Angewandte Statistik II

Dr. Uli Wannek

Skript erstellt von Alina Renz

Sommersemester 2018

Eberhard Karls Universität Tübingen
Mathematisch-Naturwissenschaftliche Fakultät
Wilhelm-Schickard-Institut für Informatik

Inhaltsverzeichnis

1 Lineare Modelle

1.1 Zufallsvariable Y

- Verteilung, Erwartungswert, Varianz, Form (Schiefe, Kurtosis,...)
- Parameter der Verteilung (μ, σ), (λ),...
  - Punktschätzer $(\widehat{\mu})$, $(\widehat{\theta})$, ...
  - Konfidenzintervall
- Zusätzlich abhängig von einer Variablen X:

$$\begin{aligned} \mathcal{E}(Y_i) &= \mu_i \\ Y_i &\sim \mathcal{N}(\mu_i, \sigma^2) \end{aligned}$$

  - mit der linearen Abhängigkeit

$$\mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$$

- Ausprägungen
  - nominal, z.B. rot/grün/blau; f/m; Städte
  - ordinal, z.B. kein/etwas/viel; Schulabschluss
  - kardinal/metrisch, z.B. Dosis, Stimulusintensität, -Abstand, -Anzahl
  - speziell dichotom, z.B. ja/nein; klein/groß; 0/1

1.2 Einfaches Lineares Modell

$$Y = \beta_0 + \beta_1 X$$

- abhängige Variable: Zufallsvariable Y
  - (mehrfache) Messung/Realisierung, ergibt Wert y_i
  - *response*
  - fehlerbehaftet
- unabhängige Variable X
  - mit Wert x_i, vom Experimentator vorgegeben, '*control*'
  - mit Wert x_i, fest, mitgemessen, '*covariate*'
  - Vorhersageparameter '*predictor*'

- Linearer Zusammenhang
  - kausale Abhängigkeit Y von X
  - Proportionalitätsfaktor β_1
  - y-Achsenabschnitt β_0 '*intercept*'
- Streuung zulassen

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- Konventionen

Schrift	Bedeutung	Beispiel
Großbuchstaben	Zufallsvariable	Y
Kleinbuchstaben	Realisierung einer Zufallsvariale, Messwert	x_i, y
fett	Vektor oder Matrix	$\mathbf{X}, \mathbf{y}, \boldsymbol{\epsilon}$
Griechisch	Parameter	β, μ
Hut $\hat{}$	Schätzer	$\widehat{\beta}_0$
Index $_i$	Index für Werte	x_i
Index $_j$	Index für Parameter	β_j
Index $^{(m)}$	Index für Iteration	$b^{(m+1)}$

- Lineares Modell - Matrix Schreibweise
  - Seien Y_i *i.i.d.* Zufallsvariablen mit normalverteilter Streuung ϵ

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i & \epsilon_i &\sim \mathcal{N}(0, \sigma^2) \\ \mathcal{E}(Y_i) = \mu_i &= \beta_0 + \beta_1 X_i & Y_i &\sim \mathcal{N}(\mu_i, \sigma^2) \end{aligned}$$

  - n-malige *unabhängige, identische* Wiederholung des Versuchs
    - Messtupel (X_i, Y_i) mit $i \in [1 \dots n]$
    - Erlaubte Streuung in Y_i
  - Abhängige Variable Y

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

  - Parametervektor β

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

    - bestimmt die Modell-Abhängigkeit $y_i \sim x_i$
  - unabhängige Variable X
    - Vektor $\mathbf{X} = [X_1, X_2, \dots, X_n]^T$
    - erweitert um den y-Achsenabschnitt *intercept*

  - Vektor $\mathbf{1} = [1, 1, \ldots, 1]^T$

- $\Rightarrow$ Designmatrix $\mathbf{X}$

$$\mathbf{X} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}$$

  * unabhängige Variablen in Spalten
  * Indikator- (Pseudo-) Variable für unabhängige Kategorien

- $\Rightarrow$ Lineares Modell

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$\mathcal{E}(\mathbf{Y}) = \mathbf{X} \quad \boldsymbol{\beta}$$

$$\mathcal{E}(Y_i) = 1 \cdot \beta_0 + X_i \cdot \beta_1$$

- ϵ Streuungen in y
  * Messfehler
  * Ungenauigkeiten
  * Residuen: Abweichungen vom Modell

$$\mathbf{y} = \mathbf{X}\,\boldsymbol{\beta} + \epsilon$$

$$\mathcal{E}(\mathbf{y}) = \mathbf{X}\,\boldsymbol{\beta}$$

- Gesucht: Parameter des Modells $\boldsymbol{\beta}$
- Lösung dieser Aufgabe:

  mittels Anpassen der Parameter durch iterative Anwendung von Matrixinversion aus Maximum-Likelihood-Prinzip / Kleinste-Quadrate-Schätzung

- Ergebnis: Parametervektor β
  * Punktschätzer $\widehat{\beta}$ mit Konfidenzintervall
  * *Signifikanz*

1.3 Additives und Interaktives Lineares Modell

- Additives Lineares Modell
  - k unabhängige Variablen X_j als Spalten der Länge n in die **Designmatrix** einfügen

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & & \ddots & & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{nk} \end{bmatrix}$$

  - den **Parametervektor** erweitern zu

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

  - ergibt das additive Lineare Modell

$$\mathcal{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$$

- Interaktives Lineares Modell
  - sind die unabhängigen Variablen X_l und X_m untereinander unabhängig, dann ist

$$x_{io} = x_{il} \cdot x_{im}$$

    eine *neue* unabhängige Variable und kann als Spalte der Designmatrix hinzugefügt werden
  - Interaktion: Beeinflussung von X_l auf X_m
  - Designmatrix mit zusätzlichem Interaktions-Term

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & \dots & x_{1,k} & x_{1,k+1} = x_{1,l} \cdot x_{1,m} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,k} & x_{n,k+1} = x_{n,l} \cdot x_{n,m} \end{bmatrix}$$

  - Schätzung der Parameter analog

$$\widehat{\boldsymbol{\beta}} = [\widehat{\beta}_0,\ \widehat{\beta}_1, \dots \widehat{\beta}_k,\ \widehat{\beta}_{lm}]^T$$

  - Formelbeschreibung in `patsy` beispielweise
    - `'y ~ x1: x2'`: beinhaltet eine Spalte mit Term $x1 * x2$ in Designmatrix
    - `'y ~ x1 * x2 + x3'`: Abkürzung für: Spalte mit Termen 1, x1, x2, x1*x2 und x3

1.4 Kennwerte Linearer Modelle

- Einzelne Messwerte
$$Y_i = 1\beta_0 + X_{i1}\beta_1 + \cdots + X_{ik}\beta_k + \epsilon_i$$
  - mit Zufall/Streuung/Rauschen ("Homoskedastizitätsannahme", (Residuen-) Varianzhomogenität)
$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$
  - Dann
$$\begin{aligned} \mathcal{E}(Y_i) &= \beta_0 + X_{i1}\beta_1 + \cdots + X_{ik}\beta_k \\ Var(Y_i) &= \sigma^2 \end{aligned}$$

- vektoriell
  - Erwartungswert
$$\mathcal{E}(\mathbf{Y}) = \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$$
  - Varianz-Kovarianz-Matrix
$$\mathbf{V}_{jk} = \mathcal{E}\Big((Y_j - \mu_j) \cdot (Y_k - \mu_k)\Big)$$
    - im unabhängigen Fall
$$\begin{aligned} Var(Y_j) &= V_{jj} = \sigma_j^2 \\ Cov(Y_j, Y_k) &= V_{jk} = 0 \quad \text{für } k \neq j \end{aligned}$$
    - im i.i.d.-Fall
$$Var(Y_j) = V_{jj} = \sigma^2$$
    - Definition:
$$Cov(Y_j, Y_k) = \mathcal{E}\Big((Y_j - \mathcal{E}(Y_j)) \cdot (Y_k - \mathcal{E}(Y_k))\Big)$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y) \cdot (x - \mathcal{E}(X)) \cdot (y - \mathcal{E}(Y))\, \mathrm{d}y\, \mathrm{d}x$$
    - daraus folgt im unabhängigen Fall (siehe oben):
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_X(x) \cdot f_Y(y) \cdot (x - \mathcal{E}(X)) \cdot (y - \mathcal{E}(Y))\, \mathrm{d}y\, \mathrm{d}x = 0 \quad q.e.d.$$

1.5 Generalisierte Lineare Modelle

- Lineares Modell

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\,\boldsymbol{\beta} + \epsilon \\ \mathcal{E}(\mathbf{y}) &= \mathbf{X}\,\boldsymbol{\beta} \end{aligned}$$

- Generalisiertes Lineares Modell mit Link-Funktion g

$$\begin{aligned} \mathcal{E}(\mathbf{Y}) &= \boldsymbol{\mu} \\ g(\boldsymbol{\mu}) &= \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} \end{aligned}$$

  - insbesondere hilfreich mit
    - kategorialer *abhängiger* Variable
    - dichotomer *abhängiger* Variable
- Spezialfall
  - Link-Funktion **Identität**

$$\eta = g(\mu) = \mu$$

  - Streuung **Normalverteilung**

$$f(\mathbf{Y}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^2}{2\sigma^2}\right)$$

  - Dann ergibt sich

$$\begin{aligned} \mathcal{E}(\mathbf{Y}) &= \mathbf{X}\boldsymbol{\beta} \\ \mathrm{Var}(\mathbf{Y}) &= \sigma^2 \end{aligned}$$

    ... das (einfache) **Lineare Modell**
- Fragestellungen
  - Das Modell ist festgelegt
    - Theorie
    - Erfahrung
    - Vorversuch
  - Die Modell-Parameter
    - sind unbekannt
    - oder dienen der Überprüfung einer Theorie
    - gilt es, aus Messungen von X_i und Y_i zu bestimmen
  - Schlussfolgerung
    - Ist Y von X abhängig? (Signifikanz)
    - Ist die Abhängigkeit stärker unter Versuchsbedingung A als unter B? (Vergleich)

1.6 Fragestellung

- Ziel: Parameter $\boldsymbol{\beta}$
- Anpassung (fit) des Linearen Modells, so dass die Residuen minimiert werden.
- Erinnerung: Homoskedastizitätsannahme der Normalverteilten Residuen.
  - Summe der Abweichungsbeträge L_1

$$S_1(\mathbf{y}, \widehat{\mathbf{y}}) = \sum_{i=1}^{n} |y_i - \widehat{y}_i|$$

  - Element der maximalen Aweichung L_∞

$$S_\infty(\mathbf{y}, \widehat{\mathbf{y}}) = max_i(|y_i - \widehat{y}_i|)$$

  - Euklidische Abstandsquadratsumme L_2

$$S_2(\mathbf{y}, \widehat{\mathbf{y}}) = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

  - Euklidische Norm: $||\mathbf{z}|| = \sqrt{S_2(\mathbf{z}, \mathbf{0})} = \sqrt{\mathbf{z}^T\mathbf{z}} = \sqrt{\sum_{i=1}^{n} z_i^2}$
  - Quadratfehlersumme

$$\text{RSS} = S_2(\mathbf{y}, \widehat{\mathbf{y}}) = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

    * Wird verwendet, wenn Gauß'sche Fehler vorhanden sind
- Gauß-Markov-Theorem
  - L_2 liefert die kleinste Varianz zu einem erwartungstreuen (*unbiased*) linearen Schätzer
  - Voraussetzung:
    * unabhängige Parameter
    * Fehler *i.i.d.* (*independently identically distributed*)
  - Nicht zwingend hier:
    * Normalverteilung

1.7 Lösung der Aufgabe

Lösung 1: Kleinste Quadrate Schätzer

- Für das Lineare Modell
$$\widehat{\mathbf{y}} = \mathcal{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$$
- Speziell: Ausgleichsgerade
$$\widehat{\mathbf{y}} = \mathcal{E}(\mathbf{Y}) = \beta_0 + \beta_1\mathbf{x}$$
  - Ansatz
$$S_2(\mathbf{y}, \widehat{\mathbf{y}}) = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i))^2 \stackrel{!}{=} min_{\beta_0,\beta_1}$$
  - führt dank einfacher Rechnung zu
$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$
$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\overline{x}$$
  - Residuenvarianz (bereits zwei Werte geschätzt, reduziert Freiheitsgrade)
$$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\widehat{\epsilon}_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2$$

Lösung 2: Matrix-Ansatz

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- Minimieren der Fehlerquadratsumme
$$S_2(\mathbf{y}, \widehat{\mathbf{y}}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \stackrel{!}{=} min_{\boldsymbol{\beta}}$$
  - führt zu
$$\mathbf{X}^T\,\mathbf{y} = \mathbf{X}^T\,\mathbf{X}\,\boldsymbol{\beta}$$
  - mit Lösung
$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\,\mathbf{X})^{-1}\mathbf{X}^T\,\mathbf{y}$$
- Mit Gewichtung
  - Minimieren der Fehlerquadratsumme mit reziprok gewichteten Varianzen
$$S_2(\mathbf{y}, \widehat{\mathbf{y}}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \stackrel{!}{=} min_{\boldsymbol{\beta}}$$
  - (Varianz-Kovarianz-Matrix $\mathbf{V}$; $\mathbf{V}_{jk} = Cov(Y_j, Y_k)$) führt zu
$$\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} = \mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta}$$

  - mit Lösung
$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$$

- Gilt für beliebige Dimensionen
  - hier mit 2x2 Matrix einfach
- Höherdimensional möglich, nur technisch schwer.
  - Dann iterativ zu bestimmen
- Numerisch instabil mit Kovarianzen
- Unlösbar oder stark fehlerbehaftet durch Gleitkommafehler
  - wenn unterbestimmt durch unglückliche Verteilung der Fehler
  - zu wenig Freiheitsgrade
- Implementiert in Python `statsmodels.ols`:
  - `pinv`: Moore-Penrose pseudoinverse
  - `qr`: Q-R-Zerlegung

Lösung 3: Maximum Likelihood Schätzer

- Ansatz über Verbund-Wahrscheinlichkeitsverteilung $f_{\boldsymbol{\theta}}(\mathbf{y})$ = Likelihood $L_{\mathbf{y}}(\boldsymbol{\theta})$
$$L(\boldsymbol{\theta}|\mathbf{y}) = f(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{N} f(y_i|\boldsymbol{\theta})$$
  - Daraus Log-Likelihood
$$l(\boldsymbol{\theta}|\mathbf{y}) := \log L(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^{N} \log f(y_i|\boldsymbol{\theta})$$
  - zu maximieren
$$l(\widehat{\boldsymbol{\theta}}) \stackrel{!}{=} max_{\boldsymbol{\theta}}$$

- Beispiel Normalverteilung
  - Lineares Modell $\quad \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad \boldsymbol{\mu} = \mathcal{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$
  - Normalverteilung $\quad f(y_i|\,\mu_i, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i-\mu_i)^2}{2\sigma^2}\right)$
  - Parametervektor $\quad \boldsymbol{\theta} = [\beta_0, \beta_1, \sigma]^T$
  - Log-Likelihood:
$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{i=1}^{N} \log f(y_i|\boldsymbol{\theta}) \\ &= \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2}\right) \\ &= -\frac{n}{2}\log 2\pi - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\Big(y_i - (\beta_0 + \beta_1 x_i)\Big)^2 \end{aligned}$$

  - Maximieren der Log-Likelihood führt zum Parametervektor-Schätzer $\widehat{\boldsymbol{\theta}} = [\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\sigma}]^T$

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$
$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$
$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$$

Vergleich der Lösungen

- Kleinste-Quadrate-Methode
  - Minimieren S_2 der Residuen
  - Findet *Kleinste-Quadrate-Schätzer (least square, LSE)* für Parameter
- Max-Likelihood-Methode
  - Maximiert Log-Likelihood
  - Findet *Max-Likelihood-Schätzer (MLE)* für Parameter
- Meist das selbe Ergenis
  - Bei Normalverteilung identisch

Anwendungsbeispiel: log(Gehirnmasse) $\sim$ log(Körpermasse)

- Designmatrix
  - Zeilen:
    - Daten der einzelnen Tiere (i)
  - Spalten:
    - unabhängige Variable 'Körpergewicht'
    - Konstante für den y-Achsenabschnitt (*intercept*) β_0
- Designmatrix mit `numpy`: `np.vstack((np.ones(len(x1)), x1)).T`
- Berechne den Punktschätzer des Parametervektors aus Designmatrix und Datenvektor

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\,\mathbf{X})^{-1}\mathbf{X}^T\,\mathbf{y}$$

1.8 Praktische Lösung mittels Python `statsmodels`

- Homepage: `http://www.statsmodels.org/stable/`
- Beschreibung
  - `GLS` = Generalized least squares regression
  - `OLS` = Ordinary least square regression
  - `GLM` = Generalized linear models
    - fit = `smf.glm(formula='log_BrainWt ~ log_BodyWt', data=animalsdata).fit()`
    - Ergebnis/Ausgabe:
      - Parametervektorschätzer
      - Standardabweichung
      - z-Wert der Gauß-Statistik
      - p-Wert dazu
      - 95%-Konfidenzintervall
- Daten interpolieren, extrapolieren
  - Modell an die Daten anpassen (fit) ergibt den Parameter-Schätzer

$$\widehat{\boldsymbol{\beta}}$$

  - Der vorhergesagte Wert $\widehat{\mathbf{y}}$ ist

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$$

$$\widehat{y}_i = (\mathbf{X}\widehat{\boldsymbol{\beta}})_i = \sum_{j=0}^{m} x_{ij}\beta_j = 1\beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + ... + x_{im}\beta_m$$

1.8.1 Python `statsmodels`

- `statsmodels.formula.glm.fit()` beschreibt ein **lineares Datenmodell**
  - Eingabe Datensatz `data =`
    - `pandas.DataFrame` mit *Variablennamen*
    - unabhängige Variablen bzw. Designmatrix
    - abhängigen Variablen
  - Eingabe Modell `formula=`
    - `patsy`-Formel mit abhängiger Variable ~ unabhängiger Variablen
    - `'y ~ x1 + x2 + x3'`
    - berücksichtigt bereits die Konstantenspalte der Designmatrix *intercept*

        - ⇒ explizit ausschließen `'~ -1'`

- `statsmodels.GLM.fit()`
  - Eingabe Daten
    - `exog`: unabhängige Variablen in Spalten der Designmatrix X
      - zusätzlich Konstante *intercept* anfügen `sm.add_constant(X)`
      - Bei *Interaktion* sind zusätzliche Spalten zu berechnen
    - `endog`: abhängige Variable, gemessene Daten y
- `statsmodels.___.fit()`
  - Ausgabe Parametervektor
    - Punktschätzer
      - Standardabweichung
      - Vertrauensintervall
      - Z-Wert der Gauß-Statistik
      - p-Wert
  - Ausgabe Statistiken und Kennzahlen
    - ...
  - Ausgabe Fit-Werte
    - `fittedvalues`: (als pandas-Daten-`Series`)
    - `resid_response`: verbleibende Fehler `(Series)`
    - `predict`(x): Zwischenwerte vorhersagen/extrapolieren
      - x als `DataFrame` mit passend benannten Spalten

1.8.2 Python `Pandas`

- Python `Pandas` für Umgang mit Daten
  - Homepage: `http://pandas.pydata.org/pandas-docs/stable/overview.html`
  - Daten aus Datei einlesen `read_csv()`
  - Variable vom Typ `DataFrame`
    - Auswahl der in Spalten enthaltenen Variablen durch Namensstring
    - Auswahl nach Kriterien, Index, Eigenschaften, ...
    - Umfangreiche Methoden
      - sortieren `sort()`
  - Beispiel: Abhängigkeit von Körpergewicht und Hirngewicht

    - Lösung? Zufällige Abweichungen zwischen Messung y_i und Modell-Vorhersage $\widehat{y}_i$
    - Residuen
    $$r_i = y_i - \widehat{y}_i$$

1.8.3 Python `Patsy`

- Designmatrix mit `patsy`
  - Homepage: `http://patsy.readthedocs.io/en/latest/overview.html`
  - Patsy erlaubt Formulierung
    - des Modells
    - der zu benutzenden Daten
  - Eingabe:
    - `y, X = patsy.dmatrices('yvar ~ xvar1 + xvar2', df)`
    - verwendet *pandas* `DataFrame df`
  - Ausgabe
    - Designmatrix x als `patsy.design_info.DesignMatrix,` $N * K$ Array, mit y-Achsenabschnittskonstante
    - Gemessene Daten y als `patsy.design_info.DesignMatrix,` $N * 1$ Array
  - Generelle Form: Innerhalb eines Strings $y \sim x$
    - links der Tilde die abhängige Variable
    - rechts die unabhängige Variablen
  - Anschaulich lassen sich die Namen der Datenfelder aus dem `DataFrame` benutzen

1.9 Ergebnis lineare Modelle in Python

- Daten lassen sich in `DataFrames` komfortabel bearbeiten
- lassen sich durch `Patsy`-Formel beschreiben
- Schätzer für Parameter lassen sich durch `statsmodels.glm` berechnen
- Rückgabewerte:
  - Kennzahlen
  - Statistik
  - Punktschätzer für Parameter (Steigung und Achsenabschnitt) und deren
  - Intervallschätzer
  - ...

1.10 Bestes Modell?

- Ein perfekt passendes Modell muss nicht das beste sein
- Gleiche Versuchsbedingung, identische Zeile in Designmatrix:
  Streuung in $\mu_{i_1} = \mu_{i_2} = \ldots$
- $\Rightarrow$ Fehler zulassen

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$$

- Theorie
- Ockham's razor

Verdichtung der Information

- Nicht von Interesse: alle einzelnen μ_i der abhängigen Variablen Y
- Von Interesse:
  - Einfluss der unabhängigen Variablen (*erklärende* Variablen, Pediktoren) X
    - kategorial
    - kontinuierlich
    - Versuchsbedingungen $i \quad i \in [1 \ldots n]$
  - zugehörige Parameter
    - modellieren X, *Gewichtung* der Einflüsse
    - Parameter $\beta_j \quad j \in [1 \ldots k] \quad k \ll n$

= das Modell

Ergebnis

- Modell = Entscheidung für Vereinfachung
- Es verbleiben Residuen

Residuen

- Verteilung der Residuen

$$\begin{aligned} Y_i &= \mathbf{X}_i^T \boldsymbol{\beta} + \epsilon_i & \epsilon_i &\sim \mathcal{N}(0, \sigma^2) \\ \mathcal{E}(Y_i) = \mu_i &= \mathbf{X}_i^T \boldsymbol{\beta} & Y_i &\sim \mathcal{N}(\mu_i, \sigma^2) \end{aligned}$$

- Anforderung an Residuen
  - Modell soll gut abbilden, 'in der Mitte' $\Rightarrow$ $\mathcal{E}(R) = 0$
  - Streuung in Verteilung hat dieselben Ursachen
    - *Lineares Modell*, *Gauß*- Verteilung: $\Rightarrow$ $Var(R) = const.$
    - Gemäß Verteilung
  - Gutes Modell erklärt Messdaten
    - Keine (wenig) Information in den Residuen:
      $\Rightarrow$ **unabhängig**, homoskedastisch
- Homoskedastizität und Unabhängigkeit
  - Systematische Abweichungen? $\Rightarrow$ Auf den Grund gehen!

1.11 Modell-Vergleich

- Quadratfehlersumme, *sum of squared residua, RSS*

$$RSS = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - (\mathbf{X}\widehat{\boldsymbol{\beta}})_i)^2$$

  - Ist eine charakteristische Kennzahl
    - Für Gauß-Verteilungen: standardisierte Quadratfehlersumme $\tilde{S} = \frac{RSS}{\sigma^2}$
    - $\tilde{S} \sim \chi^2(n-p)$
  - Abhängigkeit nur von
    - n Werten der *abhängigen* Variablen
    - n Werten der *unabhängigen* Variablen
    - p geschätzte Parameterwerte
  - je kleiner RSS, desto näher liegt das Modell an den Daten
- Schätzer für $\boldsymbol{\beta}$
  - $\widehat{\boldsymbol{\beta}}$ aus Max-Likelihood oder Kleinste-Quadtrate (k Komponenten)
- Schätzer für $\boldsymbol{\mu}$
  - $\widehat{\mu}_i = \mathbf{X}_i^T \widehat{\boldsymbol{\beta}}$ aus dem linearen Modell
- Schätzer für *Störparameter* σ^2
  - Seien y_i Normalverteilt (mindestens näherungsweise; Zentraler Grenzwertsatz) dann ist mit

$$RSS = \sum_{i=1}^{N} r_i^2 = \sum_{i=1}^{N} (y_i - (\mathbf{X}\widehat{\boldsymbol{\beta}})_i)^2$$

$$\widehat{\sigma}^2 = \frac{1}{N-p} RSS$$

  - ein erwartungstreuer Schätzer der Varianz σ^2 für das Lineare Modell

$$\mathcal{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \qquad Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$$

$$\widehat{\sigma}^2 = \frac{1}{N-p} \sum_{i=1}^{N} r_i^2 = \frac{1}{N-p} \sum_{i=1}^{N} (y_i - (\mathbf{X}\widehat{\boldsymbol{\beta}})_i)^2$$

- Verteilung der standardisierten Fehlerquadratsumme

$$\frac{RSS}{\sigma^2} \sim \chi^2(N-p)$$

  - Die Verteilung der Zufallsvariable *Schätzer der Residuen-Varianz* $\widehat{\sigma}^2$ ist dann skaliert:

$$\widehat{\sigma}^2 \sim \chi^2(\texttt{df} = N-p,\ \texttt{scale} = \frac{\sigma^2}{N})$$

  - ... unter der Nullhypothese, dass das Modell korrekt ist!
- Problem 1: Woher kennen wir das wahre σ^2?
- Problem 2: Was ergibt die Berechnung mit dem Schätzer?
- Vergleich der beiden Modelle
  - Voraussetzung: Modelle bauen aufeinander auf, Modell B ist eine Erweiterung/Verallgemeinerung des einfacheren Modells A
  - Ist Modell B (hier $p_B = 3$ Parameter) angemessen?
    * Nein $\Rightarrow$ beide Modelle verwerfen
    * Ja $\Rightarrow$ vergleiche mit Modell A
  - Ist Modell A (hier $p_A = 2$ Parameter) angemessen?
    * Nein $\Rightarrow$ wähle Modell B
    * Ja $\Rightarrow$ Vergleich mit Modell B ergibt ...

Wiederholung Tests

1. Formulierung des Problems
2. Modellannahme
   - Welcher Art sind die Daten
   - Welche Verteilung wird erwartet
3. Aufstellen der Nullhypothese und der Alternativhypothese
   - Ziel soll es sein, die Nullhypothese ablehnen zu können
   - einseitiger Test
   - zweiseitiger Test
4. Festlegen des Signifikanzniveaus
   - zulässige Irrtumswahrscheinlichkeit α
5. Teststatistik / Prüfgröße aussuchen
   - verdichtet Information aus der Stichprobe
   - Verteilung unter H_A sollte sich deutlich von der unter H_0 unterscheiden
6. Verteilungsfunktion F bestimmen
   - theoretisch bestimmbar
   - asymptotisch bestimmbar
   - Simulation
7. Verwerfungsbereich
   - Statistik: Verteilung der Prüfgröße

- Hypothese: Richtung einseitig/zweiseitig
- Signifikanzniveau: Irrtumswahrscheinlichkeit α

a) Verwerfungsbereich bestimmen
   - Wert für t der Teststatistik T aus Daten bestimmen
   - Tabelle oder berechnen oder

b) p-Wert bestimmen
   - Tabelle oder berechnen

8. Entscheidung fällen
   - t im Verwerfungsbereich: Verwerfen der Nullhypothese
   - p außerhalb α: Verwerfen der Nullhypothese
   - sonst: H_0 nicht verwerfbar

Gauß-Test / t-Test

- Neue Differenz in Kategorien = Zusätzlicher Parameter
  - Modellannahme
  - Nullhypothese: Parameter `IsMonkey` ist nicht nötig, Einfluss $\beta_1 = 0$
  - Alternativhypotehse: Parameter `IsMonkey` ist relevant, Einfluss $\beta_1 \neq 0$
  - Teststatistik standardisierte Differenz - *Gauß-Test* für $\beta_{IsMonkey}$

$$Z = \frac{\overline{X_a} - \overline{X_b}}{\sqrt{S_a^2/n_a + S_b^2/n_b}} \sim \mathcal{N}(0,1) = \varphi$$

  - Verwerfungsbereich festlegen und bestimmen
    - Zur Irrtumswahrscheinlichkeit $\alpha = 0.1\%$
  - Wert der Statistik berechnen, p-Wert
  - Ergebnis und Entscheidung
- Problem: kumulierter α-Fehler

F-Tests

- F-Test: Vergleich des Varianzenverhältnisses

$$F = \frac{SQE/(n_c - 1)}{SQR/(n - n_c)} \sim \mathcal{F}(n_c - 1, n - n_c)$$

- Siehe Varianzanalyse (ANOVA)

Vergleich der Likelihood

- Verhältnis der Likelihood $= \frac{L_A}{L_B}$
- Differenz der Log-Likelihood $log(L_A) - log(L_B) = l_A - l_B$
- Maximal mögliche Likelihood?
  - *Vollständiges* Modell $\widehat{y}_i \equiv y_i$ mit Likelihood L_V
- Deviance
  - (Doppelter) Unterschied zur Log-Likelihood des vollständigen Modells

$$D := 2(l_V - l_A)$$

1.12 Deviance

Verallgemeinert die Quadratfehlersumme von Normalverteilten Modellen.

- Anwendung: Modellvergleich
  - Voraussetzung: Modelle bauen aufeinander auf (*nested models*)
- Definition

$$D(\widehat{\boldsymbol{\theta}}; \mathbf{y}) := 2\Big(l(\tilde{\boldsymbol{\theta}}; \mathbf{y}) - l(\widehat{\boldsymbol{\theta}}; \mathbf{y})\Big)$$

  - $\mathbf{y}$ Werte der abhängigen Variable
  - $\widehat{\boldsymbol{\theta}}$ Schätzer der Parameter
  - $\tilde{\boldsymbol{\theta}}$ Schätzer der Parameter eines *vollständigen* Modells $\widehat{y}_i \equiv y_i$
- Beispiel Lineares Modell mit Normalverteilung(en)

$$l(\boldsymbol{\mu}; \mathbf{y}) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu_i)^2 - n \log\left(\sigma\sqrt{2\pi}\right)$$

$$\begin{aligned} D &= 2\Big(l(\tilde{\boldsymbol{\mu}}; \mathbf{y}) - l(\widehat{\boldsymbol{\mu}}; \mathbf{y})\Big) \\ &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \widehat{\mu}_i)^2 \end{aligned}$$

  - entspricht damit Pearsons standardisierter Quadratfehlersumme, also

$$D \sim \chi^2(n - k)$$

  - Begründung: Abhängigkeiten der $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, es verbleiben k Komponenten, Freiheitsgrade in $\boldsymbol{\beta}$
  - Verteilung $\sim \chi^2(k)$ mit Anzahl der *zusätzlichen* Parameter k zum erweiterten Modell
- auch für andere Verteilungen
  - näherungsweise χ^2-verteilt

Scaled Deviance

Streuung σ ist unbekannt

- Die angegebene *scaled Deviance* ist aus den Daten berechenbar

$$D' = \sigma^2 D = \sum_{i=1}^{n} (y_i - \mu_i)^2$$

Unterscheidung

Unterscheiden sich die beiden Modelle?

