粒子物理与核物理实验中的数据分析

陈少敏
清华大学

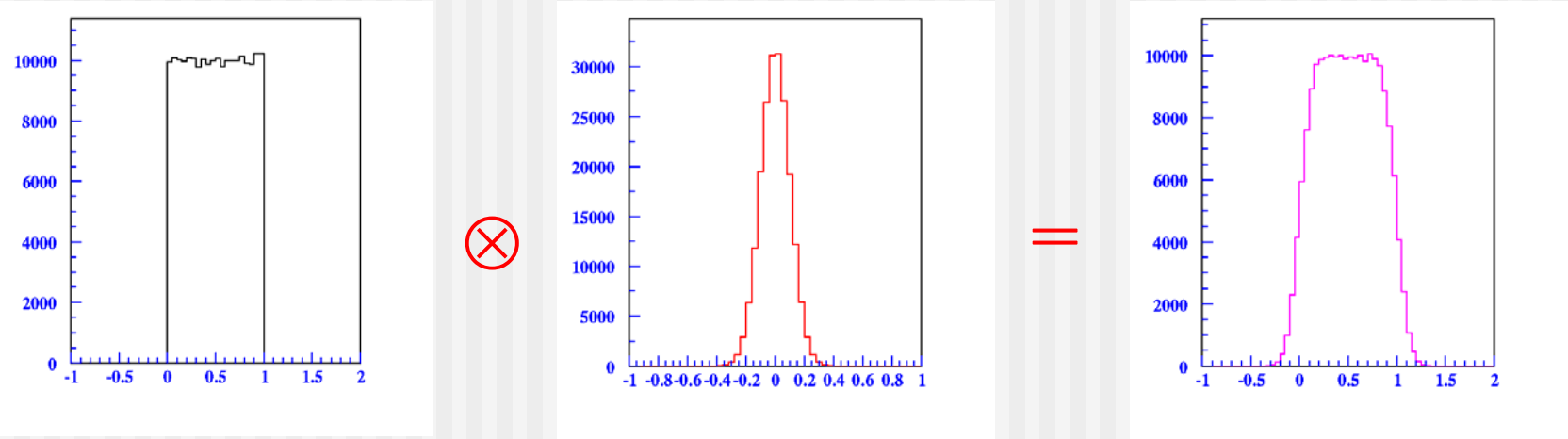
第十二讲：开拆法

本讲要点

- 数学公式,反应函数(矩阵)
- 求反应矩阵的逆
- 修正因子
- 正规化的开拆法
  a) Tikhonov 规则
  b) MaxEnt 规则
- 估计量的方差与偏置
- 正规化参数的选择
- 举例

图像还原问题

一个常见的问题：由于实验仪器的原因而出现图像变形，例如



真实分布

通过探测器模拟可以给出其影响的形式

实验观测分布

如果，已知

能否还原出不受实验仪器影响的分布？

Unfolding(开拆法)

开拆问题的表述

考虑有随机变量 y，我们的目标是要找到概率密度函数 $f(y)$

如果函数可参数化为 $f(y;\vec{\theta})$，那么确定概率密度函数，等效于

最大似然法 ➡ $\hat{\vec{\theta}}$

若无参数化形式，可通过构造直方图

$$p_j = \int_{\text{bin}\,j} f(y)dy \qquad j = 1,\ldots,M$$

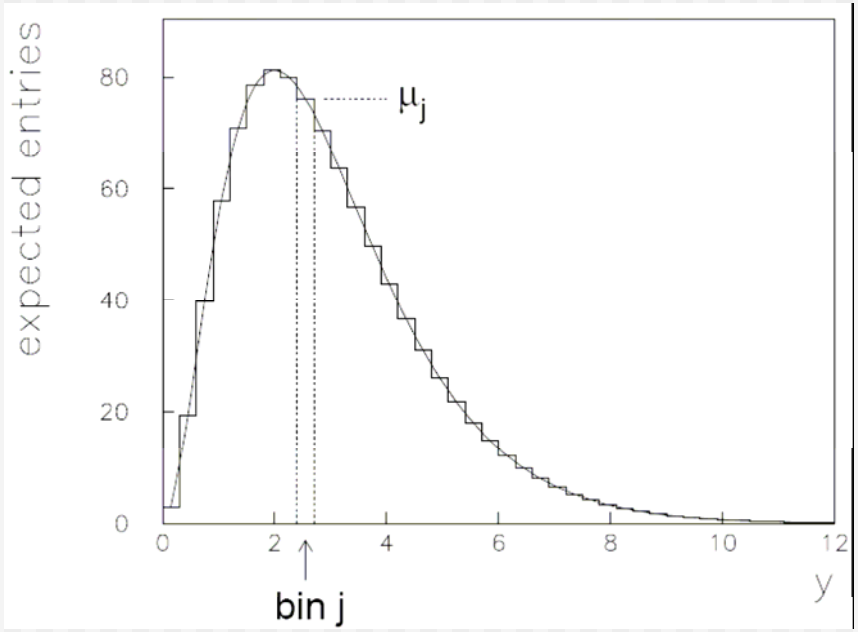
$$\mu_j = \mu_{tot}\, p_j$$ ⬅ “真实的直方图”

目标：为 μ_j(或 p_j)构造估计量

参数的数目=区间的数目M

问题：y 在测量时不可能没有误差

➡ 各区间之间填入的数目互串，导致 $f(y)$ 散开变宽。



反应矩阵

测量误差的影响：$y=$ 真值；　$x=$ 观测值

$$f_{meas}(x)=\int R(x\,|\,y)\,f_{true}(y)\,dy$$

写成离散形式

观测直方图
(期待值)

