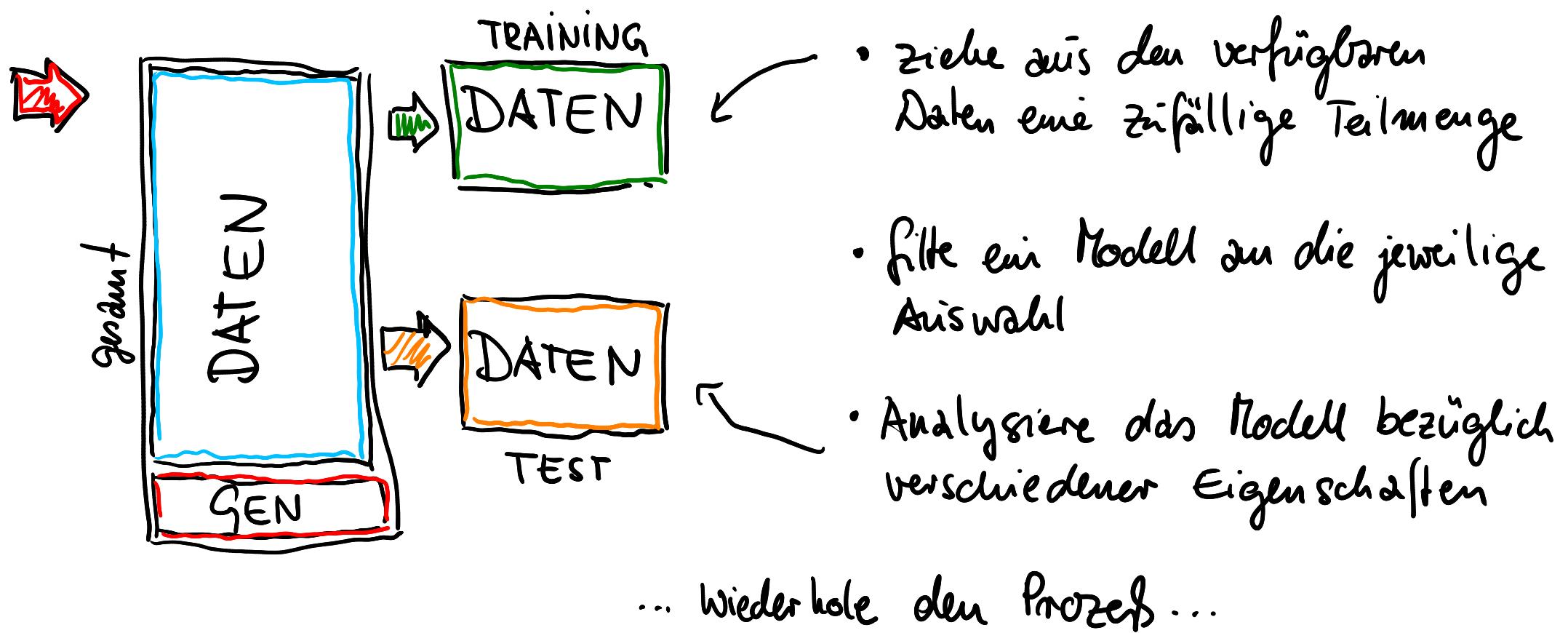


In der modernen Statistik und beim maschinellen Lernen spielen Resampling-Methoden eine große Rolle.

Was versteht man unter Resampling?



NACHTEIL: kann bei vielen Versuchen und großer Komplexität des Modells sehr rechenaufwendig werden

VORTEIL: erlaubt Einsichten in die statistischen Eigenschaften der Modellfehler, die Stabilität der Modellanzapfung etc.

Zwei wichtige Methoden:

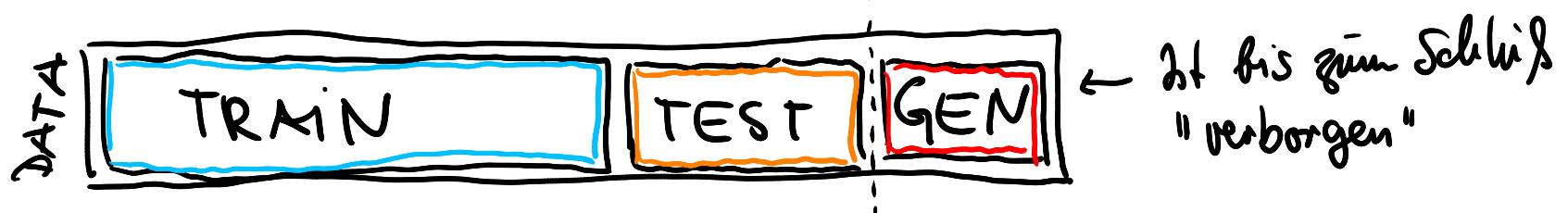
① CROSS-VALIDATION

② BOOTSTRAP

3.1) Cross-Validation

178

Schon gesehen: Daten werden üblicherweise aufgeteilt



Die Aufteilung ist normalerweise nicht a priori vorgegeben, sondern muss selbst vorgenommen werden.

Je nach Wahl der Trainings- und Testmenge unterscheiden sich auch die berechneten Maße für die Performance eines bestimmten Modells (z.T. nicht wenig !!)

Wir betrachten nun die n Beobachtungen im



(**GEN** ist außen vor, wird nur am Ende ausgewertet)

3.1.1) Validation Sampling

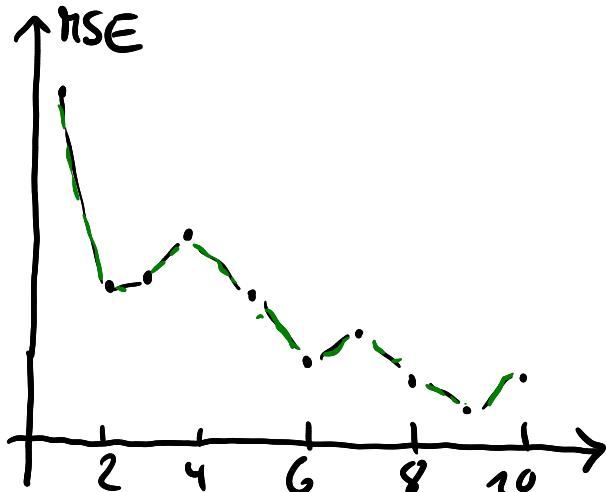
179

Überlegung: für die Anpassung und das Bewerten eines Modells stehen uns die n Beobachtungen im TRAIN + TEST zur Verfügung

→ 1 spezielle Aufteilung liefert 1 spezielles Ergebnis

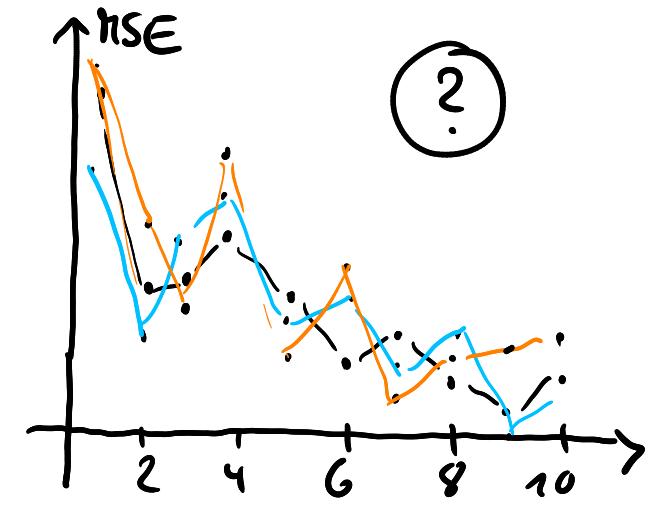
Bsp.: Regression mit Polynom

Jede Aufteilung liefert für Polynome mit unterschiedlichem Grad eine ganz spezielle Fehlercharakteristik



