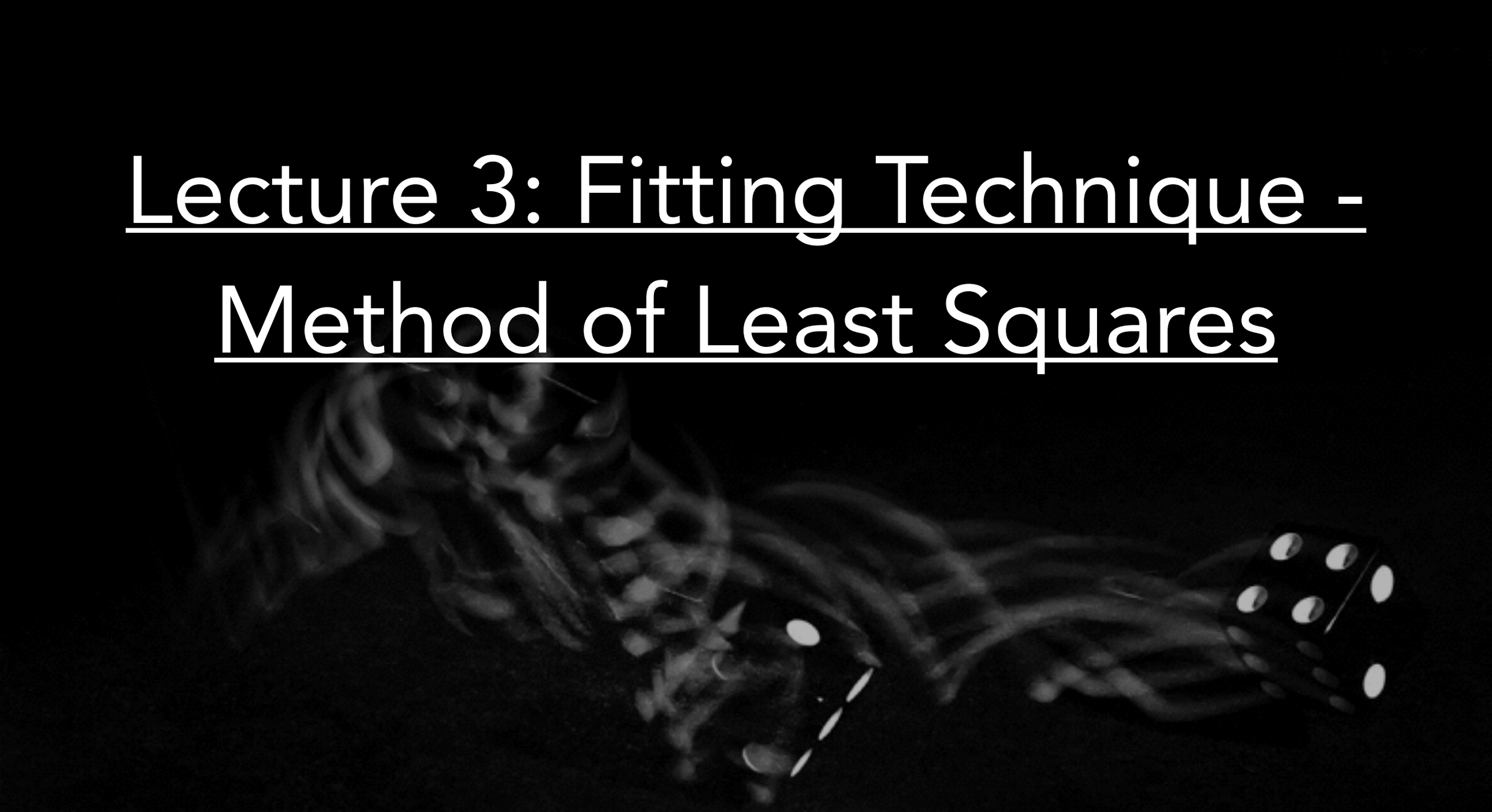
Lecture 3: Fitting Technique - Method of Least Squares



D. Jason Koskinen
koskinen@nbi.ku.dk

Advanced Methods in Applied Statistics
Feb - Apr 2016

Photo by Howard Jackman

Method of Least Squares

- In today's lecture:
  - Introduction
  - Linear Least Squares Fit
  - Least Squares method estimate of variance
  - Non-linear Least Squares
  - Least Squares as goodness-of-fit statistic
  - Least Squares on binned data (maybe)
- A lot, lot more math and analytic coverage than usual in today's slides. Should be used as reference material, but focus on **using** your least squares minimization routines.

Material from D. Grant derived entirely from lectures by A. Bellerive, G. Cowan, and D. Karlen

Method of Least Squares

- Introduction
  - Most frequently used fitting method, but has no general optimal properties that would make that the case.
  - When the parameter dependence is linear, the method produces unbiased estimators of minimum variance.
  - The method is applied as follows:
    - for observation points, x, experimental values are measured, y. The true functional form is defined by L parameters:

$$f_i = f_i(\theta_1, ..., \theta_L)$$

    - To find parameter estimates, θ, we minimize:

$$X^2 = \sum_i w_i(yi - f_i)^2$$

weight expressing accuracy of y

Method of Least Squares

- Introduction
  - The method is applied as follows (cont.):
    - In the case of constant accuracy, all w = 1.
    - If the accuracy for y is given by σ_i then $w_i = 1/\sigma_i^2$
    - If the y values represent a poisson distributed random number:

$$w_i = 1/f_i$$

    - If the observations are correlated, then the minimization function becomes:

$$X^2 = \sum_{i=1}^{N}\sum_{j=1}^{N}(y_i - f_i)V_{ij}^{-1}(y_j - f_j)$$

V_{ij} is the covariance matrix

    - where the values for independent variable(s) (generally x) are assumed to have be known precisely, i.e. no uncertainties.

