Lectures 9: Maximum likelihood II.
(nonlinear least square fits)

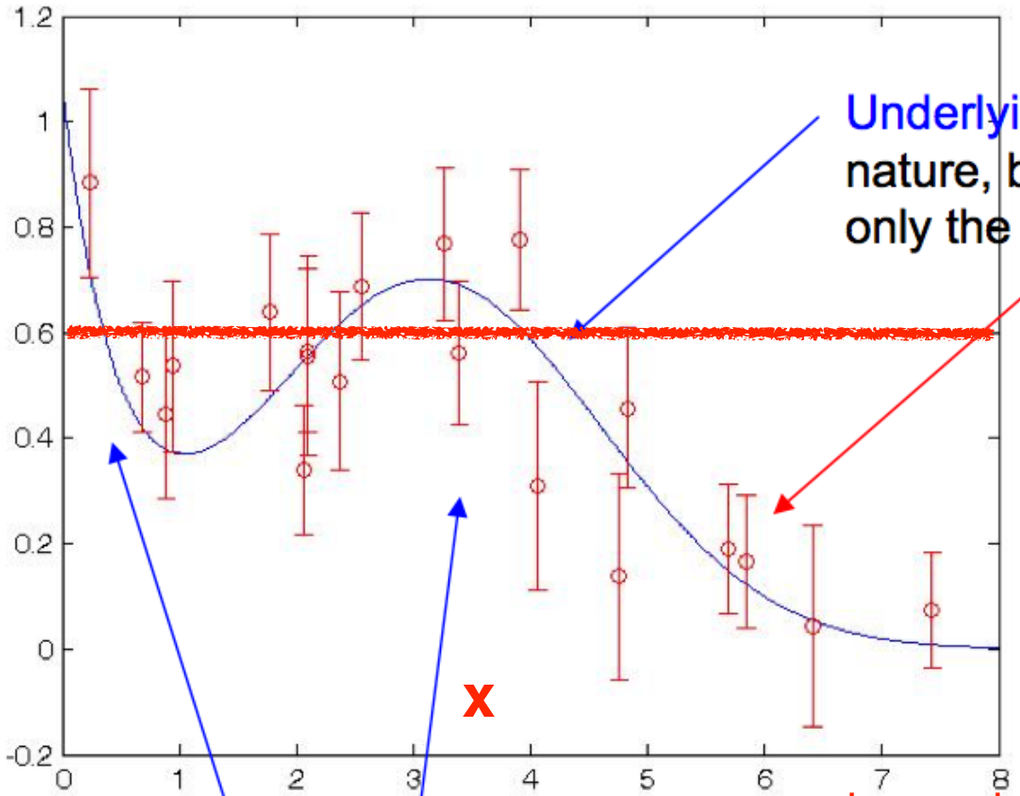
χ^2 fitting procedure!

short review of Lecture 8:

An example might be something like fitting a known functional form to data

$$f(x) = b_1 \exp(-b_2 x) + b_3 \exp\left(-\frac{1}{2}\frac{(x-b_4)^2}{b_5^2}\right)$$

measured value of 2p-0.4 as a function of x



Underlying curve is known to nature, but not to us! We see only the red data points.

Fit 5 parameters from 20 irregularly space points, with normal errors of known standard deviations.

Can we do it? How well?

increasing temperature x in some arbitrary units

for example, this rise might be an instrumental or noise effect, while this bump might be what you are really interested in

short review of Lecture 8:

example: testing coin making machine

short review of Lecture 8:

Model for motivating nonlinear least squares fitting (χ2 fitting)

Manufacturer prints coins noticing that the printing machine produces biased heads. This can be measured by tossing n coins from the batch and measuring the binomial probability p of the batch. For convenience of some analysis 2p - 0.4 is determined by measuring $2n_{head}/n - 0.4$ which turns out to be the function of the temperature where the machine operates (temperature x is recorded for the measurement). The results also depend on five parameters $b_1 \ldots b_5$ of the mechanical construction of the printing machine. A smart theorist comes up with a model how the value of p depends on the temperature x and the five parameters $b_1 \ldots b_5$:

$$f(x) = b_1 \exp(-b_2 x) + b_3 \exp\left(-\frac{1}{2}\frac{(x-b_4)^2}{b_5^2}\right)$$

f(x)=2p-0.4 is the measured value of 2p-0.4 as a function of temperature x

Manufacturer wants to determine the parameters $b_1 \ldots b_5$ so that they can operate the machine at the temperature where 2p - 0.4 = 0.6 so that p=0.5 and the coins are unbiased. This will require to fit the five parameters $b_1 \ldots b_5$ of the machine based on the available data at many temperatures. **How do we do that?**

short review of Lecture 8:

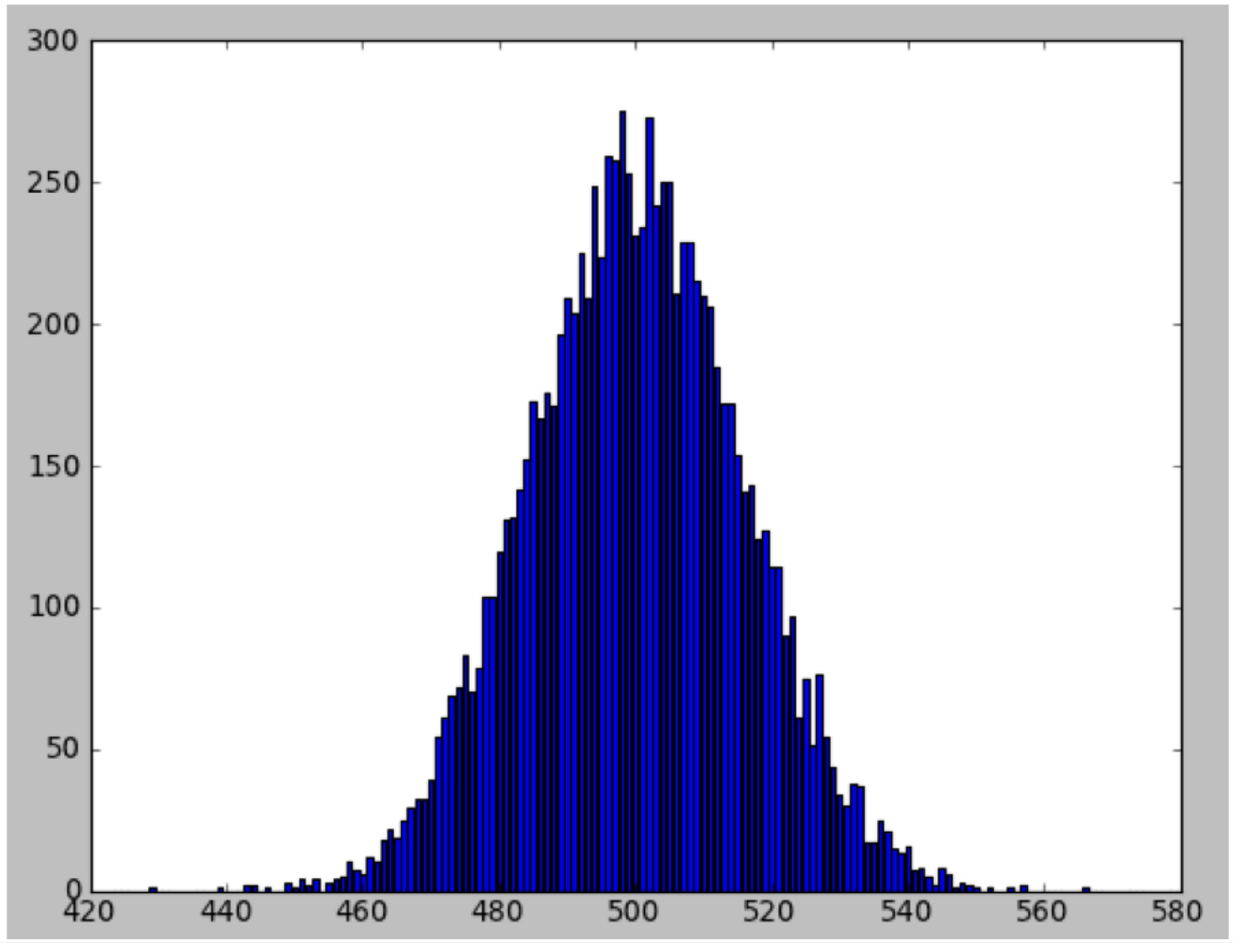
Data are collected at various temperatures x_i.

