Methoden der Statistik

Mathias Trabs, Markus Reiß

Korrekturen bitte an shevcher@student.hu-berlin.de richten

Vorläufige Version, 24. März 2015

5 Modellwahl

5.1 Allgemeine Betrachtungen

Was ist ein Modell?
$\rightsquigarrow$ mathematische Formalisierung eines realen Systems, um dieses besser zu verstehen und Berechnungen/Vorhersagen durchführen zu können. Dieses Modell ist meist stark idealisiert und approximiert die Wirklichkeit.
Die meisten Modelle beinhalten stochastische Komponenten und mittels Statistik wird der Bezug zu realen Daten hergestellt.
Was macht ein gutes stochastisches Modell aus?

- Einfachheit (sparsame Beschreibung),
- Konformität mit Daten („goodness of fit“),
- Verallgemeinerbarkeit (kann benutzt werden, um neue Daten zu beschreiben oder vorherzusagen).

Bei statistischen Modellen gibt es zwei mögliche Probleme:

- das Modell ist zu simplizistisch, Haupteffekte werden nicht erklärt („model underfit“, keine goodness of fit),
- das Modell ist zu komplex für die Daten, so dass die Verallgemeinerbarkeit auf neue Daten nur mit großen statistischen Schwankungen möglich ist („model overfit“).

Mathematisch gilt:

- Underfitting induziert einen Bias (systematische Verzerrung),
- Overfitting erhöht die stochastische Variabilität jenseits des Notwendigen.

5.1 Beispiel. Neue AIDS-Fälle in Belgien.
Modell: x: Jahr, $y(x)$: Fallzahl im Jahr x.
$y(x) \sim \text{Poiss}(\lambda_x)$ (seltenes Ereignis),
$\log(\lambda_x)$ ist ein Polynom in x vom Grad $d \rightsquigarrow$ Regression.

5.2 Akaike-Informationskriterium (AIC)

Das AIC wurde 1973 von Hirotugu Akaike (1927-2009) eingeführt.
Idee: Statistik beruht oft auf einem Likelihood-Ansatz, wofür ein spezifisches parametrisches Modell angenommen wird. Akaike hat eine Methode entwickelt, die Schätzung im linearen Modell und Wahl zwischen verschiedenen Modellen kombiniert, um die echte Situation gut zu approximieren.
Die Daten $x \in X$ in einem Stichprobenraum $(X, \mathscr{F})$ werden gemäß einer unbekannten Verteilung $\mathbb{P}$ auf $\mathscr{F}$ generiert. Wir betrachten Modelle $(X, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta_k})$ mit $\Theta_k \subseteq \mathbb{R}^k$ offen mit $k = 1, \dots, K$ und $\hat{\vartheta}_k := \operatorname{argmin}_{\vartheta\in\Theta_k} L_k(\vartheta)$ bezeichne einen MLE im Modell k ($L_k(\vartheta, x) = \frac{\mathrm{d}\,\mathbb{P}_\vartheta}{\mathrm{d}\mu}(x)$).
Ziel: Finde $\hat{k} \in \{1, \dots, K\}$, so dass $\mathbb{P}_{\hat{\vartheta}_{\hat{k}}}$ die wahre Verteilung $\mathbb{P}$ gut approximiert (beachte: $\mathbb{P}$ muss in keinem der Θ_k liegen).

Aus der Theorie der MLE ist bekannt, dass die Kullback-Leibler-Divergenz oder relative Entropie das natürliche Abstandsmaß dafür ist.

5.2 Definition. Für Wahrscheinlichkeitsmaße $\mathbb{P}$, $\mathbb{Q}$ auf $(X, \mathscr{F})$ heißt

$$KL(\mathbb{P}\,|\,\mathbb{Q}) = \begin{cases} \int \log(\frac{\mathrm{d}\,\mathbb{P}}{\mathrm{d}\,\mathbb{Q}}(x))\,\mathbb{P}(\mathrm{d}x) & \text{falls } \mathbb{P} \ll \mathbb{Q}, \\ \infty & \text{sonst} \end{cases}$$

Kullback-Leibler-Divergenz von $\mathbb{P}$ bezüglich $\mathbb{Q}$.

5.3 Lemma. *Es gilt:*

1. $KL(\mathbb{P}\,|\,\mathbb{Q}) \geq 0$ *und* $KL(\mathbb{P}\,|\,\mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}$.
2. *Bildet* $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ *eine natürliche* k*-dimensionale Exponentialfamilie,*

$$\frac{\mathrm{d}\,\mathbb{P}_\vartheta}{\mathrm{d}\mu}(x) = c(x)\exp(\langle\vartheta,\, T(x)\rangle_{\mathbb{R}^k} - \zeta(\vartheta)),$$

und ist $\vartheta_0 \in \mathrm{int}(\Theta)$*, so ist*

$$KL(\mathbb{P}_{\vartheta_0}\,|\,\mathbb{P}_\vartheta) = \zeta(\vartheta) - \zeta(\vartheta_0) - \langle\nabla_\vartheta\zeta(\vartheta_0),\, \vartheta - \vartheta_0\rangle, \qquad \vartheta \in \Theta.$$

Beweis.

1. O.B.d.A. $\mathbb{P} \ll \mathbb{Q}$

$$\Rightarrow KL(\mathbb{P}\,|\,\mathbb{Q}) = \int \underbrace{\log(\frac{\mathrm{d}\,\mathbb{P}}{\mathrm{d}\,\mathbb{Q}})\frac{\mathrm{d}\,\mathbb{P}}{\mathrm{d}\,\mathbb{Q}}(x)}_{=:f(\frac{\mathrm{d}\,\mathbb{P}}{\mathrm{d}\,\mathbb{Q}})}\,\mathbb{Q}(\mathrm{d}x)$$

für $f(x) = x\log(x)$ mit $f(0) = 0$.
Nun ist f konvex: $f''(x) = \frac{1}{x} > 0$. Jensen-Ungleichung liefert

$$KL(\mathbb{P}\,|\,\mathbb{Q}) = \int f(\frac{\mathrm{d}\,\mathbb{P}}{\mathrm{d}\,\mathbb{Q}})\mathrm{d}\,\mathbb{Q} \geq f(\int \frac{\mathrm{d}\,\mathbb{P}}{\mathrm{d}\,\mathbb{Q}}\mathrm{d}\,\mathbb{Q}) = 0.$$

2.

$$KL(\mathbb{P}_{\vartheta_0}\,|\,\mathbb{P}_\vartheta) = \mathbb{E}_{\vartheta_0}\left[\log\left(\frac{\frac{\mathrm{d}\,\mathbb{P}_{\vartheta_0}}{\mathrm{d}\mu}}{\frac{\mathrm{d}\,\mathbb{P}_\vartheta}{\mathrm{d}\mu}}\right)\right] = \zeta(\vartheta) - \zeta(\vartheta_0) - \langle\nabla_\vartheta\zeta(\vartheta_0),\, \vartheta - \vartheta_0\rangle.$$

□

5.4 Korollar.

$$KL(\mathrm{N}(\vartheta_0,\, \Sigma)|\,\mathrm{N}(\vartheta,\, \Sigma)) = \frac{1}{2}\langle\Sigma^{-1}(\vartheta - \vartheta_0),\, \vartheta - \vartheta_0\rangle$$

für ϑ, $\vartheta_0 \in \mathbb{R}^k$, $\Sigma \in \mathbb{R}^{k\times k}$ *invertierbar.*

Beweis. Nutze 2. □

5.5 Bemerkung. KL-Konvergenz $\Rightarrow$ TV-Konvergenz $\Rightarrow$ schwache Konvergenz.

Ziel: $KL(\mathbb{P} \,|\, \mathbb{P}_{\hat{\vartheta}_k}) \to \min_k !$

$KL(\mathbb{P} \,|\, \mathbb{P}_\vartheta) = \mathbb{E}_\mathbb{P}[\log \frac{\mathrm{d}\,\mathbb{P}}{\mathrm{d}\,\mathbb{P}_\vartheta}]$ ist für den Statistiker nicht zugänglich, da $\mathbb{P}$ unbekannt ist ($\mathbb{P}_\vartheta$ ist natürlich bekannt).

1. Vereinfachung: (nehme $\mathbb{P}$, $\mathbb{P}_\vartheta \ll \mu$ an)

$$KL(\mathbb{P} \,|\, \mathbb{P}_\vartheta) = \mathbb{E}_\mathbb{P}\left[\log\left(\frac{\frac{\mathrm{d}\,\mathbb{P}}{\mathrm{d}\mu}}{\frac{\mathrm{d}\,\mathbb{P}_\vartheta}{\mathrm{d}\mu}}\right)\right] = \underbrace{\mathbb{E}_\mathbb{P}\left[\log \frac{\mathrm{d}\,\mathbb{P}}{\mathrm{d}\mu}\right]}_{\text{unabh. von } \vartheta} - \mathbb{E}_\mathbb{P}\left[\log \underbrace{\frac{\mathrm{d}\,\mathbb{P}_\vartheta}{\mathrm{d}\mu}}_{=L(\vartheta)}\right].$$

Es reicht daher die Kullback-Diskrepanz (oder -Devianz)

$$d(\vartheta) := \mathbb{E}_\mathbb{P}[-2 \log L(\vartheta)]$$

zu minimieren.

2. „Naive" Idee: Ersetze den Erwartungswert unter $\mathbb{P}$ durch das Stichprobenäquivalent $-2 \log L(\vartheta,\, X)$.

Begründung: Für die Asymptotik betrachte das Modell einer mathematischen Stichprobe $X_1, \ldots, X_n$ vom Umfang n. Dann gilt

$$L_n(\vartheta,\, X_1, \ldots X_n) = \prod_{k=1}^n L_1(\vartheta,\, X_k),$$

und folglich

$$-2 \log L_n(\vartheta,\, X) = -2 \sum_{k=1}^n \log L_1(\vartheta,\, X_k).$$

Nach dem Gesetz der großen Zahlen (GdgZ) gilt dann

$$-\frac{2}{n} \log L_n(\vartheta,\, X) \underset{\mathbb{P}\text{-f.s.}}{\overset{n\to\infty}{\to}} -2\, \mathbb{E}_\mathbb{P}[\log L_1(\vartheta,\, X_1)]$$

(sofern der Erwartungswert endlich ist).

Betrachte das Problem

$$-2 \log L(\hat{\vartheta}_k,\, X) \to \min_k !$$

Wegen der Abhängigkeit von $\hat{\vartheta}_k$ und X lässt sich obiges asymptotisches Argument hier nicht mehr rechtfertigen. Beachte insbesondere

$$L(\hat{\vartheta}_k,\, X) = \max_{\vartheta \in \Theta_k} L(\vartheta,\, X).$$

Folgendes Beispiel zeigt das Problem auf:

$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\vartheta,\, E_k)$, $\vartheta \in \mathbb{R}^k$, $E_k = \mathrm{diag}(1, \ldots, 1) \in \mathbb{R}^{k\times k}$.

$\rightsquigarrow \hat{\vartheta}_k = \overline{X} = \frac{1}{n} \sum_{i=1}^n X_i$,

$$-2\log L_n(\vartheta,\, X) = -2\sum_{i=1}^{n}\Big(\log((2\pi)^{-k/2}) - \frac{||X_i-\vartheta||^2}{2}\Big),$$
$$-2\log L_n(\hat\vartheta_k,\, X) = -2\sum_{i=1}^{n}\Big(\log((2\pi)^{-k/2}) - \frac{||X_i-\overline{X}||^2}{2}\Big),$$
$$\mathbb{E}[-2\log L_n(\hat\vartheta_k,\, X)] = -2\sum_{i=1}^{n}\Big(\log((2\pi)^{-k/2}) - \frac{1}{2}k\frac{n-1}{n}\Big),$$

während

$$\begin{aligned} d(\vartheta) &= \mathbb{E}[-2\sum_{i=1}^{n}\Big(\log((2\pi)^{-k/2}) - \frac{||X_i-\vartheta||^2}{2}\Big)] \\ &= -2\sum_{i=1}^{n}\Big(\log((2\pi)^{-k/2}) - \frac{\sum_{j=1}^{k}\mathbb{E}[(X_{i,j}-\vartheta_j)^2]}{2}\Big). \end{aligned}$$

Im Fall $\mathbb{P} = \mathrm{N}(\vartheta^0,\, E_k)$ für ein $\vartheta^0 \in \mathbb{R}^k$

$$d(\vartheta) = -2\sum_{i=1}^{n}\Big(\log((2\pi)^{-k/2}) - \frac{1}{2}\sum_{j=1}^{k}((\vartheta_j^0-\vartheta_j)^2+1)\Big)$$
$$\Rightarrow \mathbb{E}[d(\hat\vartheta_k)] = -2\sum_{i=1}^{n}\Big(\log((2\pi)^{-k/2}) - \frac{1}{2}\sum_{j=1}^{k}(\frac{1}{n}+1)\Big).$$

Es gilt also

$$\mathbb{E}[-2\log L_n(\hat\vartheta_k,\, X) - d(\hat\vartheta_k)] = \Big(k\frac{n-1}{n} - k\frac{n+1}{n}\Big)n = -2k.$$

Das Prinzip der unverzerrten (=erwartungstreuen) Risikoschätzung (URE: unbiased risk estimation) suggeriert daher die Minimierung des Kriteriums

$$-2\log L_n(\hat\vartheta_k) + 2k.$$

5.6 Definition. Das Akaike-Informationskriterium (AIC) im Modell $(X,\, \mathscr{F},\, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta_k})$ ist definiert als

$$\mathrm{AIC}(k) := -2\log L_k(\hat\vartheta_k) + 2k.$$

Dann wird $\hat k = \mathrm{argmin}_k\, \mathrm{AIC}(k)$ bestimmt und $\hat\vartheta_{\hat k}$ ist der gemäß AIC ausgewählte Schätzer.

