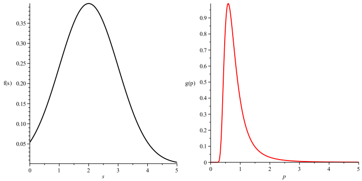
Parameter Estimation 4

A practical summary



Christoph Rosemann

DESY

20. March 2013

Parameter estimation

Common task

- Determine from measurements with uncertainties the best values of (physical) parameters
- Estimation is a mathematical procedure (!)
- Any parameter makes sense **only** within a model
- The model is encoded in the pdf of the parameters
- Wrong models deliver wrong answers!
- Uncertainties must be known: Variances and Covariances
- Distinguish between:
  - Statistical uncertainties
  - Systematic uncertainties

Choice of parametrization is crucial

Track fitting example

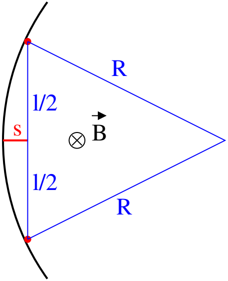
One of the most important properties in track parameter estimation:

(Transverse) Momentum p

Measurement principle:

- Momentum not a direct observable
- Usually space points with Gaussian pdf are measured
- Magnetic field B needed, Lorentz force puts charged particle on circular track:
  $F_z = F_L \Longrightarrow p = qBR$
- The radius R is also not directly measured
- Why not use radius?

Why isn't R Gaussian?



Measurement principle

- The deviation from a straight line is measured
  the sagitta s: $s \approx \frac{l^2}{8R}$ (for $l >> s$)
- Since the points are distributed Gaussian, the sagitta is as well
- With $p = qBR$ follows $p = \frac{qBl^2}{8s} = K\frac{1}{s}$

Transformation

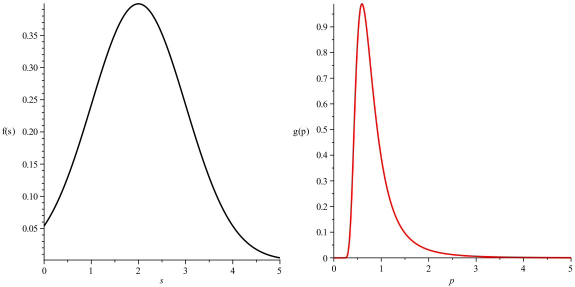
Assume sagitta Gaussian distributed:

$$f(s) = \frac{1}{\sqrt{2\pi\sigma_s^2}} \exp\left(-\frac{(s-\mu_s)^2}{2\sigma_s^2}\right)$$

Error propagation:

$$\begin{aligned} \mu_p &\approx p(\mu_s) = \frac{K}{\mu_s} \\ \sigma_p^2 &\approx \left(\frac{\mathrm{d}p}{\mathrm{d}s}\right)^2 \sigma_s^2 = \frac{p^4}{K^2}\sigma_s^2 \\ \Rightarrow g(p) &= \frac{f(s)}{\left|\frac{\mathrm{d}p}{\mathrm{d}s}\right|} = \frac{1}{\sqrt{2\pi\sigma_p^2}} \exp\left(-\frac{p^4(\frac{1}{s}-\frac{1}{\mu_s})^2}{2\sigma_p^2}\right) \end{aligned}$$

Take a look at the distributions



- Arbitrary choice for Sagitta distribution $\sigma_s = 1$, $\mu_s = 2$
- Notice: values at zero are not unlikely (negative values yield wrong charge sign)
- Result in the momentum distribution is very asymmetric

Ergo: Use parameters that are distributed like Gaussian!

Parameter estimation

Fundamental properties of estimators

Estimators can be characterized as *good* or *bad*

The characterization classes are:

- Consistency: the true value and the estimated value are equivalent

$$\lim_{n\to\infty} \hat{a} = a$$

- Bias: the expectation value is equivalent to true value

$$\langle \hat{a} \rangle = a$$

- Efficiency: small variance

The inherent accuracy of an estimator is limited!

Consistency

- Parameters are estimated from limited samples
- Any sample exhibits statistical fluctuations
- For large samples, the effect of fluctuations lessens
- If the difference between the true value and the estimated value vanishes, the estimator is **consistent**

Formal definition

An estimator is consistent, if it tends to the true value as the number of data tends to infinity:

$$\lim_{n \to \infty} \hat{a} = a$$

Bias

- For finite amounts of data the estimated parameter is unlikely to have the true value
- A good estimator has the equal chances of over- and underestimation of the true value
- Such an estimator is unbiased
- This can be expressed in terms of the expectation value of the estimator

Formal definition

An estimator is unbiased, if its expectation value is the same as the true value:

$$\langle \hat{a} \rangle = a$$

Efficiency

- The estimated value depends on the given data sample
- The fluctuations of the sample influence the estimator
- An efficient estimator exhibits a small fluctuation or spread
- The spread is measured in terms of the variance of the estimator

Formal definition

An estimator is efficient if its variance is small.

Minimum Variance Bound

(Without proof) There is a lower bound on the variance of an estimator!