反应矩阵

$$\nu_i=\sum_{j=1}^{M}R_{ij}\mu_j,\qquad i=1,\ldots,N$$

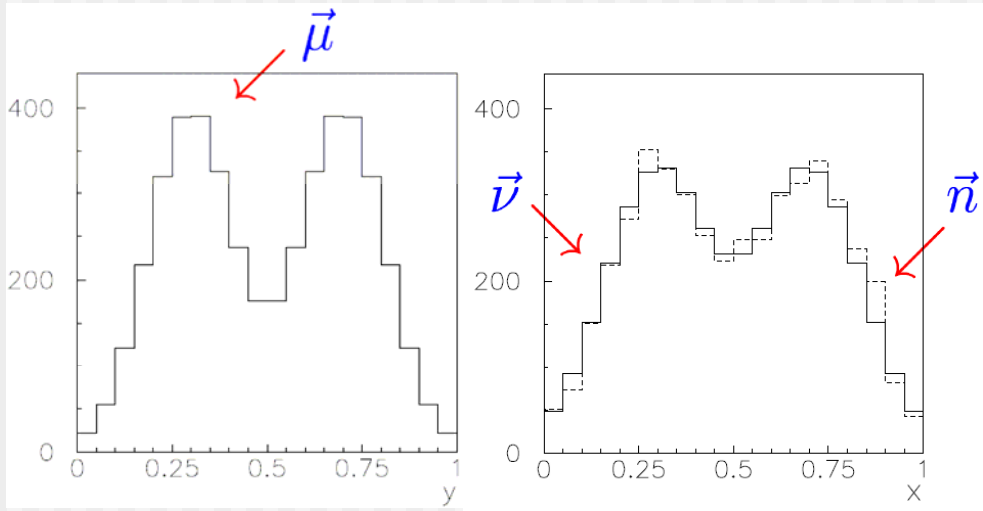
真实直方图

$R_{ij}=P$(观测值在第 i 区$\,|\,$真实值在第 j 区)

数据：$\vec{n}=(n_1,\ldots,n_N)$

这里 $\nu_i=E[n_i]$

注意：$\vec{\mu},\vec{\nu}$ 是常数，$\vec{n}$ 会受到统计涨落的影响。



效率，本底

有时侯，事例可能会不被探测到

$$\sum_{i=1}^{N} R_{ij} = \sum_{i=1}^{N} P(\text{观测值在第 } i \text{ 区} \,|\, \text{真实值在第 } j \text{ 区})$$

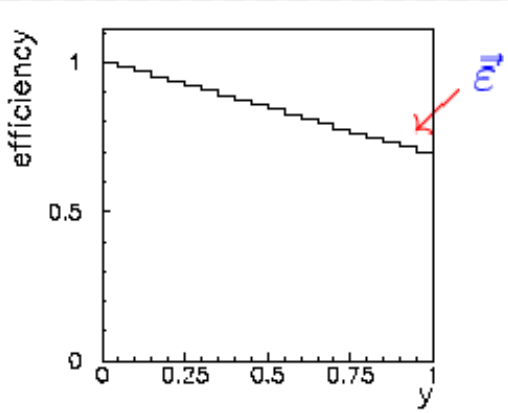
$$= P(\text{观测值在全范围} \,|\, \text{真实值在第} j \text{区})$$

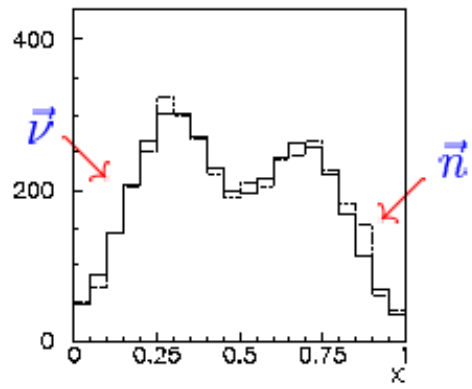
$$= \varepsilon_j(\text{效率})$$

⬅ 取决于在第 j 区的真实直方图

有时在无真实事例发生的时候，也有事例被观测到

$$\Rightarrow \nu_i = \sum_{j=1}^{M} R_{ij}\mu_j + \beta_i$$





β_i 是在观测直方图上预期的本底数目，并假设它是已知的。

各关键量总汇

“真实”直方图：$\vec{\mu}=(\mu_1,\ldots,\mu_M),\quad \mu_{tot}=\sum_{j=1}^{M}\mu_j$ ⟵ M个区间

概率：$\vec{p}=(p_1,\ldots,p_M)=\vec{\mu}/\mu_{tot}$

观测直方图的期待值：$\vec{\nu}=(\nu_1,\ldots,\nu_N)$ ⟵ N个区间

观测直方图：$\vec{n}=(n_1,\ldots,n_N)$

反应矩阵：$R_{ij}=P(\text{观测值在第 } i \text{ 区}\,|\,\text{真实值在第 } j \text{ 区})$

效率：$\varepsilon_j=\sum_{i=1}^{N}R_{ij}$　　预期的本底：$\vec{\beta}=(\beta_1,\ldots,\beta_N)$

⟶ $E[\vec{n}]=\vec{\nu}=R\vec{\mu}+\vec{\beta}$

为了找到 $\vec{\mu}$ 的估计量，需要相关的概率理论，例如：$P(n_i;\nu_i)=\frac{\nu_i^{n_i}}{n_i!}e^{-\nu_i}$

泊松分布，或关联矩阵 $V_{ij}=\mathrm{cov}[n_i,n_j]$ 以便构造 $\log L$ 或 χ^2

为什么要用开拆法

一般而言，我们并不需要开拆法，例如当比较现有理论的预期值时，最好是将探测器相应叠加到理论中去。即在预期值中包含探测器效应并与未修正的原始数据 $\vec{n}$ 相比较。

但是，不将实验数据进行开拆处理，结果发表后，有关反应矩阵的知识将不在保留。而且，开拆后的分布可以直接与各种理论的预言相比较，也可以与别的实验经过开拆以后的分布相比较。

通常开拆的结果更为有用，因为当反应矩阵变得不可恢复时，即使对实验结果可能又有了新的理论解释，也很难进行理论检验。

在粒子物理研究中，开拆法常用的领域为：

- 结构函数
- τ 的谱函数(也就是强子不变质量谱)
- 强子事例形状分布
- 粒子多重数分布
- . . .

