Daten und Modelle – die zwei Säulen der Statistik

Hansruedi Künsch
Seminar für Statistik, D-Math
ETH Zürich

Zürcher Hochschultag, 2. Februar 2012

Was ist Statistik ?

Statistik entwickelt und untersucht Methoden für

- die Erfassung von Strukturen in Daten
- die Beantwortung wissenschaftlicher Fragen basierend auf Daten

welche

- unabhängig sind vom spezifischen Anwendungsgebiet
- die Unsicherheit berücksichtigen, die durch Fehler und inhärente Variabilität dieser Daten entstehen.

Daten und Modelle

Die Statistik schlägt eine Brücke zwischen der realen Welt der Daten und der theoretischen Welt der Modelle.

Wir unterscheiden 2 Typen von Modellen:

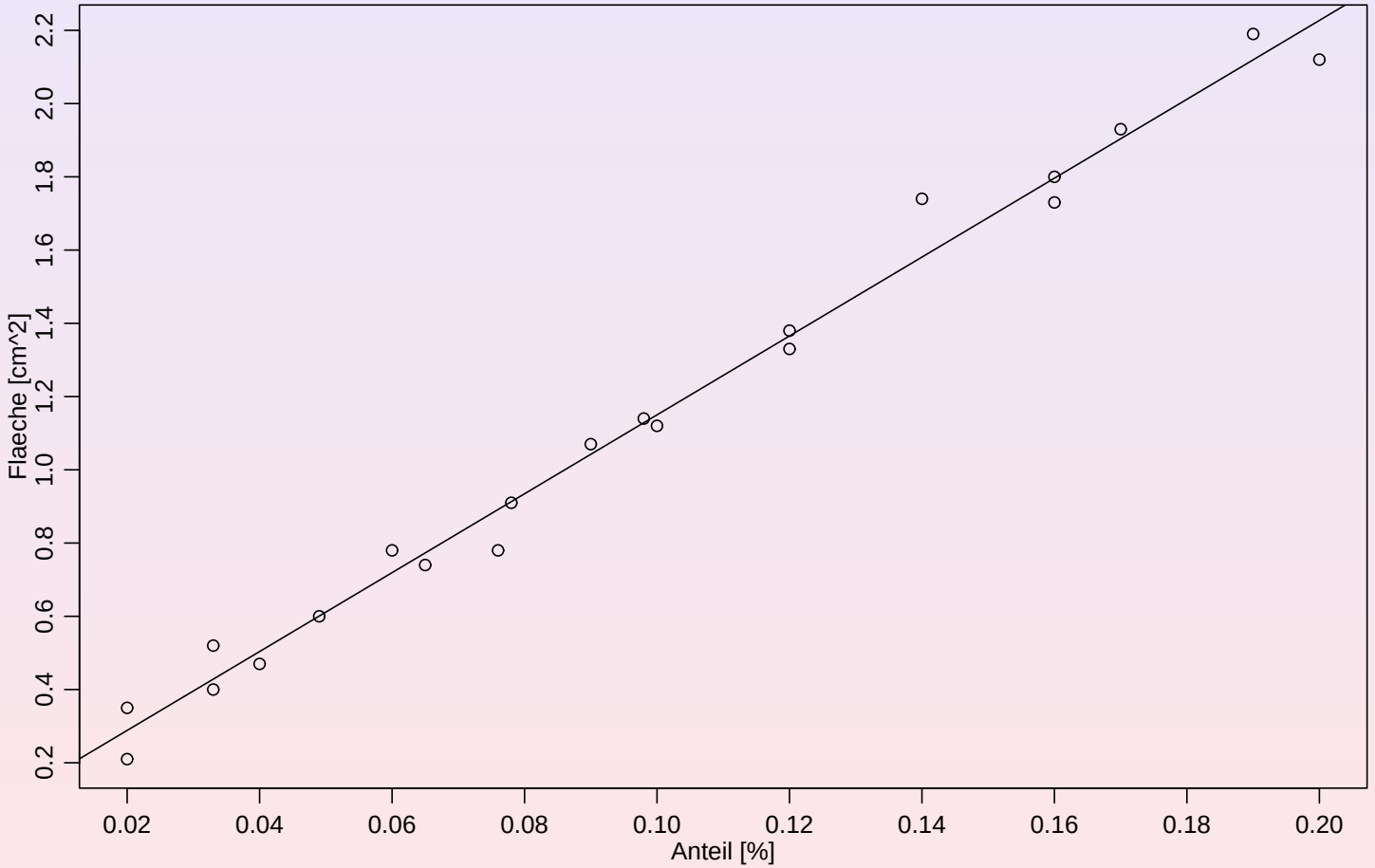
- Modelle der Substanzwissenschaften
  beschreiben Zeitentwicklung, Beziehungen zwischen Variablen
  meist deterministisch
- Statistische Modelle
  beschreiben Variabilität in den Daten
  meist stochastisch

Ziel des heutigen Vortrags: Einfache Illustration dieser Konzepte an Hand von Fragestellungen, die in der Statistik aktuell sind.
Roter Faden: Lineare Gleichungssysteme “mit Fehlern”.

Inhalt

Beispiel: Nachweis von Lebensmittelfarbstoff

y: Fläche des Peaks in einer Spektralmessung
x: Anteil des Farbstoffes.



KLeinste-Quadrate Gerade durch Punktwolke

Gegeben n Messungen (x_i, y_i) zweier Grössen, zwischen denen eine approximative lineare Beziehung besteht $y_i \approx \beta_1 + \beta_2 x_i$.

Wahl der Parameter β_1, β_2 durch Minimieren der Summe der quadratischen Abweichungen in y-Richtung:

$$(\widehat{\beta}_1, \widehat{\beta}_2) = \arg\min_{\beta_1,\beta_2} \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2$$

(Gauss 1795 (?) und 1809, Legendre 1805)

Kleinste Quadrate allgemein

Gegeben n Messungen einer Zielgrösse, die approximativ linear von $p \ll n$ erklärenden Grössen abhängt

$$y_i \approx \beta_1 + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} \quad (i = 1, \ldots, n).$$

In Matrixform $\mathbf{Y} \approx \mathbf{X}\beta$ (mit $x_{i1} = 1$).

