Introduction to Unfolding

LABORATOIRE
DE L'ACCÉLÉRATEUR
LINÉAIRE

Lectures at LAL Orsay,
17-19 December, 2018

ROYAL
HOLLOWAY
UNIVERSITY
OF LONDON

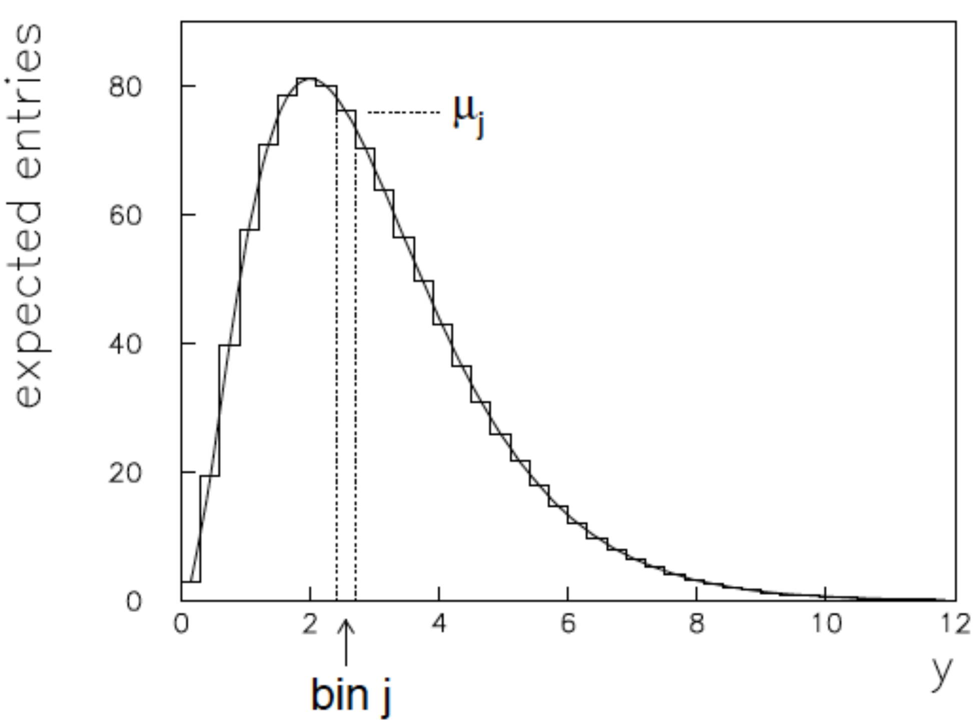
Glen Cowan
Physics Department
Royal Holloway, University of London
g.cowan@rhul.ac.uk
www.pp.rhul.ac.uk/~cowan

Unfolding: formulation of the problem

Consider a random variable y, goal is to determine pdf $f(y)$.

If parameterization $f(y;\boldsymbol{\theta})$ known, find e.g. ML estimators $\hat{\theta}$.

If no parameterization available, construct histogram:



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

$$\mu_j = \mu_{\text{tot}} p_j$$

"true" histogram

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

Migration

Effect of measurement errors: *y* = true value, *x* = observed value,

migration of entries between bins,

f(*y*) is ‘smeared out’, peaks broadened.

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

↓

discretize: data are $\mathbf{n} = (n_1, \ldots, n_N)$

$$\nu_i = E[n_i] = \sum_{j=1}^{M} R_{ij}\mu_j \;, \quad i = 1, \ldots, N$$

response matrix

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

Note $\boldsymbol{\mu}$, $\boldsymbol{\nu}$ are constants; ***n*** subject to statistical fluctuations.

Efficiency, background

Sometimes an event goes undetected:

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

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

$$= \varepsilon_j \qquad \longleftarrow \quad \text{efficiency}$$

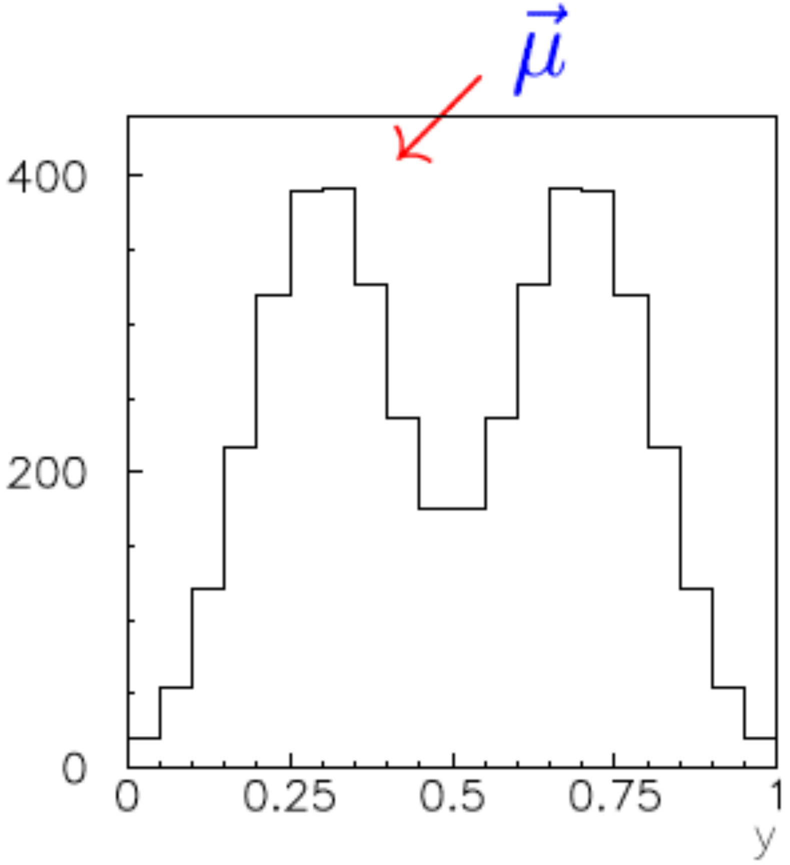
Sometimes an observed event is due to a background process:

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

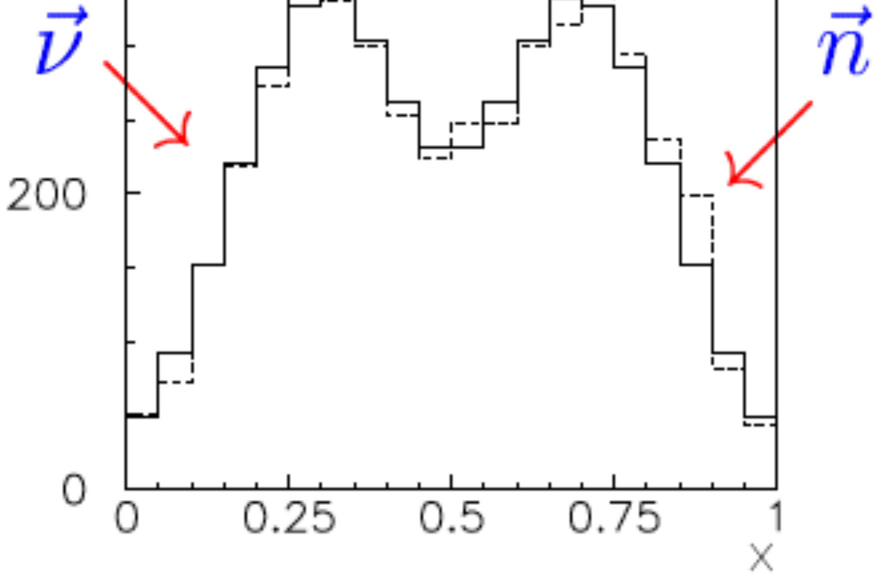
β_i = expected number of background events in *observed* histogram.

For now, assume the β_i are known.

The basic ingredients



"true"



"observed"

Summary of ingredients

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

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

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

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

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

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

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

These are related by:

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

Maximum likelihood (ML) estimator from inverting the response matrix

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

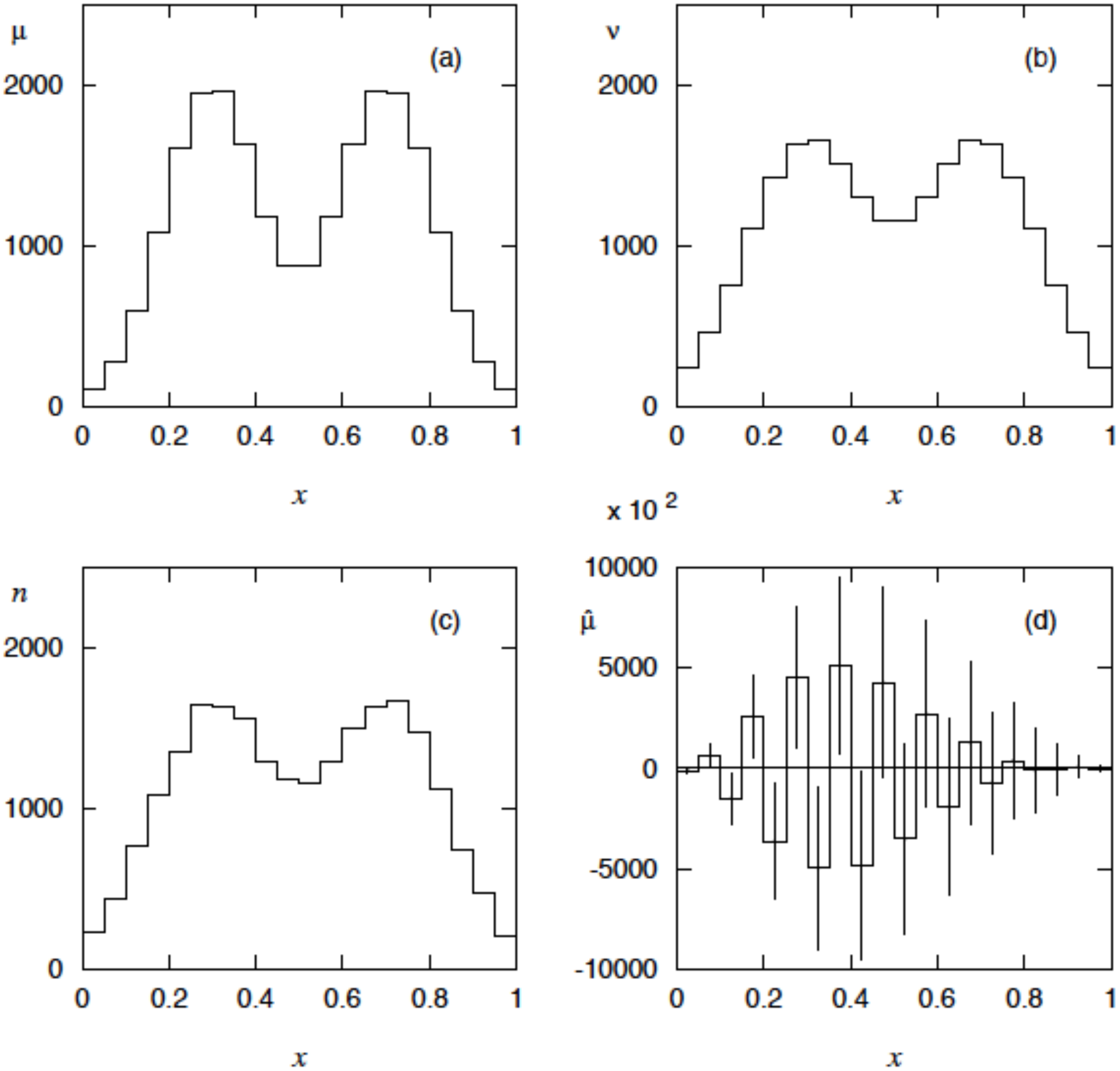
Suppose data are independent Poisson: $P(n_i; \nu_i) = \dfrac{\nu_i^{n_i}}{n_i!} e^{-\nu_i}$

So the log-likelihood is $\ln L(\boldsymbol{\mu}) = \displaystyle\sum_{i=1}^{N} (n_i \ln \nu_i - \nu_i)$

ML estimator is $\hat{\boldsymbol{\nu}} = \mathbf{n}$

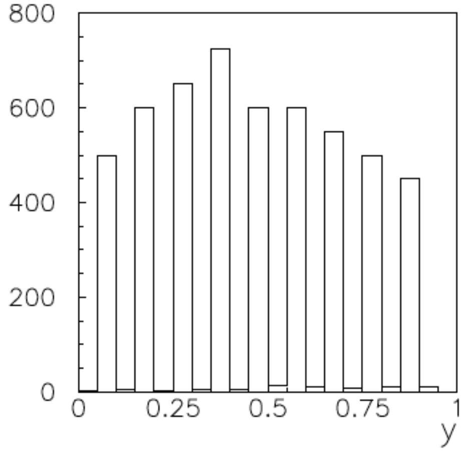
$$\longrightarrow \qquad \hat{\boldsymbol{\mu}} = R^{-1}(\mathbf{n} - \boldsymbol{\beta})$$

Example with ML solution



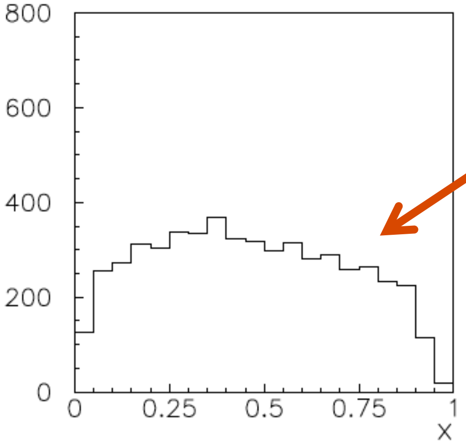
Catastrophic failure???

What went wrong?



Suppose $\boldsymbol{\mu}$ really had a lot of fine structure.

$\vec{\mu}$



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

Applying R washes this out, but leaves a residual structure:

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

But we don't have $\boldsymbol{\nu}$, only $\boldsymbol{n}$. R^{-1} "thinks" fluctuations in $\boldsymbol{n}$ are the residual of original fine structure, puts this back into $\hat{\mu}$.

ML solution revisited

For Poisson data the ML estimators are unbiased:

$$E[\hat{\boldsymbol{\mu}}] = R^{-1}(E[\mathbf{n}] - \boldsymbol{\beta}) = \boldsymbol{\mu}$$

Their covariance is:

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

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

(Recall these statistical errors were huge for the example shown.)

ML solution revisited (2)

The information inequality gives for unbiased estimators the minimum (co)variance bound:

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

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

This is the same as the actual variance! I.e. ML solution gives smallest variance among all unbiased estimators, even though this variance was huge.

In unfolding one must accept some bias in exchange for a (hopefully large) reduction in variance.