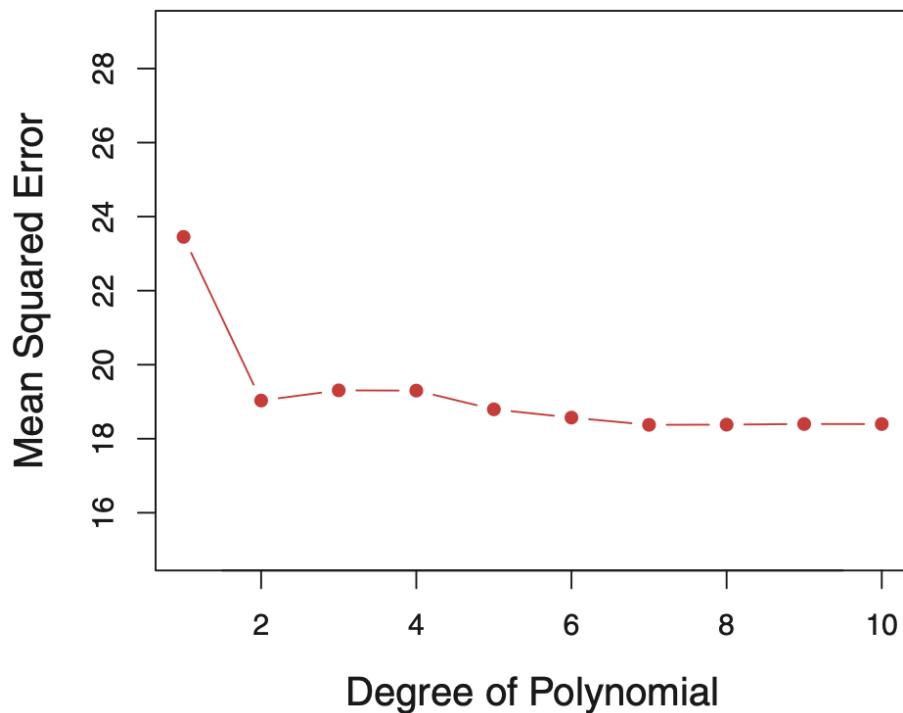
→ die Struktur sollte i.A. ähnlich aussehen

kaum aber quantitativ doch sehr unterschiedlich sein (z.B. Position min(MSE))

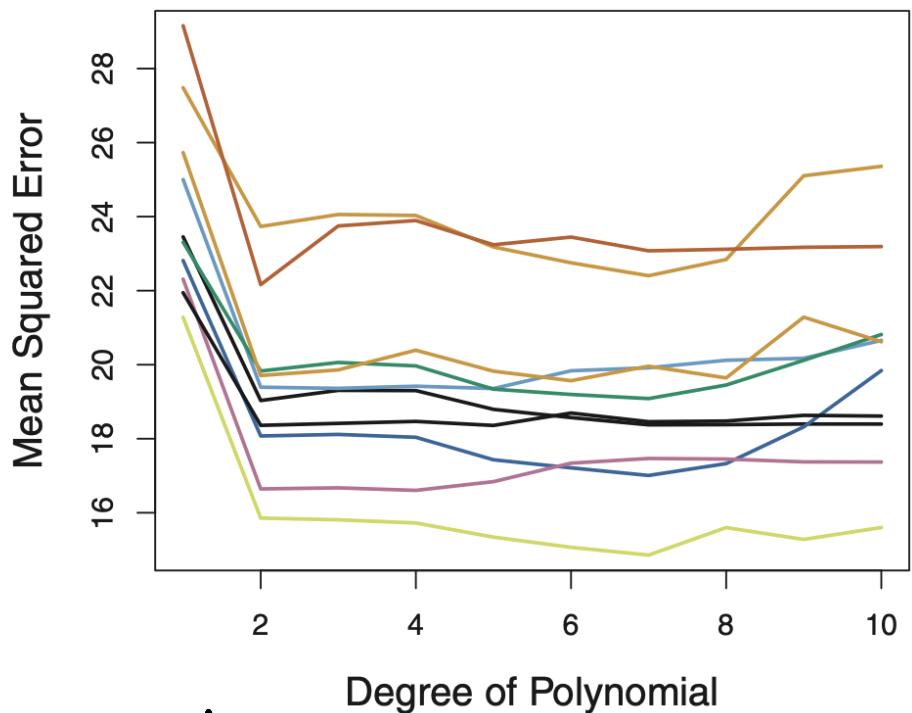


Bsp.: (Tibshirani (2014)) : Polynomiale Regression mit Polynomen von unterschiedlichen Grad (Auto-Datensatz)

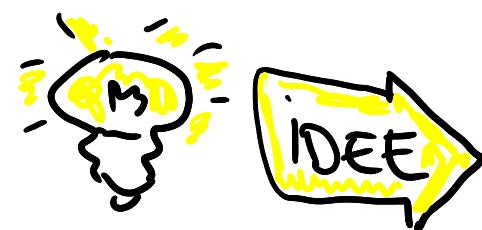
(180)



Validierungsfehler für einen einzigen Split im Training und Test



10 zufällige Splits liefern eine ähnliche Struktur in Bezug auf den Grad des verwendeten Polynoms, aber quantitativ sehr unterschiedliche Fehler



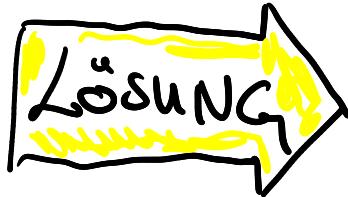
Die Modelle, die aus verschiedenen Splits der Daten resultieren können als Ensemble verstanden werden!

→ Entscheidungen (z.B. über die beste Modellkonfiguration) können über den Mittelwert oder z.B. auch als "Mehrheitsentscheidung" getroffen werden.



Was tun, wenn die Ergebnisse zu unterschiedlich sind?
Was ist dann die beste Wahl??

Beobachtung: beim Training lassen wir immer jede lange Datenpunkte nicht mitspielen (besonders kritisch bei kleinen Datensätzen)



CROSS-VALIDATION

3.1.2) Leave-One-Out Cross-Validation (LOOCV)

 ZIEL möglichst viele Daten für das Training verwenden!

Simple Lösung: n Datenpunkte $\{(x_1, y_1), \dots, (x_n, y_n)\}$ im TRAIN + TEST

- ① Lege (x_i, y_i) zur Seite
- ② fitte Modell an $\{(x_1, y_1), \dots, (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), \dots, (x_n, y_n)\}$
- ③ Validiere auf $(x_i, y_i) \Rightarrow \text{RMSE}_i$
- ④ wiederhole ①-③ für (x_i, y_i) als Validierungspunkt $\Rightarrow \text{RMSE}_i$

Wiederholt man ①-③ für alle n Datenpunkte, so erhält man $\text{RMSE}_1, \dots, \text{RMSE}_n$

 $\text{MSE}_1, \dots, \text{MSE}_m$ können als Realisierungen der Zufallsvariable MSE aufgeführt werden!

Wir interessiert der Erwartungswert der ZV MSE : $E(\text{MSE})$

Die LOOCV-Schätzung für $E(\text{MSE})$ bekommt man durch die Berechnung des Mittelwerts:

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^n \text{MSE}_i$$

Vorteile: ① mehr Daten zum Trainieren (\Rightarrow Testfehler wird nicht so stark überschätzt)

② keine Zufallskomponente durch Datensplit

Ein Datensatz liefert einen $CV_{(n)}$ für ein Modell!