- Unterschied in Deviance ΔD:

$$\Delta D(\widehat{\boldsymbol{\theta}}_A, \widehat{\boldsymbol{\theta}}_B; \mathbf{y}) = D(\widehat{\boldsymbol{\theta}}_A; \mathbf{y}) - D(\widehat{\boldsymbol{\theta}}_B; \mathbf{y}) = 2l(\widehat{\boldsymbol{\theta}}_B; \mathbf{y}) - 2l(\widehat{\boldsymbol{\theta}}_A; \mathbf{y}) > 0$$

  - $\mathbf{y}$ Werte der abhängigen Variable
  - $\widehat{\boldsymbol{\theta}}_A$ Schätzer der Parameter (k_A Stk.) des einfachen Modells
  - $\widehat{\boldsymbol{\theta}}_B$ Schätzer der Parameter (k_B Stk.) des erweiterten Modells
  - $\Delta D \geq 0$
  - Verteilung

$$\Delta D \sim \chi^2(k_B - k_A)$$

- Fisher $\mathcal{F}$-Test für Deviance
  - Betrachte das Verhältnis

$$F = \frac{D_0 - D_1}{k - q} \Big/ \frac{D_1}{n - k} \sim \mathcal{F}(k - q, n - k)$$

  - Unterschied?
    * Nullhypothese: Modell A (alle Säugetiere) ist ebenso gut wie das bessere Modell B(Affen getrennt)
    * Alternativhypothese: Modell B beschreibt den linearen Zusammenhang besser

Ergebnis

- Im Beispiel ist der Unterschied höchst signifikant ($\alpha = 0.1\%$)
  - t-Test/Gauß-Test für Parameter β_{IsMonkey}
  - Varianzanalyse für Residuen zwischen beiden Modellen
  - F-Test der Deviance zwischen beiden Modellen
- Unterschied in Deviance
  - in guter Näherung χ^2-verteilt
- Die Deviance ist eine sinnvolle Erweiterung der Pearson Quadratfehlersumme
- Konzept der Deviance gilt auch für andere Verteilungen der Exponentialfamilie

2 Generalisierte Lineare Modelle - GLM

2.1 Motivation Generalisiertes Lineares Modell

- Problemstellung
  - Jet-Piloten erfahren unter besonderes hohen Beschleunigungskräften (bezogen auf die Erdbeschleunigung g) Blackouts
- Versuch
  - Glaister und Miller (1990) erzeugten ähnliche Symptome, indem sie den Körper der Versuchspersonen einem Luftunterdruck aussetzten
- Fragestellung
  - Hängt die Ohnmacht vom Alter ab?
- Ansatz
  - Linearer fit `'symptoms ~ age'`
  - Problem: Linearer fit nicht aussagekräftig hier
- Lösung: Logit-Link
  - Wahrscheinlichkeit des Bernoulli-Ereignisses $\pi \in [0...1]$
  - Linearer Term $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$
  - Link-Funktion **logit**

$$\mathcal{E}(\mathbf{Y}) = \boldsymbol{\pi} \qquad g(\boldsymbol{\pi}) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$$
$$\mathcal{E}(\mathbf{Y}) = \boldsymbol{\pi} = g^{-1}(\mathbf{X}\boldsymbol{\beta})$$

    * logit-Funktion

$$g^{-1}(\eta) = \text{logit}(\eta) = \frac{1}{1 + e^{-\eta}}$$

    * Umkehrfunktion: logarithmisches Chancenverhältnis *log-odds-ratio*

$$\eta = g(\pi) = \ln \frac{\pi}{1 - \pi}$$

- Bernoulliverteilung
  - Wahrscheinlichkeitsverteilung des Ereignisses $y \in [0, 1]$

$$f(y|\pi) = \pi^y (1 - \pi)^{1-y}$$

$$\mathcal{E}(y) = \pi$$

- Binomialverteilung
  - Wahrscheinlichkeitsverteilung der y = *Anzahl der Erfolge* mehrerer Bernoulli-Ereignisse

$$P(y|N,\pi) = \binom{N}{y}\pi^y(1-\pi)^{(N-y)} \qquad y \in \{0\dots n\}$$

$$\mathcal{E}(y) = N\pi$$

- Ergebnis Link-Funktion: Eine Link Funktion $g(\mu)$
  - kann Anforderungen an Randbedingungen von Zufallsvariablen erfüllen
    - ∞-Problem ✓
    - Verteilung der Streuung berücksichtigen ✓
  - erweitert das Lineare Modell
    - verbindet lineare Vorhersage $\eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$
    - und zentralen Parameter der Wahrscheinlichkeitsverteilung μ_i
- Ergebnis 'Generalisiertes Lineares Modell'

$$g(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta}$$

$$\mathcal{E}(Y_i) = \mu_i = g^{-1}(\mathbf{x}_i^T\boldsymbol{\beta})$$
$$Y_i \sim f(\mu_i, \sigma^2, \dots)$$

2.1.1 Kategoriale Variable und Residuen

- Beispieldaten: Allison, Cicchetti (1976) *Sleep in mammals: ecological and constitutional correlates.* Science **194:** 732-734
  - Lineares Modell des Gehirn-Gewichts gegen das Körpergewicht
  - Interessant: Abweichungen vom Modell
    - systematisch?
    - Zufall (wie im Modell vorgesehen)?
- Ergebnis Residuen-Analyse
  - **Systematische** Abweichungen
    - Ausreißer, Auffälligkeit
    - Affen haben positive Residuen: eher *kein Zufall*
  - **Zufällige** Abweichungen
    - Verteilung gemäß Modell: Streuung
- Erweitertes Modell
  - Affen als eigene Kategorie

    * Kategoriale Variable `['IsMonkey']`
    * Anpassen der Designmatrix
    * Indikatorvariable c für Kategorie Affe `['IsMonkey']='no'` $= 0$ und `['IsMonkey']='yes'` $= 1 \Rightarrow \beta_1$

$$
\begin{aligned}
\mathcal{E}(\mathbf{Y}) &= \mathbf{X} \quad \boldsymbol{\beta} \\
\mathcal{E}(Y_i) &= 1 \cdot \beta_0 + c_i \cdot \beta_1 + X_i \cdot \beta_2 \\
\begin{bmatrix} Y_1 \\ \vdots \\ Y_a \\ Y_{a+1} \\ \vdots \\ Y_n \end{bmatrix} &= \begin{bmatrix} 1 & 0 & X_1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & X_a \\ 1 & 1 & X_{a+1} \\ \vdots & \vdots & \vdots \\ 1 & 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}
\end{aligned}
$$

- Ergebnis Kategoriale Variable
  - wirkt als Schalter
    * Wert $X_{ij} \in [0, 1]$
    * für Parameter β_j
  - Kategorien werden von `Patsy` automatisch erkannt (z.B. wenn *String*)
    * erzwingen mit `'C(variable)'`
  - fügt sich formal in Lineares Modell ein
  - erweiterbar auf mehrere Ausprägungen
    * mehrere Spalten
    * *nicht Zahlen!*

2.1.2 Modellvergleich

- Residuen der beiden Modelle
  - Modell A: $r_{Ai} = y_i - \widehat{\mu}_{Ai} = y_i - (\mathbf{X}_A \widehat{\boldsymbol{\beta}}_A)_i$
  - Modell B: $r_{Bi} = y_i - \widehat{\mu}_{Bi} = y_i - (\mathbf{X}_B \widehat{\boldsymbol{\beta}}_B)_i$
- Residuen gehören zu einem Modell
- Minimieren
  - Kleinste-Quadrate
  - Matrix Zerlegung
  - Maximum-Log-Likelihood
- Überprüfen, ob Modellvoraussetzungen erfüllt sind

  - Scatter-Plot
  - Histogramm

2.1.3 Verdichtung der Information

- Nicht von Interesse: alle einzelnen μ_i
- Von Interesse:
  - Einfluss der unabhängigen Variablen (*erklärende* Variablen, Pediktoren) X
    - kategorial
    - kontinuierlich
    - Versuchsbedingungen $i \quad i \in [1 \dots n]$
  - zugehörige Parameter
    - modellieren X, *Gewichtung* der Einflüsse
    - Parameter $\beta_j \quad j \in [1 \dots k] \quad k \ll n$

2.2 *Generalisierte* Lineare Modelle

Link-Funktion g

verbindet additiven Einfluss (η_i) der unabhängigen Variablen $\mathbf{x}_i$ auf die (erwünschte) Verteilung der abhängigen Y_i um (μ_i)

$$g(\mu_i) = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$$

Beispiel Bernoulli-Verteilung

- Exponentiell abfallende Abhängigkeit

$$P(Y_i{=}1) = e^{-\lambda t} = \pi$$
$$P(Y_i{=}0) = 1 - e^{-\lambda t} = 1 - \pi$$

- führt unter Verwendung der Link-Funktion

$$g(\pi) = \log(\pi) = -\lambda t$$

- auf eine lineare Abhängigkeit

$$g(E(Y)) = -\lambda t$$

- mit

$$\mathbf{x}_i = [t] \quad \boldsymbol{\beta} = [-\lambda]$$

- zum Generalisierten Linearen Modell

$$E(Y) = g^{-1}(x\beta)$$

Anwendung

- Biologie: Genetischer Stammbaum
- Linguistik: Abspaltung von Sprachen zum Zeitpunkt t mit gemeinsamem Wortschatz (=1) in unterschiedliche Entwicklung von Worten (=0)
- Physik: Spannung bei Kondensatorentladung über konstanten Widerstand

Modell und Fragestellung

- **Gesucht** sind die Parameter des Modells β
  - Verdichtung der Information
  - Signifikanz einer Teil-Abhängigkeit, Parameter β_j
  - Unterschiedliche Abhängigkeit bei anderen Daten
  - Unterschiedliche Modelle

2.3 Exponentialfamilie

Exponentialfamilie für Wahrscheinlichkeitsdichteverteilungen

$$f(y;\theta) = \exp\Big(a(y)b(\theta) + c(\theta) + d(y)\Big)$$

Einige wichtige bekannte Verteilungen sind Mitglied der Exponentialfamilie

- Normalverteilung
  - Parameter θ ist μ
- Binomialverteilung
  - Der einzige interessierende Parameter bei gegebenem n ist π
  - $y \in \{0 \dots n\}$
- Poissonverteilung
  - Der einzige interessierende Parameter ist λ.
  - $y \in \mathbb{N}$

Sie haben

- Gemeinsame Eigenschaften
- Gemeinsame Methoden
- und lassen sich mittels GLM-Formalismus lösen

Implementiert in `statsmodels glm`

- `Binomial ()`
- `Gamma ()`
- `Gaussian ()`
- `InverseGaussian ()`
- `NegativeBinomial ()`
- `Poisson ()`

2.3.1 Allgemeine Eigenschaften der Exponentialfamilie

- Erwartungswert
$$\mathcal{E}\left(a(Y)\right) = -\frac{c'(\theta)}{b'(\theta)}$$
- Varianz
$$Var\left(a(Y)\right) = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3}$$

2.3.2 Log-Likelihood-Funktion

- Exponentialfamilie

$$l(\theta; y) = \log(f_Y) = a(y) \cdot b(\theta) + c(\theta) + d(y)$$

Score Statistik U

- Ableiten der Log-Likelihood-Funktion nach θ ergibt die *score statistic* U, als Funktion von Y eine Zufallsvariable

$$U(\theta; y) := \frac{\mathrm{d}l(\theta; y)}{\mathrm{d}\theta} = a(y) \cdot b'(\theta) + c'(\theta)$$

- mit Erwartungswert

$$\mathcal{E}(U) = 0$$

Information $\mathcal{I}$

- Varianz von U oder *Information* $\mathcal{I}$

$$\mathcal{I} := \mathrm{Var}(U) = (b'(\theta))^2 \cdot \mathrm{Var}\left(a(y)\right) = \frac{b''(\theta)c'(\theta)}{b'(\theta)} - c''(\theta)$$

- Aus dem **Verschiebungssatz** folgt mit $\mathcal{E}(U) = 0$

$$\mathrm{Var}(U) = \mathcal{E}(U^2)$$

- Des Weiteren gilt

$$\mathcal{E}(U') = -\mathrm{Var}(U)$$

- $\Rightarrow$ Information

$$\mathcal{I} := \mathrm{Var}(U) = -\mathcal{E}(U')$$

2.3.3 Kanonische Verteilung

Verteilungen mit

$$a(Y) = Y$$

nennt man **kanonisch**

- Normalverteilung, Poissonverteilung, Binomialverteilung sind kanonisch
- Erwartungswert und Varianz für y haben eine einfache Form
- Der Parameter im zugehörigen Term $b(\theta)$ heißt **natürlicher Parameter**

Verteilung	natürlicher Parameter $b(\theta)$	Funktion $c(\theta)$	Funktion $d(y)$
Normal	$\frac{\mu}{\sigma^2}$	$-\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)$	$-\frac{y^2}{2\sigma^2}$
Binomial	$\ln(\frac{\pi}{1-\pi})$	$n\ln(1-\pi)$	$\ln\binom{n}{y}$
Poisson	$\ln\lambda$	$-\lambda$	$-\ln y!$

Natürlicher Parameter

$$f(Y;\theta) = \exp\Big(Y \cdot b(\theta) + c(\theta) + d(Y)\Big)$$

- Wählt man $b(\theta) = \theta$, dann heißt θ selbst der natürliche Parameter der Verteilung

$$f(Y;\theta) = \exp\Big(Y\theta + c(\theta) + d(Y)\Big)$$

- Möchte man diesen natürlichen Parameter selbst linear vorhersagen

$$\theta = \mathbf{X}\boldsymbol{\beta}$$

- so wird aus der allgemeinen Link-Funktion g:

$$g(\mu) = \mathbf{X}\boldsymbol{\beta}$$

- die *natürliche Link-Funktion*

$$\theta = g(\mu)$$

Verteilung	natürlicher Param. $\theta = b(\theta)$	Erwartungswert	oder $\mu = g^{-1}(\theta)$
Normal	$\theta = \frac{\mu}{\sigma^2}$	$\mu = \mu$	$\mu = \sigma^2\theta$
Binomial	$\theta = \ln(\frac{\pi}{1-\pi})$	$\mu = n\pi$	$\pi = \frac{e^\theta}{1+e^\theta}$
Poisson	$\theta = \ln\lambda$	$\mu = \lambda$	$\lambda = e^\theta$

Vereinfachungen

- Für kanonische Verteilung $a(Y) = Y$ und natürlichen Parameter $b(\theta) = \theta$ ergibt sich

$$f(Y;\theta) = \exp\Big(Y\theta + c(\theta) + d(Y)\Big)$$

- Erwartungswert

$$\mathcal{E}\left(a(Y)\right) = -\frac{c'(\theta)}{b'(\theta)}$$
$$\mathcal{E}\left(Y\right) = -c'(\theta)$$

- Varianz

$$\mathrm{Var}\left(a(Y)\right) = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3}$$
$$\mathrm{Var}\left(Y\right) = -c''(\theta)$$

Verteilung	natürlicher Param. $b(\theta)$	c	c'	c"
Normal	$\theta = \frac{\mu}{\sigma^2}$	$-\frac{\sigma^2\theta^2}{2} - \frac{1}{2}\ln(2\pi\sigma^2)$	$-\sigma^2\theta$	$-\sigma^2$
Binomial	$\theta = \ln(\frac{\pi}{1-\pi})$	$-n\ln(1+e^\theta)$	$-\frac{e^\theta}{1+e^\theta}$	$-n\frac{e^\theta}{(1+e^\theta)^2}$
Poisson	$\theta = \ln\lambda$	$-e^\theta$	$-e^\theta$	$-e^\theta$

Natürlicher Parameter und kanonischer Link

- ... ist in GLM immer für die passende Verteilung implementiert

$$\mathcal{E}(Y) = -c'(\theta)$$

- **Normal-, Poisson-** und **Binomialverteilung** haben passende Parameter
- Andere Link-Funktionen sind ebenso gut möglich

2.3.4 Zusammengesetzte Wahrscheinlichkeitsverteilung - Skalarer Parameter θ

- Ein Satz *unabhängiger, identisch verteilter* (i.i.d.) Zufallsvariabler $\mathbf{Y} = [Y_1 \dots Y_N]^T$
- mit Wahrscheinlichkeitsverteilung $f(y_i, \theta)$ aus der kanonischen Exponentialfamilie
- hat eine gemeinsame Wahrscheinlichkeitsverteilung

$$\begin{aligned} f(\mathbf{Y}, \theta) &= \prod_{i=0}^{n} \exp\Big(y_i b(\theta) + c(\theta) + d(y_i)\Big) \\ &= \exp\Big(\sum_{i=0}^{n} y_i b(\theta) + \sum_{i=0}^{n} c(\theta) + \sum_{i=0}^{n} d(y_i)\Big) \end{aligned}$$

- mit

$$\mathcal{E}(Y_i) = (\dots) = \mu$$

- wobei

$$g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$$

- als auch

$$\theta_i = fkt(\mathbf{x}_i^T \boldsymbol{\beta})$$

- mit unabhängigen $\beta_j;\ j \in [1 \dots k];\ k \ll n$

Maximum-Likelihood-Schätzung

- Für kanonische Verteilungen mit $a(y) = y$ gilt

$$\mathcal{E}(Y_i) = \mu_i \qquad g(\mu_i) = \eta_i$$

- **Gesucht:** Parameter θ
- **Ansatz: Max-Log-Likelihood**

$$l_i(\theta, y_i) = y_i \cdot b(\theta) + c(\theta) + d(y_i)$$

$$l(\theta, \mathbf{y}) = \sum_{i=0}^{n} l_i = \sum y_i b(\theta) + \sum c(\theta) + \sum d(y_i)$$

$$U = \frac{\mathrm{d}l}{\mathrm{d}\theta} \stackrel{!}{=} 0$$

- **Ziel:**
  - Parameter $\widehat{\theta}$
  - Maximum der Log-Likelihood $l_{max} = l(\widehat{\theta})$
- Numerische Lösung mittels Iteration nach Newton-Raphson
  (siehe Folien)
  - Für Mitglieder der Exponentialfamilie wird eine gute Näherung U' durch dessen Erwartungswert ersetzt

$$U' \leftarrow \mathcal{E}(U') = -\mathcal{I} = -\mathrm{Var}(U)$$

  - Damit iterative Lösung nach Newton-Raphson

$$\alpha^{(m)} = \alpha^{(m-1)} + \frac{U(\alpha^{(m-1)})}{\mathcal{I}(\alpha^{(m-1)})}$$

- Beispiel Ausfallwahrscheinlichkeit
  - Weibull-Verteilung

$$f(y, \lambda, \theta) = \frac{\lambda y^{\lambda-1}}{\theta^{\lambda}} \exp\Big(-\big(\frac{y}{\theta}\big)^{\lambda}\Big)$$

  - mit
    - $y > 0$ Zeit bis zum Ausfall
    - Parameter λ Form der Verteilung, hier $\lambda = 2$
      - $\lambda = 1$ wäre Exponentialverteilung mit konstanter Ausfallrate
      - *Rayleigh*-Verteilung; für gedächtnisbehaftete Lebensdauerverteilung
    - Parameter θ Skalierung. $\Rightarrow$ Diesen gilt es zu schätzen.
  - Darstellung als Exponentialfamilienmitglied:
    - $a(y) = y^{\lambda}$ (nicht kanonisch für $\lambda \neq 1$; wir benutzen $\lambda = 2$)

    - $b(\theta) = -\theta^{-\lambda}$
    - $c(\theta) = \log\lambda - \lambda\log\theta$
    - $d(y) = (\lambda - 1)\log y$
    - mit einem *Störparameter* λ
  - Log-Likelihood
    - damit kann U berechnet werden
    - $\mathcal{I}$ als Näherung $U' \leftarrow \mathcal{E}(U')$
      - im Falle der Weibull-Verteilung geschlossen lösbar
    - Damit **Scoring Methode**
- Ergebnis der *Score Methode*
  - Für die Verteilung aus der Exponentialfamilie

$$f_Y(y|\theta) = \exp\Big(a(y)b(\theta) + c(\theta) + d(y)\Big)$$

  - führt die iterative Anpassung des Verteilungsparameters θ durch die scoring Methode

$$\theta^{(m)} = \theta^{(m-1)} + \frac{U^{(m-1)}}{\mathcal{I}^{(m-1)}}$$

  - mit der *Score Statistik* U (erste Ableitung des Log-Likelihood)

$$U(\theta, y) := \frac{\mathrm{d}l}{\mathrm{d}\theta} = a(y)\cdot b'(\theta) + c'(\theta)$$

  - und der *Information* *Information* $\mathcal{I}$ (genäherte zweite Ableitung)

$$\mathcal{I} := \mathrm{Var}(U) = \mathcal{E}(U') = \frac{b''(\theta)c'(\theta)}{b'(\theta)} - c''(\theta)$$

  - in wenigen Schritten zum Ergebnis
  - Die Methode lässt sich auf mehrdimensionale Parametervektoren $\boldsymbol{\theta}$ erweitern.

2.3.5 Zusammengesetzte Wahrscheinlichkeitsverteilung - Parametervektor β

- Mehrdimensional: Scoring Methode iterative Lösung

$$\boldsymbol{\beta}^{(m)} = \boldsymbol{\beta}^{(m-1)} + \Big(\mathcal{I}(\boldsymbol{\beta}^{(m-1)})\Big)^{-1}\mathbf{U}(\boldsymbol{\beta}^{(m-1)})$$

  - Parameter $\alpha \Rightarrow$ Parametervektor $\boldsymbol{\beta}$
  - Score-Funktion $U \Rightarrow$ Score-Vektor $\mathbf{U}$
    - Gradientenvektor der Log-Likelihood $\mathbf{U} := \nabla l$

    - mit $U_j = \frac{\partial l}{\partial \beta_j}$
  - Information $\mathcal{I} \Rightarrow$ Informations-Matrix $\mathcal{I}$
- Modell-Parameter
  - Datentupel y_i, X_{ij}, Erwartungswerte μ_i und Verteilungs-Parameter θ_i mit $i \in [1 \dots n]$
  - Verdichtete Information in Parametervektor $\boldsymbol{\beta}$
  - Komponenten β_j mit $j \in [1 \dots p]$ mit i.A. $p \ll n$
- Ableitung für Max-Log-Likelihood-Schätzer
  - Berechnung unter Verwendung des Erwartungswerts
  - Umkehrfunktion
  - Kettenregel
  - $\Rightarrow$ 1. Teilergebnis:
    - Damit ergibt sich die vektorielle score-Funktion

$$U_j = \sum_{i=1}^{n} \Big(\frac{y_i - \mu_i}{\mathrm{Var}(Y_i)} \; x_{ij} \; \frac{\partial \mu_i}{\partial \eta_i} \Big)$$

      ausgedrückt durch zugängliche Größen
- Information

$$\mathcal{I} := \mathrm{Var}(U) = -\mathcal{E}(U')$$

  - Im mehrdimensionalen Fall ist die die Information $\mathcal{I}$ die Varianz-Kovarianz-Matrix der Score-Funktion U

$$\mathcal{I}_{jk} = \mathcal{E}(U_j \; U_k)$$

  - $\Rightarrow$ 2. Teilergebnis:
    - Damit ergibt sich die Informationsmatrix

$$\mathcal{I}_{jk} = \sum_{i=1}^{n} \frac{x_{ij} x_{ik}}{\mathrm{Var}(Y_i)} \Big(\frac{\partial \mu_i}{\partial \eta_i} \Big)^2$$

- Zwischenergebnis
  - Für die **Scoring Methode** ergibt sich

$$\mathbf{b}^{(m)} = \mathbf{b}^{(m-1)} + \left(\mathcal{I}^{(m-1)} \right)^{-1} \mathbf{U}^{(m-1)}$$

  - mit dem Schätzer für den Parametervektor

$$\mathbf{b} = [\beta_1, \dots, \beta_k]^T$$

  - der Inversen Informationsmatrix

$$\mathcal{I}^{-1}$$

  - und dem *score*-Vektor

$$\mathbf{U}$$

- Erweiterung

$$\mathcal{I}^{(m-1)} \mathbf{b}^{(m)} = \mathcal{I}^{(m-1)} \mathbf{b}^{(m-1)} + \mathbf{U}^{(m-1)}$$

2.4 IRLS

Zu lösendes Gleichungssystem

$$\mathbf{X}^T\mathbf{W}\mathbf{X}\mathbf{b}^{(m)} = \mathbf{X}^T\mathbf{W}\mathbf{z}$$

hat die selbe Form, wie die Normalengleichungen für ein lineares Modell

$$\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$$

- Vergleiche: Kleinste Quadrate Methode
- Designmatrix $\mathbf{X}$
- Gewichtungsmatrix $\mathbf{W}^{(m-1)}$
- Zielvektor $\mathbf{z}^{(m-1)}$
- Lösung muss iterativ gewonnen werden
  - Sowohl $\mathbf{z}$
  - als auch $\mathbf{W}$
  - hängen über $\boldsymbol{\mu}$ und $Var(Y_i)$ von $\mathbf{b}^{(m-1)}$ ab

2.4.1 iterative reweighted least squares, *IRLS*

- wird in `GLM` der Python `statsmodels` verwendet

Algorithmus

1. Finde einen Startwert $\mathbf{b}^{(0)}$
2. Berechne damit $\mathbf{z}$ und $\mathbf{W}$
3. Löse $\mathbf{X}^T\mathbf{W}\mathbf{X}\mathbf{b}^{(m)} = \mathbf{X}^T\mathbf{W}\mathbf{z}$

$$\mathbf{b}^{(m)} = \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\mathbf{z}$$

   - und wiederhole 2. und 3. bis
4. Abbruch bei Konvergenz

Ergebnis IRLS

$$\mathbf{X}^T\mathbf{W}^{(m-1)}\mathbf{X}\mathbf{b}^{(m)} = \mathbf{X}^T\mathbf{W}^{(m-1)}\mathbf{z}^{(m-1)}$$

- mit mehrdimensionaler *Iterative Reweighted Least Squares*-Methode lösbar

$$\mathbf{b}^{(m)} = \left(\mathbf{X}^T\mathbf{W}^{(m-1)}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}^{(m-1)}\mathbf{z}^{(m-1)}$$

- konvergiert in wenigen Schritten zum Schätzer $\mathbf{b} = \widehat{\boldsymbol{\beta}}$

2.4.2 Implementierung Python `statsmodels GLM`

- kann *Generalized Linear Models* mit verschiedenen Verteilungsfamilien aus der Exponentialfamilie
- benutzt IRLS um den Parametervektor β des Modells zu bestimmen
- liefert Ergebnis
  - `.predict`
  - `.fittedvalues`
  - `.params`
- Verwendung der Likelihood
  - Wahrscheinlichkeitsverteilung der Daten aus Sicht der Parameter
- Log-Likelihood
  - für Punkt-Schätzung von Parametern mittels Maximierung
  - für Intervall-Schätzung bei genäherter Verteilungsstatistik
  - Score Statistik $\mathbf{U}$ und
  - Informationsmatrix $\mathcal{I}$
    - IRLS

2.5 Parameter-Intervallschätzer

2.5.1 χ^2 Verteilung SSR

Beispiel: Lineares Modell mit Normalverteilung

$$E(Y_i) = \mu_i \qquad g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta} \qquad Y \sim \mathcal{N}(\mu_i, \sigma^2)$$

- Mit der Link-Funktion *Identität*
$$g(\mu_i) = \mu_i$$
- können alle Mittelwerte abgespalten werden:
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$
- mit unabhängigen $i = 1 \dots n$ Zufallsvariablen
$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$
= verbleibender additiver Zufall/Fehler/Rauschen mit bekannter Verteilung

Statistische Verteilung

Ist die Zufallsvariable X Normal-verteilt mit Erwartungswert μ und Standardabweichung σ

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

dann ist die *standardisierte* Zufallsvariable Standard-Normalverteilt:

$$\frac{X - \mathcal{E}(X)}{\mathrm{std(X)}} \sim \mathcal{N}(0, 1)$$

Gleichbedeutend mit

$$\frac{\Big(X - \mathcal{E}(X)\Big)^2}{\mathrm{Var}(X)} \sim \chi^2(1)$$

- Näherungsweise (Zentraler Grenzwertsatz) wenn eine große Anzahl n an Daten beitragen zu $X = \sum_{i=1}^{n} X_i$
- Sind mehrere Zufallsvariablen $X_i,\ i \in [1 \dots k]$, zusammengefasst im Vektor $\mathbf{X}$,
  - dann schreibt sich die standardisierte Quadratfehlersumme als
$$\Big(\mathbf{X} - \mathcal{E}(\mathbf{X})\Big)^T \mathbf{V}^{-1} \Big(\mathbf{X} - \mathcal{E}(\mathbf{X})\Big) \sim \chi^2(k)$$
  - mit der (nicht singulären, umkehrbaren) Varianz-Kovarianz-Matrix $\mathbf{V}$
- Insbesondere für i.i.d. Zufallsvariable mit $V_{i,i} = \mathrm{Var}(X); \quad V_{j,i \neq j} = 0$:
$$\frac{1}{\mathrm{Var}(X)} \sum_{i=1}^{k} \Big(X_i - \mathcal{E}(X_i)\Big)^2 \sim \chi^2(k)$$

2.5.2 χ^2 Verteilung Score-Statistik

Max-Likelihood-Schätzer für β

- GLM
$$\mathcal{E}(Y_i) = \mu_i \qquad g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta} = \eta_i$$
  - mit k Parametern β_j gewonnen per IRLS
- Vektorielle **score**-Statistik
$$U_j = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \Big(\frac{y_i - \mu_i}{\mathrm{Var}(Y_i)} \; x_{ij} \; \frac{\partial \mu_i}{\partial \eta_i} \Big)$$
  - Da $\mathcal{E}(Y_i) = \mu_i \; \forall i$, ist
$$\mathcal{E}(U_j) = 0 \quad \forall j$$
    wie bekannt
- Die Varianz-Kovarianz-Matrix für $\mathbf{U}$ ist
$$\mathcal{I}_{jk} = \mathcal{E}(U_j \; U_k)$$
  - mit
$$\begin{aligned} \mathcal{I}_{jk} &= \mathcal{E}\left(\sum_{i=1}^{n} \Big(\frac{y_i - \mu_i}{\mathrm{Var}(Y_i)} \; x_{ij} \; \frac{\partial \mu_i}{\partial \eta_i} \Big) \sum_{i=1}^{n} \Big(\frac{y_i - \mu_i}{\mathrm{Var}(Y_i)} \; x_{ik} \; \frac{\partial \mu_i}{\partial \eta_i} \Big) \right) \\ &= \sum_{i=1}^{n} \frac{\mathcal{E}\Big((Y_i - \mu_i)^2 \Big) x_{ij} x_{ik}}{\Big(\mathrm{Var}(Y_i) \Big)^2} \Big(\frac{\partial \mu_i}{\partial \eta_i} \Big)^2 \end{aligned}$$
- Damit hat die Standardisierte Quadratfehlersumme für den Score-Vektor $\mathbf{U}$, mit Erwartungswert $\mathcal{E}(U_j) = 0$ die Verteilung
$$\mathbf{U}^T \mathcal{I}^{-1} \, \mathbf{U} \sim \chi^2(k)$$
  - exakt für Normalverteilte Y, näherungsweise für große Stichproben.

Beispiel 1: Normalverteilung

- Seien Y_i i.i.d. normalverteilte Zufallsvariablen $y_i \sim \mathcal{N}(\mu, \sigma^2)$ mit bekannter Varianz σ^2 und gesuchtem Parameter μ.
$$l = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 - n \log \left(\sigma \sqrt{2\pi} \right)$$
- Die *score*-Statistik ist
$$U = \frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \mu) = \frac{n}{\sigma^2} (\overline{Y} - \mu)$$
- woraus man den Punktschätzer erhält
$$\widehat{\mu} = \overline{Y}$$

- Dann
$$\mathcal{E}(U) = \frac{1}{\sigma^2}\sum_{i=1}^{n}\Big(\mathcal{E}(Y_i) - \mu\Big) = 0$$
$$\mathrm{Var}(U) = \mathcal{I} = \frac{1}{\sigma^4}\sum_{i=1}^{n}\mathrm{Var}(Y_i) = \frac{n}{\sigma^2}$$
- Damit ist
$$\mathbf{U}^T\mathcal{I}^{-1}\mathbf{U} \sim \chi^2(p) = \frac{U^2}{\mathcal{I}} = \frac{(\overline{Y} - \mu)^2}{\sigma^2/n} \sim \chi^2(1)$$
ein exaktes Ergebnis für $\widehat{\mu}$.
- Also liegt auch das 95%-Konfidenzintervall für $\widehat{\mu}$ fest:
$$\overline{y} \pm \Phi_{(1-\alpha/2)}\frac{\sigma}{\sqrt{n}}$$
- Ergebnis
  - Mit der *Generalized Linear Models* Methode lässt sich das Konfidenzintervall für den Schätzer $\widehat{\mu}$ genauso bestimmen, wie mit klassischer Verteilungsannahme.

