Data Analysis in Astronomy and Physics

Lecture 11: Data Modelling & Parameter Estimation

M. Röllig

Introduction

The aim of model fitting is to provide most parsimonious 'best' fit of a parametric model to data. Assume we have N data Z_i following a Gaussian distribution:

Out[ ]//TraditionalForm=

$$\mathcal{P}_r(z) = \frac{1}{\sqrt{2\,\pi}\;\sigma}\exp\left[-\left(\frac{z-\mu}{2\sigma^2}\right)^2\right]$$

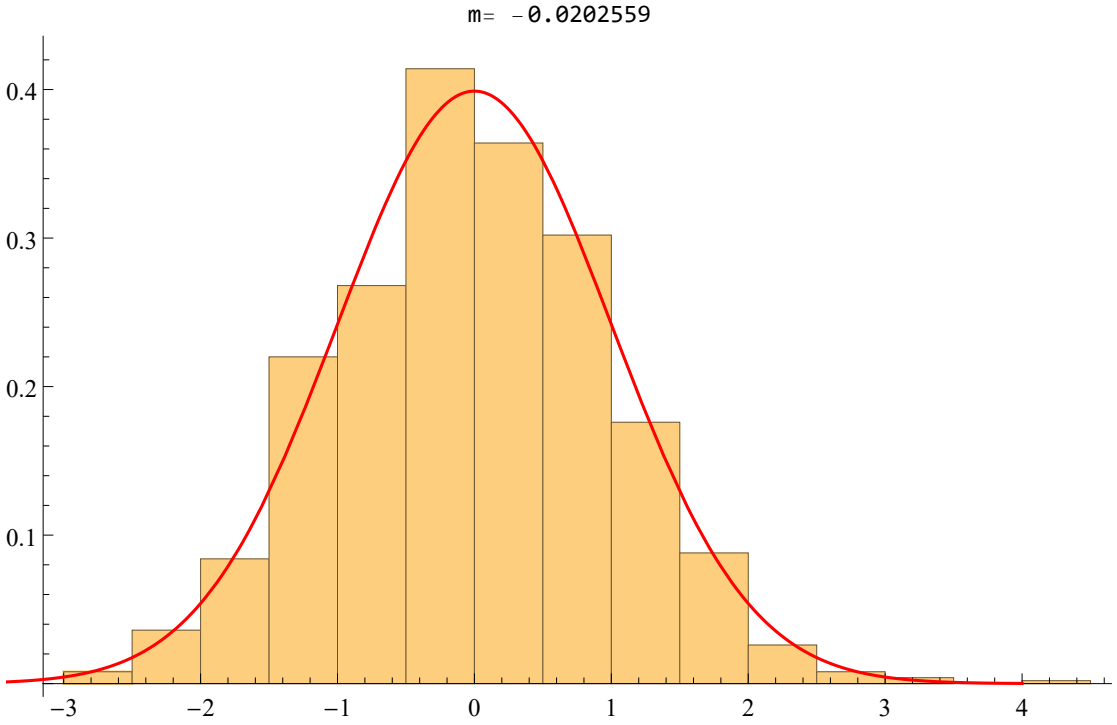
then the statistic

Out[ ]//TraditionalForm=

$$m = \frac{1}{N}\sum_{i=1}^{N} Z_i$$

is a good estimator for μ and has a known (Gaussian) distribution, which can be used to compute confidence intervals. Or, from the Bayesian point of view, we can calculate a probability distribution for μ, given the data.

Introduction



Any data-modelling procedure is just a more elaborate version of this, assuming we know the relevant probability distributions.

Introduction

Suppose our data Z_i were measured at various values of some independent variable X_i, and we believed that they were 'really' scattered, with Gaussian errors, around the underlying functional relationship

Out[•]//TraditionalForm=

$$\mu = \mu(x, \alpha_1, \alpha_2, \ldots)$$

in which $\alpha_1, \alpha_2, \ldots$ are unknown parameters (slopes, intercepts,…) of the relationship. We then have:

Out[•]//TraditionalForm=

$$\mathcal{P}(\mathrm{z}|\alpha_1,\alpha_2,\ldots) = \frac{1}{\sqrt{2\,\pi}\;\sigma}\exp\left(-\frac{(z-\mu(x,\,\alpha_1,\,\alpha_2,\,\ldots))^2}{2\,\sigma^2}\right)$$

Introduction

and, by Bayes' theorem, we have the posterior probability distribution for the parameters

Out[•]//TraditionalForm=

$$\mathcal{P}_f(\alpha_1,\alpha_2,\ldots|Z_i,\mu) \propto \prod_{i=1}^{N} \frac{1}{\sqrt{2\,\pi}\;\sigma} \exp\left(-\frac{(Z_i-\mu(x,\,\alpha_1,\,\alpha_2,\,\ldots))^2}{2\,\sigma^2}\right) \mathcal{P}_f(\alpha_1,\alpha_2,\ldots)$$

with the prior $\mathcal{P}_f(\alpha_1, \alpha_2, \ldots)$. This, at least formally, completes our task; we have a probability distribution for the parameters of our model, given the data.

It may be very time/CPU-consuming to actually compute/integrate the models and the distributions.

Maximum Likelihood

Suppose our data are described by the probability density function $f(X;\ \alpha)$, where x is a variable, and α is a parameter (maybe many parameters) characterizing the known form of f. We want to estimate α. If $X_1, X_2, \ldots, X_N$ are data, presumed independent and all drawn from f, then the likelihood function $\mathcal{L}$ is

Out[ ]//TraditionalForm=

$$\mathcal{L}(X_1,X_2, \ldots ,X_N) = f(X_1, X_2, \ldots, X_N \mid \alpha) = f(X_1 \mid \alpha)\, f(X_2 \mid \alpha) \ldots f(X_N \mid \alpha) = \prod_{}^{N} f(X_i \mid \alpha)$$

- From the classical point of view this is the probability, given α, of obtaining the data.
- From the Bayesian point of view it is proportional to the probability of α, given the data and assuming that the priors are 'diffuse'.
- Practically speaking, this means that they change little over the peaked region of the likelihood function.
- The peak value of $\mathcal{L}$ seems likely to be a useful choice of the ‘best’ estimate of α.

Maximum Likelihood

Formally, the likelihood function is a probability density function for the data sample $X_1, X_2, \ldots, X_N$, normalized in its range of definition Ω

Out[•]//TraditionalForm=

$$\int_{\Omega} \mathcal{L}(X_1, X_2, \ldots, X_N)\, dx_1 \ldots dx_N = \int_{\Omega} f(X_1, X_2, \ldots, X_N \mid \alpha)\, dx_1 \ldots dx_N = 1$$

$\mathcal{L}$ is not normalized across the α range!

Formally, the maximum-likelihood estimator (MLE) of α is $\hat{\alpha}$ = (that value of α which maximizes $\mathcal{L}(\alpha)$ for all variations of α).

Out[•]= $\mathcal{L}(X_1, X_2, \ldots, X_N \mid \hat{\alpha}) \geq \mathcal{L}(X_1, X_2, \ldots, X_N \mid \alpha) \qquad \forall \alpha$

Maximum Likelihood

The parameter can be discrete or continuous. In case of continuous α we can apply standard numerical techniques to identify $\hat{\alpha}$. Since $\mathcal{L}(x_1, x_2, \ldots, x_N) = \prod f(x_i \mid \alpha)$ is a product of many numbers smaller than 1 it can take very small values. Numerically it is advantageous to apply the logarithm to $\mathcal{L}$ and compute the so-called Log-Likelihood function:

Out[ ]//TraditionalForm=

$$\ln \mathcal{L}(X_1, X_2, \ldots, X_N) = \sum^{N} \log(f(X_i \mid \alpha))$$

Maximum Likelihood

Maximizing $\mathcal{L}$ with respect to one parameter α then usually can be done by:

Out[•]= $$\frac{\partial \mathcal{L}}{\partial \alpha} = \frac{\partial}{\partial \alpha} \sum_{i=1}^{N} \log\left(f(X_i \mid \alpha)\right) = 0 \qquad \Longrightarrow \quad \hat{\alpha}$$

Out[•]//TraditionalForm=
$$\frac{\partial^2 \mathcal{L}}{\partial \alpha^2}|_{\alpha=\hat{\alpha}} < 0$$

In the case of multiple parameters $\alpha = \alpha_1, \alpha_2, \ldots, \alpha_m$ it is

Out[•]= $$\frac{\partial \mathcal{L}}{\partial \alpha_j} = \frac{\partial}{\partial \alpha_j} \sum_{i=1}^{N} \log\left(f(X_i \mid \alpha)\right) = 0 \qquad \Longrightarrow \quad \hat{\alpha}$$

Out[•]//TraditionalForm=
$$\frac{\partial^2 \mathcal{L}}{\partial \alpha_i \, \alpha_j}|_{\alpha=\hat{\alpha}} = U_{\mathrm{ij}}(\widehat{\alpha}) \text{ negative definite}$$

U_{ij} is negative definite if all eigenvalues are < 0

Example: Estimate mean life time

The decay sequence of a radioactive probe has the probability density

Out[•]//TraditionalForm=

$$f(t|\tau) = \frac{1}{\tau}\exp\left(-\frac{t}{\tau}\right)$$

with the single parameter τ. In an experiment we measure n decays with the times t_i, $i = 1, \ldots, n$. The likelihood function of this data is

Out[•]=
$$\mathcal{L}(t_1, t_2, \ldots, t_N|\tau) = \prod_{i=1}^{n} \frac{1}{\tau}\exp\left(-\frac{t_i}{\tau}\right) \quad \Longrightarrow \quad \ln \mathcal{L}(t_1, t_2, \ldots, t_N|\tau) = \sum_{i=1}^{n}\left(-\ln(\tau) - \frac{t_i}{\tau}\right)$$

Example: Estimate mean life time

Maximising for τ

Out[•]= $\frac{\partial \ln \mathcal{L}}{\partial \tau} = \sum_{i=1}^{n} \left(-\frac{1}{\tau} - \frac{t_i}{\tau^2} \right) \quad \Longrightarrow \quad \hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i = \bar{t} \quad \text{with} \quad \frac{\partial^2 \ln \mathcal{L}}{\partial \tau^2}\Big|_{\tau=\hat{\tau}} == -\frac{n}{\hat{\tau}^2} < 0$