Correction factor method

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

$$C_i = \frac{\mu_i^{\text{MC}}}{\nu_i^{\text{MC}}}$$

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

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

Often $C_i \sim O(1)$ so statistical errors far smaller than for ML.

But the bias $b_i = E[\hat{\mu}_i] - \mu_i$ is $b_i = \left(\frac{\mu_i^{\text{MC}}}{\nu_i^{\text{MC}}} - \frac{\mu_i}{\nu_i^{\text{sig}}} \right)$

Nonzero bias unless MC = Nature.

$$\nu_i^{\text{sig}} = \nu_i - \beta_i$$

Example from Bob Cousins

Reality check on the statistical errors

Suppose for some bin *i* we have:

$$C_i = 0.1 \qquad \beta_i = 0 \qquad n_i = 100$$

→ $\hat{\mu}_i = C_i n_i = 10 \qquad \sigma_{\hat{\mu}_i} = C_i \sqrt{n_i} = 1.0$ (10% stat. error)

But according to the estimate, only 10 of the 100 events found in the bin belong there; the rest spilled in from outside.

How can we have a 10% measurement if it is based on only 10 events that really carry information about the desired parameter?

Discussion of correction factor method

As with all unfolding methods, we get a reduction in statistical error in exchange for a bias; here the bias is difficult to quantify (difficult also for many other unfolding methods).

The bias should be small if the bin width is substantially larger than the resolution, so that there is not much bin migration.

So if other uncertainties dominate in an analysis, correction factors may provide a quick and simple solution (a "first-look").

Still the method has important flaws and it would be best to avoid it.

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 $\Delta \log L$)

Regularized unfolding (2)

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

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

where λ is a Lagrange multiplier, $\partial\varphi / \partial\lambda = 0 \rightarrow \sum_{i=1}^{N} \nu_i = n_{\rm tot}$.

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

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

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

a prescription for setting α.

Tikhonov regularization

Take measure of smoothness = mean square of kth derivative,

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

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

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

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

→ setting derivatives of φ equal to zero gives linear equations.

Solution using Singular Value Decomposition (SVD).

SVD implementation of Tikhonov unfolding

A. Hoecker, V. Kartvelishvili, NIM A372 (1996) 469; (TSVDUnfold in ROOT).

Minimizes
$$\chi^2(\boldsymbol{\mu}) + \tau \sum_i \left[(\mu_{i+1} - \mu_i) - (\mu_i - \mu_{i-1})^2 \right]$$

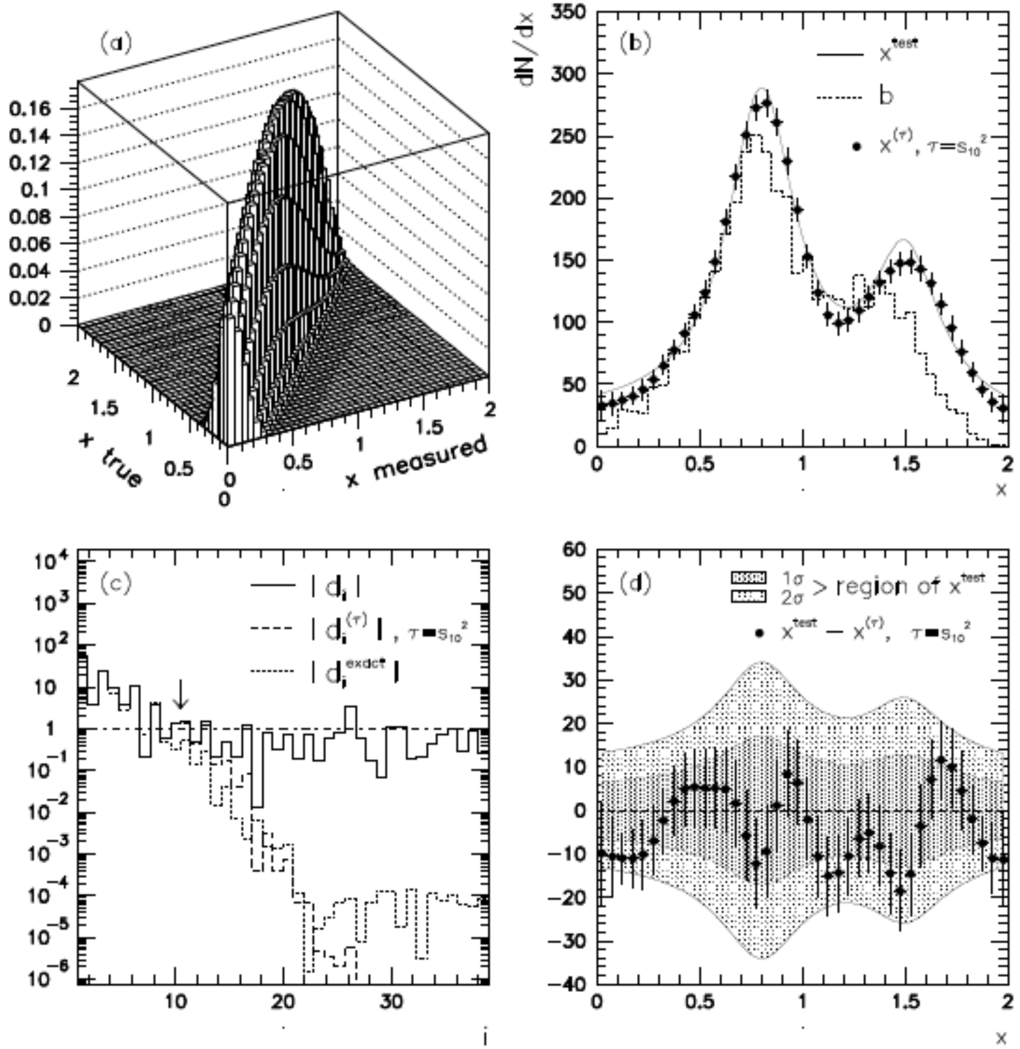
Numerical implementation using Singular Value Decomposition.

