnen Bäume geschätzt:

$$\hat{y}_{regr}(\mathbf{x}) = \hat{f}_{rf}^B(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} T_b(\mathbf{x}). \tag{2.21}$$

Dagegen wird bei der Klassifikation die sogenannte *majority vote* angewendet. Jeder der B Bäume hat eine eigene Klassenprädiktion $\hat{C}_b(\mathbf{x})$ für eine Beobachtung mit Kovariablen $\mathbf{x}$. Im Random Forest erhält diese Beobachtung jene Klasse zugewiesen, welche am häufigsten unter all den Bäumen auftritt, was sich formal darstellen lässt als

$$\hat{y}_{klass}(\mathbf{x}) = \hat{C}_{rf}^{B}(\mathbf{x}) = majority\ vote\ \{\hat{C}_b(\mathbf{x})\}_1^B. \tag{2.22}$$

Eine Besonderheit des Random Forests stellen die *out-of-bag* (OOB) Prädiktionen einer Beobachtung dar. Hierfür werden sowohl für Regression als auch für Klassifikation zur Prädiktion nicht alle B Bäume berücksichtigt, sondern nur diejenigen, deren Bootstrap-Trainingsdaten die betrachtete Beobachtung nicht enthalten. Mit diesen OOB-Prädiktionen lässt sich dann auch der OOB-Fehler des Random Forests schätzen. Sobald sich dieser stabilisiert, gibt er einen Anhaltspunkt, dass keine weiteren Iterationen bzw. Bäume für den Forest nötig sind. (Hastie et al., 2009, S. 588–593) Außerdem ist durch die Anwendung des OOB-Fehlers kein zusätzlicher Testdatensatz nötig, da der OOB-Fehler bei ausreichend hoher Anzahl an Bäumen genauso präzise ist, wie der Fehler eines Testdatensatzes mit gleicher Anzahl an Beobachtungen wie der Trainingsdatensatz. Ein weiterer Vorteil ist die Rechengeschwindigkeit, denn anders als zum Beispiel bei einer k-fachen Kreuzvalidierung, müssen nicht k Random Forests konstruiert werden, sondern nur einer, aus diesem der OOB-Fehler ermittelt werden kann. Zudem können die OOB-Prädiktionen zum Beispiel auch für die Bestimmung der Variablenwichtigkeiten eingesetzt werden. (Breiman, 2001)

Es gibt verschiedene Möglichkeiten die Variablenwichtigkeit zu messen. Ein simpler Ansatz ist zu zählen, wie häufig eine Variable als Splitkandidat in den einzelnen Bäumen eines Forests ausgewählt wurde. Etwas aufwendiger ist es dagegen das (gewichtete) Mittel aus den einzelnen Verbesserungen des Splitkriteriums durch jede Variable zu bestimmen. Dabei beschreibt zum Beispiel die *Gini Importance* die Verbesserung des Gini Index durch jede Variable. Häufig wird jedoch die *Permutation Accuracy Importance*, im Weiteren oft auch nur *Permutation Importance* genannt, angewendet, welche auf den OOB-Schätzern beruht. Dazu wird folgende Prozedur für jeden Baum und jede der p Kovariablen $\mathbf{X}_j$ wiederholt:

1. Der ursprüngliche Zusammenhang zwischen $\mathbf{X}_j$ und dem Response wird durch zufällige Permutation von $\mathbf{X}_j$ aufgehoben.

2. Auf Basis der übrigen Kovariablen und der permutierten Variable $\mathbf{X}_j$ werden die OOB-Prädiktionen des Responses erstellt und damit die Prädiktionsgüte (OOB-Fehler) ermittelt.

3. Die Prädiktionsgüte aus 2. wird mit der Prädiktionsgüte vor der Permutation durch die Differenz der beiden verglichen.

Die Wichtigkeit einer Variable $\mathbf{X}_j$ entspricht anschließend dem Mittelwert der Prädiktionsgüte-Differenzen (aus 3.) aller Bäume des Random Forests. Bei einem hohen Einfluss von $\mathbf{X}_j$ auf den Response, wird davon ausgegangen, dass die Prädiktionsgüte deutlich sinkt, wenn die permutierte Kovariable zur Prädiktion verwendet wird, was damit eine hohe Ausprägung der Variablenwichtigkeit zur Folge hat. (Strobl et al., 2007)
Bei der Wahl des Maßes für die Prädiktionsgüte können zum Beispiel die in Kapitel 2.1 vorgestellten Modellgütemaße zum Einsatz kommen. Wie der Name der Permutation Accuracy Importance jedoch schon vermuten lässt, wird für die Klassifikation meist die *Accuracy* verwendet, welche den Anteil an korrekt klassifizierten Beobachtungen innerhalb der Daten angibt. Für die Regression kommt dagegen überwiegend der MSE zum Einsatz.

2.3.2 Korrelierte Variablen

Vor allem im Zusammenhang mit der Variablenwichtigkeit werden häufig auch korrelierte Kovariablen im Random Forest untersucht.

Eine Studie von Strobl et al. (2008) definiert dafür einen Simulationsdatensatz mithilfe eines linearen Modells, der neben dem Response noch weitere zwölf Kovariablen beinhaltet. Die festgelegten Kovariableneinflüsse können Tabelle 2.2 entnommen werden. Zudem erhalten die ersten vier Variablen $\mathbf{X}_1, \ldots, \mathbf{X}_4$ eine starke Blockkorrelation, während die restlichen Kovariablen unkorreliert sind. Eine detailliertere Beschreibung eines ähnlichen Simulationsvorgehens wird in Kapitel 3.1 gegeben. Es sei darauf hingewiesen, dass die folgenden Ergebnisse sowohl für den Regressionsfall als auch für die Klassifikation gelten und zudem anhand eines Anwendungsbeispiels überprüft wurden. Außerdem kommen für die Konstruktion der Forests nicht wie üblich CART-Entscheidungsbäume zum Einsatz, sondern *Conditional Inference Trees*, die basierend auf bedingten Teststatistik-Verteilungen den Zusammenhang zwischen Response und Kovariablen messen, was ein etwas anderes Vorgehen im Splitprozess zur Folge hat (Näheres dazu in Hothorn et al. (2006)).

$\mathbf{X}_j$	$\mathbf{X}_1$	$\mathbf{X}_2$	$\mathbf{X}_3$	$\mathbf{X}_4$	$\mathbf{X}_5$	$\mathbf{X}_6$	$\mathbf{X}_7$	$\mathbf{X}_8$	...	$\mathbf{X}_{12}$
β_j	5	5	2	0	−5	−5	−2	0	...	0

Tabelle 2.2: *Regressionskoeffizienten zur Datengenerierung der Simulationsstudie von Strobl et al. (2008).*

Neben den korrelierten Variablen analysieren Strobl et al. (2008) auch den Einfluss von *mtry* auf die Auswahlhäufigkeiten der Variablen für einen Baumsplit. Dabei zeigt sich, dass gemittelt über alle Splits der Bäume für $mtry \neq 1$ die korrelierten Variablen mit schwachem oder fehlendem Einfluss ($\mathbf{X}_3$ und $\mathbf{X}_4$) häufiger ausgewählt werden als die gleichstarken unkorrelierten Variablen ($\mathbf{X}_7$ und $\mathbf{X}_8$). Dieses Phänomen lässt sich dadurch erklären, dass auch wenn eine Variable keinen oder nur einen geringen Einfluss auf den Response hat, jedoch hochkorreliert mit einer anderen Einflussvariable ist, diese bei der Splitwahl als ebenso guter Splitkandidat wie die Variable mit tatsächlichem Einfluss erscheint. Für steigendes *mtry* sinken die Auswahlhäufigkeiten für $\mathbf{X}_3$ und $\mathbf{X}_4$ jedoch, da damit auch die Wahrscheinlichkeit steigt, dass eine tatsächlich relevante Variable ebenfalls als Splitkandidat betrachtet wird.
Diese Beobachtungen haben auch Auswirkung auf die Schätzung der Variablenwichtigkeit: Die Permutation Importance aus Kapitel 2.3.1 spiegelt nicht die durch die Regressionskoeffizienten festgelegte Struktur wider. Die Variablenwichtigkeiten der korrelierten Kovariablen werden dabei deutlich überschätzt, was zum Beispiel fast dreimal so hohe Variablenwichtigkeiten für die Kovariablen $\mathbf{X}_1$ und $\mathbf{X}_2$ im Vergleich zu den gleichstarken Kovariablen $\mathbf{X}_5$ und $\mathbf{X}_6$ zur Folge hat. Für kleine *mtry*-Werte ist dieser Effekt sogar stärker ausgeprägt als für große, da die Chance, dass eine korrelierte Variable in einem Baum früh als Splitkandidat ausgewählt wird, höher ist, wenn die restlichen korrelierten Kovariablen nicht als Splitkandidaten berücksichtig werden. Allerdings stellen Strobl et al. (2008) auch fest, dass für höhere *mtry*-Werte die Variablilität der Variablenwichtigkeit steigt. Um diese Problematiken zu umgehen, schlagen Strobl et al. (2008) die Verwendung der sogenannten *Conditional Permutation Importance* vor. Außerdem erwähnen sie, dass für kleine *mtry*-Werte eine höhere Prädiktionsgüte erwartet werden kann, wobei ein Nachweis hierzu nicht angeführt wird.

Ähnliche, allerdings etwas komplexere, Simulationsdesigns verwenden auch Toloşi und Lengauer (2011). Dabei besitzen die blockkorrelierten Variablen die gleiche bzw. ähnliche Einflussstärke auf einen ausschließlich binären Response. Neben zwei verschiedenen Simulationsdesigns werden auch zwei reale Datensätze betrachtet. Aus dieser Analyse kann eine wesentliche Erkenntnis geschlossen werden: Die geschätzten Variablenwichtigkeiten sind verzerrt, was in diesem Fall bedeutet, dass je mehr Kovariablen blockkorreliert sind, desto kleiner werden deren Variablenwichtigkeiten. Dadurch können Kovariablen, die zwar einen starken Einfluss auf den Response haben, jedoch mit sehr vielen anderen Kovariablen hochkorreliert sind, mithilfe der Variablenwichtigkeit nicht als relevante Variablen erkannt werden. Der Grund für diese Verzerrung im Random Forest liegt dabei in der Randomisierung der einzelnen Bäume durch zum einen die Bootstrap-Stichproben und zum anderen die Auswahl von *mtry* Variablen im Splitprozess. Das hat zur Folge, dass die korrelierten

Variablen untereinander austauschbar als Splitkandidat eingesetzt werden können. Zudem ist intuitiv nachvollziehbar, dass sich bei der Berechnung der Variablenwichtigkeiten durch die Permutation einer relevanten korrelierten Kovariable keine deutlich schlechtere Prädiktionsgüte ergibt, da in diesem Fall zur Vorhersage die anderen korrelierten Kovariablen herangezogen werden können, welche ähnliche Informationen wie die permutierte Kovariable tragen.

Auch Genuer et al. (2008) untersuchen den Effekt von stark korrelierten Kovariablen im Random Forest. Allerdings verwenden Sie einen hochdimensionalen Simulationsdatensatz für Klassifikation, dem zusätzliche korrelierte Replikationen der relevanten Variablen hinzugefügt werden. Dabei wird, wie auch von Tolo̧si und Lengauer (2011), ein Sinken der Variablenwichtigkeit der relevanten Kovariablen beobachtet, je größer die Gruppe der korrelierten Kovariablen ist.

Eine aktuelle Zusammenfassung zu bisherigen Forschungen im Bereich der Korrelation bei Random Forests und deren Auswirkung auf die Variablenwichtigkeit geben Gregorutti et al. (2016). Außerdem werden theoretische Herleitungen für den Zusammenhang zwischen korrelierten Kovariablen und der Permutation Importance in einem additiven Regressionsmodell aufgestellt. Diese zeigen, dass die Variablenwichtigkeit sehr sensibel auf Korrelationen zwischen den Kovariablen reagiert. Im Folgenden werden nur die Ergebnisse dieser Herleitungen dargestellt. Dabei werden fünf Fälle von Korrelationen zwischen Kovariablen unterschieden, wobei die positive Korrelation zwischen diesen Variablen mit c bezeichnet ist:
Fall 1 - Zwei korrelierte Kovariablen $\mathbf{X}_1$ und $\mathbf{X}_2$ und Korrelation τ_0 mit dem Response.
Damit gilt $(\mathbf{X}_1, \mathbf{X}_2, \mathbf{Y}) \sim N_3(\mathbf{0}, \mathbf{\Sigma})$, mit

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & c & \tau_0 \\ c & 1 & \tau_0 \\ \tau_0 & \tau_0 & 1 \end{pmatrix}. \tag{2.23}$$

N_3 entspricht hier einer 3-dimensionalen Normalverteilung mit der Kovarianzmatrix $\mathbf{\Sigma}$ und $\mathbf{0} = (0, 0, 0)$, dem Vektor der Erwartungswerte.
Für steigende positive Korrelationen c sinkt die Variablenwichtigkeit der beiden Kovariablen $\mathbf{X}_1$ und $\mathbf{X}_2$.

Fall 2 - Zwei korrelierte Kovariablen und eine von diesen Variablen unabhängige Kovariable $\mathbf{X}_3$, deren Korrelation mit dem Response sich von τ_0 unterscheidet.

Damit gilt $(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{Y}) \sim N_4(\mathbf{0}, \mathbf{\Sigma})$, mit

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & c & 0 & \tau_0 \\ c & 1 & 0 & \tau_0 \\ 0 & 0 & 1 & \tau_3 \\ \tau_0 & \tau_0 & \tau_3 & 1 \end{pmatrix}. \tag{2.24}$$

Wenn c ausreichend groß ist, kann die Variablenwichtigkeit von $\mathbf{X}_3$ die Variablenwichtigkeiten von $\mathbf{X}_1$ und $\mathbf{X}_2$ übersteigen, auch wenn $\tau_3 < \tau_0$ gilt.

Fall 3 - Alle p Kovariablen sind blockkorreliert und haben Korrelation τ_0 mit dem Response.
Damit gilt $(\mathbf{X}_1, \ldots, \mathbf{X}_p, \mathbf{Y}) \sim N_{p+1}(\mathbf{0}, \mathbf{\Sigma})$, mit

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & c & \cdots & c & \tau_0 \\ c & 1 & \cdots & c & \tau_0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ c & c & \cdots & 1 & \tau_0 \\ \tau_0 & \tau_0 & \cdots & \tau_0 & 1 \end{pmatrix}. \tag{2.25}$$

Je größer die Anzahl an Kovariablen p, desto schneller sinken die Variablenwichtigkeiten der Kovariablen gegen 0.
Die Fälle 2 und 3 entsprechen damit den Beobachtungen von Tolo̧si und Lengauer (2011).

Fall 4 - p blockkorrelierte Kovariablen und q Kovariablen unabhängig von diesen, deren Korrelation mit dem Response sich untereinander und von τ_0 unterscheidet.
Damit gilt $(\mathbf{X}_1, \ldots, \mathbf{X}_p, \mathbf{X}_{p+1}, \ldots, \mathbf{X}_{p+q}, \mathbf{Y}) \sim N_{p+q+1}(\mathbf{0}, \mathbf{\Sigma})$, mit

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & \cdots & c & 0 & \cdots & 0 & \tau_0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ c & \cdots & 1 & 0 & \cdots & 0 & \tau_0 \\ 0 & \cdots & 0 & 1 & \cdots & 0 & \tau_{p+1} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 1 & \tau_{p+q} \\ \tau_0 & \cdots & \tau_0 & \tau_{p+1} & \cdots & \tau_{p+q} & 1 \end{pmatrix}. \tag{2.26}$$

Auch hierbei können die unabhängigen Variablen $\mathbf{X}_{p+1}, \ldots, \mathbf{X}_{p+q}$ stärkere Variablenwichtigkeiten als die untereinander korrelierten Kovariablen $\mathbf{X}_1, \ldots, \mathbf{X}_p$ aufweisen, auch wenn $\tau_{p+1}, \ldots, \tau_{p+q} < \tau_0$ gilt.

Fall 5 - Zwei negativ korrelierte Kovariablen $\mathbf{X}_1$ und $\mathbf{X}_2$, diese Korrelation wird mit $-\rho$ bezeichnet.
Damit gilt $(\mathbf{X}_1, \mathbf{X}_2, \mathbf{Y}) \sim N_3(\mathbf{0}, \mathbf{\Sigma})$, mit

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & -\rho & \tau_0 \\ -\rho & 1 & \tau_0 \\ \tau_0 & \tau_0 & 1 \end{pmatrix}. \tag{2.27}$$

Für steigendes ρ, was einer stärkeren negativen Korrelation entspricht, steigt auch die Variablenwichtigkeit von $\mathbf{X}_1$ und $\mathbf{X}_2$. Der Grund dafür liegt in der entgegengesetzten Wirkrichtung der Variablen, wodurch beide Variablen zur Erklärung von $\mathbf{Y}$ im Modell benötigt werden.

2.3.3 Hyperparameter mtry

Wie bereits erwähnt, kann mithilfe des Hyperparameters *mtry* die Stärke der Randomisierung innerhalb der Bäume eines Random Forests gesteuert werden. Je kleiner *mtry* gewählt wird, desto größer der Randomisierungseffekt bei der Splitwahl. Dieser Parameter hat damit Einfluss auf eine der wichtigsten Eigenschaften des Random Forest. Oft werden die empfohlenen Defaultwerte $\lfloor p/3 \rfloor$ für Regressionsmodelle und $\left\lfloor \sqrt{p} \right\rfloor$ für Klassifikationsmodelle verwendet. Allerdings ist es fragwürdig, ob diese Werte im Allgemeinen tatsächlich eine gute Wahl darstellen.

Letzteres und das Fehlen eines theoretischen Konzepts zur *mtry*-Wahl beanstanden auch Bernard et al. (2009) und führen eine Analyse mit Klassifikationsforests durch. Dabei werden zwölf verschiedene reale Datensätze mit variierender Anzahl an Beobachtungen und Variablen betrachtet, welche auch mehrkategoriale Responses beinhalten. Die Ergebnisse zeigen, dass in dreiviertel der Datensätze die Modellgüte mit dem Defaultwert für *mtry* sehr nahe an der optimalen Modellgüte liegt. Als Modellgütemaß wurde hierbei die *Accuracy* gewählt. Allerdings wurde dabei nicht getestet, ob der optimale *mtry* Wert im Vergleich zum Defaultwert zu einer signifikanten Verbesserung der Accuracy beiträgt. Damit stellt der Defaultwert jedoch nicht in allen Fällen die beste Wahl dar.
Ein möglicher Indikator für die *mtry*-Wahl ist laut Bernard et al. (2009) die Anzahl an relevanten Variablen innerhalb der Daten. Der Parameter *mtry* wirkt als Trade-Off zwischen der Performance und der Diversität der einzelnen Bäume eines Random Forests. Existieren nur wenige relevante Variablen so sinkt die Performance durch die Randomisierung im Splitprozess enorm, was den Trade-Off schwächt. Dagegen bewirken viele stark relevante Variablen, dass sich die Prädiktionen der einzelnen Bäume sehr ähneln und der Randomisierungseffekt bei der Splitwahl nachlässt. Je weniger relevante Variablen also, desto größer fällt der optimale Wert für *mtry* aus, womit die Wahrscheinlichkeit steigt,

dass im Splitprozess die unwichtigen Variablen herausgefiltert werden.

Auch Díaz-Uriarte und de Andrés (2006) untersuchen den Effekt verschiedener Hyperparamter des Random Forests auf den OOB-Fehler. Dabei liegt der Fokus auf Microarray Daten, die in Genexpressionsstudien vorliegen. Meist wird dabei versucht, aus der Vielzahl von Genen diejenigen wenigen zu identifizieren, die einen Zusammenhang mit einem Response, zum Beispiel einem bestimmtem Krankheitsbild, aufweisen. Genexpressionsdaten sind sehr speziell, weswegen viele Standardmethoden oft nicht anwendbar sind. Zum einen kann die große Anzahl an Noise Variablen problematisch sein und zum anderen bestehen diese Daten meist aus deutlich mehr Variablen als Beobachtungen, welche oft auch miteinander interagieren oder korreliert sind. All diese Besonderheiten können jedoch von einem Random Forest berücksichtigt werden, womit er auch hier als Klassifikationsmethode eingesetzt werden kann. Die Variablenwichtigkeit bietet hierbei beispielsweise eine Möglichkeit die relevanten Genvariablen zu ermitteln. Auch Díaz-Uriarte und de Andrés (2006) beobachten, dass der Defaultwert für *mtry* in Bezug auf die OOB Fehlerrate oft, jedoch nicht immer, eine gute Wahl darstellt. Neben der Anwendung des Random Forests auf vier reale Datensätze wurde auch eine sehr umfangreiche Simulationsstudie durchgeführt. Hierbei fallen die Simulationsdatensätze mit wenigen relevanten Genen auf, da bei diesen ein steigendes *mtry* die Fehlerrate etwas sinken lässt. Stomit stützen also auch diese Genexpressionsdaten die vorab genannten Thesen von Bernard et al. (2009).

Ähnliche Beobachtungen schildern auch Genuer et al. (2008), die sowohl für Regressions- als auch für Klassifikationsforests zwischen Standardproblemen ($n >> p$) und hochdimensionalen Problemen ($n << p$) unterscheiden und den OOB-Fehler für unterschiedliche *mtry*-Werte betrachten.
Für die untersuchten Standard-Regressionsdatensätze ist der Defaultwert für *mtry* oft nicht optimal, besonders wenn $\lfloor p/3 \rfloor = 1$ gilt. Für die simulierten hochdimensionalen Regressionsdatensätze sinkt der OOB-Fehler mit steigendem *mtry*, weshalb der Defaultwert für *mtry* auch hier meist nicht optimal gewählt ist.
Im Vergleich dazu liefert der *mtry*-Default für Standard-Klassifikationsdatensätze nahezu minimale OOB-Fehler, was allerdings nicht für hochdimensionale Klassifikationsdatensätze gilt. Hierbei raten auch Genuer et al. (2008) *mtry* meist deutlich größer als $\left\lfloor \sqrt{p} \right\rfloor$ zu wählen, um bei einer hohen Anzahl an Variablen p mit einer größeren Wahrscheinlichkeit die relevanten Variablen im Splitprozess auszuwählen.

3 Simulationsstudie

Im Weiteren soll untersucht werden, ob sich die in Kapitel 2.3.3 vorgestellten Beobachtungen auch in einer Simulationsstudie nachbilden und erweitern lassen. Die Grundidee des Simulationsaufbaus stammt von Hapfelmeier et al. (2012) und wurde in abgewandelter Form übernommen. Dabei ist vor allem von Interesse, wie der optimale *mtry* Wert von verschiedenen Eigenschaften eines Datensatzes abhängt.
Keine der in Kapitel 2.3.3 vorgestellten Studien verwendet Datensätze, die sowohl für Regressions- als auch für Klassifikationsmodelle vergleichbar sind. Daher beruht das Simulationsdesign in dieser Arbeit für beide Modellarten auf einem sehr ähnlichen Prinzip, welches nun detailliert beschrieben wird. Abschließend werden die Ergebnisse dieser Simulationen vorgestellt.

3.1 Simulationsdesign

3.1.1 Datensätze

Wie bereits erwähnt wurden sowohl Regressionsmodelle wie auch Klassifikationsmodelle untersucht. Die benötigten Datensätze unterscheiden sich hier lediglich in der Definition des Responses, die Kovariablen werden für beide Modelle auf die gleiche Art und Weise generiert.

Ein Datensatz besteht aus einem Response $\mathbf{Y}$ und p Kovariablen $\mathbf{X}_1, \ldots, \mathbf{X}_p$ für insgesamt N Beobachtungen. Diese Kovariablen werden zufällig aus einer multivariaten Normalverteilung mit einem Erwartungswertvektor $\boldsymbol{\mu} = \mathbf{0}$ der Länge p und Kovarianzmatrix $\boldsymbol{\Sigma}$ gezogen, es gilt also

$$(\mathbf{X}_1, \ldots, \mathbf{X}_p) \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}).$$

Jede der Variablen erhält eine Varianz gleich 1, womit alle Einträge auf der Diagonale der $p \times p$ Kovarianzmatrix $\boldsymbol{\Sigma}$ gleich 1 sind. Nach Hapfelmeier et al. (2012) entsprechen die jeweiligen Kovarianzen in diesem Fall den Korrelationen der Kovariablen, auf welche in Kapitel 3.1.3 näher eingegangen wird.
Der Einfluss jeder Kovariable auf den Response wird mithilfe eines Koeffizientenvektors

$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ bestimmt. Die verwendeten Spezifikationen können im Detail dem folgenden Kapitel 3.1.2 entnommen werden.

Ein stetiger Response für eine Beobachtung i wird daraufhin auf Basis eines linearen Modells mit den generierten Kovariablen $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ definiert:

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad \forall\ i = 1, \ldots, N. \tag{3.1}$$

$\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_N)$ entspricht dabei einem Vektor mit Fehlertermen, der einer Normalverteilung mit Erwartungswert 0 und Varianz 0.5 folgt.

Ein binärer Response für eine Beobachtung i mit den Ausprägungen 0 oder 1 wird dagegen auf Basis eines Logitmodells definiert:

$$\pi_i = P(Y = 1|\mathbf{X} = \mathbf{x}_i) = \frac{exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + exp(\mathbf{x}_i^T \boldsymbol{\beta})}, \text{ womit gilt}$$
$$y_i \sim B(\pi_i). \tag{3.2}$$

Der Response einer Beobachtung i kann demnach zufällig aus einer Bernoulliverteilung mit Wahrscheinlichkeit π_i gezogen werden.

3.1.2 Kovariableneinflüsse

Aus Kapitel 2.3.3 geht hervor, dass die Relevanz der einzelnen Kovariablen innerhalb eines Datensatzes einen starken Einfluss auf die Wahl von *mtry* hat. Durch die Anwendung realer Datensätze ist es Bernard et al. (2009) allerdings nicht möglich die Anzahl an tatsächlich relevanten Kovariablen exakt zu bestimmen und Genuer et al. (2008) stellen nur für hochdimensionale Daten den Bezug zwischen *mtry* und den relevanten Kovariablen her. Um diese These detaillierter zu untersuchen, wurde in dieser Arbeit die Anzahl an relevanten Kovariablen eines Datensatzes durch verschiedene Koeffizientenvektoren $\boldsymbol{\beta}$ gesteuert. Dabei induziert $\beta_j = 0$, $j \in \{1, \ldots, p\}$, dass die Variable $\mathbf{X}_j$ keinen Einfluss auf den Response hat. Insgesamt wurden sieben verschiedene Koeffizientenvektoren abhängig von der Anzahl an Variablen p definiert, welche in Tabelle 3.1 zusammengefasst sind.

Für $\boldsymbol{\beta}_1$ und $\boldsymbol{\beta}_2$ wurde eine feste Anzahl von ein bzw. zwei Kovariablen gewählt, die einen Einfluss auf den Response besitzen, alle restlichen Kovariablen sind für die Generierung des Responses irrelevant. Gleiches gilt auch für die Koeffizientenvektoren $\boldsymbol{\beta}_3$ und $\boldsymbol{\beta}_4$, wobei fünf relevante Kovariablen definiert wurden. Im Gegensatz zu $\boldsymbol{\beta}_4$ haben diese Kovariablen mit $\boldsymbol{\beta}_3$ allerdings nicht alle den gleichen Einfluss, sondern unterscheiden sich geringfügig, sodass zwei Kovariablen eine starke Relevanz aufweisen und drei eine moderate Relevanz.

Koeffizientenvektor $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$	Beschreibung
$\boldsymbol{\beta}_1 = (7, 0, \ldots, 0)$	Eine relevante und $p-1$ irrelevante Kovariablen.
$\boldsymbol{\beta}_2 = (7, 8, 0, \ldots, 0)$	Zwei relevante und $p-2$ irrelevante Kovariablen.
$\boldsymbol{\beta}_3 = (7, 7, 7, 20, 20, 0, \ldots, 0)$	Fünf Kovariablen mit unterschiedlicher Relevanz und $p-5$ irrelevante Kovariablen.
$\boldsymbol{\beta}_4 = (7, 7, 7, 7, 7, 0, \ldots, 0)$	Fünf Kovariablen mit gleicher Relevanz und $p-5$ irrelevante Kovariablen.
$\boldsymbol{\beta}_5 = (2, \ldots, 2, 3, \ldots, 3, 18, \ldots, 18)$	Viele Kovariablen mit geringer Relevanz und nur wenige stark relevante.
$\boldsymbol{\beta}_6 = (2, \ldots, 2, 15, \ldots, 15, 18, \ldots, 18)$	Viele Kovariablen mit starker Relevanz und nur wenige schwach relevante.
$\boldsymbol{\beta}_7 = (2, \ldots, 2, 3, \ldots, 3, 4, \ldots, 4)$	Alle Kovariablen mit ähnlicher Relevanz.

Tabelle 3.1: *Definition von sieben verschiedenen Koeffizientenvektoren für die Datengenerierung.*
Die Häufigkeiten der drei verschiedenen Koeffizientenausprägungen für $\boldsymbol{\beta}_5 - \boldsymbol{\beta}_7$ sind abhängig von der Anzahl an Variablen p. Dabei gilt für $p = 10$: 3, 4, 3; für $p = 20$: 5, 10, 5; und für $p = 50$: 13, 24, 13.

Demgegenüber stehen drei weitere Koeffizientenvektoren, für die jede einzelne Kovariable einen Beitrag zur Generierung des Responses leistet. Durch Anwendung von $\boldsymbol{\beta}_5$ existieren hauptsächlich Kovariablen mit ähnlich schwachen Einflussstärken und nur wenige stark relevante Kovariablen. Die Häufigkeit des Auftretens der Koeffizienten $2, 3$ und 18 unterscheidet sich dabei je nach Anzahl der insgesamt definierten Variablen. Gilt $p = 10$, so werden jeweils drei Kovariablen mit einer Einflussstärke von 2 und 18 definiert und vier Kovariablen erhalten eine Einflussstärke von 3. Für $p = 20$ bzw. $p = 50$ gelten ähnliche Verhältnisse: Die Einflussstärken 2 und 18 treten jeweils bei fünf bzw. 13 Kovariablen auf und die Einflussstärke 3 bei zehn bzw. 24 Kovariablen. Diese Häufigkeitsverhältnisse wurden auch für die drei verschiedenen Koeffizientenausprägungen von $\boldsymbol{\beta}_6$ und $\boldsymbol{\beta}_7$ angewendet.

Für $\boldsymbol{\beta}_6$ wurde die Einflussstärke 3 in $\boldsymbol{\beta}_5$ durch 15 ersetzt, womit vor allem stark relevante Kovariablen definiert werden und nur wenige, die kaum einen Einfluss auf den Response besitzen.

Mit $\boldsymbol{\beta}_7$ können Datensätze generiert werden, deren Kovariablen mit den Koeffizientenausprägungen 2, 3 und 4 alle einen ähnlich starken Einfluss auf den Response erkennen lassen.

Damit nimmt also die Anzahl an stark relevanten Kovariablen innerhalb der Koeffizientenvektoren von $\boldsymbol{\beta}_1$ bis $\boldsymbol{\beta}_7$ zu. Wobei für $p = 10$ eine einzige Ausnahme gilt, denn mit $\boldsymbol{\beta}_5$ werden zwei stark relevante Kovariablen weniger definiert als mit $\boldsymbol{\beta}_4$.

3.1.3 Korrelationsstrukturen

Einige Studien haben gezeigt (siehe dazu Kapitel 2.3.2), dass korrelierte Kovariablen einen erheblichen Einfluss auf den Splitprozess eines Random Forests haben können. Zudem treten auch in realen Datensätzen meist Korrelationsstrukturen auf. Deshalb wurde für die Simulationsdatensätze versucht, die von Gregorutti et al. (2016) beschriebenen Korrelationsfälle zu übernehmen. Jedoch sind die ersten beiden Fälle und der fünfte Fall für maximal drei Kovariablen ausgelegt, weswegen in dieser Arbeit nur die Fälle 3 und 4 berücksichtigt wurden. Zusätzlich wurden noch drei weitere Kovarianzmatrizen $\boldsymbol{\Sigma}$ definiert, wovon eine den Fall 4 umkehrt, und die zwei anderen an das Beispiel von Strobl et al. (2008) angelehnt sind. Die fünf betrachteten Korrelationsstrukturen für die Koeffizientenvektoren $\boldsymbol{\beta}_3$, $\boldsymbol{\beta}_5$. und $\boldsymbol{\beta}_7$ sind in Tabelle 3.2 zusammengefasst und werden im Folgenden ausführlicher beschrieben.

Kovarianz	Koeff.-Vektor	Beschreibung
$\boldsymbol{\Sigma}_1$	$\boldsymbol{\beta}_7$	Alle Kovariablen ähnlich relevant und blockkorreliert.
$\boldsymbol{\Sigma}_2$	$\boldsymbol{\beta}_5$	Nur die weniger relevanten Kovariablen mit Koeffizientenausprägung 2 und 3 sind blockkorreliert.
$\boldsymbol{\Sigma}_3$	$\boldsymbol{\beta}_5$	Nur die stark relevanten Kovariablen mit Koeffizientenausprägung 18 sind blockkorreliert.
$\boldsymbol{\Sigma}_4$	$\boldsymbol{\beta}_3$	Jeweils eine Kovariable mit Koeffizientenausprägung 20 und 0 ist blockkorreliert.
$\boldsymbol{\Sigma}_5$	$\boldsymbol{\beta}_3$	Jeweils eine Kovariable mit Koeffizientenausprägung 20, 7 und 0 ist blockkorreliert.

Tabelle 3.2: *Definition der verschiedenen Kovarianzstrukturen für die Datengenerierung mit korrelierten Kovariablen und den Koeffizientenvektoren $\boldsymbol{\beta}_3$, $\boldsymbol{\beta}_5$ und $\boldsymbol{\beta}_7$.*

Im Fall 3 von Gregorutti et al. (2016) besitzen alle Kovariablen den gleichen Einfluss auf den Response und sind untereinander mit Korrelation c blockkorreliert. Diese Vorgaben werden mit dem Koeffizientenvektor $\boldsymbol{\beta}_7 = (2, \ldots, 2, 3, \ldots, 3, 4, \ldots, 4)$ und der $p \times p$ Kovarianzmatrix

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & c & \cdots & c \\ c & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & c \\ c & \cdots & c & 1 \end{pmatrix},$$

bis auf die gering unterschiedlichen Koeffizientenausprägungen erfüllt. Die positive Korrelation zwischen den Kovariablen wird hierbei und im Weiteren mit c bezeichnet.
Im Fall 4 von Gregorutti et al. (2016) werden diejenigen Kovariablen mit den größten Koeffizientenausprägungen blockkorreliert und die restlichen relevanten Kovariablen nicht. Um diesen Fall umzukehren, müssen die weniger relevanten Kovariablen blockkorelliert sein. Diese beiden Korrelationsstrukturen können mit dem Koeffizientenvektor $\boldsymbol{\beta}_5$ und den Kovarianzmatrizen $\boldsymbol{\Sigma}_2$ und $\boldsymbol{\Sigma}_3$ realisiert werden. Dabei gilt für den umgekehrten Fall

$$\boldsymbol{\Sigma}_2 = \begin{array}{c} \boldsymbol{\beta}_5 \quad (2, \ \ldots, \ 3, \ 18, \ \ldots, \ 18) \\ \begin{pmatrix} 1 & \cdots & c & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ c & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 1 \end{pmatrix} \end{array} \tag{3.3}$$

und angelehnt an den Fall 4 kann $\boldsymbol{\Sigma}_3$ verwendet werden:

$$\boldsymbol{\Sigma}_3 = \begin{array}{c} \boldsymbol{\beta}_5 \quad (2, \ \ldots, \ 3, \ 18, \ \ldots, \ 18) \\ \begin{pmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & c \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & c & \cdots & 1 \end{pmatrix} \end{array}. \tag{3.4}$$

Strobl et al. (2008) definieren einen Simulationsdatensatz für den stark relevante, weniger relevante und irrelevante Kovariablen blockkorreliert werden. Zur Rekonstruktion dieser Variante bietet sich also $\boldsymbol{\beta}_3 = (7, 7, 7, 20, 20, 0, \ldots, 0)$ an. Um zunächst den tatsächlichen Einfluss der Korrelation einer stark relevanten und irrelevanten Kovariable zu untersuchen, wird mit der Kovarianzmatrix

$$\boldsymbol{\Sigma}_4 = \begin{array}{c} \boldsymbol{\beta}_3 \quad (7, \ 7, \ 7, \ 20, \ 20, \ 0, \ 0, \ \ldots, \ 0) \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & \vdots & \cdots & \vdots \\ 0 & 0 & 1 & 0 & 0 & 0 & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & 1 & 0 & c & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & 0 & 1 & 0 & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & c & 0 & 1 & 0 & \cdots & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 0 & \cdots & 1 \end{pmatrix} \end{array} \tag{3.5}$$

nur eine stark relevante mit einer irrelevanten Kovariable korreliert. Dabei wird im Koeffizientenvektor jeweils die erste Variable mit der entsprechenden Ausprägung als korrelierte Kovariable gewählt. Aber auch die Blockkorellation von Strobl et al. (2008) kann auf gleiche Weise mit

$$\boldsymbol{\beta}_3 \quad (7, \quad 7, \quad 7, \quad 20, \quad 20, \quad 0, \quad 0, \quad \ldots, \quad 0)$$

$$\boldsymbol{\Sigma}_5 = \begin{pmatrix} 1 & 0 & 0 & c & 0 & c & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & \vdots & \cdots & \vdots \\ 0 & 0 & 1 & 0 & 0 & 0 & \vdots & \cdots & \vdots \\ c & 0 & 0 & 1 & 0 & c & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & 0 & 1 & 0 & \vdots & \cdots & \vdots \\ c & 0 & 0 & c & 0 & 1 & 0 & \cdots & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 0 & \cdots & 1 \end{pmatrix} \tag{3.6}$$

simuliert werden.

