12. Übungsblatt zur Vorlesung **WS 2017/2018**
Statistische Methoden der Datenanalyse **Prof. W. Rhode**

Annika Burkowitz, Sebastian Bange, Alexander Harnisch

Aufgabe 36

Rechnung angehangen.

A36

a) $g = A f \quad , \quad A = \begin{pmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{pmatrix}$

b) $f = A^{-1} g \quad , \quad A^{-1} =: B$

$$B = (1-2\varepsilon)^{-1} \begin{pmatrix} 1-\varepsilon & -\varepsilon \\ -\varepsilon & 1-\varepsilon \end{pmatrix}$$

$$f_{1/2} = (1-2\varepsilon)^{-1} \left((1-\varepsilon) g_{1/2} - \varepsilon g_{2/1} \right)$$

c) Poisson => Fehler $\sqrt{n}$ => Varianz n

Also $V[g] = \begin{pmatrix} g_1 & 0 \\ 0 & g_2 \end{pmatrix}$

$$V[f] = B\, V[g]\, B^T = (1-2\varepsilon)^{-2} \begin{pmatrix} (1-\varepsilon)^2 g_1 + \varepsilon^2 g_2 & -\varepsilon(1-\varepsilon)(g_1+g_2) \\ -\varepsilon(1-\varepsilon)(g_1+g_2) & (1-\varepsilon)^2 g_2 + \varepsilon^2 g_1 \end{pmatrix}$$

d) $f_1 = 203{,}875, \; f_2 = 165{,}125 \; , \; f_1 + f_2 = 369$ ✓

$$V[f] = \begin{pmatrix} 255{,}765625 & -57{,}890625 \\ -57{,}890625 & 217{,}015625 \end{pmatrix}$$

$\sigma_{f_i} = \sqrt{(V[f])_{ii}} \quad , \quad \sigma_{f_1} = 15{,}9926741$

$\sigma_{f_2} = 14{,}7314502$

$$\rho = \frac{\mathrm{cov}(f_1, f_2)}{\sigma_{f_1} \cdot \sigma_{f_2}} = -0{,}2202532433779 6553$$

e) $f = (262, 107)^T \quad f_1 + f_2 = 369$ ✓

$V[f] = \begin{pmatrix} 2476 & -2214 \\ -2214 & 2321 \end{pmatrix}$ $\quad \sigma_{f_1} = 49{,}75942122$

$\sigma_{f_2} = 48{,}17675788$

$\rho = -0{,}92355917297586 76$

Durch Fehlkategorisierungswahrscheinlichkeit wird das Ergebnis zufälliger, dies spiegelt sich in den höheren Fehlern und der (negativen) Korrelation wieder.

f) Für $\varepsilon \to 0{,}5$ wird $V[f]$ singulär bzw. $\rho \to 1$.

Die Daten sind also vollkommen nutzlos, sie beinhalten keine Information.

Aufgabe 37

Rechnung angehangen.

a)

Es handelt sich um einen Messprozess, bei dem mit Wahrscheinlichkeit ε ein Ereignis fälschlicherweise in den nächst höheren oder niedrigeren Bin kategorisiert wird, und daher mit Wahrscheinlichkeit $1 - 2\varepsilon$ in den korrekten (bzw. $1 - \varepsilon$ für die Randbins).

c)

Die Eigenbasis hat, wie immer, den Vorteil, dass die Matrix hier diagonal ist daher lässt sie sich hier sehr leicht und stabil invertieren.

d)