Method of Least Squares

- Introduction
  - The method is applied as follows (cont.):
    - In many cases the measured value, y, can be regarded as a Gaussian random variable centered about the true value, as expected from the Central Limit Theorem as long as the total error is the sum of a large number of small contributions.
    - The deviation from the true value has the form:

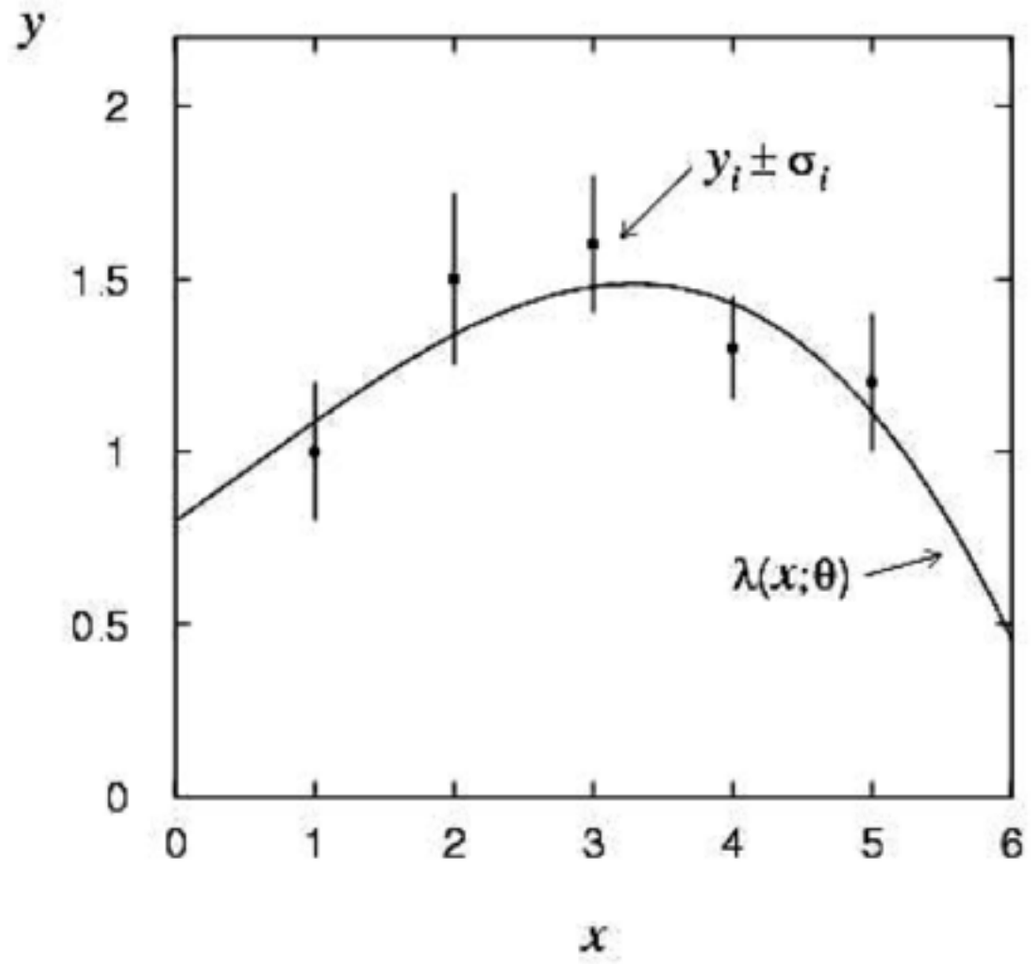
$$\epsilon_j = y_i - \lambda_i \qquad E[\epsilon_j] = 0 \qquad E[\epsilon_j^2] = V[\epsilon_j] = \sigma^2$$

    - For a set of N independent Gaussian random variables, y_i, of unknown mean, λ_i, and different but known variance σ_i^2, then the joint pdf can be written:

$$g(\vec{y}; \vec{\lambda}, \vec{\sigma}^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}(\frac{y_i - \lambda_i}{\sigma_i})^2}$$

Method of Least Squares

- Introduction
  - The method is applied as follows (cont.):
    - Again, the measurements are related to x, which is known without error. We can write the true value in terms of a function of x with unknown parameters θ: $\lambda = \lambda(x; \vec{\theta})$
  - The goal is to estimate these parameters with the least squares method, a simple evaluation of the goodness of fit of the hypothesized function above.



Method of Least Squares

- Introduction
  - The method is applied as follows (cont.):
    - The likelihood function is given by:

$$L(\vec{\theta}) = L(x;\vec{\theta}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}(\frac{y_i - \lambda_i(x;\vec{\theta})}{\sigma_i})^2}$$

- which corresponds to the log-likelihood function:

$$\ln L(\vec{\theta}) = -\frac{1}{2} n \ln 2\pi + \ln(\prod_{i=1}^{N} \sigma_i^{-1}) - \frac{1}{2}\sum_{i=1}^{N}(\frac{y_i - \lambda_i(x;\vec{\theta})}{\sigma_i})^2$$

$$\ln L(\vec{\theta}) = const - \frac{1}{2}\sum_{i=1}^{N}(\frac{y_i - \lambda(x_i;\vec{\theta})}{\sigma_i})^2$$

$$-2\ln L(\vec{\theta}) = \sum_{i=1}^{N}(\frac{y_i - \lambda(x_i;\vec{\theta})}{\sigma_i})^2 + const$$

Method of Least Squares

- Introduction
  - The method is applied as follows (cont.):
    - One may maximize lnL, or minimize:

$$\chi^2(\vec{\theta}) \equiv \sum_{i=1}^{N} \left(\frac{y_i - \lambda(x_i;\vec{\theta})}{\sigma_i}\right)^2$$

    - The errors on the estimated parameters are obtained by evaluating the corresponding one standard deviation departure from the least-squares estimate:

$$\chi^2(\vec{\theta}) = \chi^2(\tilde{\vec{\theta}}) + 1 \qquad \text{since} \qquad \chi^2(\vec{\theta}) = -2\ln L(\vec{\theta}) + const$$

    - Thus, if the measurements are Gaussian distributed, then the least square method is equivalent to the maximum likelihood method (covered on Thursday). Further, the observables will be linear functions of the parameters and follow the chi-square distribution.

