Institut für Statistik
Ludwig-Maximilians-Universität München

Erweiterung eines Testansatzes für die Variablenwichtigkeit in Random Forests

Master - Thesis

im Studiengang Biostatistik

Ender Celik

Betreuerin: M.Sc. Janitza Silke

Begutachterin: Prof. Anne-Laure Boulesteix

25.09.2015

Inhaltsverzeichnis

1. Einführung

Die *Random Forest* (RF) Methode ist eine nicht-parametrische Regressions- bzw. Klassifikationsmethode. *Random Forests* bestehen aus einer großen Menge (Ensemble) von Entscheidungsbäumen und eine Aggregation von den einzelnen Vorhersagen wird durchgeführt. Die einzelnen Entscheidungsbäume werden mit dem *C*lassification *a*nd *R*egression *T*rees - Algorithmus ($CART$) erzeugt. Man spricht von einem Klassifikationsbaum bzw. einer Klassifikationsmethode, wenn die abhängige Variable kategorial und von einem Regressionsbaum bzw. einer Regressionsmethode, wenn die abhängige Variable metrisch skaliert ist. In dieser Arbeit steht die Regressionsmethode mit dem *Random Forests*, im Vordergrund der Betrachtung.

Die *Random Forest* Methode wurde von Breiman im Jahre 2001 das erste Mal vorgestellt und erlangte in letzter Zeit in der Biostatistik und vielen anderen wissenschaftlichen Bereichen immer mehr Popularität. Die Gründe für die Popularität sind: die hohe Vorhersagegenauigkeit und dass anwendbar sind, wenn die Anzahl der Kovariablen viel größer ist als die Anzahl der Beobachtungen als auch, wenn die Kovariablen miteinander korreliert sind. Deshalb sind *Random Forests* so beliebt bei hochdimensionalen genetischen Daten (vgl. Strobl u. a., 2008 und Janitza u. Boulesteix (2015)).

Durch die große Anzahl der Entscheidungsbäume im *Random Forest* sind die Ergebnisse, im Vergleich zu einem Regressionsbaum, nicht so leicht zu interpretieren. In vielen Anwendungen interessiert man sich nicht nur für eine genaue Vorhersage, sondern gleichzeitig auch für die Identifizierung möglicher relevanter Kovariablen. Zum Beispiel ist bei der statistischen Genforschung die Identifizierung relevanter Gene von großem Interesse. Diese relevanten Gene können wertvolle Einblicke in die komplexen Mechanismen von bestimmten Erkrankungen liefern. Das *Random Forest* Verfahren berechnet zur Identifizierung möglicher relevanter Kovariablen ein sogenanntes Variablenwichtigkeitsmaß (engl. *variable importance measures* (VIM)) (vgl. Strobl u. a., 2008 und Janitza u. Boulesteix 2015).

Es gibt zwei häufig verwendete Variablenwichtigkeitsmaße: die mittlere Abnahme der Summe der quadratischen Abstände (engl. mean decrease in node impurity bzw. average impurity reduction) und die Permutation Variablenwichtigkeit (engl. permutation importance bzw. mean decrease in accuracy). In dieser Arbeit wird das Permutation Variablenwichtigkeitsmaß genauer betrachtet.

Wenn eine Kovariable keinen wesentlichen Beitrag zur Verbesserung der Prädiktion von der Zielvariable liefert, nimmt das Permutation Variablenwichtigkeitsmaß einen negativen Wert oder Werte nahe Null an. Nimmt das Permutation Variablenwichtigkeitsmaß einen positiven Wert für eine Kovariable an, bedeutet das, dass diese Kovariable einen Beitrag

zur Verbesserung der Prädiktion von der Zielvariable liefert. Allerdings kann keinesfalls angenommen werden, dass ein positiver Wert für das Permutation Variablenwichtigkeitsmaß immer auf eine relevante Kovariable verweist, da nicht festgestellt werden kann, ob der positive Wert reiner Zufall ist. Es stellt sich die Frage, ob sich ein positiver Wert für das Permutation Variablenwichtigkeitsmaß signifikant von Null unterscheidet (vgl. Janitza u. Boulesteix, 2015, S.4f).

Es werden in in dieser Arbeit drei heuristische Testansätze für hochdimensionale Daten vorgestellt, die diese Frage beantworten versuchen. Der erste heuristische Testansatz auf den in der Arbeit genauer eingegangen wird, ist ein auf Permutation basierter Test von Altmann u. a. (2010) der *p*ermutation *imp*ortance ($PIMP$) Testansatz. Ein Nachteil dieses Testansatzes ist, dass es eine sehr rechenintensive (computationale) Methode ist. Dagegen ist der neue heuristische Testansatz (engl. *novel testing approach* (NTA)) von Janitza u. Boulesteix (2015), der auf den beobachteten Variablenwichtigkeitsmaßen beruht, sehr schnell und wenig rechenintensiv. Janitza u. Boulesteix entwickelten für diesen NTA Testansatz ein neues bzw. alternatives Permutation Variablenwichtigkeitsmaß. Das erste Ziel dieser Arbeit war es dieses heuristische Testverfahren in R zu implementieren. Dafür wurde ein neues R-paket namens *vita* (variable importance testing approaches) erstellt (vgl. Anhang B). Bei der Implementation und beim Testen der Testansätze, entstand die Idee zu einem neuen heuristischen Testansatz „Null-Verteilung Approximation“ (engl. Null distribution approximation ($NuDA$)). Der $NuDA$ Testansatz ist wie der $PIMP$ Testansatz ein permutationsbasierter Test, der aber weniger rechenintensiv als der $PIMP$ Testansatz ist.

Das Hauptaugenmerk dieser Arbeit liegt in der Untersuchung dieser heuristischen Testansätze für hochdimensionale genetische Daten. Es wird zum einen untersucht, inwiefern die heuristischen Testansätze das vorgegebene Signifikanzniveau α einhalten und zum anderen wie groß die Power dieser Testansätze ist. Um diese Analysen durchführen zu können, muss bekannt sein, welche Kovariable einen Effekt bzw. keinen Effekt auf die Zielvariable hat. Da bei realen Daten diese Information nicht vorliegt, werden die Daten simuliert.

In Kapitel 2 wird es zunächst eine Einführung in den *C*lassification *a*nd *R*egression *T*rees - Algorithmus geben. Im nachfolgenden Kapitel werden die Ideen und die Entstehung des *Random Forest* Verfahrens beschrieben. Die verwendeten Variablenwichtigkeitsmaße und die heuristischen Testansätze werden in Kapitel 4 geschildert. Im Kapitel 5 werden die Simulationsstudien erklärt und deren Ergebnisse dargestellt und interpretiert.

2. Regressionsbäume

2.1. Entscheidungsbäume

Wie der Name suggeriert ist ein Entscheidungsbaum ein gerichteter Graph mit einer Baumstruktur, der von oben nach unten verläuft. Ein Entscheidungsbaum besteht aus einem Wurzelknoten (engl. root), der alle Kovariablen beinhaltet und sich an oberster Stelle in mehrere weitere Knoten (engl. *node*) bis zu den Blättern (engl. *leaf*) aufteilt. Der zu teilende Knoten wird als Elternknoten (engl. *parent node*), die aufgespaltenen Knoten werden als Kindknoten (engl. *child node*) bezeichnet. Die Elternknoten werden durch eine Entscheidungsregel bezüglich einer Kovariable in zwei Kindknoten, die in sich möglichst homogen und untereinander möglichst heterogen bezüglich der Zielvariable $\mathbf{y}$ sind, aufgeteilt. Die beiden Kindknoten werden selbst zu Elternknoten und werden in der gleichen Art und Weise in weitere Knoten unterteilt. Diese Prozedur wird solange wiederholt, bis einige Stoppregeln, die im nächsten Unterkapitel genauer beschrieben werden, angewendet werden. Ein Entscheidungsbaum ist in Abbildung 2.1 links schematisch dargestellt. Man spricht von einem Klassifikationsbaum, wenn die abhängige Variable kategorial und von einem Regressionsbaum, wenn die abhängige Variable metrisch skaliert ist. Es gibt einige Algorithmen, die automatisch Entscheidungsbäume erzeugen. Die wohl wichtigsten Algorithmen sind *Chi*-square *A*utomatic *I*nteraction *D*etectors ($CHAID$), $C4.5$ und der *C*lassification *a*nd *R*egression *T*rees - Algorithmus ($CART$) (vgl. Steiner, 2009, S. 65ff). In dieser Arbeit stehen binäre Regressionsbäume, die mit dem CART-Algorithmus erzeugt werden, im Vordergrund der Betrachtung.

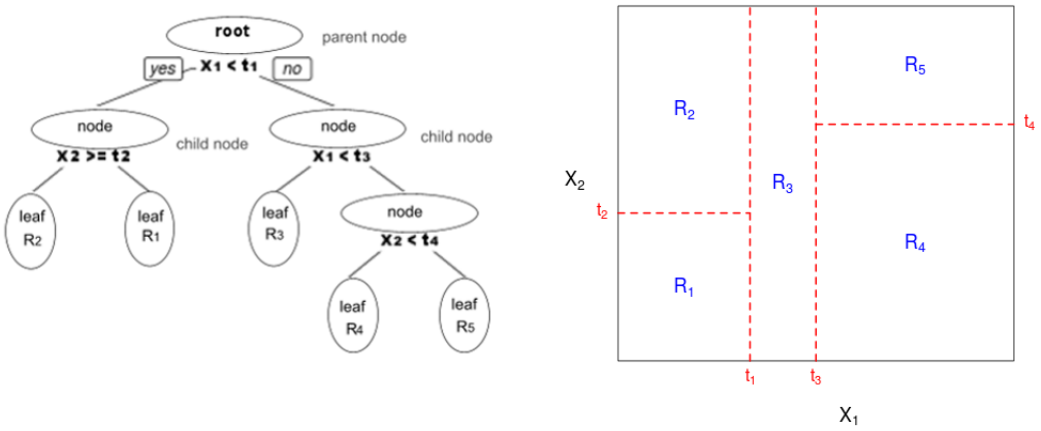

Abbildung 2.1: Entscheidungsbaum und Zerlegungen des Kovariablenraums. Links ist eine schematische Darstellung eines Entscheidungsbaums mit den dazu gehörigen Begriffen und rechts die dazugehörige Zerlegung des Kovariablenraums. (Hastie u. a., 2013, S.306).

2.2. *C*lassification *a*nd *R*egression *T*rees

Der *C*lassification *a*nd *R*egression *T*rees - Algorithmus wurde von den Statistikern Breiman, Friedman, Olshen, u. Stone (1984) entwickelt und erzeugt binäre Entscheidungsbäume. Die Lerndaten $\mathcal{L} = \{(y_i, \mathbf{x}_i), i = 1, \ldots, n\}$ beinhalten p Kovariablen $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ und die Zielvariable $\mathbf{y}$ mit n Beobachtungen. Dieser Algorithmus zerlegt den Kovariablenraum ($K \subset \mathbb{R}^p$) in p-dimensionale Hyperrechtecke. Der Algorithmus zerlegt zum Beispiel einen zweidimensionalen Kovariablenraum in Rechtecke, wie in Abbildung 2.1 rechts dargestellt. Oder anders ausgedrückt, der Kovariablenraum wird in disjunkte Untermengen aufgeteilt (vgl. Fahrmeir u. a., 1996, S. 425 ff und vgl. Hastie u. a., 2013, S. 305 f).
Der CART-Algorithmus entscheidet automatisch nach welcher Kovariable geteilt werden muss und bestimmt den Splitpunkt. Im folgenden Beispiel wird angenommen, dass eine Zerlegung des Kovariablenraums in M p-dimensionale Hyperrechtecke R_m, $m = 1, \ldots, M$ besteht und die Zielvariable y_i in jeder Region als Konstante $c_m$modelliert wird:

$$f(\mathbf{x}_i) = \sum_{m=1}^{M} c_m I\left(\mathbf{x}_i \in R_m\right) , \tag{2.1}$$

wobei $I\left(.\right)$ die Indikatorfunktion und c_m eine konstante Funktion in der jeweiligen Region R_m sind. Für die Beurteilung der Modellanpassung für einen Regressionsbaum T wird die Summe der quadratischen Abstände (eng. sum of squared errors (SSE))

$$SSE(T) = \sum_{i=1}^{n} \left(y_i - f(\mathbf{x}_i)\right)^2 = \sum_{m=1}^{M} \sum_{i=1}^{n} \left(y_i - c_m I\left(\mathbf{x}_i \in R_m\right)\right)^2 \tag{2.2}$$

verwendet. Der Algorithmus muss die optimale Zerlegung des Kovariablenraums in M p-dimensionale Hyperrechtecke und die dazugehörige Konstante c_m so bestimmen, dass die Summe der quadratischen Abstände minimiert wird. Für eine beliebige Zerlegung des Kovariablenraums in Hyperrechtecke ist die Summe der quadratischen Abstände minimal, wenn für die Konstante c_m der Mittelwert der abhängigen Variable $\mathbf{y}$ in den jeweiligen Region $R_m$eingesetzt wird:

$$\hat{c}_m = \frac{1}{n_{R_m}} \sum_{\mathbf{x}_i \in R_m} y_i \, , \tag{2.3}$$

wobei n_{R_m} die Anzahl der Beobachtungen in der Region R_m ist. Um die beste Zerlegung des Kovariablenraums in M Hyperrechtecke in Bezug auf das Minimum der Summe der quadratischen Abstände zu finden, startet der Algorithmus mit allen Beobachtungen $\mathcal{L}$ ($Wurzelknoten \hat{=} R$) und sucht die beste Split-Kovariable $\mathbf{x_{(j)}} = (x_{1j}, x_{2j}, \ldots, x_{nj})^T$ und den Splitpunkt s. Werden alle Beobachtungen bezüglich einer Kovariable $\mathbf{x_{(j)}}$ am Split-

punkt s aufgeteilt, ergeben sich zwei disjunkte Untermengen:

$$R = R_L(j,s) \cap R_R(j,s)\,,$$

wobei

$$R_L(j,s) = \left\{\mathbf{X}|\mathbf{x_{(j)}} < s\right\} \;\cap\; R_R(j,s) = \left\{\mathbf{X}|\mathbf{x_{(j)}} \geq s\right\}\,. \tag{2.4}$$

Für diese Teilung des Wurzelknoten R in zwei Knoten $R_L,\ R_T$ kann die Abnahme der Summe der quadratischen Abstände folgendermaßen berechnet werden

$$\Delta SSE(R, R_L, R_T) = SSE(R) - \left(SSE\left(R_L(j,s)\right) + SSE\left(R_R(j,s)\right)\right)\,, \tag{2.5}$$

wobei $SSE(R)$ die Summe der quadratischen Abstände im Wurzelknoten R,

$$SSE(R) = \sum_{i=1}^{n} \left(y_i - c_R I\left(\mathbf{x}_i \in R\right)\right)^2$$

und $SSE\left(R_L(j,s)\right)$ bzw. $SSE\left(R_R(j,s)\right)$ die Summe der quadratischen Abstände in den zwei disjunkte Knoten $R_L,\ R_T$,

$$SSE\left(R_k(j,s)\right) = \sum_{i=1}^{n} \left(y_i - c_{R_k} I\left(\mathbf{x}_i \in R_k\right)\right)^2, k = \{L, R\}$$

ist. Die beste Split-Kovariable $\mathbf{x_{(j)}}$ und den dazugehörigen Splitpunkt s wird berechnet über das folgendes Minimierungsproblem:

$$\min_{j,s}\left(\min_{c_L} \sum_{x_i \in R_L(j,s)} \left(y_i - c_L\right)^2 + \min_{c_R} \sum_{x_i \in R_R(j,s)} \left(y_i - c_R\right)^2\right) \tag{2.6}$$

Für eine beliebige Split-Kovariable $\mathbf{x_{(j)}}$ und einen Splitpunkt s wird die innere Minimierung durch

$$\hat{c}_L = \frac{1}{n_{R_L}} \sum_{x_i \in R_L(j,s)} y_i \;\; und \;\; \hat{c}_R = \frac{1}{n_{R_R}} \sum_{x_i \in R_R(j,s)} y_i \tag{2.7}$$

gelöst. Nachdem der Algorithmus die beste Split-Kovariable $\mathbf{x_{(j)}}$ und den dazugehörigen Splitpunkt s gefunden hat, werden die Daten in zwei disjunkte Teile aufgespalten und der Zerlegungsvorgang wird an diesen zwei Teilen wiederholt. Um den CART-Algorithmus besser zu verstehen wird in Abbildung 2.2 Schritt für Schritt die Zerlegung des Kovariablenraums für ein einfaches Regressionsproblem, mit zwei Kovariablen $\mathbf{X} = (\mathbf{x_1}, \mathbf{x_2})$ und der Zielvariable $\mathbf{y}$, dargestellt (vgl. Hastie u. a., 2013, S. 307 f).

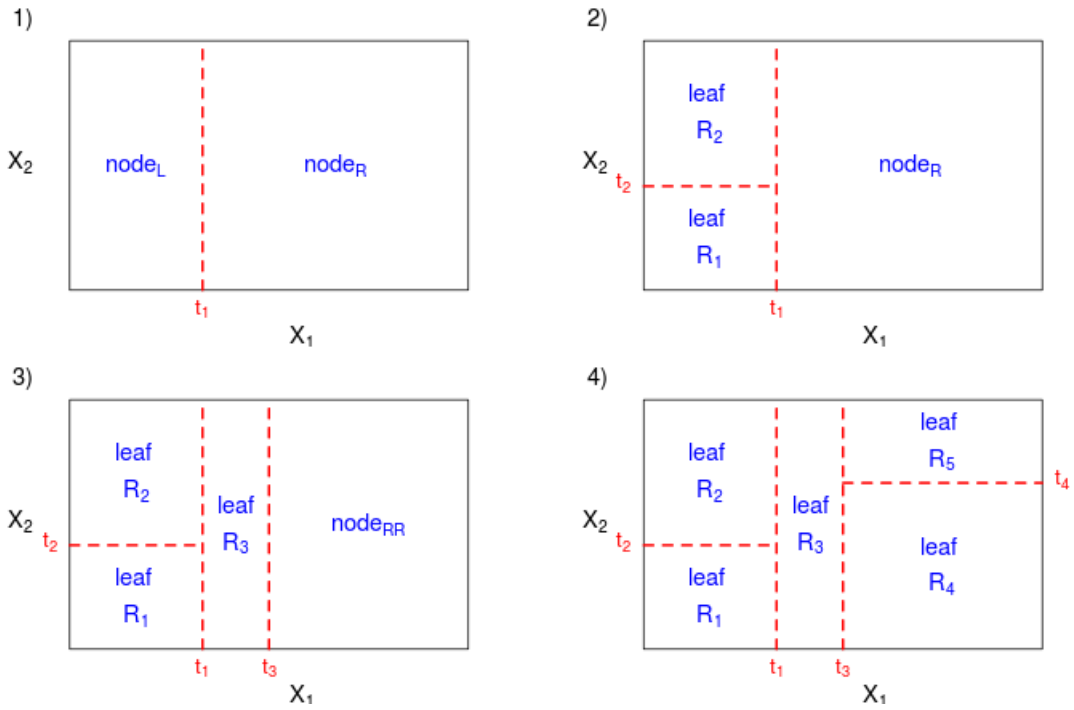

Abbildung 2.2: Schritt für Schritt die Zerlegung des Kovariablenraums

Der Algorithmus findet

1. als erste Split-Kovariable $\mathbf{x_1}$ und den dazugehören Splitpunkt bei t_1 und teilt $\mathbf{x_1}$ bei t_1 in zwei Teile, $node_L$ und $node_R$,
2. im Bereich $\mathbf{x_1} < t_1$ $(node_l)$ wird $\mathbf{x_2}$ bei t_2 in zwei Teile , $leaf\ R_1$ und $leaf\ R_2$, aufgeteilt,
3. im Bereich $\mathbf{x_1} \geq t_1$ $(node_R)$ wird $\mathbf{x_1}$ bei t_3 in zwei Teile , $leaf\ R_3$ und $node_{RR}$, aufgeteilt
4. und zuletzt wird der Bereich $\mathbf{x_1} \geq t_3$ $(node_{RR})$ $\mathbf{x_2}$ bei t_3 in zwei Teile , $leaf\ R_4$ und $leaf\ R_5$, aufgeteilt.

Würde der rekursive CART-Algorithmus den Kovariablenraum immer weiter unterteilen, dann würde in jedem Blatt nur eine Beobachtung übrig bleiben und sich damit eine Überanpassung an die Daten ergeben. Würde man zu früh abbrechen, könnten wichtige Strukturen übersehen werden. Eine Möglichkeit wäre es, die Aufspaltung des Kovariablenraums zu stoppen, sobald die Kindknoten eine bestimmte Anzahl von Beobachtungen beinhalten. Die Baumgröße bestimmt die Komplexität des Modells und ist damit ein Tuning-Parameter. Die optimale Baumgröße sollte aus den beobachteten Lerndaten ausgewählt werden und nicht eine fixe Größe sein. Eine andere mögliche Stoppregel wäre, wenn die Abnahme der Summe der quadratischen Abstände 2.5, bei der Teilung in

zwei neue Knoten, kleiner als ein vorher definierter Schwellenwert ist. Die bevorzugte Methode ist, den Regressionsbaum sehr komplex werden zu lassen und ihnen dann mit der Kosten-Komplexität-Beschneidung (engl. $cost - complexity\ pruning$) zu beschneiden (engl. *Pruning*) (vgl. Hastie u. a., 2013, S. 308). Sind die Kovariablen Kategorial, bevorzugt der CART-Algorithmus solche Kovariablen mit vielen Kategorien(vgl. Strobl u. a., 2005, S.2).

Der große Vorteil von Regressionsbäumen ist, dass die Ergebnisse, durch die graphische Darstellung, einfach zu interpretieren sind. Eine Einschränkung von Regressionsbäumen ist, dass die Vorhersagefunktion eine Treppenfunktion ist. Das Hauptproblem bei Entscheidungsbäumen ist, dass kleine Änderungen in den Lerndaten zu komplett anderen Zerlegungen des Kovariablenraums führen können und ein komplett anderer Entscheidungsbaum entsteht. Der Hauptgrund für diese Instabilität, bzw. hohe Varianz, ist die hierarchische Struktur des Verfahrens. Ein Fehler in der ersten Zerlegung des Wurzelknoten wirkt sich bis zur letzten Zerlegung aus (vgl. Hastie u. a., 2013, S. 312).

3. Random Forest

Wie am Ende des vorherigen Kapitels erklärt, können kleine Änderungen in den Daten zu ganz anderen Entscheidungsbäumen führen. Um die hohe Varianz bei instabilen Schätzern zu reduzieren wurden die Verfahren *Bootstrap aggregation* (*Bagging*) (Breiman, 1996) und ihre Erweiterung *Random Forests* (*RF*) von Breiman (2001) entwickelt. Beide Methoden erzeugen ein Ensemble von Entscheidungsbäumen und eine Aggregation von den einzelnen Schätzern wird durchgeführt. Für ein Ensemble von n_{tree} Entscheidungsbäumen, wird für jeden Baum T_b , $b \in \{1, \ldots, n_{tree}\}$ ein Zufallsvektor Θ_b, der unabhängig von den vergangenen Zufallsvektoren $\Theta_1, \ldots, \Theta_{b-1}$ ist, jedoch aus der selben Verteilung stammt, erzeugt. Mit diesem Zufallsvektor Θ_b und den Daten, wird ein Entscheidungsbaum mit der Vorhersagefunktion $f_b(\mathbf{x}, \Theta_b)$ erzeugt, wobei $\mathbf{x}$ ein Beobachtungsvektor ist. Die Vorhersagefunktion für ein Ensemble von n_{tree} Bäumen $\{T_b(\mathbf{x}, \Theta_b)\}_1^{n_{tree}}$, wird bei der Regression über den Mittelwert

$$f^{n_{tree}}(\mathbf{x}) = \frac{1}{n_{tree}} \sum_{b=1}^{n_{tree}} \left(f_b(\mathbf{x}, \Theta_b)\right) \tag{3.1}$$

der einzelnen Vorhersagefunktionen der Regressionsbäume gebildet. Bei der Klassifikation entscheidet man sich für die Klasse, die die Mehrheit der Stimmen

$$f^{n_{tree}}(\mathbf{x}) = Mehrheit\ der\ Stimmen\, \{f_b(\mathbf{x}, \Theta_b)\}_1^{n_{tree}} \tag{3.2}$$

erhält (Breiman, 2001, S. 2 und 21). Wie im vorherigen Kapitel erwähnt, werden in dieser Arbeit die Regessionsverfahren genauer betrachtet.

3.1. Bootstrap aggregation

Bei der *Bagging* Methode werden aus den Lerndaten, $\mathcal{L} = \{(y_i, \mathbf{x}_i)\, , i = 1, \ldots, n\}$, n_{tree} Bootstrap-Stichproben $\mathcal{L}^{*(b)}$, $b = 1, \ldots, n_{tree}$, gezogen. Für jede der Bootstrap-Stichproben wird ein Entscheidungsbaum (CART) T_b erstellt und die Vorhersagefunktion $\hat{f}_b(\mathbf{x}, \mathcal{L}^{*(b)})$ bestimmt. Anschließend wird die Aggregation von den einzelnen Vorhersagefunktionen durchgeführt (vgl. Formel (3.1) und Algorithmus 1). Es hat sich gezeigt, dass die *Bagging* Methode bei Modellen mit geringem Bias und einer hohen Varianz sehr gut funktioniert. Entscheidungsbäume (CART) sind für die *Bagging* Methode gut geeignet, da die Abweichung des prädiktierten vom wahren Wert bei einem nicht beschnittenen Baum relativ klein ist (vgl. Breiman, 1996, S. 1f und vgl. Hastie u. a., 2013, S. 587 f).
Die Vorhersagefunktionen $\hat{f}_b(\mathbf{x}, \mathcal{L}^{*(b)})$ der generierten Entscheidungsbäume bei der *Bagging* Methode sind identisch verteilt (i.d identically distributed), aber nicht unbedingt unab-

Algorithmus 1 *bagging* Algorithmus (vgl. Hastie u. a., 2013, S. 282)

1. Wiederhole Schritt a) und b) für b = 1, . . . , n_{tree}:
   a) Erzeuge eine Bootstrap-Stichprobe $\mathcal{L}^{*(b)} = \left(\boldsymbol{y}^{*(b)}, \boldsymbol{x}^{*(b)}\right)$: ziehe n mal mit Zurücklegen zufällig aus den beobachteten Lerndaten,$(y_i, x_{i1}, x_{i2}, \ldots, x_{ip})$ $i = 1, \ldots, n$.
   b) Lasse einen Entscheidungsbaum (CART) T_b ohne Beschneidung mit der Bootstrap-Stichprobe $\left(\boldsymbol{y}^{*(b)}, \boldsymbol{x}^{*(b)}\right)$ wachsen, bis die minimale Knotengröße n_{min} erreicht wird.
2. Ausgabe von einem Ensemble von B Entscheidungsbäumen $\{T_b\}_1^{n_{tree}}$
3. Prädiktion für einen neuen Beobachtungsvektor $\mathbf{x} = (x_1, x_2, \ldots, x_p)^T$
   a) Regession: $\hat{f}_{bagging}^{n_{tree}}(\mathbf{x}) = \frac{1}{n_{tree}} \sum_{b=1}^{n_{tree}} \left(\hat{f}_b(\mathbf{x}, \mathcal{L}^{*(b)})\right)$

hängig identisch verteilt (i.i.d), da die Entscheidungsbäume aus sehr ähnlichen Daten konstruiert werden. Darum ist der Erwartungswert des Mittelwerts der n_{tree} Vorhersagefunktionen $\hat{f}_{bagging}^{n_{tree}}$ gleich dem Erwartungswert einer beliebigen Vorhersagefunktionen aus dem Ensemble:

$$\begin{aligned} \mathrm{E}_{\mathcal{L}^{*(1)},\ldots,\mathcal{L}^{*(\mathrm{n_{tree}})}} \left[\hat{f}_{bagging}^{n_{tree}}(\mathbf{x})\right] &= \mathrm{E}_{\mathcal{L}^{*(1)},\ldots,\mathcal{L}^{*(\mathrm{n_{tree}})}} \left[\frac{1}{n_{tree}} \sum_{b=1}^{n_{tree}} \left(\hat{f}_b(\mathbf{x}, \mathcal{L}^{*(b)})\right)\right] \\ &= \frac{1}{n_{tree}} \sum_{b=1}^{n_{tree}} \mathrm{E}_{\mathcal{L}^{*(\mathrm{b})}} \left[\hat{f}_b(\mathbf{x}, \mathcal{L}^{*(b)})\right] \\ &= \mathrm{E}_{\mathcal{L}^{*(\mathrm{b})}} \left[\hat{f}_b(\mathbf{x}, \mathcal{L}^{*(b)})\right]. \end{aligned}$$

Dies bedeutet, dass der Schätzer der *Bagging* Methode unverzerrt ist. Die Varianz des Mittelwerts der n_{tree} Vorhersagefunktionen ist:

$$\mathrm{Var}_{\mathcal{L}^{*(1)},\ldots,\mathcal{L}^{*(\mathrm{n_{tree}})}} \left[\hat{f}_{bagging}^{n_{tree}}(\mathbf{x})\right] = \rho\sigma^2 + \frac{1-\rho}{n_{tree}}\sigma^2, \tag{3.3}$$

wobei σ^2 die Varianz der einzelnen Vorhersagefunktionen und ρ die positive paarweise Korrelation zwischen den Vorhersagefunktionen ist (vgl. (Hastie u. a., 2013, S.588 u. 589) und (Louppe, 2014, S.65 f)). Wenn n_{tree} groß genug ist, kann der zweite Teil der Formel (3.3) vernachlässigt werden, doch der erste Teil ist immer noch von der Korrelation ρ abhängig. Die Korrelation ρ zwischen den n_{tree} Vorhersagefunktionen erhöht die Varianz. Je größer die Korrelation ρ , desto größer wird die Varianz des Mittelwerts der n_{tree} Vorhersagefunktionen (vgl. Hastie u. a., 2013, S. 588).

3.2. Random Forests Algorithmus

Im Gegensatz zum *bagging* Verfahren hat der *Random Forests* Algorithmus zum Ziel ein Ensemble von *„de-korrelierten“* ($\rho = 0$) Entscheidungsbäumen zu erzeugen. Die heuristische Idee hinter dem *RF* Algorithmus ist die Verbesserung der Reduktion der Varianz von der *bagging* Methode durch Reduktion der Korrelation ρ zwischen den Vorhersagefunktionen, ohne die Varianz σ^2 der einzelnen Vorhersagefunktionen zu stark zu erhöhen (vgl. Formel (3.3)). Um dies zu erreichen wird der *bagging* Algorithmus 1 mit der „random feature selection“ erweitert. Im Abschnitt 2.2 auf Seite 4 wurde erklärt, wie der CART Algorithmus bei jedem Elternknoten die beste Split-Kovariable unter allen p Kovariablen sucht und eine davon auswählt. Im Gegensatz dazu, wählt man bei der „random feature selection“ bei jedem Elternknoten zuerst zufällig m_{try} Kovariablen aus den p Kovariablen aus und sucht dann unter diesen die beste Split-Kovariable. Ist die Anzahl der Beobachtungen kleiner als n_{min} in einem Kinderknoten, wird dieser Knoten nicht mehr aufgespalten (vgl. *RF* Algorithmus 2 auf der nächsten Seite Punkt (b)). Dieser zusätzliche Schritt bei der Erzeugung der Bäume führt zu unterschiedlichen bzw. zufälligen Bäumen im „Wald“ (Random Forest) und zur Reduktion der Korrelation ρ zwischen den Bäumen (vgl. Clarke u. a., 2009, S. 256 und vgl. Hastie u. a., 2013, S.588 f). Die Vorhersagefunktion für die n_{tree} Regressionsbäume wird, wie bei der *Bagging* Methode, über den Mittelwert der einzelnen Vorhersagefunktionen der Regressionsbäume gebildet (vgl. Formel (3.1) und *RF* Algorithmus 2 auf der nächsten Seite).

3.2.1. Out of Bag Daten

Jeder Baum wird beim *RF* Algorithmus 1 mit der Bootstrap-Stichprobe $\mathcal{L}^{*(t)}$ aus den Lerndaten erzeugt. Das bedeutet, dass nicht alle Beobachtungen aus den Lerndaten für die Erstellung eines Baumes verwendet werden. Die nicht gezogenen Daten werden als „Out-of-Bag“ (*OOB*) Daten bezeichnet. Beim Ziehen mit Zurücklegen werden etwa ein Drittel (36.8 %) der Daten nicht gezogen. Diese *OOB*-Daten können zur Abschätzung der Modellgüte bzw. des Generalisierungsfehlers herangezogenen werden. Der Generalisierungsfehler ist beim Random Forest für eine Regression die erwartete quadratische Verlustfunktion (vgl. Breiman, 2001, S. 21):

$$\mathrm{E}_{\mathrm{Y}\mathbf{X}}\left[(Y - f_{RF}^{n_{tree}}(\mathbf{X}))^2\right] \tag{3.4}$$

Für die Abschätzung des Generalisierungsfehlers wird für jeden Regressionsbaum $\{T_t\}_1^{n_{tree}}$ die Prädiktion für jede Zielvariable y_i, die in den *OOB*-Daten sind, mit der Vorhersage-

Algorithmus 2 *Random Forests* Algorithmus (vgl. Hastie u. a., 2013, S.588)

1. Wiederhole Schritt a bis c für t = 1, . . . , n_{tree}:
   a) Erzeuge eine Bootstrap-Stichprobe $\mathcal{L}^{*(t)} = \left(\boldsymbol{y}^{*(t)}, \boldsymbol{x}^{*(t)}\right)$: durch n-maliges Ziehen mit Zurücklegen aus den beobachteten Lerndaten,$(y_i, x_{i1}, x_{i2}, \ldots, x_{ip})$ $i = 1, \ldots, n$. Die OOB_t- Daten werden zwischengespeichert.
   b) Lasse einen Entscheidungsbaum T_b ohne Beschneidung anhand der Bootstrap-Stichprobe $\mathcal{L}^{*(t)}$ wachsen, wobei jedem Elternknoten die folgenden Schritte rekursiv wiederholt werden:
      i. Ziehe m_{try} Kovariablen zufällig ohne Zurücklegen aus den p Kovariablen $\boldsymbol{X}^{*(t)} = \left(\boldsymbol{x}_{(1)}^{*(t)}, \ldots, \boldsymbol{x}_{(p)}^{*(t)}\right)$
      ii. Wähle Split-Kovariable $\mathbf{x}_{(j)}^{*(t)}$ und den Splitpunkt s aus den m_{try} Kovariablen
      iii. und spalte den Elternknoten in zwei Kindknoten.
      iv. Wiederhole Schritt i. bis iii. bis die minimale Knotengröße n_{min} erreicht wird.
   c) Berechne anhand der OOB_t-Daten die Vorhersagefunktion $\hat{f}_t(\mathbf{x}_{oob_t}, \mathcal{L}^{*(t)})$
2. Ausgabe von einem Ensemble von B Entscheidungsbäumen $\{T_t\}_1^{n_{tree}}$
3. Prädiktion für einen neuen Beobachtungsvektor $\mathbf{x} = (x_1, x_2, \ldots, x_p)^T$
   a) Regession: $\hat{f}_{RF}^{n_{tree}}(\mathbf{x}) = \frac{1}{n_{tree}} \sum_{t=1}^{n_{tree}} \left(\hat{f}_t(\mathbf{x}, \mathcal{L}^{*(t)})\right)$
4. Berechne den mittleren quadratischen Fehler mit den OOB-Prädiktionen
   a) $MSE_{OOB} = \frac{1}{n_{tree}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

funktion

$$\hat{y}_{it} = \hat{f}_t(\mathbf{x}_{oob_{it}}, \mathcal{L}^{*(t)}) = \sum_{m_t=1}^{M_t} \hat{c}_{m_t} I\left(\mathbf{x}_{oob_{it}} \in R_{m_t}\right) \tag{3.5}$$

bestimmt, wobei R_{m_t} die jeweilige Region (p-dimensionales Hyperrechteck) und M_t die Anzahl Zerlegungen des Kovariablenraums im Regressionsbaum T_t (vgl. Kapitel 2.2 auf Seite 4). Der Generalisierungsfehler kann dann mit dem mittleren quadratischen Fehler

$$MSE_{OOB} = \frac{1}{n_{tree}} \sum_{i=1}^{n} \left(y_i - \bar{\hat{y}}_i\right)^2 \tag{3.6}$$

abgeschätzt werden, wobei $\bar{\hat{y}}_i$ der Mittelwert der OOB-Prädiktionen $\hat{y}_i$ für die i-te Beobachtung ist (vgl. Liaw u. Wiener, 2002, S. 18 u.20). Der mittlere quadratische Fehler mit den OOB-Daten ist vergleichbar mit der leave-one-out Kreuzvalidierung (vgl. Hastie u. a., 2013, S. 593). Bei den meisten Ensemble-Methoden sinkt zunächst der Generalisie-

rungsfehler mit steigender Anzahl der Mitglieder des Ensembles n_E und steigt ab einem gewissen n'_E wieder an (vgl. Cutler u. a., 2012, S. 167). Das bedeutet, dass es ab diesem n'_E zur einer Überanpassung der Daten kommt. Der Generalisierungsfehler beim Random Forest konvergiert bei steigender Anzahl der Bäume n_{tree} fast sicher gegen eine untere Schranke (vgl. Breiman, 2001, S. 21). Dies bedeutet, dass die Anzahl der Bäume n_{tree} beliebig groß gewählt werden kann, ohne dass der Generalisierungsfehler erhöht wird. Wenn die Anzahl der Bäume n_{tree} zu klein gewählt wird, schwankt der Generalisierungsfehler zu stark. Als Entscheidungshilfe, ob die Anzahl der Bäume n_{tree} groß genug gewählt ist, wird der MSE_{OOB} als Funktion von n_{tree} dargestellt. Wie man in Abbildung 3.1 erkennt, stabilisiert sich der MSE_{OOB} ab einer Baumanzahl von ca. 250 (vgl. Cutler u. a., 2012, S. 167).

