

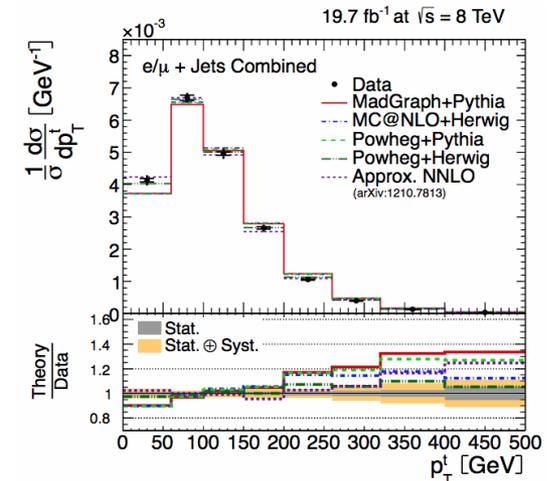
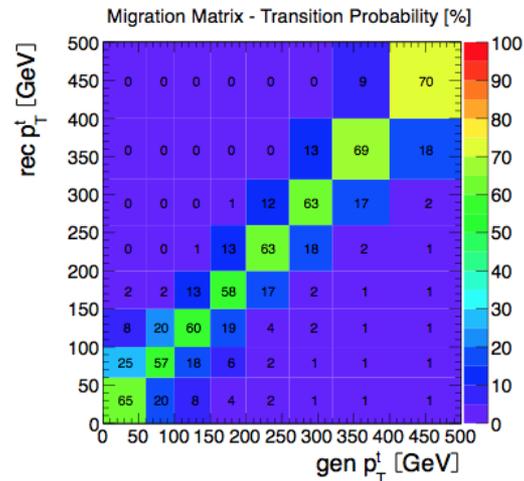
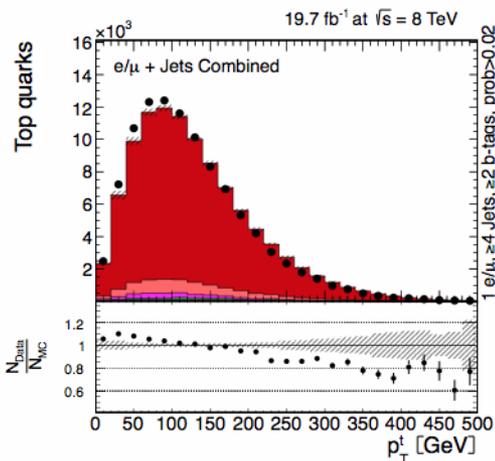
Moderne Methoden der Datenanalyse

Messen und Entfalten

Andreas B.Meyer  

18. Juli 2017

Fakultät für Physik
Institut für Experimentelle Kernphysik - IEKP



Vorlesungsprogramm

- Einführung: Überblick und grundlegende Konzepte (1)
- Einführung: Überblick und grundlegende Konzepte (2)
- Zufallszahlen und Monte-Carlo Methoden

A.MEYER

- Hypothesentests
- Parameterschätzung
- Parameterschätzung (Goodness-of-fit)
- Optimierungs- und Parametrisierungsmethoden

R.ULRICH

- Konfidenzintervalle (1)
- Konfidenzintervalle (2)
- Klassifikation (Supervised Learning und lineare Methoden)
- Klassifikation (ANN and BDT)
- Klassifikation (SVM and Deep Learning)
- Messen und Entfalten
- Systematische Unsicherheiten

A.MEYER

Messen und Entfalten

- Einleitung
- Faltungsintegrale (Beispiele)
- Exkurs: Charakteristische Funktion
- Response Matrix

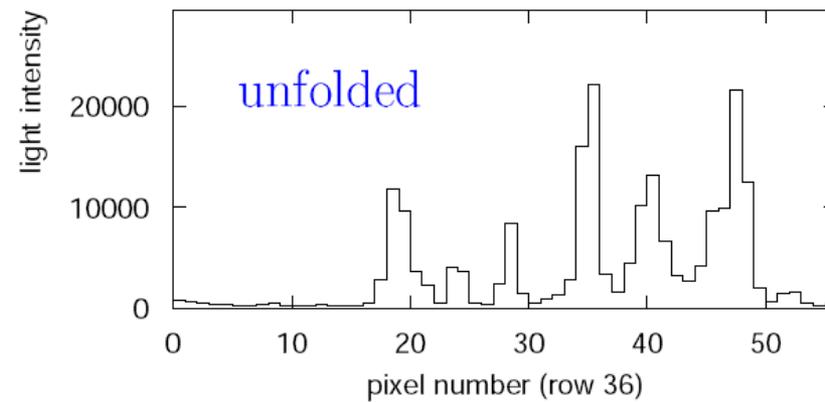
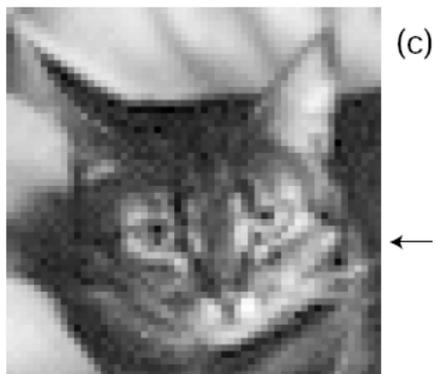
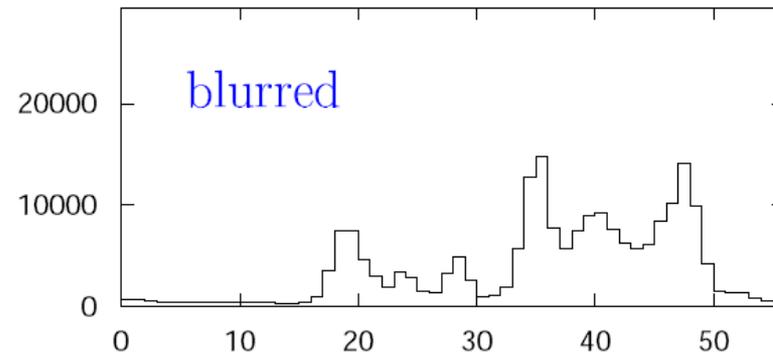
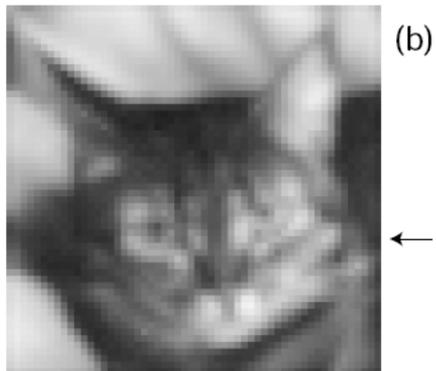
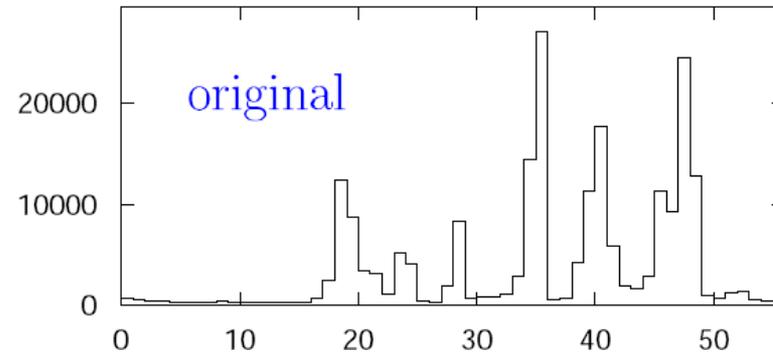
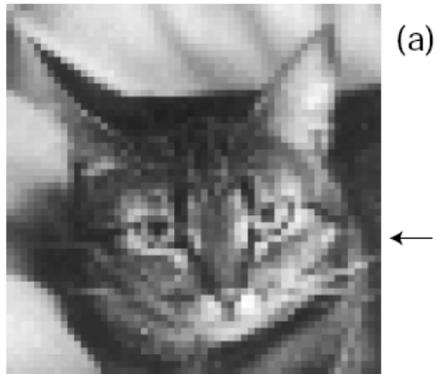
8.1 Faltung

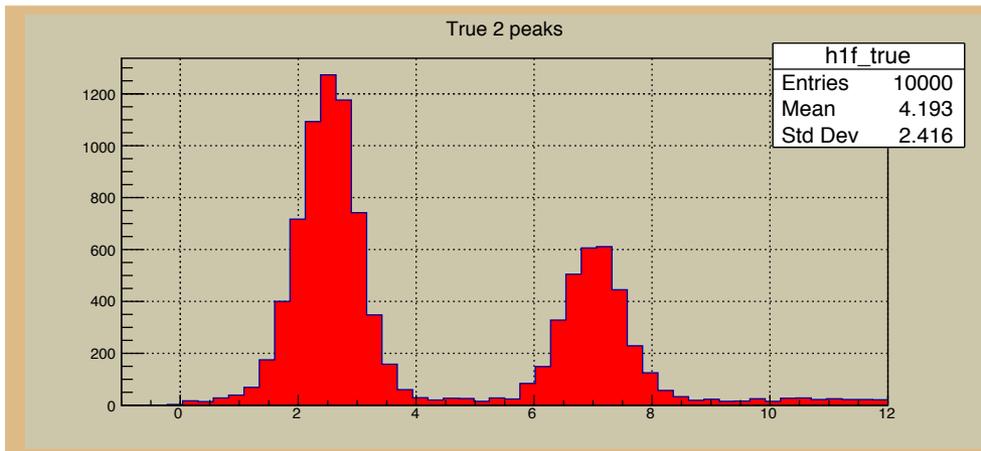
- Matrix-Inversion
- Methoden mit Regularisierung
- Iterative Methoden

8.2 Entfaltung

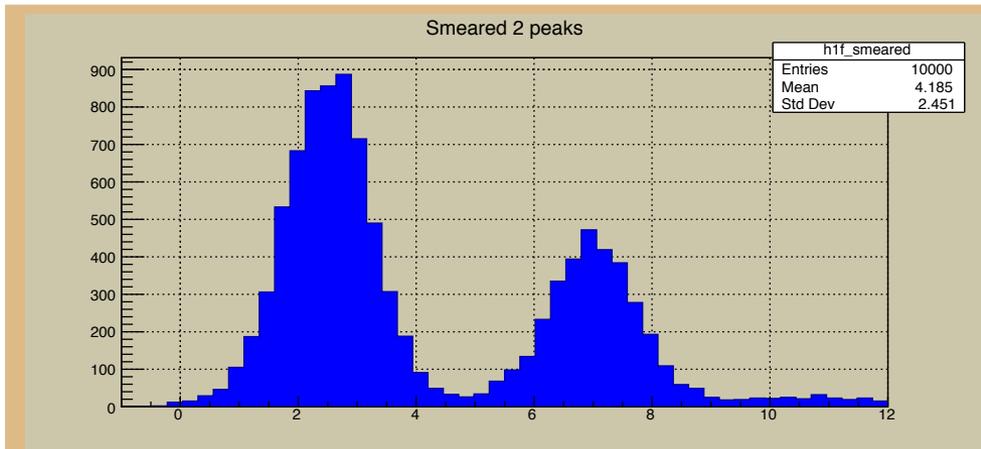
■ Literatur:

- G.Cowan, *Statistical Data Analysis (Kap.10 und 11)*
- Bohm, Zech, *Introduction to Statistics and Data Analysis for Physicists (Kap.8)*
- Blobel, Lohrmann, *Statistische und numerische Methoden der Datenanalyse (Kap.11)*
- G.D'Agostini, *A multidimensional unfolding method based on Bayes' theorem*, NIM A 362 (1995)
- S.Schmitt, *TUnfold, an algorithm for correcting migration effects in high energy physics*, JINST 7 (2012) T10003.

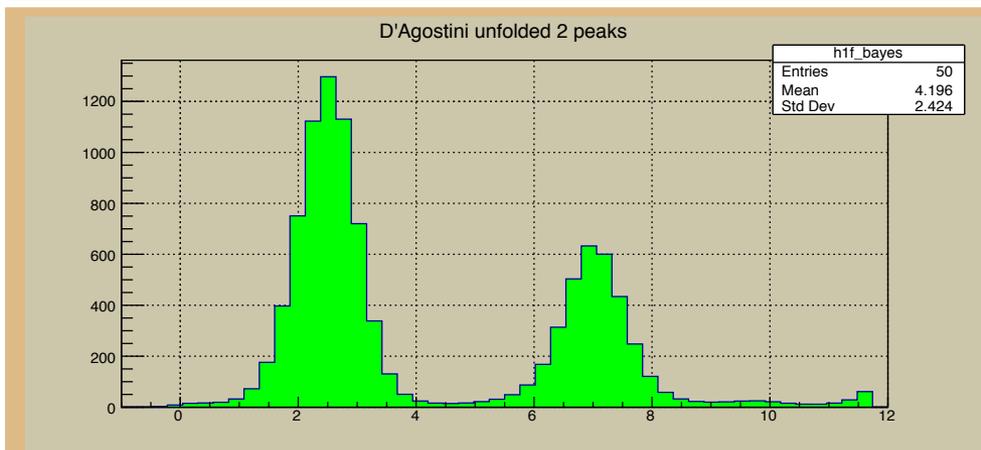




Wahre Verteilung
(messbar in perfektem Experiment)



Gemessene Verteilung
(enthält Detektoreffekte)



Entfaltete Verteilung
(Detektoreffekte herauskorrigiert)

N.B. hier mit Regularisierung

■ Ziel experimenteller Messungen:

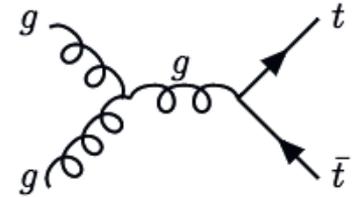
- Ermittlung und Dokumentation allgemeingültiger Beobachtungen
→ unabhängig von den technischen Eigenschaften des konkreten Experiments
- Vergleich mit anderen Experimenten
- Vergleich mit Experiment-unabhängigen theoretischen Vorhersagen

■ Durchführung einer Messung:

- Korrektur von experimentellen, Experiment-spezifischen Effekten (i.a. ermittelt durch Hilfsmessungen):
 - Subtraktion von Untergrund (z.B. falsche Ereignisse und Rauschen)
 - Kalibration → Normierung auf universelle Skala (z.B. Länge [m], Energie [GeV])
 - Nachweiswahrscheinlichkeit → Verlust oder unvollständige Messung von Signalen
 - Akzeptanzkorrektur → Korrektur für Begrenzungen des experimentell zugänglichen Bereichs
 - Auflösungseffekte → Verschmierung (engl. smearing, migration) und Verschiebung (engl. bias) gegenüber ursprünglicher Verteilung
- Bestimmung der systematischen Unsicherheiten dieser und anderer der Messung zugrundeliegenden Annahmen