5.7 Satz. *Im linearen Modell $Y = X^{(k)}\beta^{(k)} + \varepsilon$ mit $\beta^{(k)} \in \mathbb{R}^k$, $X^{(k)} \in \mathbb{R}^{n\times k}$ mit vollem Rang $k \le n$ und $\varepsilon \sim \mathrm{N}(0,\, \sigma^2 E_n)$, $k = 1, \dots, K$ gilt für $\sigma > 0$ bekannt:*

- $\hat\beta_k = (X^{(k)T}X^{(k)})^{-1}X^{(k)T}Y$ *(MLE=KQ-Schätzer),*
- $\mathrm{AIC}(k) = n\log(2\pi\sigma^2) + \frac{\mathrm{RSS}}{\sigma^2} + 2k$ *mit* $\mathrm{RSS} = ||Y - X^{(k)}\hat\beta_k||^2$ *(„residual sum of squares“),*

- *gilt im wahren Modell (= unter $\mathbb{P}$) $Y = \mu + \varepsilon$ mit einem $\mu \in \mathbb{R}^n$, so folgt*

$$\mathbb{E}[\mathrm{AIC}(k)] = \mathbb{E}[d(\hat{\beta}_k)]$$

($\rightsquigarrow$ unverzerrte Schätzung der Diskrepanz).

Beweis.

(i) $\log L(\beta^{(k)}, Y) = \log((2\pi\sigma^2)^{-n/2}) - \frac{1}{2\sigma^2}||Y - X^{(k)}\beta^{(k)}||^2$
$\Rightarrow \hat{\beta}_k =$KQ-Schätzer
$\Rightarrow \mathrm{AIC}(k) = n\log(2\pi\sigma^2) + \frac{1}{\sigma^2}||Y - X^{(k)}\hat{\beta}_k||^2 + 2k,$

(ii)

$$\begin{aligned} d(\beta) &= n\log(2\pi\sigma^2) + \frac{1}{\sigma^2}\,\mathbb{E}[||Y - X^{(k)}\beta^{(k)}||^2] \\ &= n\log(2\pi\sigma^2) + \frac{1}{\sigma^2}(||\mu - X^{(k)}\beta^{(k)}||^2 + \underbrace{\mathbb{E}[||\varepsilon||^2]}_{=\sigma^2 n}) \\ &= n(\log(2\pi\sigma^2) + 1) + \frac{||\mu - X^{(k)}\beta^{(k)}||^2}{\sigma^2}. \end{aligned}$$

Nun gilt mit der Bias-Varianz-Zerlegung

$$\begin{aligned} \mathbb{E}[d(\hat{\beta}_k)] &= n(\log(2\pi\sigma^2) + 1) + \frac{\mathbb{E}[||\mu - X^{(k)}\hat{\beta}_k||^2]}{\sigma^2} \\ &= n(\log(2\pi\sigma^2) + 1) + \frac{1}{\sigma^2}(||\mu - \underbrace{X^{(k)}(X^{(k)T}X^{(k)})^{-1}X^{(k)T}}_{=\Pi^{(k)}}\mu||^2 \\ &\qquad + \mathbb{E}[||\underbrace{X^{(k)}(X^{(k)T}X^{(k)})^{-1}X^{(k)T}\varepsilon}_{\sim \mathrm{N}(0,\,\sigma^2 E_k)}||^2]) \\ &= n(\log(2\pi\sigma^2) + 1) + \frac{1}{\sigma^2}||(E_n - \Pi^{(k)})\mu||^2 + k \end{aligned}$$

und

$$\begin{aligned} \mathbb{E}[\mathrm{AIC}(k)] &= n\log(2\pi\sigma^2) + \frac{1}{\sigma^2}(||\mu - \Pi^{(k)}\mu||^2 + \mathbb{E}[||\underbrace{\varepsilon - \Pi^{(k)}\varepsilon}_{(E_n - \Pi^{(k)})\varepsilon}||^2]) + 2k \\ &= \mathbb{E}[d(\hat{\beta}_k)], \end{aligned}$$

da $(E_n - \Pi^{(k)})$ die orthogonale Projektion auf einen $(n-k)$-dimensionalen Unterraum ist und daher $\mathbb{E}[||(E_n - \Pi^{(k)})\varepsilon||^2] = (n-k)\sigma^2$ gilt.

□

5.8 Bemerkungen.

(a) Wähle also

$$\begin{aligned} \hat{k} :&= \mathrm{argmin}_k(\underbrace{n\log(2\pi\sigma^2)}_{\text{unabh. v. } k} + \frac{\mathrm{RSS}}{\sigma^2} + 2k) \\ &= \mathrm{argmin}_k(\mathrm{RSS} + 2k\sigma^2) \\ &= \mathrm{argmin}_k(||Y - X^{(k)}\hat{\beta}_k||^2 + 2k\sigma^2). \end{aligned}$$

Das Kriterium ist auch als Mallows C_p-Kriterium im linearen Modell bekannt.

(b) Wendet man AIC(k) im Fall allgemeiner Fehlerverteilungen mit $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \sigma^2\delta_{i,j}$ an, so gilt immer noch $\mathbb{E}[\mathrm{AIC}(k)] = \mathbb{E}[d(\hat{\vartheta}_k)]$, vgl. Beweis. Beachte aber, dass MLE und AIC-Kriterium der Definition nach anders aussehen würden. Diese Robustheit gegen falsch spezifizierte Fehlerverteilungen bei Normalverteilungsannahme ist Grundlage des Pseudo-Likelihood-Ansatzes.

(c) Erwartungstreue von AIC(k) sichert natürlich noch keine guten Schätzeigenschaften. Man sollte noch die Varianz untersuchen: $\mathbb{E}[\mathrm{AIC}(k) - d(\hat{\vartheta}_k)^2] \leq ?$. Zusätzlich brauchen wir zur Analyse von $\hat{k}$ eine Gleichmäßigkeit in k, zum Beispiel $\mathbb{E}[\max_{k=1,\dots,K}(\mathrm{AIC}(k) - d(\hat{\vartheta}_k))^2] \leq ?$. Dies kann man im Fall des linearen Modells konkret analysieren. Wir schauen uns später aber einen allgemeinen Ansatz an.

5.9 Satz. *Im linearen Modell $Y = X^{(p)}\beta^{(p)} + \varepsilon$ mit $\varepsilon \sim \mathrm{N}(0, \sigma^2 E_n)$, $\beta^{(p)} \in \mathbb{R}^p$ und $X^{(p)} \in \mathbb{R}^{n\times p}$ von vollem Rang $p \leq n$ gilt für $k = p+1$, $\vartheta_k = (\beta^{(p)}, \sigma^2)$ für $\sigma > 0$ ebenfalls unbekannt:*

- *MLE $\hat{\vartheta}_k = (\hat{\beta}^{(p)}, \hat{\sigma}_k^2)$:*

$$\begin{aligned}\hat{\beta}^{(p)} &= \textit{KQ-Schätzer}\\ \hat{\sigma}_k^2 &= \frac{\mathrm{RSS}}{n} = \frac{1}{n}||Y - X^{(p)}\hat{\beta}^{(p)}||^2\end{aligned}$$

- $\mathrm{AIC}(k) = n(\log(\hat{\sigma}_k^2) + \log(2\pi) + 1) + 2k$,
- $\mathbb{E}[\mathrm{AIC}(k)] = \mathbb{E}[d(\hat{\vartheta}_k)] - 2\frac{k(k+1)}{n-k-1} = \mathbb{E}[d(\hat{\vartheta}_k)](1 + O(n^{-2}))$,
  im Fall des wahren Modells $Y = X^{(p)}\beta_0^{(p)} + \varepsilon$ mit $k = p+1$, $\varepsilon \sim \mathrm{N}(0, \sigma_0^2 E_n)$.

5.10 Bemerkung. Hier könnte und sollte man das AIC-Kriterium modifizieren, um Erwartungstreue zu sichern:

$$\widehat{\mathrm{AIC}}(k) = n(\log(\hat{\sigma}_k^2) + \log(2\pi) + 1) + 2k\Big(1 + \frac{k+1}{n-k-1}\Big).$$

Für viele weitere Modelle werden konkrete Korrekturen von AIC bei kleinem n empfohlen. Asymptotisch ($n \to \infty$) sind diese Korrekturen vernachlässigbar, wie man unter recht allgemeinen Voraussetzungen zeigen kann.

Beweis.

(a)

$$\begin{aligned}-2\log L_n(\vartheta) &= -2\log(L_n((\beta^{(p)}, \sigma^2)) = -2(\log((2\pi\sigma^2)^{-\frac{n}{2}}) - \frac{1}{2\sigma^2}||Y - X^{(p)}\beta^{(p)}||^2)\\ &= n\log(2\pi\sigma^2) + \frac{1}{\sigma^2}||Y - X^{(p)}\beta^{(p)}||^2\end{aligned}$$

MLE: $\hat{\beta}_k$ ist KQ-Schätzer.
$\frac{\partial}{\partial\sigma^2}(-2\log L_n(\vartheta)) = \frac{n}{\sigma^2} - \frac{1}{\sigma^4}||Y - X^{(p)}\beta^{(p)}||^2 \stackrel{!}{=} 0$
mit $\beta^{(p)} = \hat{\beta}_k$ gilt $\hat{\sigma}^2 = \frac{\mathrm{RSS}}{n}$, es folgt die Form des AIC.

(b)

$$d(\vartheta) = \mathbb{E}[-2\log L_n(\vartheta)] = n\log(2\pi\sigma^2) + \frac{1}{\sigma^2}(||\mu - X^{(p)}\beta^{(p)}||^2 + n\sigma_0^2)$$

für $Y = \mu + \varepsilon$, $\varepsilon \sim \mathrm{N}(0, \sigma_0^2 E_n)$.

$$\mathbb{E}[d(\hat{\vartheta}_k)] = n\,\mathbb{E}[\log(2\pi\hat{\sigma}_k^2)] + \mathbb{E}[\frac{1}{\hat{\sigma}_k^2}(||\mu - X^{(p)}\hat{\beta}_k||^2 + n\sigma_0^2)],$$
$$\mathbb{E}[\mathrm{AIC}(k)] = n\,\mathbb{E}[\log(2\pi\hat{\sigma}_k^2)] + n + 2k.$$

Es folgt:

$$\begin{aligned}\mathbb{E}[\mathrm{AIC}(k)] &= \mathbb{E}[d(\hat{\vartheta}_k)] + n + 2k - \mathbb{E}[\frac{1}{\hat{\sigma}_k^2}(||\mu - X^{(p)}\hat{\beta}_k||^2 + n\sigma_0^2)] \\ &= \mathbb{E}[d(\hat{\vartheta}_k)] + n + 2k - \mathbb{E}\left[\frac{||\mu - \Pi_p Y||^2 + n\sigma_0^2}{\frac{1}{n}(||(E_n - \Pi_p)Y||^2)}\right].\end{aligned}$$

Da $(E_n - \Pi_p)$ Orthogonalprojektion auf $\mathrm{ran}(X^{(p)})^\perp$ ist, sind $\Pi_p Y$ und $(E_n - \Pi_p)Y$ unkorreliert, und damit unabhängig.
Es folgt:

$$\begin{aligned}\mathbb{E}[\mathrm{AIC}(k)] &= \mathbb{E}[d(\hat{\vartheta}_k)] + n + 2k - \mathbb{E}[||\mu - \Pi_p Y||^2 + n\sigma_0^2]\,\mathbb{E}\left[\frac{n}{||(E_n - \Pi_p)Y||^2)}\right] \\ &= \mathbb{E}[d(\hat{\vartheta}_k)] + n + 2k - (\underbrace{||\mu - \Pi_p\mu||^2}_{=0\text{ , da }\mu = X^{(p)}\beta_0^{(p)}} + p\sigma_0^2 + n\sigma_0^2)\,\mathbb{E}\left[\frac{n}{||(E_n - \Pi_p)Y||^2)}\right]\end{aligned}$$

Problem: Was ist $\mathbb{E}\left[\frac{1}{||(E_n - \Pi_p)Y||^2)}\right]$?
Der Nenner hat nichtzentrale χ^2-Verteilung, falls $\mu \neq \Pi_p\mu$.
Unter der Annahme $\mu = X^{(p)}\beta_0^{(p)}$ gilt

$$||(E_n - \Pi_p)Y||^2 = ||(E_n - \Pi_p)\varepsilon||^2 \sim \sigma_0^2\chi^2(n-p).$$

Berechne für $Z \sim \chi^2(n-p)$ daher

$$\mathbb{E}\left[\frac{1}{Z}\right] = \int_0^\infty \frac{1}{z}\frac{2^{-\frac{n-p}{2}}}{\Gamma(\frac{n-p}{2})} z^{\frac{n-p}{2}-1} e^{-\frac{z}{2}}\,\mathrm{d}z = \frac{1}{n-p-2}$$

Es folgt

$$\mathbb{E}\left[\frac{n}{||(E_n - \Pi_p)Y||^2}\right] = \frac{n}{\sigma_0^2(n-p-2)}.$$

Einsetzen liefert die Behauptung.
Beachte nun noch, dass $\mathbb{E}[d(\hat{\vartheta}_k)] \geq cn$ gilt, so dass der Fehlerterm die Ordnung $\mathbb{E}[d(\hat{\vartheta}_k)]O(n^{-2})$ besitzt.