Method of Least Squares

- Introduction
  - This is the foundation for the LS method and it is used to look at the difference between between measured and hypothesized values, weighted by the inverse of the variance.
  - The method is applicable even when the individual measurements, y, are not Gaussian, as long as they are independent.
  - Repeated measurements can be treated as the sum:

$$y_j = \lambda + \epsilon_j$$

  - where the quantity **λ** can be determined with the quadratic sum:

$$\sum_{j=1}^{N} \epsilon_j^2 = (\lambda - y_j)^2 = min$$

Method of Least Squares

- Linear LS Fit
  - If the observables are linear functions of the unknown parameters and the weights are independent of the parameters, then the LS method has an exact solution that can be written in a closed form.
  - Consider

$$\lambda(x;\vec{\theta}) = \sum_{j=1}^{N} a_j(x)\theta_j$$

  - the a(x) terms are any linearly independent function of x such that λ is linear in the parameters θ. The a(x) are generally not linear in x, but are linearly independent of each other.
  - In this case, an analytic solution for the estimators and their variances exists. The estimators will be unbiased from the MVB condition regardless of the number of measurements and the pdf of the individual measurements.

Slight Detour - Maximum Likelihood Estimator

- Maximum Likelihood Estimator Properties
  - Unbiased
    - For finite data sets, an estimator is unbiased if its expectation value is not systematically shifted from the true parameter value and is centered around this value for all sample sizes.
    - It is natural to use the spread (dispersion) in estimates as a measure of acceptability of an estimator. Note that although $\hat{\lambda}$ may be unbiased, $t(\hat{\lambda})$ may be biased.
  - Efficient
    - if a sufficient estimator exists then the Maximum Likelihood method produces it and will give the minimum attainable variance (the Minimum Variance Bound - MVB).

$$V(t) = \frac{(\frac{\partial \tau}{\partial \lambda} + \frac{\partial B}{\partial \lambda})^2}{E[-\frac{\partial^2 \ln L}{\partial \lambda^2}]} = \frac{(\frac{\partial \tau}{\partial \lambda} + \frac{\partial B}{\partial \lambda})}{A(\lambda)}$$

Rao-Cramer Inequality

t = test statistic/estimator
λ = parameter for estimator
τ = function
B = bias
L = likelihood function

Method of Least Squares

- Linear LS Fit
  - Note that although the estimators and variances can be found analytically when it is a linear function for λ in terms of the parameters, one may use numerical methods for the estimation of the parameters.
  - Analytically:

Skip

- We begin with:

$$\lambda(x;\vec{\theta}) = \sum_{j=1}^{N} a_j(x)\theta_j = \sum_{j=1}^{m} A_{ij}\theta_j$$

- For N independent measurements, y, with N>m and errors given by σ:

$$\chi^2 = (\vec{y} - \vec{\lambda})^T V_{ij}^{-1} (\vec{y} - \vec{\lambda})$$

$$= (\vec{y} - A\vec{\theta})^T V_{ij}^{-1} (\vec{y} - A\vec{\theta})$$

diagonal matrix with terms $1/\sigma_{ii}^2$

Method of Least Squares

- Linear LS Fit
  - To find minimum chi-square, or maximum lnL, we set the derivative with respect to the parameters equal to zero:

$$\nabla\chi^2 = -2(A^T V^{-1}\vec{y} - A^T V^{-1} A\vec{\theta}) = 0$$

$$\tilde{\theta} = (A^T V^{-1} A)^{-1} A^T V^{-1}\vec{y} = B\vec{y}$$

Skip

  - Applying the error propagation formula to find the covariance matrix of the estimators:

$$U_{ij} = cov[\tilde{\theta}_i, \tilde{\theta}_j]$$

$$U = BVB^T = (A^T V^{-1} A)^{-1}$$

$$U_{ij}^{-1} = \frac{1}{2}\left(\frac{\partial^2\chi^2}{\partial\theta_i\partial\theta_j}\right)_{\vec{\theta}=\tilde{\vec{\theta}}}$$

  - This term coincides with the MVB for the inverse covariance matrix when the measurements are normally distributed.

Method of Least Squares

- Linear LS Fit
  - Therefore:

$$\ln L = -\frac{\chi^2}{2}$$

$$\chi^2(\theta) = \chi^2(\tilde{\theta}) + \frac{1}{2}\sum_{i,j=1}^{m}\left(\frac{\partial^2\chi^2}{\partial\theta_i\partial\theta_j}\right)(\theta_i - \tilde{\theta}_i)(\theta_j - \tilde{\theta}_j)$$

Skip

- Using the MVB equation for the variance, the one standard deviation contour in the parameter space where the tangents are given by $\tilde{\theta}_i \pm \Delta\tilde{\theta}_i$ from the LS estimates is:

$$\chi^2(\theta) = \chi^2(\tilde{\theta}) + 1 = \chi^2_{min} + 1$$

Method of Least Squares

- LS Fit for a polynomial
  - As a hypothesis for $\lambda(x;\vec{\theta})$, you might want to use a polynomial of order m, in the case of m+1 parameters, e.g.

