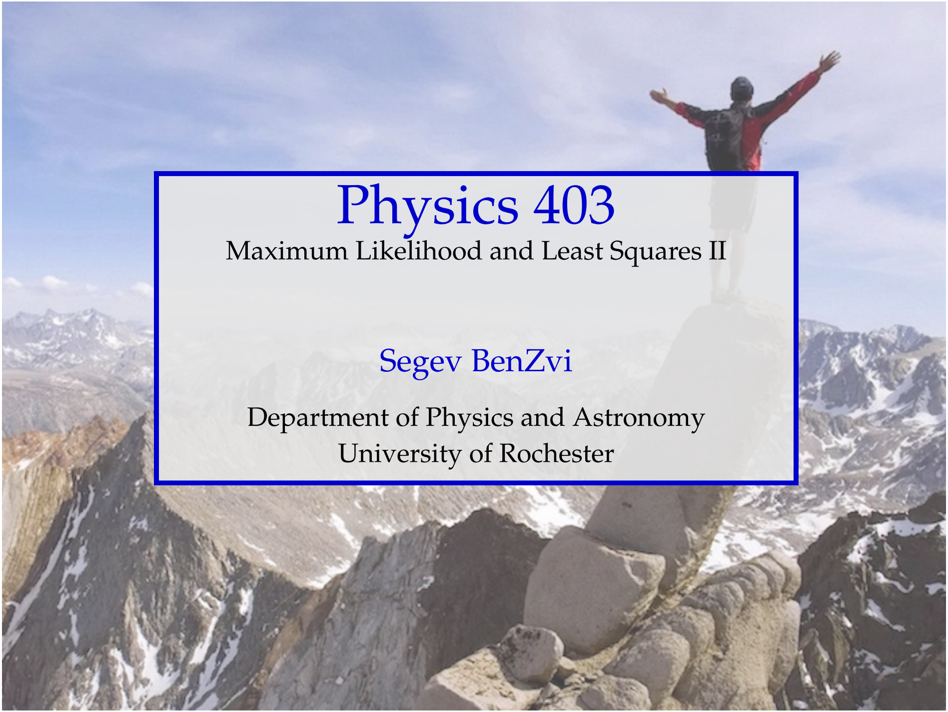


Physics 403

Maximum Likelihood and Least Squares II

Segev BenZvi

Department of Physics and Astronomy
University of Rochester

Table of Contents

Maximum Likelihood Technique

- The method of maximum likelihood is an extremely important technique used in frequentist statistics
- There is no mystery to it. Here is the connection to the Bayesian view: given parameters $\boldsymbol{x}$ and data $\boldsymbol{D}$, Bayes' Theorem tells us that

$$p(\boldsymbol{x}|\boldsymbol{D}, I) \propto p(\boldsymbol{D}|\boldsymbol{x}, I)\, p(\boldsymbol{x}|I)$$

  where we ignore the marginal evidence $p(\boldsymbol{D}|I)$
- Suppose $p(\boldsymbol{x}|I) =$ constant for all $\boldsymbol{x}$. Then

$$p(\boldsymbol{x}|\boldsymbol{D}, I) \propto p(\boldsymbol{D}|\boldsymbol{x}, I)$$

  and the best estimator $\hat{\boldsymbol{x}}$ is simply the value that maximizes the likelihood $p(\boldsymbol{D}|\boldsymbol{x}, I)$
- So the method of maximum likelihood for a frequentist is equivalent to maximizing the posterior $p(\boldsymbol{x}|\boldsymbol{D}, I)$ with uniform priors on the $\{x_i\}$.

