Signifikanztests im hochdimensionalen Kontext

Ein Überblick

Thomas Schäfer

15.12.2014

Inhaltsverzeichnis

Inhaltsverzeichnis

Inhaltsverzeichnis

Inhaltsverzeichnis

Inhaltsverzeichnis

Modelannahmen

Wir werden folgendes lineare Modell betrachten :

$$Y = X \cdot \beta + \varepsilon$$

wobei $Y \in \mathbb{R}^{n \times 1}, X \in \mathbb{R}^{n \times p}$ und $\varepsilon \in \mathbb{R}^{n \times 1}$. ε ist unabhänig von X mit i.i.d. Komponenten und Erwartungswert $= 0$. Y ist der Ergebnisvektor und X die Designmatrix.

Notation

Definition

- $S_0 := \{j;\ \beta_j^0 \neq 0,\ j = 1, ..., p\}$ wobei β^0 der wahre Parameter ist.
- $\widehat{\beta}^I(\lambda)$ ist der Lasso Schätzer für den Parameter β abhängig von λ basierend auf den Daten von $I \subset \{1, ..., n\}$.
- $\widehat{S}_\lambda(I) = \{j;\ \widehat{\beta}_j^I(\lambda) \neq 0,\ j = 1, ..., p\}$ Schätzer für S_0 basierend auf $\widehat{\beta}^I(\lambda)$.

stability selection

Sei $I \subset \{1, ..., n\}$ eine zufällig gezogene Stichprobe mit $|I| = \left\lfloor \frac{n}{2} \right\rfloor$. Dann

$$\widehat{\Pi}_K(\lambda) := \mathbb{P}[K \subset \widehat{S}_\lambda(I)]$$

mit $K \subset \{1, ..., p\}$ und $\mathbb{P}$ ist die relative Häufigkeit von $K \subset \widehat{S}_\lambda(I)$ über alle Möglichkeiten $\binom{n}{\lfloor \frac{n}{2} \rfloor}$.

stability selection

Sei $I \subset \{1, ..., n\}$ eine zufällig gezogene Stichprobe mit $|I| = \lfloor \frac{n}{2} \rfloor$. Dann

$$\widehat{\Pi}_K(\lambda) := \mathbb{P}[K \subset \widehat{S}_\lambda(I)]$$

mit $K \subset \{1, ..., p\}$ und $\mathbb{P}$ ist die relative Häufigkeit von $K \subset \widehat{S}_\lambda(I)$ über alle Möglichkeiten $\binom{n}{\lfloor \frac{n}{2} \rfloor}$. Das heißt :

$$\widehat{\Pi}_K(\lambda) = \sum_I \frac{\mathbb{1}(K \subset \widehat{S}_\lambda(I))}{\binom{n}{\lfloor \frac{n}{2} \rfloor}}$$

bzw. für $K = \{j\}$

$$\widehat{\Pi}_j(\lambda) = \sum_I \frac{\mathbb{1}(j \subset \widehat{S}_\lambda(I))}{\binom{n}{\lfloor \frac{n}{2} \rfloor}}$$

Definition

Sei $\pi_{thr} \in (0, 1]$ und $\Lambda \subset \mathbb{R}^+$. Dann

$$\widehat{S}_{stable} = \{j;\ \max_{\lambda \in \Lambda} \widehat{\Pi}_j(\lambda) \geq \pi_{thr}\}$$

Ziel: Mit einer guten Wahl von π_{thr} die Wahrscheinlichkeit für den Fehler 1. Art zu kontrollieren.
Fehler 1. Art: Der Parameter β_j^0 ist $= 0$, aber der Schätzer ist $\neq 0$.

$$\widehat{S}_\Lambda(I) = \bigcup_{\lambda \in \Lambda} \widehat{S}_\lambda(I) = \{j;\ \ \exists \lambda \in \Lambda \text{ mit } j \in \widehat{S}_\lambda(I)\}$$

$$q_\Lambda = \mathbb{E}[|\widehat{S}_\Lambda(I)|] \qquad\qquad V = |S_0^C \cap \widehat{S}_{stable}|$$

$$\widehat{S}_\Lambda(I) = \bigcup_{\lambda\in\Lambda} \widehat{S}_\lambda(I) = \{j;\ \ \exists\lambda \in \Lambda \text{ mit } j \in \widehat{S}_\lambda(I)\}$$

$$q_\Lambda = \mathbb{E}[|\widehat{S}_\Lambda(I)|] \qquad\qquad V = |S_0^C \cap \widehat{S}_{stable}|$$

Satz 1

Sei $\Lambda \subset \mathbb{R}^+$. Seien die Verteilungen von $\{1(j \in \widehat{S}_\lambda(I)), j \in S_0^c\}$ austauschbar für alle $\lambda \in \Lambda$. Sei weiter der Schätzer nicht schlechter als zufällig gewählt, dh.

$$\frac{\mathbb{E}[|S_0 \cap \widehat{S}_\Lambda|]}{\mathbb{E}[|S_0^c \cap \widehat{S}_\Lambda|]} \geq \frac{|S_0|}{|S_0^c|}$$

Dann ist für $\pi_{thr} \in (\frac{1}{2}, 1]$

$$\mathbb{E}[V] \leq \frac{1}{2\pi_{thr} - 1} \frac{q_\Lambda^2}{p}$$

Notation

$I^1,\ I^2 \subset \{1, ..., n\}$ mit $|I^1| = |I^2| = \lfloor \frac{n}{2} \rfloor$ und $I^1 \cap I^2 = \varnothing$

$$\widehat{S}^{simult,\lambda} = \widehat{S}_\lambda(I^1) \cap \widehat{S}_\lambda(I^2), \qquad \widehat{\Pi}_K^{simult}(\lambda) = \mathbb{P}[K \subset \widehat{S}^{simult,\lambda}]$$

Notation

$I^1,\ I^2 \subset \{1, ..., n\}$ mit $|I^1| = |I^2| = \lfloor \frac{n}{2} \rfloor$ und $I^1 \cap I^2 = \varnothing$

$$\widehat{S}^{simult,\lambda} = \widehat{S}_\lambda(I^1) \cap \widehat{S}_\lambda(I^2), \qquad \widehat{\Pi}_K^{simult}(\lambda) = \mathbb{P}[K \subset \widehat{S}^{simult,\lambda}]$$

Lemma 2

Sei $K \subset \{1, ..., p\}$ und $\widehat{S}_\lambda(I)$ ein Schätzer mit $|I| = \lfloor \frac{n}{2} \rfloor$.
Falls $\mathbb{P}[K \subset \widehat{S}_\lambda(I)] \leq \varepsilon$, dann $\mathbb{P}[\widehat{\Pi}_K^{simult}(\lambda) \geq \zeta] \leq \frac{\varepsilon^2}{\zeta}$
Falls $\mathbb{P}[K \subset \bigcup_{\lambda \in \Lambda} \widehat{S}_\lambda] \leq \varepsilon$ für ein $\Lambda \subset \mathbb{R}^+$, dann
$\mathbb{P}[\max_{\lambda \in \Lambda} \widehat{\Pi}_K^{simult}(\lambda) \geq \zeta] \leq \frac{\varepsilon^2}{\zeta}$