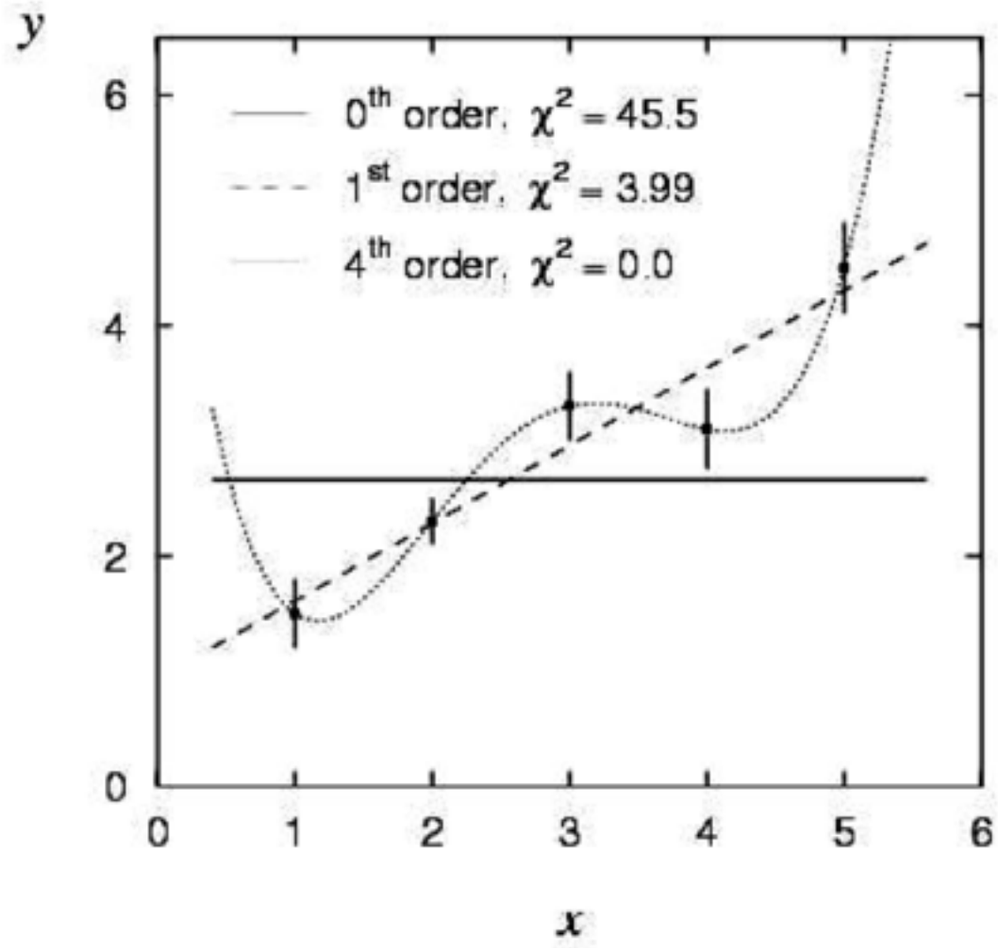
$$\lambda(x;\vec{\theta}) = \sum_{j=0}^{m} x^j \theta_j$$

  - This is a special case of the linear LS method with linearly independent weights:

$$a_j(x) = x^j$$

  - Thus, just as before,

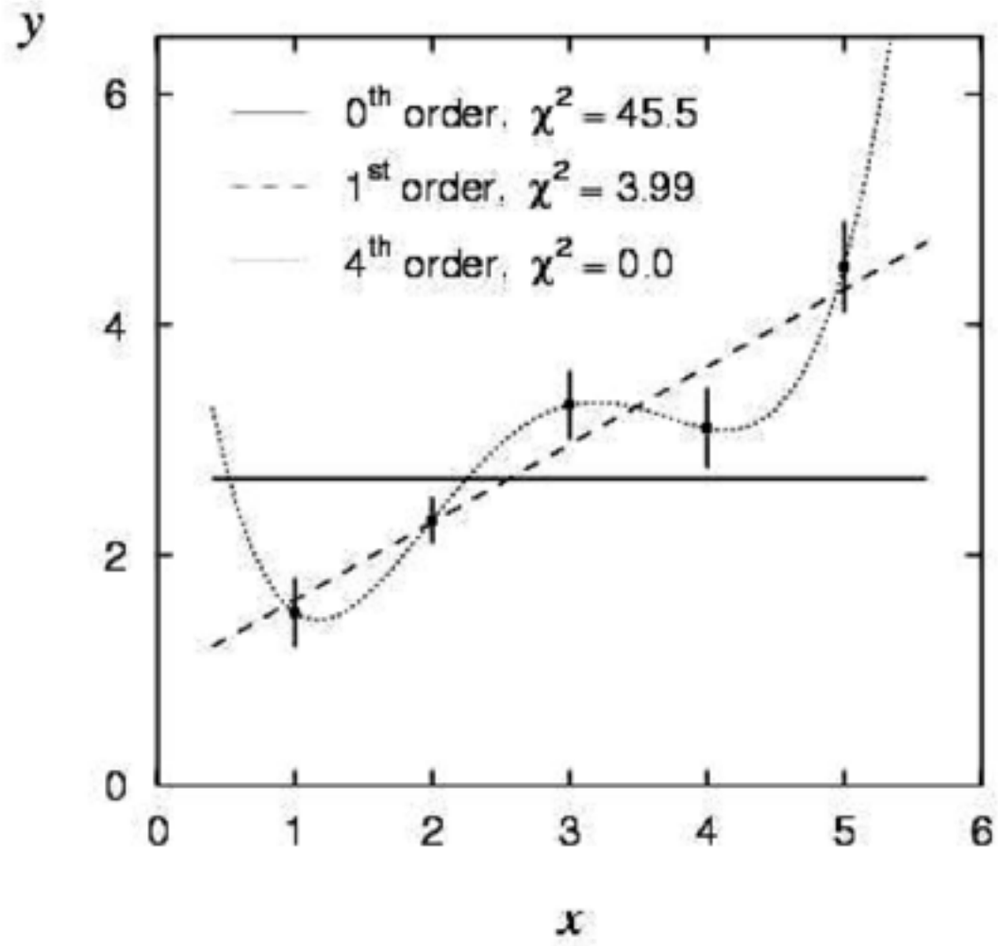
$$\chi^2 = (\vec{y} - A\vec{\theta})^T V_{ij}^{-1} (\vec{y} - A\vec{\theta})$$