Frequentist Notation

Maximum Likelihood Estimators

- Just to avoid confusion: in Cowan's book, the likelihood is written using the notation
$$\mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta})$$
where $\boldsymbol{x}$ are the data and $\boldsymbol{\theta}$ are the parameters
- Don't get thrown off. This is still equivalent to a Bayesian likelihood:
$$p(\boldsymbol{\theta}|\boldsymbol{x},I) = \frac{\mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta})\,p(\boldsymbol{\theta})}{\int d\boldsymbol{\theta}'\;\mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta}')\,p(\boldsymbol{\theta}')}$$
- I don't love the notation because it obscures the fact that $\mathcal{L}$ is a PDF, which we use to get best estimators with the tricks introduced in earlier classes. When needed, we'll denote it as $\mathcal{L}$ because L is used in Sivia for the logarithm of the posterior PDF
- In everyday applications, you will maximize $\ln\mathcal{L}$, or minimize $-\ln\mathcal{L}$

ML Estimator: Exponential PDF

Example

Consider N data points distributed according to the exponential PDF $p(t|\tau) = e^{-t/\tau}/\tau$. The log-likelihood function is

$$\ln p(D_i|\tau) = \ln \mathcal{L} = -\sum_{i=1}^{N} \left(\ln \tau + \frac{t_i}{\tau} \right)$$

Maximizing with respect to τ gives

$$\left. \frac{\partial \ln \mathcal{L}}{\partial \tau} \right|_{\hat{\tau}} = 0 \implies \hat{\tau} = \frac{1}{N} \sum_{i=1}^{N} t_i$$

It's also easy to show that

$$\mathrm{E}\left(\hat{\tau}\right) = \tau \implies \hat{\tau} \text{ is unbiased}$$

Properties of ML Estimators

- ML estimators are usually consistent ($\hat{\theta} \to \theta$)
- ML estimators are usually biased ($b = \mathrm{E}\left(\hat{\theta}\right) - \theta \neq 0$)
- ML estimators are invariant under parameter transformations:

$$\widehat{f(\theta)} = f(\hat{\theta})$$

Example

Working with $\lambda = 1/\tau$ in the exponential distribution, it's easy to show that $\hat{\lambda} = 1/\hat{\tau}$ [1].

- Due to sum of terms in $\ln\mathcal{L}$, it tends toward a Gaussian by the Central Limit Theorem, so

$$\sigma^2_{\hat{\theta}} = \left(-\left.\frac{\partial^2 \ln\mathcal{L}}{\partial\theta^2}\right|_{\hat{\theta}}\right)^{-1}$$

Minimum Variance Bound

Rao-Cramér-Frechet Inequality

Given $\mathcal{L}$ you can also put a lower bound on the variance of a ML estimator:

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

Example

For the exponential distribution,

$$\left.\frac{\partial^2 \mathcal{L}}{\partial \tau^2}\right|_{\hat{\tau}} = \frac{N}{\tau^2}\left(1 - \frac{2\hat{\tau}}{\tau}\right), \quad b = 0,$$

and so we can prove that $\hat{\tau}$ is efficient (variance is at the lower bound):

$$\operatorname{var}(\hat{\tau}) \geq \operatorname{E}\left(-\frac{N}{\tau^2}(1 - 2\hat{\tau}/\tau)\right)^{-1} = \left(-\frac{N}{\tau^2}(1 - 2\operatorname{E}(\hat{\tau})/\tau)\right)^{-1} = \frac{\tau^2}{N}$$

Variance of ML Estimators

- We can express the variance of ML estimators using the same tricks we applied to the posterior PDF: expand $\ln\mathcal{L}$ in a Taylor series about $\hat{\theta}$:

$$\ln\mathcal{L}(\theta) \approx \ln\mathcal{L}_{\max} - \frac{(\theta-\hat{\theta})^2}{2\sigma_{\hat{\theta}}^2}$$

$$\therefore \ln\mathcal{L}(\hat{\theta}\pm\sigma_{\hat{\theta}}) = \ln\mathcal{L}_{\max} - \frac{1}{2}$$