- There are different names for this:
  Cramér-Rao bound (or inequality), Fréchet inequality, MVB, CRLB
- It uses the (in the simple/unbiased form) the Likelihood function $\mathcal{L}$:

$$\sigma_{\hat{a}}^2 \leq \frac{1}{\langle (d\mathcal{L}/da)^2 \rangle}$$

- An estimator is efficient, if its variance is equal to the MVB

Characterization of Maximum Likelihood

Most important parameter estimation method

- Maximum Likelihood estimators are (usually) consistent
- Maximum Likelihood are biased (!) for small N
  for large N it becomes unbiased
- It is usually the optimal estimation in terms of the Minimum Variance Bound

Warning

- Maximum Likelihood is (usually) consistent, but biased!
- Maximum Likelihood estimators invariant under parameter transformations!:

$$\widehat{f(a)} = f(\hat{a}) \qquad e.g.: \widehat{\sigma^2} = (\hat{\sigma})^2$$

Bias example

Consider a symmetric pdf around a_0, let $\hat{a}$ be an unbiased estimator

Equal chances that $\hat{a}$ is either 10% too large or too small

- Equally possible:

$$\hat{a} = 1.1a_0 \qquad \hat{a} = 0.9a_0$$

- Now consider (non-linear) transformation $y : x \to x^2$, then

$$\hat{a}^2 = 1.21a_0^2 \qquad \hat{a}^2 = 0.81a_0^2$$

- Probability content doesn't change, equal chances that $\hat{a}^2$ is 21% larger or 19% smaller than a_0^2
- In short: the pdf becomes asymmetric and therefore biased

Relation between χ^2 and Likelihood

Likelihood definition

$$\ell(a) = -\ln \mathcal{L}(a) = -\sum_{i}^{n} \ln f(x_i; a)$$

Measurements with underlying Gaussian distribution

Estimate the mean value from:

$$f(x_i; a) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x_i - a)^2}{2\sigma_i^2}}$$

Combine both!

Relation between χ^2 and Likelihood

$$\ell(a) = -\sum_i^n \ln \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x_i-a)^2}{2\sigma_i^2}} = -\sum_i^n \ln \underbrace{\frac{1}{\sqrt{2\pi\sigma_i^2}}}_{= const} + \frac{1}{2} \underbrace{\sum_i^n \left(\frac{x_i - a}{\sigma_i}\right)^2}_{\equiv \chi^2}$$

Direct connection between Likelihood and χ^2 function

$$\ell(a) = const + \frac{1}{2}\chi^2(a)$$

Note the factor $\frac{1}{2}$

Explains the different units for uncertainty estimation:

$$\chi^2_{min} + 1 \qquad \ell_{min} + \frac{1}{2}$$

Non-linear least squares

What if $f(x;\vec{a})$ isn't linear in a_j?

Iterative solution is needed; start with first guess $\vec{a}_0$

- Use gradient:

$$grad_j(\vec{a}_0) = \left.\frac{\partial\chi^2}{\partial a_j}\right|_{\vec{a}} = \sum_i -\frac{2}{\sigma_i^2}[y_i - f(x_i;\vec{a}_0)]\frac{\partial f(x_i;\vec{a}_0)}{\partial a_j}$$

- Goal: find $\delta\vec{a}$ with:

$$grad_j(\vec{a}_o + \delta\vec{a}) = \left.\frac{\partial\chi^2}{\partial a_j}\right|_{\vec{a}+\delta\vec{a}} = 0 \quad \forall j$$

- Expand the gradient in a Taylor series, omitting higher order terms:

$$grad_j(\vec{a}_0+\delta\vec{a}) \approx grad_j(\vec{a}_0)+\sum_s \frac{\partial grad_j}{\partial a_s}\delta a_s = grad_j(\vec{a}_0)+\sum_s \frac{\partial\chi^2}{\partial a_j \partial a_s}\delta a_s$$

Non-linear least squares cont'd

- This yields an expression to find $\delta\vec{a}$ from iteration, with the matrix equation

$$\delta\vec{a} = -G^{-1}\vec{g}$$

- $\vec{g}$ is the vector of all gradients and the matrix elements are defined by

$$G_{js} = \frac{\partial\chi^2}{\partial a_j \partial a_s}$$

Solving iterative matrix equations

- Leads to another topic: numerical recipes
- In principle:
  - Start with good guess for $\vec{a}_0$ ($\sim$ 90% of all trouble)
  - Construct matrix and invert, if gradient is small enough: solution found
- Can be very tricky business
- Recommendation for all practical purposes: use library

χ^2 or Maximum Likelihood?

Maximum Likelihood pros

- Simple procedure
- For large N optimal
- Easy to use in many dimensions
- Possible to avoid information loss through binning extremely important with small data sets

Maximum Likelihood cons

- Biased (for small N)
- pdf has to be known
- No way to determine the estimation quality

χ^2 pros

- Allows simple and powerful consistency check (if variables are Gaussian distributed)
- Simple way to encode correlations, including systematic errors
- For linear models single step solution

Both: Equivalent method to ML if variables are Gaussian distributed

χ^2 cons

- Needs binning to fit a distribution to data

Many more topics

- Outlier rejection/down weighting with M-estimators
- Constraint fits; e.g. kinematic fits
- Numeric Integration
- Numerical minimization techniques
- Generating random numbers according to arbitrary distributions

Summary

- Parameter Estimation is a well defined mathematical procedure
- The results can still be ill-defined: crap in, crap out
- The two main methods were presented, (usually) clear from context what to use
- The way you formulate the problem influences the quality (and reliability) of the solution