Notation

$I^1,\ I^2 \subset \{1,...,n\}$ mit $|I^1| = |I^2| = \lfloor \frac{n}{2} \rfloor$ und $I^1 \cap I^2 = \varnothing$

$$\widehat{S}^{simult,\lambda} = \widehat{S}_\lambda(I^1) \cap \widehat{S}_\lambda(I^2), \qquad \widehat{\Pi}_K^{simult}(\lambda) = \mathbb{P}[K \subset \widehat{S}^{simult,\lambda}]$$

Lemma 2

Sei $K \subset \{1,...,p\}$ und $\widehat{S}_\lambda(I)$ ein Schätzer mit $|I| = \lfloor \frac{n}{2} \rfloor$.
Falls $\mathbb{P}[K \subset \widehat{S}_\lambda(I)] \leq \varepsilon$, dann $\mathbb{P}[\widehat{\Pi}_K^{simult}(\lambda) \geq \zeta] \leq \frac{\varepsilon^2}{\zeta}$
Falls $\mathbb{P}[K \subset \bigcup_{\lambda\in\Lambda} \widehat{S}_\lambda] \leq \varepsilon$ für ein $\Lambda \subset \mathbb{R}^+$, dann
$\mathbb{P}[\max_{\lambda\in\Lambda} \widehat{\Pi}_K^{simult}(\lambda) \geq \zeta] \leq \frac{\varepsilon^2}{\zeta}$

Lemma 3

Für ein beliebiges $K \subset \{1,...,p\}$ und eine beliebige Relation ω zu den original n Daten im zugrundeliegendem Wahrscheinlichkeitsraum Ω gilt:

$$\widehat{\Pi}_K^{simult}(\lambda) \geq 2\widehat{\Pi}_K(\lambda) - 1$$

Was bringt dieser Satz?

Ziel, dass $\mathbb{E}[V] \leq \alpha$ für eine vorher bestimmte Schranke α. Damit folgt $\mathbb{P}[V > 0] \leq \alpha$ und somit ist die Wahrscheinlichkeit mindestens eine falsche positive Wahl zu treffen kontrollierbar.

$$\mathbb{E}[V] \leq \frac{1}{2\pi_{thr} - 1} \frac{q_\Lambda^2}{p} \stackrel{!}{=} \alpha$$

Was bringt dieser Satz?

Ziel, dass $\mathbb{E}[V] \leq \alpha$ für eine vorher bestimmte Schranke α. Damit folgt $\mathbb{P}[V > 0] \leq \alpha$ und somit ist die Wahrscheinlichkeit mindestens eine falsche positive Wahl zu treffen kontrollierbar.

$$\mathbb{E}[V] \leq \frac{1}{2\pi_{thr} - 1} \frac{q_\Lambda^2}{p} \stackrel{!}{=} \alpha$$

Für $q_\Lambda^2 \leq p\alpha$ wähle $\pi_{thr} = \frac{1}{2}\left(\frac{q_\Lambda^2}{p\alpha} + 1\right)$

Was bringt dieser Satz?

Ziel, dass $\mathbb{E}[V] \leq \alpha$ für eine vorher bestimmte Schranke α. Damit folgt $\mathbb{P}[V > 0] \leq \alpha$ und somit ist die Wahrscheinlichkeit mindestens eine falsche positive Wahl zu treffen kontrollierbar.

$$\mathbb{E}[V] \leq \frac{1}{2\pi_{thr} - 1} \frac{q_\Lambda^2}{p} \stackrel{!}{=} \alpha$$

Für $q_\Lambda^2 \leq p\alpha$ wähle $\pi_{thr} = \frac{1}{2}\left(\frac{q_\Lambda^2}{p\alpha} + 1\right)$

Was tun für $q_\Lambda^2 > p\alpha$?

Was bringt dieser Satz?

Ziel, dass $\mathbb{E}[V] \leq \alpha$ für eine vorher bestimmte Schranke α. Damit folgt $\mathbb{P}[V > 0] \leq \alpha$ und somit ist die Wahrscheinlichkeit mindestens eine falsche positive Wahl zu treffen kontrollierbar.

$$\mathbb{E}[V] \leq \frac{1}{2\pi_{thr} - 1} \frac{q_\Lambda^2}{p} \stackrel{!}{=} \alpha$$

Für $q_\Lambda^2 \leq p\alpha$ wähle $\pi_{thr} = \frac{1}{2}\left(\frac{q_\Lambda^2}{p\alpha} + 1\right)$

Was tun für $q_\Lambda^2 > p\alpha$?

- Erweitere Λ s.d. q_Λ kleiner wird
- Erhöhe die Schranke α s.d. $q_\Lambda^2 \leq p\alpha$

Voraussetzung der Austauschbarkeit

Bei wahren Daten können wir die Austauschbarkeit nicht garantieren, aber reale Beispiele zeigen, dass die Schranke „vergleichsweise gut hält“.
Wir konstruieren unser Beispiel mit folgenden Parametern:

Voraussetzung der Austauschbarkeit

Bei wahren Daten können wir die Austauschbarkeit nicht garantieren, aber reale Beispiele zeigen, dass die Schranke „vergleichsweise gut hält“.
Wir konstruieren unser Beispiel mit folgenden Parametern:

Sei $n = 60$, $p = 200$ und

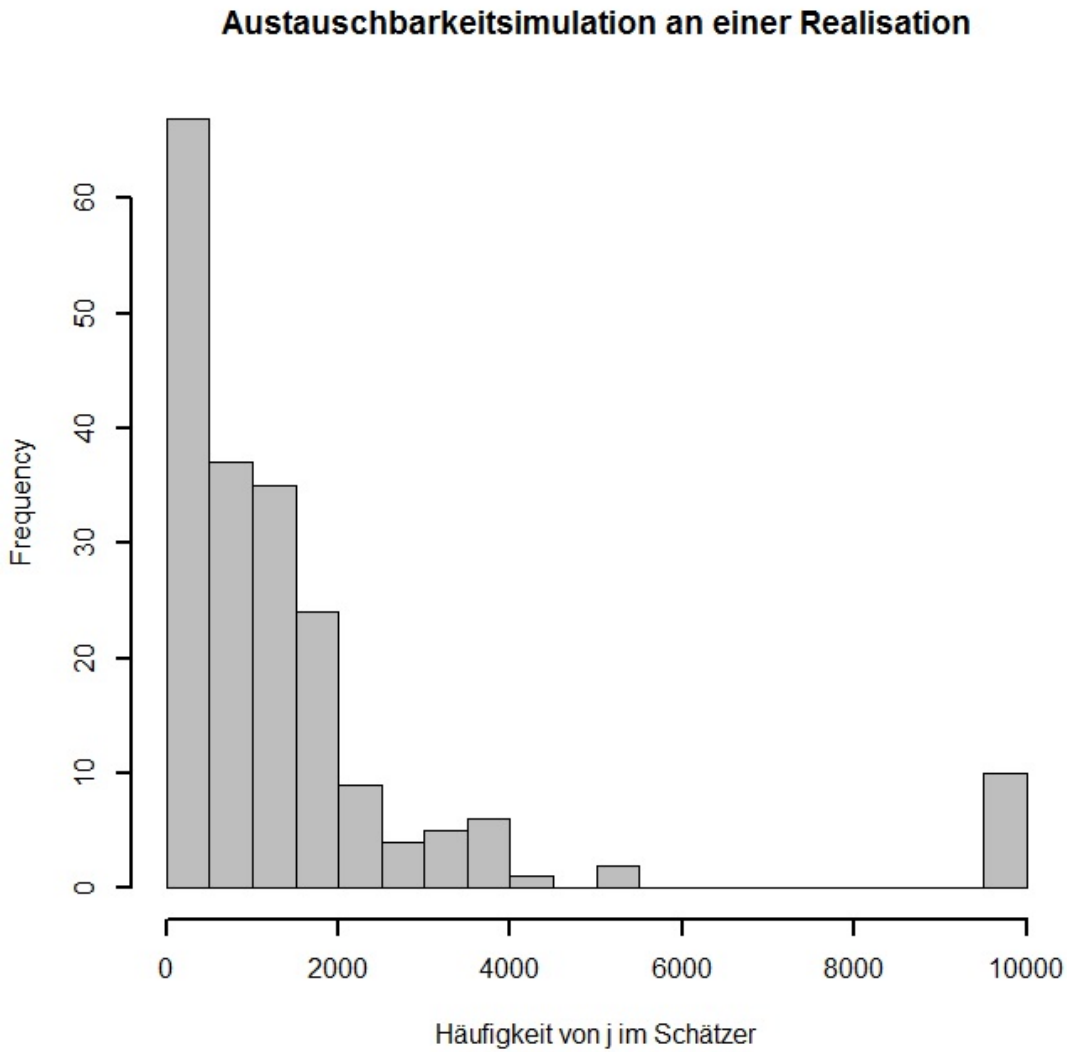
$$Y = X\beta + \varepsilon$$

wobei $Y \in \mathbb{R}^{n\times 1}$, $\varepsilon \sim \mathcal{N}_n(0, 5 \cdot \mathbb{1}) \in \mathbb{R}^{n\times 1}$ und $X \in \mathbb{R}^{n\times p}$ mit Spalten

$$X_i \sim \mathcal{N}_p(0, \Sigma) \text{ i.i.d. mit } \Sigma = \begin{pmatrix} 1 & 0.5 & \cdots & 0.5 \\ 0.5 & 1 & & \vdots \\ \vdots & & \ddots & 0.5 \\ 0.5 & \ldots & 0.5 & 1 \end{pmatrix}$$

sowie $\beta \in \mathbb{R}^{p\times 1}$ mit $\beta_j = \begin{cases} 3 & \text{falls } j = 0 \bmod 20 \\ 0 & \text{sonst} \end{cases}$

Vergleiche Beispiel 10.1 aus [BvdG11]



10.000 Durchläufe, X fest, ε zufällig in jedem Durchlauf. Häufigkeit von j in $\widehat{S}_\lambda(I)$.

single sample splitting

neues Ziel: schwächere Voraussetzungen als Austauschbarkeit. Dazu betrachten wir nun

$$Y = X\beta + \varepsilon$$

wobei $\varepsilon_1 ... \varepsilon_n$ i.i.d. $\sim \mathcal{N}(0, \sigma^2)$. Wir wollen für alle $j \in \{1, ..., p\}$ den p-Wert zum Test

$$H_{0,j} : \beta_j = 0 \qquad vs. \qquad H_{A,j} : \beta_j \neq 0$$

bestimmen.

single sample splitting Strategie

1. Teile die Daten in 2 Hälften. I_1, I_2 mit $I_1 \cap I_2 = \varnothing, |I_1| = \left\lfloor \frac{n}{2} \right\rfloor, |I_2| = n - \left\lfloor \frac{n}{2} \right\rfloor$.

single sample splitting Strategie

1. Teile die Daten in 2 Hälften. I_1, I_2 mit $I_1 \cap I_2 = \varnothing, |I_1| = \left\lfloor \frac{n}{2} \right\rfloor, |I_2| = n - \left\lfloor \frac{n}{2} \right\rfloor$.
2. Bestimme einen Schätzer $\widehat{S}(I_1)$ auf Grundlage der zugehörigen Daten (X_{I_1}, Y_{I_1}) (z.B. durch Lasso).

single sample splitting Strategie

1. Teile die Daten in 2 Hälften. I_1, I_2 mit $I_1 \cap I_2 = \varnothing, |I_1| = \left\lfloor \frac{n}{2} \right\rfloor, |I_2| = n - \left\lfloor \frac{n}{2} \right\rfloor$.
2. Bestimme einen Schätzer $\widehat{S}(I_1)$ auf Grundlage der zugehörigen Daten (X_{I_1}, Y_{I_1}) (z.B. durch Lasso).
3. Bestimme $\overline{p}_j$-Werte zur Hypothese $H_{0,j}$ mit der kleinsten Quadrate Methode auf Grundlage der Daten (X_{I_2}, Y_{I_2}) und nutze nur die Variablen aus $\widehat{S}(I_1)$.

single sample splitting Strategie

1. Teile die Daten in 2 Hälften. I_1, I_2 mit $I_1 \cap I_2 = \varnothing, |I_1| = \left\lfloor \frac{n}{2} \right\rfloor, |I_2| = n - \left\lfloor \frac{n}{2} \right\rfloor$.
2. Bestimme einen Schätzer $\widehat{S}(I_1)$ auf Grundlage der zugehörigen Daten (X_{I_1}, Y_{I_1}) (z.B. durch Lasso).
3. Bestimme $\overline{p}_j$-Werte zur Hypothese $H_{0,j}$ mit der kleinsten Quadrate Methode auf Grundlage der Daten (X_{I_2}, Y_{I_2}) und nutze nur die Variablen aus $\widehat{S}(I_1)$.
4. $\widetilde{p}_j := \begin{cases} \overline{p}_j, & \text{falls } j \in \widehat{S}(I_1) \\ 1, & sonst \end{cases}$

single sample splitting Strategie

1. Teile die Daten in 2 Hälften. I_1, I_2 mit $I_1 \cap I_2 = \varnothing, |I_1| = \left\lfloor \frac{n}{2} \right\rfloor, |I_2| = n - \left\lfloor \frac{n}{2} \right\rfloor$.
2. Bestimme einen Schätzer $\widehat{S}(I_1)$ auf Grundlage der zugehörigen Daten (X_{I_1}, Y_{I_1}) (z.B. durch Lasso).
3. Bestimme $\overline{p}_j$-Werte zur Hypothese $H_{0,j}$ mit der kleinsten Quadrate Methode auf Grundlage der Daten (X_{I_2}, Y_{I_2}) und nutze nur die Variablen aus $\widehat{S}(I_1)$.
4. $$\widetilde{p}_j := \begin{cases} \overline{p}_j, & \text{falls } j \in \widehat{S}(I_1) \\ 1, & sonst \end{cases}$$
5. korrigiere nun den Fehler des mehrfachen Testens

$$\widetilde{p}_{corr,j} := \min\{\widetilde{p}_j \cdot |\widehat{S}(I_1)|, 1\} \quad \text{für alle } j \in \{1, ..., p\}$$