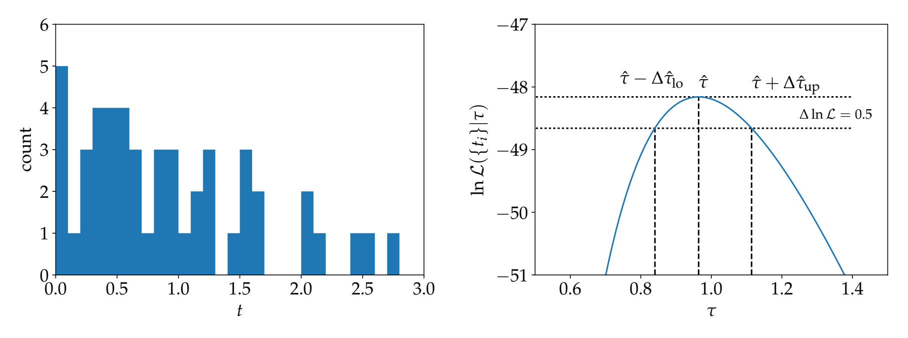
- In other words, a change in θ by one standard deviation from $\hat{\theta}$ leads to a decrease in $\ln\mathcal{L}$ by $1/2$ from its maximum value
- The definition $\Delta\ln\mathcal{L} = 1/2$ is often taken as the definition of statistical uncertainty on a parameter
- Strictly speaking this is only correct in the Gaussian limit, but it can often be a nice, reasonably accurate shortcut

Variance of ML Estimators

Realization of Exponential Data

Example

Generating 50 $\{t_i\}$ according to an exponential distribution with $\tau = 1$:

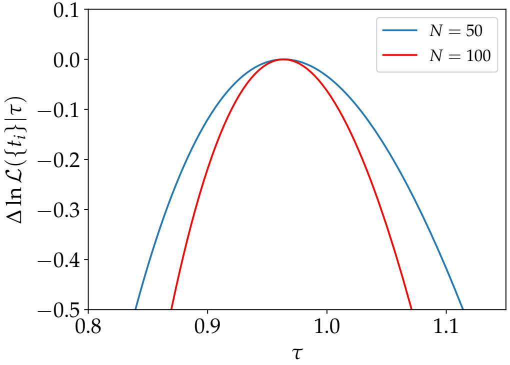


Using the criterion $\Delta \ln \mathcal{L} = 0.5$ we find $\hat{\tau} = 0.96^{+0.15}_{-0.12}$

Variance of ML Estimators

More Data

Adding more data narrows the distribution of $\mathcal{L}$, as you would expect for any PDF



The distribution also becomes more symmetric, which you would expect from the Central Limit Theorem

Asymmetric Uncertainties

- Because $\ln\mathcal{L}$ becomes increasingly parabolic with N due to the Central Limit Theorem, we can define rules of thumb for estimating variances on parameters:

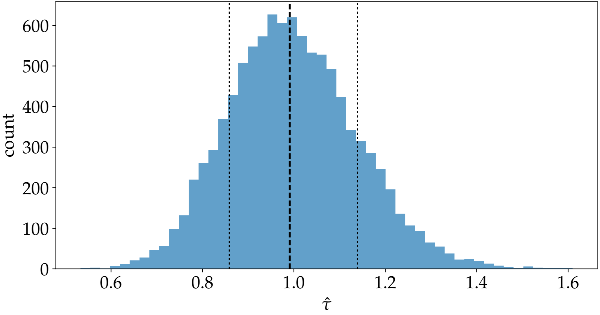
$$\ln\mathcal{L}(\theta) \approx \ln\mathcal{L}_{\max} - \frac{(\theta-\hat{\theta})^2}{2\sigma^2_{\hat{\theta}}}.$$

Range	$\Delta\ln\mathcal{L}$
1σ	$1/2\cdot(1)^2 = 0.5$
2σ	$1/2\cdot(2)^2 = 2$
3σ	$1/2\cdot(3)^2 = 4.5$

- This is done even when the likelihood isn't parabolic, producing asymmetric error bars (as we saw)
- Justification: you can reparameterize θ such that $\ln\mathcal{L}$ is parabolic, which is OK because of the invariance of the ML estimator under transformations

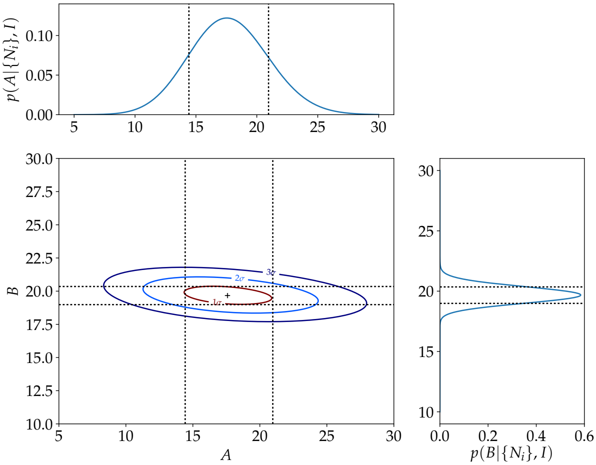
Other Approaches to Calculate Variance

- You could use $\mathcal{L}$ to estimate a central confidence interval on $\hat{\theta}$: find the 16^{th} and 84^{th} percentiles
- Monte Carlo Method: generate many random realizations of the data, maximize $\ln\mathcal{L}$ for each, and study the distribution of $\hat{\theta}$:



- From 10,000 realizations of the exponential data set, the distribution of ML estimators $\hat{\tau}$ gives $\hat{\tau} = 0.99^{+0.15}_{-0.13}$. Not bad...

ML Technique with > 1 Parameter



- For > 1 parameter:

$$\mathrm{cov}\left(x_i, x_j\right) = \left(-\frac{\partial^2 \ln \mathcal{L}}{\partial x_i \partial x_j}\bigg|_{\hat{x}_i,\hat{x}_j}\right)^{-1}$$

- Use the $\Delta \ln \mathcal{L}$ trick to get contours for $1\sigma, 2\sigma$, etc.
- Project ellipse onto each axis (i.e., marginalize) to get uncertainties in each parameter

ML Technique: Joint Confidence Intervals

Usually we want to calculate a joint likelihood on several parameters but only produce confidence intervals for individual parameters. However, if we want confidence ellipses in several parameters jointly, we need to change the $\Delta \ln \mathcal{L}$ rule a bit:

		joint parameters					
Range	p	1	2	3	4	5	6
1σ	68.3%	0.50	1.15	1.76	2.36	2.95	3.52
2σ	95.4%	2.00	3.09	4.01	4.85	5.65	6.4
3σ	99.7%	4.50	5.90	7.10	8.15	9.10	10.05

It's not very common to calculate things this way; usually we are interested in the marginal distributions of individual parameters. For more details on this, see [2].

Table of Contents

Connection to χ^2

- Suppose our data $\boldsymbol{D}$ are identical independent measurements with Gaussian uncertainties. Then the likelihood is

$$p(D_i|\boldsymbol{x},I) = \frac{1}{\sqrt{2\pi}\sigma_i}\exp\left[-\frac{(F_i - D_i)^2}{2\sigma_i^2}\right], \quad p(\boldsymbol{D}|\boldsymbol{x},I) = \prod_{i=1}^{N} p(D_i|\boldsymbol{x},I),$$

  where we defined the functional relationship between $\boldsymbol{x}$ and the ideal (noiseless) data $\boldsymbol{F}$ as

$$F_i = f(\boldsymbol{x}, i)$$

- If we define χ^2 as the sum of the squares of the normalized residuals $(F_i - D_i)/\sigma_i$, then

$$\chi^2 = \sum_{i=1}^{N}\frac{(F_i - D_i)^2}{\sigma_i^2} \implies p(\boldsymbol{D}|\boldsymbol{x},I) \propto \exp\left(-\frac{\chi^2}{2}\right)$$

Maximum Likelihood and Least Squares

- With a uniform prior on $\boldsymbol{x}$, the logarithm of the posterior PDF is

$$L = \ln p(\boldsymbol{x}|\boldsymbol{D}, I) = \ln p(\boldsymbol{D}|\boldsymbol{x}, I) = \text{constant} - \frac{\chi^2}{2}$$

- The maximum of the posterior (and likelihood) will occur when χ^2 is a minimum. Hence, the optimal solution $\hat{\boldsymbol{x}}$ is called the least squares estimate
- Least squares/maximum likelihood is used all the time in data analysis, but...
- **Note**: there is nothing mysterious or even fundamental about this; least squares is what Bayes' Theorem reduces to if:
  1. Your prior on your parameters is uniform
  2. The uncertainties on your data are Gaussian

Maximum Likelihood: Poisson Case

- Suppose that our data aren't Gaussian, but a set of Poisson counts $\boldsymbol{n}$ with expectation values $\boldsymbol{\nu}$. E.g., we are dealing with binned data in a histogram. Then the likelihood becomes