Recommendations for setting regularization parameter τ:

Transform variables so errors ~ Gauss(0,1);
number of transformed values significantly different from zero gives prescription for τ;
or base choice of τ on unfolding of test distributions.

SVD example

A. Höcker, V. Kartvelishvili, NIM A**372** (1996) 469.



Regularization function based on entropy

Shannon entropy of a set of probabilities is

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

All p_i equal → maximum entropy (maximum smoothness)

One $p_i = 1$, all others $= 0$ → minimum entropy

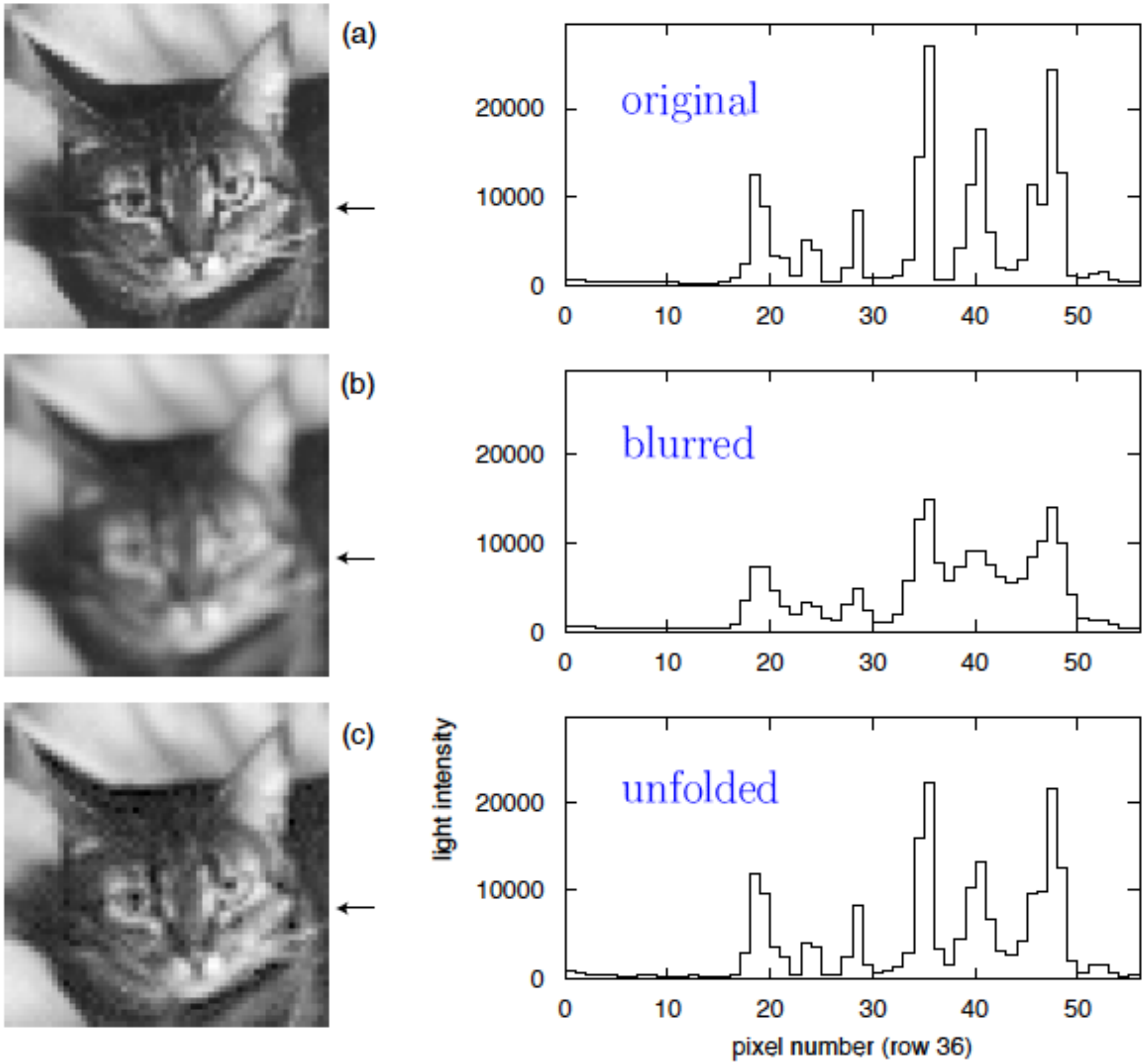
Use entropy as regularization function,

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

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

Can have Bayesian motivation: $S(\vec{\mu}) \longrightarrow$ prior pdf for $\vec{\mu}$

