Formelsammlung zur Vorlesung Schätzen und Testen II

5 Bootstrap

Einstichproben-Problem:

$$X = (X_1, \ldots, X_n) \to T(X)$$

mit $X_i \overset{i.i.d.}{\sim} F$, F unbekannt

Beobachtete Daten:

$$x = (x_1, x_2, \ldots, x_n) \to T(x)$$

Bootstrap-Stichprobe:

$$x^* = (x_1^*, x_2^*, \ldots, x_n^*) \to T(x^*)$$

Empirische Verteilungsfunktion:

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq x)$$

Bootstrap-Algorithmus zur Schätzung des Standardfehlers

1. Erzeuge B Bootstrap-Stichproben $x^{*1}, \ldots, x^{*B}$.
2. Berechne $\hat{\theta}^*(b)$, $b = 1, \ldots, B$.
3. Schätze den Standardfehler $\mathrm{se}_F(\hat{\theta}) = \sqrt{\mathrm{Var}_F(\hat{\theta})}$ durch

$$\widehat{\mathrm{se}}_B = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} \left[\hat{\theta}^*(b) - \hat{\theta}^*(\cdot) \right]^2 \right\}^{\frac{1}{2}} \quad \text{mit} \quad \hat{\theta}^*(\cdot) = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^*(b).$$

Standardfehler für die Schätzung des Korrelationskoeffizienten θ

(i) Vergleich mit der Formel für die bivariate Normalverteilung:

$$\widehat{\mathrm{se}}_{N_2(\mu,\Sigma)}(\hat{\theta}) = \frac{1 - \hat{\theta}^2}{\sqrt{n-3}}$$

(ii) Vergleich nach Fisher-Transformation:

$$\hat{\xi} = \frac{1}{2} \log \left(\frac{1+\hat{\theta}}{1-\hat{\theta}} \right) \overset{\text{approx.}}{\sim} N \left[\frac{1}{2} \log \left(\frac{1+\theta}{1-\theta} \right), \left(\frac{1}{\sqrt{n-3}} \right)^2 \right]$$

Zweistichproben-Problem für unabhängige Stichproben

$$\left.\begin{array}{l} Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} F \\ Z_1, \ldots, Z_m \overset{\text{i.i.d.}}{\sim} G \end{array}\right\} \text{unabhängig}$$

Schätzung des Standardfehlers der Schätzung für die Differenz $\theta = \mu_Y - \mu_Z$*:*

$$y^{*b} = (y_1^{*b}, \ldots, y_n^{*b}) \text{ zufällig mit Zurücklegen aus } \hat{F}_n$$
$$z^{*b} = (z_1^{*b}, \ldots, z_m^{*b}) \text{ zufällig mit Zurücklegen aus } \hat{G}_m$$

$$\widehat{\text{se}}_{F,G}(\hat{\theta}) = \text{se}_{\hat{F}_n, \hat{G}_m}(\hat{\theta}^*) \approx \widehat{\text{se}}_B = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} \left[\hat{\theta}^*(b) - \hat{\theta}^*(\cdot) \right]^2 \right\}^{\frac{1}{2}}$$

mit

$$\hat{\theta}^*(b) = \frac{1}{n} \sum_{i=1}^{n} y_i^{*b} - \frac{1}{m} \sum_{i=1}^{m} z_i^{*b}$$

und

$$\hat{\theta}^*(\cdot) = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^*(b)$$

Bootstrap im Linearen Modell

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} F, \quad \mathbb{E}(\varepsilon_i) = 0$$

Nichtparametrischer Bootstrap

1. Schritt: Berechne $\hat{\boldsymbol{\beta}}$ mit der KQ-Methode: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$.

2. Schritt: Berechne die Residuen $\hat{\boldsymbol{\varepsilon}} = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top)\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$.

3. Schritt: Setze für die empirische Verteilung $\hat{F}_n$ der Residuen eine Wahrscheinlichkeitsmasse $\frac{1}{n}$ auf $\hat{\varepsilon}_i$, $i = 1, \ldots, n$ (ohne weitere Einschränkung seien alle Residuen verschieden).

4. Schritt:
- Ziehe eine Stichprobe $\boldsymbol{\varepsilon}^* = (\varepsilon_1^*, \ldots, \varepsilon_n^*)$ mit Zurücklegen aus $\hat{F}_n$.
- Berechne „neue“ Bootstrap-Zielvariablen
$$y_i^* = \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} + \varepsilon_i^*$$
für $i = 1, \ldots, n$, d.h.
$$\boldsymbol{y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}^*.$$
- Berechne den Bootstrap-KQ-Schätzer
$$\hat{\boldsymbol{\beta}}^* = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}^*.$$

Parametrischer Bootstrap

1. Schritt: Berechne $\hat{\boldsymbol{\beta}}_{KQ}$ und $\hat{\sigma}_F^2$.

2. Schritt: Setze $y_i^* = \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_{KQ} + \varepsilon_i^*$, wobei $\varepsilon_i^* \sim N(0, \hat{\sigma}_F^2)$.

Bias-Schätzung mittels Bootstrap

$$\text{bias}_F(\hat{\theta}, \theta) = \mathbb{E}_F(\hat{\theta}) - \theta = \mathbb{E}_F(\hat{\theta}) - T(F)$$

$$\widehat{\text{bias}}_F(\hat{\theta}, \theta) = \text{bias}_{\hat{F}_n}(\hat{\theta}^*, \hat{\theta}) = \mathbb{E}_{\hat{F}_n}[\hat{\theta}^*] - T(\hat{F}_n)$$

Bias-Korrektur:

$$\overline{\theta} = 2\hat{\theta} - \hat{\theta}^*(\cdot)$$

Bootstrap-Konfidenzintervalle

Bootstrap-t-Intervall

1. Generiere B Bootstrap-Stichproben $x^{*1}, \dots, x^{*B}$.
2. Berechne
$$Z^*(b) = \frac{\hat{\theta}^*(b) - \hat{\theta}}{\widehat{\text{se}}^*(b)} \ .$$
Ordne die $Z^*(b)$ aufsteigend der Größe nach.
3. Schätze die Quantile $\hat{t}^{(\alpha)}$ und $\hat{t}^{(1-\alpha)}$ als
$$\frac{\#\left\{Z^*(b) \leq \hat{t}^{(\alpha)}\right\}}{B} = \alpha \ .$$
4. Das Bootstrap-t-Intervall zum Vertrauensgrad $1 - 2\alpha$ lautet dann
$$\left[\hat{\theta} - \hat{t}^{(1-\alpha)} \cdot \widehat{\text{se}},\ \hat{\theta} - \hat{t}^{(\alpha)} \cdot \widehat{\text{se}}\right] .$$

Bootstrap-Perzentil-Intervall

1. Ziehe
$$\begin{array}{cccl} x^{*1}, & \dots, & x^{*B} & B \text{ Bootstrap-Replikationen} \\ \downarrow & & \downarrow & \\ \hat{\theta}^*(1), & \dots, & \hat{\theta}^*(B) & \text{mit } \hat{\theta}^*(b) = T(x^{*b}). \end{array}$$
2. Ordne die $\hat{\theta}^*(b)$ der Größe nach: $\hat{\theta}^*_{(1)}, \dots, \hat{\theta}^*_{(B)}$.
3. Berechne $B\alpha$ und $B(1-\alpha)$ und bezeichne mit $\hat{\theta}_B^{*(\alpha)}$ bzw. $\hat{\theta}_B^{*(1-\alpha)}$ die Werte an den jeweiligen Positionen in der sortierten Sequenz der Bootstrap-Schätzungen. Dann ist
$$\left[\hat{\theta}_{\text{lower}}, \hat{\theta}_{\text{upper}}\right] = \left[\hat{\theta}_B^{*(\alpha)}, \hat{\theta}_B^{*(1-\alpha)}\right]$$
ein approximatives $(1 - 2\alpha)$-Konfidenzintervall.

