Topical Lectures

Fitting, Tracking and Vertexing

NIKHEF, Feb. 28-29 2007

Wouter Hulsbergen (CERN)

wouter.hulsbergen@cern.ch

program

- 6 x 45 minutes, today and tomorrow
  - 1^{st} hour: probability, statistics, least squares estimator
  - 2^{nd} hour: non-linear problems, a straight line fit, the progressive fit
  - 3^{rd} hour: interaction of particles with matter, tracking detectors
  - 4^{th} hour: track fitting
  - 5^{th} hour: track finding
  - 6^{th} hour: vertex and decay tree fitting

- slides available at http://www.cern.ch/whulsber/topicallectures

subset of recent NIKHEF theses

- van Eldik, The ATLAS muon spectrometer : calibration and pattern recognition (2007)
- Cornelissen, Track fitting in the ATLAS experiment (2006)
- Hommels, The tracker in the trigger of LHCb (2006)
- van Beuzekom, Identifying fast hadrons with silicon detectors (2006)
- Sokolov, Prototyping of Silicon Strip Detectors for the Inner Tracker of the ALICE Experiment (2006)
- van Tilburg, Track simulation and reconstruction in LHCb (2005)
- Heijboer, Track reconstruction and point source searches with ANTARES (2004)
- Hierck, Optimisation of the LHCb detector (2003)
- Vos, The ATLAS inner tracker and the detection of light supersymmetric Higgs bosons (2003)
- Peeters, The ATLAS semiconductor tracker endcap (2003)
- Visser, Muon tracks through ATLAS (2003)
- Woudstra, Precision of the ATLAS muon spectrometer (2002)
- van der Eijk, Track reconstruction in the LHCb experiment (2002)
- Hulsbergen, Track reconstruction and di-lepton production in Hera-B (2002)
- ...

Part 1

probability

least squares estimator

probability density function

- from wikipedia (stripped from the mathematical language I cannot understand)

  - the *probability density function* for a random variable X is the non-negative function $\mathcal{P} : R \longrightarrow R$ such that the probability that $X \in [a, b]$ is

$$\int_a^b \mathcal{P}(\xi)\mathrm{d}\xi$$

  - alternative formulation: if Δt is an infinitely small number, the probability that X is included within the interval $(t, t + \Delta t)$ is equal to $\mathcal{P}(t)\,\Delta t$, or:

$$\Pr(t < X < t + dt) = \mathcal{P}(t)\,\Delta t$$

- notes
  - the value of P(x) is *not* the *probability* for x; it is a *density*
  - since integrals over P represents a probability, P(x) is normalized to unity

expectation value

- expectation value for a function g(x)

$$E\left[g(x)\right]_{\mathcal{P}} = \int_{-\infty}^{\infty} g(x)\mathcal{P}(x)\mathrm{d}x$$

- less common, shorter notation $E\left[g(x)\right]_{\mathcal{P}} \equiv \langle g(x)\rangle_{\mathcal{P}}$

- some relevant properties

$$\langle\, g(x) \; + \; h(x)\,\rangle \; = \; \langle g(x)\rangle + \langle h(x)\rangle$$

$$\langle a\, g(x) \; + \; b\rangle \; = \; a\;\langle g(x)\rangle \; + \; b \qquad \text{for any } a, b \in R$$

mean, variance

- mean of P

$$\mu_x \equiv \langle x \rangle \equiv \int_{-\infty}^{\infty} x\mathcal{P}(x)\mathrm{d}x$$

- variance

$$\sigma_x^2 \equiv \mathrm{var}\,(x) \equiv \left\langle (x - \langle x \rangle)^2 \right\rangle = \left\langle x^2 \right\rangle - \langle x \rangle^2$$

- example, the gaussian distribution

$$\mathcal{P}(x)\ \mathrm{d}x = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \mathrm{d}x$$

$$\langle x \rangle = \mu \qquad\qquad \mathrm{var}\,(x) = \sigma^2$$

multi-dimensional pdfs

- two-dimensional pdf for random variables (RVs) X and Y

$$\mathcal{P}(t,s)\,\mathrm{d}t\,\mathrm{d}s \;=\; \Pr(\; t < X < t + \mathrm{d}t \;\wedge\; s < Y < s + \mathrm{d}s \;)$$

  - can be generalized to any number of RVs

- covariance

$$V_{xy} \;\equiv\; \mathrm{cov}(x,y) \;\equiv\; \langle (x - \langle x \rangle)\,(y - \langle y \rangle) \rangle$$