反应矩阵的逆

假设 $\vec{\nu} = R\vec{\mu} + \vec{\beta}$ 的逆存在: $\vec{\mu} = R^{-1}(\vec{\nu} - \vec{\beta})$

若数据是泊松分布

$$P(n_i;\nu_i) = \frac{\nu_i^{n_i}}{n_i!}e^{-\nu_i}$$

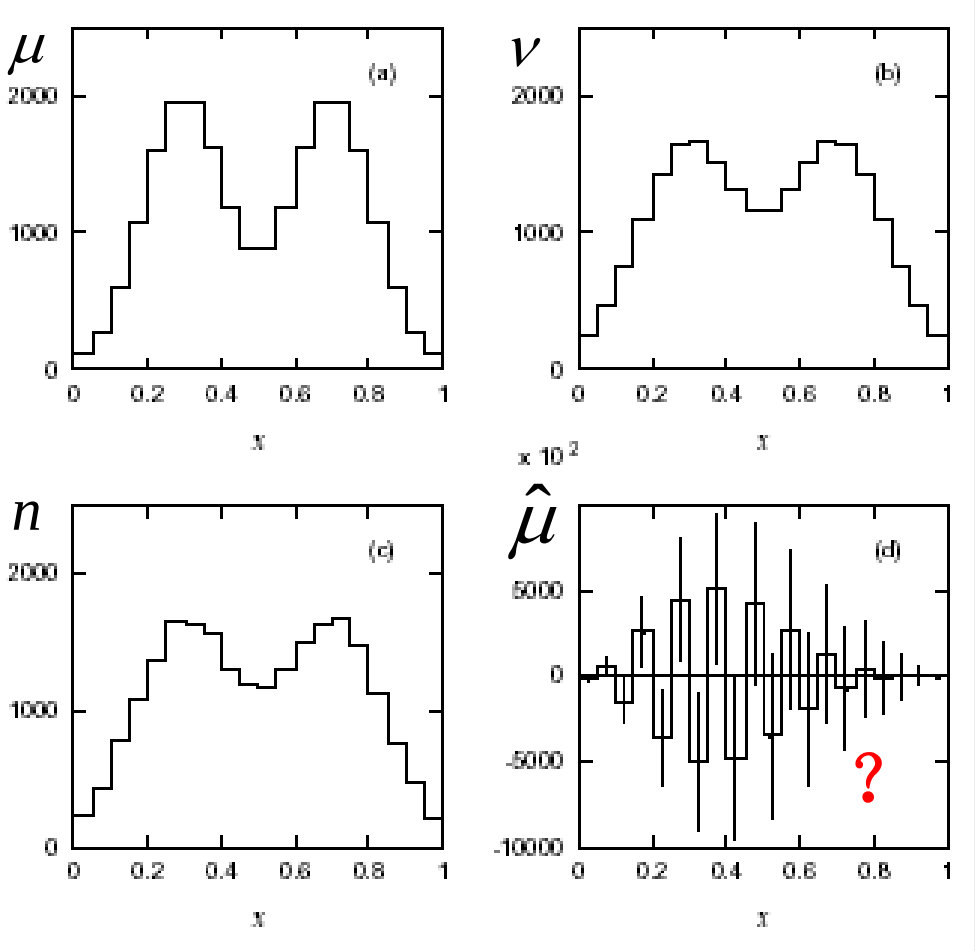
则有

$$\log L(\vec{\mu}) = \sum_{i=1}^{N}(n_i \log \nu_i - \nu_i)$$

最大似然法的估计量为

$$\hat{\vec{\nu}} = \vec{n} \Rightarrow \hat{\vec{\mu}} = R^{-1}(\vec{n} - \vec{\beta})$$

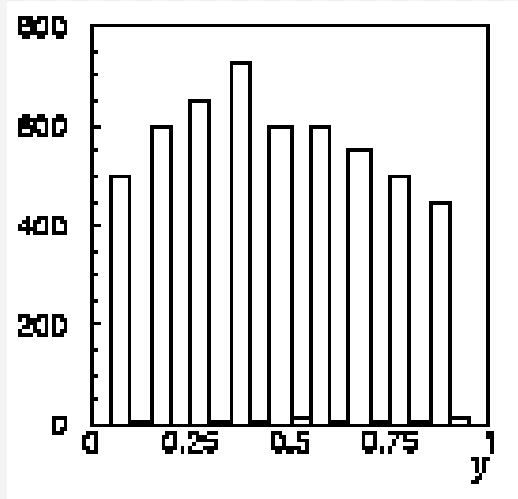
若 R 的非对角元太大，即区间宽度比分辨率要小时，会导致上式有很大的方差，以及在相邻区间产生很强的负关联。

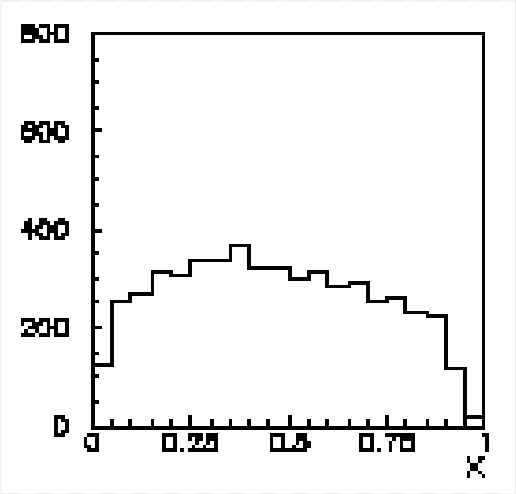


错误的原因

假设 $\vec{\mu}$ 真的有精细结构

$\vec{\mu}$ ⟶





应用R给出观测期待值时，虽然一些结构还能留下，但大部分的精细结构都被抹平了。

⟵ $\vec{\nu} = R\vec{\mu}$

应用R^{-1}到 $\vec{\nu}$ 恢复精细结构：$\vec{\mu} = R^{-1}\vec{\nu}$

但我们没有 $\vec{\nu}$ 只有 $\vec{n}$

采用观测值时，由于统计涨落的缘故，$\vec{n}$ 有不少非物理因素造成的突起。

⟶ R^{-1}“认为”这是与原来的精细结构有关，导致 $\hat{\vec{\mu}} = R^{-1}\vec{n}$ 有振荡效应。

重新研究最大似然法的解

$$E[\hat{\vec{\mu}}] = R^{-1}(E[\vec{n}] - \vec{\beta}) = \vec{\mu}$$ 是无偏的!