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

- In the limit $N \rightarrow$ large, this becomes

$$p(n_i|\nu_i, I) \propto \exp\left[-\sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{2\nu_i}\right]$$

- The corresponding χ^2 statistic is given by

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i}$$

Pearson's χ^2 Test

- The quantity

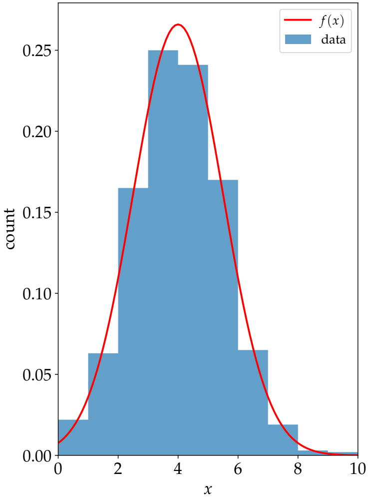
$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i}$$

  is also known as Pearson's χ^2 statistic
- Pearson's χ^2 test is a standard frequentist method for comparing histogrammed counts $\{n_i\}$ against a theoretical expectation $\{\nu_i\}$
- Convenient property: this test statistic will be asymptotically distributed like χ^2_N regardless of the actual distribution that generates the relative counts $\{n_i\}$. It is distribution free
- In practice, we can use Pearson's χ^2 to calculate a p-value

$$p(\chi^2_{\text{Pearson}} \geq \chi^2 | N)$$

- **Caveat**: the counts in each bin must not be too small; $n_i \geq 5$ for all i is a reasonable rule of thumb

Modified Least Squares



- Sometimes you will encounter a χ^2 statistic for binned data defined like this:

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - f_i)^2}{n_i}$$

- The variance is no longer the expected counts (as expected in a Poisson distribution) but the observed counts n_i. This is called modified least squares
- You don't really want this, unless you made mistakes counting n_i
- But, statistics packages may use this statistic when fitting functions to binned data

Robustness of Least Squares Algorithm

- Our definition of χ^2 as the quadrature sum (or l_2-norm) of the residuals makes a lot of calculations easy, but it isn't particularly robust. I.e., it can be affected by outliers
- **Note**: the l_1-norm

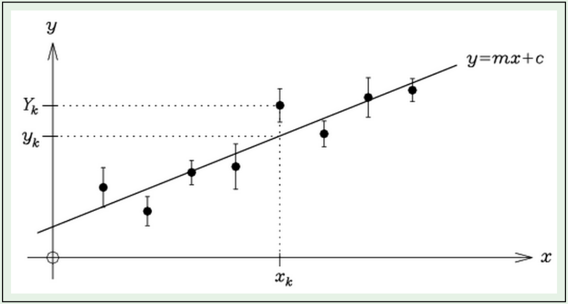
  $$l_1\text{-norm} = \sum_{i=1}^{N} \left| \frac{F_i - D_i}{\sigma_i} \right|$$

  is much more robust against outliers in the data
- This isn't used too often but if your function $f(\boldsymbol{x})$ is linear in the parameters it's not hard to calculate
- See chapter 15 of *Numerical Recipes in C* for an implementation [2]
- In Python there should be an implementation in the `statsmodels` package [3]

Application: Fitting a Straight Line to Data

Example

Suppose we have N measurements y_i with Gaussian uncertainties σ_i measured at positions x_i.



Given the straight line model $y_i = mx_i + b$, what are the best estimators of the parameters m and b?

Minimize the χ^2

Letting $F_i = mx_i + b$ and $D_i = y_i$, the χ^2 is

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

Minimizing χ^2 as a function of the parameters gives

$$\frac{\partial \chi^2}{\partial m} = \sum_{i=1}^{N} \frac{2(mx_i + b - y_i)x_i}{\sigma_i^2} \quad \text{and} \quad \frac{\partial \chi^2}{\partial b} = \sum_{i=1}^{N} \frac{2(mx_i + b - y_i)}{\sigma_i^2}$$