- correlation coefficient $\rho_{x,y} \;\equiv\; \dfrac{\mathrm{cov}(x,y)}{\sqrt{\mathrm{var}\,(x)\;\mathrm{var}\,(y)}}$

- note: $\mathrm{cov}(x,y) \;=\; \mathrm{cov}(y,x)$

$$\mathrm{var}\,(x) \;=\; \mathrm{cov}(x,x)$$

$$-1 \le \rho_{x,y} \le 1$$

covariance matrix

- covariance conveniently organized in matrix

$$V(x,y,z,\cdots) = \begin{pmatrix} V_{xx} & V_{xy} & V_{xz} & \cdots \\ V_{yx} & V_{yy} & V_{yz} & \cdots \\ V_{zx} & V_{zy} & V_{zz} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

- matrix V is symmetric and positive-definite (det(V)>=0)
- example: gaussian (normal) distribution in N dimensions

$$\mathcal{P}(x_1,\ldots,x_N)\,\mathrm{d}x_1\cdots\mathrm{d}x_N \;\propto\; \exp\left[\frac{1}{2}x^T\,V^{-1}\,x\right]\;\mathrm{d}x_1\cdots\mathrm{d}x_N$$

- where $x = (x_1,\cdots,x_N)$ and V as above

linear transformations

- if ***F*** a linear transformation such that

$$y = F\, x \qquad \text{for vectors } x \in R^n, y \in R^m \text{ and matrix } F \in R^m \times R^n$$

then

$$\langle y \rangle = F \;\langle x \rangle \qquad \text{var}(y) = F \;\text{var}(x)\; F^T$$

- this is the familiar 'error propagation'
- if the transformation is not linear, e.g. $y = f(x)$

the expressions above hold ***to first order*** *in* ***x*** with jacobian

$$F_{ij} = \frac{\partial y_i}{\partial x_j}$$

- this is just an approximation: if you want the true variance of y, you need to calculate ***var(f(x))***

linear transformation of Gaussian distribution

- example of linear transformation: for Gaussian ***P(x)***

$$\mathcal{P}(x_1, \ldots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n \;\propto\; \exp\left[\frac{1}{2}\, x^T\, V_x^{-1}\, x\right]\; \mathrm{d}x_1 \cdots \mathrm{d}x_n$$

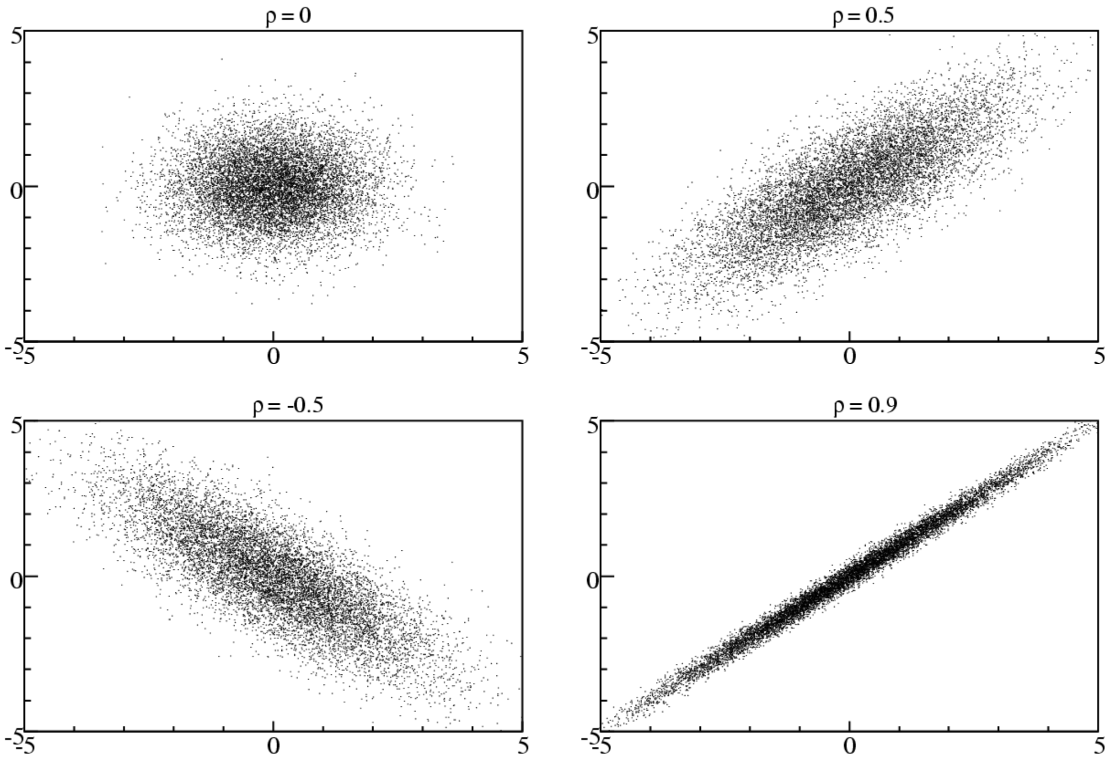
- if ***y = Fx***, then ***P(y)*** is also Gaussian

