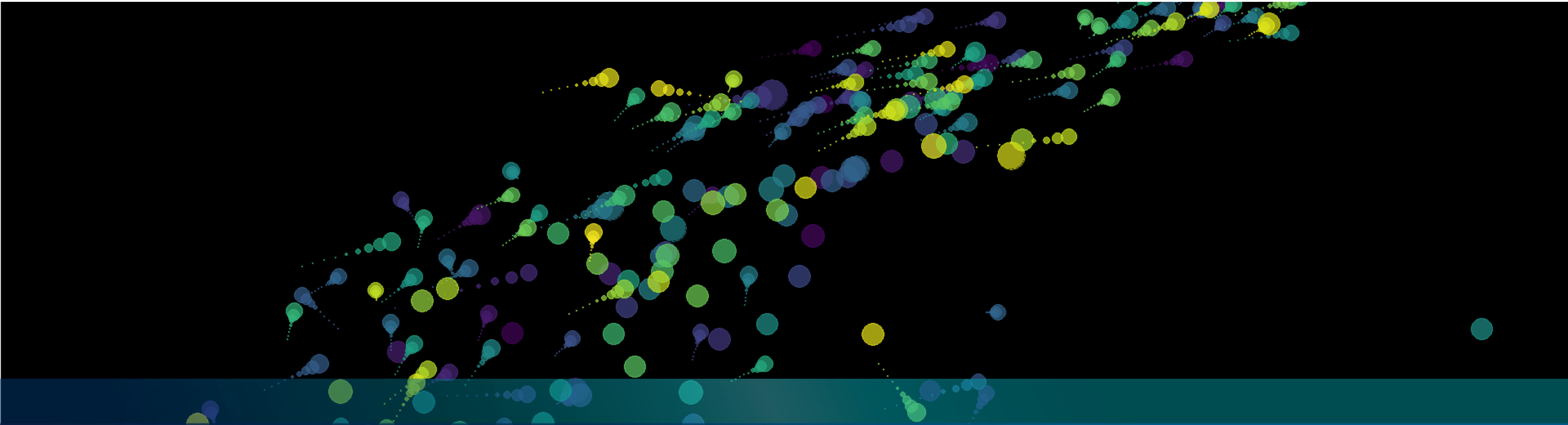
Bildungscampus

HOCHSCHULE HEILBRONN



› LASSO REGRESSION

Variablenselektion in Machine Learning

Daniel **Messner**

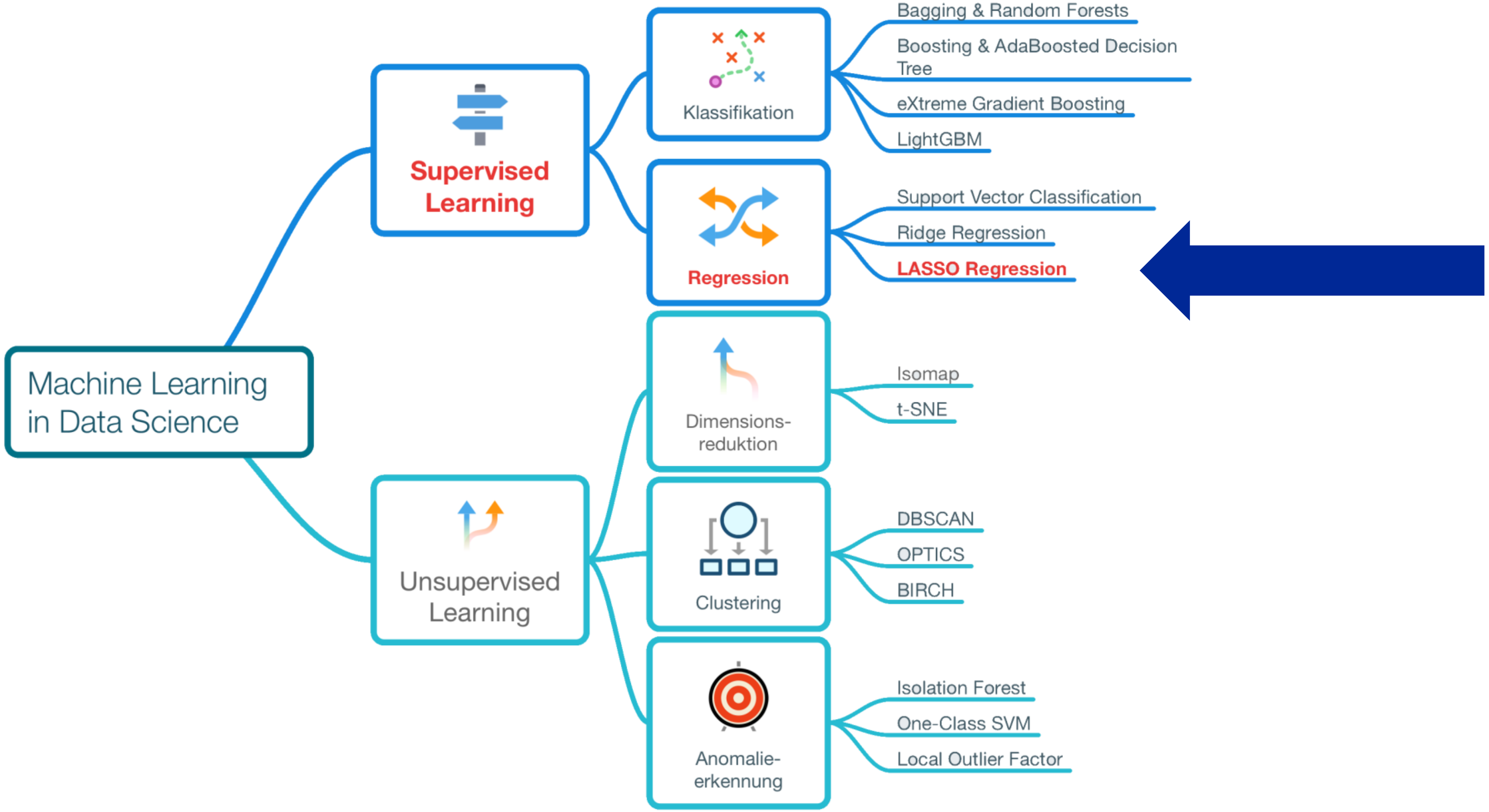
Sebastian **Kahlert**

Sebastian **Straßburg**

HOCHSCHULE HEILBRONN

GLIEDERUNG

STRUKTURELLE EINORDNUNG



HOCHSCHULE HEILBRONN

› GESCHICHTE DER LASSO REGRESSION

GESCHICHTE LASSO - REGRESSION

LASSO = **L**east **A**bsolute **S**hrinkage & **S**election **O**perator

- Operator von Robert Tibshirani im Jahr 1996 für die Parameterschätzung und gleichzeitig für die Variablen Modellselektion in der Regressionsanalyse eingesetzt [1]

- LASSO Regression bereits in der geophysikalischen Literatur von Fadil Santosa und William Symes im Jahr 1986 angewandt [2], Popularität hat jedoch Robert Tibshirani maßgeblich beigetragen

Heute kommt die LASSO Regression vor allem in der Statistik im Allgemeinen und beim maschinellen Lernen zum Einsatz.

Robert Tibshirani
Professor – Stanford University
© The Royal Society [3]

› KERNIDEE & FUNKTIONSWEISE

KERNIDEE

- Lineare Regressionsmodelle dienen zur Vorhersage einer Zielvariablen
- Dabei wird die Kausalität zwischen der Ziel- und zahlreicher Einflussvariablen überprüft
- Die Einflussvariablen können einen schwachen, starken oder gar keinen Einfluss auf die Zielvariable besitzen
- Mithilfe von zahlreichen Regressionsmethoden können die Kausalitäten untersucht werden