Method of Least Squares

- LS Fit for a polynomial
  - The illustration to the right is for different polynomial fits which are possible for least squares, including a flat constant, i.e. 0^{th} order polynomial
  - With the following data, test a similar least squares fit for the points defined by: x = (0.0, 1.0, 2.0, 3.0, 4.0, 5.0) and y = (0.0, 0.8, 0.9, 0.1, -0.8, -1.0) at different polynomial orders
  - Similar to the illustration, calculate the chi-square assuming each point has a uncertainty in y of ±0.5

*note the x and y data in the text is not the same as what is shown in the plot



Least Squares Examination

- Using your random number generator, sample from a gaussian distribution of your own choosing, i.e. width and mean, for n=10, 100, 1000, and 10000 throws and fit each with a 2^{nd} and 3^{rd} order polynomial least squares fit. Use histograms.
- Assume any negative predictions from the resulting polynomial fit are zero.
- Calculate the chi-square for each combination of trials and polynomial least squares fits
- The uncertainty is related to the expected poisson fluctuations from your samples of the random number generator.
- Plot the resultant fits and see what happens for higher order polynomial fits, e.g. order 5, 7, 8, 12, whatever, etc… as the number of data points (random number generator throws) increases

Least Squares Examination (cont.)

- From the resultant fits for higher order polynomial fits, e.g. order 5, 7, 8, 12, whatever, etc… as the number of data points (random number generator throws) increases, calculate the chi-square using the uncertainty (or weight) as the expected poisson fluctuation
- Where do the polynomial fits give a good ‘fit’ to a gaussian, even though a gaussian distribution and polynomial are not the same?
- How does the agreement change as a function of polynomial order or throws from the random number generator?

Examples of Least Squares Routines

- Non-Linear LS Fit
    - Need a numerical method to evaluate the estimators and their covariance matrix. ROOT and other packages provides this capability to fit any type of function, including that provided by a user defined routine.
    - Examples of user routine for a chi-square numerical minimization:
        - Polynomial of order m:

$$y(t) = x_0 + x_1 t + x_2 t^2 + ... + x_m t^m$$

        - Gaussian:

$$y(t) = x_1 e^{-\frac{1}{2}\left(\frac{t-x_2}{x_3}\right)^2}$$

$$x_1 = \text{amplitude}$$

$$x_2 = \text{mean}$$

$$x_3 = \text{standard deviation}$$

Examples of Least Squares Routines

- Non-Linear LS Fit
    - Examples of user routine for a chi-square numerical minimization:
        - Exponential:

$$y(t) = x_1 e^{-x_2 t}$$

        - Trigonometric:

$$y(t) = x_1 sin(x_2 t) \qquad y(t) = x_1 cos(x_2 t)$$

$$y(t) = x_1 sin(x_2 t + x3) \qquad y(t) = x_1 cos(x_2 t + x3)$$

    - Damped Oscillator

$$y(t) = x_1 e^{-x_2 t} cos(x_3 t + x4)$$

Examples of Least Squares Routines

- Non-Linear LS Fit
  - Examples of user routine for a chi-square numerical minimization:
    - Breit-Wigner:

$$y(t) = \frac{2}{\pi x_2} \frac{x_3 x_2^2}{4(t - x_1)^2 + x_2^2}$$

$$x_3 = \text{amplitude}$$

$$x_1 = \text{mean}$$

$$x_2 = \text{width}$$

- We of course want to find the relation between true values, according to some hypothesis, and measured quantities, y, at known observations with no errors, x. e.g.: $f_j(\vec{y}; \lambda) = y_j - h_j$

Method of Least Squares

- Non-Linear LS Fit
  - We need to find the minimum function:

$$\chi^2 = (\vec{y} - \vec{h}(x))^T G_y (\vec{y} - \vec{h}(x))$$

  - The convergence of a numerical (iterative) procedure will depend on if we are in an area of the phase space where the chi-square function is similar to a quadratic form. That is to say, the non-linear case first approximation is the starting point for the numerical procedure. For this simple algorithm to work we must be near the absolute minimum.

  - We expand the function around a set of first approximations for the (r) unknowns or parameters:

$$f_j = y_j - h_j \rightarrow \phi$$

$$f_j(x)_{true} = f_j(x_0) + (\frac{\partial f_i}{\partial x_i})_{\vec{x_0}}(x_i - x_{i0}) + ... = 0$$

estimates

$$\sum_{i=1}^{r}$$

Method of Least Squares

- Non-Linear LS Fit
  - We expand this about

$$\vec{x}_0 \equiv \vec{\theta}_0 = (\theta_1, ..., \theta_r)$$

$$f(y_j; \vec{\theta}) = f(y_j; \vec{\theta}_0) + \sum_{i=1}^{r} \left(\frac{\partial f_i}{\partial x_i} \right)_{\vec{\theta}_0} (x_i - \theta_i)$$

  - Which gives us:

$$\chi^2 = (\vec{c} + A\vec{\xi})^T G_y (\vec{c} + A\vec{\xi})$$

$$\vec{\xi} \equiv \vec{x} - \vec{\theta}$$

$$\vec{c}_j = f_j(\vec{y}; \vec{\theta}_0) = y_j - h_j(\vec{\theta}_0)$$

  - elements of A:

$$a_{jl} = \left(\frac{\partial f_j}{\partial x_l} \right)_{\vec{\theta}_0} = - \left(\frac{\partial h_j}{\partial x_l} \right)_{\vec{\theta}_0}$$

  - and G is the inverse variance matrix.