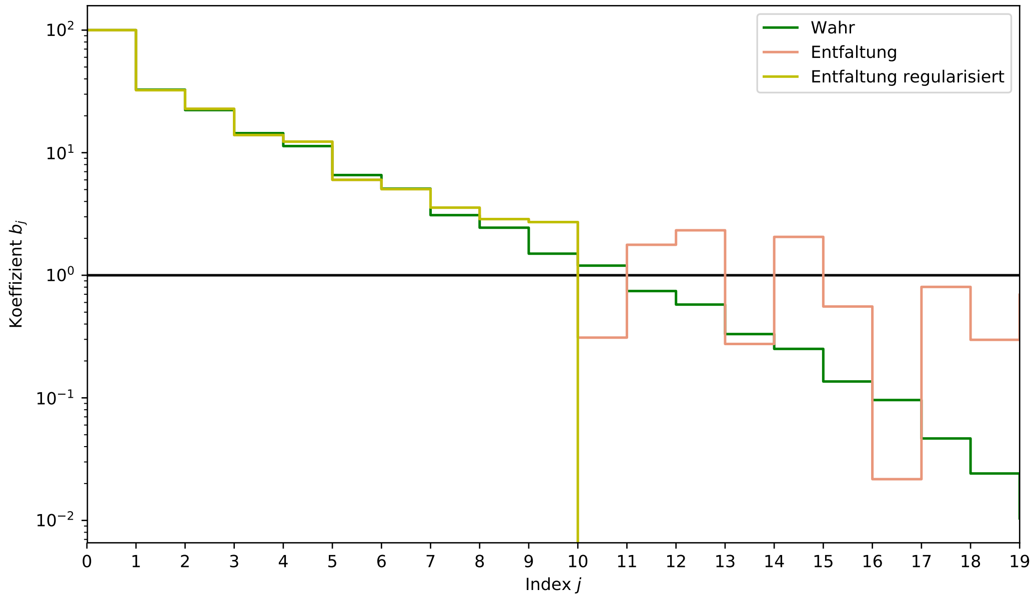

Abbildung 1: Darstellung der Koeffizienten in Eigenbasis.

Die schwachen Eigenwerte kleiner als 1 sind, wie in Abbildung 1 zu sehen, sehr instabil. Diese werden zur Regularisierung deswegen auf null gesetzt.

e)

Die Lösung mit Regularisierung hat, wie in Abbildung 2 gut erkennbar, deutlich geringere Oszillationen. Außerdem sind ihre Fehler ebenfalls deutlich geringer.

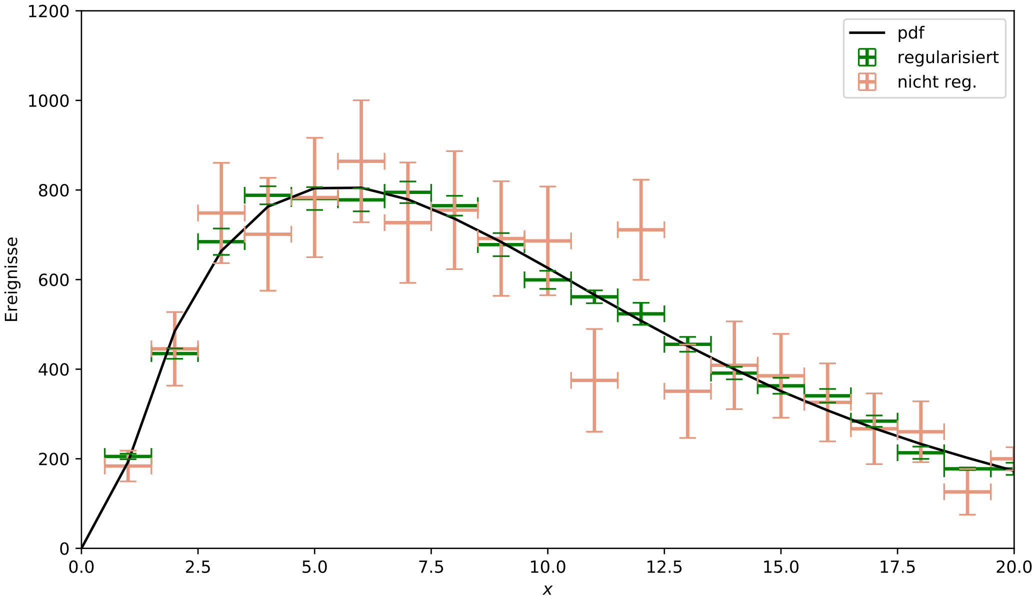

Abbildung 2: Ergebnis der Entfaltung mit und ohne Regularisierung.