Defining $w_i = 2/\sigma_i^2$ and rewriting this as a matrix equation,

$$\nabla \chi^2 = \begin{pmatrix} A & C \\ C & B \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix} - \begin{pmatrix} p \\ q \end{pmatrix} = 0$$

$$A = \sum x_i^2 w_i,\ B = \sum w_i,\ C = \sum x_i w_i,\ p = \sum x_i y_i w_i,\ q = \sum y_i w_i$$

Best Estimators of a Linear Function

- Inverting the matrix, we find that

$$\hat{m} = \frac{Bp - Cq}{AB - C^2} \quad \text{and} \quad \hat{b} = \frac{Aq - Cp}{AB - C^2}$$

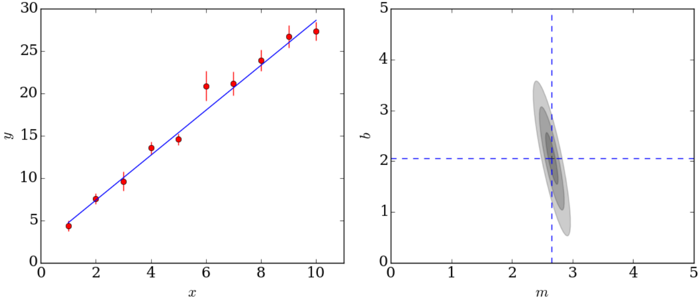
- The covariance matrix is found by evaluating $[2\nabla\nabla\chi^2]^{-1}$:

$$\begin{pmatrix} \sigma_m^2 & \sigma_{mb}^2 \\ \sigma_{mb}^2 & \sigma_b^2 \end{pmatrix} = 2\begin{pmatrix} A & C \\ C & B \end{pmatrix}^{-1} = \frac{2}{AB - C^2}\begin{pmatrix} B & -C \\ -C & A \end{pmatrix}$$

- We note that even though the data $\{y_i\}$ are independent, the parameters $\hat{m}$ and $\hat{b}$ end up anticorrelated due to the off-diagonal terms in the covariance matrix
- This makes a lot of sense, actually; wiggling the slope of the line m clearly changes the y-intercept b

LS Uncertainties

Example LS fit: best estimators $\hat{m} = 2.66 \pm 0.10$, $\hat{b} = 2.05 \pm 0.51$, $\text{cov}\,(m, b) = -0.10 \implies \rho = -0.94$, quite anti-correlated



We calculated the covariance matrix analytically, but note that we could have used a fitter with a quadratic approximation, or noted that

$$\Delta\chi^2 = -2\Delta \ln \mathcal{L}$$

$$\therefore \Delta\chi^2 = 1 \text{ from minimum} \implies 1\sigma \text{ contour}$$

Generalization: Correlated Uncertainties in Data

- So far we have been focusing on the case where uncertainties in our measurements are completely uncorrelated
- If this is not the case, then we can generalize χ^2 to

$$\chi^2 = \left(\boldsymbol{y} - \hat{\boldsymbol{y}}\right)^\top \boldsymbol{\sigma}^{-1} \left(\boldsymbol{y} - \hat{\boldsymbol{y}}\right)$$

  where $\boldsymbol{\sigma}$ is the covariance matrix of the data
- If the fit function depends linearly on the parameters,

$$y(x) = \sum_{i=1}^{m} a_i f_i(x), \qquad \hat{\boldsymbol{y}} = \boldsymbol{A} \cdot \boldsymbol{a}, \qquad A_{ij} = f_j(x_i)$$

  then

$$\begin{aligned}\chi^2 &= \left(\boldsymbol{y} - \hat{\boldsymbol{y}}\right)^\top \boldsymbol{\sigma}^{-1} \left(\boldsymbol{y} - \hat{\boldsymbol{y}}\right) \\ &= \left(\boldsymbol{y} - \boldsymbol{A} \cdot \boldsymbol{a}\right)^\top \boldsymbol{\sigma}^{-1} \left(\boldsymbol{y} - \boldsymbol{A} \cdot \boldsymbol{a}\right)\end{aligned}$$