Kreuzvalidierung und Vorhersagefehler

k-fache Kreuzvalidierung:
$$\text{CV} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{y}_i^{-k(i)}\right)^2$$

Vorhersagefehler:
$$\text{err}(x, F) \equiv \mathbb{E}_{0F}(Q(Y_0, \eta_x(\boldsymbol{Z}_0)))$$

Apparent error in sample:
$$\text{err}(x, \hat{F}_n) = \mathbb{E}_{0\hat{F}_n}(Q(Y_0, \eta_x(\boldsymbol{Z}_0)) = \frac{1}{n} \sum_{i=1}^{n} Q[y_i, \eta_x(\boldsymbol{z}_i)]$$

Plug–In–Schätzung:
$$\text{err}(x^{*b}, \hat{F}_n) = \frac{1}{n} \sum_{i=1}^{n} Q[y_i, \eta_{x^{*b}}(\boldsymbol{z}_i)]$$

Durchschnittlicher Vorhersagefehler:
$$\mathbb{E}_F[\text{err}(x, F)]$$

Approximative Bootstrap-Schätzung:
$$\hat{\mathbb{E}}_{\hat{F}_n}[\text{err}(x^*, \hat{F}_n)] = \frac{1}{B} \sum_{b=1}^{B} \frac{1}{n} \sum_{i=1}^{n} Q[y_i, \eta_{x^{*b}}(\boldsymbol{z}_i)]$$

In bootstrap-sample error:

$$\hat{\mathbb{E}}_{\hat{F}_n}[\mathrm{err}(x^*, \hat{F}_n^*)] = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{n}\sum_{i=1}^{n} Q[y_i^{*b}, \eta_{x^{*b}}(\boldsymbol{z}_i^{*b})]$$

Average optimism:

$$w(F) = \mathbb{E}_F(\mathrm{err}(x, F)) - \mathbb{E}_F(\mathrm{err}(x, \hat{F}_n))$$

Approximative Bootstrap-Schätzung:

$$\hat{w}(\hat{F}_n) = \frac{1}{Bn}\left\{\sum_{b=1}^{B}\sum_{i=1}^{n} Q[y_i, \eta_{x^{*b}}(\boldsymbol{z}_i)] - \sum_{b=1}^{B}\sum_{i=1}^{n} Q[y_i^{*b}, \eta_{x^{*b}}(\boldsymbol{z}_i^{*b})]\right\}$$

Schätzung des Vorhersagefehlers mit Bias-Korrektur:

$$\mathrm{err}(x, \hat{F}_n) + w(\hat{F}_n) \qquad \text{wird geschätzt durch} \qquad \frac{1}{n}\sum_{i=1}^{n} Q[y_i, \eta_x(\boldsymbol{z_i})] + \hat{w}(\hat{F}_n)$$

6 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

ML-Schätzung bei Fehlspezifikation

Kullback-Leibler-Distanz von g und f_θ für X stetig:

$$D(g, f_\theta) = \mathbb{E}_g\left(\log\frac{g(X)}{f(X|\theta)}\right)$$

Entropie von g:

$$-\mathbb{E}_g \log g(X) = -\int g(x)\,\log(g(x))\,dx$$

Asymptotische Eigenschaften des ML-Schätzers bei Missspezifikation

1. Konsistenz: Sei θ_0 ein (lokaler) Maximierer von

   $$\lambda(\theta) \equiv \mathbb{E}_g \log f_\theta(X|\theta)$$

   (bzw. ein Minimierer von $D(g, f_\theta)$). Unter Regularitätsannahmen (ähnlich wie bei Fisher-Regularität) existiert eine Folge $\hat{\theta}_n$ von („Quasi-") ML-Schätzern, das heißt lokalen Maximierern von

   $$\frac{1}{n}\sum_{i=1}^{n}\log f(x_i|\theta)$$

   mit

   $$\hat{\theta}_n \xrightarrow{P} \theta_0.$$

2. Asymptotische Normalität: Es gilt

   $$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(\mathbf{0}, \boldsymbol{J}_1^{-1}(\theta_0)\,\boldsymbol{I}_1(\theta_0)\,\boldsymbol{J}_1^{-1}(\theta_0)\right)$$

   mit

   $$\boldsymbol{I}_1(\theta) \equiv \mathbb{E}_g \underbrace{\left(\frac{\partial \log f(X|\theta)}{\partial\theta}\right)}_{\boldsymbol{s}_1(\theta)}\underbrace{\left(\frac{\partial \log f(X|\theta)}{\partial\theta}\right)^\top}_{\boldsymbol{s}_1(\theta)^\top}$$

   und der (Quasi-)Fisher-Information

   $$\boldsymbol{J}_1(\theta) = \mathbb{E}_g\left(-\frac{\partial^2 \log f(X|\theta)}{\partial\theta\,\partial\theta^\top}\right).$$

M-Schätzer in der robusten Statistik

Asymptotische Eigenschaften von M–Schätzern $\hat{\boldsymbol{\theta}}_M$
Unter Regularitätsvoraussetzungen, insbesondere

$$\mathbb{E}_0\, \mathbf{qs}(\boldsymbol{\theta}_0) = \mathbf{0},$$

gilt

$$\hat{\boldsymbol{\theta}}_M \overset{a}{\sim} N(\boldsymbol{\theta}_0, \boldsymbol{V}(\hat{\boldsymbol{\theta}}_M))\ .$$

Dabei ist $\boldsymbol{V}(\hat{\boldsymbol{\theta}}_M)$ definiert als

$$\boldsymbol{V}(\hat{\boldsymbol{\theta}}_M) = \boldsymbol{J}^{-1}(\hat{\boldsymbol{\theta}}_M)\boldsymbol{I}(\hat{\boldsymbol{\theta}}_M)\boldsymbol{J}^{-1}(\hat{\boldsymbol{\theta}}_M)$$

mit der empirischen (Quasi–) Fisher–Matrix

$$\boldsymbol{I}(\hat{\boldsymbol{\theta}}_M) = \sum_{i=1}^{n} \mathbf{qs}_i(\hat{\boldsymbol{\theta}}_M)\mathbf{qs}_i^\top(\hat{\boldsymbol{\theta}}_M)$$

und der (empirischen) beobachteten (Quasi–) Informationsmatrix

$$\boldsymbol{J}(\hat{\boldsymbol{\theta}}_M) = -\left.\frac{\partial\, \mathbf{qs}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_M}\ .$$

Verallgemeinerte Schätzgleichungen (Generalized Estimating Equations)

$$U(\beta,\alpha) = \sum_{i=1}^{N} \left(\frac{\partial \mu_i(\beta)}{\partial \beta}\right)^\top V_i(\beta,\alpha,\phi)^{-1}(Y_i - \mu_i(\beta)) = 0$$

Sei T_i die Anzahl der Beobachtungen an Individuum i und ϕ der Dispersionsparameter.