Messung = Bestimmen von Korrekturen und Unsicherheiten

Vergleich von Experimenten und Theorie



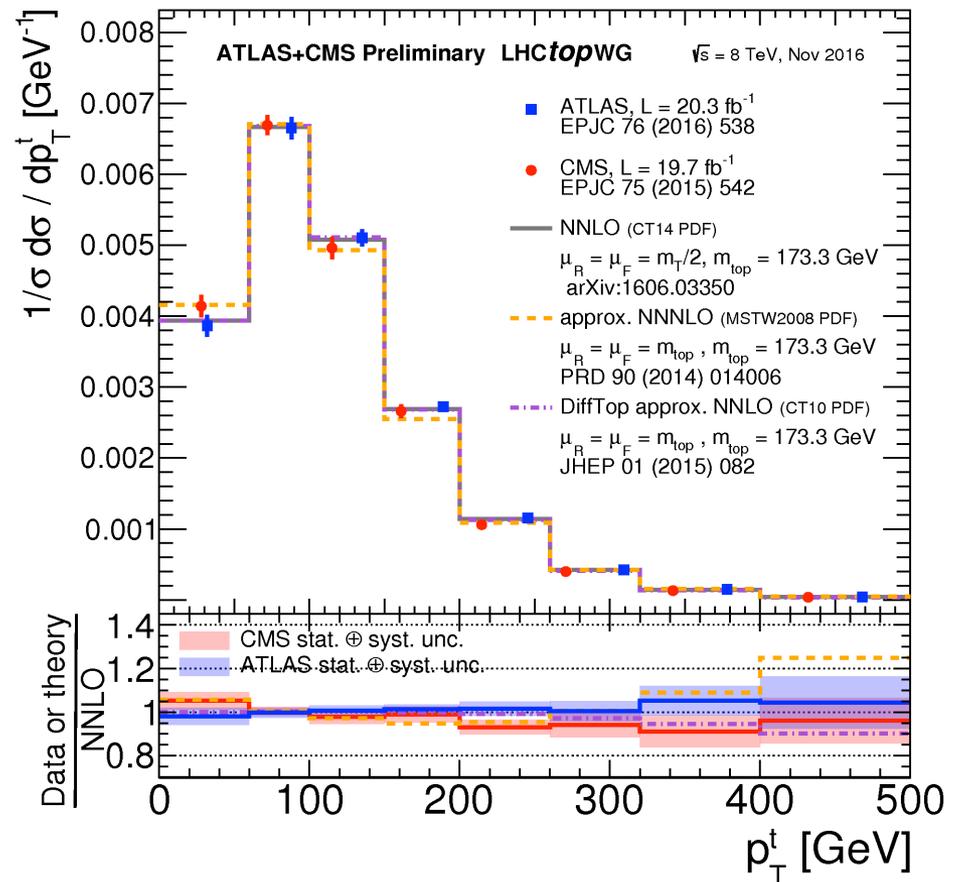
■ Beispiel:
Differentieller Wirkungsquerschnitt
als Funktion des Transversal-
Impulses von Top-Quarks
am LHC

■ Vergleich:

- CMS Experiment
- ATLAS Experiment
- Verschiedene theoretische Vorhersagen

■ Ziel: Beurteilung:

- Sind die Experimente konsistent miteinander ?
- Welche Vorhersage ist am besten?



<https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCTopWGSummaryPlots>

Faltung ↔ Entfaltung

■ Faltung (engl. convolution)

- Experimentell gemessene Verteilungen können als Faltung der all-gemeingültigen wahren Verteilungen mit Experiment-spezifischen Detektoreffekten beschrieben werden.

Wahre
Verteilung

$$f(t) \rightarrow g(s)$$

(In einem Experiment)
gemessene Verteilung

■ Entfaltung (engl. unfolding, deconvolution): Inverse Transformation

Gemessene
Verteilung

$$g(s) \rightarrow f(t)$$

Entfaltete
Verteilung

- Entfaltungen gehören i.a. zu den sogenannten “ill-posed problems”: Kleine Fluktuationen (z.B. statistische) in der gemessenen Verteilung können zu großen Schwankungen in der wahren Verteilung führen
- Lösungsansatz: Dämpfung der Schwankungen durch Regularisierung, d.h. Hinzunahme zusätzlicher Annahmen

8.1

FALTUNG

Beispiel: Lebensdauerermessung

■ Wahre Verteilung:

$$f(t) = \frac{1}{\tau} \exp\left(-\frac{t}{\tau}\right)$$

■ Detektorauflösung:

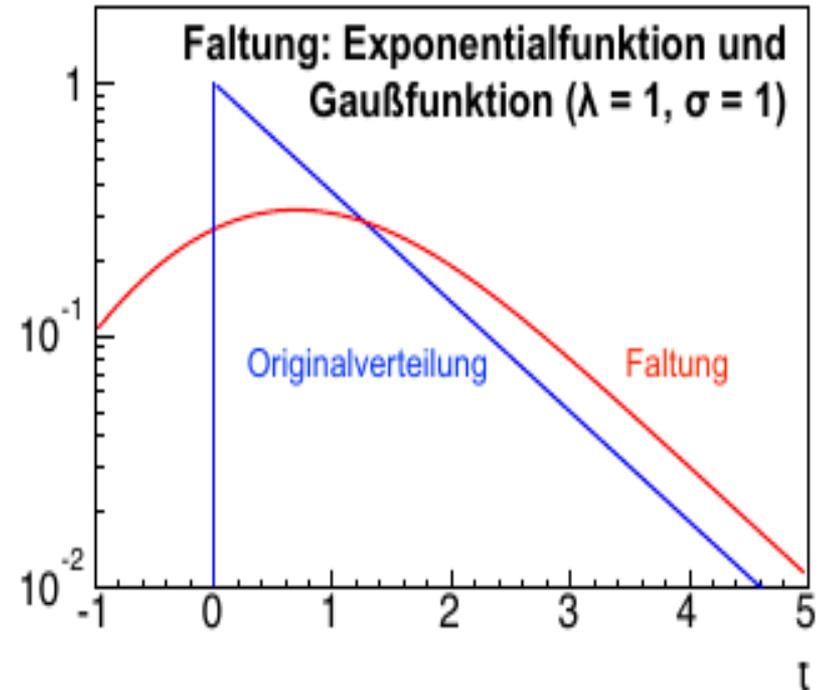
$$g(\Delta t = t_{\text{obs}} - t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\Delta t^2}{2\sigma^2}\right)$$

■ Gemessene Verteilung:

$$h(t_{\text{obs}}) = \int f(t)g(t_{\text{obs}} - t) dt$$

$$=: f \otimes g(t_{\text{obs}}) \quad \text{“Faltung”}$$

$g(t_{\text{obs}}-t)$ beschreibt Wahrscheinlichkeit t_{obs} zu messen, wenn t wahr



Messung einer Verteilung bei begrenzter Auflösung: Gewichtung der Beiträge von $f(t)$ mit $g(t_{\text{obs}}-t)$, also aus dem Intervall Δt

Faltung als Summe von Zufallszahlen

- Die PDF der Summe u zweier zufallsverteilter Größen x und y mit PDFs $f(x)$ und $g(y)$ ist gegeben durch das Faltungsintegral:

$$\begin{aligned}h(u = x + y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x) \cdot g(y) \cdot \delta(u - x - y) \, dx dy \\ &= \int_{-\infty}^{\infty} f(x) \cdot g(u - x) \, dx \\ &= \int_{-\infty}^{\infty} f(u - y) \cdot g(y) \, dy\end{aligned}$$

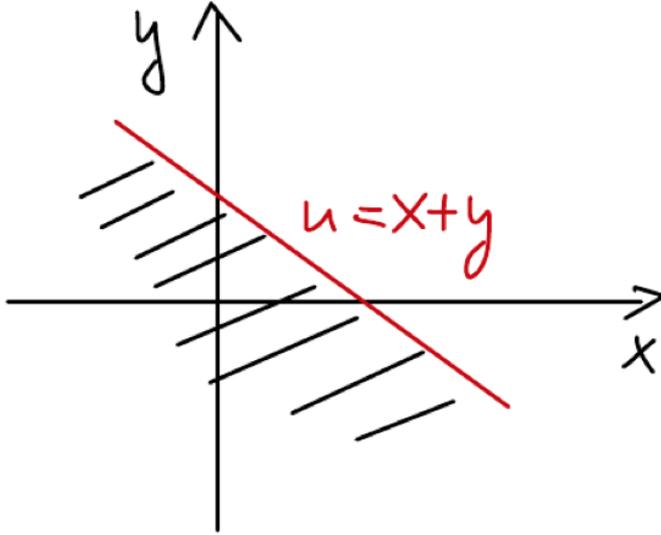
Schreibweise: $h = f \otimes g$

Faltungsintegral

Zwei unabhängige Zufallsvariable x und y , d.h. $f_U(x,y) = f_X(x) f_Y(y)$

Berechne kumulative Verteilungsfunktion für $x+y < u$:

$$\begin{aligned} F_u(u) &= P(x + y < u) = \int_{-\infty}^{\infty} f_x(x) dx \int_{-\infty}^{u-x} f_y(y) dy \\ &= \int_{-\infty}^{\infty} f_y(y) dy \int_{-\infty}^{u-y} f_x(x) dx \end{aligned}$$



$$\begin{aligned} f_u(u) &= \frac{dF_u(u)}{du} = \int_{-\infty}^{\infty} f_x(x) dx f_y(u - x) \\ &= \int_{-\infty}^{\infty} f_y(y) dy f_x(u - y) \end{aligned}$$

Beispiel: Faltung von zwei Box-Verteilungen

■ $u = x + y < 2$

$$f_x(x), f_y(y) = \begin{cases} 1, & \text{für } 0 \leq x \leq 1 \\ 0, & \text{sonst} \end{cases}$$

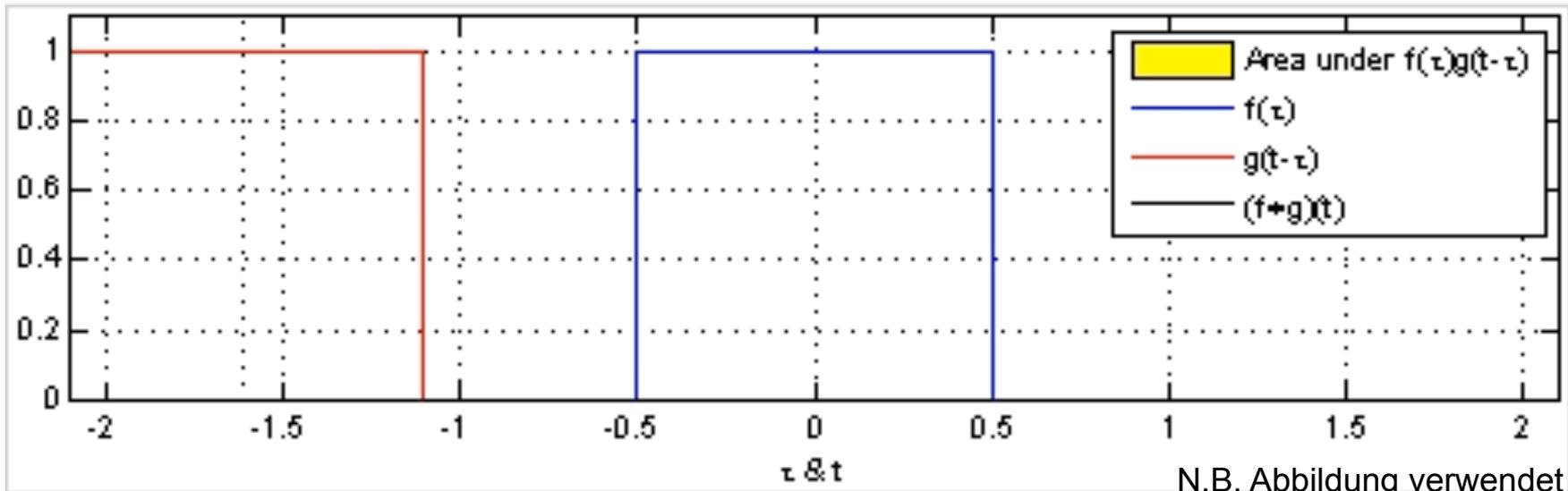
$$\begin{aligned} f(x + y) &= \int_{-\infty}^{\infty} f_x(x) f_y(u - x) dx \\ &= \int_0^1 f_y(u - x) dx = \int_{u-1}^u f_y(v) dv \quad \text{mit } v = u - x \end{aligned}$$

$$0 \leq u < 1 : \quad f_1(u) = \int_0^u f_y(v) dv = \int_0^u 1 dv = u$$

$$1 \leq u < 2 : \quad f_2(u) = \int_{u-1}^1 f_y(v) dv = \int_{u-1}^1 1 dv = 1 - (u - 1) = 2 - u$$

Dreiecksverteilung von 0 bis 2

Beispiel: Faltung von zwei Box-Verteilungen



Quelle: Wikipedia

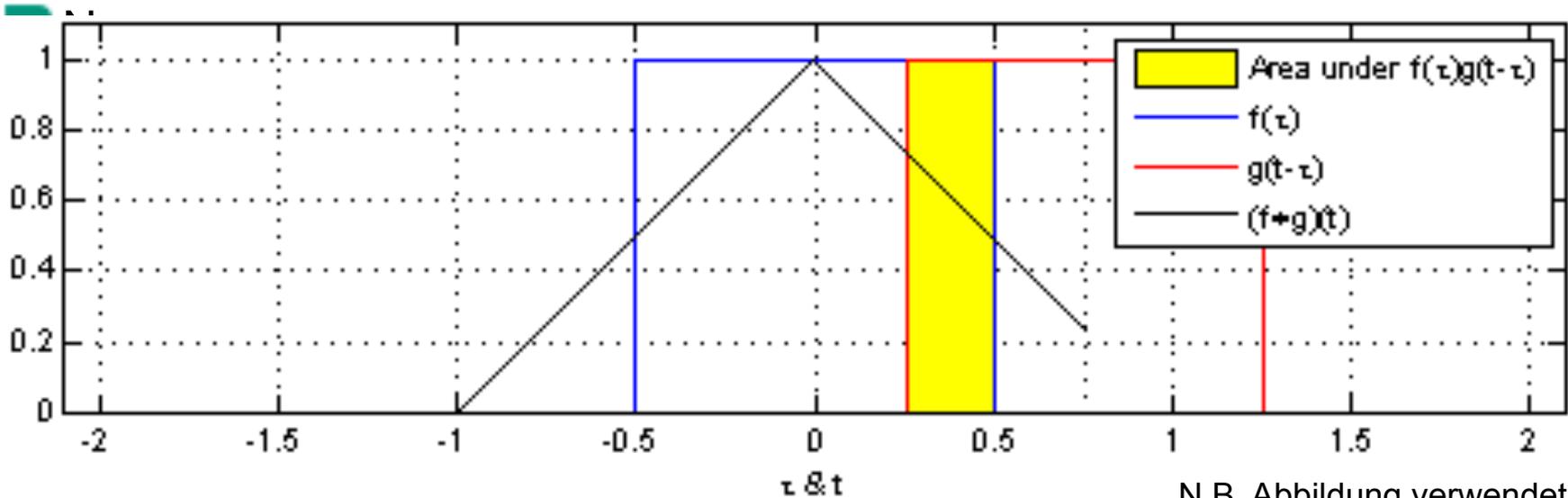
N.B. Abbildung verwendet
andere Werte von x,y
und Normierung

$$0 \leq u < 1 : f_1(u) = \int_0^u f_y(v)dv = \int_0^u 1dv = u$$

$$1 \leq u < 2 : f_2(u) = \int_{u-1}^1 f_y(v)dv = \int_{u-1}^1 1dv = 1 - (u - 1) = 2 - u$$