Der *Random Forests* Algorithmus (2) ist im R-Packet *randomForest* implementiert.

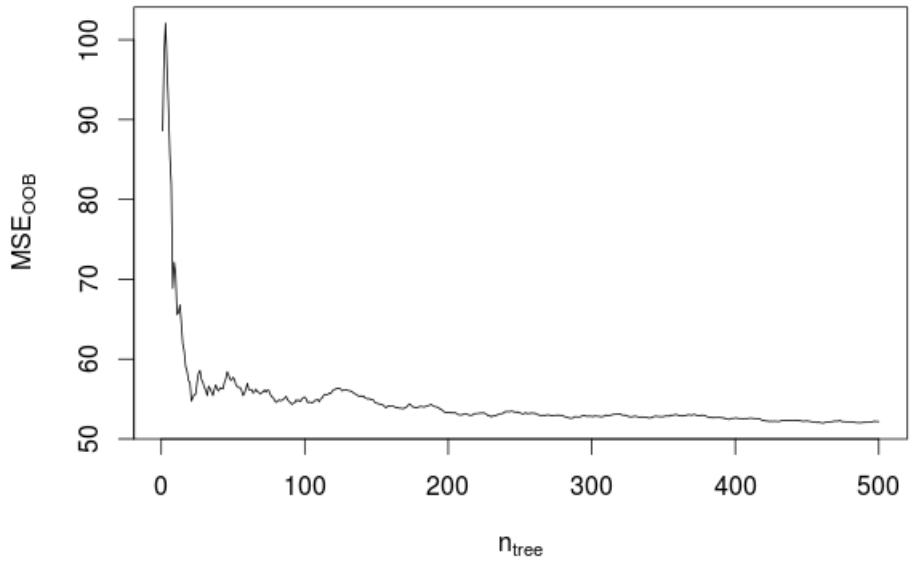

Abbildung 3.1: Die OOB mittleren quadratischen Fehler MSE_{OOB} als Funktion der Anzahl der Bäume n_{tree}

Die Anzahl der Kovariablen m_{try} die für jede Aufspaltung der Knoten verwendet wird, ist ein Tuning-Parameter. Die Standardeinstellung für die Regression ist $\lfloor p/3 \rfloor$ und für die Klassifikation $\left\lfloor \sqrt{p} \right\rfloor$ im Programm. Breiman u. Cutler (2003) schlagen vor, dass man für den Parameter m_{try} die Standardeinstellung, die Hälfte und das Doppelte der Standardeinstellung verwendet und aus diesen Parametereinstellungen, den „Wald“ mit dem geringsten mittleren quadratischen OBB Fehler auswählt (vgl. Formel (3.6)). Wenn die Anzahl der Kovariablen p sehr groß ist, aber nur wenige Kovariablen einen Effekt auf die Zielvariable haben, wird empfohlen den Parameter m_{try} größer einzustellen. Die Standar-

deinstellung für n_{min} (nodesize) ist fünf bei der Regression und eins bei der Klassifikation (vgl. Liaw u. Wiener, 2002, S. 20).
Das *Random Forests* Verfahren erreicht eine hohe Vorhersagegenauigkeit und ist sowohl anwendbar, wenn die Anzahl der Kovariablen p viel größer ist als die Anzahl der Beobachtungen n, als auch, wenn die Kovariablen miteinander korreliert sind. Der große Nachteil an dem *Random Forests* Verfahren ist, dass die Ergebnisse, durch die große Anzahl der Entscheidungsbäume nicht mehr so leicht zu interpretieren sind. Die Wichtigkeit einer Kovariable bezogen auf die Zielvariable kann nicht mehr mit einer graphischen Darstellung eines einzelnen Entscheidungsbaums dargestellt werden. Die Identifizierung relevanter Kovariablen im *Random Forests* erfolgt mit Hilfe der Variablenwichtigkeits Maße, die im nächsten Kapitel genauer beschrieben werden (vgl. Strobl u. a., 2008, S. 307 und vgl. Hastie u. a., 2013, S. 367).

4. Variablenwichtigkeit

Wie in der Einleitung bereits erwähnt, ist ein weiterer Vorteil des *Random Forest* Verfahrens, dass es zur Identifizierung möglicher relevanter Kovariablen so genannte Variablenwichtigkeitsmaße (engl. *variable importance measures* (VIM)) berechnet. Zum Beispiel ist bei der statistischen Genforschung die Identifizierung relevanter Gene von großem Interesse. Diese relevanten Gene können wertvolle Einblicke in die komplexen Mechanismen von bestimmten Erkrankungen liefern (vgl. Janitza u. Boulesteix, 2015, S. 2). Es gibt zwei häufig verwendete Variablenwichtigkeitsmaße, die mittlere Abnahme der Summe der quadratischen Abstände (engl. mean decrease in node impurity bzw. average impurity reduction) (vgl. 2.4) und die Permutation Variablenwichtigkeit (engl. permutation importance bzw. mean decrease in accuracy). In dieser Arbeit wird das Permutation Variablenwichtigkeitsmaß genauer betrachtet.

4.1. Permutation Variablenwichtigkeitsmaß

Das Permutation Variablenwichtigkeitsmaß ($PerVIM$) ist ein sehr intuitives Maß für die Variablenwichtigkeit. Um die Variablenwichtigkeit der Kovariable $\mathbf{x_{(j)}}$ in Bezug auf die Zielvariable $\mathbf{y}$ zu messen, wird das folgende Verfahren für jeden Regressionsbaum im Wald (*Random Forest*) durchgeführt.
Zunächst wird für den Regressionsbaum T_t die Prädiktion für jede Zielvariable y_i in den OOB_t Daten mit der Vorhersagefunktion 3.5 durchgeführt. Anschließend wird die mittlere quadratische Abweichung zwischen den wahren Werten und den OOB Prädiktionen berechnet:

$$MSE(OOB_t) = \frac{1}{n_{OOB_t}} \sum_{i \in OOB_t} (y_i - \hat{y}_{it})^2 , \tag{4.1}$$

wobei n_{OOB_t} die Anzahl der Beobachtungen in den OOB Daten im Regressionsbaum T_t ist. Als nächstes werden die beobachteten Werte, in den OOB_t Daten, der Kovariable $\mathbf{x_{(j)}}$ zufällig permutiert

$$\left(\mathbf{x_{(1)}}, \ldots, \mathbf{x_{(j-1)}}, \mathbf{x^{\pi}_{(j)}}, \mathbf{x_{(j+1)}}, \ldots, \mathbf{x_{(p)}}\right)$$

wobei alle anderen Kovariablen nicht verändert werden. Mit diesen modifizierten Daten werden OOB Prädiktionen für die $\hat{y}^{\pi_j}_{it}$ für $i = 1, \ldots, n$ und die mittlere quadratische Abweichung

$$MSE(OOB_t|\mathbf{x^{\pi}_{(j)}}) = \frac{1}{n_{OOB_t}} \sum_{i \in OOB_t} \left(y_i - \hat{y}^{\pi_j}_{it}\right)^2 \tag{4.2}$$

für den Regressionsbaum T_t berechnet. Wenn die Kovariable $\mathbf{x_{(j)}}$ keinen Einfluss auf die Zielvariable $\mathbf{y}$ hat($\mathbf{y} \perp \mathbf{x_{(j)}}$), sollte die Differenz zwischen den mittleren quadratischen Abweichungen $MSE(OOB_t|x^{\pi}_{(j)}) - MSE(OOB_t)$ ungefähr gleich Null sein. Die Permu-

tation der Kovariable $\mathbf{x_{(j)}}$ ändert nichts an der Ausgangssituation, das die Kovariable $\mathbf{x_{(j)}}$ keinen Einfluss auf die Zielvariable $\mathbf{y}$ hat. Im Gegensatz dazu bewirkt die Permutation der Kovariable $\mathbf{x_{(j)}}$, sollte diese einen Einfluss auf die Zielvariable $\mathbf{y}$ haben ($\mathbf{y} \not\perp \mathbf{x_{(j)}}$), dass dieser Zusammenhang zerstört wird. Die Differenz zwischen den mittleren quadratischen Abweichungen $MSE(OOB_t|x^{\pi}_{(j)}) - MSE(OOB_t)$ ist größer Null. Deshalb wird zur Beurteilung der Variablenwichtigkeit der Kovariable $\mathbf{x_{(j)}}$, in Bezug auf die Zielvariable $\mathbf{y}$ im Regressionsbaum T_t, die Differenz zwischen den mittleren quadratischen Abweichungen gebildet:

$$PerVIM_t(\mathbf{x_{(j)}}) = MSE(OOB_t|x^{\pi}_{(j)}) - MSE(OOB_t) \tag{4.3}$$

Das Variablenwichtigkeitsmaß der Kovariable $\mathbf{x_{(j)}}$, in Bezug auf die Zielvariable $\mathbf{y}$ im *Random Forest*, wird über den Durchschnitt aller $\{PerVIM_t(x^{\pi}_{(j)})\}_1^{n_{tree}}$ gebildet:

$$\begin{aligned} PerVIM(\mathbf{x_{(j)}}) &= \frac{\sum_{t=1}^{n_{tree}} PerVIM_t(\mathbf{x_{(j)}})}{n_{tree}} \\ &= \frac{1}{n_{tree}} \sum_{t=1}^{n_{tree}} \frac{1}{n_{OOB_t}} \sum_{i \in OOB_t} \left[\left(y_i - \hat{y}_{it}^{\pi_j} \right)^2 - (y_i - \hat{y}_{it})^2 \right] . \end{aligned} \tag{4.4}$$

Diese Prozedur wird für alle p Kovariablen im Datensatz durchgeführt (vgl. Algorithmus 3 und Cutler u. a. (2012, S. 168 f)). Wenn das Permutation Variablenwichtigkeitsmaß für eine Kovariable negative Werte oder Werte nahe Null annimmt, bedeutet das, dass diese Kovariable keinen wesentlichen Beitrag zur Verbesserung der Prädiktion von der Zielvariable $\mathbf{y}$ liefert. Mit anderen Worten ausgedrückt: diese Kovariable ist wahrscheinlich nicht relevant. Im Gegensatz dazu, bedeutet für eine Kovariable ein positiver Wert des Permutation Variablenwichtigkeitsmaß, dass diese bereits einen kleinen Beitrag zur Verbesserung der Prädiktion von der Zielvariable $\mathbf{y}$ liefert. Jedoch kann nicht angenommen werden, dass ein positiver Wert für das Permutation Variablenwichtigkeitsmaß immer auf eine relevante Kovariable verweist, da man nicht feststellen kann ob der positive Wert reiner Zufall ist. Es stellt sich die Frage, ob sich ein positiver Wert für das Permutation Variablenwichtigkeitsmaß sich signifikant von Null unterscheidet. In den folgenden Kapiteln werden einige Testverfahren vorgestellt, die diese Frage versuchen zu Versuchen (vgl. Janitza u. Boulesteix, 2015, S.4f).

In dieser Arbeit wird eine Kovariable als relevant bezeichnet, wenn diese die Genauigkeit der Prädiktion von der Zielvariable $\mathbf{y}$ signifikant verbessert. Diese Definition der relevanten Kovariablen umfasst auch Kovariablen, die keinen "eigenen" Einfluss auf die Zielvariable haben, aber durch Korrelation mit einer anderen Kovariable, die einen Einfluss auf die Zielvariable hat, mit der Zielvariable assoziiert sind (vgl. Janitza u. Boulesteix, 2015, S.5). *Random Forests* bevorzugen bei kategorialen Kovariablen solche mit vielen Kategorien und dies führt zu einer Verzerrung der Variablenwichtigkeitsmaße (vgl. Strobl

u. a., 2007, S. 7).

Algorithmus 3 Permutation Variablenwichtigkeitsmaß

1. Wiederhole Schritt a bis b für jeden Regressionsbaum T_t , $t = 1, \ldots, n_{tree}$:
   a) Berechne anhand der OOB_t-Daten:

   $$MSE(OOB_t) = \frac{1}{n_{OOB_t}} \sum_{i \in OOB_t} (y_i - \hat{y}_{it})^2$$

   b) Wiederhole Schritt i bis iii für $j = 1, \ldots, p$:
      i. Permutiere Kovariable $\mathbf{x_{(j)}}$
      ii. Berechne anhand der OOB_t-Daten:

      $$MSE(OOB_t|\mathbf{x}^{\pi}_{(\mathbf{j})}) = \frac{1}{n_{OOB_t}} \sum_{i \in OOB_t} \left(y_i - \hat{y}^{\pi_j}_{it}\right)^2$$

      iii. Berechne :

      $$PerVIM_t(\mathbf{x_{(j)}}) = MSE(OOB_t|x^{\pi}_{(j)}) - MSE(OOB_t)$$

2. Berechne das Permutation Variablenwichtigkeitsmaß für p Kovariablen, $j = 1, \ldots, p$:

   $$PerVIM(\mathbf{x_{(j)}}) = \frac{\sum_{t=1}^{n_{tree}} PerVIM_t(\mathbf{x_{(j)}})}{n_{tree}} \ .$$

4.2. Permutation basierter Testansatz

Der *p*ermutation *imp*ortance ($PIMP$) Testansatz von Altmann u. a. (2010) ist ein heuristischer Permutationstest. Der $PIMP$ Testansatz versucht die Verteilungen des Variablenwichtigkeitsmaßes der Kovariable $\mathbf{x_{(j)}}$abzuschätzen unter der Annahme, dass die Kovariable keinen Einfluss auf die Zielvariable $\mathbf{y}$ hat.
Zuerst wird das originale *Permutation Variablenwichtigkeitsmaß* $PerVIM(\mathbf{x_{(j)}})$ der Kovariablen abgespeichert. Anschließend permutiert man die Zielvariable S-mal und generiert für jede permutierte Zielvariable $\mathbf{y}^{\pi}_{\mathbf{s}}$, $s = 1, \ldots S$, neue *Random Forests*. Für diese S neuen *Random Forests* werden die Permutation Variablenwichtigkeitsmaße $PerVIM(\mathbf{x_{(j)}}|\mathbf{y}^{\pi}_{\mathbf{s}})$ der Kovariable berechnet. Durch die Permutation der Zielvariable, wird der Zusammenhang zwischen der Zielvariable und der Kovariable zerstört. Das bedeutet, für jede Kovariable $\mathbf{x_{(j)}}$ werden S neue Variablenwichtigkeitsmaße erzeugt. Diese Variablenwichtigkeitsmaße werden bei Altmann u. a. (2010) als „null Wichtigkeiten“ (engl. *null importances*) bezeichnet. Es wird angenommen, dass diese *null importances* Maße unabhängige Beobachtungen aus der unbekannten „Null-Verteilung“ sind. Mit der „Null-Verteilung“ ist die unbekannte Verteilung des Variablenwichtigkeitsmaßes der Kovariable $\mathbf{x_{(j)}}$ unter der

Annahme der Unabhängigkeit zwischen der Kovariable $\mathbf{x_{(j)}}$ und der Zielvariable $\mathbf{y}$ gemeint. Altmann u. a. (2010) schlagen vor, eine parametrische Verteilung für die *null importances* Maße anzupassen. In der originalen Implementation des $PIMP$ Testansatzes kann man aus einer Menge von parametrischen Verteilungen, Normal-, Lognormal- oder Gamma-Verteilung eine für die *null importances* Maße auswählen. Die Parameter für die ausgewählte Verteilung werden mit der Maximum-Likelihood-Schätzung berechnet. Die originale Implementation des $PIMP$ Testansatzes bietet die Möglichkeit, die am besten geeignete Verteilung für die *null importances* Maße mit dem Kolmogorov-Smirnov-Test automatisch zu identifizieren. Der ($PIMP$) p-Wert ist die Wahrscheinlichkeit, dass das *null importances* Maß den originalen ($PerVIM(\mathbf{x_{(j)}})$) oder einen höheren Wert annimmt, gegeben der ausgewählten „Null-Verteilung". In dieser Arbeit wurden an die *null importances* Maße nur die Normalverteilung mit Parametern μ und σ^2 angepasst. Der Erwartungswert μ und die Varianz σ^2 der Normalverteilung wurden durch den arithmetischen Mittelwert und die empirische Varianz der erzeugten *null importances* Maße geschätzt. Eine naheliegende Methode wäre die empirische Verteilung der *null importances* Maße für die „Null-Verteilung" zu verwenden. In diesem Fall ist der ($PIMP$) p-Wert der Prozentsatz aller *null importances* Maße, die größer oder gleich als das beobachtete Variablenwichtigkeitsmaß $PerVIM(\mathbf{x_{(j)}})$ sind (vgl. Altmann u. a., 2010, S. 1341f und Janitza u. Boulesteix, 2015, S. 5f). Der Algorithmus 4 beschreibt das Vorgehen beim $PIMP$ Testansatz.

Algorithmus 4 $PIMP$ Testansatz

1. Speichere die beobachteten Permutation Variablenwichtigkeitsmaße $PerVIM(\mathbf{x_{(j)}})$
2. Wiederhole Schritt a bis c für $s = 1, \dots S$:
   a) Permutierte Zielvariable $\mathbf{y_s^\pi}$,
   b) Generiere einen neuen *Random Forest mit* $\mathbf{y_s^\pi}$,
   c) Berechne die Permutation Variablenwichtigkeitsmaße $PerVIM(\mathbf{x_{(j)}}|\mathbf{y_s^\pi})$ für p Kovariablen, $j = 1, \dots, p$.
3. Berechne den ($PIMP$) p-Wert für alle p Kovariablen, $j = 1, \dots, p$:

   $$p(\mathbf{x_{(j)}}) = \frac{1}{S} \sum_{s=1}^{S} I\left(PerVIM(\mathbf{x_{(j)}}|\mathbf{y_s^\pi}) \geq PerVIM(\mathbf{x_{(j)}})\right)$$

4.3. Naiver Testansatz

Auf Grund der Definition des Permutation Variablenwichtigkeitsmaßes, wird erwartet, dass das Permutation Variablenwichtigkeitsmaß zufällig um Null verteilt ist, wenn Kovariablen keinen Einfluss auf die Zielvariable $\mathbf{y}$ haben. Bei diesem heuristischen Testansatz

von Janitza u. Boulesteix (2015) wird eine „Null-Verteilung“ für alle Kovariablen erzeugt. Die „Null-Verteilung“ für die Variablenwichtigkeitsmaße wird durch Spiegelung der beobachteten negativen Variablenwichtigkeitsmaße an der y-Achse erzeugt. Dadurch erhält man eine „Null-Verteilung“ die um Null symmetrisch verteil ist. Es werden folgende Mengen definiert:

- $M_1 = \left\{PerVIM(\mathbf{x_{(j)}})|PerVIM(\mathbf{x_{(j)}}) < 0,\ j = 1, ..., p\right\}$ sind die beobachteten negativen Variablenwichtigkeitsmaße,
- $M_2 = \left\{PerVIM(\mathbf{x_{(j)}})|PerVIM(\mathbf{x_{(j)}}) = 0,\ j = 1, ..., p\right\}$ sind die beobachteten Variablenwichtigkeitsmaße mit dem Wert Null und
- $M_3 = \left\{-PerVIM(\mathbf{x_{(j)}})|PerVIM(\mathbf{x_{(j)}}) < 0,\ j = 1, ..., p\right\} = -M_1$ ist eine hypothetische Menge, die Spiegelung der beobachteten negativen Variablenwichtigkeitsmaße M_1.

Als „Null-Verteilung“ wird die empirische Verteilungsfunktion $\hat{F}_0$ von der Vereinigungsmenge $M = M_1 \cup M_2 \cup M_3$ verwendet. Der p-Wert für die Kovariable $\mathbf{x_{(j)}}$ wird über die empirische Verteilungsfunktion $\hat{F}_0$ berechnet:

$$p(\mathbf{x_{(j)}}) = 1 - \hat{F}_0(PerVIM(\mathbf{x_{(j)}}))\,. \tag{4.5}$$

Oder über den Prozentsatz aller Variablenwichtigkeitsmaße in M, die größer oder gleich als das beobachtete Variablenwichtigkeitsmaß $PerVIM(\mathbf{x_{(j)}})$ sind:

$$p(\mathbf{x_{(j)}}) = \frac{1}{n_M} \sum_{m \in M} I\left(m \geq PerVIM(\mathbf{x_{(j)}})\right)\,, \tag{4.6}$$

wobei n_M die Anzahl der Variablenwichtigkeitsmaße in M ist. Dieser Testansatz kann nicht für alle Daten angewendet werden, da dieser eine relativ große Anzahl von Kovariablen ohne einen Effekt auf die Zielvariable benötigt, sodass die Annäherung der „Null-Verteilung“ präzise genug ist. Eine große Anzahl von Kovariablen ohne Effekt auf die Zielvariable kommen typischerweise bei genetischen Daten vor, weshalb die Anwendung dieses Testansatzes nur bei hochdimensionalen genetischen Daten Sinn macht(vgl. Janitza u. Boulesteix, 2015, S.6).

4.4. Alternatives Permutation Variablenwichtigkeitsmaß

Dieses alternative/neue Permutation Variablenwichtigkeitsmaß von Janitza u. Boulesteix (2015) basiert nicht auf den OOB Daten. Sie verwenden eine ähnliche Strategie, die

von dem Kreuzvalidierungsverfahren inspiriert ist. Zur Berechnung dieses Maßes wird als Erstes der Datensatz zufällig in k gleichgroße disjunkte Teilmengen partitioniert. Anschließend werden k *Random Forests,* $\{RF_1, RF_2, \ldots, RF_k\}$, generiert, wobei der l-te *Random Forest*, $l = 1, \ldots, k$, wird ohne die Beobachtungen der l-ten Teilmenge generiert. Für jeden *Random Forest* RF_l wird das Permutation Variablenwichtigkeitsmaß mittels der l-ten Teilmenge, die nicht für die Generierung der *Random Forests* verwendet wurden, berechnet.
Die Menge aus Beobachtungsindizes für die l-te Teilmenge wird mit $\mathcal{D}_l$ definiert. Sei Das *fold-spezifische Permutation Variablenwichtigkeitsmaß* für Kovariable $\mathbf{x_{(j)}}$, im *Random Forest* RF_l, wird folgendermaßen berechnet:

$$PerVIM(\mathbf{x_{(j)}}|\mathcal{D}_l) = \frac{1}{n_{tree}} \sum_{t=1}^{n_{tree}} \frac{1}{n_{\mathcal{D}_l}} \sum_{i \in \mathcal{D}_l} \left[\left(y_i - \hat{y}_{it}^{\pi_j} \right)^2 - (y_i - \hat{y}_{it})^2 \right], \tag{4.7}$$

wobei n_{tree} die Anzahl der Bäume im *Random Forest* RF_l, $n_{\mathcal{D}_l}$ die Anzahl der Beobachtungen in der l-ten Teilmenge, und $\hat{y}_{it}$ und $\hat{y}_{it}^{\pi_j}$ die Vorhersage der Zielvariablen y_i für $i \in \mathcal{D}_l$ durch den Regressionsbaum $T_t \in RF_l$ vor und nach der Permutation der Beobachtungen von $\mathbf{x_{(j)}}$, sind. Der Mittelwert aller *fold-spezifischen Permutation Variablenwichtigkeitsmaße* ist dann das *Kreuzvalidierungs-Variablenwichtigkeitsmaß* (engl. cross-validated variable importance):

$$CvVIM(\mathbf{x_{(j)}}) = \frac{1}{k} \sum_{l=1}^{k} PerVIM(\mathbf{x_{(j)}}|\mathcal{D}_l)\,. \tag{4.8}$$

In dieser Arbeit wurde der Datensatz zur Berechnung des Kreuzvalidierungs-*Variablenwichtigkeitsmaßes* in nur zwei disjunkte Teilmengen (k=2) unterteilt . Wenn k gleich zwei gesetzt wird, wird das *Kreuzvalidierungs-Variablenwichtigkeitsmaß* auch als *hold-out-Variablenwichtigkeitsmaß* bezeichnet (vgl. Janitza u. Boulesteix, 2015, S.6f). Im Algorithmus 5 ist Schritt für Schritt erklärt, wie das *Kreuzvalidierungs-Variablenwichtigkeitsmaß* berechnet wird.

Algorithmus 5 Kreuzvalidierungs-Variablenwichtigkeitsmaß

1. Teile den Datensatz zufällig in k gleichgroße disjunkte Teilmengen auf.

2. Wiederhole Schritt a) bis c) für $l = 1, \ldots, k$:

   a) Generiere einen neuen *Random Forest* RF_l basierend auf Beobachtungen $\{1, \ldots, n\} \setminus \mathcal{D}_l$.

   b) Berechne das *fold-spezifische Permutation Variablenwichtigkeitsmaß* in allen Regressionsbäumen $T_t \in RF_l$,$t = 1, \ldots, n_{tree}$ und für alle p Kovariablen, $j = 1, \ldots, p$:

   $$PerVIM_t(\mathbf{x_{(j)}}|\mathcal{D}_l) = \frac{1}{n_{\mathcal{D}_l}} \sum_{i \in \mathcal{D}_l} \left[\left(y_i - \hat{y}_{it}^{\pi_j} \right)^2 - (y_i - \hat{y}_{it})^2 \right].$$

   c) Berechne das *fold-spezifische Permutation Variablenwichtigkeitsmaß* für alle p Kovariablen , $j = 1, \ldots, p$, im *Random Forest* RF_l :

   $$PerVIM(\mathbf{x_{(j)}}|\mathcal{D}_l) = \frac{\sum_{t=1}^{n_{tree}} PerVIM_t(\mathbf{x_{(j)}}|\mathcal{D}_l)}{n_{tree}}$$

3. Berechne das *Kreuzvalidierungs-Variablenwichtigkeitsmaß* für alle p Kovariablen , $j = 1, \ldots, p$:

   $$CvVIM(\mathbf{x_{(j)}}) = \frac{1}{k} \sum_{l=1}^{k} PerVIM(\mathbf{x_{(j)}}|\mathcal{D}_l)$$

4.5. Neuer Testansatz

Dieser neue heuristische Testansatz (engl. *novel testing approach* (NTA)) von Janitza u. Boulesteix (2015) unterscheidet sich kaum vom naivem Testansatz (vgl. Abschnitt 4.3), der einzige Unterschied ist, dass statt dem *Permutation Variablenwichtigkeitsmaß* (4.4) das *hold-out-Variablenwichtigkeitsmaß (4.8)* (k=2) verwendet wird. Die Simulationsstudien von Janitza u. Boulesteix (2015) zeigen, dass das *hold-out-Variablenwichtigkeitsmaß* bei Unabhängigkeit zwischen den Kovariablen und der Zielvariable symmetrisch um Null verteilt ist. Das klassische *Permutation Variablenwichtigkeitsmaß* hingegen ist bei Unabhängigkeit nicht symmetrisch um Null verteilt. Dies ist auch der Grund dafür, wieso das *hold-out-Variablenwichtigkeitsmaß* verwendet wird (vgl. Janitza u. Boulesteix, 2015, S. 7f und 11). Das Vorgehen des *neuen heuristischen Testansatzes* wird im Algorithmus 6 beschrieben.

Algorithmus 6 Neuer Testansatz

1. Berechne das *hold-out-Variablenwichtigkeitsmaß* $CvVIM(x_{(j)}|k = 2)$ *für* $j = 1, \ldots, p$*(vgl.* Algorithmus 5*)*

2. Die „Null-Verteilung“ wird durch Spiegelung der Dichte für die beobachteten negativen *hold-out-Variablenwichtigkeitsmaß*e an der y-Achse erzeugt. Sie ergibt sich als empirische Verteilungsfunktion $\hat{F}_0$ für die Beobachtungen folgender Mengen::

   a) $M_1 = \Big\{PerVIM(\mathbf{x_{(j)}})|PerVIM(\mathbf{x_{(j)}}) < 0,\ j = 1, ..., p\Big\}$, die beobachteten negativen Variablenwichtigkeitsmaße,

   b) $M_2 = \Big\{PerVIM(\mathbf{x_{(j)}})|PerVIM(\mathbf{x_{(j)}}) = 0,\ j = 1, ..., p\Big\}$, die beobachteten Variablenwichtigkeitsmaße mit dem Wert Null und

   c) $M_3 = \Big\{-PerVIM(\mathbf{x_{(j)}})|PerVIM(\mathbf{x_{(j)}}) < 0\, j = 1, ..., p\Big\}$, die Spiegelung der beobachteten negativen Variablenwichtigkeitsmaße M_1.

3. Berechne den p-Wert für das *hold-out-Variablenwichtigkeitsmaß* aller p Kovariablen, $j = 1, \ldots, p$:

   $p(\mathbf{x_{(j)}}) = 1 - \hat{F}_0(CvVIM(\mathbf{x_{(j)}}|k = 2))\,.$

4.6. Kleine Simulationsstudie

Bei Janitza u. Boulesteix wurde der *NTA* und der naive Testansatz nur für kategoriale Zielvariablen untersucht. In dieser Arbeit wird untersucht, ob diese Testansätze auch für stetige den Zielvariablen funktionieren. Der *NTA* und naive Testansatz beruhen darauf, dass die verwendeten *Variablenwichtigkeitsmaße* bei Unabhängigkeit zwischen den Kovariablen und der Zielvariable symmetrisch um Null verteilt sind.
In dieser Simulationsstudie wird untersucht, ob die *Variablenwichtigkeitsmaße* (PerVIM, CvVIM(k=2)) unter der Annahme der Unabhängigkeit ($\mathbf{y} \perp \mathbf{x_{(j)}}, j = 1, \ldots, p$) immer noch symmetrisch um Null verteilt sind, wenn die Korrelation zwischen den Kovariablen ansteigt. Die Zielvariable wurde aus einer Standardnormalverteilung erzeugt, mit einem Erwartungswert Null und einer Varianz von Eins:

$$Y \sim N(0, 1) \tag{4.9}$$

Die Zielvariable ist unabhängig von den Kovariablen, $\mathbf{y} \perp \mathbf{x_{(j)}}, j = 1, \ldots, p$. Die p Kovariablen $\mathbf{X} = (X_{(1)}, X_{(2)}, \ldots, X_{(p)})^T$ wurden aus einer multivariaten Normalverteilung erzeugt:

$$\mathbf{X} \sim N_p(\mathbf{0}, \mathbf{\Sigma}) \tag{4.10}$$

wobei die Hauptdiagonale der Kovarianzmatrix $\mathbf{\Sigma}$ ausschließlich aus Einsen besteht, sodass die Kovarianzmatrix gleich der Korrelationsmatrix ist. Die Korrelation zwischen allen

Kovariablen ist konstant $\rho(X_{(j)}, X_{(k)}) = \sigma,\ j, k = 1, \ldots, p$:

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & \sigma & \sigma & \cdots & \sigma \\ \sigma & 1 & \sigma & \cdots & \sigma \\ \sigma & \sigma & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \sigma \\ \sigma & \sigma & \cdots & \sigma & 1 \end{pmatrix}. \tag{4.11}$$

Es wurden elf Datensätze mit unterschiedlichen Korrelation zwischen den Kovariablen erzeugt:

$$\sigma = 0, 0.1, 0.2, \ldots, 0.9, 1\,.$$

Die Anzahl der Beobachtungen n und die Anzahl der Kovariablen p wurden bis jetzt noch nicht definiert, da noch eine Unterscheidung zwischen zwei Fällen erfolgt.
Es wird wie folgt zwischen den beiden Fällen unterschieden:

- Fall 1: die Anzahl der Kovariablen ist *kleiner* als die Anzahl der Beobachtungen, $p = 100$, $n = 500$ und
- Fall 2: die Anzahl der Kovariablen ist *viel größer* als die Anzahl der Beobachtungen, $p = 2000$, $n = 100$.

In beiden Fällen wurde für jeden Datensatz ein *Random Forest* mit folgenden Einstellungen generiert:

- Anzahl der Regressionsbäume $n_{tree} = 1000$,
- $m_{try} = \frac{p}{3}$.

Anschließend wurden die *Variablenwichtigkeitsmaße,* das *Permutation Variablenwichtigkeitsmaß* und das *hold-out-Variablenwichtigkeitsmaß,* berechnet und die „Null-Verteilung“ grafisch als Boxplot dargestellt.

Fall 1 $p < n$

In der Abbildung 4.1 kann man sehr gut erkennen, dass beide beobachteten *Variablenwichtigkeitsmaße,* mit steigender Korrelation zwischen den Kovariablen, nicht mehr um Null verteilt sind. Wenn die Anzahl der Kovariablen *kleiner* als die Anzahl der Beobachtungen ist und die Kovariablen untereinander korreliert sind, sind die heuristischen Testansätze, neuer und naiver Testansatz, nicht mehr anwendbar. Beide Testansätze setzen voraus, dass das *Variablenwichtigkeitsmaß* unter der Annahme der Unabhängigkeit ($\mathbf{y} \perp \mathbf{x_{(j)}}, j = 1, \ldots, p$) symmetrisch um Null verteilt sein muss.

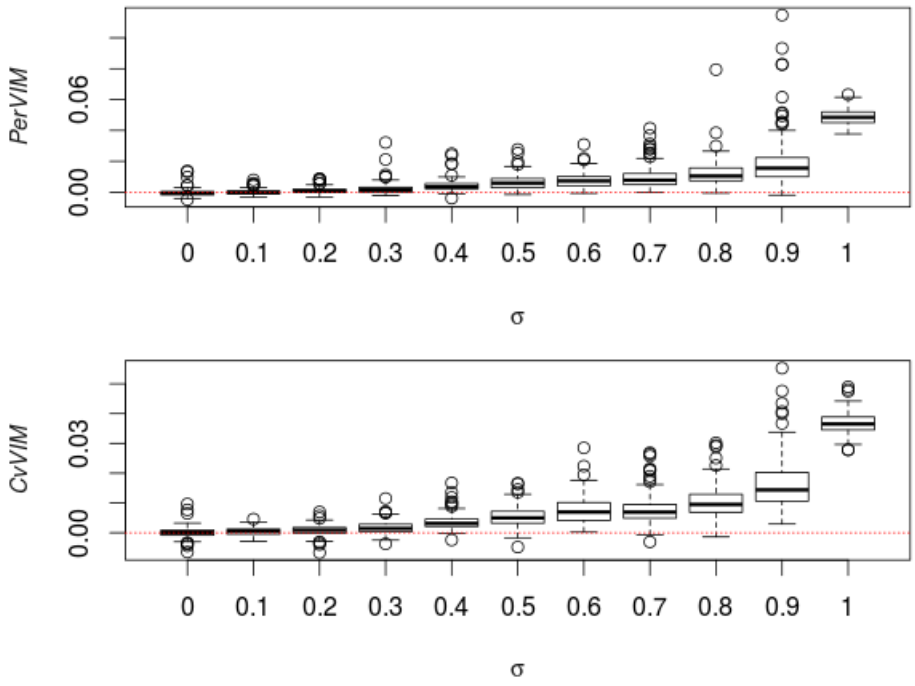

Abbildung 4.1: Boxplot der beobachteten „Null-Verteilungen“ der *Variablenwichtigkeitsmaße* in Abhängigkeit der Korrelation σ zwischen den Kovariablen im Fall 1 $p < n$: In der Abbildung sind oben die beobachteten „Null-Verteilungen“ für die *Permutation Variablenwichtigkeitsmaße* (PerVIM) und unten für die *hold-out-Variablenwichtigkeitsmaße* (CvVIM) dargestellt. Die Abszissenachse ist bei $x = 0$ als rote gepunktete Linie eingezeichnet.

Einen ähnliches Verhalten des Random Forests wurde in der Studie von Strobl u. a. (2008) beobachtet. Strobl u. a. stellten fest,dass das *Permutation Variablenwichtigkeitsmaß* für miteinander korrelierte Kovariablen größer ist als für untereinander unabhängige Kovariablen (vgl. Strobl u. a., 2008, S. 5ff).

Fall 2 $p \gg n$

Wenn die Anzahl der Kovariablen *viel größer* als die Anzahl der Stichproben ist, sind die beobachteten „Null-Verteilungen“ der *Variablenwichtigkeitsmaße* nicht mehr so stark von der Korrelation zwischen den Kovariablen abhängig. Dennoch kann man in den Abbildungen 4.3 erkennen, dass die beobachteten „Null-Verteilungen“ der *Variablenwichtigkeitsmaße* für große Korrelationen zwischen den Kovariablen nicht mehr um Null verteilt sind. Die Anwendung der heuristischen Testansätze, *NTA* und naiver Testansatz, bei hoch korrelierten Datensätzen ist näher zu untersuchen.