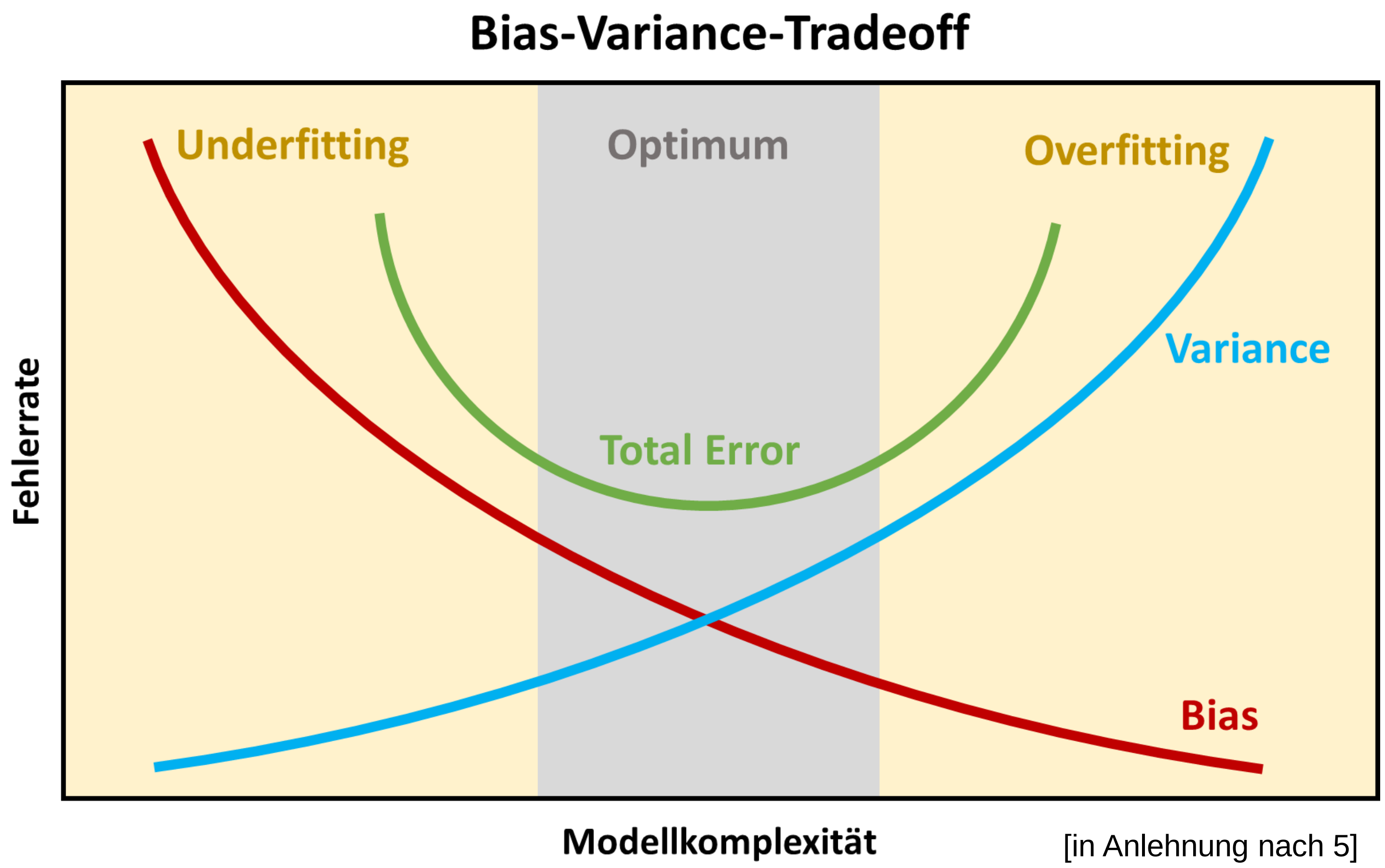
LASSO-Regression

Verfahren, bei dem die Anzahl der Einflussvariablen eines bestehenden linearen Regressionsmodells reduziert wird

KERNIDEE & FUNKTIONSWEISE
HINTERGRUND

Problematik: Wahl der optimalen Modellkomplexität

- **Bias:** statistische Verzerrung (Abweichung der Modellwerte von den Realwerten)