Dreiecksverteilung von 0 bis 2

Beispiel: Faltung von zwei Box-Verteilungen



Quelle: Wikipedia

N.B. Abbildung verwendet andere Werte von x, y und Normierung

$$0 \leq u < 1 : f_1(u) = \int_0^u f_y(v)dv = \int_0^u 1dv = u$$

$$1 \leq u < 2 : f_2(u) = \int_{u-1}^1 f_y(v)dv = \int_{u-1}^1 1dv = 1 - (u - 1) = 2 - u$$

Dreiecksverteilung von 0 bis 2

Beispiel: Gleichverteilung \otimes Gaußverteilung

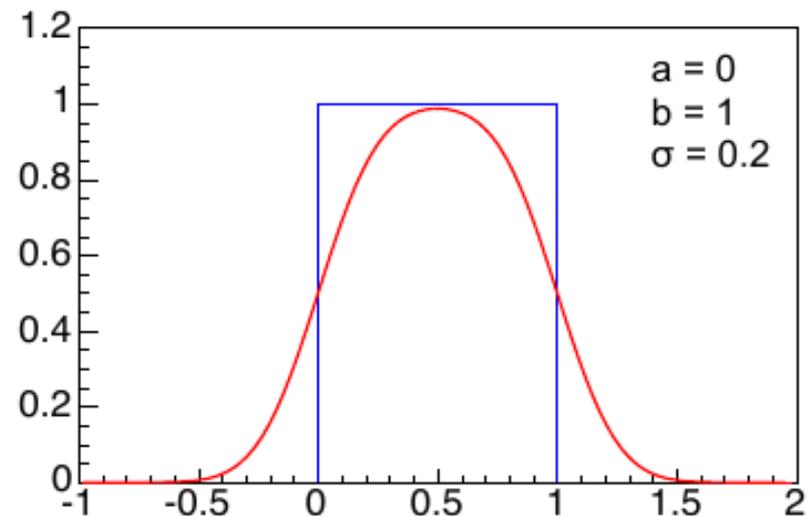
$$f_x(x) = \begin{cases} \frac{1}{b-a} & \text{für } a \leq x \leq b \\ 0, & \text{sonst} \end{cases}$$

$$f_y(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{y^2}{2\sigma^2}\right)$$

$$\begin{aligned} f(u) &= \frac{1}{\sqrt{2\pi}\sigma} \int_a^b \frac{1}{b-a} \exp\left(-\frac{(u-x)^2}{2\sigma^2}\right) dx \\ &= \frac{1}{b-a} \left(\Psi_0\left(\frac{b-u}{\sigma}\right) - \Psi_0\left(\frac{a-u}{\sigma}\right) \right) \end{aligned}$$

$$\text{mit } \Psi_0(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt$$

Verteilungsfunktion der “standardisierten Normalverteilung” (d.h. der integrierten Gauß-Funktion für $\sigma = 1$ und $\mu = 0$)



Beispiel: Exponentialverteilung \otimes Gaußverteilung

$$f_x(x) = \lambda \exp(-\lambda x)$$

$$f_y(y) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{y^2}{2\sigma^2}\right)$$

Ursprüngl. PDF

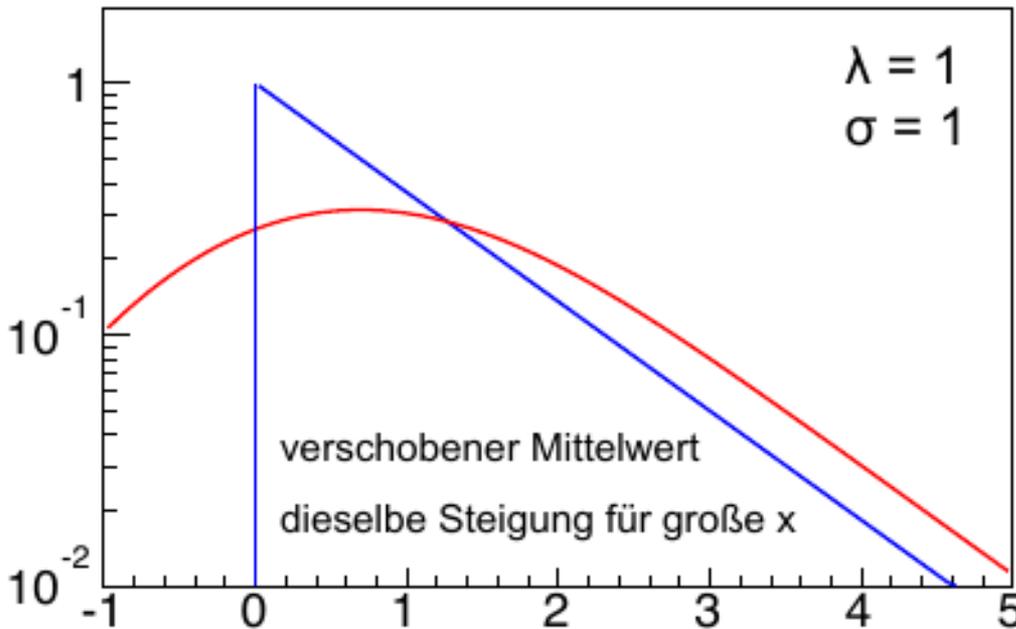
$$f_u(u) = \frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^u \exp\left[-\lambda(u-y) - \frac{y^2}{2\sigma^2}\right] dy = \frac{1}{\sqrt{2\pi\sigma}} \cdot \lambda \exp[-\lambda u]$$

$$\cdot \exp\left[\frac{\lambda^2\sigma^2}{2}\right]$$

Normierung

$$\cdot \int_{-\infty}^u \exp\left[-\frac{(y - \lambda\sigma^2)^2}{2\sigma^2}\right] dy$$

Verschiebung



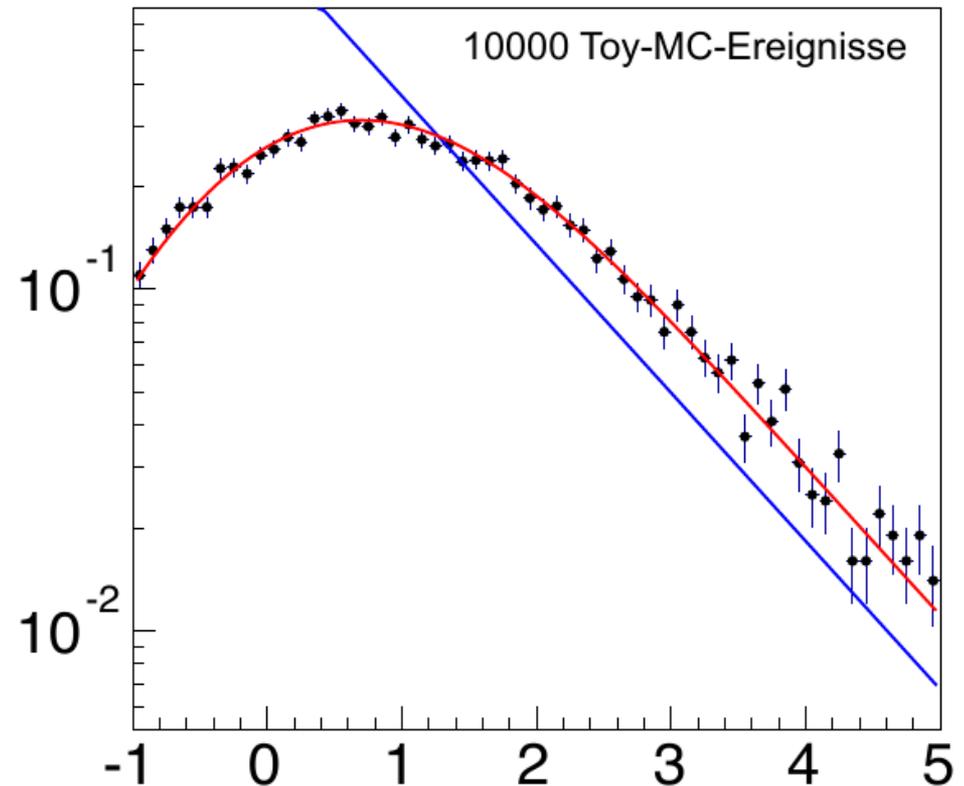
Zerfallskonstante kann für große u direkt abgelesen werden

- Mit der Monte-Carlo Methode sind Faltungen ganz einfach:
 - Ziehe x aus PDF $f_x(x)$ und y aus PDF $f_y(y)$
 - Bilde Summe $u=x+y$
 - $f_x(x) \otimes f_y(y)$ entspricht Häufigkeitsverteilung $h(u)=h(x+y)$
- Monte Carlo wird insbesondere für die Simulation des Detektorverhaltens verwendet
 - Kombination (=Summe) vieler einzelner Effekte (zufälliger Verteilungen)
 - Zentraler Grenzwertsatz: Für große Anzahl von Summanden folgt Summe einer Gauß-Verteilung unabhängig von der Form der einzelnen PDF
- Detektorsimulation erfordert viele sukzessive Faltungen: MC löst hochdimensionales Faltungsintegral → nur mit MC Methode möglich

- Analytische Methoden:
 - direktes Ausrechnen des Faltungsintegrals (ggf. numerisch)
 - Produkt der charakteristischen Funktionen

- Monte Carlo Methode:
 - Summe von Zufallszahlen x, y aus PDF $f_x(x), f_y(y)$
 - Normierte Häufigkeitsverteilung von $u = x+y$ entspricht $f_u(u)$.

Vergleich: analytische und MC-Faltung



- Detektorsimulation erfordert viele sukzessive Faltungen: MC löst hochdimensionales Faltungsintegral → nur mit MC Methode möglich

- Alternative Beschreibung der Eigenschaften von Zufallsvariablen
- Die charakteristische Funktion $\phi_x(k)$ für die Wahrscheinlichkeitsdichte $f(x)$ ist der Erwartungswert von e^{ikx}

$$\phi_x(k) = \langle \exp(ikx) \rangle = \int_{-\infty}^{\infty} e^{ikx} f(x) dx$$

- Eigenschaften:
 - $\phi_x(k)$ ist Fourier-Transformierte der PDF $f(x)$ → eindeutige Abbildung

- Rücktransformation:

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_x(k) e^{-ikx} dk$$

- Nützlich für analytische Lösungen von Problemen:
Summen von Zufallszahlen können als Produkte von Charakteristischen Funktionen geschrieben werden.

- Charakteristische Funktion der Summenverteilung von $u=x+y$:

$$\begin{aligned}\phi_u(k) &= \langle \exp(iku) \rangle = \int \exp(iku) f(u) du \\ &= \int \int \exp(ik(x+y)) f_x(x) f_y(y) dx dy \\ &= \int \exp(ikx) f_x(x) dx \int \exp(iky) f_y(y) dy \\ &= \phi_x(k) \phi_y(k)\end{aligned}$$

- Die Summe von Zufallszahlen $u=\sum x_i$ mit individuellen PDF $f_i(x_i)$ ist das Produkt der charakteristischen Funktionen
- Umgekehrt: Lösung des Faltungsintegrals durch Produktbildung $\Phi_{x+y}=\Phi_x\Phi_y$ und Fourier-Rücktransformation

■ Charakteristische Funktion = Erzeugende der Momente von Verteilungen

■ n-tes zentrales Moment einer PDF $f(x)$ (vergl. 1.Vorl am 25.4.2017)

$$\mu_n = \langle x^n \rangle = \int x^n f(x) dx$$

■ Vergleich mit partiellen Ableitungen der char. Funktion nach k :

$$\frac{\partial^n \phi_x(k)}{\partial k^n} = i^n \int x^n f(x) e^{ikx} dx \rightarrow i^n \mu_n = \frac{\partial^n \phi_x(k)}{\partial k^n} \Big|_{k=0}$$

=> n-te Ableitung entspricht (bis auf Vorfaktor i^n) dem n-ten Moment

■ Beispiel $n = 2$: Varianz

$$\begin{aligned} \frac{\partial^2 \phi_x(k)}{\partial k^2} \Big|_{k=0} &= \int -x^2 e^{ikx} f(x) dx \Big|_{k=0} = -E[x^2] \\ &= -(V[x] + \langle x \rangle^2) \end{aligned}$$

8.2 **ENTFALTUNG**

Faltung ↔ Entfaltung

■ Faltung (engl. convolution)

- Experimentell gemessene Verteilungen können als Faltung der all-gemeingültigen wahren Verteilungen mit Experiment-spezifischen Detektoreffekten beschrieben werden.

Wahre
Verteilung

$$f(t) \rightarrow g(s)$$

(In einem Experiment)
gemessene Verteilung

■ Entfaltung (engl. unfolding, deconvolution): Inverse Transformation

Gemessene
Verteilung

$$g(s) \rightarrow f(t)$$

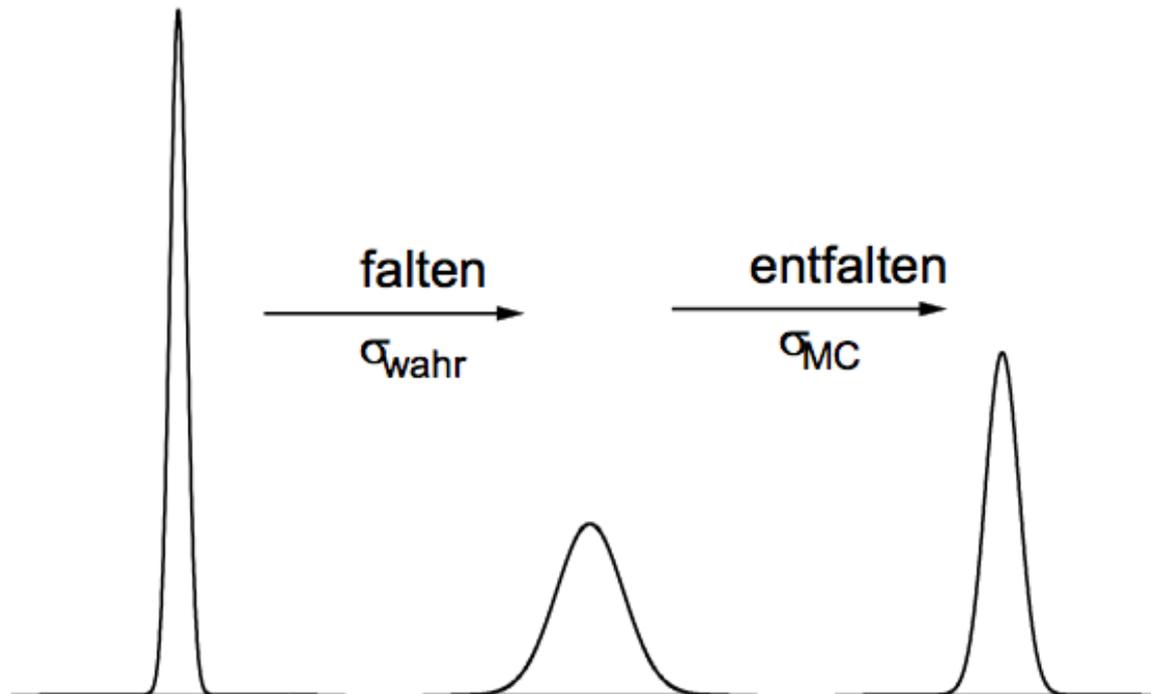
Entfaltete
Verteilung

- Entfaltungen gehören i.a. zu den sogenannten “ill-posed problems”: Kleine Fluktuationen (z.B. statistische) in der gemessenen Verteilung können zu großen Schwankungen in der wahren Verteilung führen
- Lösungsansatz: Dämpfung der Schwankungen durch Regularisierung, d.h. Hinzunahme zusätzlicher Annahmen