Bonferroni-Korrektur

Beispieldesign

Wir konstruieren unser Beispiel mit folgenden Parametern:

Sei $n = 60$, $p = 200$ und

$$Y = X\beta + \varepsilon$$

wobei $Y \in \mathbb{R}^{n\times 1}$, $\varepsilon \sim \mathcal{N}_n(0, 1 \cdot \mathbb{1}) \in \mathbb{R}^{n\times 1}$ und $X \in \mathbb{R}^{n\times p}$ mit Spalten

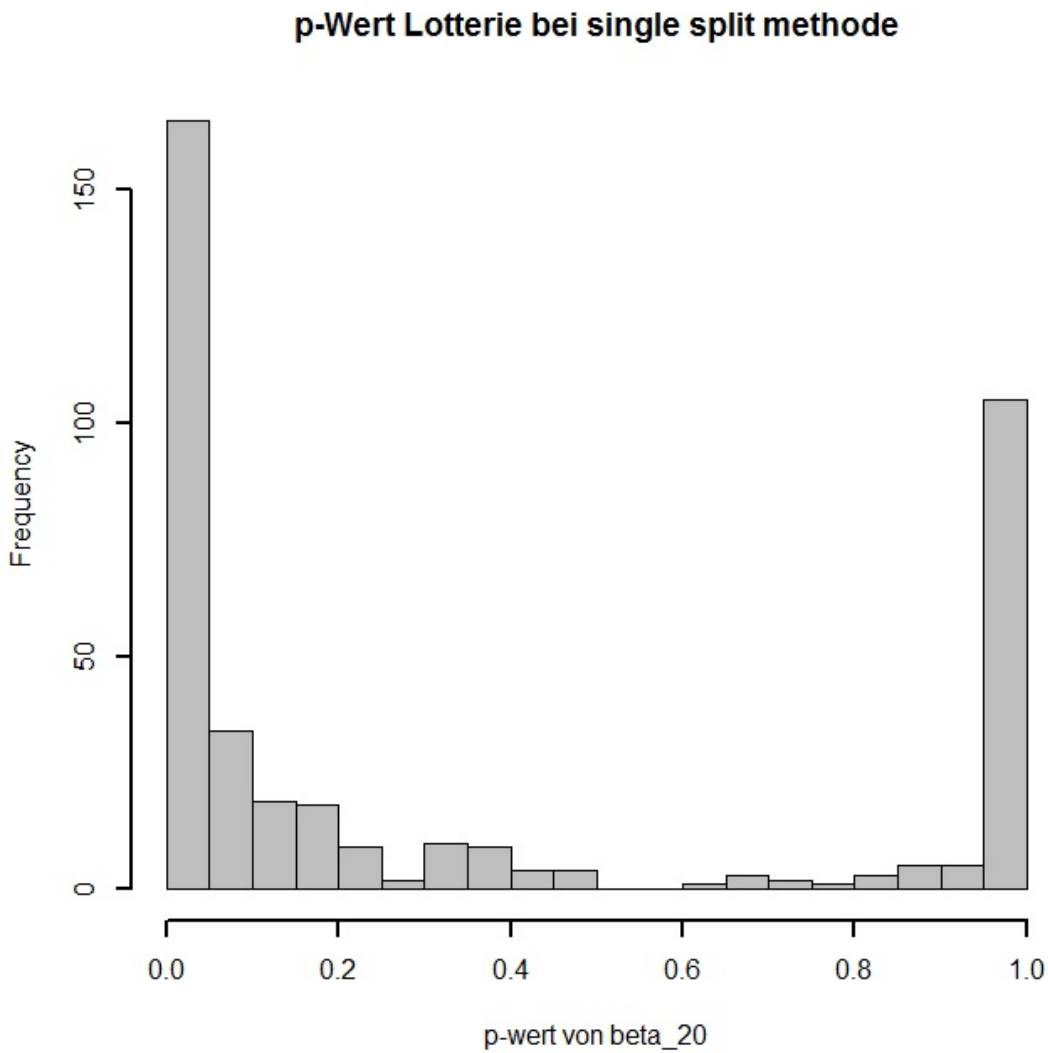
$$X_i \sim \mathcal{N}_p(0, \Sigma) \text{ i.i.d. mit } \Sigma = \begin{pmatrix} 1 & 0.5 & \cdots & 0.5 \\ 0.5 & 1 & & \vdots \\ \vdots & & \ddots & 0.5 \\ 0.5 & \ldots & 0.5 & 1 \end{pmatrix}$$

sowie $\beta \in \mathbb{R}^{p\times 1}$ mit $\beta_j = \begin{cases} 3 & \text{falls } j = 0 \bmod 20 \\ 0 & \text{sonst} \end{cases}$

Vergleiche Beispiel 10.1 aus [BvdG11]

p-Wert Lotterie

400 Durchläufe und berechne p_{20}



Annahmen

Um den Fehler kontrollieren zu können gehen wir im Folgenden von diesen Annahmen aus. Sei $\widehat{S}_m$ der Schätzer auf Grundlage von Daten $I_m \subset \{1, ..., n\}$ mit $|I_m| = m$.

Annahmen

Um den Fehler kontrollieren zu können gehen wir im Folgenden von diesen Annahmen aus. Sei $\widehat{S}_m$ der Schätzer auf Grundlage von Daten $I_m \subset \{1, ..., n\}$ mit $|I_m| = m$.

- $\lim_{n \to \infty} \mathbb{P}[\widehat{S}_{\lfloor \frac{n}{2} \rfloor} \supset S_0] = 1$

Annahmen

Um den Fehler kontrollieren zu können gehen wir im Folgenden von diesen Annahmen aus. Sei $\widehat{S}_m$ der Schätzer auf Grundlage von Daten $I_m \subset \{1, ..., n\}$ mit $|I_m| = m$.

- $\lim_{n\to\infty} \mathbb{P}[\widehat{S}_{\lfloor \frac{n}{2} \rfloor} \supset S_0] = 1$
- $\lim_{n\to\infty} \mathbb{P}[|\widehat{S}_{\lfloor \frac{n}{2} \rfloor}| < \frac{n}{2}] = 1$

Annahmen

Um den Fehler kontrollieren zu können gehen wir im Folgenden von diesen Annahmen aus. Sei $\widehat{S}_m$ der Schätzer auf Grundlage von Daten $I_m \subset \{1, ..., n\}$ mit $|I_m| = m$.