Nachteil: Das Modell muss n -mal angepasst werden und n -mal ausgewertet werden.

Bei großen Datensätzen & komplexen Modellen kann das problematisch werden ...

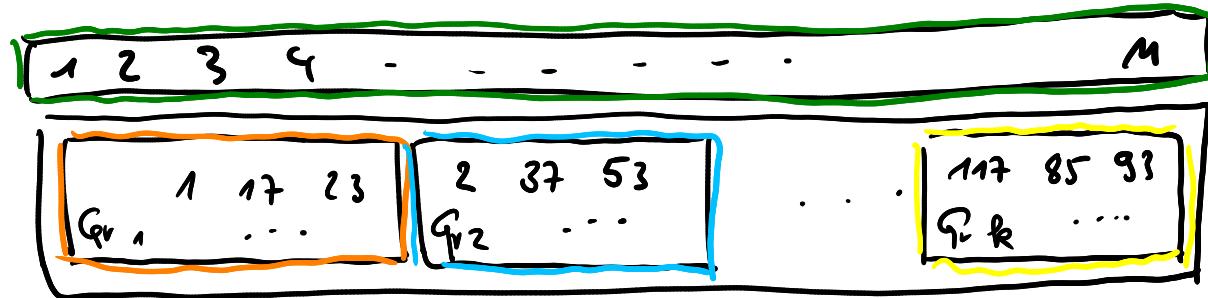


Kann man den Rechenaufwand minimieren, ohne zu viel Zufall in die Bewertung hinein zu bringen?

⇒ ... Anpassung der Gründidee

3.1.3) k-fold Cross-Validation

... wieder n Datenpunkte $\{(x_1, y_1), \dots, (x_n, y_n)\}$



- ① Teile die n Datenpunkte in k zufällig gewählte Gruppen gleicher Größe auf (k-folds)
- ② benütze Gruppen $2, 3, \dots, k$ zum Training
- ③ validiere auf Gruppe 1 $\Rightarrow \text{MSE}_1$
- ④ wiederhole ② + ③, so dass alle k Gruppen einmal zum Testen verwendet werden $\Rightarrow \{\text{MSE}_1, \text{MSE}_2, \dots, \text{MSE}_k\}$

Insgesamt:

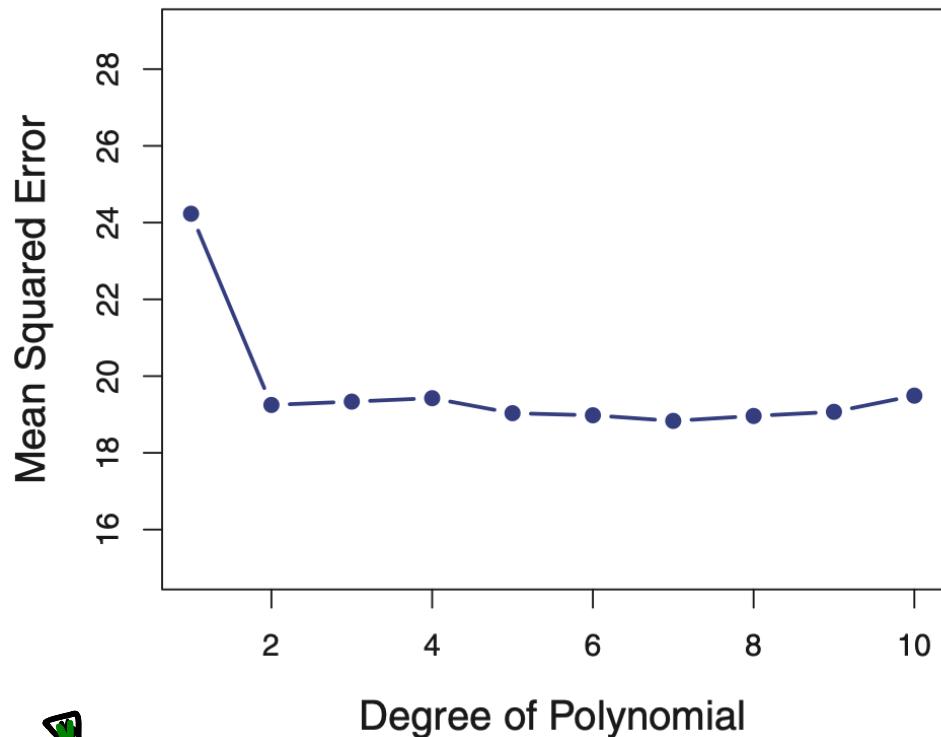
$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^k \text{MSE}_i$$

Bsp.: Kib - Datensatz

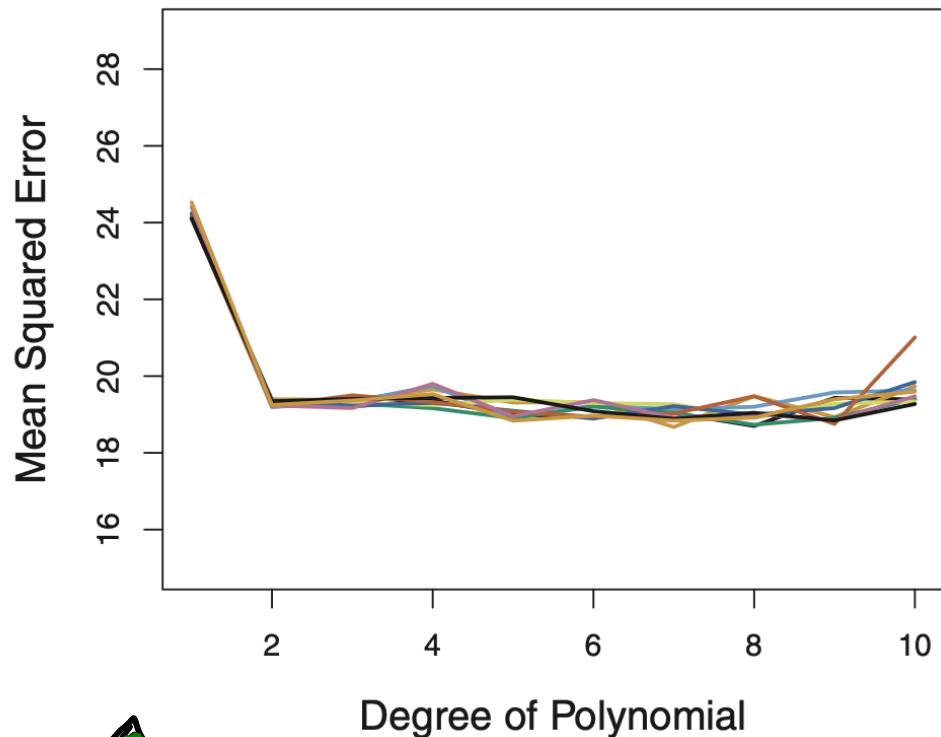
Polynomiale Regression

186

LOOCV



10-fold CV



Ergebnis der LOOCV auf den
gesamten Daten

⇒ genau 1 Fehlerkurve
(deterministisch)



wiederholte 10-fold CV mehr mal
mit jeweils verschieden belegter
Aufteilung der folds

→ g ganz leicht unterschiedliche Ergebnisse
fast kein Unterschied zu LOOCV!

- Bemerkung:
- Für eine stabile Schätzung von $E(\text{MSE})$ benützt man üblicherweise $k=5$ ($k=10$)
 $\Rightarrow 5 \times (10 \times)$ Modelle filtern \mathcal{S} auswerten
 - LOOCV ist ein Spezialfall mit $k=n$, aber maximal rechenaufwendig ↴

Überlegungen zum Bias-Variance Dilemma

Die Wahl einer einzigen (zufälligen) Validierungsmenge führt i.A. zu einer deutlichen Überschätzung des Fehlers!