Faltung \leftrightarrow Entfaltung

- Annahme Detektorauflösung: Effekt von 10% Abweichung zwischen wahrer und angenommener Auflösung

Bohm, Zech Abb. 8.2

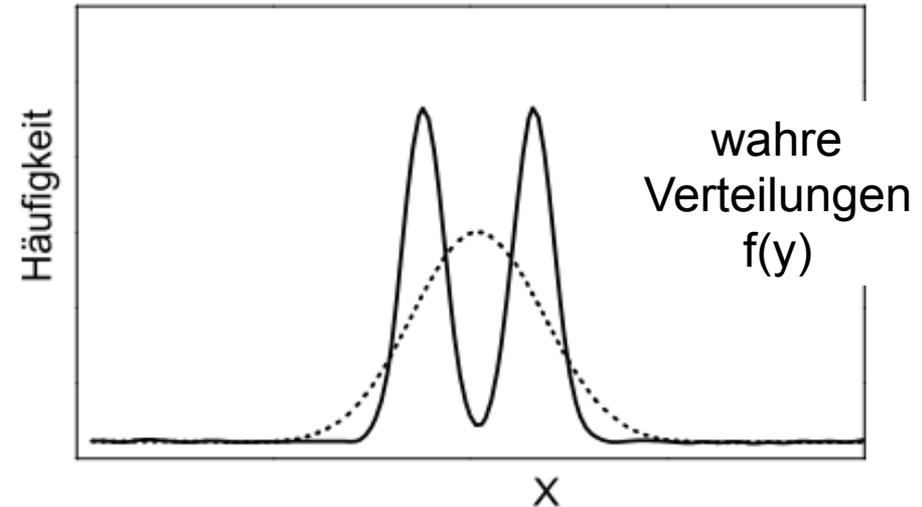


- Unsicherheiten in den Annahmen (wie z.B. der Auflösungsfunktion) werden in den systematischen Unsicherheiten berücksichtigt

Faltung \leftrightarrow Entfaltung

Bohm, Zech Abb. 8.1

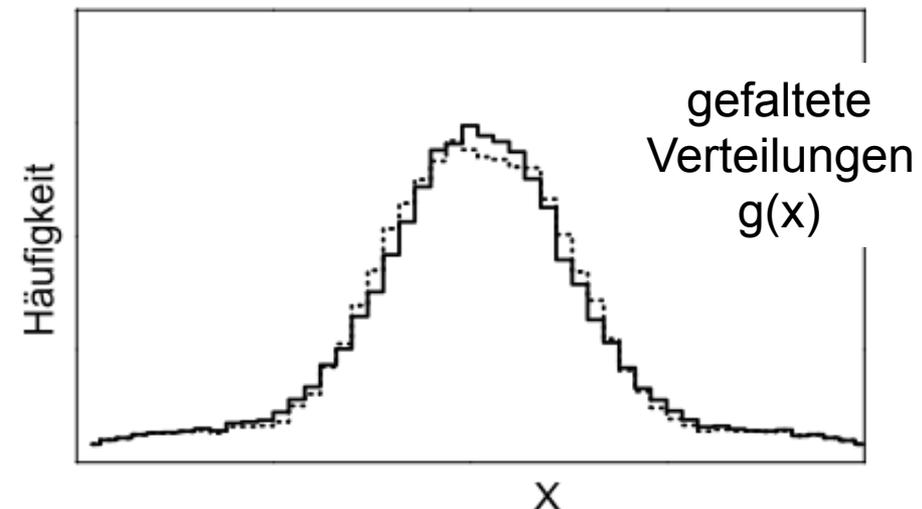
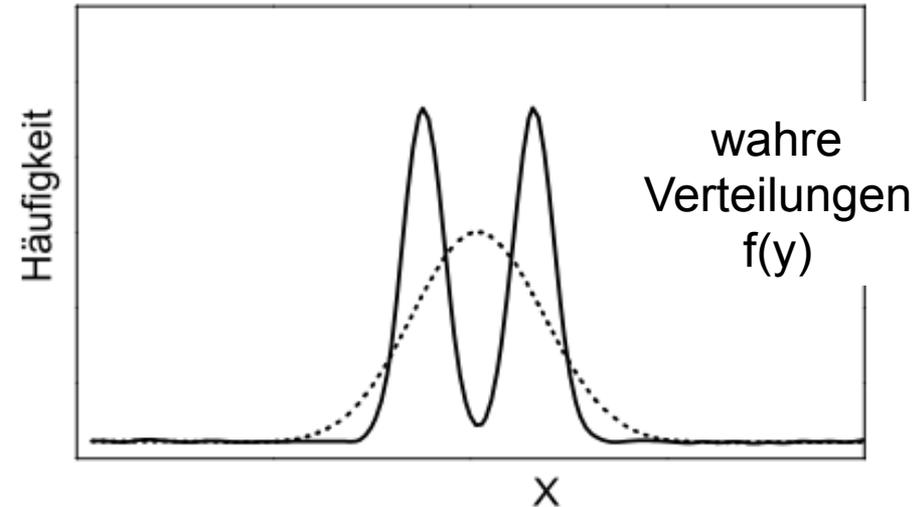
- Faltung: Informationsverlust durch Verschmierung
- Hochfrequente Strukturen (z.B. schmale Peaks) werden ausgewaschen.



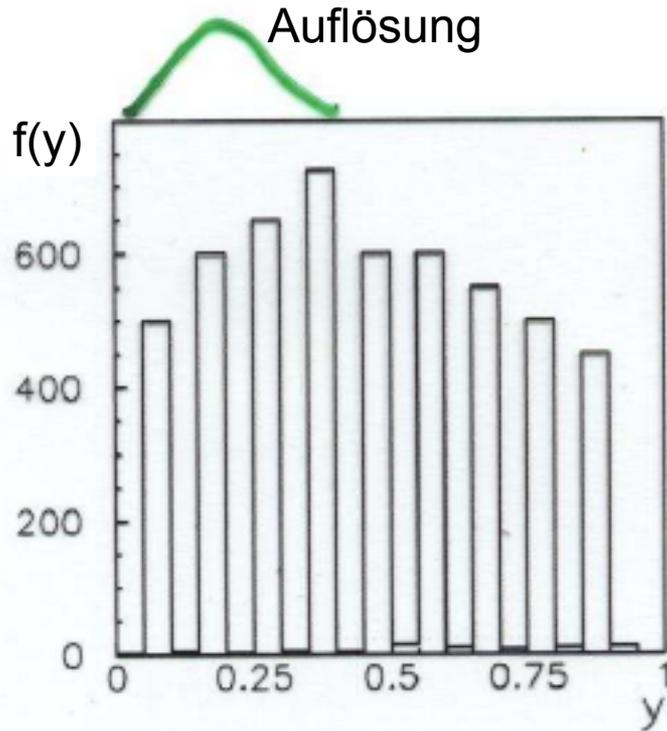
Faltung \leftrightarrow Entfaltung

Bohm, Zech Abb. 8.1

- Faltung: Informationsverlust durch Verschmierung
- Hochfrequente Strukturen (z.B. schmale Peaks) werden ausgewaschen.
- Statistische Fehler liefern Fluktuationen
- Umgekehrt können in der Entfaltung Oszillationen nicht ausgeschlossen werden \rightarrow "ill-posed problem"

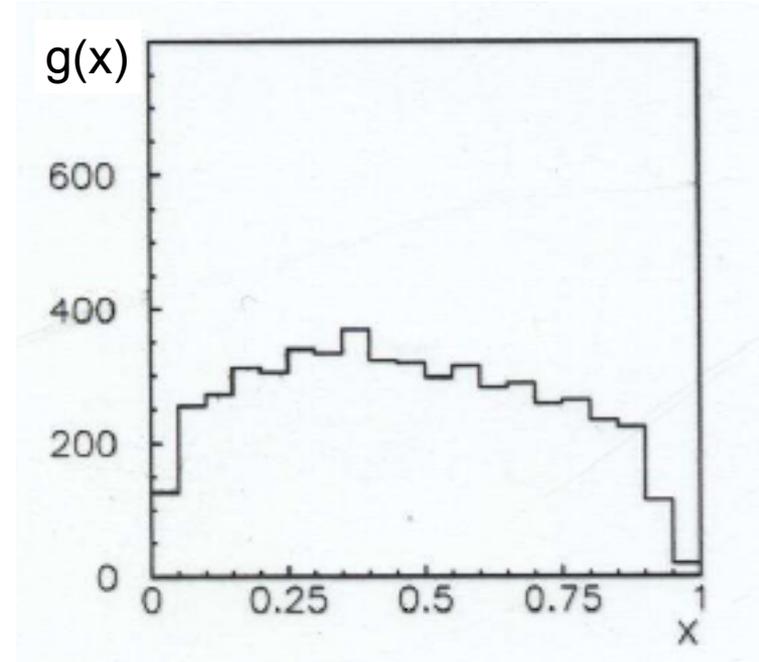


Faltung ↔ Entfaltung



Auflösungs-
funktion A
→
Faltung

←
Entfaltung
Inverse Aufl.-
funktion A^{-1}



Wenn wahre Funktion $f(x)$
Feinstruktur hätte ...

... würde diese durch Verschiebung
(also Faltung mit A) ausgewaschen

- Gemessene Verteilung $g(y)$ enthält aber zusätzlich auch statistische Fluktuationen, die vom Effekt einer Feinstruktur nicht zu unterscheiden sind
- Umgekehrt produziert die Anwendung von A^{-1} auf die gemessene Verteilung Oszillationen

Wahl der Bin-Breite

Blobel/Lohrmann Abb 11.1

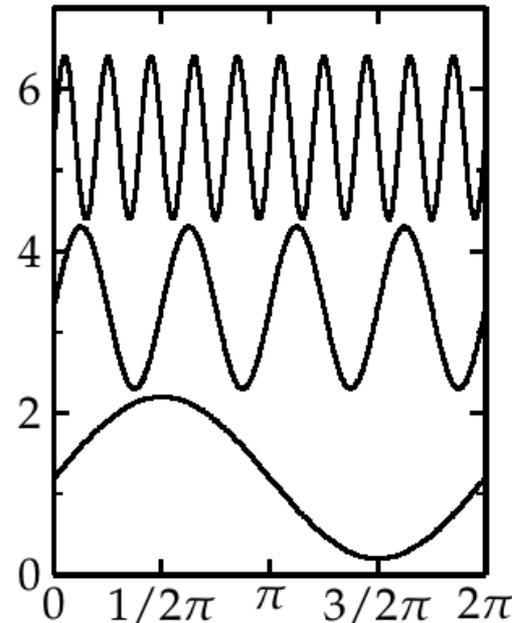
- Betrachte periodische Funktion $f(x) = a \sin(kx)$

- a : Amplitude

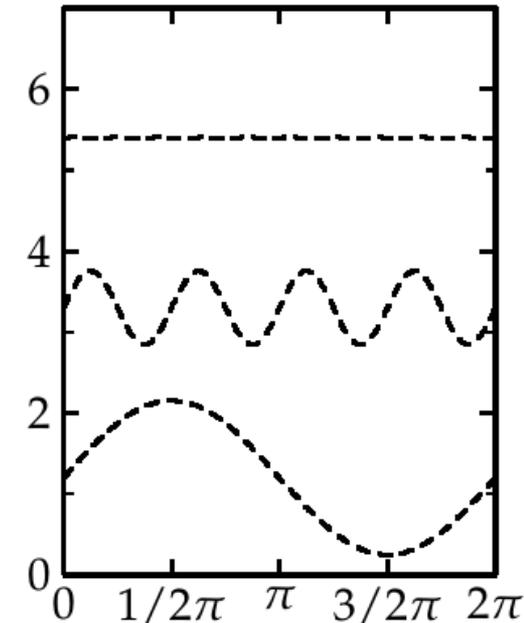
- k : Frequenz

- Gauß-förmige Detektor-Auflösung $\exp(-k^2\sigma^2/2)$

- Erwartete Messung:
 $g(x) \sim \exp(-k^2\sigma^2/2) a \sin(kx)$
 $\sim \alpha \sin(kx)$



wahre Funktionen

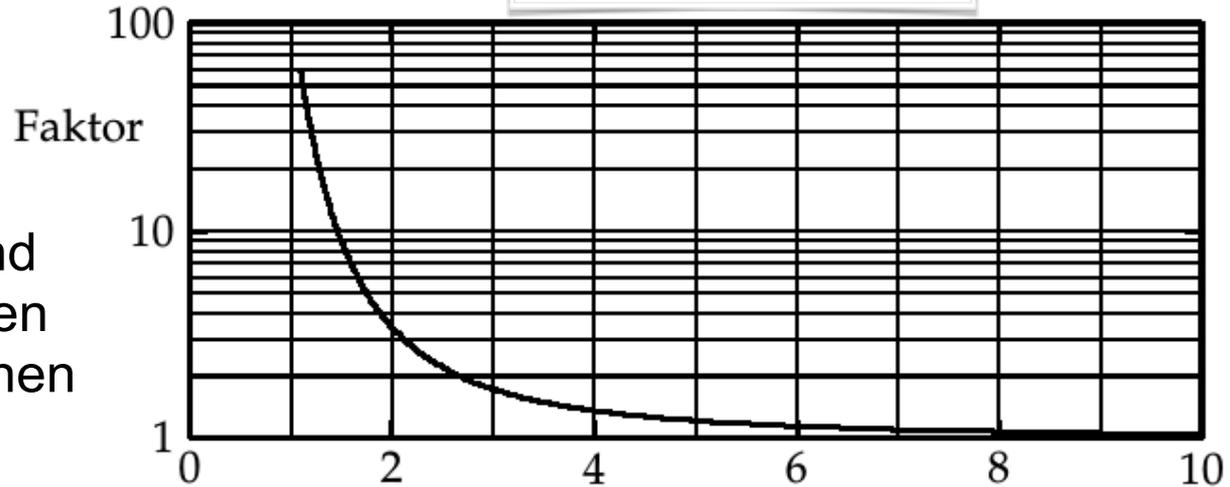


$\sin(kx) \otimes \text{Gauß}$

- Gemessene Amplitude ist vermindert um den Faktor α
- Umgekehrt: Transformation auf wahren Parameter: $(\Delta)a \sim \exp(k^2\sigma^2/2) (\Delta)\alpha$
- Faktor $\exp(k^2\sigma^2/2)$ sehr groß für große $k \rightarrow$ großes a und Δa
- D.h. kleine Bin-Breite führt zu großen Fehlern und Korrelationen

Wahl der Bin-Breite

Blobel/Lohrmann Abb 11.2



- Kleine Bin-Breite Δx und großes σ führt zu großen Fehlern und Korrelationen

- Geeignete Bin-Breite: $\Delta x \approx 2\sigma$

Vergrößerungsfaktor des statistischen Fehlers als Funktion von Intervall-Breite und Auflösung.