Beispiel 2: Binomialverteilung

- Seien Y_i i.i.d. binomialverteilte Zufallsvariablen $y_i \sim \mathcal{B}(n, \pi)$
$$l(\pi; y) = y\ln\pi + (n-y)\ln(1-\pi) + ln(\binom{n}{y})$$
- Die *score*-Statistik ist
$$U = \frac{\partial l}{\partial \pi} = \frac{Y}{\pi} - \frac{n-Y}{1-\pi} = \frac{Y - n\pi}{\pi(1-\pi)}$$
- Mit $\mathcal{E}(Y) = n\pi$ ergibt sich wieder
$$\mathcal{E}(U) = 0$$
- Mit $\mathrm{Var}(Y) = n\pi(1-\pi)$ ergibt sich
$$\mathrm{Var}(U) = \mathcal{I} = \frac{1}{\pi^2(1-\pi)^2}\mathrm{Var}(Y) = \frac{n}{\pi(1-\pi)}$$
- Damit ist
$$\frac{\mathbf{U}}{\sqrt{\mathcal{I}}} = \frac{Y - n\pi}{\sqrt{n\pi(1-\pi)}} \sim \mathcal{N}(0,1)$$
  - näherungsweise und die bekannte Normal-Näherung für binomialverteilte Zufallsvariablen

2.5.3 Allgemeine Intervallschätzung

Taylor Entwicklung

- Jede glatte Funktion f(x) läßt sich nach Taylor als Reihe ihrer Ableitungen um eine Stelle x_0 entwickeln

$$f(x) = f(x_0) + (x - x_0)\frac{\mathrm{d}f}{\mathrm{d}x}\Big|_{x=x_0} + \frac{1}{2}(x - x_0)^2\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}\Big|_{x=x_0} + \dots$$

- Taylor Entwicklung Log-Likelihood
  - (erste drei Terme) der Log-Likelihood für einen **skalaren** Parameter β um die Stelle $\beta = b$

$$\begin{aligned} l(\beta) &= l(b) + (\beta - b)U(b) + \frac{1}{2}(\beta - b)^2 U'(b) + \dots \\ &\approx l(b) + (\beta - b)U(b) - \frac{1}{2}(\beta - b)^2 \mathcal{I}(b) \end{aligned}$$

  - und für einen Parameter**vektor** $\boldsymbol{\beta}$

$$l(\boldsymbol{\beta}) \approx l(\mathbf{b}) + (\boldsymbol{\beta} - \mathbf{b})U(\mathbf{b}) - \frac{1}{2}(\boldsymbol{\beta} - \mathbf{b})^T \mathcal{I}(\mathbf{b})(\boldsymbol{\beta} - \mathbf{b})$$

  - Taylor-Entwicklung (erste zwei Terme) der *Score*-Statistik

$$\mathbf{U}(\boldsymbol{\beta}) \approx \mathbf{U}(\mathbf{b}) - \mathcal{I}(\mathbf{b})\,(\boldsymbol{\beta} - \mathbf{b})$$

Parameter β - Verteilung für ML-Schätzer

- Der Schätzer $\mathbf{b} = \widehat{\boldsymbol{\beta}}$ maximiert $l(\boldsymbol{\beta})$ mit $\mathbf{U}(\mathbf{b}) = 0$.
- Damit

$$\mathbf{U}(\boldsymbol{\beta}) = -\mathcal{I}(\mathbf{b})\,(\boldsymbol{\beta} - \mathbf{b})$$

- bzw.

$$(\mathbf{b} - \boldsymbol{\beta}) = \mathcal{I}^{-1}\mathbf{U}(\boldsymbol{\beta})$$

- Sieht man $\mathcal{I}$ als konstant an, dann ist wegen $\mathcal{E}(\mathbf{U}) = 0$ auch

$$\mathcal{E}(\mathbf{b}) = \boldsymbol{\beta}$$

  - also $\mathbf{b}$ ein (asymptotisch) erwartungstreuer Schätzer für $\boldsymbol{\beta}$

- Die Varianz-Kovarianz-Matrix für $\mathbf{b}$ ist damit

$$\mathcal{E}\Big((\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})^T\Big) = \mathcal{I}^{-1}\mathcal{E}(\mathbf{U}\mathbf{U}^T)\mathcal{I}^{-1} = \mathcal{I}^{-1}$$

  - wegen $\mathcal{I} = \mathcal{E}(\mathbf{U}\mathbf{U}^T)$ und $(\mathcal{I}^{-1})^T = \mathcal{I}^{-1}$ (Symmetrie)

- Mit dieser Varianz-Kovarianz-Matrix $\mathbf{V} = \mathcal{I}^{-1}$ ergibt sich für die standardisierte Quadratfehlersumme (asymptotisch)

$$(\mathbf{b} - \boldsymbol{\beta})^T \mathcal{I}(\mathbf{b})(\mathbf{b} - \boldsymbol{\beta}) \sim \chi^2(p)$$

  - die **Wald**-Statistik

- in der eindimensionalen Form die bekannte

$$b \sim \mathcal{N}(\beta, \mathcal{I}^{-1})$$

- Ergebnis:
  - Der Punktschätzer $\widehat{\boldsymbol{\beta}}$ des Parameters $\boldsymbol{\beta}$ ist (näherungsweise) verteilt

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}(\beta, \mathcal{I}^{-1})$$

  - mit der Informationsmatrix (Fischer Information) $\mathcal{I}$

$$\begin{aligned} \mathcal{I}_{jk} &= \mathcal{E}(U_j\, U_k) \\ &= \sum_{i=1}^{n} \frac{\mathcal{E}\Big((Y_i - \mu_i)^2\Big) x_{ij} x_{ik}}{\Big(\mathrm{Var}(Y_i)\Big)^2} \Big(\frac{\partial \mu_i}{\partial \eta_i}\Big)^2 \end{aligned}$$

Beispiel Normalverteilung

$$\mathcal{E}(Y_i) = \mu_i = \mathbf{x}_i^T \boldsymbol{\beta}\ ; \quad Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$$

- Die Information hat (Link: Identität, $\mathrm{Var}(Y_i) = \sigma^2$) die einfache Form

$$\mathcal{I}_{jk} = \sum_{i=1}^{n} \frac{x_{ij} x_{ik}}{\sigma^2}$$

- oder

$$\mathcal{I} = \frac{1}{\sigma^2} \mathbf{X}^T \mathbf{X}$$

- Für die rechte Seite von $\mathbf{X}^T \mathbf{W} \mathbf{X} \mathbf{b}^{(m)} = \mathbf{X}^T \mathbf{W} \mathbf{z}$ war

$$z_i = \sum_{k=1}^{p} x_{ik} b_k^{(m-1)} + (y_i - \mu_i)$$

- Da $\mu_i\Big|_{b^{(m-1)}} = \mathbf{x}_i^T \mathbf{b}^{(m-1)} = \sum_{k=1}^{p} x_{ik} b_k^{(m-1)}$ verbleibt

$$z_i = y_i$$

- Damit wird das zu lösende LGS $\mathbf{X}^T \mathbf{W} \mathbf{X} \mathbf{b}^{(m)} = \mathbf{X}^T \mathbf{W} \mathbf{z}$ zu

$$\frac{1}{\sigma^2} \mathbf{X}^T \mathbf{X} \mathbf{b} = \frac{1}{\sigma^2} \mathbf{X}^T \mathbf{y}$$

- also
$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$
  - der aus `OLS` bekannte Maximum-Likelihood-Schätzer für den Parametervektor $\boldsymbol{\beta}$
- Punktschätzer
  - Mit $\mathbf{y} \sim mv\mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbb{1})$
  - wird der Erwartungswert
$$\mathcal{E}(\mathbf{b}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$
  - und $\mathbf{b}$ ein erwartungstreuer Schätzer für $\boldsymbol{\beta}$
$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$
- Intervallschätzer
  - Mit
$$\begin{aligned}(\mathbf{b} - \boldsymbol{\beta}) &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} - \boldsymbol{\beta} \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\end{aligned}$$
  - folgt die Varianz-Kovarianz-Matrix für $\mathbf{b}$
$$\begin{aligned}\mathcal{E}\Big((\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})^T\Big) &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathcal{E}\Big((\mathbf{y} - \mathbf{X}\boldsymbol{\beta})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\Big)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\Big(\mathrm{Var}(\mathbf{y})\Big)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \mathcal{I}^{-1}\end{aligned}$$
    - wie bereits bekannt

Vergleich klassische Statistik

- Sind die Messwerte $\mathbf{y}$ normalverteilt, sind es auch die Schätzer für die Parameter des linearen Modells, also
$$\mathbf{b} \sim \mathcal{N}(\boldsymbol{\beta}, \mathcal{I}^{-1})$$
- oder für die standardisierte Quadratfehlersumme
$$(\mathbf{b} - \boldsymbol{\beta})^T\mathcal{I}\,(\mathbf{b} - \boldsymbol{\beta}) \sim \chi^2(k)$$
- wie aus der klassischen Statistik bekannt.
- Der GLM-Formalismus reproduziert die von der Normalverteilung bekannten Ergebnisse
  - Diese sind näherungsweise für andere Verteilungen anwendbar
  - Methoden sind implementiert in `statsmodels GLM`

2.6 Logistische Regression

Bernoulli-Experiment

- Ereignis A *tritt ein* oder *tritt nicht ein* $\Omega = \{A, \bar{A}\}$
- Binäre Zufallsvariable Z '*Indikatorvariable*' kann nur Werte $\omega \in \{0, 1\}$ annehmen

$$Z = \begin{cases} 1 & \text{wenn A zutrifft} \\ 0 & \text{wenn A nicht zutrifft} \end{cases}$$

- Beispiele
  - Münzwurf: Kopf / Zahl
  - Produktion: innerhalb Toleranz / Ausschuss
  - Geburten: Mädchen / Jungen
  - Psychophysik: gesehen / nicht-gesehen

Bernoulli-Verteilung

- Die Wahrscheinlichkeit für A sei π

$$\begin{aligned} P(A) &= P(Z=1) = & \pi \\ P(\bar{A}) &= P(Z=0) = & 1-\pi \end{aligned}$$

- Schreibweise

$$P(Z) = \pi^Z (1-\pi)^{1-Z}$$

- n unabhängige Zufallsvariablen $Z_1 \dots Z_n$ mit Einzel-Wahrscheinlichkeiten $P(Z_i) = \pi_i$ haben eine gemeinsame Verbund-Wahrscheinlichkeitsverteilung

$$\prod \pi_j^{Z_j} (1-\pi_j)^{(1-Z_j)} = \exp\Big[\sum_{j=1}^{n} Z_j \log\Big(\frac{\pi_j}{1-\pi_j}\Big) + \sum_{j=1}^{n} \log\Big(1-\pi_j\Big)\Big]$$

  - welche Mitglied der kanonischen Exponentialfamilie ist
- Im 1. Spezialfall gleicher Wahrscheinlichkeiten

$$\pi_i = \pi$$

  - ergibt sich für die Zufallsvariable *Anzahl der Erfolge*

$$Y = \sum_{i=1}^{n} Z_i$$

Die Binomialverteilung

$$P(Y=y) = \binom{n}{y}\pi^y(1-\pi)^{n-y}$$

- wobei $y \in [0 \dots n]$
- Die Log-Likelihood ist

$$l(\pi; y) = y\ln\Big(\frac{\pi}{1-\pi}\Big) + n\ln\Big(1-\pi\Big) + \ln\binom{n}{y}$$

- im 2. allgemeineren Fall mit N Kategorien
  - Bei N kategorial unterschiedlichen Wahrscheinlichkeiten summiert sich die Log-Likelihood

$$l(\pi_1, \dots, \pi_N; y_1, \dots, y_N) = \sum_{j=1}^{N}\Big(y_j\ln\Big(\frac{\pi_j}{1-\pi_j}\Big) + n_j\ln\Big(1-\pi_j\Big) + \ln\binom{n_j}{y_j}\Big)$$

  - mit in jeder Kategorie j
    - y_j Erfolge
    - $n_j - y_j$ Misserfolge

2.7 Toleranzverteilung

- Beschreiben der *Erfolgsrate* als *Generalisiertes Lineares Modell*
  - Zufallsvariable $P_j = \frac{Y_j}{n_j}$
  - mit Erwartungswert $\mathcal{E}(Y_j) = n_j\pi_j \quad \Rightarrow \quad \mathcal{E}(P_j) = \pi_j$
  - sei abhängig von erklärenden Variablen/Kategorien.

$$g(\pi_j) = \mathbf{x}_i^T\boldsymbol{\beta}$$

2.7.1 Lineares Modell

$$\pi_j = \mathbf{x}_i^T\boldsymbol{\beta}$$

- unangemessen (Siehe Beispiel Piloten-Ohnmacht: Grenzwertüberschreitung, Seite 24)
- Begrenzung: Beschränken auf eine Verteilungsfunktion (cdf)

$$\pi = F(t) = \int_{-\infty}^{t} f(s)\mathrm{d}s$$

  - $f(s) \geq 0$ nicht-negative Wahrscheinlichkeit
  - $\int_{-\infty}^{\infty} = 1$ Normierung
  - *Toleranzverteilung*

2.7.2 Beschränkt-Lineares Modell / Rechteckverteilung

- Wählt man als Toleranzverteilung die Rechteckverteilung

$$f(s) = \begin{cases} \frac{1}{c_2 - c_1} & \text{wenn } c_1 \leq s \leq c_2 \\ 0 & \text{sonst} \end{cases}$$

- dann ist π kummulativ in x

$$\pi = \int_{c_1}^{x} f(s)\mathrm{d}s = \frac{x - c_1}{c_2 - c_1} \qquad \text{für } c_1 \leq s \leq c_2$$

- bzw.

$$\pi = \beta_0 + \beta_1 x$$

- mit $\beta_0 = \frac{-c_1}{c_2 - c_1}$ und $\beta_1 = \frac{1}{c_2 - c_1}$
- wird selten benutzt

2.7.3 Probit

- Wählt man als Toleranzverteilung die Normalverteilung

$$f(s) = \mathcal{N}(\mu, \sigma^2)$$

- erhält man

$$\pi = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

- Damit erhält man die gewünschte lineare Abhängigkeit von x

$$g(\pi) = \Phi^{-1}(\pi) = \beta_0 + \beta_1 x$$

- mit $\beta_0 = \frac{-\mu}{\sigma}$ und $\beta_1 = \frac{1}{\sigma}$
- Dieses Modell kommt häufig in Biologie und Sozialwissenschaften vor
- Interpretation: versteckte Variable

2.7.4 Logistisches Modell / Logit

- Wählt man als Toleranzverteilung

$$f(s) = \frac{\beta_1 \exp(\beta_0 + \beta_1 s)}{\Big(1 + \exp(\beta_0 + \beta_1 s)\Big)^2}$$

- womit

$$\pi = \int_{-\infty}^{x} f(s)\mathrm{d}s = \frac{\beta_1 \exp(\beta_0 + \beta_1 s)}{1 + \exp(\beta_0 + \beta_1 s)}$$

- dann erhält man die gewünschte lineare Abhängigkeit von x mittels *Logit-Funktion*

$$g(\pi) = \ln \frac{\pi}{1-\pi} = \beta_0 + \beta_1 x$$

- welche als logarithmisches Chancenverhältnis '*Log-odds-ratio*' interpretiert werden kann
- Wird oft benutzt, vor allem für Binomial-verteilte Daten, deren *natürliche* Link-Funktion

2.7.5 Extremwertverteilung / c-log-log

- Wählt man als Toleranzverteilung die Extremwertverteilung

$$f(s) = \beta_1 \exp\Big((\beta_0 + \beta_1 s) - \exp(\beta_0 + \beta_1 s)\Big)$$

- damit

$$\pi = 1 - \exp\Big(-\exp(\beta_0 + \beta_1 t)\Big)$$

- Mit der komplementären Log-Log-Funktion erhält man die gewünschte lineare Abhängigkeit

$$g(\pi) = \log\Big(-\log(1-\pi)\Big) = \beta_0 + \beta_1 x$$

- Dies ist die *komplementäre-Log-Log-Funktion*

2.8 Beispiele

2.8.1 LD50

Käfer wurden einem Gift ausgesetzt, woran sie in Abhängigkeit von Konzentration $\log_{10} \frac{mg}{l}$ starben [Bliss, 1935]

Fragestellung: Was ist die 50%-Lethaldosis *LD50*

- Ab welcher Dosis sterben mehr als die Hälfte der Käfer?

Lösung mittels Logit-Link

- Ergebnis: Der Max-Likelihood-Parameter-Vektor-Schätzer ist

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} -60.7 \\ 34.3 \end{bmatrix} \qquad \widehat{\boldsymbol{\beta}}_{95\%CI} = \begin{bmatrix} -70.9 \cdots -50.6 \\ 28.6 \ldots 40.0 \end{bmatrix}$$

- $\Rightarrow$ LD50 kann nun aus dem Modell ausgegeben werden

Frage: Vertrauensbereich der LD50?

- Wie verlässlich ist die LD50-Angabe?
- Lösung: Suche das 95% Konfidenzintervall
- Für normalverteilte Werte gilt das Gauß'sche Fehlerfortpflanzungsgesetz
  - Hier nicht
- Ausweg: 'worst case'

Diskrepanz zwischen Daten und Modell

- Grund: Parameter korrelieren
- Ursache:
  - Lineare Vorhersage für η geht vom Ursprung aus: Konzentration $log(dose) = 0$
    - Eine kleine Änderung in der Steigung bewirkt eine große Änderung von η
    - Eine Änderung im Achsenabschnitt bewirkt eine Verschiebung
    - Fit benötigt Variation beider
      - breite Randverteilung
- Abhilfe
  - Modellparametrisierung entkoppeln durch Zentrieren
  - Im Mittelpunkt der Datenwolke sind Steigung und Achsenabschnitt unabhängig

$$x \to x - x_0$$

- Ergebnis Zentrierung

  - Der Achsenabschnitt verschwindet wie erwartet
    - sein 95%-Konfidenzintervall ist deutlich enger
    - er ist nicht signifikant
  - Einziger linearer Parameter β_1 behält seinen ursprünglichen Wert
    - ebenso sein 95%-Konfidenzintervall
- Interpretation als Verdünnungsreihe
  - Ausgangskonzentration eines Gifts ρ_0
  - Unabhängige Variable: Verdünnungsfaktor zB. halbieren je Schritt x:

$$\rho_x = \rho_0 \frac{1}{2^x}$$

  - Logarithmieren linearisiert die Abhängigkeit von x

$$\ln(\rho_x) = \ln(\rho_0) - x \ln(2)$$

Zwischenergebnis

- Parameterschätzer und Konfidenzintervall durch GLM bestimmen
- Konfidenzintervall für Daten
  - wenn bekannt: bei Normalverteilung t-verteilt
  - wenn unbekannt:
    - simulieren
    - 'worst case' Abschätzung durch **unabhängige Parametrisierung**
  - Logit-Modell beschreibt Dosisabhängigkeit gut
    - logarithmische Abhängigkeit sorgt für Linearität in der Verdünnungsreihe

Andere Link-Funktion

- Probit
- Extremwertverteilung (komplementäre-log-log)
- Kein direkter Vergleich der drei Modelle möglich!
  - Siehe Kapitel Modellvergleich und Deviance (ab Seite 19)
  - Die Entscheidung für ein Modell muss aus der Theorie kommen
    - Ob das Modell dann angemessen ist, das kann getestet werden

2.8.2 Wahrnehmungsexperiment Wahrnehmungsschwelle

- Kategoriale abhängige Variable
  - Dichotome Variable
    - Ja / Nein - Experiment
    - Merkmal liegt vor / liegt nicht vor
- Experimente zur Wahrnehmungsschwelle
  - Gabor-Muster
  - Kontrast $\in \{0 \dots 1\}$
    $[100\%, 20\%, 4\%, 0.8\%]$
  - Streifenbreite spatial frequency: x cpd
- Durchführung
  - 12 Kontraste
  - 20 Wiederholungen jeweils
- Gesucht: Wahrnehmungsschwelle
  - Festgelegte Schwelle z.B. 75%
- Sinnvolle Darstellung der dichotomen Daten
  - Anteil der korrekten Antworten in Prozent
- Was ist die Wahrnehmungsschwelle?
  - Zufall: Rauschen, Zwinkern, Aufmerksamkeit, Müdigkeit
  - Kein Zufall: Adaption, Individualität (Genetik?), ...
  - Modell
  - Anpassen
  - Auswerten

Generalisiertes Lineares Modell

- Unabhängige Variable:
  - Kontrast
  - individuelle Versuchsperson
  - Umgebungshelligkeit
  - Streifenmuster (Breite, Winkel,...)
  - ...
- Abhängige Variable

  - Binomialverteilung 0/1
  - Anteil Antwort 'percent correct'
- Modellparameter
  - y-Achsenabschnitt
  - Abhängigkeit vom *Kontrast*
  - Logit-Link
- Gesucht
  - Wahrnehmungsschwelle

2.8.3 Wahrnehmungsexperiment Rezeptive Felder

- ja / nein Experiment
  - Signalentdeckungstheorie
- Erzwungene Alternative
  - Bei mehreren Auswahlmöglichkeiten geht die Antwort-Wahrscheinlichkeit nicht auf Null, sondern startet von einem 'Zufalls'-Niveau $\frac{1}{N_a}$
  - Solch ein Modell muss gesondert erstellt werden
    - Beispielweise *Psignifit*
- Beispiel-Experiment: Drei Versuchspersonen sollen Störungen erkennen
  - Störungen bestehen aus helligkeitsgleichen grob gefilterten Strukturen
  - Fixation auf Bildmitte
  - Störung hat verschiedene Durchmesser
  - Abstand der Störung zur Bildmitte variiert, 'Exzentrizität'
  - Störung wird an vier möglichen Positionen gezeigt
- Daten
  - Spalten
    - Versuchspersonenkürzel, Versuchsnummer, innerhalb Versuchs-Block Nummer, Bildausrichtung
    - Exzentrizität in ° vom Fixationspunkt
    - Durchmesser der Störung in log 10 Pixel
    - Ort, an dem Störung gezeigt wurde und Antwort der Versuchsperson
    - **correct**, wenn richtig
  - Zeilen
    - jeweils ein Versuch $\sim$ Designmatrix

- Daten und Modell
  - abhängige Variable: Antworten
    - Dichotom 'richtig' / 'falsch'
  - unabhängige Variablen:
    - Exzentrizität
    - Größe '*patch size*'
    - Versuchsperson
  - Generalisiertes Lineares Modell
    - Binomialverteilung
    - Logistische Linkfunktion *logit*
    - keine Beeinflussung (Interaktion) zwischen den unabhängigen Variablen
- Ergebnis
  - Einfluss der Stimulusgröße
    - Je größer ein Stimulus (bei gleicher Exzentrizität), desto besser wird er erkannt
    - Je-desto ist nicht linear, sondern vermittels Link-Funktion *logit*
  - Einfluss der Exzentrizität
    - Je weiter in der Peripherie ein Stimulus (gleiche Größe) gezeigt wird, desto schlechter wird er erkannt

3 Principal Components Analysis - PCA

Ziele der Hauptkomponentenanalyse - PCA

- Wichtigste Informationen aus Daten extrahieren
- Unwichtige Daten verwerfen
- Beschreibung der Daten vereinfachen
- Struktur in den Daten erkennen

3.1 Lineare Abhängigkeit

Lineares Modell

$$\mathcal{E}(Y) = \mathbf{X}\boldsymbol{\beta}$$

- unabhängige Variable X, vorhersagende Variable
  - Designmatrix $\mathbf{X}$
- abhängige Variable Y, gemessene Größe
  - Erwartungswert abhängig von vorhersagenden Variablen, Streuung, Messfehler, Zufall (modellabhängig)

Beispiel: Testat-Punkte und Klausur-Punkte

- Lineare Abhängigkeit $y \sim x \neq x \sim y$
- Ergebnis: ein lineares Modell ist angemessen, wenn
  - Kausale Abhängigkeit bekannt
  - Fehler *nur* in abhängiger Variable

Problem:

- Absolute Benotung?
- Reihenfolge: Wenn beide Zufallsvariablen gleichberechtigt sind...
  - welche Reihenfolge würden wir dann annehmen?
  - und wie diese sinnvoll bestimmen?

3.2 Multivariate Verteilung

von (zwei) gleichberechtigten Zufallsvariablen

3.2.1 Projektion

auf eine (neue, bestmögliche) Zufallsvariable

- Projektionen beispielsweise auf
  - x-Achse (1:0)
  - y-Achse (0:1)
  - 45° Diagonale (1:1)
  - 60° Diagonale
- Zwischenergebnis Pojektion
  - Skalarprodukt Vektor $\mathbf{x}$ mit Vektor $\mathbf{e}$ ergibt Koordinate x' in Bezug auf Vektor $\mathbf{e}$
  - Koordinate $x' \cdot \mathbf{e}$ beschreibt Projektion von $\mathbf{x}$ auf $\mathbf{e}$
  - Information über zu $\mathbf{e}$ senkrechte Richtung wird ignoriert
- Python: Matrix-Multiplikation mit x-y-Koordinaten
  - `np.vstack((x, y)).T`
  - `np.dot()`

3.2.2 Varianz

- Kennzahl für Streuung einer Variablen
- Empirische Varianz für einen Datensatz X_i mit $i \in [1 \dots n]$

$$\mathrm{Var}(X) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

- Varianz einer Zufallsvariablen

$$\mathrm{Var}(X) = \mathcal{E}\Big(\big(X - \mathcal{E}(X)\big)^2\Big)$$

- Schätzer der Varianz
  - $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$ für die Stichprobenvarianz $Var(X) = \sigma^2$
  - $\tilde{S}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2$ für die empirische Varianz $Var(X) = \sigma^2$
- Zwei Variable X und Y
  - $\mathrm{Var}(X),\ \mathrm{Var}(Y)$.

3.2.3 Kovarianz

- Die Kovarianz zweier verbundener Zufallsvariablen X_1 und X_2 mit gemeinsamer Verteilung $f(x_1, x_2)$ ist

$$\begin{aligned}\mathrm{Cov}(X_1, X_2) &= \mathcal{E}\Big(\big(X_1 - \mathcal{E}(X_1)\big)\big(X_2 - \mathcal{E}(X_2)\big)\Big) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)\cdot(x - \mathcal{E}(X))\cdot(y - \mathcal{E}(Y))\,\mathrm{d}y\,\mathrm{d}x\end{aligned}$$

  $\Rightarrow$ Kennzahl für *linearen* Zusammenhang $X_1 \sim X_2$

- 'je-desto'
- Gewichtung in Quadranten
- Schätzer der Kovarianz, empirische Kovarianz

$$\widehat{C}_{X_1,X_2} = \frac{1}{n-1}\sum_{i=1}^{n}\Big(\big(X_1 - \overline{X}_1\big)\big(X_2 - \overline{X}_2\big)\Big)$$

Eigenschaften der Kovarianz

- Verschiebungssatz
$$\mathrm{Cov}(X_1, X_2) = \mathcal{E}(X_1 \cdot X_2) - \mathcal{E}(X_1)\cdot\mathcal{E}(X_2)$$
- Symmetrie
$$\mathrm{Cov}(X_1, X_2) = \mathrm{Cov}(X_2, X_1)$$
- Varianzen
$$Var(X_i) = \mathrm{Cov}(X_i, X_i)$$
  $\Rightarrow$ Varianzen sind Diagonalelemente der Kovarianzmatrix
- Kovarianz-Matrix von n-dimensionaler Zufallsvariable $\mathbf{X}$ und ihrem Erwartungswert $\boldsymbol{\mu}$

$$\mathrm{Cov}(\mathbf{X}) = \mathcal{E}[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T] = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1,X_2) & \dots & \mathrm{Cov}(X_1,X_n) \\ \mathrm{Cov}(X_1,X_2) & \mathrm{Var}(X_2) & \dots & \mathrm{Cov}(X_2,X_n) \\ \vdots & & \ddots & \vdots \\ \mathrm{Cov}(X_1,X_n) & \mathrm{Cov}(X_2,X_n) & \dots & \mathrm{Var}(X_n) \end{pmatrix}$$

- Unter linearer Transformation $X_1' = a_1X_1 + b_1$ und $X_2' = a_2X_2 + b_2$

$$\mathrm{Cov}(X_1', X_2') = a_1 \cdot a_2 \cdot \mathrm{Cov}(X_1, X_2)$$

Linearkombination

$$Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n$$

- Dann ist der Erwartungswert

$$\begin{aligned}\mathcal{E}(Y) &= a_1\mathcal{E}(X_1) + a_2\mathcal{E}(X_2) + \cdots + a_n\mathcal{E}(X_n)\\ &= \sum_{i=1}^{n} a_i\mathcal{E}(X_i)\end{aligned}$$

- und die Varianz

$$\begin{aligned}\mathrm{Var}(Y) &= \mathcal{E}\Big(\big(Y - \mathcal{E}(Y)\big)^2\Big)\\ &= a_1^2Var(X_1) + a_2^2Var(X_2) + \cdots + a_n^2Var(X_n)\\ &\quad + 2a_1a_2\mathrm{Cov}(X_1, X_2) + 2a_1a_3\mathrm{Cov}(X_1, X_3) + \ldots\\ &= \sum_{i=1}^{n} a_i^2Var(X_i) + 2\sum_{j=1}^{n}\sum_{i=1}^{j-1} a_ia_j\mathrm{Cov}(X_i, X_j)\\ &= \sum_{j=1}^{n}\sum_{i=1}^{n} a_ia_j\mathrm{Cov}(X_i, X_j)\end{aligned}$$

- Erinnerung:

  Bei *unabhängigen* Zufallsvariablen haben sich die Varianzen addiert

Ergebnis

- Kovarianz beschreibt einen linearen Zusammenhang
- Kein linearer Zusammenhang ⇒ Kovarianz (nahe) 0

3.2.4 Anwendung Reduktion der Dimension

Ein linearer Zusammenhang erlaubt

- Datenreduktion durch Ersetzung
- bei (geringem) Informationsverlust

Fragestellung: Welcher Zusammenhang?

Lösungsansatz

- Suche Linearkombination aus X und Y, sodass restliche Fehler/ Informationsverlust minimal werden

Ziel

- maximale Varianz gewünscht
- minimale Varianz ausblenden

3.2.5 Korrelation

Voraussetzung: Gleichberechtigte Variablen

- Gemeinsame Variation: Kovarianz
- Linearer Zusammenhang

Empirischer Korrelationskoeffizient

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2(Y_i - \bar{Y})^2}}$$

Korrelationskoeffizient zweier Zufallsvariablen

$$\rho = \rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

- invariant unter Skalierung

3.2.6 Korrelationstest

Seien $(X_i,\ Y_i) \quad i \in \{1 \dots n\}$ unabhänge, gemeinsam normalverteilte Zufallsvariablen

Nullhypothese *unabhängig, unkorreliert*

(a) $H_0: \quad \rho_{XY} = 0 \qquad H_1: \quad \rho_{XY} \neq 0$
(b) $H_0: \quad \rho_{XY} = 0 \qquad H_1: \quad \rho_{XY} < 0$
(c) $H_0: \quad \rho_{XY} = 0 \qquad H_1: \quad \rho_{XY} > 0$

- Teststatistik

$$T = \frac{r_{XY}}{\sqrt{1 - r_{XY}^2}}\sqrt{n-2}$$

- Verteilung unter $H_0:\ \rho_{XY} = 0$

$$T \sim t(n-2)$$

- Ablehnungsbereich

(a) $|T| > t_{1-\alpha/2(n-2)}$
(b) $T < -t_{1-\alpha(n-2)}$
(c) $T > t_{1-\alpha(n-2)}$

Allgemeine Nullhypothese

(a) $H_0: \quad \rho_{XY} = \rho_0 \qquad H_1: \quad \rho_{XY} \neq \rho_0$
(b) $H_0: \quad \rho_{XY} \geq \rho_0 \qquad H_1: \quad \rho_{XY} < \rho_0$
(c) $H_0: \quad \rho_{XY} \leq \rho_0 \qquad H_1: \quad \rho_{XY} > \rho_0$

- Teststatistik

$$Z = \frac{1}{2}(ln\frac{1+r_{XY}}{1-r_{XY}} - ln\frac{1+\rho_0}{1-\rho_0})\sqrt{n-3}$$

- Verteilung unter $H_0: \ \rho_{XY} = \rho_0$ approximativ $(n > 25)$

$$Z \sim \mathcal{N}(0,1)$$

- Ablehnungsbereich

(a) $|Z| > z_{1-\alpha/2}$
(b) $Z < -z_{1-\alpha}$
(c) $Z > z_{1-\alpha}$

3.2.7 Zweidimensionale Normalverteilung

Mit den Parametern

- $\mu_x = \mathcal{E}(X)$
- $\mu_y = \mathcal{E}(Y)$
- $\sigma_x^2 = Var(X)$
- $\sigma_y^2 = Var(Y)$
- $\rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_x \sigma_y}$

ergibt sich

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} exp\{\arg\}$$

mit

$$\arg = -\frac{1}{2(1-\rho^2)}[(\frac{x-\mu_x}{\sigma_x})^2 + 2\rho\frac{x-\mu_x}{\sigma_x}\frac{y-\mu_y}{\sigma_y} + (\frac{y-\mu_y}{\sigma_y})^2]$$

3.2.8 (lineare) Unabhängigkeit

Für gemeinsam **normal**verteilte Zufallsvariablen X und Y gilt:

- X und Y sind unabhängig $\Leftrightarrow$ unkorreliert $C = 0$

Für beliebig verteilte Zufallsvariablen X und Y gilt:

- X und Y sind unabhängig $\Rightarrow$ Korrelation $C = 0$
- X und Y sind linear unabhängig $\Leftrightarrow$ Korrelation $C = 0$

Höherdimensionale Normalverteilung

- Analog mittels Kovarianzmatrix, jedoch unanschaulich

3.3 Datenreduktion

3.3.1 Beispiel: Gewichte von Säugetieren

- Gleichberechtigte Variablen X_1 log-Körpergewicht und X_2 log-Gehirnmasse
- Statistik
  - Grundgesamtheit: `data.describe()`
- Abhängigkeit der Gewichte: `body vs. brain`
  - *Eine* Variable?
  - Größte Varianz
  - Mischung von zwei Variablen

$$Y = a \cdot X_1 + b \cdot X_2$$

    mit X_1: log(BodyWt) und X_2: log(BrainWt)
  - = Projektion auf Unterraum Y
    - eindimensional
  - maximale Varianz?

$$\begin{aligned} \mathrm{Var}(Y) &= \mathcal{E}\Big(\big(Y - \mathcal{E}(Y)\big)^2\Big) \\ &... = a^2\mathrm{Var}(X_1) + b^2\mathrm{Var}(X_2) + 2 \cdot a \cdot b \cdot \mathrm{Cov}(X_1, X_2) \end{aligned}$$

3.3.2 Datenreduktion auf eine Dimension 'allgemeines Gewicht'

$$y_i = d_1 \cdot x_{i1} + d_2 \cdot x_{i2} = \mathbf{d} \cdot \mathbf{x}_i$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} d_1 & d_2 \end{pmatrix} \begin{pmatrix} x_{11} & x_{12} \dots x_{1n} \\ x_{21} & x_{22} \dots x_{2n} \end{pmatrix}$$

- mit der (hier) 1×2 Projektionsmatrix $\mathbf{D} = \left(\mathbf{d}^T\right)$, allgemein:

$$\mathbf{Y} = \mathbf{D}\,\mathbf{X}$$

- Unterraum (hier eine Dimension) erlaubt Datenreduktion
- *Bester* Unterraum enthält maximale Varianz
  - Richtungsvektor
- Projektion in Unterraum
  - auf Richtungsvektor
  - hier: 2D $\rightarrow$ 1D

- Rekonstruktion mittels Umkehrprojektion
  - aus Richtungsvektor
- Mittelwertskorrektur für direkten Vergleich

3.3.3 Datenreduktion - weitere Dimension(en)

- Beispiel: Schlafdauer
  - Nun drei Variablen: Gehirngewicht, Körpergewicht und Schlafdauer
  - Maximale Varianz?
- Verschiebung: Zentrieren

$$\mathbf{a} := \mathbf{v} - \overline{\mathbf{v}}$$

  - originale Daten $\mathbf{v}$
  - zentrierte Daten $\mathbf{a}$
  - Löst Mittelwert-Problem von vorhin
  - Vereinfacht Berechnungen

$$\mathbf{A} := \mathbf{V} - \overline{\mathbf{v}}$$

3.4 Kovarianzmatrix $\mathrm{Cov}(X_i, X_k)$

$$C_{ik} = \frac{1}{m}\sum_{j=1}^{m}(v_{ij} - \overline{v}_i)(v_{kj} - \overline{v}_k) = \frac{1}{m}\sum_{j=1}^{m} a_{ij}a_{kj}$$

Mittels Datenmatrix $\mathbf{A}$ aus Spalten $\mathbf{a}_i$, so dass

$$\mathbf{C} = \frac{1}{m}\mathbf{A}\mathbf{A}^T$$

Projektion

aller Daten $\mathbf{v}$ auf beliebige Richtung $\mathbf{w}$ ergibt

$$\mathcal{E}(\mathbf{w}\cdot\mathbf{v}) = \frac{1}{m}\sum_{j=1}^{m}\mathbf{w}\cdot\mathbf{v}_j = \mathbf{w}\cdot\sum_{j=1}^{m}\frac{1}{m}\mathbf{v}_j = \mathbf{w}\cdot\overline{\mathbf{v}}$$

und

$$\begin{aligned}\mathrm{Var}(\mathbf{w}\cdot\mathbf{v}) &= \frac{1}{m}\sum_{j=1}^{m}(\mathbf{w}\cdot\mathbf{v}_j - \mathbf{w}\cdot\overline{\mathbf{v}})^2 = \frac{1}{m}\sum_{j=1}^{m}(\mathbf{w}\cdot(\mathbf{v}_j - \overline{\mathbf{v}}))^2 \\ &= \frac{1}{m}\sum_{j=1}^{m}(\mathbf{w}\cdot\mathbf{a}_j)^2 = \frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{n}\sum_{k=1}^{n} w_i a_{ij} a_{kj} w_k \\ &= \frac{1}{m}\mathbf{w}\cdot\mathbf{A}\mathbf{A}^T\mathbf{w} = \mathbf{w}\cdot\mathbf{C}\mathbf{w}\end{aligned}$$

3.5 Singularwertzerlegung

singular value decomposition, SVD

Daten

- $m \times n$ (hier im Bsp. Tiere $\times$ Variablen)
- Unterraum $\mathbf{K} < \mathbf{M}$ mit Hauptkomponenten $i \in [1 \dots K]$
  - Meist $K \ll n$
- Aber nicht nur sortiert sondern auch noch gedreht
- Finde Unterraum mit größter Varianz!