A 37

c) $g = Af$, $A = U \cdot D \cdot U^T$ => U hat die EVs als Spalten

$\Leftrightarrow \underbrace{U^T g}_{\tilde{g}} = \underbrace{U^T U}_{1} \underbrace{D}_{\tilde{A}} \underbrace{U^T f}_{\tilde{f}}$, $\tilde{A} = D$

$\Leftrightarrow \tilde{g} = D \tilde{f}$, , $\tilde{g} = c$, $\tilde{f} = b$ laut Aufgabe

$\Leftrightarrow c = Db$

d) $V[g] = \mathrm{diag}(s)$ (Poisson Fehler)

$\Rightarrow V[c] = U^T V[s] U$

$\Rightarrow V[b] = D^{-1} U^T V[s] U D^{-1}$ $\left((D^{-1})^T = D^{-1} \right)$

e) $f = Ub$

$\Rightarrow V[f] = U V[b] U^T$

Aufgabe 38

a)

In den Histogrammen ist eindeutig erkennbar, dass das Attribut *size* die beste Korrelation mit dem Zielattribut *energy_true* aufweist, daher wird *size* als Attribut zur Entfaltung gewählt.

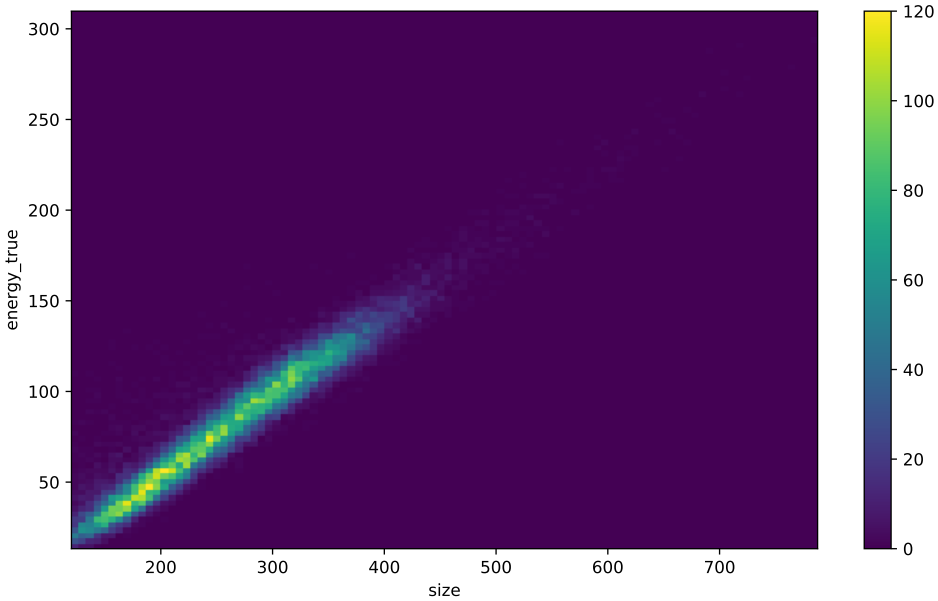

Abbildung 3: Zwei-dimensionales Histogramm für die Attribute *energy_true* und *size*.