n_i是独立的泊松变量时，$\mathrm{cov}[n_k, n_l] = \delta_{kl}\nu_k$

计算估计量的方差

$$U_{ij} = \mathrm{cov}[\hat{\mu}_i, \hat{\mu}_j] = \sum_{k,l=1}^{N} (R^{-1})_{ik}(R^{-1})_{jl}\,\mathrm{cov}[n_k, n_l] = \sum_{k=1}^{N} (R^{-1})_{ik}(R^{-1})_{jk}\nu_k$$

利用RCF边界做无偏估计量

$$(U^{-1})_{kl} = -E\left[\frac{\partial^2 \log L}{\partial\mu_k \partial\mu_l}\right] = \sum_{i=1}^{N} \frac{R_{ik}R_{il}}{\nu_i}$$

倒数后给出

$$U_{ij} = \sum_{i=1}^{N} (R^{-1})_{ik}(R^{-1})_{jk}\nu_k$$

即使最大似然法在各无偏估计中给出的方差最小。但得到的方差可能仍然很大。

→ 为了减小方差，必须引入一些偏置量

策略：接受小的偏置量(系统误差)以换取大幅减小方差(统计误差)。

简单方法：修正因子法

对 $\vec{\mu},\vec{\nu}$ 做相同的分区，并取 $\hat{\mu}_i = C_i(n_i - \beta_i)$ ，$C_i = \dfrac{\mu_i^{MC}}{\nu_i^{MC}}$ (修正因子)

ν_i^{MC} 与 μ_i^{MC} 是来自无本底情况下的蒙特卡罗模拟结果。

$$U_{ij} = \mathrm{cov}[\hat{\mu}_i, \hat{\mu}_j] = C_i^2 \,\mathrm{cov}[n_i, n_j]$$

通常 $C_i \approx O(1)$ ，因此方差不会被放大。但偏置 $b_i = E[\hat{\mu}_i] - \mu_i$ 为

$$b_i = \left(\frac{\mu_i^{MC}}{\nu_i^{MC}} - \frac{\mu_i}{\nu_i^{sig}} \right) \nu_i^{sig}, \text{这里} \nu_i^{sig} = \nu_i - \beta_i。$$

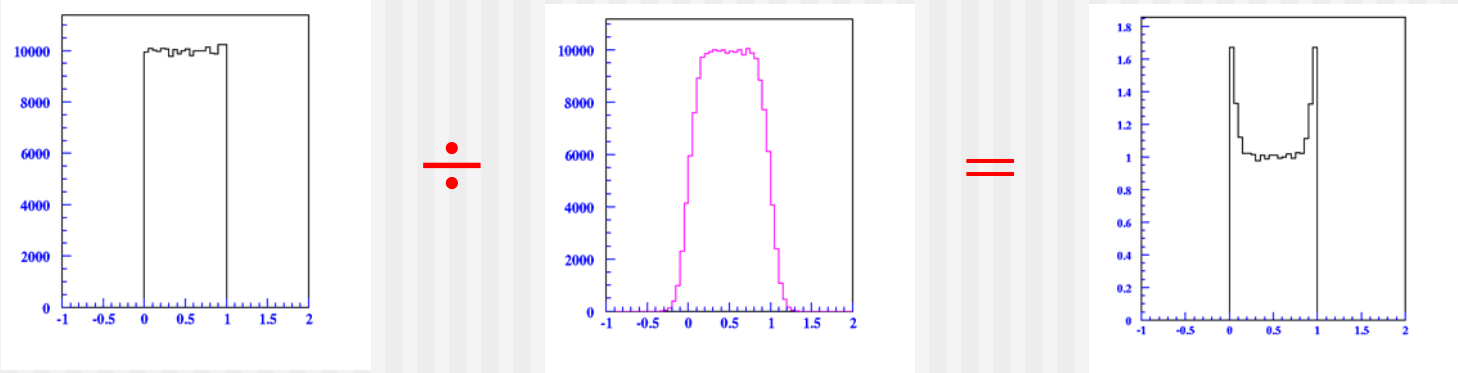
除非模拟采用的模型无误，否则上式不为零，需要考虑对应的系统误差。

注意：该偏置量存在把 $\hat{\vec{\mu}}$ 拉向 $\vec{\mu}^{MC}$的倾向，造成模型检验的困难。

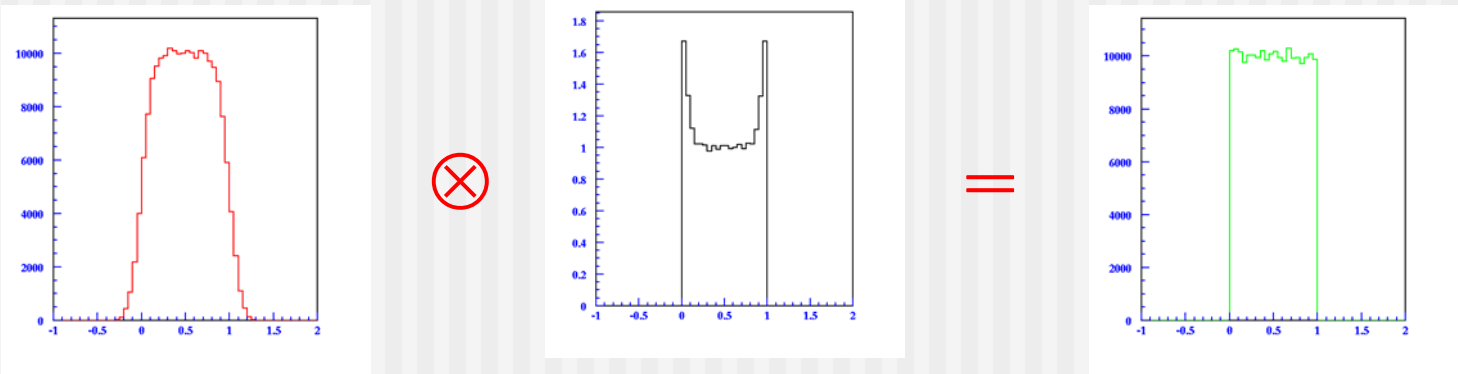
1）如果分区宽度 ≥ 几倍的分辨率，结果不会太坏。
2）实际应用中，该方法常用于事例形状变量的分布研究中。