Bestimmung von β durch Minimierung von $||\mathbf{Y} - \mathbf{X}\beta||_2^2 \rightsquigarrow$ Normalgleichung $\mathbf{X}^T\mathbf{Y} = (\mathbf{X}^T\mathbf{X})\widehat{\beta}$.

Für später: $\mathbf{X}^T\mathbf{X}$ enthält die Skalarprodukte der Spalten von $\mathbf{X}$.

Das lineare Modell der Statistik

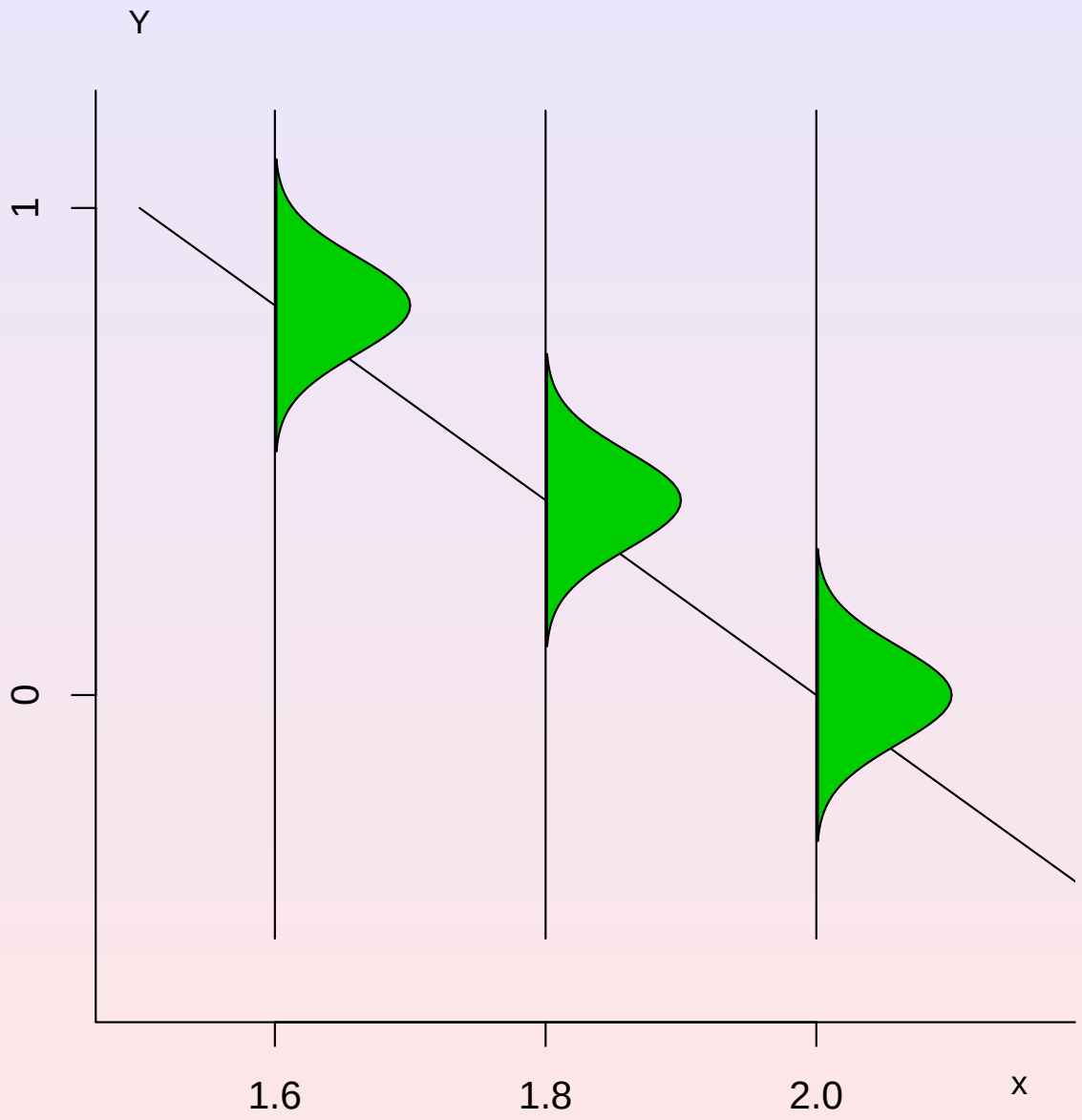
Annahme: Es gibt eine exakte lineare Beziehung mit unbekanntem wahren Parameter β_0. Abweichungen der Messpunkte davon sind zufällig:

$$\mathbf{Y} = \mathbf{X}\beta_0 + \boldsymbol{\epsilon},$$

wobei für den Fehler $\boldsymbol{\epsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$ gilt

- $\mathbb{E}(\varepsilon_i) = 0$ (Fehler im Mittel gleich Null),
- $\mathrm{Var}(\varepsilon_i)$ konstant (gleiche Streuung der Fehler),
- $\varepsilon_1, \ldots, \varepsilon_n$ stochastisch unabhängig,
- ε_i normalverteilt ($\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$).

Illustration des Modells



Verteilung von $\widehat{\beta}$

Wenn ϵ zufällig ist, dann sind es auch $\mathbf{Y}$ und $\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. Unter obigen Annahmen ist $\widehat{\beta}$ ebenfalls normalverteilt mit $\mathbb{E}(\widehat{\beta}) = \beta_0$ und $\text{Cov}(\widehat{\beta}) = \text{Var}(\varepsilon_i)(\mathbf{X}^T\mathbf{X})^{-1}$.

Damit kann man

- Die Genauigkeit der Schätzung $\widehat{\beta}$ angeben (Vertrauensintervalle).
- Testen, ob einzelne Komponenten von $\beta_0 = 0$.
- Die Unsicherheit von Vorhersagen $\mathbf{X}_{\text{neu}}\widehat{\beta}$ für zusätzliche Beobachtungen quantifizieren.