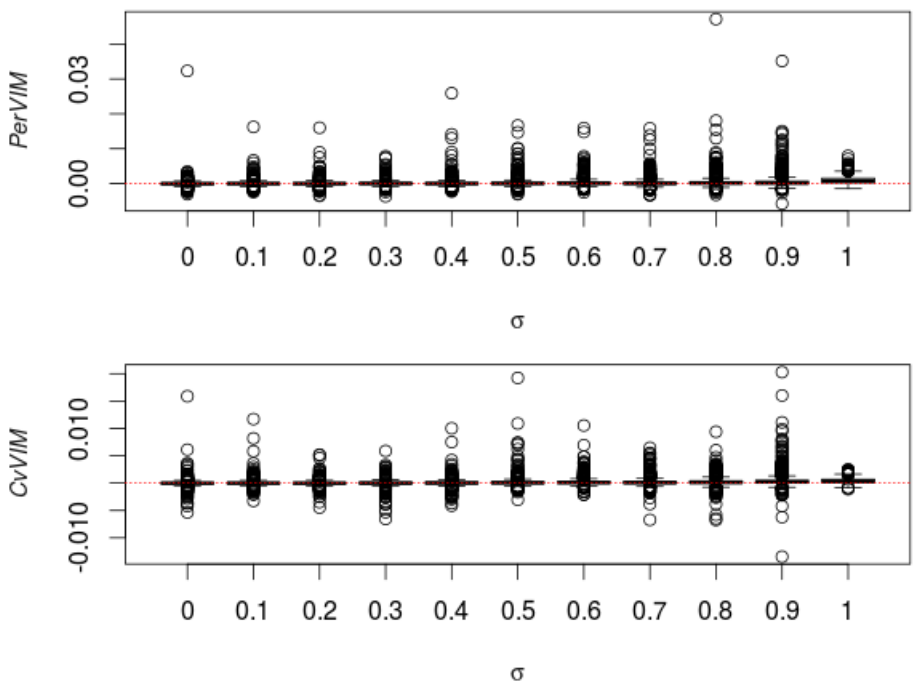

Abbildung 4.2: Boxplot der beobachteten „Null-Verteilungen“ der *Variablenwichtigkeitsmaße* in Abhängigkeit der Korrelation σ zwischen den Kovariablen im Fall 2 $p \gg n$: In der Abbildung sind oben die beobachteten „Null-Verteilungen“ für die *Permutation Variablenwichtigkeitsmaße* (PerVIM) und unten für die *hold-out-Variablenwichtigkeitsmaße* (CvVIM) dargestellt. Die Abszissenachse ist bei $y = 0$ als rote gepunktet Linie eingezeichnet.

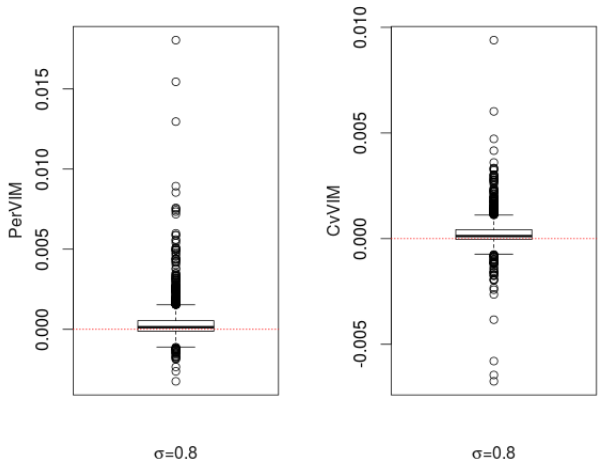

Abbildung 4.3: Boxplot der beobachteten „Null-Verteilungen“ der *Variablenwichtigkeitsmaße* bei eine der Korrelation von $\sigma = 0.8$ zwischen den Kovariablen im Fall 2 $p \gg n$: Links ist die „Null-Verteilung“ des *Permutation Variablenwichtigkeitsmaßes* (PerVIM) und rechts ist die „Null-Verteilung“ des *hold-out-Variablenwichtigkeitsmaßes* (CvVIM) dargestellt. Die Abszissenachse ist bei $y = 0$ als rot gepunktete Linie eingezeichnet.

4.7. Die Null-Verteilung Approximation

Die Idee zu diesem neuen heuristischen Testansatz „Null-Verteilung Approximation“ (engl. Null distribution approximation ($NuDA$)) entstand nach der „kleinen“ Simulationsstudie Abschnitt 4.6. Dieser heuristische Testansatz verbindet die Ideen vom $PIMP$ Testansatz und dem naivem Testansatz. Bei diesem heuristischen Testansatz wird eine „Null-Verteilung“ für alle Kovariablen, wie beim naivem Testansatz, erzeugt. Als Erstes wird das originale *Permutation Variablenwichtigkeitsmaß* $PerVIM(\mathbf{x_{(j)}})$ der Kovariablen abgespeichert. Anschließend permutiert man die Zielvariable und generiert für die permutierte Zielvariable $\mathbf{y}^{\pi}$ einen neuen *Random Forest*. Für diesen neuen *Random Forest* werden die *Permutation Variablenwichtigkeitsmaße* $PerVIM(\mathbf{x_{(j)}}|\mathbf{y}^{\pi})$ der Kovariablen berechnet. Wie beim $PIMP$ Testansatz wird durch die Permutation der Zielvariable der Zusammenhang zwischen der Zielvariable und der Kovariable zerstört. Es wird angenommen, dass die erzeugten Permutation Variablenwichtigkeitsmaße $PerVIM(\mathbf{x_{(j)}}|\mathbf{y}^{\pi})$ aus der unbekannten „Null-Verteilung“ stammen. Möchte man noch mehr Ziehungen aus der „Null-Verteilung“ erhalten, wird diese beschriebene Prozedur solange wiederholt, bis die gewünschte Anzahl N erreicht ist. Als „Null-Verteilung“ wird die empirische Verteilungsfunktion $\hat{F}_0$ von den $N \times p$ erzeugten Permutation Variablenwichtigkeitsmaße $PerVIM(\mathbf{x_{(j)}}|\mathbf{y}^{\pi})$ verwendet. Der p-Wert für die Kovariable $\mathbf{x_{(j)}}$, wird wie beim naivem Testansatz, über die empirische Verteilungsfunktion $\hat{F}_0$ berechnet (vgl. Formel 4.5 bzw. Gleichung (4.6)).

Algorithmus 7 *NuDA* Testansatz

1. Speichere die beobachteten Permutation Variablenwichtigkeitsmaße $PerVIM(x_{(j)})$ für alle Kovariablen $j = 1, \ldots, p$

2. Setze $l = 0$

3. Wiederhole Schritt a) bis d) solange $l < N$ (N die Anzahl der Ziehungen aus der„Null-Verteilung“) ist:

   a) Permutiere Zielvariable $\mathbf{y}^{\pi}$

   b) Generiere einen neuen *Random Forest mit* $\mathbf{y}^{\pi}$

   c) Berechne die Permutation Variablenwichtigkeitsmaße $PerVIM(\mathbf{x_{(j)}}|\mathbf{y}^{\pi})$ für p Kovariablen , $j = 1, \ldots, p$

   d) Setze $l = l + p$.

4. Als „Null-Verteilung“ wird die empirische Verteilungsfunktion $\hat{F}_0$ von den $N \times p$ erzeugten Permutation Variablenwichtigkeitsmaße $PerVIM(\mathbf{x_{(j)}}|\mathbf{y}^{\pi})$ verwendet.

5. Berechne den p-Wert für alle p Kovariablen, $j = 1, \ldots, p$:

   $p(\mathbf{x_{(j)}}) = 1 - \hat{F}_0(PerVIM(\mathbf{x_{(j)}}))$

5. Simulationsstudien

Wie in der Einleitung bereits erwähnt, liegt das Hauptaugenmerk der Arbeit in der Untersuchung, ob die heuristischen Testansätze (vgl. Kapitel 4) für hochdimensionale genetische Daten geeignet sind. Es wird zum einen untersucht, inwiefern die heuristischen Testansätze das vorgegebene Signifikanzniveau α einhalten und zum anderen wie groß die Power dieser Testansätze ist. Um diese Analysen durchzuführen, muss bekannt sein welche Kovariable einen Effekt bzw. keinen Effekt auf die Zielvariable hat. Da bei realen Daten diese Information nicht vorliegt, werden die Daten simuliert. Da der Random Forest bei kategorialen Kovariablen solche mit vielen Kovariablen bevorzugt und dies zur Verzerrung der Variablenwichtigkeitsmaße führen kann, werden in der Simulation nur stetige Kovariablen verwendet (vgl. Strobl u. a., 2007, S. 7).

5.1. Was soll untersucht werden?

Im Abschnitt 4.1 stellte man sich die Frage, ob sich ein positiver Wert für das Variablenwichtigkeitsmaß signifikant von Null unterscheidet. Anders ausgedrückt: man möchte testen, ob die Kovariable $X_{(j)}$ unabhängig von der Zielvariable Y ist. Die Hypothesen für dieses Testproblem sind folgendermaßen definiert:

$$\begin{aligned} H_0 &: \; X_{(j)} \textit{ und } Y \textit{ sind unabhängig voneinander}, \\ H_1 &: \; X_{(j)} \textit{ und } Y \textit{ sind abhängig voneinander}. \end{aligned} \tag{5.1}$$

Auf Grund der Korrelation zwischen den Kovariablen in hochdimensionalen genetischen Daten, treffen diese Hypothesen nicht ganz zu. Die Hypothesen für dieses Testproblem sind folgendermaßen definiert:

$$\begin{aligned} H_0 &: \; X_{(j)} \textit{ und } (Y, Z) \textit{ sind unabhängig voneinander}, \\ H_1 &: \; X_{(j)} \textit{ und } (Y, Z) \textit{ sind abhängig voneinander}, \end{aligned} \tag{5.2}$$

wobei $Z = X_{(1)}, \ldots, X_{(j-1)}, X_{(j+1)}, \ldots, X_{(p)}$ die restlichen Kovariablen sind. Wenn die Kovariablen miteinander korreliert sind, kann keine sichere Aussage darüber getroffen werden, warum die Null-Hypothese verworfen wurde. Es gibt zwei Szenarien, wieso die Null-Hypothese verworfen wird. Das erste Szenario sagt aus, dass die Kovariable $X_{(j)}$ und die Zielvariable Y abhängig voneinander sind. Oder das zweite, dass die Kovariable $X_{(j)}$ und die restlichen Kovariablen Z abhängig voneinander sind (vgl. Strobl u. a., 2008, S. 6).

5.2. Studie I

In der ersten Studie wird untersucht, ob die Testverfahren das vorgegebene Signifikanzniveau α einhalten. Anders ausgedrückt, die Wahrscheinlichkeit für den Fehler 1. Art (Null-Hypothese wird verworfen, obwohl die Null-Hypothese wahr ist) sollte das vorgegebene Signifikanzniveau α nicht überschreiten:

$$P(Fehler\ 1.Art) \leq \alpha\,.$$

Darum müssen die Daten entsprechend der Null-Hypothese simuliert werden. Die Kovariablen und Zielvariable sind unabhängig voneinander. Das Signifikanzniveau α ist in allen Simulationen 0.05.

5.2.1. Datengenerierender Prozess für Studie I

Wie bereits erwähnt, ist die Zielvariable unabhängig von den Kovariablen, $\mathbf{y} \perp \mathbf{x_{(j)}}, j = 1, \ldots, p$. Die Zielvariable wurde aus einer Standardnormalverteilung mit einem Erwartungswert Null und einer Varianz von Eins erzeugt:

$$Y \sim N(0, 1)\,. \tag{5.3}$$

Bei den Kovariablen wird zwischen zwei Fällen unterschieden:

- Im *ersten* Fall sind die Kovariablen untereinander nicht korreliert.
- Im *zweiten* Fall sind manche Kovariablen untereinander korreliert.

Fall 1

Jede der p Kovariablen $\mathbf{X} = (X_{(1)}, X_{(2)}, \ldots, X_{(p)})^T$ wurde unabhängig voneinander aus einer Standardnormalverteilungen mit Erwartungswert Null und Varianz von Eins erzeugt:

$$X_{(j)} \overset{i.i.d}{\sim} N(0, 1), j = 1, \ldots, p\,. \tag{5.4}$$

Fall 2

Die p Kovariablen werden in drei gleich große Teile unterteilt:

$$\begin{aligned}
\mathbf{X}_{p_1} &= \quad (X_{(1)}, \ldots, X_{(p_1)})^T \quad &, p_1 = \left\lceil \frac{p}{3} \right\rceil, \\
\mathbf{X}_{p_2} &= (X_{(p_1+1)}, \ldots, X_{(p_2)})^T &, p_2 = p_1 + \left\lceil \frac{p}{3} \right\rceil, \\
\mathbf{X}_{p_3} &= (X_{(p_2+1)}, \ldots, X_{(p_3)})^T &, p_3 = p.
\end{aligned} \tag{5.5}$$

Das erste Drittel und das zweite Drittel der Kovariablen $\mathbf{X}_{p_1}, \mathbf{X}_{p_2}$ wurden aus einer multivariaten Normalverteilung erzeugt:

$$\mathbf{X}_{p_1} \sim N_{p_1}\left(\mathbf{0}_{p_1}, \mathbf{\Sigma}_{p_1}\right) \; und \; \mathbf{X}_{p_2} \sim N_{p_2}\left(\mathbf{0}_{p_2}, \mathbf{\Sigma}_{p_2}\right) , \tag{5.6}$$

wobei die Hauptdiagonale der Kovarianzmatrixen $\mathbf{\Sigma}_{p_1}, \mathbf{\Sigma}_{p_2}$ aus Einsen besteht, sodass die Kovarianzmatrix gleich der Korrelationsmatrix ist. Die Korrelation zwischen allen Kovariablen in der ersten Kovarianzmatrix $\mathbf{\Sigma}_{p_1}$ ist konstant $\rho(X_{(j)}, X_{(k)}) = 0.3,\ j, k = 1, \ldots, p_1$ und $j \neq k$:

$$\mathbf{\Sigma}_{p_1} = \begin{pmatrix} 1 & 0.3 & 0.3 & \cdots & 0.3 \\ 0.3 & 1 & 0.3 & \cdots & 0.3 \\ 0.3 & 0.3 & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0.3 \\ 0.3 & 0.3 & \cdots & 0.3 & 1 \end{pmatrix} . \tag{5.7}$$

In der zweiten Kovarianzmatrix $\mathbf{\Sigma}_{p_2}$ ist die Korrelation zwar höher, aber auch konstant zwischen allen Kovariablen $\rho(X_{(j)}, X_{(k)}) = 0.8,\ j, k = (p_1 + 1), \ldots, p_2$ und $j \neq k$:

$$\mathbf{\Sigma}_{p_2} = \begin{pmatrix} 1 & 0.8 & 0.8 & \cdots & 0.8 \\ 0.8 & 1 & 0.8 & \cdots & 0.8 \\ 0.8 & 0.8 & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0.8 \\ 0.8 & 0.8 & \cdots & 0.8 & 1 \end{pmatrix} . \tag{5.8}$$

Jede Kovariable $X_{(j)} \in \mathbf{X}_{p_3}$ im dritten Drittel wird aus einer Standardnormalverteilung mit einem Erwartungswert Null und einer Varianz von Eins erzeugt:

$$X_{(j)} \overset{i.i.d}{\sim} N(0, 1), j = (p_2 + 1), \ldots, p\,. \tag{5.9}$$

Es wurden in beiden Fällen 200 Datensätze (D) mit 100 Beobachtungen (n), 2000 Kovariablen (p) und 200 Datensätze mit 100 Beobachtungen, 8000 Kovariablen erstellt. Einen Überblick über die erstellten Daten kann man in der Tabelle 5.1 finden. Für die Erstellung der Datensätze wurde eine neue Funktion ($DataStudyI$) in R geschrieben (vgl. Anhang B). Die 200 Datensätze werden parallel mehreren Prozessoren gleichzeitig erzeugt. Zu beachten ist, dass in dieser Funktion $DataStudyI$ nicht der Standard-Zufallsgenerator verwendet worden ist, sondern der $L'Ecuyer - CMRG$ ($set.seed(1982)$). Dieser Zufallsgenerator sichert die Reproduzierbarkeit der Simulationen bzw. der erzeugten Datensätze beim parallelen Rechnen (vgl. McCallum u. Weston, 2011, S. 56). Auch alle anderen implementierten Funktionen verwenden diesen Zufallsgenerator.

Studie I	p	n	D	$\rho(X_{(j)}, X_{(k)})$
$Fall\ 1$	2000	100	200	Nein
	8000	100	200	Nein
$Fall\ 2$	2000	100	200	Ja
	8000	100	200	Ja

Tabelle 5.1: Überblick über die erstellten Datensätze in der *ersten* Studie im Fall 1 und Fall 2, wobei p die Anzahl der Kovariablen, n die Anzahl der Beobachtungen, D die Anzahl der generierten Datensätze sind. Des weiteren ist zu sehen ob die Kovariablen untereinander korreliert sind $\rho(X_{(j)}, X_{(k)})$.

5.2.2. Aufbau der Simulation

Der Ablauf der Simulation in der *Studie I* und die verwendeten Parameter, werden im Algorithmus(8) erklärt. Jeder generierte *Random Forests* besteht aus 1000 (n_{tree}) Regressionsbäumen. Der $PIMP$ Testansatz wurde wegen der hohen Rechenzeit nur mit 100 (S) Wiederholungen ausgeführt. Und die Anzahl der Ziehungen aus der„Null-Verteilung“ N beim $NuDA$ Testansatz wurde auf $N = 2p$ gesetzt. Alle Methoden wurden für drei verschiedene Einstellungen des Parameters m_{try} durchgeführt:

$$m_{try} = \frac{p}{3}\ ,\ m_{try} = \frac{p}{5} \text{ und } m_{try} = \frac{p}{10}\,.$$

Der p-Wert mit dem $PIMP$ Testansatz wird auf zwei Arten, nicht-parametrisch und parametrisch, berechnet.

Algorithmus 8 Aufbau der Simulation

1. Wiederhole Schritt a bis f für alle 200 Datensätze mit dem Parameter (n_{tree}= 1000,m_{try},$S = 100$, $k = 2$, $N = 2p$)
   a) Generiere einen neuen *Random Forest,*
   b) Berechne die *Permutation Variablenwichtigkeitsmaße* $PerVIM(\mathbf{x_{(j)}})$ für p Kovariablen , $j = 1, \ldots, p$,
   c) Berechne die *hold-out-Variablenwichtigkeitsmaße* $CvVIM(\mathbf{x_{(j)}}|k = 2)$ für p Kovariablen , $j = 1, \ldots, p$,
   d) Berechne den p-Wert für das *hold-out-Variablenwichtigkeitsmaß* aller p Kovariablen, $j = 1, \ldots, p$ mit dem *neuen* Testansatz,
   e) Berechne den p-Wert für das *Permutation Variablenwichtigkeitsmaß* aller p Kovariablen, $j = 1, \ldots, p$ mit dem $NuDA$ Testansatz,
   f) Berechne den p-Wert für das *Permutation Variablenwichtigkeitsmaß* aller p Kovariablen, $j = 1, \ldots, p$ mit dem $PIMP$ Testansatz,
   g) Speichere alle Ergebnisse

5.2.3. Evaluationsgrößen der Studie I

Für jeden Testansatz wird die Wahrscheinlichkeit für den Fehler 1. Art durch die relative Häufigkeit der Fälle, in denen die Null-Hypothese fälschlicherweise verworfen wurde, geschätzt:

$$\alpha(\mathbf{x_{(j)}}) = \frac{1}{D}\sum_{d=1}^{D} I\left(p(\mathbf{x_{(j)}}) \leq 0.05\right),\ j = 1, ..., p\,, \tag{5.10}$$

wobei D die Anzahl der generierten Datensätze ist. Die Wahrscheinlichkeiten für den Fehler 1. Art der verschiedenen heuristischen Testansätze werden für verschiedene Parameter Einstellungen mithilfe von Boxplots miteinander verglichen und ggf. durch zusammenfassende Tabellen (Mittelwert, Median) ergänzt. In dieser Studie wird des Weiteren untersucht, ob die *Variablenwichtigkeitsmaße* (PerVIM, CvVIM(k=2)) symmetrisch um Null verteilt sind.

5.3. Ergebnisse der Studie I

5.3.1. Die „Null-Verteilung“

Wie in der „kleinen“ Simulationsstudie in Abschnitt 4.6 auf Seite 21, wird als Erstes untersucht, ob die *Permutation Variablenwichtigkeitsmaße* symmetrisch um Null verteilt sind. In der ersten Studie, im Fall 1 und Fall 2, ist die Zielvariable unabhängig von den Kovariablen. Gemäß der Definition der *Variablenwichtigkeitsmaße* wird erwartet, dass das *Permutation Variablenwichtigkeitsmaß* und das *hold-out-Variablenwichtigkeitsmaß* zufällig um Null verteilt sind. Gerade der naive und der neue heuristische Testansatz beruhen auf dieser Annahme (vgl. Abschnitt 4.3 und 4.5). Ist die „Null-Verteilung“ der *Variablenwichtigkeitsmaße* nicht symmetrisch um Null verteilt, führt dies zu verzerrten Testresultaten.

In der Abbildung 5.1 und 5.2 sind die beobachteten „Null-Verteilungen“ der *Variablenwichtigkeitsmaße* mit der Parametereinstellung $m_{try} = \frac{p}{3}$ dargestellt. In der Abbildung 5.1 kann man sehr gut erkennen, dass das *hold-out-Variablenwichtigkeitsmaß* in beiden Fällen, die Kovariablen sind unabhängig identisch verteilt und die Kovariablen sind nicht unabhängig voneinander, symmetrisch um Null verteilt ist. Im Gegensatz dazu, ist das *Permutation Variablenwichtigkeitsmaß* (PerVIM) in beiden Fällen nicht exakt symmetrisch um Null verteilt. Der rechte Tail der beobachteten „Null-Verteilungen“ der *Permutation Variablenwichtigkeitsmaße* ist länger als der linke. Aus diesem Grund könnte der heuristische naive Testansatz mit dem *Permutation Variablenwichtigkeitsmaß* nicht für die Überprüfung der Hypothesen 5.1 bzw. 5.2 geeignet sein. Das vorgegebene Signifikanzniveau α könnte nicht eingehalten werden. Die beobachteten „Null-Verteilungen“ der *Variablenwichtigkeitsmaße* ändern sich kaum für die Parametereinstellungen $m_{try} = \frac{p}{5}$

und $m_{try} = \frac{p}{10}$ (siehe Anhang A.1.1).

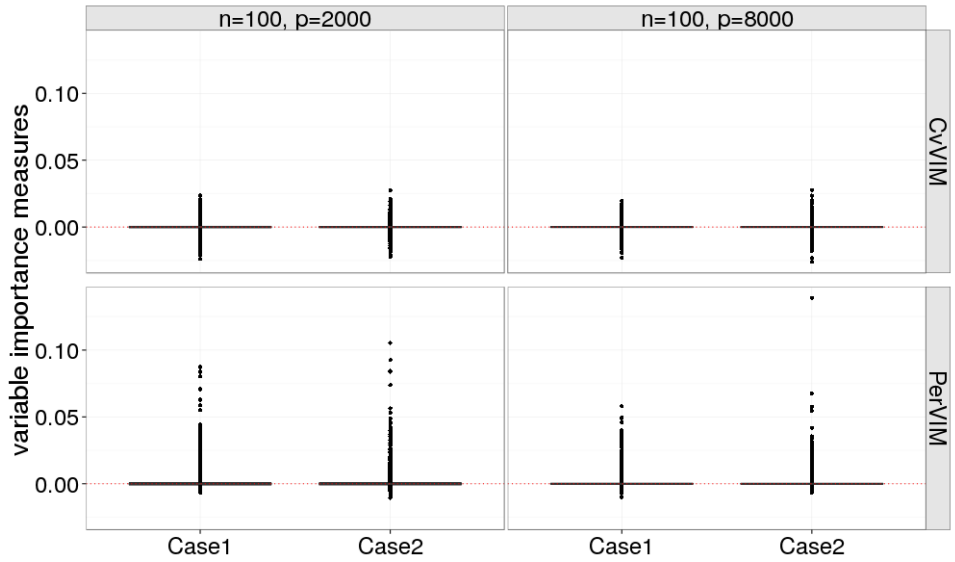

Abbildung 5.1: Boxplot der beobachteten „Null-Verteilungen“ der *Variablenwichtigkeitsmaße* mit der Parametereinstellung $m_{try} = \frac{p}{3}$ für die verschiedenen Fälle: Oben sind die beobachteten „Null-Verteilungen“ für die *hold-out-Variablenwichtigkeitsmaße* (CvVIM) und unten die *Permutation Variablenwichtigkeitsmaße* (PerVIM) dargestellt. Links sind die beobachteten „Null-Verteilungen“ für 200 Datensätze mit 100 Beobachtungen (n) , 2000 Kovariablen (p) und rechts mit 100 Beobachtungen, 8000 Kovariablen dargestellt. Die Abszissenachse ist bei $y = 0$ als rot gepunktete Linie eingezeichnet.

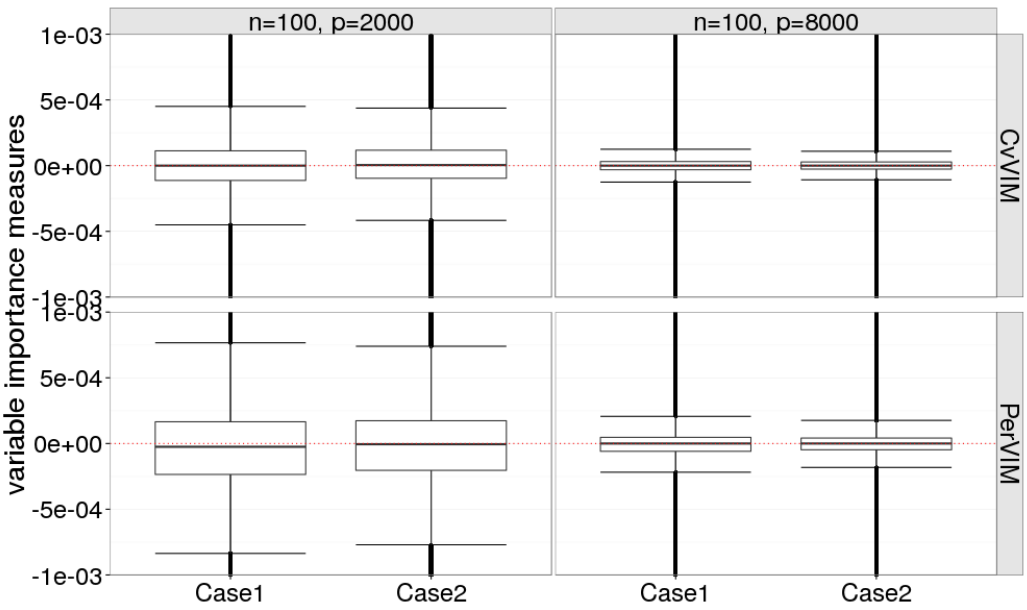

Abbildung 5.2: Boxplot der beobachteten „Null-Verteilungen“ der *Variablenwichtigkeitsmaße* für den gezoomten Wertebereich [-0.001 bis 0.001] der y-Achse mit der Parametereinstellung $m_{try} = \frac{p}{3}$ für die verschiedenen Fälle: Oben sind die beobachteten „Null-Verteilungen“ für die *hold-out-Variablenwichtigkeitsmaße* (CvVIM) und unten die *Permutation Variablenwichtigkeitsmaße* (PerVIM) dargestellt. Links sind die beobachteten „Null-Verteilungen“ für 200 Datensätze mit 100 Beobachtungen (n) , 2000 Kovariablen (p) und Rechts mit 100 Beobachtungen, 8000 Kovariablen dargestellt. Die Abszissenachse ist bei $y = 0$ als rot gepunktete Linie eingezeichnet.

5.3.2. Der Fehler 1. Art

Die Wahrscheinlichkeiten für den Fehler 1. Art $\alpha(\mathbf{x}_{(\mathbf{j})})$ (vgl. Gleichung (5.10)) der heuristischen Testansätze, werden separat für alle Kovariablen p für jeden Fall betrachtet.

Fall 1

Die folgende Abbildung 5.3 zeigt die Wahrscheinlichkeiten $\alpha(\mathbf{x}_{(\mathbf{j})})$ für den Fehler 1. Art im *ersten* Fall der heuristischen Testansätze, die Kovariablen sind unabhängig identisch verteilt und unabhängig von der Zielvariable, in einem Boxplot:

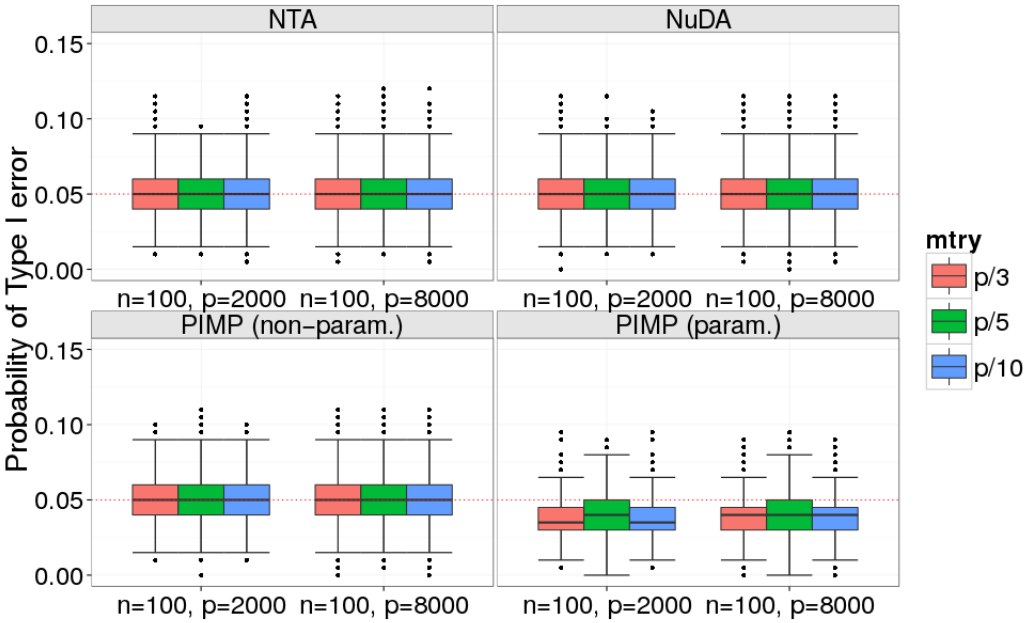

Abbildung 5.3: Boxplot der Wahrscheinlichkeiten für den Fehler 1. Art der Testansätze für die p Kovariablen im *ersten* Fall in der *ersten* Studie und in den beiden Datensätzen ($n = 100$, $p = 2000$ und $n = 100$, $p = 8000$): Oben links der NTA Testansatz, oben rechts $NuDA$ Testansatz und unten recht und links der $PIMP$ Testansatz (nicht-parametrisch und parametrisch). Das vorgegebene Signifikanzniveau $\alpha = 0.05$ ist als rot gepunktete Linie eingezeichnet.

Für den Datensatz mit 2000 Kovariablen und 100 Beobachtungen sind die mittleren und medianen Wahrscheinlichkeiten für den Fehler 1. Art $\alpha(\mathbf{x}_{(j)})$ im *ersten* Fall der heuristischen Testansätze in der Tabelle 5.2 aufgeführt. Da sich die Ergebnisse für den Datensatz mit 8000 Kovariablen und 100 Beobachtungen kaum voneinander unterscheiden, sind diese im Anhang zu finden (vgl. Anhang A.1.2). Der Median der Wahrscheinlichkeiten für den Fehler 1. Art der heuristischen Testansätze, NTA, $NuDA$ und $PIMP$ Testansatz (nicht-parametrisch) ist bei allen drei gleich dem vorgegebenen Signifikanzniveau $\alpha = 0.05$. Auch die Verteilungen der Wahrscheinlichkeiten für den Fehler 1. Art dieser heuristischen Testansätze unterscheiden sich kaum voneinander. Sie sind alle zufällig um das Signifikanzniveau 0.05 verteilt.

	p/3		p/5		p/10	
n=100, p=2000	Mittelwert	Median	Mittelwert	Median	Mittelwert	Median
NTA	0.05003	0.050	0.05040	0.05	0.05062	0.050
NuDA	0.04946	0.050	0.05020	0.05	0.04954	0.050
PIMP(nicht-param.)	0.04907	0.050	0.04969	0.05	0.04907	0.050
PIMP (param.)	0.03761	0.035	0.03950	0.04	0.03761	0.035

Tabelle 5.2: Mittelwert und Median der $p = 2000$ Wahrscheinlichkeiten für den Fehler 1. Art im *ersten* Fall der heuristischen Testansätze. ($n = 100$).

Nur der Median der Wahrscheinlichkeiten für den Fehler 1. Art für den parametrischen $PIMP$ Testansatz ist deutlich kleiner als das vorgegebenen Signifikanzniveau. Im Vergleich zu den anderen heuristischen Testansätzen, ist dieser deutlich konservativer. Die Wahrscheinlichkeiten für den Fehler 1. Art sind für fast alle Kovariablen deutlich kleiner als das vorgegebenen Signifikanzniveau $\alpha = 0.05$. Der Grund dafür ist, dass für die „Null-Verteilungen“ die angenommene Normalverteilung nicht zutrifft. Der implementierte Kolmogorov-Smirnov-Test lehnt die Null-Hypothese oft ab. Zum Beispiel wurde für die 200 Datensätze mit 100 Beobachtungen (n) , 2000 Kovariablen (p) mit der Parametereinstellungen $m_{try} = \frac{p}{3}$ rund 33%, $m_{try} = \frac{p}{5}$ rund 37% und $m_{try} = \frac{p}{10}$ 44% die Annahme, dass die *null importances* Maße normalverteilt sind, abgelehnt ($\alpha = 0.1$).

Fall 2

Für den *zweiten* Fall, die Kovariablen sind miteinander korreliert aber unabhängig von der Zielvariable, ändert sich das Verhalten des heuristischen NTA Testansatzes (vgl. Abbildung 5.4). Der Median und der Mittelwert der Wahrscheinlichkeiten für den Fehler 1. Art in der Tabelle 5.3 des heuristischen NTA Testansatzes, ist minimal aber unwesentlich größer als das vorgegebene Signifikanzniveau $\alpha = 0.05$.

	p/3		p/5		p/10	
n=100, p=2000	Mittelwert	Median	Mittelwert	Median	Mittelwert	Median
NTA	0.05371	0.055	0.05420	0.055	0.05409	0.055
NuDA	0.04963	0.050	0.04970	0.050	0.04932	0.050
PIMP(nicht-param.)	0.04874	0.050	0.04883	0.050	0.04874	0.050
PIMP (param.)	0.03825	0.035	0.03958	0.040	0.03825	0.035

Tabelle 5.3: Mittelwert und Median der $p = 2000$ Wahrscheinlichkeiten für den Fehler 1. Art im *zweiten* Fall der heuristischen Testansätze. ($n = 100$).

Dies bedeutet, es werden im Durchschnitt ein wenig mehr Kovariablen als signifikant abhängig von der Zielvariable erkannt. Anders ausgedrückt: die Null-Hypothese wird öfters abgelehnt. Aber dieser Anstieg der Wahrscheinlichkeit für den Fehler 1. Art ist geringfügig. Das ist verwunderlich, da die beobachteten „Null-Verteilungen“ symmetrisch verteilt sind (vgl. Abschnitt 5.3.1). Die anderen heuristischen Testansätze verhalten sich fast gleich wie beim ersten Fall.