例子：脉冲形状的还原

- 将理论(真实)的直方图除以受实验仪器影响的直方图得到修正因子



- 将观测直方图乘以修正因子直方图得到理论(真实)的直方图



正规化的开拆法

考虑“合理的”估计量，使得某些 $\Delta\log L$ 满足

$$\log L(\vec{\mu}) \geq \log L_{\max} - \Delta\log L$$

$\Delta\log L$描述了数据$\vec{n}$ 与期待值$\vec{\nu}$ 之间的“距离”。

$\vec{\mu}$ 估计量可通过将下式求最大值，选出最“光滑的”一个来构造

$$\Phi(\vec{\mu}) = \alpha \log L(\vec{\mu}) + S(\vec{\mu})$$

$S(\vec{\mu}) =$ 正则化函数(光滑的量度)

$\alpha =$ 正则化参量(选择给出欲求的$\Delta\log L$)

另外，要求开拆后对总事例数的估计为无偏的

$$\sum_{i=1}^{N} \nu_i = \sum_{i,j} R_{ij}\mu_{ij} = n_{tot}$$

因$\vec{\nu} = R\vec{\mu} + \beta$,
所以是$\vec{\mu}$ 的函数

在约束情况下将下式求最大值

$$\varphi(\vec{\mu},\lambda) = \alpha \log L(\vec{\mu}) + S(\vec{\mu}) + \lambda\left[n_{tot} - \sum_{i=1}^{N}\nu_i\right]$$

λ :拉格朗日乘子

正规化的开拆法(续)

显然，

$$\partial\varphi / \partial\lambda = 0 \quad \rightarrow \quad \sum_{i=1}^{N} \nu_i = n_{tot}$$

$\alpha = 0$ 给出最光滑的解(数据无关)

$\alpha \rightarrow \infty$ 给出最大似然解(方差可能太大)

需要正规化函数 $S(\vec{\mu})$ 与如何取 α 值的方案。

a) Tikhonov 规则
b) MaxEnt 规则

所得到的估计量的好坏由它们的偏置和方差来判断。

Tikhonov 规则

取光滑度等于第 k 阶导数均值的平方，有

$$S[f_{true}(y)] = -\int \left(\frac{d^k f_{true}(y)}{dy^k} \right)^2 dy, \text{这里} k = 1,2...$$

通常取 k=2，使得 S 约等于曲率平方的平均值。对直方图而言，也就是

$$S(\vec{\mu}) = -\sum_{i=1}^{M-2} (-\mu_i + 2\mu_{i+1} - \mu_{i+2})^2$$

Sov. Math.5(1963)1035

注意：2 阶导数对直方图的第一和最后的区间没有很好的定义。

如果在 $\log L = -\frac{1}{2}\chi^2$ 下，采用Tikhonov(k=2)规则，

$$\varphi(\vec{\mu}, \lambda) = -\frac{\alpha}{2}\chi^2(\vec{\mu}) + S(\vec{\mu})$$ 是 μ_i 的二次项

令 φ 的导数为零，给出线性方程。

在高能物理界现有好几个现成的程序：
RUN，Blobel, SVD, Höcker,…

最大熵(MaxEnt)规则

另一种表征光滑度的方法基于熵。对于一组概率而言，它表示为

$$H = -\sum_{i=1}^{M} p_i \log p_i$$

Ann. Rev. Astron. Astrophys.24 (1986)127

所有p_i相等意味着熵最大(最光滑)

有一个$p_i = 1$，其它为零，则意味着熵最小

用熵作为正规化函数，

$$S(\vec{\mu}) = H(\vec{\mu}) = -\sum_{i=1}^{M} \frac{\mu_i}{\mu_{tot}} \log \frac{\mu_i}{\mu_{tot}}$$

$$\propto \log(\mu_{tot}\text{填入M个区间中各种可能的总数})$$

有时侯，根据贝叶斯统计 $S(\vec{\mu}) \to \vec{\mu}$ 的先验概率密度函数(?)

这里，我们坚持采用经典近似：估计量的好坏由偏置，方差来判断。

注意：熵并不取决于区间的顺序。

$\hat{\vec{\mu}}$ 的方差与偏置

一般来说，决定 $\hat{\vec{\mu}}(\vec{n})$ 的方程是非线性的。在 $\vec{n}_{obs}$ 附近展开 $\hat{\vec{\mu}}(\vec{n})$

$$\hat{\vec{\mu}}(\vec{n}) \approx \hat{\vec{\mu}}_{obs} - A^{-1}B(\vec{n} - \vec{n}_{obs}),$$

$$A_{ij} = \begin{cases} \dfrac{\partial^2 \varphi}{\partial \mu_i \partial \mu_j}, & i, j = 1, \ldots, M, \\ \dfrac{\partial^2 \varphi}{\partial \mu_i \partial \lambda} = -1, & i = 1, \ldots, M,\ j = M+1, \\ \dfrac{\partial^2 \varphi}{\partial \lambda^2} = 0, & i = M+1,\ j = M+1, \end{cases}$$

$$B_{ij} = \begin{cases} \dfrac{\partial^2 \varphi}{\partial \mu_i \partial n_j}, & i = 1, \ldots, M,\ j = 1, \ldots, N, \\ \dfrac{\partial^2 \varphi}{\partial \lambda \partial n_j} = 1, & i = M+1,\ j = 1, \ldots, N. \end{cases}$$