At each temperature x_i the value $y_i = 2n^{(i)}_{heads}/n - 0.4$ is measured to approximate $2p - 0.4$ from n coin tosses

But y_i has some error e_i

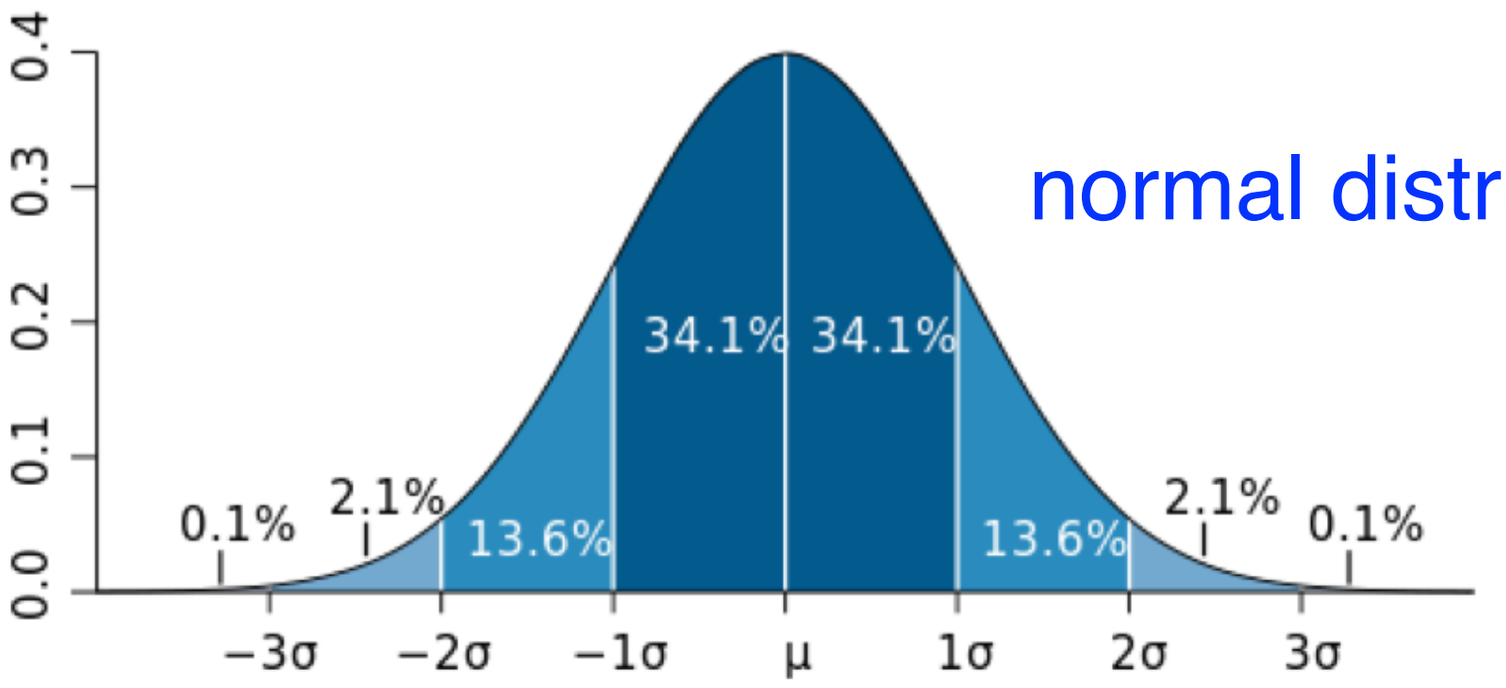
What is the error?

short review of Lecture 8:



central limit theorem

10,000 trials of 1,000 tosses



normal distribution (bell curve)

short review of Lecture 8:

Weighted Nonlinear Least Squares Fitting
a.k.a. χ^2 Fitting
a.k.a. Maximum Likelihood Estimation of Parameters (MLE)
a.k.a. Bayesian parameter estimation
(with uniform prior and maybe some other normality assumptions)

these are not all exactly identical,
but they're real close!

$y_i = y(\mathbf{x}_i|\mathbf{b}) + e_i$ measured values supposed to be a model, plus an error term

$e_i \sim N(0, \sigma_i)$ the errors are Normal, either independently

$\mathbf{e} \sim N(0, \mathbf{\Sigma})$ or else with errors correlated in some known way (e.g., multivariate Normal)

We want to find the parameters of the model **b** from the data.