Die beschriebenen fünf Kovarianzmatrizen sind sehr verschieden, weswegen die Möglichkeit besteht, dass sie keine eindeutigen Interpretationen in Bezug auf den Einfluss der Korrelationen auf *mtry* zulassen. Daher wurden zusätzlich noch weitere Kovarianzmatrizen für die Koeffizientenvektoren mit einer ($\boldsymbol{\beta}_1$) bzw. fünf ($\boldsymbol{\beta}_4$) relevanten Kovariablen definiert, welche in Tabelle 3.3 aufgelistet sind. Mit diesen Kovarianzmatrizen kann unter anderem explizit untersucht werden, welche Auswirkungen die Blockkorrelation von irrelevanten Kovariablen hat.

Kovarianz	Koeff.-Vektor	Beschreibung
$\boldsymbol{\Sigma}_{6\_a}$	$\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_4$	Die relevanten und eine bestimmte Anzahl an a irrelevanten Kovariablen sind blockkorreliert.
$\boldsymbol{\Sigma}_7$	$\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_4$	Nur die irrelevanten Kovariablen sind blockkorreliert.
$\boldsymbol{\Sigma}_8$	$\boldsymbol{\beta}_4$	Die Hälfte der relevanten und die Hälfte der irrelevanten Kovariablen sind blockkorreliert.

Tabelle 3.3: *Definition der verschiedenen Kovarianzstrukturen für die Datengenerierung mit korrelierten Kovariablen und den Koeffizientenvektoren $\boldsymbol{\beta}_1$ und $\boldsymbol{\beta}_4$.*

Mit $\boldsymbol{\Sigma}_{6\_a}$ wird neben den relevanten Kovariablen in $\boldsymbol{\beta}_1$ und $\boldsymbol{\beta}_4$ zusätzlich eine bestimmte Anzahl an a irrelevanten Kovariablen blockkorreliert. Für a kann jeder ganzzahlige Wert größer oder gleich 0 gewählt werden. Zum Beispiel ist $\boldsymbol{\Sigma}_{6\_2}$ mit $\boldsymbol{\beta}_1$ folgendermaßen

definiert:

$$\boldsymbol{\Sigma}_{6\_2} = \begin{array}{c} \boldsymbol{\beta}_1 \quad (7, \quad 0, \quad 0, \quad 0, \quad \ldots, \quad 0) \\ \begin{pmatrix} 1 & c & c & 0 & \cdots & 0 \\ c & 1 & c & \vdots & \ddots & \vdots \\ c & c & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 1 \end{pmatrix} \end{array}. \tag{3.7}$$

Werden nur die irrelevanten Kovariablen blockkorreliert, ergibt sich auch für $\boldsymbol{\beta}_1$ und $\boldsymbol{\beta}_4$ mit der Kovarianzmatrix $\boldsymbol{\Sigma}_7$ die gleiche Darstellungsweise wie bereits für $\boldsymbol{\Sigma}_3$ in Definition (3.4), allerdings mit einem geringeren Anteil an unkorrelierten Variablen.
Die letzte Kovarianzmatrix $\boldsymbol{\Sigma}_8$ ist ähnlich zu $\boldsymbol{\Sigma}_5$, bei der eine irrelevante Kovariable zusätzlich zu zwei unterschiedlich relevanten Kovariablen blockkorreliert wird. Für $\boldsymbol{\Sigma}_8$ werden jedoch zwei gleich relevante Kovariablen und die Hälfte aller irrelevanten Kovariablen des Koeffizientenvektors $\boldsymbol{\beta}_4$ blockkorreliert. Entspricht die Hälfte einer ungeraden Zahl, wird die nächstkleinere ganze Zahl angewendet, womit sich $\boldsymbol{\Sigma}_8$ für zehn Kovariablen beispielsweise zu

$$\boldsymbol{\Sigma}_8 = \begin{array}{c} \boldsymbol{\beta}_4 \quad (7, \quad 7, \quad 7, \quad 7, \quad 7, \quad 0, \quad 0, \quad 0, \quad 0, \quad 0) \\ \begin{pmatrix} 1 & c & 0 & 0 & 0 & c & c & 0 & 0 & 0 \\ c & 1 & 0 & 0 & 0 & c & c & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & \vdots & \vdots & \vdots \\ c & c & 0 & 0 & 0 & 1 & c & 0 & \vdots & \vdots \\ c & c & 0 & 0 & 0 & c & 1 & 0 & \vdots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & 1 & 0 & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & 1 & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & 1 \end{pmatrix} \end{array} \tag{3.8}$$

ergibt.

3.1.4 Implementierung

Alle Analysen wurden mit der statistischen Software **R**, Version 3.2.3, durchgeführt (R Core Team, 2015). Die Datensätze wurden wie vorab beschrieben generiert. Dabei variiert die Anzahl an Beobachtungen N zwischen 500 und 1000, wobei anzumerken ist, dass durch eine zu geringe Anzahl an Beobachtungen, wie zum Beispiel $N = 100$, die definierten Kovarianzstrukturen nicht ausreichend exakt eingehalten werden können. Die Anzahl an Variablen p wurde auf 10, 20 und 50 festgelegt. Außerdem wurden für jedes Szenario 500 Datensätze erstellt und die Auswertungen dieser jeweils gemittelt. Ein Szenario ist dabei durch die Responseart, die Anzahl an Beobachtungen N, die Anzahl an Kovariablen p, den Koeffizientenvektor $\boldsymbol{\beta}$ und die Kovarianzmatrix $\boldsymbol{\Sigma}$ mit entsprechender Korrelation c definiert.
Es wurde darauf verzichtet hochdimensionale Daten mit $n << p$ zu definieren, da Díaz-Uriarte und de Andrés (2006) bereits eine detaillierte Simulationsstudie dazu durchgeführt haben, bei der ebenfalls der OOB-Fehler in Abhängigkeit von *mtry* dokumentiert ist.

Charakteristik	Ausprägungen	Bedingung
Responseart	metrisch, binär	
Koeffizientenvektor $\boldsymbol{\beta}$	$\boldsymbol{\beta}_1$ - $\boldsymbol{\beta}_7$	
Anzahl an Kovariablen p	$10, 20, 50$	
Anzahl an Beobachtungen N	$500, 1000$	
Kovarianzmatrix $\boldsymbol{\Sigma}$	$\boldsymbol{\Sigma}_1$	$\boldsymbol{\beta}_7$
	$\boldsymbol{\Sigma}_2$, $\boldsymbol{\Sigma}_3$	$\boldsymbol{\beta}_5$
	$\boldsymbol{\Sigma}_4$, $\boldsymbol{\Sigma}_5$	$\boldsymbol{\beta}_3$
	$\boldsymbol{\Sigma}_{6\_0}$	$\boldsymbol{\beta}_4$
	$\boldsymbol{\Sigma}_{6\_2}$	$\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_4$
	$\boldsymbol{\Sigma}_{6\_4}$, $\boldsymbol{\Sigma}_{6\_7}$	$\boldsymbol{\beta}_1$, $p = 10$
	$\boldsymbol{\Sigma}_{6\_9}$, $\boldsymbol{\Sigma}_{6\_15}$	$\boldsymbol{\beta}_1$, $p = 20$
	$\boldsymbol{\Sigma}_{6\_24}$, $\boldsymbol{\Sigma}_{6\_39}$	$\boldsymbol{\beta}_1$, $p = 50$
	$\boldsymbol{\Sigma}_7$	$\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_4$
	$\boldsymbol{\Sigma}_8$	$\boldsymbol{\beta}_4$
Korrelation c	$0, 0.3, 0.9$	
	0.6	$\boldsymbol{\Sigma}_6$, $\boldsymbol{\Sigma}_7$, $\boldsymbol{\Sigma}_8$

Tabelle 3.4: *Charakteristiken eines Szenarios und die gewählten Ausprägungen für die durchgeführte Simulationsstudie. Einige Ausprägungen für die Kovarianzmatrizen und Korrelationen wurden nur für bestimmte Szenarien verwendet, diese sind in der letzten Spalte gekennzeichnet.*

Tabelle 3.4 fasst die gewählten Charakteristiken der Szenarien zusammen und zusätzlich gibt Tabelle 3.5 Aufschluss über die Anzahl an Szenarien für eine Responseart unter verschiedenen Bedingungen. So wurden zum Beispiel mit allen sieben Koeffizientenvektoren aus Tabelle 3.1 Datensätze ohne Korrelation zwischen den Kovariablen ermittelt, was für jede Responseart zu 42 verschiedenen Szenarien für die Bedingung $c = 0$ führt.

Bedingung	# Ausprägungen je Charakteristik					# Szenarien
	p	N	$\boldsymbol{\beta}$	$\boldsymbol{\beta} + \boldsymbol{\Sigma}$	c	
$c = 0$	3	2	7	-	-	42
$\boldsymbol{\Sigma}_1$ - $\boldsymbol{\Sigma}_5$	3	2	-	5	2	60
$\boldsymbol{\Sigma}_6$ - $\boldsymbol{\Sigma}_8$	3	2	-	8	3	144
Gesamtanzahl Szenarien für eine Responseart						246

Tabelle 3.5: *Anzahl an Szenarien für eine Responseart unter verschiedenen Bedingungen der Charakteristiken.*

Für die Kovarianzmatrizen und deren entsprechende Korrelationen wurden etwas komplexe Kombinationen der verschiedenen Ausprägungen angewendet, welche im Weiteren zwar angesprochen, deren Bedeutungen jedoch erst bei den Auswertungen in Kapitel 3.2 deutlich werden.
Die Korrelation c aller Kovarianzmatrizen wurde auf die Werte 0.3 und 0.9 festgelegt, wobei für $\boldsymbol{\Sigma}_6$ bis $\boldsymbol{\Sigma}_8$ noch zusätzlich $c = 0.6$ hinzugenommen wurde. Wie bereits im vorherigen Kapitel angesprochen, wurde nicht jede Kovarianzmatrix auf jeden Koeffizientenvektor angewendet. Um nicht nochmals auf jede einzelne dieser Kombinationen einzugehen, sei auf Tabelle 3.4 verwiesen. Dabei stellt die Kovarianzmatrix $\boldsymbol{\Sigma}_{6\_a}$ allerdings eine Besonderheit dar, da mit ihr nicht nur eine feste Anzahl an irrelevanten Kovariablen zusätzlich zu den relevanten Kovariablen blockkorreliert wird, sondern diese Anzahl auch abhängig von p gewählt wurde. So werden hierbei immer 50 bzw. 80 Prozent aller irrelevanten Kovariablen für den Koeffizientenvektor $\boldsymbol{\beta}_1$ zusätzlich blockkorreliert. Dies muss berücksichtigt werden, wenn die Anzahl an Szenarien für $\boldsymbol{\Sigma}_6$ - $\boldsymbol{\Sigma}_8$ in Tabelle 3.5 bestimmt wird. Insgesamt sind acht Kombinationen aus $\boldsymbol{\beta} + \boldsymbol{\Sigma}$ für jedes p definiert worden, da die Kovarianzmatrizen $\boldsymbol{\Sigma}_{6\_4}$ - $\boldsymbol{\Sigma}_{6\_39}$ für die einzelnen p nur jeweils zwei Ausprägungen darstellen. Damit wurden für eine Responseart 144 verschiedene Szenarien mit den Kovarianzmatrizen $\boldsymbol{\Sigma}_6$ - $\boldsymbol{\Sigma}_8$ erstellt. Dies führt zu insgesamt 246 verschiedenen Szenarien für eine Responseart.

Aufgrund der hohen Anzahl an zu untersuchenden Datensätzen wurden die Berechnungen parallel auf einem Server durchgeführt, wofür das Package `parallelMap` (Bischl und Lang, 2015, Version 1.3) verwendet wurde. Der kombinierte multiple rekursive Zufallszahlengenerator von L'Ecuyer (1999) stellt dabei während der Seedspezifikation die Reproduzierbarkeit der Ergebnisse sicher. Ergänzend dazu ist es mit dem Package `mlr` (Bischl et al., 2016, Version 2.11) allgemein möglich, diverse maschinelle Lernverfahren (sogenannte *Learner*) in **R** zu nutzen. Die Verfahren können mit den darin bereitgestellten Funktionen auf relativ einfache Weise auch parallel implementiert und ausgewertet werden. Dabei übersteigt die Funktionalität des `mlr`-Packages oft die der zugrundeliegenden Basisfunktionen, so sind zum Beispiel auch Parametertuning oder Variablenselektion benutzerfreundlich umsetzbar.

Für die Regressions- und Klassifikationsforest wurden die zwei Learner *regr.ranger* und *classif.ranger* verwendet. Diese greifen auf das Package `ranger` (Wright und Ziegler, 2017, Version 0.8.0) zu, welches die schnellste und speicherplatzeffizienteste Implementierung eines Random Forests bereitstellt. Es wurden damit Random Forests mit 500 Bäumen gebildet, da Probst und Boulesteix (2017) gezeigt haben, dass sich die Struktur des OOB-Fehlers mit einer höheren Anzahl an Bäumen nicht mehr beachtlich ändert. Für alle Parameter außer *mtry* wurden die Defaultwerte verwendet. Das bedeutet unter anderem, dass sich in jeder Node eines Baumes mindestens eine (Klassifikation) bzw. fünf (Regression) Beobachtungen befinden. Außerdem wird als Splitkriterium der Gini Index (Klassifikation) bzw. die minimale Varianz des Response in den entstehenden Unterräumen (Regression) angewendet. Für jeden *mtry* Wert zwischen 1 und p wurde daraufhin ein Random Forest gebildet und die mittlere OOB-Prädiktionsgüte aus den 500 Wiederholungen für jedes Szenario ermittelt. Diese Prädiktionsgüte kann dann in Abhängigkeit von *mtry* als Kurve dargestellt werden, welche im Weiteren als *OOB-Kurve* bezeichnet wird. Um eine möglichst glatte Schätzung der OOB-Prädiktionsgüte zu erzielen, ist es nötig die Anzahl der Wiederholungen ausreichend groß zu wählen. In Abbildung A.1 kann beispielhaft für ein Szenario der Kurvenverlauf bei 50, 500 und 1000 Wiederholungen verglichen werden. Damit wurde überprüft, dass 500 Wiederholungen eine gute Wahl sind, denn mit 1000 Wiederholungen ergibt sich kein glatterer Kurvenverlauf.

Wie schon in Kapitel 2.3.1 angesprochen, kann die OOB-Prädiktionsgüte eines Random Forests mit verschiedenen Maßen bestimmt werden. Das Package *mlr* hat neben den in Kapitel 2.1 vorgestellten Modellgütemaßen eine Vielzahl an weiteren Maßen implementiert. Um herauszufinden, welches Maß unter diesen ein möglichst eindeutiges optimales *mtry* extrahieren kann, wurden verschiedene relevante Maße auf zwei Szenarien angewendet. Beispiele für die unterschiedlichen maßabhängigen Kurven der OOB-Prädiktionsgüte

liefern die Abbildungen in Anhang A.2. Dabei zeigt sich, dass sich die Performancemaße anhand ihrer OOB-Kurvenverläufe und den damit verbundenen optimalen *mtry* Werten in zwei Gruppen einteilen lassen. In der ersten Gruppe hat der Random Forest die optimalste Prädiktionsgüte meist für einen kleinen Wert von *mtry*. Das Optimum lässt sich bei diesen Performancemaßen durch einen relativ eindeutigen „Knick“ im Kurvenverlauf erkennen. Das *AUC* und $Kendall's\ \tau$ wurden repräsentativ für diese Maße ausgewählt und im Weiteren verwendet. In der zweiten Gruppe ist dieser Knick im Kurvenverlauf nicht mehr zu erkennen. Damit lässt sich das Optimum dieser Performancemaße auch nicht mehr so eindeutig anhand des Kurvenverlaufs bestimmen und die optimale Prädiktionsgüte liegt meist bei höheren *mtry* Werten als mit Maßen aus der ersten Gruppe. Der *Brier Score* und der *MSE* wurden repräsentativ für diese Gruppe von Performancemaßen ausgewählt und im Weiteren verwendet.
Welches Modellgütemaß verwendet wird, sollte spezifisch anhand der vorliegenden Daten entschieden werden, da keine allgemein gültigen Richtlinien existieren, in welchen Situationen welches Maß Verwendung finden sollte. Eine grobe Vorgabe zur Auswertung von Klassifikationsmethoden mit dem *BrierScore* oder dem *AUC* geben allerdings Hernández-Orallo et al. (2012): Demnach müssen für eine konkrete Wahl aus verschiedenen Performancemaßen zwei Faktoren berücksichtigt werden. Das sind zum einen die Einsatzbedingungen des Modells wie Missklassifizierungskosten und/oder Klassenverteilungen und zum anderen auf welche Art und Weise die Klassenzuteilung stattfindet(zum Beispiel ab welchem Schwellenwert der Prädiktionswahrscheinlichkeit eine Beobachtung $y = 1$ zugewiesen bekommt). Wenn allerdings keine Informationen über die Einsatzbedingungen zur Verfügung stehen, sollte auf das *AUC* zurückgegriffen werden. Oft sind diese Bedingungen jedoch voraussichtlich nach der Evaluierung, wenn das Modell im Einsatz ist und unter Umständen weiterentwickelt wird, bekannt. In diesem Fall und wenn zusätzlich davon ausgegangen werden kann, dass das Modell zuverlässige Prädiktionen ermittelt, wird der *Brier Score* empfohlen. Möglicherweise lassen sich die theoretischen Herleitungen von Hernández-Orallo et al. (2012) auch auf die in Kapitel 2.1.1 vorgestellten Modellgütemaße für die Regression übertragen. Dazu bedarf es allerdings einer genaueren Untersuchung, auf die im Rahmen dieser Arbeit nicht eingegangen werden kann.

Die für diese Arbeit generierten Datensätze berücksichtigen nur lineare Einflussgrößen (siehe Kapitel 3.1.2). Es wurde allerdings auch überprüft, ob sich das Verhalten der OOB-Prädiktionsgüte in Abhängigkeit von *mtry* ändert, wenn nicht-lineare Kovariablen aufgenommen werden. Hierfür gibt es verschiedene Möglichkeiten diese zu definieren, die zwei verwendeten Ansätze werden nun kurz vorgestellt.
Zum einen stellt das **R**-Package `mlbench` (Leisch und Dimitriadou, 2010, Version 2.1-1) eine Funktion zur Verfügung, mit der Daten für das sogenannte *Friedman 1* Regressions-

problem generiert werden können. Dabei werden zehn unabhängige Kovariablen aus einer Gleichverteilung auf dem Intervall $[0, 1]$ gezogen, wobei nur fünf von diesen den Response $\mathbf{y}$ auf folgende Weise definieren:

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

Dabei gilt für die Fehlerterme $\boldsymbol{\epsilon} \sim N(0, \sigma)$. Die damit generierten Datensätze beinhalten somit drei nicht-lineare, zwei lineare und fünf irrelevante Einflussgrößen. Sowohl σ, die Varianz der Fehlerterme, als auch die Anzahl an Beobachtungen werden vom Benutzer festgelegt.
Zum anderen wurden Datensätze ähnlich wie in Kapitel 3.1.2 definiert. Die generierten Kovariablen $\mathbf{x}_1, \ldots, \mathbf{x}_p$ wurden lediglich vor der Responsebildung mit einer Polynomfunktion 3. Grades transformiert. In den Modellgleichungen (3.1) und (3.2) werden somit die einzelnen Kovariablen $\mathbf{x}_i, i = 1, \ldots, p$, nur durch $(\mathbf{x}_i)^3$ ersetzt. Diese Transformation kann daher für Regressions- und auch für Klassifikationsdatensätze durchgeführt werden.
Anhang A.3 zeigt eine Auswahl der betrachteten OOB-Kurven für diese beiden Ansätze und ermöglicht einen Vergleich mit den analogen Szenarien auf Basis linearer Einflussgrößen. Dabei wird deutlich, dass sich der Verlauf und auch die optimalen *mtry* Werte nicht wesentlich unterscheiden, weswegen diese zusätzliche Eigenschaft innerhalb der Daten nicht weiter verfolgt wurde.

Das genaue Vorgehen für die Simulation eines Szenarios wird in Algorithmus 2 als Pseudocode für $W = 500$ Wiederholungen zusammengefasst.

Algorithmus 2: Erstellung einer OOB-Kurve für ein Szenario.

für *Response* $\in \{metrisch, binär\}$

- **für** $mtry = 1$ **bis** p
  - **für** $w = 1$ **bis** 500
    1. Generiere einen Datensatz entsprechend N, p, $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$ und c.
    2. Definiere einen *mlr*-Task abhängig vom Response.
    3. Bilde mit diesem Task einen entsprechenden Random Forest Learner.
    4. Trainiere auf Basis des Learners einen Random Forest mit *mtry* als Anzahl zufällig ausgewählter Variablen in jedem Baumsplit und extrahiere die OOB-Prädiktionen des Forest.
    5. Ermittle die gewünschten Performancemaße aus diesen OOB-Prädiktionen (*AUC* und *Brier Score* für Klassifikation bzw. $Kendall's\ \tau$ und MSE für Regression).
    6. Gebe die Performancemaße zurück.
  - Ermittle für jedes Performancemaß individuell den Mittelwert über alle Wiederholungen und die einzelnen *mtry* Werte.
- Stelle die erhaltenen mittleren Performancemaße als OOB-Kurve dar, dabei liegen die *mtry* Werte auf der x-Achse und die Performancemaße auf der y-Achse.

3.2 Ergebnisse

Im Folgenden werden die wichtigsten Erkenntnisse aus der Simulation der verschiedenen Szenarien dargestellt. Dabei werden die Ergebnisse aus den generierten Datensätzen mit metrischem Response getrennt von den Ergebnissen der Datensätze mit binärem Response betrachtet. Zudem wird innerhalb dieser Szenarien nach der Korrelationsstruktur der Kovariablen unterschieden.

3.2.1 Regression

Unkorrellierte Kovariablen

Die OOB-Kurven für die verschiedenen Regressionsszenarien wurden wie in Algorithmus 2 beschrieben ermittelt.
Beispielhaft zeigt Abbildung 3.1 den Verlauf der gemittelten Performancemaße $Kendall's\ \tau$ und MSE in Abhängigkeit des Parameters *mtry* für zwei verschiedene Szenarien, die sich

nur durch ihren Koeffizientenvektor unterscheiden. Hierbei werden diejenigen Random Forests verglichen, welche die wenigsten ($\boldsymbol{\beta}_1$) und die meisten ($\boldsymbol{\beta}_7$) relevanten Kovariablen in den Daten beinhalten. Die simulierten Datensätze der Szenarien bestehen aus $N = 1000$ Beobachtungen und $p = 10$ Kovariablen, wobei diese unkorreliert sind und damit die Kovarianzmatrix $\boldsymbol{\Sigma}$ als Einheitsmatrix mit Dimension 10×10 definiert ist. Die Quadrate markieren das Optimum der jeweiligen Maße und die gestrichelte graue Linie den Defaultwert für *mtry* bei dieser Konfiguration.

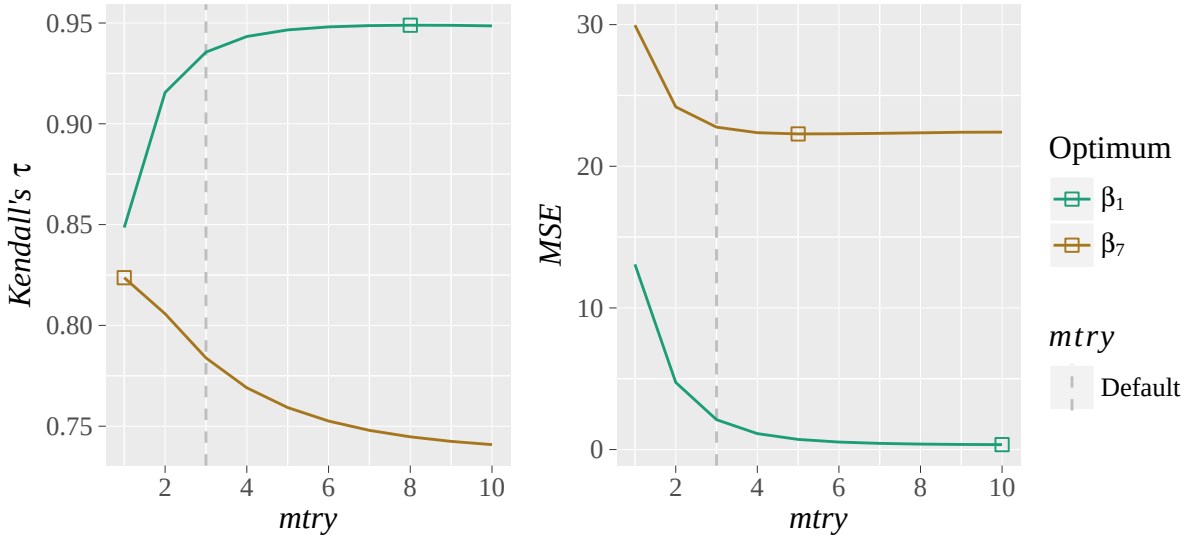


Abbildung 3.1: *OOB-Kurven der Performancemaße Kendall's τ und MSE für Regressionsszenarien mit* 1000 *Beobachtungen,* 10 *unkorrelierten Kovariablen und zwei verschiedene Koeffizientenvektoren $\boldsymbol{\beta}_1$ (eine relevante Kovariable) und $\boldsymbol{\beta}_7$ (nur relevante Kovariablen).*

Zunächst fällt auf, dass der Defaultwert von *mtry* in diesen beiden Beispielen nicht die beste Wahl darstellt, denn die *mtry* Werte an den Optima weichen teilweise deutlich davon ab. Außerdem ist gut zu erkennen, dass für einen bestimmten Koeffizientenvektor je nach Performancemaß auch verschiedene *mtry* Parameter gewählt werden sollten, um einen Random Forest mit optimaler Performance zu erhalten. Für $\boldsymbol{\beta}_1$ ist dieser Unterschied nur gering, nachdem *Kendall's* τ durch ein *mtry* von 8 optimal ausfällt und der *MSE* für $mtry = 10$. Für $\boldsymbol{\beta}_7$ fällt der Unterschied jedoch etwas größer aus, da *Kendall's* τ für $mtry = 1$ optimal ist und der *MSE* für $mtry = 5$.
Es muss jedoch auch erwähnt werden, dass zum Beispiel für *Kendall's* τ im Szenario mit $\boldsymbol{\beta}_1$ das Maximum nicht so eindeutig ausgeprägt ist wie mit $\boldsymbol{\beta}_7$. Mit $\boldsymbol{\beta}_1$ liefern alle Random Forests mit einem *mtry* größer oder gleich 6 eine sehr ähnliche Performance. Da Gleiches auch am Kurvenverlauf des *MSE* zu erkennen ist, ist es empfehlenswert den besten *mtry* Wert nicht am Kurvenoptimum zu wählen, sondern eine geringfügige Anpassung vorzunehmen.

Im Weiteren wird daher der kleinste $mtry$ Wert, für den das Perfomancemaß eine Abweichung von maximal 0.5 % zum Optimum hat, als optimales $mtry$ bezeichnet. Sei p die Anzahl an Kovariablen und P_{mtry} die Performance eines Random Forests mit Parameter $mtry$, $mtry \in [1, p]$. Um mithilfe dieser Werte das optimale $mtry$ zu bestimmen, wurden folgende zwei Schritte durchgeführt:

1. Je nachdem, ob das Performancemaß minimiert oder maximiert werden soll, wird für jede Ausprägung P_{mtry}, $mtry \in [1, p]$, folgendes Verhältnis ermittelt

$$v_{mtry} = \begin{cases} \dfrac{P_{opt}}{P_{mtry}}, & \text{falls } P_{opt} = min(P_1, \ldots, P_p) \\ \dfrac{P_{mtry}}{P_{opt}}, & \text{falls } P_{opt} = max(P_1, \ldots, P_p). \end{cases} \tag{3.9}$$

2. Daraufhin wird jenes optimale $mtry$ gesucht, das die Gleichung

$$mtry_{opt} = min\ \{mtry \mid 0.995 \leq v_{mtry} \leq 1\} \tag{3.10}$$

   erfüllt.

Falls demnach keines der Performancemaße eine Abweichung von 0.5 % einhält, wird der $mtry$ Wert am Optimum P_{opt} gewählt. Außerdem wird durch die Einschränkung, dass das Verhältnis v_{mtry} kleiner oder gleich 1 sein muss, sichergestellt, dass das optimale $mtry$ nicht größer gewählt wird als der $mtry$ Wert am Optimum der Kurve. Insgesamt liefert diese Anpassung natürlich etwas kleinere Werte für $mtry$ als das Optimum, jedoch birgt das den Vorteil, dass somit rechensparsamere Modelle bevorzugt werden, die trotzdem eine ähnlich gute Performance liefern. Der gewählte Schwellenwert für die untere Intervallgrenze von v_{mtry} sollte zwischen 0 und 1 liegen und nicht zu klein gewählt werden, damit die optimalen $mtry$ Werte nicht zu stark geschrumpft werden. Ein Wert von 0.995 erschien bei Betrachtung der resultierenden optimalen $mtry$ Werte für die Regressionsszenarien als sinnvoll.

Für die eben vorgestellten Szenarien ergeben sich nur leicht veränderte optimale $mtry$ Werte wie Abbildung 3.2 zeigt. Da der MSE im Vergleich zu $Kendall's\ \tau$ einen relativ großen Wertebereich besitzt, ist es hierbei selten der Fall, dass eine Abweichung kleiner 0.5 % vom Optimum eintritt. Deswegen bringt die Anpassung für den MSE nur für eine hohe Anzahl an Kovariablen deutliche Unterschiede im optimalen $mtry$ mit sich. Die analogen Abbildungen der OOB-Kurven aller betrachteten Szenarien können im elektronischen Anhang, im Unterordner *„Zusätzliche_Grafiken“*, aufgerufen werden.

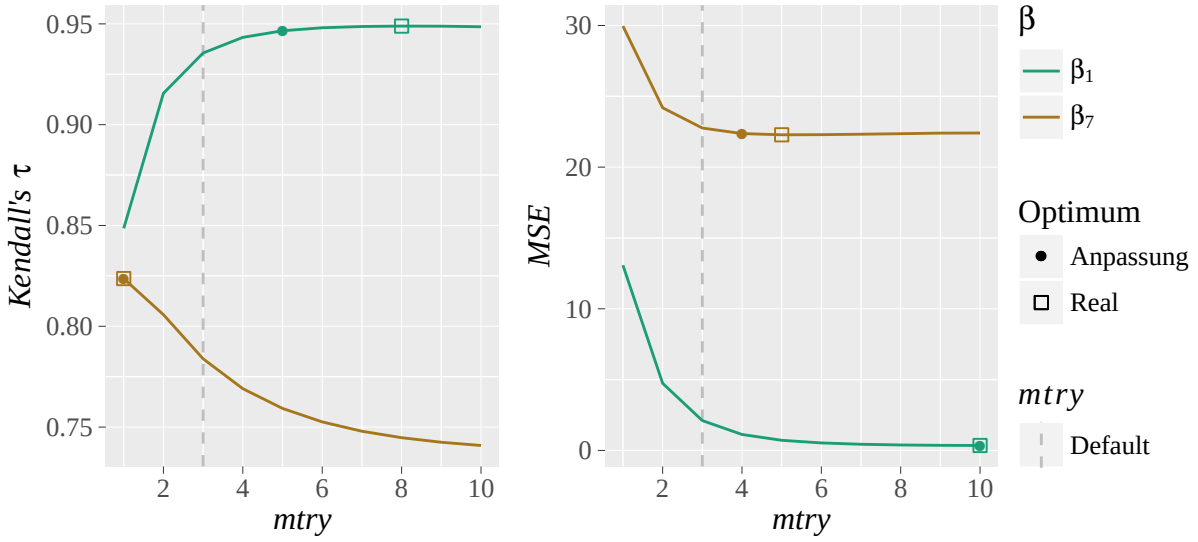
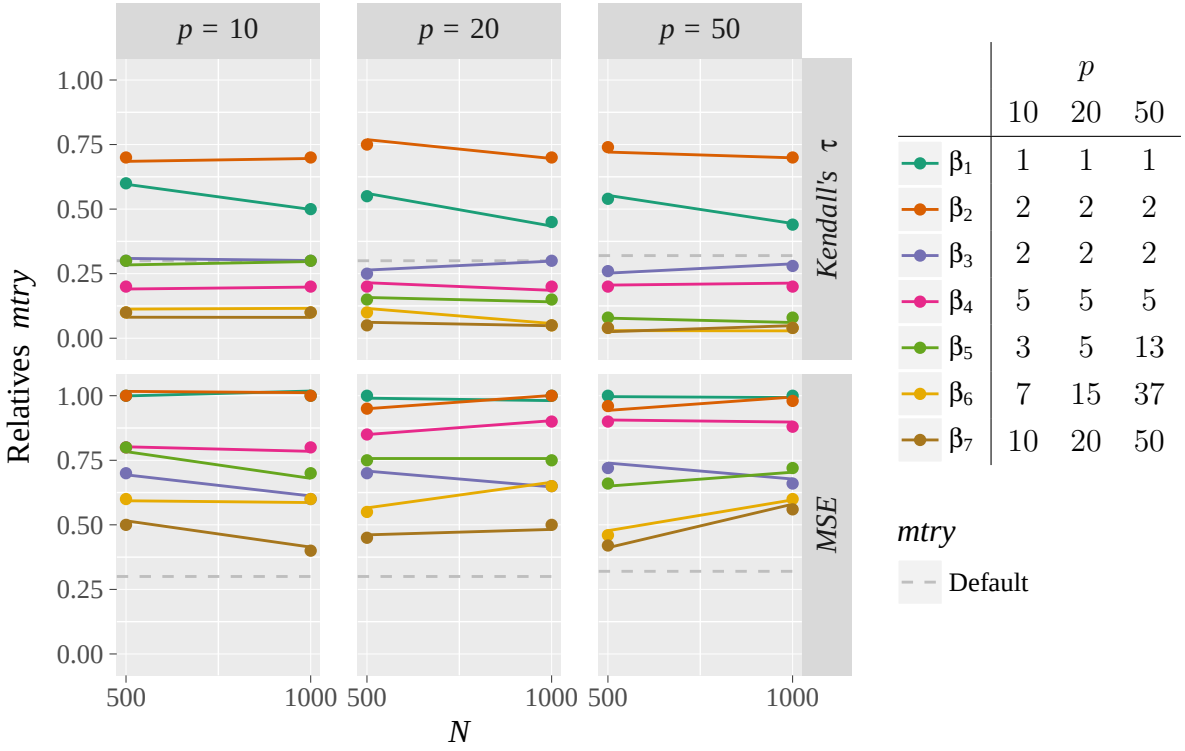


Abbildung 3.2: *OOB-Kurven der Performancemaße Kendall's τ und MSE für Regressionsszenarien mit* 1000 *Beobachtungen,* 10 *unkorrelierten Kovariablen und zwei verschiedenen Koeffizientenvektoren $\boldsymbol{\beta}_1$ (eine relevante Kovariable) und $\boldsymbol{\beta}_7$ (nur relevante Kovariablen). Zusätzlich sind die mtry Werte am Optimum und nach der Anpassung gekennzeichnet.*

Nun ist von Interesse, wie das optimale *mtry* für die verschiedenen Szenarien ohne korrelierte Kovariablen ausfällt. Abbildung 3.3 fasst diese zusammen und unterscheidet auch hier wieder zwischen *Kendall's* τ und dem MSE als Performancemaß.
Durch die vorab beschriebene Anpassung des optimalen *mtrys* konnte erreicht werden, dass auch für den MSE deutlichere Strukturen für die verschiedenen Koeffizientenvektoren zu erkennen sind und die optimalen *mtry* Werte nicht für fast alle Koeffizientenvektoren bei p liegen (Abbildung A.7 zeigt die entsprechenden *mtry* Werte an den Optima der OOB-Kurven).
Da die Anzahl an Kovariablen innerhalb der Szenarien zwischen 10, 20 und 50 variiert, ist hier zur besseren Vergleichbarkeit nicht der absolute *mtry* Wert auf der y-Achse angetragen, sondern der relative. Die Reihenfolge der Koeffizientenvektoren in der Legende entspricht der Anzahl an stark relevanten Variablen innerhalb der Datensätze mit $p = 50$, von den wenigsten am Anfang bis zu den meisten am Ende der Liste. Da sich für einige der Koeffizientenvektoren Änderungen im optimalen *mtry* über N überdecken, wurde den Verbindungslinien in der Höhe eine geringe zufällige Variation aufaddiert, wodurch diese besser sichtbar sind.



	p		
	10	20	50
β_1	1	1	1
β_2	2	2	2
β_3	2	2	2
β_4	5	5	5
β_5	3	5	13
β_6	7	15	37
β_7	10	20	50

Abbildung 3.3: *Optimale relative mtry Werte für alle 42 betrachteten Regressionsszenarien ohne korrelierte Kovariablen getrennt nach den verwendeten Performancemaßen und der Anzahl an Variablen p.*
Die Tabelle rechts gibt die Anzahl an stark relevanten Kovariablen für die einzelnen Koeffizientenvektoren in Abhängigkeit von p an.