- LOOCV liefert eine nicht-„gebiaste“ Schätzung des Testfehlers
(jede Trainingsmenge hat ja $(n-1)$ Teilglieder, für 1 Datensatz $\Rightarrow 1$ LOOCV-Wert)
- die k -fold CV schneidet natürlich schlechter ab im Bezug auf den Bias, ist aber immer noch besser als mit nur einer einzigen Auswahl.

BIAS - REDUKTION

→ LOOCV gewinnt natürlich
... aber die Varianz ist ja auch noch wichtig !!

wie sieht damit aus?

- LOOCV: wir trainieren n Modelle mit fast identischer Trainingsmenge (nur 1 Pkt. verschieden)
⇒ Outputs sind hochgradig (positiv) korreliert
- k-fold CV: stärkere Unterschiede in den Trainingsdaten
⇒ deutliche Reduktion der Korrelation der Outputs

Was bedeutet das?

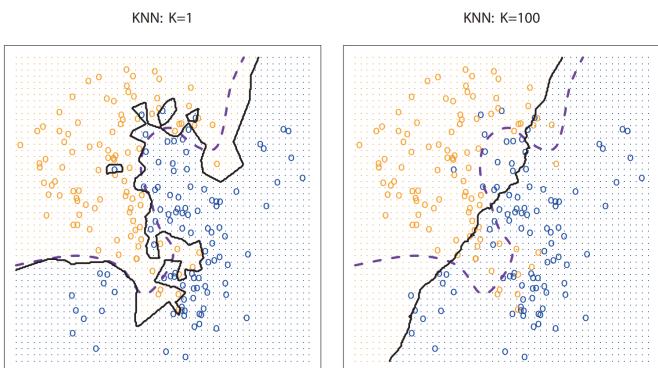
STATISTIK

185

Der Mittelwert vieler stark korrelierter Größen hat eine größere Varianz als der von solchen, die nicht so stark korrelieren.

- ⇒ Die Schätzung des Modellfehlers mit Hilfe der LOOCV hat eine größere Varianz als die der k-fold CV. (je kleiner k , desto kleiner...)
- ⇒ Die Wahl von $k=5$ ($k=10$) liefert bei diesem trade-off (empirisch) die besten Ergebnisse!

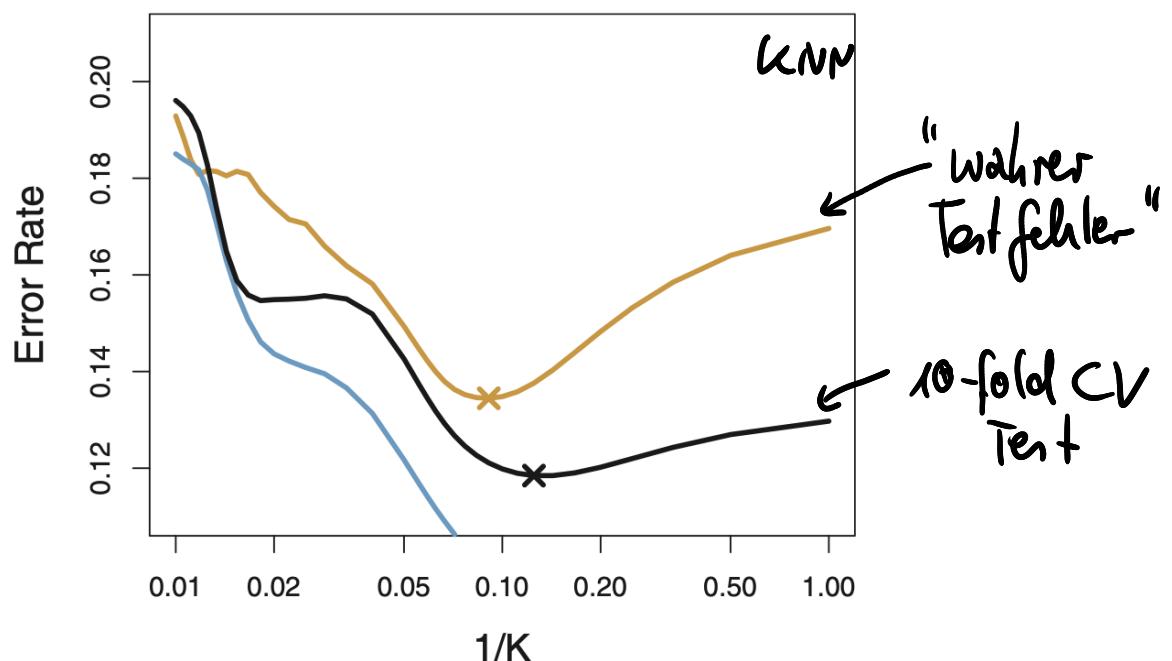
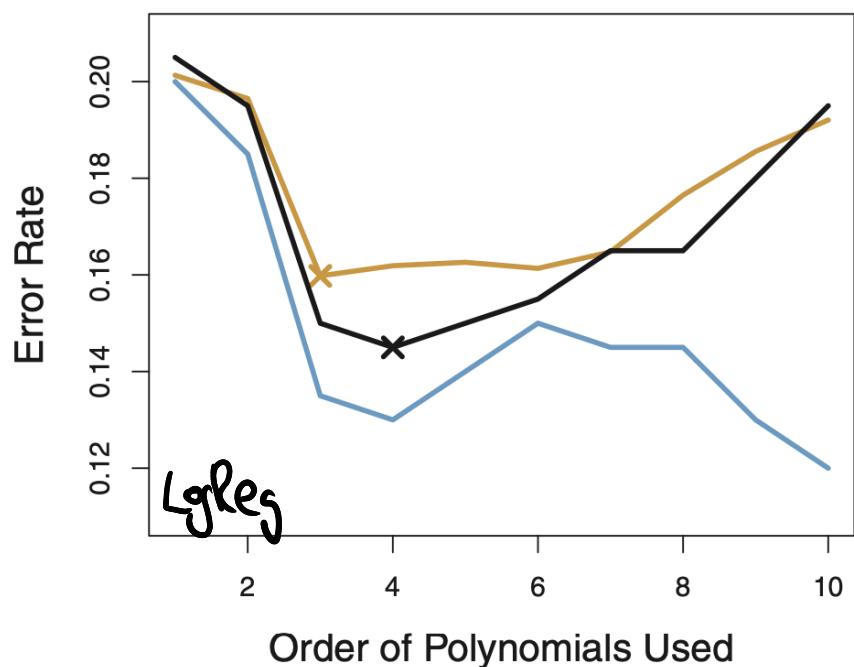
Bsp.: bekannte Bayes-Décision Boundaries



250

- Lösung mit
- multiplikativer Logistische Funktion
(Polynomiale Features)
 - KNN - variiere $W^K(x)$
Metaparameter k

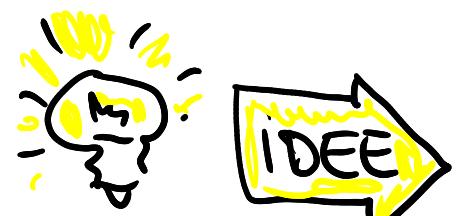
↓
über die Cross Validation kommen wir den
bestmöglichen Meta-Parametern schon sehr nahe !!



3.2) Bootstrap

In der (klassischen) Statistik setzt man Bootstrap - Verfahren ein um die Unsicherheit quantitativ abzuschätzen, die mit Schätzfunktionen T_θ (\hat{ZV}) verbunden sind.