- **Variance:** Streuung der Daten um den Mittelwert (je stärker die Daten um den Mittelwert streuen, desto höher ist die Varianz)
- **Total Error:** Summe des Bias und der **Variance** [4]



[in Anlehnung nach 5]

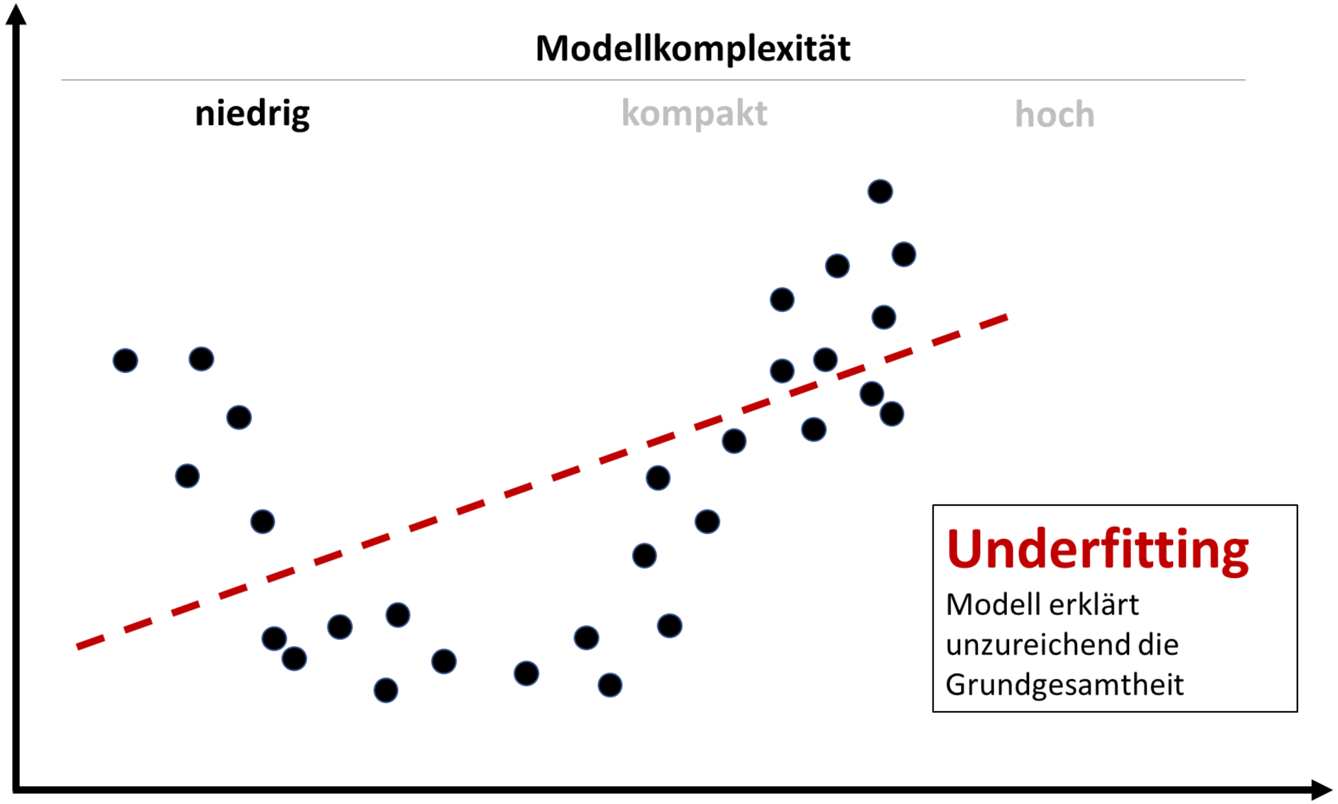
KERNIDEE & FUNKTIONSWEISE
HINTERGRUND

Underfitting:

- Niedrige Modellkomplexität
- Unzureichende Erklärung der Grundgesamtheit

Overfitting: [4]

- Hohe Modellkomplexität
- Sehr gute Erklärung der Trainingsdaten, jedoch hohe Fehlerraten bei Testdaten
- Schlechte Interpretierbarkeit und Anwendung für weitere Berechnungen



[in Anlehnung nach 6]

KERNIDEE & FUNKTIONSWEISE

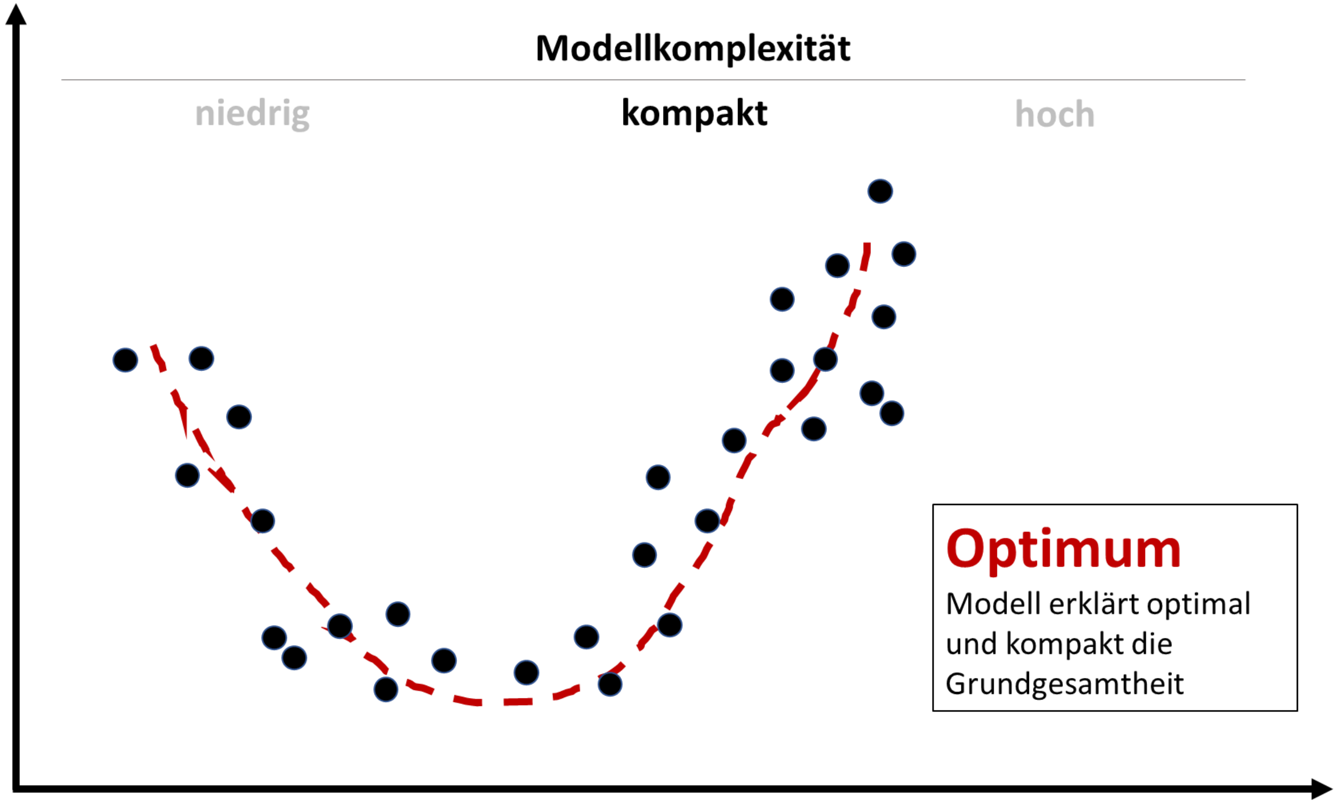
HINTERGRUND

Underfitting:

- Niedrige Modellkomplexität
- Unzureichende Erklärung der Grundgesamtheit

Overfitting: [4]

- Hohe Modellkomplexität
- Sehr gute Erklärung der Trainingsdaten, jedoch hohe Fehlerraten bei Testdaten
- Schlechte Interpretierbarkeit und Anwendung für weitere Berechnungen



[in Anlehnung nach 6]