Mittels Residuenanalyse kann man auch einzelne der obigen Annahmen an ϵ überprüfen.

Vorhersage

$\mathbf{X}\widehat{\beta}$ = Vorhersage von n neuen, unabhängigen Beobachtungen mit den gleichen Werten der erklärenden Variablen.
$\mathbf{X}(\widehat{\beta} - \beta_0)$ = Fehler der Vorhersage wegen Parameterschätzung.

Wenn Modell korrekt:

$$\frac{1}{n}\mathbb{E}(\|\mathbf{X}(\widehat{\beta} - \beta_0)\|_2^2) = \frac{p}{n}\mathrm{Var}(\varepsilon_i).$$

(Quadratischer Vorhersagefehler proportional zu Anzahl unbekannter Parameter).

Kreuzvalidierung

Empirische Bestimmung des Vorhersagefehlers: Schätze β aus einem Teil der Daten ("Trainingsdaten") und validiere die Vorhersage auf dem Rest ("Testdaten").

Effiziente Implementation (jede Beobachtung einmal bei den Test- und einmal bei den Trainingsdaten).

- $(I_1, \ldots, I_K)$ = Partition von $\{1, 2, \ldots, n\}$ in K (=5 oder 10) etwa gleichgrosse Teilmengen.
- $\widehat{\beta}^{(-j)}$ = Schätzung basierend auf $((y_i, x_i); i \notin I_j)$.
- Mittlerer quadratischer Kreuzvalidierungsfehler

$$\frac{1}{n} \sum_{j=1}^{K} \sum_{i \in I_j} (y_i - x_i^T \widehat{\beta}^{(-j)})^2.$$

Inhalt

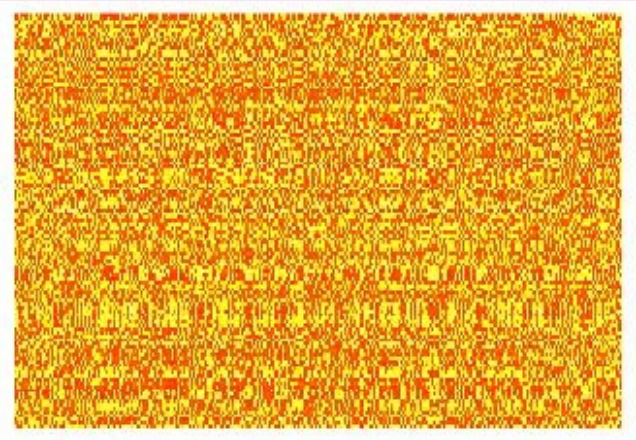
Beispiel I: Riboflavinproduktion

Zusammenarbeit von Peter Bühlmann mit Firma DSM

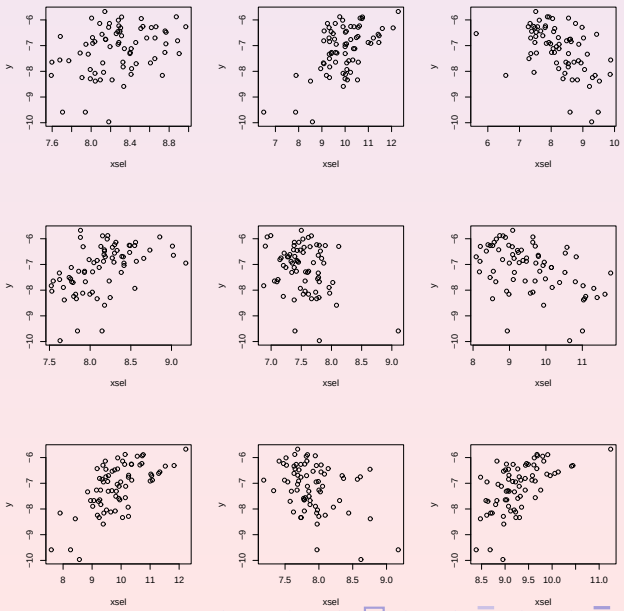
Ziel: Verbesserung der Riboflavin-Produktionsrate von Bacillus Subtilis mit Hilfe von genetischem Engineering.

Zielgrösse Y: (log) Riboflavin-Produktionsrate
Erklärende Grössen X: (log) Expression von $p = 4088$ Genen
Stichprobengrösse $n = 115 \ll p$.

