Lecture 10 outline

Unfolding

1. **Mathematical formulation, response function (matrix)**
2. **Inverting the response matrix**
3. **Correction factors**
4. **Regularized unfolding**
   (a) Tikhonov
   (b) MaxEnt
5. **Variance and bias of the estimators**
6. **Choice of the regularization parameter**
7. **Some examples**

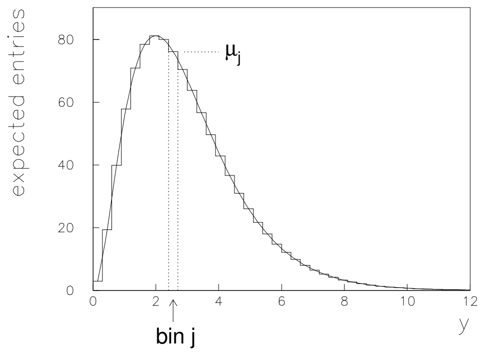
Formulation of the unfolding problem

Consider random variable y. Goal: determine pdf $f(y)$.

If parametrization $f(y;\vec{\theta})$ known,

maximum likelihood $\rightarrow \hat{\vec{\theta}}$

If no parametrization available, construct histogram:



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

$$\mu_j = \mu_{\mathrm{tot}}\, p_j \leftarrow \text{the ‘true histogram’}$$

The goal: construct estimators for the μ_j (or p_j).

$\rightarrow$ number of parameters = number of bins, M

The problem: y cannot be measured without error.

$\rightarrow$ migration of entries between bins

$\rightarrow$ $f(y)$ is ‘smeared out’, peaks broadened

Response matrix

Effect of measurement errors: $y =$ true value

$x =$ observed value

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

$\downarrow$ discretize

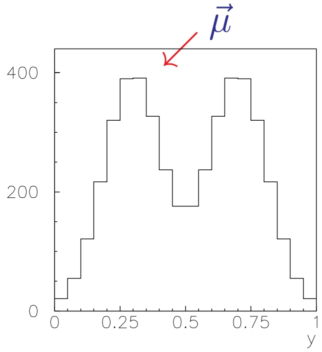
$$\nu_i = \sum_{j=1}^{M} R_{ij}\, \mu_j\,, \quad i = 1, \ldots, N$$

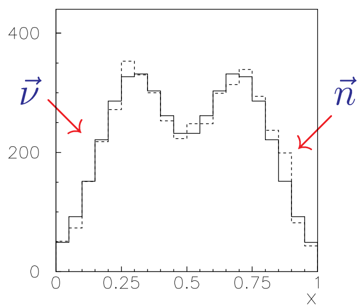
observed histogram (expec. val.) — response matrix — true histogram

$$R_{ij} = P(\text{observed in bin } i \,|\, \text{true value in bin } j)$$

The data: $\vec{n} = (n_1, \ldots, n_N)$, where $\nu_i = E[n_i]$.

N.B. $\vec{\mu}$, $\vec{\nu}$ constants, $\vec{n}$ subject to statistical fluctuations.





Efficiency, background

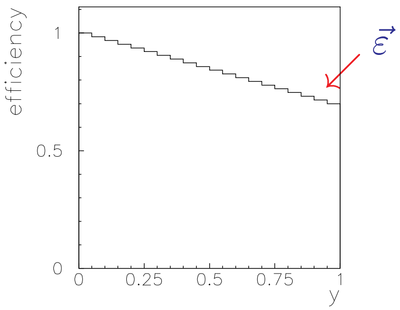
Sometimes an event goes undetected,

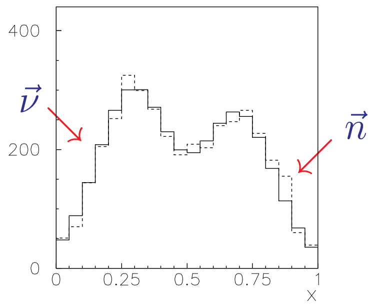
$$\sum_{i=1}^{N} R_{ij} = \sum_{i=1}^{N} P(\text{observed in bin } i \,|\, \text{true value in bin } j)$$

$$= P(\text{observed anywhere} \,|\, \text{true value in bin } j)$$

$$= \varepsilon_j \qquad (\text{efficiency})$$

N.B. ε_j depends on bin j of *true* histogram.