φ为非正规的似然函数

G. Cowan, Statistical Data Analysis, Oxford University Press(1998)

$\hat{\vec{\mu}}$ 的方差与偏置(续)

利用误差传递得到协方差 $U_{ij} = \mathrm{cov}[\hat{\mu}_i, \hat{\mu}_j],$

$$U = CVC^T \qquad \text{这里} \qquad C = A^{-1}B,$$

以及对偏置的估计量，$b_i = E[\hat{\mu}_i] - \mu_i,$

$$\hat{b}_i = \sum_{j=1}^{N} C_{ij}(\hat{\nu}_j - n_j) = \sum_{j=1}^{N} \frac{\partial \hat{\mu}_i}{\partial n_j}(\hat{\nu}_j - n_j),$$

此处 $\hat{\vec{\nu}} = R\hat{\vec{\mu}} + \vec{\beta}.$ 而且通常情况下 $\hat{\vec{\nu}} \neq \vec{n}$

正规化参数α的选取

α 决定了置于数据的权重大小以便能与光滑度相比较， α =0给出最大的光滑估计值，并与数据无关。因此虽然方差为零，但有明显的偏置。而取大的α ，则回到高度振荡无偏的最大似然解。

为了在偏置与方差之间达到最大平衡:选择α使均值误差的平方最小

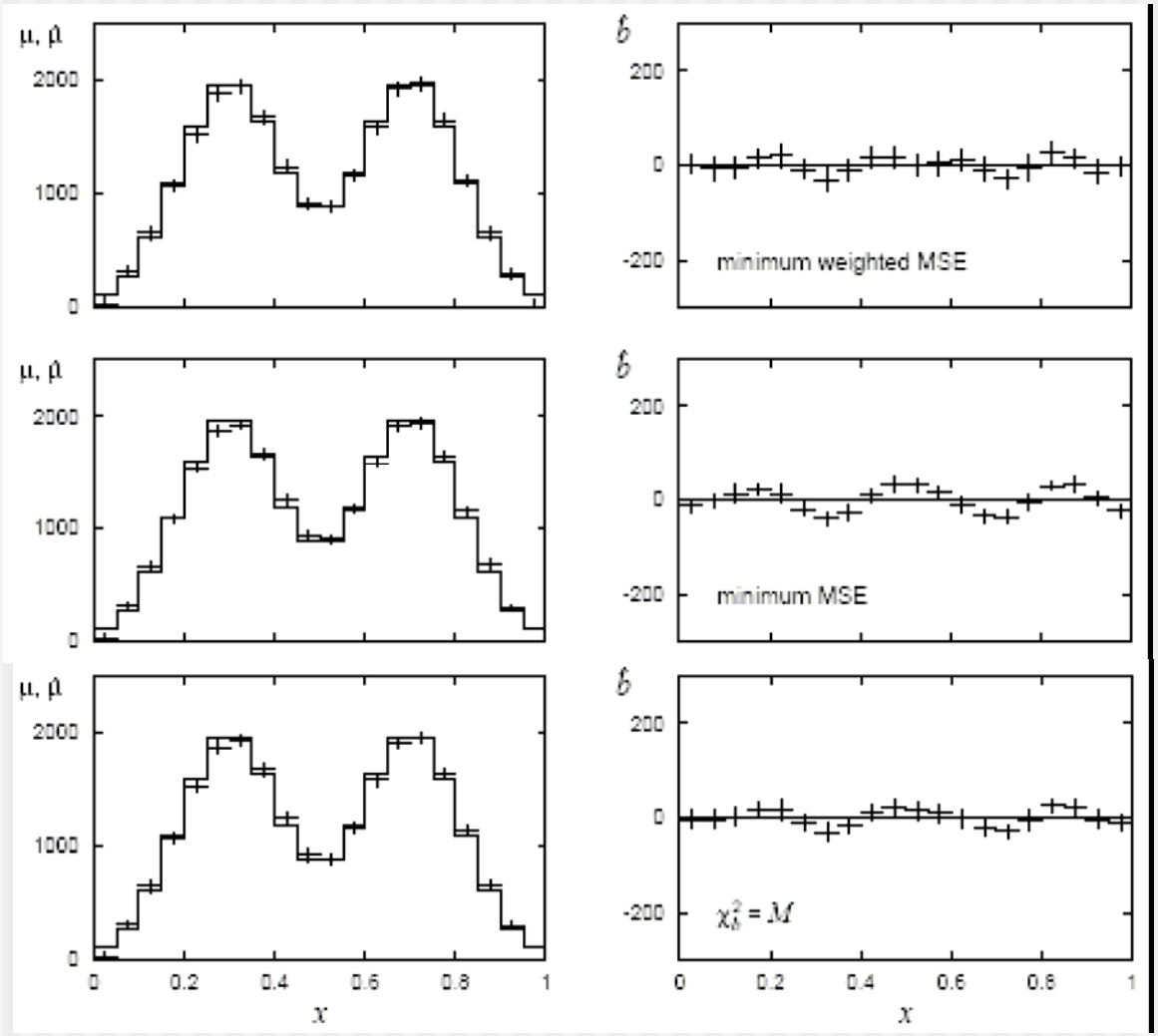
$$MSE = \frac{1}{M}\sum_{i=1}^{M}\left(U_{ii} + \hat{b}_i^2\right), \quad \text{或} \quad Weighted\ MSE = \frac{1}{M}\sum_{i=1}^{M}\frac{U_{ii} + \hat{b}_i^2}{\hat{\mu}_i}.$$