Genexpressionsdaten



Y gegen 9 “vernünftige” Gene



Temperatur-Rekonstruktion 1000-2000

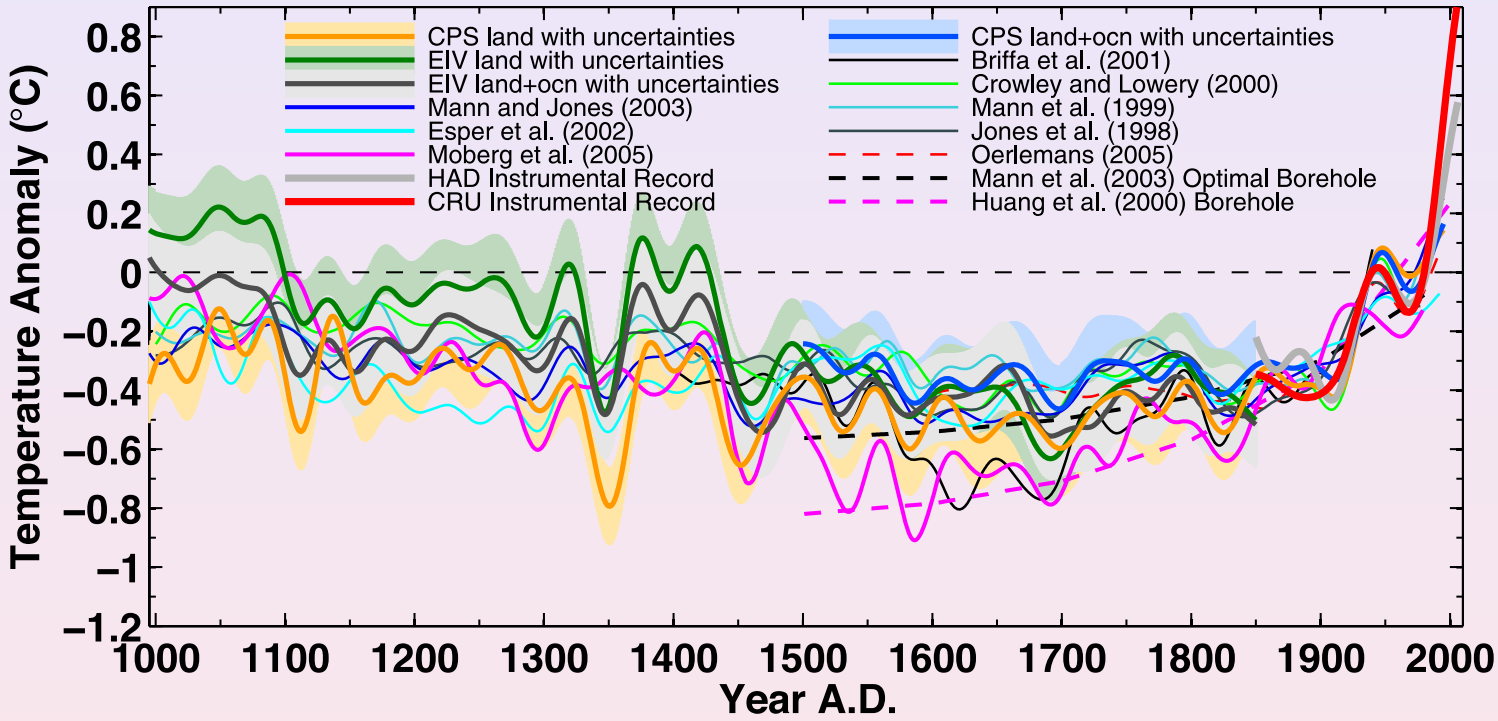


FIG. 2. *Various reconstructions of Northern Hemisphere temperatures over the last with 95% confidence intervals. Source: Mann et al.* (2008).

Beispiel II: Rekonstruktion des Paläoklimas

Ziel: Rekonstruktion der globalen mittleren Oberfächentemperatur der letzten 1000 Jahre auf Grund von Proxy-Daten (Baumringe, Eisbohrkerne etc.)

Zielgrösse Y: Mittlere globale Temperatur, berechnet aus Temperaturmessstationen, 1850-1998
Erklärende Variablen X: $p = 1209$ Proxy-Zeitreihen
Stichprobengrösse $n = 149 \ll p$.

Proxy-Zeitreihen gehen weiter zurück als Messstationen $\rightsquigarrow$ Verwendung des Modells zur Rekonstruktion.

Führte vor kurzem zu wissenschaftlichen Kontroversen , siehe McShane and Wyner, Annals of Applied Statistics, Vol. 5, 2011, 5-123, (erwähnt im Wallstreet Journal).

Beispiel III: Nichtparametrische Regression

Nicht näher spezifizierter Zusammenhang zwischen Zielgrösse und einer erklärenden Variablen $x \in \mathbb{R}$:

$$y_i = f(x_i) + \varepsilon_i.$$

Entwickle f bezüglich einer Basis von Funktionen $(\psi_j(.); j = 1, 2 \ldots)$ (Polynome, trigonometrische Funktionen, Wavelets, B-splines):

$$y_i = \sum_{j=1}^{p} \beta_j \psi_j(x_i) + \varepsilon_i.$$

Häufig ist $p = n$ oder $p > n$.