Maximum Likelihood discussion

Fitting is usually presented in frequentist, MLE language.
But one can equally well think of it as Bayesian:

$$P(\mathbf{b}|\{y_i\}) \propto P(\{y_i\}|\mathbf{b})P(\mathbf{b})$$

$$\propto \prod_i \exp\left[-\tfrac{1}{2}\left(\frac{y_i - y(\mathbf{x}_i|\mathbf{b})}{\sigma_i}\right)^2\right] P(\mathbf{b})$$

$$\propto \exp\left[-\tfrac{1}{2}\sum_i \left(\frac{y_i - y(\mathbf{x}_i|\mathbf{b})}{\sigma_i}\right)^2\right] P(\mathbf{b})$$

$$\propto \exp[-\tfrac{1}{2}\chi^2(\mathbf{b})]P(\mathbf{b})$$

Now the idea is: Find (somehow!) the parameter value $\mathbf{b}_0$ that minimizes χ^2 .

For linear models, you can solve linear "normal equations" or, better, use Singular Value Decomposition. See NR3 section 15.4

In the general nonlinear case, you have a general minimization problem, for which there are various algorithms, none perfect.

Those parameters are the MLE. (So it is Bayes with uniform prior.)

Maximum Likelihood discussion

Nonlinear fits are often easy in MATLAB (or other high-level languages) **if** you can make a reasonable starting guess for the parameters:

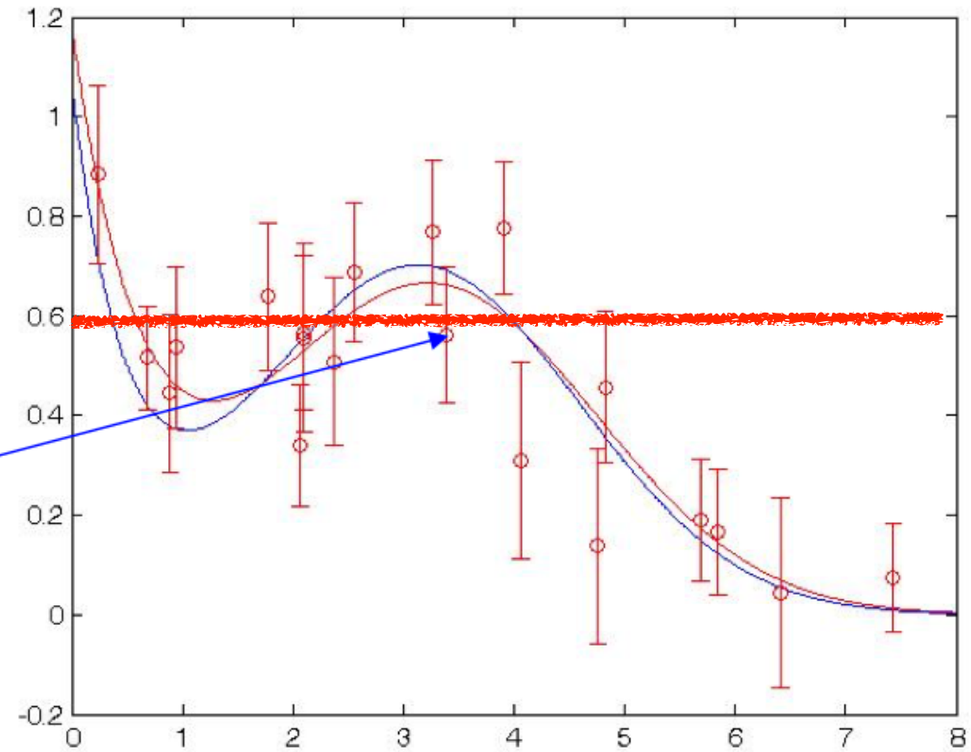
$$y(x|\mathbf{b}) = b_1 \exp(-b_2 x) + b_3 \exp\left(-\frac{1}{2}\frac{(x-b_4)^2}{b_5^2}\right)$$

$$\chi^2 = \sum_i \left(\frac{y_i - y(x_i|\mathbf{b})}{\sigma_i}\right)^2$$

```
ymodel = @(x,b) b(1)*exp(-b(2)*x)+b(3)*exp(-(1/2)*((x-b(4))/b(5)).^2)
chisqfun = @(b) sum(((ymodel(x,b)-y)./sig).^2)
bguess = [1 2 .5 3 1.5]
bfit = fminsearch(chisqfun,bguess)
xfit = (0:0.01:8);
yfit = ymodel(xfit,bfit);
```

bfit = 1.1235 1.5210 0.6582
3.2654 1.4832

Suppose that what we really care about is the area of the bump, and that the other parameters are “nuisance parameters”.



increasing temperature x
in some arbitrary units

Maximum Likelihood parameter errors?