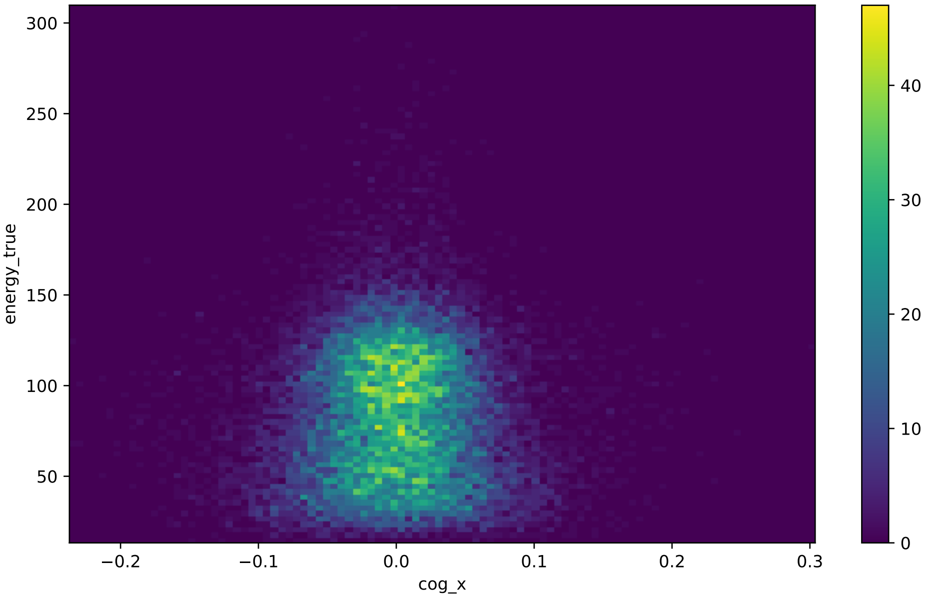

Abbildung 4: Zwei-dimensionales Histogramm für die Attribute *energy_true* und *cog_x*.

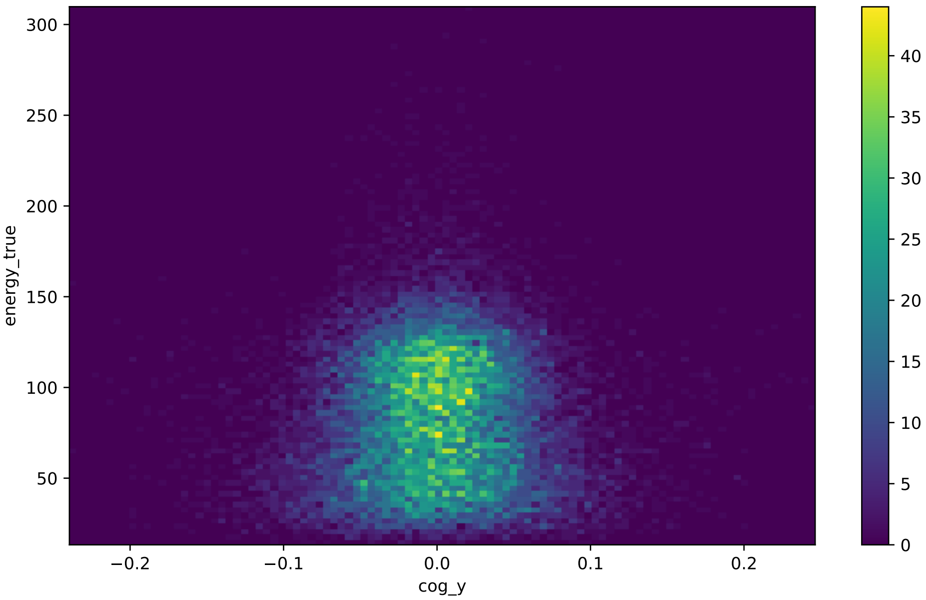

Abbildung 5: Zwei-dimensionales Histogramm für die Attribute *energy_true* und *cog_y*.