Method of Least Squares

- Non-Linear LS Fit
  - Note that the second derivative must of course be positive at the minimum when the chi-square is of quadratic form:

$$\frac{1}{2}\left(\frac{\partial^2\chi^2}{\partial x_1 \partial x_2}\right)_{\vec{x}=\tilde{\theta}} = A^T G_y A > 0$$

  - A step beyond the simple iterative method, known as step-size reduction, applies the fact that on each side of the minimum the first derivative changes sign and the second derivative is positive.

Method of Least Squares

- LS Fit as a goodness-of-fit
  - The value of the chi-square minimum reflects the agreement between data and hypothesis and can thus be used as a goodness-of-fit test statistic:

$$\chi^2_{min} = \sum_{i=1}^{N} \frac{(y_i - \lambda(x_i;\hat{\theta}))^2}{\sigma_i^2}$$

  - where our hypothesized function form is given by λ.
  - If the hypothesis is correct, then the test statistic, t, follows the chi-square pdf:

$$f(t;n_d) = \frac{1}{2^{n_d/2}\Gamma(n_d/2)} t^{n_d/2-1} e^{-t/2}$$

  - where n_d is the number of data points - number of fitted parameters.

Method of Least Squares

- LS Fit as a goodness-of-fit
    - The chi-square pdf has an expectation value equal to the number of degrees of freedom such that if the minimum chi-square is approximately the number of degrees of freedom then the fit is considered “good.”
    - We can find the p-value, here the probability of obtaining a chi-square minimum as large as the one measured, or higher, if the hypothesis is correct:

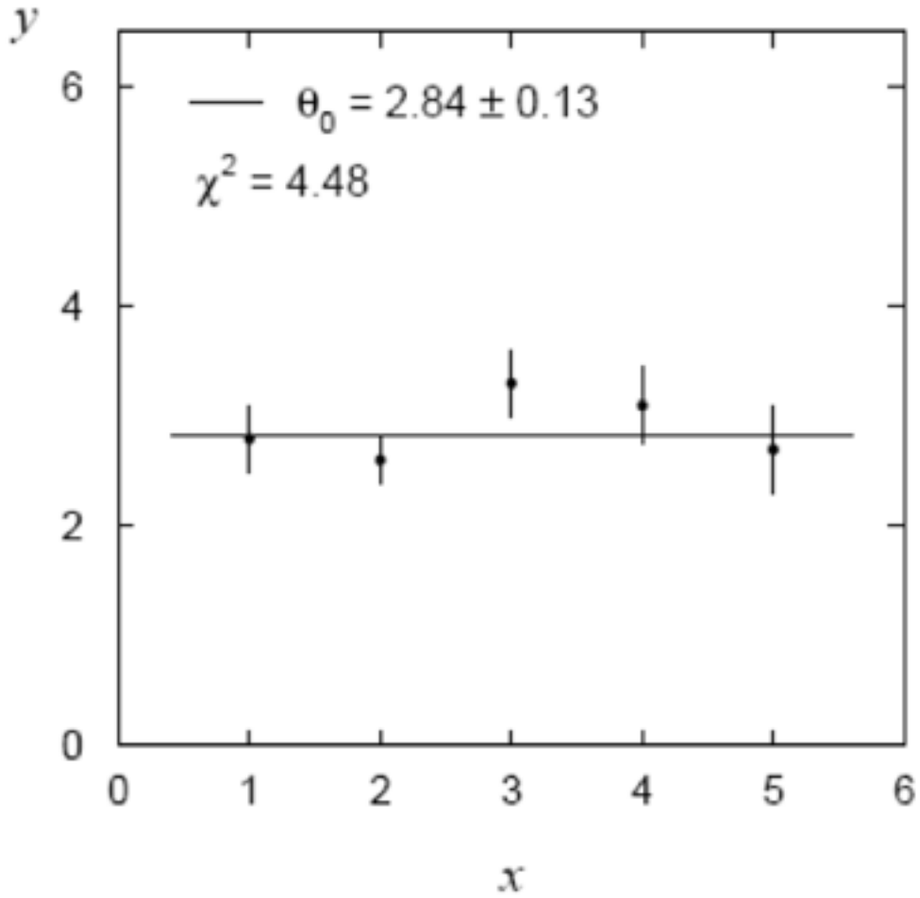
$$p = \int_{\chi^2_{min}}^{\infty} f(t; n_d) dt$$

- From the polynomial fit example:
    - 2 parameter fit: $\chi^2_{min} = 3.99 \qquad n_d = 5 - 2 = 3 \qquad p = 0.263$
    - 1 parameter fit: $\chi^2_{min} = 45.5 \qquad n_d = 5 - 1 = 4 \qquad p = 3.1 \times 10^{-9}$

*results from illustration in slide 16

Method of Least Squares

- LS Fit as a goodness-of-fit vs. statistical errors
    - It is important to note that a small statistical error does not imply a good fit, nor does a good fit imply small statistical errors.
    - The curvature of the chi-square near its minimum is related to the statistical errors
    - The value of the chi-square minimum is the goodness-of-fit.
    - For horizontal line fit, move the data points (transform), keeping the errors on the points the same.



Variance is same as previously, so the chi-square minimum is now "good"

Method of Least Squares

- LS Fit as a goodness-of-fit vs. statistical errors
    - The variance of the estimator (statistical error) tells us that if the experiment were repeated many times, the width of the distribution of the estimates, but not if the hypothesis, is correct.
    - The p-value tells us that if the hypothesis is correct, and the experiment repeated many times, what fraction of those will give equal or worse agreement between data and hypothesis according to the chi-square minimum test statistic.
    - Thus, a low p-value may indicate the hypothesis is wrong, due to systematic error.