Diagonalisieren der Kovarianzmatrix C

- Wegen Symmetrie existiert
$$\mathbf{C} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$$
- mit orthonormaler Matrix $\mathbf{U}$ und Diagonalmatrix $\mathbf{\Lambda}$
- Dabei sind Eigenvektoren $\mathbf{u}_i$ von $\mathbf{C}$ in den Spalten von $\mathbf{U}$:
$$\mathbf{C}\mathbf{u}_i = \lambda_i \mathbf{u}_i$$
- und Eigenwerte λ_i mit
$$\sigma_i^2 = Var(\mathbf{u}_i \cdot \mathbf{v}) = \mathbf{u}_i \cdot \mathbf{C}\mathbf{u}_i = \mathbf{u}_i \cdot \lambda_i \mathbf{u}_i = \lambda_i$$
- $\Rightarrow$ Die Diagonalmatrix $\mathbf{\Lambda}$ enthält als Eigenwerte die projizierten Varianzen

3.6 Hauptkomponenten

Erste Hauptkomponente

Unterraum mit
$$\max(\text{Var}) = \max(\mathbf{w} \cdot \mathbf{C}\mathbf{w})$$
hat größte Varianz

- in Richtung des (normierten) Eigenvektors $\mathbf{u}_{(\mathbf{1})}$
- zum größten Eigenwert $\lambda_{(1)} = \max(\lambda_i)$

$\Rightarrow$ Erste Hauptkomponente, 1^{st} *principal component*

Beschränkung

- Auf Unterraum orthogonal zur ersten Hauptkomponente
- Wird von restlichen Eigenvektoren aufgespannt (Orthogonalsystem, Symmetrie von $\mathbf{C}$)

Zweite Hauptkomponente

Im Unterraum **ohne** die erste Hauptkomponente entspricht dann die verbleibende maximale Varianz dem

- zweitgrößten Eigenwert $\lambda_{(2)} = \max(\lambda_{i \neq (1)})$
- in Richtung des zugehörigen Eigenvektors $\mathbf{u}_{(\mathbf{2})}$

Und so weiter ...

- Praktischerweise Abschneiden ab $\lambda_{(r)} <$ Schwelle
- Verbleibender Unterraum hat kaum Beitrag zur Varianz
- Abbruchkriterium beispielsweise durch Test auf Signifikanz

3.7 Pipeline PCA

- Daten in Datenmatrix $n \times m$
$$\mathbf{v} = (v_{ij})$$
- Mittelwerte der Variablen $\bar{\mathbf{v}} = \frac{1}{m} \sum_{j=1}^{m} \mathbf{v}_j$ abziehen
$$\mathbf{a} = \mathbf{v} - \bar{\mathbf{v}}$$
- Dann Varianz in den zentrierten Variablen i
$$\mathrm{Var}_i = \frac{1}{m} \sum_{j=1}^{m} a_{ij}^2$$
- Kovarianzmatrix $\mathbf{C}$ aus Datenmatrix $\mathbf{A}$ bestimmen
$$\mathbf{C} = \frac{1}{m} \mathbf{A}\mathbf{A}^T$$
- Diagonalisieren
$$\mathbf{C} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$$
- Sortiere Eigenvektoren $\mathbf{u}_i$ der Größe der Eigenwerte λ_i nach
$$\lambda_{(1)} \geq \lambda_{(2)} \geq \ldots$$

- Abschneiden nach Unterschreiten einer Schwelle für Eigenwerte
- Erste Eigenvektoren spannen Unterraum $\mathbf{U}'$ mit jeweils größtmöglicher Varianz auf
- Hauptkomponenten aus Projektion in $\mathbf{U}'$

$$\boldsymbol{\alpha} = \mathbf{U}'\mathbf{a}$$

- Grenzen der PCA
  - Faktoren → Korrelation
  - de-korreliert → nicht unabhängig

3.8 Beispiele

3.8.1 Beispiel Säugetiere

- Daten
  - $m = 58$ Tiere j: Elefant, ...
  - $n = 3$ Variablen i: BodyWt, BrainWt, SleepTime
- Daten zentrieren: Mittelwerte der Variablen abziehen
- Dann Varianz in den zentrierten Variablen i

$$\mathrm{Var}_i = \frac{1}{m}\sum_{j=1}^{m} a_{ij}^2$$

- Kovarianzmatrix

$$\mathbf{C} = \frac{1}{m}\mathbf{A}\mathbf{A}^T$$

  - Erinnerung: Variable i gegen Variable k über alle Tiere $j = 1 \ldots m$

$$C_{ik} = \frac{1}{m}\sum_{j=1}^{m}(v_{ij} - \overline{v}_i)(v_{kj} - \overline{v}_k) = \frac{1}{m}\sum_{j=1}^{m} a_{ij}a_{kj}$$

- Diagonalisieren
- Sortieren
- Unterraum U' mit zwei größten Eigenwerten in Spalten

3.8.2 Beispiel Bilder

- Bilder als Vektoren
  - Bildvektor $\mathbf{v} = (v_1, v_2, \ldots, v_n)^T \in \mathbb{R}^n$ Pixel für Pixel in einer langen Zeile $1 \ldots n$
  - Mehrere m Bilder $\mathbf{v}_j$
  - Damit Bilderdatenbank
    - $m \times n$ Daten-Array, z.B. $(m = 20 \text{ Bilder}) \times (n = 256 \cdot 256 = 64K \text{ Pixel})$
    - $= 1280K$ Werte v_{ij}
- Projektion
  - Erwartungswert:
$$\mathcal{E}(\mathbf{w} \cdot \mathbf{v}) = \mathbf{w} \cdot \overline{\mathbf{v}}$$
  - Mittelwertsbild:
$$\overline{\mathbf{v}} = \frac{1}{m} \sum_{j=1}^{m} \mathbf{v}_j$$
  - Differenzbilder '*Karrikaturen*':
$$\mathbf{a}_j = \mathbf{v}_j - \bar{\mathbf{v}}$$
  - Varianz in einem Pixel i:
$$\mathrm{Var}_i = \frac{1}{m} \sum_{j=1}^{m} a_i^2$$
  - Varianz unter Projektion
$$\mathrm{Var}(\mathbf{w} \cdot \mathbf{v}_j) = \mathbf{w} \cdot \mathbf{C}\mathbf{w}$$
- Kovarianzmatrix
  - Pixel i vs. Pixel k über alle Bilder $j = 1 \ldots m$
$$C_{ik} = \frac{1}{m} \sum_{j=1}^{m} (v_{ij} - \overline{v}_i)(v_{kj} - \overline{v}_k) = \frac{1}{m} \sum_{j=1}^{m} a_{ij} a_{kj}$$
  - Kann dargestellt werden mittels Datenmatrix $\mathbf{A}$ aus Spalten $\mathbf{a}_i$, so dass
$$\mathbf{C} = \frac{1}{m} \mathbf{A}\mathbf{A}^T$$
  - Aufgabe: Finde Unterraum mit größter Varianz
  - Lösung: Diagonalisieren der Kovarianzmatrix $\mathbf{C}$
- Diagonalisieren
  - Daten $M = 30$ Bilder $\times$ $N = 9$ Pixel.
  - Unterraum $K < M, N$ mit Hauptkomponenten $i = 1 \ldots K$
- Sortieren

  - Sortiere Eigenwerte (Varianzen) der Größe nach

$$\sigma_1 \geq \sigma_2 \geq \dots$$

  - Dann ist
    * $\mathbf{u}_1$ die Richtung größter Variation
    * $\mathbf{u}_2$ die Richtung größter Variation im dazu orthogonalen Unterraum
    * ...
    * $\sigma_m = 0$ (nur wenn Daten zentriert: Verlust eines Freiheitsgrades)
- Abschneiden nach h Dimensionen
  - Das sind die Hauptkomponenten
  - Sie spannen einen Unterraum (Hyperebene) in den Daten auf
  - Beispiel $h = 5$
- Ergebnis Rekonstruktion
  - Aus den wichtigsten Hauptkomponenten lassen sich die Bilder wiederherstellen

3.9 Komponenten

- Projektion auf Unterraum aus Hauptkomponentenvektoren $\mathbf{u}_i$:

$$\boldsymbol{\alpha} = \mathbf{U}'\mathbf{a}$$

  - Gewichtung der Hauptkomponenten(vektoren) im Bild $\mathbf{a}$
  - Koordinaten des Bildes im Unterraum U'
  - Dimension h

3.10 Separation und Interpretation

- Daten sehen nach der PCA separiert aus
- Interpretation: Verschiedene Pixel in den beiden ersten Hauptkomponenten
- PCA-Ergebnisse
  - Im Unterraum der Hauptkomponenten
  - Für $\mathbf{v} = \overline{\mathbf{v}} + \sum_{i=1}^{m-1} \alpha_i \mathbf{u}_i$ gilt

$$\begin{aligned} \mathcal{E}(\alpha_i) &= 0 \\ \mathrm{Var}(\alpha_i) &= \sigma_i^2 \\ \mathrm{Cov}(\alpha_i, \alpha_j) &= 0 \qquad i \neq j \end{aligned}$$

Whitening

Standardisieren in mehreren Dimensionen.

Erinnerung: Momente von Wahrscheinlichkeitsverteilungen

- Nulltes Moment = 1 (Normierung)
- Erstes Moment = 0 (Erwartungswert)
- Zweites Moment = 1 (Varianz)

3.11 Korrelationskoeffizientenmatrix

Anstatt der Kovarianz wird die **Korrelation** verwendet

- $\Rightarrow$ Skalierungsinvariantes Problem
- $\Rightarrow$ SVD *de-korreliert*

SVD auf C anwenden

- Hohe Dimension n
- Eigenwerte sind sehr unterschiedlich
- Rang der Matrix ist $< min(m, n)$

$\Rightarrow$ SVD auf $\mathbf{A}$ anwenden

$$\mathbf{A} = \mathbf{U}\mathbf{W}\mathbf{V}^T$$

$$\mathbf{C} = \frac{1}{m}\mathbf{A}\mathbf{A}^T = \frac{1}{m}\mathbf{U}\mathbf{W}\mathbf{V}^T\mathbf{V}\mathbf{W}\mathbf{U}^T = \frac{1}{m}\mathbf{U}\mathbf{W}^2\mathbf{U}^T$$

- Dann sind die Spalten von $\mathbf{U}$ Eigenvektoren $\mathbf{u_i}$
- und $\sigma_i = \frac{1}{\sqrt{m}} w_i$
- Für eine symmetrische Matrix sind die Eigenvektoren orthogonal

$$\mathbf{u}_i \cdot \mathbf{u}_j = 0 \quad \forall\, i \neq j$$

$$|\mathbf{u}_i| = 1$$

- Sortiere Eigenwerte (Varianzen) der Größe nach

$$\sigma_1 \geq \sigma_2 \geq \dots$$

  - $\mathbf{u_1}$ die Richtung größter Variation
  - $\mathbf{u_2}$ die Richtung größter Variation im dazu orthogonalen Unterraum
  - ...
  - $\sigma_m = 0$ (nur wenn Daten zentriert: Verlust eines Freiheitsgrades)
- Abschneiden nach h Dimensionen

  - Hauptkomponenten spannen einen Unterraum (Hyperebene) in den Daten auf

$$\mathbf{v} = \overline{\mathbf{v}} + \sum_{i=1}^{h} \alpha_i \mathbf{u}_i$$

    mit $h < m - 1$

  - mit den Projektionen

$$\alpha_i = \mathbf{u_i} \cdot \mathbf{a} = \mathbf{u_i} \cdot (\mathbf{v} - \overline{\mathbf{v}})$$

Kumulierte Varianz

$$\sigma^2_{\text{accumulated}} = Var(|\mathbf{v} - \overline{\mathbf{v}}|^2) = \sum_{i=1}^{h} \sigma_i^2$$

- Optimum!

3.12 Python sklearn PCA

- **S**cipy tool**k**it for machine **learn**ing
  - `http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html`
- Methoden:

`fit(X[, y])`	`Fit the model with X.`
`fit_transform(X,[, y])`	`Fit the model with X and apply the dimensionality reduction on X.`
`get_covariance`	`Compute data covariance with the generative model.`
`get_params([deep])`	`Get parameters for this estimator.`
`get_precision()`	`Compute data precision matrix with the generative model.`
`inverse_transform(X)`	`Transform data back to its original space, i.e.,`
`score(X[, y])`	`Return the average log-likelihood of all samples`
`set_params(**params)`	`Set the parameters of this estimator`
`transform(X)`	`Apply the dimensionality reduction on X`

- Daten:
  - Dimension so wählen, wie sie co-variieren sollen
  - `Parameters: X: array-like, shape (n_samples, n_features)`
- PCA:

```
from sklearn.decomposition import PCA
X =  np.array(v.T)
pca = PCA(n_components=6)
pca.fit(v.T)
```

- Formalitäten

  - Transponieren
    * `RuntimeError: we assume data in a is organized with numrows>numcols`
    * ⇒ Daten gegebenenfalls transponieren
    * Ergebnis ist fast gleich
    * Rücktransformation mittels PCA auch transponieren
  - Daten übergeben:
    ```
    DeprecationWarning: Passing 1d arrays as data is
    deprecated in 0.17 and will raise ValueError in 0.19.
    Reshape your data either using X.reshape(-1, 1) if your data has a single
    feature  or X.reshape(1, -1) if it contains a single sample.
    ```

3.13 Bildanalyse Natürlicher Bilder

- Auswertung von vielen 32x32 Pixel Bild-Ausschnitten
  - PC_0: gleichmäßige Fläche hell/dunkel (Mittelwert)
  - PC_1: oben hell, unten dunkel ('Himmel')
  - PC_2: links dunkel, rechts hell (z.B.)
  - PC_3: oben rechts, unten links hell, sonst dunkel
  - ...
  - PC_{m-1}: Rauschen hoher räumlicher Frequenz
- Literatur: Hyvärinen, Hurri, Hoyer: Natural Image Statistics - A Probabilistic Approach to Early Computational Vision. Springer 2009

3.14 Gesichtserkennung und Rekonstruktion

- Anwendung Gesichter
  - Gesichts-Bilder als Trainings-Datensatz
  - Berechnung eines 'Durchschnitts-Gesichts'
  - Sogar ein teilweise überdecktes Gesicht kann wiederhergestellt werden
  - Literatur: Turk, Pentland: Eigenfaces for Recognition **JCogNeurosci Vol3**.1 1991
- 3D Gesichtserkennung
  - Datenbank von 3D-Scans
  - Gesichtsmodell
  - Anpassung
  - Probe

- Identität
- Literatur:
  * Blanz, Vetter: IEEE Transactions on Pattern Analysis and Machine Intelligence **25, 9** 2003
  * Blanz, Vetter: A Morphable Model for the Synthesis of 3D Faces. T. SIGGRAPH'99 Conference Proceedings

4 Independent Components Analysis - ICA

4.1 Cocktailparty Stimm-Separation

4.1.1 Menschen und Computer

- Gesichter auseinanderhalten
  - unterschiedliche Blickwinkel
  - unterschiedliche Beleuchtungssituationen
  - unterschiedliche Gesichtsausdrücke
  - Komponentenzerlegung
  - Speichern und vergleichen
  - Lernfähig, erweiterbar
- Stimmung ablesen
  - Lachen, Trauer, Wut, ...
  - unabhängig von der Person
  - andere Komponente im Gesichts-Raum

4.1.2 Geräusche - Cocktailparty Problem

- Menschen können einzelne Stimmen auseinanderhalten
  - weil sie klassifizieren können
- Andere Lösung
  - Mathematik LGS: so viele Variablen bestimmen, wie (unabhängige) Gleichungen
  - Ein Mikrophon im Raum reicht also nicht
  - Aber zwei für 2 Geräusche usw...
- Mehrere Quellsignale und Mischungen daraus
  - Können wir diese trennen?
  - Suche *unabhängige* Komponenten → 'independent component analysis'

4.2 PCA nicht geeignet

- Originalquellen erscheinen unabhängig, Mischungen allerdings nicht
  - Wir haben nur die Mischungen
  - Wie trennen wir diese?
  - PCA Hauptkomponenten?
- PCA
  - Varianz ist maximiert
  - Kovarianzmatrix ist diagonal
  - Durch PCA gefundene Signale sind dekorreliert
- Keine Lösung!
  - Ursprüngliche Quellsignale nicht gefunden
- unkorreliert $\neq$ unabhängig
  - Nur unkorreliert: PCA funktioniert *nicht*

4.3 Frage: Entmischung

- Beispiel: Zwei Stimmen und zwei Mikrofone
- Abhängigkeit: Die beiden Mischungen sind abhängig
- Statistik: Verteilungen sehen ähnlich aus
- Ziel: Entmischung
  - Unabhängigkeit
  - Komplexität
  - Nicht-Normalverteilt
  - Erinnerung: *Zentraler Grenzwertsatz*

4.4 Zentraler Grenzwertsatz

- Quell-Daten
  - I Quellsignal-Vektoren $\mathbf{s}_i$ der Länge N:

$$\mathbf{s}_i = (s_{i1}, s_{i2}, \ldots, s_{iN})$$

  - zusammen in der Datenmatrix

$$\mathbf{S} = \begin{pmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \\ \vdots \\ \mathbf{s}_I \end{pmatrix}$$

- Mischungs-Daten
  - J Mischungsvektoren $\mathbf{x}_j$ der Länge N:

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_J \end{pmatrix}$$

  - entstanden aus den Quellsignalen

$$\mathbf{x}_j = \mathbf{a}_j \cdot \mathbf{S} = a_{j1}\mathbf{s}_1 + a_{j2}\mathbf{s}_2 + a_{j3}\mathbf{s}_3 + \ldots$$

  - mit den Mischungskoeffizienten $\mathbf{a}_j = (a_1, a_2, a_3, ...)$
  - Hier im Bsp.:

$$\mathbf{x}_1 = a_{11} \cdot \mathbf{s}_1 + a_{12} \cdot \mathbf{s}_2$$
$$\mathbf{x}_2 = a_{21} \cdot \mathbf{s}_1 + a_{22} \cdot \mathbf{s}_2$$

- Matrix-Schreibweise

$$\mathbf{X} = \mathbf{A} \cdot \mathbf{S}$$

  - Dabei hat $\mathbf{A}$ die Dimension $J \times I$
  - die I Quelldaten $\mathbf{S}$ die Dimension $I \times N$
  - und die J Mischdaten $\mathbf{X}$ die Dimension $J \times N$
- Unser Beispiel

$$\mathbf{A} = \begin{pmatrix} 0.4 & 0.9 \\ 0.7 & 0.5 \end{pmatrix}$$

- Problem
  - $\mathbf{A}$ ist leider unbekannt
  - $\mathbf{S_i}$ sind unbekannt: gesucht!

- Lösung?
  - Wüssten wir die Mischungsmatrix $\mathbf{A}$, könnten wir die Entmischungs-Matrix für den Fall $I = J$ berechnen
    $$\mathbf{A}^{-1}$$
  - Wüssten wir die Quelldaten, könnten wir die Entmischungs-Matrix für den Fall $I = J$ berechnen (überbestimmt)
- Anzahl Komponenten / Datensätze
  - Problem bei $J < I$
    - Weniger Mischungs-Datensätze als Quell-Signale lassen sich normalerweise nicht trennen
    - Im Bsp. also: Anzahl der Mikrofone $\geq$ Anzahl der zu extrahierenden Stimmen
  - Praxis: mehr Datensätze $J > I$
    - z.B: im EEG > 10 Elektroden und < 5 Signale
  - I bekannt
    - mittels PCA vorfiltern, um Dimension auf J zu reduzieren
  - I unbekannt
    - Es verbleiben eventuell restliche Dimensionen, die nur noch Rauschen enthalten
    - Lösung:
      - Festlegen einer Schwelle
      - Vorverarbeiten mittels PCA
  - Lösung: Quellsignale erscheinen im Mischungsgraphen als Orientierung (gemeinsame Ursache)
    - Die Richtungen der Quellen s_i sind in der Grafik angedeutet
- Entmischen
  $$\mathbf{S} = \mathbf{W}\mathbf{X}$$
  - mit den Datenreihen der Mischungen $\mathbf{X}_j$ und den (rekonstruierten) Quellen $\mathbf{S}_i$
  - Die Entmischungsmatrix $\mathbf{W}$ hat die Dimension $I \times J$ und enthält die Gewichtung der I Quellsignale in den J Mischdaten
  - Mischung war
    $$\mathbf{X} = \mathbf{A}\mathbf{S}$$
- Ansatz: *Un*-Normal-Verteilung
  - Da Mischungen *Normal*-verteilt(er) sind, suche nach Entmischungen mit
    - möglichst nicht-gauß-förmigen Verteilungen $\mathbf{s}$
    - unterschiedlichen höheren Momenten

Einschub: Momente einer Verteilung

- Die Zufallsvariable X habe die Wahrscheinlichkeitsdichte $f_x(x)$ und sei o.b.d.A. zentriert $\mu = 0$
- Normierung
$$\int_{x=-\infty}^{\infty} f_x(x)\,\mathrm{d}x = 1$$
- erstes Moment: Erwartungswert $\mathcal{E}(X)$
$$\mu = \mathcal{E}\{X\} = \int_{x=-\infty}^{\infty} x \cdot f_x(x)\,\mathrm{d}x = 0$$
- zweites Moment: Varianz $\mathcal{E}\left((X-\mu)^2\right)$
$$\sigma^2 = \mathcal{E}\{X^2\} = \int_{x=-\infty}^{\infty} x^2 \cdot f_x(x)\,\mathrm{d}x$$
- drittes Moment: Schiefe (*Skewness*) $\mathcal{E}\left(\frac{(X-\mu)^3}{\sigma^3}\right)$
$$\mathcal{E}\{X^3\} = \int_{x=-\infty}^{\infty} x^3 \cdot f_x(x)\,\mathrm{d}x$$
  - beschreibt die Asymmetrie der Verteilung von X
- viertes Moment (zentriert):
$$\mathcal{E}\{X^4\} = \int_{-\infty}^{\infty} (x)^4 \cdot f(x)\,\mathrm{d}x \;= m_4$$
  - ergibt die (allgemeine) **Kurtosis**
$$\mathrm{K}(X) = \mathcal{E}\Big(\frac{(X-\mu)^4}{(\sigma^2)^2}\Big) - 3$$
  - beschreibt, wie spitz oder flach die Verteilung verläuft
  - *Normal*verteilung hat Kurtosis $K = 0$
  - spitzere Verteilungen '*super Gauß*' $K > 0$
- Momente einer gemeinsamen Verteilung
  - Das Signal x hat die Verteilung $f_x(x)$, das Signal y hat die Verteilung $f_y(y)$
  - Sind x und y stochastisch unabhängig
    - dann (und nur dann) zerfällt die gemeinsame Verteilung (*joint distribution*) in das Produkt aus den einzelnen Randverteilungen (*marginal distributions*):
$$f_{xy}(x,y) = f_x(x) \cdot f_y(y)$$
  - Kovarianz:

  * beschreibt **lineare** Abhängigkeit der beiden Verteilungen

$$\mathrm{Cov}(x,y) = \mathcal{E}(x \cdot y) = \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} f_x(x) \cdot f_y(y) \cdot x \cdot y \,\mathrm{d}x \,\mathrm{d}y$$

- Unabhängigkeit
  * Allgemein sind zwei Verteilungen unabhängig, wenn *alle* Momente faktorisieren:

$$\mathcal{E}(x^p \cdot y^q) = \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} f_x(x) \cdot f_y(y) \cdot x^p \cdot y^q \,\mathrm{d}x \,\mathrm{d}y = \mathcal{E}(x^p) \cdot \mathcal{E}(y^q)$$

- Kurtosis als Beispiel
  * Signal y der Länge N

$$K = \frac{\frac{1}{N}\sum_{t=1}^{N}(y_t - \bar{y})^4}{\left(\frac{1}{N}\sum_{t=1}^{N}(y_t - \bar{y})^2\right)^2} - 3$$

$$\begin{aligned}\mathcal{E}(x^4) \cdot \mathcal{E}(y^4) &= \mathcal{E}((a_{11}s + a_{12}t)^4) \cdot \mathcal{E}((a_{21}s + a_{22}t)^4) \\ &= c_1 \cdot \mathcal{E}(s^4) \cdot \mathcal{E}(t^4) + c_2 \cdot f(a_i, s, t)\end{aligned}$$

4.5 Projection Pursuit

- Suche Maximum der Kurtosis innerhalb der Mischung
- Durchprobieren aller möglichen Entmischungen
- Stimmt der gefundene Vektor mit der Richtung aus der (uns unbekannten) Mischungsmatrix überein?
  - Erste Richtung gefunden
- Wie weiter?
  - Bestimme unabhängigen Unterraum davon
  - Suche nächstes Maximum der Kurtosis
  - usw.
- Mathematik:
  - Gram Schmidt Orthogonalisierung
- Senkrechter Unterraum
  - Senkrecht = (Kovarianz = 0)
- Ergebnis:
  - Damit wären die Quellsignale $\mathbf{y_1} \sim \mathbf{s_2}$ und $\mathbf{y_2} \sim \mathbf{s_1}$ aus den Mischungen ($\mathbf{x_1}$, $\mathbf{x_2}$) extrahiert
- Weiteres Vorgehen bei mehreren Dimensionen

  - Beschränken auf Unterraum senkrecht zur ersten Komponente
  - Finde darin nächste (=zweite) unabhängige Komponente
  - Beschränken auf Unterraum senkrecht zur ersten und zweiten Komponente
  - Finde darin nächste unabhängige Komponente
  - ...

4.5.1 Zusammenfassung der unabhängigen Komponenten

- Erstelle Raum der Mischsignale
- Suche Projektionsrichtung in den Mischungen, die die Unabhängigkeit maximiert
  - das ist nach dem *Zentralen Grenzwertsatz* die am wenigsten *Normal*-verteilte
  - dafür eignet sich die Kurtosis
  - Rückprojektion entspricht der ersten *unabhängigen* Komponente
- Rekursiv durch Unterräume
  - findet weitere unabhängige Komponenten
  - Abbruch
    * wenn genug Komponenten (Dimension)
    * wenn Schwelle für Kurtosis unterschritten
- Ergebnis: unabhängige Komponenten (extremer) Kurtosis
  - jedoch nicht deren ursprüngliches Verhältnis
- *blind source separation*
  - Separation: zerlegen in ursprüngliche Bestandteile
  - Quellen: Vermutung, dass unabhängige Bestandteile die Quelle bzw. Ursache der Mischungen sind
  - blind: Keine Information über die zugrundeliegenden Quelldaten bekannt
    * Stimmlange
    * Statistische Verteilung
    * *parameterfrei*

4.6 ICA

4.6.1 Unabhängigkeit

- Die *Independent Component Analysis* berechnet im Gegensatz zur *Projection Pursuit* alle Dimensionen parallel
  - Vorteil: robuster
- Unabhängigkeit
  - Die Quellsignale s_i sollen unabhängig sein
  - Alle Quellsignale s_i sollen dieselbe Wahrscheinlichkeitsverteilung $p_s(s)$ haben
  - Innerhalb der Signale sollen die Einzelwerte unabhängig sein
    - (ungeordnet in t, keine versteckte Abhängigkeit)
- Gemeinsam *Normal*-verteilte Variablen
  - ... sind uninteressant für die ICA!
  - Haben nur zweite Momente
    - keine höheren, wie Kurtosis
  - Können maximal *ko*-variant sein
    - per *Whitening* reduzierbar
- Was ist unabhängig?
  - Zufallsvariablen
  - Keine Struktur in der gemeinsamen Wahrscheinlichkeitsverteilungsdichte
  - Bedingte Verteilung = Randverteilung

$$p_X(X=x|Y=y_1) = p_X(X=x)$$

  - gemeinsame Wahrscheinlichkeitsdichte zerfällt in Produkt der einzelnen

$$p_{XY}(X=x, Y=y) = p_X(X=x) \cdot p_Y(Y=y)$$

  - *Alle* Momente zerfallen

$$\mathcal{E}(x^p \cdot y^q) = \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} f_x(x) \cdot f_y(y) \cdot x^p \cdot y^q \,\mathrm{d}x \,\mathrm{d}y = \mathcal{E}(x^p) \cdot \mathcal{E}(y^q)$$

- Gegenbeispiel:
  - Zwei Sinuswellen unterschiedlicher Phase
    - Zerfallen **nicht** in ihre Momente
  - Zwei Sinuswellen unterschiedlicher Frequenz
    - Zerfallen in ihre Momente

4.6.2 Fragestellung der ICA

- Quellsignale $\mathbf{S}$:
  - unabhängig!
  - unbekannt, gesucht
- Mischungsmatrix $\mathbf{A}$:
  - unbekannt, gesucht
  - Mischungsmatrix $\mathbf{A}$ hat aus Quellsignalen $\mathbf{s}$ Mischungen erzeugt $\mathbf{s} \xrightarrow{A} \mathbf{x}$:

$$\mathbf{x} = \mathbf{A}\mathbf{s}$$

- Mischsignale $\mathbf{x}$
  - als einzige bekannt
- Entmischungsmatrix $\mathbf{W}$

$$\mathbf{s} = \mathbf{W}^*\mathbf{x}$$

  - wäre $\mathbf{W}^* = \mathbf{A}^{-1}$ bekannt, ließen sich Quellsignale berechnen
  - unbekannt, gesucht
- Mischung
  - Entstehung der Mischungen $\mathbf{x}$ aus den Quellen $\mathbf{s}$:

$$\mathbf{x} = \mathbf{A}\mathbf{s}$$

  - Gesucht: Umkehrung

$$\mathbf{s} = \mathbf{W}^*\mathbf{x}$$

  - Wahrscheinlichkeitsdichteverteilung von $\mathbf{s}$ ist

$$p_s(\mathbf{x})$$

  - daraus die der Mischungen

$$p_x(\mathbf{x}) = p_s(\mathbf{s}) \left| \frac{\partial \mathbf{s}}{\partial \mathbf{x}} \right| \qquad = p_s(\mathbf{s}) \left| \mathbf{W}^* \right|$$

  - Für nicht-optimale Entmischungsmatrix $\mathbf{W}$ ist die nicht-optimale Lösung $\mathbf{y} = \mathbf{W}\mathbf{x}$ und

$$p_x(\mathbf{x}) = p_s(\mathbf{W}\mathbf{x}) \left| \mathbf{W} \right|$$

- Wenn unabhängig...
  - dann Gesamtwahrscheinlichkeitsverteilung

$$p_s(\mathbf{s}) = \prod_{t=1}^{N} \prod_{i=1}^{I} p_s(s_i(t))$$

4.6.3 Maximum Likelihood

$$p_x(\mathbf{x}) = p_s(\mathbf{W}\mathbf{x})\,|\mathbf{W}| =: L(\mathbf{W})$$

- Likelihoodfunktion L, die es zu maximieren gilt:

$$L(\mathbf{W}) = \prod_{i=1}^{M} p_s(\mathbf{w}_i^T\mathbf{x})\,|\mathbf{W}|$$

- mit unabhängigen Quellsignalen

$$L(\mathbf{W}) = \prod_{i=1}^{M}\prod_{t=1}^{N} p_s(\mathbf{w}_i^T\mathbf{x}^t)\,|\mathbf{W}|$$

- Log-Likelihood zerfällt in Summe

$$l(\mathbf{W}) := \ln L(\mathbf{W}) = \sum_{i=1}^{M}\sum_{t=1}^{N} \ln p_s(\mathbf{w}_i^T\mathbf{x}^t) + N \ln|\mathbf{W}|$$

- Aufgabe:
  - Maximiere l

$$f(\mathbf{W}, p_s, \mathbf{x})$$

- Lösung:
  - 'Modell'-Verteilung: pdf p_s
  - Gradientenmethode: maximieren

4.6.4 Modellvergleich 'cdf matching'

- Wähle Verteilung p_s
- '*cdf-matching*'
  - Durch Anwendung erhält man eine *Gleich*verteilung
    - maximale Unabhängigkeit
    - maximale Komplexität
    - maximale '*Entropie*'
- Beispiele
  - Bilder mit hellen Flecken: schiefe Verteilung
  - Sprachsignale: spitze Verteilung
  - Spitze *super-Gaussian* Verteilung ist $p_s = 1 - \tanh^2(\mathbf{s})$.