KERNIDEE & FUNKTIONSWEISE

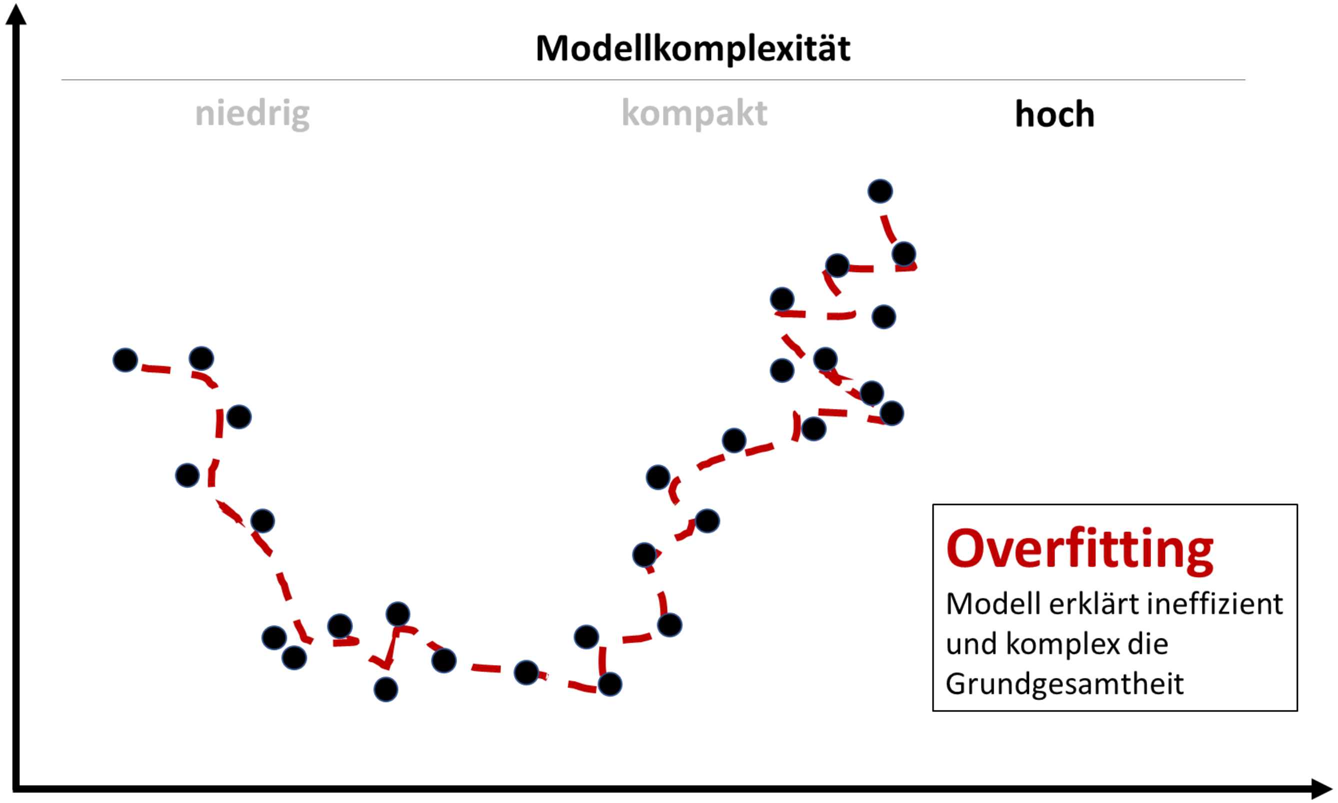
HINTERGRUND

Underfitting:

- Niedrige Modellkomplexität
- Unzureichende Erklärung der Grundgesamtheit

Overfitting: [4]

- Hohe Modellkomplexität
- Sehr gute Erklärung der Trainingsdaten, jedoch hohe Fehlerraten bei Testdaten
- Schlechte Interpretierbarkeit und Anwendung für weitere Berechnungen