Werden zum Beispiel die relativen optimalen *mtry* Werte der einzelnen Szenarien mit *Kendall's* τ verglichen wird deutlich, dass die *mtry* Werte über die Anzahl an Beobachtungen hinweg vergleichsweise konstant ausfallen. Allerdings ist die Anzahl an stark relevanten Kovariablen ein Einflussfaktor auf das optimale *mtry*. Die Reihenfolge der Koeffizientenvektoren in der Grafik entspricht fast exakt der Reihenfolge in der Legende, was bedeutet, je größer die Anzahl an stark relevanten Kovariablen ist, desto kleiner wird das relative optimale *mtry*.
Eine Ausnahmen stellt allerdings der Koeffizientenvektor $\boldsymbol{\beta}_1$ mit nur einer relevanten Kovariable dar. Für diesen ist das optimale *mtry* kleiner als für $\boldsymbol{\beta}_2$ mit zwei relevanten Kovariablen. Um diese Tatsache näher zu untersuchen, vergleicht Abbildung 3.4 den Verlauf der einzelnen OOB-Kurven der ersten vier Koeffizientenvektoren für *Kendall's* τ und den MSE. Hierbei ist zu erkennen, dass sich *Kendall's* τ über *mtry* hinweg zwischen $\boldsymbol{\beta}_1$ und $\boldsymbol{\beta}_2$ sehr ähnlich verhält. Erst wenn eine höhere Anzahl an relevanten Kovariablen berücksichtigt wird, wie mit $\boldsymbol{\beta}_3$ oder $\boldsymbol{\beta}_4$, ändert sich der Verlauf der Kurve, was zur Folge hat, dass sich ein eindeutigeres Optimum an einem geringeren *mtry* ausbildet. Der MSE zeigt dazu im Vergleich für $\boldsymbol{\beta}_1$ und $\boldsymbol{\beta}_2$ nur einen geringen Unterschied in den Kurvenverläufen, sodass daraus geschlossen werden kann, dass es für die Wahl von *mtry* nicht

von Bedeutung ist, ob nun ein oder zwei relevante Kovariablen existieren. Aus diesem Grund und weil die Rangfolge der beiden Vektoren $\boldsymbol{\beta}_1$ und $\boldsymbol{\beta}_2$ die einzige Ausnahme für *Kendall's* τ darstellt, kann trotzdem davon ausgegangen werden, dass je größer die Anzahl an stark relevanten Kovariablen ist, desto kleiner ist auch das optimale *mtry*.
Dies ist keine überraschende Erkenntnis, da bereits Bernard et al. (2009) empfohlen haben, *mtry* bei nur sehr wenigen relevanten Kovariablen höher zu setzen, um die Wahrscheinlichkeit zu steigern, dass auch die wenigen wichtigen Kovariablen im Splitprozess Berücksichtigung finden.
Jedoch sind nicht nur die stark relevanten Kovariablen von Bedeutung, sondern auch die weniger relevanten, wie $\boldsymbol{\beta}_2 = (7, 8, 0, \ldots, 0)$ und $\boldsymbol{\beta}_3 = (7, 7, 7, 20, 20, 0, \ldots, 0)$ zeigen. Durch drei zusätzliche, weniger relevante Kovariablen stellt sich für $\boldsymbol{\beta}_3$ ein deutlich geringerer optimaler *mtry* Wert heraus, als für $\boldsymbol{\beta}_2$.

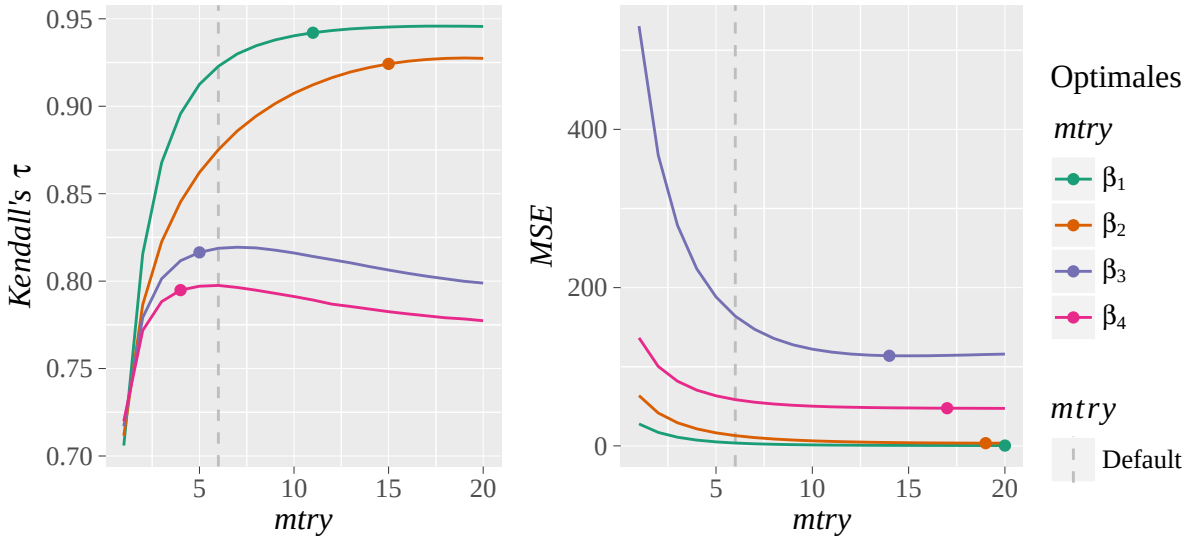


Abbildung 3.4: *OOB-Kurven der Performancemaße Kendall's* τ *und MSE für Regressionsszenarien mit* 500 *Beobachtungen,* 20 *unkorrelierten Kovariablen und den Koeffizientenvektoren* $\boldsymbol{\beta}_1$ *bis* $\boldsymbol{\beta}_4$.

Die ersten vier Koeffizientenvektoren besitzen über p hinweg alle die gleiche Anzahl an relevanten Kovariablen. Dabei lässt sich mit *Kendall's* τ als Performancemaß gut erkennen, dass der relative optimale *mtry* Wert über p hinweg sehr ähnlich ist. Denn ob nun, wie bei $\boldsymbol{\beta}_2$, zwei aus zehn Kovariablen relevant sind oder zwei aus 20 Kovariablen, führt in beiden Fällen zu einem relativen *mtry* von ca. 0.5. Das heißt, bei nur sehr wenigen relevanten Kovariablen innerhalb der Daten scheint anhand dieser Ergebnisse die absolute Anzahl an stark relevanten Kovariablen eine wichtige Rolle zu spielen und nicht unbedingt deren Anteil innerhalb der p Kovariablen.
Vergleichbares zeigt sich auch an $\boldsymbol{\beta}_5$ und $\boldsymbol{\beta}_6$: Je größer dabei die Anzahl an Variablen p, desto mehr stark relevante Kovariablen werden berücksichtigt, diese stellen allerdings

jeweils einen ähnlichen Anteil innerhalb der p Kovariablen dar. In Abbildung 3.3 ist für diese beiden Koeffizientenvektoren mit ähnlichem Anteil an relevanten Kovariablen das optimale relative *mtry* nicht konstant für alle p, sondern sinkt mit steigendem p und damit mit steigender Anzahl an stark relevanten Kovariablen.

Werden die optimalen *mtry* Werte für $Kendall's\ \tau$ und für den MSE verglichen, können auf Basis dieser Szenarien zwei Unterschiede ausgemacht werden: Zum einen liegt das optimale *mtry* mit dem MSE als Performancemaß immer über dem Defaultwert, was für $Kendall's\ \tau$ nicht gilt. Zum anderen variieren die *mtry* Werte mit dem MSE für einzelne Koeffizientenvektoren in Abhängigkeit der Anzahl an Beobachtungen. Diese Variationen belaufen sich allerdings auf maximal 10%, was absolut gesehen nur einer *mtry*-Änderung zwischen 1 ($p = 10$) und 5 ($p = 50$) entspricht. Diese kleinen Unterschiede zwischen verschiedenen N kommen für den MSE vor allem durch das Fehlen eines eindeutigen Optimums zustande, was bereits Abbildung 3.4 gezeigt hat. Damit sind auch leichte Abweichungen vom optimalen *mtry* für den MSE denkbar, mit denen sich trotzdem eine vergleichbare Modellperformance ergibt. Die fehlenden eindeutigen Optima für den MSE, können auch für die insgesamt vergleichsweise hohen optimale *mtry* Werte verantwortlich sein.
Wie auch schon mit $Kendall's\ \tau$ entspricht auch hierbei die Reihenfolge der Koeffizientenvektoren nahezu der Reihenfolge in der Legende, also der Anzahl an stark relevanten Kovariablen, wobei für $\boldsymbol{\beta}_3$, $\boldsymbol{\beta}_4$ und $\boldsymbol{\beta}_5$ einzelne Abweichungen zu erkennen sind.

Korrellierte Kovariablen mit $\boldsymbol{\Sigma}_1$ - $\boldsymbol{\Sigma}_5$

Wenn im Weiteren von Szenarien mit (block)korrelierten Kovariablen gesprochen wird, ist im Allgemeinen nicht gemeint, dass alle Kovariablen miteinander korreliert sind, sondern, dass eine Kovarianzmatrix ungleich der Einheitsmatrix definiert wurde. Da diese Korrelationen unter anderem auch Einfluss auf die Variablenwichtigkeit haben (siehe Kapitel 2.3.2) werden auch einige Grafiken dazu vorgestellt. Wie dabei die mittlere relative Variablenwichtigkeit bestimmt und die entsprechenden *mtry* Werte ausgewählt wurden beschreibt Anhang A.4.

Ähnlich zu den vorhergehenden Datensätzen mit unkorrelierten Kovariablen stellt Abbildung 3.5 die optimalen *mtry* Werte der Szenarien für die Koeffizientenvektoren $\boldsymbol{\beta}_3$, $\boldsymbol{\beta}_5$ und $\boldsymbol{\beta}_7$ mit den Kovarianzmatrizen $\boldsymbol{\Sigma}_1$ bis $\boldsymbol{\Sigma}_5$ dar. Da die Verläufe kaum von der Anzahl an Variablen p beeinflusst werden, sind diese beispielhaft nur für $p = 20$ dargestellt, die weiteren Abbildungen können dem Anhang A.6 entnommen werden. Mit diesen

Abbildungen kann zum einen untersucht werden, wie sich das optimale *mtry* für verschiedene Korrelationsstärken verhält und zum anderen aber auch, ob Unterschiede zwischen den Performancemaßen existieren. Die einzelnen Koeffizientenvektoren unterscheiden sind farblich und je nach verwendeter Kovarianzmatrix sind die Szenarien durch eine niedrigere Farbintensität von der Ausgangssituation ohne Korrelationen abgesetzt.

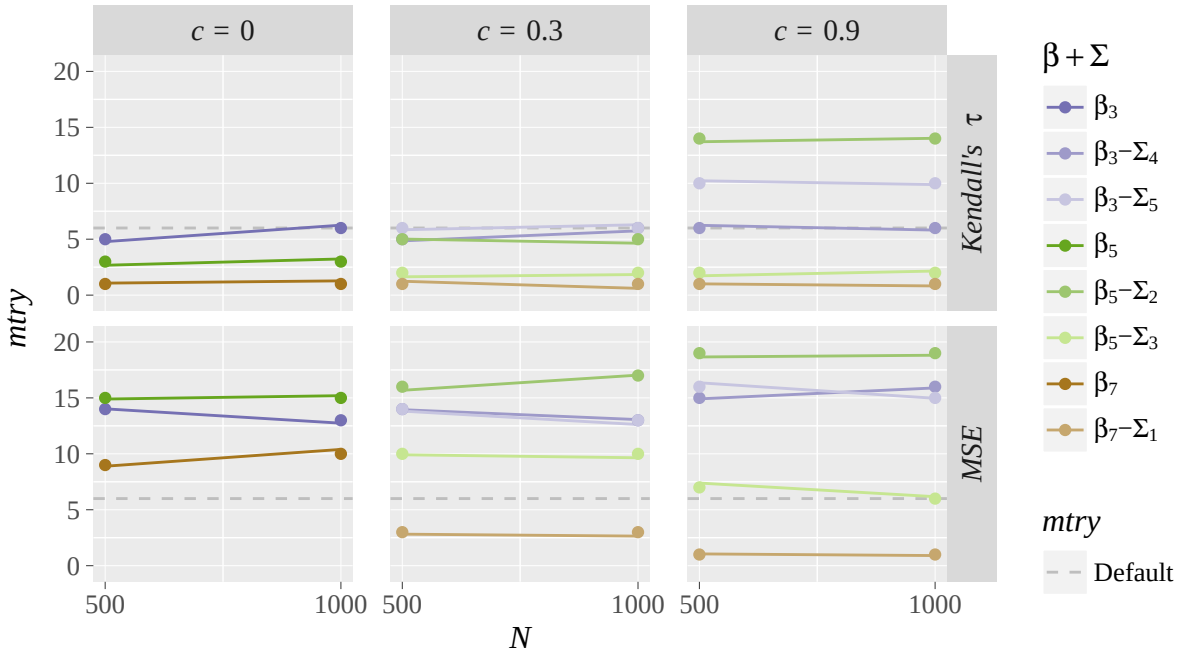


Abbildung 3.5: *Optimale mtry Werte für alle 20 betrachteten Regressionsszenarien mit $p = 20$, $\boldsymbol{\beta}_3$, $\boldsymbol{\beta}_5$ oder $\boldsymbol{\beta}_7$ und korrelierten Kovariablen getrennt nach den verwendeten Performancemaßen und Korrelationen c. In der linken Spalte sind zum Vergleich die optimalen mtry Werte für die analogen Szenarien ohne korrelierte Kovariablen angetragen.*
Die Definitionen der einzelnen Koffizientenvektoren und Kovarianzmatrizen sind in den Tabellen 3.1 und 3.2 zusammengefasst.

Beginnend mit dem Koeffizientenvektor $\boldsymbol{\beta}_3 = (7, 7, 7, 20, 20, 0, \ldots 0)$ ist für $c = 0.3$ mit keiner der beiden Kovarianzmatrizen und für keines der Performancemaße eine deutliche Veränderung des optimalen *mtry* zu erkennen (im Vergleich zum analogen Szenario mit unkorrelierten Kovariablen). Das ändert sich jedoch, wenn die Korrelation auf $c = 0.9$ ansteigt. Diese Korrelation bewirkt laut Strobl et al. (2008), dass vor allem für kleinere *mtry* die korrelierten Kovariablen als Splitkandidaten bevorzugt werden und dadurch die Variablenwichtigkeiten von weniger relevanten oder gar irrelevanten Kovariablen überschätzt werden. Wie in Abbildung 3.3 beobachtet, gilt für Szenarien mit unkorrelierten Kovariablen: Je mehr relevante Kovariablen existieren, desto kleiner wird das optimale *mtry*. Wenn nun durch Korrelation innerhalb der Daten mehr Kovariablen als relevant erkannt werden, könnte vermutet werden, dass das optimale *mtry* im Vergleich zum unkorrelierten Szenario kleiner ausfällt. Dies bestätigt sich allerdings nicht. Denn für $\boldsymbol{\Sigma}_4$, wenn also nur

eine stark relevante und eine irrelevante Kovariable korreliert sind, bleibt das optimale relative *mtry* für *Kendall's* τ zwar noch konstant, steigt aber fast auf das Doppelte an, sobald noch eine weniger relevante Kovariable zusätzlich blockkorreliert wird ($\boldsymbol{\Sigma}_5$). Der *MSE* hingegen steigt für diese beiden Korrelationsstrukturen nur geringfügig an. Diese Beobachtungen widersprechen somit auch Strobl et al. (2008), die für das Szenario mit $\boldsymbol{\Sigma}_5$ und einem kleinen *mtry* eine höhere Prädiktionsgüte erwartet hätten. Möglicherweise lassen sich allerdings die Erkenntnisse aus genannter Studie nicht direkt auf Random Forests basierend auf CART-Entscheidungsbäume wie in dieser Arbeit anwenden.
Wie Abbildung 3.6 der Variablenwichtigkeiten für diese Szenarien zeigt, können tatsächlich Abweichungen für die Variablenwichtigkeiten der drei korrelierten Kovariablen (1, 4 und 6) im Vergleich zum unkorrelierten Szenario ausgemacht werden. Die Abweichungen der Variablen 1 und 6 können jedoch durch ein größeres *mtry* verringert werden. Warum erst die Hinzunahme einer weniger relevanten korrelierten Kovariable ($\boldsymbol{\Sigma}_5$) das optimale *mtry* ansteigen lässt, kann anhand der beschriebenen Szenarien und Random Forest Eigenschaften nicht sicher geklärt werden.

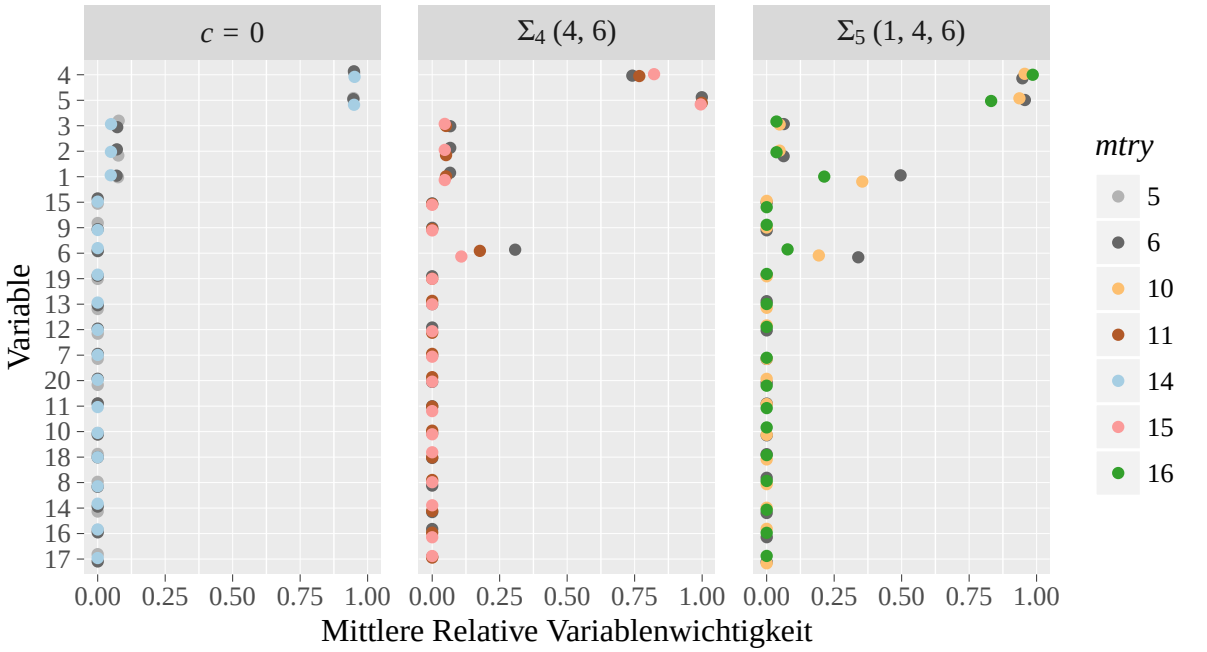


Abbildung 3.6: *Mittlere relative Permutation Importance über* 500 *Wiederholungen mit den Spezifikationen: Metrischer Response,* $\boldsymbol{\beta}_3 = (7, 7, 7, 20, 20, 0, \ldots, 0)$, $N = 500$, 20 *unkorrelierte Kovariablen bzw. Kovarianzmatrizen* $\boldsymbol{\Sigma}_4$ *und* $\boldsymbol{\Sigma}_5$ *mit* $c = 0.9$. *Die dabei jeweils blockkorrelierten Kovariablen sind im Titel gekennzeichnet. Die verschiedenen mtry Werte je Szenario entsprechen dem Defaultwert und den optimalen mtry Werten für die Performancemaße Kendall's* τ *und MSE.*

Auch für den Koeffizientenvektor $\boldsymbol{\beta}_5 = (2, \ldots, 2, 3, \ldots, 3, 18, \ldots, 18)$ wurden zwei verschiedene Kovarianzmatrizen verwendet. $\boldsymbol{\Sigma}_2$ blockkorreliert dabei die 15 weniger relevanten Ko-

variablen. Das hat zur Folge, dass bereits bei einer Korrelation von 0.3 ein etwas höherer *mtry* Wert (im Vergleich zum unkorreliertem Szenario) die optimale Modellperformance in Bezug auf *Kendall's* τ liefert. Dieser Effekt verstärkt sich sogar deutlich für $c = 0.9$. Ist dagegen nur eine Blockkorrelation der stark relevanten Kovariablen mit der Kovarianzmatrix $\mathbf{\Sigma}_3$ definiert, hat dies für *Kendall's* τ kaum Auswirkungen auf das optimale *mtry*, unabhängig von der Stärke der Korrelation. Da dies dem Fall 4 von Gregorutti et al. (2016) entspricht, wird erwartet, dass die Variablenwichtigkeit der unkorrelierten, weniger relevanten Kovariablen überschätzt werden könnte. Abbildung 3.7 vergleicht die zugehörigen Variablenwichtigkeiten für das unkorrelierte Szenario und die beiden Korrelationsszenarien. Entgegen den Erwartungen steigen die Variablenwichtigkeiten der weniger relevanten Kovariablen mit $\mathbf{\Sigma}_3$ für keines der betrachteten *mtry* an. Sind jedoch nur die weniger relevanten Kovariablen blockkorreliert ($\mathbf{\Sigma}_2$), ist für alle betrachteten *mtry* eine deutliche Überschätzung dieser Variablenwichtigkeiten zu erkennen, wobei diese Überschätzung für ein großes *mtry* am kleinsten ist. Dies könnte auch Grund für das etwas größere optimale *mtry* für dieses Szenario sein.

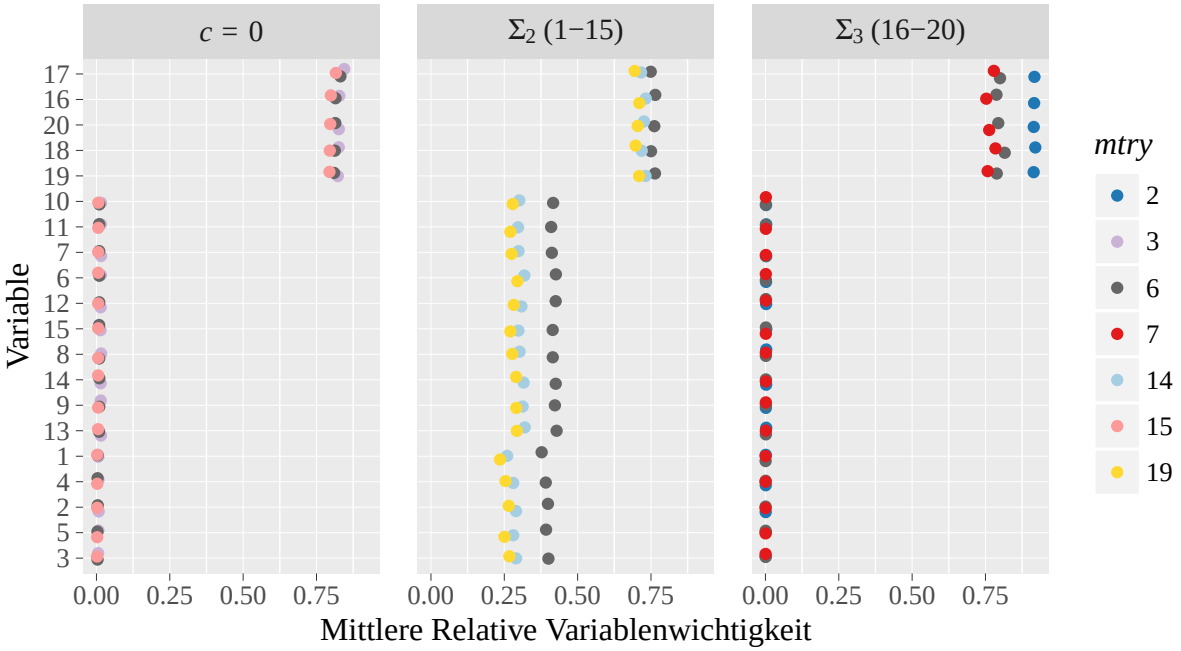


Abbildung 3.7: *Mittlere relative Permutation Importance über 500 Wiederholungen mit den Spezifikationen: metrischer Response,* $\boldsymbol{\beta}_5 = (2, \ldots, 2, 3, \ldots, 3, 18, \ldots, 18)$, $N = 500$, *20 unkorrelierte Kovariablen bzw. Kovarianzmatrizen* $\mathbf{\Sigma}_2$ *und* $\mathbf{\Sigma}_3$ *mit* $c = 0.9$. *Die dabei jeweils blockkorrelierten Kovariablen sind im Titel gekennzeichnet. Die verschiedenen mtry Werte je Szenario entsprechen dem Defaultwert und den optimalen mtry Werten für die Performancemaße Kendall's* τ *und MSE.*

Im Gegensatz zu *Kendall's* τ nimmt der optimale *mtry* Wert mit dem *MSE* für eine steigende Korrelation c ab, wenn nur die stark relevanten Kovariablen blockkorreliert sind

($\boldsymbol{\Sigma}_3$). In diesem Fall nähern sich die Variablenwichtigkeiten für das kleinste $mtry = 2$ am stärksten an die Variablenwichtigkeiten des unkorrelierten Szenarios an, was vielleicht das kleinere optimale $mtry$ für den MSE induziert. Wohingegen die Korrelation der weniger relevanten Kovariablen, wie auch schon mit $Kendall's\ \tau$, ebenfalls einen Anstieg des optimalen $mtry$ in Bezug auf den MSE verursacht. Dieses Verhalten führt auch dazu, dass sich die optimalen $mtry$ mit den beiden Performancemaßen für diesen Koeffizientenvektor und $c = 0.9$ nicht mehr so stark unterscheiden.

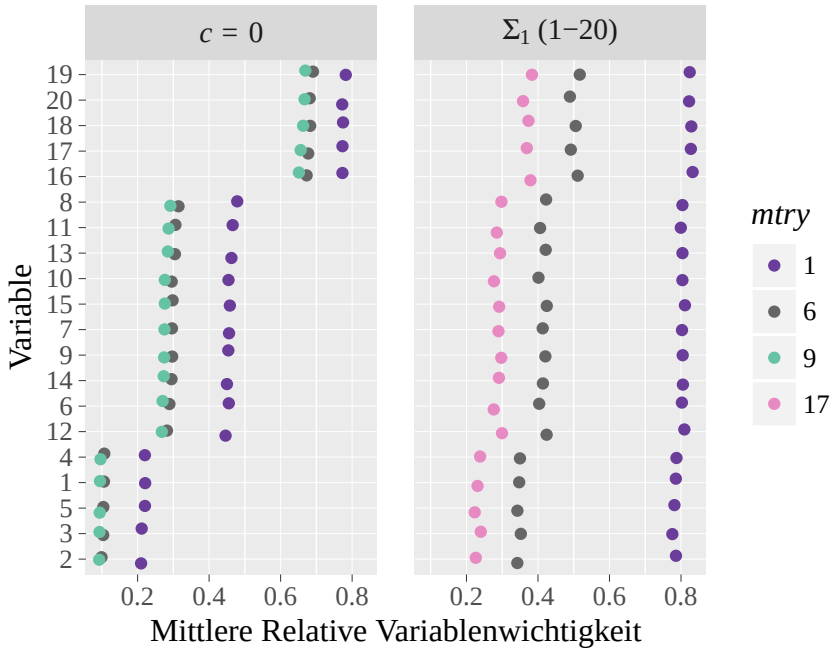


Abbildung 3.8: *Mittlere relative Permutation Importance über 500 Wiederholungen mit den Spezifikationen: metrischer Response, $\boldsymbol{\beta}_7 = (2, \ldots, 2, 3, \ldots, 3, 4, \ldots, 4)$, $N = 500$, 20 unkorrelierte Kovariablen bzw. Kovarianzmatrix $\boldsymbol{\Sigma}_1$ mit $c = 0.9$. Die dabei blockkorrelierten Kovariablen sind sind im Titel gekennzeichnet. Die verschiedenen mtry Werte je Szenario entsprechen dem Defaultwert und den optimalen mtry Werten für die Performancemaße $Kendall's\ \tau$ und MSE.*

Für den letzten betrachteten Koeffizientenvektor $\boldsymbol{\beta}_7 = (2, \ldots, 2, 3, \ldots, 3, 4, \ldots, 4)$ wurde nur eine Korrelationsstruktur mit der Kovarianzmatrix $\boldsymbol{\Sigma}_1$ berücksichtigt. Dabei werden alle Kovariablen, welche in diesem Fall eine ähnliche Einflussstärke besitzen, blockkorreliert. Durch die leicht verschiedenen Koeffizientenausprägungen wird nicht exakt der Fall 3 von Gregorutti et al. (2016) abgebildet, weswegen hier zu erwarten ist, dass die Variablenwichtigkeiten der Kovariablen mit Koeffizientenausprägungen 2 wie im Fall 4 etwas überschätzt werden. Gleichzeitig wird aber auch die Variablenwichtigkeit der Kovariablen mit Koeffizientenausprägungen 4 unterschätzt, wodurch sich die Auswahlhäufigkeiten aller Kovariablen angleichen. Siehe dazu auch Abbildung 3.8 der simulierten Variablenwichtigkeiten. Allerdings sollten diese Unterschiede in den Auswahlhäufigkeiten im Vergleich zum

unkorrelierten Szenario für die sehr ähnlichen Koeffizientenausprägungen keinen gravierenden Einfluss auf die Modellperformance haben, wodurch ein ähnliches optimales *mtry* erwartet wird. Für *Kendall's* τ trifft diese Theorie auch bei steigender Korrelation c zu, denn das optimale *mtry* bleibt konstant bei einem Wert von 1. Jedoch ist für den MSE mit steigendem c ein deutlich kleineres *mtry* als im unkorrelierten Szenario zu bevorzugen, wodurch sich bei hoher Korrelation in diesem Szenario wieder ähnliche optimale *mtry* Werte für beide Performancemaße ergeben.

Für die Unterschiede zwischen *Kendall's* τ und dem MSE muss allerdings auch berücksichtigt werden, dass sich mit *Kendall's* τ im Vergleich zum MSE bereits für die unkorrelierten Szenarien kleinere optimale *mtry* Werte ergeben, womit natürlich für dieses Performancemaß keine großen Veränderungen hinsichtlich eines noch kleineren *mtry* beobachtet werden können. Allerdings lassen sich die sinkenden optimalen *mtry* Werte mit dem MSE zum Beispiel für $\boldsymbol{\beta}_5$ und $\boldsymbol{\Sigma}_3$ oder $\boldsymbol{\beta}_7$ und $\boldsymbol{\Sigma}_1$ mit den betrachteten Random Forest Eigenschaften nicht mit Sicherheit erklären. Denn es gilt beispielsweise für $\boldsymbol{\beta}_5$ und $\boldsymbol{\Sigma}_3$ nicht, dass in diesem Szenario eine höhere Anzahl an relevanten Kovariablen erkannt wird, weswegen ein kleines *mtry* nachvollziehbar wäre.
Jedoch haben diese Analysen gezeigt, dass die Theorien von Gregorutti et al. (2016) und Strobl et al. (2008) eine guten Ansatz liefern und verschiedene Kovarianzstrukturen auch einen Einfluss auf das optimale *mtry* besitzen können. Möglicherweise existieren durch die Korrelation einiger Kovariablen noch weitere Effekte, die bisher nicht berücksichtigt wurden und anhand der vorliegenden Szenarien nicht deutlich werden.

Korrellierte Kovariablen mit $\boldsymbol{\Sigma}_6$ - $\boldsymbol{\Sigma}_8$

Für die Koeffizientenvektoren $\boldsymbol{\beta}_1 = (7, 0 \ldots, 0)$ und $\boldsymbol{\beta}_4 = (7, 7, 7, 7, 7, 0, \ldots, 0)$ wurden noch weitere Szenarien mit verschiedenen Kovarianzmatrizen definiert, um etwas strukturierter zu untersuchen, ob ein eindeutiger Effekt der Korrelation irrelevanter Kovariablen auf das optimale *mtry* nachgewiesen werden kann.
Für $\boldsymbol{\beta}_1$ und 20 Kovariablen sind dies die Kovarianzmatrizen $\boldsymbol{\Sigma}_{6\_2}$, $\boldsymbol{\Sigma}_{6\_9}$ und $\boldsymbol{\Sigma}_{6\_15}$, mit denen neben der relevanten Kovariable jeweils 2, 9 bzw. 15 der irrelevanten Kovariablen zusätzlich blockkorreliert werden. Außerdem kann mit $\boldsymbol{\Sigma}_7$ überprüft werden, wie sich das optimale *mtry* verhält, wenn nur die irrelevanten Kovariablen blockkorreliert sind. Abbildung 3.9 stellt die optimalen *mtry* in Abhängigkeit der verwendeten Korrelationen $c \in \{0.3, 0.6, 0.9\}$ und der Performancemaße *Kendall's* τ und MSE für $p = 20$ dar. Die optimalen *mtry* Werte für $p = 10$ und $p = 50$ sind im Anhang A.7.1 zu finden.

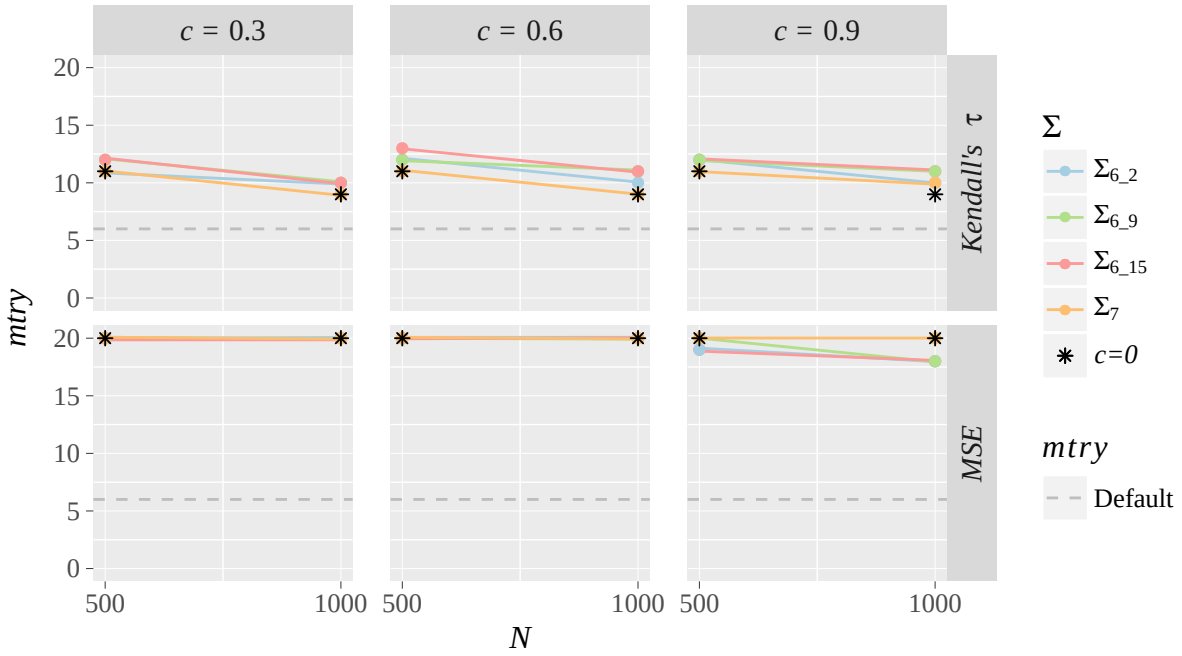


Abbildung 3.9: *Optimale mtry Werte für Regressionsszenarien mit β_1 und $p = 20$ korrelierten Kovariablen getrennt nach den verwendeten Performancemaßen und Korrelationen c. Zusätzlich sind in jeder Grafik die optimalen mtry Werte für die analogen Szenarien ohne korrelierte Kovariablen ergänzt.*
Die Definitionen der einzelnen Kovarianzmatrizen sind in Tabelle 3.3 zusammengefasst.