Example of entropy-based unfolding



Estimating bias and variance

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

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

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

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

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

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

Choosing the regularization parameter

$\alpha = 0 \rightarrow \hat{\vec{\mu}}$ 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,

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

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

G. Cowan, Statistical Data Analysis, OUP (1998) Ch. 11

Choosing the regularization parameter (2)

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

$$\Delta\chi^2 = 2\Delta \log L = N$$

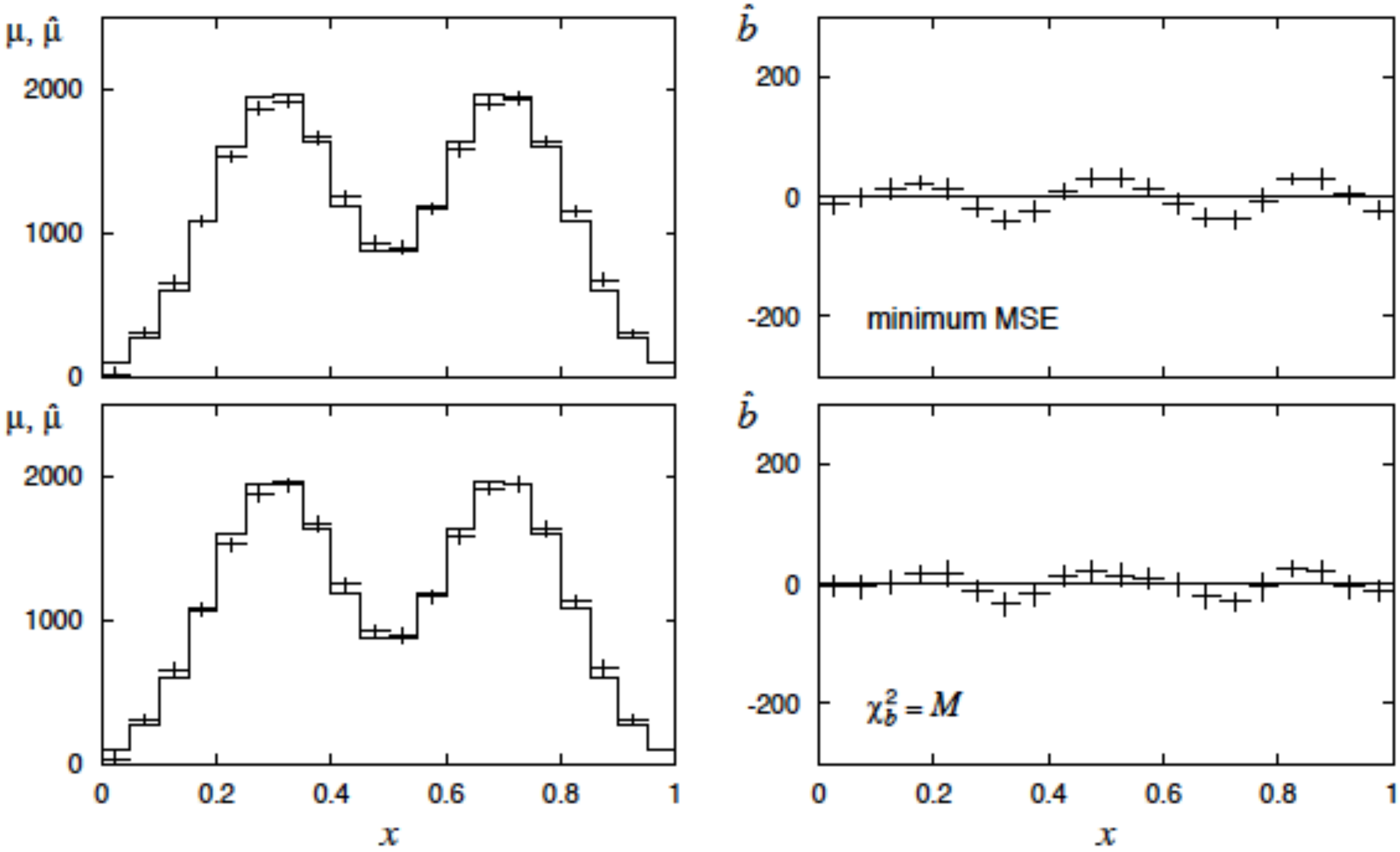
Or require that bias be consistent with zero to within its own error,