⇒ Auch nützlich bei ML - Verfahren (hier wird ja ein Prädiktor geschätzt)



Betrachte eine $ZV X$ mit Verteilung $F_\theta(x)$

Bsp.: Schätzung für den Erwartungswert $E(x)$

wahre, unbekannte Verteilung

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$$

für Stichprobe (X_1, \dots, X_n) mit Realisierung $(x_1, \dots, x_n) \in \mathbb{R}^n$

? aus Daten



Wie sicher \ unsicher ist die Schätzung?

Wenn wir die echte (unbekannte) Verteilung kennen würden könnten, wir beliebig viele Samples der Länge n (simulierte Stichproben) aus dieser Verteilung generieren und daraus jeweils $\hat{\mu}_i$ berechnen: $\hat{\mu}_1, \hat{\mu}_2, \dots$

⇒ Die Varianz bzw. die Standardabweichung der $\hat{\mu}_i$ liefert dann eine gute Abschätzung für die Unsicherheit.



In der Realität können wir so nicht vorgehen, weil wir $F_\theta(x)$ ja nicht kennen.

BOOTSTRAP simuliert diese Vorgehensweise auf Basis der vorliegenden Beobachtungen ...

AUSGANGSLAGE: n Datenpunkte $x_1, \dots, x_n \rightarrow$ Schätzung für \hat{Q} ist das Ziel

193

- • wähle zufällig k Datenpunkte aus ("Ziehen mit Zurücklegen")
 - berechne erste Schätzung \hat{Q}_1 aus diesem Sample
 - ziehe neue Menge mit k Punkten $\rightarrow \hat{Q}_2$
 - wiederhole den Prozess N mal $\Rightarrow \hat{Q}_1, \hat{Q}_2, \dots, \hat{Q}_N$
 - berechne die Standardabweichung der $\hat{Q}_1, \dots, \hat{Q}_N$ mit

$$SE_N(\hat{Q}) = \sqrt{\frac{1}{N-1} \sum_{j=1}^N (\hat{Q}_j - \bar{\hat{Q}})^2}$$

Standardfehler

$$\bar{\hat{Q}} = \frac{1}{N} \sum_{j=1}^N \hat{Q}_j$$

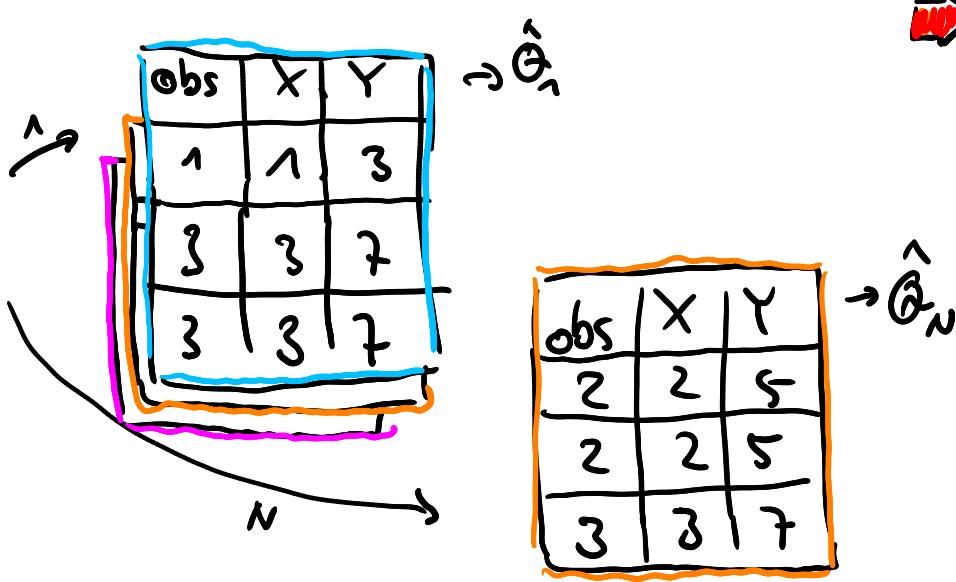
kaum als Standardfehler der Schätzung interpretiert werden

Bemerkung:

- da bei Bootstrap - Verfahren "mit Zurücklegen" gezogen wird können einzelne Beobachtungen öfter auftauchen
- bei multivariaten Datensätzen $(X_1, X_2, \dots, X_p, Y)$ ist ein Datenpunkt gegeben als $(x_{1i}, x_{2i}, \dots, x_{pi}, y_i)$ \leftarrow nicht "mixen"
- die Anzahl der Ziehungen N kann wegen Zahlen "mit Zurücklegen" auch bei kleinen Datensätzen groß werden.

obs	X	Y
1	1	3
2	2	5
3	3	7

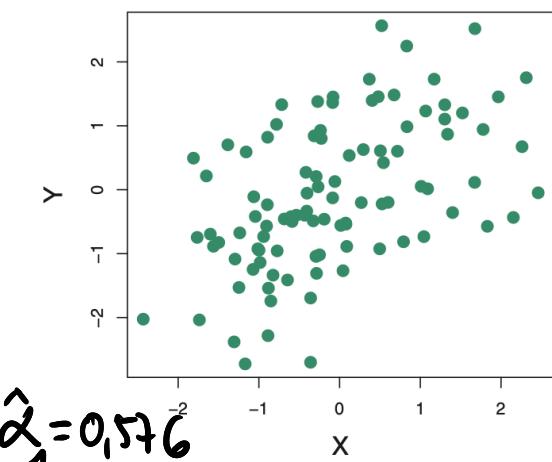
DATA



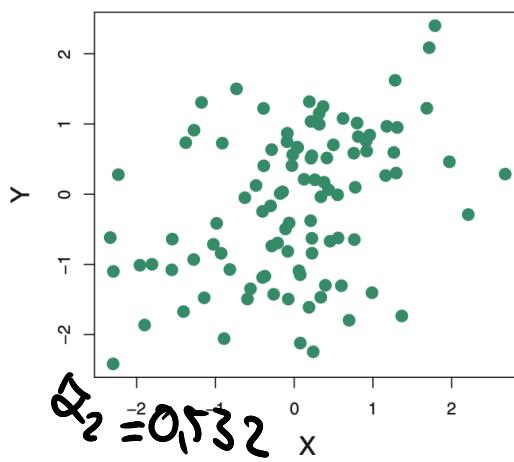
$$\hat{Q} = \frac{1}{N} \sum_{j=1}^N \hat{Q}_j$$

approximiert den wahren Q aus der Stichprobe!

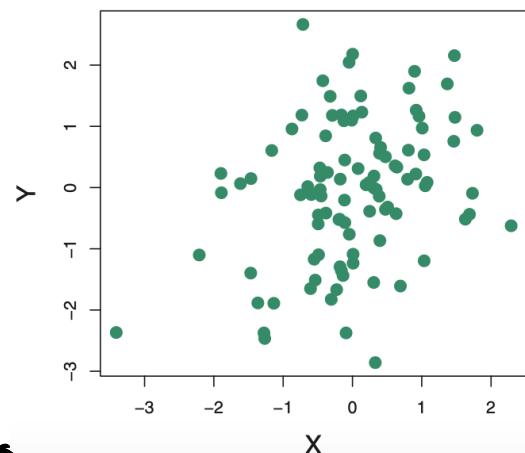
Bsp.: 4 Bootstrap Samples



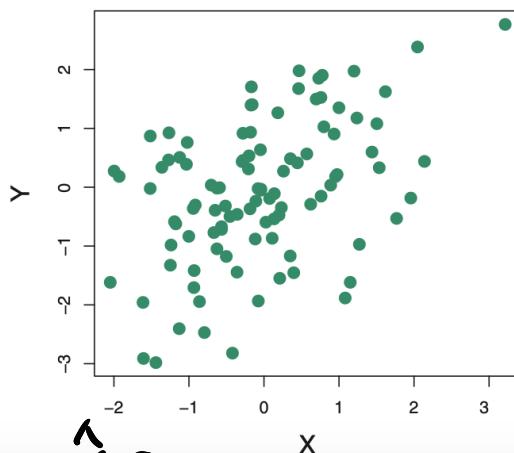
$$\hat{\alpha}_1 = 0,576$$



$$\hat{\alpha}_2 = 0,532$$



$$\hat{\alpha}_3 = 0,657$$



$$\hat{\alpha}_4 = 0,651$$

← jedes Bootstrap-Sample liefert eine eigene Schätzung für den unbekannten Parameter α