- $\left(\frac{\partial \mu_i}{\partial \beta}\right)^T$ ist eine $p \times T_i$ Matrix (GLM: $p \times 1$ Vektor)
- Y_i und μ_i sind $T_i \times 1$ Vektoren (GLM: Skalare)
- V_i ist die $T_i \times T_i$ Arbeitskovarianzmatrix (GLM: Skalar)

$$V_i(\beta,\alpha,\phi) = \begin{pmatrix} c_{i11}(\beta,\alpha,\phi) & c_{i12}(\beta,\alpha,\phi) & \cdots & c_{i1T_i}(\beta,\alpha,\phi) \\ c_{i21}(\beta,\alpha,\phi) & c_{i22}(\beta,\alpha,\phi) & \cdots & c_{i2T_i}(\beta,\alpha,\phi) \\ \vdots & \vdots & \vdots & \vdots \\ c_{iT_i1}(\beta,\alpha,\phi) & c_{iT_i2}(\beta,\alpha,\phi) & \cdots & c_{iT_iT_i}(\beta,\alpha,\phi) \end{pmatrix}$$

  mit

$$c_{itt'}(\beta,\alpha,\phi) = \begin{cases} \text{Cov}(Y_{it}, Y_{it'}; \beta,\alpha,\phi), & \text{falls} \quad t \neq t' \\ \text{Var}(Y_{it}; \beta,\phi), & \text{falls} \quad t = t'. \end{cases}$$

- Bei Verwendung einer "Arbeitskorrelation" berechnet sich die "Arbeitskovarianz" als

$$V_i(\beta,\alpha,\phi) = \phi \boldsymbol{A}_i(\beta)^{\frac{1}{2}} \boldsymbol{R}_i(\alpha) \boldsymbol{A}_i(\beta)^{\frac{1}{2}}$$

  mit der Diagonalmatrix der Varianzfunktionen

$$\boldsymbol{A}_i(\beta) = \text{diag}[v(\mu_{i1}(\beta)), v(\mu_{i2}(\beta)), \ldots, v(\mu_{iT_i}(\beta))]$$

Varianzschätzung mittels Sandwich-Schätzer
Sei

$$F = \sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \beta}\right)^\top V_i^{-1} \left(\frac{\partial \mu_i}{\partial \beta}\right),$$

wobei V_i die modellbasierte Kovarianzmatrix ist, die sich aus den Annahmen über Varianz und Korrelation ergibt, und

$$G = \sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \beta}\right)^\top V_i^{-1} \text{Cov}(Y_i) V_i^{-1} \left(\frac{\partial \mu_i}{\partial \beta}\right),$$

wobei $\mathrm{Cov}(Y_i)$ die wahre Kovarianzmatrix ist. Dann gilt:

$$\widehat{\mathrm{Cov}}(\widehat{\beta}) = \widehat{F}^{-1}\widehat{G}\widehat{F}^{-1}$$

ist die empirirsche konsistente Schätzung der Kovarianzmatrix von $\widehat{\beta}$. $\widehat{F}$ und $\widehat{G}$ sind die empirischen Matrizen, die sich durch Einsetzen von $\widehat{\beta}, \widehat{\alpha}$ und $\widehat{\phi}$ ergeben. $\mathrm{Cov}(Y_i)$ wird konsistent geschätzt durch die Matrix $(y_i - \mu_i)(y_i - \mu_i)^\top$. Im Idealfall gilt:

$$\mathrm{Cov}(Y_i) = V_i.$$

Dann ist GEE effizient und man erhält

$$\mathrm{Cov}(\widehat{\beta}) = F^{-1}$$

bzw. die Schätzung

$$\widehat{\mathrm{Cov}}(\widehat{\beta}) = \widehat{F}^{-1}.$$

Eigenschaften der GEE Schätzung

- Verwendet man als "Arbeitskorrelation" die Unabhängigkeit, so spricht man vom IEE Schätzer (independence estimating equations). Man erhält

  $$V_i(\beta, \alpha, \phi) = V_i(\beta, \phi) = \phi \boldsymbol{A}_i(\beta)$$

  und die Schätzgleichung reduziert sich dann auf

  $$U(\beta) = \sum_{i=1}^{N} \left(\frac{\partial \mu_i(\beta)}{\partial \beta}\right)^\top V_i(\beta, \phi)^{-1}(Y_i - \mu_i(\beta)) = 0.$$

  Die IEE Schätzung $\widehat{\beta}_I$ ist unter schwachen Voraussetzungen konsistent und asymptotisch normalverteilt, wenn ϕ konsistent geschätzt wird und das Modell für den Erwartungswert korrekt spezifiziert ist.

- Wird als "Arbeitskorrelation" nicht die Unabhängigkeit gewählt, so ist die GEE Schätzung $\widehat{\beta}_G$ unter schwachen Vorraussetzungen konsistent und asymptotisch normalverteilt, wenn ϕ und α $N^{\frac{1}{2}}$-konsistent geschätzt werden.

Quantilregression

Bedingte Quantilsfunktion:

$$Q_\tau(y|\boldsymbol{X} = \boldsymbol{x}) = F^{-1}_{Y|\boldsymbol{X}=\boldsymbol{x}}(\tau|\boldsymbol{X} = \boldsymbol{x}) = y_\tau(\boldsymbol{x})$$

$$Q_\tau(y|\boldsymbol{X} = \boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}_\tau$$

Modellformel:

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau + \varepsilon_{\tau i}$$

mit unabhängigen, aber möglicherweise heteroskedastischen $\varepsilon_{\tau i}$ und

$$F_{\varepsilon_{\tau i}}(0) = \int_{-\infty}^{0} f(\varepsilon_{\tau i})\, d\varepsilon_{\tau i} = \tau$$

Entscheidungstheoretischer Ansatz

Check-Funktion:

$$\rho_\tau(u) = u \cdot (\tau - I(u < 0)), \quad \tau \in (0, 1)$$

Zielfunktion:

$$\underset{\boldsymbol{\beta}_\tau \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau)$$

Erwarteter Verlust:

$$\mathbb{E}_{F_Y}[\rho_\tau(y - \hat{y})]$$

Satz. *Der erwartete Verlust wird minimiert durch* $\hat{y} = F_Y^{-1}(\tau)$.