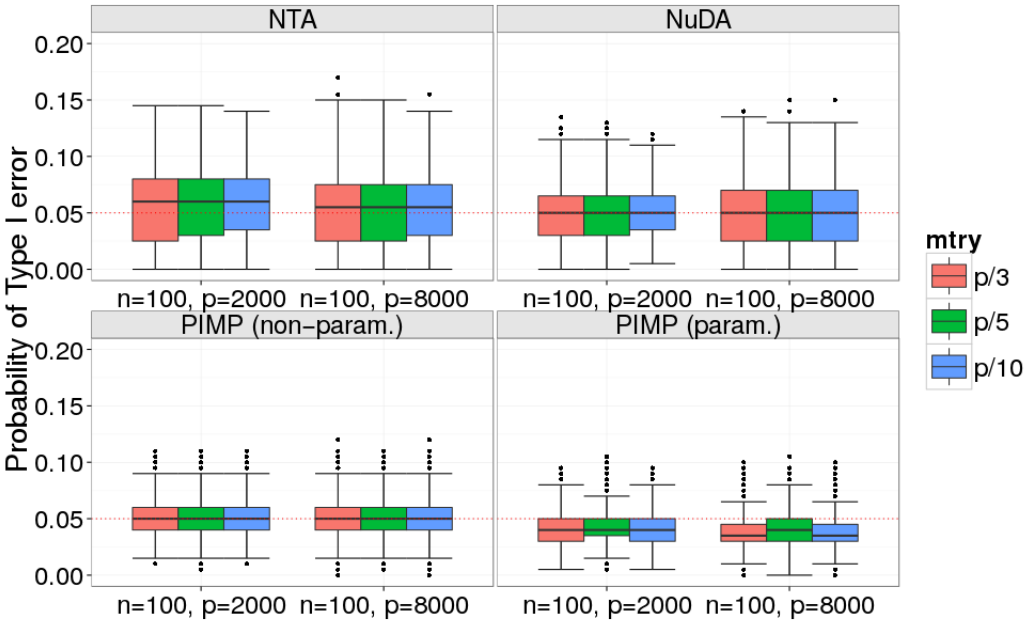

Abbildung 5.4: Boxplot der Wahrscheinlichkeiten für den Fehler 1. Art der der Testansätze für die p Kovariablen im *zweiten* Fall und in beiden Datensätzen ($n = 100, p = 2000$ und $n = 100, p = 8000$): Oben links der *neue* Testansatz (NTA), oben rechts $NuDA$ Testansatz und unten recht und links der $PIMP$ Testansatz (nicht-parametrisch und parametrisch). Das vorgegebene Signifikanzniveau $\alpha = 0.05$ ist als rot gepunktete Linie eingezeichnet.

5.4. Studie II

In der zweiten Studie wird untersucht, wie groß die *Power* (Macht, Trennschärfe) der Testverfahren ist. Die *Power* eines Tests ist die Wahrscheinlichkeit, die Null-Hypothese abzulehnen, falls die Alternativ-Hypothese wahr ist. Anders ausgedrückt, die Gegenwahrscheinlichkeit zur Wahrscheinlichkeit für den Fehler 2. Art β (Null-Hypothese wird beibehalten, obwohl die Null-Hypothese falsch ist). Die *Power* eines guten statistischen Tests sollte möglichst groß sein, bei einem vorgegebenen Signifikanzniveau α. Um die Power der Testverfahren zu bestimmen, müssen die Daten entsprechend der Alternativ-Hypothese simuliert werden. Die Kovariablen und Zielvariable sind abhängig voneinander. Des Weiteren wird, wie in der *ersten* Studie, untersucht, ob die Testverfahren das vorgegebene Signifikanzniveau α einhalten. Infolgedessen müssen die Daten so simuliert werden, dass einige Kovariablen abhängig und manche unabhängig von der Zielvariable sind.

5.4.1. Datengenerierender Prozess für Studie II

Wie bereits erwähnt, ist die Zielvariable abhängig, aber auch unabhängig, von einigen Kovariablen. Die Zielvariablen wurden entsprechend dem linearen Modell mit p Kovariablen erzeugt:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i,\ i = 1, \ldots, n\,, \tag{5.11}$$

wobei die Störvariablenϵ_i unabhängig voneinander aus einer Standardnormalverteilung mit einem Erwartungswert Null und einer Varianz von 0.25 erzeugt wurden:

$$\epsilon_i \overset{i.i.d}{\sim} N(0, 0.25),\ i = 1, \ldots, n\,. \tag{5.12}$$

Bei den Kovariablen wird, wie in der *ersten* Studie, zwischen zwei Fällen unterschieden:

- Im *ersten* Fall sind die Kovariablen untereinander nicht korreliert.
- Im *zweiten* Fall sind die Daten untereinander korreliert.

Fall 1

Wie in der *ersten* Studie im *ersten* Fall, wurde jede der p Kovariablen $\mathbf{X} = (X_{(1)}, \ldots, X_{(p)})^T$ unabhängig voneinander aus einer Standardnormalverteilung mit Erwartungswert Null und einer Varianz von Eins erzeugt (vgl. Gleichung (5.4)). Die Regressionskoeffizienten $\beta_1, \ldots, \beta_p$ wurden im *ersten* Fall folgendermaßen definiert: Die ersten 120 Regressionskoeffizienten $\beta_1, \ldots, \beta_{120}$ sind ungleich Null und die restlichen Regressionskoeffizienten $\beta_{121}, \ldots, \beta_p$ sind gleich Null. Das bedeutet, dass die Zielvariable von den ersten 120 Kovariablen abhängig und von den restlichen Kovariablen unabhängig ist. Für

die Regressionskoeffizienten der ersten 120 Kovariablen wird die achtstellige Zahlenfolge $(-3, -2, -1, -0.5, 0.5, 1, 2, 3)$ 15-mal repliziert. Einen Überblick über die gewählten Regressionskoeffizienten gibt die folgende Tabelle.

$X_{(j)}$	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$	$X_{(6)}$	$X_{(7)}$	$X_{(8)}$	$\cdots$	$X_{(117)}$	$X_{(118)}$	$X_{(119)}$	$X_{(120)}$	$X_{(121)}$	$\cdots$	$X_{(p)}$
β_j	-3	-2	-1	-0.5	0.5	1	2	3	$\cdots$	0.5	1	2	3	0	$\cdots$	0

Tabelle 5.4: Regressionskoeffizienten in der *zweiten* Studie im *ersten* Fall

Fall 2

Die p Kovariablen werden in drei gleich große Teile unterteilt, $\mathbf{X}_{p_1}, \mathbf{X}_{p_2}, \mathbf{X}_{p_3}$ (vgl. Gleichung (5.5)). Wie in der *ersten* Studie im *zweiten* Fall, wird das erste und zweite Drittel der Kovariablen $\mathbf{X}_{p_1}, \mathbf{X}_{p_2}$ aus einer multivariaten Normalverteilung (vgl. Gleichung (5.6)) mit unterschiedlichen Kovarianzmatrixen $\mathbf{\Sigma}_{p_1}, \mathbf{\Sigma}_{p_2}$erzeugt (vgl. Gleichung (5.7) und Gleichung (5.8)). Im dritten Drittel wird jede Kovariable $X_{(j)} \in \mathbf{X}_{p_3}$ aus einer Standardnormalverteilung mit einem Erwartungswert Null und einer Varianz von Eins erzeugt (vgl. Gleichung (5.9)). Die Regressionskoeffizienten $\beta_1, \ldots, \beta_p$ wurden wie die Kovariablen in drei gleich große Teile unterteilt:

$$\begin{aligned}
\boldsymbol{\beta}_{p_1} &= \quad (\beta_1, \ldots, \beta_{p_1})^T \quad &, p_1 = \left\lceil \frac{p}{3} \right\rceil, \\
\boldsymbol{\beta}_{p_2} &= (\beta_{p_1+1}, \ldots, \beta_{p_2})^T \quad &, p_2 = p_1 + \left\lceil \frac{p}{3} \right\rceil, \\
\boldsymbol{\beta}_{p_3} &= (\beta_{p_2+1}, \ldots, \beta_{p_3})^T \quad &, p_3 = p.
\end{aligned} \tag{5.13}$$

Die ersten 40 Regressionskoeffizienten in jedem Drittel der Regressionskoeffizienten $\boldsymbol{\beta}_{p_1}$, $\boldsymbol{\beta}_{p_2}$, $\boldsymbol{\beta}_{p_1}$ sind ungleich Null und die restlichen Regressionskoeffizienten $\beta_{121}, \ldots, \beta_p$ sind gleich Null. Für das erste und zweite Drittel der Kovariablen $\mathbf{X}_{p_1}$, $\mathbf{X}_{p_2}$ bedeutet dies, dass die Zielvariable von den ersten 40 Kovariablen in jedem Drittel direkt abhängig und von den restlichen Kovariablen durch die Korrelation $\rho(X_{(j)}, X_{(k)})$ zwischen den Kovariablen abhängig ist. Im dritten Drittel sind nur die ersten 40 Kovariablen von der Zielvariable abhängig. Die restlichen Kovariablen sind unabhängig. Für die Regressionskoeffizienten der ersten 40 Kovariablen in jedem Drittel wird die achtstellige Zahlenfolge $(-3, -2, -1, -0.5, 0.5, 1, 2, 3)$ 5-mal repliziert. In der Tabelle 5.5 sind die gewählten Regressionskoeffizienten im *zweiten* Fall übersichtlich dargestellt. Es wurden im *zweiten* Fall noch zusätzlich Daten mit anderen Regressionskoeffizienten generiert. In diesem Fall 2B werden die Regressionskoeffizienten der ersten 40 Kovariablen in jedem Drittel mit der Zahlenfolge $(3, 2, 1, 0.5, 0.5, 1, 2, 3)$ erzeugt. Der Grund dafür wird später im Abschnitt 5.5 erklärt.

$X_{(j)} \in \mathbf{X}_{p_1}$	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$	$X_{(6)}$	$X_{(7)}$	$X_{(8)}$	$\cdots$	$X_{(37)}$	$X_{(38)}$	$X_{(39)}$	$X_{(40)}$	$X_{(41)}$	$\cdots$	$X_{(p_1)}$
$\beta_j \in \boldsymbol{\beta}_{p_1}$	-3	-2	-1	-0.5	0.5	1	2	3	$\cdots$	0.5	1	2	3	0	$\cdots$	0

$X_{(j)} \in \mathbf{X}_{p_2}$	$X_{(p_1+1)}$	$X_{(p_1+2)}$	$X_{(p_1+3)}$	$X_{(p_1+4)}$	$X_{(p_1+5)}$	$X_{(p_1+6)}$	$X_{(p_1+7)}$	$X_{(p_1+8)}$	$\cdots$	$X_{(p_1+37)}$	$X_{(p_1+38)}$	$X_{(p_1+39)}$	$X_{(p_1+40)}$	$X_{(p_1+41)}$	$\cdots$	$X_{(p_2)}$
$\beta_j \in \boldsymbol{\beta}_{p_2}$	-3	-2	-1	-0.5	0.5	1	2	3	$\cdots$	0.5	1	2	3	0	$\cdots$	0

$X_{(j)} \in \mathbf{X}_{p_3}$	$X_{(p_2+1)}$	$X_{(p_2+2)}$	$X_{(p_2+3)}$	$X_{(p_2+4)}$	$X_{(p_2+5)}$	$X_{(p_2+6)}$	$X_{(p_2+7)}$	$X_{(p_2+8)}$	$\cdots$	$X_{(p_2+37)}$	$X_{(p_2+38)}$	$X_{(p_2+39)}$	$X_{(p_2+40)}$	$X_{(p_2+41)}$	$\cdots$	$X_{(p)}$
$\beta_j \in \boldsymbol{\beta}_{p_3}$	-3	-2	-1	-0.5	0.5	1	2	3	$\cdots$	0.5	1	2	3	0	$\cdots$	0

Tabelle 5.5: Regressionskoeffizienten in der *zweiten* Studie im *zweiten* Fall

In beiden Fällen wurden auch in der *zweiten* Studie 200 Datensätze (D) mit 100 Beobachtungen (n), 2000 Kovariablen (p) und 200 Datensätze mit 100 Beobachtungen, 8000 Kovariablen erstellt (vgl. Tabelle 5.1). Für die Erstellung der Datensätze wurde eine neue Funktion ($DataStudyII$) in R geschrieben (vgl. Anhang B).

5.4.2. Aufbau der Simulation

Der Ablauf der Simulation und die verwendeten Parameter sind in der *zweiten* Studie gleich wie in der *ersten* Studie (vgl. Abschnitt 5.2.2 und Algorithmus 8).

5.4.3. Evaluationsgrößen der Studie I

Die Power der Testansätze wird durch die relative Häufigkeit der Fälle, in denen die Null-Hypothese zu Recht verworfen wurde, geschätzt:

$$power(\mathbf{x_{(j)}}) = \frac{1}{D} \sum_{d=1}^{D} I\left(p(\mathbf{x_{(j)}}) \leq 0.05\right), \tag{5.14}$$

wobei D die Anzahl der generierten Datensätze ist. Welche Kovariablen von der Zielvariable abhängig sind, muss im Vorhinein bekannt sein, um die Power der Testansätze bestimmen zu können. Wie im Abschnitt 5.1 auf Seite 27 erklärt, unterscheiden sich die Hypothesen, sobald die Kovariablen untereinander korreliert sind. Im *ersten* Fall sind die p Kovariablen voneinander unabhängig und die Zielvariable ist von den ersten 120 Kovariablen abhängig (vgl. 5.1). Im *zweiten* Fall sind nicht alle p Kovariablen voneinander unabhängig, aber es ist bekannt, welche der Kovariablen abhängig von der Zielvariable und welche Kovariable mit den anderen Kovariablen korreliert sind (vgl. 5.2). In beiden Fällen wird die Power der Testansätze in Abhängigkeit der absoluten Regressionskoeffizienten 0.5, 1, 2, 3 untersucht. Des Weiteren wird für jeden Testansatz die Wahrscheinlichkeit für den Fehler 1. Art wie in der ersten Studie untersucht (vgl. Gleichung (5.10)). Durch den datengenerierenden Prozess ist in beiden Fällen bekannt, welche der Kovariablen unabhängig von der Zielvariable sind.
Die Power und die Wahrscheinlichkeiten für den Fehler 1. Art der verschiedenen heuristischen Testansätze, werden für verschiedene Parametereinstellungen mithilfe von Boxplots

miteinander verglichen und ggf. durch zusammenfassende Tabellen (Mittelwert, Median) ergänzt.

5.5. Ergebnisse der Studie II

5.5.1. Power und Fehler 1. Art

Die Power (vgl. Gleichung (5.14)) und die Wahrscheinlichkeiten für den Fehler 1. Art (vgl. Gleichung (5.10)) der heuristischen Testansätze, werden separat für jeden Fall betrachtet.

Power im Fall 1

Die folgende Abbildung 5.5 zeigt die Power der heuristischen Testansätze in Abhängigkeit der absoluten Regressionskoeffizienten $|\beta_j| \in \{0.5, 1, 2, 3\}$ im *ersten* Fall: Die Kovariablen sind unabhängig identisch verteilt und die ersten 120 Kovariablen sind abhängig von der Zielvariable:

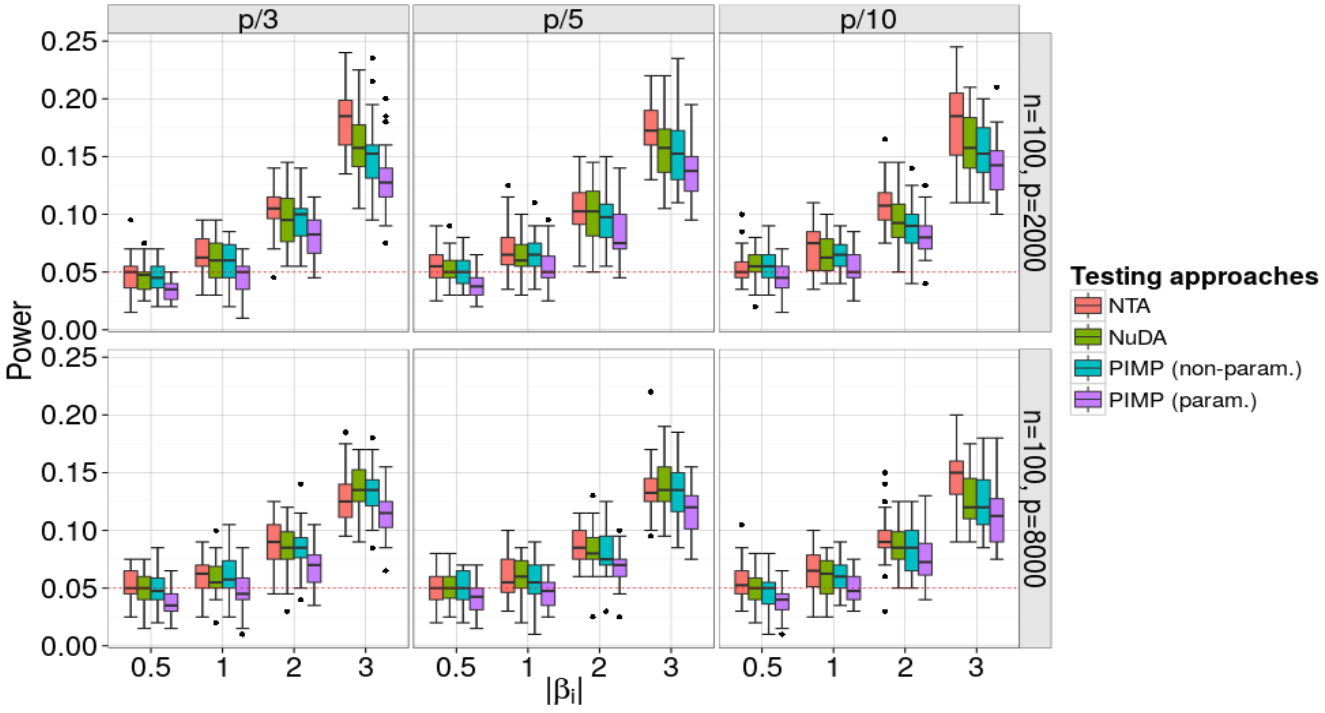

Abbildung 5.5: Boxplot der Power der Testansätze in Abhängigkeit der absoluten Regressionskoeffizienten $|\beta_j| \in \{0.5, 1, 2, 3\}$ im *ersten* Fall und für beide Datensätze: Oben im Datensatz mit $n = 100$, $p = 2000$ und unten mit $n = 100$, $p = 8000$. Links mit der Parametereinstellung $m_{try} = \frac{p}{3}$, in der Mitte $m_{try} = \frac{p}{5}$ und rechts $m_{try} = \frac{p}{10}$. Die Power der Testansätze, NTA, $NuDA$ und $PIMP$ (nicht-parametrisch und parametrisch) ist in verschiedenen Farben eingezeichnet. Das vorgegebene Signifikanzniveau $\alpha = 0.05$ ist als rot gepunktete Linie eingezeichnet.

Die mittlere und mediane Power im *ersten* Fall der heuristischen Testansätze mit der Parametereinstellung $m_{try} = \frac{p}{3}$, ist in der Tabelle 5.6 für den Datensatz mit 2000 Kova-

riablen und in der Tabelle 5.7 für den Datensatz mit 8000 Kovariablen aufgeführt. Da sich die Ergebnisse mit den Parametereinstellungen $m_{try} = \frac{p}{5}$ und $m_{try} = \frac{p}{10}$ sich kaum unterscheiden, sind diese im Anhang zu finden (vgl. Anhang A.2.1). Die Power aller vier Testansätze steigt mit der absoluten zunehmenden Größe der Regressionskoeffizienten in beiden Datensätzen und für alle Parametereinstellungen von m_{try}. Anders ausgedrückt, die Testansätze entdecken, die Abhängigkeit zwischen den Kovariablen und der Zielvariable besser mit einem großen Effekt ($|\beta_j| = 3$) als mit kleinen Effekten ($|\beta_j| = 0.5$). Im Datensatz mit 2000 Kovariablen und 100 Beobachtungen ist die mediane Power des NTA Testansatzes etwas größer als der anderen Testansätze. Die mediane Power der Testansätze, $NuDA$ und $PIMP$ (nicht-parametrisch) unterscheiden sich kaum voneinander.

	$\|\beta_j\| = 0.5$		$\|\beta_j\| = 1$		$\|\beta_j\| = 2$		$\|\beta_j\| = 3$	
n=100, p=2000	Mittelwert	Median	Mittelwert	Median	Mittelwert	Median	Mittelwert	Median
NTA	0.04883	0.0500	0.06567	0.0625	0.10367	0.1050	0.1818	0.1850
NuDA	0.04500	0.0475	0.05950	0.0600	0.09550	0.0950	0.1580	0.1575
PIMP (nicht-param.)	0.04633	0.0450	0.05750	0.0600	0.09533	0.1000	0.1520	0.1525
PIMP (param.)	0.03450	0.0350	0.04417	0.0500	0.08100	0.0825	0.1312	0.1275

Tabelle 5.6: Mittelwert und Median der Power im *ersten* Fall der heuristischen Testansätze mit der Parametereinstellung $m_{try} = \frac{p}{3}$, aufgeteilt in die absoluten Regressionskoeffizienten $|\beta_j| \in \{0.5, 1, 2, 3\}$ ($n = 100$, $p = 2000$).

Im Datensatz mit 8000 Kovariablen und 100 Beobachtungen ist der Unterschied der medianen Power zwischen den heuristischen Testansätzen, NTA, $NuDA$ und $PIMP$ (nicht-parametrisch), sehr gering. Die Power der Testansätze ist im Datensatz mit 8000 Kovariablen etwas niedriger als wie im Datensatz mit 2000 Kovariablen. Der parametrische $PIMP$ Testansatz hat durchwegs die kleinste Power in beiden Datensätzen. Wie in Abschnitt 5.3.2 auf Seite 33 schon erklärt, wird vermutet, dass die angenommene Normalverteilung für die „Null-Verteilungen“ nicht zutrifft.

	$\|\beta_j\| = 0.5$		$\|\beta_j\| = 1$		$\|\beta_j\| = 2$		$\|\beta_j\| = 3$	
n=100, p=8000	Mittelwert	Median	Mittelwert	Median	Mittelwert	Median	Mittelwert	Median
NTA	0.05350	0.0500	0.06067	0.0625	0.08917	0.090	0.1297	0.125
NuDA	0.04967	0.0500	0.05850	0.0550	0.08367	0.085	0.1352	0.135
PIMP (nicht-param.)	0.04950	0.0475	0.06167	0.0575	0.08483	0.085	0.1323	0.135
PIMP (param.)	0.03683	0.0350	0.04867	0.0450	0.06800	0.070	0.1138	0.115

Tabelle 5.7: Mittelwert und Median der Power im *ersten* Fall der heuristischen Testansätze mit der Parametereinstellung $m_{try} = \frac{p}{3}$, aufgeteilt in die absoluten Regressionskoeffizienten $|\beta_j| \in \{0.5, 1, 2, 3\}$. ($n = 100$, $p = 8000$).

Fehler 1. Art im Fall 1

In der Abbildung 5.6 sind die Wahrscheinlichkeiten für den Fehler 1. Art der heuristischen Testansätze im *ersten* Fall der *zweiten* Studie als Boxplot dargestellt. Aus dieser Abbildung erkennt man, dass im Datensatz mit 2000 Kovariablen, die medianen Wahrscheinlichkeiten für den Fehler 1. Art der heuristischen Testansätze $NuDA$ und $PIMP$ (nicht-parametrisch) kleiner als das vorgegebene Signifikanzniveau $\alpha = 0.05$ sind.

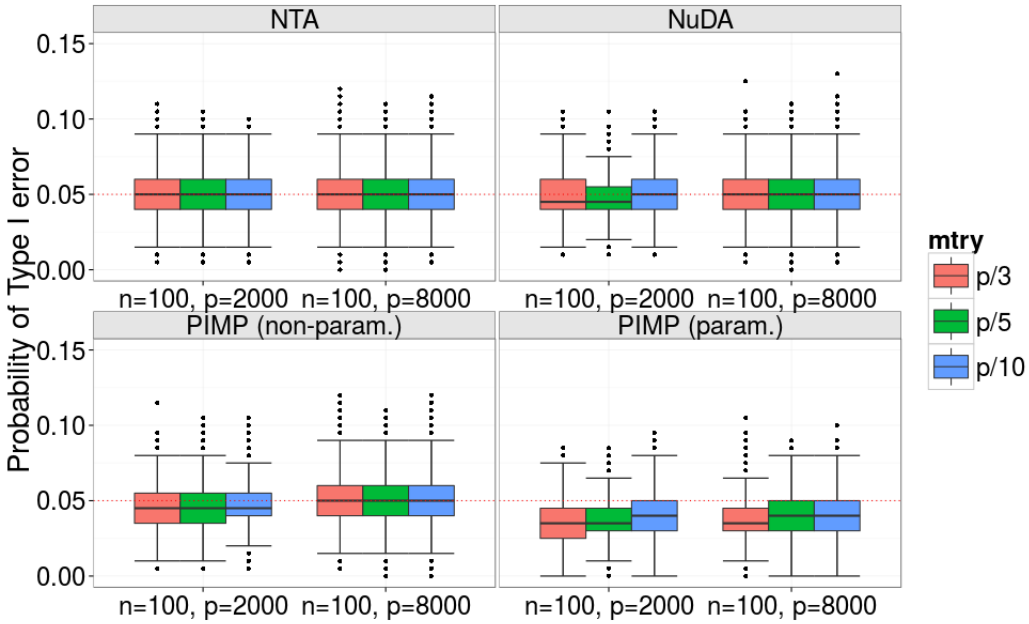

Abbildung 5.6: Boxplot der Wahrscheinlichkeiten für den Fehler 1. Art der Testansätze für die Kovariablen mit Regressionskoeffizienten $\beta_j = 0$ im *ersten* Fall in der *zweiten* Studie in beiden Datensätzen ($n = 100$, $p = 2000$ und $n = 100$, $p = 8000$): Oben links der NTA Testansatz, oben rechts $NuDA$ Testansatz und unten recht und links der $PIMP$ Testansatz (nicht-parametrisch und parametrisch). Das vorgegebene Signifikanzniveau $\alpha = 0.05$ ist als rot gepunktete Linie eingezeichnet.

Aus der Tabelle 5.8 ist ersichtlich, dass diese Abweichung minimal ist. Die anderen heuristischen Testansätze unterscheiden sich kaum von den Ergebnissen aus der *ersten* Studie (vgl. 5.2 auf Seite 34).

	p/3		p/5		p/10	
n=100, p=2000	Mittelwert	Median	Mittelwert	Median	Mittelwert	Median
NTA	0.05087	0.050	0.05071	0.050	0.04989	0.050
NuDA	0.04793	0.045	0.04781	0.045	0.04851	0.050
PIMP (nicht-param.)	0.04748	0.045	0.04760	0.045	0.04773	0.045
PIMP (param.)	0.03630	0.035	0.03772	0.035	0.04028	0.040

Tabelle 5.8: Mittelwert und Median der Wahrscheinlichkeiten für den Fehler 1. Art der Kovariablen mit Regressionskoeffizienten $\beta_j = 0$ im *ersten* Fall in der *zweiten* Studie der heuristischen Testansätze ($n = 100$, $p = 2000$).

Power im Fall 2

Im *zweiten* Fall wurden die p Kovariablen in drei gleich große Teile unterteilt, $\mathbf{X}_{p_1}$, $\mathbf{X}_{p_2}$, $\mathbf{X}_{p_3}$. Die ersten beiden Drittel der Kovariablen $\mathbf{X}_{p_1}, \mathbf{X}_{p_2}$ sind untereinander korreliert und im dritten Drittel $\mathbf{X}_{p_3}$ sind die Kovariablen unabhängig identisch verteilt. Jeweils die ersten 40 Kovariablen in jedem Drittel sind abhängig von der Zielvariable (vgl. Abschnitt 5.4.1 auf Seite 37). Die Abbildung 5.7 zeigt in einem Boxplot die Power der heuristischen Testansätze in Abhängigkeit der absoluten Regressionskoeffizienten $|\beta_j| \in \{0.5, 1, 2, 3\}$ im *zweiten* Fall. Wie im *ersten* Fall, steigt die Power aller vier Testansätze in beiden Datensätzen und für alle Parametereinstellung von m_{try} mit zunehmender Größe der Regressionskoeffizienten an. Die Streuung der Power der vier Testansätze ist aber viel größer als wie im ersten Fall.

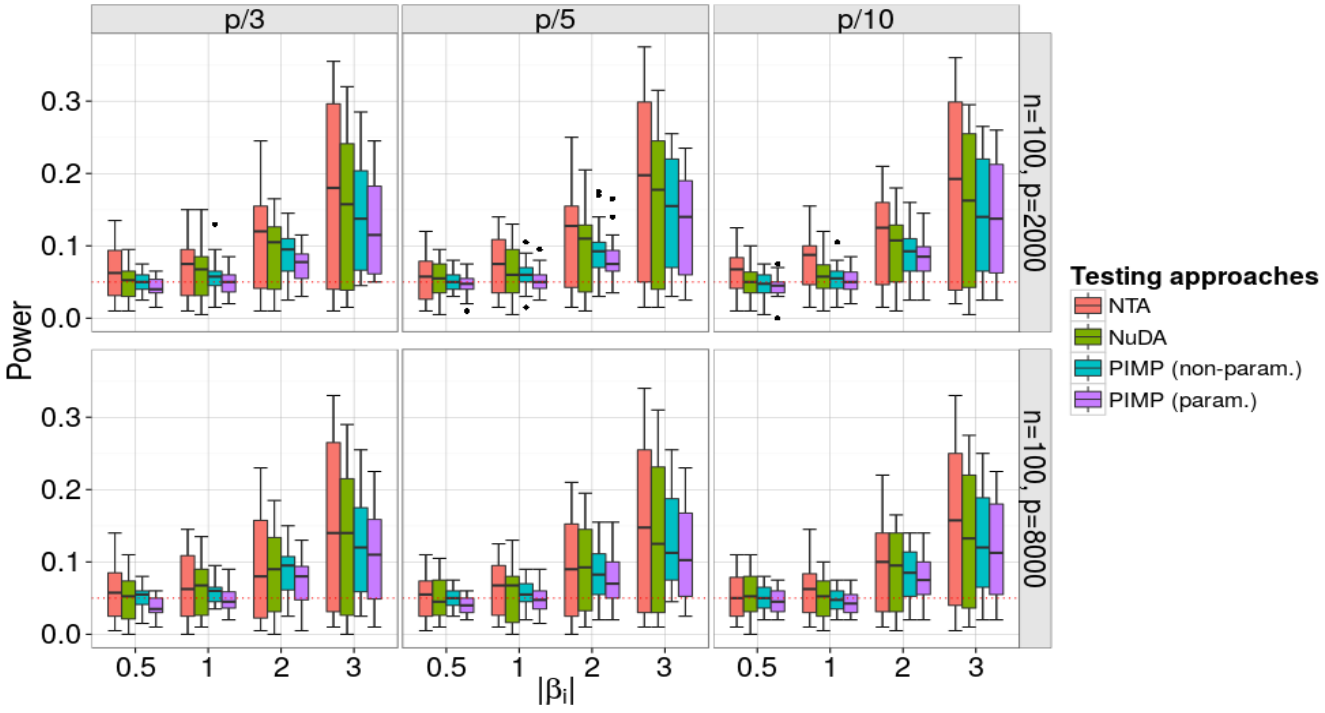

Abbildung 5.7: Boxplot der Power der Testansätze in Abhängigkeit der absoluten Regressionskoeffizienten $|\beta_j| \in \{0.5, 1, 2, 3\}$ im *zweiten* Fall und für beide Datensätze: Oben im Datensatz mit $n = 100$, $p = 2000$ und unten mit $n = 100$, $p = 8000$. Links mit der Parametereinstellung $m_{try} = \frac{p}{3}$, in der Mitte $m_{try} = \frac{p}{5}$ und rechts $m_{try} = \frac{p}{10}$. Die Power der Testansätze, NTA, $NuDA$ und $PIMP$ (nicht-parametrisch und parametrisch) ist in verschiedenen Farben eingezeichnet. Das vorgegebene Signifikanzniveau $\alpha = 0.05$ ist als rot gepunktete Linie eingezeichnet.

Um diese Streuung besser verstehen zu können, wurde in der Abbildung 5.8 separat für die drei Kovariablengruppen ($\mathbf{X}_{p_1}$, $\mathbf{X}_{p_2}$, $\mathbf{X}_{p_3}$) die Power der heuristischen Testansätze in Abhängigkeit der absoluten Größe der Regressionskoeffizienten $|\beta_j| \in \{0.5, 1, 2, 3\}$ mit der Parametereinstellung $m_{try} = \frac{p}{3}$ anhand von Boxplots dargestellt. Aus dieser

Abbildung 5.8 ist ersichtlich, dass die Power aller vier Testansätze mit zunehmender Korrelation zwischen den Kovariablen in beiden Datensätzen sinkt. Aufgrund der geringen Unterschiede in den Ergebnissen mit den Parametereinstellungen $m_{try} = \frac{p}{5}$ und $m_{try} = \frac{p}{10}$, sind diese im Anhang zu finden (vgl. Anhang A.2.1).

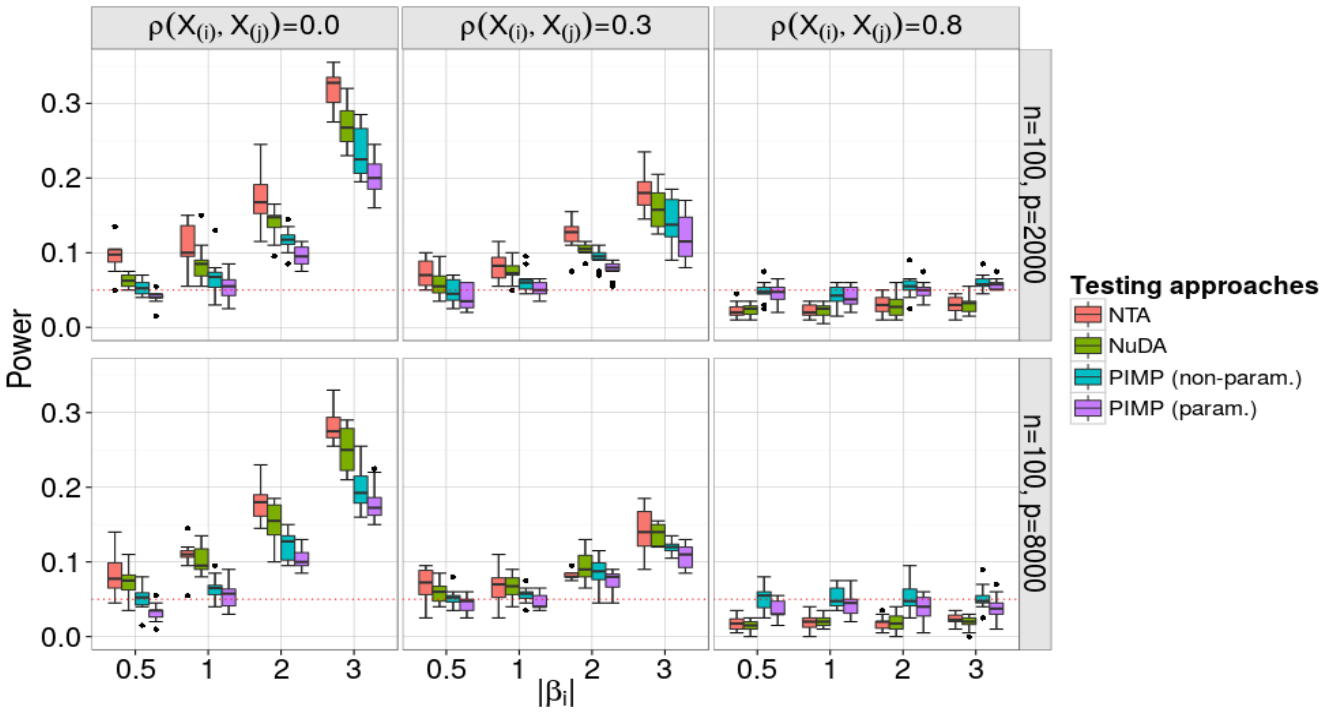

Abbildung 5.8: Boxplot der Power der Testansätze in Abhängigkeit der absoluten Regressionskoeffizienten $|\beta_j| \in \{0.5, 1, 2, 3\}$ im *zweiten* Fall mit der Parametereinstellung $m_{try} = \frac{p}{3}$ in beiden Datensätzen: Oben für den Datensatz mit $n = 100, p = 2000$ und unten für $n = 100, p = 8000$. Links für jede Kovariable $\mathbf{x}_{(j)} \in \mathbf{X}_{p_3} \wedge |\beta_j| \neq 0$ mit $\rho(X_{(j)}, X_{(k)}) = 0.0$, $j, k = p_2 + 1, \ldots, p_3$, $j \neq k$, in der Mitte für jede Kovariable $\mathbf{x}_{(j)} \in \mathbf{X}_{p_1} \wedge |\beta_j| \neq 0$ mit $\rho(X_{(j)}, X_{(k)}) = 0.3$, $j, k = 1, \ldots, p_1$, $j \neq k$ und rechts für jede Kovariable $\mathbf{x}_{(j)} \in \mathbf{X}_{p_2} \wedge |\beta_j| \neq 0$ mit $\rho(X_{(j)}, X_{(k)}) = 0.8$, $j, k = (p_1 + 1), \ldots, p_2$, $j \neq k$. Das vorgegebene Signifikanzniveau $\alpha = 0.05$ ist als rot gepunktete Linie eingezeichnet.