Method of Least Squares

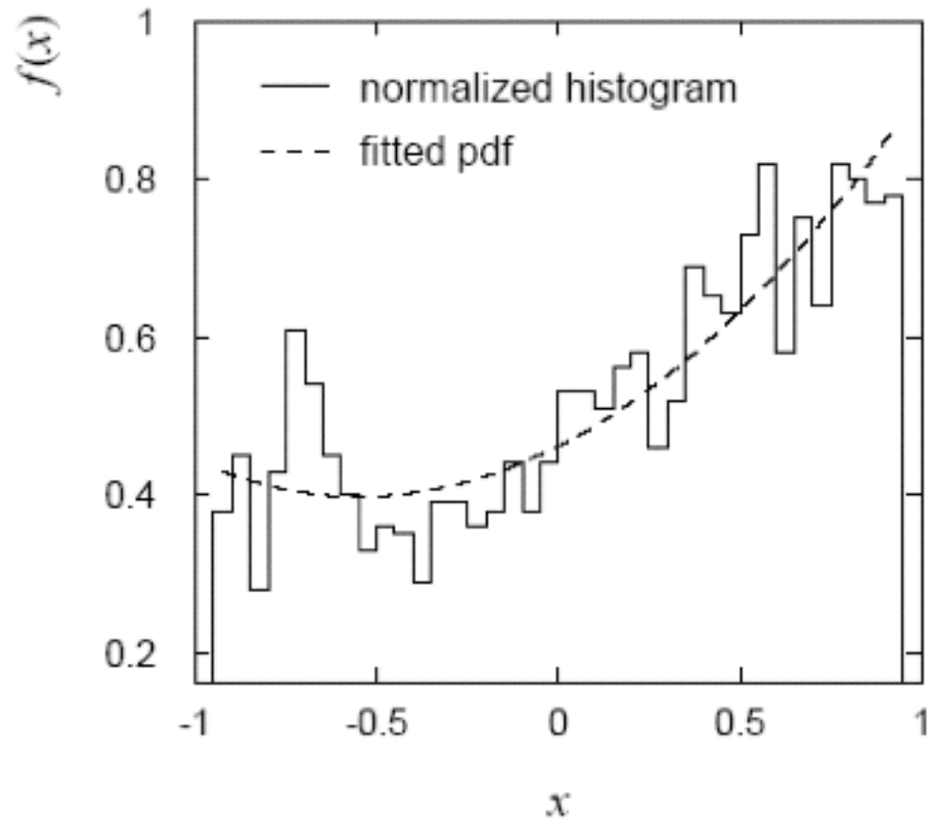
- LS method with binned data
  - If the data is split into N bins, with bin i containing n_i entries, there is a probability for an event to populate, p_i, that bin. Our hypothesized pdf is:

$$f(x;\vec{\theta})$$

  - The expected number of events in each bin is given by:

$$\lambda_i(\vec{\theta}) = n\int_{x_i^{min}}^{x_i^{max}} f(x;\vec{\theta})dx = np_i(\vec{\theta})$$

$$n = \sum_{i=1}^{N} n_i$$



Method of Least Squares

- LS method with binned data
    - Now for the fit we minimize

$$\chi^2(\vec{\theta}) = \sum_{i=1}^{N} \frac{(y_i - \lambda_i(\vec{\theta}))^2}{\sigma_i^2}$$

    - where our variances are not known a priori. We treat the y terms as Poisson random variables and, in place of the true variance, take either:

$$\sigma_i^2 = \lambda_i(\vec{\theta}) \text{ (Least Square method)}$$

$$\sigma_i^2 = y_i \text{ (Modified Least Square method)}$$

    - Note that the modified least squares is sometimes easier to compute, but the chi-square minimum statistic no longer follows the chi-square pdf if some of the bins have few or no entries.

Method of Least Squares

- LS method with binned data
    - We lose a degree of freedom because of the normalization condition:

$$\sum n_i = n$$

    - such that the chi-square minimum statistic will follow:

$$f(\chi^2, N-1-L)$$

    - assuming the model consists of L independent parameters.
    - It is NOT correct to fit for the normalization, e.g.

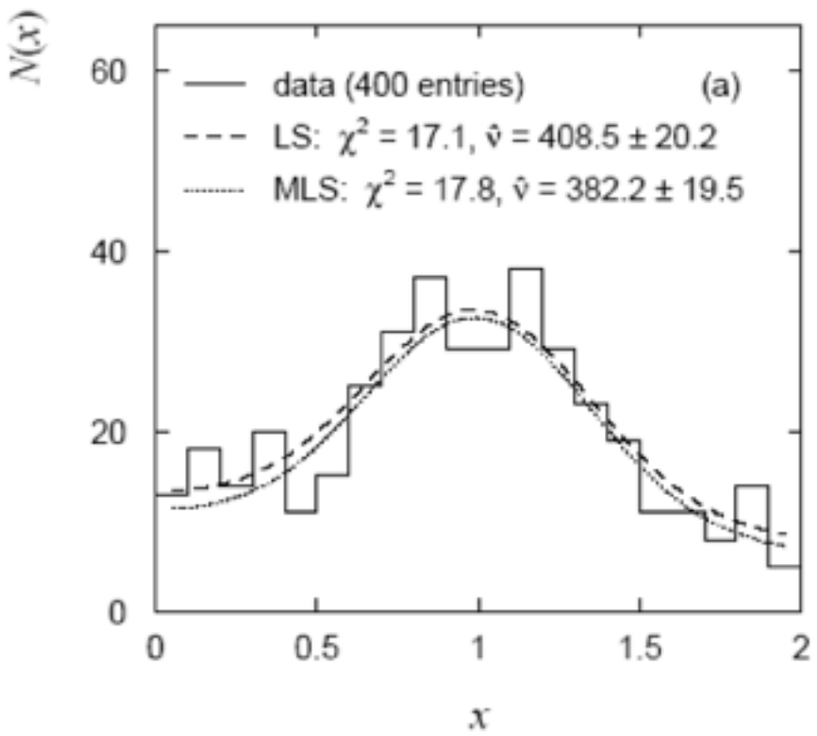
$$\lambda_i(\vec{\theta}, \nu) = \nu \int_{x_i^{min}}^{x_i^{max}} f(x; \vec{\theta}) dx = \nu p_i(\vec{\theta})$$