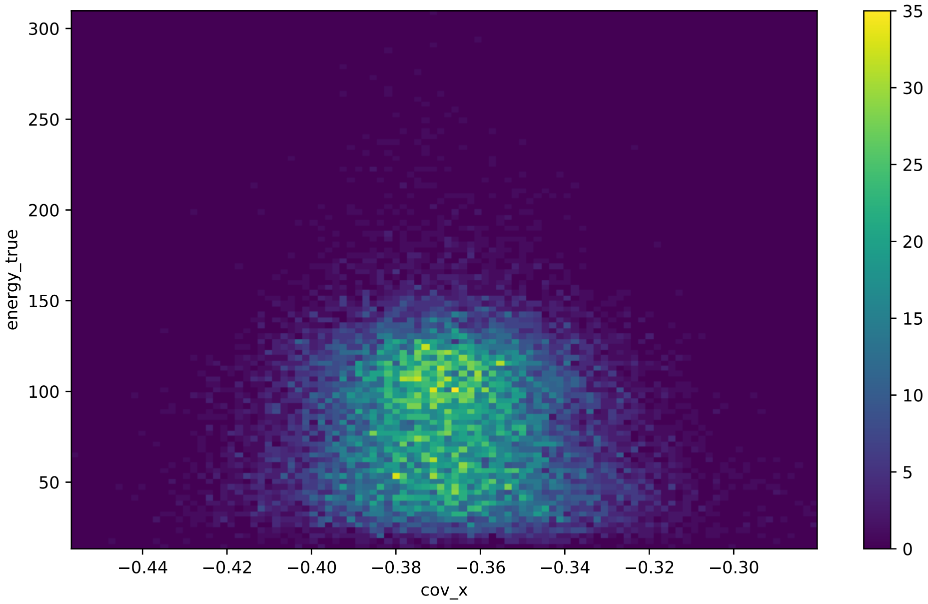

Abbildung 6: Zwei-dimensionales Histogramm für die Attribute *energy_true* und *cov_x*.

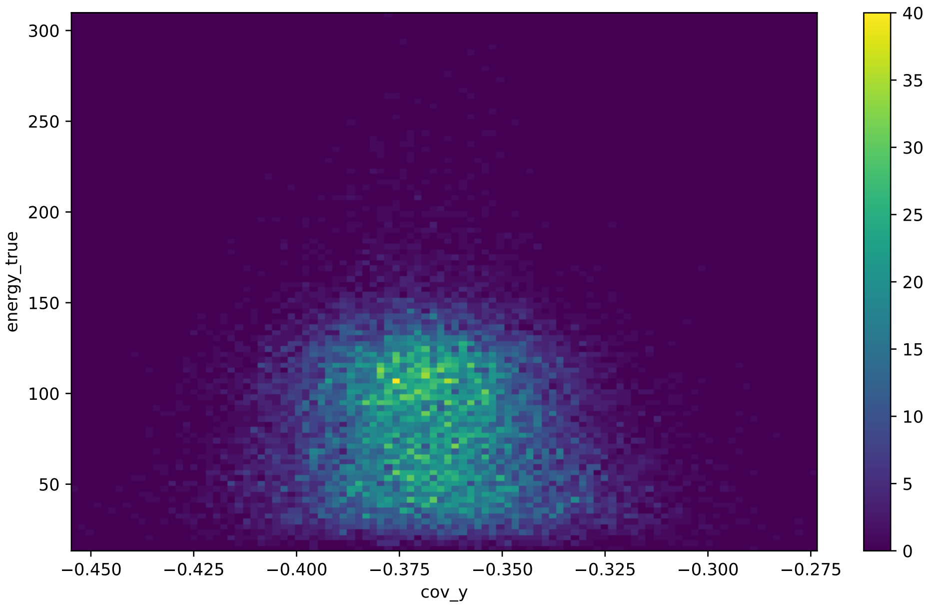

Abbildung 7: Zwei-dimensionales Histogramm für die Attribute *energy_true* und *cov_y*.