The maximum likelihood estimate (MLE) of the mean life time is the arithmetic mean of the measured times:

Example: Estimate Gaussian parameters

A sample $X_1, X_2, \ldots, X_N$ drawn from a normal distribution $\mathcal{N}(\mu,\sigma)$ has the likelihood function

Out[•]= $$\mathcal{L}(x_1,x_2,\ldots,x_N\,|\,\mu,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \quad \Longrightarrow \quad \ln \mathcal{L}(t_1,t_2,\ldots,t_N\,|\,\tau) = \frac{1}{2}\sum_{i=1}^{n}\left(-\ln\left(\sigma^2\right) - \ln\left(2\pi\right) - \frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

Maximizing for

Out[•]= $$\frac{\partial \ln \mathcal{L}}{\partial \mu} = \sum_{i=1}^{n} \frac{x_i-\mu}{\sigma^2} = 0$$

Out[•]= $$\frac{\partial \ln \mathcal{L}}{\partial \sigma^2} = \sum_{i=1}^{n} \left(-\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(x_i-\mu)^2\right) = 0$$

Example: Estimate Gaussian parameters

Solving the system yields

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

Out[•]=

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \hat{\mu}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

The maximum likelihood estimate of the mean is again the arithmetic mean. The MLE of the variance is biased because:

Out[•]//TraditionalForm=

$$\mathrm{E}\left(\widehat{\sigma^2}\right) = \left(1 - \frac{1}{n}\right)\sigma^2$$

converging to σ^2 for large n.

MLE and Regression

Consider the regression line, for which we have values of Y_i measured at given values of the independent variable X_i. Our model is

Out[•]//TraditionalForm=

$$y(a,\ b) = a\,x + b$$

and assuming that the Y_i have a Gaussian scatter, each term in the likelihood product is

Out[•]//TraditionalForm=

$$\mathcal{L}_i(y \mid a,\ b) = \exp\left(-\frac{(Y_i - (a\,X_i + b))^2}{2\,\sigma^2}\right)$$

i.e. the residuals are $(Y_i$ – model$)$, and our model has the free parameters $(a,\ b)$. Maximizing the log of the likelihood products then yields

Out[•]=

$$\frac{\partial \sum}{\partial a} = -2 \sum \left(Y_i - a - b\,X_i\right) = 0$$
$$\frac{\partial \sum}{\partial b} = -2 \sum X_i \left(Y_i - a - b\,X_i\right) = 0$$

MLE and Regression

from which two equations in two unknowns we get the well-known

Out[•]= $a = \frac{\sum_{i=1}^{N} Y_i \left(X_i - \overline{X}\right)}{\sum_{i=1}^{N} \left(X_i - \overline{X}^2\right)}$ and $b = \overline{Y} - a\,\overline{X}$

With this simple maximum-likelihood example, we have accidentally derived the standard OLS, the ordinary least squares estimate of *y* on the independent variable *x*.

Bayesian likelihood analysis

Bayes' theorem says, for the model parameter (a vector in general) $\vec{\alpha}$ and data X_i

Out[ ]//TraditionalForm=

$$\mathcal{P}_f(\vec{\alpha} \mid X_i) \propto \mathcal{L}(\vec{\alpha} \mid X_i)\, \mathcal{P}_f(\vec{\alpha})$$

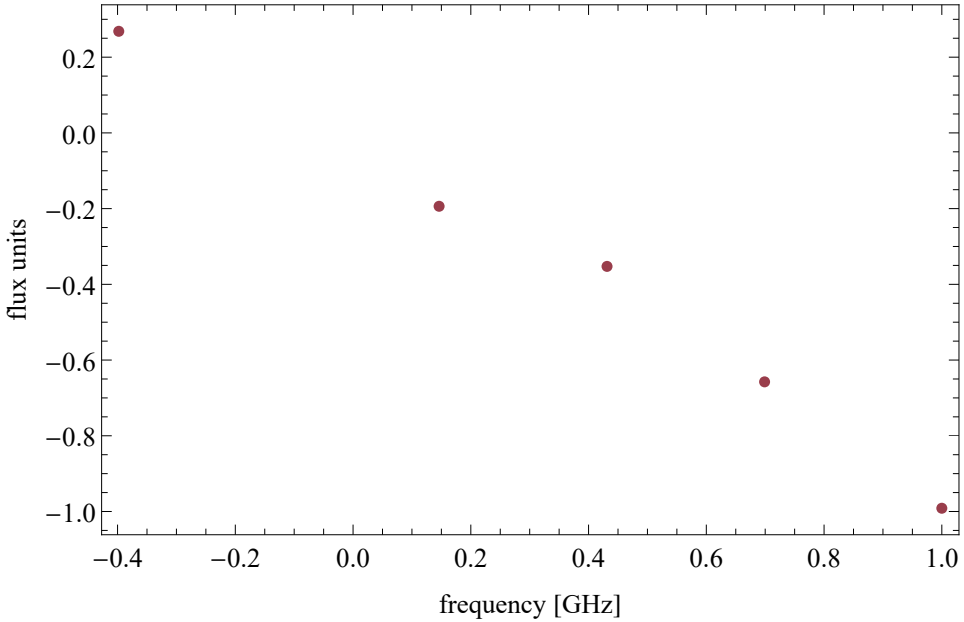
However, given the posterior probability of $\vec{\alpha}$, we may choose to emphasize properties other than the most probable $\vec{\alpha}$, we may only be interested in the probability that it exceeds a certain value, for example.

Two great strengths of the Bayesian approach are the ability to deal with nuisance parameters via marginalization, and the use of the evidence or Bayes factor to choose between models.

Example: Flux density measurements

Let us suppose we have flux density measurements at 0.4,1.4, 2.7, 5 and 10 GHz. The corresponding data are 1.855, 0.640, 0.444, 0.22 and 0.102 flux units

Out[•]=



Let us label the frequencies as f_i and the data as S_i. These follow a power law of slope –1, but have a 10% Gaussian noise added. The noise level is denoted ϵ, and the model for the flux density as a function of frequency is $k\, f^{-\gamma}$. Assuming we know the noise level and distribution, each term in the likelihood product is of the form.

Out[•]//TraditionalForm=

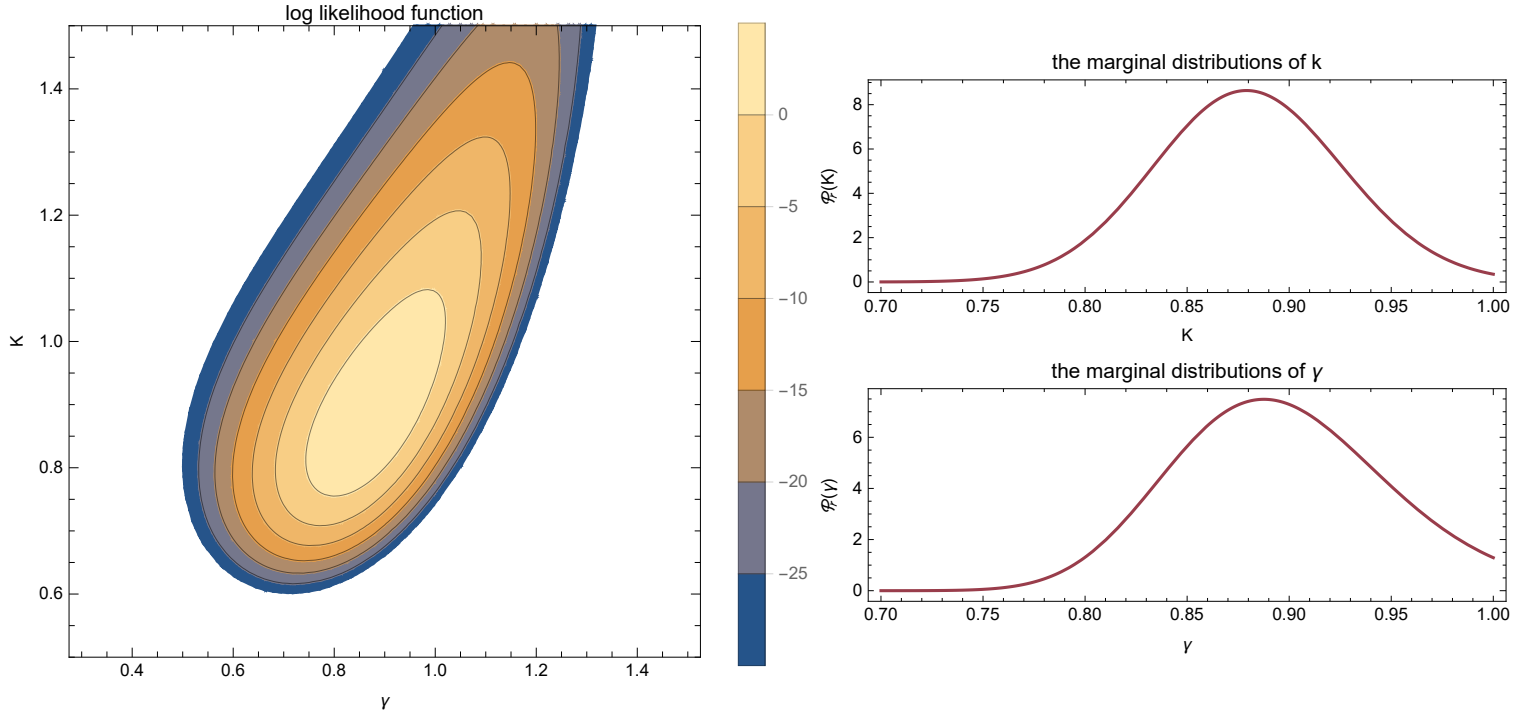
$$\frac{1}{\sqrt{2\pi}\,\epsilon\, k\, f_i^{-\gamma}} \exp\left(-\frac{(S_i - k\, f_i^{-\gamma})^2}{2\,(\epsilon\, k\, f_i^{-\gamma})^2}\right)$$

We are actually fitting a Gaussian to the errors!