Sometimes we observe something when no true event occurred,

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

$\beta_i =$ expected number of background events in *observed* histogram.

For now, assume $\vec{\beta}$ is known.

Summary of ingredients

'true' histogram: $\vec{\mu} = (\mu_1, \ldots, \mu_M)$, $\mu_{\rm tot} = \sum_{j=1}^{M} \mu_j$

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

expectation values for observed histogram: $\vec{\nu} = (\nu_1, \ldots, \nu_N)$

observed histogram: $\vec{n} = (n_1, \ldots, n_N)$

response matrix: $R_{ij} = P(\text{observed in bin } i \,|\, \text{true value in bin } j)$

efficiencies: $\varepsilon_j = \sum_{i=1}^{N} R_{ij}$

expected background: $\vec{\beta} = (\beta_1, \ldots, \beta_N)$

These are related by:

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

To find estimators for $\vec{\mu}$, we need probability law, e.g.

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

or covariance matrix,

$$V_{ij} = \text{cov}[n_i, n_j],$$

in order to construct likelihood function or χ^2.

Why unfold?

Often unfolding not needed, e.g. when

comparing to prediction of existing theory, better to

'fold' theory with detector response,

i.e. include detector effects in its prediction,

compare this with uncorrected ('raw') data $\vec{n}$.

→ simpler, more robust.

But, 'folding' theory with detector effects requires response matrix,

usually this knowledge not retained after publication of result.

Unfolded distribution can be compared directly to:

predictions of theories,

unfolded results from other experiments.

Usually unfolded result is more useful, since new theories

may be invented when response matrix is long gone.

In HEP often unfold:

structure functions $F_2(x, Q^2)$,

τ spectral functions (hadronic mass distributions),

hadronic event-shape distributions,

particle multiplicity distributions.

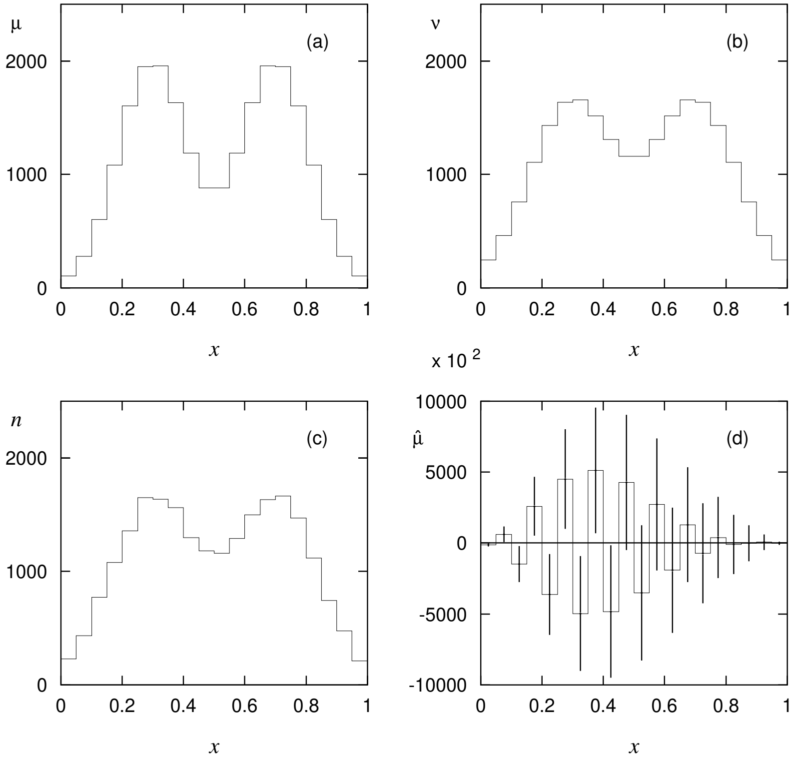
Inverting the response matrix

Assume $\vec{\nu} = R\vec{\mu} + \vec{\beta}$ can be inverted: $\vec{\mu} = R^{-1}(\vec{\nu} - \vec{\beta})$