$$\mathcal{P}(y_1, \ldots, y_m)\, \mathrm{d}y_1 \cdots \mathrm{d}y_m \;\propto\; \exp\left[\frac{1}{2}\, y^T\, V_y^{-1}\, y\right]\; \mathrm{d}y_1 \cdots \mathrm{d}y_m$$

with $V_y \;=\; F\, V_x\, F^T$

- in other words
  - linear transformation of Gaussian PDF is still Gaussian PDF
  - if X is sum of Gaussian Rvs, X is itself a Gaussian RV

- example: x and y gaussian distributed with unit variance



- correlation tells about the sign of the *direction* of the slope and how *squeezed* the distribution is

- sizes of half the major and minor axis of the 'ellipse' correspond to eigenvalues of covariance matrix V

central limit theorem

- central limit theorem

Consider sum of N random variables

$$S = x_1 + x_2 + \cdots + x_N$$

If x_i independent and distributed according to a pdf $\mathcal{P}(x)$ with finite mean μ_x and variance V_x, then

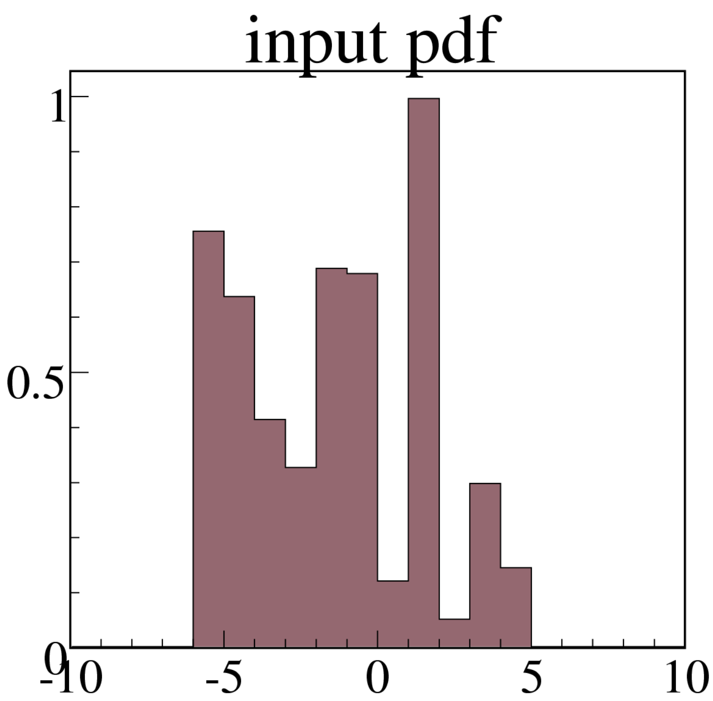
$$\mu_S = N\mu_x \qquad V_S = NV_x$$

In the limit for large N the distribution for S approaches a normal distribution with mean μ_S and variance V_S.