Quasi-ML-Ansatz

Asymmetrische Laplace-Verteilung (ALD):

$$Y \sim \text{ALD}(\mu, \sigma, \tau)$$

mit $-\infty < y < \infty$, $\mu \in \mathbb{R}$, $\sigma > 0$ und $\tau \in (0,1)$. Die Dichtefunktion der ALD lautet

$$f_Y(y) = \frac{\tau(1-\tau)}{\sigma} \cdot \exp\left\{-\rho_\tau\left(\frac{y-\mu}{\sigma}\right)\right\}$$

Quasi-Likelihood:

$$\text{QL}(\boldsymbol{\beta}_\tau) \propto \frac{1}{\sigma}\exp\left\{-\sum_{i=1}^{n}\rho_\tau\left(\frac{y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}_\tau}{\sigma}\right)\right\}$$

Eigenschaften der Quantilregression

- *Äquivarianz:*

$$\hat{\boldsymbol{\beta}}_\tau(a\boldsymbol{y}, \boldsymbol{X}) = a\hat{\boldsymbol{\beta}}_\tau(\boldsymbol{y}, \boldsymbol{X}),$$
$$\hat{\boldsymbol{\beta}}_\tau(\boldsymbol{y}, \boldsymbol{X}\mathbf{A}) = \mathbf{A}^{-1}\hat{\boldsymbol{\beta}}_\tau(\boldsymbol{y}, \boldsymbol{X})$$

- *Asymptotische Verteilung:*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}_\tau) \to N(0, \tau(1-\tau)\mathbf{H}^{-1}(\tau)\mathbf{J}(\tau)\mathbf{H}^{-1}(\tau))$$

  mit

$$\begin{aligned}\mathbf{J}(\tau) &= \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top, \\ \mathbf{H}(\tau) &= \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top \cdot f_i(y_{i\tau})\end{aligned}$$

7 Non- und Semiparametrische Inferenz

Kerndichteschätzung

Definition (Histogramm)**.** *Sei $x_1, \dots, x_n$ eine i.i.d. Stichprobe einer stetigen Zufallsvariable X mit Dichte f. Dann heißt der Schätzer*

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n}\sum_{j\in\mathbb{Z}} I_{B_j}(x_i)I_{B_j}(x)$$

Histogramm mit Klassenbreite (Bandweite) $h > 0$ und Ursprung x_0.

Definition (Kerndichteschätzer)**.** *Der Schätzer*

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-x_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n}K_h(x-x_i)$$

mit

$$K_h(u) := \frac{1}{h}K\left(\frac{u}{h}\right)$$

heißt Kerndichteschätzer *mit Kern K (bzw. K_h) und Bandweite $h > 0$.*

- Dreieckskern*: $K(u) = (1-|u|)I_{[-1,1]}(u)$,*
- Epanechnikovkern*: $K(u) = \frac{3}{4}(1-u^2)I_{[-1,1]}(u)$,*
- Normalkern*: $K(u) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}u^2)$.*

Erwartungswert, Varianz, Bias, MSE und MISE des Kerndichteschätzers

$$\mathbb{E}(\hat{f}_h(x)) = \frac{1}{h}\int\limits_{\mathbb{R}} K\left(\frac{x-y}{h}\right) f(y)\,dy$$

$$\mathrm{Var}(\hat{f}_h(x)) = \frac{1}{nh^2}\int\limits_{\mathbb{R}} K^2\left(\frac{x-y}{h}\right) f(y)\,dy - \frac{1}{n}\mathbb{E}(\hat{f}_h(x))^2$$

$$\mathrm{bias}(\hat{f}_h(x)) = \mathbb{E}(\hat{f}_h(x)) - f(x) = \frac{1}{h}\int\limits_{\mathbb{R}} K\left(\frac{x-y}{h}\right) f(y)\,dy - f(x)$$

$$\mathrm{MSE}(\hat{f}_h(x)) = \frac{1}{nh^2}\int\limits_{\mathbb{R}} K^2\left(\frac{x-y}{h}\right) f(y)\,dy - \frac{1}{n}\,\mathbb{E}(\hat{f}_h(x))^2 + \left(\frac{1}{h}\int\limits_{\mathbb{R}} K\left(\frac{x-y}{h}\right) f(y)\,dy - f(x)\right)^2$$

$$\mathrm{MISE}(\hat{f}_h) = \int\limits_{\mathbb{R}} \mathrm{MSE}(\hat{f}_h(x))\,dx.$$

Konsistenz des Kerndichteschätzers

Satz (Satz von Parzen)**.** *Sei $R(x)$, $x \in \mathbb{R}$, eine (messbare) Funktion mit den Eigenschaften*

1. $\sup\limits_{x\in\mathbb{R}} |R(x)| < \infty$ *(d.h. $R(x)$ ist beschränkt),*
2. $\int\limits_{\mathbb{R}} |R(x)|\,dx < \infty$,
3. $|x|R(x) \to 0$ *für* $|x| \to \infty$.

Sei weiterhin $g(x)$, $x \in \mathbb{R}$, eine (messbare) Funktion mit $\int\limits_{\mathbb{R}} |g(x)|\,dx < \infty$. Betrachte die Folge

$$g_n(x) = \frac{1}{h_n}\int\limits_{\mathbb{R}} R\left(\frac{x-y}{h_n}\right) g(y)\,dy,$$

wobei h_n eine Folge ist mit $\lim_{n\to\infty} h_n = 0$. Dann gilt für jeden Stetigkeitspunkt x von g

$$g_n(x) \to g(x)\int\limits_{\mathbb{R}} R(s)\,ds$$

falls $n \to \infty$.

Satz (Konsistenz des Kerndichteschätzers)**.** *Sei f stetig. Dann gilt*

$$\mathbb{E}(\hat{f}_{h_n}(x)) \to f(x),$$

falls die Bandweite h_n für $n \to \infty$ gegen Null konvergiert. $\hat{f}_{h_n}(x)$ ist also asymptotisch erwartungstreu. Falls $nh_n \to \infty$ für $n \to \infty$, dann gilt

$$\mathrm{Var}(\hat{f}_{h_n}(x)) \to 0.$$

Damit ist $\hat{f}_{h_n}(x)$ konsistent.

Konvergenzordnung des MISE

Satz. *Sei f mindestens zweimal stetig differenzierbar, f'' beschränkt, f und f'' quadratintegrierbar. Sei h_n eine Folge mit $h_n \to 0$ und $nh_n \to \infty$ für $n \to \infty$. Unter Verwendung der Abkürzungen $\int\limits_{\mathbb{R}} g^2(s)\,ds = ||g||_2^2$ und $\mu_2(g) = \int\limits_{\mathbb{R}} g(s)s^2\,ds$ für eine Funktion g gilt:*

1. $\mathrm{Var}(\hat{f}_{h_n}(x)) = \dfrac{1}{nh_n}||K||_2^2 f(x) + o\left(\dfrac{1}{nh_n}\right)$ *bzw.*

   $$\int\limits_{\mathbb{R}} \mathrm{Var}(\hat{f}_{h_n}(x))\,dx = \frac{1}{nh_n}||K||_2^2 + o\left(\frac{1}{nh_n}\right).$$

2. $\text{bias}(\hat{f}_{h_n}(x)) = \frac{{h_n}^2}{2}\mu_2(K)f''(x) + o({h_n}^2)$ *bzw.*

$$\int_{\mathbb{R}} \text{bias}^2(\hat{f}_{h_n}(x))\,dx = \frac{{h_n}^4}{4}\mu_2^2(K)||f''||_2^2 + o({h_n}^4).$$

3. $\text{MISE}(\hat{f}_{h_n}) = \frac{1}{nh_n}||K||_2^2 + \frac{{h_n}^4}{4}\mu_2^2(K)||f''||_2^2 + o\Big(\frac{1}{nh_n} + {h_n}^4\Big).$

Asymptotische Verteilung und Konfidenzintervalle

Satz (Asymptotische Verteilung)**.** *$f''(x)$ existiere; es gelte $h_n = cn^{-1/5}$. Dann ist der Kern-Dichteschätzer $\hat{f}_{h_n}(x)$ asymptotisch normalverteilt,*

$$n^{\frac{2}{5}}\left\{\hat{f}_{h_n}(x) - f(x)\right\} \xrightarrow{d} N\Big(\underbrace{\frac{c^2}{2}f''(x)\;\mu_2(K)}_{b_x},\underbrace{c^{-1}f(x)||K||_2^2}_{v_x^2}\Big)$$

für $n \to \infty$.