4 Prototypen für f, Donoho und Johnstone (1994)

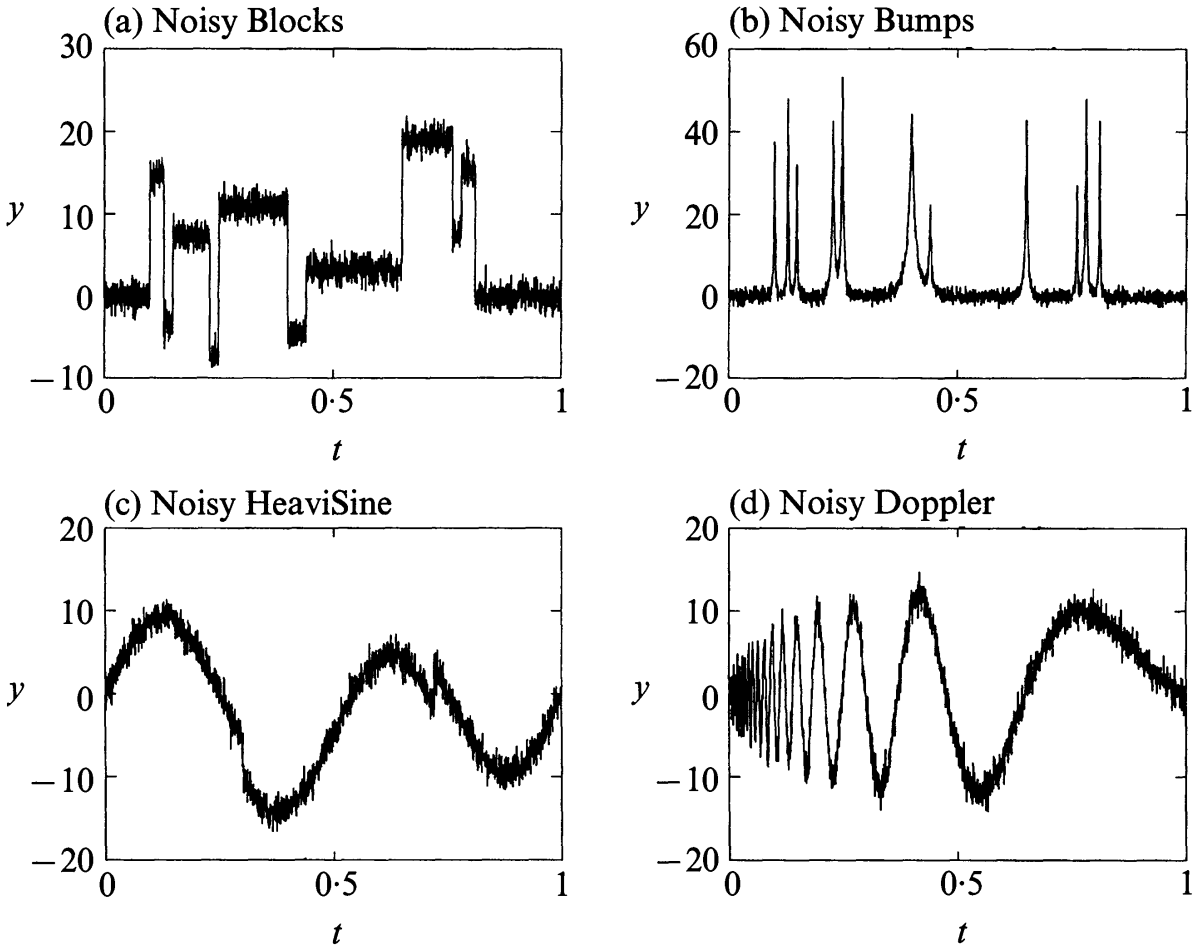


Fig. 3. Four functions with Gaussian white noise, $\sigma = 1$, with f rescaled to have signal-to-noise ratio, $\text{SD}(f)/\sigma = 7$.

Regularisierung

Gleiches Modell wie zuvor $\mathbf{Y} = \mathbf{X}\beta_0 + \epsilon$,
aber jetzt $p \approx n$ oder $p \gg n$.

Wenn Rang($\mathbf{X}$) $= n$, dann gibt es Parameter mit perfektem fit:
$\mathbf{Y} = \mathbf{X}\beta$. Schlecht gestelltes Problem.

Gesucht sind Parameter, die einen vernünftigen Kompromiss zwischen gutem fit und einfacher Form des Zusammenhangs erreichen. Wir messen Komplexität des Zusammenhangs durch eine (Pseudo)-Norm $||\beta||$, Dafür müssen wir $\mathbf{Y}$ und die Spalten von $\mathbf{X}$ zuerst zentrieren und standardisieren.

Regularisierung mit q-Normen

Regularisierte Schätzung

$$\widehat{\beta}(\lambda, q) = \text{argmin}_{\boldsymbol{\beta}} \left(n^{-1} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \, \|\boldsymbol{\beta}\|_q \right).$$

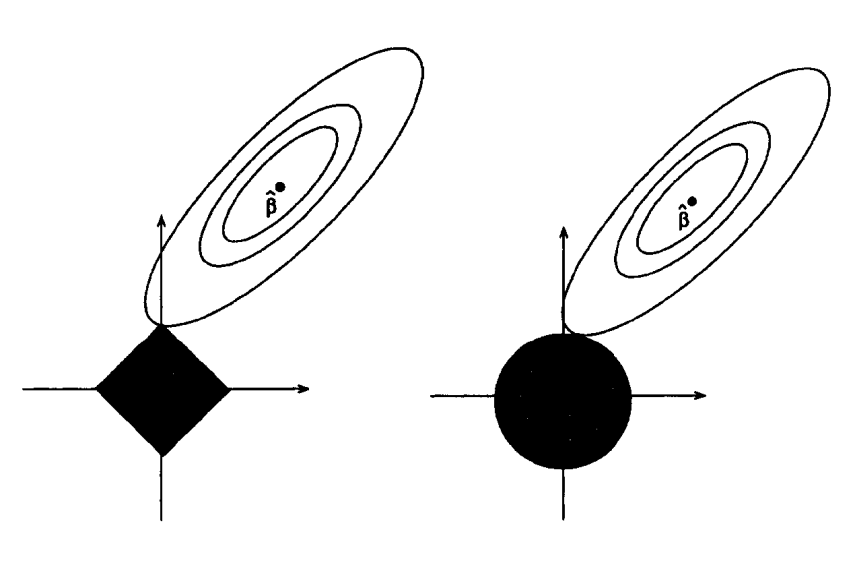
mit der q-Norm

$$||\boldsymbol{\beta}||_q^q = \sum_{j=1}^{p} |\beta_j|^q \quad (0 < q < \infty).$$

Für $q \leq 1$: Lösung setzt einzelne Komponenten von $\widehat{\beta}$ auf Null ($\rightsquigarrow$ Variablenselektion).
Für $q \geq 1$: Zielfunktion ist konvex ($\rightsquigarrow$ Optimierung machbar in hohen Dimensionen).

$q = 1$ erfüllt beides (Tibshirani, 1996, LASSO = Least Absolute Shrinkage and Selection Operator).

Geometrische Veranschaulichung



Situation für $q = 1$ (links) und $q = 2$ (rechts). In hohen Dimensionen werden die Unterschiede ausgeprägter.

Spezialfall: Orthogonale Spalten

Wenn $p \leq n$ und $n^{-1}\mathbf{X}^T\mathbf{X} = I$ (z.B. bei Wavelets), dann

$$\arg\min_{\beta}\left(n^{-1}\|\mathbf{Y}-\mathbf{X}\beta\|_2^2+\lambda\,\|\beta\|_1\right) = \arg\min_{\beta}\sum_{j=1}^{p}(\beta_j^2-2\beta_j Z_j+\lambda|\beta_j|)$$

wobei $\mathbf{Z} = n^{-1}\mathbf{X}^T\mathbf{Y}$ = Kleinste Quadrate Lösung.

LASSO schrumpft in diesem Fall die Kleinste-Quadrate-Lösung nichtlinear gegen Null:

$$\widehat{\beta}_j(\lambda) = \begin{cases} Z_j-\lambda/2 & \text{wenn} \quad Z_j > \lambda/2 \\ 0 & \text{wenn} \quad |Z_j| \leq \lambda/2 \\ Z_j+\lambda/2 & \text{wenn} \quad Z_j < -\lambda/2 \end{cases}$$

(sogenanntes soft-thresholding).