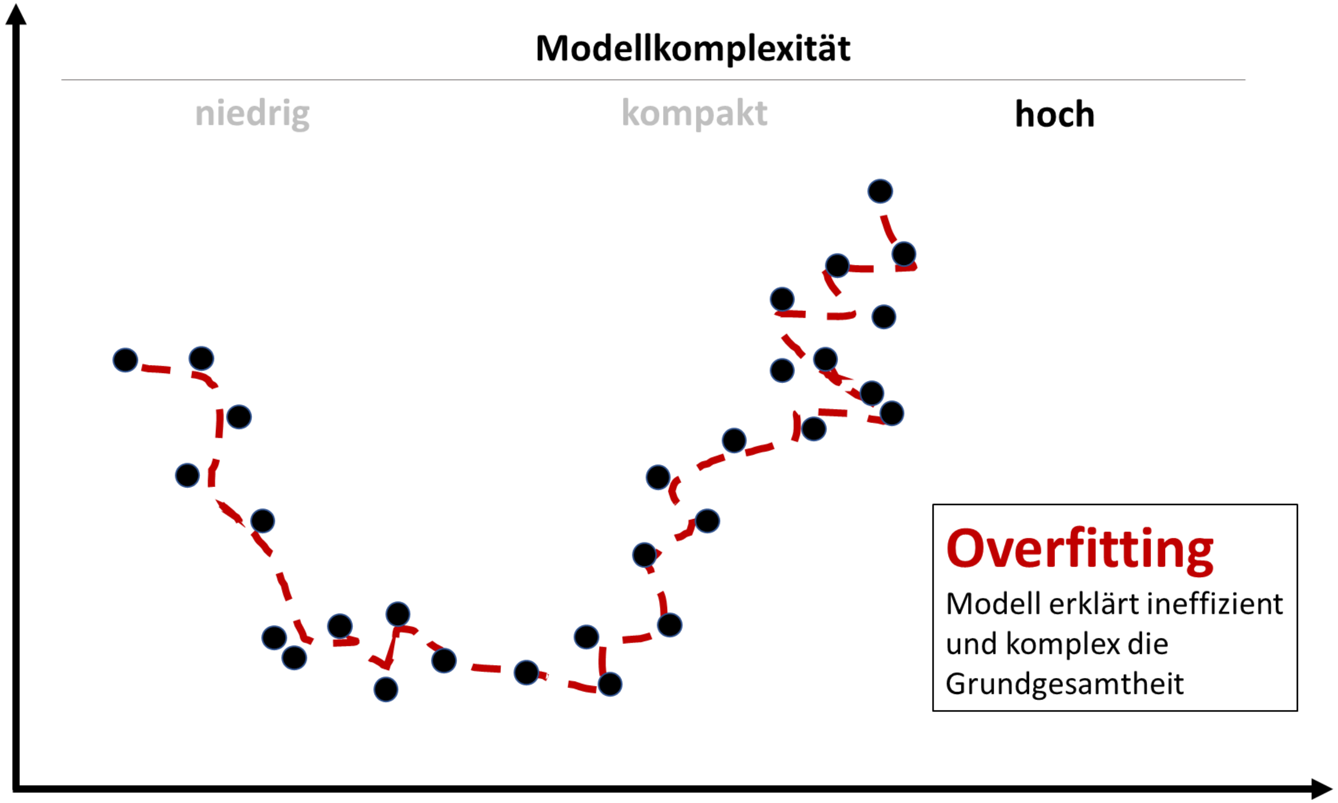
[in Anlehnung nach 6]

KERNIDEE & FUNKTIONSWEISE

HINTERGRUND

Verringerung der Modellkomplexität:

- **Dimensionsreduktion** (Verschmelzung von ähnlich strukturierten Variablen)
- **Variablenselektion** (Selektion einzelner Variablen unter Berücksichtigung bestimmter Kriterien)
- **Regression Shrinkage** (Skalierung der Regressionsparameter)



[in Anlehnung nach 6]

KERNIDEE & FUNKTIONSWEISE
FORMULA

- **Ziel der LASSO Regression**: Minimierung der Summe der quadratischen Abweichungen + Skalierung der Parameter
  - Minimierung ermöglicht die bestmögliche Regression zwischen der Einfluss- und Zielvariablen
  - Skalierung mittels Schrumpfungsterm $\lambda \sum_{j=1}^{p} |\beta_j|$

  → Skalierung aller Parameter mit Ausnahme des konstanten Parameters [7, S. 219]
- **Formel der LASSO Regression** [7, S. 219]

$$\underbrace{\sum_{i=1}^{n}(y_i - \beta_0 \sum_{j=1}^{p} \beta_j x_{ij})^2}_{\text{Berechnung der Parameter}} + \lambda \underbrace{\sum_{j=1}^{p} |\beta_j|}_{\text{Schrumpfungsterm}}$$

$\boldsymbol{n} \rightarrow Gesamtanzahl\ der\ Variablen$

$\boldsymbol{p} \rightarrow Gesamtzahl\ der\ Werte\ einer\ Variable$

$\boldsymbol{\beta_0} \rightarrow$ konstanter Paramter

$\boldsymbol{\beta_j} \rightarrow Steigungsparamter\ einer\ Variable$

$\boldsymbol{y_i} \rightarrow Zielvariable$

$\boldsymbol{x_i} \rightarrow Einflussvariablen$

$\boldsymbol{\lambda} \rightarrow Sensibilitätsparameter\ für\ Schrunpfungsterm$

KERNIDEE & FUNKTIONSWEISE

FORMULA

Regressionsmodell

$$y = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \beta_3 \times x_3 + \beta_4 \times x_4 + \dots + \beta_n \times x_n + \epsilon$$

Lineare Regression

$$y = 13.0806 + \boxed{0.0074} \times x_1 - \boxed{1.1457} \times x_2 - \boxed{0.1730} \times x_3 + \boxed{0.0029} \times x_4$$