Suppose data are Poisson: $P(n_i; \nu_i) = \dfrac{\nu_i^{n_i}}{n_i!} e^{-\nu_i}$

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

ML estimator is $\hat{\vec{\nu}} = \vec{n} \;\rightarrow\; \hat{\vec{\mu}} = R^{-1}(\vec{n} - \vec{\beta})$.

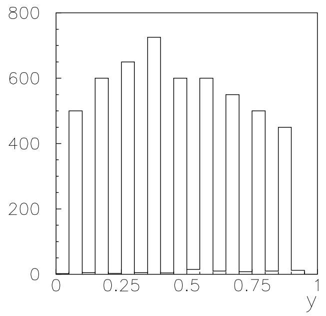


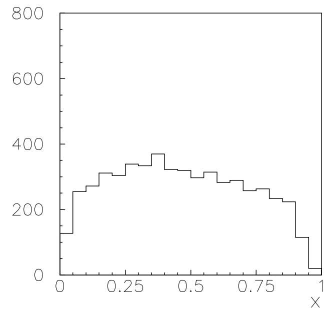
Catastrophic failure (?!)

What went wrong?

Suppose $\vec{\mu}$ really had fine structure:

$\vec{\mu} \rightarrow$





Applying R washes this out, but leaves residual structure.

$\leftarrow \vec{\nu} = R\vec{\mu}$

Applying R^{-1} to $\vec{\nu}$ puts the fine structure back: $\vec{\mu} = R^{-1}\vec{\nu}$.

But we don't have $\vec{\nu}$, only $\vec{n}$.

$\vec{n}$ *has* small bumps due to statistical fluctuations.

$\rightarrow R^{-1}$ 'thinks' this is residual of original fine structure,

$\hat{\vec{\mu}} = R^{-1}\vec{n}$ winds up getting huge 'fine structure'.

ML solution revisited

$$E[\hat{\vec{\mu}}] = R^{-1}(E[\vec{n}] - \vec{\beta}) = \vec{\mu} \;\rightarrow\; \text{unbiased!}$$

Compute variance of estimators,

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

$$= \sum_{k=1}^{N} (R^{-1})_{ik}\,(R^{-1})_{jk}\,\nu_k$$

Recall RCF bound for unbiased estimators,

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

Inverting gives

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

→ ML estimator has minimum variance among unbiased estimators.

But this variance was huge!

→ to reduce variance, we must introduce some bias.

Strategy: accept small bias (systematic error) in exchange for large reduction in variance (statistical error).

Correction factor method

Use equal binning for $\vec{\mu}$, $\vec{\nu}$ and take $\hat{\mu}_i = C_i(n_i - \beta_i)$, where

$$C_i = \frac{\mu_i^{\mathrm{MC}}}{\nu_i^{\mathrm{MC}}} \quad \text{(correction factor)}$$

ν_i^{MC} and μ_i^{MC} from Monte Carlo simulation (no background).

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

Usually $C_i \approx O(1)$, so variances don't blow up.

But the bias $b_i = E[\hat{\mu}_i] - \mu_i$ is

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

Need to include systematic error due to MC dependence.