- In der Praxis: 2σ häufig nicht praktikabel
- Verwendung von Zusatzannahmen → Regularisierung
 - z.B. Annahme von kontinuierlichem Verlauf
 - Unterdrückung hochfrequenter Anteile in $f(x)$
 - Führt zu Verzerrung (Bias) → Kontrolle
 - Funktioniert i.a. für Reinheit und Stabilität oberhalb von 40-50%

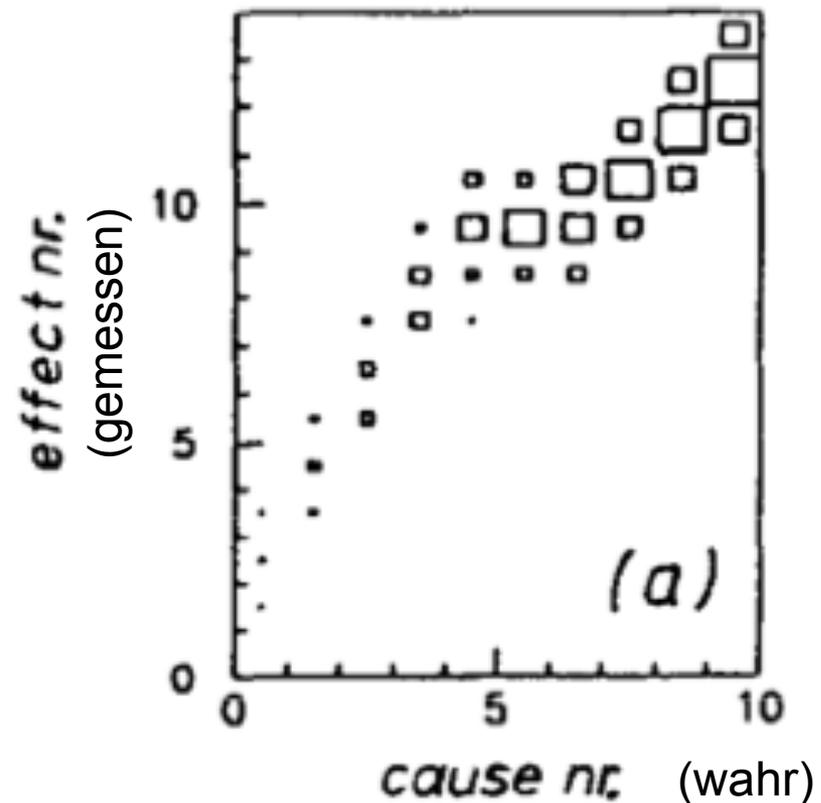
Entfaltung: Bestimme Auflösungsfunktion

- Allgemein hängt die Auflösung vom Messwert ab

$$g(x) = \int A(x, y) \cdot f(y) dy$$

“Fredholmsche Integralgleichung”

- y : wahrer Wert mit PDF f_y
- x : gemessener Wert
- $A(x, y)$: Auflösungsfunktion
(nicht mehr notwendigerweise von der Form $A(y-x)$)



aus: G.D'Agostini, 1995

- In diskretisierter Form:

Migrationsmatrix A_{ij} beschreibt 2D-Häufigkeitsverteilung:

“wahrer Wert x in Bin i , gemessener Wert y in Bin j ”

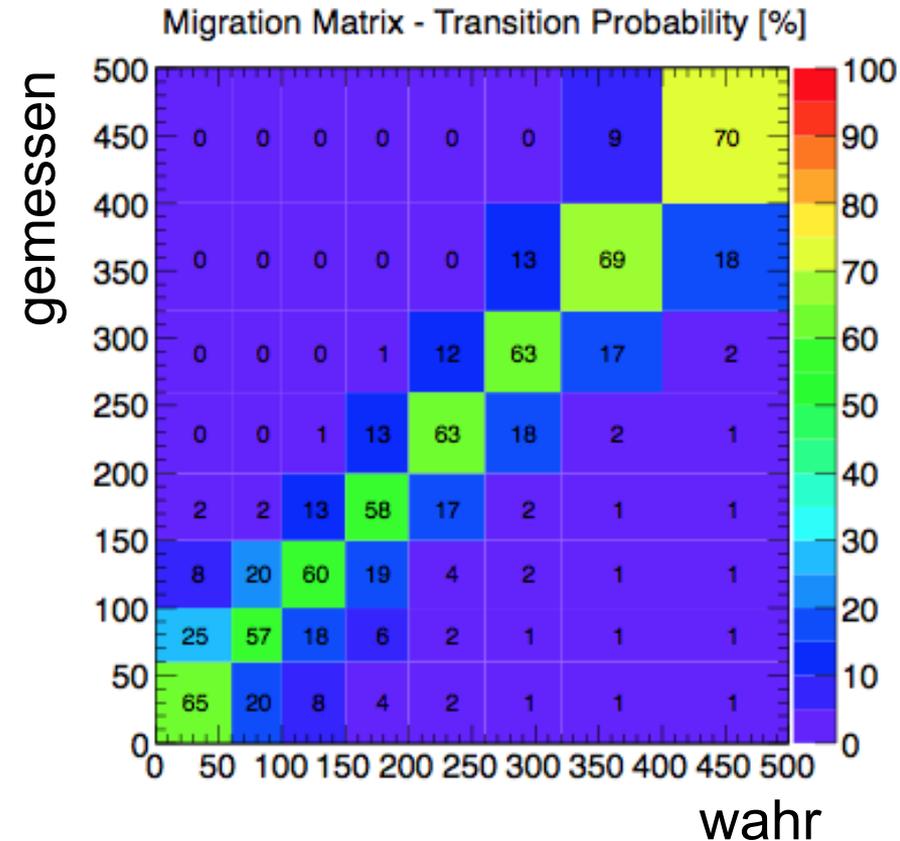
Migrationsmatrix

Faltung in Matrixschreibweise:

$$g_i = \sum_{j=1}^m A_{ij} f_j + b_i$$

$$g_i = \int_{\text{bin } i} g(x) dx; \quad f_j = \int_{\text{bin } j} f(y) dy$$

$$A_{ij} = \frac{\int_{\text{bin } i} dx \int_{\text{bin } j} dy A(x|y) \epsilon(y) f(y)}{\int_{\text{bin } j} f(y) dy}$$



- $A_{ij} = m \times n$ Migrationsmatrix (auch Antwortmatrix oder Response Matrix)
- Bedingte Wahrscheinlichkeit: $P(\text{beobachtet in Bin } j \mid \text{wahrer Wert in Bin } i)$
- Nebendiagonalelemente: “Verschmierungseffekte” durch Auflösung:
Wenn Auflösung \ll Binbreite, dann ist Matrix (so gut wie) diagonal

Migrationsmatrix

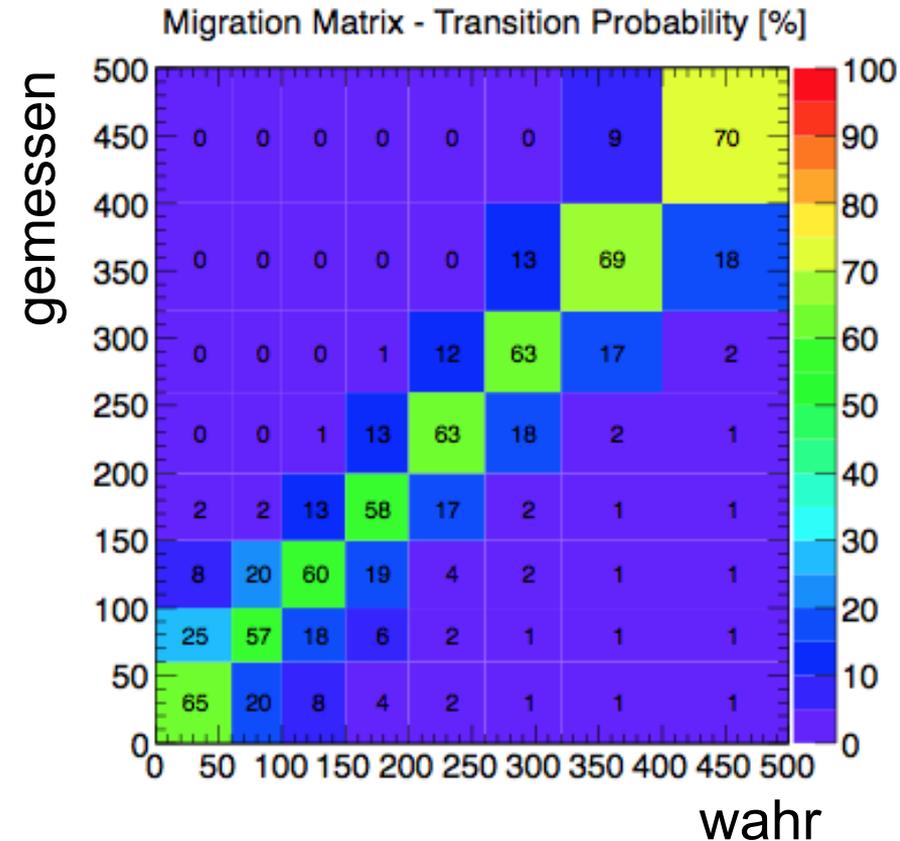
- Faltung in Matrixschreibweise:

$$g = Af + b$$

- 1D Histogramme:

- Gemessene Verteilung: g
- Untergrund: b
- Wahre Verteilung: f

- 2D Migrationsmatrix A_{ij}
(i.a. erzeugt aus MC, enthält
Detektorverhalten, Kalibration,
Auflösungen etc. aus Hilfsmessungen)

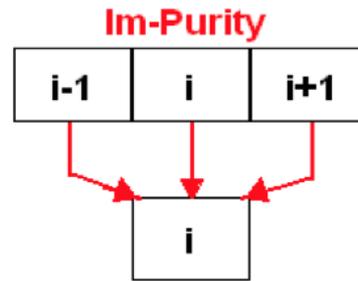


- $A_{ij} = m \times n$ Migrationsmatrix (auch Antwortmatrix oder Response Matrix)
- Bedingte Wahrscheinlichkeit: $P(\text{beobachtet in Bin } j \mid \text{wahrer Wert in Bin } i)$
- Nebendiagonalelemente: “Verschmierungseffekte” durch Auflösung:
Wenn Auflösung \ll Binbreite, dann ist Matrix (so gut wie) diagonal

Reinheit, Stabilität

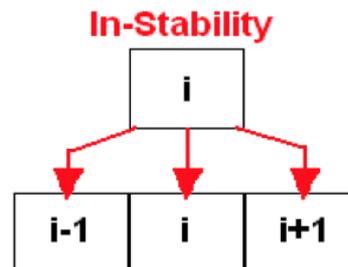
- Wähle Binbreite so, dass Stabilität und Reinheit genügend groß sind
- Reinheit: Anteil der gemessenen Einträge, die tatsächlich aus dem gleichen Bin stammen.

$$P_i = \frac{N_i^{\text{rec}} |_{\text{gen in } i}}{N_i^{\text{rec}}}$$

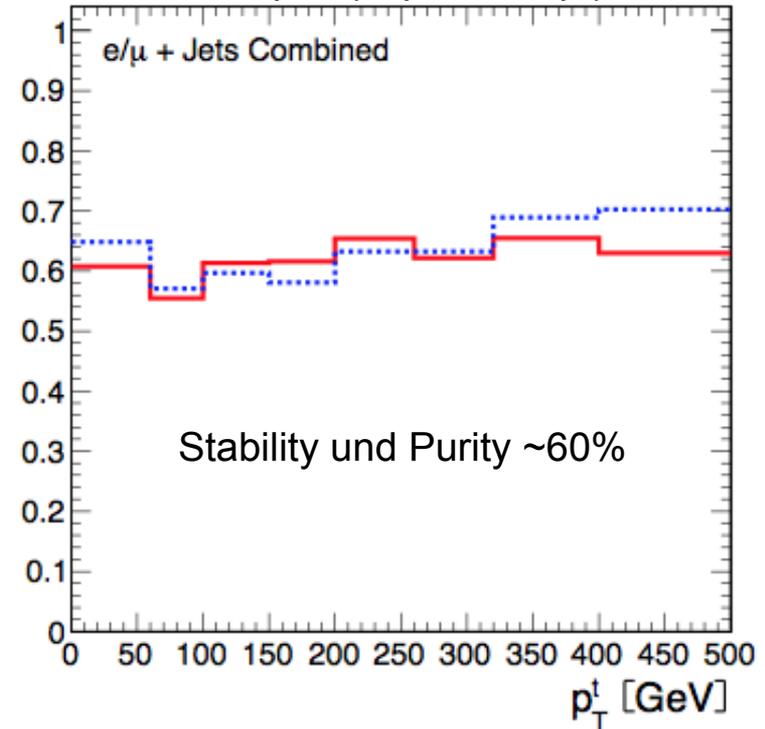


- Stabilität: Anteil der korrekt gemessenen Einträge in einem Bin

$$S_i = \frac{N_i^{\text{gen}} |_{\text{rec in } i}}{N_i^{\text{gen}} |_{\text{rec}}}$$



Beispiel (Top-Quark pt)



Empfohlen: 90% → 2σ

In der Praxis sind 2σ häufig nicht erreichbar. In regularisierter Entfaltung sind i.a. Werte ab 40-50% akzeptabel. Muss für jede Verteilung einzeln geprüft werden !

Entfaltung durch Matrix-Inversion

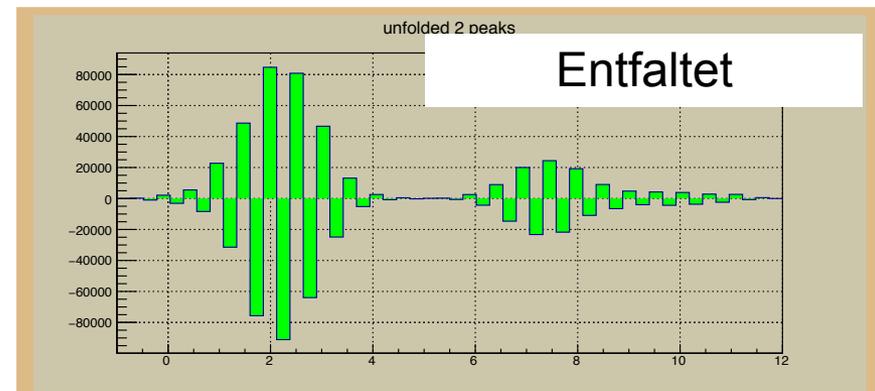
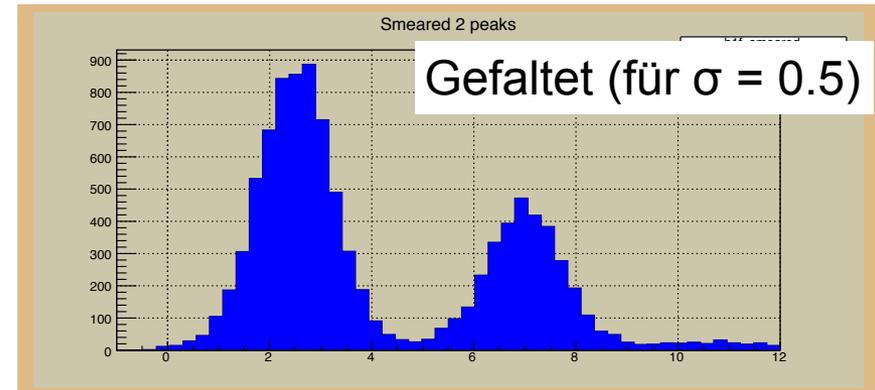
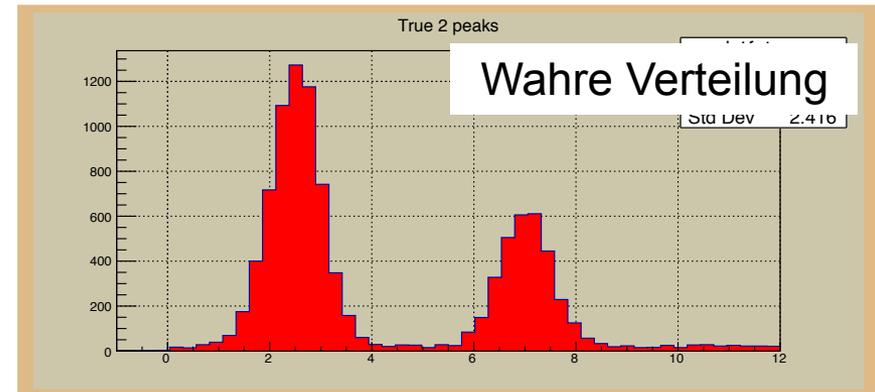
- Inversion der Migrations-Matrix

$$\mathbf{g} = \mathbf{A}\mathbf{f} + \mathbf{b}$$

$$\Rightarrow \mathbf{f} = \mathbf{A}^{-1}(\mathbf{g} - \mathbf{b})$$

- Anwendung von \mathbf{A}^{-1} auf die gemessene Verteilung: i.a. starke Oszillationen, d.h. benachbarte Bins sind antikorreliert