Example: Flux density measurements

The likelihood is therefore a function of k and γ. At this point, there are at least two possibilities for further analysis. We may wish to know which pairs of (k, γ) are, say, 90 per cent probable.

Out[ ]=



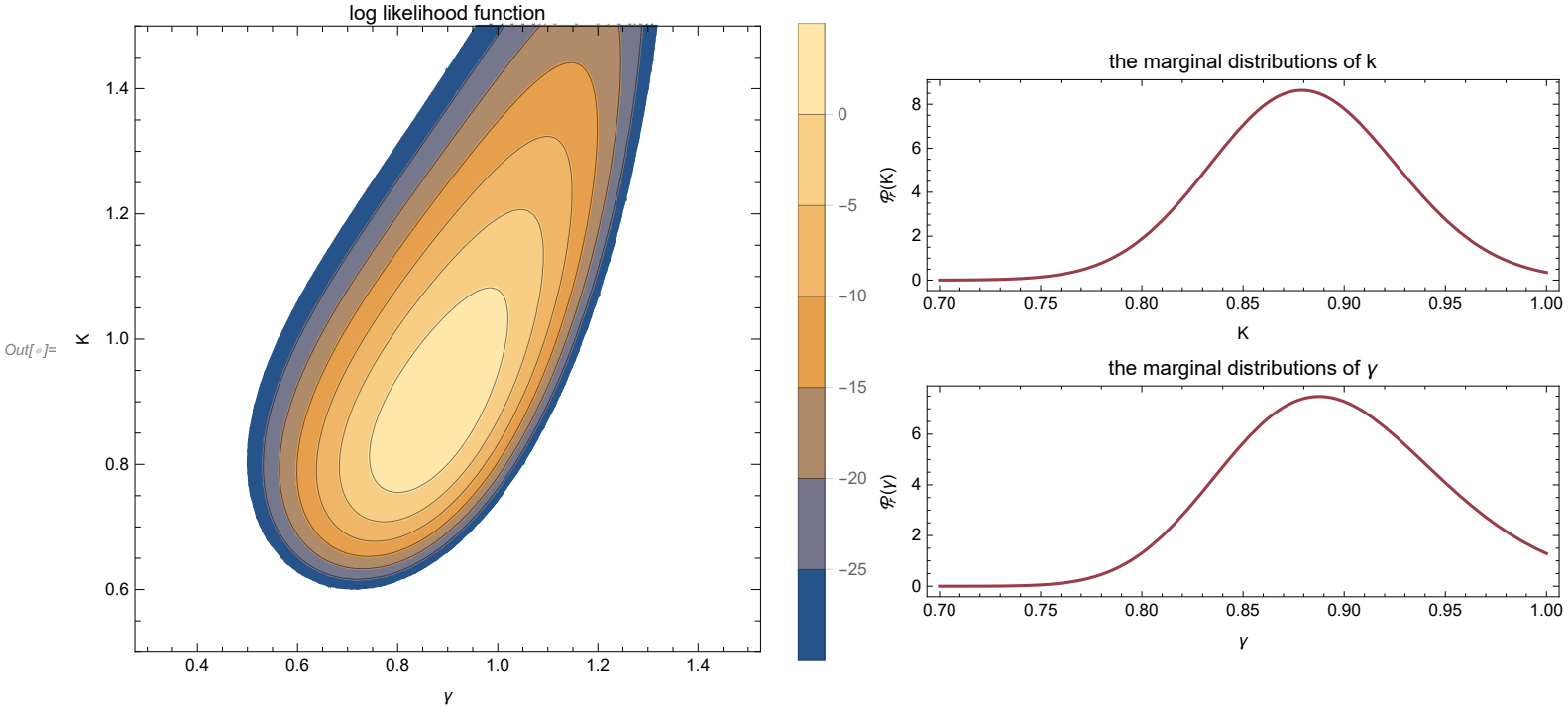
Another possibility is to ask for the probability of, say, k regardless of γ. So we have a posterior probability $\mathcal{P}(k, \gamma \mid S_i)$ and we form

Out[ ]//TraditionalForm=

$$\mathcal{P}(k \mid S_i) = \int \mathcal{P}(k, y \mid S_i)\, d\gamma$$

Example: Flux density measurements

Out[ ]=



The (marginal) probability distributions for *k* and *γ* are also shown in the Figure.

Marginalization can be a very useful technique. Often we are not interested in all the parameters we need to estimate to make a model. If we were investigating radio spectra, for instance, we would want to marginalize out k in our example. We may also have to estimate instrumental parameters as part of our modelling process, but at the end we marginalize them out in order to get answers which do not depend on these parameters.

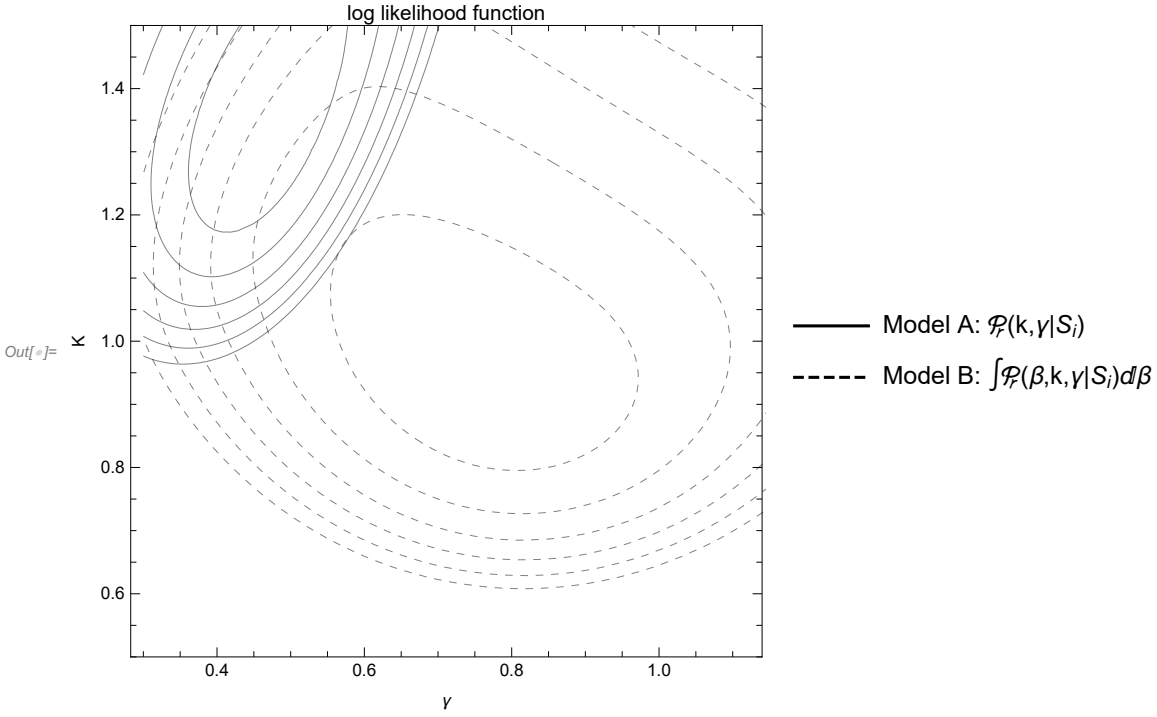
Example: Flux density measurements II

In our radio spectrum example we will add (somewhat artificially) an offset of 0.4 flux units to each measurement. This has the effect of flattening the spectrum quite markedly. We will calculate two possibilities. **Model A** is the simple one we assumed before, with no offsets built in. **Model B** uses a model for the flux densities of the form $\beta + k f^{-\gamma}$. Each likelihood term is then

Out[•]//TraditionalForm=

$$\frac{1}{\sqrt{2\pi}\,\epsilon\, k\, f_i^{-\gamma}} \exp\left(-\frac{(S_i - (\beta + k\, f_i^{-\gamma}))^2}{2\,(\epsilon\, k\, f_i^{-\gamma})^2}\right)$$

Out[•]=

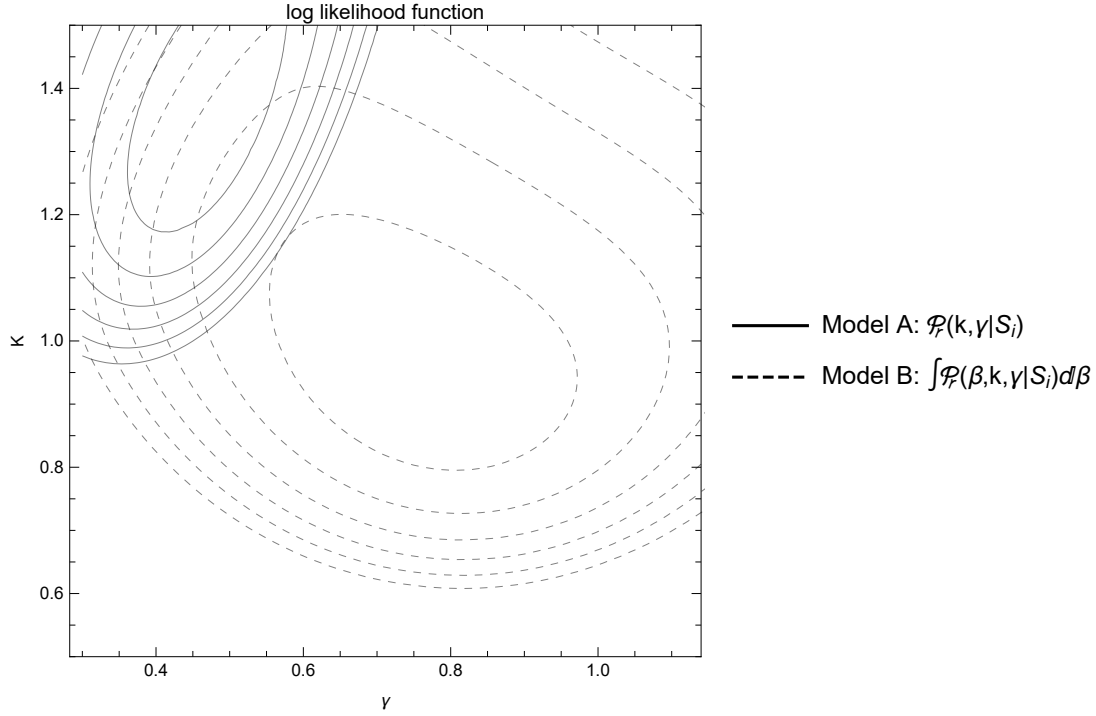


Example: Flux density measurements II

We also suppose that we have some suspicion of the existence of this offset, so we place a prior on β of mean 0.4, standard deviation ϵ. **Model B** therefore returns a posterior distribution for k, γ and β. We are not actually interested in β (although an instrumental scientist might be) so we marginalize it out. The likelihoods from the two models are shown in the Fig. above, and it is clear that the more complex model does a better job of recovering the true parameters.