N.B. bias tends to pull $\hat{\vec{\mu}}$ towards $\vec{\mu}^{\mathrm{MC}}$

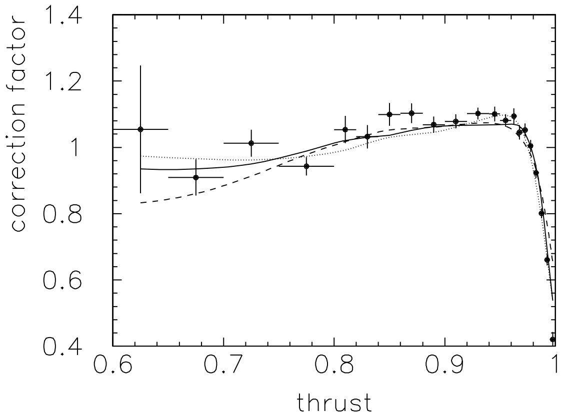
→ hard to test models.

Not too bad if bin width $\geq$ several times resolution.

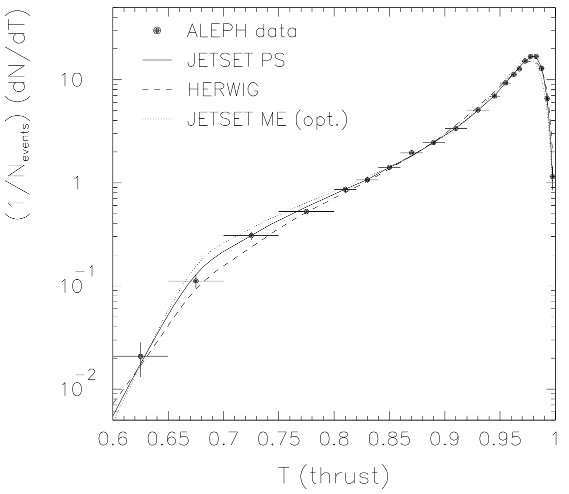
Often used for distributions of event-shape variables.

Example of correction factors in $e^+e^- \to$ hadrons

The correction factors:



The unfolded distribution compared to model predictions:



Regularized unfolding

Consider 'reasonable' estimators such that for some $\Delta \log L$,

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

Out of these estimators, choose the 'smoothest', by maximizing

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

$S(\vec{\mu})$ = regularization function (measure of smoothness),

α = regularization parameter (choose to give desired $\Delta \log L$)

In addition require $\sum_{i=1}^{N} \nu_i = \sum_{i,j} R_{ij}\mu_j = n_{\text{tot}}$, i.e. maximize

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

where λ is a Lagrange multiplier,

$$\partial\varphi/\partial\lambda = 0 \to \sum_{i=1}^{N} \nu_i = n_{\text{tot}}.$$

$\alpha = 0$ gives smoothest solution (ignores data!),

$\alpha \to \infty$ gives ML solution (variance too large).

We need: regularization function $S(\vec{\mu})$,

a prescription for setting α.

Goodness of resulting estimators judged by their bias and variance.

Tikhonov regularization

Take measure of smoothness = mean square of kth derivative,

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

Often take $k = 2$, $\rightarrow$ $S \approx$ mean squared curvature.

For histogram this becomes (e.g. for $k = 2$),

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

N.B. 2nd derivative not well defined for first and last bins.

If we use Tikhonov ($k = 2$) with $\log L = -\frac{1}{2}\chi^2$,

$$\varphi(\vec{\mu}, \lambda) = -\frac{\alpha}{2}\chi^2(\vec{\mu}) + S(\vec{\mu}) \quad \text{quadratic in } \mu_i,$$

$\rightarrow$ setting derivatives of φ equal to zero gives linear equations.

Several programs available for use in HEP:

`RUN`, Blobel

`SVD`, Höcker and Kartvelishvili

Regularization function based on entropy (MaxEnt)

Shannon entropy of a set of probabilities is

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

All p_i equal $\rightarrow$ maximum entropy (maximum smoothness)

One $p_i = 1$, all others $= 0$ $\rightarrow$ minimum entropy

Use entropy as regularization function,

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

$$\propto \log(\text{number of ways to arrange } \mu_{\rm tot} \text{ entries in } M \text{ bins})$$

Sometimes motivated by Bayesian statistics,

$$S(\vec{\mu}) \rightarrow \text{prior pdf for } \vec{\mu} \quad (?)$$

Here stay with classical approach:

goodness of estimator judged by bias, variance.