- Korrekt bei Betrachtung der vollen Fehlermatrix $V[f]$ (unanschaulich)



Beispiel: `fold_main.cxx`

Matrix-Inversion: Rechnung für zwei Bins

- Gemessen: g Migrationsmatrix: A wahre Verteilung: f

$$g = \begin{pmatrix} g_1 \\ g_2 \end{pmatrix} \quad A = \begin{pmatrix} 1 - \epsilon & \epsilon \\ \epsilon & 1 - \epsilon \end{pmatrix} \quad f = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}$$

- $\epsilon = 0$: Perfekte Messung, triviales Ergebnis:

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

- Inversion der Matrix zur Bestimmung von f und $V[f]$:

$$f = A^{-1}g \quad V[f] = A^{-1}V[g](A^{-1})^T$$

- Berechnen von A^{-1} :

$$A^{-1} = \frac{1}{1 - 2\epsilon} \begin{pmatrix} 1 - \epsilon & -\epsilon \\ -\epsilon & 1 - \epsilon \end{pmatrix}$$

Matrix-Inversion: Rechnung für zwei Bins

- Gemessen: g Migrationsmatrix: A wahre Verteilung: f

$$g = \begin{pmatrix} g_1 \\ g_2 \end{pmatrix} \quad A = \begin{pmatrix} 1 - \epsilon & \epsilon \\ \epsilon & 1 - \epsilon \end{pmatrix} \quad f = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}$$

$$f = A^{-1}g \quad V[f] = A^{-1}V[g](A^{-1})^T$$

- Berechnen von f und $V[f]$:

$$f_1 = \frac{1}{1 - 2\epsilon} [(1 - \epsilon)g_1 - \epsilon g_2] \quad f_2 = \frac{1}{1 - 2\epsilon} [-\epsilon g_1 + (1 - \epsilon)g_2]$$

$$V[f] = \frac{1}{(1 - 2\epsilon)^2} \begin{pmatrix} (1 - \epsilon)^2 g_1 + \epsilon^2 g_2 & -\epsilon(1 - \epsilon)(g_1 + g_2) \\ -\epsilon(1 - \epsilon)(g_1 + g_2) & (1 - \epsilon)^2 g_2 + \epsilon^2 g_1 \end{pmatrix}$$

- Negative Nicht-Diagonalelemente in $V[f]$:
- Für $\epsilon \gg 0$ werden $\epsilon(1-\epsilon)$ Terme groß \rightarrow Bins sind antikorreliert
- Für $\epsilon = 0.5$: vollständiger Informationsverlust $\rightarrow A$ singular \rightarrow keine Messung

■ Bestimmung von Parametern durch Anpassung an die gemessenen Daten

- Vorteil: vermeidet tatsächliches Entfalten
- Nachteile: erfordert Theorievorhersage
- kein Vergleich mit anderen Daten oder Theorien möglich

■ “Bin-by-Bin” Korrektur

- Verwende einfache Korrekturfaktoren (insbes. wenn A_{ij} hinreichend diagonal)

$$g_i = \sum_{j=1}^m A_{ij} f_j + b_i \rightarrow g_i = c_i f_i + b_i$$

- Korrekturfaktoren aus MC-Simulation

$$c_i^{MC} = (g_i^{MC} - b_i^{MC}) / f_i^{MC} \quad \text{so dass:} \quad f_i = (g_i - b_i) / c_i^{MC}$$

- Eigenschaften:

- c_i hängen von Form der MC-Verteilungen ab (kann überprüft werden)
- Ergebnis liefert keine Information über Korrelationen zwischen Bins

Entfaltung durch Parametrisierung

■ Betrachte $g = Af$ $g_i = \sum_{j=1}^m A_{ij} f_j$

■ Lineares Gleichungssystem (überbestimmt, falls $n > m$)

■ Ermitteln der Lösung f durch χ^2 -Minimierung (vergl. Parameterschätzung)

$$\chi^2 = (g - Af)^T V^{-1} (g - Af)$$

■ $V[f]^{-1}$ ist die inverse Kovarianzmatrix der Datenverteilung
 => Gesucht: f_0 , für das χ^2 minimal wird.

■ Minimierung von χ^2 führt zum gleichen Ergebnis wie Matrix-Inversion.

■ **Regularisierung:** Einführung eines Penalty-Terms $S(f)$ unterdrückt unerwünschte Fluktuationen:

$$\chi_{reg}^2 = (g - Af)^T V^{-1} (g - Af) + \tau \cdot S(f)$$

■ Tikhonov Regularisierung

“Globale Krümmung”

$$S_K(f) = - \int \left(\frac{d^k f(y)}{dy^k} \right)^2 dy$$

■ Für Histogramme (diskret) und nur für benachbarte Bins:

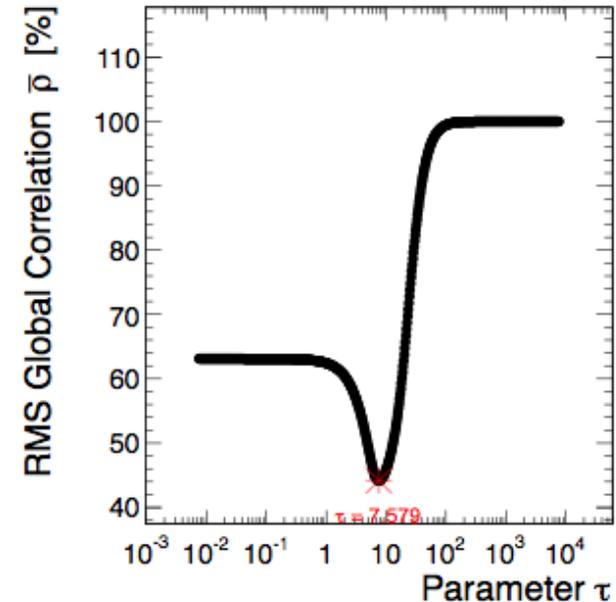
$$S_2 = \sum_{j=1}^{m-2} (-f_j + 2f_{j+1} - f_{j+2})^2$$

■ Wahl der Stärke τ , häufig durch Minimierung der globalen Korrelation, definiert als Summe über Korrelationen zwischen allen Bins

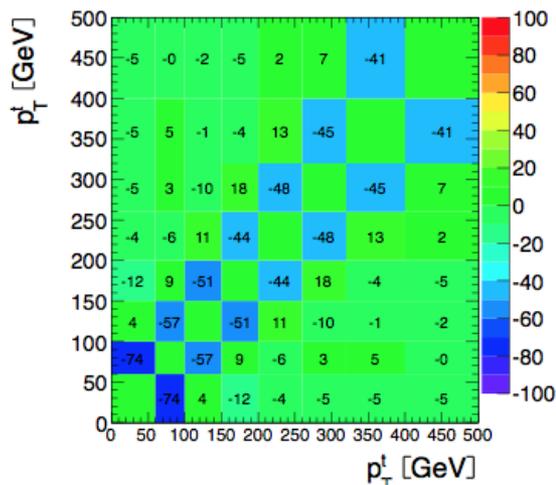
$$\sum_j \rho_j / m \quad \text{mit} \quad \rho_j = \sqrt{1 - \frac{1}{(V[f]^{-1})_{jj} (V[f])_{jj}}}$$

Beispiel: Transversal-Impulsverteilung von Top-Quarks

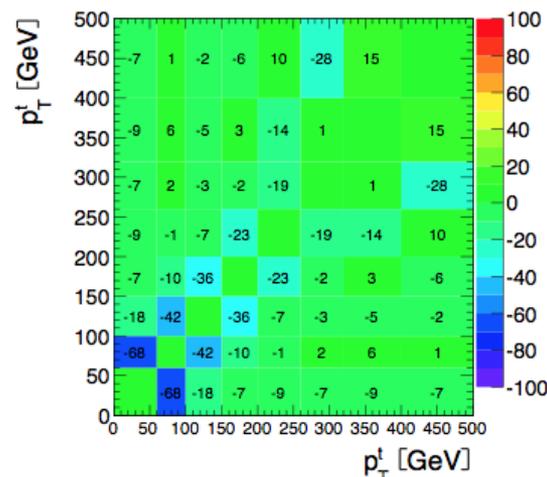
- Optimale Wahl von τ : Reduktion der (negativen) Korrelation zwischen benachbarten Bins auf ein vernünftiges Niveau
- Starke Regularisierung: Positive Korrelation zwischen benachbarten Bins



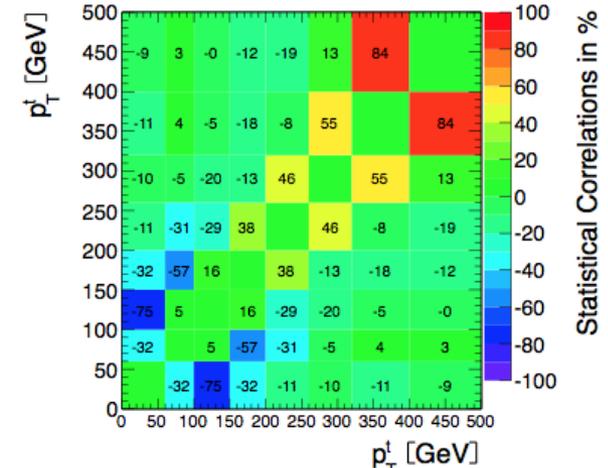
$\tau = 0.00$



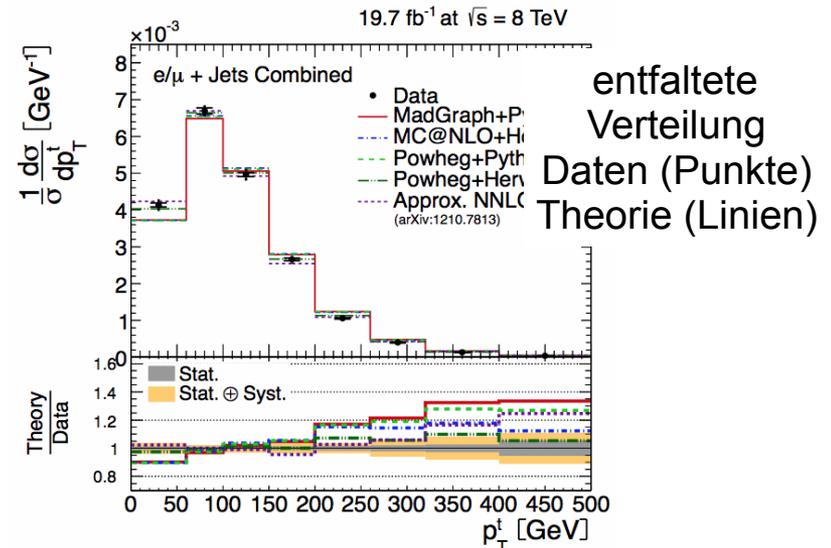
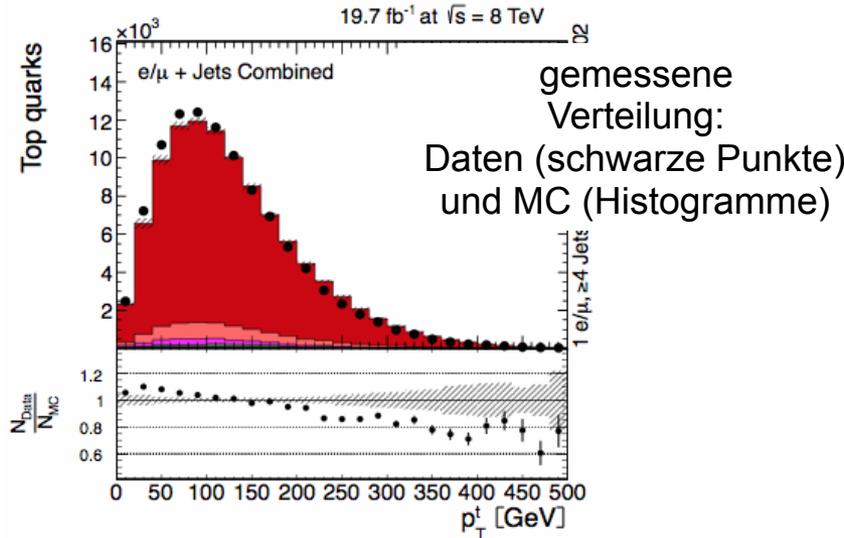
$\tau_d = 7.57$



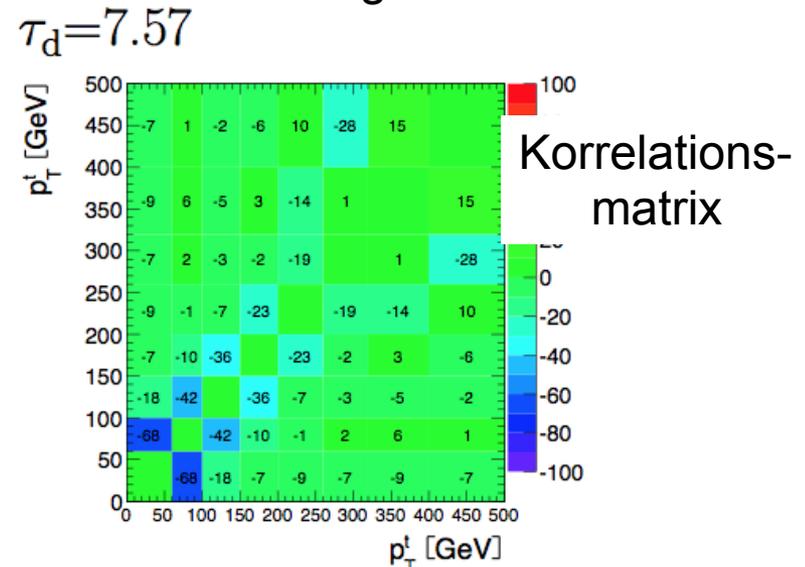
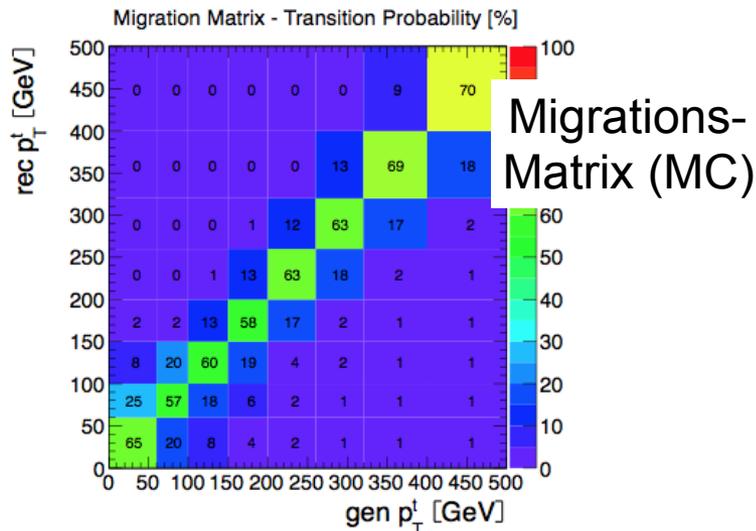
$\tau = 25.00$