Mit *Kendall's* τ ist für diese Szenarien durch die Hinzunahme von blockkorrelierten, irrelevanten Kovariablen nur ein geringfügiger Anstieg des optimalen *mtry* im Vergleich zum unkorrelierten Szenario ($c = 0$) zu erkennen. Dabei kann jedoch kein deutlicher Unterschied zwischen den einzelnen Kovarianzmatrizen $\mathbf{\Sigma}_6$ ausgemacht werden. Die Korrelation der irrelevanten Kovariablen ($\mathbf{\Sigma}_7$) bewirkt dagegen keine Veränderung des optimalen *mtry* im Vergleich zum unkorrelierten Szenario. Die Mehrzahl der optimalen *mtry* mit dem *MSE* liegt bei 20, womit sich auch hier kein Einfluss der Kovarianzmatrizen erkennen lässt. Die einzelnen OOB-Kurven für *Kendall's* τ und den *MSE* in Abbildung 3.10 verstärken diesen Eindruck, denn für keine der genannten Kovarianzstrukturen mit Korrelation $c = 0.9$ tritt eine Änderung im Kurvenverlauf ein.
Bei nur einer relevanten Kovariable in den Daten ist demnach der Einfluss der Korrelation sehr gering und zeigt keine bedeutenden Auswirkungen auf das optimale *mtry*.

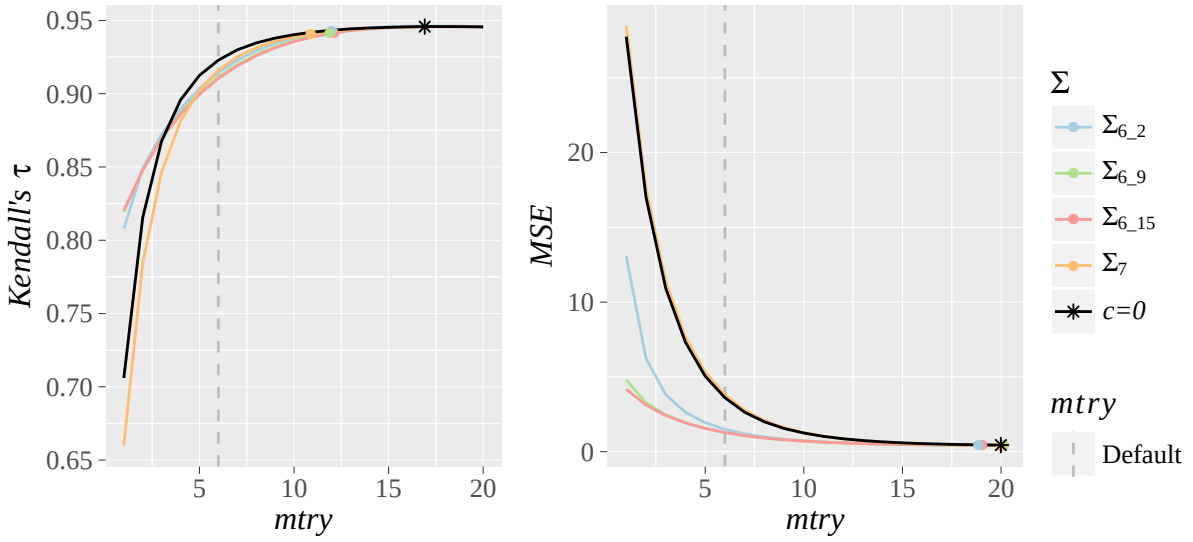


Abbildung 3.10: *OOB-Kurven der Performancemaße Kendall's τ und MSE für Regressionsszenarien mit* 500 *Beobachtungen, 20 korrelierte Kovariablen ($c = 0.9$) und Koeffizientenvektor $\boldsymbol{\beta}_1$.*

Andere Erkenntnisse liefern dagegen die Korrelationsstrukturen des Koeffizientenvektors $\boldsymbol{\beta}_4$ mit fünf relevanten Kovariablen. Hierbei wurden ebenfalls vier verschiedene Kovarianzmatrizen angewendet: Mit $\boldsymbol{\Sigma}_{6\_0}$ werden nur die relevanten Kovariablen blockkorreliert. Diese Struktur wird durch zwei weitere irrelevante Kovariablen ergänzt, indem $\boldsymbol{\Sigma}_{6\_2}$ angewendet wird. Mit $\boldsymbol{\Sigma}_7$ können nur die irrelevanten Kovariablen blockkorreliert definiert werden und $\boldsymbol{\Sigma}_8$ kombiniert die Korrelation von relevanten und irrelevanten Kovariablen, wobei jeweils die Hälfte der relevanten ($\hat{=}$ 2 Kovariablen) und die Hälfte der irrelevanten Kovariablen ($\hat{=}$ 7 Kovariablen für $p = 20$) blockkorreliert werden. Die dabei mit den Performancemaßen *Kendall's* τ und MSE resultierenden optimalen *mtry* sind für die Szenarien mit $p = 20$ Kovariablen und verschiedenen Korrelationen c in Abbildung 3.11 dargestellt. Die optimalen *mtry* für die analogen Szenarien mit $p = 10$ und $p = 50$ sind im Anhang A.7.1 dargestellt.

Mit *Kendall's* τ ist für eine steigende Korrelation kaum eine Änderung im optimalen *mtry* im Vergleich zum unkorrelierten Szenario zu erkennen, wenn nur die relevanten ($\boldsymbol{\Sigma}_{6\_0}$) oder auch zusätzlich dazu noch zwei weitere irrelevante Kovariablen ($\boldsymbol{\Sigma}_{6\_2}$) blockkorreliert sind. Auch die alleinige Korrelation der irrelevanten Kovarialben ($\boldsymbol{\Sigma}_7$) lässt das optimale *mtry* nur für eine sehr starke Korrelation von $c = 0.9$ geringfügig anwachsen. Sind jedoch jeweils die Hälfte der relevanten als auch die Hälfte der irrelevanten Kovariablen blockkorreliert ($\boldsymbol{\Sigma}_8$) wird das optimale *mtry* für eine steigende Korrelation deutlich größer. Liegt es im unkorrelierten Szenario noch bei $mtry = 4$, so ist es zum Beispiel mit $c = 0.9$ und $N = 500$ bei $mtry = 14$.

Für den MSE ergeben sich allerdings andere Effekte für die einzelnen Kovarianzmatrizen:

So sinkt das optimale *mtry* deutlich, wenn nur die relevanten ($\mathbf{\Sigma}_{6\_0}$) oder auch zusätzlich dazu noch zwei weitere irrelevante Kovariablen ($\mathbf{\Sigma}_{6\_2}$) blockkorreliert sind. Dagegen ändert es sich kaum, wenn nur die irrelevanten ($\mathbf{\Sigma}_7$) oder auch jeweils die Hälfte der irrelevanten und relevanten ($\mathbf{\Sigma}_8$) Kovariablen korreliert sind.

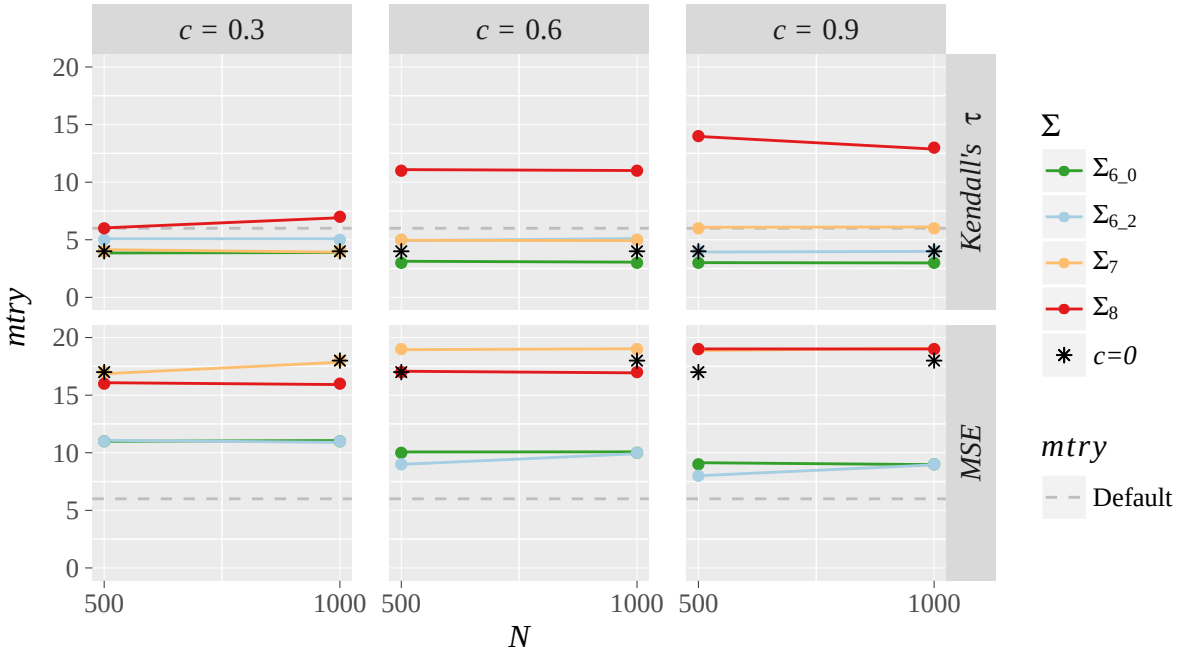


Abbildung 3.11: *Optimale mtry Werte für Regressionsszenarien mit $\boldsymbol{\beta}_4$ und $p = 20$ korrelierten Kovariablen getrennt nach den verwendeten Performancemaßen und Korrelationen c. Zusätzlich sind in jeder Grafik die optimalen mtry Werte für die analogen Szenarien ohne korrelierte Kovariablen ergänzt.*
Die Definitionen der einzelnen Kovarianzmatrizen sind in Tabelle 3.3 zusammengefasst.

Werden noch die Variablenwichtigkeiten der 500 Wiederholungen berücksichtigt, welche Abbildung 3.12 darstellt, so sind die Variablenwichtigkeiten der fünf korrelierten Kovariablen mit $\mathbf{\Sigma}_{6\_0}$ und dem kleinsten $mtry = 3$ am größten. Dies kann vielleicht eine mögliche Ursache für das gesunkene optimale *mtry* mit dem *MSE* sein. Theoretisch wird dagegen ein ähnliches optimales *mtry* wie im unkorrelierten Szenario erwartet, denn auch im unkorrelierten Fall werden die relevanten Kovariablen am häufigsten ausgewählt und diese Auswahlhäufigkeiten sollten sich durch die Blockkorrelation nicht stark ändern. Die Variablenwichtigkeiten mit $\mathbf{\Sigma}_7$ ähneln sich für alle betrachteten *mtry*, allerdings ist mit einem relativ geringem $mtry = 6$ eine leichte Überschätzung der korrelierten irrelevanten Kovariablen zu erkennen. Werden jeweils die Hälfte der relevanten und irrelevanten Kovarialben blockkorreliert ($\mathbf{\Sigma}_8$), so führt das dazu, dass die Variablenwichtigkeiten der unkorrelierten relevanten Kovariablen etwas unterschätzt werden und dagegen die korrelierten irrelevanten Kovariablen überschätzt werden. Diese Überschätzung fällt für ein

größeres *mtry* jedoch kleiner aus, was für das höhere optimale *mtry* spricht.

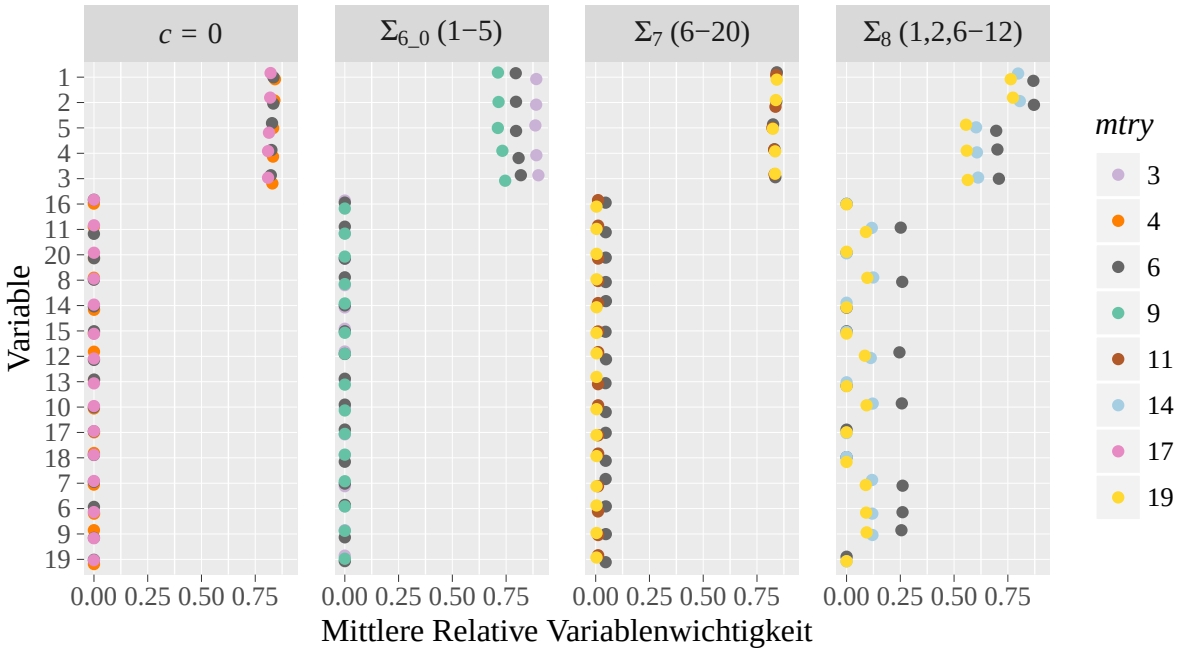


Abbildung 3.12: *Mittlere relative Permutation Importance über 500 Wiederholungen mit den Spezifikationen: metrischer Response, $\boldsymbol{\beta}_4 = (7, 7, 7, 7, 7, 0, \ldots, 0)$, $N = 500$, 20 unkorrelierte Kovariablen bzw. Kovarianzmatrizen $\boldsymbol{\Sigma}_6$ bis $\boldsymbol{\Sigma}_8$ mit $c = 0.9$. Die dabei blockkorrelierten Kovariablen sind in den Titeln gekennzeichnet. Die verschiedenen mtry Werte je Szenario entsprechen dem Defaultwert und den optimalen mtry Werten für die Performancemaße Kendall's τ und MSE.*

3.2.2 Klassifikation

Unkorrellierte Kovariablen

Auf gleiche Weise wie für die Regressionsszenarien wurden auch die OOB-Kurven für die verschiedenen Klassifikationsszenarien nach Algorithmus 2 ermittelt. Da sich die allgemeinen Interpretationen für die optimalen *mtry* zwischen den beiden Responsearten stark ähneln werden im Weiteren vor allem Auffälligkeiten angesprochen.

In Abbildung 3.13 sind beispielhaft für zwei Szenarien die OOB-Kurven auf Basis des *AUC* und des *Brier Scores* dargestellt. Dabei werden wie auch schon für die Regressionsszenarien diejenigen Random Forests verglichen, welche die wenigsten ($\boldsymbol{\beta}_1$) und die meisten ($\boldsymbol{\beta}_7$) relevanten Kovariablen in den Daten beinhalten. Ebenso bestehen die Datensätze dieser beiden Szenarien aus $N = 1000$ Beobachtungen und $p = 10$ unkorrelierten Kovariablen. Das angetragene optimale *mtry* wurde hierbei wieder über die Anpassung

aus den Gleichungen (3.9) und (3.10) ermittelt (Abbildung A.8 fasst für alle Koeffizientenvektoren die *mtry* Werte ohne Anpassung, also am Optimum, zusammen). Mit einem unteren Schwellenwert von 0.995 für das Verhältnis v_{mtry} in (3.10) werden jedoch die optimalen *mtry* Werte besonders für das *AUC* etwas zu stark geschrumpft. Daher wurde dieser Wert für die Klassifikationsszenarien auf 0.999 angehoben, womit nun das kleinste *mtry*, dessen Performancemaß eine Abweichung von maximal 0.1% zum Optimum besitzt, als optimales *mtry* bezeichnet wird.

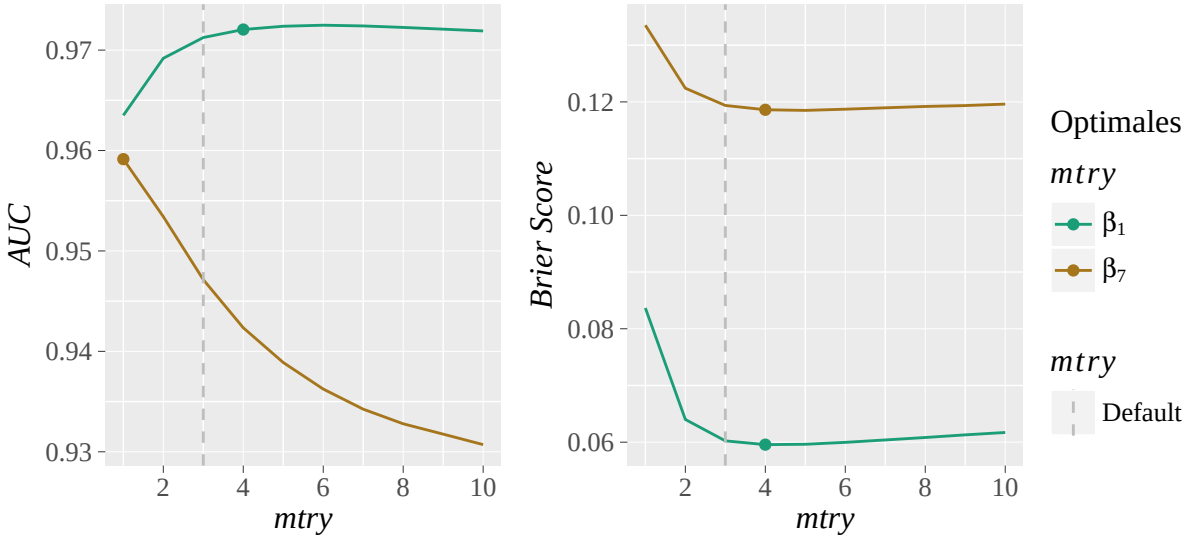


Abbildung 3.13: *OOB-Kurven der Performancemaße AUC und Brier Score für binären Response mit* 1000 *Beobachtungen,* 10 *unkorrelierten Kovariablen und zwei verschiedenen Koeffizientenvektoren* $\boldsymbol{\beta}_1$ *(eine relevante Kovariablen und* $\boldsymbol{\beta}_7$ *(nur relevante Kovariablen).*

Je nach residuen- oder rangbasiertem Performancemaß ähneln die Kurvenverläufe dabei stark den Regressionsverläufen. Die Abweichungen der optimalen *mtry* Werte vom Defaultwert fallen jedoch für die beiden betrachteten Szenarien nicht besonders groß aus. Das liegt daran, dass die optimalen *mtry* Werte mit dem *AUC* alle sehr klein sind und nahe am Defaultwert liegen und $\boldsymbol{\beta}_1$ für den *Brier Score* eine Ausnahme darstellt, was beides in Abbildung 3.14 deutlich wird. Hier sind die optimalen *mtry* für alle sieben betrachteten Koeffizientenvektoren dargestellt. Die Unterschiede im optimalen *mtry* zwischen den einzelnen Koeffizientenvektoren sind mit dem *AUC* teilweise sehr gering. Trotzdem ist eine Tendenz zu erkennen, denn bei ein oder zwei relevanten Kovariablen ($\boldsymbol{\beta}_1$ und $\boldsymbol{\beta}_2$) ist das optimale *mtry* am größten und dagegen ist es für eine große Anzahl an relevanten Kovarialben ($\boldsymbol{\beta}_6$ und $\boldsymbol{\beta}_7$) sehr klein. Damit gilt auch für diese Szenarien, je mehr relevante Kovariablen existieren, desto kleiner wird das optimale *mtry*.

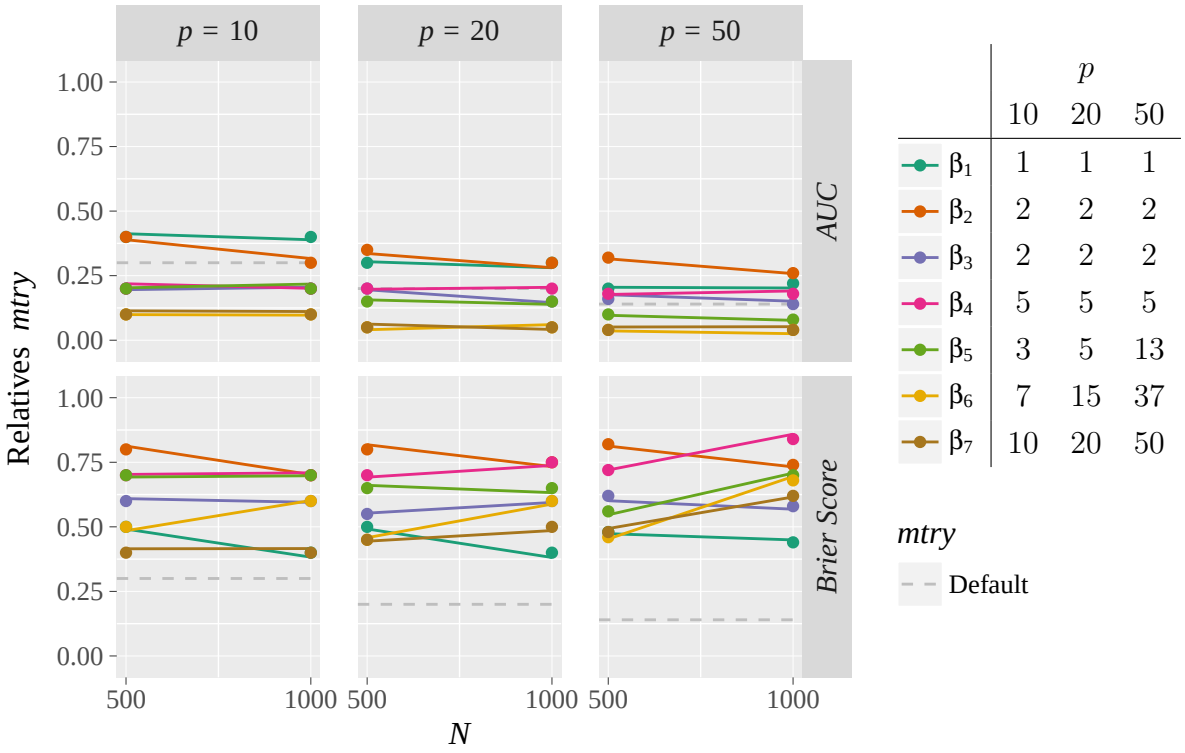


Abbildung 3.14: *Optimale relative mtry Werte für alle 42 betrachteten Klassifikationsszenarien ohne korrelierte Kovariablen getrennt nach den verwendeten Performancemaßen und der Anzahl an Variablen p.*
Die Tabelle rechts gibt die Anzahl an stark relevanten Kovariablen für die einzelnen Koeffizientenvektoren in Abhängigkeit von p an.

Im Vergleich dazu fällt es in Abbildung 3.14 schwer mit dem *Brier Score* solch eindeutige Aussagen zu treffen. Eine Erklärung für die stark schwankenden optimale *mtry* Werte liefert Abbildung 3.15. Denn hier ist gut zu erkennen, dass sich der grundsätzliche Kurvenverlauf der OOB-Kurven mit dem *Brier Score* für keinen der vier betrachteten Koeffizientenvektoren gravierend ändert. Dagegen ist mit dem *AUC* bereits bei zwei stark relevanten und drei weniger relevanten Kovariablen ($\boldsymbol{\beta}_3$) ein konkretes Optimum auszumachen. Dadurch, dass dieses Optimum jedoch für ein verhältnismäßig kleines *mtry* angenommen wird, ergibt sich für dieses Szenario mit dem *AUC* kein Unterschied im optimalen *mtry* zu $\boldsymbol{\beta}_4$ mit fünf stark relevanten Kovariablen.

Insgesamt kann also für diese Klassifikationsszenarien nicht so augenscheinlich wie für die Regressionsszenarien nachgewiesen werden, welche Auswirkung die Relevanz der einzelnen Kovariablen auf das optimale *mtry* besitzt. Jedoch ist die gleiche Tendenz zu erkennen: Die optimalen *mtry* sind für Szenarien mit wenigen relevanten Kovariablen meist höher als für Szenarien mit vielen ähnlich relevanten Kovariablen.

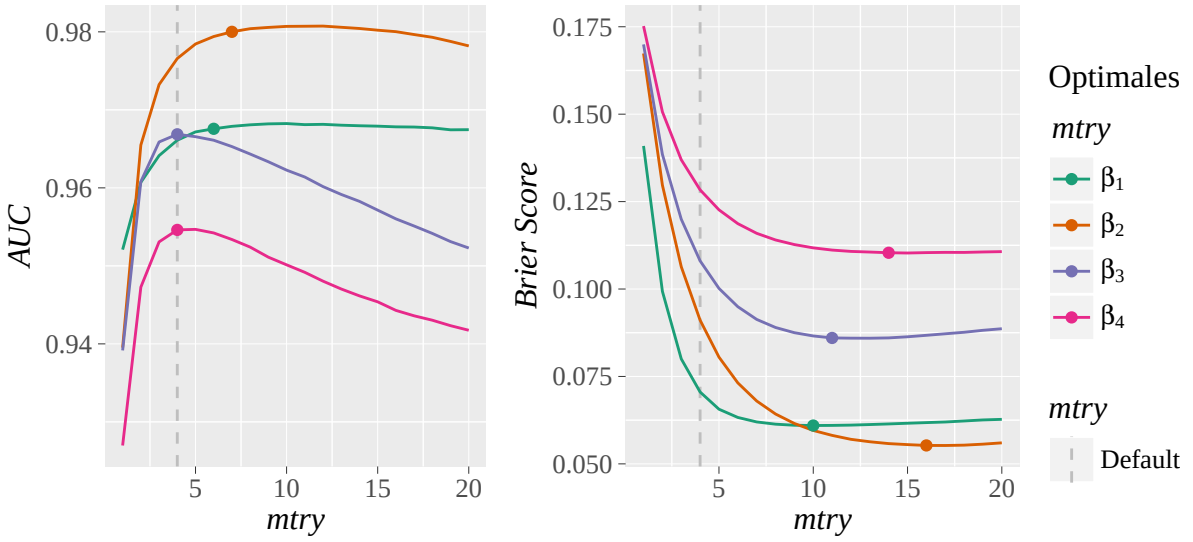


Abbildung 3.15: *OOB-Kurven der Performancemaße AUC und Brier Score für Klassifikationsszenarien mit* 500 *Beobachtungen,* 20 *unkorrelierten Kovariablen und den Koeffizientenvektoren* $\boldsymbol{\beta}_1$ *bis* $\boldsymbol{\beta}_4$.

Korrellierte Kovariablen mit $\boldsymbol{\Sigma}_1$ - $\boldsymbol{\Sigma}_5$

Auch für die Klassifikationsszenarien wurden die Kovarianzmatrizen $\boldsymbol{\Sigma}_1$ bis $\boldsymbol{\Sigma}_5$ für die Koeffizientenvektoren $\boldsymbol{\beta}_3$, $\boldsymbol{\beta}_5$ und $\boldsymbol{\beta}_7$ angewendet. Die dabei resultierenden optimalen *mtry* für $p = 20$ sind in Abbildung 3.16 dargestellt. Auch hier sind in der linken Spalte die Szenarien ohne korrelierte Kovariablen abgebildet und die verwendeten Kovarianzmatrizen sind durch verschiedene Farbintensitäten den jeweiligen Koeffizientenvektoren zuordenbar. Die Abbildungen für $p = 10$ und $p = 50$ sind im Anhang A.6.2 zu finden.
Abhängig von den verwendeten rang- bzw. residuenbasierten Performancemaßen zeigen sich mit diesen Kovarianzstrukturen innerhalb der Daten bei steigender Korrelation c sehr ähnliche Veränderungen im optimalen *mtry* wie schon für die Regressionsszenarien. Tabelle 3.6 stellt die Ergebnisse der beiden Responsearten gegenüber. Dadurch ergeben sich die gleichen Interpretationen wie zuvor in Kapitel 3.2.1. Analoge Abbildungen für die entsprechenden Variablenwichtigkeiten der Klassifikationsszenarien wie in Kapitel 3.2.1 können Anhang A.8 entnommen werden.

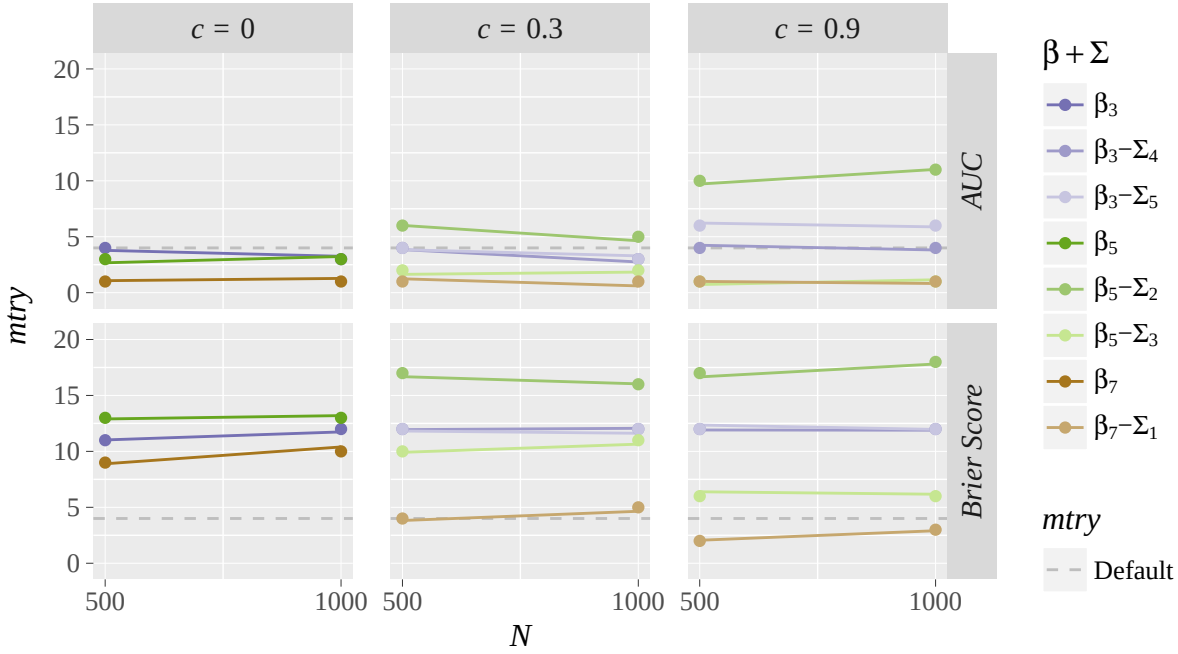


Abbildung 3.16: *Optimale mtry Werte für alle 20 betrachteten Klassifikationsszenarien mit $p = 20$, $\boldsymbol{\beta}_3$, $\boldsymbol{\beta}_5$ oder $\boldsymbol{\beta}_7$ und korrelierten Kovariablen getrennt nach den verwendeten Performancemaßen und Korrelationen c. In der linken Spalte sind zum Vergleich die optimalen mtry Werte für die analogen Szenarien ohne korrelierte Kovariablen angetragen.*
Die Definitionen der einzelnen Koffizientenvektoren und Kovarianzmatrizen sind in den Tabellen 3.1 und 3.2 zusammengefasst.

	rangbasierte Performancemaße		residuenbasierte Performancemaße	
	Klassifikation *AUC*	Regression *Kendall's τ*	Klassifikation *Brier Score*	Regression *MSE*
$\boldsymbol{\beta}_3 - \boldsymbol{\Sigma}_4$	→	→	→	(→)
$\boldsymbol{\beta}_3 - \boldsymbol{\Sigma}_5$	(→)	↗	→	(→)
$\boldsymbol{\beta}_5 - \boldsymbol{\Sigma}_2$	↗	↗	↗	↗
$\boldsymbol{\beta}_5 - \boldsymbol{\Sigma}_3$	(→)	(→)	↘	↘
$\boldsymbol{\beta}_7 - \boldsymbol{\Sigma}_1$	→	→	↘	↘

Tabelle 3.6: *Gegenüberstellung der Veränderungen im optimalen mtry bei steigender Korrelation c für die Klassifikations- und Regressionsszenarien mit Kovarianzmatrizen $\boldsymbol{\Sigma}_1$ bis $\boldsymbol{\Sigma}_5$. Ein gleichbleibendes mtry ist dabei mit → gekennzeichnet, ein steigendes mtry mit ↗ und ein sinkendes mtry mit ↘. Ist das Ansteigen oder Absinken durch eine steigende Korrelation c im Vergleich zum unkorrelierten Szenario nur sehr gering (mtry Differenz ≤ 3), ist dies durch (→) dargestellt.*

Korrellierte Kovariablen mit Σ_6 - Σ_8

Auch für die Szenarien mit binären Response und Kovarianzmtarizen Σ_6 - Σ_8 für die Koeffizientenvektoren $\beta_1 = (7, 0, \ldots, 0)$ und $\beta_4 = (7, 7, 7, 7, 7, 0, \ldots, 0)$ zeigen sich in Abbildung 3.17 und 3.18 ähnliche Veränderungen im optimalen *mtry* wie auch schon bei den jeweiligen Regressionsszenarien mit $p = 20$ Kovariablen. Für $p = 10$ und $p = 50$ sind die optimalen *mtry* im Anhang A.7.2 abgebildet. Außerdem sind die Variablenwichtigkeiten für β_4 in Abbildung A.24 ergänzt. Aufgrund der geringen Unterschiede zwischen den Responsearten, welche in Tabelle 3.7 zusammengefasst sind, gelten somit die Ergebnisse und Interpretationen aus Kapitel 3.2.1 ebenso für die entsprechenden Klassifikationsszenarien.

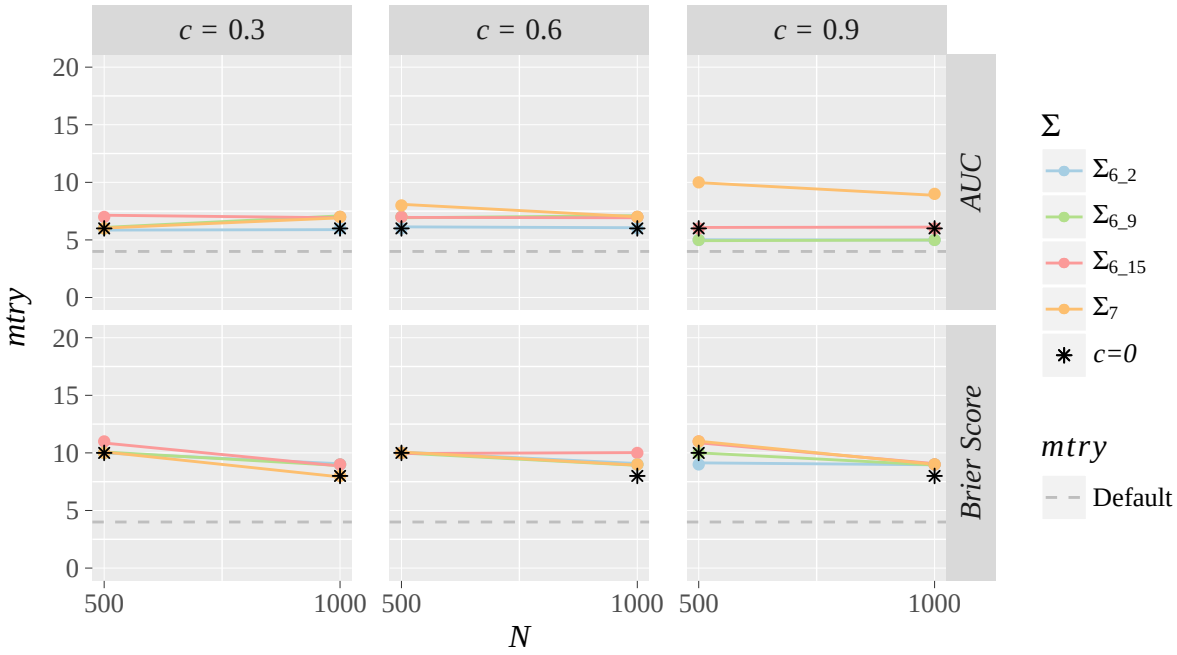


Abbildung 3.17: *Optimale mtry Werte für Klassifikationsszenarien mit $p = 20$, β_1 und korrelierten Kovariablen getrennt nach den verwendeten Performancemaßen und Korrelationen c. Zusätzlich sind in jeder Grafik die optimalen mtry Werte für die analogen Szenarien ohne korrelierte Kovariablen ergänzt.*
Die Definitionen der einzelnen Kovarianzmatrizen sind in Tabelle 3.3 zusammengefasst.