或要求偏置不大于它自身的估计方差 $\hat{W}_{ii}$。它可以找到α的值使得

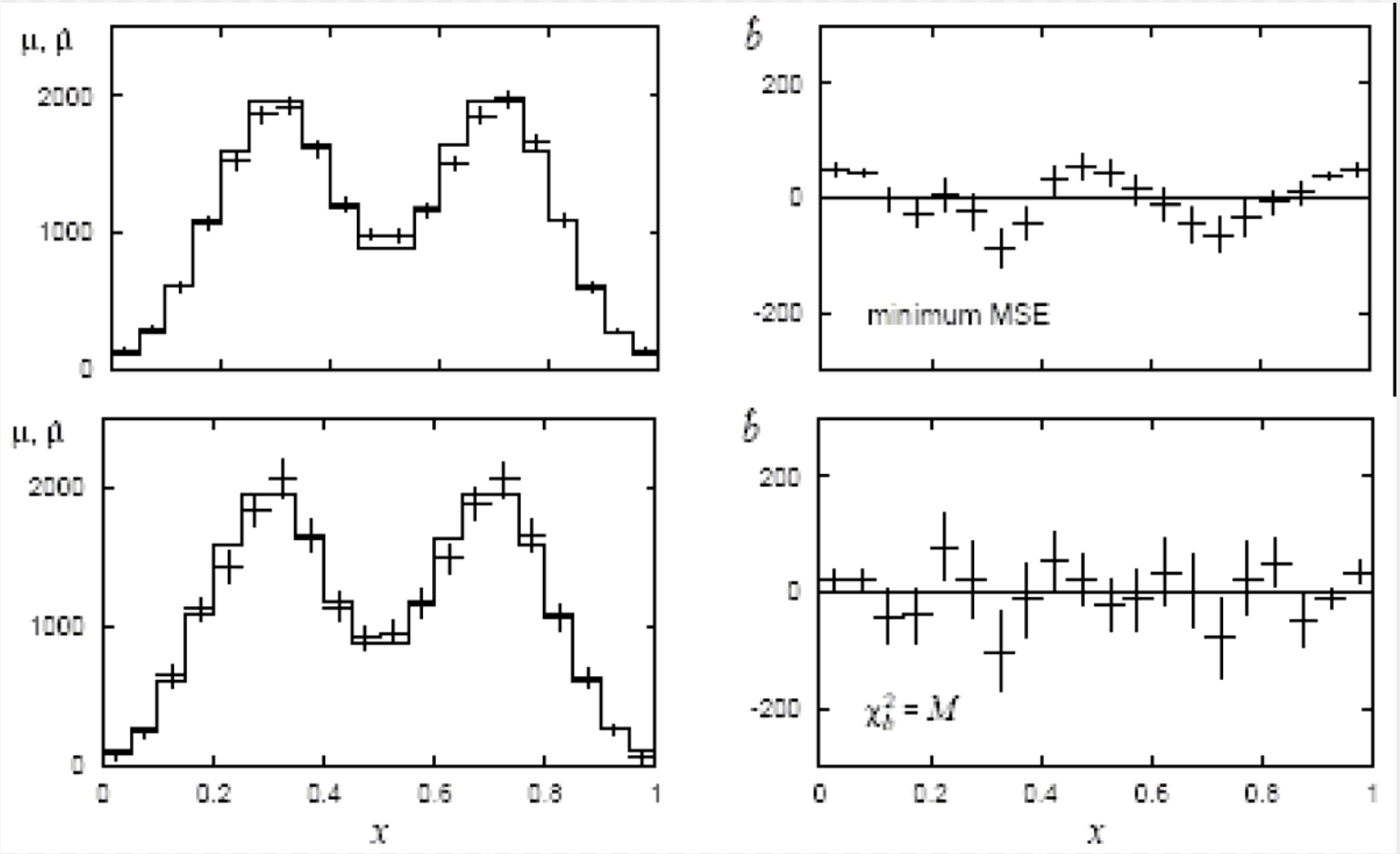
$$\chi_b^2 = \sum_{i=1}^{M}\frac{\hat{b}_i^2}{\hat{W}_{ii}} = M \quad \text{这里} \quad \hat{W}_{ij} = \mathrm{cov}[\hat{b}_i, \hat{b}_j].$$