4.6.5 Gradientenmethode

- Finde Entmischungsmatrix $\mathbf{W}$, welche Log-Likelihood $l(\mathbf{W}, \mathbf{x})$ unter gegebenen Daten $\mathbf{x}$ maximiert
  - brute force
    - Siehe Beispiel zuvor
- Passend gewählte Verteilungsfunktion
  - *pattern matching*
  - Rechen-Vereinfachung
- Dann ist der Gradient bestimmbar aus
  - Daten-Matrix $\mathbf{x}$
  - (testweiser) Entmischungs-Matrix $\mathbf{W}$
- Und das Optimum der Entmischungsmatrix $\mathbf{W}$ kann mit der Gradientenmethode schrittweise angenähert werden
- Gradientenmethode

$$\frac{1}{N}\, l = \frac{1}{N} \sum_{i=1}^{M} \sum_{t=1}^{N} \ln p_s(w_i^T x^t) + \ln |\mathbf{W}|$$

  - Dazu benötigen wir die Gradientenmatrix mit den Einträgen

$$\frac{\partial l}{N \partial \mathbf{W}_{ij}} = \mathcal{E}\Big(\sum_{i=1}^{M} \frac{\partial \ln g'(y_i)}{\partial \mathbf{W}_{ij}} \Big) + \frac{\partial \ln |\mathbf{W}|}{\partial \mathbf{W}_{ij}} \tag{4.1}$$

  - cdf

$$g$$

  - pdf

$$p_s = g'$$

  - testweise Entmischung

$$\mathbf{y}_i = \mathbf{w}_i^T \mathbf{x}$$

  - *erster* Term
    - Kettenregel

$$\frac{\partial \ln g'(y_i)}{\partial \mathbf{W}_{ij}} = \frac{1}{g'(y_i)} \frac{\partial g'(y_i)}{\partial \mathbf{W}_{ij}}$$

    - $\mathbf{y} = \mathbf{W}\mathbf{x}$

$$\frac{\partial g'(y_i)}{\partial \mathbf{W}_{ij}} = \frac{\partial g'(y_i)}{\partial y_i} \frac{\partial y_i}{\partial \mathbf{W}_{ij}} = g''(y_i) \cdot x_j$$

    - $\Rightarrow$

$$\mathcal{E}\Big(\sum_{i=1}^{M} \frac{\partial \ln g'(y_i)}{\partial \mathbf{W}_{ij}} \Big) = \mathcal{E}\Big(\sum_{i=1}^{M} \frac{g''(y_i)}{g'(y_i)} x_j \Big)$$

  - *zweiter* Term
    * Es gilt
$$\frac{\partial \ln |\mathbf{W}|}{\partial \mathbf{W}_{ij}} = (\mathbf{W}^T)^{-1}{}_{ij}$$
  - Beide Terme eingesetzt in (4.1)
    * Abkürzung $\Psi(y_i) := \frac{g''(y_i)}{g'(y_i)}$
$$\frac{\partial l}{N \partial \mathbf{W}_{ij}} = \mathcal{E}\Big(\sum_{i=1}^{M} \Psi(y_i) x_j\Big) + (\mathbf{W}^T)^{-1}{}_{ij}$$
    * Vektorschreibweise: Jakobi/Gradientenmatrix (Dimension $M \times M$)
$$\nabla \frac{l}{N} = (\mathbf{W}^T)^{-1} + \mathcal{E}\Big(\Psi(\mathbf{y}^t)[x^t]^T\Big)$$
    * Erwartungswert
$$\mathcal{E}\Big(\Psi(\mathbf{y}^t)[x^t]^T\Big) = \frac{1}{N}\sum_{t=1}^{N} \Psi(\mathbf{W}\mathbf{x}^\mathbf{t})[x^t]^T$$
  - Gradientenmethode
$$\begin{aligned} \mathbf{W}_{neu} = & \quad \mathbf{W}_{alt} + \eta \nabla l \\ = & \quad \mathbf{W}_{alt} + \eta \left((\mathbf{W}_\mathbf{alt}{}^T)^{-1} + \tfrac{1}{N} \textstyle\sum_{t=1}^{N} \Psi(\mathbf{W}_\mathbf{alt}\mathbf{x}^\mathbf{t})[x^t]^T \right) \end{aligned}$$
    * mit passend gewählter Schrittweite η.

- Modellverteilung - Beispiel
$$p_s = 1 - \tanh^2(\mathbf{s})$$
  - cdf
$$g(\mathbf{y}^t) = \tanh(\mathbf{y}^t)$$
  - $g' = 1 - \tanh^2$ und $g'' = -2 \tanh g'$
$$\Psi(\mathbf{W}\mathbf{x}^t) = \frac{g''}{g'} = -2 \tanh(\mathbf{W}\mathbf{x}^t)$$

4.6.6 fastICA

- Implementiert eine Art *Newton Iteration*
- Dekorrelieren des neuen angenäherten Unterraums in jedem Schritt
  - Konvergiert quadratisch (oft kubisch)
    * im Vergleich zur Gradientenmethode (=linear)
  - keine Schrittweitenanpassung nötig
  - Verteilungsfunktion g unkritisch

- Literatur: Hyvarinen, Oja: Independent Component Analysis: Algorithms and Applications. Neural Networks, **13(4-5)**, 2000 (pp. 411-430)
- Die Methode *fast*-ICA ist implementiert in
  `http://scikit-learn.org/stable/auto_examples/decomposition/plot_ica_blind_source_separation.html`

4.7 Python sklearn FastICA

- **S**cipy tool**k**it for machine **learn**ing
- `FastICA`

```
from sklearn.decomposition import FastICA  # import
myica = FastICA(n_components=N)            # initialize
myica.fit(X)                               # call: fit the model to X
                                           -or-
Y = myica.fit_transform(X)                 # fit and recover the
                                           sources from X
```

```
Parameter
-----------
X : array-like, shape (n_samples, n_features)
Training data, where n_samples is the number of samples
and n_features is the number of features.

Returns
-----------
X_new : array-like, shape (n_samples, n_components)
```

- Optimale Argumente zu FastICA:

```
fun: (optional)  string or function (value, derivative).
     (Default: 'logcosh')

        def my_g(x):
                return x**3, 3*x**2

whiten: (optional)   boolean
```

- FastICA-Ergebnisse
  - Parameter:
    `components_ : array, shape (n_components, n_features) The unmixing matrix.`
    `mixing_ : array, shape (n_features, n_components) The mixing matrix.`
  - Funktionen:
    `transform(X, y=None, copy=True)`
    `Recover the sources from X (apply the unmixing matrix)`
    `X : array-like, shape(n_samples, n_features) Input Data to transform`
    `copy : bool (optional)`

```
    X_new : array-like, shape (n_samples, n_components) Return value found
sources
```

  - Liste der Mölichkeiten per Autovervollständig
    `ica.<TAB>`
  - Hilfe
    `ica.mixing_?`

4.8 Unabhängige Verteilung

- Ergebnis:
  - $\mathbf{W}$ enthält Entmischungsvektoren $\mathbf{w}_i$

$$\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \dots)^T$$

$$\mathbf{y} = \mathbf{W}\mathbf{x} = \mathbf{s}$$

- Problem-Anpassung
  - Bei komplexen Signalen spielt es durchaus eine wichtige Rolle, welche Verteilung p_s man annimmt
  - Man kann p_s aus Trainings-Daten punktuell schätzen
    - z.B. durch die mittlere Entfernung der nächsten Nachbarn
- Verteilung
  - Randverteilung der Misch- und Quellsignale
  - Verbundwahrscheinlichkeitsverteilung der Mischungen und Quellsignale
    - joint probability density function

4.9 Zusammenfassung ICA

- ICA findet *unabhängige* Signale $\mathbf{s}$ in Daten $\mathbf{x}$, die durch $\mathbf{A}$ linear gemischt wurden

$$\mathbf{x} = \mathbf{A}\mathbf{s}$$

- erlaubt die Zerlegung in *unabhängige* Quellsignale

$$\mathbf{y} = \mathbf{W}^*\mathbf{x}$$

- (maximiert die Entropie)
- Modellverteilung der Quellsignale *cdf-matching*
  - funktioniert auch mit ähnlichen Verteilungen
  - Beispiel *high curtotic* $\text{cdf} = \tanh(x)$

- Entmischungsmatrix z.B. per Gradientenmethode
- Zeitlicher Verlauf in den Daten (Sortierung) spielt keine Rolle
- *blind source separation*
  - keine Einschränkung der Daten
  - keine Modellvorgabe - außer *cdf-matching*
  - kann aus Trainings-Daten gelernt werden
- Einschränkungen
  - ICA hat größeren Rechenaufwand als PCA
    - Vor allem bei hohen Dimensionen
  - Signale dürfen nicht *normalverteilt* sein
    - keinerlei auswertbare Information nach Dekorrelation (PCA)
  - Entmischungsmatrix muss invertierbar sein $N \times N$
    - Ausweg: Pseudo-Inverse
    - Ausweg: Dimensionsreduzierung durch Vorbehandlung der PCA
  - Skalierung und Vorzeichen der Quellsignale bleiben unbestimmt

4.10 Anwendungen

Bilder zu den Anwendungen in den Folien zu *PCA*

zeitliche und räumliche ICA

- zeitlich = tICA
  - M gemischte Signale $\mathbf{x}_i$ der Länge N
  - $\mathbf{x}_i(t)$ Filmbilder mit i=Pixel-Nummer
- räumlich = sICA
  - Bildersammlung $\mathbf{x}^T$
- zeitlich und räumlich = stICA
  - Kombiniert sICA und tICA

Magnetocardiographie

- Literatur: Stone; Independent Component Analysis; MIT press 2004

EEG

- Ausschneiden von Blinzelartefakten und Rauschen
- Zeitlicher Verlauf der Komponenten
- Literatur: Tzyy-Ping Jung & Scott Makeig auf `https://sccn.ucsd.edu/~jung/Site/EEG_artifact_removal.html`

Funktionelle Magnetresonanztomographie (fMRT)

- Vorabscans
- Funktionelle Analyse
- Literatur: Stone; Independent Component Analysis; MIT press 2004

Gesichtserkennung

- Literatur: Draper, Baek, Bartlett, Beveridgea; Recognizing Faces with PCA and ICA; 2003

Natürliche Bilder

- Literatur: Hyvärinen, Hurri, Hoyer; Natural Image Statistics; Springer 2009

5 Bayes-Statistik

5.1 Satz von Bayes & Schlussfolgerung

Beispiel WG

- Mitbewohner
  - Ludger ist penibel
  - Erik ist gutmütig
  - Michael lässt seine Sachen herumliegen, drückt sich um das Müllruntertragen
- Sie kommen nach Hause, die Küche ist ein Saustall ⇒ Wer wars?
- Kalendereintrag
  - Michael ist seit 2 Wochen im Urlaub ⇒ Wer wars?
- Kühlschrank-Notiz
  - von Ludger: 'Sorry, musste dringend weg, mache später sauber' ⇒ Wer wars?

Satz von Bayes

- Thomas Bayes (1702 - 1761)

$$p(A \mid B) = \frac{p(B \mid A)\, p(A)}{p(B)}$$

- verknüpft bedingte Wahrscheinlichkeit $p(A|B)$ zweier Zufallsvariablen A und B
- mit bedingter Wahrscheinlichkeit $p(B|A)$
- Verbundwahrscheinlichkeiten $p(A, B) = p(A|B)p(B)$ und $p(A, B) = p(B|A)p(A)$

	krank	gesund	Summe
Test +	99	4.995	5.094
Test -	1	94.905	94.906
Summe	100	99.900	100.000

$\Rightarrow 99/5.094 = 1{,}9\%$

Beispiel Bluttest

- Treffer-Rate von 99% bei vorliegender Krankheit
- Fehler-Rate von 5% für gesunde Person
- Die Krankheit ist relativ selten in der Bevölkerung: $0,1\%$
- Wie wahrscheinlich ist es, dass man bei positivem Test die Krankheit hat?
- Mathematisch mit der Regel von Bayes

$$
\begin{aligned}
p(\text{Krank}=\text{ja}|\text{Test}=\text{pos}) &= \frac{p(\text{Test}=\text{pos}|\text{Krank}=\text{ja})p(\text{Krank}=\text{ja})}{p(\text{Test}=\text{pos})} \\
&= \frac{p(\text{Test}=\text{pos}|\text{Krank}=\text{ja})p(\text{Krank}=\text{ja})}{p(\text{Test}=\text{pos},\text{Krank}=\text{ja}) + p(\text{Test}=\text{pos},\text{Krank}=\text{nein})} \\
&= \frac{p(\text{Test}=\text{pos}|\text{Krank}=\text{ja})p(\text{Krank}=\text{ja})}{p(\text{Test}=\text{pos}|\text{Krank}=\text{ja})p(\text{Krank}=\text{ja}) + p(\text{Test}=\text{pos}|\text{Krank}=\text{nein})p(\text{Krank}=\text{nein})} \\
p(\text{Krank}=\text{ja}|\text{Test}=\text{pos}) &= \frac{99\% \cdot 0.1\%}{99\% \cdot 0.1\% + 5\% \cdot 99.9\%} \\
&= \frac{0.00099}{0.00099 + 0.04995} = 1.9\%
\end{aligned}
$$

- Ergebnis
  - Reihenuntersuchung ／
  - bei Vorhandensein von Symptomen ✓
- Wiederholung des Tests
  - Sie haben ein positives Testergebnis erhalten, wissen jedoch nun, dass sie dennoch nur zu 1,9% krank sind
  - Sie wiederholen den Test und erhalten ein negatives Ergebnis
    1. Wie sehr beruhigt sie das?
    2. Was wäre im Falle eines (zweiten) positiven Ergebnisses?

Beispiel Haarfarbe & Augenfarbe

- Bedingte Wahrscheinlichkeit: Augenfarbe je Haarfarbe
  - Satz von Bayes:
$$p(H|A) = \frac{p(A|H)p(H)}{p(A)}$$
- Bedingte Wahrscheinlichkeit: Haarfarbe je Augenfarbe

%	schwarz	brünett	rot	blond	Randverteilung
braun	11	20	4	1	37
blau	3	14	3	16	36
nuss	3	9	3	2	16
grün	1	5	2	3	11
Randverteilung	18	48	12	21	100

%	schwarz	brünett	rot	blond
blau	17	30	25	76
	100	100	100	100

%	schwarz	brünett	rot	blond	
blau	8	39	8	45	100

Bayes'sches Schlussfolgern

$$\begin{aligned} p(B|A) &= \frac{p(A|B)p(B)}{p(A)} \\ &= \frac{p(A|B)p(B)}{\sum_{B'} p(A|B')p(B')} \end{aligned}$$

5.2 Bayes Statistik

Was nennt man Bayes Statistik?

- Nicht (nur) *Satz von Bayes*
- Statistische Behandlung von Parametern

Frequentistische Statistik

- Wahrer Parameter θ
- Streuung, Rauschen, Zufall
- Gesetz der großen Zahl, Hauptsatz der Statistik
- Schätzer $\hat{\theta}$, Vertrauensbereich, Konfidenzintervall
- Nullhypothesen-Signifikanztest (NHST)
  - z.B. $T = \frac{\hat{X}-\mu_0}{\hat{S}_X}\sqrt{n} \sim t(n-1)$ unter H_0

Bayes Statistik

- Wahrer Parameter θ
- Wissen über den wahren Parameter (als Verteilung $p(\theta)$)

$$p(\theta \mid \mathrm{D}) = \frac{p(\mathrm{D} \mid \theta)\, p(\theta)}{p(\mathrm{D})}$$

- Bezeichnungen

$$posterior = \frac{likelihood \cdot prior}{evidence}$$

- mit Normierung im Nenner

$$p(D) = \sum_{\theta'} p(\mathrm{D} \mid \theta')\, p(\theta')$$

bzw.

$$p(D) = \int_{\theta'=-\infty}^{\infty} p(\mathrm{D} \mid \theta')\, p(\theta')\, \mathrm{d}\theta' \; = \mathcal{E}(p(\mathrm{D} \mid \theta')\, p(\theta'))$$

Prinzipielles Vorgehen

- Vorwissen, Prior $p(\theta)$
- Messung: Daten D
- Posterior: $p(\theta|D)$
- $\Rightarrow$ neues, verbessertes **Wissen** $p(\theta)$

5.2.1 Reihenfolge der Datenerhebung

- Messungen D_1 und D_2
- Spielt die Reihenfolge eine Rolle?
  - $p(\theta|D_1, D_2)$ und $p(\theta|D_2, D_1)$
- Voraussetzung: Unabhängigkeit der Messungen
  - Likelihood $p(D_1, D_2|\theta) = p(D_1|\theta) \cdot p(D_2|\theta)$
    - $*$ $\Rightarrow$ Reihenfolge spielt keine Rolle für Likelihood
  - Posterior

$$p(\theta|D_1, D_2) = \frac{p(D_1, D_2|\theta)p(\theta)}{\sum_{\theta'} p(D_1, D_2|\theta')p(\theta')} = \frac{p(D_1|\theta)p(D_2|\theta)p(\theta)}{\sum_{\theta'} p(D_1|\theta')p(D_2|\theta)p(\theta')}$$

$$= \frac{p(D_2|\theta)p(D_1|\theta)p(\theta)}{\sum_{\theta'} p(D_2|\theta')p(D_1|\theta)p(\theta')}$$

$$= p(\theta|D_2, D_1)$$

$*$ $\Rightarrow$ Reihenfolge spielt keine Rolle für Posterior

5.3 Dichotome Daten

Exakte mathematische Behandlung am Beispiel dichotomer Daten

- Münzwurf steht stellvertretend für alle Bernoulli-Experimente
  - Sozialwissenschaften: Umfrage ja/nein
  - Biologie: Merkmal vorhanden/nicht vorhanden, Mädchengeburten
  - Physik: Spin up/down
  - Psychophysik: Reiz gesehen/nicht gesehen
  - Medizin: Behandlung wirkt/wirkt nicht
- Eigenschaften
  - genau 2 Möglichkeiten
  - schließen sich gegensätzlich aus
  - benötigen keine Metrik ('größer', 'Abstand')
  - interessierende Größe: jeweilige Häufigkeit
    - Parameter θ

5.3.1 Bernoulli-Experimente

$$Y = \begin{cases} 1 & \text{für Ergebnis 'Kopf'} \\ 0 & \text{für Ergebnis 'nicht Kopf' = 'Zahl'} \end{cases}$$

Wahrscheinlichkeit

$$p(Y) = \begin{cases} \theta & \text{für Ergebnis 'Kopf', y=1} \\ 1-\theta & \text{für Ergebnis 'Zahl', y=0} \end{cases}$$

Bernoulli-Verteilung

$$p(y|\theta) = \theta^y \cdot (1-\theta)^{1-y}$$

Mehrere Würfe

- i.i.d.
- N Wiederholungen
- darunter z mal Kopf

$$p(y_i|\theta) = \theta^{y_i} \cdot (1-\theta)^{1-y_i}$$

- Für ein erhaltenes Ereignis $\mathbf{y} = (y_1, y_2, \ldots, y_N)^T$ unabhängiger Einzelereignisse y_i, darunter die Anzahl z positiver Einzelereignisse, multiplizieren sich die Wahrscheinlichkeiten zur

Binomial-Verteilung

$$\begin{aligned} p(\mathbf{y}|\theta) &= \prod_{i=1}^{N} p(y_i|\theta) \\ &= \prod_{i=1}^{N} \theta^{y_i}(1-\theta)^{(1-y_i)} \\ &= \theta^{\sum_i y_i}(1-\theta)^{\sum_i (1-y_i)} \\ &= \theta^{z}(1-\theta)^{N-z} \end{aligned}$$

Ein Münz-Beispiel

mit 11 diskreten unterschiedlichen Münzen

- mit 11 unterschiedlichen Wahrscheinlichkeiten für *Kopf*:

$$\theta_j \in [0.0,\, 0.1, \ldots 1.0]$$

- Wir haben eine Münze davon, wissen nicht welche
- Experiment: einmaliger Münzwurf
  - Ergebnis: Kopf $y_{i=1} = 1$
- Bei angenommen eher fairem Prior ist das Ergebnis für den Posterior:

thetas	[0.00 0.10 0.20 0.30 0.40 0.50 0.60 0.70 0.80 0.90 1.00]
prior	[0.00 0.04 0.08 0.12 0.16 0.20 0.16 0.12 0.08 0.04 0.00]
likelihood	[0.00 0.10 0.20 0.30 0.40 0.50 0.60 0.70 0.80 0.90 1.00]
posterior	[0.00 0.01 0.03 0.07 0.13 0.20 0.19 0.17 0.13 0.07 0.00]

Kontinuierliche Auswahl an Münzen

- Gibt ähnliches Bild wie diskrete Münzen

5.4 Einflüsse der Beiträge

5.4.1 Einfluss des Stichprobenumfangs (data)

Zwei verschiedene Stichproben-Umfänge (siehe Abbildung 5.1)

- 4 vs. 40 Münzwürfe
- jeweils 25% mal Kopf

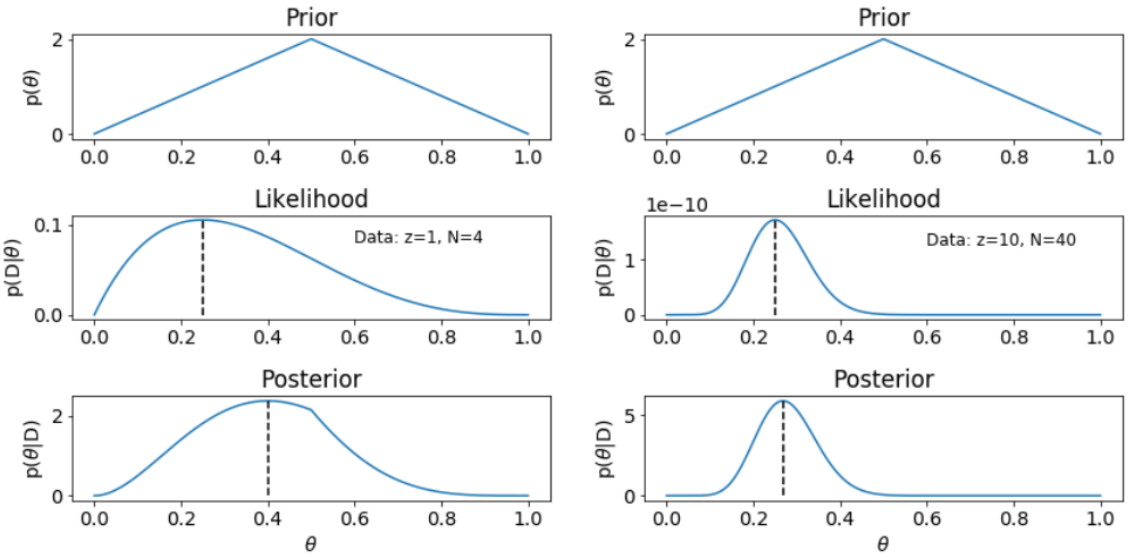

Abbildung 5.1: Einfluss des Stichprobenumfangs

5.4.2 Einfluss des Vorwissens (prior)

Zwei verschiedene Prior-Verteilungen (siehe Abbildung 5.2)

- spitz und flach

5.4.3 Schlussfolgerung

- Prior & Daten ⇒ Posterior
- Prior
  - Vorwissen über den Parameter
  - je schärfer/besser das Vorwissen desto größer sein Einfluss
  - Ausschließen von Möglichkeiten durch Nullsetzen
- Likelihood
  - Daten aus Versuchen *unter der Annahme* eines Parameters
  - je mehr Daten/schärfere Likelihood, desto größer deren Einfluss

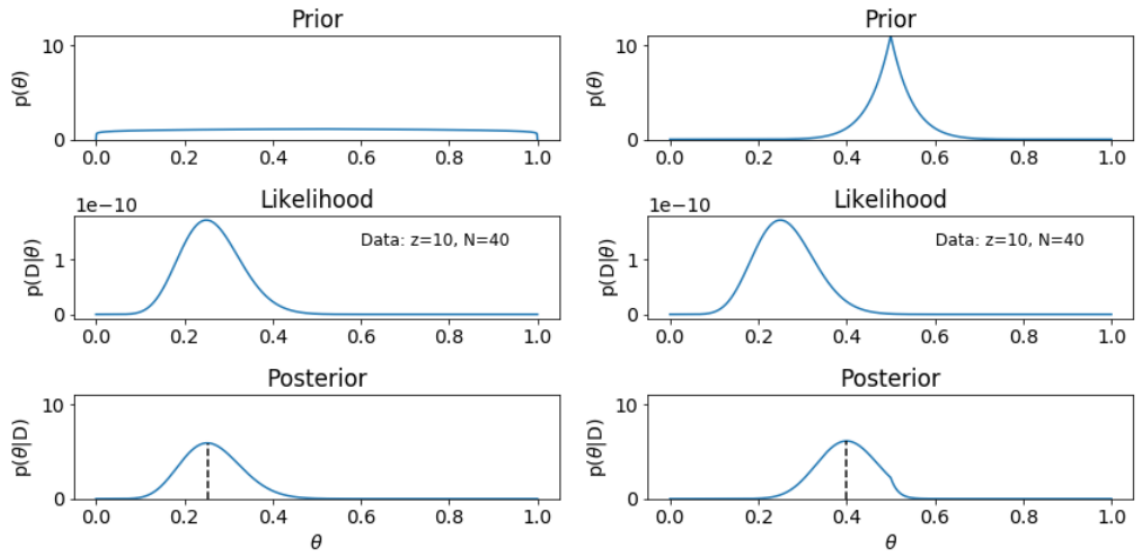

Abbildung 5.2: Einfluss des Vorwissens

- Posterior
  - Neujustierung der Erkenntnis über die Verteilung des Parameters
  - damit über Punktschätzer, Intervallschätzer, ...
  - Kreislauf möglich (Unabhängigkeit der Reihenfolge)

5.5 Parameter

- Theoretische Verteilung des Parameters θ
  - $p(D|\theta)$
  - $p(\theta)$
  - $p(\theta|D)$
- Likelihood = Wahrscheinlichkeit $p(D|\theta)$ für einen Daten-Vektor von N Bernoulli-Experimenten
- Bernoulli-Likelihood

$$p(z|\theta) = \theta^z \cdot (1-\theta)^{1-z}$$

- Modell für Parameter
  - Bereich $\theta \in [0, 1]$
  - Ziel: vor und nach Anwendung der Bayes Schlussfolgerung sollte ähnliche Form von Formel herauskommen
    - *'conjugate prior'*
  - $\Rightarrow$ Potenzen von θ und $(1-\theta)$
- *konjugierte Priors* erlauben
  - geschlossene Formel

  - exakte Berechnung
  - Einbeziehung der Daten (Anzahl)
  - unabhängig von der Datenerfassung (Reihenfolge)
  - Interpretation des Priors als Vorversuche
  - lassen nicht jedes Modell zu; für komplexe Modelle ungeeignet
  - Anderes Beispiel für konjugierte Priors: Gauß-Verteilung

5.6 Beta-Verteilung

$$p(\theta|a,b) = \text{beta}(\theta|a,b) = \theta^{a-1}(1-\theta)^{b-1}/B(a,b)$$

- mit Normierungsfaktor *Beta-Funktion* B

$$B(a,b) = \int_0^1 \theta^{a-1}(1-\theta)^{b-1}\,\mathrm{d}\theta$$

  - $a, b \in \mathbb{R} > 0$
- Beta-Verteilung ist in `Python stats` vordefiniert
- Prior
  - aus Tabelle gewünschte Vorauswahl für θ aussuchen
  - Beispielweise
    * (4,4) für relativ faire Münze
    * (0.1, 2) für Zahl-lastige Münze
    * (1, 1) komplettes *nicht*-Wissen
    * (0.5, 0.5) wenn Ränder wahrscheinlicher als fair sind: Münze aus dem Zauberladen

5.6.1 Eigenschaften

- Erwartungswert

$$\mu = \frac{a}{a+b}$$

- Modus
  - nur möglich für $a > 1$ und $b > 1$

$$\omega = \frac{a-1}{a+b-2}$$

- Streuung
  - nur sinnvoll für $a > 1$ und $b > 1$

$$\sigma = \sqrt{\mu(1-\mu)/(a+b+1)}$$

5.7 Vorwissen und Prior

5.7.1 Festlegen eines Priors gemäß Vorwissen

- Beispielweise aus μ und σ

$$a = \mu\Big(\frac{\mu(1-\mu)}{\sigma^2} - 1\Big)$$
$$b = (1-\mu)\Big(\frac{\mu(1-\mu)}{\sigma^2} - 1\Big)$$

- Beta-Posterior
  - Anwendung der Bayes-Regel mit Prior und Versuchsergebnis

$$\begin{aligned} p(\theta \,|\, z, N) &= \frac{p(z, N \,|\, \theta) p(\theta)}{p(z, N)} \\ &= \theta^z (1-\theta)^{N-z} \theta^{a-1} (1-\theta)^{b-1} \Big/ B(a,b) p(z, N) \\ &= \theta^{z+a-1} (1-\theta)^{N-z+b-1} \Big/ B(a,b) p(z, N) \\ &= \theta^{z+a-1} (1-\theta)^{N-z+b-1} \Big/ B(z+a, N-z+b) \end{aligned}$$

- Ergebnis Posterior

$$p(\theta \,|\, z, N) = \frac{\theta^{z+a-1}(1-\theta)^{N-z+b-1}}{B(z+a, N-z+b)}$$

  - Geschlossene Formel
  - Beta-Ansatz (*c*onjugate prior) erhält Form
    - beliebig erweiterbar: Beta-Verteilung bleibt erhalten
  - Interpretation a, b als vorherige Würfe:
    - Prior mit $a \times$ Kopf und $b \times$ Zahl
- Eigenschaften
  - Posterior-Formparameter bestehen aus Summe aus *Prior* (a, b) und *Daten* $(z, N-z)$
  - besonders praktisch, wenn immer weiter ....
  - Ergebnis für größere N sofort abschätzbar, nicht sukzessiv nötig
  - Reihenfolge der Ergebnisse spielt keine Rolle