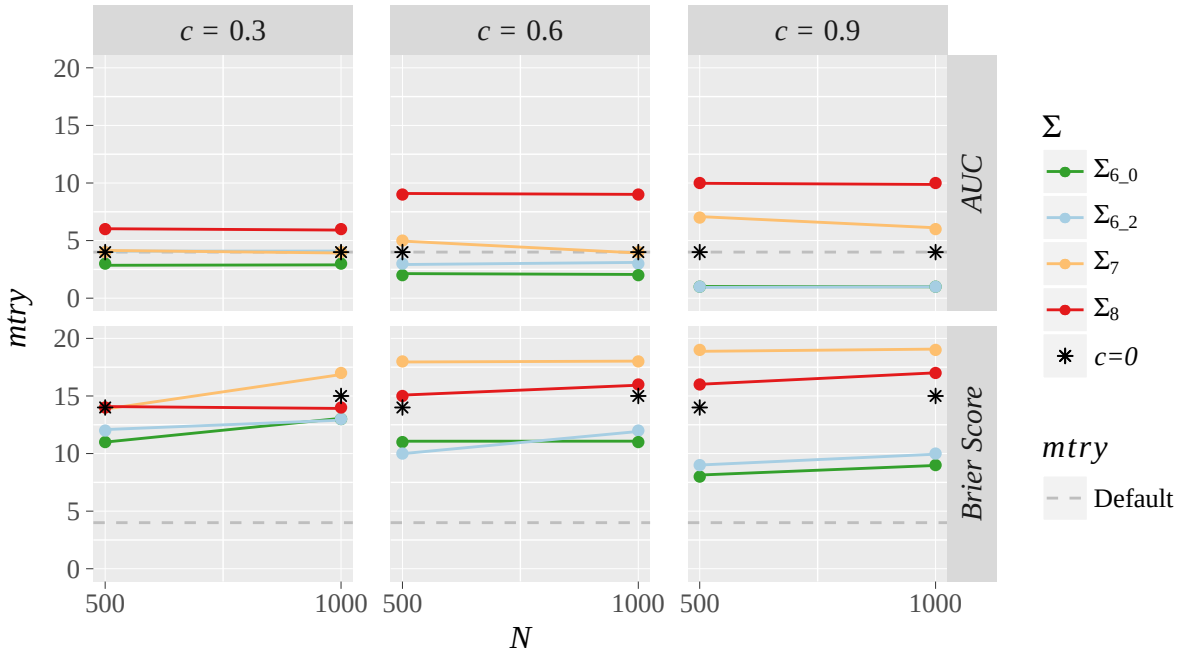


Abbildung 3.18: *Optimale mtry Werte für Klassifikationsszenarien mit $p = 20$, $\boldsymbol{\beta}_4$ und korrelierten Kovariablen getrennt nach den verwendeten Performancemaßen und Korrelationen c. Zusätzlich sind in jeder Grafik die optimalen mtry Werte für die analogen Szenarien ohne korrelierte Kovariablen ergänzt.*
Die Definitionen der einzelnen Kovarianzmatrizen sind in Tabelle 3.3 zusammengefasst.

	rangbasierte Performancemaße		residuenbasierte Performancemaße	
	Klassifikation *AUC*	Regression *Kendall's τ*	Klassifikation *Brier Score*	Regression *MSE*
$\boldsymbol{\beta}_1 - \boldsymbol{\Sigma}_{6\_2}$	$(\rightarrow)$	$(\rightarrow)$	$\rightarrow$	$(\rightarrow)$
$\boldsymbol{\beta}_1 - \boldsymbol{\Sigma}_{6\_9}$	$(\rightarrow)$	$(\rightarrow)$	$\rightarrow$	$(\rightarrow)$
$\boldsymbol{\beta}_1 - \boldsymbol{\Sigma}_{6\_15}$	$\rightarrow$	$(\rightarrow)$	$(\rightarrow)$	$(\rightarrow)$
$\boldsymbol{\beta}_1 - \boldsymbol{\Sigma}_7$	$\nearrow$	$\rightarrow$	$(\rightarrow)$	$\rightarrow$
$\boldsymbol{\beta}_4 - \boldsymbol{\Sigma}_{6\_0}$	$(\rightarrow)$	$(\rightarrow)$	$\searrow$	$\searrow$
$\boldsymbol{\beta}_4 - \boldsymbol{\Sigma}_{6\_2}$	$(\rightarrow)$	$\rightarrow$	$\searrow$	$\searrow$
$\boldsymbol{\beta}_4 - \boldsymbol{\Sigma}_7$	$(\rightarrow)$	$(\rightarrow)$	$\nearrow$	$(\rightarrow)$
$\boldsymbol{\beta}_4 - \boldsymbol{\Sigma}_8$	$\nearrow$	$\nearrow$	$(\rightarrow)$	$(\rightarrow)$

Tabelle 3.7: *Gegenüberstellung der Veränderungen im optimalen mtry bei steigender Korrelation c für die Klassifikations- und Regressionsszenarien mit $\boldsymbol{\beta}_1$ bzw. $\boldsymbol{\beta}_4$ und Kovarianzmatrizen $\boldsymbol{\Sigma}_6$ bis $\boldsymbol{\Sigma}_8$. Ein gleichbleibendes mtry ist dabei mit $\rightarrow$ gekennzeichnet, ein steigendes mtry mit $\nearrow$ und ein sinkendes mtry mit $\searrow$. Ist das Ansteigen oder Absinken durch eine steigende Korrelation c im Vergleich zum unkorrelierten Szenario nur sehr gering (mtry Differenz ≤ 3), ist dies durch $(\rightarrow)$ dargestellt.*

4 Empfehlungen zur mtry Wahl

In den vorherigen Kapiteln wurde deutlich, dass der Random Forest Hyperparamter *mtry* sehr stark von der Anzahl an relevanten Kovariablen innerhalb eines Datensatzes und der Korrelation der Kovariablen abhängt. Daher ist es empfehlenswert, besonders diese beiden Eigenschaften eines Datensatzes zu berücksichtigen, wenn ein Wert für *mtry* festgelegt werden soll. Ebenfalls sollte allerdings auch das gewählte Modellgütemaß Beachtung finden, da Kapitel 3.2 auch gezeigt hat, dass die Performance eines Random Forests je nach Maß für unterschiedliche *mtry* Werte optimiert werden kann.
Im Folgenden werden Möglichkeiten vorgestellt, mit denen diese Eigenschaften bestimmt werden können. Außerdem wird anhand zweier Beispiele überprüft, ob sich damit ein nahezu optimales *mtry* bestimmen lässt.

4.1 Messung der Korrelation und Relevanz von Kovariablen

Die Stärke des Zusammenhangs von Kovariablen lässt sich mit verschiedenen Korrelationsmaßen bestimmen. Neben dem in Kapitel 2.1.1 vorgestellten $Kendall's\ \tau$ beschreiben Fahrmeir et al. (2006, S. 135 - 146) auch zwei weitere Zusammenhangsmaße für metrische Variablen: Der Bravais-Person-Korrelationskoeffizient misst demnach lineare Zusammenhänge von Variablen. Dagegen kann mit dem Spearman-Korrelationskoeffizient die Stärke des monotonen Zusammenhangs zweier Variablen ermittelt werden. Natürlich besteht mit diesen Korrelationsmaßen nicht nur die Möglichkeit, den Zusammenhang zweier Kovariablen zu messen, sondern auch den Zusammenhang der einzelnen Kovariablen mit einem Response. Werden diese Korrelationskoeffizienten für jede Kovariable ermittelt, entsteht meist ein erster Eindruck, welche der Kovariablen einen Einfluss auf die Zielgröße besitzt und damit eine möglicherweise relevante Kovariable darstellt.

Für Datensätze mit binären Variablen eignen sich die genannten Koeffizienten allerdings nicht. Um eine Assoziation zwischen metrischen Kovariablen und kategorialem Response messen zu können, müssen daher andere Maße eingesetzt werden. Ein Beispiel hierfür ist die *Mutual Information*. Diese ist zwischen zwei stetigen Zufallsvariablen $\mathbf{X}$ und $\mathbf{Z}$ nach

Cover und Thomas (1991, S. 231-232) über die gemeinsame Dichte $f(x, z)$ definiert als

$$I(X; Z) = \int_Z \int_X f(x, z) log \frac{f(x, z)}{f(x)f(z)} dxdz. \tag{4.1}$$

Damit lässt sich die Mutual Information auch für zwei diskrete Variablen mit

$$I(X; Z) = \sum_Z \sum_X f(x, z) log \frac{f(x, z)}{f(x)f(z)} \tag{4.2}$$

darstellen. Dabei gilt allgemein $I(X; Z) \geq 0$ und $I(X; Z) = 0$, falls $\mathbf{X}$ und $\mathbf{Z}$ unabhängige Zufallsvariablen sind.
Die Mutual Information gibt damit die durchschnittliche Menge an Information über eine Variable $\mathbf{X}$ an, die durch $\mathbf{Z}$ vorhergesagt werden kann (Cellucci et al., 2005). Um dieses Maß für eine metrische und eine kategoriale Variable anwenden zu können, muss eine der beiden Variablen transformiert werden. Meist wird dabei die stetige Kovariable diskretisiert. Cellucci et al. (2005) empfehlen dafür, den Wertebereich der Variable in n_P gleich große Partitionen aufzuteilen. Mit der Anzahl an Beobachtungen N ist n_P dabei als größte ganze Zahl definiert, die folgende Gleichung erfüllt:

$$n_P \leq \sqrt{\frac{N}{5}}. \tag{4.3}$$

In **R** lässt sich die beschriebene Mutual Information für zwei diskrete bzw. diskretisierte Variablen mit der Funktion *mi.plugin* aus dem `entropy` Package (Hausser und Strimmer, 2014, Version 1.2.1) berechnen.

Eine Alternative, mit welcher ebenfalls die Relevanz der einzelnen Kovariablen festgestellt werden kann, stellt die Variablenwichtigkeit eines Random Forests dar. Kapitel 3.2 zeigt von einigen der simulierten Datensätze die Permutation Importance für verschiedene *mtry* Werte. Für die Szenarien ohne korrelierte Kovariablen liefern die betrachteten *mtry* dabei kaum einen Unterschied in der Rangfolge der Variablenwichtigkeiten (siehe dazu beispielsweise Abbildungen 3.6, 3.7 und 3.8 für $c = 0$). Daher kann in diesen Fällen die jeweilige Relevanz der Kovariablen im Verhältnis zu den restlichen Kovariablen aus einem Random Forest mit beliebigem *mtry* bestimmt werden. Je kleiner dabei *mtry* gewählt wird, desto geringer ist der Rechenaufwand für den Random Forest.
Problematisch ist allerdings das Auftreten von korrelierten Kovariablen. Denn wie bereits in Kapitel 3.2 dargestellt, bestätigen sich die Erkenntnisse von Strobl et al. (2008) und die Korrelationen wirken sich auf die Permutation Importance aus. Die Variablenwichtigkeiten werden dabei vor allem für korrelierte irrelevante Kovariablen überschätzt und die Wichtigkeiten von korrelierten relevanten Kovariablen werden möglicherweise unterschätzt. Diese Effekte verstärken sich auch für eine steigende Korrelation zwischen den

Kovariablen.
Demnach ist die Permutation Importance kein besonders gut geeignetes Mittel relevante Kovariablen zu ermitteln, wenn sehr hohe Korrelationen zwischen den Kovariablen auftreten. Aber auch die von Strobl et al. (2008) vorgeschlagene *Conditional Permutation Importance* in Kombination mit den *Conditional Inference Trees* kann die beschriebene Überschätzung nicht gänzlich eliminieren, senkt diese jedoch beachtlich, wie auch Abbildung A.25 zu entnehmen ist. Es muss allerdings im Einzelfall abgewägt werden, ob der deutlich größere Rechenaufwand des Conditional Inference Forest für eine manchmal nur sehr geringe Verbesserung der Variablenwichtigkeiten in Kauf genommen wird.

4.2 Anwendungsbeispiele

In der Praxis sollte der Rechenaufwand zur Bestimmung eines Modellparameters natürlich so gering wie möglich gehalten werden. Daher wird im Folgenden anhand zweier Beispieldatensätze überprüft, ob die Korrelations- bzw. Assoziationsmaße zwischen den Kovariablen und dem Response bereits eine ausreichend gute Tendenz für die Wahl von *mtry* liefern.
Die beiden Beispieldatensätze stammen von der Onlineplattform *OpenML* (Vanschoren et al., 2013), welche unter anderem frei zugängliche Datensätze für das maschinelle Lernen aus den unterschiedlichsten Quellen bereitstellt. Mit dem **R**-Package `OpenML` (Casalicchio et al., 2017, Version 1.7) ist es möglich, diese Datensätze in einem **R** kompatiblen Format herunterzuladen.

4.2.1 Regressionsdaten

Der erste betrachtete Datensatz, mit der OpenML-ID 308, wird als *puma32H* (Rasmussen et al., 1996) bezeichnet. Dieser beinhaltet 8192 Beobachtungen und 33 stetige Variablen. Die Daten wurden während einer realistischen Simulation der Dynamiken eines Roboterarms mit der Produktbezeichnung *Puma 560* erhoben. Die Winkelbeschleunigung einer der Verbindungen des Roboterarms stellt dabei den Response dar, welche durch verschiedene Eigenschaften wie zum Beispiel Winkelpositionen, Drehmomente und Geschwindigkeiten vorhergesagt werden kann.

Abbildung 4.1 visualisiert für eine Auswahl der im Datensatz vorhandenen Variablen den Korrelationskoeffizienten nach Spearman. Der Korrelationsplot aller Variablen ist in Abbildung A.26 ergänzt. Dabei tritt zwischen keiner der 32 Kovariablen ein messbarer monotoner Zusammenhang auf. Auch der Zusammenhang des Responses (hier mit Y gekennzeichnet) mit den Kovariablen ergibt lediglich für die Kovariable *tau4* einen erhöhten

Korrelationskoeffizienten. Damit existiert scheinbar nur eine Kovariable, die einen nennenswerten Einfluss auf den Response besitzt.
Mit der Simulationsstudie aus Kapitel 3 hat sich gezeigt, dass bei einer geringen Anzahl an relevanten Kovariablen das optimale *mtry* deutlich über dem Defaultwert liegt. Der Datensatz *puma32H* lässt sich sehr gut mit den Simulationsszenarien des Koeffizientenvektors $\boldsymbol{\beta}_1 = (7, 0, \ldots, 0)$ und 20 bzw. 50 unkorrelierten Kovariablen aus Kapitel 3.2.1 vergleichen. Für $N = 1000$ ergibt sich dabei ein optimales relatives *mtry* von 0.45 mit dem Performancemaß *Kendall's* τ und ein optimales relatives *mtry* von 1 mit dem MSE. Das spricht in diesem Beispiel mit 32 Kovariablen für ein *mtry* von 14 bzw. 32, was ebenfalls über dem Defaultwert von $mtry = 10$ liegt.

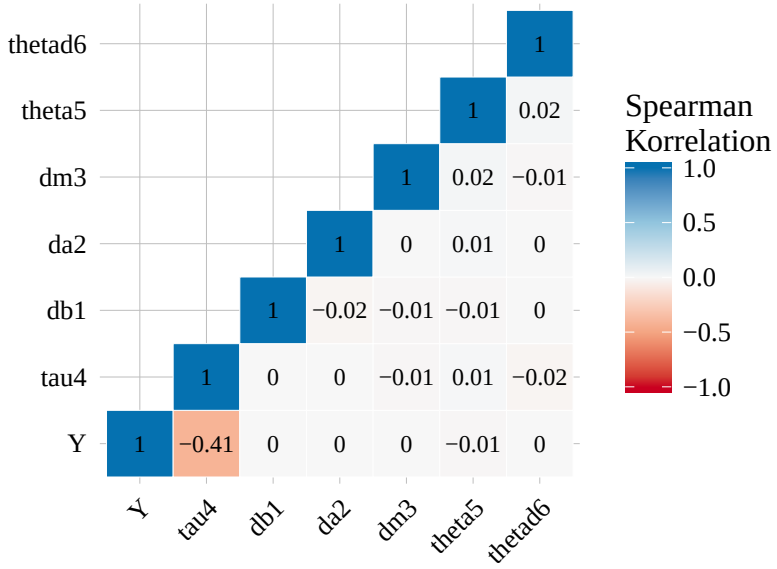


Abbildung 4.1: *Korrelationsplot einer Auswahl an stetigen Kovariablen und des Responses Y des puma32H Datensatzes. Der Korrelationsplot aller Variablen ist im Anhang A.26 ergänzt.*

Um diese Werte zu überprüfen, wurde für jeden *mtry* Wert im Intervall $[1, 32]$ ein Random Forest mit 500 Bäumen gefittet und dessen Performance sowohl mit *Kendall's* τ als auch mit dem MSE ermittelt. Daraus ergeben sich die bereits bekannten OOB-Kurven, die in Abbildung 4.2 für den Datensatz *puma32H* dargestellt sind. Die Optima der OOB-Kurven liegen hierbei mit *Kendall's* τ bei $mtry = 20$ und mit dem MSE bei $mtry = 27$. Wird jedoch die gleiche Anpassung für das optimale *mtry* durchgeführt wie auch schon bei der Simualitionsstudie (Gleichungen (3.9) und (3.10)), liegt das optimale *mtry* für *Kendall's* τ bei 15 und für den MSE bei 24. Die Random Forests mit diesen optimalen *mtry* Werten besitzen damit eine um maximal 0.5% vom Optimum abweichende Prädiktionsgüte.

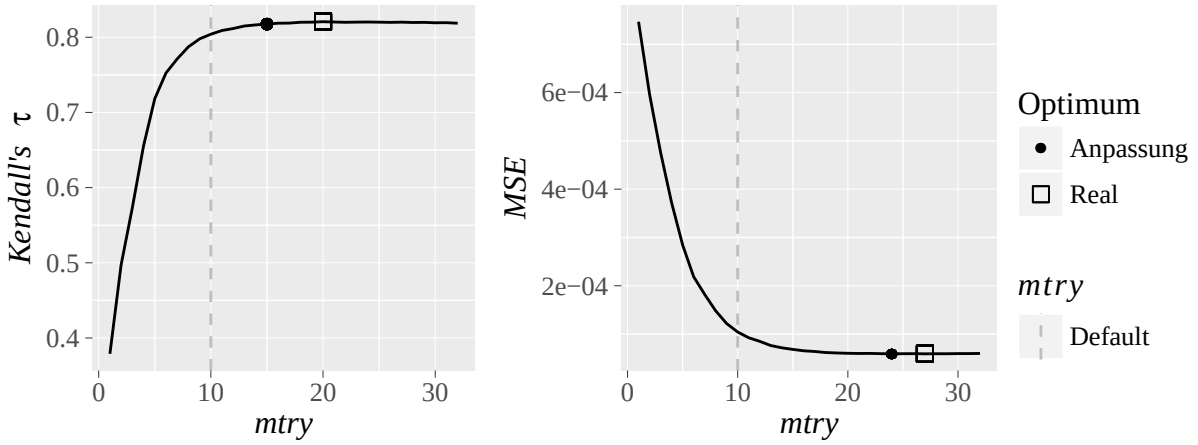


Abbildung 4.2: *OOB-Kurven der Performancemaße Kendall's τ und MSE für den puma32H Datensatz.*

Für *Kendall's* τ liegt somit das vorab durch die Korrelationskoeffizienten ermittelte $mtry = 14$ nur geringfügig unterhalb des wahren optimalen $mtry$. Dagegen wurde der $mtry$ Wert für den MSE deutlich zu groß gewählt. Der Kurvenverlauf für den MSE lässt allerdings bereits ab einem $mtry$ von etwa 17 ein Plateau erkennen, womit auch das ermittelte $mtry = 32$ eine ähnliche Prädiktionsgüte wie das optimale $mtry$ liefert. Eventuell könnte für den MSE eine etwas stärkere Anpassung für das optimale $mtry$ vorgenommen werden, wodurch sich in diesem Beispiel kleinere optimale $mtry$, sehr ähnlich denen zu *Kendall's* τ ergeben.

Abschließend sollen nun die Korrelationskoeffizienten nach Spearman zwischen dem Response und den Kovariablen mit der Variablenwichtigkeit eines Random Forests verglichen werden. Da für die Relevanz einer Variable die Richtung des Zusammenhangs keine Bedeutung hat, sind die absoluten Spearman-Korrelationen in Abbildung 4.3 angetragen. Die Rangfolge der Variablen entspricht dabei den Variablenwichtigkeiten, welche aus einem Random Forest mit dem optimalen $mtry$ für den MSE, $mtry = 24$, entnommen sind. Sowohl mit der Permutation Importance als auch mit dem Spearman Korrelationskoeffizient wird der Variable *tau4* die größte Relevanz zugewiesen. Überraschend ist die Schätzung der Variablenwichtigkeit für *theta5*, denn diese Kovariable ist weder mit dem Response noch mit der relevanten Kovariable *tau4* korreliert (siehe Abbildung 4.1) und besitzt dennoch eine verhältnismäßig hohe Variablenwichtigkeit. Aufgrund der fehlenden Fachkenntnisse über die Daten, kann an dieser Stelle die Plausibilität der beiden Maße nicht überprüft werden, da nicht bekannt ist, ob sich *theta5* oder auch eine der anderen Kovariablen tatsächlich auf die Winkelbeschleunigung einer Verbindung des Roboterarms auswirken.

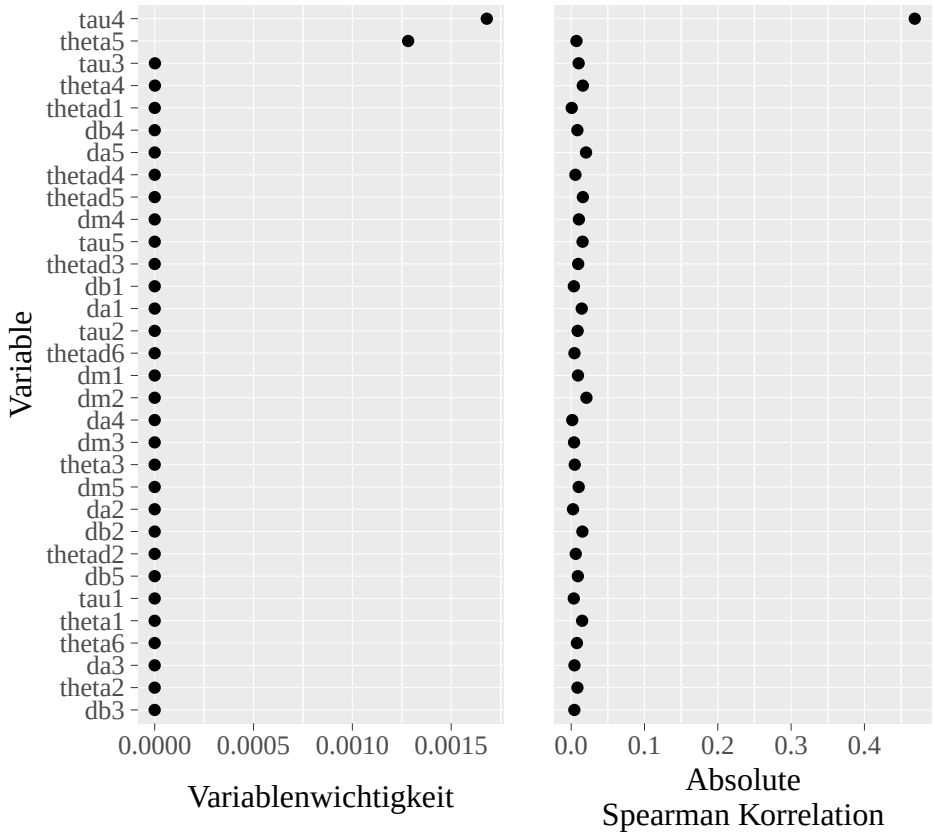


Abbildung 4.3: *Vergleich der Variablenwichtigkeit eines Random Forests mit mtry = 24 und der Spearmankorrelation aller Kovariablen des puma32H Datensatzes.*

4.2.2 Klassifikationsdaten

Der zweite betrachtete Datensatz, mit der OpenML-ID 1510, ist der *Breast Cancer Wisconsin (Diagnostic)* Datensatz (Lichman, 2013), auch *wdbc* genannt. Für die 569 Beobachtungen existieren 30 stetige Kovariablen und ein binärer Response. Die Kovariablen wurden aus digitalisierten Bildern einer Feinnadelbiopsie der Brust ermittelt. Diese beinhalten 10 verschiedene Eigenschaften von jeweils drei Zellkernen, wie zum Beispiel den Radius, die Kompaktheit oder die Standardabweichung der Graustufenwerte. Einige der Kovariablen sind allerdings auch auf Basis anderer vorliegender Kovariablen definiert, darunter unter anderem die Fläche oder der Umfang. Der binäre Response entspricht der Prognose des Gewebes (gutartig oder bösartig), welche auf Basis der Zellkerneigenschaften vorhergesagt werden kann.

Mit dieser Datenstruktur ist zu erwarten, dass einige der Kovariablen stark korreliert sind. Dies bestätigen auch die Korrelationskoeffizienten nach Spearman für eine Auswahl an 11 Kovariablen in der folgenden Abbildung 4.4. Beispielsweise besitzen die Kovariablen *V21,*

V23 und *V24* eine starke Blockkorrelation von fast 1. Die Korrelationskoeffizienten aller 30 Kovariablen sind im Anhang A.27 ergänzt.

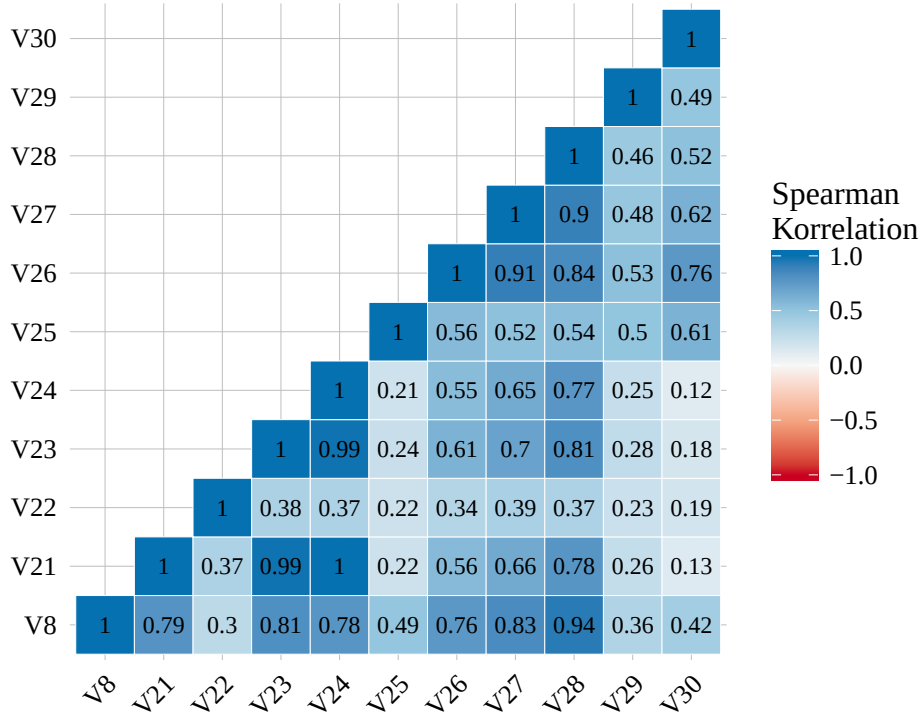


Abbildung 4.4: *Korrelationsplot einer Auswahl an stetigen Kovariablen des wdbc Datensatzes. Der Korrelationsplot aller Kovariablen ist im Anhang A.27 ergänzt.*

Um die Relevanz der einzelnen Kovariablen zu bestimmen, kam nun die in Kapitel 4.1 beschriebene Mutual Information zum Einsatz. Die damit erhaltenen Assoziationen der Kovariablen und des Responses sind in Abbildung 4.5 visuell dargestellt. Eine Mutual Information gleich 0 bedeutet allgemein die Unabhängigkeit zweier Variablen. In diesem Beispiel existieren nur sehr wenige Kovariablen, die eine Mutual Information nahe 0 besitzen und dagegen 24 Kovariablen mit einer Mutual Information größer 0.05. Damit lassen sich die Kovariablen grob in folgende vier Gruppen einteilen: 10 Kovariablen mit starker Relevanz, 5 mit moderater Relevanz, 9 mit geringer Relevanz und 6 Kovariablen mit kaum einer Relevanz für den Response.

Keines der betrachteten Simulationsszenarien in Kapitel 3.2.2 berücksichtigt exakt diese Kovariablenstruktur, am besten lässt sie sich wohl mit $\boldsymbol{\beta}_6 = (2, \ldots, 2, 15, \ldots, 15, 18, \ldots, 18)$ für $p = 20$ bzw. $p = 50$ und $N = 500$ vergleichen. Für unkorrelierte Kovariablen liegt dabei das optimale relative *mtry* mit dem *AUC* bei 0.05 bzw. 0.02 und für den Brier Score bei 0.45 (vgl. Abbildung 3.14).
Wie der Korrelationsplot in Abbildung 4.4 zeigt, tritt allerdings in diesem Beispiel eine

nicht unbedeutende Korrelation zwischen einigen der stark relevanten Kovariablen *V8, V21, V23* und *V28* auf.

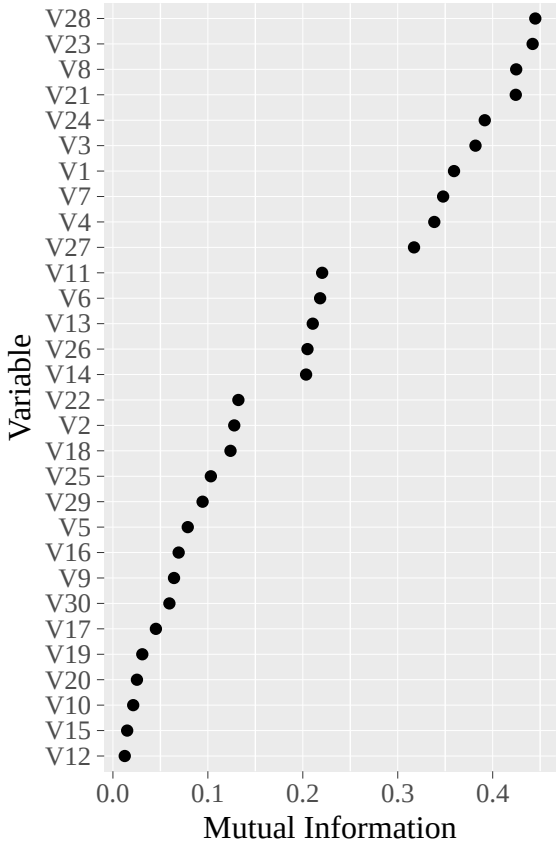


Abbildung 4.5: *Mutual Information aller stetigen Kovariablen mit dem Response des wdbc Datensatzes.*

In der Simulationsstudie wurden mit den Kovarianzmatrizen $\boldsymbol{\Sigma}_1$ und $\boldsymbol{\Sigma}_3$ jeweils die relevanten Kovariablen blockkorreliert. Dies hat laut Tabelle 3.6 mit dem AUC kaum einen Einfluss auf das optimale $mtry$ im Vergleich zum unkorrelierten Szenario hat. Jedoch ist demnach davon auszugehen, dass das optimale $mtry$ für den $BrierScore$ etwas geringer als im unkorrelierten Szenario ist. In Abbildung 3.16 sinkt das optimale $mtry$ für $c = 0.9$, $N = 500$ und $p = 20$ im Simulationsszenario mit $\boldsymbol{\beta}_5 - \boldsymbol{\Sigma}_3$ von $mtry = 13$ im unkorrelierten Szenario auf $mtry = 6$ und für $\boldsymbol{\beta}_7 - \boldsymbol{\Sigma}_1$ von $mtry = 9$ auf $mtry = 3$. Somit liegen beide optimalen $mtry$ Werte nahe am Defaultwert ($mtry = 4$) dieser Szenarien.
Da weder die Koeffizientenvektoren noch die Kovarianzmatrizen dieser Korrelationsszenarien vergleichbar mit der vorherrschenden Datenstruktur des *wdbc* Datensatzes sind, kann nur eine grobe Schätzung getätigt werden, dass das optimale $mtry$ mit dem *Brier Score* wahrscheinlich nahe am Defaultwert liegt und mit dem AUC bei 1 ($\lfloor 0.05 \cdot 30 \rfloor = 1$).

Auch für dieses Beispiel wurden die OOB-Kurven mithilfe von Random Forests mit 500 Bäumen für alle $mtry$ Werte im Intervall $[1, 30]$ ermittelt. Dabei ergeben sich die in Abbil-

dung 4.6 dargestellten Kurvenverläufe. Das Optimum des *AUC* liegt bei $mtry = 3$, wobei das optimale *mtry* durch die nachträgliche Anpassung wie in Gleichung (3.10) und einen Schwellenwert von 0.999 für Klassifikationsszenarien, bei $mtry = 1$ liegt. Ein Random Forest mit diesem optimalen *mtry* besitzt damit eine Prädiktionsgüte, die maximal 0.1% vom Optimum abweicht. Diese Anpassung ergibt für den *Brier Score* $mtry = 4$, was auch dem *mtry* Wert am Optimum entspricht.

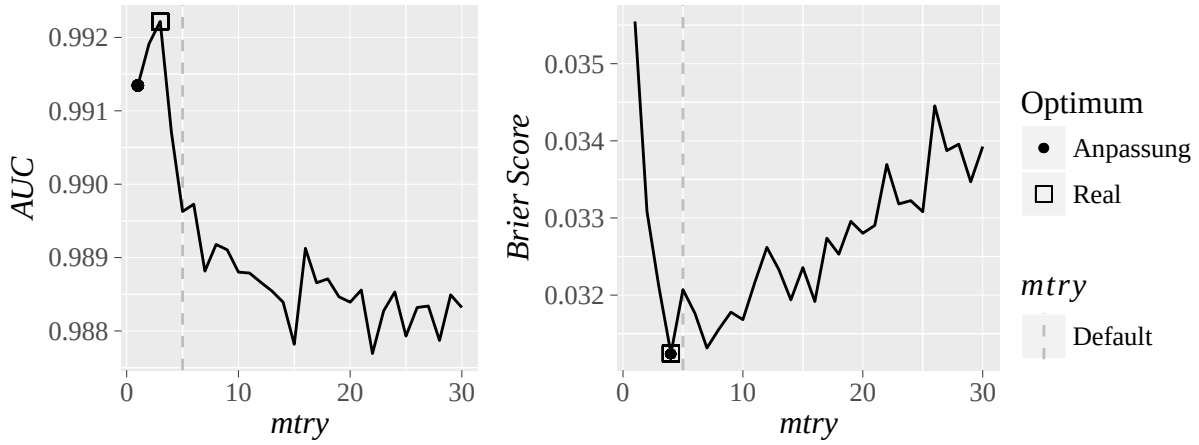


Abbildung 4.6: *OOB-Kurven der Performancemaße AUC und BrierScore für den wdbc Datensatz.*

Somit wurden die vorab mit der Mutual Information und den Korrelationskoeffizienten der Kovariablen festgelegten *mtry* Werte sehr gut gewählt. Das optimale *mtry* mit dem *AUC* liegt mit 1 exakt auf der Einschätzung und da der Defaultwert in diesem Beispiel $mtry = 5$ ist, weicht das tatsächliche optimale $mtry = 4$ für den *Brier Score* kaum davon ab.