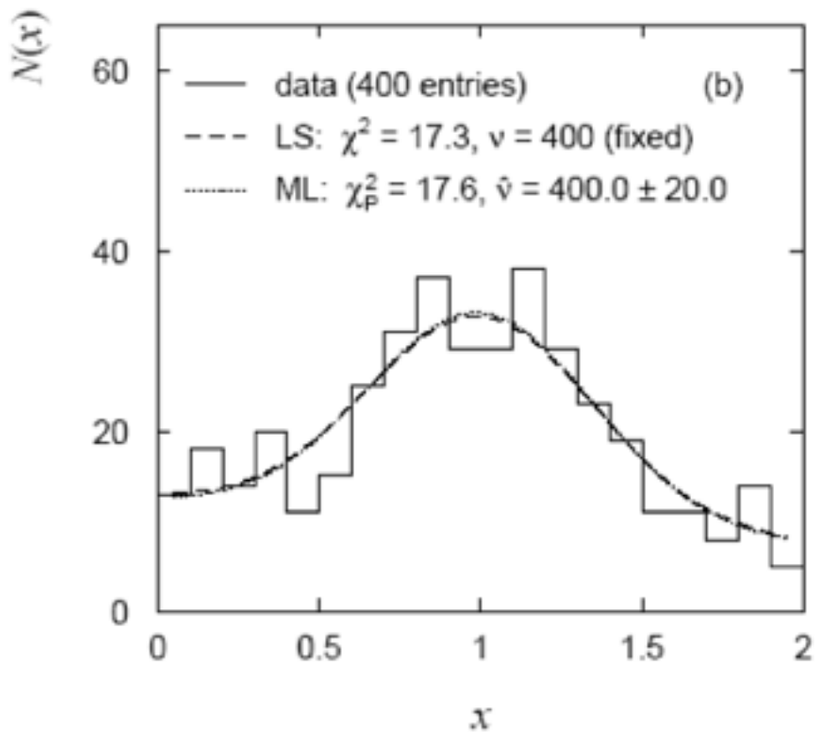
    - The estimator for n, $\hat{\nu}$, will be bad.

$$\hat{\nu}_{LS} = n + \frac{\chi^2_{min}}{2} \qquad \hat{\nu}_{MLS} = n - \chi^2_{min}$$

Method of Least Squares

- LS method with binned data
    - Normalization example: n=400 entries in N=20 bins.
    - The expected chi-square minimum is near N-m which means the relative error in the estimated normalization is large when N is large and n is small.
    - Ultimately get n directly from the data for LS method, or use a maximum likelihood.





Method of Least Squares

- LS method with binned data
    - Choices of binning is critical. Two common choices are:
        - equal width
        - equal probability
    - It is important not to choose the binning in order to make the chi-square minimum as small as possible! Doing so would cause the statistic to no longer follow the chi-square distribution.
    - It is necessary to have several entries (>5) in each bin so that the statistic approximates a standard normal distribution.

Method of Least Squares

- Combining measurements with LS method
    - The LS method may be used to obtain the weighted average of N measurements of the true value λ.
    - Given measurements, y, the variance, assumed to be known, is:

$$\sigma_i^2 = V[y_i]$$

    - For uncorrelated measurements:

$$\chi^2(\lambda) = \sum_{i=1}^{N} \frac{(y_i - \lambda)^2}{\sigma_i^2}$$

    - and, as usual, we solve for the first derivative equated to zero.

$$\hat{\lambda} = \frac{\sum_{i=1}^{N} y_i/\sigma_i^2}{\sum_{j=1}^{N} 1/\sigma_j^2} \qquad V[\hat{\lambda}] = \frac{1}{\sum_{i=1}^{N} 1/\sigma_i^2}$$

Method of Least Squares

- Combining measurements with LS method
  - If the covariance between measurements is , then minimize: $cov[y_i, y_j] = V_{ij}$

$$\chi^2(\lambda) = \sum_{i,j=1}^{N} (y_i - \lambda)(V^{-1})_{ij}(y_j - \lambda)$$

$$\hat{\lambda} = \sum_{i=1}^{N} w_i y_i \qquad\qquad w_i = \frac{\sum_{j=1}^{N}(V^{-1})_{ij}}{\sum_{k,l=1}^{N}(V^{-1})_{kl}}$$

$$V[\hat{\lambda}] = \sum_{i,j=1}^{N} w_i V_{ij} w_j$$

  - The least square estimate has zero bias and minimum variance according to the Gauss-Markov theorem.

Method of Least Squares

- Using LS with biased data samples
  - It may happen that some data samples will not reflect the true distribution due to, for instance, unequal detection efficiency for each event. To deal with this it is best to modify the theoretical model to account for the detection efficiency. In doing so, no modification of the least squares minimization is necessary. If that is not possible you can either
    - Modify the events in a bin, n_i: If the detection efficiency for event j in bin i is:

$$\epsilon_{ij} \rightarrow n_i^{'} = \sum_{j=1}^{n_i} 1/\epsilon ij \qquad \text{and minimize} \qquad \chi^2 = \sum_{i=1}^{n_i} (n_i^{'} - f_i)^2 / f_i$$

    - Modify $f_i : \; f_i^{'} = f_i D_i \qquad D_i = n_i^{-1} \sum_{j=1}^{N} \epsilon_{ij}$

$$\text{and minimize} \quad \chi^2 = \sum_{i=1}^{N} (n_i - f_i^{'})^2 / f_i^{'}$$

  - Both work well when the variation of the weights is small, otherwise the uncertainty of the estimates are not well defined.