In the real world, of course, we do not have the truth available to guide us as to our choice of **model A** or **model B**.

Out[ ]=



Decide between models

Suppose we are choosing between **model A** and **model B** and we believe they are the only possibilities. The prior probability of A is, say, p_A and of B is p_B. The posterior probability of the parameters α, given data X_i, is

Out[ ]//TraditionalForm=

$$\mathcal{P}_r(\alpha \mid X_i, A, B) = \frac{p_A \mathcal{L}(X_i \mid \alpha, A) \mathcal{P}_r(\alpha \mid A) + p_B \mathcal{L}(X_i \mid \alpha, B) \mathcal{P}_r(\alpha \mid B)}{\mathcal{P}_r(X_i)}$$

where we are emphasizing which model enters the various likelihoods. $\mathcal{P}_r(X_i)$ is the normalizing factor which ensures that the posterior distribution is properly normalized; its calculation usually involves a multidimensional integral (see example). $\mathcal{P}_r(\alpha|\mathrm{A})$ is the prior on α in **model A**, similarly for **B**.

Decide between models

The posterior odds on **model A**, compared to **model B**, are then simply

Out[•]//TraditionalForm=

$$\frac{\int_\alpha p_A \, \mathcal{L}(X_i \mid \alpha, A) \, \mathcal{P}r(\alpha \mid A)}{\int_\alpha p_B \, \mathcal{L}(X_i \mid \alpha, B) \, \mathcal{P}r(\alpha \mid B)}$$

in which we have to integrate over the range of parameters appropriate to each model. This is worth the effort because we get a straightforward answer to the question: which of **A** or **B** would it be better to bet on?

Example: Decide between models

In the previous two examples we have worked out the likelihood functions, which we abbreviate $\mathcal{L}(X_i \mid k, \gamma, A)$ for **model A** and similarly for **model B**. In **model B** we also have a prior on the offset β, which is

Out[•]//TraditionalForm=

$$\mathcal{P}_I(\beta \mid B) = \frac{1}{\sqrt{2\,\pi}\,\epsilon} \exp\left(-\frac{(\beta - 0.4)^2}{2\,\epsilon^2}\right)$$

We then form the ratio of the integrals

Out[•]//TraditionalForm=

$$p_A \int dk \int d\gamma\, \mathcal{L}(X_i \mid k, \gamma, A)$$

and

Out[•]//TraditionalForm=

$$p_B \int dk \int d\gamma \int d\beta\, \mathcal{L}(X_i \mid k, \gamma, B)\, \mathcal{P}_I(\beta \mid B)$$

Example: Decide between models

Let's take $p_A = p_B$, an agnostic prior state; note we have implicitly assumed uniform priors on k and γ. Performing the integrations, we get odds on **B** compared to **A**:

In[•]:=
```
norm2
-----
norm
```

Out[•]=
```
8.55786
```

Another way of looking at this is that we would have had to have been prepared to offer prior odds of 8:1 against the existence of the offset, for the posterior odds to have been even.

Error estimation for MLE

Errors or uncertainties in the parameter estimation with ML can only be given under certain circumstances. Generally we need the full covariance matrix.

Direct Method

The direct method gives the variance of the parameter estimates $\hat{\alpha} = \hat{\alpha}(x_1, \ldots, x_n)$ if we have many measurements with samples $(x_1, \ldots, x_n)$:

Out[ ]//TraditionalForm=

$$V_{ij}(\alpha) = \int (\hat{\alpha}_i - \alpha_i)\,(\hat{\alpha}_j - \alpha_j)\,\mathcal{L}(x_1, \ldots, x_n \mid \alpha)\,dx_1 \ldots dx_n$$

here, $\alpha = (\alpha_1, \ldots, \alpha_m)$ is the 'true' set of parameters and $\hat{\alpha}(x_1, \ldots, x_n)$ are the estimates for one sample. The samples, over which we integrate follow the probability density $\mathcal{L}(x_1, \ldots, x_n)$.

This method requires the knowledge of the true parameter set α and $\mathcal{L}(x_1, \ldots, x_n \mid \alpha)$. After an experiment, one usually does not know either.

Error estimation for MLE

Practical Method

Practically, we take $\mathcal{L}(x_1, \ldots, x_n \mid \alpha)$ for a fixed sample $(x_1, \ldots, x_n)$ as probability density for α. Then we find for the variance matrix

Out[•]//TraditionalForm=

$$V_{ij}(\alpha) = \frac{\int (\alpha_i - \hat{\alpha}_i)\,(\alpha_j - \hat{\alpha}_j)\,\mathcal{L}(x_1, \ldots, x_n \mid \alpha)\,dx_1 \ldots dx_n}{\int \mathcal{L}(x_1, \ldots, x_n \mid \alpha)\,dx_1 \ldots dx_n}$$

Here, $\hat{\alpha}$ is the MLE, that was estimated from the measured sample $(x_1, \ldots, x_n)$. The denominator is for normalization.

Error estimation for MLE

Series expansion around Max

The log-likelihood function can be expanded around $\alpha = \hat{\alpha}$:

Out[•]//TraditionalForm=

$$\log\mathcal{L}(x_1, \ldots, x_n \mid \alpha) = \log\mathcal{L}\left(x_1, \ldots, x_n \mid \widehat{\alpha}\right) + \left(\alpha - \widehat{\alpha}\right) \frac{\partial \log\mathcal{L}}{\partial \alpha}\Big|_{\alpha=\hat{\alpha}} + \frac{1}{2}\left(\alpha_i - \widehat{\alpha_i}\right)\left(\alpha_j - \widehat{\alpha_j}\right) \frac{\partial^2 \log\mathcal{L}}{\partial \alpha_i \, \partial \alpha_j}\Big|_{\alpha=\hat{\alpha}} + \ldots$$

The first derivative vanishes (maximum condition). The second derivatives are collected in

Out[•]//TraditionalForm=

$$V_{ij}^{-1} = -\frac{\partial^2 \log\mathcal{L}}{\partial \alpha_i \, \partial \alpha_j}\Big|_{\alpha=\hat{\alpha}}$$

Series expansion around Max

Hence, we find in the vicinity of the maximum

Out[•]//TraditionalForm=

$$\log\mathcal{L}(x_1, \ldots, x_n \mid \alpha) = \log\mathcal{L}_{\max} - \frac{1}{2}\left(\alpha - \hat{\alpha}\right)^{\mathsf{T}} V^{-1} \left(\alpha - \hat{\alpha}\right)$$

and for the likelihood function

Out[•]//TraditionalForm=

$$\mathcal{L}(x_1, \ldots, x_n \mid \alpha) = \mathcal{L}_{\max} \exp\left(-\frac{1}{2}\left(\alpha - \hat{\alpha}\right)^{\mathsf{T}} V^{-1} \left(\alpha - \hat{\alpha}\right)\right)$$

Hence, if the likelihood function is approximately Gaussian, we can estimate variance using the second derivatives. Practically it is often assumed that the likelihood function follows a multi-normal distribution. In the case of uncorrelated parameters, V^{-1} is diagonal and the parameter variance is

Out[•]//TraditionalForm=

$$\sigma_i^2 = \frac{1}{V_{\text{ii}}^2} = \left(-\frac{\partial^2 \log\mathcal{L}}{\partial\alpha_i\,\partial\alpha_j}\Big|_{\alpha=\hat{\alpha}}\right)^{-1}$$

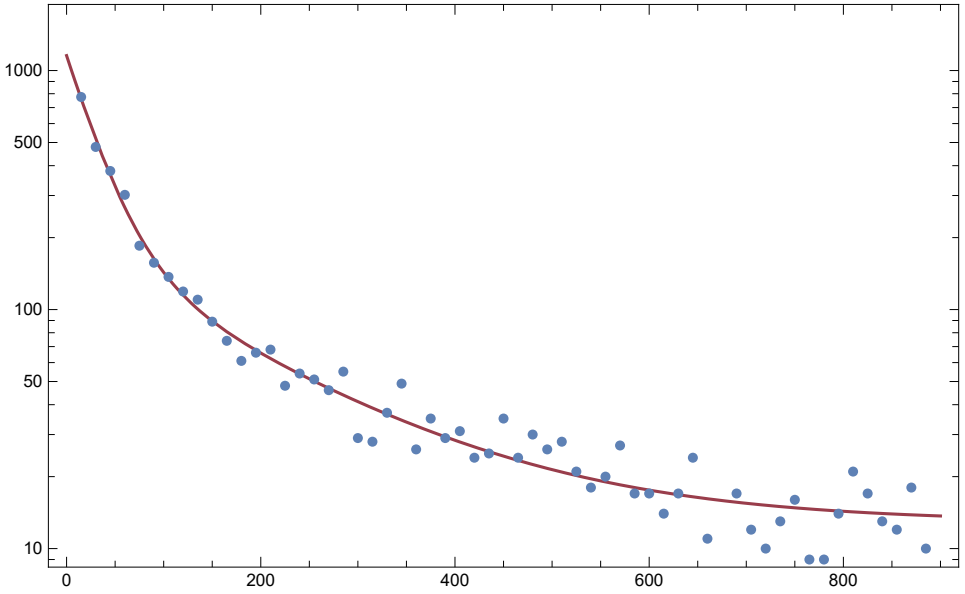
Chi-square minimization

The method of least squares and multiple regression are restricted to fitting function that are linear in the parameters. This limitation is imposed by the fact that, in general, minimizing χ^2 can yield a set of coupled equations that are linear in the m unknown parameters only if the fitting functions $y(x)$ are themselves linear in the parameters. If this is not the case we need to perform a *nonlinear fitting*.