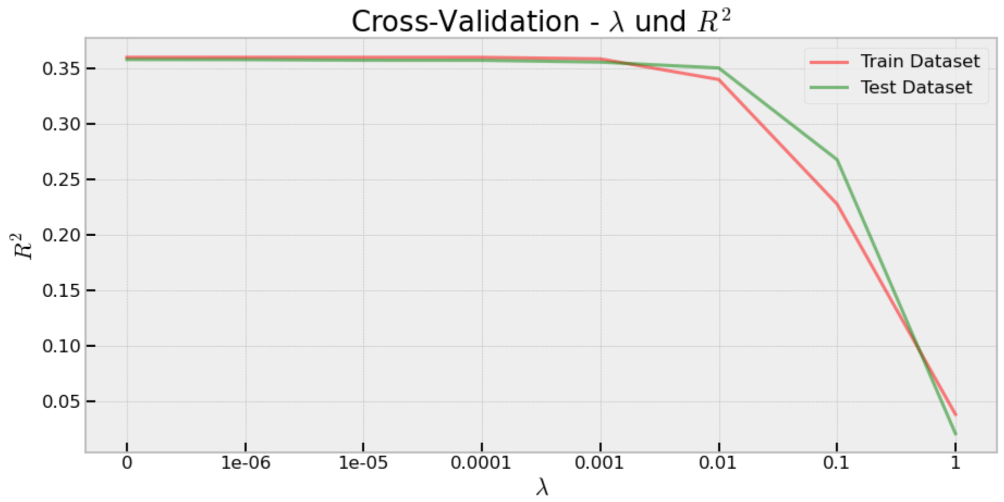
$$\boxed{\lambda \sum_{j=1}^{p} |\beta_j|}$$

LASSO Regression

$$y = 13.0806 - \boxed{0.0000} \times x_1 - 1.1449 \times x_2 - 0.1670 \times x_3 - \boxed{0.0000} \times x_4$$

Variablen-selektion

λ-Wert



KERNIDEE & FUNKTIONSWEISE

MODELLGÜTE & MODELLAUSWAHL

Bestimmtheitsmaß $\boldsymbol{R^2}$

- Je höher $\boldsymbol{R^2}$, desto besser erklärt das Modell die Regression multivariater Dimensionen.
- Je höher die Streuung der Werte, desto ungenauer wird die Regression und desto niedriger ist auch $\boldsymbol{R^2}$ [8, S. 84]

$$R^2 = \frac{SSE}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Aikaike Informationskriterium (AIC)

- Vergleich verschiedener Modelle mithilfe eines Strafterms
- Je höher die Anzahl der Parameter, desto sensitiver agiert der Strafterm [8, S. 333]

$$AIC = 2 \times k - 2 \ln \mathcal{L}$$

Bayessche Informationskriterium (BIC)

- weiterer Strafterm (ähnliche Funktionsweise wie AIC)
- Härtere Bestrafung bei hoher Anzahl an Parameter als bei AIC [8, S. 333]

$$BIC = \ln U \times k - 2 \ln \mathcal{L}$$

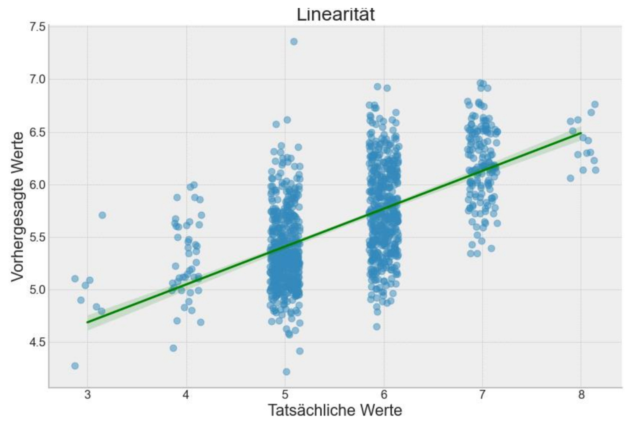
› ANWENDUNGS-VORAUSSETZUNGEN

HOCHSCHULE HEILBRONN

ANWENDUNGSVORAUSSETZUNGEN

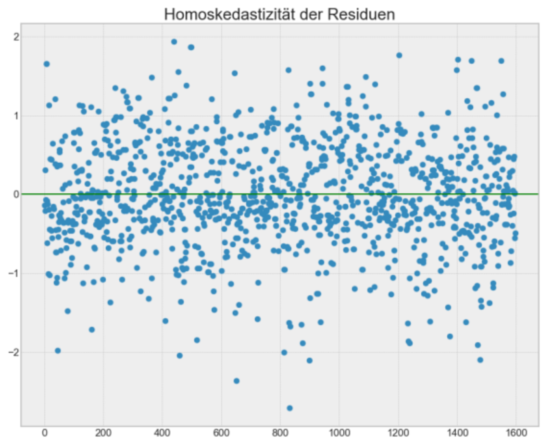
1 Linearität und korrekte Spezifizierung

- alle Variablen dürfen nur als 1. Grad vorliegen (keine Potenzierung der Werte)
- klare Festlegung der Ziel- und Einflussvariable(n)