Abschließend ist auch für dieses Beispiel ein Vergleich der Permutation Importance und der Mutual Information sehr interessant. Abbildung 4.7 stellt diese gegenüber, wobei die Rangfolge der Kovariablen in dieser Grafik den Variablenwichtigkeiten eines Random Forests mit 500 Bäumen und $mtry = 1$ (dem optimalen *mtry* mit dem *AUC*) entspricht. Durch die Berücksichtigung dieser Rangfolge, sind die einzelnen Werte der Mutual Information nicht ihrem Betrag nach angeordnet, womit die beiden Maße die Kovariablen nicht gleichermaßen relevant einschätzen. Jedoch besitzen die zehn Kovariablen, die mit der Mutual Information als stark relevant gruppiert wurden, auch eine hohe Variablenwichtigkeit (≥ 0.015). Damit stellt nur die Variable *V7* eine Ausnahme dar, die zwar ebenfalls eine Variablenwichtigkeit größer als 0.015 hat, deren Mutual Information diese allerdings nicht als stark relevante Kovariable gruppieren lässt. Für Variablen mit einer Variablenwichtigkeit kleiner als 0.015 können dagegen keine eindeutigen Übereinstimmungen mit den Gruppen der Mutual Information ausgemacht werden. Auch für dieses Beispiel kann ohne

inhaltliche Fachkenntnisse die Plausibilität beider Maße an dieser Stelle nicht überprüft werden.

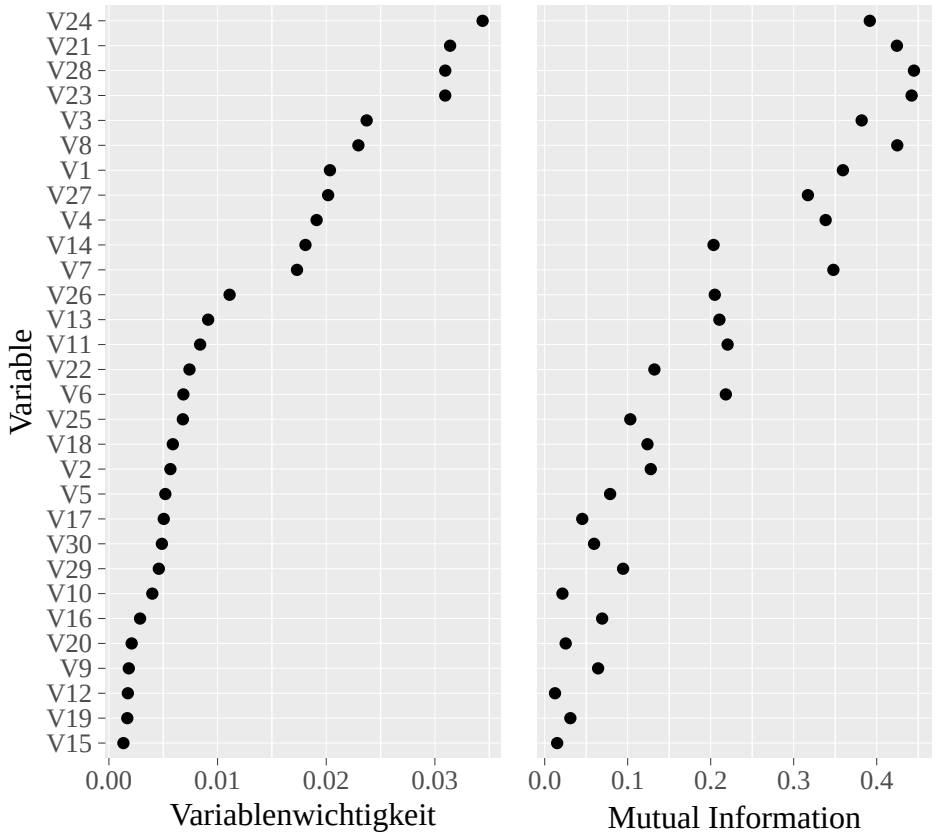


Abbildung 4.7: *Vergleich der Variablenwichtigkeit eines Random Forests mit mtry* $= 1$ *und der Mutual Information aller Kovariablen mit dem Response des wdbc Datensatzes.*

5 Fazit

Ziel dieser Arbeit war es, den Einfluss des Hyperparamteters *mtry* auf Random Forests zu untersuchen. Dafür wurde eine umfangreiche Simulationsstudie sowohl für Regressions- als auch für Klassifikationsdatensätze umgesetzt. Basis dieser Simulationsstudie waren unter anderem Thesen von Bernard et al. (2009), die für Klassifikationsdaten einen Einfluss der Anzahl an relevanten Kovariablen auf das optimale *mtry* festgestellt hatten. Außerdem aber auch Analysen von Strobl et al. (2008) und Gregorutti et al. (2016), die Auswirkungen korrelierter Kovariablen im Random Forest auf die Variablenwichtigkeit entdeckten, wodurch ebenfalls ein Einfluss auf das optimale *mtry* denkbar ist.

Insgesamt wurden 492 verschiedene Szenarien für die Simulationsstudie definiert, wobei diese sich für jede Responseart in drei Gruppen aufteilen lassen: Darunter die Szenarien ohne korrelierte Kovariablen, die Szenarien angelehnt an oben genannte Thesen mit den Kovarianzmatrizen $\mathbf{\Sigma}_1$ bis $\mathbf{\Sigma}_5$ und weitere Szenarien mit den Kovarianzmatrizen $\mathbf{\Sigma}_6$ bis $\mathbf{\Sigma}_8$ für detailliertere Analysen. Diese Aufteilung erleichtert das Zusammenfassen der jeweiligen Ergebnisse.
Grundsätzlich müssen alle Ergebnisse unter Berücksichtigung der Modellgütemaße betrachtet werden. Denn es hat sich gezeigt, dass je nach verwendetem rang- oder residuenbasierten Modellgütemaß, auch verschiedene optimale *mtry* Werte für ein Szenario in Betracht gezogen werden müssen. Dabei ist für die einzelnen Szenarien das optimale *mtry* mit den residuenbasierten Modellgütemaßen meist größer als mit den rangbasierten Maßen.

Die Szenarien mit unkorrelierten Kovariablen nehmen insbesondere die Relevanz der einzelnen Kovariablen in Augenschein. Durch die flexible Gestaltung der Simulationsdatensätze konnte vorab nicht nur festgelegt werden, welche Kovariablen einen relevanten Einfluss auf den Response besitzen sollen, sondern auch wie groß dieser tatsächlich ist. Damit bestätigt sich für die entsprechenden Regressionsszenarien eindeutig, dass mit steigender Anzahl an relevanten Kovariablen ein geringerer Wert für *mtry* ausreichend ist, um eine hohe Prädiktionsgüte sicherzustellen. Dabei zeigten nicht nur die stark relevanten, sondern auch die weniger relevanten Kovariablen einen Einfluss auf das optimale *mtry*.

Diese Beobachtungen lassen sich intuitiv erklären, da bei wenigen relevanten Kovariablen mit einem hohen *mtry* die Wahrscheinlichkeit größer ist, dass auch diese relevanten Kovariablen im Splitprozess berücksichtigt werden. Bei vielen ähnlich relevanten Kovariablen reicht dagegen schon ein kleines *mtry* aus, da nahezu jede Kovariable die gleiche Information trägt und die explizite Kovariablenauswahl im Splitprozess damit an Bedeutung verliert.
Auch die Klassifikationsszenarien lieferten ähnliche Erkenntnisse, wobei die OOB-Kurven der Performancemaße häufig ein Plateau zeigen, weswegen nicht immer eindeutige Optima und damit keine eindeutigen optimalen *mtry* festgestellt werden konnten. Allerdings besitzen die eben beschriebenen Grenzfälle mit vielen bzw. wenigen relevanten Kovariablen auch hierbei die kleinsten bzw. größten optimalen *mtry*.

Für die Korrelationsszenarien war vor allem von Interesse, wie sich die optimalen *mtry* bei steigender Korrelation der Kovariablen im Vergleich zu den jeweiligen unkorrelierten Szenarien verändern. Dabei zeigten sich zwischen den beiden Responsearten keine bedeutenden Unterschiede, weswegen die beobachteten Ergebnisse sowohl für die Regressionsszenarien als auch für die Klassifikationsszenarien gelten.
Bei Korrelation der relevanten Kovariablen ($\mathbf{\Sigma}_1$ und $\mathbf{\Sigma}_3$) ergeben sich verhältnismäßig kleine optimale *mtry*. Da mit den rangbasierten Performancemaßen das optimale *mtry* bereits für die unkorrelierten Szenarien sehr gering ist, kann nur für die residuenbasierten Maße ein Sinken des optimalen *mtry* aufgrund der Korrelation beobachtet werden. Dagegen steigt das optimale *mtry* unabhängig vom verwendeten Performancemaß, wenn nur die Kovariablen mit geringer Relevanz blockkorreliert sind ($\mathbf{\Sigma}_2$). Die Korrelation von jeweils einer stark relevanten, weniger relevanten und irrelevanten Kovariable ($\mathbf{\Sigma}_4$ und $\mathbf{\Sigma}_5$) hat, unabhängig vom Performancemaß, kaum einen Einfluss auf das optimale *mtry* im Vergleich zu den unkorrelierten Szenarien.
Mit den weiteren Kovarianzmatrizen $\mathbf{\Sigma}_6$ und $\mathbf{\Sigma}_8$, standen die Auswirkungen von korrelierten, irrelevanten Kovariablen auf das optimale *mtry* im Zentrum der Analysen. Bei nur einer relevanten Kovariable ließen sich, abgesehen von einer Ausnahme, keine Unterschiede im optimalen *mtry* durch die definierten Kovarianzmatrizen erkennen. Dagegen zeigten sich für fünf relevante Kovariablen abhängig vom Performancemaß Veränderungen im optimalen *mtry*. Bei steigender Korrelation zwischen den relevanten und bis zu zwei irrelevanten Kovariablen ($\mathbf{\Sigma}_6$) ergeben sich verhältnismäßig kleine optimale *mtry*. Wie auch schon vorab beschrieben, ist das optimale *mtry* der rangbasierten Perfomancemaße bereits für die unkorrelierten Szenarien sehr gering, weshalb auch hier nur für die residuenbasierten Maße deutlich kleinere *mtry* zu beobachten sind. Wenn nur die irrelevanten Kovariablen blockkorreliert sind ($\mathbf{\Sigma}_7$), hat dies kaum Auswirkungen auf das optimale *mtry*. Dagegen ist für die rangbasierten Performancemaße ein deutlicher Anstieg des op-

timalen *mtry* zu erkennen, wenn sowohl die Hälfte der relevanten als auch die Hälfte der irrelevanten Kovariablen korreliert ist ($\mathbf{\Sigma}_8$). Dieser Anstieg ist für die residuenbasierten Perfomancemaße dagegen nicht zu erkennen, was auch an den bereits hohen optimalen *mtry* der unkorrelierten Szenarien liegen kann.

Zusammenfassend muss also für die Veränderungen im optimalen *mtry* bei steigender Korrelation der Kovariablen besonders zwischen den rang- und residuenbasierten Perfomancemaßen unterschieden werden. Für die rangbasierten Maße ist bei diesen Simulationsszenarien kein Sinken des optimalen *mtry* zu beobachten, da bereits für die unkorrelierten Szenarien verhältnismäßig kleine *mtry* als optimal gelten. Allerdings erhöht sich das optimale *mtry*, wenn nur die weniger relevanten ($\mathbf{\Sigma}_2$) oder sowohl einige der relevanten als auch irrelevanten Kovariablen korreliert sind ($\mathbf{\Sigma}_5$, $\mathbf{\Sigma}_8$).
Für die residuenbasierten Maße steigt dagegen das optimale *mtry*, wenn nur weniger relevante oder nur irrelevante Kovariablen korreliert sind ($\mathbf{\Sigma}_2$, $\mathbf{\Sigma}_7$). Im Gegensatz dazu sinkt es, wenn alle relevanten und nur wenige der irrelevanten Kovariablen blockkorreliert sind ($\mathbf{\Sigma}_1$, $\mathbf{\Sigma}_3$, $\mathbf{\Sigma}_6$).
Die Beobachtungen mit $\mathbf{\Sigma}_2$ lassen sich zum Beispiel damit erklären, dass die Auswahlhäufigkeiten der korrelierten Kovariablen im Splitprozess vor allem für kleinere *mtry* ansteigen (Strobl et al., 2008). Jedoch können anhand der beschriebenen Szenarien und Random Forest Eigenschaften die Auswirkungen der Kovarianzmatrizen auf das optimale *mtry* nicht mit einer allgemeingültigen These aufgeklärt werden. Daher sind besonders in Bezug auf die korrelierten Kovariablen weitere Untersuchungen nötig, um detailliertere Aussagen zu den Gründen der jeweiligen Veränderungen treffen zu können.

Außerdem sind auch diverse Erweiterungen der Simulationsstudie denkbar. Mit dem bisherigen Design wurden nur normalverteilte Variablen berücksichtigt, womit keine allgemeingültigen Aussagen getroffen werden können. Daher könnte die Variablengenerierung auch auf Basis verschiedener anderer Verteilungen betrachtet werden. Außerdem sind die ausschließlich metrischen Kovariablen eine Einschränkung, da auch kategoriale Einflussgrößen eventuell einen Einfluss auf das optimale *mtry* haben können. Dabei ist auch die Aufnahme von mehrkategorialen Kovariablen, ebenso wie mehrkategorialen Responses für die Klassifikationsszenarien möglich. Zudem können Interaktionsstrukturen zwischen den Kovariablen und auch Ausreißer innerhalb der einzelnen Kovariablen neue Erkenntnisse liefern.
Durch die gewählte Anpassung des optimalen *mtry*, wird in den meisten Szenarien nicht das *mtry* am Optimum der OOB-Kurve als optimal angesehen, sondern ein kleineres. Das führt zu einem rechensparsameren Modell, welches dabei aber gleichzeitig eine Modellperformance sehr nahe am Optimum besitzt. Bereits bei diesen Auswertungen hat sich

jedoch gezeigt, dass es vielleicht auch sinnvoll ist, den Schwellenwert für diese Anpassung je nach Performancemaß und nicht abhängig von der Responseart zu wählen. Da dieser Schwellenwert maßgeblich die Ergebnisse beeinflusst, sind sicherlich auch alternative Konzepte zum Auffinden des optimalen *mtry* interessant.

Insgesamt kann geschlussfolgert werden, dass bei besonders wenigen oder vielen relevanten Kovariablen innerhalb der Daten und bei starken Korrelationen zwischen den Kovariablen, die bekannten Defaultwerte meist suboptimal sind und die *mtry* Wahl entscheidend für eine gute Modellperformance ist.
Die Anwendungsbeispiele haben gezeigt, dass zur Wahl von *mtry* die Relevanz der Kovariablen herangezogen werden kann, welche zum Beispiel anhand von Korrelationskoeffizienten oder Assoziationsmaßen, wie der Mutual Information, bestimmt werden. Um damit auf ein optimales *mtry* zu schließen, können Präzedenzfälle herangezogen werden, wie zum Beispiel die Szenarien aus der Simulationsstudie in dieser Arbeit. Jedoch ist dieses Vorgehen vor allem für die betrachteten Grenzfälle mit sehr wenigen oder sehr vielen relevanten Kovariablen besonders geeignet, da diese Variablenstrukturen mit den Zusammenhangsmaßen gut zu unterscheiden sind. Falls viele möglicherweise irrelevante Kovariablen vorab festgestellt werden (wie im zweiten Anwendungsbeispiel, Kapitel 4.2.2), sollte außerdem ein weiterer Ansatz in Erwägung gezogen werden. Statt einen Random Forest mit großem *mtry* zu ermitteln, kann durch eine Variablenselektion einerseits die Modellierung enorm vereinfacht werden und andererseits verringert sich daraufhin in den meisten Situationen der kritische Einfluss der *mtry* Wahl. Wenn jedoch anderweitige, komplexere Strukturen innerhalb der Daten auftreten, ist es denkbar, dass diese mithilfe der Zusammenhangsmaße nicht zuverlässig erkannt werden und die Einteilung in einen Präzedenzfall nicht mehr möglich ist. In dieser Situation ist es grundsätzlich ratsam auf die bereits bekannten Defaultwerte zurückzugreifen, die in vielen Fällen ebenfalls zu einer guten Modellperformance führen (Bernard et al., 2009; Díaz-Uriarte und de Andrés, 2006).

Literaturverzeichnis

Bernard, S., L. Heutte und S. Adam (2009). Influence of Hyperparameters on Random Forest Accuracy. In: J. A. Benediktsson, J. Kittler, und F. Roli (Hrsg.), *Multiple Classifier Systems. MCS 2009. Lecture Notes in Computer Science, vol 5519*, 171–180. Springer Berlin, Heidelberg.

Bischl, B. und M. Lang (2015). *parallelMap: Unified Interface to Parallelization BackEnds*. R package version 1.3.

Bischl, B., M. Lang, L. Kotthoff, J. Schiffner, J. Richter, E. Studerus, G. Casalicchio und Z. M. Jones (2016). mlr: Machine Learning in R. *Journal of Machine Learning Research 17*(170), 1–5.

Breiman, L. (2001). Random Forests. *Machine Learning 45*, 5–32.

Brier, G. W. (1950). Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review 78*, 1–3.

Casalicchio, G., J. Bossek, M. Lang, D. Kirchhoff, P. Kerschke, B. Hofner, H. Seibold, J. Vanschoren und B. Bischl (2017). OpenML: An R Package to Connect to the Machine Learning Platform OpenML. *Computational Statistics*, 1–15.

Cellucci, C. J., A. M. Albano und P. E. Rapp (2005). Statistical Validation of Mutual Information Calculations: Comparison of Alternative Numerical Algorithms. *Physical Review E 71*, 066208.

Cover, T. M. und J. A. Thomas (1991). *Elements of Information Theory*. New York: Wiley.

Díaz-Uriarte, R. und S. A. de Andrés (2006). Gene Selection and Classification of Microarray Data Using Random Forest. *BMC Bioinformatics 7*, 3.

Duroux, R. und E. Scornet (2016). Impact of Subsampling and Ppruning on Random Forests. ArXiv e-prints. arXiv:1603.04261 [math.ST].

Fahrmeir, L., R. Künstler, I. Pigeot und G. Tutz (2006). *Statistik: der Weg zur Datenanalyse* (6. Aufl.). Berlin u. a.: Springer.

Fawcett, T. (2006). An Introduction to ROC Analysis. *Pattern Recognition Letters 27*, 861–874.

Genuer, R., J.-M. Poggi und C. Tuleau (2008). Random Forests: Some Methodological Insights. Forschungsbericht RR-6729, INRIA.

Gregorutti, B., B. Michel und P. Saint-Pierre (2016). Correlation and Variable Importance in Random Forests. *Statistics and Computing 27*, 659–678.

Hanley, J. A. und B. J. McNeil (1982). The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve. *Radiology 143*, 29–36.

Hapfelmeier, A., T. Hothorn, K. Ulm und C. Strobl (2012). A New Variable Importance Measure for Random Forests with Missing Data. *Statistics and Computing 24*, 21–34.

Hastie, T., R. Tibshirani und J. Friedman (2009). *The Elements of Statistical Learning* (2. Aufl.). New York: Springer.

Hausser, J. und K. Strimmer (2014). *entropy: Estimation of Entropy, Mutual Information and Related Quantities*. R package version 1.2.1.

Hernández-Orallo, J., P. Flach und C. Ferri (2012). A Unified View of Performance Metrics: Translating Threshold Choice Into Expected Classification Loss. *Journal of Machine Learning Research 13*, 2813 – 2869.

Hothorn, T., K. Hornik und A. Zeileis (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics 15*, 651–674.

L'Ecuyer, P. (1999). Good Parameters and Implementations for Combined Multiple Recursive Random Number Generators. *Operations Research 47*, 159–164.

Leisch, F. und E. Dimitriadou (2010). *mlbench: Machine Learning Benchmark Problems*. R package version 2.1-1.

Lichman, M. (2013). UCI Machine Learning Repository [`http://archive.ics.uci.edu/ml`]. Irvine, CA: University of California, School of Information and Computer Sciences.

Probst, P. und A.-L. Boulesteix (2017). To Tune or not to Tune the Number of Trees in Random Forest? ArXiv e-prints. arXiv:1705.05654 [stat.ML].

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rasmussen, C. E., R. M. Neal, G. Hinton, D. van Camp, M. Revow, Z. Ghahramani, R. Kustra und R. Tibshirani (1996). Delve: Data for Evaluating Learning in Valid Experiments [`http://www.cs.toronto.edu/~delve/`]. Toronto, Ontario, Canada: University of Toronto.

Rosset, S., C. Perlich und B. Zadrozny (2006). Ranking-Based Evaluation of Regression Models. *Knowledge and Information Systems 12*, 331–353.

Roulston, M. S. (2007). Performance Targets and the Brier Score. *Meteorological Applications 14*, 185–194.

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin und A. Zeileis (2008). Conditional Variable Importance for Random Forests. *BMC Bioinformatics 9*, 307.

Strobl, C., A.-L. Boulesteix, A. Zeileis und T. Hothorn (2007). Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinformatics 8*, 25.

Toloşi, L. und T. Lengauer (2011). Classification with Correlated Features: Unreliability of Feature Ranking and Solutions. *Bioinformatics 27*, 1986–1994.

Vanschoren, J., J. N. van Rijn, B. Bischl und L. Torgo (2013). OpenML: Networked Science in Machine Learning. *SIGKDD Explorations 15*, 49–60.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wright, M. N. und A. Ziegler (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software 77*, 1–17.

A Allgemeiner Anhang

A.1 OOB-Kurven für verschiedene Anzahl an Wiederholungen

Die folgende Abbildung vergleicht die OOB-Kurven eines Beispielszenarios für verschiedene Anzahl an Wiederholungen W.
Mit $W = 50$ zeigt sich dabei eine etwas raue Kurve, womit diese Anzahl an Wiederholungen noch zu gering ist. Dagegen sind die Unterschiede im Verlauf zwischen den Kurven mit $W = 500$ und $W = 1000$ sehr gering und 500 Wiederholungen scheinen ausreichend zu sein.

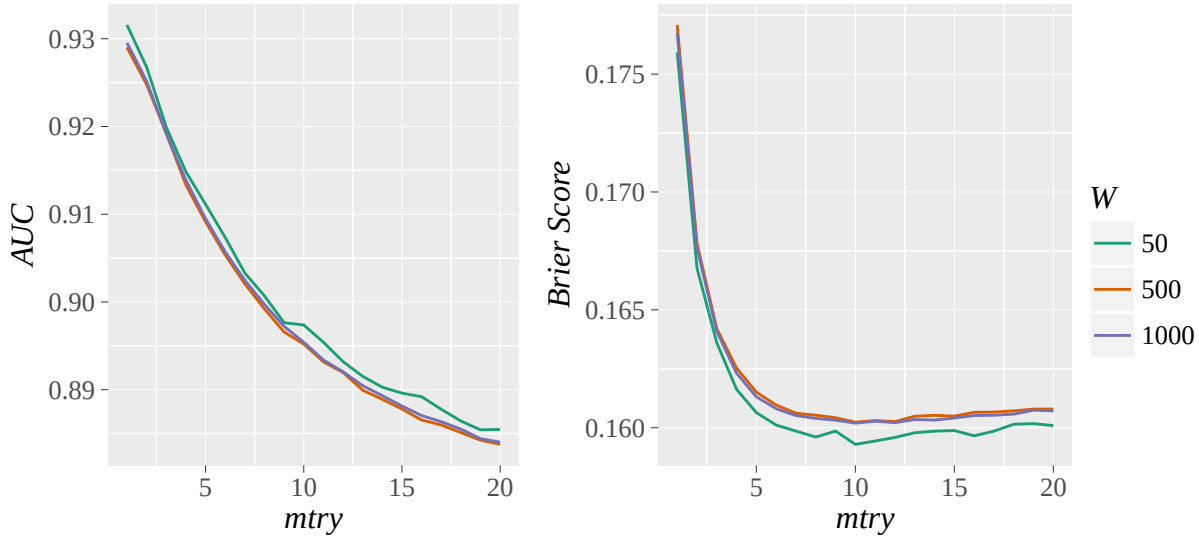


Abbildung A.1: *OOB-Kurven der Performancemaße AUC und Brier Score für binären Response mit* 500 *Beobachtungen,* 20 *unkorrelierten Kovariablen, Koeffizientenvektor* $\boldsymbol{\beta}_7 = (2, \ldots, 2, 3, \ldots, 3, 4, \ldots, 4)$ *und unterschiedlicher Anzahl an Wiederholungen* W.

A.2 OOB-Kurven für verschiedene Performancemaße

Beispielhaft sind im Folgenden die OOB-Kurven verschiedener Performancemaße für das Szenario mit Koeffizientenvektor $\boldsymbol{\beta}_3 = (7, 7, 7, 20, 20, 0, \ldots, 0)$, $N = 500$ Beobachtungen und $p = 10$ unkorrelierten Kovariablen für jeweils einen numerischen und binären Response dargestellt. Die Abkürzungen und Beschreibungen der verwendeten Performancemaße können dem **R**-Package `mlr` entnommen werden.

A.2.1 Regression

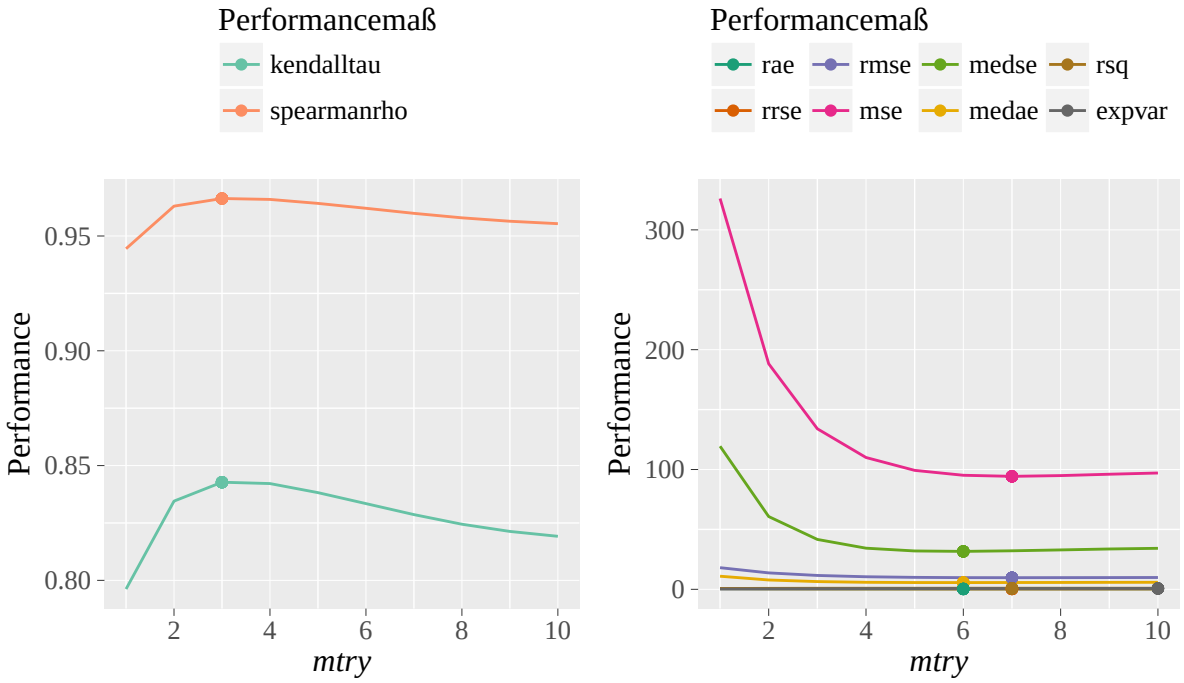


Abbildung A.2: *OOB-Kurven für verschiedene Regressions-Performancemaße.*

Die linken OOB-Kurven besitzen ein ausgeprägtes Optimum bei $mtry = 3$, das sich durch einen „Knick“ der OOB-Kurven zeigt. Dieser „Knick“ ist für die rechten OOB-Kurven nicht vorhanden, deren Optima bei etwas höheren $mtry \geq 6$ liegen. Da sich die Wertebereiche der rechten Performancemaße sehr stark unterscheiden, überdecken sich einige Kurven in dieser Grafik. Allerdings sind detailliertere Verläufe im elektronischen Anhang im Unterordner *„Zusätzliche_Grafiken“* dargestellt.
Exemplarisch wurden die Maße *Kendall's* τ und MSE für diese beiden Gruppen gewählt.

A.2.2 Klassifikation

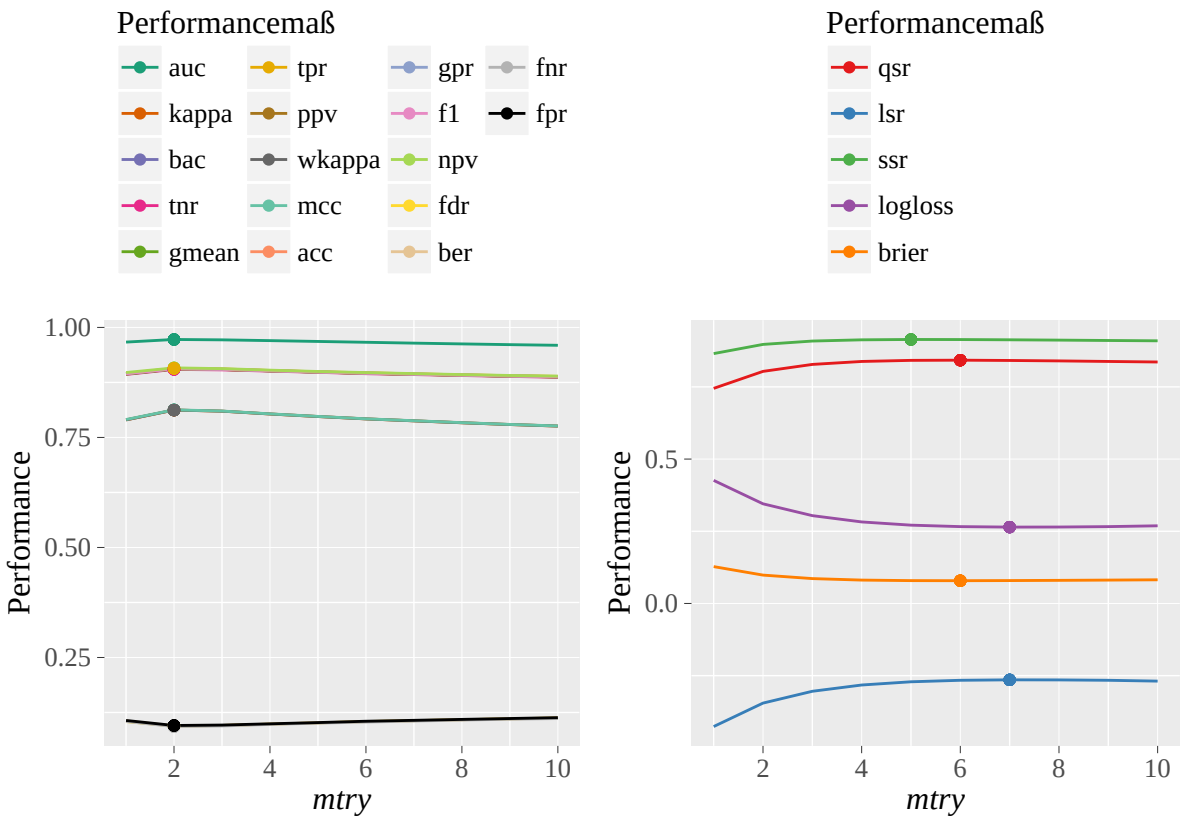


Abbildung A.3: *OOB-Kurven für verschiedene Klassifikations-Performancemaße.*

Die linken OOB-Kurven besitzen ihr Optimum bei $mtry = 2$, dagegen besitzen die rechten OOB-Kurven alle ihr Optimum bei einem größerem $mtry \geq 5$. Aufgrund der verschiedenen Wertebereiche der linken Perfomancemaße sind hier die beschriebenen „Knicke“ im Kurvenverlauf nur marginal zu erkennen. Allerdings sind detailliertere Verläufe im elektronischen Anhang im Unterordner *„Zusätzliche_Grafiken“* dargestellt.
Exemplarisch wurden die Maße *AUC* und *Brier Score* für die beiden Gruppen gewählt.

A.3 OOB-Kurven für nicht-lineare Einflussvariablen

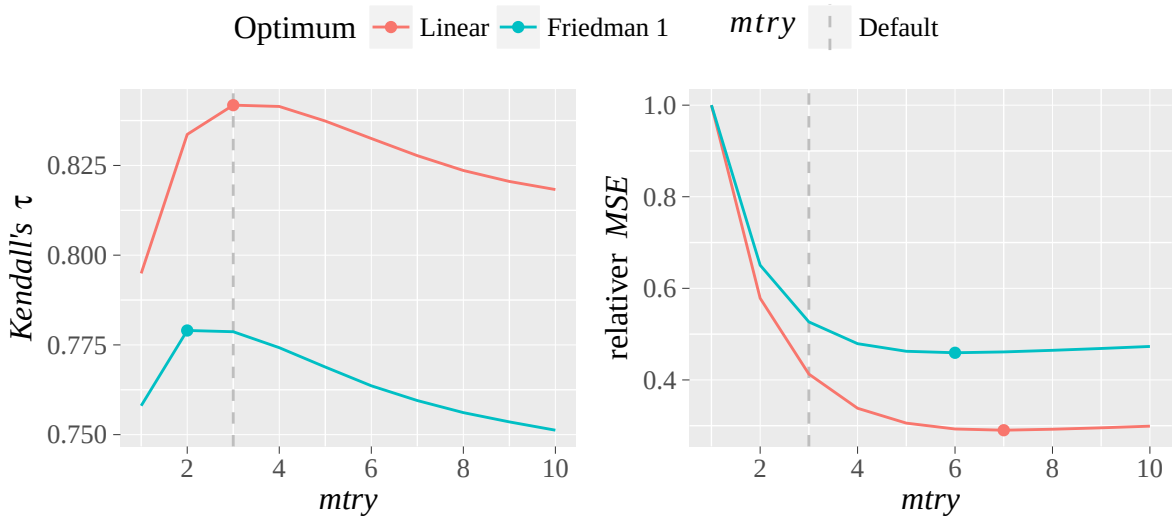


Abbildung A.4: *OOB-Kurven des Friedman 1 Regressionsproblems.*

Die Kurven in Abbildung A.4 sind nach Algorithmus 2 entstanden, wobei in Schritt 1 die Daten mithilfe der **R**-Funktion mlbench.friedman1 generiert wurden. Für die Übergabeparameter gilt $n = 500$ und $sd = 0.5$ (Anzahl an Beobachtungen und Varianz der Fehlerterme).

Da kein Koeffizientenvektor mit vergleichbar relevanten Kovariablen definiert wurde, ist beispielhaft der Verlauf der OOB-Kurve für einen linearen Einfluss mit dem Koeffizientenvektor $\boldsymbol{\beta}_3 = (7, 7, 7, 20, 20, 0, \ldots, 0)$, $p = 10$ und $N = 500$ Beobachtungen dargestellt. Es ist jedoch trotzdem sehr gut zu erkennen, dass sich die Kurvenverläufe und *mtry* am Optimum weder für $Kendall's\ \tau$ noch für den MSE stark unterscheiden. Anzumerken ist, dass hier für eine bessere Vergleichbarkeit der Wertebereiche, die Werte des MSE durch die jeweiligen Maxima geteilt wurden.