5.7.2 Erwartungswert

1. Prior
$$\mu_{\text{Prior}} = \mathcal{E}(\theta) = \frac{a}{a+b}$$
2. Likelihood-Daten
$$\mu_{\text{Daten}} = \frac{z}{N}$$
3. Posterior
$$\begin{aligned}\mu_{\text{Posterior}} &= \frac{z+a}{N+a+b} \\ &= \frac{z}{N}\frac{N}{N+a+b} + \frac{a}{a+b}\frac{a+b}{N+a+b} \\ &= \mu_{\text{Daten}}\frac{N}{N+a+b} + \mu_{\text{Prior}}\frac{a+b}{N+a+b}\end{aligned}$$

Ergebnis Erwartungswert

$$\mu_{\text{Posterior}} \in \left[\mu_{\text{Daten}} \dots \mu_{\text{Prior}}\right]$$

- gewichtet mit den relativen Mengen-Verhältnissen
  - mehr Daten (N): geringeres Gewicht des Priors
  - stärkerer Prior (a+b): geringeres Gewicht der Daten
- a und b des Priors repräsentieren den Ausgang und die Anzahl der Vorversuche

5.7.3 Koordinatentransformation

- Treffer/Gesamtzahl
  - z als Anzahl von N:
$$a = z + 1 \qquad b = N - z + 1$$
  - z' als Anteil von N:
$$a = Nz' + 1 \qquad b = N(1 - z') + 1$$
- Verhältnis / Standardabweichung
  - mit der Standardabweichung (sinnvoll für $s > 0.289$ bzw. $a, b \geq 1$)
$$a = \mu(\frac{\mu(1-\mu)}{s^2} - 1)$$
$$b = (1-\mu)(\frac{\mu(1-\mu)}{s^2} - 1)$$

5.7.4 Vorwissen im Prior

1. Kein Vorwissen
   - Keine Vorversuche $N = 0$ und $z = 0$: $p(\theta|a,b) = \text{beta}(\theta|a,b) = \theta^{a-1}(1-\theta)^{b-1}/B(a,b)$
     - $a = 1,\ b = 1$
     - flache Wahrscheinlichkeitsverteilung $p(\theta) = 1$
     - Versuchsergebnis (Daten) bestimmen alleine den Posterior
2. Starkes Vorwissen
   - Münze ist neu, direkt aus der Prägeanstalt
     - $\theta = 0,5$ - als ob bereits 100 mal geworfen: $a = 51,\ b = 51$
3. Schwaches Vorwissen
   - Münze ist zweifelhaft
     - $\theta = 0,75$ - breite Verteilung: $a = 3,\ b = 7$ (2 von 8)

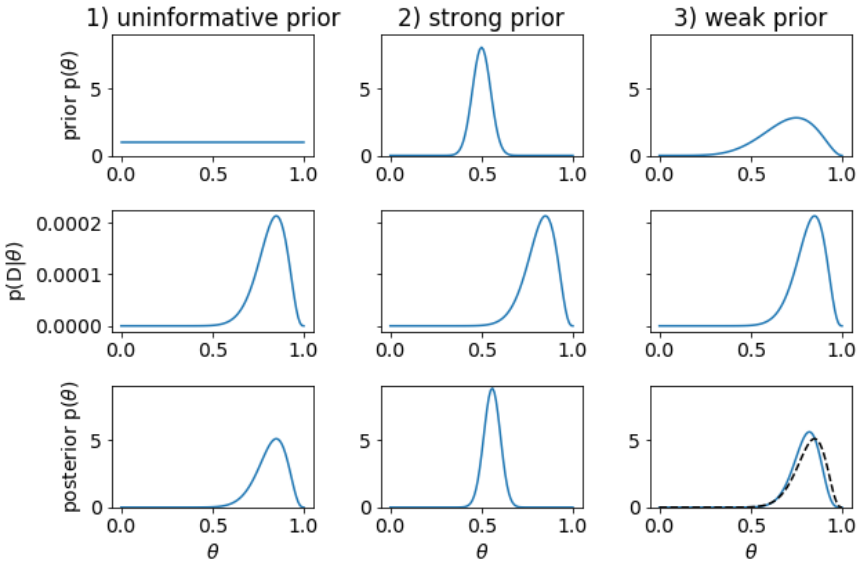

Abbildung 5.3: Vorwissen im Prior

5.8 Grenzen der Methode *conjugate priors*

bisher:

- geschlossen lösbar
- Lösung sofort zugänglich
- reichhaltige Auswahl an Priors

Speziell: Prior mono-modal/bi-modal (aufgrund Beta-Verteilung)

- Es gibt nur einen zentralen Peak
- oder zwei fest am Rand an 0 und 1

5.8.1 Beispiel: zwei Trick-Münzen aus der Spiegelgasse

- Entweder Kopf zu 25%
- oder Zahl zu 25%

Damit Posterior

- mehr Wahrscheinlichkeit Richtung 75%

Das ist jedoch eine unrealistische Lösung:

- 50% häufiger als 25%?
- Tal bei 30% und 70%
- entspricht nicht dem Modell

5.9 MCMC

5.9.1 Einleitung

Posterior-Zufallsstichproben am Beispiel dichotomer Daten

Ziel

- Stichproben aus dem Posterior gewinnen
- Statistik: Punktschätzer, Intervallschätzer

Problem

- Mathematisch geschlossene Lösung
  - nicht immer anwendbar
- Numerische Gitter-Berechnung
  - Bayes Schlussfolgern erlaubt Berechnung des Posteriors
  - Benötigt jedoch das Integral *evidence* über alle Parameterkombinationen bei numerischer Berechnung mit hinreichend vielen Stützstellen
  - für komplexe Modelle nicht in endlicher Zeit berechenbar (z.B. 100^{100})

Lösung

- Markov Chain Monte Carlo Methode *MCMC*
  - Stichproben aus Posterior-Verteilung in einer Markov-Kette
  - Daraus Erwartungswert und *Credible Interval* schätzen

Markov Chain Monte Carlo Methode

Vorgehensweise

- (zufällige aber zielgerichtete und repräsentative) Stichprobe (nicht das vollständige Gitter)
  - 'Monte Carlo'
  - gemäß der Posterior-Verteilung(!)

Voraussetzungen MCMC

- Berechenbarkeit des Priors $p(\theta)$
  - für jeden Parameter $\theta \in \mathbb{R}$
- Berechenbarkeit der Likelihood $p(D|\theta)$
  - für jedes Datum D und jeden Parameter θ

Vorteil

- Dafür kann die evidence (Normierung, Skalierungsfaktor) übergangen werden

Ergebnis

- Posterior $p(\theta|D)$ Stichproben
- Daraus Abschätzen
  - Erwartungswert *Mean Posterior*
  - Modus *Maximal A Posterior* (MAP)
  - *Credibility-Intervall* (CI) oder *Highest Density Interval* (HDI)
- Beachten
  - Keine p sollten exakt Null sein, sonst Zu-/Durchgang schwierig
- Beweis
  - Transformationsmatrix ist unter der Ziel-Verteilung stabil

5.9.2 Vereinfachter Metropolis-Algorithmus

Eigenschaften

- diskrete Möglichkeiten
- eine Dimension für Parameter θ
- konstante Schrittweite: 1

Beispiel: Ein Versicherungsvertreter möchte auf einer Kette von Inseln Kunden gleich häufig besuchen

- Jede Insel hat einen Bevölkerungsanteil θ_i
- Per Blick ist abends abzuschätzen, wie viel höher der Bevölkerungsanteil auf einer der beiden Nachbarinseln ($i-1$ bzw. $i+1$) ist
- Am Morgen wird (wenn lohnenswert) eine Nachbarinsel besucht

1. Richtungsentscheidung
   - Wähle zufällig mit $p = 0.50$ rechte oder linke Insel / kleineren oder größeren Index des Parametervektors θ aus
   - $\Rightarrow$ Dies liefert den *Kandidaten*

$$\theta_{\text{Kandidat}} = \begin{cases} \theta[i_{\text{aktuell}} - 1] \\ \theta[i_{\text{aktuell}} + 1] \end{cases}$$

2. Sprung-Wahrscheinlichkeit

   a) Wenn $p(\theta_{\text{Kandidat}}) > p(\theta_{\text{aktuell}})$ dann gehe zu Kandidat

$$q_{Sprung} = 1$$

   b) Wenn $p(\theta_{\text{aktuell}}) \geq p(\theta_{\text{Kandidat}})$ dann gehe proportional zum Wahrscheinlichkeits(dichte)-Verhältnis zum Kandidaten

$$q_{Sprung} = \frac{p(\theta_{\text{Kandidat}})}{p(\theta_{\text{aktuell}})}$$

c) sonst bleibe

$$q_{Sprung} = 1 - \frac{p(\theta_{\text{Kandidat}})}{p(\theta_{\text{aktuell}})}$$

3. Keine Berechnung der *evidence* nötig

Ergebnis:

- Kette von Sprüngen 'Markov Chain'
- Wahrscheinlichkeit des Aufenthaltes ≡ Wahrscheinlichkeit der Punkte
- funktioniert
  - Es gibt einen stabilen Zustand
  - Dieser stabile Zustand repräsentiert die Verteilung
- Stichprobe aus der Posterior-Verteilung
  - für den Parameter θ
  - (nicht für Daten!)
- Auswerten der Posterior-Verteilung
  - Erwartungswert *Mean Posterior*
    - Mittelwert
  - Modus *Maximum A Posteriority* (MAP)
    - Histogramm
    - KDE
    - Modell-Anpassung
  - *Credible Interval* (CI)

$$p(\theta \in \text{CI}) >= 1 - \alpha$$

Vergleich NHST

- Verteilung von Daten unter Parameter
- Likelihood
- Punktschätzer: Maximum Likelihood Estimator (MLE)
- Intervallschätzer: Konfidenzintervall

$$\hat{\mu} - t_{1-\alpha/2}(n-1)\frac{\hat{\sigma}}{\sqrt{n}} \;\leq\; \mu \;\leq\; \hat{\mu} + t_{1-\alpha/2}(n-1)\frac{\hat{\sigma}}{\sqrt{n}}$$

  - überdeckt zu $1 - \alpha$ den wahren Parameter
  - Voraussetzung: Normalverteilte Stichprobe
    - Punktschätzer Erwartungswert

$$\hat{\mu} = \overline{x}$$

- Punktschätzer Varianz

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

5.9.3 Kontinuierlicher Metropolis-Algorithmus

Metropolis-Algorithmus (nach Metropolis, Rosenbluth, Rosenbluth, Teller & Teller 1953)

Metropolis	**vereinfachter**	**kontinuierlicher**
Dimension	$\mathbb{R}$	$\mathbb{R}$
Ziele	$\theta_i;\ i \in \mathbb{N}$	$\theta \in \mathbb{R}$
Schrittweite	$\Delta i \in [-1, 0, +1]$	$d \sim \mathcal{N}(0, \sigma^2) \in \mathbb{R}$

Voraussetzung

- Verteilung $p(\theta)$
  - berechenbar $\forall \theta_i$
  - keine Bereiche mit $p \equiv 0$

Durchführung

1. Sprungweite und -richtung

$$\Delta\theta \sim \mathcal{N}(\mu = 0, \sigma^2)$$
$$\theta_{\text{new}} = \theta_{\text{cur}} + \Delta\theta$$

2. Wahrscheinlichkeit dafür

$$p_{\text{move}} = \min\left(1, \frac{P(\theta_{\text{new}})}{P(\theta_{\text{cur}})}\right) = \min\left(1, \frac{p(D|\theta_{\text{new}})p(\theta_{\text{new}})}{p(D|\theta_{\text{cur}})p(\theta_{\text{cur}})}\right)$$

   - und $p_{\text{move}} = 0$ wenn außerhalb des erlaubten Parameter-Bereichs

3. Verteilung
   - beispielsweise

$$p(D \,|\, \theta_{\text{x}}) = \text{Bernoulli}(z, N \,|\, \theta_{\text{x}})$$
$$p(\theta_{\text{x}}) = \text{beta}(\theta_{\text{x}} \,|\, a, b)$$

4. Abbruch-Kriterium
   - Wenn genügend unabhängige Samples vorliegen

Verschiedene Sprungweitenverteilungen

- $\sigma \in [0.02; 0.2; 2.0]$
- unwissender Prior $\text{beta}(\theta|1, 1)$
- Daten $N = 20$, $z = 14$

Ergebnis

- Metropolis Algorithmus funktioniert meistens
- Schrittweite ist manchmal kritisch
  - z.B. Verteilung mit einem schmalen und einem breiten Peak
- Näherung an Posterior gut
  - immer noch nicht perfekt; mit 50.000 Grid-Berechnungen wäre die Genauigkeit viel besser gewesen
- im N-dimensionalen sieht das ganz anders aus

5.9.4 Mehrdimensionaler Metropolis-Algorithmus

Fragestellung

- Medizin: zwei Gruppen Patienten, eine bekommt Plazebo, eine ein neues Medikament
  - Wirkt das Medikament
- Verhaltensforschung: zwei Gruppen Versuchspersonen, eine spielt Ballspiele, eine spielt Tetris
  - Auswirkung auf Konzentrationstest gelöst/nicht gelöst?
- Münzwurf: zwei verschiedene Münzen mit θ_1 und θ_2

Voraussetzungen

- Unabhängigkeit der Daten (wie bisher)
- Unabhängigkeit der Parameter θ_1 und θ_2

Metropolis	**vereinfachter**	**verallgemeinerter**
Dimension	$\mathbb{R}^1$	$\mathbb{R}^N$
Ziele	$\theta_i;\ i \in \mathbb{N}$	$\theta \in \mathbb{R}^N$
Schrittweite	$\Delta i \in [-1, 0, +1]$	$\mathbf{d} \sim mv\mathcal{N}(\mathbf{0}, \boldsymbol{\sigma}^2) \in \mathbb{R}^N$

Gemeinsame Verteilung

- Aus Unabhängigkeit folgt
$$p(\theta_1, \theta_2) = p(\theta_1)p(\theta_2)$$
- Normierung
$$\int_{\theta_1}\int_{\theta_2} p(\theta_1, \theta_2)\ \mathrm{d}\theta_2\mathrm{d}\theta_1 = 1$$

Daten

- Tupel gemeinsamer Merkmale (2 Münzen, 2 Gruppen,...)
  - je Versuchsdurchführung
  - z.B. 32/48 und 24/52

- (Nicht unbedingt paarweise erhoben
  - zwei Münzen gleichzeitig geworfen
  - Erinnerung: t-Test unabhängig oder gepaart)
- Unabhängigkeit bedeutet:
  - $p(y_1 \,|\, \theta_1, \theta_2) = p(y_1 \,|\, \theta_1)$ und $p(y_2 \,|\, \theta_1, \theta_2) = p(y_2 \,|\, \theta_2)$
- Mehrfache Durchführung
  - $z_1 = \sum_{i=1}^{N_1} y_{1i}$ und $z_2 = \sum_{i=1}^{N_2} y_{2i}$
- Datensatz
  - $D = \{z_1, N_1, z_2, N_2\}$
- Likelihood

$$\begin{aligned} p(D \,|\, \theta_1, \theta_2) &= \prod_{y_{1i} \in D_1} p(y_{1i} \,|\, \theta_1, \theta_2) \prod_{y_{2j} \in D_2} p(y_{2j} \,|\, \theta_1, \theta_2) \\ &= \theta_1^{z_1} (1-\theta_1)^{N_1 - z_1} \theta_2^{z_2} (1-\theta_2)^{N_2 - z_2} \end{aligned}$$

Bayes Schlussfolgerung

- Posterior

$$\begin{aligned} p(\theta_1, \theta_2 \,|\, D) &= p(D \,|\, \theta_1, \theta_2) p(\theta_1, \theta_2) \Big/ p(D) \\ &= p(D \,|\, \theta_1, \theta_2) p(\theta_1, \theta_2) \Big/ \int_{\theta_1'} \int_{\theta_2'} p(D \,|\, \theta_1', \theta_2') p(\theta_1', \theta_2') \, \mathrm{d}\theta_1' \mathrm{d}\theta_2' \end{aligned}$$

Exakte Lösung unter *conjugate Prior* Beta-Verteilung

- Posterior

$$\begin{aligned} p(\theta_1, \theta_2 \,|\, D) &= p(D \,|\, \theta_1, \theta_2) p(\theta_1, \theta_2) \Big/ p(D) \\ &= \frac{\theta_1^{z_1} (1-\theta_1)^{N_1 - z_1} \theta_2^{z_2} (1-\theta_2)^{N_2 - z_2} \theta_1^{a_1 - 1} (1-\theta_1)^{b_1 - 1} \theta_2^{a_2 - 1} (1-\theta_2)^{b_2 - 1}}{p(D) \; B(a_1, b_1) B(a_2, b_2)} \end{aligned}$$

- mit, wegen Normierung

$$p(D) \; B(a_1, b_1) B(a_2, b_2) = B(z_1 + a_1, N_1 - z_1 + b_1) B(z_2 + a_2, N_2 - z_2 + b_2)$$

- also für den Posterior wieder eine Beta-Verteilung

Ergebnis

- Analog zum eindimensionalen Fall

5.10 Gibbs Sampling

Metropolis Algorithmus 2D

- analog zum eindimensionalen Fall
  - bivariate Normalverteilung zur Sprungvorhersage
  - kann Kovarianz haben, wenn Daten korreliert
- Problem
  - Enge Verteilungen werden schlecht erreicht

Gibbs Sampling

- Autoren: Geman & Geman 1984
- Bekannt nach dem Physiker J. W. Gibbs: Statistische Mechanik und Thermodynamik
- Spezialfall des Metropolis-Hastings-Algorithmus

Gemeinsamkeiten

- Random Walk
- Markov Chain (unabhängige Vorgeschichte)

Unterschied

- Jeder Schritt nur entlang eines Parameters, anderer fest; meist zyklisch
  - Bedingte Wahrscheinlichkeitsverteilung $p(\theta_i|\{\theta_{j\neq i}\}, D)$
  - Form bekannt $\Rightarrow$ direkte Zufallsauswahl
  - Anspringen (es entfällt kein Schritt)

Vorteil

- Anwendbar, wenn gesamte Verbundwahrscheinlichkeit nicht bestimmt werden kann $p(\{\theta_i\}|D)$,
  - lediglich die bedingte Wahrscheinlichkeit $p(\theta_i|\{\theta_{j\neq i}\}, D)$
  - diese dafür bekannt
- effektiver, da (fast) alle Schritte zählen
- gleiches Ergebnis im Limes

Einschränkung

- nicht anwendbar, wenn bedingte Wahrscheinlichkeit nicht bestimmbar
- oder keine Zufallszahlen daraus gezogen werden können

Nachteile Metropolis/Gibbs

- Ausgehend von aktuellem Parametervektor mit symmetrischer Sprungvorhersage
  - Multimodale Verteilung ineffektiv abgetastet
  - Korrelierte Parameter ineffektiv erreichbar

  - Schwänze der Verteilung beibehalten
- Schlechte Konvergenz
  - Parameter dürfen nicht zu sehr korrelieren
    * sonst zu kleine Schritte, nicht in Diagonalrichtung möglich

Literatur

- Gelfand & Smith 1990
- McGrayne 2011
- Bolstad 2009
  - Mathematik zum Metropolis-Hastings Algorithmus

5.10.1 BUGS

- **B**ayesian-inference **U**sing **G**ibbs **S**ampling
- Erste weit verbreitete Implementierung
  - *Winbugs* 1997
- Heute `openBUGS`: `http://www.openbugs.net/w/FrontPage`

5.10.2 JAGS

- **J**ust **A**nother **G**ibbs **S**ampler
- Auch für Mac, Linux, Unix, ...
- 2003

5.10.3 emcee

- pure-Python implementation of *Goodman & Weare's Affine Invariant Markov chain Monte Carlo (MCMC) Ensemble sampler*
- `http://dan.iel.fm/emcee/current/`

5.11 Hamilton HMC

Adaptiert an Form der Verteilung

- Sprungziel-Wahrscheinlichkeit in Richtung des Modus erhöht: Gradient
- Weite der Sprünge an Breite/Höhe der Verteilung angepasst

Sprungziel

- Anleihe aus der Physik: Hamilton-Operator
  - Potential $-log(posterior)$
  - Impuls
  - Bewegung
  - Stopp nach bestimmter Zeit

Sprung-Entscheidung

$$p_{\text{accept}} = \min\Big(1,\ \frac{p(\theta_{\text{proposed}}|D)p(\phi_{\text{proposed}})}{p(\theta_{\text{current}}|D)p(\phi_{\text{current}})}\Big)$$

- ähnlich wie Metropolis-Algorithmus, jedoch gewichtet gemäß dem Impuls *phi*
  - erwartet wird (Energieerhaltungssatz) ein Verhältnis von $p_{\text{accept}} = 1$
  - zufällige Abweichungen durch Diskretisierung des Bewegungswegs

Ergebnis

- HMC bildet eine Makrov-Kette
  - Kleine Schrittweite *phi* $\Rightarrow$ gute Näherung
  - Große Schrittweite $\Rightarrow$ grobe Annäherung
- Anpassen an Posterior
  - Schrittweite ϵ und Schrittanzahl
- Anpassen an Statistik
  - 65% Akzeptanz-Rate hat sich empirisch als sinnvoll herausgestellt
  - *Burn in*-Schritt zu Beginn der Adaption

Problem

- Wenn Schrittlänge zu lang ist

5.12 NUTS

- **N**o **U**-**T**urn **S**ampler
- Nachteil aller *random walks*
  - Um Entfernung D zurückzulegen braucht es D^2 Schritte
  - Am Ende eines Bereichs
    - viele Samples
    - dreht um, also *negativer* Fortschritt (Wiederholung)
- Also
  - definiere Umkehrpunkt als einen iterativen Schritt mit Komponente auf Start zu → Stop
- Literatur
  - Hoffman, Gelman: `https://arxiv.org/abs/1111.4246`
  - `https://www.youtube.com/watch?v=oMNXRYRNj_M`
- Implementierung: Stan
  - Siehe Kapitel 5.14 auf Seite 111

5.13 Ziele eines guten Samples

1. Repräsentative Samples
   - Dafür gibt es keinen 100% Test
   - Pfad der Kette anschauen *trace plot*
     - Abhängigkeit vom Startwert?
     - Verwaiste Pfade?
     - Nebenmaxima?
     - gleiche Konvergenz mit anderen Zufallszahlen?
   - *Burn in* abschneiden
   - Gut sind
     - im *trace plot* überlappende Pfade ohne Abzweige
       - Konvergenz → Überlapp $\not\Rightarrow$ Überlapp → Konvergenz
     - Überlappender *density plot* (zur Glättung) in allen späteren Teilpfaden
     - ANOVA unterschiedlicher Ketten
       - *shrink factor* (*Gelman Ruby factor*) kleiner als 1,1
2. Ausreichende Länge

- Genauigkeit und Stabilität für MLE und HDI
- Messen der Klumpenbildung durch Autokorrelation
  - große Werte nahe 1 für kleine Lags bedeuten Klumpenbildung
  - Schritte sind nicht unabhängig, effektive Kettenlänge kürzer als Gesamtlänge
  - *effective sample size* (ESS) bewertet Autokorrelation
  - ESS muss nicht so hoch sein für Mean/MAP
  - ESS muss hoch sein für HDI, da an der Grenze ja per Definition selten samples sind
    - Daumenregel: 10.000 effektive Samples für das 95% HDI
  - *Monte Carlo Standard Error* (*MSCE*) analog zum Mittelwertsfehler

$$\text{MCSE} = \frac{SD}{\sqrt{ESS}}$$

3. Effizienz
   - in endlicher Zeit berechenbar
     - wir haben z.B. eine ganze Woche Rechenzeit für Priors auf Cluster mit 4 cuda-Karten
   - Hinweise:
     - Parallelisierung auf CPU-Kerne
     - Gibbs statt Metropolis Sampler (wenn angemessen)
     - Hamiltonian (wenn angemessen)
     - Umparametrisieren des Modells zur Korrelationsvermeidung
       - Beispiel: unabhängige μ und δ anstatt $\alpha = \mu + \delta$ und $\beta = \mu - \delta$

5.14 Stan

Ziel:

- Stichprobe aus dem Posterior gewinnen
- Daraus Statistik
  - Punktschätzer
  - Intervallschätzer

Vorgehensweise:

- Markov Chain Monte Carlo Methode *MCMC*: Hamiltonian Monte Carlo
  - *no U-Turn sampler* (NUTS)
- effektive Stichproben aus Posterior-Verteilung
  - umgeht Problem der Ineffektivität (Korrelation, Schwänze der Verteilung, Mehrfachberechnung)
  - adaptiert an Verteilung (unterschiedliche Form)

Der Name *Stan*

- Benannt nach Stanislaw Ulam, einem der Entwickler der Monte Carlo Methode in den 1940ern
- 'Sampling Through Adaptive Neighborhoods'

Link und Literatur

- `http://mc-stan.org/`
- `http://mc-stan.org/users/documentation/index.html`

5.14.1 PyStan

Webseiten

- `https://pystan.readthedocs.io/en/latest/`
- `http://mc-stan.org/users/interfaces/pystan.html`
- `https://pypi.python.org/pypi/pystan`

Voraussetzungen:

- Compiler (Betriebssystem): `gcc, gcc-c++`
- C für Python (conda): `cython-0.25.2`

Verwendung

- `import pystan`

Modell-String bildet kompletten Modellaufbau ab

- `data`; optional `transposed data`
- `parameters`; optional `transposed parameters`
- `model`

Daten

- als `dictionary`
- Skalar, Vektor, Matrix

Initialisieren und Kompilieren

- `stanmodel = pystan.StanModel(model_code=..., ...)`

Berechnen der Posterior-Markovkette

- `fit = stanmodel.sampling(data=..., iter=..., warmup=..., chains=..., n_jobs=..., ...)`

Ergebnis Posterior aller Parameter

- Statistik mit `print(fit)`
- Graphiken mit `fit.plot(['param1', 'param2', ...])`
- Numerisch mit `fit.extract()`

usw.

- `fit.<TAB>`
- `help(fit)`

5.15 PyStan-Beispiele

5.15.1 Eine Münze

Modell

Prior-Verteilung $p(\theta)$

- Beta-Verteilung

$$\theta \sim \mathrm{Beta}(a, b)$$

- mit normalisierender Beta-Funktion

$$p(\theta \mid a, b) = \theta^{a-1}(1-\theta)^{b-1} \;/\; \mathrm{B}(a, b)$$

Daten:

$$z, N$$

Likelihood-Funktion:

$$p(D|\theta)$$

- Bernoulli

$$y_i \sim \text{Bernoulli}(\theta)$$
$$p(\mathbf{y} \mid \theta) = \theta^z (1-\theta)^{N-z}$$

Graphische Modellbeschreibung

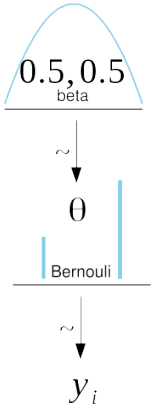

Abbildung 5.4: Im Beispiel hängt das Ergebnis des Münzwurfs vom Bernoulli-Parameter θ ab, dieser entstamme einer Beta-Verteilung

Modell als String

wird durch (Py)Stan interpretiert:

```
mycoinmodel = """
data {
        int<lower=0> Ntotal;            // number of tosses
        int y[Ntotal];                  // data 0=tails, 1=heads;
                                           has to be supplied
}
parameters {
        real<lower=0, upper=1> theta;   // the (restricted)
                                           parameter of interest
}
model {
        theta ~ beta( 0.5, 0.5 );       // prior for parameter;
                                           a more tricky one
        y ~ bernoulli( theta );         // vectorized likelihood for data
                                           built-in bernoulli
}             """
```

Optional im String

- `transformed parameters`

- `transformed data`
- `generated quantities`

Implementierte Verteilungen

Tabelle 5.1: Vorinstallierte Verteilungen in PyStan

`bernoulli`	`bernoulli_logit`	`beta`	`beta_binomial`
`binomial`	`binomial_logit`	`cauchy`	`chi_square`
`exponential`	`gamma`	`logistic`	`multinomial`
`multi_normal`	`normal`	`pareto`	`student_t`
`uniform`	`...`		

Prior

- a und b der Beta-Verteilung

Daten und Likelihood

- z.B. aus `pandas dataframes`
- Daten in einem `dictionary`
  - Namen müssen wie im Modell-String sein

Daten

als `dictionary`

- Namen gemäß Modell-String
- Werte: Array, Skalar

PyStan Aufruf

- `fit1 = pystan.stan(model_code=mycoinmodel, data=mydata, iter=1000, chains=4)`
- Kompilieren → Berechnung → Auswertung
  - `print(fit1)`:

```
Inference for Stan model: anon_model_472cfc0b457697b5a983d1c.
4 chains, each with iter=1000; warmup=500; thin=1;
post-warmup draws per chain=500, total post-warmup draws=2000.

        mean se_mean     sd   2.5%    50%  97.5%  n_eff   Rhat
theta    0.3  2.6e-3   0.07   0.19    0.3   0.45    678    1.0
lp__   -31.88    0.03   0.74 -34.12 -31.59 -31.32    689    1.0

Samples were drawn using NUTS at Thu Feb 15 12:09:04 2018.
For each parameter, n_eff is a crude measure of effective sample
size,and Rhat is the potential scale reduction factor on split
chains (at convergence, Rhat=1).
```

Auswertung

- Graphische Darstellung
  - `myplot = fit.plot()`
- Extrahieren der Ketten
  - `myresult = fit1.extract(permuted=False)`
  - Ketten können einzeln geplottet oder als Histogramm (flattened) dargestellt werden
- Erneutes fitten mit anderen Optionen möglich
  - `fit2 = pystan.stan(fit=fit1, data=mydata, iter=100+wup, chains=6, warmup=wup)`

Zzusammenfassung PyStan

- Eingabe
  - Modell-String
    - Prior
    - Datenverteilung ($\Rightarrow$ Likelihood)
    - $\Rightarrow$ Posterior
  - Daten
- Ausgabe
  - Posterior Markov-Kette(n)
  - $\Rightarrow$ Auswertung
- Ausführungs-Varianten `help(pystan.stan)`

5.15.2 Zwei Münzen

Kommen zwei Münzen aus unterschiedlichen Prägeanstalten?

Unterscheiden sich zwei Kategorien dichotomer Daten?

- N1 und N2
- θ_1 für N1 und θ_2 für N2
- y1 und y2 (Verteilungen)
- Modell wird wieder als String übergeben
- Zusatzfrage: unabhängig?

Ergebnis: gemeinsamer Münzwurf

- Gemeinsame Verteilung günstiger Posterior-Stichproben
- HDI als 95% Credible Interval für die Differenz
  - beinhaltet die Nullhypothese $\Delta\theta = 0$
- $\Rightarrow$ Kein Unterschied
  - Mehr Daten ...