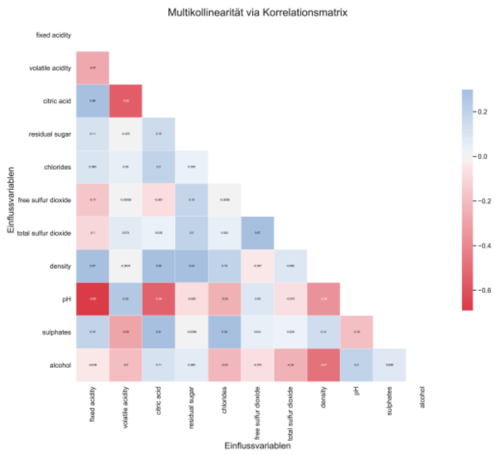
2 Homoskedastizität

- Variation der Residuen ist gleichmäßig ausgeprägt
- Heteroskedastizität führt zu ungleichmäßigen Variation der Residuen → Verzerrung des Modells (Underfitting)



3 keine Multikollinearität

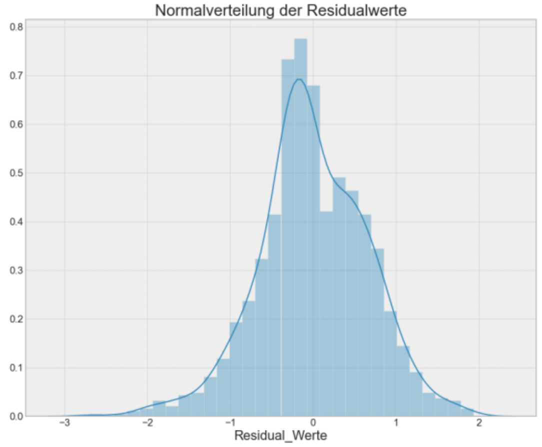
- Korrelationen innerhalb der Einflussvariablen beträgt: $Cor(X) < 1$
- bei perfekter Korrelation $Cor(X) = |1|$ → vollkommende lineare Abhängigkeit → Redundanz innerhalb des Modells (Overfitting)



ANWENDUNGSVORAUSSETZUNGEN

4 Normalverteilung der Residuen

- Residuen müssen normalverteilt sein (gleichmäßige Verteilung der Residuen um den Wert 0)
- bei keiner Normalverteilung → Verzerrung des Modells



5 keine Autokorrelation

- Korrelation der Residuen beträgt: $Cor(E) = 0$
- bei Autokorrelation werden Residuen bei Veränderung der Einflussvariable(n) linear miterklärt → Overfitting [8, S. 98, 111]

› STÄRKEN & SCHWÄCHEN

HHN HOCHSCHULE HEILBRONN

STÄRKEN UND SCHWÄCHEN

Merkmal	Stärken	Schwächen
Variablenselektion	✓ Schrumpfung von nutzlosen Variablen auf 0 ✓ Ausschluss aus dem Modell	X willkürliche Auswahl von Merkmalen bei mehreren korrelierenden Merkmalen X möglicher Informationsverlust und geringere Genauigkeit des Modells
Einfachheit	✓ Einfache Interpretation des Modells durch Variablenselektion	X starke Automatisierung des Modells vernachlässigt Modellanpassungen
Informationsgehalt	✓ Verringerung der Varianz ✓ Anpassung der Modellkomplexität zwischen Verzerrung und Varianz	X wichtige bzw. relevante Variablen könnten auch ausgeschlossen werden
Rechenzeit	✓ Niedrigdimensionelle Datensätze erfordern weniger Rechenzeiten	X Hochdimensionelle Datensätze erfordern mehr Rechenzeiten

[9, 10]

› DEMONSTRATION

› FRAGEN & ANTWORTEN

HOCHSCHULE HEILBRONN

REFERENZEN I

[1] Tibshirani, Robert (1996). “Regression Shrinkage and Selection via the Lasso”. In: Journal of the Royal Statistical Society 58, S. 267–288.

[2] Santosa, Fadil und William W. Symes* (1986). "Linear Inversion of Band-Limited Reflection Seismograms". In: SIAM Journal on Scientific and Statistical Computing 7.4, S. 1307-1330.

[3] The Royal Society (2019) - URL: https://royalsociety.org/people/robert-tibshirani-14130/.

[4] Singh, Seema (2018). Understanding the Bias-Variance Tradeoff, Hrsg. von Towards Data Science. (URL: https://towardsdatascience.com/understandingthe-bias-variance-tradeoff-165e6942b229)

[5] Deng, Bai-Chuan u. a.* (2015). “A new strategy to prevent overfitting in partial least squares models based on model population analysis”. In: Analytica Chimica Acta 880, S. 35.

REFERENZEN II

[6] Rashidi, Hooman H u. a.* (2019). “Artificial Intelligence and Machine Learning in Pathology: The Present Landscape of Supervised Methods”. In: Academic Pathology 6, S. 11.

[7] James, Gareth u. a. (2013). An Introduction to Statistical Learning. Bd. 103. Springer New York.

[8] Backhaus, Klaus u. a. (2016). Multivariate Analysemethoden. Berlin, Heidelberg: Springer Berlin Heidelberg.

[9] Pereira, Jose Manuel, Mario Basto und Amelia Ferreira da Silva* (2016). "The Logistic Lasso and Ridge Regression in Predicting Corporate Failure". In: Procedia Economics and Finance 39, S. 634 - 641.

[10] Fonti, Valeria und Eduard Belitser (2017). Feature Selection using LASSO.

Mimicking the herd invites regression to the mean.

Charlie Munger