- $\lim_{n\to\infty} \mathbb{P}[\widehat{S}_{\lfloor \frac{n}{2} \rfloor} \supset S_0] = 1$
- $\lim_{n\to\infty} \mathbb{P}[|\widehat{S}_{\lfloor \frac{n}{2} \rfloor}| < \frac{n}{2}] = 1$
- Sei $\Sigma(I_m) = m^{-1} X_{I_m}^T X_{I_m}$. Sei $\Sigma(I_m)_{S,S}$ die zugehörige $|S| \times |S|$ Teilmatrix zu $S \subset \{1, ..., p\}$. Annahme:

  $\Lambda_{min}(\Sigma(I_m)_{S,S}) > 0$ für alle S mit $|S| < \frac{n}{2}$

  für alle I_m mit $|I_m| = m = n - \left\lfloor \frac{n}{2} \right\rfloor$

  mit $\Lambda_{min}(A)$ kleinster Eigenvektor von A.

$$\widehat{S}_{single-split}(\alpha) := \{j : \widetilde{p}_j \leq \alpha\}$$
$$V_{single-split}(\alpha) := |\widehat{S}_{single-split}(\alpha) \cap S_0^c|$$

Lemma 4

Unter vorherigen Annahmen gilt für $\alpha \in (0, 1)$

$$\limsup_{n \to \infty} \mathbb{P}[V_{single-split}(\alpha) > 0] \leq \alpha$$

Beweis s. Kapitel 11.1 aus [BvdG11]

multi sample splitting

Idee: Wir führen das „sample splitting“ Verfahren oft hintereinander aus.
Für $b = 1, ..., B$

multi sample splitting

Idee: Wir führen das „sample splitting“ Verfahren oft hintereinander aus.
Für $b = 1, ..., B$

- Spalte die Daten zufällig in I_1^b und I_2^b zu fast gleich großen Teilen.

multi sample splitting

Idee: Wir führen das „sample splitting“ Verfahren oft hintereinander aus.
Für $b = 1, ..., B$

- Spalte die Daten zufällig in I_1^b und I_2^b zu fast gleich großen Teilen.
- Benutze nur I_1^b um $\widehat{S}^b := \widehat{S}(I_1^b)$ zu bestimmen.

multi sample splitting

Idee: Wir führen das „sample splitting“ Verfahren oft hintereinander aus.
Für $b = 1, ..., B$

- Spalte die Daten zufällig in I_1^b und I_2^b zu fast gleich großen Teilen.
- Benutze nur I_1^b um $\widehat{S}^b := \widehat{S}(I_1^b)$ zu bestimmen.
- Benutze I_2^b zur Bestimmung der p-Werte
  $\widetilde{p}^b_{corr,j} = \min(\widetilde{p}^b_j |\widehat{S}^b|, 1)$ für $j = 1, ..., p$

multi sample splitting

Idee: Wir führen das „sample splitting“ Verfahren oft hintereinander aus.
Für $b = 1, ..., B$

- Spalte die Daten zufällig in I_1^b und I_2^b zu fast gleich großen Teilen.
- Benutze nur I_1^b um $\widehat{S}^b := \widehat{S}(I_1^b)$ zu bestimmen.
- Benutze I_2^b zur Bestimmung der p-Werte
  $\widetilde{p}^b_{corr,j} = \min(\widetilde{p}^b_j |\widehat{S}^b|, 1)$ für $j = 1, ..., p$
- Führe die p-Werte zusammen.

multi sample splitting

Idee: Wir führen das „sample splitting“ Verfahren oft hintereinander aus.
Für $b = 1, ..., B$

- Spalte die Daten zufällig in I_1^b und I_2^b zu fast gleich großen Teilen.
- Benutze nur I_1^b um $\widehat{S}^b := \widehat{S}(I_1^b)$ zu bestimmen.
- Benutze I_2^b zur Bestimmung der p-Werte
  $\widetilde{p}^b_{corr,j} = \min(\widetilde{p}^b_j |\widehat{S}^b|, 1)$ für $j = 1, ..., p$
- Führe die p-Werte zusammen.

Frage: Wie verbinde ich sinnvoll diese vielen p-Werte?

Zusammenführen der p-Werte

Betrachte die empirische γ Quantilfunktion $q_\gamma(\cdot)$. Sei $x = (x_1, x_2, ..., x_n)$ aufsteigend sortiert. Dann ist

$$q_\gamma(x) = \begin{cases} \frac{1}{2}(x_{n\cdot\gamma} + x_{n\cdot\gamma+1}), & \text{falls } n \cdot \gamma \text{ ganzzahlig.} \\ x_{\lceil n\cdot\gamma \rceil}, & \text{sonst} \end{cases}$$

Zusammenführen der p-Werte

Betrachte die empirische γ Quantilfunktion $q_\gamma(\cdot)$. Sei $x = (x_1, x_2, ..., x_n)$ aufsteigend sortiert. Dann ist

$$q_\gamma(x) = \begin{cases} \frac{1}{2}(x_{n\cdot\gamma} + x_{n\cdot\gamma+1}), & \text{falls } n\cdot\gamma \text{ ganzzahlig.} \\ x_{\lceil n\cdot\gamma \rceil}, & \text{sonst} \end{cases}$$

Sei $B = 10$ und $p_{corr,j} = (0.02, 0.05, 0.1, 0.2, 0.5, 0.5, 1, 1, 1, 1)$. Dann

$$q_{5\%}(0.02, 0.05, 0.1, 0.2, 0.5, 0.5, 1, 1, 1, 1) = 0.02$$

$$q_{50\%}(0.02, 0.05, 0.1, 0.2, 0.5, 0.5, 1, 1, 1, 1) = 0.5$$

Zusammenführen der p-Werte

Für γ definiere:

$$Q_j(\gamma) = \min\{q_\gamma\left(\{\frac{\widetilde{p}^b_{corr,j}}{\gamma}; b = 1, ..., B\}\right), 1\}$$

mit $q_\gamma(\cdot)$ als empirische γ Quantilfunktion.