Approximatives (1-α)- Konfidenzintervall

$$\left[\hat{f}_h(x) - \frac{h^2}{2}f''(x)\;\mu_2(K) - z_{1-\frac{\alpha}{2}}\sqrt{\frac{f(x)||K||_2^2}{nh}}\;,\;\hat{f}_h(x) - \frac{h^2}{2}f''(x)\;\mu_2(K) + z_{1-\frac{\alpha}{2}}\sqrt{\frac{f(x)||K||_2^2}{nh}}\;\right]$$

Integrated Squared Error (ISE)

$$\text{ISE}(h) = \int_{\mathbb{R}}(\hat{f}_h(x) - f(x))^2\,dx = \int_{\mathbb{R}}\hat{f}_h^2(x)\,dx - 2\int_{\mathbb{R}}\hat{f}_h(x)f(x)\,dx + \int_{\mathbb{R}}f^2(x)\,dx$$

Kreuzvalidierungsfunktion

$$\text{CV}(h) = \frac{1}{n^2h}\sum_{i=1}^{n}\sum_{j=1}^{n}(K\star K)\left(\frac{x_j - x_i}{h}\right) - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{h,i}(x_i)$$

Bayesianische nichtparametrische Dichteschätzung

Dirichlet-Verteilung

Dirichlet–Dichte:

$$p(\pi|\alpha) = \frac{\Gamma(\sum_{i=1}^m \alpha_i)}{\prod_{i=1}^m \Gamma(\alpha_i)}\pi_1^{\alpha_1-1}\pi_2^{\alpha_2-1}\cdots\pi_m^{\alpha_m-1}\;,$$

wobei $\pi_m = 1 - \sum_{k=1}^{m-1}\pi_k$. Kurz:

$$\pi \sim \text{Diri}(\alpha_1, \ldots, \alpha_m)\;.$$

Es gilt mit $\overline{\alpha} := \alpha_1 + \ldots + \alpha_m$:

$$\begin{aligned}
\mathbb{E}(\pi_i) &= \frac{\alpha_i}{\overline{\alpha}} \\
\mathbb{E}(\pi_i^2) &= \frac{\alpha_i(\alpha_i+1)}{\overline{\alpha}(\overline{\alpha}+1)} \\
\mathbb{E}(\pi_i\pi_j) &= \frac{\alpha_i\alpha_j}{\overline{\alpha}(\overline{\alpha}+1)} \qquad i \neq j\;.
\end{aligned}$$

Äquivalente Definition: $Z_1, Z_2, \ldots, Z_m$ seien unabhängige Gamma$(\alpha_i, 1)$–verteilte Zufallsvariablen, $\alpha_i > 0$. Dann gilt für $\pi = (\pi_1, \ldots, \pi_m)$ mit

$$\pi_i = \frac{Z_i}{\sum_{j=1}^m Z_j}\;:$$

$$\pi \sim \text{Diri}(\alpha_1, \alpha_2, \ldots, \alpha_m)\;.$$

Dirichlet-Prozesse

Definition (Dirichlet-Prozess). *Ein* Dirichlet-Prozess (DP) *ist ein RPM G auf* $(\Omega, \mathcal{A}) :\overset{def}{\Leftrightarrow}$
Für jede finite Partition $(A_1, \ldots, A_m)$ *von* Ω *mit* $A_j \in \mathcal{A}$ *ist der Zufallsvektor* $(G(A_1), \ldots, G(A_m))$ *dirichlet-verteilt mit*

$$(G(A_1), \ldots, G(A_m)) \sim \mathit{Diri}(\alpha_0\, G_0(A_1), \ldots, \alpha_0\, G_0(A_m)),$$

kurz:

$$\boxed{G \sim \mathit{DP}(\alpha_0, G_0).}$$

G_0 *heißt* Basis-Verteilung (base measure), α_0 *heißt* Präzisionsparameter *und bestimmt die Varianz um* $\mathbb{E}(G)$.

Konjugiertheit des DP

Sei $G \sim \mathrm{DP}(\alpha_0, G_0)$ und $\theta_1, \ldots, \theta_n \mid G \overset{\text{i.i.d.}}{\sim} G$. Dann ist die Posteriori

$$G \mid \theta_1, \ldots, \theta_n \sim \mathrm{DP}\Big(\alpha_0 + n,\ \frac{\alpha_0}{\alpha_0 + n}\, G_0 + \frac{1}{\alpha_0 + n} \sum_{i=1}^{n} \delta_{\theta_i}\Big).$$

Polya–Urnen–Repräsentation eines DP

Die a posteriori prädiktive Verteilung lautet

$$\theta_n \mid \theta_1, \ldots, \theta_{n-1}, \alpha_0, G_0 \ \sim\ \frac{\alpha_0}{n - 1 + \alpha_0}\, G_0 + \frac{1}{n - 1 + \alpha_0} \sum_{j=1}^{n-1} \delta_{\theta_j}.$$

Steckerlbruch–Repräsentation eines DP

$$G(A) = \sum_{k=1}^{\infty} \pi_k\, \delta_{\phi_k}(A) \qquad \text{mit} \qquad \phi_k \overset{\text{i.i.d.}}{\sim} G_0$$

für beliebiges $A \in \mathcal{A}$ und

$$\pi_k = \beta_k \prod_{j=1}^{k-1} (1 - \beta_j) \qquad \text{mit} \qquad \beta_j \overset{\text{i.i.d}}{\sim} \mathrm{Beta}(1, \alpha_0).$$

Dabei gilt $\sum\limits_{k=1}^{\infty} \pi_k = 1$ sowie $\pi_1 = \beta_1$

Trunkierter Dirichlet–Prozess (TDP)

$$G_T(\cdot) = \sum_{k=1}^{T} \pi_k \delta_{\phi_k}(\cdot), \qquad \pi_T := 1 - \sum_{k=1}^{T-1} \pi_k$$

DP–Mischungen (DPM)

- θ_i ist latenter, mit Datenpunkt x_i assoziierter Parameter.
- Der (trunkierte) DP wird benutzt, um eine Priori für die θ_i zu konstruieren

$$f(x_i) = \sum_{k=1}^{d} \pi_k\, \phi\big(x_i \mid \underbrace{\mu_k, \sigma_k^2}_{\theta_k}\big)$$