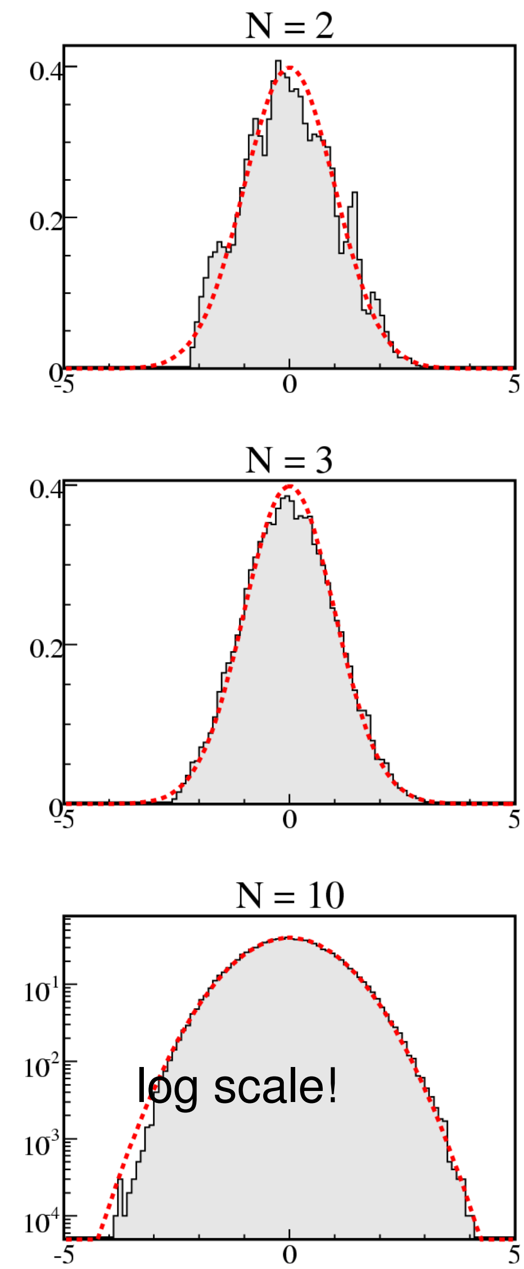
- why is this important for us?
  - if error on measurement is sum of many small contributions, it is approximately gaussian distributed
  - if we extract <N parameters from N measurements, their errors are usually more Gaussian then those on original measurements

CLT in action

starting from an arbitrary PDF



generated distribution of $(S - \mu_S)/\sqrt{V_S}$



note: used finite number of samplings
(10000). in reality distributions
even more gaussian!

estimators

- suppose we have
  - a data set $\{x_i\}$
  - a model with unknown parameters α
- a *statistic* is any function of the data that does not depend on α
- an *estimator* for α is a statistic whose value estimates α
- some important properties of estimators
  - **consistency**: estimator is consistent if it approaches true value with more data
  - **bias**: difference between expectation value of estimate and α
  - **efficiency**: ratio between variance of estimate and best possible variance of any estimate for α

method of maximum likelihood

- given
  - set of independent measurements $\{x_i\}$
  - 'model' which gives the pdf for each x_i: $\mathcal{P}_i(x_i;\alpha)\mathrm{d}x_i$
- define the **likelihood function**

$$\mathcal{L}(\alpha;x) = \prod_i \mathcal{P}_i(x_i;\alpha)$$

- maximum likelihood estimate of α is the value α_{ML} for which $\mathscr{L}$ is maximum
- it can be proven that if an efficient estimator exists, then α_{ML} is efficient
  - that means that there exists no estimator with smaller variance
  - (that does not mean that there exists no estimator with smaller bias)

method of maximum likelihood

- in applications we usually deal with the **log** of the likelihood function, because it is easier to add than to multiply

$$\ln \mathcal{L}(\alpha; x) \;=\; \sum_i \ln \mathcal{P}_i(x_i; \alpha)$$

- covariance matrix may be estimated from

$$V \;=\; \left[E\left(-\frac{\partial^2 \ln \mathcal{L}}{\partial \alpha^2} \right) \right]^{-1}$$

  - don't need to believe this now: will derive later for gaussian case

- most commonly, solution found with generic minimization algorithm, like MINUIT

- NOT HERE: we do not use MINUIT in track and vertex fitting

method of least squares

- consider N independent measurements with Gaussian PDF

measurement ***i***

measurement model

$$\mathcal{P}_i\left(m_i\,;\,x\right) = \frac{1}{\sqrt{2\pi}} \exp\left[\frac{1}{2}\left(\frac{m_i - h_i(x)}{\sigma_i}\right)^2\right]$$

model parameters

uncertainty in measurement ***i***

- note: change of variable names
  - till now mostly followed PDG
  - from now on use notations closer to tracking literature

method of least squares

- consider N independent measurements with Gaussian PDF

$$\mathcal{P}_i\,(m_i\,;\,x) = \frac{1}{\sqrt{2\pi}} \exp\left[\frac{1}{2}\left(\frac{m_i - h_i(x)}{\sigma_i}\right)^2\right]$$

- define the **chi-square**

$$\chi^2 \equiv \sum_i \left(\frac{m_i - h_i(x)}{\sigma_i}\right)^2 = -2\ln\mathcal{L} + \text{constant}$$