Example

Measurements of beta decay of short lived silver isotopes $_{47}Ag^{108}$ and $_{47}Ag^{110}$ measured in 15-s intervals. The data points do not fall on a straight line because the probability function that describes the process is a sum of two exponential functions plus a constant background:

$$y(x_i) = a_1 + a_2 \exp(-t / a_4) + a_3 \exp(-t / a_5)$$



a1 → 12.9302
a2 → 970.319
a3 → 171.646
a4 → 30.3899
a5 → 166.128

Series expansion around Max

We can generalize the probability function, or likelihood-function, to any number of parameters,

Out[ ]//TraditionalForm=

$$\mathcal{P}_r(a_1, a_2, \ldots, a_m) = \prod \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{1}{2} \sum \left(\frac{y_i - y(x_i)}{\sigma_i}\right)^2\right)$$

Chi-square minimization

and maximize the likelihood with respect to the parameters by minimizing the exponent, the **goodness-of-fit parameter χ^2 :**

Out[ ]//TraditionalForm=

$$\chi^2 = \Sigma\left(\frac{y_i - y\,(x_i)}{\sigma_i}\right)^2$$

where x_i and y_i are the measured variables, σ_i is the uncertainty in y_i and $y(x_i)$ are values of the function calculated at x_i. The optimum value of the parameters a_j are obtained by minimizing χ^2 simultaneously

Chi-square minimization

Out[ ]//TraditionalForm=

$$\frac{\partial \chi^2}{\partial a_j} = \frac{\partial}{\partial a_j} \sum \left(\frac{y_i - y(x_i)}{\sigma_{\text{i}}} \right)^2 = 0$$

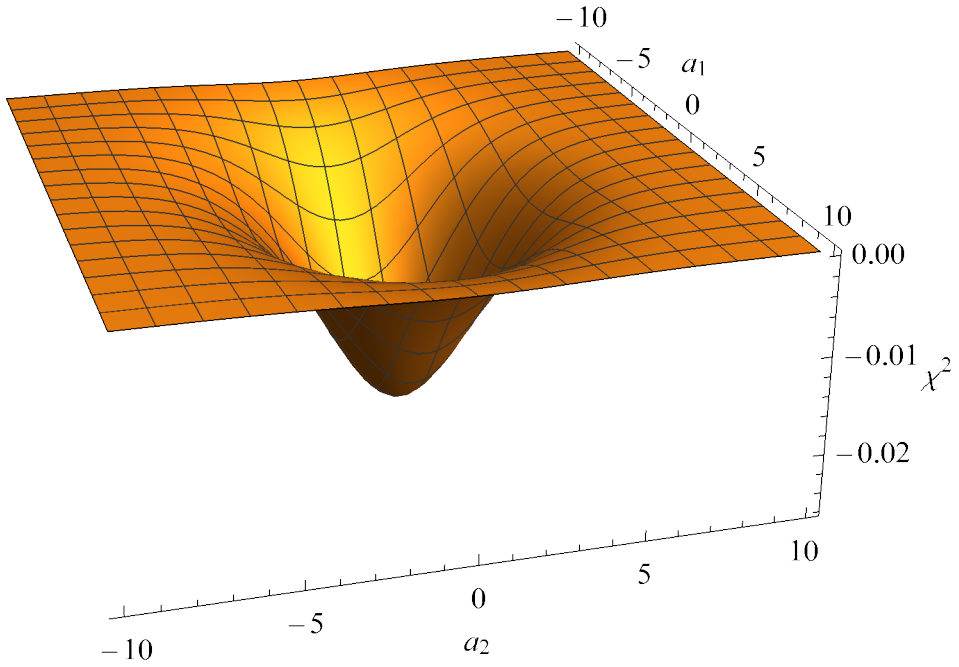
Out[ ]//TraditionalForm=

$$= 2 \sum \left(\frac{1}{\sigma_i^2} (y_i - y(x_i)) \frac{\partial y(x_i)}{\partial a_j} \right)$$

This will yield m coupled equations in the m unknown parameters a_j. If these equations are not linear in all the parameters, we must, in general, treat χ^2 as a continuous function of the m parameters, describing a hypersurface in an m-dimensional space, and search that space for the appropriate minimum.

Chi-square minimization

Out[ ]=



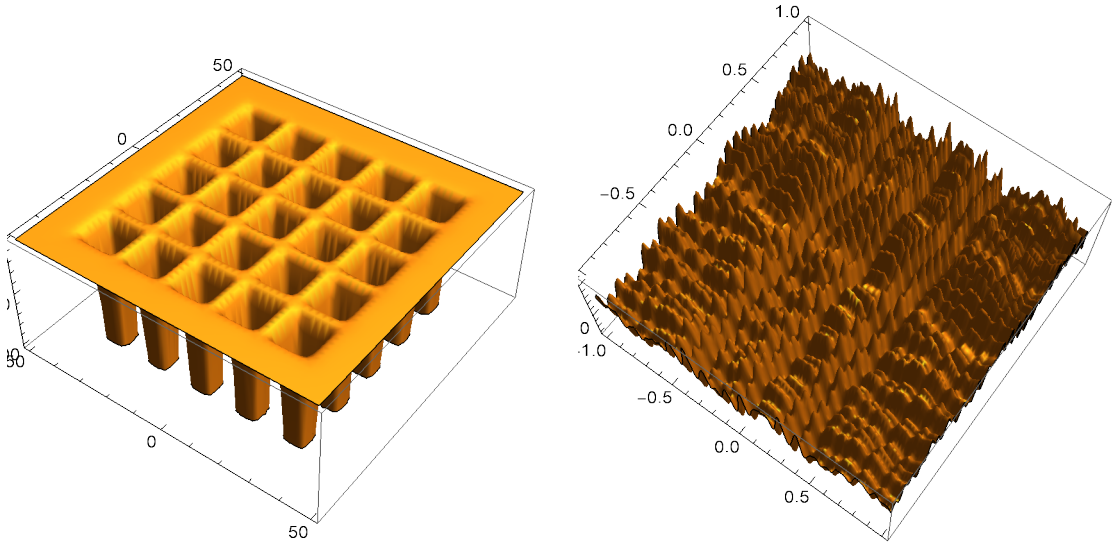
Several strategies exist to minimize χ^2:

- brute force: scanning of the parameter space, Monte Carlo, ...
- grid-search method
- series expansion of χ^2 near minimum
- gradient search methods
- more sophisticated algorithms (e.g. Levenberg-Marquardt, ...)

Chi-square minimization

There are some important points to keep in mind when minimizing nonlinear functions.

- **Starting values:** Most of the methods require a starting value for the a_i. Because of the nonlinear nature of the problem, an unfortunate choice of initial values can slow down numerical convergence to the minimum or even make it impossible.
- **Multiple solutions/Local Minima:** For an arbitrary function there may be more than one minimum of the χ^2 function. A 'bad' starting value may drive the algorithm into one of these side-minima. There are algorithms less vulnerable to becoming stuck in local minima (e.g. simulated annealing).



Chi-square minimization

- **Bounds on parameters:** From a particular set of starting values, the search may converge toward solutions that are physically unreasonable: In the example above, negative values for the parameters are not acceptable. In addition, a current trial value of one parameter may the limit the possible determination of other parameters. For example, if a_2 becomes very small (or 0), a_4 cannot be determined at all. Placing limits on the range of possible parameter values may prevent this. Care should be taken that the final value of any parameter is not one of these artificial values!

Chi-square minimization

Brute force

One possibility is to simply scan the parameter space within the given limits. This is feasible if you have some prior knowledge about the range of viable parameter values and about the magnitude of the $\partial \chi / \partial a_i$. Of course, the gridding in a_i needs to be fine enough to capture significant variation in χ^2.

If χ^2 is changing only slowly with the a_i the danger of missing any minimum is small.

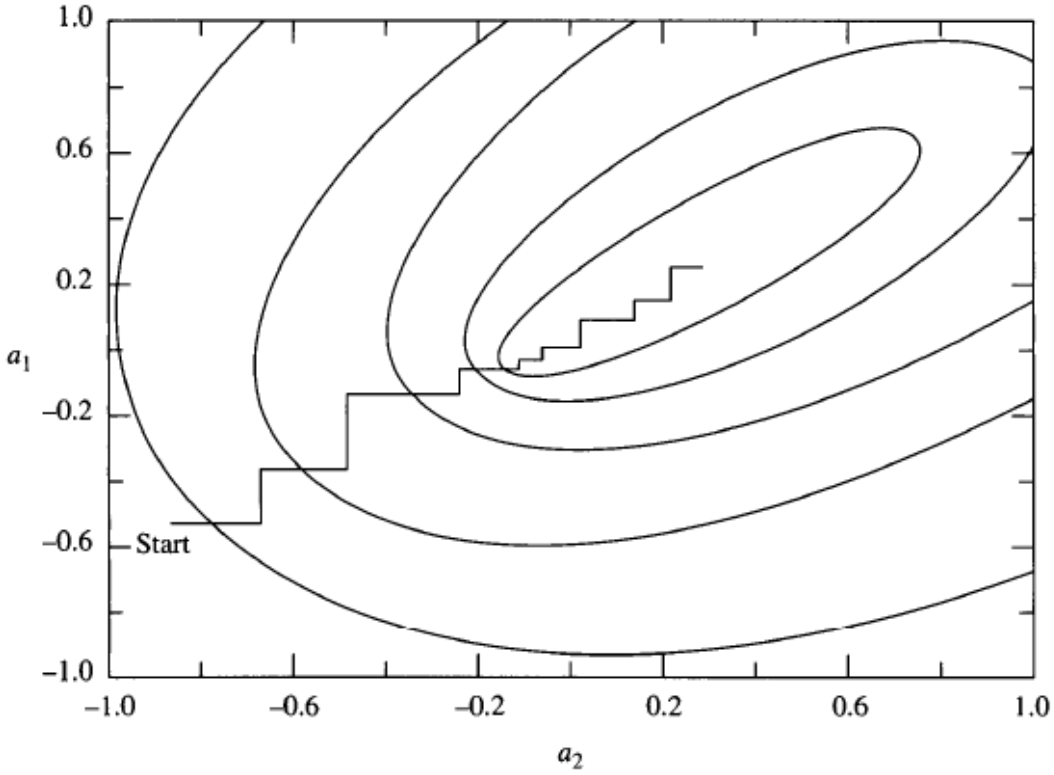
This approach is progressively less efficient with increasing number of parameters.