Im dritten Drittel der Kovariablen $\mathbf{X}_{p_3}$, die Kovariablen sind unabhängig identisch verteilt ($\rho(X_{(j)}, X_{(k)}) = 0.0$, $j, k = p_2 + 1, \ldots, p_3$ und $j \neq k$), ist die Power aller vier Testansätze sogar größer als in der *ersten* Studie. Zum Beispiel steigt im ersten Datensatz (n=100, p=2000) für eine Kovariable mit einem kleinen Effekt $|\beta_j| = 0.5$ die mediane Power des *NTA* Testansatzes von 0.05 auf 0.0975 an (vgl. Tabelle 5.9 auf der nächsten Seite und 5.6 auf Seite 41). Ebenfalls im Datensatz mit 8000 Kovariablen ist die mediane Power der Testansätze größer als in der *ersten* Studie (vgl. Tabelle 5.10 auf der nächsten Seite und 5.7 auf Seite 41). Wie in der *ersten* Studie nimmt die Power aller vier Testansätze im Datensatz mit 8000 Kovariablen im Vergleich zum Datensatz mit 2000 Kovariablen ab. Des Weiteren ist zu beobachten, dass die mediane Power des *NTA* Testansatzes

größer als die der anderen Testansätze ist. Die mediane Power der Testansätze $NuDA$ und $PIMP$ (nicht-parametrisch) unterscheiden sich im Vergleich zum *ersten* Fall sehr voneinander. Der $NuDA$ Testansatz hat eine größere mediane Power als der $PIMP$ (nicht-parametrisch) Testansatz. Der parametrische $PIMP$ Testansatz hat durchwegs die kleinste Power.

Median (n=100,p=2000)	$\rho(X_{(j)}, X_{(k)}) = 0.0$				$\rho(X_{(j)}, X_{(k)}) = 0.3$				$\rho(X_{(j)}, X_{(k)}) = 0.8$			
$\lvert\beta_j\rvert$	0.5	1	2	3	0.5	1	2	3	0.5	1	2	3
NTA	0.0975	0.1000	0.168	0.328	0.070	0.0825	0.128	0.180	0.0200	0.0200	0.0300	0.0300
NuDA	0.0625	0.0850	0.147	0.268	0.055	0.0725	0.105	0.158	0.0250	0.0250	0.0275	0.0325
PIMP (non-param.)	0.0525	0.0675	0.117	0.225	0.045	0.0600	0.095	0.138	0.0475	0.0425	0.0550	0.0575
PIMP (param.)	0.0400	0.0550	0.095	0.200	0.035	0.0500	0.080	0.115	0.0475	0.0375	0.0500	0.0575

Tabelle 5.9: Median der Power im *zweiten* Fall der heuristischen Testansätze mit der Parametereinstellung $m_{try} = \frac{p}{3}$, aufgeteilt in die drei Kovariablengruppen mit den unterschiedlichen Korrelationen zwischen den Kovariablen und in die absoluten Regressionskoeffizienten $|\beta_j| \in \{0.5, 1, 2, 3\}$ ($n = 100$, $p = 2000$).

Im ersten Drittel der Kovariablen $\mathbf{X}_{p_1}$, die Kovariablen sind untereinander korreliert mit 0.3, ist die Power aller vier Testansätze fast gleich groß wie in der *ersten* Studie. Die mediane Power des NTA Testansatzes ist etwas größer als die der anderen Testansätze im Datensatz mit 2000 Kovariablen. Dieser Unterschied der medianen Power zwischen den heuristischen Testansätzen NTA, $NuDA$ und $PIMP$ (nicht-parametrisch), wird im Datensatz mit 8000 Kovariablen geringer.
Für die Kovariablen, die untereinander eine Korrelation von $\rho(X_{(j)}, X_{(k)}) = 0.8$, $j, k = (p_1 + 1), \ldots, p_2$ und $j \neq k$ aufweisen, sinkt die Power der vier Testansätze unter bzw. auf das vorgegebene Signifikanzniveau $\alpha = 0.05$. Das bedeutet, dass die vier Testansätze in diesem Drittel nur zufällig eine von der Zielvariable abhängige Kovariable entdecken. Auffallend ist, dass die mediane Power der Testansätze NTA und $NuDA$ sogar deutlich kleiner als das vorgegebene Signifikanzniveau $\alpha = 0.05$ wird.

Median (n=100,p=8000)	$\rho(X_{(j)}, X_{(k)}) = 0.0$				$\rho(X_{(j)}, X_{(k)}) = 0.3$				$\rho(X_{(j)}, X_{(k)}) = 0.8$			
$\lvert\beta_j\rvert$	0.5	1	2	3	0.5	1	2	3	0.5	1	2	3
NTA	0.0775	0.1100	0.1800	0.2750	0.0725	0.0700	0.0800	0.14	0.0175	0.0200	0.0200	0.0225
NuDA	0.0750	0.0950	0.1550	0.2500	0.0600	0.0675	0.0900	0.14	0.0150	0.0200	0.0175	0.0200
PIMP (non-param.)	0.0525	0.0650	0.1275	0.1925	0.0525	0.0575	0.0875	0.12	0.0550	0.0475	0.0475	0.0475
PIMP (param.)	0.0350	0.0575	0.1000	0.1725	0.0475	0.0400	0.0800	0.11	0.0300	0.0450	0.0400	0.0375

Tabelle 5.10: Median der Power im *zweiten* Fall der heuristischen Testansätze mit der Parametereinstellung $m_{try} = \frac{p}{3}$, aufgeteilt in die drei Kovariablengruppen mit den unterschiedlichen Korrelation zwischen den Kovariablen und in die absoluten Regressionskoeffizienten $|\beta_j| \in \{0.5, 1, 2, 3\}$ ($n = 100$, $p = 8000$).

Im *zweiten* Fall wird außerdem untersucht, wie groß die Power der vier heuristischen Testverfahren ist, wenn die Kovariablen nur über die Korrelation $\rho(X_{(j)}, X_{(k)})$ zwischen den Kovariablen mit der Zielvariable abhängig sind. Wie in Abschnitt 5.1 auf Seite 27 erklärt, kann die Null-Hypothese 5.2 auf Seite 27 verworfen werden, wenn eine Kovaria-

ble $X_{(j)}$ und die restlichen Kovariablen $Z = X_{(1)}, \ldots, X_{(j-1)}, X_{(j+1)}, \ldots, X_{(p)}$ abhängig voneinander sind. In der Abbildung 5.9 ist die Power der vier heuristischen Testansätze für die Kovariablen aus den Kovariablengruppen $\mathbf{X}_{p_1}, \mathbf{X}_{p_2}$ und mit einen Regressionskoeffizienten von $\beta_j = 0$ im Datensatz mit 2000 Kovariablen und 100 Beobachtungen als Boxplot dargestellt. Da sich die Ergebnisse für den Datensatz mit 8000 Kovariablen und 100 Beobachtungen kaum voneinander unterscheiden, sind diese im Anhang aufgeführt (vgl. Anhang A.2.1).

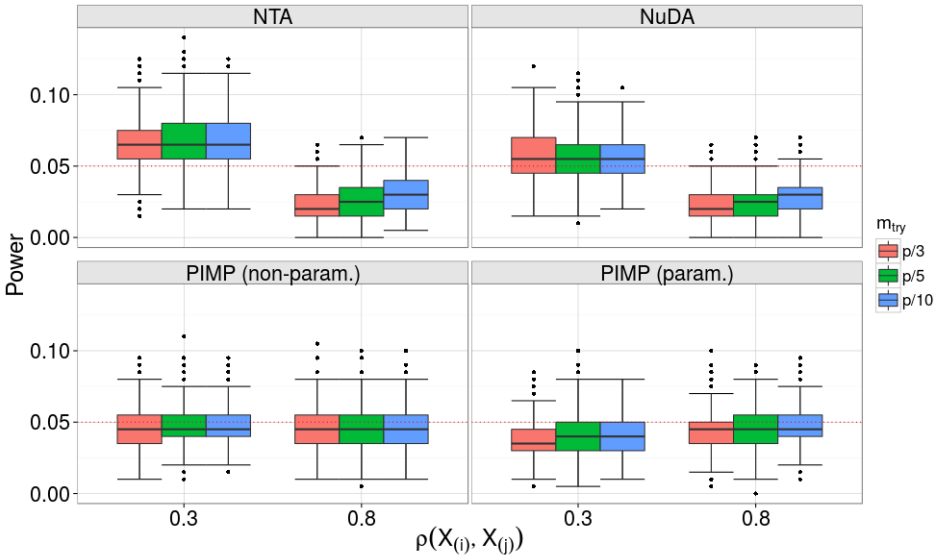

Abbildung 5.9: Boxplot der Power für die Kovariablen aus den Kovariablengruppen $\mathbf{X}_{p_1}, \mathbf{X}_{p_2} \wedge$ $\beta_j = 0$ der Testansätze im Datensatz mit $n = 100$, $p = 2000$: Auf der x-Achse sind die zwei Kovariablengruppen $\mathbf{X}_{p_1}, \mathbf{X}_{p_2}$ dargestellt, wobei 0.3 auf das erste Drittel der Kovariablen $\mathbf{X}_{p_1}$ mit $\rho(X_{(j)}, X_{(k)}) = 0.3,\ j, k = 1, \ldots, p_1,\ j \neq k$ und 0.8 auf zweite Drittel der Kovariablen $\mathbf{X}_{p_2}$ mit $\rho(X_{(j)}, X_{(k)}) = 0.8,\ j, k = (p_1 + 1), \ldots, p_2,\ j \neq k$ verweist. Oben links der NTA Testansatz, oben rechts der $NuDA$ Testansatz und unten recht, und links der $PIMP$ Testansatz(nicht-parametrisch und parametrisch). Das vorgegebene Signifikanzniveau $\alpha = 0.05$ ist als rot gepunktete Linie eingezeichnet.

Für die Kovariablen aus der Kovariablengruppe $\mathbf{X}_{p_1}$ mit einen Regressionskoeffizienten von $\beta_j = 0$ und mit einer Korrelation zwischen den Kovariablen von 0.3, ist die Power der heuristischen Testansätze, NTA, $NuDA$ und $PIMP$ (nicht-parametrisch), nicht wesentlich größer bzw. kleiner als das vorgegebene Signifikanzniveau $\alpha = 0.05$. Der parametrische $PIMP$ Testansatz hat durchwegs die kleinste Power in dieser Kovariablengruppe. Die mediane Power der Testansätze, NTA und $NuDA$ wird sogar deutlich kleiner als das vorgegebene Signifikanzniveau $\alpha = 0.05$ für die Kovariablen aus der Kovariablengruppe $\mathbf{X}_{p_2}$ mit einen Regressionskoeffizienten von $\beta_j = 0$ und mit einer Korrelation zwischen den Kovariablen von 0.8. Die $PIMP$ Testansätze (nicht-parametrisch und parametrisch) haben ungefähr die gleiche Power wie in der anderen Kovariablengruppe (vgl. Tabelle 5.11). Allerdings ist die Power der vier Testansätze in beiden Kovariablengruppen nie

deutlich größer als das vorgegebene Signifikanzniveau $\alpha = 0.05$. Diese Ergebnisse verwundern nicht, da die Power der vier Testansätze in beiden Kovariablengruppen für eine Kovariable mit einem kleinen Effekt $|\beta_j| = 0.5$ schon sehr gering war.

Median (n=100,p=2000)	$\rho(X_{(j)}, X_{(k)}) = 0.3$			$\rho(X_{(j)}, X_{(k)}) = 0.8$		
m_{try}	$p/3$	$p/5$	$p/10$	$p/3$	$p/5$	$p/10$
NTA	0.065	0.065	0.065	0.020	0.025	0.030
NuDA	0.055	0.055	0.055	0.020	0.025	0.030
PIMP (non-param.)	0.045	0.045	0.045	0.045	0.045	0.045
PIMP (param.)	0.035	0.040	0.040	0.045	0.045	0.045

Tabelle 5.11: Median der Power im *zweiten* Fall in der *zweiten* Studie der heuristischen Testansätze für die Kovariablen aus den Kovariablengruppen $\mathbf{X}_{p_1}, \mathbf{X}_{p_2}$ und einen Regressionskoeffizienten von $\beta_j = 0$ ($n = 100$, $p = 2000$).

In der „kleinen“ Simulationsstudie in Abschnitt 4.6 auf Seite 21 wurde beobachtet, dass die beobachteten „Null-Verteilungen“ der *Variablenwichtigkeitsmaße,* mit steigender Korrelation zwischen den Kovariablen, nicht mehr um Null verteilt sind. Auch in der Studie von Strobl u. a. (2008) wurde beobachtet, dass das *Permutation Variablenwichtigkeitsmaß* für Kovariablen, die miteinander korreliert sind (egal ob abhängig oder nicht abhängig von der Zielvariable) größer ist als für Kovariablen die untereinander sind unabhängig, sind (vgl. Strobl u. a., 2008, S. 5ff). Wegen diesen beobachteten Eigenschaften hätte man erwartet, dass die Power der heuristischen Testansätze in der Kovariablengruppe $\mathbf{X}_{p_2}$ mit der Korrelation zwischen den Kovariablen von 0.8 am größten ist. Die Power aller vier Testansätze sinkt mit zunehmender Korrelation zwischen den Kovariablen in beiden Datensätzen und steigt nicht an. Was ist in dieser *zweiten* Studie passiert? Es wird angenommen, dass die symmetrische Anordnung der Regressionskoeffizienten $(-3, -2, -1, -0.5, 0.5, 1, 2, 3)$ in den Kovariablengruppen diesen beschriebenen Effekt umgedreht hat. Um diese Annahme zu bestätigen, wurde noch einmal eine Simulation mit modifizierten Daten in der *zweiten* Studie im *zweiten* Fall durchgeführt. Bei diesen modifizierten Daten wurden die Regressionskoeffizienten der ersten 40 Kovariablen in jedem Drittel mit dieser achtstelligen Zahlenfolge $(3, 2, 1, 0.5, 0.5, 1, 2, 3)$ erzeugt (vgl. Abschnitt 5.4.1 auf Seite 37).

Power im Fall 2B

Wie am Ende des vorherigen Abschnitts erklärt, sind die Regressionskoeffizienten der ersten 40 Kovariablen in jedem Drittel für die Kovariablengruppen $\mathbf{X}_{p_1}$, $\mathbf{X}_{p_2}$, $\mathbf{X}_{p_3}$ durch Replikation der Zahlenfolge $(3, 2, 1, 0.5, 0.5, 1, 2, 3)$ erzeugt worden. Die Abbildung 5.10 zeigt in einem Boxplot die Power der heuristischen Testansätze in Abhängigkeit der Regressionskoeffizienten $\beta_j \in \{0.5, 1, 2, 3\}$ im Fall 2B. Nur die mediane Power des NTA Testansatzes steigt mit zunehmender Größe der Regressionskoeffizienten an. Während

die mediane Power der anderen Testansätze sehr nah am Signifikanzniveau ist. Wie im *zweiten* Fall, ist die Streuung der Power der vier Testansätze sehr groß.

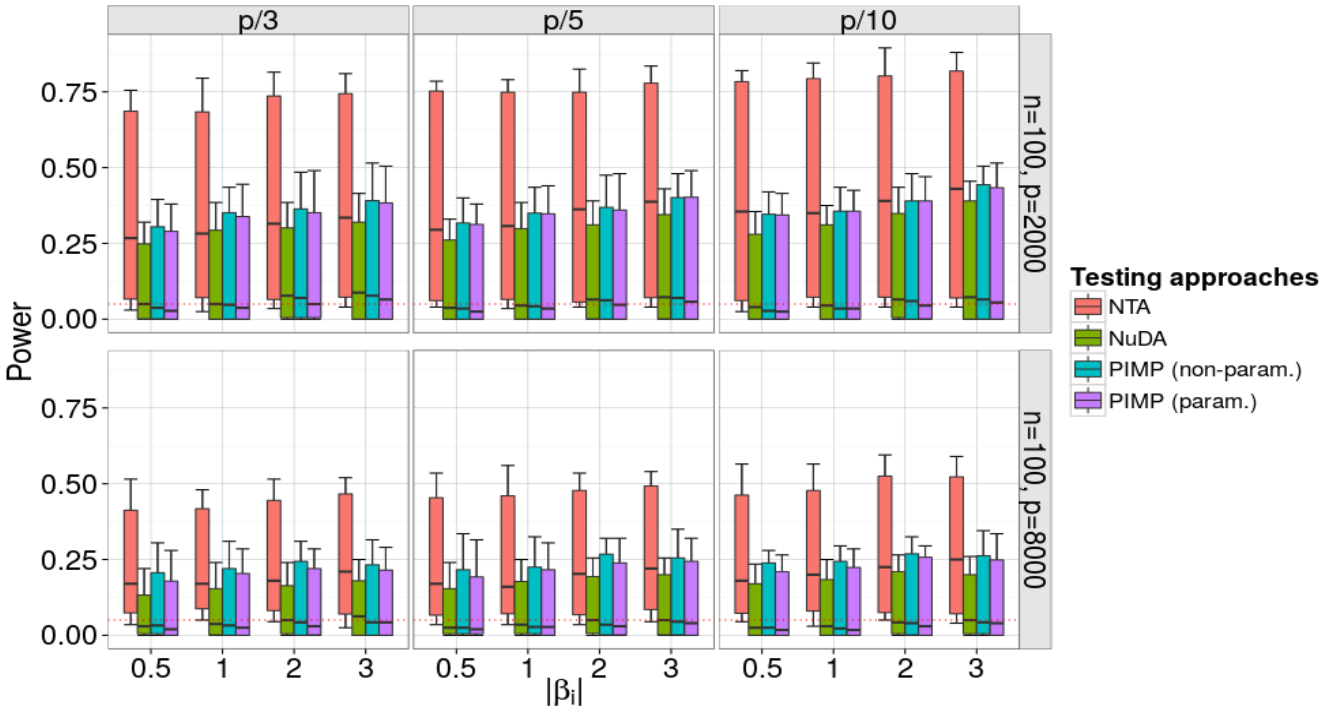

Abbildung 5.10: Boxplot der Power der Testansätze in Abhängigkeit der absoluten Regressionskoeffizienten $\beta_j \in \{0.5, 1, 2, 3\}$ im Fall 2B und für beide Datensätze: Oben im Datensatz mit $n = 100$, $p = 2000$ und unten mit $n = 100$, $p = 8000$. Links mit der Parametereinstellung $m_{try} = \frac{p}{3}$, in der Mitte $m_{try} = \frac{p}{5}$ und rechts $m_{try} = \frac{p}{10}$. Die Power der Testansätze, NTA, $NuDA$ und $PIMP$ (nicht-parametrisch und parametrisch), ist in verschiedenen Farben eingezeichnet. Das vorgegebene Signifikanzniveau $\alpha = 0.05$ ist als rot gepunktete Linie eingezeichnet.

Die folgende Abbildung 5.11 zeigt in einem Boxplot die Power der heuristischen Testansätze für die drei Kovariablengruppen separat in Abhängigkeit der Regressionskoeffizienten $\beta_j \in \{0.5, 1, 2, 3\}$ mit der Parametereinstellung $m_{try} = \frac{p}{3}$:

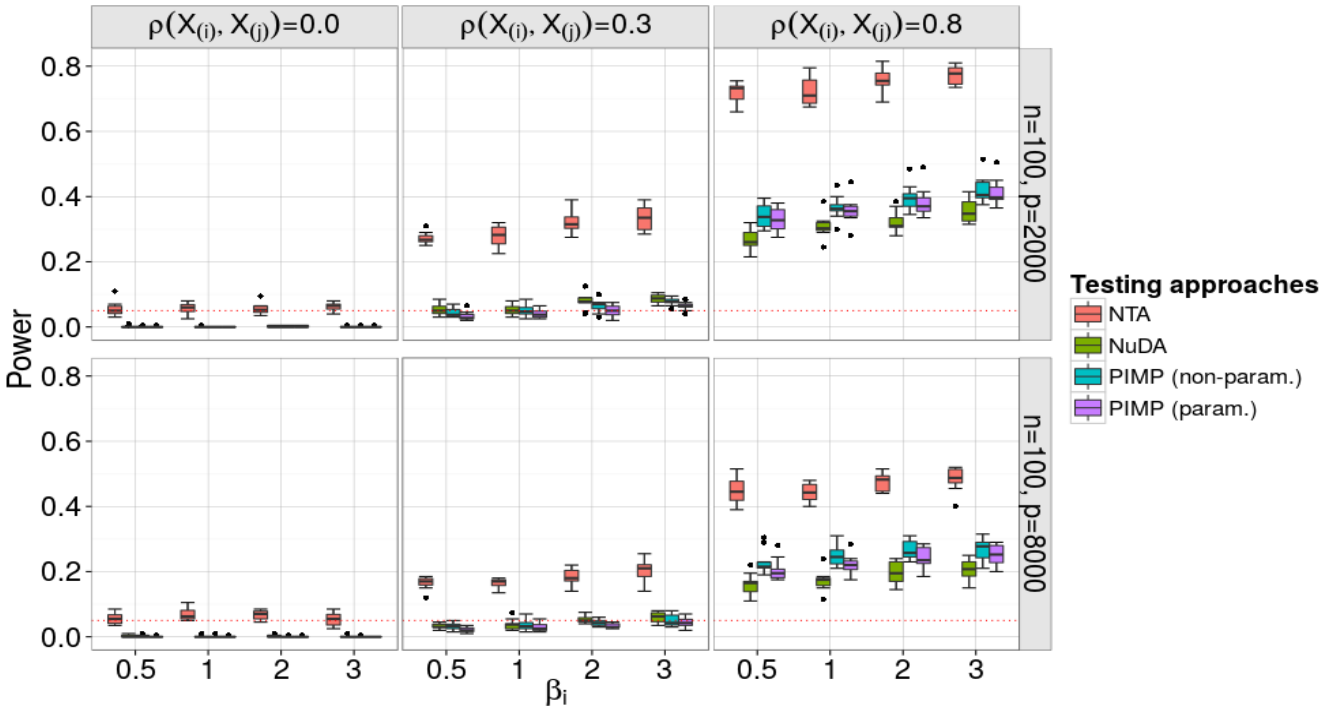

Abbildung 5.11: Boxplot der Power der Testansätze in Abhängigkeit der Regressionskoeffizienten $\beta_j \in \{0.5, 1, 2, 3\}$ im Fall 2B mit der Parametereinstellung $m_{try} = \frac{p}{3}$ in beiden Datensätzen : Oben für den Datensatz mit $n = 100$, $p = 2000$ und unten mit $n = 100$, $p = 8000$. Links für jede Kovariable $\mathbf{x}_{(j)} \in \mathbf{X}_{p_3} \wedge |\beta_j| \neq 0$ mit $\rho(X_{(j)}, X_{(k)}) = 0.0$, $j, k = p_2+1, \ldots, p_3$, $j \neq k$, in der Mitte für jede Kovariable $\mathbf{x}_{(j)} \in \mathbf{X}_{p_1} \wedge |\beta_j| \neq 0$ mit $\rho(X_{(j)}, X_{(k)}) = 0.3$, $j, k = 1, \ldots, p_1$, $j \neq k$ und rechts für jede Kovariable $\mathbf{x}_{(j)} \in \mathbf{X}_{p_2} \wedge |\beta_j| \neq 0$ mit $\rho(X_{(j)}, X_{(k)}) = 0.8$, $j, k = (p_1 + 1), \ldots, p_2$, $j \neq k$. Das vorgegebene Signifikanzniveau $\alpha = 0.05$ ist als rot gepunktete Linie eingezeichnet.

Mit diesen Parametereinstellungen für die Regressionskoeffizienten $(3, 2, 1, 0.5, 0.5, 1, 2, 3)$ steigt in beiden Datensätzen die Power aller vier Testansätze mit zunehmender Korrelation zwischen den Kovariablen an. In der Kovariablengruppe $\mathbf{X}_{p_3} \wedge \beta_j \neq 0$, die Kovariablen sind unabhängig identisch verteilt ($\rho(X_{(j)}, X_{(k)}) = 0.0$, $j, k = p_2 + 1, \ldots, p_3$ und $j \neq k$), ist die mediane Power der Testansätze $NuDA$ und $PIMP$ (nicht-parametrisch und parametrisch) für alle Regressionskoeffizienten praktisch gleich Null. Die mediane Power des NTA Testansatzes ist etwas größer als das vorgegebene Signifikanzniveau $\alpha = 0.05$ (vgl. Tabelle 5.12 und Tabelle 5.13). Die Testansätze, $NuDA$ und $PIMP$ (nicht-parametrisch und parametrisch) entdecken nicht einmal zufällig eine von der Zielvariable abhängige Kovariable in dieser Kovariablengruppe.

Median (n=100,p=2000)	$\rho(X_{(j)}, X_{(k)}) = 0.0$				$\rho(X_{(j)}, X_{(k)}) = 0.3$				$\rho(X_{(j)}, X_{(k)}) = 0.8$			
$\lvert\beta_j\rvert$	0.5	1	2	3	0.5	1	2	3	0.5	1	2	3
NTA	0.05	0.06	0.0525	0.065	0.2675	0.2825	0.3150	0.3350	0.732	0.710	0.755	0.778
NuDA	0.00	0.00	0.0000	0.000	0.0500	0.0500	0.0775	0.0875	0.260	0.302	0.310	0.347
PIMP (non-param.)	0.00	0.00	0.0000	0.000	0.0375	0.0475	0.0700	0.0775	0.338	0.362	0.395	0.405
PIMP (param.)	0.00	0.00	0.0000	0.000	0.0275	0.0375	0.0500	0.0650	0.328	0.355	0.370	0.398

Tabelle 5.12: Median der Power im Fall 2B der heuristischen Testansätze mit der Parametereinstellung $m_{try} = \frac{p}{3}$, aufgeteilt in die drei Kovariablengruppen mit den unterschiedlichen Korrelationen zwischen den Kovariablen und in die absoluten Regressionskoeffizienten $\beta_j \in \{0.5, 1, 2, 3\}$ ($n = 100$, $p = 2000$).

Für die Kovariablen, die untereinander eine Korrelation von $\rho(X_{(j)}, X_{(k)}) = 0.3$, $j, k = 1, \ldots, p_1$,und $j \neq k$ aufweisen, ist nur die mediane Power des NTA Testansatzes deutlich größer als das vorgegebene Signifikanzniveau, während die anderen Testansätze, $NuDA$ und $PIMP$ (nicht-parametrisch und parametrisch), sehr nah am Signifikanzniveau liegen.
Im zweiten Drittel der Kovariablen $\mathbf{X}_{p_2}$, die Kovariablen sind untereinander korreliert mit 0.8, ist die mediane Power aller vier Testansätze deutlich größer als das vorgegebene Signifikanzniveau. Die Power des NTA Testansatzes ist deutlich größer als die der anderen Testansätze. In dieser Kovariablengruppe ist die Power beider $PIMP$ Testansätze (nicht-parametrisch und parametrisch) sogar größer als die des $NuDA$ Testansatzes.

Median (n=100,p=8000)	$\rho(X_{(j)}, X_{(k)}) = 0.0$				$\rho(X_{(j)}, X_{(k)}) = 0.3$				$\rho(X_{(j)}, X_{(k)}) = 0.8$			
$\lvert\beta_j\rvert$	0.5	1	2	3	0.5	1	2	3	0.5	1	2	3
NTA	0.055	0.0625	0.07	0.055	0.1700	0.1700	0.1800	0.2100	0.445	0.4425	0.4825	0.4875
NuDA	0.000	0.0000	0.00	0.000	0.0300	0.0375	0.0500	0.0625	0.165	0.1750	0.1950	0.2075
PIMP (non-param.)	0.000	0.0000	0.00	0.000	0.0325	0.0325	0.0425	0.0425	0.215	0.2450	0.2575	0.2775
PIMP (param.)	0.000	0.0000	0.00	0.000	0.0200	0.0250	0.0300	0.0425	0.195	0.2200	0.2350	0.2525

Tabelle 5.13: Median der Power im Fall 2B der heuristischen Testansätze mit der Parametereinstellung $m_{try} = \frac{p}{3}$, aufgeteilt in die drei Kovariablengruppen mit den unterschiedlichen Korrelationen zwischen den Kovariablen und in die absoluten Regressionskoeffizienten $\beta_j \in \{0.5, 1, 2, 3\}$ ($n = 100$, $p = 8000$).

Die Ergebnisse mit den Parametereinstellungen $m_{try} = \frac{p}{5}$ und $m_{try} = \frac{p}{10}$ sind im Anhang zu finden (vgl. Anhang A.2.1 auf Seite 74). Es fällt auf, dass die Power der Testansätze in den Kovariablengruppen mit einer Korrelation zwischen den Kovariablen mit einem kleinerem m_{try} größer wird. Dieser Effekt ist beim NTA Testansatz am stärksten ausgeprägt (vgl. Abbildung 5.12).

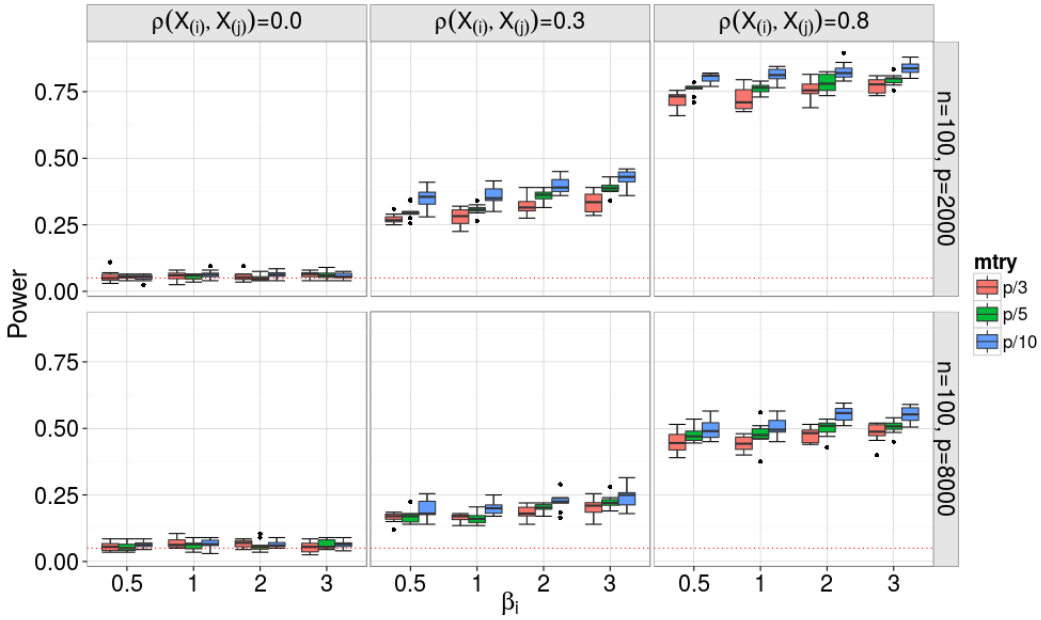

Abbildung 5.12: Boxplot der Power des NTA Testansatzes in Abhängigkeit der Regressionskoeffizienten $\beta_j \in \{0.5, 1, 2, 3\}$ im Fall 2B mit der Parametereinstellung $m_{try} = \frac{p}{3}$,$m_{try} = \frac{p}{5}$ und $m_{try} = \frac{p}{10}$ in beiden Datensätzen : Oben für den Datensatz mit $n = 100$, $p = 2000$ und unten für $n = 100$, $p = 8000$. Links für jede Kovariable $\mathbf{x}_{(j)} \in \mathbf{X}_{p_3} \wedge |\beta_j| \neq 0$ mit $\rho(X_{(j)}, X_{(k)}) = 0.0$, $j, k = p_2 + 1, \ldots, p_3$, $j \neq k$, in der Mitte für jede Kovariable $\mathbf{x}_{(j)} \in \mathbf{X}_{p_1} \wedge |\beta_j| \neq 0$ mit $\rho(X_{(j)}, X_{(k)}) = 0.3$, $j, k = 1, \ldots, p_1$, $j \neq k$ und rechts für jede Kovariable $\mathbf{x}_{(j)} \in \mathbf{X}_{p_2} \wedge |\beta_j| \neq 0$ mit $\rho(X_{(j)}, X_{(k)}) = 0.8$, $j, k = (p_1 + 1), \ldots, p_2$, $j \neq k$. Das vorgegebene Signifikanzniveau $\alpha = 0.05$ ist als rot gepunktete Linie eingezeichnet.

Die Abbildung 5.9 bildet die Power der vier heuristischen Testansätze im Fall 2B für die Kovariablen aus den Kovariablengruppen $\mathbf{X}_{p_1}, \mathbf{X}_{p_2}$ und mit einen Regressionskoeffizienten von $\beta_j = 0$ im Datensatz mit 2000 Kovariablen und 100 Beobachtungen als Boxplot ab. Die Ergebnisse für den Datensatz mit 8000 Kovariablen unterscheiden sich kaum von den Ergebnissen mit 2000 Kovariablen, weshalb sie im Anhang zu finden sind (vgl. Anhang A.2.1).

Median (n=100,p=2000)	$\rho(X_{(j)}, X_{(k)}) = 0.3$			$\rho(X_{(j)}, X_{(k)}) = 0.8$		
m_{try}	$p/3$	$p/5$	$p/10$	$p/3$	$p/5$	$p/10$
NTA	0.260	0.290	0.325	0.715	0.755	0.805
NuDA	0.045	0.040	0.035	0.275	0.290	0.315
PIMP (non-param.)	0.035	0.035	0.030	0.345	0.355	0.365
PIMP (param.)	0.030	0.030	0.025	0.335	0.345	0.365

Tabelle 5.14: Median der Power im Fall 2B in der *zweiten* Studie der heuristischen Testansätze für die Kovariablen aus den Kovariablengruppen $\mathbf{X}_{p_1}, \mathbf{X}_{p_2}$ und einen Regressionskoeffizienten von $\beta_j = 0$ ($n = 100$, $p = 2000$).

Für die Kovariablen aus der Kovariablengruppe $\mathbf{X}_{p_1}$ mit einen Regressionskoeffizienten von $\beta_j = 0$ und mit einer Korrelation zwischen den Kovariablen von 0.3, ist die Power

der heuristischen Testansätze $NuDA$ und $PIMP$ (nicht-parametrisch und parametrisch) kleiner als das vorgegebene Signifikanzniveau $\alpha = 0.05$. Nur die Power des NTA Testansatzes ist deutlich größer als das vorgegebene Signifikanzniveau. Die mediane Power aller vier Testansätze ist deutlich größer als das vorgegebene Signifikanzniveau $\alpha = 0.05$ für die Kovariablen aus der Kovariablengruppe $\mathbf{X}_{p_2}$ mit einen Regressionskoeffizienten von $\beta_j = 0$ und mit einer Korrelation zwischen den Kovariablen von 0.8. Der NTA Testansatz hat mit Abstand die größte Power in beiden Gruppen.
Die Annahme, dass die symmetrische Anordnung der Regressionskoeffizienten (-3, -2, -1, -0.5, 0.5, 1, 2, 3) in den Kovariablengruppen den Korrelationseffekt vermindert bzw. eliminiert, ist mit dieser Simulation bestätigt. Es fällt aber auf, dass die Power aller vier Testansätze in der Kovariablengruppe $\mathbf{X}_{p_3} \wedge \beta_j \neq 0$ mit keiner Korrelation zwischen den Kovariablen unter bzw. sehr nah am Signifikanzniveau ist.

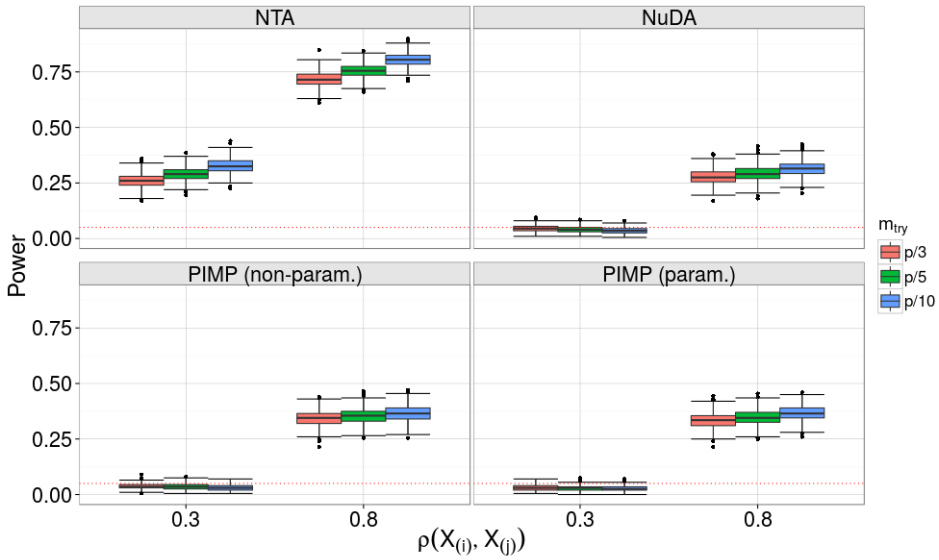

Abbildung 5.13: Boxplot der Power im Fall 2B für die Kovariablen aus den Kovariablengruppen $\mathbf{X}_{p_1}, \mathbf{X}_{p_2} \wedge\ \beta_j = 0$ der Testansätze im Datensatz mit $n = 100$, $p = 2000$: Auf der x-Achse sind die zwei Kovariablengruppen $\mathbf{X}_{p_1}, \mathbf{X}_{p_2}$ dargestellt, wobei 0.3 auf das erste Drittel der Kovariablen $\mathbf{X}_{p_1}$ mit $\rho(X_{(j)}, X_{(k)}) = 0.3$, $j, k = 1, \ldots, p_1$, $j \neq k$ und 0.8 auf zweite Drittel der Kovariablen $\mathbf{X}_{p_2}$ mit $\rho(X_{(j)}, X_{(k)}) = 0.8$, $j, k = (p_1 + 1), \ldots, p_2$, $j \neq k$ verweist. Oben links der NTA Testansatz, oben rechts der $NuDA$ Testansatz und unten rechts und links der $PIMP$ Testansatz (nicht-parametrisch und parametrisch). Das vorgegebene Signifikanzniveau $\alpha = 0.05$ ist als rot gepunktete Linie eingezeichnet.