Zusammenführen der p-Werte

Für γ definiere:

$$Q_j(\gamma) = \min\{q_\gamma\left(\{\frac{\widetilde{p}^b_{corr,j}}{\gamma}; b = 1, ..., B\}\right), 1\}$$

mit $q_\gamma(\cdot)$ als empirische γ Quantilfunktion.
Wir wählen eine angemessen Wert indem wir eine untere Schranke $\gamma_{min} \in (0, 1)$ wählen und definieren:

$$p_j := \min\left\{(1 - \log(\gamma_{min})) \inf_{\gamma \in (\gamma_{min}, 1)} Q_j(\gamma), 1\right\}$$

Zusammenführen der p-Werte

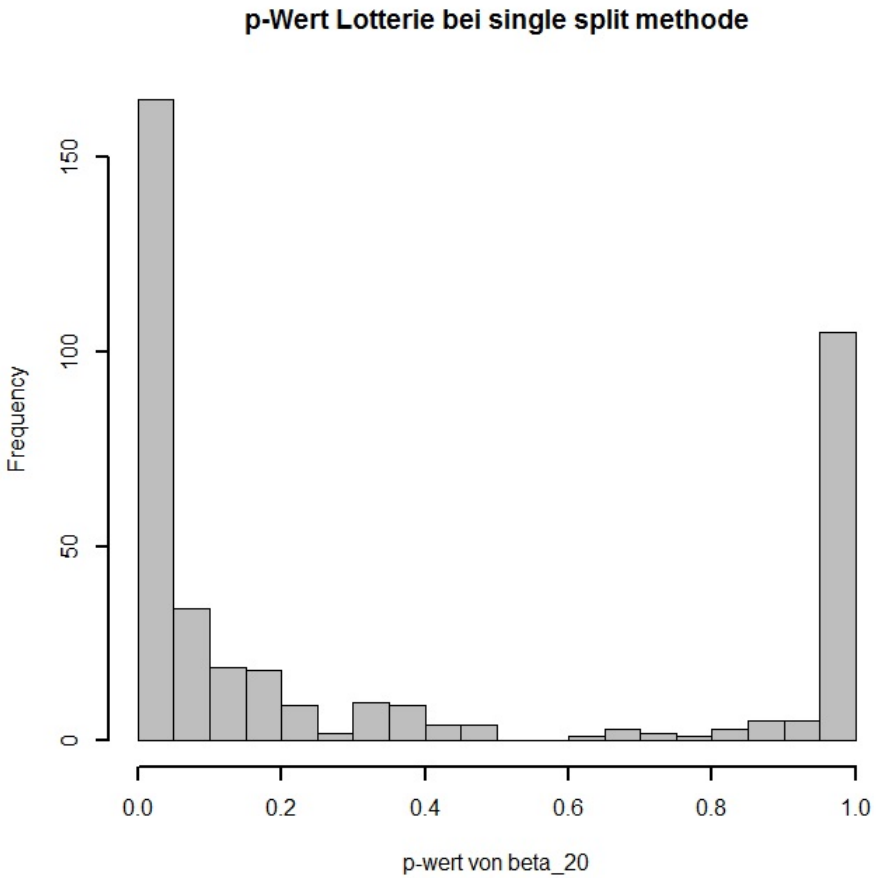
Für γ definiere:

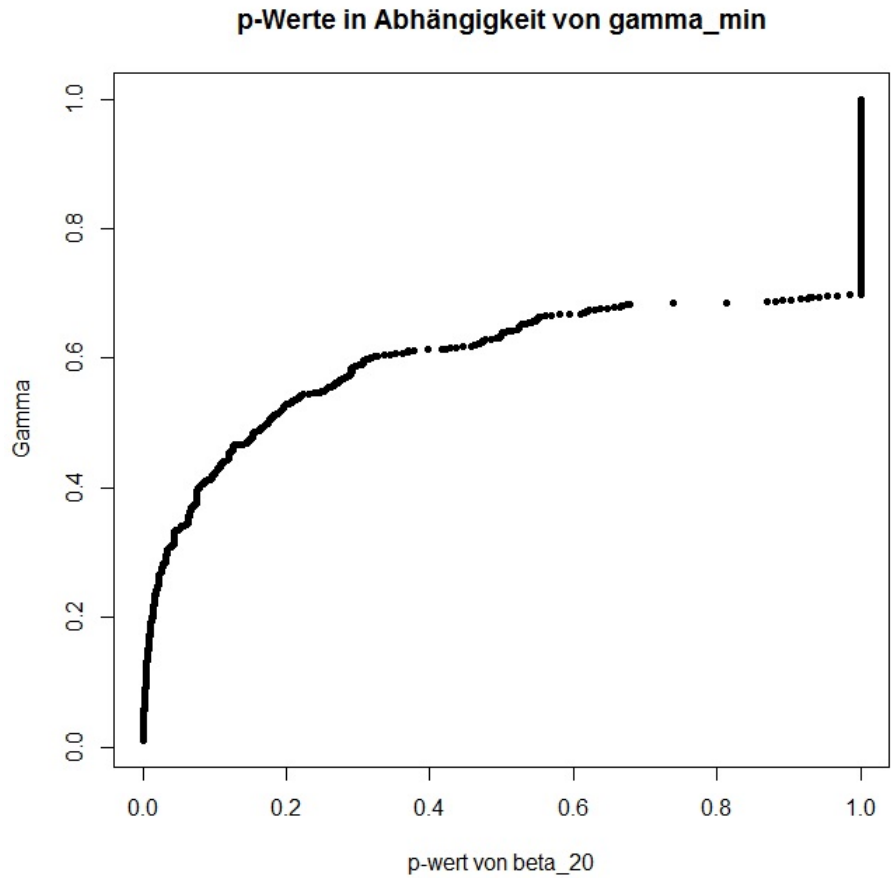
$$Q_j(\gamma) = \min\{q_\gamma\left(\{\frac{\widetilde{p}^b_{corr,j}}{\gamma}; b = 1, ..., B\}\right), 1\}$$

mit $q_\gamma(\cdot)$ als empirische γ Quantilfunktion.
Wir wählen eine angemessen Wert indem wir eine untere Schranke $\gamma_{min} \in (0, 1)$ wählen und definieren:

$$p_j := \min\left\{(1 - \log(\gamma_{min})) \inf_{\gamma \in (\gamma_{min}, 1)} Q_j(\gamma), 1\right\}$$

$1 - \log(\gamma_{min})$ kontrolliert die Suche nach einem optimalen Gamma.


p-Wert Lotterie bei single split methode
Frequency
0
50
100
150
0.0
0.2
0.4
0.6
0.8
1.0
p-wert von beta_20


p-Werte in Abhängigkeit von gamma_min
Gamma
0.0
0.2
0.4
0.6
0.8
1.0
p-wert von beta_20

$$\widehat{S}_{multi-split}(\alpha) := \{j : p_j \leq \alpha\}$$
$$V_{multi-split}(\alpha) := |\widehat{S}_{multi-split}(\alpha) \cap S_0^c|$$

Satz 5

Sei $Y = X\beta + \varepsilon$ lineares Model mit festem Design und standardnormalverteiltem Fehler. Sei zudem B die Anzahl der multi-split Durchführungen fest. Dann gilt für alle $\gamma_{min} \in (0, 1)$

$$\limsup_{n \to \infty} \mathbb{P}[V_{multi-split}(\alpha) > 0] \leq \alpha$$

Beispieldesign

Wir konstruieren unser Beispiel mit folgenden Parametern:

Sei $n = 60$, $p = 200$ und

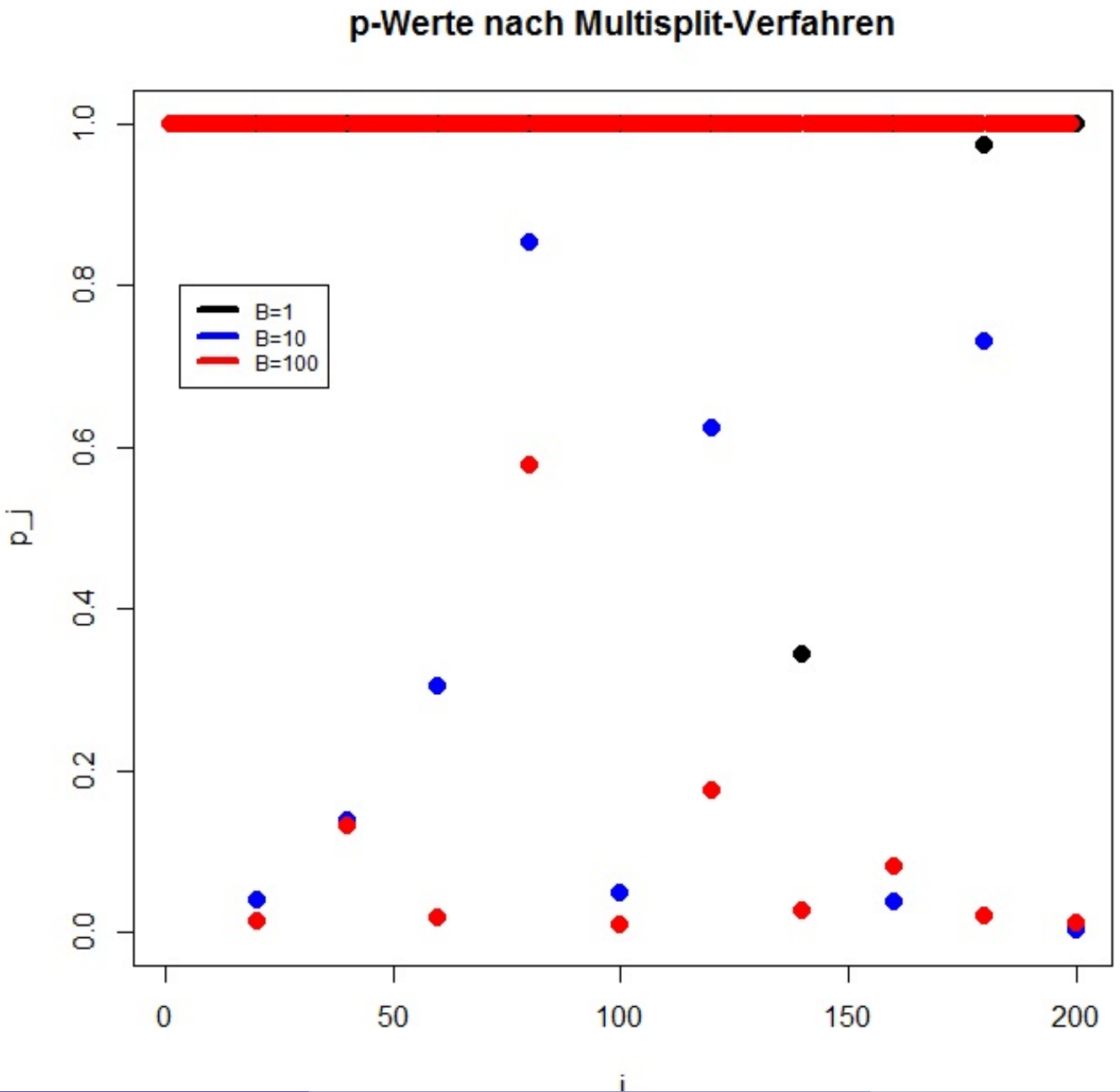
$$Y = X\beta + \varepsilon$$

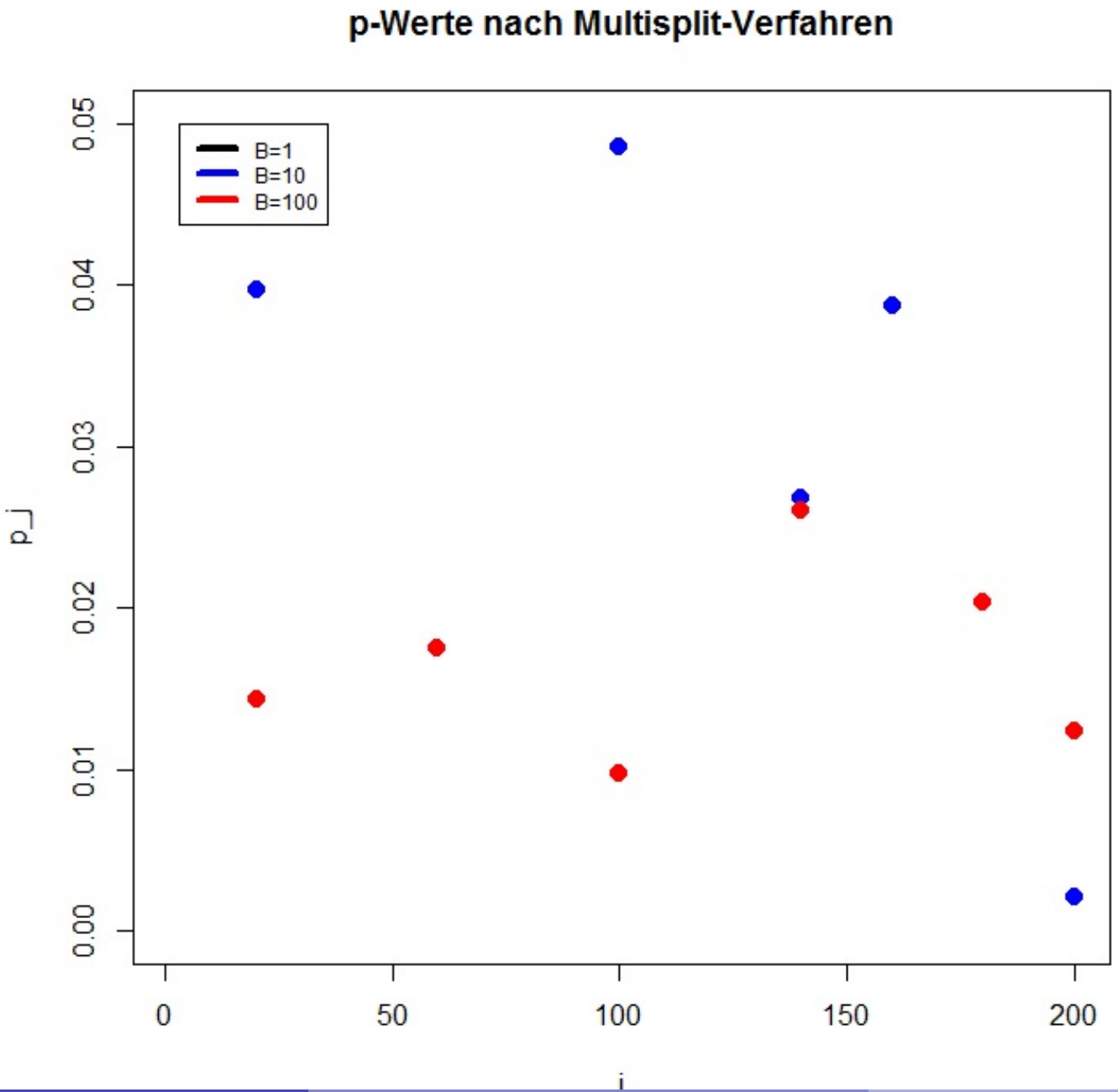
wobei $Y \in \mathbb{R}^{n\times 1}$, $\varepsilon \sim \mathcal{N}_n(0, \mathbb{1}) \in \mathbb{R}^{n\times 1}$ und $X \in \mathbb{R}^{n\times p}$ mit Spalten