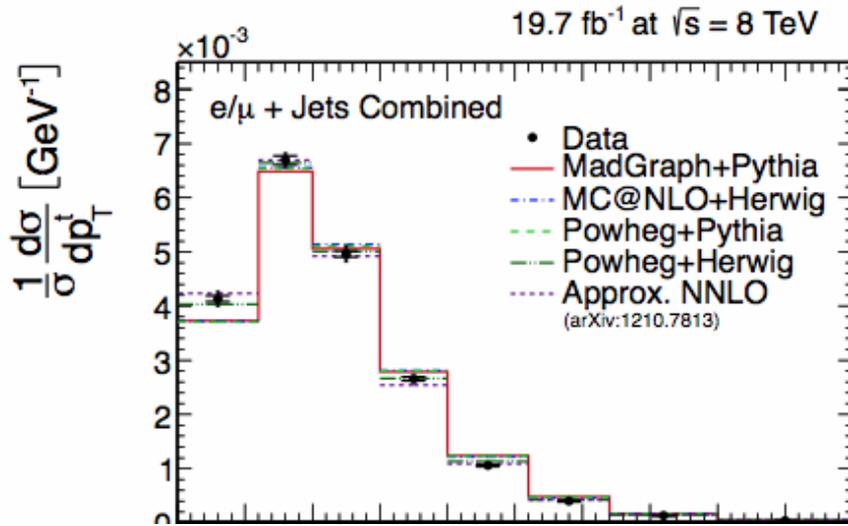
Beispiel: Transversal-Impulsverteilung von Top-Quarks



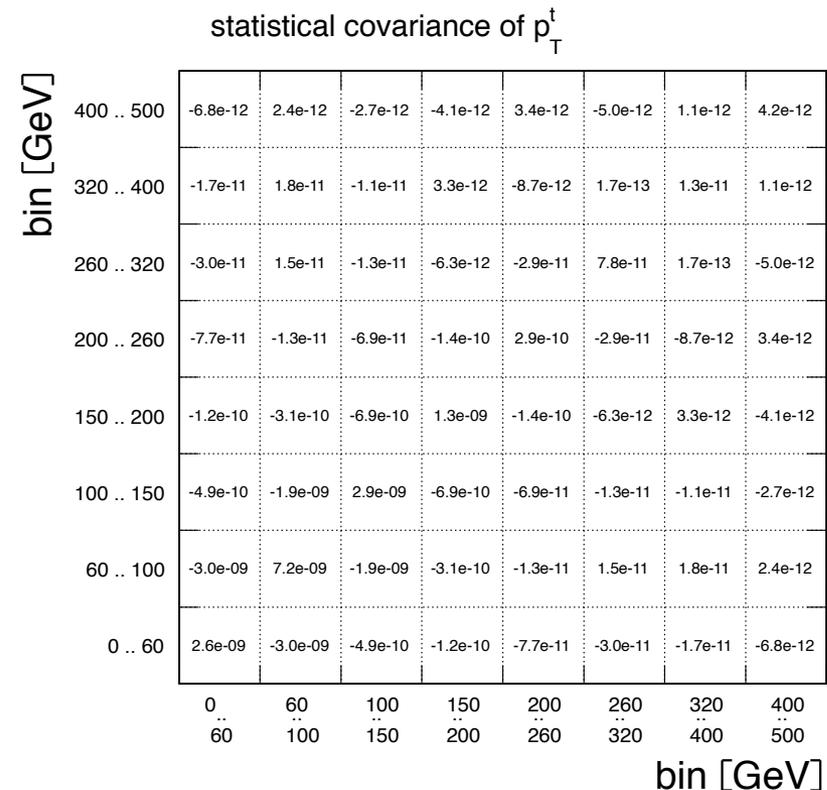
Eingabe $\xrightarrow{\text{Entfaltung}}$ Ausgabe



Beispiel: Transversal-Impulsverteilung von Top-Quarks



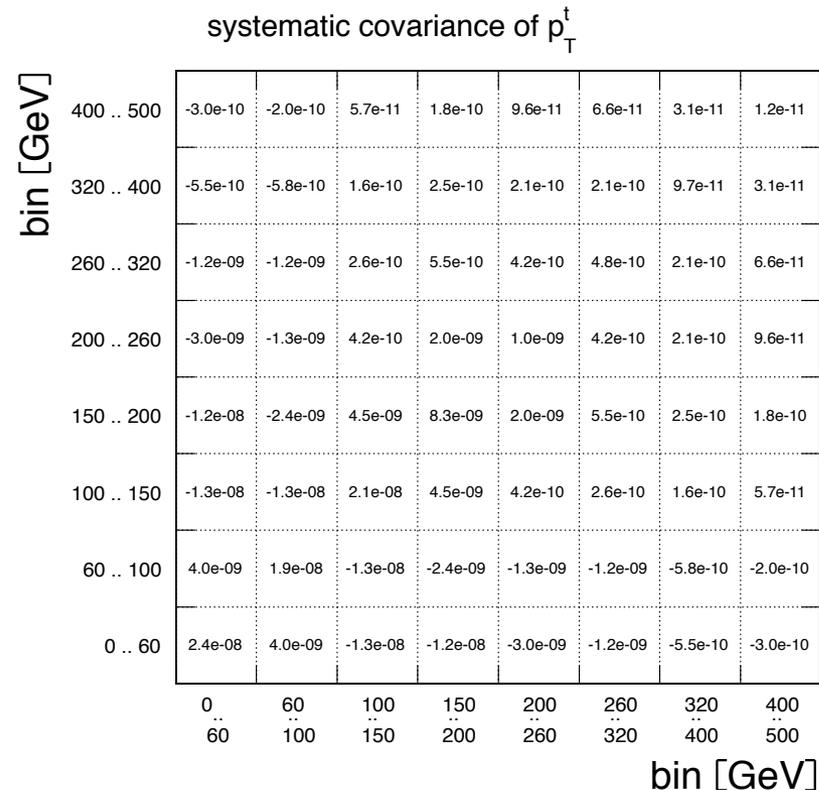
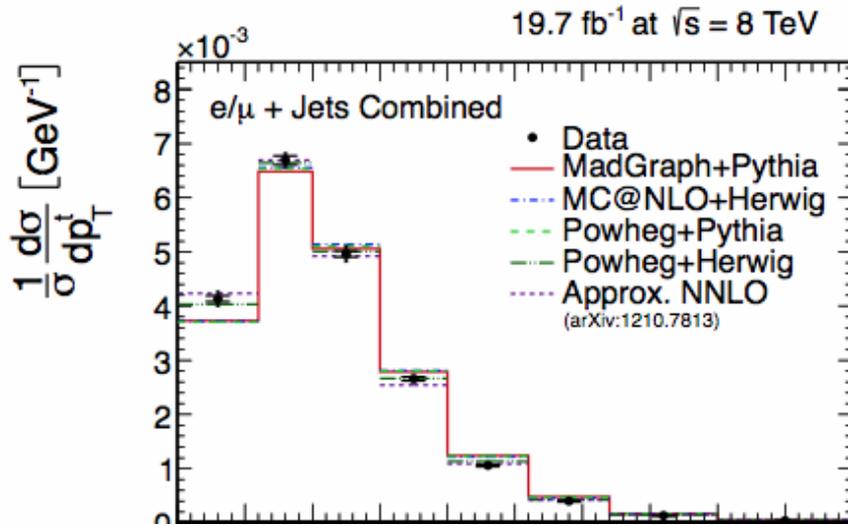
p_T^t [GeV]	$(\frac{1}{\sigma} \frac{d\sigma}{dX})_{MC}$	$(\frac{1}{\sigma} \frac{d\sigma}{dX})_{exp}$	stat. [%]	syst. [%]	tot. [%]
0 to 60	0.003732	0.004137	1.2	3.6	3.8
60 to 100	0.006487	0.006692	1.3	1.9	2.3
100 to 150	0.005063	0.004962	1.1	3.0	3.2
150 to 200	0.002786	0.002660	1.3	3.5	3.8
200 to 260	0.001238	0.001057	1.6	3.2	3.6
260 to 320	0.000484	0.000399	2.2	5.6	6.0
320 to 400	0.000173	0.000130	2.8	7.7	8.1
400 to 500	0.000050	0.000037	5.5	9.5	11.0



Ergebnis wird publiziert als: 1d-Diagramm, Tabelle und Kovarianz- oder Korrelations-Matrix (statistische und systematische Fehler)

Abbildungen aus Doktorarbeit M. Görner, Uni Hamburg, 2014
 Publiziert in CMS Collab: Eur.Phys.J. C 75 (2015) 542

Beispiel: Transversal-Impulsverteilung von Top-Quarks

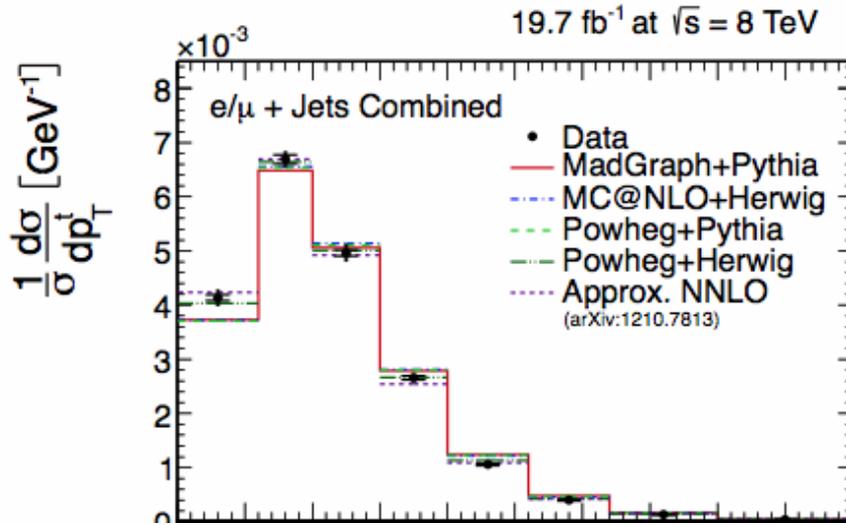


p_T^t [GeV]	$(\frac{1}{\sigma} \frac{d\sigma}{dX})_{MC}$	$(\frac{1}{\sigma} \frac{d\sigma}{dX})_{exp}$	stat. [%]	syst. [%]	tot. [%]
0 to 60	0.003732	0.004137	1.2	3.6	3.8
60 to 100	0.006487	0.006692	1.3	1.9	2.3
100 to 150	0.005063	0.004962	1.1	3.0	3.2
150 to 200	0.002786	0.002660	1.3	3.5	3.8
200 to 260	0.001238	0.001057	1.6	3.2	3.6
260 to 320	0.000484	0.000399	2.2	5.6	6.0
320 to 400	0.000173	0.000130	2.8	7.7	8.1
400 to 500	0.000050	0.000037	5.5	9.5	11.0

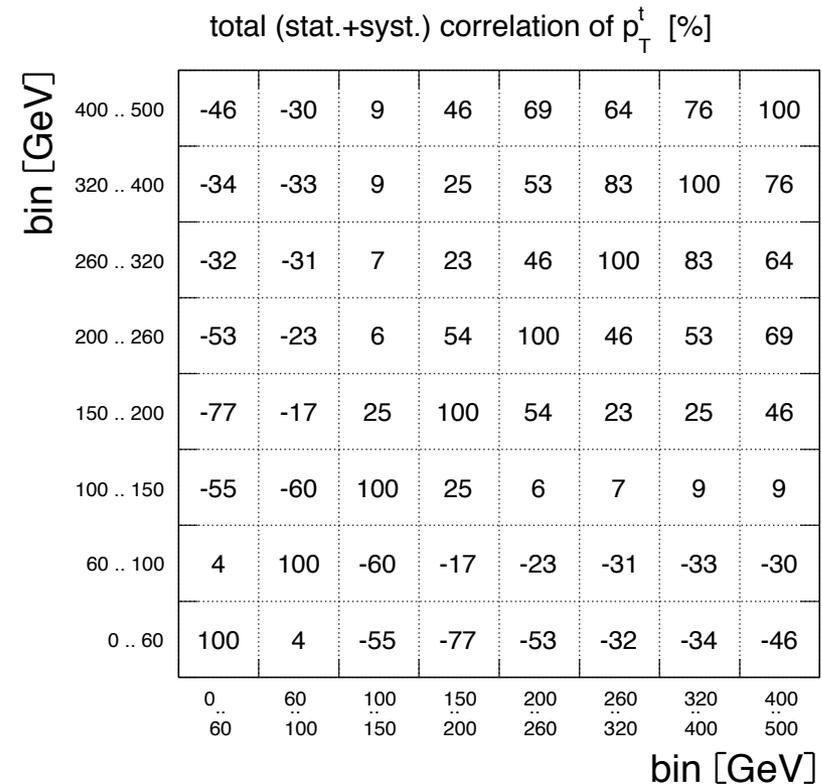
Ergebnis wird publiziert als: 1d-Diagramm, Tabelle und Kovarianz- oder Korrelations-Matrix (statistische und systematische Fehler)

Abbildungen aus Doktorarbeit M. Görner, Uni Hamburg, 2014
 Publiziert in CMS Collab: Eur.Phys.J. C 75 (2015) 542

Beispiel: Transversal-Impulsverteilung von Top-Quarks



p_T^t [GeV]	$(\frac{1}{\sigma} \frac{d\sigma}{dX})_{MC}$	$(\frac{1}{\sigma} \frac{d\sigma}{dX})_{exp}$	stat. [%]	syst. [%]	tot. [%]
0 to 60	0.003732	0.004137	1.2	3.6	3.8
60 to 100	0.006487	0.006692	1.3	1.9	2.3
100 to 150	0.005063	0.004962	1.1	3.0	3.2
150 to 200	0.002786	0.002660	1.3	3.5	3.8
200 to 260	0.001238	0.001057	1.6	3.2	3.6
260 to 320	0.000484	0.000399	2.2	5.6	6.0
320 to 400	0.000173	0.000130	2.8	7.7	8.1
400 to 500	0.000050	0.000037	5.5	9.5	11.0



Ergebnis wird publiziert als: 1d-Diagramm, Tabelle und Kovarianz- oder Korrelations-Matrix (statistische und systematische Fehler)