□

5.11 Definition. Das wahre Modell P sei in der Familie $(P_\vartheta)_{\vartheta\in\Theta_k}$ für ein $k \in \{1,\dots,K\}$ enthalten, k_0 sei das minimale k mit dieser Eigenschaft. Dann heißt eine Modellwahl konsistent, falls $P(\hat{k} = k_0) \to 1$, wenn der Stichprobenumfang n gegen ∞ konvergiert.
Ist das wahre Modell nicht in $(P_\vartheta)_{\vartheta\in\Theta_k}$ enthalten, so heißt eine Modellwahl asymptotisch effizient, falls das Modell den Vorhersagefehler für $n \to \infty$ minimiert, das heißt im linearen Modell

$$\frac{\mathbb{E}[||X^{(\hat{k})}\hat{\beta}^{(\hat{k})} - \mu||^2]}{\min\limits_{k=1,\dots,K}\mathbb{E}[[[X^{(k)}\hat{\beta}^{(k)} - \mu||^2]} \overset{n\to\infty}{\to} 1.$$

5.3 Das Bayessche Informationskriterium (BIC)

5.12 Definition. In der Situation des vorigen Abschnitts setze

$$\mathrm{BIC}(k) = -2\log L_k(\hat{\vartheta}_k) + log(n)k,$$

wobei wir ein Produktmodell vom Stichprobenumfang n ab jetzt voraussetzen.

5.13 Bemerkung. Ab $n \geq 8$ gilt $\mathrm{BIC}(k) \geq \mathrm{AIC}(k)$, das heißt $\hat{k} = \mathrm{argmin}_k \mathrm{BIC}(k)$ führt zu einer Auswahl von kleineren Modellen als AIC.

Bayessche Herleitung: Im Modell $(X, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta_k})$, $\Theta_k \subseteq \mathbb{R}^k$ sei π_k eine a-priori-Lebesguedichte für ϑ auf Θ_k, $k = 1,\dots,K$. Die Modelle $\Theta_1,\dots,\Theta_K$ seien gemäß einer nichtinformativen a-priori-Verteilung jeweils mit Wahrscheinlichkeit $\frac{1}{K}$ gewählt.
Die a-posteriori-Verteilung von (Modell k, Parameter $\vartheta \in \Theta_k$) besitzt nach der Bayesformel die Dichte (bezüglich Zählmaß $\otimes$ Lebesguemaß)

$$f(k,\, \vartheta_k|X) = \frac{\frac{1}{K}\pi_k(\vartheta_k)L_k(\vartheta_k,\, X)}{\sum_{k=1}^K \int \frac{1}{K}\pi_{k'}(\vartheta'_{k'})L_k(\vartheta'_{k'},\, X)\mathrm{d}\vartheta'_{k'}} = c_X\pi_k(\vartheta_k)L_k(\vartheta_k,\, X).$$

Durch Ausintegrieren ergibt sich die a-posteriori-Wahrscheinlichkeit für das Modell $k = \kappa$

$$\mathbb{P}(k = \kappa|X) = \int c_X\pi_\kappa(\vartheta_\kappa)L_\kappa(\vartheta_\kappa,\, X)\mathrm{d}\vartheta_\kappa.$$

Im Bayesschen Sinne sollte nun das Modell $\hat{\kappa}$ gewählt werden, welches $\mathbb{P}(k = \kappa\,|X)$ über $\kappa = 1,\dots,K$ maximiert (MAP: maximum-a-posteriori).
Wir führen asymptotische Näherungen durch im Modell von n i.i.d. Beobachtungen mit $n \to \infty$. Dann gilt:

$$L_k(\vartheta_k,\, X) = \prod_{i=1}^n L_k^{(1)}(\vartheta_k,\, X_i) = \exp(\sum_{i=1}^n l_k(\vartheta_k,\, X_i)).$$

Beachte: $\max\limits_{\vartheta_k\in\Theta_k} L_k(\vartheta_k,\, X) = \exp(\sum\limits_{i=1}^n l_k(\hat{\vartheta}_k,\, X_i))$.
Ziel: $-2\log(\mathbb{P}(k = \kappa|X)) \to \min\limits_\kappa! \Leftrightarrow -2\log(\int L_k(\vartheta_k,\, X)\pi_k(\vartheta_k)\mathrm{d}\vartheta_k) \to \min\limits_\kappa!$

Entwickle $L_\kappa(\cdot\,, X)$ um $\hat{\vartheta}_\kappa$ (Taylor):

$$\begin{aligned} L_\kappa(\vartheta_\kappa, X) &\approx \exp(\sum_{i=1}^{n}(l_\kappa(\hat{\vartheta}_\kappa, X_i) + \underbrace{\nabla_\vartheta l_\kappa(\hat{\vartheta}_\kappa, X_i)(\vartheta_\kappa - \hat{\vartheta}_\kappa)}_{=0,\text{ falls } (\vartheta_\kappa \mapsto l_\kappa(\vartheta_\kappa)) \in C^2} \\ &\qquad + \frac{1}{2}\langle \nabla^2_\vartheta l_\kappa(\hat{\vartheta}_\kappa, X_i)(\vartheta_\kappa - \hat{\vartheta}_\kappa), \vartheta_\kappa - \hat{\vartheta}_\kappa\rangle_{\mathbb{R}^\kappa})) \\ &= L_\kappa(\hat{\vartheta}_\kappa, X)\exp(-\frac{n}{2}\langle I_n(\hat{\vartheta}_\kappa)(\vartheta_\kappa - \hat{\vartheta}_\kappa), \vartheta_\kappa - \hat{\vartheta}_\kappa\rangle) \end{aligned}$$

mit $I_n(\vartheta) := \frac{1}{n}\sum_{i=1}^{n} \nabla^2_\vartheta l_\kappa(\vartheta, X_i) =$ „beobachtete Fisher-Informationsmatrix“. Es gilt nach dem Gesetz der großen Zahlen:

$$I_n(\vartheta) \overset{\text{f.s.}}{\to} \mathbb{E}[-\nabla^2_\vartheta l_\kappa(\vartheta, X_1)] =: I(\vartheta).$$

Unter Regularitätsannahmen an l_κ:

$$I_n(\hat{\vartheta}_\kappa) \overset{\mathbb{P}}{\to} I(\vartheta_{\kappa,0}), \qquad \vartheta_{\kappa,0} := \operatorname{argmin}_{\vartheta_\kappa \in \Theta_\kappa} d(\vartheta_\kappa).$$

$$\begin{aligned} &\rightsquigarrow \text{minimiere } -2\log(\int_{\mathbb{R}^k} \exp(-\frac{n}{2}\langle I(\vartheta_{\kappa,0})(\vartheta_\kappa - \hat{\vartheta}_\kappa), \vartheta_\kappa - \hat{\vartheta}_\kappa\rangle) L_\kappa(\hat{\vartheta}_\kappa)\pi(\vartheta_\kappa)\mathrm{d}\vartheta_\kappa) \\ &= -2\log L_\kappa(\hat{\vartheta}_\kappa) - 2\log(\int \varphi_{\hat{\vartheta}_\kappa, (nI)^{-1}(\vartheta_{\kappa,0})}(\vartheta_\kappa)(2\pi)^{\kappa/2}\det(nI(\vartheta_{\kappa,0}))^{-1/2}\pi(\vartheta_\kappa)\mathrm{d}\vartheta_\kappa) \\ &\approx -2\log L_\kappa(\hat{\vartheta}_\kappa) - 2\log((2\pi)^{\kappa/2}n^{-\kappa/2}\det(I(\vartheta_{\kappa,0}))^{-1/2}\pi_\kappa(\hat{\vartheta}_\kappa)) \\ &= -2\log L_\kappa(\hat{\vartheta}_\kappa) + \kappa\log(n) + \log(\det(I(\vartheta_{\kappa,0}))) - 2\log\pi_\kappa(\hat{\vartheta}_\kappa). \end{aligned}$$

($\varphi_{\mu,\Sigma}$ bezeichnet die Dichte von $\mathrm{N}(\mu, \Sigma)$). Für $n \to \infty$ bleiben die Terme $\kappa\log(2\pi)$, $\log(\det(I(\vartheta_{\kappa,0})))$ und $2\log\pi_\kappa(\hat{\vartheta}_\kappa)$ beschränkt, während $L_\kappa(\hat{\vartheta}_\kappa)$ und $\kappa\log(n)$ in n wachsen. Dies führt auf die Wahl von BIC als

$$\mathrm{BIC}(\kappa) := -2\log L_\kappa(\hat{\vartheta}_\kappa) + \kappa\log(n).$$

Man kann häufig zeigen, dass BIC ein konsistentes, aber nicht asymptotisch effizientes Modellwahlkriterium ist (genau im Gegensatz zu AIC).

5.4 Allgemeine Analyse der penalisierten Kleinste-Quadrate-Methode

(vgl. Pascal Massart: Concentration inequalities and model selection)

Motivation: Im linearen Modell $Y = X^{(k)}\beta^{(k)} + \varepsilon$ mit $\beta^{(k)} \in \mathbb{R}^k$, $X^{(k)} \in \mathbb{R}^{n\times k}$ mit vollem Rang und $\varepsilon \sim \mathrm{N}(0, \sigma^2 E_n)$ haben wir gesehen, dass

$$\hat{k}^{\mathrm{AIC}} = \operatorname{argmin}_k(\mathrm{RSS} + 2\sigma^2 k) = \operatorname{argmin}_k(\|Y - X^{(k)}\hat{\beta}^{(k)}\|^2 + 2\sigma^2 k)$$

gilt und daher auch

$$\hat{k}^{\mathrm{BIC}} = \operatorname{argmin}_k(\mathrm{RSS} + \log(n)\sigma^2 k) = \operatorname{argmin}_k(\|Y - X^{(k)}\hat{\beta}^{(k)}\|^2 + \log(n)\sigma^2 k).$$

Setzt man $\mu_k := X^{(k)}\beta^{(k)} \in \mathrm{ran}(X^{(k)}) \subseteq \mathbb{R}^n$ sowie $\hat{\mu}_k := X^{(k)}\hat{\beta}^{(k)} = \Pi_k Y$, Π_k Orthogonalprojektion auf $\mathrm{ran}(X^{(k)})$, so gilt also im k-ten Modell:

$$Y = \mu_k + \varepsilon \text{ mit } \mu_k \in S_k := \mathrm{ran}(X^{(k)}),$$
$$\hat{k} = \mathrm{argmin}_k(\underbrace{\|Y - \Pi_{S_k} Y\|^2 + \mathrm{Pen}(\dim(S_k))}_{(*)})$$

mit einem Strafterm $\mathrm{Pen}(\dim(S_k))$, der nur von der Dimension von S_k, n und σ^2 abhängt, z.B. $\mathrm{Pen}(d) = 2\sigma^2 d$ für AIC oder $\mathrm{Pen}(d) = \log(n)\sigma^2 d$ für BIC. Das Kriterium in $(*)$ heißt penalisierter Kleinste-Quadrate-Ansatz.
Wir werden eine sogenannte Orakelungleichung der folgenden Form beweisen:

$$\|\hat{\mu}_{\hat{k}} - \mu\|^2 \leq C \min(\|\hat{\mu}_k - \mu\|^2 + \mathrm{Pen}(\dim(S_k))) + \text{ (kleiner) Fehlerterm.}$$

Dies zeigt, dass der Vorhersagefehler (= linke Seite) bei (korrekter) Modellwahl maximal ein festes Vielfaches des Orakelvorhersagefehlers (bei Kenntnis des besten Modells) und des Strafterms ist.
$\rightsquigarrow$ asymptotische Effizienz
Allgemeiner: lineare Unterräume $S_m \subseteq \mathbb{R}^n$, $\dim S_m = d_m$, $m = 1, \ldots, M$,

$$\hat{m} = \mathrm{argmin}_{m=1,\ldots,M}(\|Y - \underbrace{\hat{S}_m}_{\Pi_m Y}\|^2 + \mathrm{Pen}(d_m)).$$

5.14 Theorem. *Für lineare Unterräume $S_m \subseteq \mathbb{R}^n$, $m = 1, \ldots, M$ mit $\dim(S_m) = d_m$ und für das (wahre) Modell $Y = \mu + \varepsilon$, $\mu \in \mathbb{R}^n$, $\varepsilon \sim \mathrm{N}(0, \sigma^2 E_n)$ setze $\hat{\mu}_m = \Pi_m Y$ mit $\Pi_m : \mathbb{R}^n \to S_m$ Orthogonalprojektion (das heißt, $\hat{\mu}_m$ ist der KQ-Schätzer), $\mu_m = \Pi_m \mu$ und wähle*

$$\hat{m} = \mathrm{argmin}_m(\|Y - \hat{\mu}_m\|^2 + \mathrm{Pen}(d_m))$$

mit $\mathrm{Pen}(d_m) \geq K\sigma^2(d_m + 1)$ für ein $K > 1$. Dann gilt für beliebiges $\kappa \in (0, \sqrt{K} - 1)$, $\tau > 0$ mit Wahrscheinlichkeit $1 - \varepsilon_{\kappa,\tau}$,

$$\varepsilon_{\kappa,\tau} = \Big(\sum_{m=1}^{M} e^{-d_m \kappa^2/2}\Big) e^{-\tau/2},$$

die Orakelungleichung:

$$\|\hat{\mu}_{\hat{m}} - \mu\|^2 \leq C(K, \kappa)(\min_{1 \leq m \leq M}(\|\mu_m - \mu\|^2 + \mathrm{Pen}(d_m)) + \sigma^2\tau)$$

für beliebiges μ mit einer Konstanten $C(K, \kappa) > 0$
($C(K, \kappa) \to \infty$ für $\kappa \uparrow \sqrt{K} - 1$).