- Formalisiert:

$$\begin{aligned} x_i \mid \theta_i &\overset{\text{ind}}{\sim} f(x_i \mid \theta_i) \qquad (\text{bzw.} \overset{\text{ind}}{\sim} F\,(x_i \mid \theta_i)) \\ \theta_i \mid G &\overset{\text{i.i.d}}{\sim} G \\ G &\sim (\mathrm{T})\mathrm{DP}(\alpha_0, G_0) \end{aligned}$$

  für $i = 1, \ldots, n$

Bayes-Inferenz mit DPM-Priori

- *Allgemeines hierarchisches Modell:*

$$\begin{array}{rcll} x_i \mid \theta_i, \delta & \overset{\text{ind}}{\sim} & f\left(x_i \mid \theta_i, \delta\right), & i = 1, \ldots, n \\ \theta_i \mid G_{(T)} & \overset{\text{i.i.d.}}{\sim} & G_{(T)} & \\ \delta & \sim & p(\delta) & \delta \text{ endlich-dimensionaler Parameter} \\ G_{(T)} & \sim & (T)\text{DP}(\alpha_0, G_0) & \end{array}$$

- *Wahl von f:*

  1. $x_i \mid \theta_i, \delta \overset{\text{ind}}{\sim} N(\mu_i, \tau)$
  2. $x_i \mid \theta_i \overset{\text{ind}}{\sim} N(\mu_i, \tau_i)$, $\theta_i = (\mu_i, \tau_i)$
  3. $\boldsymbol{x}_i \mid \theta_i \overset{\text{ind}}{\sim} \text{MVN}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$
  4. $\boldsymbol{x}_i \mid \theta_i \overset{\text{ind}}{\sim} \text{MVN}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$

- *Repräsentation von (T)DP durch (T)SB:*

$$G_T(\cdot) = \sum_{k=1}^{T} \pi_k \, \delta_{\phi_k}(\cdot), \qquad \phi_k \overset{\text{i.i.d.}}{\sim} G_0$$

  Dabei ist $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_T)$ ein trunkierter SB-Prozess:

$$\begin{array}{rcll} \pi_1 & = & \beta_1 & \\ \pi_k & = & (1-\beta_1)(1-\beta_2) \cdot \cdots \cdot (1-\beta_{k-1})\beta_k & \text{für } k = 2, \ldots, T-1 \\ \pi_T & = & 1 - \pi_1 - \cdots - \pi_{T-1} & \end{array}$$

  und

$$\beta_1, \ldots, \beta_{T-1} \overset{\text{i.i.d.}}{\sim} \text{Beta}(1, \alpha_0).$$

Glättung und semiparametrische Regression

Glättung für normalverteilte Zeitreihen

Betrachte die Zeitreihe $\boldsymbol{y} = (y_1, \ldots, y_t, \ldots, y_n)^\top$ mit glattem Trend $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)^\top$.

- **Klassische Glättung**

  Modell:

$$y_t = \gamma_t + \varepsilon_t, \qquad t = 1, \ldots, n$$

  Differenzen erster und d-ter Ordnung:

$$\Delta^1 \gamma_t = \gamma_t - \gamma_{t-1}\ , \quad \Delta^d \gamma_t = \Delta^{d-1}\gamma_t - \Delta^{d-1}\gamma_{t-1}\ , \quad d = 2, 3, \ldots$$

  Differenzenmatrizen:

$$\boldsymbol{D}_1^{(n)} = \begin{pmatrix} -1 & 1 & & 0 \\ & \ddots & \ddots & \\ 0 & & -1 & 1 \end{pmatrix} \in \mathbb{R}^{(n-1)\times n}\ , \quad \boldsymbol{D}_d^{(n)} = \boldsymbol{D}_{d-1}^{(n-1)} \boldsymbol{D}_1^{(n)} \in \mathbb{R}^{(n-d)\times n}$$

  Strafmatrizen:

$$\boldsymbol{K}_d = \boldsymbol{D}_d^{(n)\top} \boldsymbol{D}_d^{(n)} \in \mathbb{R}^{n\times n}, \qquad \text{rg}(\boldsymbol{K}_d) = n - d$$

  Straffunktion:

$$\text{pen}(\boldsymbol{\gamma}) = \sum_{t=d+1}^{n} (\Delta^d \gamma_t)^2 = \boldsymbol{\gamma}^\top \boldsymbol{K}_d \boldsymbol{\gamma}$$

Penalisiertes KQ-Kriterium:

$$\mathrm{PKQ}(\boldsymbol{\gamma}) = (\boldsymbol{y} - \boldsymbol{\gamma})^\top (\boldsymbol{y} - \boldsymbol{\gamma}) + \lambda \boldsymbol{\gamma}^\top \boldsymbol{K} \boldsymbol{\gamma}$$

Penalisierter KQ-Schätzer:

$$\hat{\boldsymbol{\gamma}}_{\mathrm{PKQ}} = (\boldsymbol{I} + \lambda \boldsymbol{K})^{-1} \boldsymbol{y}$$

- **Bayesianische Glättung**

Likelihood:

$$f(\boldsymbol{y}|\boldsymbol{\gamma}) \sim N(\boldsymbol{\gamma}, \sigma^2 \boldsymbol{I})$$

Glattheits-Priori:

$$p(\boldsymbol{\gamma}) \propto (\tau^2)^{-\mathrm{rg}(\boldsymbol{K})/2} \exp\left(-\frac{1}{2\tau^2} \boldsymbol{\gamma}^\top \boldsymbol{K} \boldsymbol{\gamma}\right)$$

Posteriori-Erwartungswert:

$$\mathbb{E}(\boldsymbol{\gamma}|\boldsymbol{y}) = \hat{\boldsymbol{\gamma}}_{\mathrm{PKQ}} \quad \text{mit} \quad \lambda := \frac{\sigma^2}{\tau^2}$$

Glättung für nicht-normalverteilte Zeitreihen

Sei $y_t|\gamma_t \sim$ einfache Exponentialfamilie.

- **Klassische Glättung**

Penalisierte log-Likelihood:

$$l_{\mathrm{pen}}(\boldsymbol{\gamma}) = l(\boldsymbol{\gamma}) - \frac{\lambda}{2} \boldsymbol{\gamma}^\top \boldsymbol{K} \boldsymbol{\gamma}$$

- **Bayesianische Glättung**

Likelihood:

$$f(\boldsymbol{y}|\boldsymbol{\gamma}) = \prod_{t=1}^{n} f(y_t|\gamma_t)$$

Glattheits-Priori:

$$p(\boldsymbol{\gamma}) \propto (\tau^2)^{-\mathrm{rg}(\boldsymbol{K})/2} \exp\left(-\frac{1}{2\tau^2} \boldsymbol{\gamma}^\top \boldsymbol{K} \boldsymbol{\gamma}\right)$$

Posteriori-Modus:

$$\underset{\boldsymbol{\gamma}}{\operatorname{argmax}}\Big(\log f(\boldsymbol{y}|\boldsymbol{\gamma}) + \log p(\boldsymbol{\gamma})\Big) = \underset{\boldsymbol{\gamma}}{\operatorname{argmax}}\Big(l(\boldsymbol{\gamma}) - \frac{1}{2\tau^2} \boldsymbol{\gamma}^\top \boldsymbol{K} \boldsymbol{\gamma}\Big) = \underset{\boldsymbol{\gamma}}{\operatorname{argmax}}\, l_{\mathrm{pen}}(\boldsymbol{\gamma}) \quad \text{mit } \lambda := \frac{1}{\tau^2}$$

P-Splines

Definition (Spline-Funktionen, Polynom-Splines). *Eine Funktion* $f : [a, b] \to \mathbb{R}$ *heißt* (Polynom-) Spline *vom Grad* $l \geq 0$ *zu den Knoten* $a \leq \kappa_0 < \kappa_1 < \ldots < \kappa_{M-1} < \kappa_M \leq b$ $\overset{def}{\Leftrightarrow}$

1. $f(x)$ *ist* $(l-1)$*-mal stetig differenzierbar,*
2. $f(x)$ *ist ein Polynom vom Grad* l *für* $x \in [\kappa_m, \kappa_{m+1})$, $m = 0, \ldots, M-1$.