Anwendung des Lassos: Wavelet-thresholding

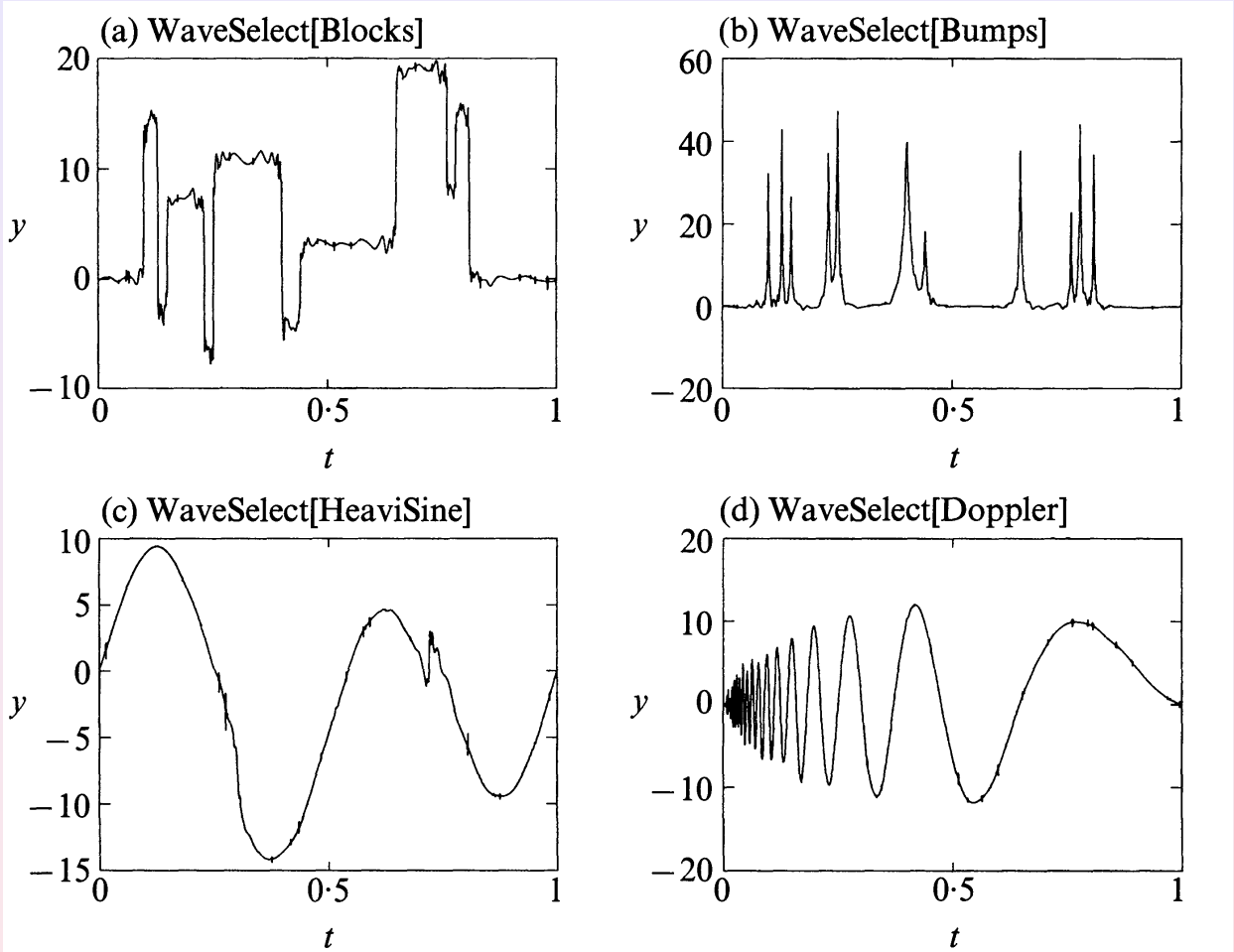
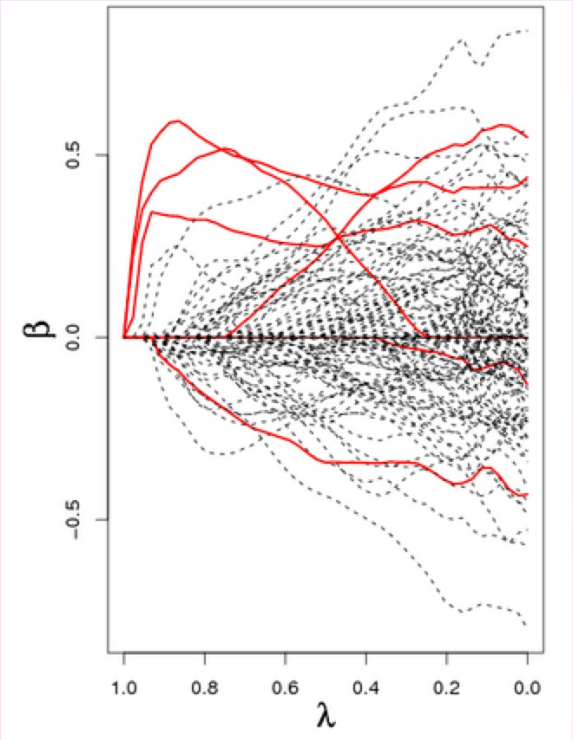


Fig. 7. RiskShrink reconstruction using soft thresholding and $\lambda = \lambda_n^*$. Mimicking an oracle while relying on the data alone.

Anwendung des Lassos: Vitaminproduktion

Ergebnisse für die Originaldaten nicht öffentlich.

Experiment zur Überprüfung der Methodik: Wähle 6 “vernünftige” Gene, permutiere die restlichen 4082 Gene zufällig $\rightsquigarrow$ keine Assoziation mehr zur Zielvariable.
Plotte $\widehat{\beta}_j$ gegen λ (rot für die 6 ausgewählten Gene, die einen Einfluss auf die Vitaminproduktion haben).



Anwendung des Lassos: Paläoklima

Zur Erinnerung: $\mathbf{Y}$ = Oberflächentemperatur 1850-1998
$\mathbf{X}$ = 1209 Proxy-Zeitreihen (Baumringe, Pollen etc.).