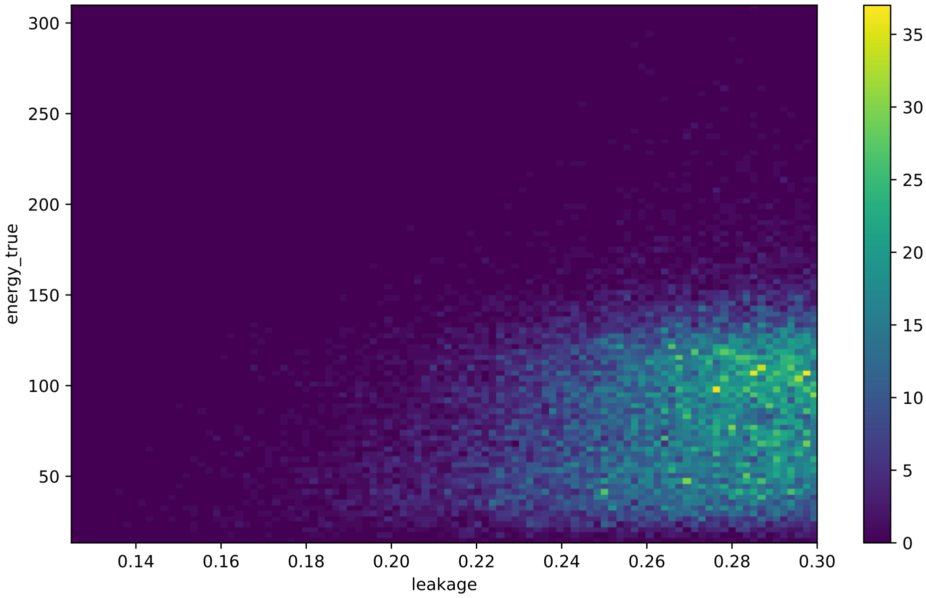

Abbildung 8: Zwei-dimensionales Histogramm für die Attribute *energy_true* und *leakage*.

b)

Die Migrationsmatrix A ist in Abbildung 9 dargestellt.

e)

Leider zu wenig Zeit, wir benutzen stattdessen einfach *scipy* zum Minimieren.

f)

Der Vergleich der verschiedenen Regularisierungsstärken in Abbildung 10 zeigt, dass die Regularisierung mit $\tau = 0{,}001$ am angemessensten erscheint.

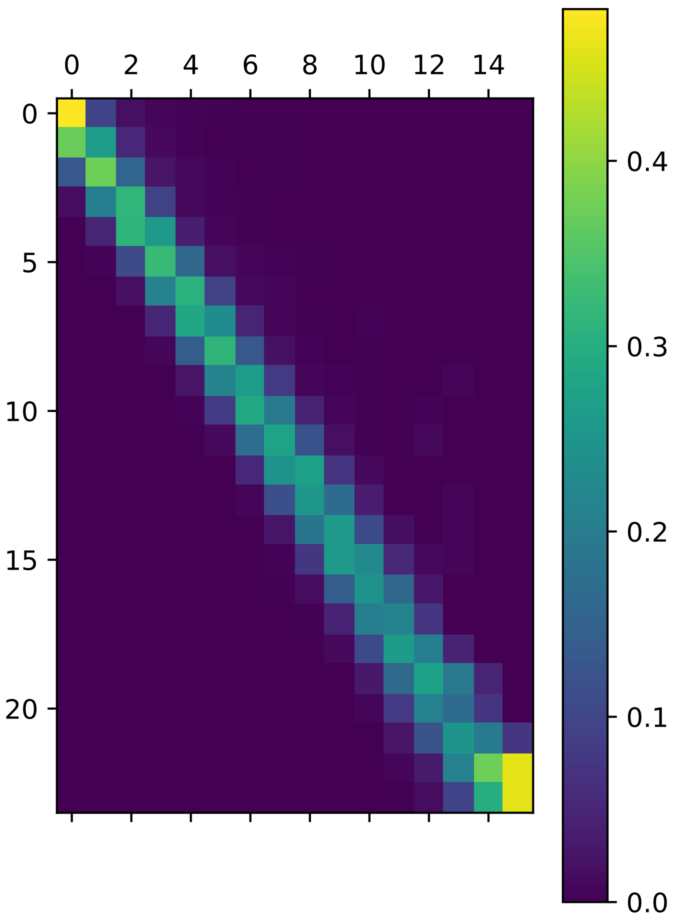

Abbildung 9: Grafische Darstellung der Migrationsmatrix.