N.B. Entropy does not depend on order of bins.

G. Cowan – Computing and Statistical Data Analysis – University of London Postgraduate Lectures

Variance and bias of $\hat{\vec{\mu}}$

In general, the equations determining $\hat{\vec{\mu}}(\vec{n})$ are nonlinear.

Expand $\hat{\vec{\mu}}(\vec{n})$ about $\vec{n}_{\rm obs}$ (observed data set),

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

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

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

Use error propagation to get covariance $U_{ij} = {\rm cov}[\hat{\mu}_i, \hat{\mu}_j]$,

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

and estimators for the bias, $b_i = E[\hat{\mu}_i] - \mu_i$,

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

where $\hat{\vec{\nu}} = R\hat{\vec{\mu}} + \vec{\beta}$. (N.B. $\hat{\vec{\nu}} \neq \vec{n}$.)

Choosing the regularization parameter α

$\alpha = 0 \rightarrow \hat{\vec{\mu}}$ maximally smooth (ignores data).

$\alpha \rightarrow \infty \rightarrow$ ML solution (no bias, very large variance).

Possible criteria for best trade-off between bias and variance:

Minimize mean squared error,

$$\text{MSE} = \frac{1}{M}\sum_{i=1}^{M}(U_{ii} + \hat{b}_i^2)\,, \quad \text{or}$$

$$\text{MSE}' = \frac{1}{M}\sum_{i=1}^{M}\frac{U_{ii} + \hat{b}_i^2}{\hat{\mu}_i}.$$

Or look at changes in χ^2 from unregularized (ML) solution,

$$\Delta\chi^2 = 2\Delta\log L = N\,, \quad \text{or}$$

$$\Delta\chi^2_{\rm eff} = (\hat{\vec{\nu}} - \vec{n})^T\, R\,C\,V^{-1}(R\,C)^T\,(\hat{\vec{\nu}} - \vec{n}) = 1\,.$$

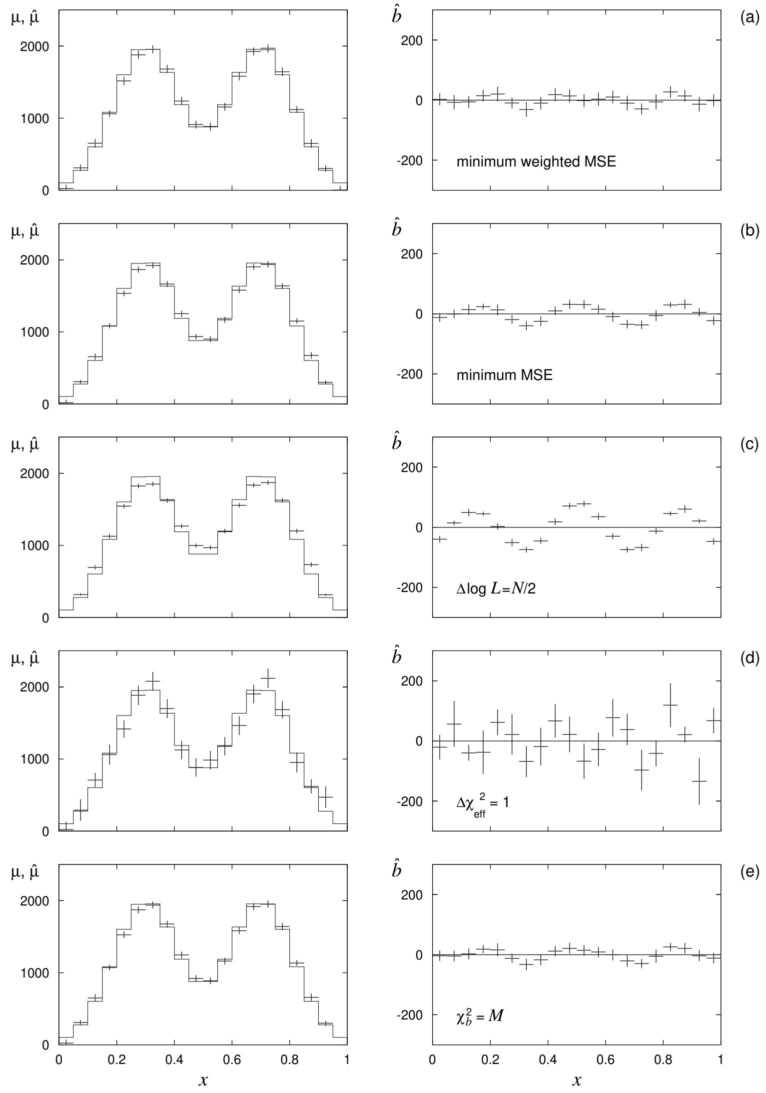
Or require that bias be consistent with zero to within its own error,