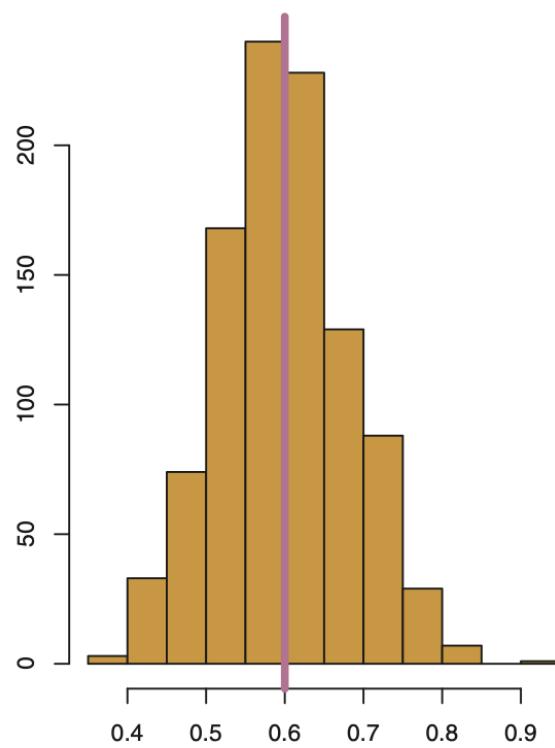
← Wiederhole den Prozess

1000 x

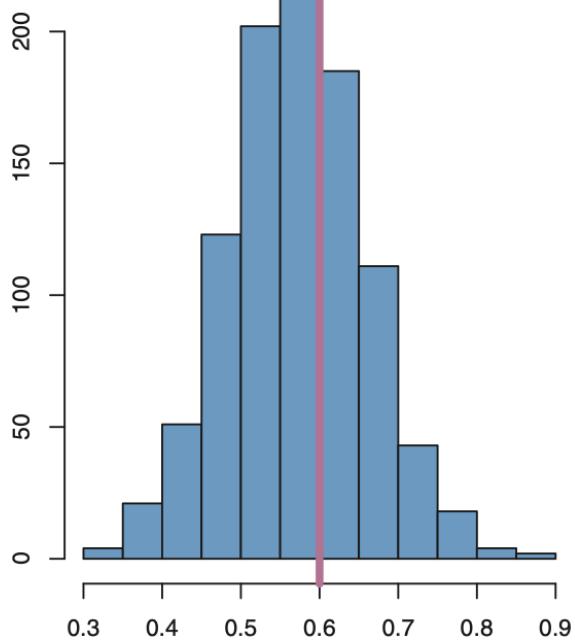
$$\Rightarrow \hat{\alpha} = \frac{1}{1000} \sum_{j=1}^{1000} \hat{\alpha}_j^1$$

aus $\hat{\sigma}_Y^1, \hat{\sigma}_X^1, \hat{\sigma}_{XY}^1$
...

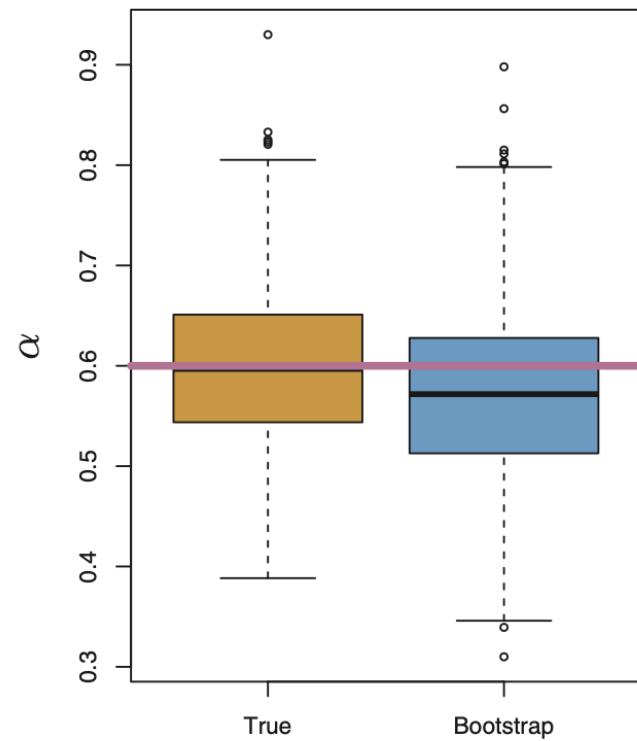
$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}^2}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}$$

 α

ein Histogramm von Schätzungen
für α aus 1000 Samples die
aus der WÄHREN (Verteilung)
gezogen wurden



Samples aus
bootstrap

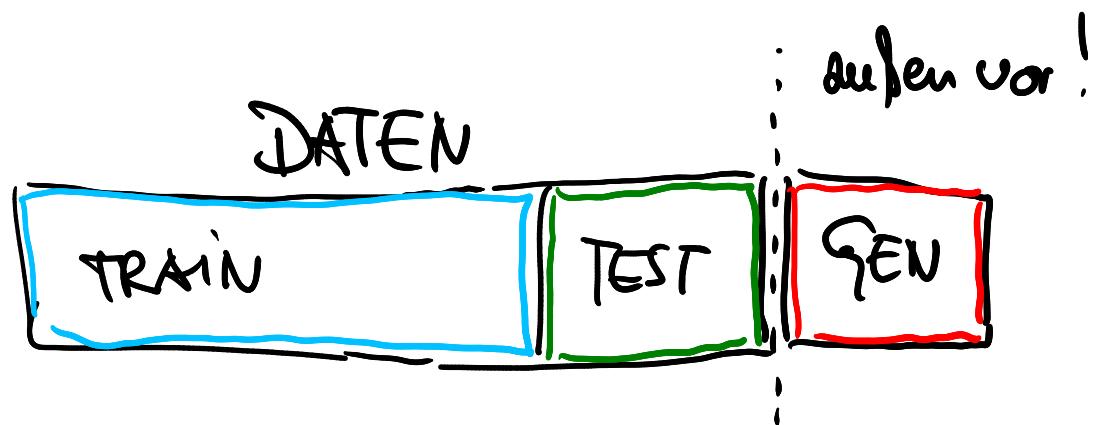


Boxplot mit
Vergleich
"wahr" vs. bootstrap

$\alpha=0.6$ ist der
wahre (unbek.) Wert

ZUSAMMENFASSUNG:

Daten Split erlaubt Anpassung von Rechenparametern durch Auswertung des Test-Fehlers



5-fold bzw. 10-fold

① Validation Set Approach

1 zufälliger Split \Rightarrow 1 MSE

- Vorteil: wenig Aufwand

- Nachteil: starke Abhängigkeit von der zuf. Aufteilung

② LOOCV

deterministischer Fehler

- Vorteil: mehr Daten, solide Fehlerstat.