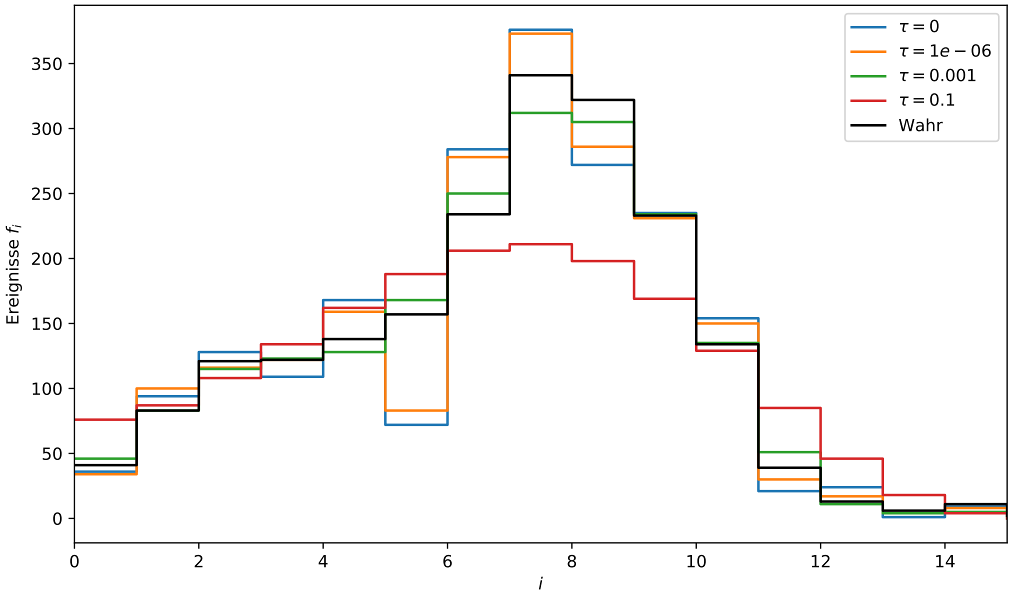

Abbildung 10: Ergebnis der Entfaltung mit verschiedenen Regularisierungsstärken τ.

A38

b) $g = A f$ g: size ; f = true energy
24 bins, 120-500 ; 16 bins 15 bis 200

Also A ist 24 × 16 Matrix

c) $p(k) = \frac{\lambda^k e^{-\lambda}}{k!}$

g_i seien die Messwerte

$\lambda_i = A_i f$ also Likelihood:

$$L = \prod_{i=1}^{24} \frac{\lambda_i^{g_i} e^{-\lambda_i}}{g_i!} = \prod_{i=1}^{24} \frac{(A_i \cdot f)^{g_i} e^{-A_i \cdot f}}{g_i!}$$

$$F = -\log L = \sum_{i=1}^{24} \left[\log(g_i!) - g_i \log(A_i f) + A_i \cdot f \right]$$

d) $F_{reg} = F + \frac{\tau}{2} \| \Gamma f \|^2$

$$\frac{\partial F_{reg}}{\partial f_i} = -\sum_{j=1}^{24} \left(g_j \frac{A_{ji} f_i}{A_j \cdot f} + A_{ji} f_i \right) + \frac{\partial}{\partial f_i} \frac{\tau}{2} \left(\sum_{j=1}^{16} (\Gamma_j f_i)^2 \right)$$

$$= -\sum_{j=1}^{24} A_{ji} \left(f_i - \frac{g_j}{A_j f} \right) + \tau \sum_{j=1}^{16} (\Gamma_j f) \cdot \Gamma_{ji} f_i$$

$$\frac{\partial^2 F}{\partial f_i \partial f_j} = \frac{\partial}{\partial f_j} \left(\frac{\partial F}{\partial f_i} \right) = + \sum_{k=1}^{24} \left[\frac{g_k}{A_{kj} f_j^2} + A_{kj} \delta_{ij} \right]$$
$$+ \tau \left(\sum_{k=1}^{16} \Gamma_{kj} f_j \Gamma_{ki} f_i + \delta_{ij} \sum_{k=1}^{16} (\Gamma_k f) \Gamma_{kj} \right)$$

mit Sicherheit falsch, ich versuch gar nicht erst das zu implementieren/debuggen leider zu wenig Zeit.