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

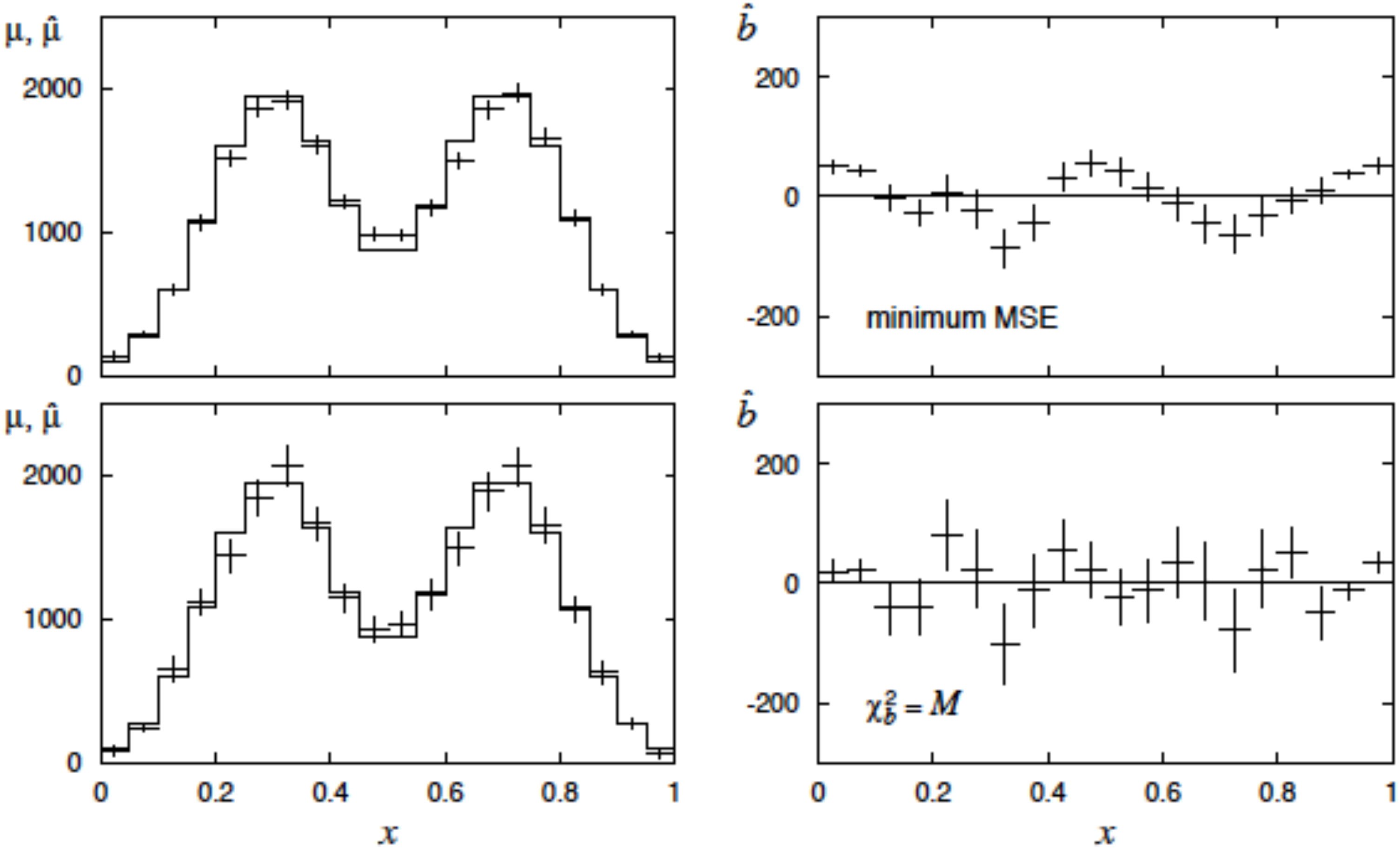
i.e. if bias significantly different from zero, we would subtract it;

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

Some examples with Tikhonov regularization



Some examples with entropy regularization



Stat. and sys. errors of unfolded solution

In general the statistical covariance matrix of the unfolded estimators is not diagonal; need to report full

$$U_{ij} = \mathrm{cov}[\hat{\mu}_i, \hat{\mu}_j]$$

But unfolding necessarily introduces biases as well, corresponding to a systematic uncertainty (also correlated between bins).

This is more difficult to estimate. Suppose, nevertheless, we manage to report both U_{stat} and U_{sys}.

To test a new theory depending on parameters $\boldsymbol{\theta}$, use e.g.

$$\chi^2(\boldsymbol{\theta}) = (\boldsymbol{\mu}(\boldsymbol{\theta}) - \hat{\boldsymbol{\mu}})^T (U_{\mathrm{stat}} + U_{\mathrm{sys}})^{-1} (\boldsymbol{\mu}(\boldsymbol{\theta}) - \hat{\boldsymbol{\mu}})$$

Mixes frequentist and Bayesian elements; interpretation of result can be problematic, especially if U_{sys} itself has large uncertainty.

Folding

Suppose a theory predicts *f*(*y*) → $\boldsymbol{\mu}$ (may depend on parameters $\boldsymbol{\theta}$).