5.15 Bemerkung. Das m^*, mit dem das Minimum auf der rechten Seite angenommen wird, nennt man auch Orakelmodell. Wegen

$$\mathbb{E}[\|\hat{\mu}_m - \mu\|^2] = \|\mu_m - \mu\|^2 + \sigma^2 d_m$$

ist für K nahe 1

$$\|\mu_m - \mu\|^2 + \mathrm{Pen}(d_m) \approx \mathbb{E}[\|\hat{\mu}_m - \mu\|^2],$$

so dass $\|\mu_{m^*} - \mu\|^2 + \mathrm{Pen}(d_{m^*})$ nahe am Orakelfehler $\min_m \mathbb{E}[\|\hat{\mu}_m - \mu\|^2]$ liegt. Mit $\tau \sim d_{m^*}$ ist der „Restterm“ $\sigma^2\tau$ von der Ordnung $\mathrm{Pen}(d_{m^*})$ (oder kleiner). Dann gilt

$$\varepsilon_{\kappa,\tau} = \underbrace{\Big(\sum_m e^{-d_m\kappa^2/2}\Big)}_{(*)} e^{-d_{m^*}/2}.$$

Falls das Orakelmodell asymptotisch groß ist ($d_{m^*} \to \infty$) und der Term $(*)$ beschränkt bleibt, so gilt die Orakelungleichung mit sehr großer Wahrscheinlichkeit („with overwhelming probability“).

Beweis. $m^* \in \{1, \dots, M\}$ sei beliebig.

1. Per definitionem gilt:

   $$\|Y - \hat{\mu}_{\hat{m}}\|^2 + \mathrm{Pen}(d_{\hat{m}}) \le \|Y - \hat{\mu}_{m^*}\|^2 + \mathrm{Pen}(d_{m^*}) \le \|Y - \mu_{m^*}\|^2 + \mathrm{Pen}(d_{m^*}).$$

   Benutze für $x \in \mathbb{R}^n$:

   $$\begin{aligned}
   &\|Y - x\|^2 = \|\mu - x\|^2 + \|\varepsilon\|^2 - 2\langle \varepsilon,\, x - \mu\rangle \\
   &\Rightarrow \|\mu - \hat{\mu}_{\hat{m}}\|^2 + \|\varepsilon\|^2 - 2\langle \varepsilon,\, \hat{\mu}_{\hat{m}} - \mu\rangle + \mathrm{Pen}(d_{\hat{m}}) \\
   &\qquad\qquad \le \|\mu - \mu_{m^*}\|^2 + \|\varepsilon\|^2 - 2\langle \varepsilon,\, \mu_{m^*} - \mu\rangle + \mathrm{Pen}(d_{m^*}) \\
   &\Rightarrow \|\hat{\mu}_{\hat{m}} - \mu\|^2 \le \|\mu_{m^*} - \mu\|^2 + \mathrm{Pen}(d_{m^*}) + 2\langle \varepsilon,\, \hat{\mu}_{\hat{m}} - \mu_{m^*}\rangle - \mathrm{Pen}(d_{\hat{m}}).
   \end{aligned}$$

   Notiz: $\hat{\mu}_{\hat{m}} - \mu_{m^*} \in \mathrm{span}(S_{\hat{m}}, \mu_{m^*}) =: S^*_{\hat{m}}$ mit $\dim S^*_{\hat{m}} \le d_{\hat{m}} + 1$. Für $\hat{m} = m'$ deterministisch

   $$\langle \varepsilon,\, \hat{\mu}_{\hat{m}} - \mu_{m^*}\rangle^2 \le \|\hat{\mu}_{\hat{m}} - \mu_{m^*}\|^2 \underbrace{\Big(\sup_{s \in S^*_{\hat{m}}} \frac{\langle \varepsilon,\, s\rangle}{\|s\|}\Big)^2}_{= \|\Pi_{S^*_{m'}}\varepsilon\|^2}$$

   und $\mathbb{E}[\|\Pi_{S^*_{m'}}\varepsilon\|^2] \le \sigma^2(d_{m'} + 1)$.

2. Sei $Z \sim \chi^2(p)$ verteilt, das heißt, $Z = \sum_{i=1}^p X_i^2 = \|X\|^2$ mit $X = (X_1, \dots, X_p)^T \sim \mathbb{N}(0,\, E_p)$. Für $\rho > 1$:

   $$\begin{aligned}
   &\mathbb{P}(Z \ge \rho p) \le \frac{\mathbb{E}[e^{\lambda Z}]}{e^{\lambda\rho p}} \overset{(\ddot{\mathrm{U}})}{=} (1 - 2\lambda)^{-p/2} e^{-\lambda\rho p} \overset{\lambda = \frac{\rho-1}{2\rho}}{=} e^{-\frac{p}{2}(\rho - 1 - \log\rho)} \\
   &\Rightarrow \mathbb{P}(Z \ge \underbrace{\Big(1 + \kappa + \sqrt{\frac{\tau}{p}}\Big)^2}_{=\rho} p) \\
   &\qquad \le \exp\Big(-\frac{p}{2}\Big(\Big(\kappa + \sqrt{\frac{\tau}{p}}\Big)^2 + 2\Big(\kappa + \sqrt{\frac{\tau}{p}}\Big) - 2\log\Big(1 + \kappa + \sqrt{\frac{\tau}{p}}\Big)\Big)\Big) \\
   &\qquad \overset{\log(1+h) \le h}{\le} \exp\Big(-\frac{p}{2}\Big(\kappa + \sqrt{\frac{\tau}{p}}\Big)^2\Big) \le \exp\Big(-\frac{p}{2}\kappa^2 - \frac{\tau}{2}\Big).
   \end{aligned}$$

3. Schreibe

$$\langle \varepsilon,\, \hat\mu_{\hat m} - \mu_{m^*}\rangle \le \|\hat\mu_{\hat m} - \mu_{m^*}\| \sup_{s\in S^*_{\hat m}} \frac{\langle \varepsilon,\, s\rangle}{\|s\|}$$
$$\le \|\hat\mu_{\hat m} - \mu_{m^*}\|\sigma((1+\kappa)\sqrt{d_{m^*}+1} + \max_{1\le m\le M}(\frac{1}{\sigma}\|\Pi^*_m\varepsilon\| - (1+\kappa)\sqrt{d_m+1}))$$

(Π^*_m ist die Projektion auf S^*_m). Beachte $\|\Pi_m\varepsilon\| = \sigma\sqrt{Z_m}$ mit $Z_m \sim \chi^2(d_m+1)$. Es folgt

$$\mathbb{P}(\max_{1\le m\le M}(\underbrace{\|\Pi_m\varepsilon\|\frac{1}{\sigma}}_{\sim\sqrt{Z_m}} - (1+\kappa)\sqrt{d_m+1}) \ge \sqrt{\tau})$$
$$\overset{\text{Bonferroni}}{\le} \sum_{m=1}^M \mathbb{P}(\underbrace{\sqrt{Z_m} \ge (1+\kappa)\sqrt{d_m+1}+\sqrt{\tau}}_{\Leftrightarrow Z_m \ge ((1+\kappa)\sqrt{d_m+1}+\sqrt{\tau})^2})$$
$$\le \sum_{m=1}^M \exp\Big(-\frac{\kappa^2}{2}(d_m+1) - \frac{\tau}{2}\Big) = \varepsilon_{\kappa,\tau}.$$

Somit gilt mit Wahrscheinlichkeit $\ge 1-\varepsilon_{\kappa,\tau}$:

$$\max_m \Big(\frac{\|\Pi_m\varepsilon\|}{\sigma} - (1+\kappa)\sqrt{d_m+1}\Big) < \sqrt{\tau}.$$

Erhalte auf diesem Ereignis

$$\langle \varepsilon,\, \hat\mu_{\hat m} - \mu_{m^*}\rangle \le \|\hat\mu_{\hat m} - \mu_{m^*}\|\sigma((1+\kappa)\underbrace{\sqrt{d_{\hat m}+1}}_{\le\sqrt{\frac{\mathrm{Pen}(d_{\hat m})}{K\sigma^2}}} + \sqrt{\tau}).$$

4. Wir haben jetzt:

$$\|\hat\mu_{\hat m} - \mu\|^2 \le \|\mu_{m^*} - \mu\|^2 + \mathrm{Pen}(d_{m^*}) - \mathrm{Pen}(d_{\hat m})$$
$$+ 2\underbrace{\|\hat\mu_{\hat m} - \hat\mu_{m^*}\|}_{\le\|\hat\mu_{\hat m}-\mu\|+\|\hat\mu_{m^*}-\mu\|}\Big(\frac{(1-\kappa)}{\sqrt{K}}\sqrt{\mathrm{Pen}(d_{\hat m})} + \sigma\sqrt{\tau}\Big).$$

Benutze $2AB \le \eta A^2 + \eta^{-1}B^2$, $\eta > 0$, und erhalte

$$\|\hat\mu_{\hat m} - \mu\|^2 \le (1+\eta_2^{-1}+\eta_4^{-1})\|\mu_{m^*} - \mu\|^2 + \mathrm{Pen}(d_{m^*}) + (\eta_3+\eta_4)\tau\sigma^2$$
$$+ \Big((\eta_1+\eta_2)\frac{(1+\kappa)^2}{K} - 1\Big)\mathrm{Pen}(d_{\hat m}) + (\eta_1^{-1}+\eta_3^{-1})\|\hat\mu_{\hat m} - \mu\|^2.$$

Wähle $\eta_1^{-1}+\eta_3^{-1} < 1$, $\eta_1+\eta_2 = \frac{K}{(1+\kappa)^2} > 1$, $\eta_4 = 1$. Dann

$$\|\hat\mu_{\hat m} - \mu\|^2 \le C(\eta_1,\dots,\eta_4)(\|\mu_{m^*} - \mu\|^2 + \mathrm{Pen}(d_{m^*}) + \sigma^2\tau).$$

□

5.16 Beispiel. Variablenselektion (subset selection)

$$Y_i = \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i \qquad i = 1, \ldots, n, \qquad \varepsilon \sim \mathrm{N}(0, \sigma^2 E_n)$$

Ziel: Wähle aus $\{x_{\cdot 1}, \ldots, x_{\cdot k}\}$ die aktiven Kovariablen aus.
Also $S_m = \mathrm{span}\{x_{\cdot j} | j \in J_m\}$ mit $J_m \subseteq \{1, \ldots, k\}$ sowie $d_m = |J_m|$, $M = 2^k$.
Wie in der Übung gezeigt, wählt man $\kappa^2 \sim \log k \sim \log\log M$ und erhält damit $\mathrm{Pen}(d_m) \sim \sigma^2 d_m \log k$.

5.5 Die Lassomethode

Wir betrachten das multiple Regressionsmodell

$$Y_i = \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i \qquad i = 1, \ldots \ n, \quad \varepsilon \sim \mathrm{N}(0, \sigma^2 E_n).$$

Ist k sehr groß (im Vergleich zu n), dann wird der Kleinste-Quadrate-Schätzer sehr instabil (die Varianz wächst), und es kann bei vielen Parametern zu Interpolation der Daten kommen.
Wir streben eine Modellierung der Daten durch eine sparsame Darstellung (sparsity) mit wenigen β_j an. Dafür führen wir einen Strafterm ein, der bewirkt, dass viele β_j nullgesetzt werden.

$$\tilde{\beta} := \mathrm{argmin}_{b \in \mathbb{R}^k} \{\|Y - Xb\|^2 + \lambda |b|_{l^0}\}$$

mit $|b|_{l^0} = |\{j | b_j = 0\}|$ und $\lambda > 0$.
Dies ist jedoch ein nichtkonvexes Optimierungsproblem, welches für große k NP-vollständig , also algorithmisch quasi unmöglich ist.
Lösung: konvexe Relaxation.