G. Cowan, Statistical Data Analysis, Oxford University Press(1998)
M。 Schmelling, NIM A340(1994)400

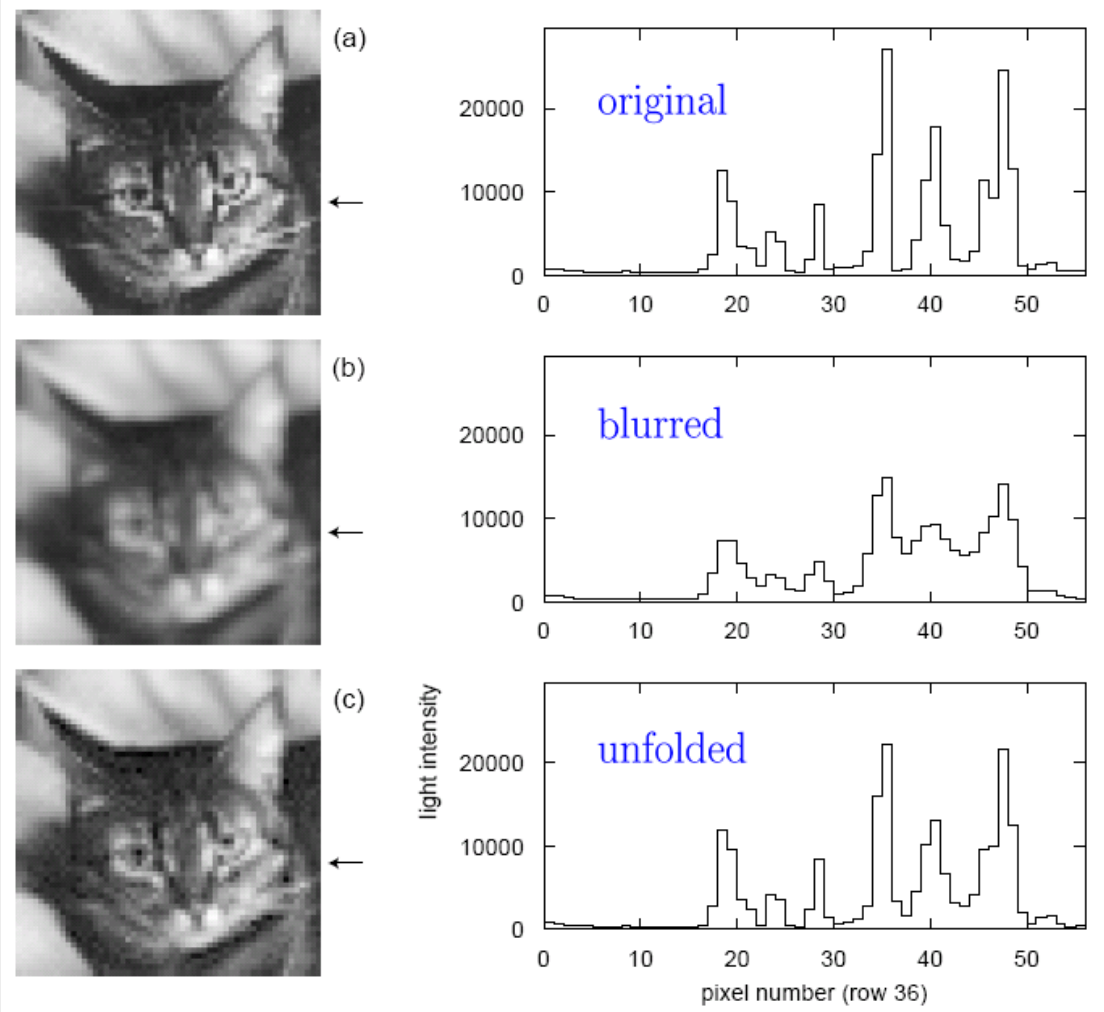
例子:Tikhonov规则($k=2$)



例子:最大熵(MaxEnt)规则



一个在图像处理中的最大熵例子



最大熵值方法常用于天文观测图像的重建,与点源的偏置较小,易于推广到两维以上的情况。

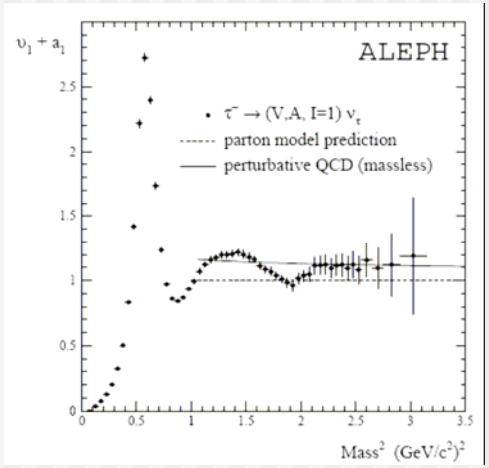
例子： τ 的谱函数

为了测定奇异夸克质量，实验上可采用比较$\tau \to X(ud)\nu_\tau$与$\tau \to X(us)\nu_\tau$中，X的质量平方差

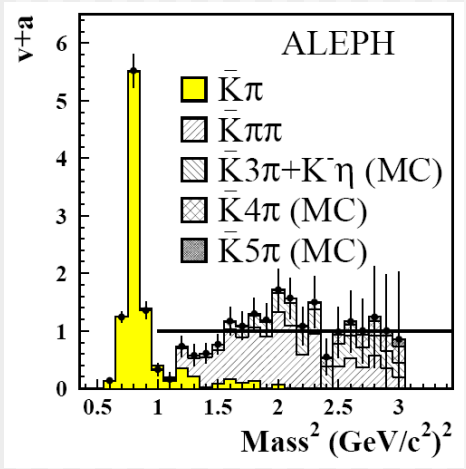
$$\Delta^{00} = M^2(ud) - M^2(us) \Rightarrow m_s$$

Eur. Phys. J. *C11: 599, 1999*

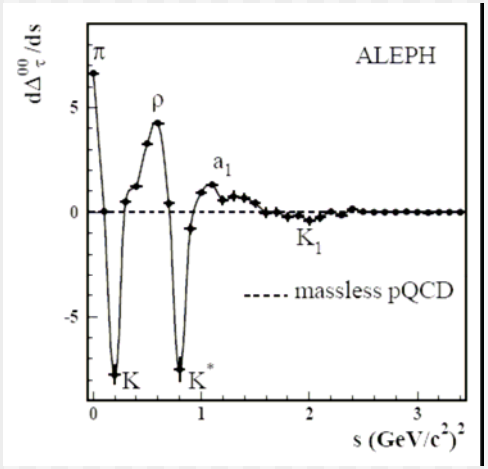
由于探测器对两者影响各不相同，因此，需要用开拆法求出“真实”分布。



Tikhonov 规则



修正因子方法



小结

1. 数学上的原理

真实直方图$\vec{\mu}$，数据 $\vec{n}$ 以及其期待值 $\vec{\nu}$，满足 $\vec{\mu} = R\vec{\nu} + \vec{\beta}$,目标是构造 $\vec{\mu}$ 的估计量。

2. 求反应矩阵的逆

有很大的振荡行为(及大的方差)，但在各种无偏解中具有零偏置与最小的方差。

3. 修正因子

$C_i = \mu_i^{MC} / \nu_i^{MC}$, 方法又快又简单。

4. 正规的开拆过程 Tikhonov:从第 k阶导数的均方值中进行光滑处理

MaxEnt:从$H = -\sum_i p_i \log p_i$熵中进行光滑处理

5. 估计量的方差与偏置

在求解过程中采用了线性近似，因而不是无偏的

6. 正规化参数的选择

无最好的方案，可以采用χ^2 = M(区间总数)的方案。

7. 例子

只要探测器的响应可知，就一定可以得到真实的分布