- the value x-hat for which the chi-square is minimum is called the **least squares estimator (LSE)**
- as you can see above, if the measurements are distributed normally around their true values, the LSE is the maximum likelihood estimator

method of least squares

- so, minimizing the chi-square is well motivated for 'Gaussian' errors

- there is another motivation: the **Gauss-Markov theorem** states that for a **linear** model, the LSE is **efficient** for (almost) any error distribution
  - there is no *linear* estimator with smaller variance

- because it is a good illustration of the concepts we have just introduced, we now prove the Gauss-Markov theorem
  - first we rewrite the chi-square in matrix notation
  - then we linearize it, extract the LSE and its variance
  - finally, we prove the theorem

chi-square in matrix notation

- rewrite chi-square using covariance matrix for measurements

vector of measurements

measurement model

(diagonal) measurement covariance matrix

$$\chi^2 = \sum_i \left(\frac{m_i - h_i(x)}{\sigma_i} \right)^2 = (m - h(x))^T \, V^{-1} \, (m - h(x))$$

vector of 'residuals'

- condition that chi-square is minimum, can now be written as

$$0 = \frac{\mathrm{d}\chi^2}{\mathrm{d}x} = -2 \frac{\mathrm{d}h(x)}{\mathrm{d}x}^T V^{-1} \, (m - h(x))$$

derivative matrix

- for N measurements and M parameters, derivative is NxM matrix

LSE for a linear model

- in many fit applications derivative of h(x) varies slowly with respect to measurement errors
- therefore, consider linear measurement model

$$h(x) = h_0 + H\,x$$

where the derivative matrix $H \equiv \frac{\mathrm{d}h(x)}{\mathrm{d}x}$ is constant

- condition that chi-square derivative vanishes, becomes

$$\frac{\mathrm{d}\chi^2}{\mathrm{d}x} = -2\,H^T V^{-1}\,(m - h_0 - Hx) = 0$$

which has a solution

$$\hat{x} = \left(H^T V^{-1} H\right)^{-1}\,H^T V^{-1}\,(m - h_0)$$

- this is the LSE for linear models. it is called a **linear estimator**, because it is a linear function of the measurements

bias and variance of the LSE

- provided that the measurements are unbiased and have variance V

$$\langle m \rangle = m^{\text{true}} \equiv h_0 + H x^{\text{true}} \qquad \text{var}(m) \equiv V$$

- we find that the bias of the LSE is zero

$$\begin{aligned} \langle \hat{x} - x^{\text{true}} \rangle &= \left(H^T V^{-1} H\right)^{-1} H^T V^{-1} (\langle m \rangle - h_0 - H x^{\text{true}}) \\ &= 0 \end{aligned}$$

- and that its variance is

$$\text{var}(\hat{x}) = \text{var}\left(\left(H^T V^{-1} H\right)^{-1} H^T V^{-1} (m - h_0)\right)$$

drop constants

$$= \text{var}\left(\left(H^T V^{-1} H\right)^{-1} H^T V^{-1} m\right)$$

$\text{var}(Ax) = A\ \text{var}(x)\ A^T$

$$= \left(H^T V^{-1} H\right)^{-1} H^T V^{-1}\ \text{var}(m)\ V^{-1} H \left(H^T V^{-1} H\right)^{-1}$$

var(m)=V

$$= \left(H^T V^{-1} H\right)^{-1}$$

other linear estimators

- we now simplify things a bit, without loss of generality
  - choose $h(x_0)=0$ by absorbing constants in measurements
  - choose V = 1 by scaling measurements to have unit variance
- the LSE then becomes

$$\hat{x} = \left(H^T H\right)^{-1} H^T m \qquad \mathrm{var}\left(x\right) = \left(H^T H\right)^{-1}$$

- now take an arbitrary other linear estimator

$$\hat{x}' = Am$$

- again, without loss of generality rewrite it as

$$\hat{x}' = \left(\left(H^T H\right)^{-1} H^T + B\right) m$$

Gauss-Markov theorem

- for the bias and variance of A we obtain

$$\langle \hat{x}' - x^{\text{true}} \rangle = BHx^{\text{true}}$$

$$\text{var}(\hat{x}') = (H^T H)^{-1} + BH(H^T H)^{-1} + (H^T H)^{-1} H^T B^T + BB^T$$

- so, if we require the estimator to be unbiased for any true x, then BH=0 and therefore

$$\text{var}(\hat{x}') = (H^T H)^{-1} + BB^T$$

variance of LSE

pos-def matrix

- this completes our 'proof' of the Gauss-Markov theorem: if the data are unbiased and uncorrelated and the model is linear, then the LSE is unbiased and there is no linear unbiased estimator with smaller variance