- Nachteil: großer Rechenaufwand

③ k-fold CV

Schätzung $E(\text{MSE})$

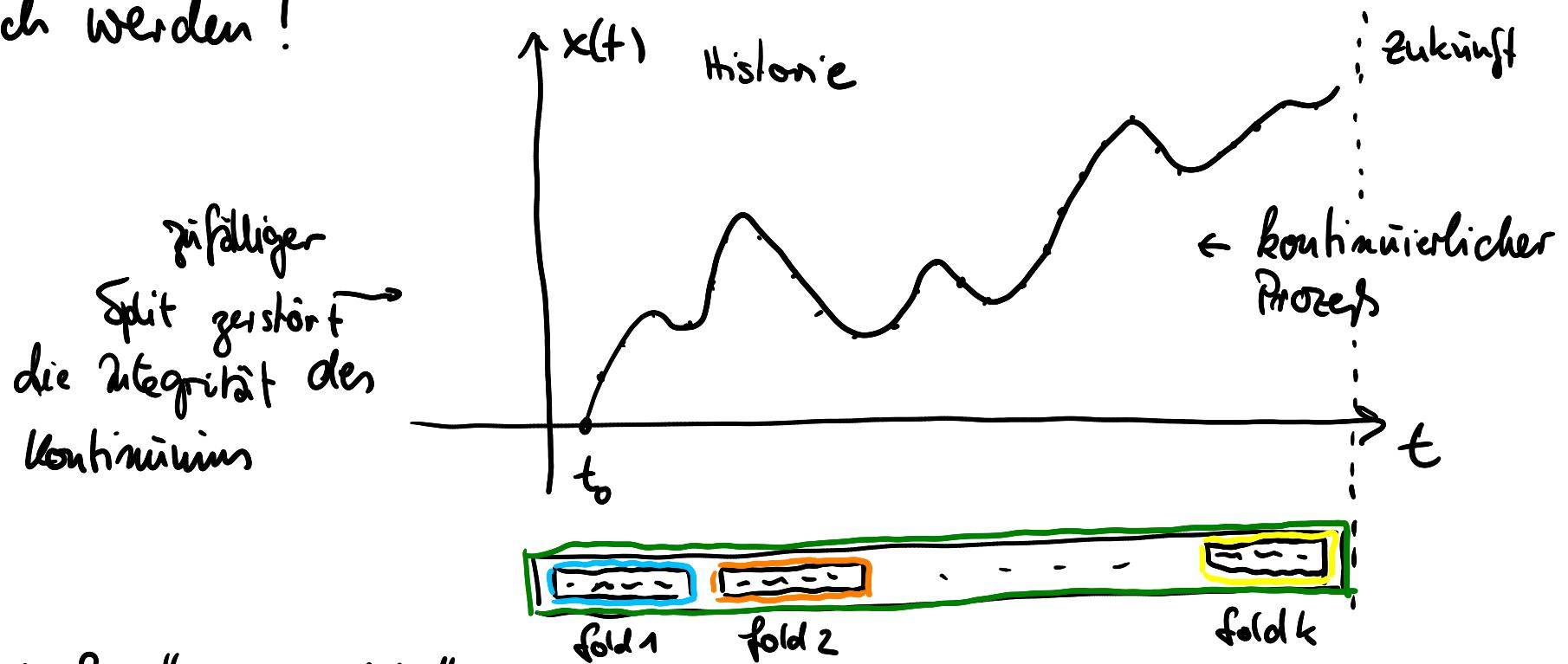
- Vorteil: stabile Bewertung der Unsicherheit

- vertretbarer Aufwand

- Nachteil: ?

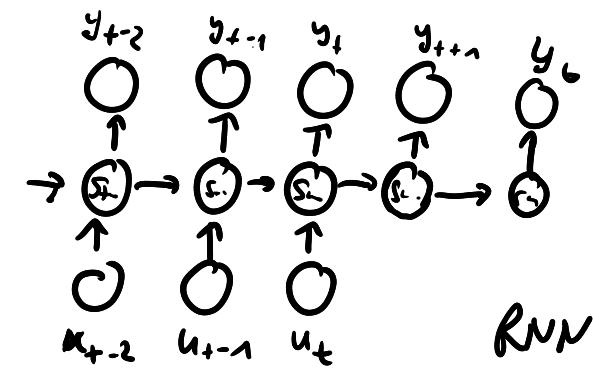
Ergebnis:

Bei "echten" Zeitreihenmodellen können die Konzepte der Cross-Validation und des Bootstrapping unter Umständen richtig problematisch werden!



... das gilt auch für "ausgeworfene" Samples !!

→ Besonders schlimm bei Modellen mit autoregressivem Charakter !

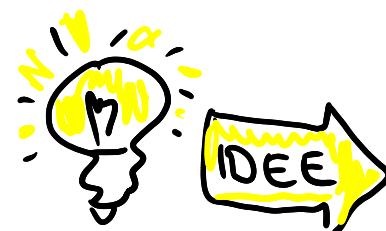


RNN

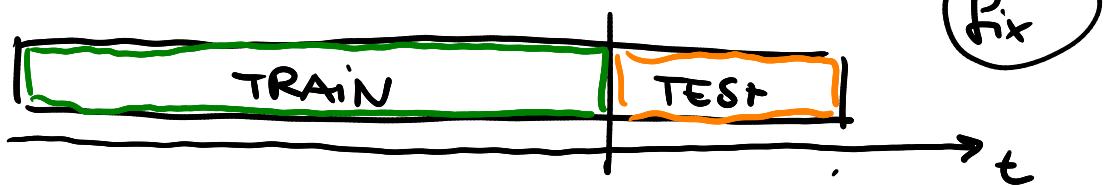
→ Lösungen sind in der Literatur unterrepräsentiert (Forschungs-gap) (199)

Ein Beispiel :

"Evaluating time series forecasting models :
an empirical study on performance estimation methods"
Cerqueira et.al. (2020), Machine Learning 105



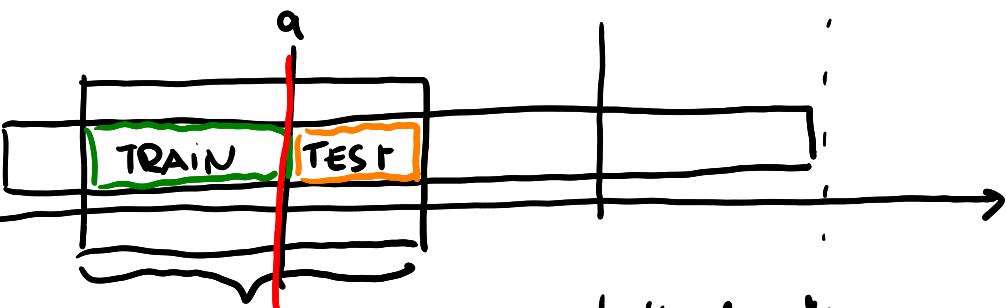
Suche Möglichkeit k-fold CV zu machen, ohne Zeitkontinuum
zu zerstören !



repeated holdout:

wähle a aus dem verfügbaren Bereich ...

- Vorgänger \Rightarrow TRAIN
- Nachfolger \Rightarrow TEST



Zusammenhang innerhalb der Menge
ist nicht gestört !

Es gibt dann
strategisch
unterschiedliche
Möglichkeiten...