$$X_i \sim \mathcal{N}_p(0, \Sigma) \text{ i.i.d. mit } \Sigma = \begin{pmatrix} 1 & 0.5 & \cdots & 0.5 \\ 0.5 & 1 & & \vdots \\ \vdots & & \ddots & 0.5 \\ 0.5 & \ldots & 0.5 & 1 \end{pmatrix}$$

sowie $\beta \in \mathbb{R}^{p\times 1}$ mit $\beta_j = \begin{cases} 3 & \text{falls } j = 0 \bmod 20 \\ 0 & \text{sonst} \end{cases}$

Vergleiche Beispiel 10.1 aus [BvdG11]

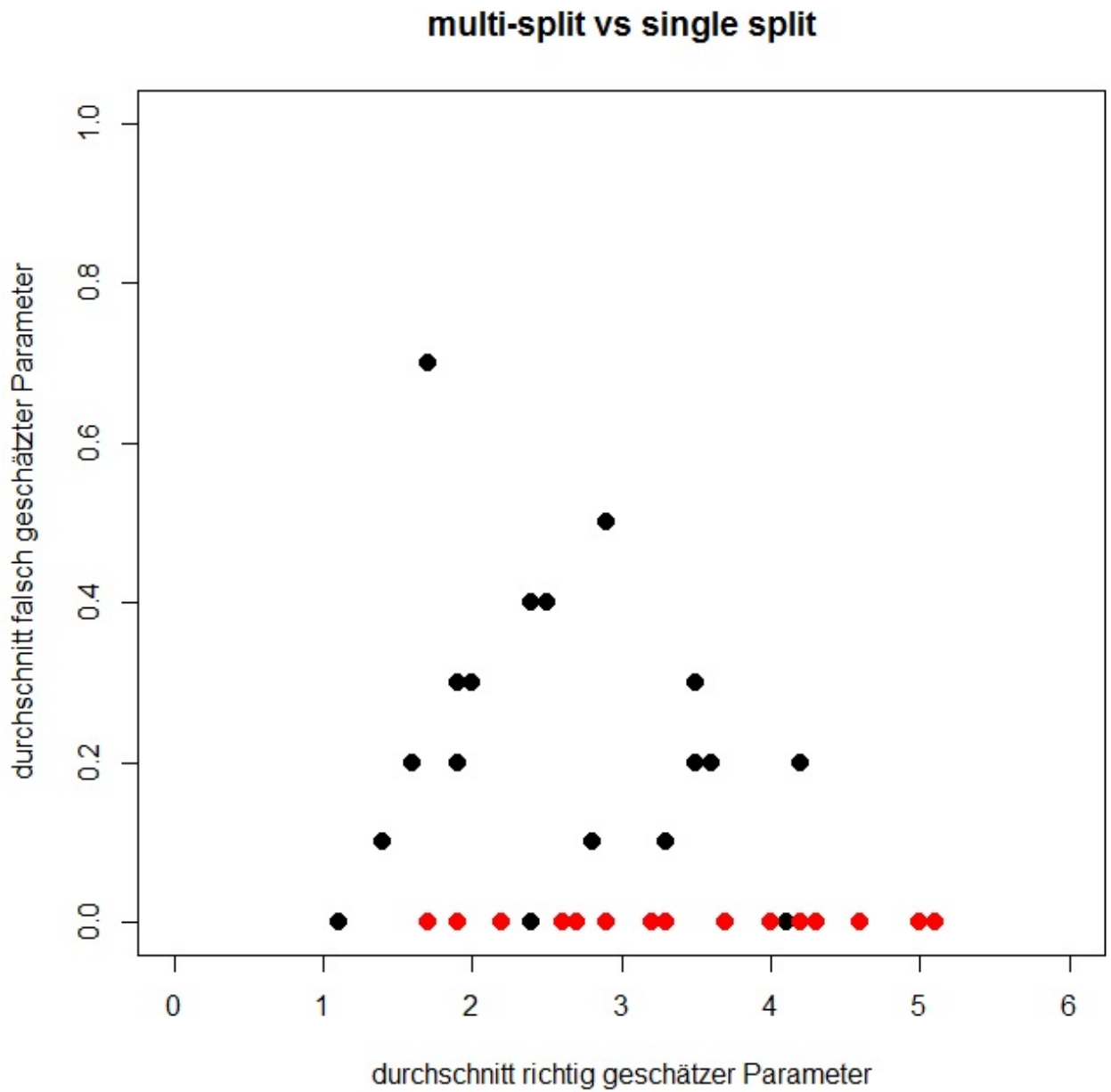

p-Werte nach Multisplit-Verfahren
B=1
B=10
B=100
p_j
i
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200


p-Werte nach Multisplit-Verfahren
B=1
B=10
B=100
p_j
0.00
0.01
0.02
0.03
0.04
0.05
0
50
100
150
200
i

single vs. multi split Verfahren

Untersuche wie oft das Verfahren korrekte bzw. falsche p-Werte ausgibt.
$L = 20$ Durchläufe

- Erstelle ein Modell X wie zuvor, $\varepsilon \sim \mathcal{N}_n(0, 10 \cdot \mathbb{1})$ und β mit 6 zufälligen Einträgen $= 4$
- Maximal 12 Einträge im Schätzer $\neq 0$
- Führe 10 mal das single split Verfahren und 10 mal das multi split Verfahren mit $B = 10$ durch.
- Bilde Mittelwert über Anzahl richtiger bzw. falscher Parameter mit p-Werte $< 30\%$.


multi-split vs single split
durchschnitt falsch geschätzter Parameter
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
6
durchschnitt richtig geschätzer Parameter

$B = 200$, $p_{7375} = 0.1344832$ sowie $p_{7069} = 0.745386$.