$$\chi^2_b = \sum_{i=1}^{M}\frac{\hat{b}_i^2}{W_{ii}} = M \quad \text{where } W_{ij} = {\rm cov}[\hat{b}_i, \hat{b}_j].$$

i.e. if bias significantly different from zero, we would subtract it;

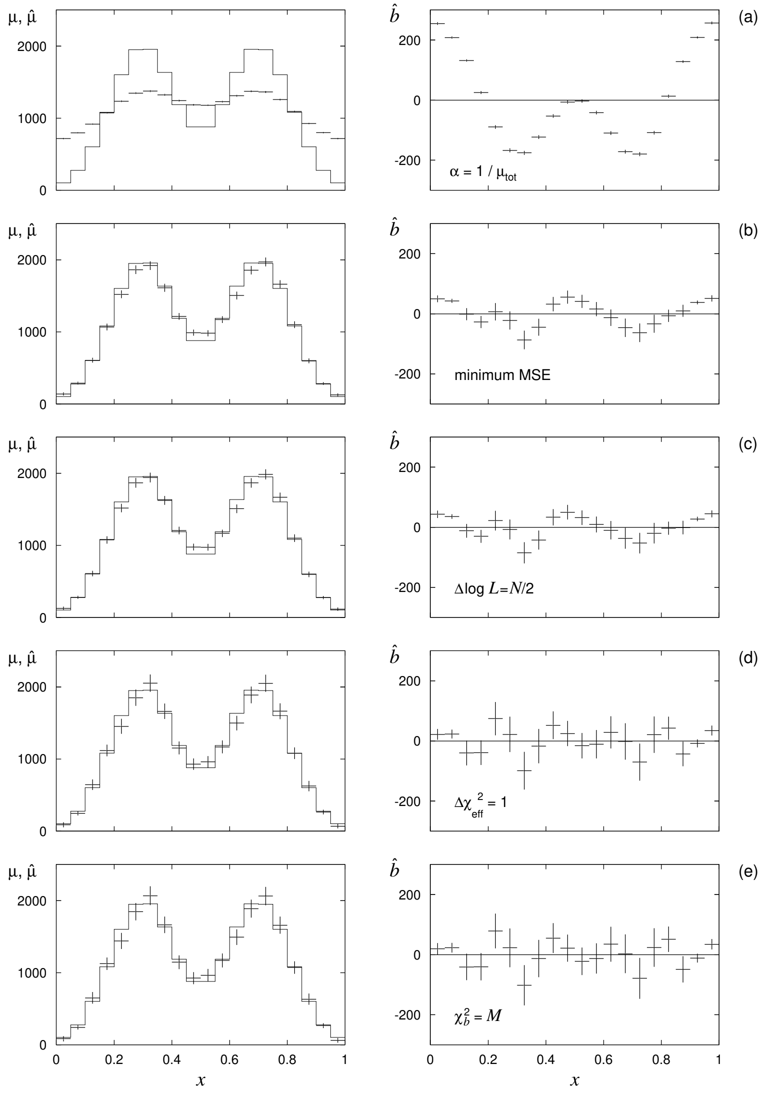
$\rightarrow$ equivalent to going to smaller $\Delta\log L$ or larger α (less bias).