5.17 Definition. Im obigen Modell ist der Lasso-Schätzer $\hat{\beta}$ definiert als

$$\hat{\beta} := argmin_{b \in \mathbb{R}^k} \{\|Y - Xb\|^2 + \lambda |b|_{l^1}\}$$

mit $|b|_{l^1} = \sum_{j=1}^{k} |b_j|$ und $\lambda > 0$.

5.18 Lemma (Fundamentale Ungleichung)**.** *Für jedes $\beta^* \in \mathbb{R}^k$ gilt im obigen Modell*

$$\|X\hat{\beta} - X\beta\|^2 + \lambda |\hat{\beta}|_{l^1} \leq \|X\beta^* - X\beta\|^2 + 2\langle \varepsilon, X(\hat{\beta} - \beta^*)\rangle + \lambda |\beta^*|_{l^1}.$$

Beweis. Aus

$$\|Y - X\hat{\beta}\|^2 + \lambda |\hat{b}|_{l^1} \leq \|Y - X\beta^*\|^2 + \lambda |\beta^*|_{l^1}$$

folgt

$$\|X\beta\|^2 - 2\langle Y, X\hat{\beta}\rangle + \|X\hat{\beta}\|^2 + \lambda |\hat{\beta}|_{l^1} \leq \|X\beta\|^2 - 2\langle Y, X\beta^*\rangle + \|X\beta^*\|^2 + \lambda |\beta^*|_{l^1}.$$

Aus $Y = X\beta + \varepsilon$ folgt die Behauptung.

□

Gäbe es den stochastischen Fehlerterm nicht, würde das Lemma bereits implizieren, dass der Vorhersagefehler von $\hat{\beta}$ mindestens so gut ist wie der des Orakelschätzers β^* bezüglich l^1-penalisiertem Verlust.
Den stochastischen Fehlerterm schätzen wir ab durch

$$|\langle \varepsilon, X(\hat{\beta}-\beta^*)\rangle| = |\langle X^T\varepsilon, \hat{\beta}-\beta^*\rangle| \leq (\sum_{i=1}^{k} |\hat{\beta}_i - \beta_i^*|) \max_{j=1,\dots,k} |(X^T\varepsilon)_j|,$$

wobei $X^T\varepsilon \sim \mathrm{N}(0, \sigma^2 X^T X)$.

5.19 Satz. *Für $\beta^* \in \mathbb{R}^k$ sei $S_* = \{j | \beta_j^* \neq 0\}$ und allgemein sei $|b|_S = \sum_{j\in S} |b_j|$ für $b \in \mathbb{R}^k$, $S \subseteq \{1,\dots,k\}$.*
Auf dem Ereignis

$$\mathscr{G} := \{ \max_{j=1,\dots,k} |(X^T\varepsilon)_j| \leq \frac{\lambda}{8}\} \in \mathcal{B}_{\mathbb{R}^n}$$

gilt

$$\|X(\hat{\beta}-\beta)\|^2 + \lambda|\hat{\beta}|_{S_*^c} \leq \frac{5}{3}\|X(\beta^*-\beta)\|^2 + \frac{25}{12}\frac{\lambda^2|S_*|}{\varphi_x},$$

wenn $\varphi_x = \lambda_{\min}(X^TX)$ der kleinste Eigenwert von X^TX ist und wir annehmen, dass $\varphi_x > 0$.

Beweis. Auf $\mathscr{G}$ gilt:

$$4\|X(\hat{\beta}-\beta)\|^2 + 4\lambda|\hat{\beta}|_{l^1} \leq \lambda|\hat{\beta}-\beta^*|_{l^1} + 4\|X(\beta^*-\beta)\|^2 + 4\lambda|\beta^*|_{l^1}.$$

Aus $|\beta|_{l^1} = |\beta|_{S_*^c} + |\beta|_{S_*}$ und $|\beta^*|_{S_*^c} = 0$ folgt

$$4\|X(\hat{\beta}-\beta)\|^2 + 4\lambda|\hat{\beta}|_{S_*^c} \leq -4\lambda|\hat{\beta}|_{S_*} + \lambda|\hat{\beta}-\beta^*|_{l^1} + 4\|X(\beta^*-\beta)\|^2 + 4\lambda|\beta^*|_{S_*}.$$

Mit der inversen Dreiecksungleichung folgt

$$4\|X(\hat{\beta}-\beta)\|^2 + 3\lambda|\hat{\beta}|_{S_*^c} \leq 5\lambda|\hat{\beta}-\beta^*|_{S_*} + 4\|X(\beta^*-\beta)\|^2.$$

Um $|\hat{\beta}-\beta^*|_{S_*}$ abzuschätzen, verwenden wir $|\beta|_S^2 \overset{\text{CSU}}{\leq} |S|\sum_{j\in S}|\beta_j|^2$ und $\|X\beta\|^2 = \langle X^TX\beta, \beta\rangle \geq \varphi_x\|\beta\|^2$, wobei nach Voraussetzung gilt

$$\varphi_x = \min_{b\in\mathbb{R}^k}\Big\{\frac{\langle X^TXb, b\rangle}{\langle b, b\rangle}\Big\} > 0.$$

Damit folgt

$$\|\hat{\beta}-\beta^*\|_{S_*} \leq \sqrt{\frac{|S_*|}{\varphi_x}}\|X(\hat{\beta}-\beta^*)\|.$$

Mit $\|X(\hat{\beta}-\beta^*)\|_{S_*} \leq \|X(\hat{\beta}-\beta)\| + \|X(\beta^*-\beta)\|$ und $AB \leq (A/2)^2 + B^2$ für A, $B \in \mathbb{R}$ folgt

$$\|X(\hat{\beta}-\beta)\|^2 + \lambda|\hat{\beta}|_{S_*^c} \leq \frac{5}{3}\|X(\beta^*-\beta)\|^2 + \frac{25}{12}\frac{\lambda^2|S_*|}{\varphi_x}.$$

□

Der Satz kann als verallgemeinerte Version der Bias-Varianz-Zerlegung aufgefasst werden. Der stochastische Fehler wächst in der Dimension der aktiven Menge S_k, aber fällt in λ. Schließlich müssen wir $\lambda > 0$ so wählen, dass $P(\mathscr{G})$ klein ist.

5.20 Lemma. *Für $\lambda = 8\tau\sigma\sqrt{log(k)\lambda_{\max}(X^TX)}$ mit $\tau > 2$ gilt:*

$$P(\mathscr{G}^c) \leq \sqrt{2}\exp(-(\frac{\tau^2}{4} - 1)\log k).$$

Beweis. Definiere $Z_j = (X^T\varepsilon)_j \sim \mathrm{N}(0, \sigma_j^2)$ mit

$$\sigma_j^2 = \sigma^2(X^TX)_{jj} \leq \sigma^2\lambda_{\max}(X^TX) =: \lambda_{\max}$$

für $j = 1, \ldots, n$.
Aus $\mathbb{E}[\exp(\frac{Z^2}{4})] = \sqrt{2}$ für $Z \sim \mathrm{N}(0,1)$ folgt

$$\begin{aligned} P(\mathscr{G}^c) = P\Big(\max_{j=1,\ldots,k}|Z_i| > \frac{\lambda}{8}\Big) &\leq P\Big(\exp\Big(\max_{j=1,\ldots,k}\frac{|Z_i|^2}{4\sigma_i^2}\Big) > \exp\Big(\frac{(\frac{\lambda}{8})^2}{4\sigma_{\max}^2}\Big)\Big) \\ &\overset{\text{Markov}}{\leq} \exp\Big(-\frac{\lambda^2}{4\cdot 64\sigma_{\max}^2}\Big)\,\mathbb{E}\Big[\exp\Big(\max_{j=1,\ldots,k}\frac{|Z_j|^2}{4\sigma_j^2}\Big)\Big] \\ &\leq \exp\Big(-\frac{\lambda^2}{4\cdot 64\sigma_{\max}^2}\Big)\sum_{i=1}^{k}\mathbb{E}\Big[\exp\Big(\frac{|Z_i|^2}{4\sigma_i^2}\Big)\Big] = \sqrt{2}\exp\Big(-\frac{\lambda^2}{256\sigma_{\max}^2} + \log k\Big). \end{aligned}$$

Die Behauptung folgt durch explizite Wahl von λ. □

5.21 Korollar. *Falls $\lambda_{\max}(X^TX) > 0$, gilt mit einer Wahrscheinlichkeit größer als $1-\sqrt{2}\exp((\frac{-\tau^2}{4}-1)\log k)$ für den Lassoschätzer $\hat{\beta}$ mit λ aus dem vorherigen Lemma und $\tau > 2$:*

$$\|X(\hat{\beta} - \beta)\|^2 \leq \frac{5}{3}\inf_{\beta^* \in \mathbb{R}^k}\left\{\|X(\beta^* - \beta)\|^2 + 80\tau^2|S_*|\sigma^2\log(k)\frac{\lambda_{max}(X^TX)}{\lambda_{min}(X^TX)}\right\}.$$

5.6 Kreuzvalidierung (CV)

Erste Idee:
Gegeben die Beobachtungen $(Y_1, \ldots, Y_n)$, spalte sie in eine Trainingsmenge $(Y_1, \ldots, Y_m)$, $m < n$, und eine Validierungsmenge $(Y_{m+1}, \ldots, Y_n)$ auf. Konstruiere die statistische Methode (zum Beispiel Schätzer) anhand der Trainingsmenge und bestimme ihre Güte (zum Beispiel Risiko) anhand der Vorhersage auf der Validierungsmenge. Auf diese Weise kann ein Tuningsparameter (zum Beispiel Modelldimension) ausgewählt werden, indem die Güte auf der Validierungsmenge maximiert (also der Fehler minimiert) wird.
Dieser Ansatz heißt auch „Validierungsansatz“ oder „hold out“.

Zweite Idee: (Verfeinerung, Symmetrisierung)
Setze $p := n - m$. Dann wird bei „hold out“ die Methode auf die ersten $n - p$

Daten angewendet und mit den letzten p Daten validiert. Nun kann man dies genauso mit anderen Teilmengen der Kardinalität $n-p$ beziehungsweise p machen und die Fehlerschätzung durch Mittelung der entsprechenden einzelnen Fehler versuchen zu verbessern oder zumindest symmetrischer durchzuführen.

a) „V-fold cross-validation (V-fold CV)“
Teile die Daten in V Blöcke mit jeweils circa $\frac{n}{V}$ Beobachtungen auf. Für $v = 1, \dots, V$ wähle Block v zum Validieren des Verfahrens, das mit allen Blöcken außer v trainiert wurde. Mittele die jeweiligen Fehler über $v = 1, \dots, V$.

b) „Leave-p-out cross-validation (Lpo-CV)“
Wähle Mengen $S \subseteq \{1, \dots, n\}$ der Kardinalität p und benutze S als Validierungs- und S^c als Trainingsmenge. Mittlere die Fehler über verschiedene S (müssen nicht disjunkt sein; oft wird S zufällig ausgewählt).

Wir werden uns hier auf die Leave-one-out-Kreuzvalidierung (Loo-CV) konzentrieren. Dies ist V-fold CV mit $V = n$ beziehungsweise Lpo-CV mit $p = 1$ und allen einelementigen Teilmengen $S \subseteq \{1, \dots, n\}$. Die statistische Methode beruht dann auf $(n-1)$ Beobachtungen, so dass der Fehler bei n Beobachtungen nur marginal kleiner ist. Andererseits wird die Fehlerschätzung im Allgemeinen eine größere Varianz haben als in den Fällen $V < n$ oder $p > 1$.

Im gewöhnlichen linearen Modell $Y = X^{(k)}\beta^{(k)} + \varepsilon$ mit $\varepsilon \sim \mathrm{N}(0, \sigma^2 E_n)$, $\beta^{(k)} \in \mathbb{R}^k$, $X^{(k)} \in \mathbb{R}^{n \times k}$ vom Rang k und $k = 1, \dots, K$ (wie bei AIC, BIC oben) sei $\hat{\beta}_k^{(-i)}$ der Kleinste-Quadrate-Schätzer basierend auf $\{Y_1, \dots, Y_n\} \backslash \{Y_i\}$:

$$\hat{\beta}_k^{(-i)} = \mathrm{argmin}_{\beta \in \mathbb{R}^k} \sum_{\substack{j=1 \\ j \neq i}}^{n} (Y_j - (X^{(k)}\beta)_j)^2$$

Dann gilt

$$\hat{\beta}_k^{(-i)} = (X_{-i}^T X_{-i})^{-1} X_{-i}^T Y$$

mit $X_{-i} = (E_n - E_{ii})X^{(k)}$, wobei E_{ii} die Matrix mit einer 1 an Stelle (i,i) und Nullen sonst ist. Dadurch ist X_{-i} die Matrix, die aus X entsteht, indem man die i-te Zeile von X gleich 0 setzt, insbesondere ist $X_{-i}^T Y$ unabhängig von Y_i.

Der Vorhersagefehler von Y_i ist dann $(Y_i - (X^{(k)}\hat{\beta}_k^{(-i)})_i)^2$.