Given the response matrix *R* and expected background $\boldsymbol{\beta}$, this predicts the expected numbers of observed events:

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

From this we can get the likelihood, e.g., for Poisson data,

$$L(\mathbf{n}|\boldsymbol{\nu}) = \prod_{i=1}^{N} \frac{\nu_i^{n_i}}{n_i!} e^{-\nu_i}$$

And using this we can fit parameters and/or test, e.g., using the likelihood ratio statistic

$$q = -2 \ln \frac{L(\mathbf{n}|\boldsymbol{\nu})}{L(\mathbf{n}|\hat{\boldsymbol{\nu}})} \sim \chi^2_N$$

Versus unfolding

If we have an unfolded spectrum and full statistical and systematic covariance matrices, to compare this to a model $\boldsymbol{\mu}$ compute likelihood

$$L(\hat{\boldsymbol{\mu}}|\boldsymbol{\mu}) \sim e^{-\chi^2/2}$$

where

$$\chi^2 = (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})^T (U_{\rm stat} + U_{\rm sys})^{-1} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})$$

Complications because one needs estimate of systematic bias $U_{\rm sys}$.

If we find a gain in sensitivity from the test using the unfolded distribution, e.g., through a decrease in statistical errors, then we are exploiting information inserted via the regularization (e.g., imposed smoothness).

ML solution again

From the standpoint of testing a theory or estimating its parameters, the ML solution, despite catastrophically large errors, is equivalent to using the uncorrected data (same information content).

There is no bias (at least from unfolding), so use

$$\chi^2(\boldsymbol{\theta}) = (\boldsymbol{\mu}(\boldsymbol{\theta}) - \hat{\boldsymbol{\mu}}_{\mathrm{ML}})^T U_{\mathrm{stat}}^{-1} (\boldsymbol{\mu}(\boldsymbol{\theta}) - \hat{\boldsymbol{\mu}}_{\mathrm{ML}})$$

The estimators of $\boldsymbol{\theta}$ should have close to optimal properties: zero bias, minimum variance.

The corresponding estimators from any unfolded solution cannot in general match this.

Crucial point is to use full covariance, not just diagonal errors.

Unfolding discussion

Unfolding can be a minefield and is not necessary if goal is to compare measured distribution with a model prediction.

Even comparison of uncorrected distribution with *future* theories not a problem, as long as it is reported together with the expected background and response matrix.

In practice complications because these ingredients have uncertainties, and they must be reported as well.

Unfolding useful for getting an actual estimate of the distribution we think we've measured; can e.g. compare ATLAS/CMS.

Model test using unfolded distribution should take account of the (correlated) bias introduced by the unfolding procedure.

Finally...

Estimation of parameters is usually the “easy” part of statistics:

Frequentist: maximize the likelihood.

Bayesian: find posterior pdf and summarize (e.g. mode).

Standard tools for quantifying precision of estimates:
Variance of estimators, confidence intervals,...

But there are many potential stumbling blocks:

bias versus variance trade-off (how many parameters to fit?);

goodness of fit (usually only for LS or binned data);

choice of prior for Bayesian approach;

unexpected behaviour in LS averages with correlations,...

Extra slides

Variance of estimators from information inequality

The information inequality (RCF) sets a lower bound on the variance of any estimator (not only ML):

$$V[\hat{\theta}] \geq \left(1 + \frac{\partial b}{\partial \theta}\right)^2 \Big/ E\left[-\frac{\partial^2 \ln L}{\partial \theta^2}\right]$$

Minimum Variance Bound (MVB)

$(b = E[\hat{\theta}] - \theta)$

Often the bias *b* is small, and equality either holds exactly or is a good approximation (e.g. large data sample limit). Then,

$$V[\hat{\theta}] \approx -1 \Big/ E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right]$$

Estimate this using the 2nd derivative of ln *L* at its maximum:

$$\widehat{V}[\hat{\theta}] = -\left(\frac{\partial^2 \ln L}{\partial \theta^2}\right)^{-1}\Bigg|_{\theta=\hat{\theta}}$$

Information inequality for *n* parameters

Suppose we have estimated *n* parameters $\vec{\theta} = (\theta_1, \ldots, \theta_n)$.

The (inverse) minimum variance bound is given by the Fisher information matrix:

$$I_{ij} = -E\left[\frac{\partial^2 \ln L}{\partial\theta_i \, \partial\theta_j}\right] = -\int P(\mathbf{x}|\boldsymbol{\theta}) \frac{\partial^2 \ln P(\mathbf{x}|\boldsymbol{\theta})}{\partial\theta_i \, \partial\theta_j} \, d\mathbf{x}$$

The information inequality then states that $V - I^{-1}$ is a positive semi-definite matrix, where $V_{ij} = \mathrm{cov}[\hat{\theta}_i, \hat{\theta}_j]$. Therefore

$$V[\hat{\theta}_i] \geq (I^{-1})_{ii}$$

Often use I^{-1} as an approximation for covariance matrix, estimate using e.g. matrix of 2nd derivatives at maximum of *L*.