Fehler 1. Art im Fall 2

In der Kovariablengruppe $\mathbf{X}_{p_3}$ sind die Kovariablen unabhängig identisch verteilt ($\rho(X_{(j)}, X_{(k)}) = 0.0$, $j, k = p_2 + 1, \ldots, p_3$ und $j \neq k$) und mit einem Regressionskoeffizienten von $\beta_j = 0$ sind diese Kovariablen auch unabhängig von der Zielvariable. In der Abbildung 5.14 sind die Wahrscheinlichkeiten für den Fehler 1. Art der heuristischen Testansätze für die Kovariablen $\mathbf{X}_{p_3} \wedge \beta_j = 0$ im zweiten Fall der *zweiten* Studie als Boxplot dargestellt.

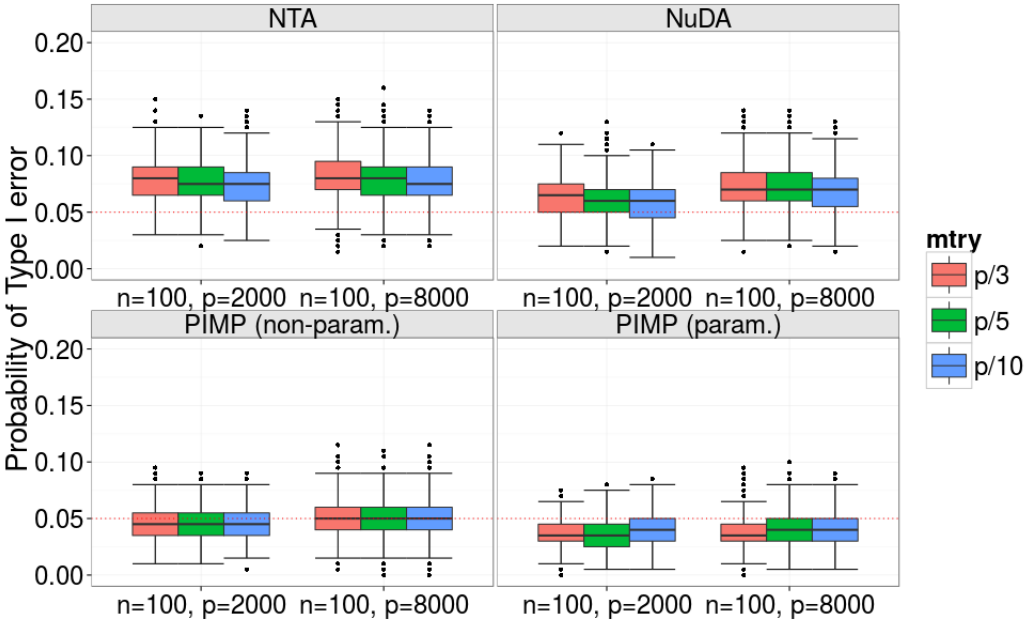

Abbildung 5.14: Boxplot der Wahrscheinlichkeiten für den Fehler 1. Art der Testansätze für die Kovariablen in der Kovariablengruppe $\mathbf{X}_{p_3}$ mit Regressionskoeffizienten $\beta_j = 0$ im *zweiten* Fall der *zweiten* Studie in beiden Datensätzen ($n = 100$, $p = 2000$ und $n = 100$, $p = 8000$): Oben links der NTA Testansatz, oben rechts $NuDA$ Testansatz und unten recht und links der $PIMP$ Testansatz (nicht-parametrisch und parametrisch). Das vorgegebene Signifikanzniveau $\alpha = 0.05$ ist als rot gepunktete Linie eingezeichnet.

Aus dieser Abbildung und der Tabelle 5.15 kann man erkennen, dass in beiden Datensätzen die medianen Wahrscheinlichkeiten für den Fehler 1. Art der heuristischen Testansätze NTA und $NuDA$ größer als das vorgegebene Signifikanzniveau $\alpha = 0.05$ sind. Dies bedeutet, es werden systematisch zu viele Kovariablen als signifikant abhängig von der Zielvariable erkannt. Die mediane Wahrscheinlichkeit für den Fehler 1. Art des $PIMP$ (nicht-parametrisch) Testansatzes ist im Datensatz mit 2000 Kovariablen unwesentlich kleiner als das vorgegebene Signifikanzniveau $\alpha = 0.05$. Nur der Median der Wahrscheinlichkeiten für den Fehler 1. Art für den parametrischen $PIMP$ Testansatz ist, wie in der ersten Studie, deutlich kleiner als das vorgegebenen Signifikanzniveau. Da sich die Ergebnisse für den Datensatz mit 8000 Kovariablen kaum von den Ergebnissen mit 2000 Kovariablen unterscheiden, sind diese im Anhang zu finden (vgl. Anhang A.2.1).

	p/3		p/5		p/10	
n=100, p=2000	Mittelwert	Median	Mittelwert	Median	Mittelwert	Median
NTA	0.07926	0.080	0.07702	0.075	0.07462	0.075
NuDA	0.06392	0.065	0.06136	0.060	0.05920	0.060
PIMP (non-param.)	0.04712	0.045	0.04554	0.045	0.04706	0.045
PIMP (param.)	0.03658	0.035	0.03680	0.035	0.03997	0.040

Tabelle 5.15: Mittelwert und Median der Wahrscheinlichkeiten für den Fehler 1. Art für die Kovariablen in der Kovariablengruppe $\mathbf{X}_{p_3}$ mit Regressionskoeffizienten $\beta_j = 0$ im *zweiten* Fall in der *zweiten* Studie der heuristischen Testansätze ($n = 100$, $p = 2000$).

Fehler 1. Art im Fall 2B

Da die mediane Power der Testansätze $NuDA$ und $PIMP$ (nicht-parametrisch und parametrisch) in der Kovariablengruppe $\mathbf{X}_{p_3}$ für alle Regressionskoeffizienten praktisch gleich Null ist, verwundert es auch nicht, dass die Wahrscheinlichkeiten für den Fehler 1. Art dieser heuristischen Testansätze praktisch Null ist (vgl. Abbildung 5.15 und Tabelle 5.16)

	p/3		p/5		p/10	
n=100, p=2000	Mittelwert	Median	Mittelwert	Median	Mittelwert	Median
NTA	0.0583946	0.055	0.0569169	0.055	0.0562859	0.055
NuDA	0.0010064	0.000	0.0008466	0.000	0.0004073	0.000
PIMP (non-param.)	0.0006949	0.000	0.0005831	0.000	0.0003355	0.000
PIMP (param.)	0.0004313	0.000	0.0003674	0.000	0.0002636	0.000

Tabelle 5.16: Mittelwert und Median der Wahrscheinlichkeiten für den Fehler 1. Art für die Kovariablen in der Kovariablengruppe $\mathbf{X}_{p_3}$ mit Regressionskoeffizienten $\beta_j = 0$ im Fall 2B in der *zweiten* Studie der heuristischen Testansätze ($n = 100$, $p = 2000$).

Im Datensatz mit 2000 Kovariablen ist die mediane Wahrscheinlichkeit für den Fehler 1. Art des NTA Testansatzes unwesentlich größer als das vorgegebene Signifikanzniveau $\alpha = 0.05$. Während die mediane Wahrscheinlichkeit für den Fehler 1. Art des NTA Testansatzes im Datensatz mit 8000 Kovariablen einiges größer als das vorgegebene Signifikanzniveau $\alpha = 0.05$ ist.

	p/3		p/5		p/10	
n=100, p=8000	Mittelwert	Median	Mittelwert	Median	Mittelwert	Median
NTA	0.0583946	0.055	0.0569169	0.055	0.0562859	0.055
NuDA	0.0010064	0.000	0.0008466	0.000	0.0004073	0.000
PIMP (non-param.)	0.0006949	0.000	0.0005831	0.000	0.0003355	0.000
PIMP (param.)	0.0004313	0.000	0.0003674	0.000	0.0002636	0.000

Tabelle 5.17: Mittelwert und Median der Wahrscheinlichkeiten für den Fehler 1. Art für die Kovariablen in der Kovariablengruppe $\mathbf{X}_{p_3}$ mit Regressionskoeffizienten $\beta_j = 0$ im Fall 2B in der *zweiten* Studie der heuristischen Testansätze ($n = 100$, $p = 8000$).

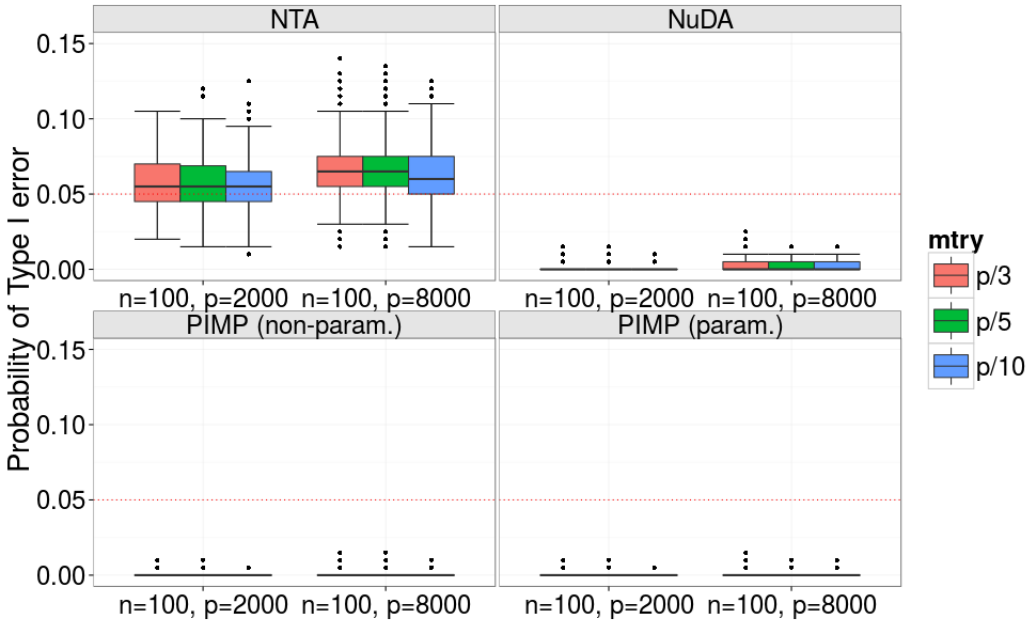

Abbildung 5.15: Boxplot der Wahrscheinlichkeiten für den Fehler 1. Art der Testansätze für die Kovariablen in der Kovariablengruppe $\mathbf{X}_{p_3}$ mit Regressionskoeffizienten $\beta_j = 0$ im Fall 2B in der *zweiten* Studie in beiden Datensätzen ($n = 100, p = 2000$ und $n = 100, p = 8000$): Oben links der NTA Testansatz, oben rechts $NuDA$ Testansatz und unten recht und links der $PIMP$ Testansatz (nicht-parametrisch und parametrisch). Das vorgegebene Signifikanzniveau $\alpha = 0.05$ ist als rot gepunktete Linie eingezeichnet.

6. Zusammenfassung und Ausblick

In dieser Arbeit wurde untersucht, ob die heuristischen Testansätze, NTA, $NuDA$,$PIMP$ (parametrisch und nicht-parametrisch), für hochdimensionale genetische Daten geeignet sind.

In der *ersten* Simulationsstudie wurde untersucht, ob die Testverfahren das vorgegebene Signifikanzniveau α einhalten und ob die *Permutation Variablenwichtigkeitsmaße* symmetrisch um Null verteilt sind. In der *zweiten* Simulationsstudie wurde analysiert, wie groß die *Power* dieser heuristischen Testverfahren ist und ob diese Testverfahren weiterhin das vorgegebene Signifikanzniveau α einhalten. Um diese Analysen durchführen zu können, musste im Vorhinein bekannt sein, welche Kovariable einen Effekt bzw. keinen Effekt auf die Zielvariable hat. Da bei realen Daten diese Information nicht vorliegt, wurden die Daten simuliert.

Die Daten in der *ersten* Simulationsstudie wurden entsprechend der Null-Hypothese, die Kovariablen und Zielvariable sind unabhängig voneinander, erzeugt. Bei der Generierung der Daten wurde bei den Kovariablen zwischen zwei Fällen unterschieden. Im *ersten* Fall waren die Kovariablen untereinander nicht korreliert. Im *zweiten* Fall waren die Daten untereinander korreliert (vgl. Abschnitt 5.2.1). In der *zweiten* Simulationsstudie wurden die Zielvariablen entsprechend dem linearen Modell generiert. Um die *Power* und die Wahrscheinlichkeiten für den Fehler 1. Art der Testansätze gleichzeitig untersuchen zu können, wurden die Daten so erzeugt, dass manche Kovariablen abhängig und manche unabhängig von der Zielvariable waren. Wie bereits in der ersten Studie, wurde bei den Kovariablen zwischen zwei Fällen, untereinander korreliert und nicht korreliert, unterschieden (vgl. Abschnitt 5.4.1).

Für beide Simulationsstudien und Fälle wurden 200 Datensätze (D) mit 100 Beobachtungen (n), 2000 Kovariablen (p) und 200 Datensätze mit 100 Beobachtungen, 8000 Kovariablen erzeugt. Für jeden Datensatz wurde ein *Random Forest,* bestehend aus 1000 Regressionsbäumen, generiert. Für alle p Kovariablen wurde das *Permutation Variablenwichtigkeitsmaß* und das *hold-out-Variablenwichtigkeitsmaß* berechnet. Anschließend wurde für jede Kovariable der p-Wert mit den verschiedenen Testansätzen, NTA, $NuDA$,$PIMP$ (parametrisch und nicht-parametrisch), berechnet. Diese Berechnungen wurden mit drei verschiedenen Einstellungen des Parameters m_{try} ($\frac{p}{3}$, $\frac{p}{5}$, $\frac{p}{10}$) durchgeführt (vgl. Abschnitt 8).

In beiden Simulationsstudien wurde für jeden Testansatz die Wahrscheinlichkeit für den Fehler 1. Art durch die relative Häufigkeit der Fälle, in denen die Null-Hypothese fälschlicherweise verworfen wurde, geschätzt. In der *zweiten* Simulationsstudie wurde die Power der Testansätze durch die relative Häufigkeit der Fälle, in denen die Null-Hypothese zu

Recht verworfen wurde, geschätzt. Die Power und die Wahrscheinlichkeiten für den Fehler 1. Art der verschiedenen heuristischen Testansätze, wurden mithilfe von Boxplots für verschiedene Parametereinstellungen miteinander verglichen und ggf. durch zusammenfassende Tabellen (Mittelwert, Median) ergänzt.
Im Kapitel 5.3 wurden die Ergebnisse der *ersten* Simulationsstudie beschrieben. Es konnte gezeigt werden, dass das *hold-out-Variablenwichtigkeitsmaß* in beiden Fällen, die Kovariablen sind unabhängig identisch verteilt und die Kovariablen sind nicht unabhängig voneinander, symmetrisch um Null verteilt ist. Im Gegensatz dazu, war das *Permutation Variablenwichtigkeitsmaß* (PerVIM) in beiden Fällen nicht exakt symmetrisch um Null verteilt. Des Weiteren war der Median der Wahrscheinlichkeiten für den Fehler 1. Art in beiden Fällen der bei allen drei heuristischen Testansätze, NTA, $NuDA$ und $PIMP$ (nicht-parametrisch) gleich bzw. unwesentlich größer dem vorgegebenen Signifikanzniveau $\alpha = 0.05$. Nur der Median der Wahrscheinlichkeiten für den Fehler 1. Art beim parametrischen $PIMP$ Testansatz war deutlich kleiner als das vorgegebenen Signifikanzniveau. Der Grund dafür war, dass die angenommene Normalverteilung für die „Null-Verteilungen“ nicht zutreffend war.
Die Ergebnisse der *zweiten* Simulationsstudie wurden im Kapitel 5.5 beschrieben. Im ersten Fall, die Kovariablen sind unabhängig identisch verteilt, war zu beobachten, dass die Power aller vier Testansätze mit der zunehmender absoluten Größe der Regressionskoeffizienten anstieg. Die mediane Power des NTA Testansatzes war etwas größer als der anderen Testansätze. Während die mediane Power der Testansätze, $NuDA$ und $PIMP$ (nicht-parametrisch) fast gleich war. Der parametrische $PIMP$ Testansatz hatte durchwegs die kleinste Power, was auf die nicht zutreffende Normalverteilung für die „Null-Verteilungen“ zurückzuführen ist. Im *ersten* Fall in der *zweiten* Simulationsstudie war die Wahrscheinlichkeiten für den Fehler 1. Art der heuristischen Testansätze nicht auffallend anders als wie die in der *ersten* Studie.
Im zweiten Fall der *zweiten* Simulationsstudie war zu beobachten, dass die Power aller vier Testansätze wie bereits *ersten* Fall, mit der zunehmender absoluten Größe der Regressionskoeffizienten angestiegen ist. Auch war zu sehen, dass die Streuung der Power aller vier Testansätze viel größer war. Es wurde beobachtet, dass die Power aller vier Testansätze mit zunehmender Korrelation zwischen den Kovariablen sinkte. Man hätte erwartet, dass die Power der heuristischen Testansätze mit zunehmender Korrelation zwischen Kovariablen ansteigen würde. Die Gründe dafür waren: in der „kleinen“ Simulationsstudie in Abschnitt 4.6 auf Seite 21 wurde beobachtet, dass die beobachteten „Null-Verteilungen“ der *Variablenwichtigkeitsmaße* mit steigender Korrelation zwischen den Kovariablen nicht mehr um Null verteilt sind. Als weiteren Grund sprach dafür, dass in der Studie von Strobl u. a. (2008) beobachtet wurde, dass das *Permutation Variablenwichtigkeitsmaß* für mit-

einander korrelierte Kovariablen größer ist als für untereinander unabhängige Kovariablen (vgl. Strobl u. a., 2008, S. 5ff). Es wurde angenommen, dass in den Kovariablengruppen die symmetrische Anordnung der Regressionskoeffizienten $(-3, -2, -1, -0.5, 0.5, 1, 2, 3)$ diesen beschriebenen Effekt umgedreht haben. Um diese Annahme zu bestätigen, wurde noch einmal eine Simulation in der *zweiten* Studie im *zweiten* Fall mit neuen Parametereinstellungen für die Regressionskoeffizienten $(3, 2, 1, 0.5, 0.5, 1, 2, 3)$ durchgeführt. Mit diesen Parametereinstellungen für die Regressionskoeffizienten $(3, 2, 1, 0.5, 0.5, 1, 2, 3)$ war zu beobachten, dass die Power aller vier Testansätze mit der zunehmender Korrelation zwischen den Kovariablen anstieg. Die getroffene Annahme war bestätigt. In diesem Fall 2B war auffallend, dass die Power der Testansätze $NuDA$ und $PIMP$ (nicht-parametrisch und parametrisch) in der Kovariablengruppe, mit keiner Korrelation zwischen Kovariablen, für alle Regressionskoeffizienten praktisch gleich null war. Die mediane Power des NTA Testansatzes war in allen Kovariablengruppen deutlich größer als die der anderen Testansätze. Der NTA Testansatz war auch der einzige Testansatz, der relevante Kovariablen die keinen "eigenen" Einfluss auf die Zielvariable haben, mit einer deutlich größeren Power als das vorgegebene Signifikanzniveau entdecken konnte. Im *zweiten* Fall und im Fall 2B war die mediane Wahrscheinlichkeit für den Fehler 1. Art des NTA Testansatzes größer als das vorgegebene Signifikanzniveau. Die mediane Wahrscheinlichkeit für den Fehler 1. Art der Testansätze $NuDA$ und $PIMP$ (nicht-parametrisch und parametrisch) war im Fall 2B sehr nah an Null.

Die Power aller Testansätze war allgemein sehr niedrig. Im *zweiten* Fall und im Fall 2B war sie in den Kovariablengruppen sehr unterschiedlich groß. Von diesen vier Testansätzen lieferte der nicht-permutationsbasierte NTA Testansatz von Janitza u. Boulesteix (2015) die besten Ergebnisse.

Es stellt sich die Frage, ob die datengenerierenden Prozesse auch ähnliche Daten, wie die beobachteten reellen hochdimensionalen genomischen Daten, in der „realen Welt“ erzeugen. Bei Janitza u. Boulesteix (2015) heißt es:“There is common consensus in the literature that it very difficult – if not impossible – to simulate realistic complex data structures which capture all the patterns and sources of variability that are generated by a real biological system.“ Darum sollten die Ergebnisse dieser Simulationsstudien nicht verallgemeinert werden. Eine Simulationsstudie wie bei Janitza u. Boulesteix (2015), welche reelle genomische Daten verwenden, wäre noch empfehlenswert.

Es wäre auch möglich, ein anderes Maß für die Variablenwichtigkeit in den vier Testansätzen zu verwenden, wie zum Beispiel das *bedingte Permutation Variablenwichtigkeitsmaß* von Strobl u. a. (2008). Das *bedingte Permutation Variablenwichtigkeitsmaß* ist für miteinander korrelierte Kovariablen, die nicht viel größer sind als die untereinander unabhängige Kovariablen (vgl. Strobl u. a., 2008). Die Testansätze könnten evaentuell mit

dem *bedingten Permutation Variablenwichtigkeitsmaß,* Kovariablen, die abhängig von der Zielvariable sind, in allen Kovariablengruppen mit gleicher Power entdecken.

Abbildungsverzeichnis

Tabellenverzeichnis

Literatur

[Altmann u. a. 2010] Altmann, André ; Toloşi, Laura ; Sander, Oliver ; Lengauer, Thomas: Permutation importance: a corrected feature importance measure. In: *Bioinformatics* 26 (2010), Nr. 10, S. 1340–1347

[Breiman u. a. 1984] Breiman, L. ; Friedman, J.H. ; Olshen, R. A. ; Stone, C. J.: *Classification and Regression Trees*. Belmont, California : Wadsworth, 1984

[Breiman 1996] Breiman, Leo: Bagging Predictors. In: *Machine Learning* 24 (1996), August, Nr. 2, S. 123–140. – ISSN 0885–6125

[Breiman 2001] Breiman, Leo: Random Forests. In: *Machine Learning* 45 (2001), Oktober, Nr. 1, S. 5–32. – ISSN 0885–6125

[Breiman u. Cutler 2003] Breiman, Leo ; Cutler, Adele: Manual: Setting up, using, and understanding Random Forests v4.0 / University of California, Berkeley. Version: 2003. `https://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf`. 2003. – Forschungsbericht

[Clarke u. a. 2009] Clarke, B. ; Fokoue, E. ; Zhang, H.H.: *Principles and Theory for Data Mining and Machine Learning*. Springer New York, 2009 (Springer Series in Statistics). – ISBN 9780387981352

[Cutler u. a. 2012] Cutler, Adele ; Cutler, D. R. ; Stevens, John R.: Random Forests. In: *Ensemble Machine Learning: Methods and Applications*, Springer New York, 2012. – ISBN 9781441993267, S. 157 – 176

[Fahrmeir u. a. 1996] Fahrmeir, L. ; Brachinger, W. ; Hamerle, A. ; Tutz, G.: *Multivariate statistische Verfahren*. de Gruyter, 1996

[Hastie u. a. 2013] Hastie, T. ; Tibshirani, R. ; Friedman, J.: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer New York, 2013 (Springer Series in Statistics). – ISBN 9780387216065

[Janitza u. Boulesteix 2015] Janitza, Silke ; Boulesteix, Anne-Laure: *A computationally fast variable importance test for random forests for high-dimensional data*. June 2015

[Liaw u. Wiener 2002] Liaw, Andy ; Wiener, Matthew: Classification and Regression by randomForest. In: *R News* 2 (2002), Nr. 3, 18–22. `http://CRAN.R-project.org/doc/Rnews/`

[Louppe 2014] Louppe, G.: Understanding Random Forests: From Theory to Practice. In: *ArXiv e-prints* (2014), Juli

[McCallum u. Weston 2011] McCallum, E. ; Weston, S.: *Parallel R.* O'Reilly Media, 2011 (Data analysis in the distributed world). – ISBN 9781449309923

[Steiner 2009] Steiner, V.: *Modellierung des Kundenwertes: Ein branchenübergreifender Ansatz.* Gabler Verlag, 2009 (Gabler Edition Wissenschaft). – ISBN 9783834916266

[Strobl u. a. 2005] Strobl, Carolin ; Boulesteix, Anne-Laure ; Augustin, Thomas: *Unbiased split selection for classification trees based on the Gini Index.* `http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-1833-1`. Version: 2005 (sfb386)

[Strobl u. a. 2008] Strobl, Carolin ; Boulesteix, Anne-Laure ; Kneib, Thomas ; Augustin, Thomas ; Zeileis, Achim: Conditional variable importance for random forests. In: *BMC Bioinformatics* 9 (2008), Nr. 1, 307. `http://dx.doi.org/10.1186/1471-2105-9-307`. – DOI 10.1186/1471–2105–9–307. – ISSN 1471–2105

[Strobl u. a. 2007] Strobl, Carolin ; Boulesteix, Anne-Laure ; Zeileis, Achim ; Hothorn, Torsten: Bias in random forest variable importance measures: Illustrations, sources and a solution. In: *BMC Bioinformatics* 8 (2007), Nr. 1, 25. `http://dx.doi.org/10.1186/1471-2105-8-25`. – DOI 10.1186/1471–2105–8–25. – ISSN 1471–2105

A. Anhang

A.1. Grafiken und Tabellen zu Studie I

Im Anhang A.1 befinden sich die ergänzenden Abbildungen und Tabellen zu Kapitel 5.3(Beschreibung der Ergebnisse in der Studie I).

A.1.1. „Null-Verteilung“

Die beobachteten „Null-Verteilungen“ der *Variablenwichtigkeitsmaße* für die Parametereinstellungen $m_{try} = \frac{p}{5}$ und $m_{try} = \frac{p}{10}$.

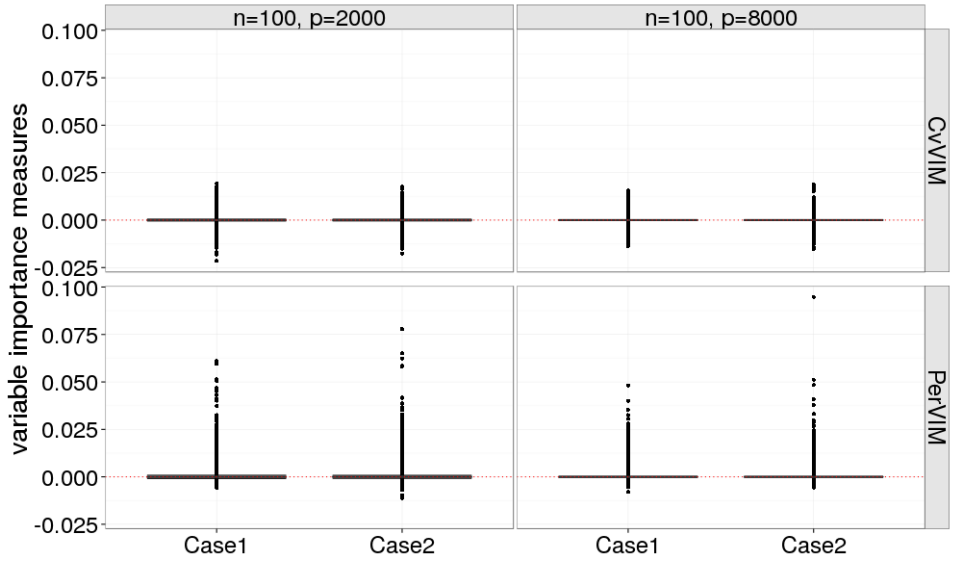

Abbildung A.1: Boxplot der beobachteten „Null-Verteilungen“ der *Variablenwichtigkeitsmaße* mit der Parametereinstellung $m_{try} = \frac{p}{5}$ für die verschiedenen Fälle: Oben sind die beobachteten „Null-Verteilungen“ für die *hold-out- Variablenwichtigkeitsmaße* (CvVIM) und unten die *Permutation Variablenwichtigkeitsmaße* (PerVIM) dargestellt. Links sind die beobachteten „Null-Verteilungen“ für 200 Datensätze mit 100 Beobachtungen (n) , 2000 Kovariablen (p) und rechts mit 100 Beobachtungen, 8000 Kovariablen dargestellt. Die Abszissenachse ist bei $y = 0$ als rot gepunktete Linie eingezeichnet.

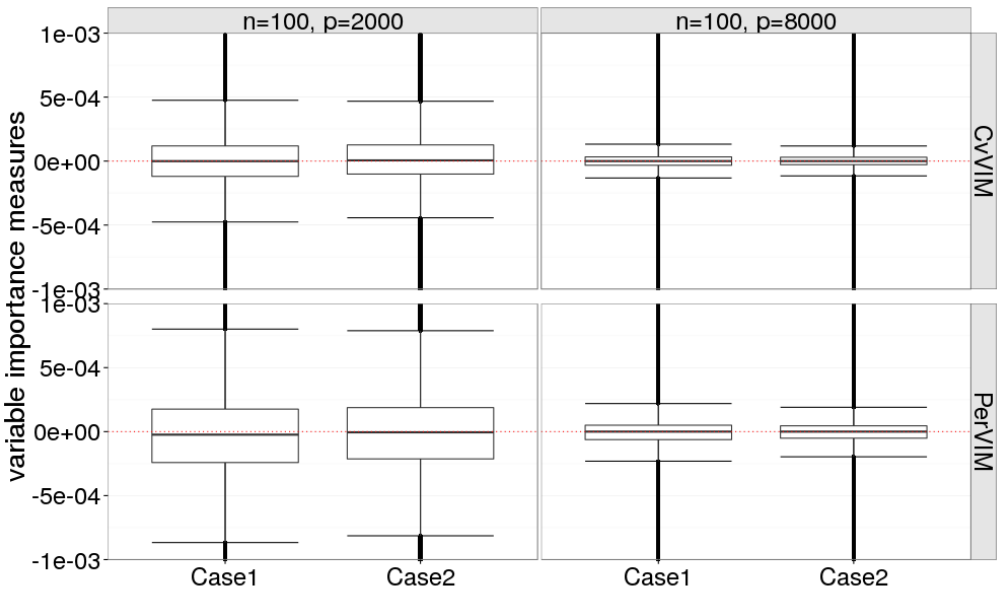

Abbildung A.2: Boxplot der beobachteten „Null-Verteilungen“ der *Variablenwichtigkeitsmaße* für den gezoomten Wertebereich [-0.001 bis 0.001] der y-Achse mit der Parametereinstellung $m_{try} = \frac{p}{5}$ für die verschiedenen Fälle: Oben sind die beobachteten „Null-Verteilungen“ für die *hold-out- Variablenwichtigkeitsmaße* (CvVIM) und unten die *Permutation Variablenwichtigkeitsmaße* (PerVIM) dargestellt. Links sind die beobachteten „Null-Verteilungen“ für 200 Datensätze mit 100 Beobachtungen (n) , 2000 Kovariablen (p) und rechts mit 100 Beobachtungen, 8000 Kovariablen dargestellt. Die Abszissenachse ist bei $y = 0$ als rot gepunktete Linie eingezeichnet.

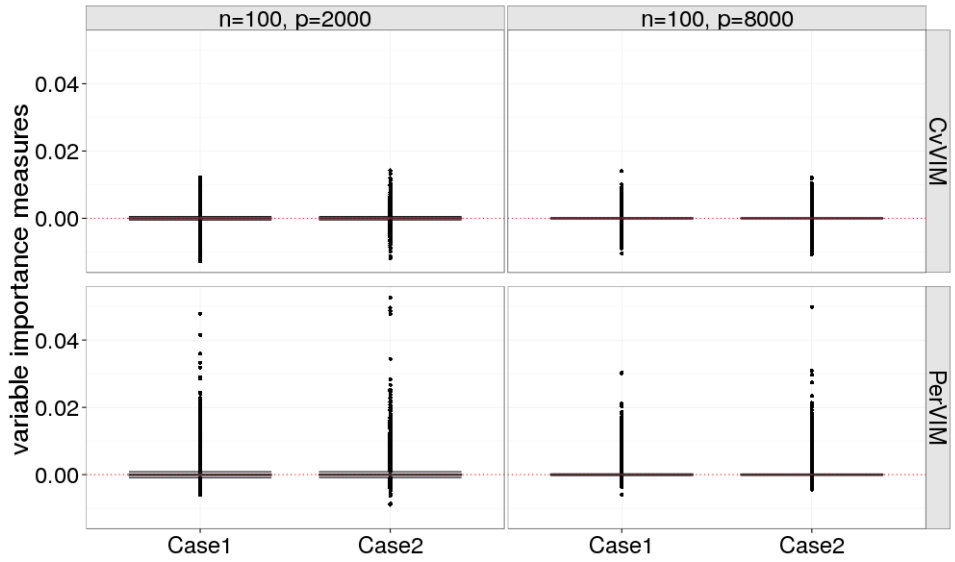

Abbildung A.3: Boxplot der beobachteten „Null-Verteilungen“ der *Variablenwichtigkeitsmaße* mit der Parametereinstellung $m_{try} = \frac{p}{10}$ für die verschiedenen Fälle: Oben sind die beobachteten „Null-Verteilungen“ für die *hold-out- Variablenwichtigkeitsmaße* (CvVIM) und unten die *Permutation Variablenwichtigkeitsmaße* (PerVIM) dargestellt. Links sind die beobachteten „Null-Verteilungen“ für 200 Datensätze mit 100 Beobachtungen (n) , 2000 Kovariablen (p) und rechts mit 100 Beobachtungen, 8000 Kovariablen dargestellt. Die Abszissenachse ist bei $y = 0$ als rote gepunktete Linie eingezeichnet.

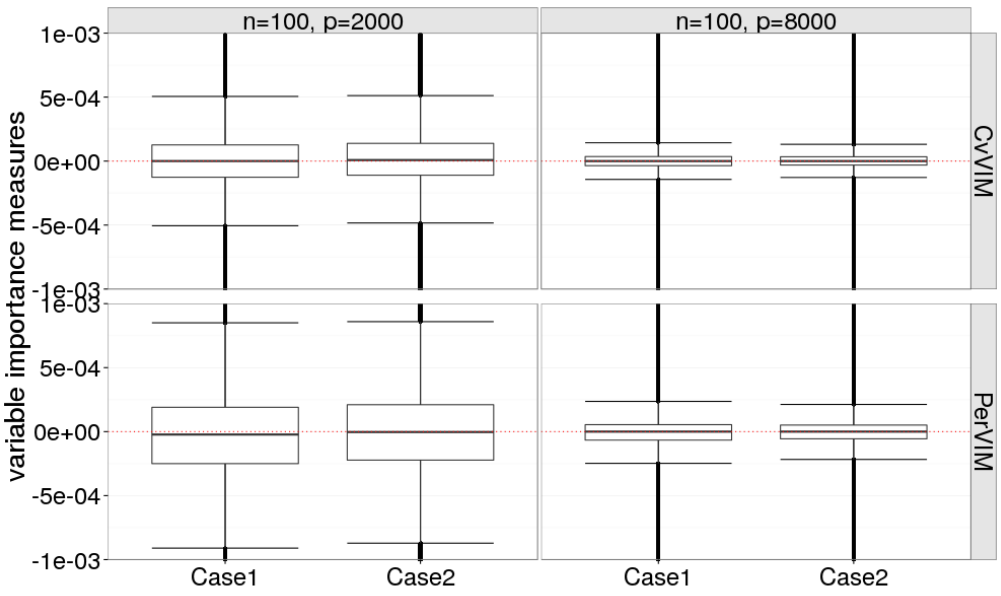

Abbildung A.4: Boxplot der beobachteten „Null-Verteilungen“ der *Variablenwichtigkeitsmaße* für den gezoomten Wertebereich [-0.001 bis 0.001] der y-Achse mit der Parametereinstellung $m_{try} = \frac{p}{10}$ für die verschiedenen Fälle: Oben sind die beobachteten „Null-Verteilungen“ für die *hold-out-Variablenwichtigkeitsmaße* (CvVIM) und unten die *Permutation Variablenwichtigkeitsmaße* (PerVIM) dargestellt. Links sind die beobachteten „Null-Verteilungen“ für 200 Datensätze mit 100 Beobachtungen (n) , 2000 Kovariablen (p) und rechts mit 100 Beobachtungen, 8000 Kovariablen dargestellt. Die Abszissenachse ist bei $y = 0$ als rot gepunktete Linie eingezeichnet..