McShane und Wyner (2011) untersuchen Vorhersagefehler

$$\left(\frac{1}{30} \sum_{i=k}^{k+29} (y_i - \widehat{y}_i^{(-k)})^2 \right)^{1/2} \qquad (k = 1850, \ldots, 1969),$$

wobei $\widehat{y}_i^{(-k)}$ = verschiedene Vorhersagen von y_i ohne Verwendung der Jahre $k, k+1, \ldots, k+29$:

- $\widehat{y}_i^{(-k)}$ = arithmetisches Mittel der verbleibenden 119 Jahre.
- $\widehat{y}_i^{(-k)}$ = Zeitreihenvorhersage ohne Proxies.
- $\widehat{y}_i^{(-k)} = x_i^T \widehat{\beta}^{(-k)}(\widehat{\lambda}^{(-k)})$ (Lasso mit kreuzvalidiertem λ).
- Wie oben, aber mit “Pseudo-Proxies” = künstliche Zeitreihen, die unabhängig sind von (y_i), aber ähnliche statistische Eingeschaften haben wie die echten Proxies.

Boxplots der Vorhersagefehler

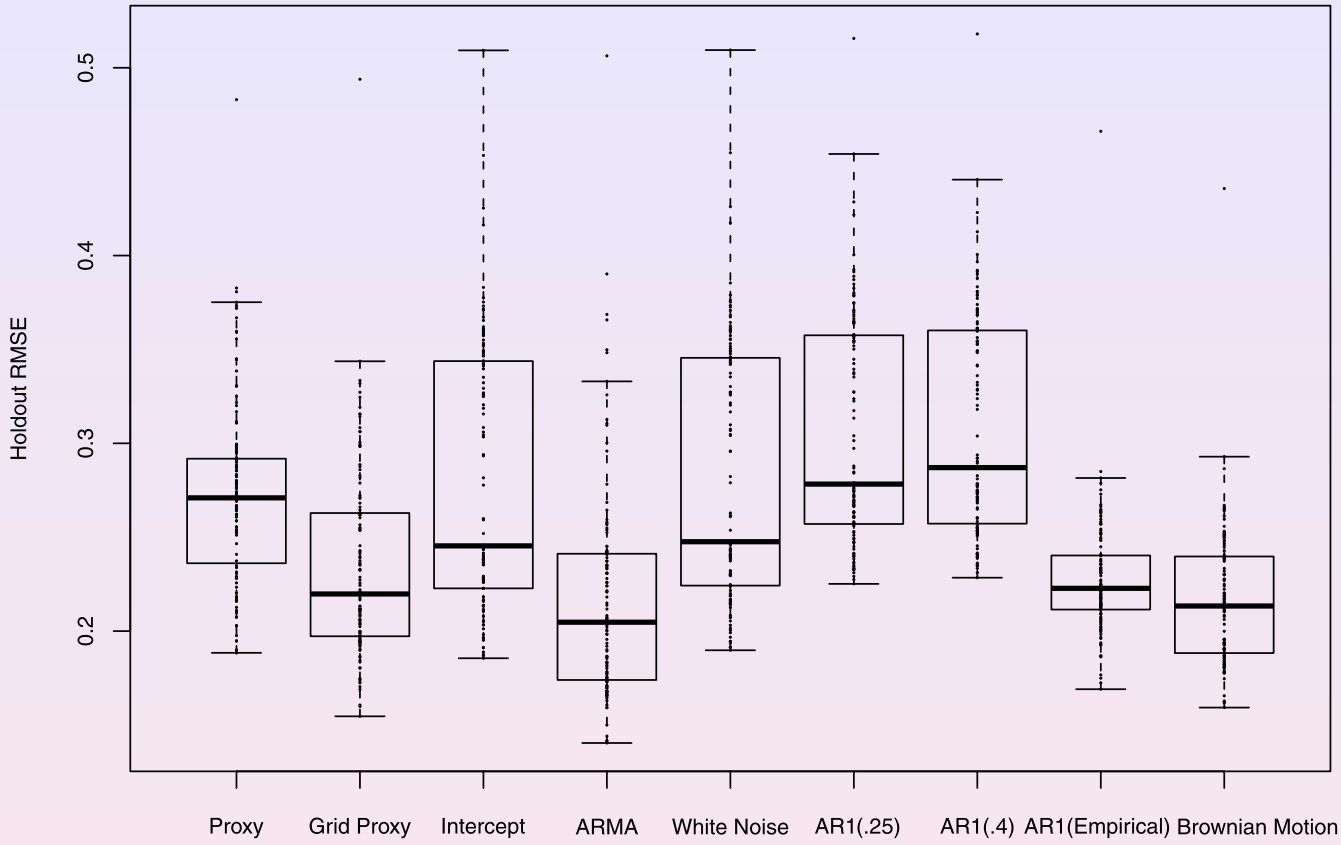


FIG. 9. *Cross-validated RMSE on* 30*-year holdout blocks for various models fit to proxies and pseudo-proxies. The procedures used to generate the Proxy, Intercept, and ARMA boxplots are discussed in Section* 3.2*. The procedures used to generate the White Noise, AR*1*, and Brownian motion boxplots are discussed in Section* 3.3*. The procedure used to generate the Grid Proxy boxplot is discussed in Section* 3.6*.*

"Proxies do not predict temperature significantly better than random series generated independently of temperature."

Argumente der Kritiker dieser Studie

- Lasso ungeeignet für diese Daten: Reduktion auf wenige Proxies hat keine wissenschaftliche Basis. Man braucht alle Proxies, da diese meistens nur etwas über die lokale Temperatur in einem Jahr aussagen.
- Koeffizienten für Proxies der selben Art (z.B. alle Baumringproxies) sollten ähnlich sein ($\rightsquigarrow$ andere Art der Regularisierung).
- Gewisse Proxies enthalten nur Information über langfristige Änderungen der Temperatur, und es gibt Zeitverzögerungen.
- Falsche Richtung des linearen Modells: Wahre Temperatur bestimmt Proxies, nicht umgekehrt.