Insgesamt erhalten wir als Fehlerschätzung:

$$\mathrm{CV}(k) = \sum_{i=1}^{n} (Y_i - (X^{(k)}\hat{\beta}_k^{(-i)})_i)^2.$$

Wähle nun das Modell $\hat{k}$ mit

$$\hat{k} = \mathrm{argmin}_{k=1,\dots,K} \mathrm{CV}(k).$$

5.22 Bemerkung. Würde man ganz $\|Y - X^{(k)}\hat{\beta}_k^{(-i)}\|^2$ zur Schätzung nehmen, so wäre im Allgemeinen Overfitting die Folge, da dieselben Daten für Training und Validierung benutzt werden.

5.23 Lemma. *Es gilt*

$$\mathrm{CV}(k) = \|(E_n - \tilde{\Pi}_k)^{-1}(E_n - \Pi_k)Y\|^2$$

mit $\tilde{\Pi}_k = \mathrm{diag}((\Pi_k)_{11}, \ldots, (\Pi_k)_{nn})$ *und* $\Pi_k = X^{(k)}(X^{(k)T}X^{(k)})^{-1}X^{(k)T}$.

Beweis. (Skizze).

$$\begin{aligned}
\mathrm{CV}(k) &= \|Y - \sum_{i=1}^{n} E_{ii}X^{(k)}\hat{\beta}_k^{(-i)}\|^2 \\
&= \|Y - \Pi_k Y + (\sum_{i=1}^{n} E_{ii}\Pi_k Y) - \sum_{i=1}^{n} E_{ii}X^{(k)}(X^{(k)T}(E_n - E_{ii})X^{(k)})^{-1}X^{(k)T}(E_n - E_{ii})Y\|^2 \\
&= \|Y - \Pi_k Y - \sum_{k=1}^{n} E_{ii}X^{(k)}(X^{(k)T}(E_n - E_{ii})X^{(k)})^{-1}X^{(k)T}(E_n - E_{ii})(Y - \Pi_k Y)\|^2 \\
&\overset{X^{(k)T}(Y-\Pi_k Y)=0}{=} \|\{E_n + \underbrace{\sum_{i=1}^{n} E_{ii}X^{(k)}(X^{(k)T}(E_n - E_{ii})X^{(k)})^{-1}X^{(k)T}E_{ii})}_{=\mathrm{diag}((X^{(k)}(X^{(k)T}(E_n - E_{ii})X^{(k)})^{-1}X^{(k)T})_{ii}}\}(Y - \Pi_k Y)\|^2
\end{aligned}$$

□

Beachte nun: Π_k ist Orthogonalprojektion vom Rang k, das heißt Π_k ist symmetrisch mit Eigenwerten 1 (k-fach) und 0 ($(n-k)$-fach). Folglich gilt $\mathrm{Spur}(\Pi_k) = \mathrm{tr}(\Pi_k) = k$. Außerdem:

$$(\Pi_k)_{ii} = \langle \Pi_k e_i,\, e_i\rangle = \langle \Pi_k^2 e_i,\, e_i\rangle = \|\Pi_k e_i\|^2 \in [0,\, 1].$$

Der Mittelwert der inversen Gewichte $(1 - (\Pi_k)_{ii})$ ist gerade

$$\frac{1}{n}\sum_{i=1}^{n}(1 - (\Pi_k)_{ii}) = 1 - \frac{1}{n}\,\mathrm{tr}(\Pi_k) = 1 - \frac{k}{n}.$$

Je größer k ist, desto größer sind „im Schnitt“ auch die Gewichte $(1-(\Pi_k)_{ii})^{-1}$. Diese Überlegungen sowie Rechenzeitaspekte führen im linearen Modell auf folgende Variante der Loo-CV:

5.24 Definition. Im linearen Modell (wie oben) ist das verallgemeinerte CV-Kriterium (GCV: generalised CV) gegeben durch

$$\hat{k}^{\mathrm{GCV}} = \mathrm{argmin}_k \sum_{i=1}^{n}\Big(\frac{Y_i - (\Pi_k Y)_i}{1 - \frac{k}{n}}\Big)^2 = \mathrm{argmin}_k \Big(1 - \frac{k}{n}\Big)^{-2} \underbrace{\|Y - (\Pi_k Y)\|^2}_{\mathrm{RSS}_k}.$$

Was ist der Zusammenhang mit bekannten Modellwahlkriterien im Fall des linearen Modells?
Beachte, dass für $\frac{k}{n}$ klein $\frac{1}{(1-\frac{k}{n})^2} \approx 1 + 2\frac{k}{n}$ gilt. Es gilt dann also

$$\left(1-\frac{k}{n}\right)^{-2}\|Y-\Pi_k Y\|^2 \approx \|Y-\Pi_k Y\|^2+2k\frac{\|Y-\Pi_k Y\|^2}{n} \overset{n \text{ groß}}{\approx} \|Y-\Pi_k Y\|^2+2k\sigma^2.$$

Asymptotisch gilt, wie man zeigen kann, dass $\hat{k}^{\mathrm{GCV}}$ und $\hat{k}^{\mathrm{AIC}}$ die gleichen Eigenschaften besitzen (sogar $\mathbb{P}(\hat{k}^{\mathrm{GCV}} = \hat{k}^{\mathrm{AIC}}) \to 1$ für $n \to \infty$ im Allgemeinen).
Falls mit $X^{(k)} = \begin{pmatrix} x_1^{(k)} \\ \vdots \\ x_n^{(k)} \end{pmatrix}$, $(x_i^{(k)})^T \in \mathbb{R}^k$, gilt

$$\max_{1\le i\le n} (x_i^{(k)})^T((X^{(k)})^T X^{(k)})^{-1} x_i^{(k)} \overset{n\to\infty}{\to} 0$$

gleichmäßig in k, so sind auch $\hat{k}^{\mathrm{CV}}$, $\hat{k}^{\mathrm{GCV}}$, $\hat{k}^{\mathrm{AIC}}$ asymptotisch gleich.
Man kann auch zeigen, dass Leave-p-out-Kreuzvalidierung asymptotisch äquivalent ist (im linearen Modell) zum penalisierten KQ-Kriterium

$$\hat{k} = \mathrm{argmin}_k(\|Y - \underbrace{X^{(k)}\hat{\beta}_k}_{\Pi_k Y}\|^2(1+\lambda_n k))$$

mit $\lambda_n = \frac{n}{n-p} + 1$. Gilt also $\frac{p}{n} \to 0$, so ergibt sich AIC, für $\frac{p}{n} \to 1$ (Trainingsmenge klein, Validierungsmenge sehr groß), folgt $\lambda_n \to \infty$. Im Fall $\lambda_n = \log n$ ergibt sich z.B. BIC. (Quelle: Jun Shao: An asymptotic theory for linear model selection. Statistica Sinica, 1997.)

6 Dimensionsreduktion

6.1 Hauptkomponentenanalyse

(Principal components analysis: PCA)
Motivation:

- Stilanalyse in der Literatur,
- Schrifterkennung.

Idee: Gegeben sind p-dimensionale Daten $X_1, \dots, X_n \in \mathbb{R}^p$. Wir wollen diese Datenpunkte durch einen affinen Unterraum von viel kleinerer Dimension q ($q \ll p$, $q \ll n$) beschreiben, das heißt durch Orthogonalprojektion auf diesen Unterraum.
Sei dazu $f : \mathbb{R}^q \to \mathbb{R}^p$ eine (linear) affine Funktion, das heißt, $f(v) = \mu + Av$, $\mu \in \mathbb{R}^p$, $A \in \mathbb{R}^{p\times q}$, wobei A eine orthogonale Matrix ist ($A^TA = E_q$ oder $A = (a_1, \dots, a_q)$ mit $a_1, \dots, a_q$ Orthonormalsystem (ONS), die einen Unterraum aufspannen). Beste Annäherung an die Daten wird über das Kleinste-Quadrate-Kriterium definiert:

$$\sum_{i=1}^n \|X_i - \mu - Av_i\|^2 \to \min_{\mu,\, v,\, A}!$$

Die Kleinste-Quadrate-Methode liefert für festes A die optimalen Werte

$$\hat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \hat{v}_i = A^T(X_i - \overline{X}).$$

Wir müssen also noch lösen

$$\sum_{i=1}^{n} \|(X_i - \overline{X}) - AA^T(X_i - \overline{X})\|^2 \to \min_A!$$

(für festes q).
Im Folgenden sei o.B.d.A. $\overline{X} = 0$ (sonst betrachte $\tilde{X}_i = X_i - \overline{X}$). Wir erhalten also

$$\sum_{i=1}^{n} \|X_i - AA^T X_i\|^2 \to \min_A!$$

AA^T ist eine Orthogonalprojektion vom Rang q. Schreibe nun

$$X = \begin{pmatrix} X_1^T \\ \vdots \\ X_n^T \end{pmatrix} \in \mathbb{R}^{n\times p}.$$

Dann gilt: X^TX ist symmetrisch und positiv semidefinit; $X^TX = WD^2W^T$ mit $W \in \mathbb{R}^{p\times p}$ orthogonal, $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$ mit $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$. Wir lösen

$$\sum_{i,j}((X^T)_{i,j} - (AA^TX)_{i,j})^2 \to \min_A!$$

Das heißt, wir minimieren

$$\begin{aligned}
&\operatorname{tr}((E_p - AA^T)X^T((E_p - AA^T)X^T)^T) \\
&\quad = \operatorname{tr}((E_p - AA^T)X^TX(E_p - AA^T)) = \operatorname{tr}(X^TX(E_p - AA^T)) \\
&\quad = \operatorname{tr}(WD^2W^T(E_p - AA^T)) = \operatorname{tr}(D^2 \underbrace{(E_p - (W^TA)(W^TA)^T)}_{\text{Orth.-Proj., Rang } p-q}) \to \min_A!
\end{aligned}$$

Für Π Orthogonalprojektion vom Rang $p - q$ gilt:

$$\min_{\Pi} \operatorname{tr}(D^2\Pi) = \min_{\Pi} \operatorname{tr} \begin{pmatrix} \lambda_1^2\Pi_{11} & & * \\ & \ddots & \\ * & & \lambda_p^2\Pi_{pp} \end{pmatrix}$$

Wir wissen: $\sum_{i=1}^{p} \Pi_{ii} = p - q$, $\Pi_{ii} \geq 0$, $\Pi_{ii} \leq 1$, so dass

$$\min_{\Pi} \operatorname{tr}(D^2\Pi) = \min_{\Pi} \sum_{i=1}^{p} \lambda_i^2\Pi_{ii} \geq \sum_{i=q+1}^{p} \lambda_i^2.$$

Das Minimum wird angenommen für

$$(W^T\hat{A})(W^T\hat{A})^T = \operatorname{diag}(\underbrace{1, \ldots, 1}_{q} \underbrace{0, \ldots, 0}_{p-q}).$$

Wähle dazu $\hat{A} = (w_1 w_2 \dots w_q)$ (mit $W = (w_1 \dots w_q w_{q+1} \dots w_p)$).
Alles zusammen ergibt dann folgenden Satz:

6.1 Satz. *Es gilt mit* $\hat{\mu} = \overline{X}$, $\hat{v}_i = (w_1 \dots w_q)^T (X_i - \overline{X})$, $\hat{A} = (w_1 \dots w_q)$*:*

$$(\hat{\mu}, (\hat{v}_i), \hat{A}) = \operatorname{argmin}_{\mu, (v_i), A} \sum_{i=1}^{n} \|X_i - \mu - A v_i\|^2$$

sowie

$$\min_{\mu, (v_i), A} \sum_{i=1}^{n} \|X_i - \mu - A v_i\|^2 = \sum_{i=q+1}^{p} \lambda_i^2.$$

6.2 Definition. Sind $w_1, \dots, w_p \in \mathbb{R}^p$ die Eigenvektoren von $X^T X$ zu $\lambda_1 \geq \dots \geq \lambda_p \geq 0$, so heißt w_i i-te Hauptkomponente.