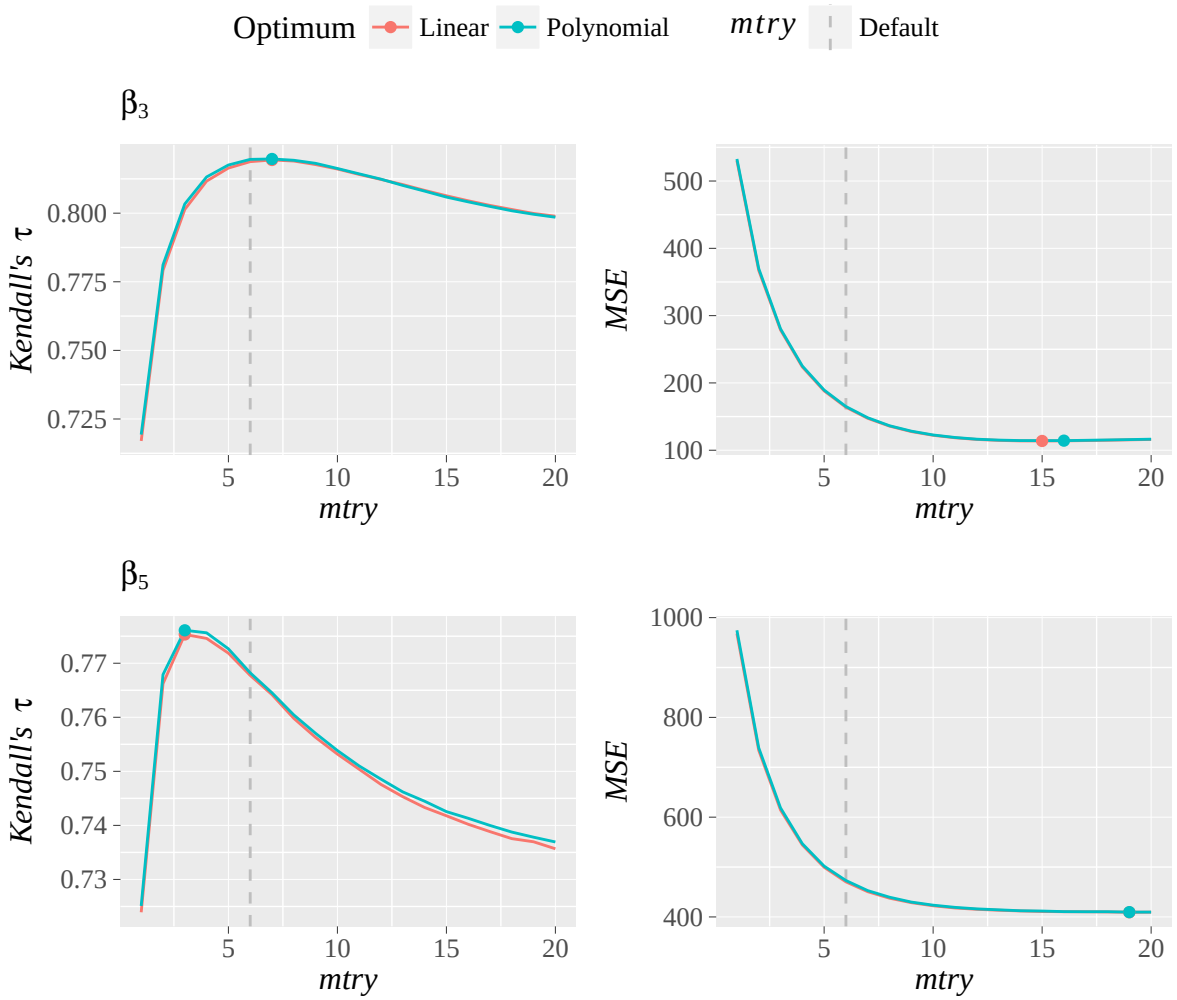


Abbildung A.5: *OOB-Kurven für polynomiale und lineare Einflussvariablen eines stetigen Responses.*

Abbildung A.5 stellt die OOB-Kurven für polynomiale und lineare Einflussvariablen eines stetigen Responses gegenüber. Die Datensätze bestehen aus 500 Beobachtungen, 20 unkorrelierten Kovariablen und den zwei Koeffizientenvektoren $\boldsymbol{\beta}_3 = (7, 7, 7, 20, 20, 0, \ldots, 0)$ und $\boldsymbol{\beta}_5 = (2, \ldots, 2, 3, \ldots, 3, 18, \ldots, 18)$.
Mit beiden Performancemaßen sind für die jeweiligen Koeffizientenvektoren kaum Unterschiede zwischen den OOB-Kurven für lineare und polynomiale Einflussvariablen zu erkennen.

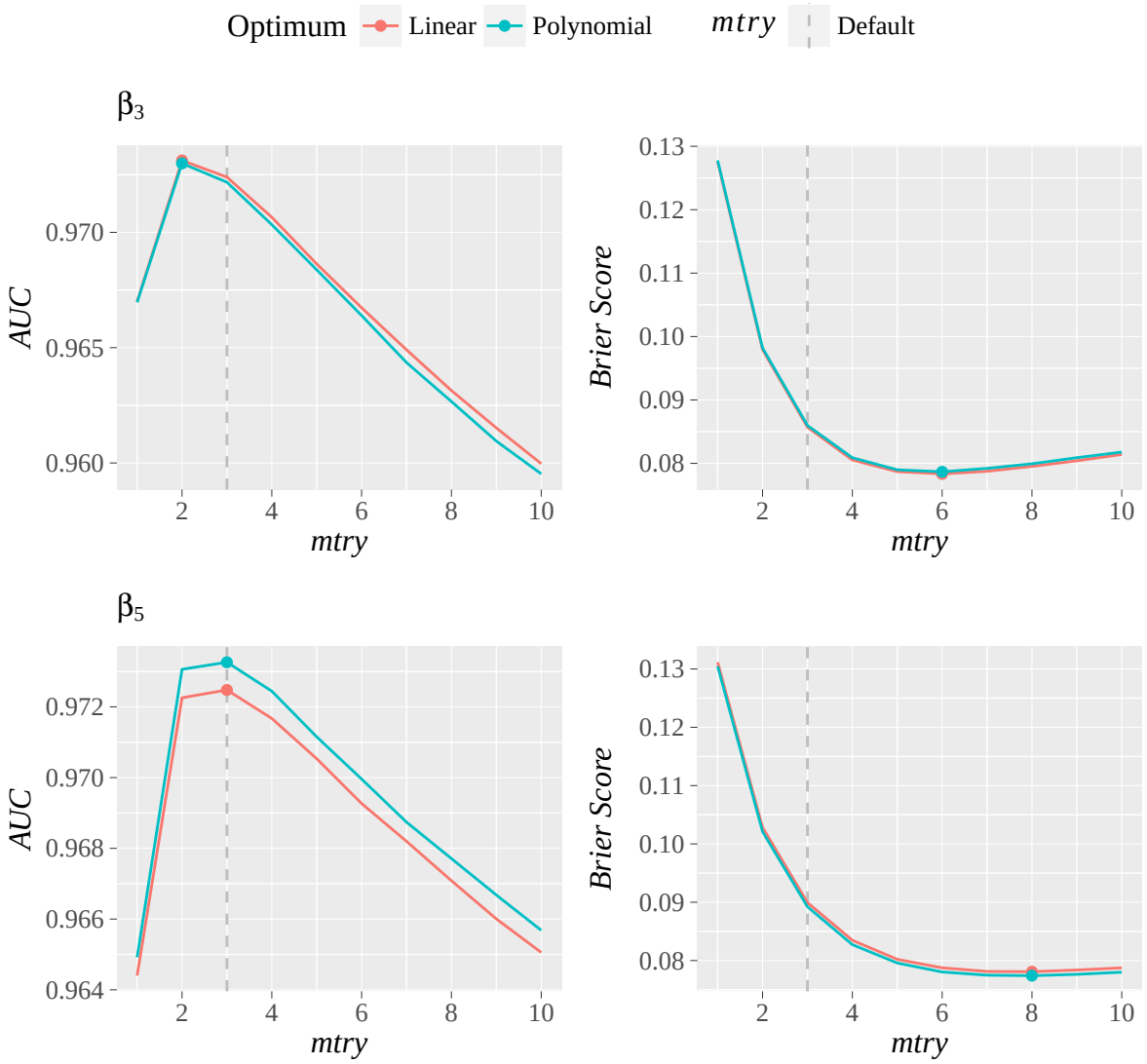


Abbildung A.6: *OOB-Kurven für polynomiale und lineare Einflussvariablen eines binären Responses.*

Die Abbildung A.6 stellt die OOB-Kurven für polynomiale und lineare Einflussvariablen eines binären Responses gegenüber. Die Datensätze bestehen aus 500 Beobachtungen, 10 unkorrelierten Kovariablen und den zwei Koeffizientenvektoren $\boldsymbol{\beta}_3 = (7, 7, 7, 20, 20, 0, \ldots, 0)$ und $\boldsymbol{\beta}_5 = (2, \ldots, 2, 3, \ldots, 3, 18, \ldots, 18)$.
Mit beiden Performancemaßen sind für die jeweiligen Koeffizientenvektoren kaum Unterschiede zwischen den OOB-Kurven für lineare und polynomiale Einflussvariablen zu erkennen.

A.4 Beschreibung der mittleren relativen Variablenwichtigkeit und der drei gewählten mtry Parameter

Um die Variablenwichtigkeiten der 500 Wiederholungen eines Szenarios zusammenfassen zu können, mussten für jede Wiederholung zuerst die absoluten Werte der Wichtigkeiten durch die maximale Wichtigkeit geteilt werden. Die damit erhaltenen relativen Variablenwichtigkeiten konnten daraufhin über alle 500 Wiederholungen gemittelt werden.
Die Variablenwichtigkeiten für ein Szenario wurden anschließend an Schritt 4 von Algorithmus 2 für drei festgelegte Werte $mtry_1$, $mtry_2$ und $mtry_3$ bestimmt und ebenfalls ausgegeben. Dabei wurden jeweils diejenigen $mtry$ gewählt, welche die zwei jeweiligen Performancemaße optimieren (nach der Anpassung aus Gleichung (3.10)) und der Defaultwert. Traten für ein Szenario zwei dieser $mtry$ Werte wiederholt auf, wurde eine Fallunterscheidung vorgenommen, um einen davon verschiedenen dritten $mtry$ Wert festzulegen. Dafür wurde der Wertebereich $[1, p]$ der möglichen $mtry$ in drei gleichgroße Quantile aufgeteilt. Den unterschiedlichen Werten $mtry_1$ und $mtry_2$ konnten dann die jeweiligen Quantile q_1 und q_2 zugeteilt werden. Auf Basis dieser beiden Quantile wurde daraufhin folgendermaßen $mtry_3$ festgelegt:

- Unterscheiden sich q_1 und q_2, so entspricht $mtry_3$ dem ganzzahligen Mittelwert innerhalb des verbliebenen Quantils, womit $q_3 \neq (q_1, q_2)$.
- Gleichen sich dagegen q_1 und q_2 und die beiden $mtry$ Werte entstammen dem
  - 1. oder 2. Quantil, so entspricht $mtry_3$ dem ganzzahligen Mittelwert innerhalb des 3. Quantils.
  - 3. Quantil, so entspricht $mtry_3$ dem ganzzahligen Mittelwert innerhalb des 1. Quantils.

A.5 Relative optimale mtry Werte unter Berücksichtigung des Optimums

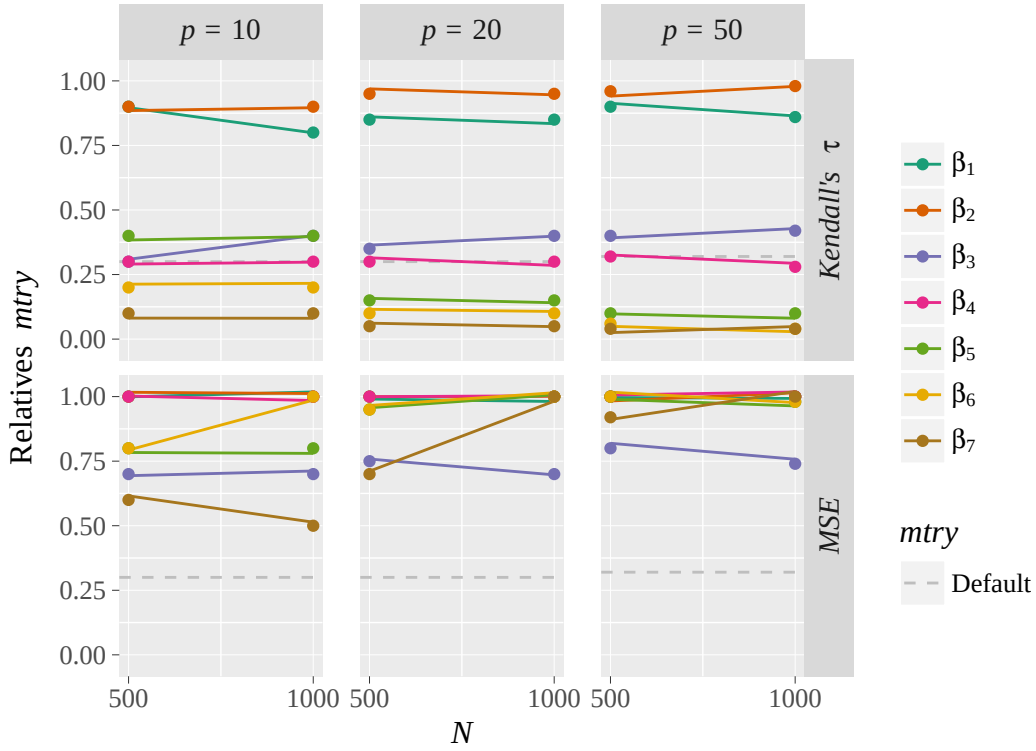


Abbildung A.7: *Relative mtry Werte am Optimum (ohne Anpassung) der Regressionsszenarien ohne korrelierte Kovariablen.*

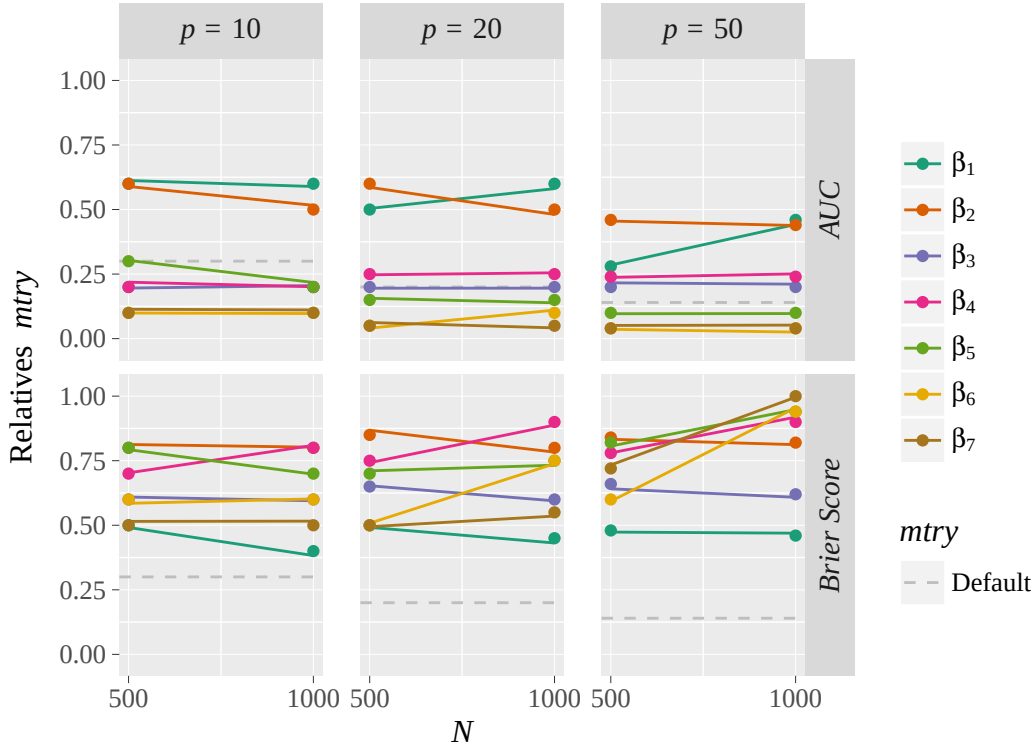


Abbildung A.8: *Relative mtry Werte am Optimum (ohne Anpassung) der Klassifikationsszenarien ohne korrelierte Kovariablen.*

A.6 Weitere Ergebnisse für optimale mtry Werte mit korrelierten Kovariablen ($\boldsymbol{\Sigma}_1$ bis $\boldsymbol{\Sigma}_5$)

Nachfolgend jeweils für alle Regressions- und Klassifikationsszenarien mit $p = 10$ und $p = 50$ blockkorrelierten Kovariablen ($\boldsymbol{\Sigma}_1$ bis $\boldsymbol{\Sigma}_5$) die optimalen *mtry* Werte, getrennt nach den verwendeten Performancemaßen und Korrelationen $c \in \{0.3, 0.9\}$. In der linken Spalte sind jeweils zum Vergleich die Werte für die analogen Szenarien ohne Korrelationen dargestellt. Die Definitionen der Koeffizientenvektoren und Kovarianzmatrizen sind in den Tabellen 3.1 und 3.2 zusammengefasst.

A.6.1 Regression

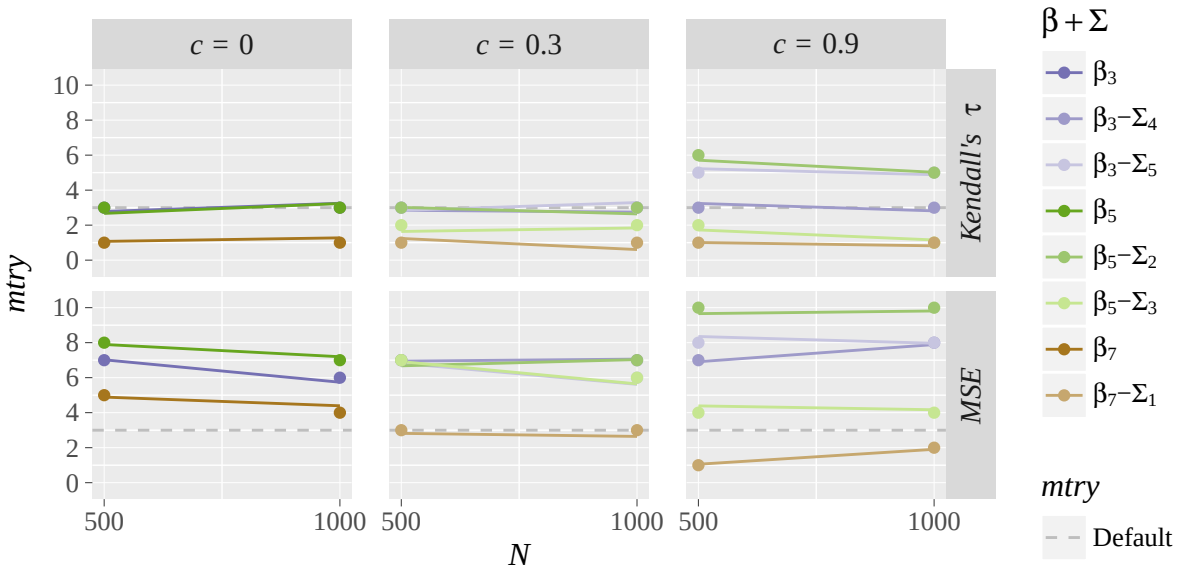


Abbildung A.9: *Optimale mtry Werte für Regressionsszenarien mit $p = 10$ und Kovarianzmatrizen $\boldsymbol{\Sigma}_1$ bis $\boldsymbol{\Sigma}_5$.*

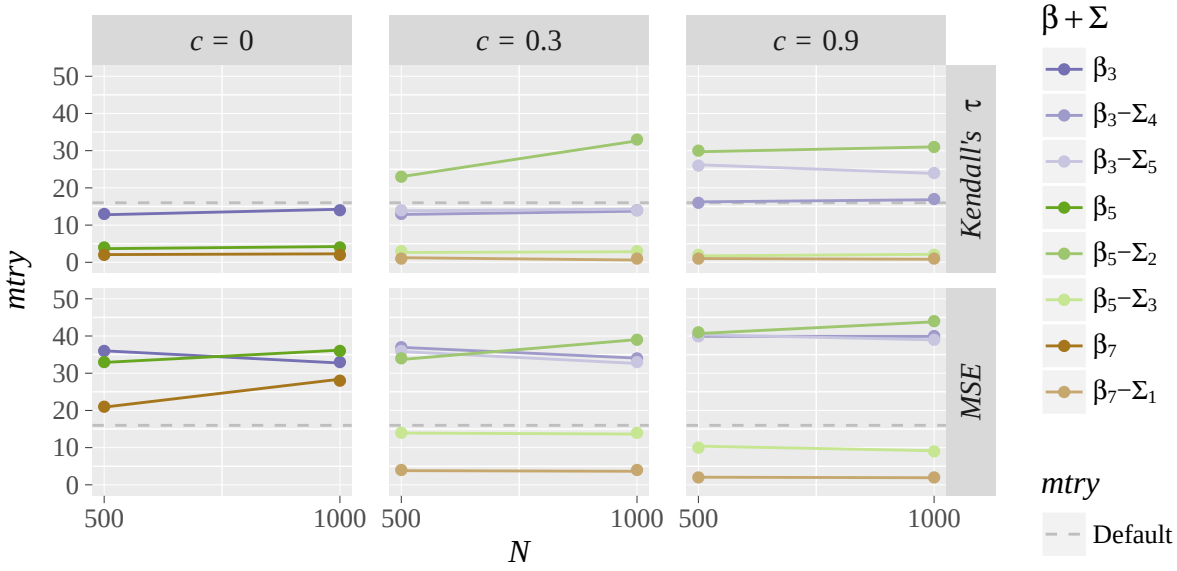


Abbildung A.10: *Optimale mtry Werte für alle Regressionsszenarien mit $p = 50$ und Kovarianzmatrizen $\boldsymbol{\Sigma}_1$ bis $\boldsymbol{\Sigma}_5$.*

A.6.2 Klassifikation

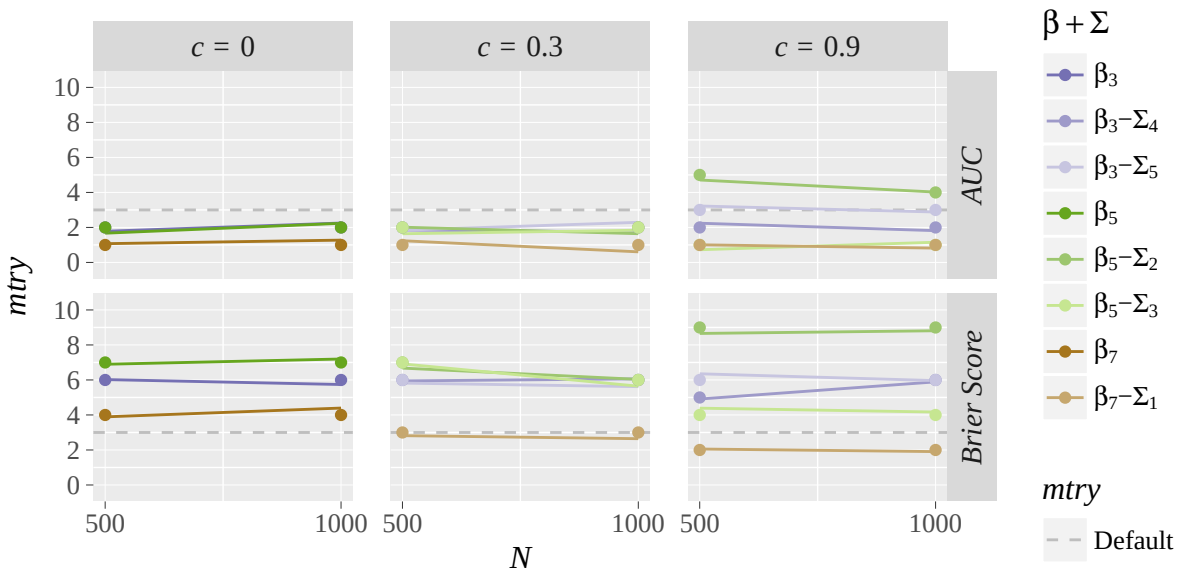


Abbildung A.11: *Optimale mtry Werte für Klassifikationsszenarien mit* $p = 10$ *und Kovarianzmatrizen* $\mathbf{\Sigma}_1$ *bis* $\mathbf{\Sigma}_5$.

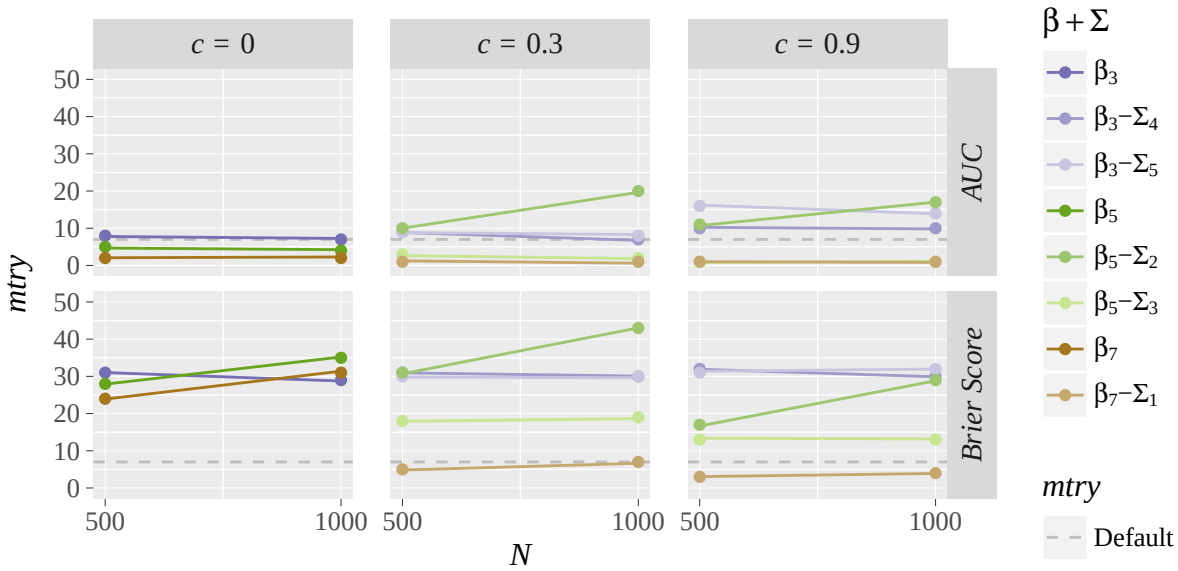


Abbildung A.12: *Optimale mtry Werte für alle Klassifikationsszenarien mit* $p = 50$ *und Kovarianzmatrizen* $\mathbf{\Sigma}_1$ *bis* $\mathbf{\Sigma}_5$.

A.7 Weitere Ergebnisse für optimale mtry Werte mit korrelierten Kovariablen (Σ_6 bis Σ_8)

Nachfolgend jeweils für alle Regressions- und Klassifikationsszenarien mit $p = 10$ und $p = 50$ blockkorrelierten Kovariablen (Σ_6 bis Σ_8) die optimalen *mtry* Werte, getrennt nach den verwendeten Performancemaßen und Korrelationen $c \in \{0.3, 0.6, 0.9\}$. In jeder Grafik sind die optimalen *mtry* Werte für die analogen Szenarien ohne korrelierte Kovariablen ergänzt. Die Definitionen der Kovarianzmatrizen sind in Tabelle 3.3 zusammengefasst.

A.7.1 Regression

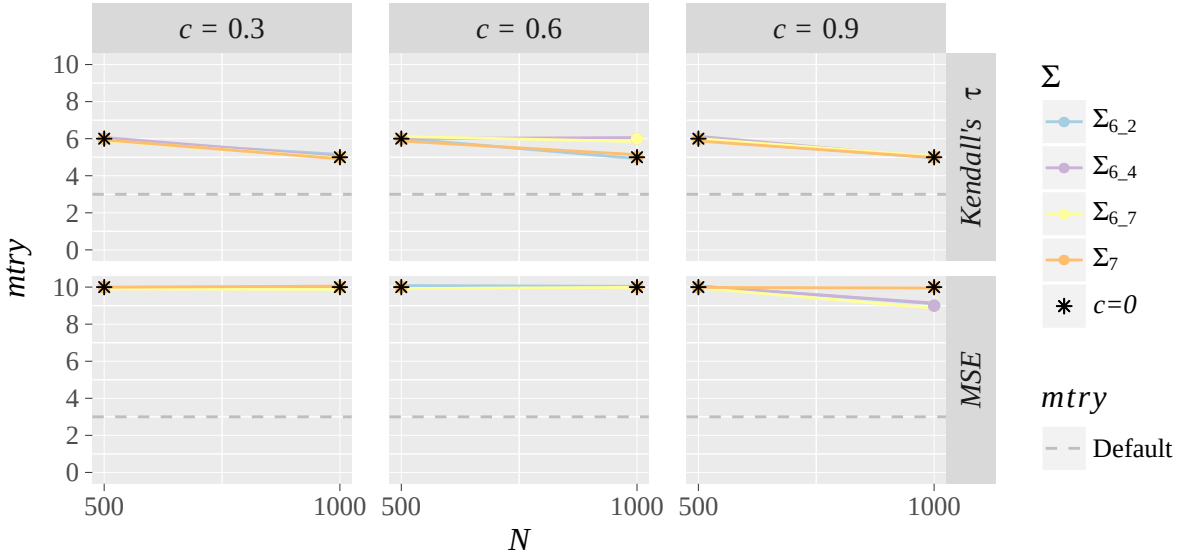


Abbildung A.13: *Optimale mtry Werte für Regressionsszenarien mit $\beta_1 = (7, 0, \ldots, 0)$, $p = 10$ Kovariablen und Kovarianzmatrizen Σ_6 und Σ_7.*

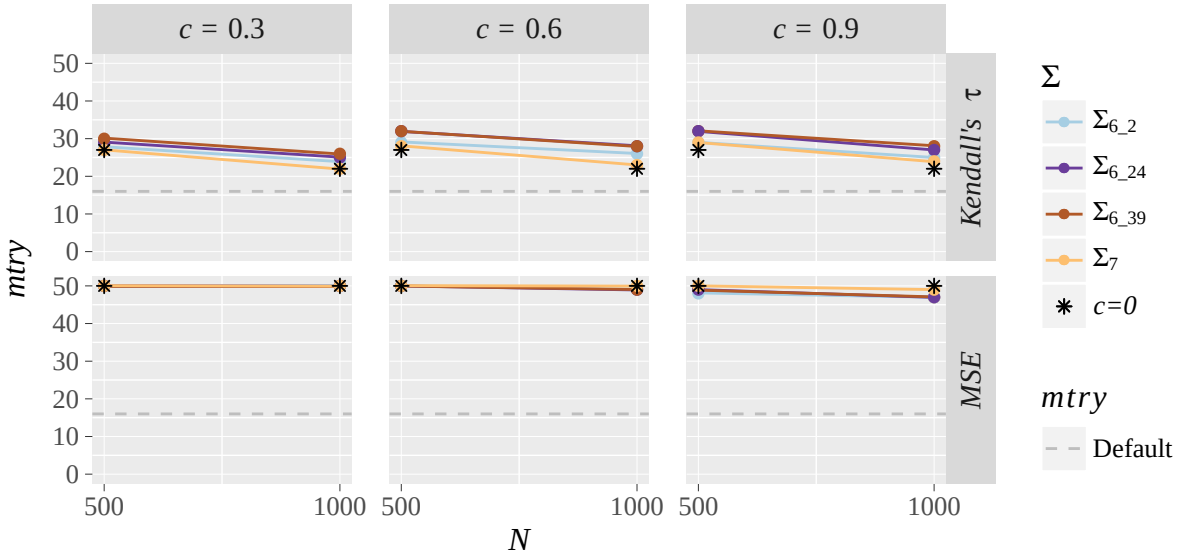


Abbildung A.14: *Optimale mtry Werte für Regressionsszenarien mit $\beta_1 = (7, 0, \ldots, 0)$, $p = 50$ Kovariablen und Kovarianzmatrizen Σ_6 und Σ_7.*

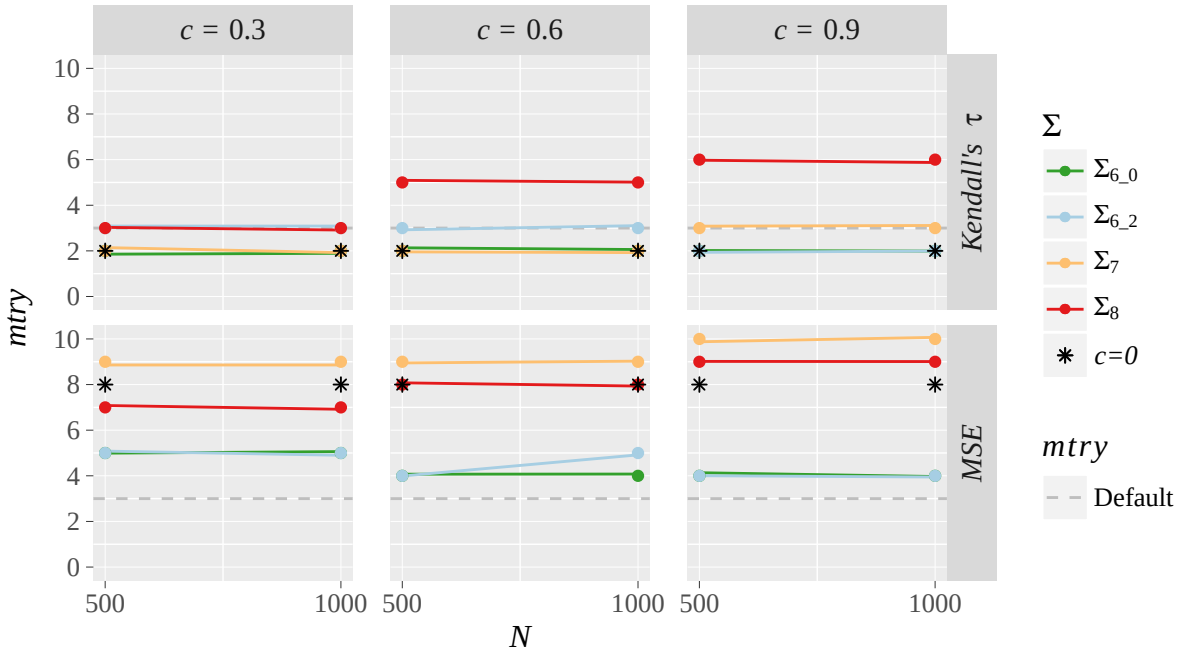


Abbildung A.15: *Optimale mtry Werte für Regressionsszen. mit $\boldsymbol{\beta}_4 = (7, 7, 7, 7, 7, 0, \ldots, 0)$, $p = 10$ Kovariablen und Kovarianzmatrizen $\boldsymbol{\Sigma}_6$ bis $\boldsymbol{\Sigma}_8$.*

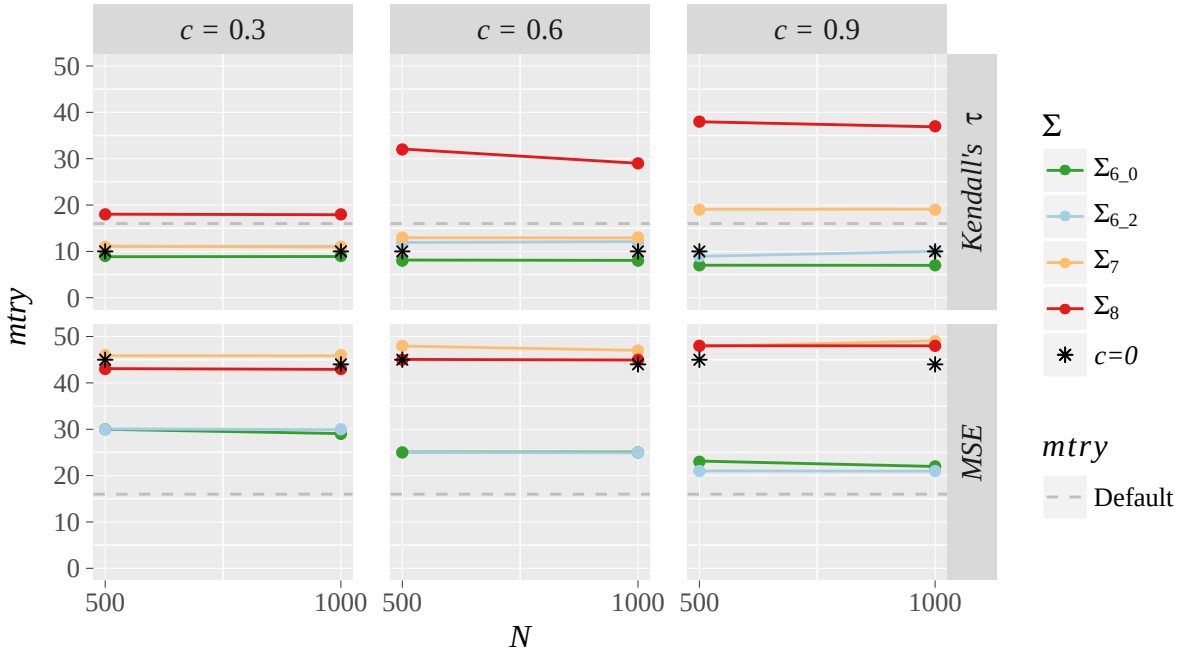


Abbildung A.16: *Optimale mtry Werte für Regressionsszenarien mit $\boldsymbol{\beta}_4 = (7, 7, 7, 7, 7, 0, \ldots, 0)$, $p = 50$ Kovariablen und Kovarianzmatrizen $\boldsymbol{\Sigma}_6$ bis $\boldsymbol{\Sigma}_8$.*

A.7.2 Klassifikation

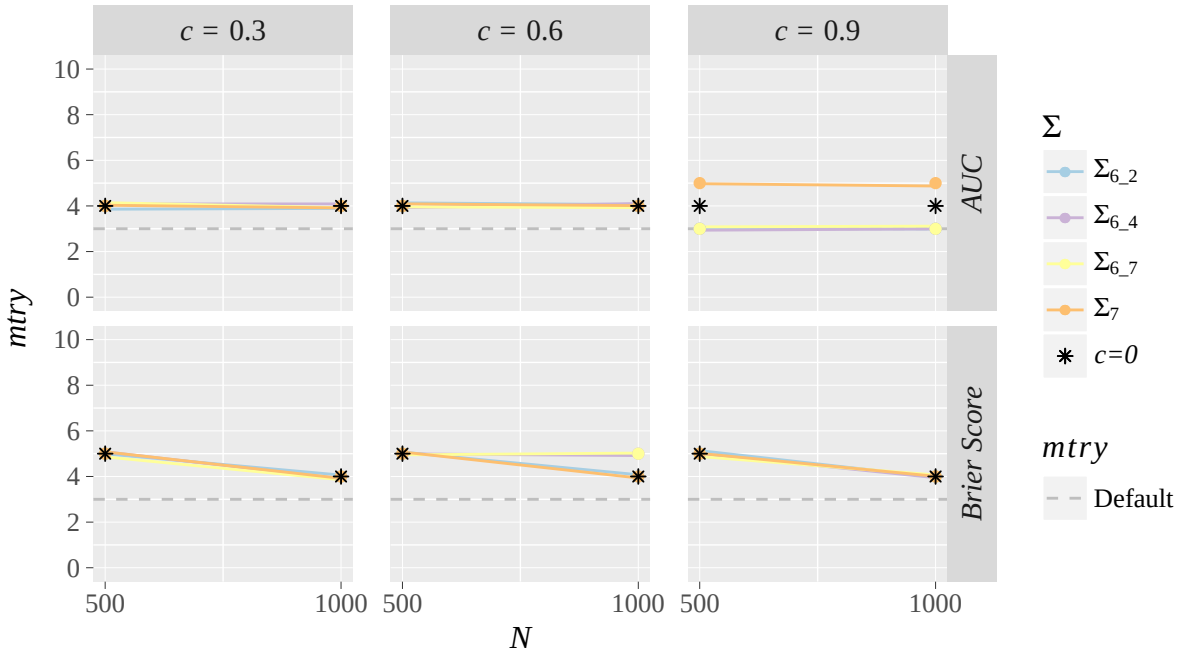


Abbildung A.17: *Optimale mtry Werte für Klassifikationsszenarien mit* $\boldsymbol{\beta}_1 = (7, 0, \ldots, 0)$, $p = 10$ *Kovariablen und Kovarianzmatrizen* $\boldsymbol{\Sigma}_6$ *und* $\boldsymbol{\Sigma}_7$.