5.16 Hierarchische Modelle

5.16.1 Einleitung

- Mehrere Modellparameter
- Gemeinsame Grundlage
- Gegenseitige Abhängigkeit
  - Kopplung zwischen Hierarchie-Ebenen
  - Kopplung innerhalb Hierarchie-Ebene
- Beispiel: Heilungserfolg im Krankenhaus
  - Krankenhäuser: Heilungschance ω_j
    - Ausstattung: Motivation der Belegschaft, Ausbildungsniveau, ...
  - Ärzte-Teams: Erfolgsrate θ_i
    - Erfahrung, individuelle Ausbildung, eingespielt, ...
  - Patient: wird geheilt y_j mit Wahrscheinlichkeit θ_i
- Beispiel: Psychophysischer Effekt
  - Versuchsbedingung (z.B. Stimuluskontrast) ω beeinflusst Leistung der Versuchspersonen
  - Versuchspersonen haben Antwortwahrscheinlichkeit θ_i
  - Antwortverhalten y_{ij} bei mehrfacher Wiederholung
  - $\rightarrow$ Interesse liegt nicht auf den individuellen Leistungen θ, sondern auf der Beeinflussung ω

Beispiel: Münzprägeanstalt

- Prägeanstalt produziert Münzen mit Parameter ω
- Jede Münze hat eine Wahrscheinlichkeit θ_i für Kopf, abhängig von Prägemethode (ω)
- Bernoulli-Zufallsexperiment mit Wahrscheinlichkeit $p_{Kopf} = \theta_i$
- Siehe Abbildung 5.5

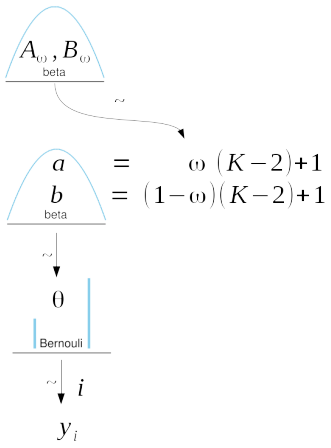

Abbildung 5.5: Münzwurf wie zuvor; die Parameter a und b für die Beta-Verteilung für θ entstammen über einen weiteren Parameter ω einer zusätzlichen Hierarchie-Ebene, ebenfalls einer Beta-Verteilung (nach: Kruschke: Doing Bayesian Data Analysis, 2nd. AP (2015) Fig. 9.1)

Gemeinsamer Parameterraum '*joint parameter space*'

- D: Daten Versuchsergebnis
- θ: Parameter für Verteilung der Daten
  - θ_i für verschiedene Münzen
- ω: Parameter für Verteilung der θ_i
- Gemeinsame Verteilung $p(\theta, \omega, D)$
- Der Posterior ergibt sich nach Bayes zu

$$\begin{aligned} p(\theta, \omega \mid D) &= \frac{p(D|\theta, \omega)\, p(\theta, \omega)}{p(D)} \\ &= \frac{p(D|\theta)\, p(\theta|\omega)\, p(\omega)}{p(D)} \end{aligned}$$

Take home: Hierarchische Modelle

- übergeordnete Parameter
- ohne *direkten* Einfluss auf die gemessene Zufallsvariable

5.16.2 Abhängigkeit

Beispiel: Münze aus Münzprägeanstalt

Likelihood

$$y_i \sim \text{Bernoulli}(\theta)$$

Prior für θ

$$\theta \sim \text{beta}(a, b)$$

- Hängt ab von Parametrisierung a, b
  - anstatt a und b könnte auch
  - *Erwartungswert* μ und *Standardabweichung* σ oder
  - *Modus* $\omega = \frac{a-1}{a+b-2}$ und *Konzentration* $\kappa = a + b$ gewählt werden:

$$\theta \sim \text{beta}\Big(\omega(\kappa - 2) + 1,\ (1 - \omega)(\kappa - 2) + 1\Big)$$

Prior für ω

$$p(\omega) = \text{beta}(\omega | A_\omega, B_\omega)$$

- mit Konstanten A_ω und B_ω
- Der Modus für ω liegt dann bei $\frac{A_\omega - 1}{A_\omega + B_\omega - 2}$

(Prior für κ)

- Konstante $\kappa = 100$

Damit von oben

$$\begin{aligned} p(\theta, \omega \mid D) &= \frac{p(D|\theta, \omega)\, p(\theta, \omega)}{p(D)} && \text{\# Bayes} \\ &= \frac{p(D|\theta)\, p(\theta|\omega)\, p(\omega)}{p(D)} && \text{\# Hierarchie} \end{aligned}$$

Lösung

- *Nicht* mathematisch geschlossen lösbar
- Numerisch per Gitter-Näherung

Schwache Abhängigkeit

- Sei ω unbekannt, etwas eingeschränkt um 0.5
- Sei θ nur schwach von ω abhängig
  - aber breit verteilt

Starke Abhängigkeit

- Sei ω unbekannt, breit verteilt
- Sei θ stark von ω abhängig
  - Münze einer Prägemethode untereinander ziemlich ähnlich
  - Münze unterschiedlicher Prägemethode unterscheiden sich

Ausgangssituation

- Gleiche Daten, gleiche Likelihood
- Prior stark gekoppelt/ schwach gekoppelt
  - breiter/schmaler
  - Bedingte Wahrscheinlichkeitsverteilung unterschiedlich, schmal/ähnlich, breit

Gemeinsamkeiten

- breite Prior-Randverteilung über θ $\Rightarrow$ Posterior-Randverteilung über θ ähnlich

Unterschiede

- Bedingte Posterior Verteilung $p(\theta|\omega)$:
  - passt sich an Likelihood an
  - wird vom Prior bestimmt
- Posterior Randverteilung $p(\omega)$ deutlich/kaum beeinflusst
  - obwohl Posterior im gekoppelten Fall breiter bleibt

'Take home': Abhängigkeit

- Abhängigkeit zwischen Parametern unterschiedlicher Hierarchie-Ebenen
  - erlaubt Rückschlüsse auf sonst unzugänglichen Parameter der oberen Ebene

5.16.3 Kopplung

Hierarchisches Modell am Beispiel zweier Münzen aus einer Prägeanstalt

- Zusätzliche Dimension: mehrere θs
- Unterschiedliche Daten: Likelihood
- Mehrere Posteriors für θs
- einen gemeinsamen Posterior für ω

Schwache Kopplung

- θ_s sind von ω weitgehend unabhängig.

Starke Kopplung

- θ_s hängen stark von ω ab.

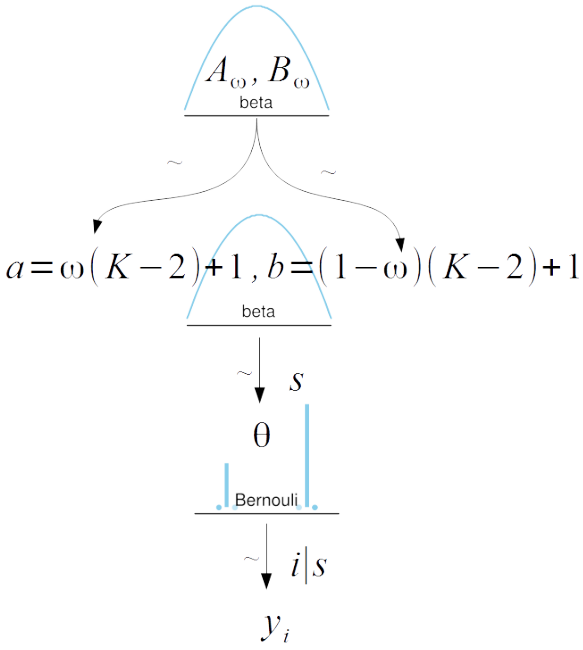

Abbildung 5.6: Beispiel mehrerer Münzen, nach Kruschke

Ergebnis

Kopplung durch 'Konzentration'-Parameter κ der Beta-Verteilung (äquivalent Anzahl der Daten im Vorwissen)

$$\kappa = a + b$$

erzwingt gemeinsame Betrachtung der Datensätze

Vergleich

- Kopplung erzwingt gemeinsame Posterior-Verteilung
- Kopplung erlaubt Aussage über obere Hierarchie-Ebene ω

'Take home': Kopplung

- Mehrere Parameter aus derselben unteren Hierarchie-Ebene können über die übergeordnete Ebene gekoppelt sein
- erlaubt Rückschlüsse
  - Posterior untere Stufe $p(\theta_i)$ beeinflusst durch Likelihood/Kopplung
  - Posterior obere Stufe $p(\omega)$
  - Abhängigkeit

5.16.4 Berechnung mit *Markov Chain Monte Carlo* Methode

Probleme mit Numerik bei Näherungsrechnung mit Gitter-Methode

- Graphiken entstanden mit 50 Stützstellen je Parameter
- Drei Parameter benötigen $50^3 = 125.000$ Berechnungen
- Vier Parameter schon 6 Millionen;
- Fünf Parameter gehen mit 312 Millionen schon nicht mehr zu berechnen

$\Rightarrow$ daher andere Methode: random walk *Markov Chain Monte Carlo*

Beispiel: Eine Münze aus der Prägeanstalt - starke Abhängigkeit

- Modell als String
- Ergebnis der Gitter-Näherungslösung wurde bestätigt

Überprüfen des Priors

- sinnvoll zur Überprüfung abhängiger Priors
  - Zwischenebenen in Hierarchie
  - abgeleitete Parameter
  - Differenzen

Ergebnis

- Modell mit enger Bindung von θ an ω reproduziert
- Prior, obwohl breite Randverteilung, beschreibt Abhängigkeit
- Posterior für θ von Daten (Likelihood) verschoben
- Posterior für ω durch Kopplung verschoben

Beispiele der schwachen und starken Kopplung

- schwache Kopplung: $\kappa = 5$
- starke Kopplung: $\kappa = 75$

Literatur:

- J. K. Kruschke: "Doing Bayesian Data Analysis, 2nd. A Tutorial with R, JAGS and Stan". Academic Press (2015)
- Bilder freundlicherweise zur Verfügung gestellt von J. K. Kruschke `https://sites.google.com/site/doingbayesiandataanalysis/figures`
- Graphiken selbst erstellen `https://github.com/tinu-schneider/DBDA_hierach_diagram/blob/master/README.md`

'Take home': Hierarchische Modelle mit MCMC

- Interpretation 'folgt aus'
- keine direkte Abhängigkeit der y von *Meta* Parametern ω
- Abhängigkeit θ von ω spielt wichtige Rolle für Ergebnis
- einfacher und schneller zu berechnen als vollständig verknüpfte Modellformulierung

5.17 Modellvergleich

5.17.1 Einleitung

Fragestellung:

Sie haben zwei Modelle zur Auswahl und Daten gemessen: welches Modell beschreibt die Daten besser?

Beispiele:

- Temperaturabhängigkeit des elektrischen Widerstands einer Kohleschicht
  - Linear?
  - Polynom?
  - Exponentiell?
- Lineares Modell
  - mit oder ohne Gruppenunterteilung
- Münzprägeanstalten
  - Münze aus Prägeanstalt A oder B
  - Modell-1: aus A oder Modell-2: aus B

Hierarchisches Modell

- Daten y
- Parameter θ
- Modellauswahl m # jetzt neu und verbessert
- Likelihood
  - Wahrscheinlichkeit für das Auftreten der Daten

$$p_m(y|\theta_m, m)$$

- Prior innerhalb eines Modells
  - Wahrscheinlichkeit für das Auftreten der Parameter

$$p_m(\theta_m|m)$$

- Prior zur Auswahl des Modells

$$p(m)$$

Bayes Regel

$$p(\theta_1, \theta_2, \theta_3, \ldots, m|D) = \frac{p(D|\theta_1, \theta_2, \theta_3, \ldots, m)\, p(\theta_1, \theta_2, \theta_3, \ldots, m)}{\sum_m \int p(D|\theta_1, \theta_2, \theta_3, \ldots, m)\, p(\theta_1, \theta_2, \theta_3, \ldots, m)\, \mathrm{d}\theta_m}$$

$$= \frac{\prod_m p(D|\theta_1, \theta_2, \theta_3, \ldots, m)\, p_m(\theta_m|m)\, p(m)}{\sum_m \int \prod_m p(D|\theta_1, \theta_2, \theta_3, \ldots, m)\, p_m(\theta_m|m)\, p(m)\, \mathrm{d}\theta_m}$$

Hierarchisches Modell

1. Modellauswahl m
2. Für jedes Modell Prior-Parameterverteilung
3. Für jedes Modell Prior-Datenverteilung
4. Daten

Modellauswahl m

Randverteilung (*Marginal*) von m sagt etwas darüber aus, wie wahrscheinlich welches Modell ist

- m ist diskret
- $p(m)$ die Wahrscheinlichkeit jedes Modells

$$p(m|D) = \frac{p(D|m)\, p(m)}{\sum_m p(D|m)\, p(m)}$$

- Das ist gefragt

Likelihood unter Modellauswahl m

- marginalisiert über alle Parameter θ

$$p(D|m) = \int p_m(D|\theta_m, m)\, p_m(\theta_m|m)\, \mathrm{d}\theta_m$$

- Diese Likleihood enthält den
  - **Prior** der **Parameter** innerhalb dieses Modells
  - **Likelihood** der Daten unter dem Modell
- ausintegriert (*marginalisiert*) als Randverteilung des ganzen Modells
- ... daher kann das gesamte Modell stak von den gewählten Priors der Modellparameter abhängen

5.17.2 Bayes-Faktor

Vergleich zweier Modelle

$$\frac{p(m=1|D)}{p(m=2|D)} = \frac{p(D|m=1)}{p(D|m=2)} \frac{p(m=1)}{p(m=2)} \frac{/ \sum_m p(D|m)\, p(m)}{/ \sum_m p(D|m)\, p(m)}$$

- Kürzen

$$\frac{\sum_m p(D|m)\, p(m)}{\sum_m p(D|m)\, p(m} = 1$$

- Vorwissen Modell-Prior

$$\frac{p(m=1)}{p(m=2)}$$

Bayes-Faktor (BF)

$$BF := \frac{p(D|m=1)}{p(D|m=2)}$$

- verschiebt die a-priori Wahrscheinlichkeit für die Modelle $\frac{p(m=1)}{p(m=2)}$
- durch Vergleich der *Modell-Likelihoods*
- Daumenregel zur Auswertung

  $BF > 3$ bzw. $BF < \frac{1}{3}$ gelten nach Harold Jeffreys als 'substantiell':

BF	Strength of evidence
$< 10^0$	negative
$10^0 \cdots 10^{1/2}$	barely worth mentioning
$10^{1/2} \cdots 10^1$	substantial
$10^1 \cdots 10^{3/2}$	strong
$10^{3/2} \cdots 10^2$	very strong
$> 10^2$	decisive

- Literatur:

  H. Jeffreys: The Theory of Probability (3 ed.). Oxford. p. 432 (1961)

5.17.3 (stellvertretendes) Beispiel: Münze aus zwei Prägeanstalten

- Modell-1: Münze aus Anstalt #1 haben $\omega = 0.25$
- Modell-2: Münze aus Anstalt #2 haben $\omega = 0.75$

Frage

Woher stammt eine vorliegende Münze?

Daten

- Nach $N = 9$ Würfen kommt $z = 6x$ Kopf

Prior

- ω_m siehe oben
- $\kappa = 12$
- gleichbedeutend mit
  - $\theta_1 \sim \text{beta}(3.5, 8.5)$ und $\theta_2 \sim \text{beta}(8.5, 3.5)$

Posterior m?

5.17.4 Lösung 1: formal

Mathematisch geschlossen lösbar (siehe *conjugate priors* in Kapitel 5.5 auf Seite 94)

- mit Beta-Funktion B (nicht beta-Verteilung beta)

$$p(D|m) = p(z, N) = \frac{\mathrm{B}(z + a, N - z + b)}{\mathrm{B}(a, b)} \quad (*)$$

- Bayes-Faktor

$$BF = \frac{0.000499}{0.002339} = 0.213 < \frac{1}{3}$$

- Unter der Annahme des Unwissens $p(m{=}1) = p(m{=}2) = \frac{1}{2}$ ergibt sich daraus

$$BF = \frac{p(m{=}1|D)}{p(m{=}2|D)} = \frac{p(m{=}1|D)}{1 - p(m{=}1|D)} \quad \Rightarrow \quad p(m{=}1|D) = 0.176\,, \quad p(m{=}2|D) = 0.824$$

Ergebnis:

- Die Münze kommt sehr wahrscheinlich aus der Anstalt #2,
- beschrieben durch Modell #2

Posterior?

- ... für m je nach Prior für m
- ... für θ haben wir damit noch nicht

'Take home': Lösung 1

Der Bayes Faktor

$$\frac{p(m=1|D)}{p(m=2|D)}$$

- kann direkt mathematisch geschlossen gelöst werden über konjugierte Funktionen

$$p(D|m) = p(z, N) = \frac{\text{beta}(z+a, N-z+b)}{\text{beta}(a, b)}$$

- erspart das Integral

$$p(D|m) = \int p_m(D|\theta_m, m)\, p_m(\theta_m|m)\, \mathrm{d}\theta_m$$

- sagt nichts über die Posterior-Verteilung für θ aus
  - θ wurde marginalisiert

5.17.5 Lösung 2: vollständige Gitter-Näherung

ω kann auch als kontinuierlicher Parameter angesehen werden

- erlaubt sind beide Werte der Prägeanstalten $[\omega_1, \omega_2]$

Prior

- Randverteilung '*marginal*' über ω hat zwei Spitzen
  - bei den beiden möglichen Werten
- Randverteilung '*marginal*' über θ hat zwei Höcker
  - um die beiden möglichen Werte ω_i
- Verbundwahrscheinlichkeit aus 1:1 Modell-Prior

Likelihood

- Daten **nur** in direkter Abhängigkeit von θ
- nicht direkt abhängig vom Modell mit Modell-Parameter ω

Posterior

- verschiebt Gewichte des Modellparameters ω
  - Verhältnis der Höhe = *Bayes-Faktor*
- Je nach Modell Verteilung der Parameter θ
- Insgesamt deren Randverteilung
  - wenn nur nach θ gefragt ist, unabhängig vom Modell(!)

5.17.6 Lösung 3a: diskrete MCMCs auf die beiden einzelnen Modelle

- Schritt 1: Integral = gewichteter Durchschnitt

$$\int f(\theta)p(\theta)\,\mathrm{d}\theta \approx \frac{1}{N}\sum_{\theta_i\sim p(\theta)} f(\theta_i)$$

  - da Häufigkeit $\sim$ Dichte

- Schritt 2: Modell-Likelihood

$$\begin{aligned} p(D) &= \int p(D|\theta)p(\theta)\,\mathrm{d}\theta \\ &\approx \frac{1}{N}\sum_{\theta_i\sim p(\theta)} p(D|\theta_i) \end{aligned}$$

  - Also Werte aus dem *Prior* ziehen und die Wahrscheinlichkeiten aufsummieren

- Lösung
  - Dafür ist die Markov-Kette gut: Stichprobe aus Posterior-Verteilung
  - Likelihood berechnen und aufsummieren

- Problem
  - Wahrscheinlichkeiten sind meist sehr klein
  - Genauigkeit der Computer beim Aufsummieren nicht ausreichend

Mathematischer Trick

Satz von Bayes

$$p(\theta|D) = \frac{p(D|\theta)\,p(\theta)}{p(D)}$$

Daraus

$$\frac{1}{p(D)} = \frac{p(\theta|D)}{p(D|\theta)\,p(\theta)}$$

Mit einer (vorerst) beliebigen, normierten Wahrscheinlichkeitsverteilung $h(\theta)$ ergibt sich

$$\frac{1}{p(D)} = \frac{p(\theta|D)}{p(D|\theta)\,p(\theta)}\int h(\theta)\,\mathrm{d}\theta$$

Weil die *evidence* $p(D)$ nicht von θ abhängt (!) und damit 1.) eine Konstante ist, 2.) der Bruch für alle θ gilt (Trick Teil I):

$$\frac{1}{p(D)} = \int \frac{h(\theta)}{p(D|\theta)\,p(\theta)} p(\theta|D)\,\mathrm{d}\theta$$

Damit näherungsweise

$$\frac{1}{p(D)} \approx \frac{1}{N}\sum_{\theta_i\sim p(\theta|D)} \frac{h(\theta_i)}{p(D|\theta_i)\,p(\theta_i)}$$

Wähle h so, dass es der zu erwartenden *Posterior*-Verteilung entspricht (Trick Teil II)

- Ähnliche Werte in Zähler und Nenner entschärfen das Problem der Genauigkeit
- beta-Verteilung für Bernoulli-Experimente
- Wähle Parameter der Verteilung h gemäß der Erwartung der θ des *Posteriors*
  - Mittelwert und Streuung
  - Posterior gegebenenfalls vorab aus repräsentativem Datensatz bestimmen
  - (nicht sehr kritisch)
- Kann dann vollständig MCMC-Methode ausnutzen: weniger, wenn nicht dicht, aber sinnvoll gewichtet

Ergebnis MCMC Modellvergleich

Die Likelihood für Versuchsergebnis-Daten $z = 6x$ Kopf in $N = 9$ Würfen beträgt (wie theoretisch berechnet)

- für Modell-1 ($\theta_1 = 0.25$): $p(D|m=1) = 0.002338$
- für Modell-2 ($\theta_2 = 0.75$): $p(D|m=2) = 0.000499$

Daraus errechnet sich der **Bayes-Faktor** zu

- $BF = \frac{0.002338}{0.000499} = 4.7$

womit sich die *posterior* Wahrscheinlichkeit der Modelle (nach *prior* 50%-50%) ergibt zu

- Modell-1: 83%
- Modell-2: 17%

'Take home': diskrete MCMCs

- wähle Vergleichsverteilung h so ähnlich wie möglich zur erwartenden *Posterior*-Verteilung
  - z.B. aus Trainings-Datensatz
- Löse obige Summe und invertiere
- $\Rightarrow$ *Likelihood* für das Modell
- Weiter für Posterior $p(\theta)$ wie bisher: MCMC darauf anwenden

5.17.7 Lösung 3b: MCMC auf das gesamte hierarchische Modell

Zu vergleichende Modelle als oberste Hierarchie einbauen in ein Gesamt-Modell

- Prior für Modellauswahl-Parameter m
- Prior für Parameterverteilungen ω je Modell
- Prior für Datenverteilung θ je Modell
- Berechne gesamtes Modell
- Werte Parameter m aus

Prior-Abhängigkeit

- uninformativer Prior für die Modelle
- Lernen durch Anpassung an Teile der Daten
- Auswerten mit Rest der Daten

ABER kein *int* Parameter in *Stan*

5.17.8 Vorhersage treffen mit alternativen Modellen

A) bestes Modell
   - das beste Modell suchen
   - dessen Vorhersage bestimmen

B) alle Modelle
   - alle Modelle gleichberechtigt
   - gewichten gemäß Wahrscheinlichkeit / Modell-Posterior
   - Vorhersagen mitteln

5.17.9 Komplexität

Zwei Prägeanstalten: A produziert faire Münzen, B alle möglichen

- Versuch: eine Münze wählen und werfen
  - $N = 20$
- Ergebnis 1)
  - $z_1 = 11$
  - Berechnet man m, so ist $m_A >> m_B$
- Ergebnis 2)
  - $z_2 = 15$
  - Berechnet man m, so ist $m_A << m_B$

Warum?

- ... wo doch 50%-50% bei *beiden* Prägeanstalten möglich ist?

Ergebnis

- B zahlt den Preis der höheren Komplexität durch weit verstreute Priors
- Gute Vergleiche mit ähnlich-informierten Priors für alternative Modelle

5.17.10 Abhängigkeit vom Prior

Posterior Parameter je nach Parameter-Prior

Beispiel: Drei Prägeanstalten A, B, C

A produziert faire, B und C allerhand Münzen

- 'Fairer' Prior für A

$$\theta_A \sim \mathbf{beta}(500, 500)$$

- 'Allerhand' Priors: 'gleich' für B und '*Haldane*' für C

$$\theta_B \sim \mathbf{beta}(1, 1) \qquad \theta_C \sim \mathbf{beta}(0.01, 0.01)$$

Das Versuchsergebnis sei

$$z = 65 \text{ von } N = 100$$

Auswertung

- Posterior

$$\theta_B \sim \mathbf{beta}(66, 36) \qquad \theta_C \sim \mathbf{beta}(65.01, 35.01)$$

- 95%-HDI

$$[0.554,\ 0.738] \qquad [0.556,\ 0.742]$$

- $\Rightarrow$ sehr ähnlich

Ergebnis

Posterior Modell nach Bayes-Faktor je nach Parameter-Prior

$$\mathrm{BF}_{AB} = 5.728 \qquad \mathrm{BF}_{AC} = 0.125$$

$\Rightarrow$ Kontrovers!

Ausweg

Generiere einen *Zwischen*-Prior mit einem Teil der Daten

Gleiches Beispiel wie oben:

- Versuchsergebnis
$$z = 65 \text{ von } N = 100$$
- Erster Teil der Daten
$$z = 6 \text{ von } N = 10$$
- Prior
$$\theta'_B \sim \mathbf{beta}(7, 5) \qquad \theta'_C \sim \mathbf{beta}(6.01, 4.01)$$
- Zweiter Teil der Daten
$$z = 59 \text{ von } N = 90$$
- Ergebnis
$$\mathrm{BF}_{AB'} = 0.0557 \qquad \mathrm{BF}_{AC'} = 0.0575$$
- $\Rightarrow$ plausibel

5.17.11 *'Take home':* MCMC eines gesamten hierarchischen Modells

- Mischung von Modellen
  - universeller Posterior zur Vorhersage
- Komplexität
  - Modellvergleich gut nur bei ähnlich-informativen Priors
- Prior-Abhängigkeit
  - entschärfen durch Anlernen mit Teil-Daten
- Bemerkung: Beta-Verteilung ist ***nicht*** immer passend bei Bernoulli-Experimenten
  - aber lehrreich, da Vergleich Geschlossene Lösung (Lösung 1) / Gitter Näherungslösung (Lösung 2) / MCMC (Lösung 3a und 3b) möglich

5.18 Vergleich zu frequentistischer Statistik

Experimente: Wieder stellvertretend für alle Ja/Nein Experimente der Münzwurf

1. Experiment: $24x$ Werfen
   - $7x$ Kopf von 24 Würfen
   - Verwerfungsbereich (mit Irrtumswahrscheinlichkeit): $z < 7$ oder $z > 17$
   - p-Wert: 0.064 $\Rightarrow$ Nullhypothese wird nicht verworfen
2. Experiment: Werfen bis $7x$ Kopf
   - Verwerfungsbereich (mit Irrtumswahrscheinlichkeit): $N < 8$ oder $N > 20$
   - p-Wert: 0.017 $\Rightarrow$ Nullhypothese wird verworfen
3. Experiment: 2 Minuten werfen
   - Abstand zweier Würfe sei *Poisson*-verteilt mit $\lambda = 5$

Vergleich zwischen Versuch 1) und Versuch 2) mit unterschiedlicher Intention

- gleiche Versuchsdaten: $N = 24$ Würfe, davon $z = 7$ Köpfe
- Versuch 1)
  - Intention: feste Anzahl werfen, wie viele Erfolge darunter?
  - $\Rightarrow$ Nullhypothese: 'Münze ist fair' wird **nicht verworfen**
- Versuch 2)
  - Intention: werfen, bis feste Anzahl Erfolge, wie oft insgesamt geworfen?
  - $\Rightarrow$ Nullhypothese 'Münze ist fair' wird **verworfen**

Frequentistische Statistik Das Beispiel ($z = 7, N = 24$) hat

- Punktschätzer
$$\hat{\theta} = \frac{z}{N} = 0.2917$$
- Vertrauensintervalle
$$\begin{aligned} \theta_{N=24} &\in [0.126,\, 0.511] \quad \text{auf} \quad \alpha = 5\% \\ \theta_{z=7} &\in [0.126,\, 0.484] \quad \text{auf} \quad \alpha = 5\% \\ \theta_{t=2min} &\in [0.135,\, 0.497] \quad \text{auf} \quad \alpha = 5\% \end{aligned}$$
- p-Werte unter Nullhypothese $\theta = \frac{1}{2}$
$$\begin{aligned} p{-}val_{N=24} &= 0.064 \quad \text{auf} \quad \alpha = 5\% \\ p{-}val_{z=7} &= 0.025 \quad \text{auf} \quad \alpha = 5\% \\ p{-}val_{t=2min} &= 0.049 \quad \text{auf} \quad \alpha = 5\% \end{aligned}$$

5.19 Versuchs-Intention

Frequentistische Statistik

- Obwohl das Versuchsergebnis das Selbe ist, ist die Schlussfolgerung abhängig von der Intention des Versuchs
- Relevanz für Praktiker
  - Situation entschärft für große N
  - Problem bleibt bestehen für kleine N
  - Problem kann für bestimmte Verteilungen dramatisch sein

Bayes-Statistik

- Bayes Schlussfolgerung hängt nicht von Intention ab, sondern von der *Likelihood*
- Diese ist für alle Versuchsintentionen die Selbe
- Mit der mathematisch geschlossenen Lösung für den Posterior θ ergibt sich
  - HDI [0.125, 0.474]
  - enthält Nullhypothese $\theta_0 = \frac{1}{2}$ nicht
- mit PyStan MCMC ergibt sich
  - HDI [0.126, 0.468]
- Ergebnis
  - gleiche Versuchsdaten wie oben
    - $N = 24$ Würfe
    - $z = 7$ Köpfe
  - von Intention unabhängig
    - Posterior kann mit jedem Einzelergebnis erneuert werden
  - Versuchsauswertung
    - Posterior direkt interpretierbar als Verlässlichkeit von θ: durch $p(\theta|D)$
    - Posterior HDI enthält den Parameter der Nullhypothese **nicht**
    - $\Rightarrow$ Nullhypothese '*Münze ist fair* wird **verworfen**
- MAP und HDI hängen vom Prior ab
  - *maximum-a-posterior* als Punktschätzer
  - *credible interval* HDI
  - Vorteil: man kann sein Vorwissen im Prior weiterverwenden

Prior

Ominöser Prior?

- Man muss such auf Prior einigen
  - Theorie (Verteilungen ...)
  - Vorwissen durch andere Veröffentlichungen
  - Vorversuche
- Selbst wenn '*agree to disagree*'
  - dann kann man beide Varianten berechnen
- Ähnliches Problem bei NHST
  - welche Verteilung ist angemessen?

Unwissender Prior?

- *indifferent* Prior
  - $p(\theta) = const.$
  - Bernoulli-Experiment: $p(\theta) = 1 = \text{beta}(1, 1)$
- Jeffreys Prior
  - invariant unter Koordinatentransformation
  - Bernoulli-Experiment: $p(\theta) = \text{beta}(\frac{1}{2}, \frac{1}{2})$
- Haldane Prior
  - als ob *keinerlei* Vorwissen
  - Bernoulli-Experiment: $p(\theta) = \text{beta}(\epsilon, \epsilon)$

Literatur:

- Jeffreys: An Invariant Form for the Prior Probability in Estimation Problems. Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences. 186 (1007): 453–461. JSTOR 97883 (1946) (`http://rspa.royalsocietypublishing.org/content/186/1007/453`)
- Haldane: A note on inverse probability. Mathematical Proceedings of the Cambridge Philosophical Society. 28: 55–61 (1932) (`https://www.cambridge.org/core/journals/mathematical-proceedings-of-the-cambridge-philosophical-society/article/note-on-inverse-probability/1BC33DBEA96916D0A31998DCD`

5.20 Entscheidung mit Bayes-Statistik

HDI

- Fällt der zu testende Parameter-Wert θ_x in das *credible interval*, den Bereich der höchsten Dichte *HDI*? $\theta_x \in$ HDI

Bayes Faktor

- beschreibt der gesuchte Parameter θ_x die Daten besser als andere?
- Vergleich mit einem weiten Bereich von möglichen Parametern

5.20.1 HDI und ROPE

ROPE = *'Region of Practical Equivalence'* definiert denjenigen Bereich an Parametern

- der für die Anwendung relevant ist
- ob Unterschiede in θ sich auf das Verhalten auswirken

Beispiel: Münze für Fußballanstoß darf ROPE = [0.45...0.55] haben

- Definition: Nullhypothese

Sind die Daten mit der ROPE vereinbar?

- Irrtumswahrscheinlichkeit α z.B. $\alpha = 5\%$

Verwerfen

- Verwerfen der Nullhypothese genau dann, wenn $(1-\alpha)$-HDI und ROPE keinen Überlapp zeigen

Akzeptieren

- Akzeptiere die Nullhypothese genau dann, wenn gesamtes $(1-\alpha)$-HDI innerhalb der ROPE

Graubereich

- ROPE innerhalb breitem HDI
  - zu ambitionierte Vorgabe
  - zu wenig Daten

sub-optimal

- HDI liegt innerhalb ROPE, aber HDI enthält θ_0 selbst nicht
  - ROPE zu weit
  - Modell könnte besser sein (Prior)

Entscheidung

- Jede Entscheidung reduziert die vorliegenden Daten/Berechnungen auf einen binären Wert

  - beispielsweise die Posterior-Verteilung auf verwerfen/nicht-verwerfen
- In der Posterior-Verteilung steckt jedoch mehr Information
  - Damit kann jeder *seine* Entscheidung fällen
- Äquivalent aus den Statistik-Grundlagen 'Frequentistische Statistik'
  - Punkt (Ort, Schätzer)
  - Intervall (Breite)
  - Verteilung (Form)
- Wenn insbesondere Stichproben erhoben werden (z.B. durch MCMC)
  - spielt die Grenze der Intervalle eine Rolle
  - variiert selbst
- Beispielsweise MCMC
  - 10.000 *effective sample size*
  - $\Rightarrow$ SD des HDI ist $\sim$5% der SD des Posteriors (bei Normalverteilung)

5.20.2 HDI und ETI

Equally Tailed Interval = ETI

- 95%-ETI schneidet auf beiden Seiten 2,5% der Wahrscheinlichkeit ab
- andere Gewichtung: absolute Wahrscheinlichkeit
- ETI lässt sich leichter berechnen ('ppf')
- ETI ist invariant unter Transformation
  - Für HDI sinnvolle Bedeutung der Parameter wählen

HDI einer (angepassten) theoretischen Verteilung

- gemäß Modell an Posterior Samples anfitten
- daraus HDI berechnen

5.20.3 Weitere Entscheidungen

Parameter-Relation: z.B. Differenz

- Aus den Randverteilungen lassen sich keine Schlüsse ziehen, aus dem Gesamt-Modell durchaus

Mehrfach-Vergleiche

- Kumulierte α-Fehler sind problematisch (Erinnerung Angewandte Statistik I)
- Bayes-Statistik beschreibt ein (Gesamt-) Modell
- $\Rightarrow$ unproblematisch

Posterior Vorhersage

- Bayes-Statistik bestimmt die relative Verteilung der Parameter
- Guter Schätzer?
  - Ob der MAP-Schätzer ein guter Schätzer ist, muss getestet werden
  - Beispiel: Wir gehen von einer 1% oder 99% Trick-Münze aus
    - Ergebnis ist 30/40
    - Beide Modelle sind schlecht
    - Das 99%-Modell ist jedoch viel weniger schlecht als das 1%-Modell
  - Sehen Daten aus dem geschätzten Modell den gemessenen ähnlich?
    - $\Rightarrow$ Dann ist das Modell angemessen

5.20.4 Entscheidung durch Modell-Vergleich

Eine andere Art der Fragestellung: Ist ein Modell mit einem spezifischen Prior (*Nullhypothese*) besser als eines mit einem uninformativen Prior?

