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

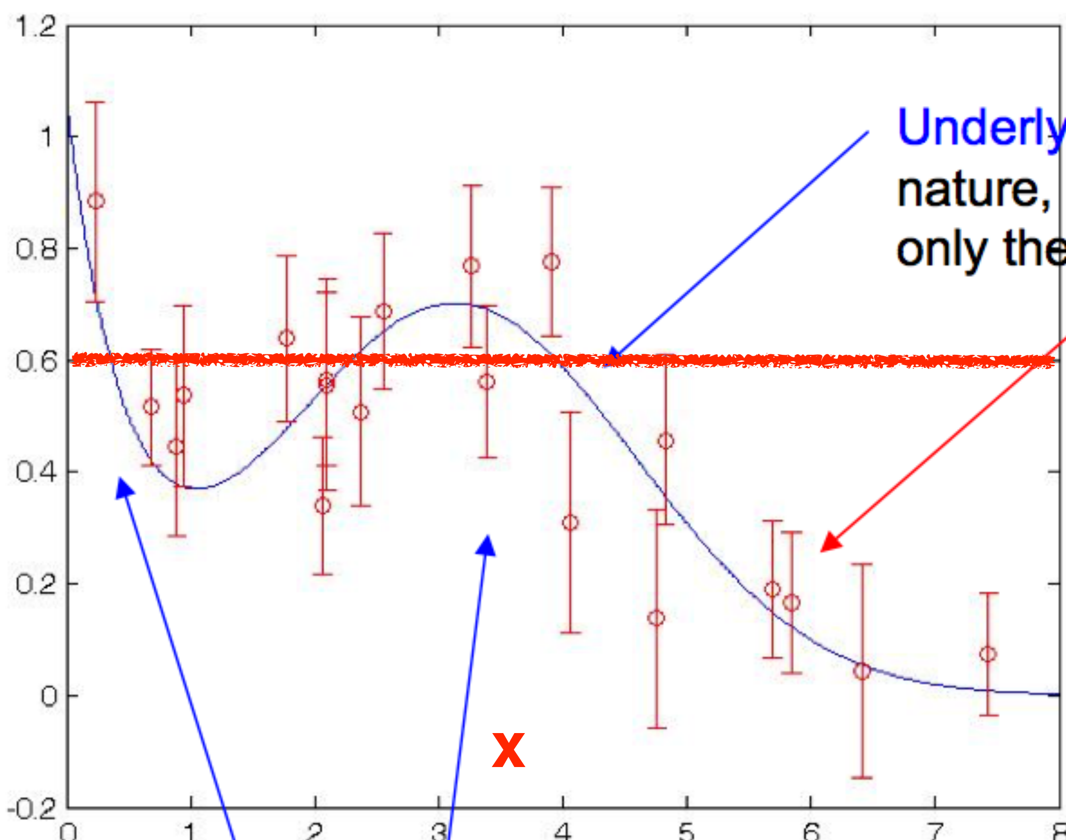
χ^2 fitting procedure!

from 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

from Lecture 8: Maximum Likelihood discussion

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

$$
\begin{aligned}
P(\mathbf{b}|\{y_i\}) &\propto P(\{y_i\}|\mathbf{b})P(\mathbf{b}) \\
&\propto \prod_i \exp\left[-\tfrac{1}{2}\left(\frac{y_i - y(\mathbf{x}_i|\mathbf{b})}{\sigma_i}\right)^2\right] P(\mathbf{b}) \\
&\propto \exp\left[-\tfrac{1}{2}\sum_i \left(\frac{y_i - y(\mathbf{x}_i|\mathbf{b})}{\sigma_i}\right)^2\right] P(\mathbf{b}) \quad \text{prior set to } P(b) = 1 \\
&\propto \exp[-\tfrac{1}{2}\chi^2(\mathbf{b})]P(\mathbf{b})
\end{aligned}
$$

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.)

from Lecture 8: **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:

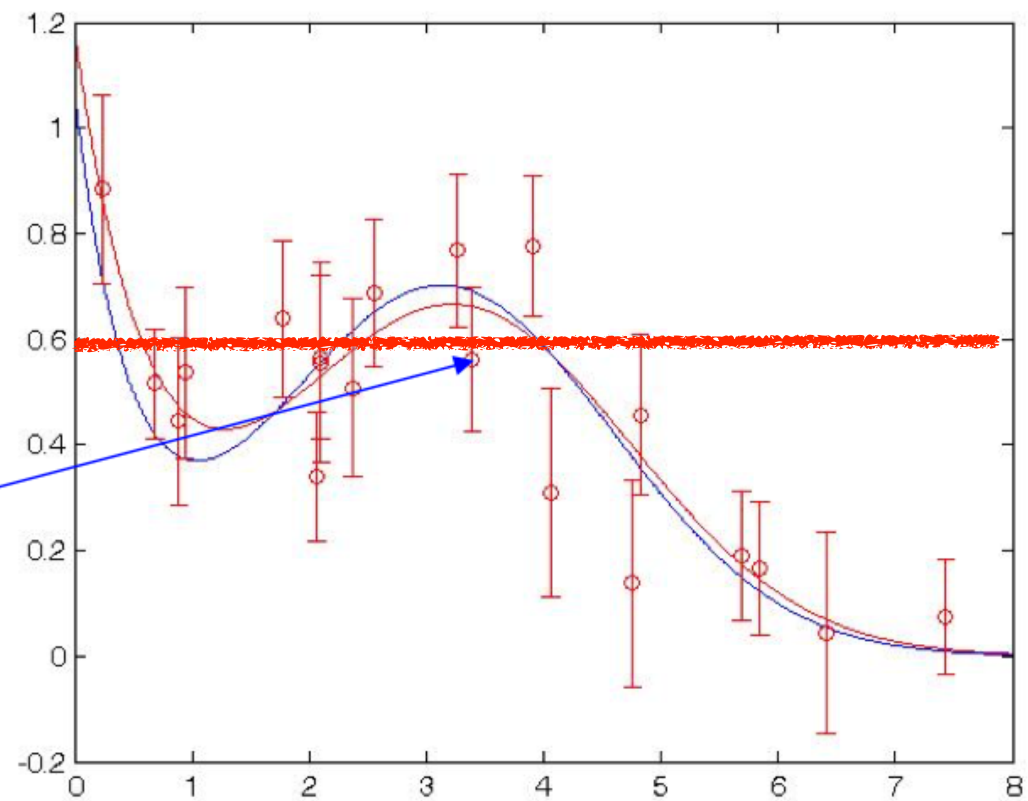
$$y(x|\mathbf{b}) = b_1 \exp(-b_2 x) + b_3 \exp\left(-\frac{1}{2}\frac{(x-b_4)^2}{b_5^2}\right)$$