Beachte: Hierbei sind $w_1, \dots, w_p$ Eigenvektoren von $\tilde{X}^T \tilde{X}$ mit $\tilde{X} = \begin{pmatrix} (X_1 - \overline{X})^T \\ \vdots \\ (X_n - \overline{X})^T \end{pmatrix}$ zu Eigenwerten $\lambda_1^2 \geq \lambda_2^2 \geq \dots \geq \lambda_p^2 \geq 0$.
Die Hauptkomponentenanalyse erlaubt daher eine für viele Anwendungen unabdingbare Dimensionsreduktion, das heißt Datenkompression, und findet gleichzeitig die signifikanten Charakteristiken („feature extraction“) und Strukturen in den Daten. Erweiterungen sind meist nichtlinear, wie zum Beispiel „manifold learning“, „kernel PCA“.
Wie wir jetzt sehen werden, kann PCA äquivalent auch dadurch beschrieben werden, dass die ersten q Hauptkomponenten die empirische Varianz von $X_1, \dots, X_n$ am besten (unter allen q-dimensionalen Unterräumen) erklären.
Für eine statistische Analyse nehmen wir nun an, dass $X_1, \dots, X_n$ i.i.d. gemäß einer p-dimensionalen Verteilung gezogen werden mit $X_i \in L^2$, $\mu = \mathbb{E}[X_i] \in \mathbb{R}^p$, Kovarianzmatrix $\Sigma = \mathbb{E}[X_i X_i^T]$.
Dann ist der empirische Erwartungswert $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i = \overline{X}$, die empirische Kovarianzmatrix $\hat{\Sigma}_n = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})(X_i - \overline{X})^T \in \mathbb{R}^{p \times p}$.
Mit $\tilde{X}_i = X_i - \overline{X}$ sowie $\tilde{X} = \begin{pmatrix} \tilde{X}_1^T \\ \vdots \\ \tilde{X}_n^T \end{pmatrix}$ gilt dann:

$$\hat{\Sigma}_n = \frac{1}{n-1} \sum_{i=1}^{n} \tilde{X}_i \tilde{X}_i^T = \frac{1}{n-1} (\tilde{X}_1 \dots \tilde{X}_n) \begin{pmatrix} \tilde{X}_1^T \\ \vdots \\ \tilde{X}_n^T \end{pmatrix} = \frac{1}{n-1} \tilde{X}^T \tilde{X}.$$

Damit bilden $w_1, \dots, w_p \in \mathbb{R}^p$ eine Orthogonalbasis von Eigenvektoren von $\hat{\Sigma}_n$ zu Eigenwerten $\frac{1}{n-1}\lambda_1^2 \geq \dots \geq \frac{1}{n-1}\lambda_p^2 \geq 0$. Man sagt auch, dass $W = (w_1 \dots w_p)$ die Daten dekorreliert („whitening“).

6.3 Satz. *In dieser Notation gilt mit* $S^{p-1} = \{v \in \mathbb{R}^p \,|\, \|v\| = 1\}$*:*

(a) $w_1 = \operatorname{argmax}_{v\in S^{p-1}} \langle \hat{\Sigma}_n v,\, v\rangle = \operatorname{argmax}_{v\in S^{p-1}} \sum_{i=1}^{n} \langle \tilde{X}_i,\, v\rangle^2,$

(b) $w_2 = \operatorname{argmax}_{\substack{v\in S^{p-1}\\ v\perp w_1}} \langle \hat{\Sigma}_n v,\, v\rangle = \operatorname{argmax}_{\substack{v\in S^{p-1}\\ v\perp w_1}} \sum_{i=1}^{n} \langle \tilde{X}_i,\, v\rangle^2$ *und allgemein*

(c) $w_{k+1} = \operatorname{argmax}_{\substack{v\in S^{p-1}\\ v\perp w_j,\, j\le k}} \langle \hat{\Sigma}_n v,\, v\rangle = \operatorname{argmax}_{\substack{v\in S^{p-1}\\ v\perp w_j,\, j\le k}} \sum_{i=1}^{n} \langle \tilde{X}_i,\, v\rangle^2.$

Beweis. Dies ist die Variationscharakterisierung der Eigenvektoren symmetrischer positiv semidefiniter Matrizen.

(a) Wir verwenden $\tilde{X}^T\tilde{X} = WD^2W^T$:

$$\begin{aligned}(n-1)\langle \hat{\Sigma}_n v,\, v\rangle =& \langle WD^2W^Tv,\, v\rangle = \langle D^2W^Tv,\, W^Tv\rangle \\ &\le \lambda_1^2\|W^Tv\|^2 = \lambda_1^2\|v\|^2 = \lambda_1^2 \text{ für } v\in S^{p-1}.\end{aligned}$$

Nun gilt aber

$$(n-1)\langle \hat{\Sigma}_n w_1,\, w_1\rangle = \langle D^2 \underbrace{W^Tw_1}_{=e_1},\, W^Tw_1\rangle = \lambda_1^2.$$

Beachte noch:

$$(n-1)\langle \hat{\Sigma}_n v,\, v\rangle = v^T(\tilde{X}^T\tilde{X})v = \sum_{i=1}^{n}\langle \tilde{X}_i,\, v\rangle^2.$$

(b) Für $v\in S^{p-1}$, $v\perp w_1$ gilt $v = (E_p - w_1w_1^T)v$ wegen

$$w_1w_1^Tv = w_1\langle w_1,\, v\rangle = 0.$$

Wir rechnen:

$$\begin{aligned}&(E_p - w_1w_1^T)W = W - w_1w_1^T(w_1\dots w_p)\\ &= W - w_1(\underbrace{w_1^Tw_1}_{=1},\, \underbrace{w_1^Tw_2}_{=0},\dots,\underbrace{w_1^Tw_p}_{=0}) = W\operatorname{diag}(0,\,1,\dots,1).\end{aligned}$$

Folglich gilt

$$\begin{aligned}(n-1)\langle v,\, \hat{\Sigma}_n v\rangle &= \langle (E_p - w_1w_1^T)v,\ \underbrace{\tilde{X}^T\tilde{X}}_{=WD^2W^T}\ (E_p - w_1w_1^T)v\rangle\\ &= v^TW\underbrace{\operatorname{diag}(0,\,1,\dots,1)D^2\operatorname{diag}(0,\,1,\dots,1)}_{\operatorname{diag}(0,\,\lambda_2^2,\dots,\lambda_p^2)}W^Tv.\end{aligned}$$

Es folgt wie oben: $(n-1)\langle v,\, \hat{\Sigma}_n v\rangle \le \lambda_2^2$ und $(n-1)\langle w_2,\, \hat{\Sigma}_n w_2\rangle = \lambda_2^2$.

(c) analog.

□

Die Hauptkomponenten sind also die Richtungen mit maximaler empirischer Varianz. Die Dimension q wird häufig so gewählt, dass ein gewisses Perzentil (z. B. 50%, 75%, 90%) der erklärten empirischen Varianz $\sum_{i=1}^{q} \lambda_i^2$ bezüglich der Gesamtvarianz $\sum_{i=1}^{p} \lambda_i^2$ erreicht wird.
In manchen Anwendungen ist p viel größer als n, z. B. in der Gesichts-/ Bilderkennung mit $p = 1024^2$ Graustufenpixeln. Dann ist die Diagonalisierung von $X^T X \in \mathbb{R}^{p \times p}$ viel zu aufwändig (selbst das Auffinden der q größten Eigenwerte). Beachte nun: Für $XX^T \in \mathbb{R}^{n \times n}$ mit Eigenwerten μ_i^2, $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n \geq 0$, und Eigenvektoren $v_1, \ldots, v_n$ gilt

$$X^T X X^T v_i = \mu_i^2 X^T v_i,$$

und folglich ist $w_i = X^T v_i \in \mathbb{R}^p$ Eigenvektor von $X^T X$ zum Eigenwert $\lambda_i^2 = \mu_i^2$ ist.
Finde nun Eigenwerte $\lambda_1^2, \cdots, \lambda_q^2$ mit Eigenvektoren $v_1, \ldots, v_q$ von XX^T und berechne $w_i = X^T v_i$.
Es gilt allgemein ($Xw_i = \lambda_i^2 v_i$, w_i normalisieren):

$$Xw = \sum_{i=1}^{p} \lambda_i \langle w, w_i \rangle v_i$$

für $w \in \mathbb{R}^p, w_i = \frac{X^T v_i}{\|X^T v_i\|}$. Dies ist die sogenannte Singulärwertzerlegung (SVD) von $X \in \mathbb{R}^{n \times p}$.

6.2 Stabilistische Analyse von PCA

Ein wesentliches Ziel von PCA ist es, die Daten gut zu approximieren durch niedrigdimensionale Unterräume. Wir betrachten dazu den Rekonstruktionsfehler (Approximationsfehler) bei zukünftigen Beobachtungen $X_{n+1}, \ldots$ unter der Annahme, dass $(X_i)_{i \geq 1}$ i.i.d. im $\mathbb{R}^d$ verteilt sind: Wie groß ist dann

$$\mathbb{E}[\|X_{n+1} - \overline{X}_n - \hat{A}_n \hat{A}_n^T (X_{n+1} - \overline{X}_n)\|^2],$$

wobei $\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ und $\hat{A}_n$ wie oben aus den Daten $X_1, \ldots, X_n$ ermittelt wird?
Dazu betrachten wir zunächst den Orakelfehler:

$$\min_{\mu, A} \mathbb{E}[\|X_{n+1} - \mu - AA^T (X_{n+1} - \mu)\|^2] = \min_{\substack{\mu \in \mathbb{R}^p \\ \Pi}} \mathbb{E}[\|X_{n+1} - \mu - \Pi(X_{n+1} - \mu)\|^2],$$

mit dem zweiten Minimum über alle Orthogonalprojektionen Π vom Rang q.

Ist $X_{n+1} \in L^2$ mit $\mathrm{Cov}(X_{n+1}) = \mathbb{E}[(X_{n+1} - \mu)(X_{n+1} - \mu)^T] = \Sigma \in \mathbb{R}^{p \times p}$ und $\mu = \mathbb{E}[X_{n+1}]$, so gilt mit dem gleichen Beweis wie oben, dass der Orakelfehler gleich

$$\mathbb{E}[\|X_{n+1} - \mu - \Pi_q (X_{n+1} - \mu)\|^2]$$

ist mit Π_q Orthogonalprojektion auf $\text{span}(w_1, \ldots, w_q)$ mit Eigenvektoren $w_1, \ldots, w_q$ von Σ zu den q größten Eigenwerten $\lambda_1^2, \ldots, \lambda_q^2$.
Wie oben, beweist man auch für Σ statt $\hat{\Sigma}_n$:

6.4 Satz. *Es gilt:*

$$\text{Var}(\langle X_{n+1}, w_1 \rangle) = \max_{v \in S^{p-1}} \text{Var}(\langle X_{n+1}, v \rangle),$$

die Varianz ist also in Richtung w_1 maximal, und für $2 \leq k \leq p$ gilt

$$\text{Var}(\langle X_{n+1}, w_k \rangle) = \max_{\substack{v \in S^{p-1} \\ v \perp w_j \forall j \leq k-1}} \text{Var}(\langle X_{n+1}, v \rangle).$$

Also spannen $w_1, \ldots, w_q$ den Unterraum der Dimension q auf, in dem die Varianz des projizierten Zufallsvektor am größten ist. Wir fragen uns nun, wie nahe der Rekonstruktionsfehler von PCA am Orakelfehler ist.
Der Einfachheit halber betrachten wir den Fall ohne Zentrierung:

$$\Sigma = \mathbb{E}[X_{n+1} X_{n+1}^T],$$

$$X = \begin{pmatrix} X_1^T \\ \vdots \\ X_n^T \end{pmatrix}.$$

Führe folgende Notation ein:

$$U_q = \{U \subseteq \mathbb{R}^p \,|U \text{ Unterraum}, \dim(U) = q\}.$$

Für $V \in U_q$ definiere den empirischen Rekonstruktionsfehler

$$\text{R}_n(V) = \frac{1}{n} \sum_{i=1}^{n} \|X_i - \Pi_V X_i\|^2$$

mit Π_V der Orthogonalprojektion auf V.
Mit $\mathbb{P}_n$ sei das empirische Maß von $X_1, \ldots, X_n$ bezeichnet:

$$\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i},$$

wir schreiben

$$\mathbb{P}_n[f(X)] = \frac{1}{n} \sum_{i=1}^{n} f(X_i),$$

wobei X einen Zufallsvektor im $\mathbb{R}^p$ unter der Verteilung $\mathbb{P}_n$ bezeichnet. Damit gilt dann

$$\text{R}_n(V) = \mathbb{P}_n[\|X - \Pi_V X\|^2] = \mathbb{P}_n[\|\Pi_{V^\perp} X\|^2] = \mathbb{P}_n[X^T \Pi_{V^\perp} X] = \mathbb{P}_n[\text{tr}(\Pi_{V^\perp} \overline{\text{X}})]$$

mit $\overline{\mathrm{X}} = XX^T \in \mathbb{R}^{p\times p}$.
Dies entspricht (im zentrierten Fall)

$$\mathrm{R}_n(V) = \mathrm{tr}(\Pi_{V^\perp}\hat{\Sigma}_n),$$

beachte dabei $\mathbb{P}_n(\overline{\mathrm{X}}) = \frac{1}{n}\sum_{i=1}^n (X_{i,k}X_{i,l})_{k,l=1,\dots,p}$.

Im Modell erhalten wir entsprechend $\mathrm{R}(V) := \mathbb{E}[\|X - \Pi_V X\|^2]$ (wie vorher $\mathbb{E}[\|X_{n+1} - \Pi_V X_{n+1}\|^2]$) und folglich

$$\mathrm{R}(V) = \mathbb{P}[\|X - \Pi_V X\|^2] = \mathbb{P}[\mathrm{tr}(\Pi_{V^\perp}\overline{\mathrm{X}})] = \mathrm{tr}(\Pi_{V^\perp}\mathbb{P}[\overline{\mathrm{X}}]).$$

Beachte, dass $\mathbb{P}[\overline{\mathrm{X}}] = \Sigma$ gilt im Fall $\mu = 0$.
Wir erhalten also, dass PCA $\hat{V}_q = \mathrm{argmin}_{V\in U_q}\mathrm{R}_n(V)$ und das Orakel $V_q = \mathrm{argmin}_{V\in U_q}\mathrm{R}(V)$ auswählt.