A.1.2. Der Fehler 1.Art

Fall 1

Tabelle der mittleren und medianen Wahrscheinlichkeiten für den Fehler 1.Art $\alpha(\mathbf{x_{(j)}})$ für den Datensatz mit 8000 Kovariablen und 100 Beobachtungen im *ersten* Fall in der *ersten* Studie der heuristischen Testansätze:

	p/3		p/5		p/10	
n=100, p=8000	Mittelwert	Median	Mittelwert	Median	Mittelwert	Median
NTA	0.04989	0.05	0.04992	0.05	0.05015	0.05
NuDA	0.05010	0.05	0.04995	0.05	0.04967	0.05
PIMP (nicht-param.)	0.04963	0.05	0.04939	0.05	0.04963	0.05
PIMP (param.)	0.03831	0.04	0.03909	0.04	0.03831	0.04

Tabelle A.1: Mittelwert und Median der $p = 8000$ Wahrscheinlichkeiten für den Fehler 1.Art im *ersten* Fall der heuristischen Testansätze. ($n = 100$)

Fall 2

Tabelle der mittleren und medianen Wahrscheinlichkeiten für den Fehler 1.Art $\alpha(\mathbf{x_{(j)}})$ für den Datensatz mit 8000 Kovariablen und 100 Beobachtungen im *zweiten* Fall in der

ersten Studie der heuristischen Testansätze:

	p/3		p/5		p/10	
n=100, p=8000	Mittelwert	Median	Mittelwert	Median	Mittelwert	Median
NTA	0.05371	0.055	0.05420	0.055	0.05409	0.055
NuDA	0.04963	0.050	0.04970	0.050	0.04932	0.050
PIMP (nicht-param.)	0.04874	0.050	0.04883	0.050	0.04874	0.050
PIMP (param.)	0.03825	0.035	0.03958	0.040	0.03825	0.035

Tabelle A.2: Mittelwert und Median der $p = 8000$ Wahrscheinlichkeiten für den Fehler 1.Art im *zweiten* Fall der heuristischen Testansätze. ($n = 100$)

A.2. Grafiken und Tabellen zu Studie II

Im Anhang A.2 befinden sich die ergänzenden Abbildungen und Tabellen zu Kapitel 5.5(Beschreibung der Ergebnisse in der Studie II).

A.2.1. Power und der Fehler 1. Art

Fall 1

Tabelle für die mittlere und mediane Power für den Datensatz mit 2000 Kovariablen und 100 Beobachtungen im *ersten* Fall der heuristischen Testansätze mit der Parametereinstellung $m_{try} = \frac{p}{5}$ und $m_{try} = \frac{p}{10}$:

	$\|\beta_j\| = 0.5$		$\|\beta_j\| = 1$		$\|\beta_j\| = 2$		$\|\beta_j\| = 3$	
n=100, p=2000	Mittelwert	Median	Mittelwert	Median	Mittelwert	Median	Mittelwert	Median
NTA	0.05533	0.0550	0.07083	0.065	0.10517	0.1025	0.1750	0.1725
NuDA	0.05150	0.0500	0.06333	0.060	0.10017	0.1025	0.1578	0.1575
PIMP (nicht-param.)	0.05100	0.0500	0.06450	0.065	0.09767	0.0975	0.1555	0.1525
PIMP (param.)	0.03867	0.0375	0.05417	0.050	0.08333	0.0750	0.1353	0.1375

Tabelle A.3: Mittelwert und Median der Power im *ersten* Fall der heuristischen Testansätze mit der Parametereinstellung $m_{try} = \frac{p}{5}$, aufgeteilt in die absoluten Regressionskoeffizienten $|\beta_j| \in \{0.5, 1, 2, 3\}$. ($n = 100$, $p = 2000$).

	$\|\beta_j\| = 0.5$		$\|\beta_j\| = 1$		$\|\beta_j\| = 2$		$\|\beta_j\| = 3$	
n=100, p=2000	Mittelwert	Median	Mittelwert	Median	Mittelwert	Median	Mittelwert	Median
NTA	0.05367	0.050	0.06967	0.0750	0.10767	0.1075	0.1803	0.1850
NuDA	0.05483	0.055	0.06550	0.0625	0.09450	0.0925	0.1620	0.1575
PIMP (nicht-param.)	0.05367	0.055	0.06383	0.0650	0.08967	0.0900	0.1555	0.1525
PIMP (param.)	0.04517	0.045	0.05333	0.0500	0.08150	0.0800	0.1425	0.1425

Tabelle A.4: Mittelwert und Median der Power im *ersten* Fall der heuristischen Testansätze mit der Parametereinstellung $m_{try} = \frac{p}{10}$, aufgeteilt in die absoluten Regressionskoeffizienten $|\beta_j| \in \{0.5, 1, 2, 3\}$. ($n = 100$, $p = 2000$).

Tabelle für die mittlere und mediane Power für den Datensatz mit 8000 Kovariablen und 100 Beobachtungen im *ersten* Fall der heuristischen Testansätze mit der Parametereinstellung $m_{try} = \frac{p}{5}$ und $m_{try} = \frac{p}{10}$:

	$\lvert\beta_j\rvert = 0.5$		$\lvert\beta_j\rvert = 1$		$\lvert\beta_j\rvert = 2$		$\lvert\beta_j\rvert = 3$	
n=100, p=8000	Mittelwert	Median	Mittelwert	Median	Mittelwert	Median	Mittelwert	Median
NTA	0.04917	0.0500	0.06100	0.0550	0.08633	0.085	0.1363	0.1325
NuDA	0.05133	0.0500	0.05817	0.0600	0.08350	0.080	0.1390	0.1350
PIMP (nicht-param.)	0.05033	0.0500	0.05433	0.0550	0.08217	0.075	0.1342	0.1350
PIMP (param.)	0.04100	0.0425	0.04583	0.0475	0.06933	0.070	0.1197	0.1200

Tabelle A.5: Mittelwert und Median der Power im *ersten* Fall der heuristischen Testansätze mit der Parametereinstellung $m_{try} = \frac{p}{5}$, aufgeteilt in die absoluten Regressionskoeffizienten $|\beta_j| \in \{0.5, 1, 2, 3\}$. ($n = 100$, $p = 8000$).

	$\lvert\beta_j\rvert = 0.5$		$\lvert\beta_j\rvert = 1$		$\lvert\beta_j\rvert = 2$		$\lvert\beta_j\rvert = 3$	
n=100, p=2000	Mittelwert	Median	Mittelwert	Median	Mittelwert	Median	Mittelwert	Median
NTA	0.05533	0.0525	0.06483	0.0650	0.09333	0.0900	0.1458	0.1500
NuDA	0.04800	0.0500	0.05983	0.0625	0.08700	0.0850	0.1288	0.1200
PIMP (nicht-param.)	0.04517	0.0500	0.06100	0.0600	0.08533	0.0850	0.1255	0.1200
PIMP (param.)	0.03933	0.0400	0.04967	0.0475	0.07700	0.0725	0.1132	0.1125

Tabelle A.6: Mittelwert und Median der Power im *ersten* Fall der heuristischen Testansätze mit der Parametereinstellung $m_{try} = \frac{p}{10}$, aufgeteilt in die absoluten Regressionskoeffizienten $|\beta_j| \in \{0.5, 1, 2, 3\}$. ($n = 100$, $p = 8000$).

Tabelle der mittleren und medianen Wahrscheinlichkeiten für den Fehler 1.Art $\alpha(\mathbf{x}_{(\mathbf{j})})$ für den Datensatz mit 8000 Kovariablen und 100 Beobachtungen im *ersten* Fall in der *zweiten* Studie der heuristischen Testansätze:

	p/3		p/5		p/10	
n=100, p=8000	Mittelwert	Median	Mittelwert	Median	Mittelwert	Median
NTA	0.04995	0.050	0.05016	0.05	0.04982	0.05
NuDA	0.04936	0.050	0.04982	0.05	0.04976	0.05
PIMP (nicht-param.)	0.04905	0.050	0.04927	0.05	0.04920	0.05
PIMP (param.)	0.03763	0.035	0.03895	0.04	0.04034	0.04

Tabelle A.7: Mittelwert und Median der Wahrscheinlichkeiten für den Fehler 1. Art der Kovariablen mit Regressionskoeffizienten $\beta_j = 0$ im *ersten* Fall in der *zweiten* Studie der heuristischen Testansätze. ($n = 100$, $p = 8000$).

Fall 2

Boxplot der Power der Testansätze für die drei Kovariablengruppen in Abhängigkeit der absoluten Regressionskoeffizienten $|\beta_j| \in \{0.5, 1, 2, 3\}$ im *zweiten* Fall mit der Parametereinstellung $m_{try} = \frac{p}{5}$:

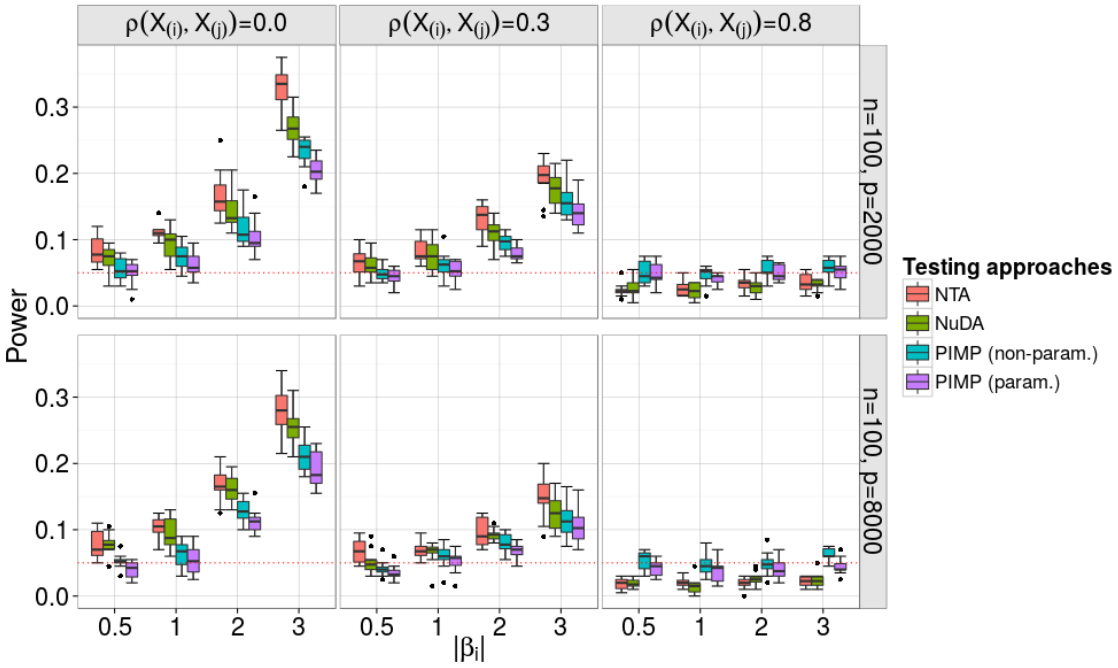

Abbildung A.5: Boxplot der Power der Testansätze in Abhängigkeit der absoluten Regressionskoeffizienten $|\beta_j| \in \{0.5, 1, 2, 3\}$ im *zweiten* Fall mit der Parametereinstellung $m_{try} = \frac{p}{5}$ in beiden Datensätzen : Oben für den Datensatz mit $n = 100, p = 2000$ und unten mit $n = 100, p = 8000$. Links für jede Kovariable $\mathbf{x}_{(j)} \in \mathbf{X}_{p_3} \wedge |\beta_j| \neq 0$ mit $\rho(X_{(j)}, X_{(k)}) = 0.0$, $j, k = p_2 + 1, \ldots, p_3,\ j \neq k$, in der Mitte für jede Kovariable $\mathbf{x}_{(j)} \in \mathbf{X}_{p_1} \wedge |\beta_j| \neq 0$ mit $\rho(X_{(j)}, X_{(k)}) = 0.3$, $j, k = 1, \ldots, p_1,\ j \neq k$ und rechts für jede Kovariable $\mathbf{x}_{(j)} \in \mathbf{X}_{p_2} \wedge |\beta_j| \neq 0$ mit $\rho(X_{(j)}, X_{(k)}) = 0.8,\ j, k = (p_1 + 1), \ldots, p_2,\ j \neq k$. Das vorgegebene Signifikanzniveau $\alpha = 0.05$ ist als rot gepunktete Linie eingezeichnet.

Tabelle für die mediane Power im Datensatz mit 2000 Kovariablen und 8000 Kovariablen und im *zweiten* Fall der heuristischen Testansätze mit der Parametereinstellung $m_{try} = \frac{p}{5}$:

Median (n=100,p=2000)	$\rho(X_{(j)}, X_{(k)}) = 0.0$				$\rho(X_{(j)}, X_{(k)}) = 0.3$				$\rho(X_{(j)}, X_{(k)}) = 0.8$			
$\lvert\beta_j\rvert$	0.5	1	2	3	0.5	1	2	3	0.5	1	2	3
NTA	0.0775	0.1100	0.158	0.335	0.0675	0.0750	0.1375	0.198	0.0225	0.0250	0.035	0.0325
NuDA	0.0750	0.1000	0.133	0.268	0.0575	0.0750	0.1125	0.177	0.0225	0.0225	0.030	0.0325
PIMP (non-param.)	0.0525	0.0750	0.107	0.240	0.0475	0.0625	0.0975	0.155	0.0450	0.0525	0.050	0.0575
PIMP (param.)	0.0525	0.0575	0.095	0.203	0.0450	0.0525	0.0750	0.140	0.0425	0.0450	0.045	0.0550

Tabelle A.8: Median der Power im *zweiten* Fall der heuristischen Testansätze mit der Parametereinstellung $m_{try} = \frac{p}{5}$, aufgeteilt in die drei Kovariablengruppen mit den unterschiedlichen Korrelation zwischen den Kovariablen und in die absoluten Regressionskoeffizienten $|\beta_j| \in \{0.5, 1, 2, 3\}$ $(n = 100,\ p = 2000)$.

Median (n=100,p=8000)	$\rho(X_{(j)}, X_{(k)}) = 0.0$				$\rho(X_{(j)}, X_{(k)}) = 0.3$				$\rho(X_{(j)}, X_{(k)}) = 0.8$			
$\lvert\beta_j\rvert$	0.5	1	2	3	0.5	1	2	3	0.5	1	2	3
NTA	0.0700	0.1050	0.165	0.280	0.0675	0.0675	0.0900	0.148	0.0200	0.0200	0.0200	0.0225
NuDA	0.0775	0.0875	0.160	0.255	0.0475	0.0700	0.0925	0.125	0.0175	0.0150	0.0250	0.0225
PIMP (non-param.)	0.0525	0.0675	0.128	0.210	0.0400	0.0600	0.0775	0.112	0.0600	0.0450	0.0475	0.0600
PIMP (param.)	0.0425	0.0525	0.113	0.182	0.0325	0.0575	0.0700	0.103	0.0450	0.0425	0.0375	0.0400

Tabelle A.9: Median der Power im *zweiten* Fall der heuristischen Testansätze mit der Parametereinstellung $m_{try} = \frac{p}{5}$, aufgeteilt in die drei Kovariablengruppen mit den unterschiedlichen Korrelation zwischen den Kovariablen und in die absoluten Regressionskoeffizienten $|\beta_j| \in \{0.5, 1, 2, 3\}$ $(n = 100,\ p = 8000)$.

Boxplot der Power der Testansätze für die drei Kovariablengruppen in Abhängigkeit der absoluten Regressionskoeffizienten $|\beta_j| \in \{0.5, 1, 2, 3\}$ im *zweiten* Fall mit der Parametereinstellung $m_{try} = \frac{p}{10}$:

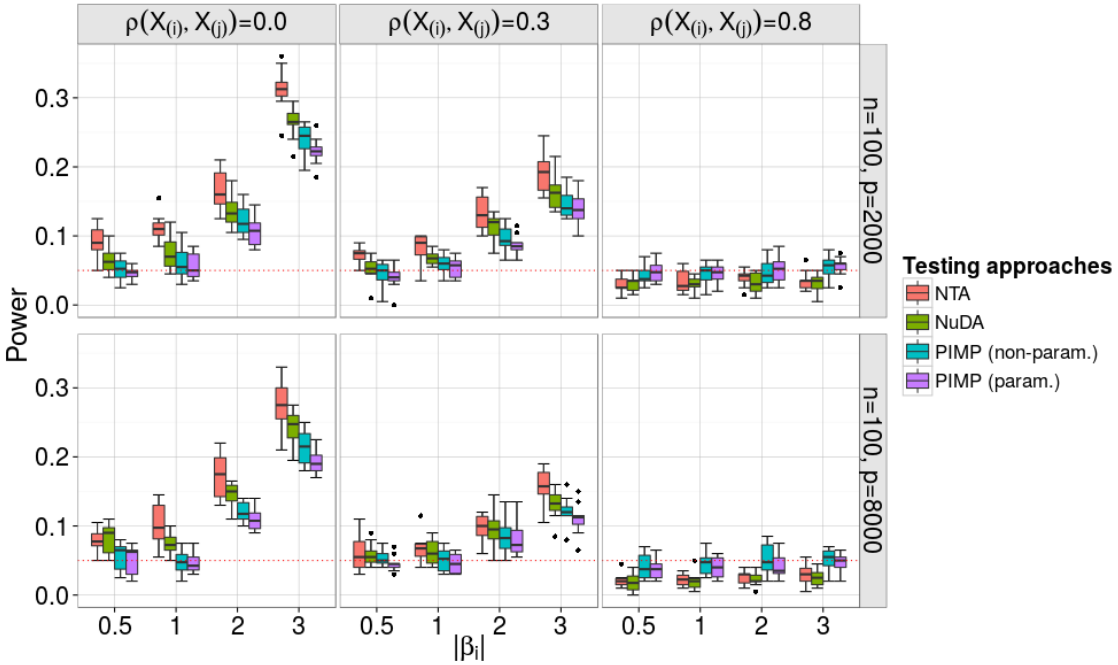

Abbildung A.6: Boxplot der Power der Testansätze in Abhängigkeit der absoluten Regressionskoeffizienten $|\beta_j| \in \{0.5, 1, 2, 3\}$ im *zweiten* Fall mit der Parametereinstellung $m_{try} = \frac{p}{10}$ in beiden Datensätzen : Oben für den Datensatz mit $n = 100, p = 2000$ und unten mit $n = 100, p = 8000$. Links für jede Kovariable $\mathbf{x}_{(j)} \in \mathbf{X}_{p_3} \wedge |\beta_j| \neq 0$ mit $\rho(X_{(j)}, X_{(k)}) = 0.0$, $j, k = p_2 + 1, \ldots, p_3,\ j \neq k$, in der Mitte für jede Kovariable $\mathbf{x}_{(j)} \in \mathbf{X}_{p_1} \wedge |\beta_j| \neq 0$ mit $\rho(X_{(j)}, X_{(k)}) = 0.3,\ j, k = 1, \ldots, p_1,\ j \neq k$ und rechts für jede Kovariable $\mathbf{x}_{(j)} \in \mathbf{X}_{p_2} \wedge |\beta_j| \neq 0$ mit $\rho(X_{(j)}, X_{(k)}) = 0.8,\ j, k = (p_1 + 1), \ldots, p_2,\ j \neq k$. Das vorgegebene Signifikanzniveau $\alpha = 0.05$ ist als rot gepunktete Linie eingezeichnet.

Tabelle für die mediane Power im Datensatz mit 2000 Kovariablen und 8000 Kovariablen und im *zweiten* Fall der heuristischen Testansätze mit der Parametereinstellung $m_{try} = \frac{p}{5}$:

Median (n=100,p=2000)	$\rho(X_{(j)}, X_{(k)}) = 0.0$				$\rho(X_{(j)}, X_{(k)}) = 0.3$				$\rho(X_{(j)}, X_{(k)}) = 0.8$			
$\lvert\beta_j\rvert$	0.5	1	2	3	0.5	1	2	3	0.5	1	2	3
NTA	0.0900	0.110	0.160	0.312	0.0750	0.0900	0.1300	0.193	0.0250	0.0275	0.0425	0.0350
NuDA	0.0625	0.070	0.133	0.265	0.0525	0.0675	0.1200	0.163	0.0350	0.0300	0.0300	0.0350
PIMP (non-param.)	0.0525	0.055	0.117	0.245	0.0500	0.0600	0.0925	0.140	0.0375	0.0500	0.0425	0.0575
PIMP (param.)	0.0475	0.050	0.108	0.223	0.0400	0.0575	0.0850	0.138	0.0475	0.0475	0.0525	0.0600

Tabelle A.10: Median der Power im *zweiten* Fall der heuristischen Testansätze mit der Parametereinstellung $m_{try} = \frac{p}{10}$, aufgeteilt in die drei Kovariablengruppen mit den unterschiedlichen Korrelation zwischen den Kovariablen und in die absoluten Regressionskoeffizienten $|\beta_j| \in \{0.5, 1, 2, 3\}$ $(n = 100,\ p = 2000)$.

Median (n=100,p=8000)	$\rho(X_{(j)}, X_{(k)}) = 0.0$				$\rho(X_{(j)}, X_{(k)}) = 0.3$				$\rho(X_{(j)}, X_{(k)}) = 0.8$			
$\lvert\beta_j\rvert$	0.5	1	2	3	0.5	1	2	3	0.5	1	2	3
NTA	0.0775	0.0975	0.175	0.275	0.055	0.0675	0.1000	0.158	0.0200	0.0225	0.0300	0.030
NuDA	0.0900	0.0725	0.150	0.247	0.055	0.0600	0.0950	0.133	0.0175	0.0200	0.0200	0.025
PIMP (non-param.)	0.0650	0.0475	0.117	0.215	0.050	0.0525	0.0825	0.120	0.0375	0.0475	0.0475	0.055
PIMP (param.)	0.0625	0.0425	0.108	0.190	0.045	0.0450	0.0725	0.113	0.0375	0.0400	0.0350	0.050

Tabelle A.11: Median der Power im *zweiten* Fall der heuristischen Testansätze mit der Parametereinstellung $m_{try} = \frac{p}{10}$, aufgeteilt in die drei Kovariablengruppen mit den unterschiedlichen Korrelation zwischen den Kovariablen und in die absoluten Regressionskoeffizienten $|\beta_j| \in \{0.5, 1, 2, 3\}$ $(n = 100,\ p = 8000)$.

Boxplot der Power der Testansätze für die Kovariablen aus den Kovariablengruppen $\mathbf{X}_{p_1}, \mathbf{X}_{p_2}$ und mit einen Regressionskoeffizienten von $\beta_j = 0$ im Datensatz mit 8000 Kovariablen und 100 Beobachtungen:

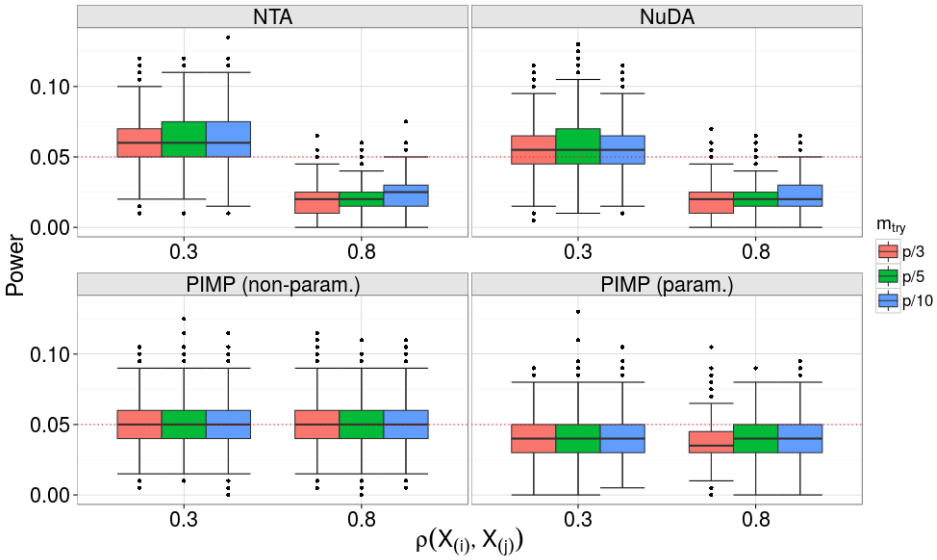

Abbildung A.7: Boxplot der Power für die Kovariablen aus den Kovariablengruppen $\mathbf{X}_{p_1}, \mathbf{X}_{p_2} \wedge$ $\beta_j = 0$ der Testansätze im Datensatz mit $n = 100$, $p = 8000$: Auf der x-Achse sind die zwei Kovariablengruppen $\mathbf{X}_{p_1}, \mathbf{X}_{p_2}$ dargestellt, wobei 0.3 auf das erste Drittel der Kovariablen $\mathbf{X}_{p_1}$ mit $\rho(X_{(j)}, X_{(k)}) = 0.3$, $j, k = 1, \ldots, p_1$, $j \neq k$ und 0.8 auf zweite Drittel der Kovariablen $\mathbf{X}_{p_2}$ mit $\rho(X_{(j)}, X_{(k)}) = 0.8$, $j, k = (p_1 + 1), \ldots, p_2$, $j \neq k$ verweist. Oben links der NTA Testansatz, oben rechts $NuDA$ Testansatz und unten recht und links der $PIMP$ Testansatz(nichtparametrisch und parametrisch). Das vorgegebene Signifikanzniveau $\alpha = 0.05$ ist als rot gepunktete Linie eingezeichnet.

Tabelle für die mediane Power im Datensatz mit 8000 Kovariablen und im *zweiten* Fall der heuristischen Testansätze für die Kovariablen aus den Kovariablengruppen $\mathbf{X}_{p_1}, \mathbf{X}_{p_2}$ und mit einen Regressionskoeffizienten von $\beta_j = 0$:

Median (n=100,p=8000)	$\rho(X_{(j)}, X_{(k)}) = 0.3$			$\rho(X_{(j)}, X_{(k)}) = 0.8$		
m_{try}	$p/3$	$p/5$	$p/10$	$p/3$	$p/5$	$p/10$
NTA	0.060	0.060	0.060	0.020	0.02	0.025
NuDA	0.055	0.055	0.055	0.020	0.02	0.020
PIMP (non-param.)	0.050	0.050	0.050	0.050	0.05	0.050
PIMP (param.)	0.040	0.040	0.040	0.035	0.04	0.040

Tabelle A.12: Median der Power im *zweiten* Fall in der *zweiten* Studie der heuristischen Testansätze mit den Kovariablen aus den Kovariablengruppen $\mathbf{X}_{p_1}, \mathbf{X}_{p_2}$ und einen Regressionskoeffizienten von $\beta_j = 0$ ($n = 100$, $p = 8000$).

Tabelle der mittleren und medianen Wahrscheinlichkeiten für den Fehler 1.Art $\alpha(\mathbf{x_{(j)}})$ im Datensatz mit 8000 Kovariablen und 100 Beobachtungen und im *ersten* Fall in der *zweiten* Studie der heuristischen Testansätze für die Kovariablen $\mathbf{X}_{p_3} \wedge \beta_j = 0$:

	p/3		p/5		p/10	
n=100, p=8000	Mittelwert	Median	Mittelwert	Median	Mittelwert	Median
NTA	0.08234	0.080	0.07954	0.08	0.07812	0.075
NuDA	0.07323	0.070	0.07192	0.07	0.06910	0.070
PIMP (non-param.)	0.04894	0.050	0.04868	0.05	0.04884	0.050
PIMP (param.)	0.03765	0.035	0.03879	0.04	0.04084	0.040

Tabelle A.13: Mittelwert und Median der Wahrscheinlichkeiten für den Fehler 1. Art für die Kovariablen in der Kovariablengruppe $\mathbf{X}_{p_3}$ mit Regressionskoeffizienten $\beta_j = 0$ im *zweiten* Fall in der *zweiten* Studie der heuristischen Testansätze.($n = 100$, $p = 8000$).

Fall 2B

Boxplot der Power der Testansätze für die drei Kovariablengruppen in Abhängigkeit der absoluten Regressionskoeffizienten $\beta_j \in \{0.5, 1, 2, 3\}$ im Fall 2B mit der Parametereinstellung $m_{try} = \frac{p}{5}$:

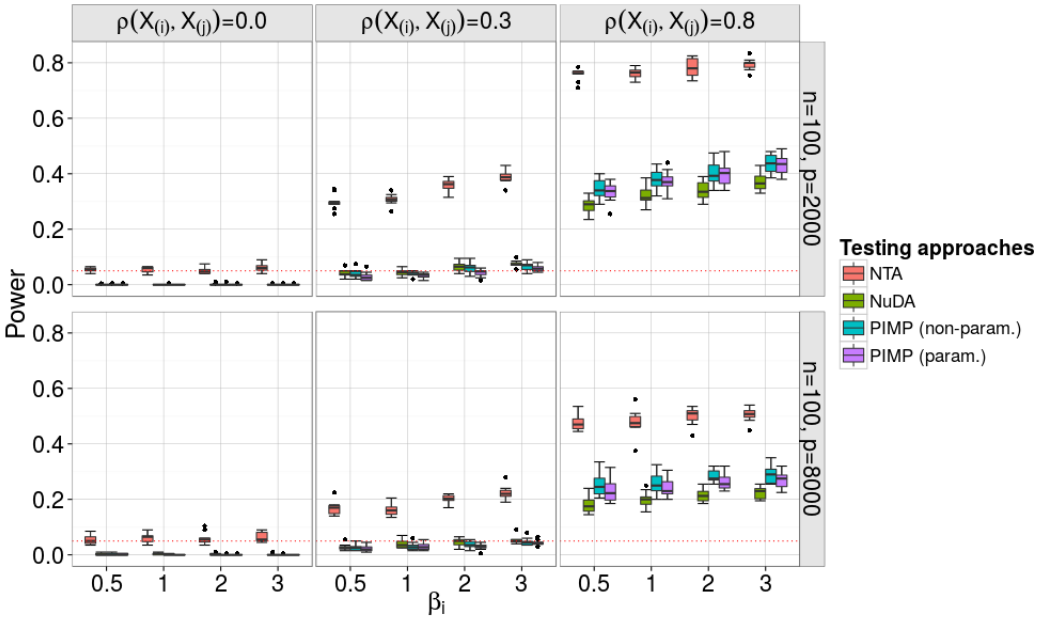

Abbildung A.8: Boxplot der Power der Testansätze in Abhängigkeit der Regressionskoeffizienten $\beta_j \in \{0.5, 1, 2, 3\}$ im Fall 2B mit der Parametereinstellung $m_{try} = \frac{p}{5}$ in beiden Datensätzen : Oben für den Datensatz mit $n = 100$, $p = 2000$ und unten mit $n = 100$, $p = 8000$. Links für jede Kovariable $\mathbf{x}_{(j)} \in \mathbf{X}_{p_3} \wedge |\beta_j| \neq 0$ mit $\rho(X_{(j)}, X_{(k)}) = 0.0$, $j, k = p_2+1, \ldots, p_3$, $j \neq k$, in der Mitte für jede Kovariable $\mathbf{x}_{(j)} \in \mathbf{X}_{p_1} \wedge |\beta_j| \neq 0$ mit $\rho(X_{(j)}, X_{(k)}) = 0.3$, $j, k = 1, \ldots, p_1$, $j \neq k$ und rechts für jede Kovariable $\mathbf{x}_{(j)} \in \mathbf{X}_{p_2} \wedge |\beta_j| \neq 0$ mit $\rho(X_{(j)}, X_{(k)}) = 0.8$, $j, k = (p_1 + 1), \ldots, p_2$, $j \neq k$. Das vorgegebene Signifikanzniveau $\alpha = 0.05$ ist als rot gepunktete Linie eingezeichnet.

Tabelle für die mediane Power im Datensatz mit 2000 Kovariablen und 8000 Kovariablen und im Fall 2B der heuristischen Testansätze mit der Parametereinstellung $m_{try} = \frac{p}{5}$:

Median (n=100,p=2000)	$\rho(X_{(j)}, X_{(k)}) = 0.0$				$\rho(X_{(j)}, X_{(k)}) = 0.3$				$\rho(X_{(j)}, X_{(k)}) = 0.8$			
$\lvert\beta_j\rvert$	0.5	1	2	3	0.5	1	2	3	0.5	1	2	3
NTA	0.055	0.06	0.0475	0.06	0.2950	0.3075	0.3625	0.3875	0.762	0.765	0.780	0.800
NuDA	0.000	0.00	0.0000	0.00	0.0375	0.0450	0.0650	0.0725	0.290	0.312	0.335	0.365
PIMP (non-param.)	0.000	0.00	0.0000	0.00	0.0350	0.0425	0.0625	0.0700	0.340	0.378	0.393	0.438
PIMP (param.)	0.000	0.00	0.0000	0.00	0.0250	0.0350	0.0475	0.0575	0.338	0.370	0.402	0.435

Tabelle A.14: Median der Power im Fall 2B der heuristischen Testansätze mit der Parametereinstellung $m_{try} = \frac{p}{5}$, aufgeteilt in die drei Kovariablengruppen mit den unterschiedlichen Korrelation zwischen den Kovariablen und in die absoluten Regressionskoeffizienten $\beta_j \in \{0.5, 1, 2, 3\}$ ($n = 100$, $p = 2000$).

Median (n=100,p=8000)	$\rho(X_{(j)}, X_{(k)}) = 0.0$				$\rho(X_{(j)}, X_{(k)}) = 0.3$				$\rho(X_{(j)}, X_{(k)}) = 0.8$			
$\lvert\beta_j\rvert$	0.5	1	2	3	0.5	1	2	3	0.5	1	2	3
NTA	0.05	0.065	0.055	0.055	0.170	0.1600	0.203	0.220	0.470	0.475	0.510	0.508
NuDA	0.00	0.005	0.000	0.000	0.025	0.0350	0.050	0.050	0.175	0.200	0.212	0.230
PIMP (non-param.)	0.00	0.000	0.000	0.000	0.025	0.0275	0.035	0.045	0.245	0.250	0.275	0.290
PIMP (param.)	0.00	0.000	0.000	0.000	0.020	0.0275	0.030	0.040	0.223	0.230	0.255	0.275

Tabelle A.15: Median der Power im Fall 2B der heuristischen Testansätze mit der Parametereinstellung $m_{try} = \frac{p}{5}$, aufgeteilt in die drei Kovariablengruppen mit den unterschiedlichen Korrelation zwischen den Kovariablen und in die absoluten Regressionskoeffizienten $\beta_j \in \{0.5, 1, 2, 3\}$ ($n = 100$, $p = 8000$).

Boxplot der Power der Testansätze für die drei Kovariablengruppen in Abhängigkeit der absoluten Regressionskoeffizienten $\beta_j \in \{0.5, 1, 2, 3\}$ im Fall 2B mit der Parametereinstellung $m_{try} = \frac{p}{10}$:

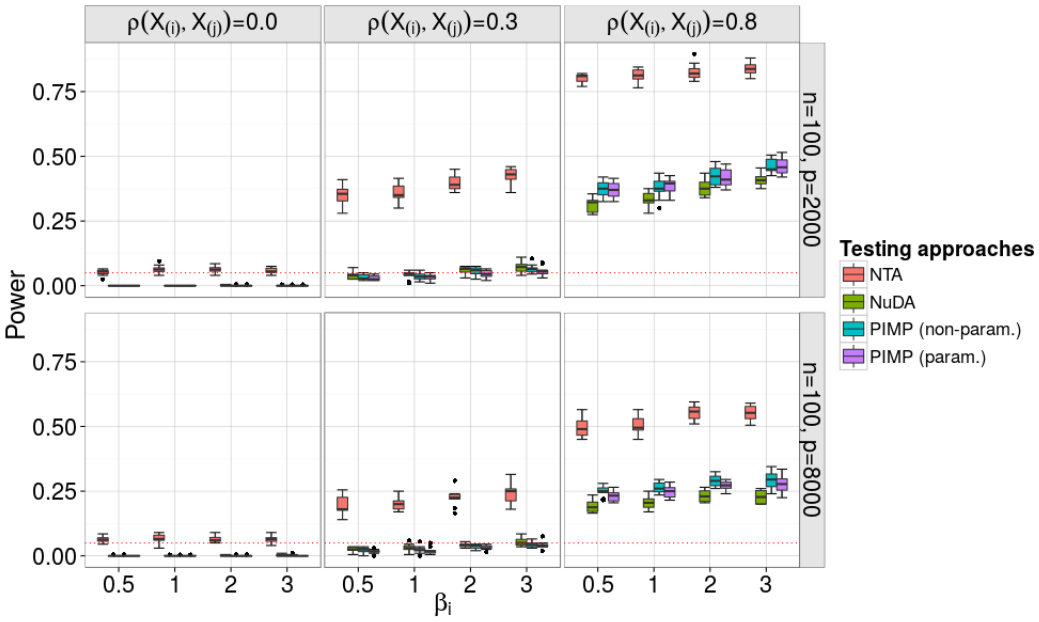

Abbildung A.9: Boxplot der Power der Testansätze in Abhängigkeit der Regressionskoeffizienten $\beta_j \in \{0.5, 1, 2, 3\}$ im Fall 2B mit der Parametereinstellung $m_{try} = \frac{p}{10}$ in beiden Datensätzen : Oben für den Datensatz mit $n = 100$, $p = 2000$ und unten mit $n = 100$, $p = 8000$. Links für jede Kovariable $\mathbf{x}_{(j)} \in \mathbf{X}_{p_3} \wedge |\beta_j| \neq 0$ mit $\rho(X_{(j)}, X_{(k)}) = 0.0$, $j, k = p_2+1, \ldots, p_3$, $j \neq k$, in der Mitte für jede Kovariable $\mathbf{x}_{(j)} \in \mathbf{X}_{p_1} \wedge |\beta_j| \neq 0$ mit $\rho(X_{(j)}, X_{(k)}) = 0.3$, $j, k = 1, \ldots, p_1$, $j \neq k$ und rechts für jede Kovariable $\mathbf{x}_{(j)} \in \mathbf{X}_{p_2} \wedge |\beta_j| \neq 0$ mit $\rho(X_{(j)}, X_{(k)}) = 0.8$, $j, k = (p_1 + 1), \ldots, p_2$, $j \neq k$. Das vorgegebene Signifikanzniveau $\alpha = 0.05$ ist als rot gepunktete Linie eingezeichnet.