Definition (Trunkierte-Potenz- (truncated power, TP-) Basis). $\left\{B_1^{(l)}(x), \ldots, B_K^{(l)}(x)\right\}$ *ist* TP-Basis *vom Grad* l $\overset{def}{\Leftrightarrow}$

$$B_1^{(l)}(x) = 1, B_2^{(l)}(x) = x, \ldots, B_{l+1}^{(l)}(x) = x^l$$

$$B_{l+2}^{(l)}(x) = (x - \kappa_1)_+^l, \ldots, B_K^{(l)}(x) = (x - \kappa_{M-1})_+^l$$

mit

$$(x - \kappa_k)_+^l = \begin{cases} (x - \kappa_k)^l & \text{für } x \geq \kappa_k \\ 0 & \text{sonst.} \end{cases}$$

Definition (B-Spline-Basis). $\left\{B_1^{(l)}(x), \ldots, B_K^{(l)}(x)\right\}$ *ist B-Spline-Basis vom Grad* $l \overset{def}{\Leftrightarrow}$

1. *Jede Basisfunktion ist stückweises,* $(l-1)$*-mal stetig differenzierbares, nichtnegatives Polynom vom Grad* l *über* $l-2$ *benachbarten Knotenpunkten, sonst ist* $B_k(x) = 0$.

2. *Die Basisfunktionen sind so normiert, dass*

$$\sum_{k=1}^{K} B_k^{(l)}(x) = 1 \quad \textit{für alle } x \; .$$

P-Spline-Schätzung bei normalverteilten Zielvariablen

Betrachte $y_i = f(x_i) + \varepsilon_i$ mit unabhängigen $\varepsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, n$, mit $f(x)$ approximiert durch

$$f(x) = \sum_{k=1}^{K} \gamma_k B_k(x) \; , \quad \{B_k(x)\} \text{ eine Spline-Basis} \; .$$

Dies führt zum linearen Modell

$$\boldsymbol{y} = \boldsymbol{Z\gamma} + \boldsymbol{\varepsilon} \; , \quad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

mit Designmatrix

$$\boldsymbol{Z} = \begin{pmatrix} B_1(x_1) & \cdots & B_K(x_1) \\ \vdots & \ddots & \vdots \\ B_1(x_n) & \cdots & B_K(x_n) \end{pmatrix} .$$

- **Klassische Glättung**

  Wahl der Penalisierung:

  – *Bei TP-Basis:*

$$\text{pen}(\boldsymbol{\gamma}) = \sum_{k=l+2}^{K} \gamma_k^2 = \boldsymbol{\gamma}^\top \tilde{\boldsymbol{K}} \boldsymbol{\gamma}$$

$$\tilde{\boldsymbol{K}} = \text{diag}(\boldsymbol{0}_{l+1}, \boldsymbol{1}_{M-1}) = \begin{bmatrix} 0 & & & & & 0 \\ & \ddots & & & & \\ & & 0 & & & \\ & & & 1 & & \\ & & & & \ddots & \\ 0 & & & & & 1 \end{bmatrix}$$

$$\hat{\gamma} = \underset{\boldsymbol{\gamma}}{\text{argmin}} \, \text{PKQ}(\boldsymbol{\gamma}) = (\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \tilde{\boldsymbol{K}})^{-1} \boldsymbol{Z}^\top \boldsymbol{y}$$

  – *Bei B-Spline-Basis*:

$$\text{pen}(\boldsymbol{\gamma}) = \sum_{k=d+1}^{K} (\Delta^d \gamma_k)^2$$

$$\hat{\gamma} = (\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{K})^{-1} \boldsymbol{Z}^\top \boldsymbol{y}$$

- **Bayesianische Glättung**

  Likelihood:

$$\boldsymbol{y} | \boldsymbol{\gamma} \sim N(\boldsymbol{Z\gamma}, \sigma^2 \boldsymbol{I})$$

  Prioriverteilungen:

  – *Bei TP-Basis:*

$$\begin{aligned} p(\gamma_k) \quad &\propto \quad \text{const (oder schwach informativ)} \; , \; k = 1, \ldots, l+1, \\ \gamma_k \quad &\overset{i.i.d.}{\sim} \quad N(0, \tau^2) \; , \; k = l+2, \ldots, K \end{aligned}$$

– *Bei B-Spline-Basis:*

$$p(\boldsymbol{\gamma}) \propto (\tau^2)^{-\mathrm{rg}(\boldsymbol{K})/2} \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\gamma}^\top \boldsymbol{K}\boldsymbol{\gamma}\right)$$

Posterioriverteilung:

$$\boldsymbol{\gamma}|\boldsymbol{y} \sim N(\boldsymbol{\mu}_{\boldsymbol{\gamma}}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}})$$

$$\boldsymbol{\mu}_{\boldsymbol{\gamma}} = \mathbb{E}(\boldsymbol{\gamma}|\boldsymbol{y}) = \left(\boldsymbol{Z}^\top\boldsymbol{Z} + \frac{\sigma^2}{\tau^2}\boldsymbol{K}\right)^{-1} \boldsymbol{Z}^\top\boldsymbol{y} \quad \text{und} \quad \boldsymbol{\Sigma}_{\boldsymbol{\gamma}} = \mathrm{Cov}(\boldsymbol{\gamma}|\boldsymbol{y}) = \sigma^2\left(\boldsymbol{Z}^\top\boldsymbol{Z} + \frac{\sigma^2}{\tau^2}\boldsymbol{K}\right)^{-1}$$

P-Spline-Schätzung bei nicht-normalverteilten Zielvariablen

Sei $y|f \sim$ einfache Exponentialfamilie:

$$\mathbb{E}(y|f(x)) = h(f(x))$$

$$f(x) = \sum \gamma_k B_k(x) \quad \text{bzw.} \quad \boldsymbol{f} = \boldsymbol{Z}\boldsymbol{\gamma}$$

- **Klassische Glättung**

$$\begin{aligned} l_{\text{pen}}(\boldsymbol{\gamma}) &= l(\boldsymbol{\gamma}) - \frac{\lambda}{2}\boldsymbol{\gamma}^\top\boldsymbol{K}\boldsymbol{\gamma} \\ \boldsymbol{s}_{\text{pen}}(\boldsymbol{\gamma}) &= \boldsymbol{s}(\boldsymbol{\gamma}) - \lambda\boldsymbol{K}\boldsymbol{\gamma} \\ \boldsymbol{F}_{\text{pen}}(\boldsymbol{\gamma}) &= \boldsymbol{F}(\boldsymbol{\gamma}) + \lambda\boldsymbol{K} \end{aligned}$$