How accurately are the fitted parameters determined?
As Bayesians, we would **instead** say, what is their posterior distribution?

Taylor series:

$$-\tfrac{1}{2}\chi^2(\mathbf{b}) \approx -\tfrac{1}{2}\chi^2_{\mathbf{min}} - \tfrac{1}{2}(\mathbf{b}-\mathbf{b}_0)^T \left[\tfrac{1}{2}\frac{\partial^2\chi^2}{\partial\mathbf{b}\partial\mathbf{b}}\right](\mathbf{b}-\mathbf{b}_0)$$

So, while exploring the χ^2 surface to find its minimum, we must also calculate the Hessian (2nd derivative) matrix at the minimum.

Then

$$P(\mathbf{b}|\{y_i\}) \propto \exp\left[-\tfrac{1}{2}(\mathbf{b}-\mathbf{b}_0)^T\boldsymbol{\Sigma}_b^{-1}(\mathbf{b}-\mathbf{b}_0)\right] P(\mathbf{b})$$

with

$$\boldsymbol{\Sigma}_b = \left[\tfrac{1}{2}\frac{\partial^2\chi^2}{\partial\mathbf{b}\partial\mathbf{b}}\right]^{-1}$$

covariance (or "standard error") matrix of the fitted parameters

Notice that if (i) the Taylor series converges rapidly and (ii) the prior is uniform, then the posterior distribution of the $\mathbf{b}$'s is multivariate Normal, a very useful CLT-ish result!

χ^2 distribution

Let's talk more about **chi-square**.
Recall that a t-value is (by definition) a deviate from $\mathrm{N}(0,1)$

χ^2 is a "statistic" defined as the sum of the squares of n independent t-values.

$$\chi^2 = \sum_i \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2, \qquad x_i \sim \mathrm{N}(\mu_i, \sigma_i)$$

$\mathrm{Chisquare}(\nu)$ is a distribution (special case of Gamma), defined as

$$\chi^2 \sim \mathrm{Chisquare}(\nu), \qquad \nu > 0$$

$$p(\chi^2)d\chi^2 = \frac{1}{2^{\frac{1}{2}\nu}\Gamma(\frac{1}{2}\nu)}(\chi^2)^{\frac{1}{2}\nu-1}\exp\left(-\tfrac{1}{2}\chi^2\right)d\chi^2, \qquad \chi^2 > 0$$

The important theorem is that $\chi 2$ is in fact distributed as Chisquare.

Let's prove it.

Γ function 1-pager

In mathematics, the **gamma function** (represented by the capital Greek letter **Γ**) is an extension of the factorial function, with its argument shifted down by 1, to real and complex numbers. That is, if n is a positive integer:

$$\Gamma(n) = (n-1)!.$$

The gamma function is defined for all complex numbers except the non-positive integers. For complex numbers with a positive real part, it is defined via a convergent improper integral:

$$\Gamma(z) = \int_0^{\infty} x^{z-1} e^{-x}\,\mathrm{d}x. \qquad \Gamma\left(\tfrac{1}{2}\right) = \sqrt{\pi}$$

$$\Gamma\left(\tfrac{1}{2}+n\right) = \frac{(2n)!}{4^n n!}\sqrt{\pi} = \frac{(2n-1)!!}{2^n}\sqrt{\pi} = \binom{n-\frac{1}{2}}{n} n!\sqrt{\pi}$$

$$\Gamma\left(\tfrac{1}{2}-n\right) = \frac{(-4)^n n!}{(2n)!}\sqrt{\pi} = \frac{(-2)^n}{(2n-1)!!}\sqrt{\pi} = \frac{\sqrt{\pi}}{\binom{-\frac{1}{2}}{n} n!}$$