A typical application is a χ^2 function with model values $y(x_i)$ being the result of complex, time-consuming computations. In this case it is common to compute the model on a fixed parameter grid and tabulate them. Using these tabulated results, the χ^2 can be easily calculated for any given set of measured data. The accuracy of the minimum depends on the granularity of the underlying parameter grid.

Monte-Carlo minimization is also an application of the brute force approach. The χ^2 hypersurface is computed at a large set of random parameters.

Chi-square minimization

Grid-Search Method



If the variation of χ^2 with each parameter a_j is not very sensitive to the values of the other parameters, then the optimum values can be obtained most simply by minimizing χ^2 w.r.t. each of the a_j separately:

Chi-square minimization

Grid-Search Method

1. Select starting values a_j and step/increment size Δa_j for each parameter and calculate χ^2 with the starting parameters.
2. Increment one parameter a_j by $\pm \Delta a_j$ and calculate χ^2, where the sign is chosen so that χ^2 decreases.
3. Repeat step 2 until χ^2 stops decreasing and begins to increase. The increase in χ^2 indicates that the search has crossed a ravine and started up the other side.
4. Use the last three values of a_j (which bracket the minimum) and the associated values of χ^2 to determine the minimum of the parabola, which passes through the three points.
5. Repeat to minimize χ^2 w.r.t. each parameter in turn.
6. Continue to repeat the procedure until the last iteration yields a predefined negligibly small decrease in χ^2.

If the variations of χ^2 with the parameters are strongly correlated, then the approach to the minimum may be very slow.

Chi-square minimization

Gradient-Search Method - Steepest descent

The search could be improved if the zigzag path in the grid-search example would have been replaced by a more direct vector toward the appropriate minimum. In the gradient-search method of least squares, all the parameters a_j are changed simultaneously, such that the direction of travel is along the gradient of χ^2.

Out[ ]//TraditionalForm= $$\nabla \chi^2 = \sum_{j=1}^{n} \frac{\partial \chi^2}{\partial a_j} \hat{a}_j$$

where $\hat{a}_j$ indicates a unit vector in the direction of the a_j coordinate axis. Estimating the partial derivatives numerically gives

Out[ ]//TraditionalForm= $$\nabla \chi^2{}_j = \frac{\partial \chi^2}{\partial a_j} \simeq \frac{\chi^2(a_j + f\,\Delta a_j) - \chi^2(a_j)}{f\,\Delta a_j}$$

where f is the fraction of the step size Δa_j by which a_j in order to determine the derivative.

Chi-square minimization

Expansion Method

We expand χ^2 to second order in the parameters about a local minimum χ_0^2 where $a_j = a_j^{'}$. This approximates the χ^2 hypersurface by a parabolic surface.

Out[•]//TraditionalForm=

$$\chi^2 \simeq \chi_0^2 + \sum_{j=1}^{m} \frac{\partial \chi_0^2}{\partial a_j} \delta \mathrm{a}_j + \frac{1}{2} \sum_{k=1}^{m} \sum_{j=1}^{m} \frac{\partial^2 \chi_0^2}{\partial a_j \, \partial a_k} \delta \mathrm{a}_j \delta \mathrm{a}_k$$

Here we define $\delta \mathrm{a}_j = a_j - a_j^{'}$, and χ_0^2 is given by

Out[•]//TraditionalForm=

$$\chi_0^2 = \Sigma \left(\frac{1}{\sigma_i^2} (y_i - \mathrm{y}'[x_i])^2 \right)$$

Expansion Method

where $y'(x_i)$ is the value of the function when $\delta a_j = 0$. Minimising χ^2 w.r.t. the increment δa_j leads to

Out[•]= $\frac{\partial \chi^2}{\partial (\delta a_k)} = \frac{\partial \chi_0^2}{\partial a_k} + \sum_{j=1}^{m} \frac{\partial^2 \chi_0^2}{\partial a_j \partial a_k} \delta a_j = 0 \qquad k=1,\ldots,m$

The result is a set of m linear equations in δa_j that we can write as

Out[•]= $\beta_k - \sum_{j=1}^{m} \delta a_j \alpha_{jk} = 0 \qquad k=1,\ldots,m$

with

Out[•]= $\beta_k = -\frac{1}{2} \frac{\partial \chi_0^2}{\partial a_k}$ and $\alpha_{jk} = \frac{1}{2} \frac{\partial^2 \chi_0^2}{\partial a_j \partial a_k}$

Expansion Method

(The factors ±1/2 are included for agreement with the conventional definitions of these quantities). We can treat the equations as a matrix equation:

Out[•]//TraditionalForm=

$$\beta = \delta \mathrm{a}\, \alpha$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\delta}\mathbf{a}$ are row matrices and $\boldsymbol{\alpha}$ is a symmetric matrix of order m (curvature matrix). The $\delta\alpha$ can be obtained by matrix inversion.

We see from the equations above that the calculation of the elements of β and α requires the knowledge of the first and second derivatives of χ^2. If the analytic derivatives are not available, or difficult to compute, then they can be approximated by the method of finite differences. Using forward differences we find:

Out[•]//TraditionalForm=

$$\frac{\partial \chi_0^2}{\partial a_j} \simeq \frac{\chi_0^2(a_j + \Delta \mathrm{a}_j,\, a_k) - \chi_0^2(a_j,\, a_k)}{\Delta \mathrm{a}_j}$$

Expansion Method

Out[ ]//TraditionalForm=

$$\frac{\partial^2 \chi_0^2}{\partial^2 a_j} \simeq \frac{4\left(\chi_0^2(a_j,\, a_k) - 2\,\chi_0^2\left(a_j + \frac{\Delta a_j}{2},\, a_k\right) + \chi_0^2(a_j + \Delta a_j,\, a_k)\right)}{(\Delta a_j)^2}$$

Out[ ]//TraditionalForm=

$$\frac{\partial^2 \chi_0^2}{\partial a_j\, \partial a_k} \simeq \frac{1}{\Delta a_j\, \Delta a_k}\left(\chi_0^2(a_j,\, a_k) - \chi_0^2(a_j + \Delta a_j,\, a_k) - \chi_0^2(a_j,\, a_k + \Delta a_k) + \chi_0^2(a_j + \Delta a_j,\, a_k + \Delta a_k)\right)$$

The intervals Δa_j should be chosen to be large enough to avoid roundoff errors but small enough to furnish reasonably accurate values of the derivatives near the minimum.

Chi-square minimization

Levenberg-Marquardt Method

A disadvantage of the expansion method is that even though it converges rapidly to the minimum from points nearby (region of trust) it cannot be relied on to approach with any accuracy from points outside the region where the χ^2 hypersurface is approximately parabolic.

In contrast, the gradient search is ideally suited for approaching the minimum from far away, but does not converge rapidly near the minimum.

The method by Marquardt (1963) and Levenberg (1944) combines the best features of the two approaches.

Out[•]= $\beta = \delta \mathbf{a}\, \alpha'$ with $\left(\alpha'_{jk} = \begin{cases} \alpha_{jk}\,(1+\lambda) & j = k \\ \alpha_{jk} & j \neq k \end{cases}\right)$

Levenberg-Marquardt Method

If λ is very small, the equation above is similar to the solution developed from the Taylor expansion. If λ is very large, the diagonal terms of the curvature matrix dominate and the matrix equation degenerates into m separate equations

Out[ ]//TraditionalForm=

$$\beta_j \simeq \lambda \, \delta \mathrm{a}_j \, \alpha_{\mathrm{jj}}$$

which yield the vector increment δa in the same direction as the vector β.

The initial value of the constant factor should be chosen small enough to take advantage of the analytical solution, but large enough that χ^2 decreases.

Chi-square minimization

Levenberg-Marquardt Method

The recipe given by Marquardt is:

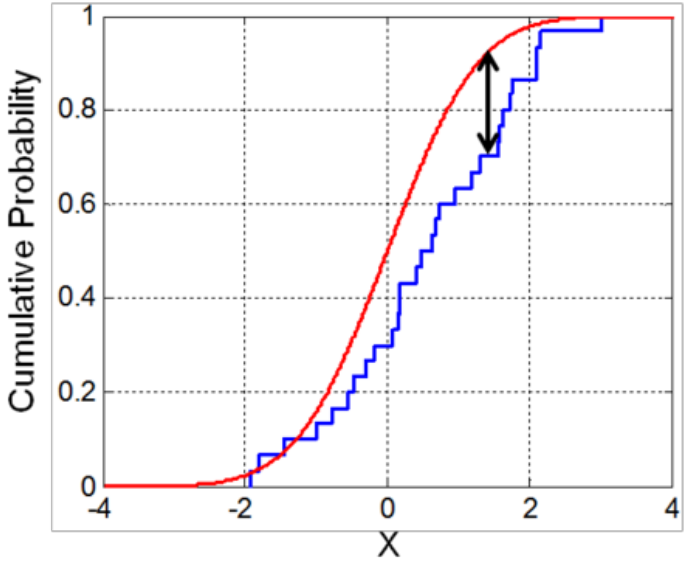
1. Compute $\chi^2(a)$
2. Start initially with λ=0.001
3. Compute δa and $\chi^2(a + \delta a)$ with this choice of λ
4. If $\chi^2(a + \delta a) > \chi^2(a)$ increase λ by a factor of 10 and repeat step 3.
5. If $\chi^2(a + \delta a) < \chi^2(a)$ decrease λ by a factor of 10, consider $a' = a + \delta a$ to be the new starting point, and return to step 3, substituting a' for a.

At the conclusion of the search, the inverse ϵ of the final value of the curvature matrix α is treated as the error matrix, and the errors in the parameters are obtained from the square roots of the diagonal terms.

Kolmogorov-Smirnov Test

The Kolmogorov–Smirnov test (K–S test or KS test) is a nonparametric test of the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution (one-sample K–S test), or to compare two samples (two-sample K–S test).