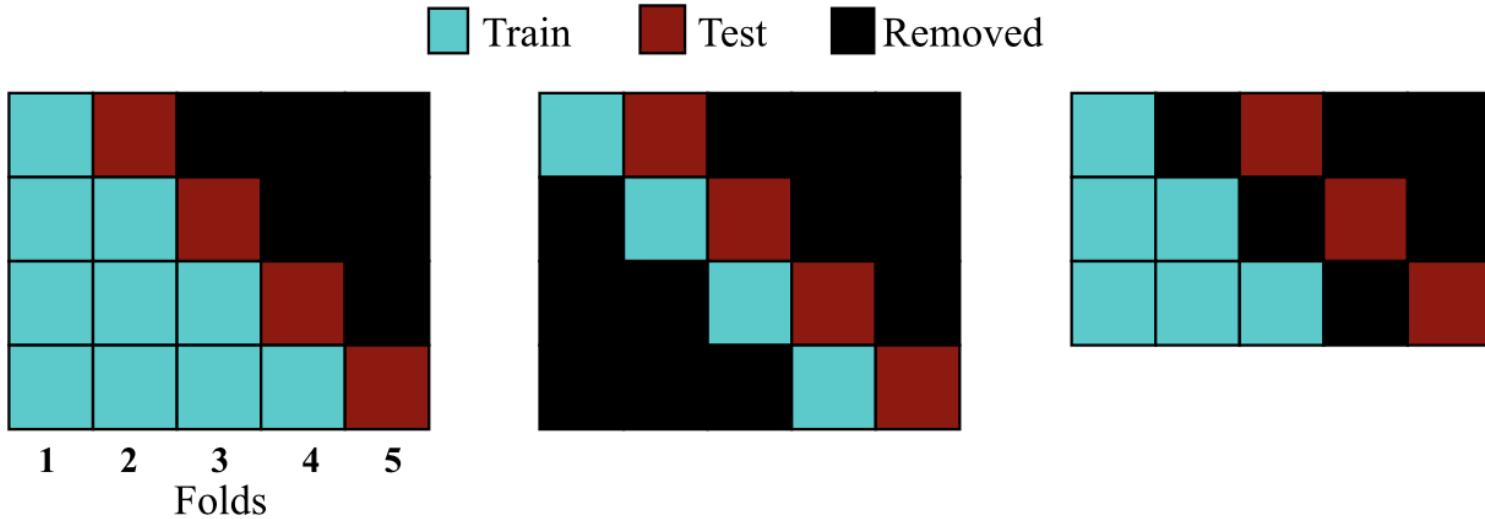


Fig. 3 Variants of prequential approach applied in blocks for performance estimation. This strategy can be applied using a growing window (left, right), or a sliding window (middle). One can also introduce a gap between the training and test sets

→ TEST-Block ist
zeitlich konsistent!

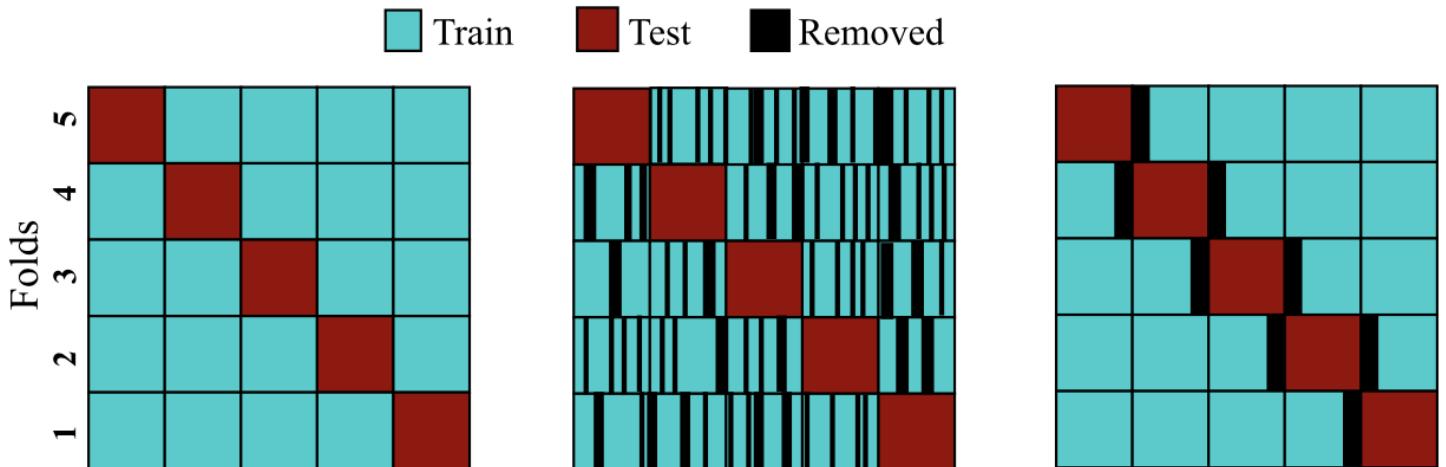


Fig. 4 Variants of cross-validation estimation procedures

Au einem Beispiel sieht man
die Unterschiede der verschiedenen Strategien

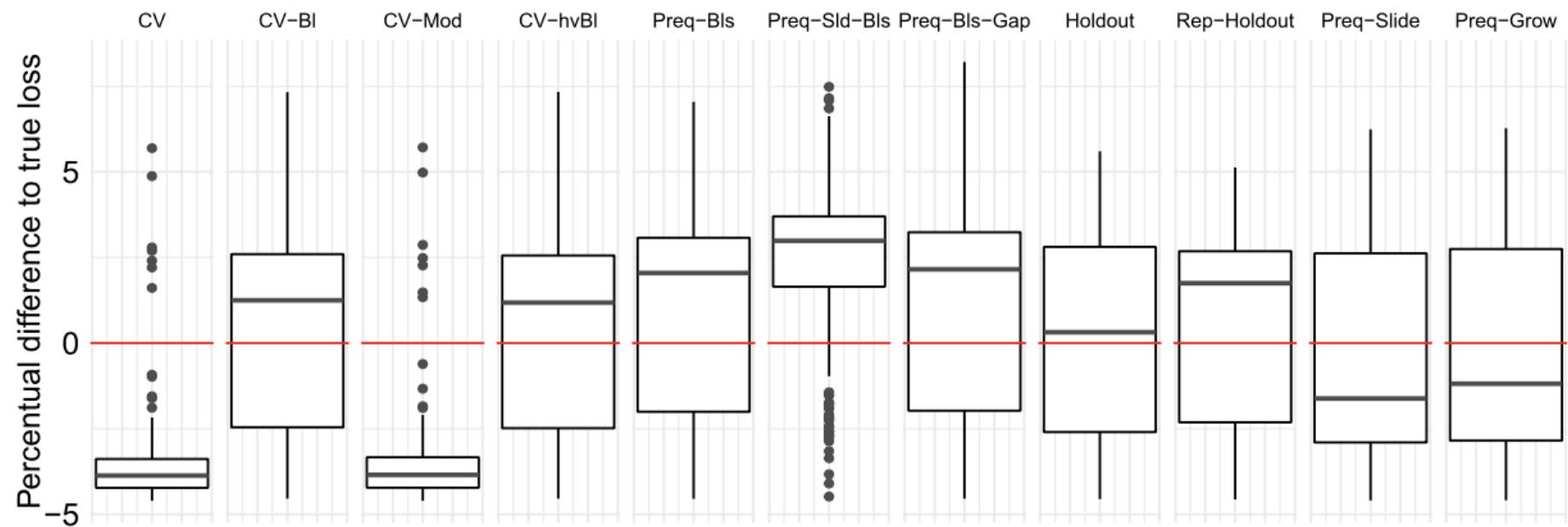


Fig. 15 Log percentage difference of the estimated loss relative to the true loss for each estimation method in the real-world case study, and using the RBR learning method. Values below the zero line represent under-estimations of error. Conversely, values above the zero line represent over-estimations of error