Ziel ist nun die Abschätzung des „excess risk“ $\mathrm{R}(\hat{V}_q) - \mathrm{R}(V_q)$:

6.5 Lemma. *Es gilt*

$$\mathrm{R}(\hat{V}_q) \le \mathrm{R}(V_q) + 2\|\,\mathrm{R}_n - \mathrm{R}\,\|_\infty$$
$$\mathrm{R}(\hat{V}_q) \le \mathrm{R}(V_q) + \sup_{V\in U_q}(\mathrm{R}_n - \mathrm{R}) + \sup_{V\in U_q}(\mathrm{R} - \mathrm{R}_n)$$

mit $\|\,\mathrm{R}_n - \mathrm{R}\,\|_\infty = \sup_{V\in U_q}(|\,\mathrm{R}_n(V) - R(V)|)$.

Beweis.

$$\begin{aligned}\mathrm{R}(\hat{V}_q) &\le \mathrm{R}_n(\hat{V}_q) + \|\,\mathrm{R} - \mathrm{R}_n\,\|_\infty \le \mathrm{R}_n(V_q) + \|\,\mathrm{R} - \mathrm{R}_n\,\|_\infty\\ &\le \mathrm{R}(V_q) + 2\|\,\mathrm{R}_n - \mathrm{R}\,\|_\infty,\end{aligned}$$

die zweite Ungleichung analog.

□

Wir schätzen wie folgt weiter ab:

$$\begin{aligned}\sup_{V\in U_q}(\mathrm{R}_n - \mathrm{R}) &= \sup_{V\in U_q}(\mathbb{P}_n - \mathbb{P})[\mathrm{tr}(\Pi_{V^\perp}\overline{\mathrm{X}})] = \sup_{V\in U_q}\mathrm{tr}((E_p - \Pi_V)(\mathbb{P}_n - \mathbb{P})[\overline{\mathrm{X}}])\\ &\le \mathrm{tr}((\mathbb{P}_n - \mathbb{P})[\overline{\mathrm{X}}]) + \sup_{V\in U_q}|\,\mathrm{tr}(\Pi_V(\mathbb{P}_n - \mathbb{P})[\overline{\mathrm{X}}])|\\ &= \mathrm{tr}((\mathbb{P}_n - \mathbb{P})[\overline{\mathrm{X}}]) + \sup_{V\in U_q}|\langle\Pi_V, (\mathbb{P}_n - \mathbb{P})[\overline{\mathrm{X}}]\rangle_{HS}|\end{aligned}$$

mit dem Skalarprodukt $\langle A, B\rangle_{HS} = \mathrm{tr}(AB^T)$. Es gilt $\langle A, B\rangle_{HS} = \sum_{i,j=1}^p A_{i,j}B_{i,j}$ und $\|A\|_{HS} = \langle A, A\rangle_{HS}$ ist die Hilbert-Schmidt- oder Frobeniusnorm von A. Diese ist invariant bezüglich orthogonaler Transformationen (Wahl einer Orthonormalbasis).
Somit erhalten wir mit der Cauchy-Schwarz-Ungleichung

$$\sup_{V\in U_q}(\mathrm{R}_n - \mathrm{R}) \le \mathrm{tr}((\mathbb{P}_n - \mathbb{P})[\overline{\mathrm{X}}]) + \underbrace{\sup_{V\in U_q}\|\Pi_V\|_{HS}}_{=\sqrt{q}\text{ für }\mathrm{diag}(1,\dots,1,0,\dots,0)}\|(\mathbb{P}_n - \mathbb{P})[\overline{\mathrm{X}}]\|_{HS}$$
$$= (\mathbb{P}_n - \mathbb{P})[\|X\|^2_{\mathbb{R}^p}] + \sqrt{q}\|(\mathbb{P}_n - \mathbb{P})[\overline{\mathrm{X}}]\|_{HS}$$

und folglich

$$\mathbb{E}[\sup_{V\in U_q}(\mathrm{R}(V)-\mathrm{R}_n(V))] \le \sqrt{q}\,\mathbb{E}[\|(\mathbb{P}_n-\mathbb{P})[\overline{\mathrm{X}}]\|_{HS}],$$
$$\mathbb{E}[\sup_{V\in U_q}(\mathrm{R}_n(V)-\mathrm{R}(V))] \le \sqrt{q}\,\mathbb{E}[\|(\mathbb{P}_n-\mathbb{P})[\overline{\mathrm{X}}]\|_{HS}].$$

Aus dem Lemma folgt dann mit der Cauchy-Schwarz-Ungleichung

$$\mathbb{E}[\mathrm{R}(\hat{V}_q)] \le \mathrm{R}(V_q) + 2\sqrt{q}\,\mathbb{E}[\underbrace{\|(\mathbb{P}_n-\mathbb{P})[\overline{\mathrm{X}}]\|_{HS}^2}_{\sum_{k,\,l=1}^p((\mathbb{P}_n-\mathbb{P})[X_kX_l])^2}]^{1/2}$$

mit X_i der i-ten Zeile von $\overline{\mathrm{X}}$.
Nun ist

$$\mathbb{E}[(\mathbb{P}_n-\mathbb{P})[X_kX_l]] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^n(X_k^{(i)}X_l^{(i)}-\mathbb{E}[X_kX_l])] = 0,$$

$$\mathrm{Var}((\mathbb{P}_n-\mathbb{P})[X_kX_l]) = \frac{1}{n}\mathrm{Var}(X_kX_l) = \frac{1}{n}(\mathbb{E}[X_k^2X_l^2]-E[X_kX_l]^2),$$

wir erhalten:

$$\mathbb{E}[\|(\mathbb{P}_n-\mathbb{P})[\overline{\mathrm{X}}]\|_{HS}^2] = \frac{1}{n}\sum_{k,\,l=1}^p(\mathbb{E}[X_k^2X_l^2]-\mathbb{E}[X_kX_l]^2) = \frac{1}{n}(E[\underbrace{\|X\|^4}_{(\sum X_k^2)^2}]-\|\,\mathbb{E}[\underbrace{XX^T}_{(X_kX_l)_{k,\,l}}]\|_{HS}^2).$$

Wir haben somit bewiesen:

6.6 Satz. *Es gilt*

$$\mathbb{E}[\mathrm{R}(\hat{V}_q)-\mathrm{R}(V_q)] \le \frac{2\sqrt{q}}{\sqrt{n}}(\mathbb{E}[\|X\|^4]-\|\,\mathbb{E}[XX^T]\|_{HS}^2)^{1/2}.$$

6.7 Bemerkung. Die linke Seite beschreibt das excess risk von $\hat{V}_q$ (gewonnen durch $\mathbb{P}(A)$) bezüglich dem Orakel V_q. Dieses konvergiert mit Rate $\frac{1}{\sqrt{n}}$ gegen Null für Stichprobenumfang $n\to\infty$. Die Schranke ist aber sogar nicht-asymptotisch und wächst mit $\sqrt{q}$ in der Dimension des zu wählenden Unterraums. Die Dimension p erscheint nur implizit über die Momente von X.

Im Fall $X\sim\mathrm{N}(0,\Sigma)$, $\Sigma\in\mathbb{R}^{p\times p}$ gilt:

$$\begin{aligned}\mathbb{E}[\|X\|^4]-\|\,\mathbb{E}[XX^T]\|_{HS}^2 &= \sum_{k,\,l=1}^p(\mathbb{E}[X_k^2]\,\mathbb{E}[X_l]^2+\mathbb{E}[X_kX_l]^2) = \sum_{k,\,l=1}^p(\Sigma_{kk}\Sigma_{ll}+\Sigma_{kl}^2)\\ &= (\mathrm{tr}(\Sigma))^2+\|\Sigma\|_{HS}^2 = \|(\lambda_i(\Sigma))_i\|_{\ell_1}^2+\|(\lambda_i(\Sigma))_i\|_{\ell_2}^2\end{aligned}$$

Es gibt allerdings bei Eigenwerten das Phänomen der Superkonvergenz, das heißt (unter Zusatzannahmen) eine Konvergenzrate von n^{-1} statt $n^{-1/2}$.

Man kann zeigen: (Blanchard, Bousquet, Zwald 2007)
Falls die Matrix $\Sigma := \mathbb{E}[XX^T] \in \mathbb{R}^{p\times p}$ Eigenwerte

$$\lambda_1^2 \geq \lambda_2^2 \geq \cdots \geq \lambda_q^2 > \lambda_{q+1}^2 \geq \ldots \lambda_p^2 \geq 0$$

besitzt („spectral gap“ bei λ_q, λ_{q+1}), so gilt:

$$\mathbb{E}\left[\sup_{V\in U_q} \frac{\langle \Pi_V - \Pi_{V_q}, (\mathbb{P}-\mathbb{P}_n)[\overline{\mathrm{X}}]\rangle_{HS}^2}{\mathrm{R}(V)-\mathrm{R}(V_q)}\right] \leq c\frac{1}{n(\lambda_q^2-\lambda_{q+1}^2)}\operatorname{Var}\left(\sum_{k,l=1}^{p} X_k X_l\right).$$

Damit folgt:

$$\begin{aligned}\mathbb{E}[\mathrm{R}(\hat{V}_q)-\mathrm{R}(V)] &\leq \mathbb{E}\left[\left(\sup_{V\ inU_q} \frac{\langle \Pi_V - \Pi_{V_q}, (\mathbb{P}-\mathbb{P}_n)[\overline{\mathrm{X}}]\rangle_{HS}}{\sqrt{\mathrm{R}(V)-\mathrm{R}(V_q)}}\right)\sqrt{\mathrm{R}(\hat{V}_q)-\mathrm{R}(V)}\right] \\ &\overset{\text{CSU}}{\leq} \sqrt{M}\,\mathbb{E}[\mathrm{R}(\hat{V}_q)-\mathrm{R}(V)]^{1/2}\end{aligned}$$

und somit

$$\mathbb{E}[\mathrm{R}(\hat{V}_q)-\mathrm{R}(V)] \leq M.$$

6.3 Independent Component Analysis (ICA)

PCA approximiert $X_1, \ldots, X_n$ durch einen q-dimensionalen Unterraum. Im Orakelfall (Verteilungen X bekannt) und $\mathbb{E}[X] = 0$ berechnet man $X = A^{(q)}S^{(q)} \in \mathbb{R}^p$ mit deterministischer orthogonaler Matrix $A^{(q)} \in R^{p\times q}$ und $\mathbb{E}[S^{(q)}] = 0$, $\mathbb{E}[S^{(q)}S^{(q)T}] = \mathbb{E}[XX^T] = \Sigma$.
O.B.d.A. gelte $\Sigma = E_p$ (multipliziere sonst X mit $\Sigma^{-1/2}$, das sogenannte „prewhitening“ von X; bei Daten multipliziere mit $\hat{\Sigma}_n^{-1/2}$).
PCA sucht als Hauptkomponenten Richtungen mit größter Varianz. $A^{(q)}$ ist nicht eindeutig, $\tilde{A}^{(q)} = A^{(q)}U$ mit U orthogonal führt auf $\tilde{S}^{(q)} = U^T S^{(q)}$ mit derselben Kovarianzmatrix.
Idee von ICA:
Versuche Zufallsvariablen $S_1, \ldots, S_q \in \mathbb{R}$ zu finden, die unabhängig sind (nicht nur unkorreliert) im Fall $\Sigma = E_p$, $S^{(q)} = (S_1, \ldots, S_q)^T$. Falls $S^{(q)}$ nicht normalverteilt ist, wird dadurch $A^{(q)}$ eindeutig bestimmt.
Ziel:
Finde $A^{(q)}$, so dass $S^{(q)} = A^{(q)T}X$ unabhängige Koordinaten besitzt. Minimiere dazu den Kullback-Leibler-Abstand zwischen einem Zufallsvektor $Y \in \mathbb{R}^q$ und dem Produktmaß seiner Koordinaten $Y_1, \ldots, Y_p$: $\mathrm{KL}(\mathbb{P}_n^Y \mid \bigotimes_{k=1}^q \mathbb{P}^{Y_k})$.
Für A orthogonal ergibt sich:

$$\mathrm{KL}(\mathbb{P}_n^{A^TX} \mid \bigotimes_{k=1}^{q} \mathbb{P}^{(A^TX)_k}) = \underbrace{\sum_{k=1}^{p} \int_{\mathbb{R}} -\mathbb{P}_{(A^TX)_k}(y) \log \mathbb{P}_{(A^TX)_k}(y) dy}_{\to \min!} + \underbrace{\int_{\mathbb{R}^p} \mathbb{P}_X(y) \log \mathbb{P}_X(y) dy}_{\text{unabhängig von } A}.$$

Es ist bekannt, dass die Entropie $\int_{\mathbb{R}} -\mathbb{P}_{(A^TX)_k}(y) \log \mathbb{P}_{(A^TX)_k}(y) dy$ maximal ist bei Varianz 1 für den Fall der Normalverteilung. Es werden also gerade Richtungen gesucht, die nicht gaußsch sind (Paradigmen, Gaußsche Verteilung hat keine Struktur).