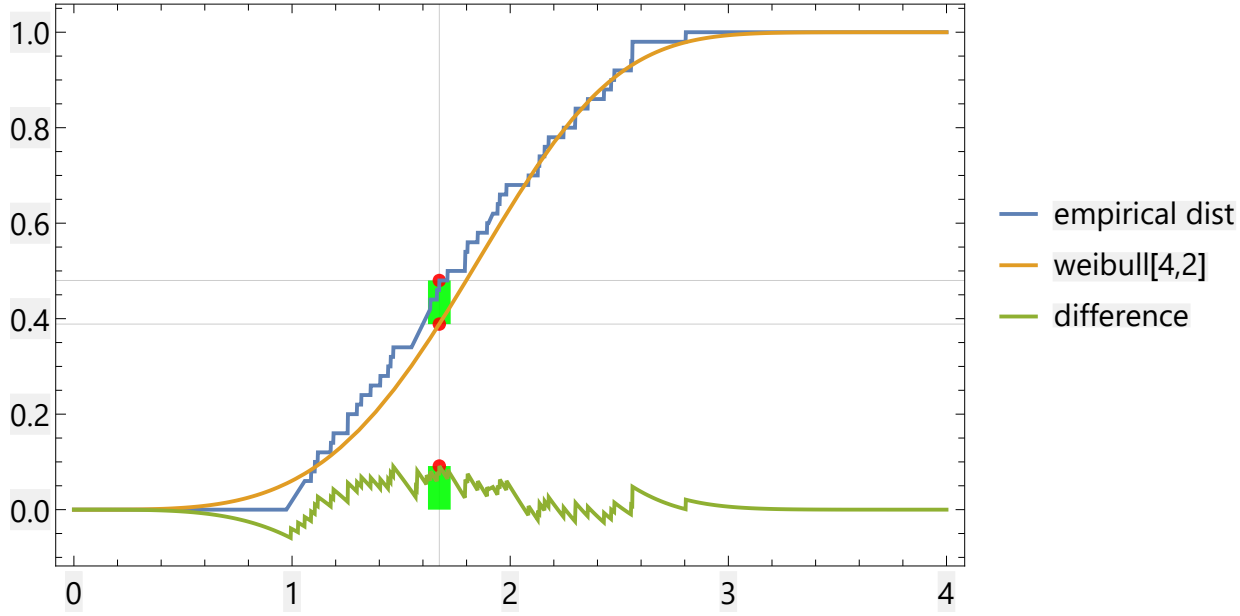
The Kolmogorov–Smirnov statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution, or between the empirical distribution functions of two samples.



Kolmogorov-Smirnov Test

	Statistic	P-Value
Kolmogorov-Smirnov	0.091354	0.763993

Out[ ]=



Empirical Distribution Function

In statistics, the empirical distribution function, or empirical cdf, is the cumulative distribution function associated with the empirical measure of the sample. This cdf is a step function that jumps up by 1/n at each of the n data points.

Out[ ]//TraditionalForm=

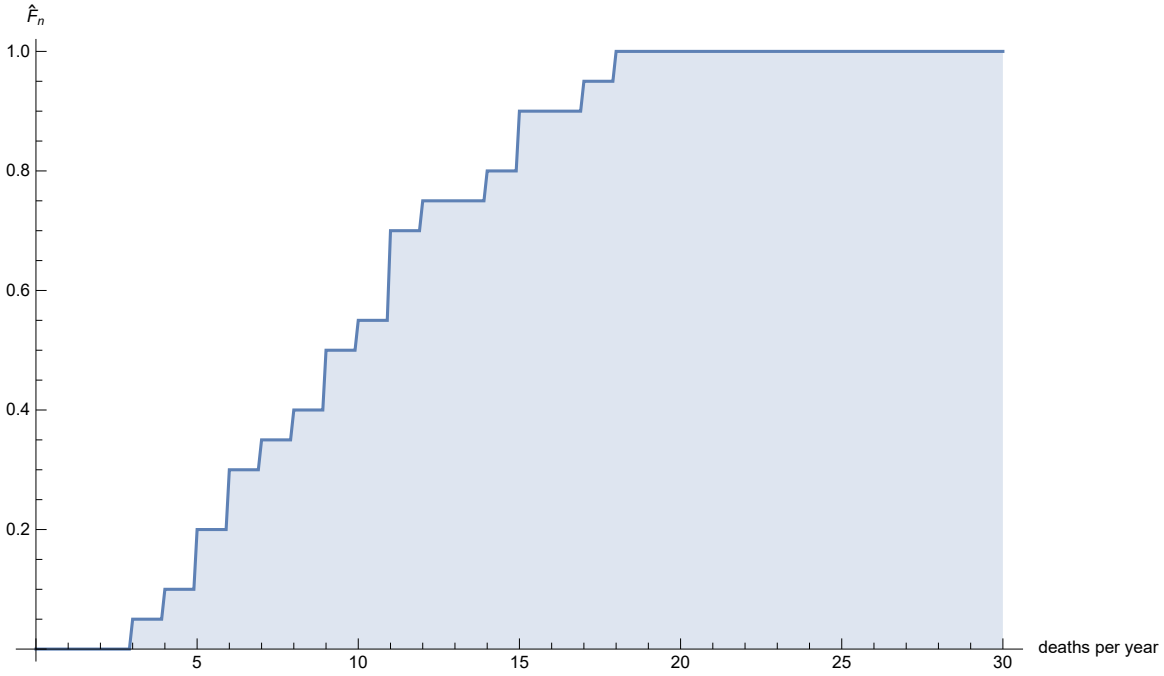
$$\widehat{F}_n(t) = \frac{\text{number of elements in the sample} \leq \text{t}}{n} = \frac{1}{n}\sum_{i=1}^{n} 1\,\{x_i \leq \text{t}\}$$

where $1\{A\}$ is the indicator of event A. For large n $\hat{F}_n(t)$ converges to $F(t)$, the true underlying CDF.

Example:

Between 1875 and 1894, 196 Prussian soldiers from 14 cavalry regiments died after being kicked by horses:

75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94
3	5	7	9	10	18	6	14	11	9	5	11	15	6	11	17	12	15	8	4



Kolmogorov-Smirnov Test

The Kolmogorov–Smirnov statistic for a given theoretical cumulative distribution function F(t) is

Out[•]//TraditionalForm=

$$D_n = \sup_t \left|\widehat{F}_n(t) - F(t)\right|$$

where $\sup_t$ is the supremum of the set of distances.

The goodness-of-fit test or the Kolmogorov–Smirnov test is constructed by using the critical values of the Kolmogorov distribution. The null hypothesis is rejected at level α if

Out[•]= $\sqrt{n}\ D_n > K_\alpha$ with $\mathbb{P}(K \leq K_\alpha) = 1 - \alpha$

Kolmogorov-Smirnov Test

using the CDF of of the Kolmogorov distribution

Out[•]//TraditionalForm=

$$\mathcal{P}_r(K \le t) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}\exp\left(-2\,k^2\,t^2\right) = \frac{\sqrt{2\,\pi}}{t}\sum_{k=1}^{\infty}\exp\left(-\frac{(2\,k-1)^2\,\pi^2}{\left(8\,t^2\right)}\right)$$

The critical value for large n is

Out[•]//TraditionalForm=

$$K_{\alpha}^{\text{crit}} > \frac{1}{\sqrt{n}}\sqrt{\left(-\frac{1}{2}\ln\left[\frac{\alpha}{2}\right]\right)}$$

A table with the critical values of D_n can be found on the lecture website or at http://www.mathematik.uni-kl.de/~schwaar/Exercises/Tabellen/table_kolmogorov.pdf

Kolmogorov-Smirnov Test

The Kolmogorov–Smirnov test may also be used to test whether two underlying one-dimensional probability distributions differ. In this case, the Kolmogorov–Smirnov statistic is

Out[ ]//TraditionalForm=

$$D_{n,n'} = \sup_t \left| \widehat{F}_{1,n}(t) - \widehat{F}_{2,n'}(t) \right|$$

where $\hat{F}_{1,n}$ and $\hat{F}_{2,n'}$ are the empirical distribution functions of the first and the second sample respectively. The null hypothesis is rejected at level α if

Out[ ]//TraditionalForm=

$$D_n > c(\alpha) \sqrt{\frac{n + n'}{n\, n'}}$$

Kolmogorov-Smirnov Test

The values of $c(\alpha)$ are given in the table below for each level of α:

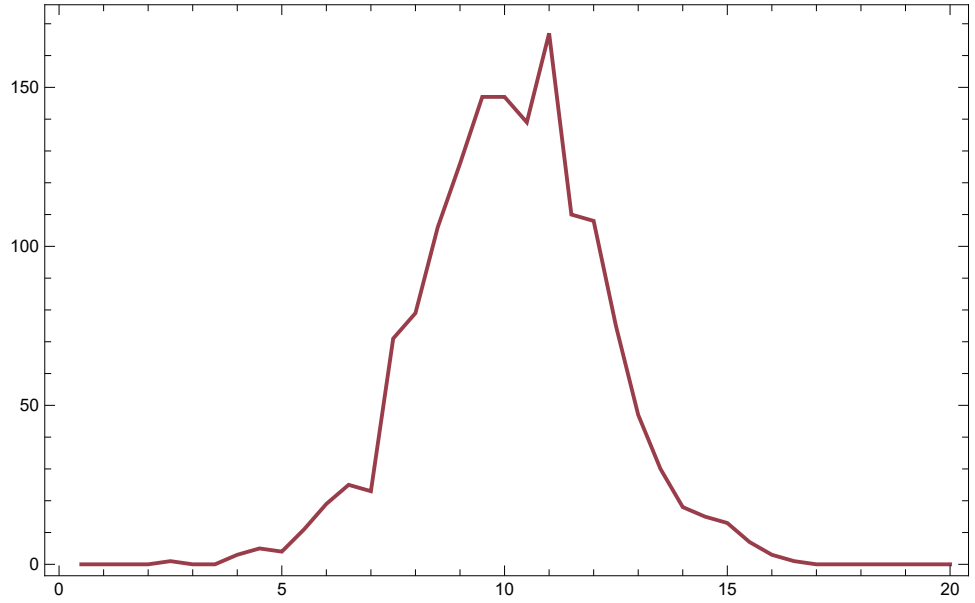
α	0.10	0.05	0.025	0.01	0.005	0.001
c (α)	1.22	1.36	1.48	1.63	1.73	1.95

Note that the two-sample test checks whether the two data samples come from the same distribution. This does not specify what that common distribution is (e.g. whether it's normal or not normal).