- **Bayesianische Glättung**

  Das Beobachtungsmodell $f(y|\boldsymbol{\gamma})$ ist durch den GLM-Typ definiert. Bei voller Bayes-Inferenz wählt man a priori $\tau^2 \sim \mathrm{IG}(a, b)$ für $\lambda = \frac{1}{\tau^2}$. Voll bedingte Dichten:

$$\begin{aligned} f(\boldsymbol{\gamma}|\tau^2, \boldsymbol{y}) &\propto \prod_{i=1}^{n} \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi}\right) \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\gamma}^\top\boldsymbol{K}\boldsymbol{\gamma}\right) \\ \tau^2|\boldsymbol{\gamma}, \boldsymbol{y} &\sim \mathrm{IG}\left(a + \frac{\mathrm{rg}(\boldsymbol{K})}{2}, b + \frac{1}{2}\boldsymbol{\gamma}^\top\boldsymbol{K}\boldsymbol{\gamma}\right) \end{aligned}$$

Glättungs-Splines (Smoothing-Splines)

Definition (Natürliche Splines)**.** *Eine Funktion f ist ein* natürlicher (kubischer) Spline *zu den Knoten* $a < \kappa_1 < \ldots < \kappa_m < b \overset{def}{\Leftrightarrow}$

1. *$f(x)$ ist (kubischer) Polynom–Spline zur obigen Knotenmenge.*
2. *$f(x)$ genügt $f''(a) = f''(b) = 0$, d.h. $f(x)$ ist linear in den Intervallen $[a, \kappa_1]$ und $[\kappa_m, b]$.*

Modell:

$$y_i = f(x_i) + \varepsilon_i \quad \text{mit unabhängigen} \quad \varepsilon_i \sim (0, \sigma^2)$$

bzw.

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

Gesucht:

$$\hat{f} = \underset{f \in C^2}{\operatorname{argmin}} \, \mathrm{PKQ}(f)$$

Penalisiertes KQ-Kriterium:

$$\mathrm{PKQ}(f) = \sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda\int(f''(x))^2dx$$

Straffunktion:

$$\mathrm{pen}(\boldsymbol{\gamma}) = \int(f''(z))^2dz = \sum_{j=1}^{K}\sum_{k=1}^{K}\gamma_j\gamma_k\int B_j''(z)B_k''(z)dz = \boldsymbol{\gamma}^\top\boldsymbol{K}\boldsymbol{\gamma}$$

mit

$$\boldsymbol{K} = (K_{jk}) \quad \text{und} \quad K_{jk} = \int B_j''(z)B_k''(z)dz$$

Strukturiert additive Modelle

AM und GAM

- Gegeben:
  - Response $y_i, \quad i = 1, \dots, n$ aus Normalverteilung (additives Modell) oder einfacher Exponentialfamilie (generalisiertes additives Modell)
  - Kovariablen $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ip})'$ mit linearem Einfluss
  - Kovariablen $z_{iq}, \dots, z_{iq}$ mit nonparametrisch modelliertem Einfluss
- **Additives Modell:**
$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + f_1(z_{i1}) + \dots + f_q(z_{iq}) + \varepsilon_i = \eta_i + \varepsilon_i, \qquad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2).$$
- **Generalisiertes additives Modell:**
$$\mathbb{E}(y_i|\eta_i) = \mu_i = h(\eta_i),$$
$$\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta} + f_1(z_{i1}) + \dots + f_q(z_{iq}).$$

Modelle mit variierenden Koeffizienten

- Gegeben:
  - Response y_i
  - Kovariablen $\boldsymbol{x}_i$ mit linearem Effekt
  - Kovariablen $z_{i1}, \dots, z_{iq}$ mit nonparametrischem Effekt
  - Kovariablen $u_{i1}, \dots, u_{iq}$, deren Effekt mit z variiert
  - ansonsten selbe Struktur wie (G)AM
- **Modell mit variierenden Koffizienten (VCM):**
$$\eta_i^{VCM} = \boldsymbol{x}_i'\boldsymbol{\beta} + u_{i1}f_1(z_{i1}) + \dots + u_{iq}f_q(z_{iq}).$$
- ein (G)AM erhält man als Spezialfall mit $u_1 = u_2 = \dots = u_q = 1$

8 Modellwahl

AIC

- Wahres Modell:
$$f_{wahr}(\boldsymbol{y})$$
- Approximierende Modellklasse:
$$f(\boldsymbol{y}|\boldsymbol{\theta})$$

AIC Kriterium:
$$\text{AIC} = -2\log\big(f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{ML})\big) + 2\dim(\hat{\boldsymbol{\theta}}_{ML}).$$

BIC

- Gegeben: Daten $\boldsymbol{y}$, Likelihood $f(\boldsymbol{y}|\boldsymbol{\theta})$ und Priori $p(\boldsymbol{\theta})$.
- Modellwahl mit BIC entspricht Modellwahl basierend auf marginaler Likelihood $f(\boldsymbol{y})$:
$$f(\boldsymbol{y}) = \int_{\boldsymbol{\Theta}} f(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}.$$
- Approximation:
$$\log(f(\boldsymbol{y})) \approx \log\big(f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{ML})\big) - \frac{1}{2}\log(n)\cdot\dim(\hat{\boldsymbol{\theta}}_{ML}).$$

BIC Kriterium:
$$\text{BIC} = -2\log\big(f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{ML})\big) + \log(n)\cdot\dim(\hat{\boldsymbol{\theta}}_{ML})$$

DIC

- Gegeben: (burn-in-bereinigtes) MCMC-Sample $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_L$.
- unstandardisierte Devianz: $D(\boldsymbol{\theta}) = -2\log\big(f(\boldsymbol{y}|\boldsymbol{\theta})\big)$
- Komplexitätsmaß: $p_D = \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}})$

DIC Kriterium:

$$\text{DIC} = D(\bar{\boldsymbol{\theta}}) + 2p_D = -2\log\big(f(\boldsymbol{y}|\boldsymbol{\theta})\big) + 2p_D,$$

wobei $D(\bar{\boldsymbol{\theta}})$ die Devianz am geschätzten Posteriori-Erwartungswert bezeichnet und $\overline{D(\boldsymbol{\theta})}$ die mittlere Devianz über das MCMC-Sample.

Bayes-Faktoren

- Gegeben: 2 Modelle, bezeichnet mit M_k und M_l, mit jeweiligen Parametervektoren $\boldsymbol{\theta}_k$ und $\boldsymbol{\theta}_l$.
- Vergleich zwischen M_k und M_l basierend auf marginaler Likelihood:

$$p(\boldsymbol{y}|M_k) = \int p(\boldsymbol{y}|\boldsymbol{\theta}_k, M_k)p(\boldsymbol{\theta}_k|M_k)\mathrm{d}\boldsymbol{\theta}_k,$$

analog für M_l.

Bayes-Faktor zugunsten von M_k:

$$B_{kl} = \frac{p(\boldsymbol{y}|M_k)}{p(\boldsymbol{y}|M_l)}$$