Exact Solution to Linear Least Squares

- This is the case of linear least squares; the LS estimators of the $\{a_i\}$ are unbiased, efficient, and can be solved analytically
- The general solution:

$$
\begin{aligned}
\chi^2 &= \left(\boldsymbol{y} - \boldsymbol{A}\cdot\boldsymbol{a}\right)^\top \boldsymbol{\sigma}^{-1} \left(\boldsymbol{y} - \boldsymbol{A}\cdot\boldsymbol{a}\right) \\
\boldsymbol{a} &= (\boldsymbol{A}^\top\boldsymbol{\sigma}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{\sigma}^{-1}\cdot\boldsymbol{y} \\
\mathrm{cov}\left(\hat{a}_i, \hat{a}_j\right) &= (\boldsymbol{A}^\top\boldsymbol{\sigma}^{-1}\boldsymbol{A})^{-1}
\end{aligned}
$$

- In practice one still minimizes numerically, because the matrix inversions in the analytical solution can be computationally expensive and numerically unstable
- Nice property: if uncertainties are Gaussian and the fit function is linear in the m parameters, then $\chi^2 \sim \chi^2_{N-m}$. But often these assumptions are broken, e.g., when using binned data with low counts

Nonlinear Least Squares

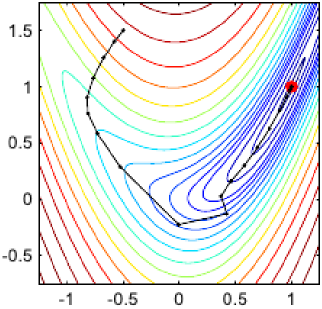
- If $y(x)$ is nonlinear in the parameters, we can try to approximate χ^2 as quadratic and use Newton's Method:

$$\boldsymbol{a}_{n+1} = \boldsymbol{a}_n - [\boldsymbol{H}(\boldsymbol{a}_n)]^{-1} \nabla \chi^2(\boldsymbol{a}_n)$$

- But, this could be a poor approximation to the function, so we could also try to use steepest descent:

$$\boldsymbol{a}_{n+1} = \boldsymbol{a}_n - \gamma_n \nabla \chi^2(\boldsymbol{a}_n)$$

- **Levenberg-Marquardt Algorithm**: use steepest descent far from the minimum, then switch to using the Hessian [2]. Basis of `scipy.optimize.curve_fit`



χ^2 and Goodness of Fit

- Because $\chi^2 \sim \chi^2_{N-m}$ if several conditions are satisfied, it can be used to estimate the goodness of fit
- Basic idea: the outcome of Linear Least Squares is the value χ^2_{min}. Goodness of fit comes from calculating the p-value

$$p(\chi^2 \geq \chi^2_{\text{min}} | N, m)$$

- This tail probability tells us how unlikely it is to have observed our data *given the model* and its best fit parameters
- Recall the warning about p-values: they are biased against the null hypothesis that the model is correct, and can lead you to spuriously reject a model
- The 5σ rule applies, because we're not dealing with a proper posterior PDF

ML and Goodness of Fit

- The ML technique does not provide a similar goodness of fit parameter because there is no standard reference distribution to compare to
- Suggested approach: estimate paramaters with ML, but calculate goodness of fit by binning the data and using χ^2
- **Note**: be careful about assuming that your χ^2 statistic actually follows a χ^2 distribution. Remember that this is true only for linear models with Gaussian uncertainties
- This isn't the 1920s. Use simulation to model the distribution of your χ^2 statistic and calculate p-values from that distribution

Summary

- The maximum likelihood (ML) method and the least squares (LS) method are very popular techniques for parameter estimation and are easy to implement
- Generally it's better to use the ML technique if you have the PDFs of the measurements. Your estimators will be biased though it's not an issue in the large N limit
- If your problem is linear in the parameters and you have Gaussian uncertainties, you can use LS. Advantage: closed form solutions and a measure of the goodness of fit
- Uncertainties on estimators:

Error	$\Delta \ln \mathcal{L}$	$\Delta\chi^2$
1σ	0.5	1
2σ	2	4
3σ	4.5	9

References I

[1] Glen Cowan. *Statistical Data Analysis*. New York: Oxford University Press, 1998.

[2] W. Press et al. *Numerical Recipes in C*. New York: Cambridge University Press, 1992. URL: http://www.nr.com.

[3] *Statsmodels*. URL: http://statsmodels.sourceforge.net/.