- Dann wird die Nullhypothese nicht verworfen

Obiges Beispiel ($z = 7, N = 24$)

- Nullhypothese $H_0: \ \theta_0 = \frac{1}{2}$

$$p(z, N|M_0) = \theta_0^z (1 - \theta_0)^{N-z}$$

- Alternativhypothese H_A:

$$p(z, N|M_A) = B(z + a_A, N - z + b_A) \Big/ B(a_A, b_A)$$

Bayes-Faktor

$$\frac{p(z,N|M_A)}{p(z,N|M_0)} = \frac{B(z+A_A, N-z+b_A)\Big/B(a_A,b_A)}{\theta_0^z(1-\theta_0)^{N-z}}$$

$$\frac{p(z,N|M_{alt})}{p(z,N|M_{null})} = \begin{cases} 3.7227 & \text{for } a_{alt}=2,\ \ b_{alt}=4 \\ 1.9390 & \text{for } a_{alt}=b_{alt}=1.000 \\ 0.4211 & \text{for } a_{alt}=b_{alt}=0.100 \\ 0.0481 & \text{for } a_{alt}=b_{alt}=0.010 \\ 0.0049 & \text{for } a_{alt}=b_{alt}=0.001 \end{cases}$$

Entscheidungen treffen

1. Bayes-Faktor
   - Der Bayes-Faktor ändert sich stark mit dem Prior
     - nicht verwerfen von H_0 für (a,b) mit $a=b<0.01$
     - verwerfen von H_0 für $(a=2,b=4)$
2. Posterior
   - Sieht man sich die Posterior-Verteilung und die HDIs an, so unterscheiden die sich gar nicht so sehr
   - alle würden die Nullhypothese verwerfen
3. Übergeordnetes Gesamt-Modell: **Ein** Hierarchisches Modell
   - Modell-Vergleich
     - übergeordneter Modell-Parameter
     - entschiedet sich für/gegen das Null-Modell
     - Bayes-Faktor
   - Parameter-Schätzung
     - vergleicht Null-Parameter mit unvoreingenommenem Prior
     - zeigt Posterior
     - HDI $\in$ ROPE?
   - Beide Varianten des einen Modells können ausgewertet werden
     - müssen nicht übereinstimmen
     - Anwendungsabhängig, was bevorzugen
     - Null-Parameter muss Bedeutung haben (Theorie, Literatur, Entscheidung, ...)
     - Meist ist Posterior (Parameter-Schätzung) aussagekräftiger
   - Einschränkungen
     - ROPE
       - muss eng sein

    - muss angemessen gewählt werden (Genauigkeit)
  - Prior und Posterior Verteilung sollten glatt sein im Bereich
  - Näherung $\Leftrightarrow$ exakt = @Punkt

'Savage-Dickey Methode'

Verschachtelte Modelle (*nested models*)

0) Einfaches Modell mit Parameter $\theta = \theta_0$ (Nullhypothese)

A) Erweitertes Modell mit freiem Parameter θ (Alternative Hypothese $\theta \neq \theta_0$)

Gesucht: Bayes-Faktor

Die ***Savage-Dickey Methode***

$$BF_{0A} = \frac{p(D|H_0)}{p(D|H_A)} = \frac{p(\theta = \theta_0|D, H_A)}{p(\theta = \theta_0|H_A)}$$

vergleicht nur im das erweiterten Modell A) den Posterior mit dem Prior für den interessierenden Parameterwert $\theta = \theta_0$

- BF findet einen Widerspruch zum (scharfen) Vorwissen (Nullhypothese)
- BF bestätigt Nullhypothese (Nähe Likelihood)

Beweis:

Mit $p(D|H_0) =: p_0(D)$

$$p_0(D) = \int p_0(D|\Psi)p_0(\Psi)\mathrm{d}\Psi$$
$$p_0(D) = \int p_A(D|\Psi, \phi = \phi_0)p_A(\Psi|\phi = \phi_0)\mathrm{d}\Psi = p_A(D|\phi = \phi_0)$$
$$p_0(D) = \frac{p_A(\phi = \phi_0|D)p_A(D)}{p_A(\phi = \phi_0)}$$

und damit

$$BF_{0A} = \frac{p_0(D)}{p_A(D)} = \frac{p_A(\phi = \phi_0|D)}{p_A(\phi = \phi_0)}$$

Literatur

- Wagenmakers, Lodewyckx, Kuriyal, Grasman: Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. (2010)

5.21 Tests

5.21.1 Trennschärfe

- Macht (*power*) eines Tests: kann das Ziel erreicht werden
- Ziele
  - Verwerfen der Nullhypothese
    - ROPE außerhalb 95% HDI
  - Akzeptieren der Nullhypothese
    - 95% HDI innerhalb ROPE
  - Präzision
    - 95% HDI schmaler als geforderte Genauigkeit
- Problematik
  - Wie in der klassischen Statistik verbleibt eine Irrtumswahrscheinlichkeit von 5%
  - Daten sind Zufallsvariablen und können auch zufällig
    - 24 von 30 mal Kopf zeigen, *obwohl* die Münze fair ist
    - Placebo wirkungsvoller als Medikament in getesteter Gruppe
    - ...
- Vorsichtsmaßnahmen
  - Rauschen weitestgehend vermeiden
    - Zufall einhalten
    - Einflüsse ausschließen (Magnetfeld, Kontrollgruppe an Patienten angleichen)
    - Effekt verstärken (Labor)
- Anzahl Messdaten
  - Störeinflüsse mitteln sich heraus
  - Mittelwertsfehler

5.21.2 Simulation

- Parameter
  - aus Gesamtraum
  - aus Theorie
  - aus Vorexperimenten
- Mögliche Daten simulieren
  - 'Ziehen' aus Verteilungen (Parameter)
  - Entsprechend nachher der Versuchsdurchführung
- Bayes-Statistik anwenden
  - Posterior
  - Schätzer
  - HDI
- Ergebnis des simulierten Posteriors wie benötigt?
  - ROPE & HDI
- oft wiederholen
  - Ziel meist erreicht?
  - Bayes-Statistik darauf anwenden
- Entscheidung
  - Nein?
    - Stichprobenumfang erhöhen
  - Ja?
    - Versuch genau so durchführen
    - Messen und auf Daten anwenden

5.21.3 Stichprobenumfang

- Klassische Statistik
  - Varianz wird mit steigender Datenanzahl kleiner ('Gesetz der großen Zahlen')
- Bayes Statistik
  - Posterior wird mit steigender Datenanzahl schmaler und überstimmt jeden Prior
  - Bayes-Faktor im Modellvergleich soll bestimmte Höhe erreichen

5.21.4 Abbruch-Kriterium

- Daten sammeln bis zum Abbruchkriterium N ist Standard
- Daten sammeln bis zum Abbruchkriterium 'Nullhypothese verwerfen' = $p < 0.05$
  - führt zu Ablehnung der Nullhypothese in 100% der Fälle(!)
  - Grund: es finden sich selbst wenn die Nullhypothese zutrifft in 5% der Fälle Ausnahmen, die ein Verwerfen rechtfertigen würden
  - MCMC Ergebnisse sind immer 'biased' zu einem Schwanz der Verteilung hin
    - da Anhäufungen von Extremwerten bevorzugt werden
    - und die entgegengesetzen Extremwerte nach dem Abbruch keine Chance mehr bekommen
- Ausweg
  - NHST
    - Festlegen des Stichprobenumfangs nach vorheriger Bestimmung der Macht des Tests
    - Registrieren des Versuchs
  - Bayes
    - Festlegen des Stichprobenumfangs anhand geforderter Genauigkeit
    - Genauigkeit ist nicht vom Wert beeinflusst
      - Ausnahmen: Poisson-Statistik, Beta-Verteilung (nur leicht, daher Ergebnis trotzdem anwendbar)

5.21.5 Daten-Modell-Vergleich

- Posterior liefert Parameter
- Simuliere Daten mit diesen Parametern
- Sehen diese simulierten Daten so aus wie die gemessenen?
  - Ja
    - Modell ist angemessen
    - Parameter verwendbar
  - Nein
    - Modell ist unangemessen
    - Parameter sinnlos

5.22 *'Take home'*-Messages

credible intervals, HDI

- die Posterior-Verteilung $p(\theta|D)$ ist die interessante Größe
- das Posterior 95%-HDI enthält zu 95% den wahren Wert θ_0
  - kann eine Nullhypothese ausschließen
  - kann durch Vergleich mit ROPE die Nullhypothese *für praktische Belange* akzeptieren
- verlangt Prior / erlaubt Vorwissen
  - Einfluss auf Breite und Lage des HDI
  - jeder nach seiner Fasson

Konfidenzintervale der Frequentistischen Statistik

- können eine Nullhypothese verwerfen
- können nicht die Nullhypothese akzeptieren
- sagen nichts über Wahrscheinlichkeit des wahren Parameters aus
  - schon gar nicht über eine Verteilung des wahren Parameters
- setzen Normalverteilung voraus (bei t-Test)
- hängen ab von der Intention, Datenerhebungsstrategie
- erlauben keine Abschätzung der Macht des Tests

Bayes-Faktor

- Kann zu Entscheidung verwendet werden
  - vergleicht Likelihoods unter Nullhypothese mit Alternativ-Modell
- Starke Abhängigkeit von Prior-Auswahl
  - Prior sinnvoll zur Fragestellung wählen
- Überprüfen, ob Ziel erreicht
  - Liegt die Nullhypothese in der Nähe des Posterior HDI?
- Oft ist Kriterium HDI vs. ROPE sinnvoller

5.23 Generalisierte Lineare Modelle mit Bayes

5.23.1 Eine kontinuierliche Variable - Beispiel Intelligenztest *IQ*

Daten - Zufallsvariable Y

- natürliche Streuung, Abweichungen, Rauschen, Messfehler, ...
- Wahrscheinlichkeitsverteilung

$$\begin{aligned} Y &\sim \mathcal{N}(\mu, \sigma^2) \\ \mathcal{E}(Y) &= \mu \end{aligned}$$

Modell mit Modellparametern

- Beispielsweise Normalverteilung der IQ-Messwerte in der Bevölkerung

$$p(y|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(y-\mu)^2}{2\sigma^2}\Big)$$

Likelihood für Datensatz

- Versuche/Messwiederholungen sind voneinander unabhängig *i.i.d.*
- $D = \{y_i\}$

$$L = p(D|\mu, \sigma) = \prod_{i=1}^{N} p(y_i|\mu, \sigma)$$

Satz von Bayes

$$p(\mu, \sigma|D) = \frac{p(D|\mu, \sigma)\, p(\mu, \sigma)}{\int_\mu \int_\sigma p(D|\mu, \sigma)\, p(\mu, \sigma)\, \mathrm{d}\sigma\, \mathrm{d}\mu}$$

Prior

- Vorwissen, Theorie, allgemein akzeptiert, ... Beispielsweise:
  - μ: Mittelwert bei 100 (Definition!); streut zwischen 0 und 200
    - $\mu \sim \mathcal{N}(100, 100)$
  - σ: aus langjähriger Erfahrung schwankend, vielleicht 10; kann aber nahe 0 bis 100 sein
    - $\sigma \sim \mathcal{U}(1/100; 100)$

Posterior

- Mathematisch geschlossene Lösung für μ
  - Annahme: $\sigma = const. = \sigma_L$
  - Konjugierte Prior für Gauß-Verteilung: wieder Gauß-Verteilung
    - Beweis per Produkt zweier Gauß-Verteilungen ist eine Gauß-Verteilung

  - Likelihood

$$p(D|\mu,\sigma) = \prod_{i=1}^{N} p(y_i|\mu,\sigma)$$

    * mit

$$p(y_i|\mu,\sigma) = \mathcal{N}(\mu,\sigma_L^2)$$

  - Prior:

$$p(\mu) = \mathcal{N}(\mu_{prior},\sigma_{prior}^2)$$

  - Ergebnis Posterior

$$\mu_{posterior} = \frac{1/\sigma_L^2}{1/\sigma_L^2 + 1/\sigma_{prior}^2}\mu_{L'} + \frac{1/\sigma_{prior}^2}{1/\sigma_L^2 + 1/\sigma_{prior}^2}\mu_{prior}$$
$$\sigma_{posterior} = \frac{1}{1/\sigma_L + 1/\sigma_{prior}}$$

- Mathematisch geschlossene Lösung für σ
  - Annahme: $\mu = const. = \mu_L$
  - konjugierte Prior: Gamma-Verteilung für $1/\sigma$

Zutaten für Bayes

- Datenvektor y_i (Länge N)
- Mittelwert $\mathbf{y}$ (Schätzer $\overline{y}$)
- Standardabweichung y (Störparameter, Schätzer s)
- robuste Schätzung (akzeptiert Ausreißer)
  - Student t (statt Normalverteilung)
  - mit zusätzlichem Parameter *Freiheitsgrade* ν
- $\Rightarrow$ Parameter μ, σ, ν

Daraus Modell für PyStan

- als String
- MCMC mit PyStan

Ergebnis MCMC Posteriors:

- Posterior μ
- Posterior ν
- Posterior σ

Ergebnis: Eine Verteilung

- Posterior beschreibt passende Verteilung an Daten
  - Modell ist angemessen

- Ausreißer werden durch t-Verteilung beschrieben
  - Abweichung von Normalverteilung mit Freiheitsgrade-Parameter $\nu = 5.6$ (Median)
- Erwartungswert der Verteilung '*smart drug*' über dem Durchschnitt 'ohne'
  - Punktschätzer für $\mu = 107$
  - Breite der Verteilung μ: Mittelwertsfehler innerhalb 95%-HDI [101.7, 112.6]
  - Breite der Verteilung der Daten: Parameter $\sigma = 20$ mit $[t_{0.025} = 58,\ t_{0.975} = 157]$

5.23.2 Kategoriales Modell - Beispiel Intelligenztest *IQ*

Vergleich zweier Gruppen/Kategorien: Plazebo vs. '*smart drug*'

Kategorien können mit einem Modell angepasst werden

- Gruppierte Daten y_{ij}
- Indikator-Variable *'category'* mit Inhalt j
- Parametervektor $\boldsymbol{\mu} = [\mu_j]$

Posterior liegt als Verbundverteilung für gesamten Parametervektor vor und erlaubt daher einen direkten Vergleich durch die Differenz beider Ketten

- Die Differenz $\Delta\,\mu = \mu_{smart\ drug} - \mu_{Placebo}$ ist positiv und von Null verschieden
- das 95%-HDI für $\Delta\,\mu$ liegt außerhalb einer *ROPE* von 0 ± 1

5.23.3 Lineares Modell

Siehe Kapitel 1 ab Seite 4

Erweiterung der Kategorien auf beliebig viele: Abhängigkeit von einem (oder mehreren) unabhängigen (Vorhersage-) Variablen X

Beispiel:

1. Intelligenztest: IQ-Messung
   - Verteilung
2. Intelligenztest: IQ-Vergleich zweier Bedingungen
   - zwei Verteilungen
   - zwei Kategorien '*placebo*' und '*smart drug*'
3. Intelligenztest: IQ in Abhängigkeit von der Konzentration eines Dopings
   - kontinuierlicher Vorhersage-Parameter
   - Verteilung für IQ ändert sich kausal

Wahrscheinlichkeitsverteilung der abhängigen Zufallsvariablen Y

- Ursache: Streuung, Rauschen, Messfehler, ... = Zufall

$$\begin{aligned}\mathcal{E}(Y_i) &= \mu_i \\ Y_i &\sim \mathcal{N}(\mu_i, \sigma^2)\end{aligned}$$

- mit der Linearen Abhängigkeit von unabhängigen Variablen X_j

$$\mu_i = \mathbf{x}_i^T \boldsymbol{\beta} = \sum_{j=1}^{N_j} \beta_j x_{ji}$$

- mit der Generalisierten Linearen Abhängigkeit

$$g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$$

Mathematisches Modell: Designmatrix und Parametervektor

- **Designmatrix** mit k unabhängigen Variablen X_j in Spalten der Länge n (Anzahl der Messwerte)

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & & \ddots & & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{nk} \end{bmatrix}$$

- **Parametervektor**

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

- **Generalisiertes Lineares Modell**

$$g\big(\mathcal{E}(\mathbf{Y})\big) = g\big(\boldsymbol{\mu}\big) = \mathbf{X}\boldsymbol{\beta}$$

$$Y_i \sim f(y_i; \boldsymbol{\theta}_i) \quad \text{z.B.} \quad \mathcal{N}(\mu_i, \sigma_i^2)$$

Varianten unabhängiger Variablen

- Nominal, eine Kategorie mit Faktor β_0
- Nominal, mehrere Kategorien mit Faktoren $\beta_i \quad i \in \mathbb{N}^+$ und Indikatorvariablen in Designmatrix
- Eine metrische Variable $\beta_0 + \beta_1 x$
- Mehrere metrische Variablen $\beta_0 + \sum_j \beta_j x_j$
- mit Interaktion $\beta_0 + \sum_j \beta_j x_j + \sum_{jk} \beta_{jk} x_j x_k [+...]$
- Ordinale Variable(n) ($\rightarrow$ nominal; aber Reihenfolge spielt Rolle)

Beispiel Link-Funktion g

$$g(\eta) = \text{logistic}(\eta) = \frac{1}{1 + e^{-\eta}}$$

- ausgedrückt durch unabhängige Variable X

$$g(x; \beta_0, \beta_1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

- umparametrisiert mittels *gain* γ und *threshold* θ :

$$g(x; \gamma, \theta) = \frac{1}{1 + e^{-\gamma(x-\theta)}}$$

Anwendung Tierdaten: Gehirngewicht

- Zum Vergleich mit frequentistischer Statistik, siehe GLM (Kapitel 2, ab Seite 24)

Wie mit Bayes-Statistik?

$$p(\beta_0, \beta_1, \sigma, \nu | D) = \frac{p(D|\beta_0, \beta_1, \sigma, \nu) p(\beta_0, \beta_1, \sigma, \nu)}{\int\int\int\int p(D|\beta_0, \beta_1, \sigma, \nu) p(\beta_0, \beta_1, \sigma, \nu) \, \mathrm{d}\beta_0 \, \mathrm{d}\beta_1 \, \mathrm{d}\sigma \, \mathrm{d}\nu}$$

Wie mit MCMC aus PyStan?

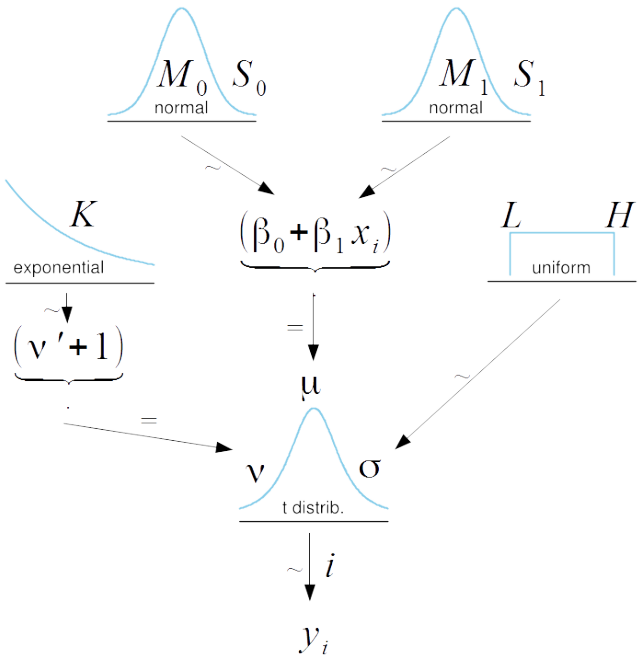

Abbildung 5.7: Model mit Abhängigkeit für robuste lineare Regression, nach: Kruschke: Doing Bayesian Data Analysis, 2nd. AP (2015)

- Parameter von Interesse:
  - Parametervektor β
    - besteht ein linearer Zusammenhang zwischen X und Y? $\Rightarrow$ *gain* $\beta_1 \neq 0$
    - ist der lineare Zusammenhang proportional? $\Rightarrow$ *intercept* $\beta_0 = 0$
  - Streuung der Daten σ
  - Abweichung von Normalverteilung (robust gegen Ausreißer)

  - Üblicherweise **nicht** μ_i (wird von $\boldsymbol{\beta}$ und X vorhergesagt)
  - Vorgabe für konstante Prior-Parameter M_i, S_i, K, L, H.

Ergebnis und Vergleich mit GLM: Parametervektor β

- Bayes-Posterior mittels PyStan
- GLM Fit-Ergebnis für Parameter
- $\Rightarrow$ Sehr gute Übereinstimmung

Ergebnis Störparameter

- Abweichung von der Normalverteilung
- Mittelwert von $\nu = 33.47$ mit Standardabweichung $s_\nu = 28.54$
  - $\Rightarrow$ die Abweichung von der Normalverteilung spielt keine Rolle und kann vernachlässigt werden.
- Streuung
  - Mittelwert von $\sigma = 0.29$ mit Standardabweichung $s_\sigma = 0.03$
  - $\Rightarrow$ die Daten (Gehirngewicht) streuen um den lineare vorhergesagten Erwartungswert ($\hat{=}$ Faktor 2)

Ergebnis Lineares Modell

- Datenverteilung wird sehr gut beschrieben
- Vergleich mit KQ/IRLS Linearem Modell
  - Werte für Parametervektor stimmen überein
  - Streuung der Parameter stimmt mit Konfidenzintervall überein
  - Streuung der Daten stimmt mit Konfidenzintervall überein

5.23.4 Lineares Modell mit Kategorien

- Modell-String
- Ergebnis
  - PyStan
  - Vergleich mit GLM
  - $\Rightarrow$ sehr gute Übereinstimmung

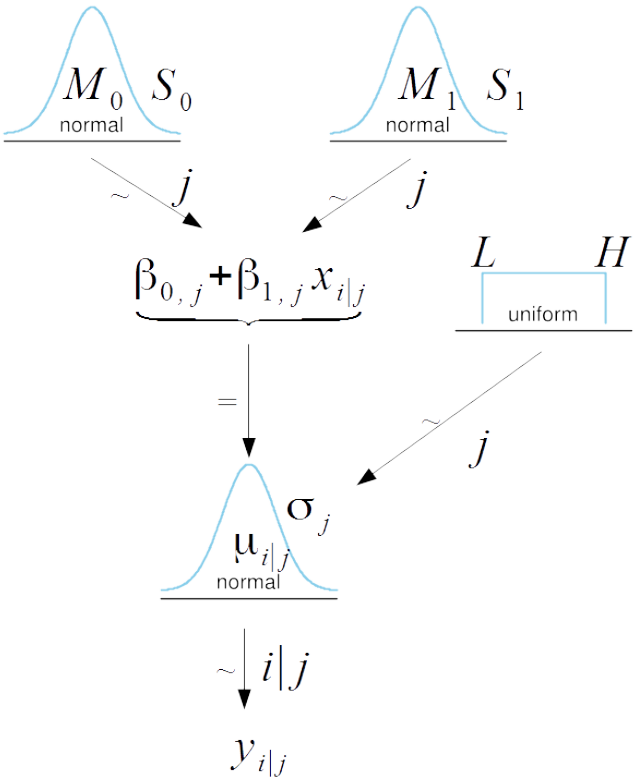

Abbildung 5.8: Lineares Modell mit Kategorien, nach: Kruschke: Doing Bayesian Data Analysis, 2nd. AP (2015)

5.23.5 Hierarchisches Lineares Modell mit Kategorien

Ergebnis

- Parametervektor
  - Die Parameter werden durch den Posterior gut geschätzt
  - Abstand zwischen $\beta_{0,j}$
  - Steigung (war bereits ähnlich)
  - Streuung $\sigma_{0,j}$ nähert sich an
- Effekt durch Hierarchie
  - Einschränkung von $\beta_{0,monkey}$
- Kopplung
  - *Kaum* Kopplung zwischen den Ebenen
  - Große Streuung in *zbetamu*
    - große Freiheit für *beta* in jeder Kategorie
  - Dennoch leichte Kopplung:
    - Einengung $\beta_{1,\text{monkey}}$
- Problem

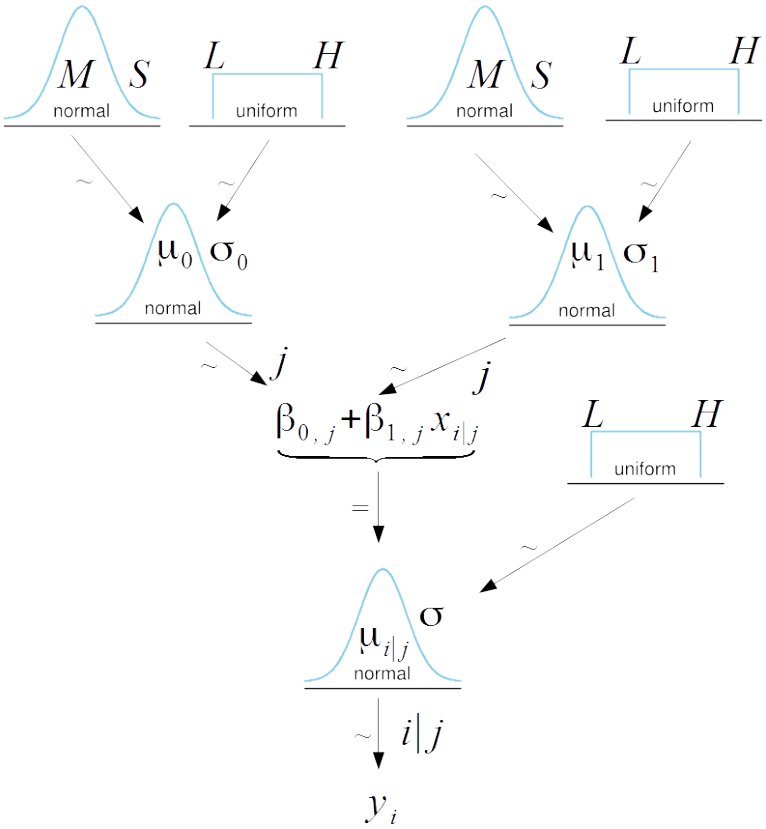

Abbildung 5.9: Hierarchisches Lineares Modell mit Kategorien, nach: Kruschke: Doing Bayesian Data Analysis, 2nd. AP (2015)

  - Es steckt kaum Information in den **nur 2** Kategorien
  - die obere Hierarchieebene wird gar nicht ausgenutzt
    - zu viele Freiheitsgrade
  - $\beta_{0,\sigma}$ und $\beta_{1,\sigma}$ sind daher beliebig
  - $\beta_{0,\mu}$ und $\beta_{1,\mu}$ sind ebenfalls sehr breit verteilt
- Folge
  - Hamiltonian MCMC Posterior-Spaziergang stößt an Rand der Verteilung
    - Gradient nicht mehr stetig
    - Ablehnung von Sprüngen
  - $\Rightarrow$ Warnmeldung
- Lösung des Problems
  - Mehr Daten nötig
    - Anzahl der Kategorien $\geq$ Anzahl Parameter in oberster Hierarchieebene
    - Elefanten, Affen, Nagetiere, Katzen, ...

Mehrere Kategorien mit Hierarchischem Modell

- Anderes Beispiel: Gruppe von Versuchspersonen

  - Versuchspersonen als (ähnliche, aber leicht unterschiedliche) Individuen (j)
  - individuelle Einzelergebnisse (i)
- Ergebnis
  - Individuen zeigen ähnliches Verhalten
    * erlaubt sind (hier im Beispiel) individuelle Steigerung und individuelles Level
  - Hierarchisches Modell verbindet Individuen als Mitglieder der Gruppenpopulation
    * Summe der Individuen erlaubt Rückschlüsse auf Eigenschaften der Gruppe an sich
    * Gruppeneigenschaften erlauben Rückschlüsse auf einzelne Individuen
  - Obacht
    * Anpassung nur möglich, wenn ausreichend Daten (hier Kategorien) verfügbar

5.23.6 *Generalisierte* Lineare Modelle

Bereits gesehen in GLM (Kapitel 2, ab Seite 24)

- Link Funktion in GLM
- Anwendung einer Link-Funktion mit Bayes-Statistik
  - Psychometrische Daten *logistische Regression*: Psignifit

Funktionen in Stan

- Link-Funktionen
  - `logit(x)`
  - `inv_logit(x)`
  - `inv_cloglog(x)`
- Verteilungen
  - `y ~ bernoulli_logit( alpha + beta * x )`
  - `y ~ bernoulli( inv_logit( alpha + beta * x[n] ) )`
    * `# equivalent, but less efficient and less arithmetically stable`

5.23.7 '*Take-Home*': GLM und Bayes-Statistik

- **PyStan-MCMC** und **GLM-IRLS**-Ergebnisse stimmen sehr gut überein
- Methode der Wahl hängt von Zweck der Auswertung ab
- Vorteile durch Bayes-Statistik können ausgenutzt werden
  - Posterior-Verteilung der Zufallsvariablen des Parametervektors
  - Vergleich von Parametern
  - Zugriff auf übergeordnete Gruppenvariablen in hierarchischem Modell
  - Einblick in Verteilungen
- flexiblere Modelle möglich
- Berechnung aufwändiger

6 Literatur

(Links unterlegt)

6.0.1 (Frequentistische) Statistik

- Fahrmeir, Künstler, Pigeot, Tutz: Statistik. Springer-Verlag Berlin Heidelberg, 6. Auflage (2007)
- Stahel: Statistische Datenanalyse. Vieweg&Sohn, 5. Auflage (2008)

6.0.2 Python und Notebooks

- A Crash Course in Python for Scientists
- zu `matplotlib`
- zu `Pandas`

6.0.3 Generalized Linear Models

- GLM statistics course by Tom Wallis
- Dobson, Barnett: An Introduction to Generalized Linear Models. Chapman&Hall/CRC, 3rd ed. (2008)
- McCullagh, Nelder: Generalized Linear Models. Chapman&Hall/CRC, 2nd ed. (1989)

6.0.4 PCA

- Hyvärinen, Hurri, Hoyer: Natural Image Statistics - A Probabilistic Approach to Early Computational Vision. Springer (2009)
- Abdi, Williams: Principal component analysis. Wiley (2010)
- Novembre et.al.: Genes mirror geography within Europe. Nature 456(7218) 98–101 (2008). doi:10.1038/nature07331
- Turk, Pentland: Eigenfaces for Recognition; JCogNeurosci Vol3.1 (1991)
- Blanz, Vetter: Face recognition based on fitting a 3D morphable model; IEEE Transactions on Pattern Analysis and Machine Intelligence 25, 9 (2003)
- Blanz, Vetter: A Morphable Model for the Synthesis of 3D Faces. T. SIGGRAPH' (1999) Conference Proceedings

6.0.5 ICA

- Stone: Independent Component Analysis, MIT press (2004)
- Hyvärinen, Hurri, Hoyer: Natural Image Statistics, Springer (2009)
- Hyvärinen, Oja: Independent Component Analysis: Algorithms and Applications. Neural Networks, 13(4-5) 411-430 (2000)
- Hyvärinen Homepage

6.0.6 Bayes

- J. K. Kruschke: Doing Bayesian Data Analysis, 2nd. A Tutorial with R, JAGS and Stan. Academic Press (2015)
- Beispielprogramme zum Kruschke-Buch
- Stan documentation
- PyStan documentation
- Stan's source-code repository
- über die Wahl von Priors

Psignifit

- Schütt, Harmeling, Macke and Wichmann: Painfree and accurate bayesian estimation of psychometric functions for (potentially) overdispersed data. Vision research, 122:105–123 (2016)
- Software: `https://github.com/wichmann-lab/psignifit`
- Kontsevich and Tyler: Bayesian Adaptive Estimation of Psychometric Slope and Threshold. Vision Research, 39(16):2729–2737 (1999)
- Cavagnaro, Pitt and Myung: Model discrimination through adaptive experimentation. Psychonomic Bulletin & Review, 18(1):204–210 (2011)
- Shen and Richards: An updated maximum-likelihood procedure: Thresholds, slopes, and lapses of attention. Journal of the accoustical society of america, 132(2):957–967 (2012)
- Watson and Pelli: Quest: A bayesian adaptive psychometric method. Perception & psychophysics, 33(2):113–120 (1983)
- Prins: The psi-marginal adaptive method: How to give nuisanceparameters the attention they deserve (no more, no less). Journal of Vision, 13(7):3–3 (2013)
- Lesmes, Lu, Baek and Albrigh:. Bayesian adaptive estimation of the contrast sensitivity function: The quick CSF method. Journal of Vision, 10(3):17–17 (2010)
- Watson: QUEST+: A general multidimensional Bayesian adaptive psychometric method-Watson. Journal of Vision, 17(3):10–10 (2017)

- Stefanie Otto: Vergleichende Simulation adaptiver, psychometrischer Verfahren zur Schätzung von Wahrnehmungsschwellen. Magisterarbeit (2009)

6.0.7 Kausalität

- Peters, Janzing, Schölkopf: Elements of Causal Inference. MIT press (2017)