Theorie für Lasso: Vorhersage

Modell $\mathbf{Y} = \mathbf{X}b\beta_0 + \boldsymbol{\epsilon}$. Wie gross ist $n^{-1}\|\mathbf{X}(\widehat{\beta}(\lambda) - \beta_0)\|_2^2$?

Satz: Wenn die ε_i unabhängig und $\mathcal{N}(0, \sigma^2)$-verteilt sind, und wenn $\lambda \geq 2\lambda_0 = 4\sigma\sqrt{(u^2 + 2\log p)/n}$, dann gilt mit Wahrscheinlichkeit $\geq 1 - \exp(-u^2/2)$

$$n^{-1}\|\mathbf{X}(\widehat{\beta} - \beta_0)\|_2^2 \leq \frac{3}{2}\lambda\|\beta_0\|_1.$$

Fehler klein, falls $\log p \ll n$ und $\|\beta_0\|_1$ klein.

Beweisskizze

Nach Definition von $\widehat{\beta}$ gilt

$$n^{-1}\|\mathbf{Y}-\mathbf{X}\widehat{\beta}\|_2^2+\lambda\|\widehat{\beta}\|_1 \leq n^{-1}\|\mathbf{Y}-\mathbf{X}\beta_0\|_2^2+\lambda\|\beta_0\|_1.$$

Einsetzen von $\mathbf{Y}=\mathbf{X}\beta_0+\epsilon$ und Ausmultiplizieren ergibt

$$n^{-1}\|\mathbf{X}(\widehat{\beta}-\beta_0)\|_2^2+\lambda\|\widehat{\beta}\|_1 \leq 2n^{-1}\epsilon^T\mathbf{X}(\widehat{\beta}-\beta_0)+\lambda\|\beta_0\|_1.$$

Ferner

$$|\epsilon^T\mathbf{X}(\widehat{\beta}-\beta_0)| \leq \|\mathbf{X}^T\epsilon\|_\infty\|\widehat{\beta}-\beta_0\|_1.$$

Wenn $2n^{-1}\|\mathbf{X}^T\epsilon\|_\infty \leq \lambda_0 \leq \frac{1}{2}\lambda$, dann

$$n^{-1}\|\mathbf{X}(\widehat{\beta}-\beta_0)\|_2^2 \leq \lambda_0\|\widehat{\beta}-\beta_0\|_1+\lambda\|\beta_0\|_1-\lambda\|\widehat{\beta}\|_1 \leq \frac{3}{2}\lambda\|\beta_0\|_1.$$

Das Maximum von normalverteilten Zufallsvariablen

Noch abzuschätzen:

$$\mathbb{P}(n^{-1}\|\mathbf{X}^T\epsilon\|_\infty \leq \lambda_0) = \mathbb{P}(\max_j |Z_j| > \lambda_0),$$

wobei $Z_j = n^{-1}\sum_{i=1}^n x_{ij}\varepsilon_i$. Wir benutzen, dass jedes $Z_j \sim \mathcal{N}(0, \sigma^2/n)$ (weil Spalten von $\mathbf{X}$ normiert).

$$\begin{aligned}\mathbb{P}(\max_j |Z_j| > \lambda_0) &\leq p\,\mathbb{P}(|Z_1| > \lambda_0) = 2p\,(1 - \Phi(\lambda_0\sqrt{n}/\sigma)) \\ &\leq 2p\,\sigma\frac{\exp(-n\lambda_0^2/(2\sigma^2))}{\sqrt{2\pi}\sqrt{n}\lambda_0}\end{aligned}$$

weil $1 - \Phi(x) \leq \phi(x)/x$. Beachte: λ_0 wächst mit p und kompensiert so den Faktor p.

Stärkere Resultate: Orakel-Ungleichung

Annahme: Die meisten wahren Koeffizienten sind Null

$$s_0 = |S_0| = |\{j;\ \beta_{0,j} \neq 0\}| \ll n.$$

Falls S_0 bekannt, verwende Kleinste Quadrate nur mit den Variablen aus S_0 (ohne Regularisierung). Dann

$$n^{-1} \|\mathbf{X}(\widehat{\beta} - \beta_0)\|_2^2 \leq C\sigma^2 \frac{s_0}{n}.$$

mit beliebig grosser Wahrscheinlichkeit, sobald C genügend gross.

Welchen Preis müssen wir bezahlen dafür, dass wir S_0 nicht kennen ? Unter Bedingungen an $\mathbf{X}$ kommt nur ein zusätzlicher Faktor $\log p$ dazu.

Exakte Formulierung: Siehe Peter Bühlmann und Sara van de Geer, Statistics for High-Dimensional Data, Springer 2011.

Rückblick

- Das einfache Modell einer linearen Beziehung plus Fehler ist reichhaltiger als es auf den ersten Blick scheint.
- Motiviert von vielen Anwendungen untersucht die Statistik heute Situationen, wo man mehr Variablen als Beobachtungseinheiten hat.
- In solchen Situationen ist Regularisierung ein zentrales Konzept.
- Der Unterschied zwischen der 1-Norm und der 2-Norm ist wesentlich, obwohl beide Normen in endlichen Dimensionen äquivalent sind.
- Kreuzvalidierung und die Simulation unter geeigneten "Nullmodellen" sind wichtige Methoden zur Quantifizierung von Unsicherheit.
- Die "richtige" Anwendung und Interpretation statistischer Modelle ist eine Herausforderung !

Ausblick: Weiterbildung im Juni

Thema:	Bayes-Statistik und stochastische Simulation
Leitung:	HRK und Corinne Dahinden (KS Zug)
Ort und Zeit:	Mittwoch 6. Juni 2012, 9:30-17:30, ETH
Ziel 1:	Ansatz der Bayes-Statistik kennenlernen, und ihn umsetzen für den Unterricht im Fall der Binomialverteilung.
Ziel 2:	Einblick in einige Anwendungen von stochastischer Simulation.
Ziel 3:	Beispiel einer stochastischen Simulation auf Ihrem eigenen Laptop implementieren.
Anmeldung:	bis 6. April auf http://www.webpalette.ch/