Examples with Tikhonov regularization ($k = 2$)

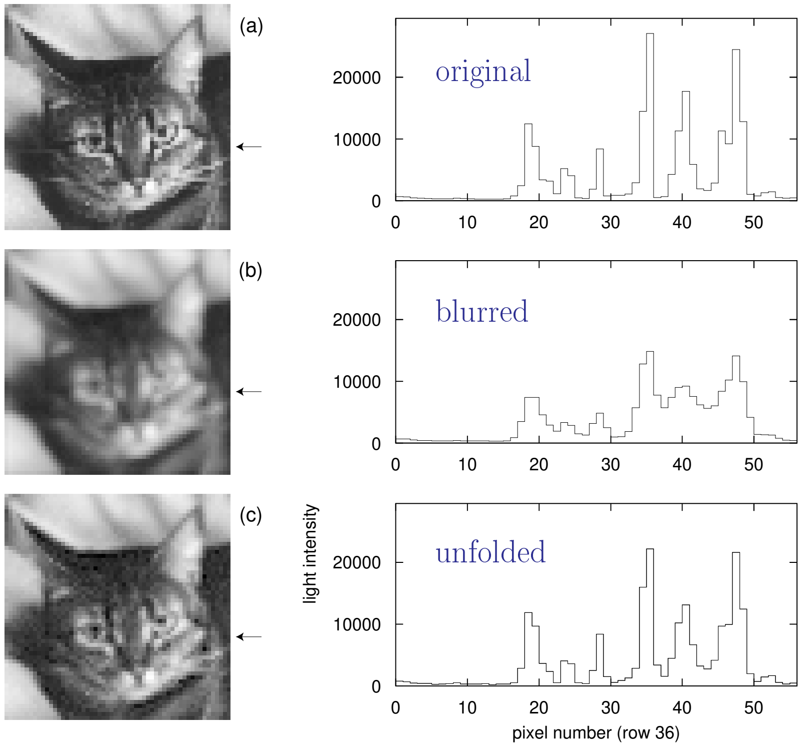


N.B. solution not always positive.

Examples with MaxEnt regularization



A MaxEnt example with image reconstruction (Newton)



MaxEnt often used in astronomical image reconstruction,

only small bias against point sources (peaks),

easy to generalize to two (or more) dimensions.

Lecture 10 summary

Unfolding

1. **Mathematical formulation:**

   true histogram: $\vec{\mu} = (\mu_1, \ldots, \mu_M)$

   data: $\vec{n} = (n_1, \ldots, n_N)$

   expectation values of $\vec{n}$: $\vec{\nu} = (\nu_1, \ldots, \nu_N)$

   $\vec{\mu} = R\vec{\nu} + \vec{\beta}$

   Goal: construct estimators for $\vec{\mu}$

2. **Inverting the response matrix:** huge oscillations (large variance) but zero bias and minimum variance among unbiased solutions.

3. **Correction factors:** quick and simple.

4. **Regularized unfolding:**

   Tikhonov: smoothness from mean square kth derivative.

   MaxEnt: smoothness from entropy $H = -\sum_i p_i \log p_i$.

5. **Variance and bias of the estimators:** based on linearized approximation to solution.

6. **Choice of the regularization parameter:** no clear winner (but $\chi^2_b = M$ my favorite).

7. **Examples:** anything where structure smeared out, detector response is known, no parametrization of true distribution available.