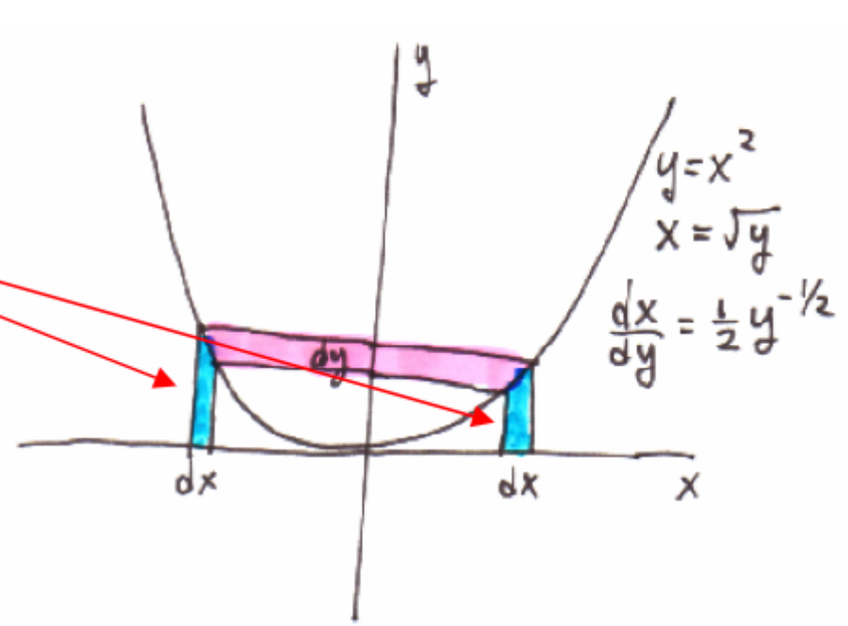
χ^2 distribution

Prove first the case of ν=1:

Suppose $p_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \quad \Rightarrow \quad x \sim \mathrm{N}(0,1)$

and $y = x^2$

$p_Y(y)\,dy = 2p_X(x)\,dx$



So, $p_Y(y) = y^{-1/2} p_X(y^{1/2}) = \frac{1}{\sqrt{2\pi y}} e^{-\frac{1}{2}y}$

$\sim$ Chisquare(1)

χ^2 distribution

To prove the general case for integer ν, compute the characteristic function

$$\chi^2 \sim \text{Chisquare}(\nu), \qquad \nu > 0$$

$$p(\chi^2)d\chi^2 = \frac{1}{2^{\frac{1}{2}\nu}\Gamma(\frac{1}{2}\nu)}(\chi^2)^{\frac{1}{2}\nu-1}\exp\left(-\tfrac{1}{2}\chi^2\right)d\chi^2, \qquad \chi^2 > 0$$

characteristic function by Fourier transformation:

$(1-2i*t)^{-\nu/2}$

Since we already proved that ν=1 is the distribution of a single t^2-value, this proves that the general ν case is the sum of ν t^2-values.

χ^2 distribution Maximum Likelihood parameter errors?

How accurately are the fitted parameters determined?
As Bayesians, we would **instead** say, what is their posterior distribution?

Taylor series:

$$-\tfrac{1}{2}\chi^2(\mathbf{b}) \approx -\tfrac{1}{2}\chi^2_{\mathbf{min}} - \tfrac{1}{2}(\mathbf{b}-\mathbf{b}_0)^T \left[\tfrac{1}{2}\frac{\partial^2\chi^2}{\partial\mathbf{b}\partial\mathbf{b}}\right](\mathbf{b}-\mathbf{b}_0)$$

So, while exploring the χ^2 surface to find its minimum, we must also calculate the Hessian (2nd derivative) matrix at the minimum.

Then

$$P(\mathbf{b}|\{y_i\}) \propto \exp\left[-\tfrac{1}{2}(\mathbf{b}-\mathbf{b}_0)^T\boldsymbol{\Sigma}_b^{-1}(\mathbf{b}-\mathbf{b}_0)\right] P(\mathbf{b})$$

with

$$\boldsymbol{\Sigma}_b = \left[\tfrac{1}{2}\frac{\partial^2\chi^2}{\partial\mathbf{b}\partial\mathbf{b}}\right]^{-1}$$