Example: Test for normality

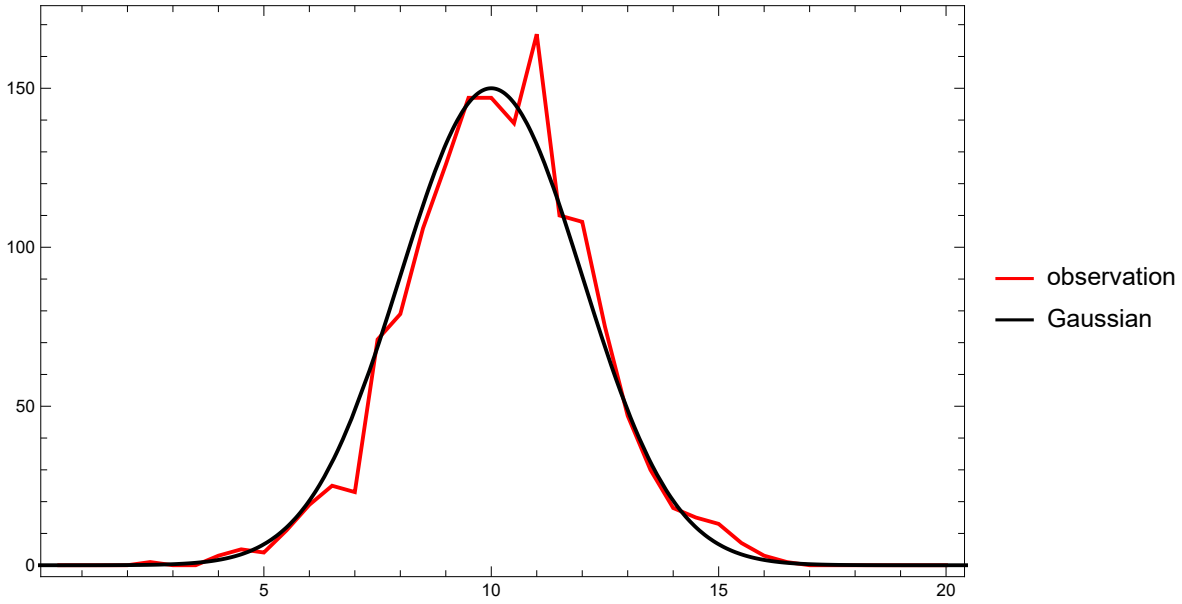
Observing a spectral emission line gives the following data:

Out[ ]=



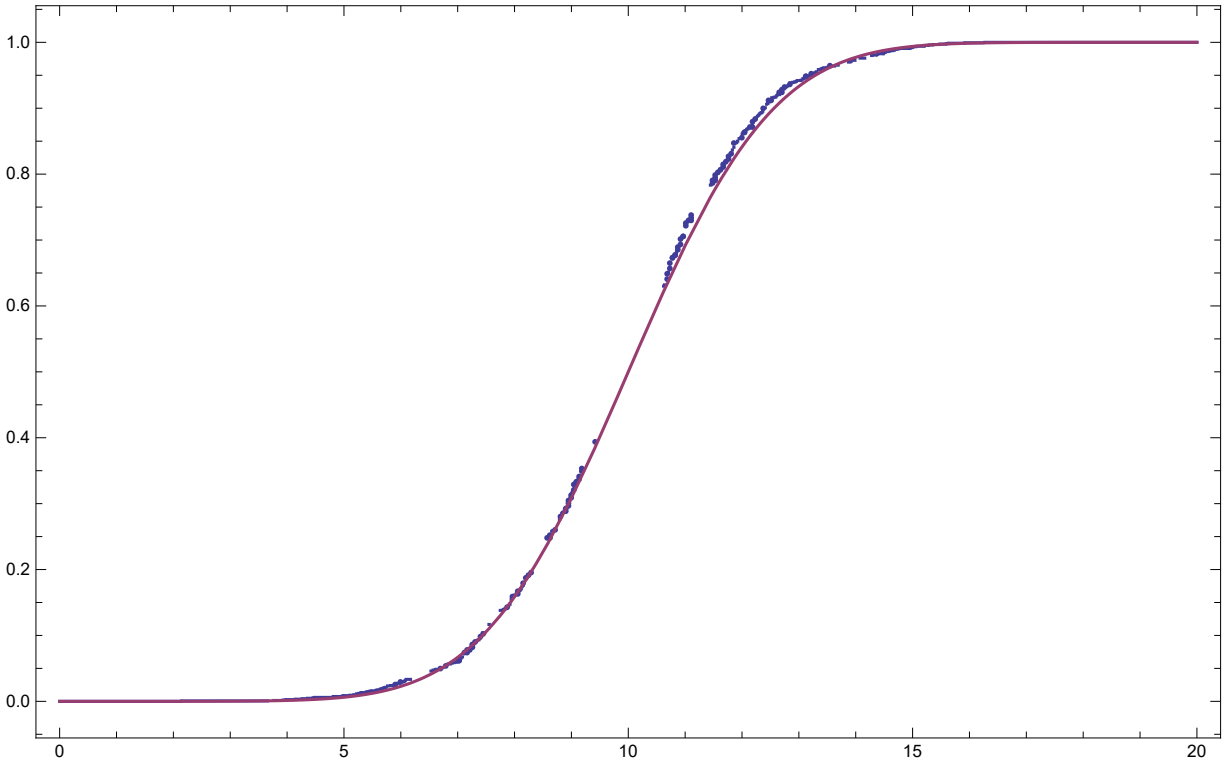
Example: Test for normality

We assume the emission line has Gaussian shape centered on 10 with width 2:



Comparing the empirical c.d.f. with the expected CDF of the Normal distribution:

Example: Test for normality




```
MaxValue[{Abs[CDF[NormalDistribution[10, 2], x] - CDF[emdist, x]], 0 < x < 20}, {x}]
```

```
0.0281366
```

Example: Test for normality

The critical value for large n is

Out[ ]//TraditionalForm=

$$K_\alpha^{\text{crit}} > \frac{1}{\sqrt{n}} \sqrt{\left(-\frac{1}{2} \ln\left[\frac{\alpha}{2}\right]\right)}$$

```
1.358
─────
√1500
```

```
0.0350634
```

Example: Test for normality

Our test statistic is smaller than the critical value, indicating a p-value > α. In fact, the exact p-value is:

```
KolmogorovSmirnovTest[emiss1, NormalDistribution[10, 2]]
```

```
0.182419
```

Therefore we cannot reject the null hypothesis, that our observed data is consistent with a Gaussian of mean 10 and width 2.

Example: Test for normality

The Kolmogorov–Smirnov test has some enormous advantages over the chi-square test.:

- It treats the individual observations separately, and no information is lost because of grouping.
- It works for small samples; for very small samples it is the only alternative. For intermediate sample sizes it is more powerful.
- Finally, note that as described here, the Kolmogorov–Smirnov test is non-directional or two-tailed, as is the chi-square test.

Cramér-von Mises Test

The Cramér–von Mises (CvM) statistic, T_{CvM}, measures the sum of the squared differences between $\hat{F}_n$ and F,

Out[ ]//TraditionalForm=

$$T_{\mathrm{CvM}} = n \int_{-\infty}^{\infty} \left(\hat{F}_n[t] - F[t]\right)^2 d F(t) = \frac{1}{12\,n} + \sum_{i=1}^{n} \left(\frac{2\,i-1}{2\,n} - F(X_i)\right)^2$$

It captures both global and local differences between the data and model, and thus often performs better than the KS test. Two-sample KS and CvM tests are also commonly used, where the e.d.f. of one sample is compared to the e.d.f. of another sample rather than the c.d.f of a model distribution where X_i is the i-th entry when $X_1, X_2, \ldots, X_n$ are placed in increasing order.

Anderson-Darling Test

But here again a limitation is seen: by construction, the cumulative e.d.f. and model converge at zero at low x values and at unity at high x values, so that differences between the distributions are squeezed near the ends of the distributions. Consistent sensitivity across the full range of x is provided by the Anderson–Darling (AD) statistic A^2_{AD}, a weighted variant of the CvM statistic:

Out[ ]//TraditionalForm=

$$A^2_{\mathrm{AD}} = n \sum_{i=1}^{n} \frac{(1/\mathrm{n} - F(X_i))^2}{F(X_i)\,(1 - F(X_i))}$$

Stephens (1974) has found that the Anderson–Darling test is more effective than other e.d.f. tests, and is particularly better than the Kolmogorov–Smirnov test, under many circumstances. Astronomers Hou et al. (2009), comparing the χ^2, KS and AD tests in a study of galaxy group dynamics, also validated that the AD test has the best performance.

Anderson-Darling Test

While the Anderson–Darling test was historically used to test a dataset for normality, there is no reason why it cannot compare a dataset with any parametric model.

Tables are available for AD critical values for different sample sizes n, or they can be obtained from bootstrap resampling. The limiting distribution for large n has a complicated expression. Like the KS and CvM tests, the AD test distribution is independent of the distribution of X_i provided the distribution of X_i is continuous.

Here the results for our example:

Out[•]=

	Statistic	P-Value
Kolmogorov-Smirnov	0.0281366	0.182419
Cramér-von Mises	0.218348	0.234576
Anderson-Darling	1.35984	0.21377

Further reading

- The AstroStat Slog: www.harvard.edu/astrostat/slog/groundtruth.info/AstroStat/slog/tag/model-selection/index.html
- C. Blake's lecture on : Lecture 3 : hypothesis testing and model-fitting http://astronomy.swin.edu.au/~cblake/StatsLecture3.pdf
- K. Levenberg, "A method for the solution of certain problems in least squares, Quart. Appl. Math., 1944, Vol. 2, pp. 164–168.
- D. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," SIAM J. Appl. Math., 1963, Vol. 11, pp. 431–441, doi:10.1137/0111030

Init