Tabelle für die mediane Power im Datensatz mit 2000 Kovariablen und 8000 Kovariablen und im Fall 2B der heuristischen Testansätze mit der Parametereinstellung $m_{try} = \frac{p}{10}$:

Median (n=100,p=2000)	$\rho(X_{(j)}, X_{(k)}) = 0.0$				$\rho(X_{(j)}, X_{(k)}) = 0.3$				$\rho(X_{(j)}, X_{(k)}) = 0.8$			
$\lvert\beta_j\rvert$	0.5	1	2	3	0.5	1	2	3	0.5	1	2	3
NTA	0.055	0.0625	0.0625	0.055	0.3550	0.350	0.390	0.4300	0.810	0.812	0.820	0.837
NuDA	0.000	0.0000	0.0000	0.000	0.0400	0.045	0.065	0.0725	0.323	0.330	0.375	0.407
PIMP (non-param.)	0.000	0.0000	0.0000	0.000	0.0275	0.035	0.060	0.0650	0.375	0.375	0.422	0.450
PIMP (param.)	0.000	0.0000	0.0000	0.000	0.0250	0.035	0.045	0.0550	0.370	0.395	0.410	0.458

Tabelle A.16: Median der Power im Fall 2B der heuristischen Testansätze mit der Parametereinstellung $m_{try} = \frac{p}{10}$, aufgeteilt in die drei Kovariablengruppen mit den unterschiedlichen Korrelation zwischen den Kovariablen und in die absoluten Regressionskoeffizienten $\beta_j \in \{0.5, 1, 2, 3\}$ ($n = 100$, $p = 2000$).

Median (n=100,p=8000)	$\rho(X_{(j)}, X_{(k)}) = 0.0$				$\rho(X_{(j)}, X_{(k)}) = 0.3$				$\rho(X_{(j)}, X_{(k)}) = 0.8$			
$\lvert\beta_j\rvert$	0.5	1	2	3	0.5	1	2	3	0.5	1	2	3
NTA	0.0625	0.065	0.06	0.065	0.1800	0.2000	0.2250	0.2500	0.490	0.495	0.558	0.552
NuDA	0.0000	0.000	0.00	0.000	0.0250	0.0300	0.0425	0.0500	0.188	0.205	0.230	0.227
PIMP (non-param.)	0.0000	0.000	0.00	0.000	0.0250	0.0225	0.0400	0.0425	0.250	0.260	0.290	0.295
PIMP (param.)	0.0000	0.000	0.00	0.000	0.0175	0.0175	0.0300	0.0400	0.232	0.250	0.273	0.278

Tabelle A.17: Median der Power im Fall 2B der heuristischen Testansätze mit der Parametereinstellung $m_{try} = \frac{p}{10}$, aufgeteilt in die drei Kovariablengruppen mit den unterschiedlichen Korrelation zwischen den Kovariablen und in die absoluten Regressionskoeffizienten $\beta_j \in \{0.5, 1, 2, 3\}$ ($n = 100$, $p = 8000$).

Boxplot der Power der Testansätze im Fall 2B für die Kovariablen aus den Kovariablengruppen $\mathbf{X}_{p_1}, \mathbf{X}_{p_2}$ und mit einen Regressionskoeffizienten von $\beta_j = 0$ im Datensatz mit 8000 Kovariablen und 100 Beobachtungen:

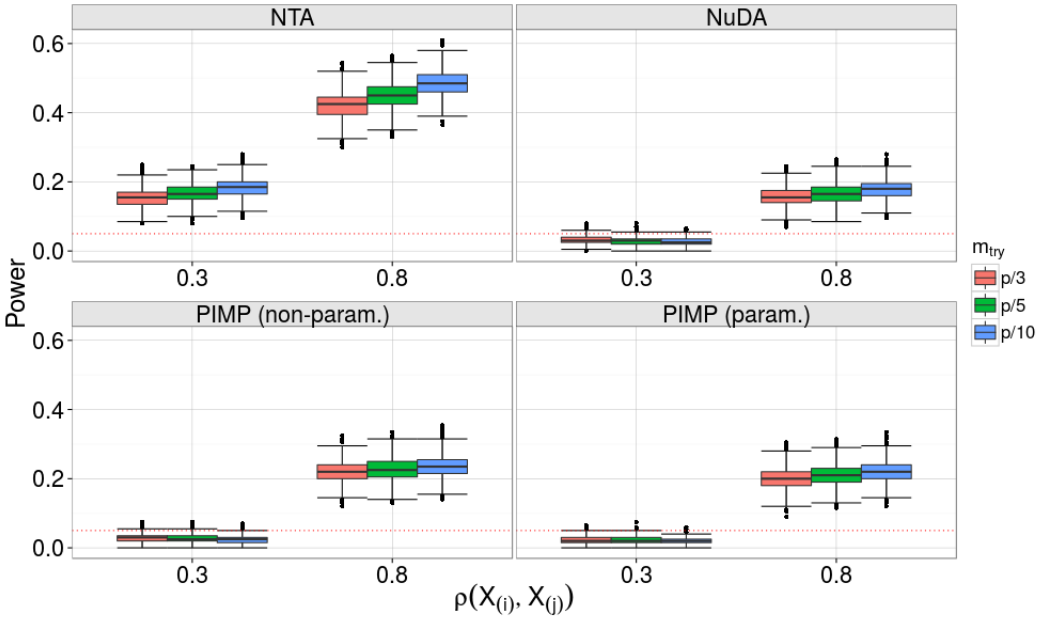

Abbildung A.10: Boxplot der Power im Fall 2B für die Kovariablen aus den Kovariablengruppen $\mathbf{X}_{p_1}, \mathbf{X}_{p_2} \wedge\ \beta_j = 0$ der Testansätze im Datensatz mit $n = 100$, $p = 8000$: Auf der x-Achse sind die zwei Kovariablengruppen $\mathbf{X}_{p_1}, \mathbf{X}_{p_2}$ dargestellt, wobei 0.3 auf das erste Drittel der Kovariablen $\mathbf{X}_{p_1}$ mit $\rho(X_{(j)}, X_{(k)}) = 0.3$, $j, k = 1, \ldots, p_1$, $j \neq k$ und 0.8 auf zweite Drittel der Kovariablen $\mathbf{X}_{p_2}$ mit $\rho(X_{(j)}, X_{(k)}) = 0.8$, $j, k = (p_1 + 1), \ldots, p_2$, $j \neq k$ verweist. Oben links der NTA Testansatz, oben rechts $NuDA$ Testansatz und unten recht und links der $PIMP$ Testansatz(nicht-parametrisch und parametrisch). Das vorgegebene Signifikanzniveau $\alpha = 0.05$ ist als rot gepunktete Linie eingezeichnet.

Tabelle für die mediane Power im Datensatz mit 8000 Kovariablen und im Fall 2B der heuristischen Testansätze für die Kovariablen aus den Kovariablengruppen $\mathbf{X}_{p_1}, \mathbf{X}_{p_2}$ und mit einen Regressionskoeffizienten von $\beta_j = 0$:

Median (n=100,p=8000)	$\rho(X_{(j)}, X_{(k)}) = 0.3$			$\rho(X_{(j)}, X_{(k)}) = 0.8$		
m_{try}	$p/3$	$p/5$	$p/10$	$p/3$	$p/5$	$p/10$
NTA	0.155	0.165	0.185	0.425	0.450	0.485
NuDA	0.030	0.030	0.025	0.155	0.165	0.180
PIMP (non-param.)	0.030	0.025	0.025	0.220	0.225	0.235
PIMP (param.)	0.020	0.020	0.020	0.200	0.210	0.220

Tabelle A.18: Median der Power im Fall 2B in der *zweiten* Studie der heuristischen Testansätze für die Kovariablen aus den Kovariablengruppen $\mathbf{X}_{p_1}, \mathbf{X}_{p_2}$ und einen Regressionskoeffizienten von $\beta_j = 0$ ($n = 100$, $p = 8000$).

B. Elektronischer Anhang

```
************************************************************************
************************************************************************
************************************************************************
**                      This readme file contains a brief             **
**                       description of the implemented               **
**                       R-scripts. For further comments              **
**                       and explantions please have a                **
**                       look at the particular R-file.               **
************************************************************************
************************************************************************

All calculations including the simulations have been done under Linux
(Debian 3.11.8-11 x86_64).


R Version of this Study: 3.1.0 (2014-04-10), nickname: Spring Dance


All used packages must be installed, with the command:
'install.packages("example", dependencies = TRUE)'

"example" is a placeholder for the name of the package.
The following packages are used:
** Name **                                                  **Version**
- rpart                                                      4.1.10
- rpart.plot                                                 1.5.2
- randomForest                                               4.6.10
- mnormt                                                     1.5.3
- parallel                                                   3.1.0
(parallel: mclapply -> implementation do not work under Windows)
- ggplot2                                                    1.0.0
- reshape                                                    0.8.5
- plyr                                                       1.8.1
- Hmisc                                                      3.16.0


************************************************************************
Package: vita
Type: Package
Title: Variable importance testing approaches
Version: 0.1

The package will be installed as:
```

```
install.packages("~/R/vita_0.1.tar.gz", repos = NULL, type = "source")

Implemented R functions:

compVarImp ..... Compute permutation variable importance measure
CVPVI ..... Cross-validated permutation variable
                importance measure
CVPVI.default ..... Cross-validated permutation variable
                importance measure
NTA ..... Novel testing approach
NTA.default ..... Novel testing approach
NuDA ..... Null distribution approximation
NuDA.default ..... Null distribution approximation
NuDaTest ..... NuDA testing approach
NuDaTest.default ..... NuDA testing approach
PIMP ..... PIMP-algorithm for the permutation variable
                importance measure
PIMP.default ..... PIMP-algorithm for the permutation variable
                importance measure
PimpTest ..... PIMP testing approach
PimpTest.default ..... PIMP testing approach
print.CVPVI ..... Cross-validated permutation variable
                importance measure
print.NTA ..... Novel testing approach
print.NuDA ..... Null distribution approximation
print.NuDaTest ..... NuDA testing approach
print.PIMP ..... PIMP-algorithm for the permutation variable
                importance measure
print.PimpTest ..... PIMP testing approach
print.summary.NTA ..... Summarizing the outcomes of novel testing
                approach
print.summary.NuDaTest ... Summarizing NuDA-algorithm outcomes
print.summary.PimpTest ... Summarizing PIMP-algorithm outcomes
summary.NTA ..... Summarizing the outcomes of novel testing
                approach
summary.NuDaTest ..... Summarizing NuDA-algorithm outcomes
summary.PimpTest ..... Summarizing PIMP-algorithm outcomes
VarImpCVI ..... Fold-specific permutation variable importance
                measure

All implemented R functions are described in detail in the Help files:

foo is a placeholder for the name of the R function
help.start()   # general help
help(foo)      # help about function foo
```

```
?foo           # same thing
************************************************************************


************************************************************************
*Be careful with the filepaths in the R-scripts, you have to
*adjust them yourself, otherwise the r-scripts wont run
************************************************************************

The structure of the "R" folder is corresponding to the structure of
the chapters of the report and "~" represents a "R-Script", "?"
represents a "txt file", "#" represents a "data file" and "+" represents
a "graphic file":


========================================================================
-> /R/Kap2                              [Kapitel 2]
========================================================================

   ~ plot_kap2.R ..................... Creation of figures;
                                       Abbildung 2.1, Abbildung 2.2
   + Ab2_1.png ....................... Decision Tree and
                                       dissections of the variable space
   + Ab2_2.png ....................... Step by step dissections of the
                                       variable space

========================================================================
-> /R/Kap3                              [Kapitel 3.2.1]
========================================================================

   ~ plot_kap3.R ..................... Creation of figure;
                                       Abbildung 3.1
   + Ab3_1.png ....................... Out-of-bag mean squared error

========================================================================
-> /R/Kap4                              [Kapitel 4.4]
========================================================================

   ~ sim_kap4.R ...................... Small simulation study;
                                       Abbildung 4.1
                                       Abbildung 4.2
                                       Abbildung 4.3
   + Ab_4p100.png .................... Boxplot: p=100, n=500
   + Ab_4p2000.png ................... Boxplot: p=2000, n=100
   + Ab_4p200k08.png ................. Boxplot: p=2000, n=100,
```

```
                                        COV(x_i,x_j)=0.8
   # sim4.RData ..... results of the simulations

================================================================================

-> /R/kap5                              [Kapitel 5]

================================================================================


--------------------------------------------------------------------------------

-> R/Kap5/Data/DataStudyI               [Kapitel 5.2.1]

--------------------------------------------------------------------------------

   ? Start.txt ..... commands for Batch Execution of
                                        r-scripts (DataStudyI.R)
   ~ DataStudyI.R ..... DGP: Study I and

                                        Case I:
                                        DATA 1 ... n = 100, p = 2000
                                        # DI.I.1.rds
                                        DATA 2 ... n = 100, p = 8000
                                        # DI.I.2.rds

                                        Case II:
                                        DATA 1 ... n = 100, p = 2000
                                        # DI.II.1.rds
                                        DATA 2 ... n = 100, p = 8000
                                        # DI.II.2.rds

   ? DataStudyI.Rout ..... output of the r-script
                                        "DataStudyI.R"

--------------------------------------------------------------------------------

-> /R/Kap5/SimStudyI/...                [Kapitel 5.2.2]
                                        [Kapitel 5.2.3]

--------------------------------------------------------------------------------


--------------------------------------------------------------------------------

-> /R/Kap5/SimStudyI/DI.I.1             ( Study I & Case I & DATA 1 )

--------------------------------------------------------------------------------

   ? Start.txt ..... commands for Batch Execution of
                                        r-scripts (est_StudyI_I_1.R)
   ~ est_StudyI_I_1.R ..... see Algorithm 8

                                        Generate 200 new random forest:
                                        # reg.rf_DI.I.1.RData
                                        Save original PerVIM measures:
```

```
                                    # VarImp_DI.I.1.RData
                                    NTA:
                                    Compute and save the
                                    original CvVIM(k=2) measures
                                    and p-values NTA
                                    # nta_DI.I.1.RData
                                    NuDA:
                                    Compute and Save the
                                    Null importance
                                    distribution approximations
                                    and p-values NuDA
                                    # nuda_DI.I.1.RData
                                    PIMP:
                                    Compute and Save the
                                    S permuted VarImp measures
                                    for each predictor variables and
                                    p-values PIMP (non-param.) and
                                    p-values PIMP (param.) and the
                                    p-values of the Kolm.-Smirnov Tests
                                    # pimp_DI.I.1.RData

   ? est_StudyI_I_1.Rout ..... output of the r-script
                                    "est_StudyI_I_1.R"
   ~ eval_StudyI_I_1.R ..... Evaluation of the computed/
                                    estimated data from the results of
                                    "est_StudyI_I_1.R"
   # Eval_DI.I.1.RData ..... save results of
                                    "eval_StudyI_I_1.R

-------------------------------------------------------------------------
-> /R/Kap5/SimStudyI/DI.I.2          ( Study I & Case I & DATA 2 )
-------------------------------------------------------------------------

   ? Start.txt ..... commands for Batch Execution of
                                    r-scripts (est_StudyI_I_2.R)
   ~ est_StudyI_I_2.R ..... see Algorithm 8

                                    Generate 200 new random forest:
                                    # reg.rf_DI.I.2.RData
                                    Save original PerVIM measures:
                                    # VarImp_DI.I.2.RData
                                    NTA:
                                    Compute and save the
                                    original CvVIM(k=2) measures
                                    and p-values NTA
```

```
                                    # nta_DI.I.2.RData
                                    NuDA:
                                    Compute and Save the
                                    Null importance
                                    distribution approximations
                                    and p-values NuDA
                                    # nuda_DI.I.2.RData
                                    PIMP:
                                    Compute and Save the
                                    S permuted VarImp measures
                                    for each predictor variables and
                                    p-values PIMP (non-param.) and
                                    p-values PIMP (param.) and the
                                    p-values of the Kolm.-Smirnov Tests
                                    # pimp_DI.I.2.RData

  ? est_StudyI_I_2.Rout ..... output of the r-script
                                    "est_StudyI_I_2.R"
  ~ eval_StudyI_I_2.R ..... Evaluation of the computed/
                                    estimated data from the results of
                                    "est_StudyI_I_2.R"
  # Eval_DI.I.2.RData ..... save results of
                                    "eval_StudyI_I_2.R

-------------------------------------------------------------------------
-> /R/Kap5/SimStudyI/DI.II.1          ( Study I & Case II & DATA 1 )
-------------------------------------------------------------------------

  ? Start.txt ..... commands for Batch Execution of
                                    r-scripts (est_StudyI_II_1.R)
  ~ est_StudyI_II_1.R ..... see Algorithm 8

                                    Generate 200 new random forest:
                                    # reg.rf_DI.II.1.RData
                                    Save original PerVIM measures:
                                    # VarImp_DI.II.1.RData
                                    NTA:
                                    Compute and save the
                                    original CvVIM(k=2) measures
                                    and p-values NTA
                                    # nta_DI.II.1.RData
                                    NuDA:
                                    Compute and Save the
                                    Null importance
                                    distribution approximations
```

```
                                    and p-values NuDA
                                    # nuda_DI.II.1.RData
                                    PIMP:
                                    Compute and Save the
                                    S permuted VarImp measures
                                    for each predictor variables and
                                    p-values PIMP (non-param.) and
                                    p-values PIMP (param.) and the
                                    p-values of the Kolm.-Smirnov Tests
                                    # pimp_DI.II.1.RData

  ? est_StudyI_II_1.Rout ..... output of the r-script
                                    "est_StudyI_II_1.R"
  ~ eval_StudyI_II_1.R ..... Evaluation of the computed/
                                    estimated data from the results of
                                    "est_StudyI_II_1.R"
  # Eval_DI.II.1.RData ..... save results of
                                    "eval_StudyI_II_1.R

-------------------------------------------------------------------------
-> /R/Kap5/SimStudyI/DI.II.2        ( Study I & Case II & DATA 2 )
-------------------------------------------------------------------------

  ? Start.txt ..... commands for Batch Execution of
                                    r-scripts (est_StudyI_II_2.R)
  ~ est_StudyI_II_2.R ..... see Algorithm 8

                                    Generate 200 new random forest:
                                    # reg.rf_DI.II.2.RData
                                    Save original PerVIM measures:
                                    # VarImp_DI.II.2.RData
                                    NTA:
                                    Compute and save the
                                    original CvVIM(k=2) measures
                                    and p-values NTA
                                    # nta_DI.II.2.RData
                                    NuDA:
                                    Compute and Save the
                                    Null importance
                                    distribution approximations
                                    and p-values NuDA
                                    # nuda_DI.II.2.RData
                                    PIMP:
                                    Compute and Save the
                                    S permuted VarImp measures
```

```
                                    for each predictor variables and
                                    p-values PIMP (non-param.) and
                                    p-values PIMP (param.) and the
                                    p-values of the Kolm.-Smirnov Tests
                                    # pimp_DI.II.2.RData

   ? est_StudyI_II_2.Rout ..... output of the r-script
                                    "est_StudyI_II_2.R"
   ~ eval_StudyI_II_2.R ..... Evaluation of the computed/
                                    estimated data from the results of
                                    "est_StudyI_II_2.R"
   # Eval_DI.II.2.RData ..... save results of
                                    "eval_StudyI_II_2.R

-------------------------------------------------------------------------
-> /R/Kap5/SimStudyI/               [Kapitel 5.3]
-------------------------------------------------------------------------

   ~ eval_StudyI.R ..... Evaluation of Study I
                                    1) null imp. distrib. mtry=p/3
                                      Abbildung 5.1
                                    2) null imp. distrib. mtry=p/3,
                                      zoomed Abbildung 5.2
                                    3) null imp. distrib. mtry=p/5
                                       Abbildung A.1
                                    4) null imp. distrib. mtry=p/5,
                                       zoomed Abbildung A.2
                                    5) null imp. distrib. mtry=p/10
                                       Abbildung A.3
                                    6) null imp. distrib. mtry=p/10,
                                       zoomed Abbildung A.4
                                    7) Probability of Type I error,
                                       Study I Case 1
                                       Abbildung 5.3,
                                       Tabelle 2,
                                       Tabelle 18
                                    8) Probability of Type I error,
                                       Study I Case 2
                                       Abbildung 5.4,
                                       Tabelle 3,
                                       Tabelle 19
   + density.p3.png ..... null imp. distrib. mtry=p/3
                                    Abbildung 5.1
   + density.p3.z.png ..... null imp. distrib. mtry=p/3,
                                    zoomed Abbildung 5.2
```

```
+ density.p5.png ..... null imp. distrib. mtry=p/5
                       Abbildung A.1
+ density.p5.z.png ... null imp. distrib. mtry=p/5,
                       zoomed Abbildung A.2
+ density.p10.png .... null imp. distrib. mtry=p/10
                       Abbildung A.3
+ density.p10.z.png .. null imp. distrib. mtry=p/10,
                       zoomed Abbildung A.4
+ type.I.err.S1.C1.png . Probability of Type I error,
                       Study I Case 1
                       Abbildung 5.3
+ type.I.err.S1.C2.png . Probability of Type I error,
                       Study I Case 2
                       Abbildung 5.4
```

-> R/Kap5/Data/DataStudyII [Kapitel 5.4.1]

```
? Start.txt ..... commands for Batch Execution of
                  r-scripts (DataStudyII.R)
~ DataStudyII.R .. DGP: Study II and
                   ------------------------
                   BetaSet = c(-3, -2, -1, -0.5,
                               0.5, 1, 2, 3)
                   Case I:
                   DATA 1 ... n = 100, p = 2000
                   # DII.I.1B.rds
                   DATA 2 ... n = 100, p = 8000
                   # DII.I.2B.rds

                   Case II:
                   DATA 1 ... n = 100, p = 2000
                   # DII.II.1B.rds
                   DATA 2 ... n = 100, p = 8000
                   # DII.II.2B.rds

                   ------------------------
                   BetaSet = c(3, 2, 1, 0.5,
                               0.5, 1, 2, 3)
                   Case 2B:
                   DATA 1 ... n = 100, p = 2000
                   # DII.II.1B2.rds
                   DATA 2 ... n = 100, p = 8000
                   # DII.II.2B2.rds
```

```
   ? DataStudyII.Rout ..... output of the r-script
                            "DataStudyII.R"
```

```
-> /R/Kap5/SimStudyI/...    [Kapitel 5.4.2]
                            [Kapitel 5.4.3]
```

```
-> /R/Kap5/SimStudyII/CaseI/DII.I.1B      ( Study II & Case I & DATA 1 )
```

```
   ? Start.txt ..... commands for Batch Execution of
                     r-scripts (est_StudyII_I_1B.R)
   ~ est_StudyII_I_1B.R ..... see Algorithm 8

                     Generate 200 new random forest:
                     # reg.rf_DII.I.1B.RData
                     Save original PerVIM measures:
                     # VarImp_DII.I.1B.RData
                     NTA:
                     Compute and save the
                     original CvVIM(k=2) measures
                     and p-values NTA
                     # nta_DII.I.1B.RData
                     NuDA:
                     Compute and Save the
                     Null importance
                     distribution approximations
                     and p-values NuDA
                     # nuda_DII.I.1B.RData
                     PIMP:
                     Compute and Save the
                     S permuted VarImp measures
                     for each predictor variables and
                     p-values PIMP (non-param.) and
                     p-values PIMP (param.) and the
                     p-values of the Kolm.-Smirnov Tests
                     # pimp_DII.I.1B.RData

   ? est_StudyII_I_1B.Rout ..... output of the r-script
                                 "est_StudyII_I_1B.R"
   ~ eval_StudyII_I_1B.R ..... Evaluation of the computed/
                               estimated data from the results of
                               "est_StudyII_I_1B.R"
   # Eval_DII.I.1B.RData ..... save results of
```

```
                                    "eval_StudyII_I_1B.R
```

-> /R/Kap5/SimStudyII/CaseI/DII.I.2B (Study II & Case I & DATA 2)

```
? Start.txt ..... commands for Batch Execution of
                  r-scripts (est_StudyII_I_2B.R)
~ est_StudyII_I_2B.R ..... see Algorithm 8

                  Generate 200 new random forest:
                  # reg.rf_DII.I.2B.RData
                  Save original PerVIM measures:
                  # VarImp_DII.I.2B.RData
                  NTA:
                  Compute and save the
                  original CvVIM(k=2) measures
                  and p-values NTA
                  # nta_DII.I.2B.RData
                  NuDA:
                  Compute and Save the
                  Null importance
                  distribution approximations
                  and p-values NuDA
                  # nuda_DII.I.2B.RData
                  PIMP:
                  Compute and Save the
                  S permuted VarImp measures
                  for each predictor variables and
                  p-values PIMP (non-param.) and
                  p-values PIMP (param.) and the
                  p-values of the Kolm.-Smirnov Tests
                  # pimp_DII.I.2B.RData

? est_StudyII_I_2B.Rout ..... output of the r-script
                  "est_StudyII_I_2B.R"
~ eval_StudyII_I_2B.R ..... Evaluation of the computed/
                  estimated data from the results of
                  "est_StudyII_I_2B.R"
# Eval_DII.I.2B.RData ..... save results of
                  "eval_StudyII_I_2B.R
```

-> /R/Kap5/SimStudyII/CaseII/DII.II.1B (Study II & Case II & DATA 1)

```
? Start.txt ..........commands for Batch Execution of
                      r-scripts (est_StudyII_II_1B.R)
~ est_StudyII_II_1B.R ..see Algorithm 8

                      Generate 200 new random forest:
                      # reg.rf_DII.II.1B.RData
                      Save original PerVIM measures:
                      # VarImp_DII.II.1B.RData
                      NTA:
                      Compute and save the
                      original CvVIM(k=2) measures
                      and p-values NTA
                      # nta_DII.II.1B.RData
                      NuDA:
                      Compute and Save the
                      Null importance
                      distribution approximations
                      and p-values NuDA
                      # nuda_DII.II.1B.RData
                      PIMP:
                      Compute and Save the
                      S permuted VarImp measures
                      for each predictor variables and
                      p-values PIMP (non-param.) and
                      p-values PIMP (param.) and the
                      p-values of the Kolm.-Smirnov Tests
                      # pimp_DII.II.1B.RData

? est_StudyII_II_1B.Rout ..output of the r-script
                      "est_StudyII_II_1B.R"
~ eval_StudyII_II_1B.R ....Evaluation of the computed/
                      estimated data from the results of
                      "est_StudyII_II_1B.R"
# Eval_DII.II.1B.RData ....save results of
                      "eval_StudyII_II_1B.R
```

-> /R/Kap5/SimStudyII/CaseII/DII.II.2B (Study II & Case II & DATA 2)

```
? Start.txt ..........commands for Batch Execution of
                      r-scripts (est_StudyII_II_2B.R)
~ est_StudyII_II_2B.R ..see Algorithm 8
```

```
                                    Generate 200 new random forest:
                                    # reg.rf_DII.II.2B.RData
                                    Save original PerVIM measures:
                                    # VarImp_DII.II.2B.RData
                                    NTA:
                                    Compute and save the
                                    original CvVIM(k=2) measures
                                    and p-values NTA
                                    # nta_DII.II.2B.RData
                                    NuDA:
                                    Compute and Save the
                                    Null importance
                                    distribution approximations
                                    and p-values NuDA
                                    # nuda_DII.II.2B.RData
                                    PIMP:
                                    Compute and Save the
                                    S permuted VarImp measures
                                    for each predictor variables and
                                    p-values PIMP (non-param.) and
                                    p-values PIMP (param.) and the
                                    p-values of the Kolm.-Smirnov Tests
                                    # pimp_DII.II.2B.RData

  ? est_StudyII_II_2B.Rout ..... output of the r-script
                                    "est_StudyII_II_2B.R"
  ~ eval_StudyII_II_2B.R ..... Evaluation of the computed/
                                    estimated data from the results of
                                    "est_StudyII_II_2B.R"
  # Eval_DII.II.2B.RData ..... save results of
                                    "eval_StudyII_II_2B.R
```

-> /R/Kap5/SimStudyII/Case2B/DII.II.1B2 (Study II & Case 2B & DATA 1)

```
  ? Start.txt ..... commands for Batch Execution of
                                    r-scripts (est_StudyII_II_1B2.R)
  ~ est_StudyII_II_1B2.R ..... see Algorithm 8

                                    Generate 200 new random forest:
                                    # reg.rf_DII.II.1B2.RData
                                    Save original PerVIM measures:
                                    # VarImp_DII.II.1B2.RData
                                    NTA:
```

```
                                    Compute and save the
                                    original CvVIM(k=2) measures
                                    and p-values NTA
                                    # nta_DII.II.1B2.RData
                                    NuDA:
                                    Compute and Save the
                                    Null importance
                                    distribution approximations
                                    and p-values NuDA
                                    # nuda_DII.II.1B2.RData
                                    PIMP:
                                    Compute and Save the
                                    S permuted VarImp measures
                                    for each predictor variables and
                                    p-values PIMP (non-param.) and
                                    p-values PIMP (param.) and the
                                    p-values of the Kolm.-Smirnov Tests
                                    # pimp_DII.II.1B2.RData

? est_StudyII_II_1B2.Rout ..... output of the r-script
                                    "est_StudyII_II_1B2.R"
~ eval_StudyII_II_1B2.R ..... Evaluation of the computed/
                                    estimated data from the results of
                                    "est_StudyII_II_1B2.R"
# Eval_DII.II.1B2.RData ..... save results of
                                    "eval_StudyII_II_1B2.R

-----------------------------------------------------------------------
-> /R/Kap5/SimStudyII/Case2B/DII.II.2B2  ( Study II & Case 2B & DATA 2 )
-----------------------------------------------------------------------

? Start.txt ..... commands for Batch Execution of
                                    r-scripts (est_StudyII_II_2B2.R)
~ est_StudyII_II_2B2.R ..... see Algorithm 8

                                    Generate 200 new random forest:
                                    # reg.rf_DII.II.2B2.RData
                                    Save original PerVIM measures:
                                    # VarImp_DII.II.2B2.RData
                                    NTA:
                                    Compute and save the
                                    original CvVIM(k=2) measures
                                    and p-values NTA
                                    # nta_DII.II.2B2.RData
                                    NuDA:
```

```
                                    Compute and Save the
                                    Null importance
                                    distribution approximations
                                    and p-values NuDA
                                    # nuda_DII.II.2B2.RData
                                    PIMP:
                                    Compute and Save the
                                    S permuted VarImp measures
                                    for each predictor variables and
                                    p-values PIMP (non-param.) and
                                    p-values PIMP (param.) and the
                                    p-values of the Kolm.-Smirnov Tests
                                    # pimp_DII.II.2B2.RData

   ? est_StudyII_II_2B2.Rout ..... output of the r-script
                                    "est_StudyII_II_2B2.R"
   ~ eval_StudyII_II_2B2.R ....... Evaluation of the computed/
                                    estimated data from the results of
                                    "est_StudyII_II_2B2.R"
   # Eval_DII.II.2B2.RData ....... save results of
                                    "eval_StudyII_II_2B2.R

-------------------------------------------------------------------------
-> /R/Kap5/SimStudyII/...          [Kapitel 5.4]
-------------------------------------------------------------------------


-------------------------------------------------------------------------
-> /R/Kap5/SimStudyII/CaseI/       ( Study II & Case I )
-------------------------------------------------------------------------

   ~ eval_StudyII_I.R ............ Evaluation of Study II & Case I
                                    1) Boxplot  Power
                                       Abbildung 5.5
                                    2) Table Power,mtry=p/3, p=2000
                                       Tabelle 6
                                    3) Table Power, mtry=p/5, p=2000
                                       Tabelle 20
                                    4) Table Power, mtry=p/10, p=2000
                                       Tabelle 21
                                    5) Table Power, mtry=p/3, p=8000
                                       Tabelle 7
                                    6) Table Power, mtry=p/5, p=8000
                                       Tabelle 22
                                    7) Table Power, mtry=p/10, p=8000
                                       Tabelle 23
```

```
                                    8) Probability of Type I error,
                                       Study I Case 1
                                       Abbildung 5.6,
                                       Tabelle 8,
                                       Tabelle 24
   + Power.SII.C1.png ..................Boxplot  Power
                                       Study II Case 1
                                       Abbildung 5.5
   + type.I.err.SII.C1.png .............Probability of Type I error,
                                       Study II Case 1
                                       Abbildung 5.6

-------------------------------------------------------------------------
-> /R/Kap5/SimStudyII/CaseII/          ( Study II & Case II )
-------------------------------------------------------------------------

   ~ eval_StudyII_II.R .................Evaluation of Study II & Case II
                                    1) Boxplot  Power
                                       Abbildung 5.7
                                    2) Boxplot  Power, mtry=p/3
                                       Abbildung 5.8,
                                       Tabelle 9,
                                       Tabelle 10
                                    3) Boxplot  Power, mtry=p/5
                                       Abbildung A.5,
                                       Tabelle 25,
                                       Tabelle 26
                                    4) Boxplot  Power, mtry=p/10
                                       Abbildung A.6,
                                       Tabelle 27,
                                       Tabelle 28
                                    5) Boxplot  Power, beta=0, p=2000
                                       Abbildung 5.9,
                                       Tabelle 11
                                    6) Boxplot  Power, beta=0, p=8000
                                       Abbildung A.7,
                                       Tabelle 29
                                    7) Probability of Type I error,
                                       Study I Case II
                                       Abbildung 5.14,
                                       Tabelle 15,
                                       Tabelle 30
   + Power.SII.C2.png ..................Boxplot  Power
                                       Study II Case II
                                       Abbildung 5.7
```

```
+ Power.SII.C2.p3.png ..... Boxplot  Power, mtry=p/3
                            Study II Case II
                            Abbildung 5.8
+ Power.SII.C2.p5.png ..... Boxplot  Power, mtry=p/5
                            Study II Case II
                            Abbildung A.5
+ Power.SII.C2.p10.png .... Boxplot  Power, mtry=p/10
                            Study II Case II
                            Abbildung A.6
+ Power.SII.C2.0.D1.png ... Boxplot  Power, beta=0, p=2000
                            Study II Case II
                            Abbildung 5.9
+ Power.SII.C2.0.D2.png ... Boxplot  Power, beta=0, p=8000
                            Study II Case II
                            Abbildung A.7
+ type.I.err.SII.C2.png ... Probability of Type I error,
                            Study II Case II
                            Abbildung 5.14

-------------------------------------------------------------

-> /R/Kap5/SimStudyII/Case2B/      ( Study II & Case 2B )

-------------------------------------------------------------

~ eval_StudyII_II2.R ..... Evaluation of Study I & Case 2B
                           1) Boxplot  Power
                              Abbildung 5.10
                           2) Boxplot  Power, mtry=p/3
                              Abbildung 5.11,
                              Tabelle 12,
                              Tabelle 13
                           3) Boxplot  Power, mtry=p/5
                              Abbildung A.8,
                              Tabelle 31,
                              Tabelle 32
                           4) Boxplot  Power, mtry=p/10
                              Abbildung A.9,
                              Tabelle 33,
                              Tabelle 34
                           5) Boxplot  Power, NTA
                              Abbildung 5.12
                           6) Boxplot  Power, beta=0, p=2000
                              Abbildung 5.13,
                              Tabelle 14
                           7) Boxplot  Power, beta=0, p=8000
                              Abbildung A.10,
```

Tabelle 35
8) Probability of Type I error,
Study I Case 2B
Abbildung 5.15,
Tabelle 16,
Tabelle 17

+ Power.SII.C2.2.png ..... Boxplot Power
Study II Case 2B
Abbildung 5.10

+ Power.SII.C2.p3.2.png ..... Boxplot Power, mtry=p/3
Study II Case 2B
Abbildung 5.11

+ Power.SII.C2.p5.2.png ..... Boxplot Power, mtry=p/5
Study II Case 2B
Abbildung A.8

+ Power.SII.C2.p10.2.png ..... Boxplot Power, mtry=p/10
Study II Case 2B
Abbildung A.9

+ Power.SII.C2.NTA.2.png ..... Boxplot Power, NTA
Study II Case 2B
Abbildung 5.12

+ Power.SII.C2.0.D1.2.png ..... Boxplot Power, beta=0, p=2000
Study II Case 2B
Abbildung 5.13

+ Power.SII.C2.0.D2.2.png ..... Boxplot Power, beta=0, p=8000
Study II Case 2B
Abbildung A.10

+ type.I.err2.SII.C2.png ..... Probability of Type I error,
Study II Case 2B
Abbildung 5.15