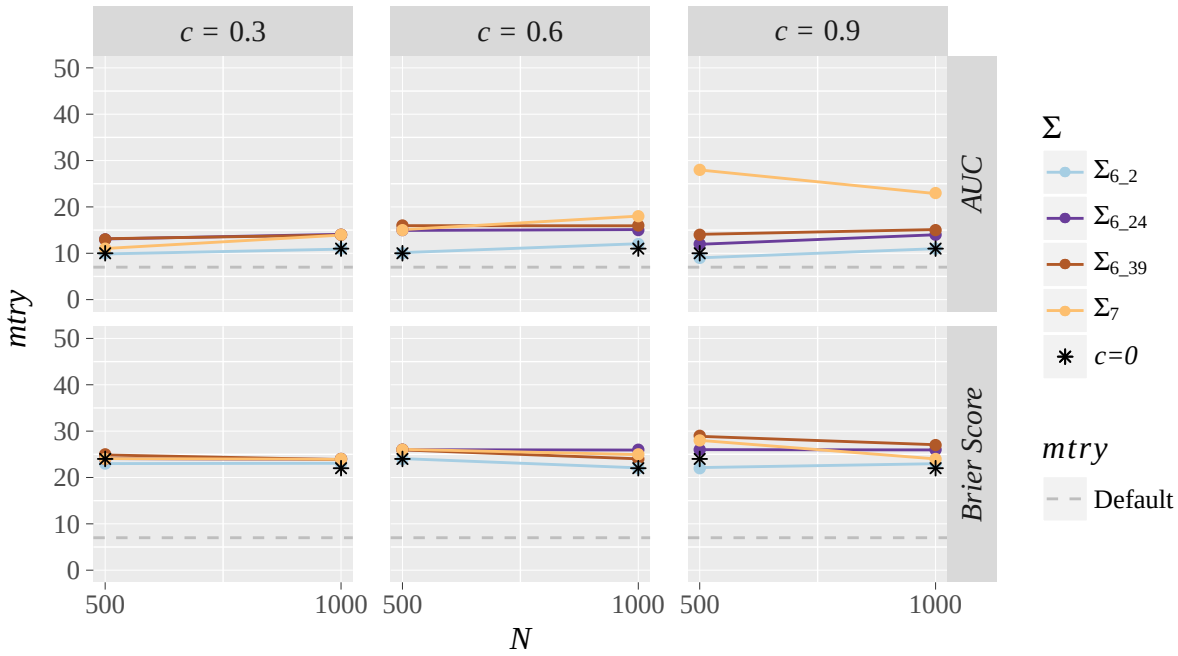


Abbildung A.18: *Optimale mtry Werte für Klassifikationsszenarien mit* $\boldsymbol{\beta}_1 = (7, 0, \ldots, 0)$, $p = 50$ *Kovariablen und Kovarianzmatrizen* $\boldsymbol{\Sigma}_6$ *und* $\boldsymbol{\Sigma}_7$.

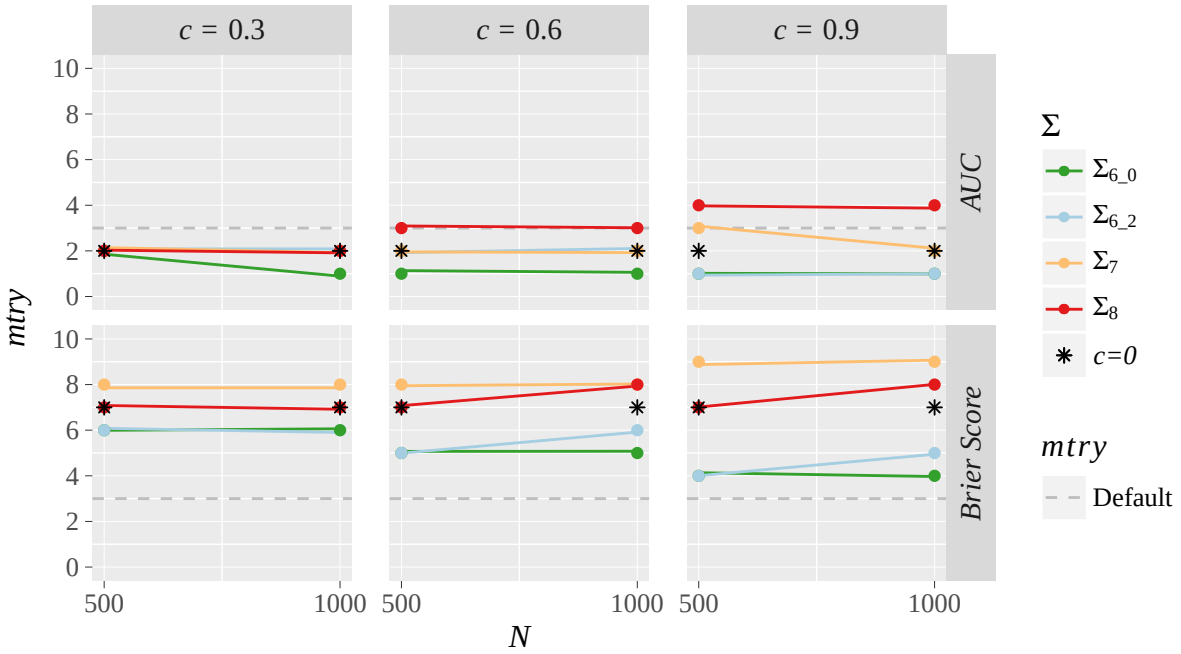


Abbildung A.19: *Optimale mtry Werte für Klassifikationsszenarien mit $\boldsymbol{\beta}_4 = (7, 7, 7, 7, 7, 0, \ldots, 0)$, $p = 10$ Kovariablen und Kovarianzmatrizen $\boldsymbol{\Sigma}_6$ bis $\boldsymbol{\Sigma}_8$.*

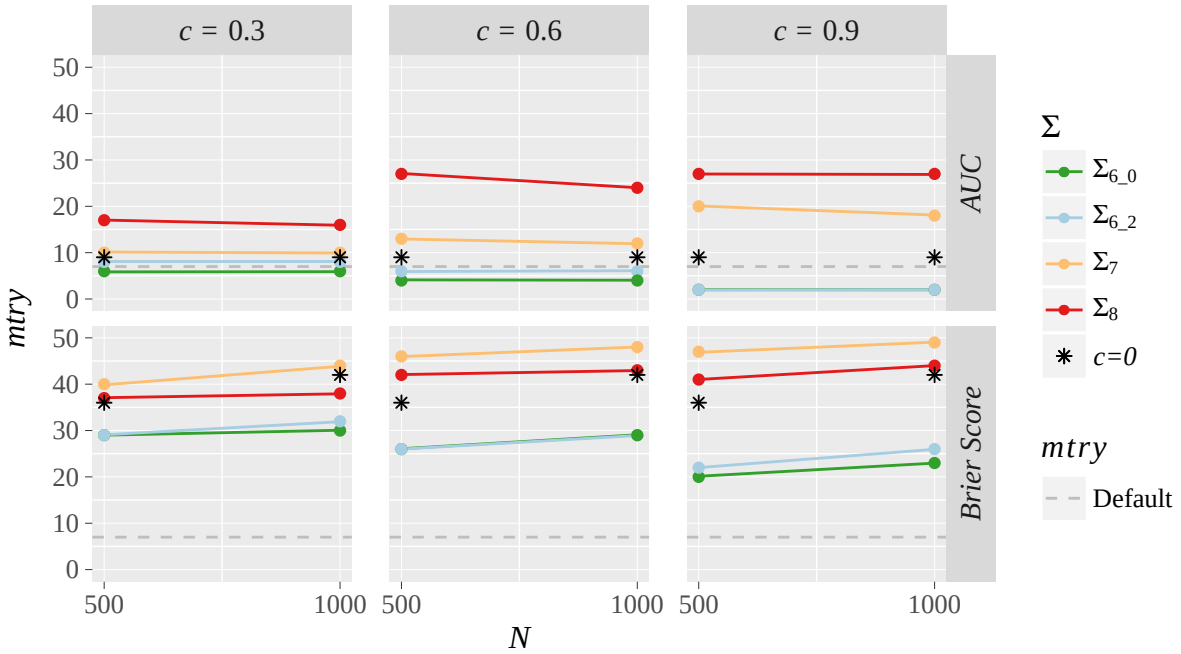


Abbildung A.20: *Optimale mtry Werte für Klassifikationsszenarien mit $\boldsymbol{\beta}_4 = (7, 7, 7, 7, 7, 0, \ldots, 0)$, $p = 50$ Kovariablen und Kovarianzmatrizen $\boldsymbol{\Sigma}_6$ bis $\boldsymbol{\Sigma}_8$.*

A.8 Variablenwichtigkeiten der Klassifikationsszenarien mit den Kovarianzmatrizen Σ_1 bis Σ_8

Die folgenden Abbildungen stellen beispielhaft die Variablenwichtigkeiten der 500 Wiederholungen verschiedener Klassifikationsszenarien mit $N = 500$, 20 Kovariablen und den Kovarianzmatrizen $\mathbf{\Sigma}_1$ bis $\mathbf{\Sigma}_8$ dar. Die dabei jeweils mit $c = 0.9$ blockkorrelierten Kovariablen sind im Titel gekennzeichnet. Die verschiedenen *mtry* Werte je Szenario entsprechen dem Default und den optimalen *mtry* für die Performancemaße *AUC* und *Brier Score*.

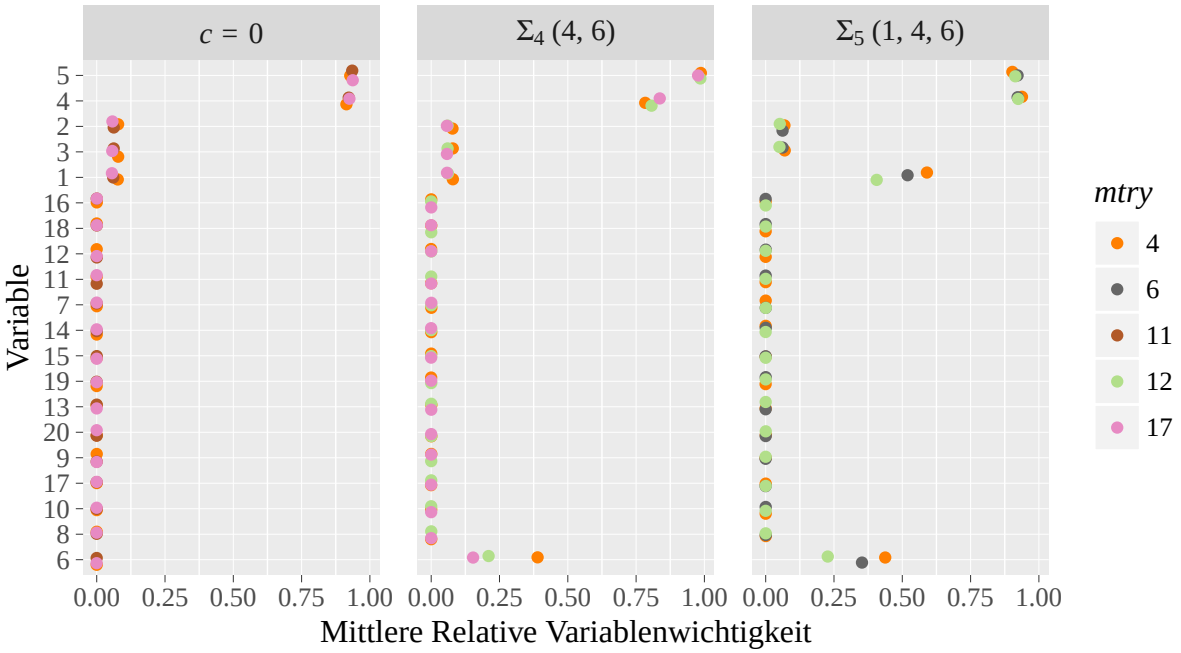


Abbildung A.21: *Mittlere relative Permutation Importance der Klassifikationsszenarien mit* $\boldsymbol{\beta}_3 = (7, 7, 7, 20, 20, 20, 0, \ldots, 0)$, $\mathbf{\Sigma}_4$ *und* $\mathbf{\Sigma}_5$.

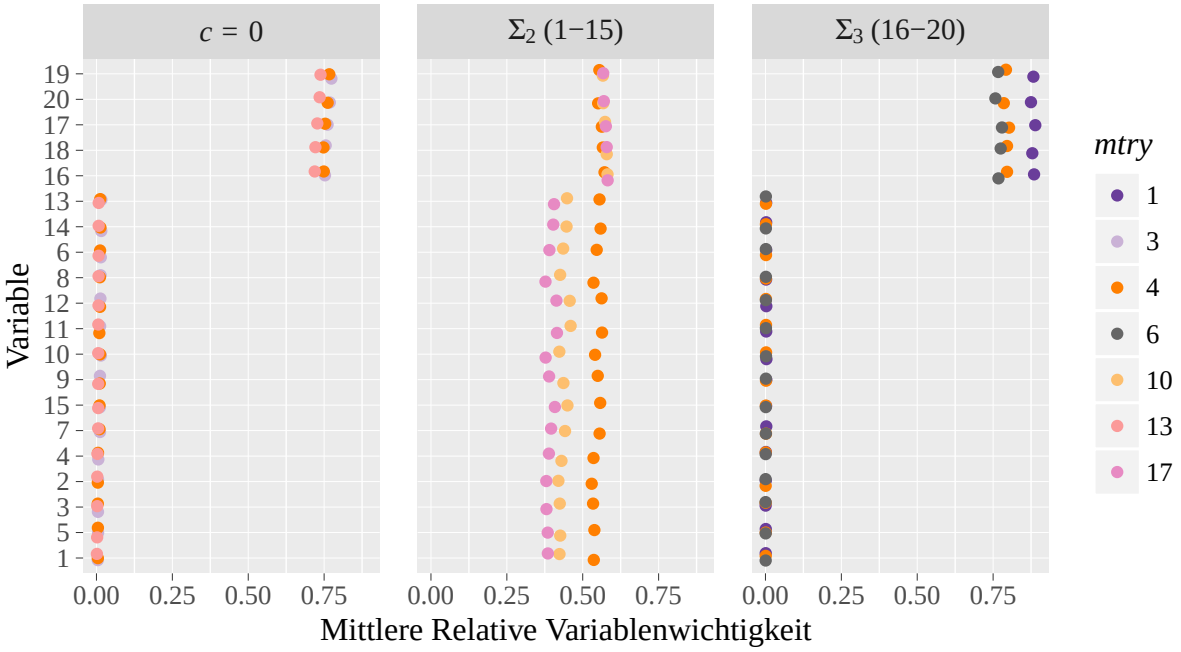


Abbildung A.22: *Mittlere relative Permutation Importance der Klassifikationsszenarien mit* $\boldsymbol{\beta}_5 = (2, \ldots, 2, 3, \ldots, 3, 18, \ldots, 18)$, $\mathbf{\Sigma}_2$ *und* $\mathbf{\Sigma}_3$.

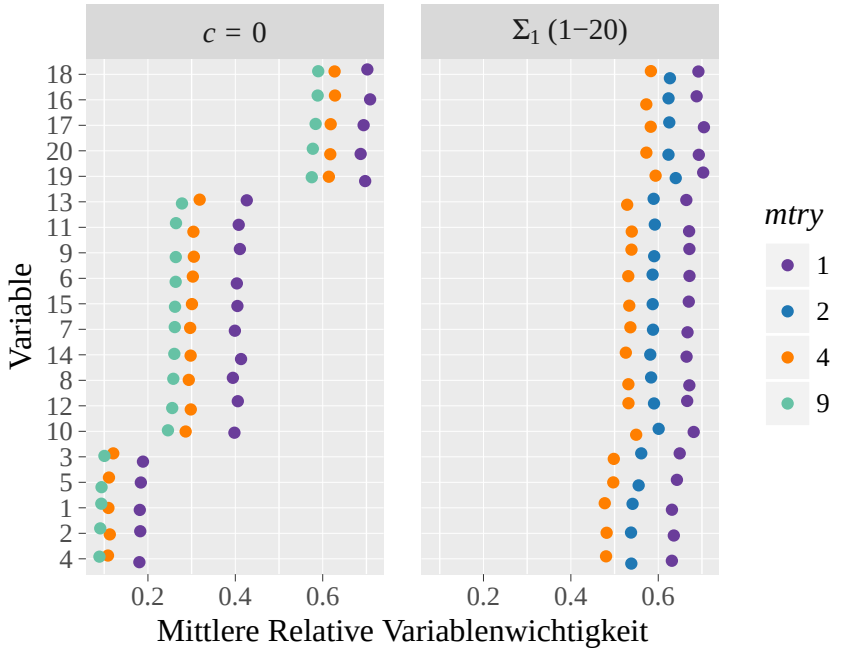


Abbildung A.23: *Mittlere relative Permutation Importance der Klassifikationsszenarien mit* $\boldsymbol{\beta}_7 = (2, \ldots, 2, 3, \ldots, 3, 4, \ldots, 4)$ *und* $\boldsymbol{\Sigma}_1$.

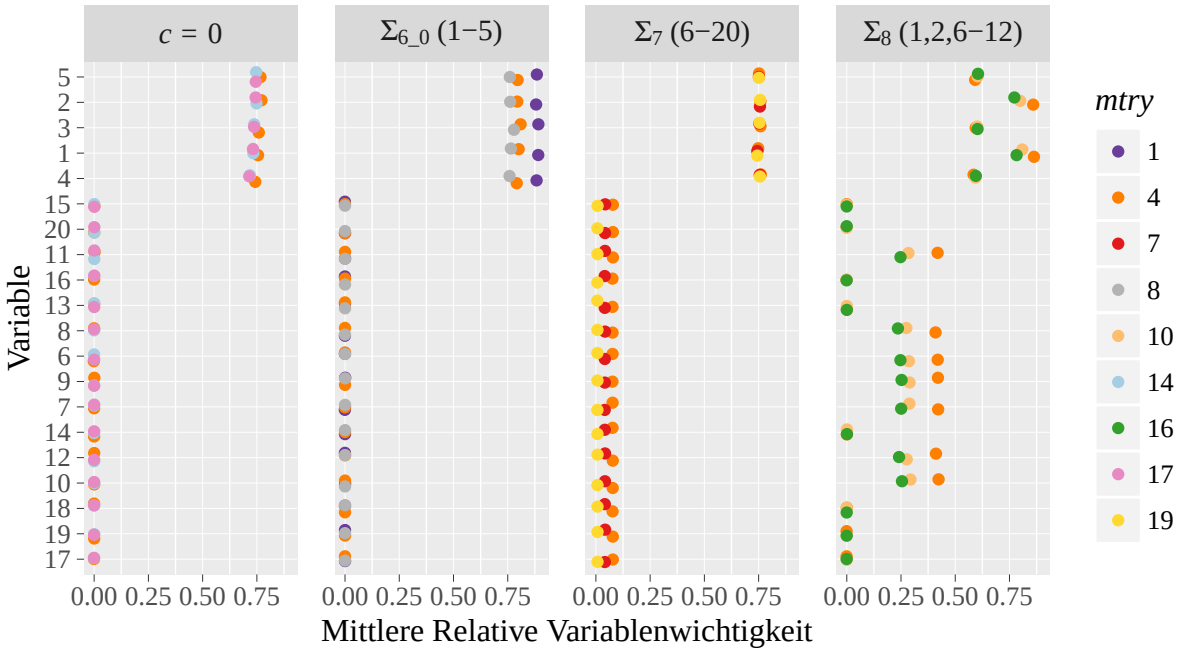


Abbildung A.24: *Mittlere relative Permutation Importance der Klassifikationsszenarien mit* $\boldsymbol{\beta}_4 = (7, 7, 7, 7, 7, 0 \ldots, 0)$ *und Kovarianzmatrizen* $\boldsymbol{\Sigma}_6$ *bis* $\boldsymbol{\Sigma}_8$.

A.9 Korrelierten Kovariablen: Vergleich zweier Variablenwichtigkeiten

In der folgenden Abbildung wird die Conditional Importance aus einem Conditional Inference Forest (**R**-Funktion cforest aus dem Package `party` (Strobl et al., 2008, Version 1.2-2)) mit der Permutation Importance aus einem CART-Forest (**R**-Funktion ranger) verglichen.
Der Conditional Inference Forest liefert vor allem mit $c = 0.9$ deutlich geringere Variablenwichtigkeiten für die korrelierten Kovariablen, die näher an der Relevanz der Kovariablen im unkorrelierten Szenario liegen. Das ist zwar für die Kovariablen 1 und 6 angemessen, jedoch ist auch die Variablenwichtigkeit der korrelierten, stark relevanten Kovariable 4 deutlich gesunken. Damit enspricht die Relevanz von Kovariable 4 nicht mehr der Relevanz der Kovariable 5, trotz gleicher Koeffizientenausprägung.

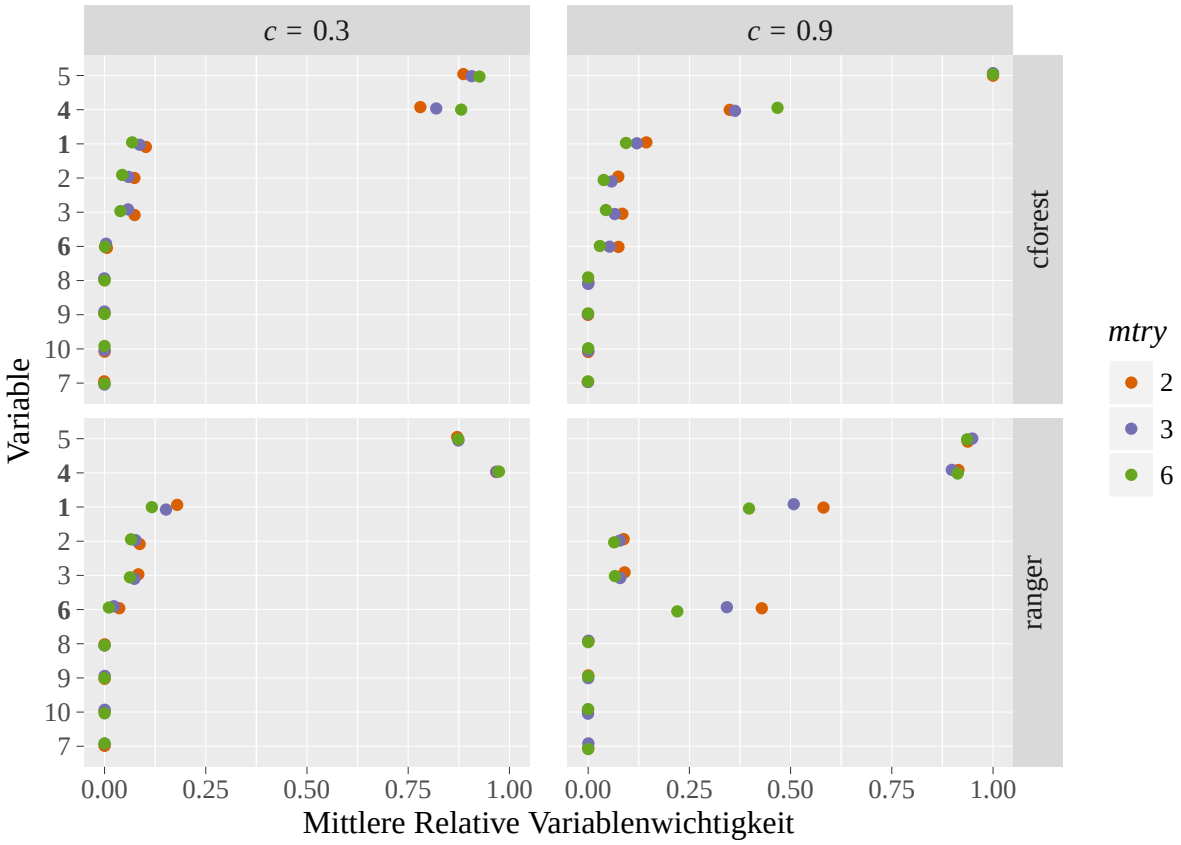


Abbildung A.25: *Vergleich der Conditional Importance aus einem Conditional Inference Forest mit der Permutation Importance aus einem CART-Forest.*
Dabei wurden zwei Szenarien mit folgender Spezifikation verwendet: Binärer Response, $p = 10$, $N = 500$, $\boldsymbol{\beta}_3 = (7, 7, 7, 20, 20, 0, \ldots, 0)$, Kovarianzmatrix $\boldsymbol{\Sigma}_5$ mit $c \in (0.3, 0.9)$. Die dabei blockkorrelierten Kovariablen sind durch Fettdruck an der y-Achse gekennzeichnet.

A.10 Korrelationsplots aller stetigen Variablen der Anwendungsbeispiele

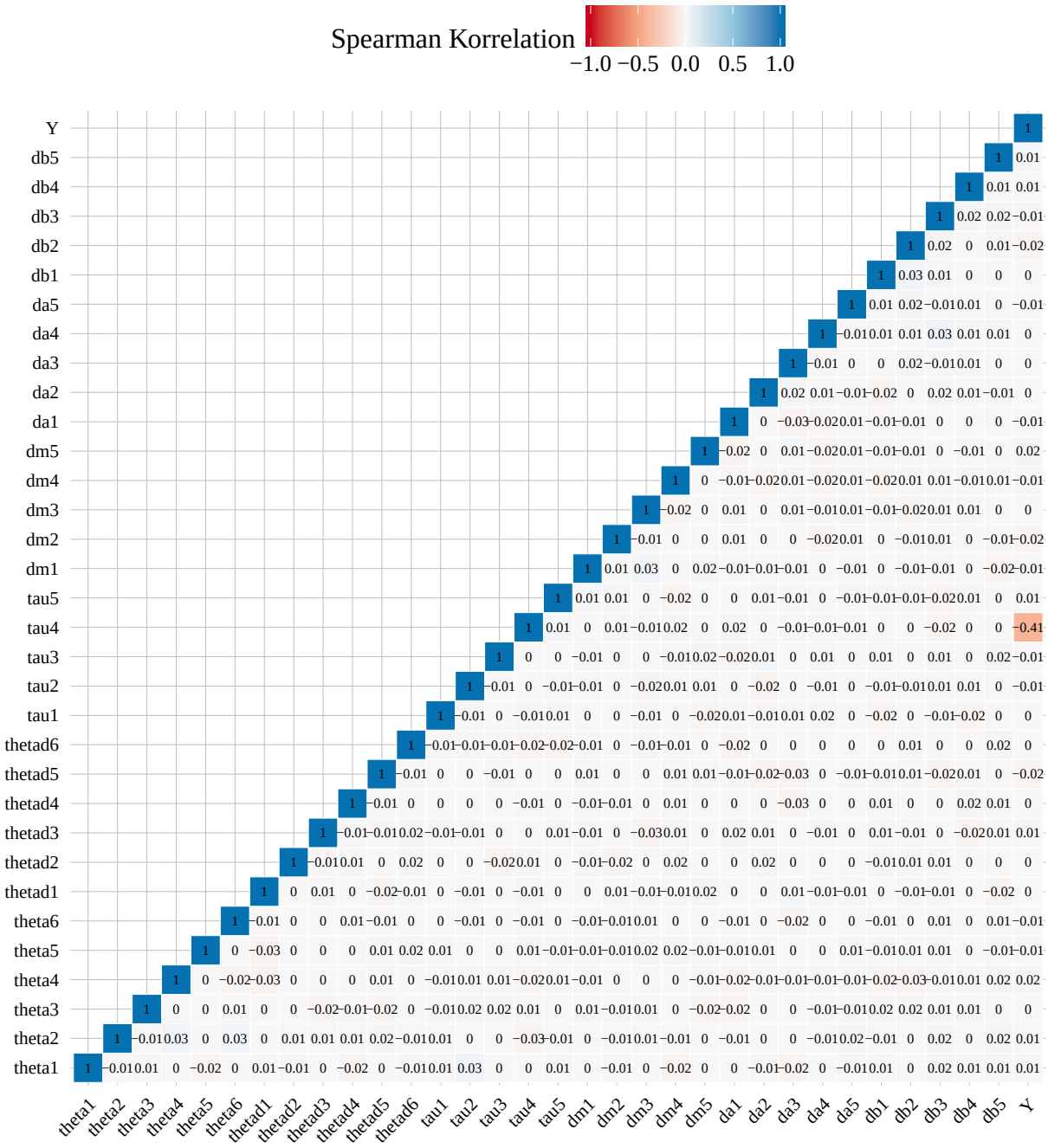


Abbildung A.26: *Korrelationsplot der stetigen Kovariablen und des Responses Y des puma32H Datensatzes.*

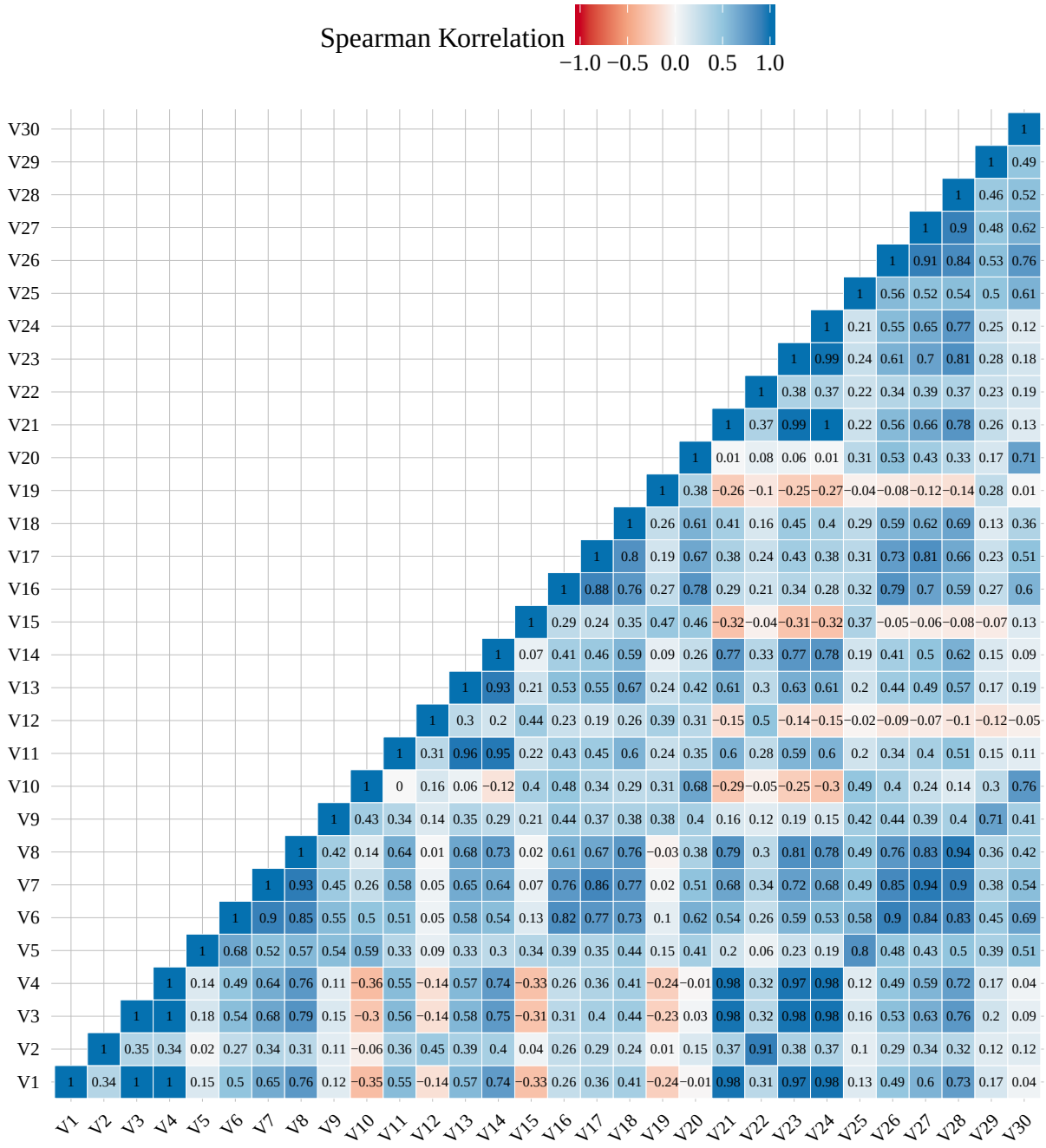


Abbildung A.27: *Korrelationsplot der stetigen Kovariablen des wdbc Datensatzes.*

B Elektronischer Anhang

Der elektronische Anhang umfasst neben einer Readme-Datei die drei Ordner *„Ergebnisse“*, *„Masterarbeit“* und *„Skripte“*:

- **Ergebnisse:** Enthält einige Unterordner.
  - **Simulationsordner 01 bis 09:** Für jedes Simulationsskript gibt es einen Unterordner mit den jeweiligen rds-Dateien der Ergebnisse. Die Ordner haben dabei die gleiche Nummerierung, wie die entsprechenden Simulationsskripte. Die Dateibenennung ist im jeweiligen Simulationsskript am Anfang beschrieben.
  - **Spezifikationen:** Enthält rds-Dateien mit den jeweiligen Spezifikationen der Szenarien für die Simulationsstudie.
  - **Zusätzliche_Grafiken:** Enthält pdf-Dateien mit Abbildungen zu einigen OOB-Kurven, die nicht in dieser Arbeit verwendet wurden.
- **Masterarbeit:** Enthält die vorliegende Arbeit im pdf-Format und einen Unterordner, mit allen in dieser Arbeit verwendeten Abbildungen im pdf-Format.
- **Skripte:** Enthält drei Unterordner mit Syntax-Dateien der Statistiksoftware **R**.
  - **Funktionen**: Enthält Skripte, in denen die Koeffizientenvektoren und Kovarianzmatrizen aus den Tabellen 3.1, 3.2 und 3.3 definiert werden, sowie Funktionen, die während der Simulation aufgerufen werden.
  - **Grafiken**: Enthält Skripte, mit denen die Abbildungen in dieser Arbeit reproduziert werden können. Die Dateien haben dabei die gleiche Nummerierung, wie die entsprechende Simulationsskripte, mit denen die Ergebnisse erstellt worden sind.
  - **Simulationen**: Enthält Skripte, mit denen die Ergebnisse der Simulationsstudie aus dieser Arbeit reproduziert werden können. Die Dateien sind nach der Reihenfolge ihrer Ausführung von 01 bis 09 nummeriert.

Eidesstattliche Erklärung

Ich versichere hiermit, dass ich die vorliegende Masterarbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

München, 30. Januar 2018

Myriam Hatz