Iterative Methode zur Entfaltung

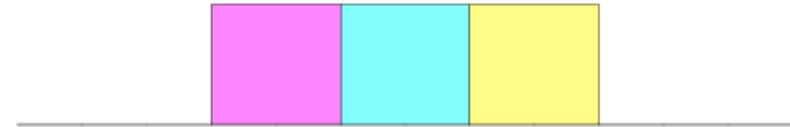
- Beginne mit beliebigem Ansatz $f_{(0)}$
z.B. flache oder gemessene Verteilung
- Bestimme $g_{(0)} = A \cdot f_{(0)}$ und vergleiche in
jedem Bin $g_{(0)}$ mit Beobachtung \hat{g}
- Modifiziere $f_{(0)} \rightarrow f_{(1)} = \hat{g}/g_{(0)} \cdot f_{(0)}$

- Iterativ bis Abbruchkriterium:
 - Vergleiche $A \cdot f_{(i)}$ mit \hat{g}
 - Modifiziere: $f_{(i)} \rightarrow f_{(i+1)} = \hat{g}/g_{(i)} \cdot f_{(i)}$

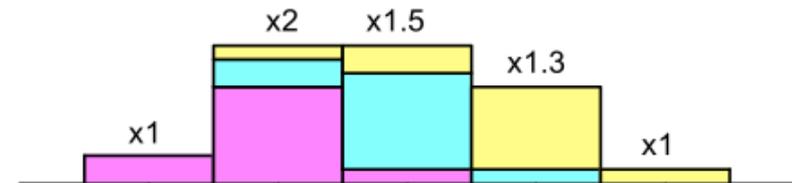
- Vorteil des iterativen Ansatz: keine
numerischen Schwierigkeiten mit
Matrixinversion

- Im Limit vieler Iterationen erhält man das
Ergebnis der Matrix-Inversion (also ohne
Regularisierung)

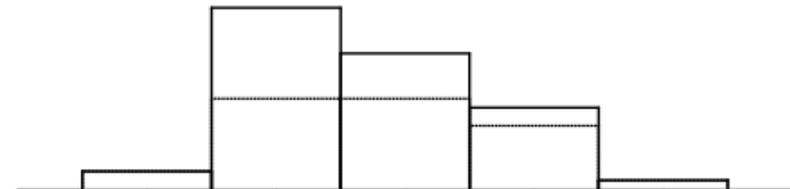
a) Startverteilung



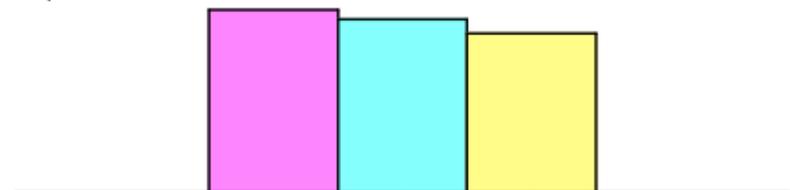
b) gefaltete Starvert.



c) beobachtete Daten



d) 1. Iteration



G.d'Agostini, NIM A 362 (1995)

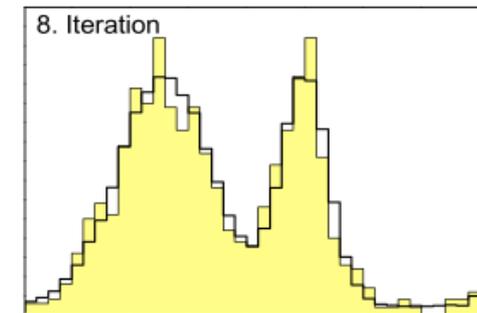
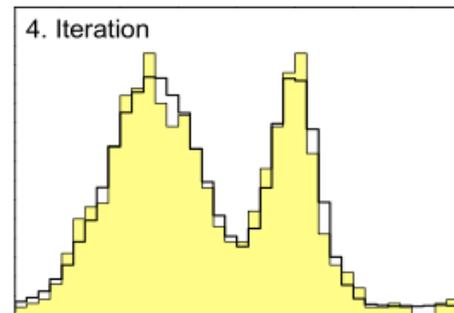
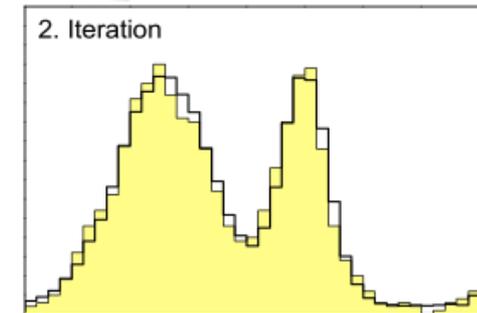
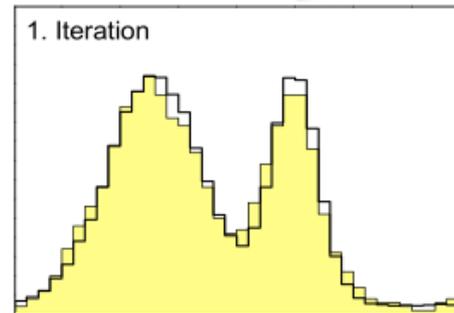
Iterative Methode zur Entfaltung

■ Mögliche Abbruchkriterien:

■ χ^2 -Kompatibilität
zwischen \hat{g} und $g_{(i)}$
→ meist bereits erfüllt, wenn
noch positive Korrelationen
zwischen Bins vorhanden sind

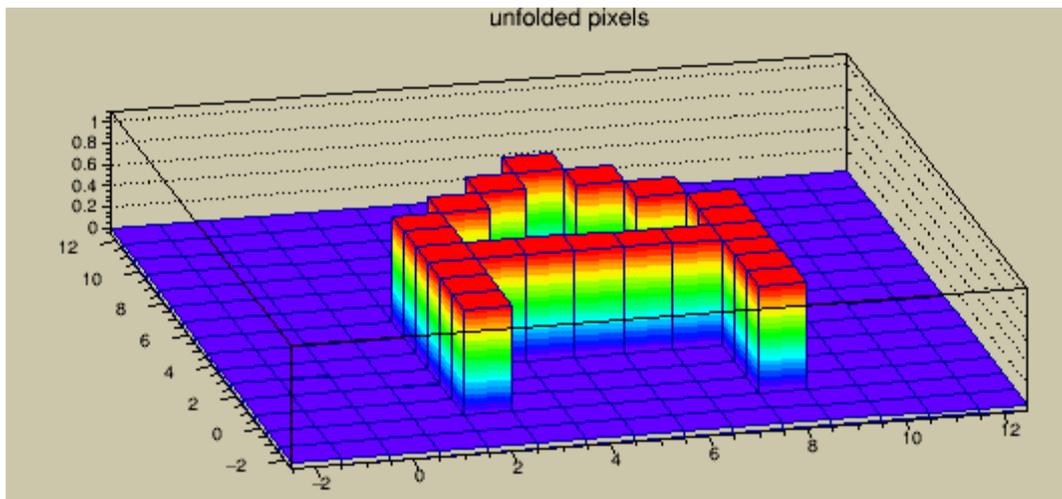
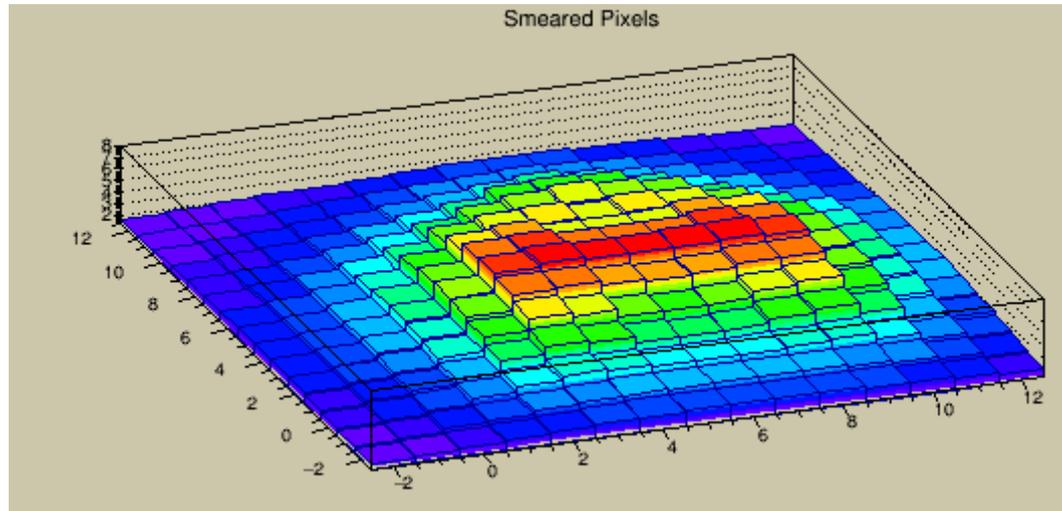
■ Betrachte Diagonalelemente
der Kovarianzmatrix:
Verlange, dass Fehler σ_{jj} in
jedem Bin von $g_{(i)}$ größer ist
als im entsprechenden Bin
von \hat{g}
→ i.a. konservativere Wahl

Bohm/Zech Kap. 8

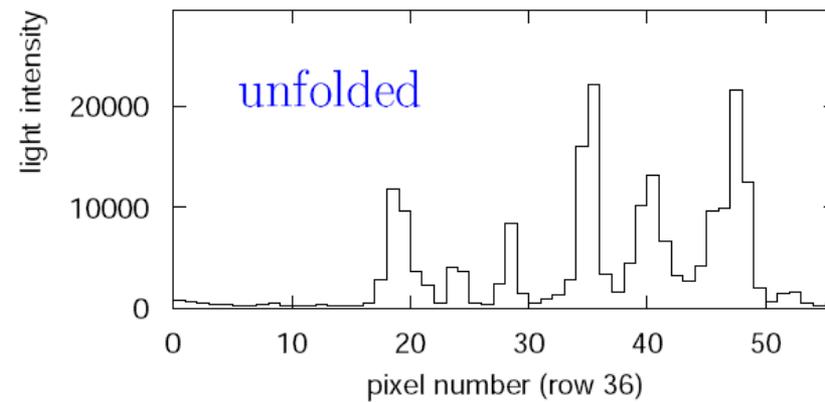
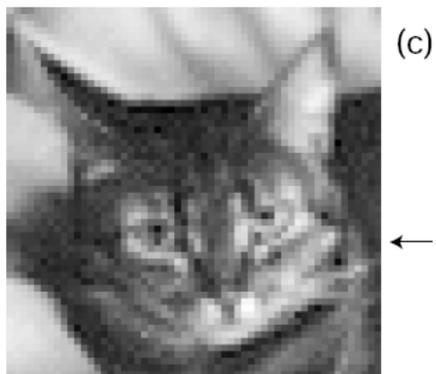
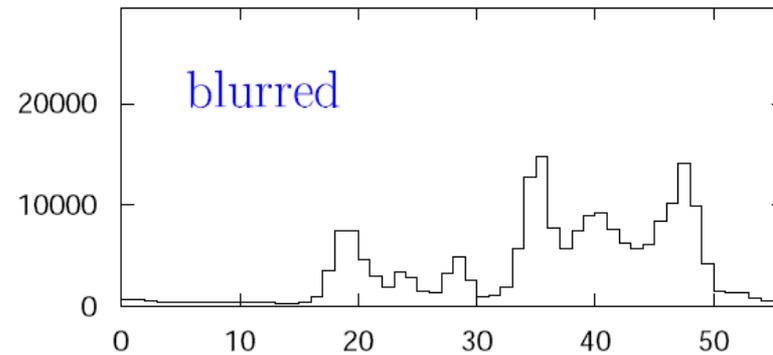
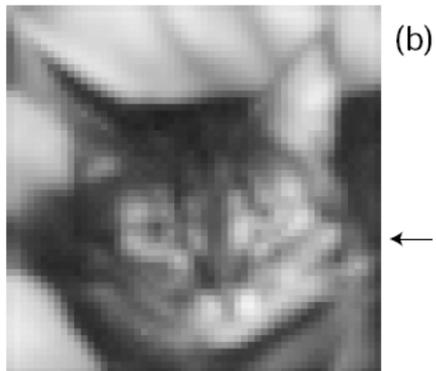
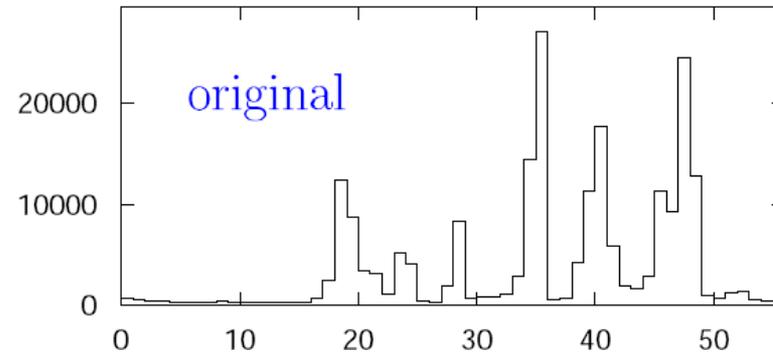
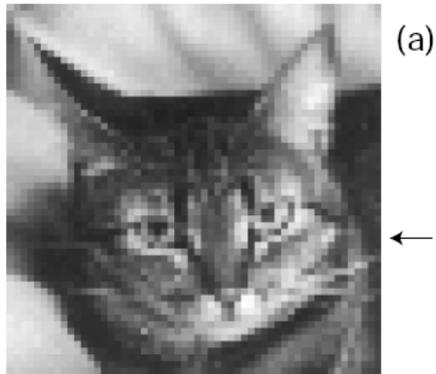


Beispiel hier: Abbruch bereits nach 2. Iteration
(anhand von χ^2 Kriterium)

- In Daten mit hoher Statistik sind Signalstrukturen häufig sehr scharf
 - Astronomie: Punktförmig
 - Spektroskopie: scharfe Linien
 - Allgemein Mustererkennung
- Prinzip der maximalen Entropie (MaxEnt)
 - „Shannon-Jaynes Entropie“ oder „Cross-Entropie“:
 $S \sim \log(\text{Anzahl möglicher Anordnungen der Ereignisse})$, anschaulich:
 - Gleichverteilung der Daten über alle Bins favorisiert. Unterschiede zwischen Bins werden unterdrückt, unabhängig vom relativen Abstand der Bins.
 - Lösungen, die scharfe Maxima über konstantem Untergrund liefern werden bevorzugt.
 - Kann auch relativ zu Referenzverteilung definiert werden \rightarrow Bias null, wenn Daten gleich Referenzverteilung
 - In Teilchenphysik meist nicht nützlich, da Detektoraufösungen Korrelationen zwischen benachbarten Bins verursachen.



Beispiel: `fold_main.cxx`



■ Faltung

- beschreibt Verschmierung der Messung durch Auflösungseffekte
- Summe von Zufallszahlen → Monte-Carlo Methoden oder auch:
- Produkte von charakteristischen Funktionen

■ Entfaltung:

- Ill-posed Problem, kleine Fluktuationen können zu großen Abweichungen führen. Hochfrequente Anteile werden verstärkt.
- Matrix-Inversion: i.a. große negative Korrelationen zwischen benachbarten Bins.
- Lösung ohne Entfaltung: z.B. Bin-by-bin Korrektur
- Entfaltung mit Regularisierung
 - Durch Parametrisierung und χ^2 -Minimierung (TUnfold)
 - Iterative Methoden (D'Agostini)
 - Cross-Entropie