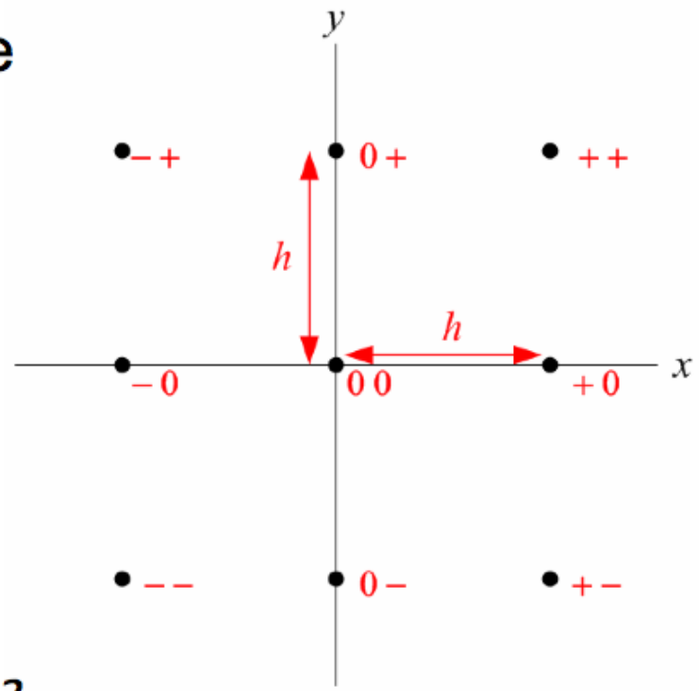
covariance (or "standard error") matrix of the fitted parameters

Notice that if (i) the Taylor series converges rapidly and (ii) the prior is uniform, then the posterior distribution of the $\mathbf{b}$'s is multivariate Normal

χ^2 distribution Maximum Likelihood parameter errors?

Numerical calculation of the Hessian by finite difference

$$\frac{\partial^2 f}{\partial x \partial y} \approx \frac{1}{2h}\left(\frac{f_{++} - f_{-+}}{2h} - \frac{f_{+-} - f_{--}}{2h}\right)$$
$$= \frac{1}{4h^2}\left(f_{++} + f_{--} - f_{+-} - f_{-+}\right)$$




```
bfit = 1.1235     1.5210     0.6582     3.2654     1.4832
```

```
chisqfun = @(b) sum(((ymodel(x,b)-y)./sig).^2)
h = 0.1;
unit = @(i) (1:5) == i;
hess = zeros(5,5);
for i=1:5, for j=1:5,
        bpp = bfit + h*(unit(i)+unit(j));
        bmm = bfit + h*(-unit(i)-unit(j));
        bpm = bfit + h*(unit(i)-unit(j));
        bmp = bfit + h*(-unit(i)+unit(j));
        hess(i,j) = (chisqfun(bpp)+chisqfun(bmm)…
          -chisqfun(bpm)-chisqfun(bmp))./(2*h)^2;
    end
end
covar = inv(0.5*hess)
```

This also works for the diagonal components. Can you see how?

χ^2 distribution Maximum Likelihood parameter errors?

For our example, $y(x|\mathbf{b}) = b_1 \exp(-b_2 x) + b_3 \exp\left(-\frac{1}{2}\frac{(x-b_4)^2}{b_5^2}\right)$

```
bfit =
    1.1235    1.5210    0.6582    3.2654    1.4832
hess =
   64.3290  -38.3070   47.9973  -29.0683   46.0495
  -38.3070   31.8759  -67.3453   29.7140  -40.5978
   47.9973  -67.3453  723.8271  -47.5666  154.9772
  -29.0683   29.7140  -47.5666   68.6956  -18.0945
   46.0495  -40.5978  154.9772  -18.0945   89.2739
covar =
    0.1349    0.2224    0.0068   -0.0309    0.0135
    0.2224    0.6918    0.0052   -0.1598    0.1585
    0.0068    0.0052    0.0049    0.0016   -0.0094
   -0.0309   -0.1598    0.0016    0.0746   -0.0444
    0.0135    0.1585   -0.0094   -0.0444    0.0948
```

This is the covariance structure of all the parameters, and indeed (at least in CLT normal approximation) gives their entire joint distribution!

The standard errors on each parameter separately are $\sigma_i = \sqrt{C_{ii}}$

```
sigs =
    0.3672    0.8317    0.0700    0.2731    0.3079
```

But why is this, and what about two or more parameters at a time (e.g. b_3 and b_5)?