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

```
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

χ^2 distribution Maximum Likelihood parameter errors?

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

Taylor series:

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

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

Then

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

with

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

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

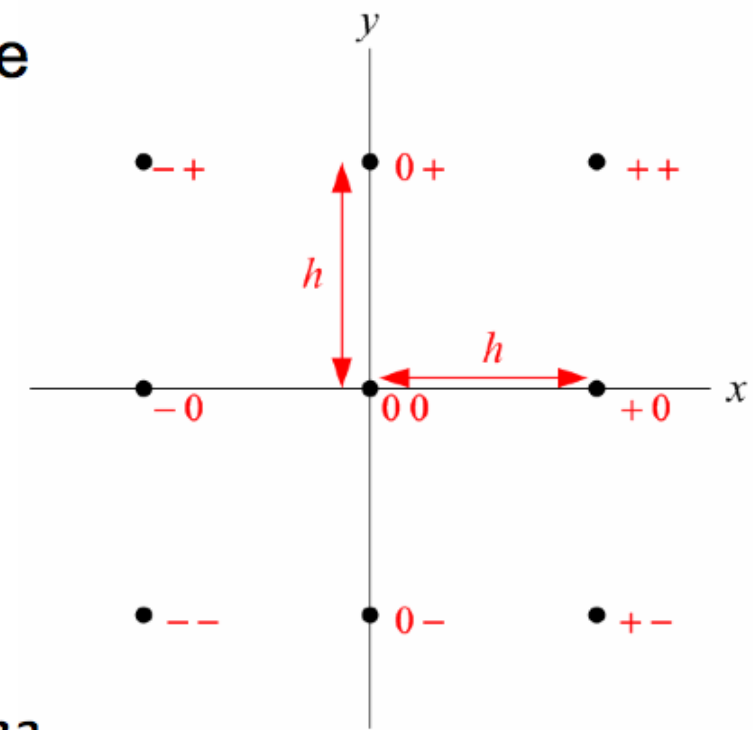
Notice that if (i) the Taylor series converges rapidly and (ii) the prior is uniform, then the posterior distribution of the $\mathbf{b}$'s is multivariate Normal

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)?

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.

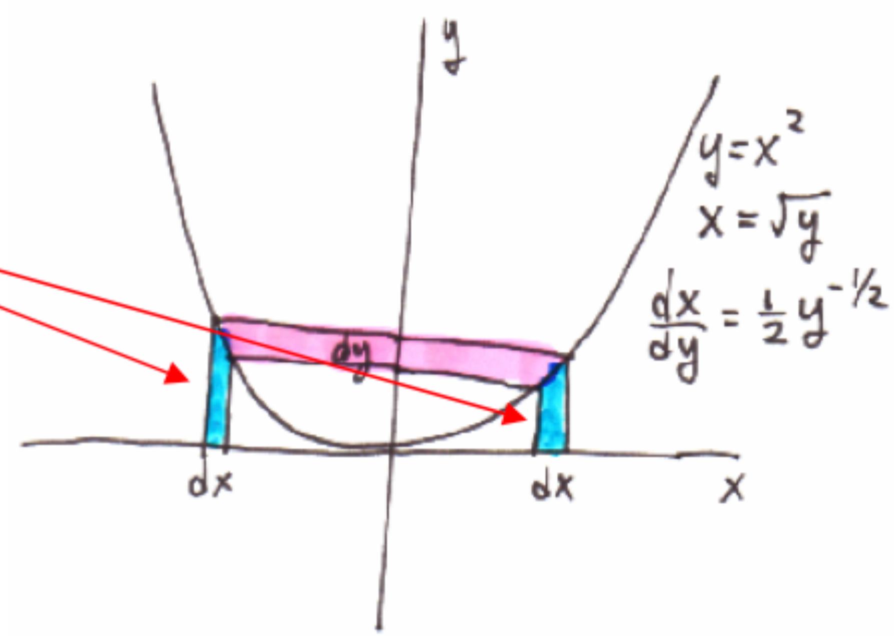
χ^2 distribution

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 $$y = x^2$$

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



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

$$\sim \mathrm{Chisquare}(1)$$

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:

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

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 goodness of fit

we have assumed that, for some value of the parameters $\mathbf{b}$
the model $y(\mathbf{x}_i|\mathbf{b})$ is correct

Suppose that the model $y(\mathbf{x}_i|\mathbf{b})$ does fit. This is the null hypothesis.

Then the "statistic" $\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - y(\mathbf{x}_i|\mathbf{b})}{\sigma_i} \right)^2$ is the sum of N t^2-values.

(not quite)

So, if we imagine repeated experiments (which Bayesians refuse to do), the statistic should be distributed as $\text{Chisquare}(N)$.

If our experiment is <u>very unlikely</u> to be from this distribution, we consider the model to be disproved. In other words, <u>it is a p-value test</u>.

χ^2 distribution (from Lecture 9)

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

$$y = x^2$$

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

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

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

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

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

χ^2 distribution Maximum Likelihood marginalized parameters

We can Marginalize or Condition uninteresting parameters. (Different things!)

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

Marginalize: (this is usual) Ignore (integrate over) uninteresting parameters.

In $\Sigma_b = \left[\frac{1}{2}\frac{\partial^2\chi^2}{\partial\mathbf{b}\partial\mathbf{b}}\right]^{-1}$ submatrix of *interesting* rows and columns is new Σ_b

Special case of one variable at a time: Just take diagonal components in Σ_b

Covariances are pairwise expectations and don't depend on whether other parameters are "interesting" or not.

Condition: (this is rare!) Fix uninteresting parameters at specified values.

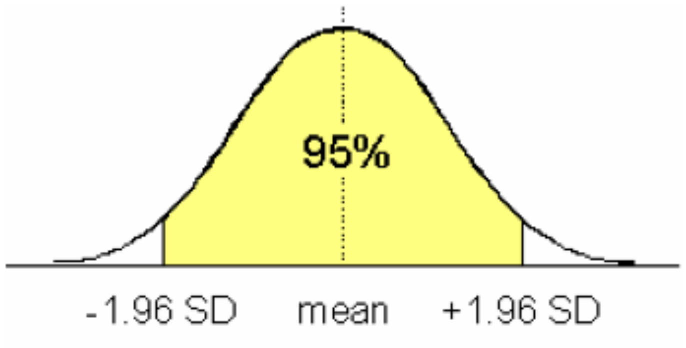
In $\Sigma_b^{-1} = \left[\frac{1}{2}\frac{\partial^2\chi^2}{\partial\mathbf{b}\partial\mathbf{b}}\right]$ submatrix of *interesting* rows and columns is new Σ_b^{-1}

Take matrix inverse if you want their covariance Σ_b

(If you fix parameters at other than $\mathbf{b}_0$, the mean also shifts – exercise for reader!)

confidence intervals

The variances of *one parameter* at a time imply confidence intervals as for an ordinary 1-dimensional normal distribution:

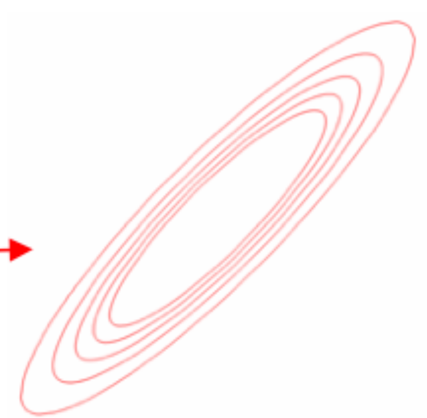


(Remember to take the square root of the variances to get the standard deviations!)

If you want to give confidence regions for *more than one parameter* at a time, you have to decide on a shape, since any shape containing 95% (or whatever) of the probability is a 95% confidence region!

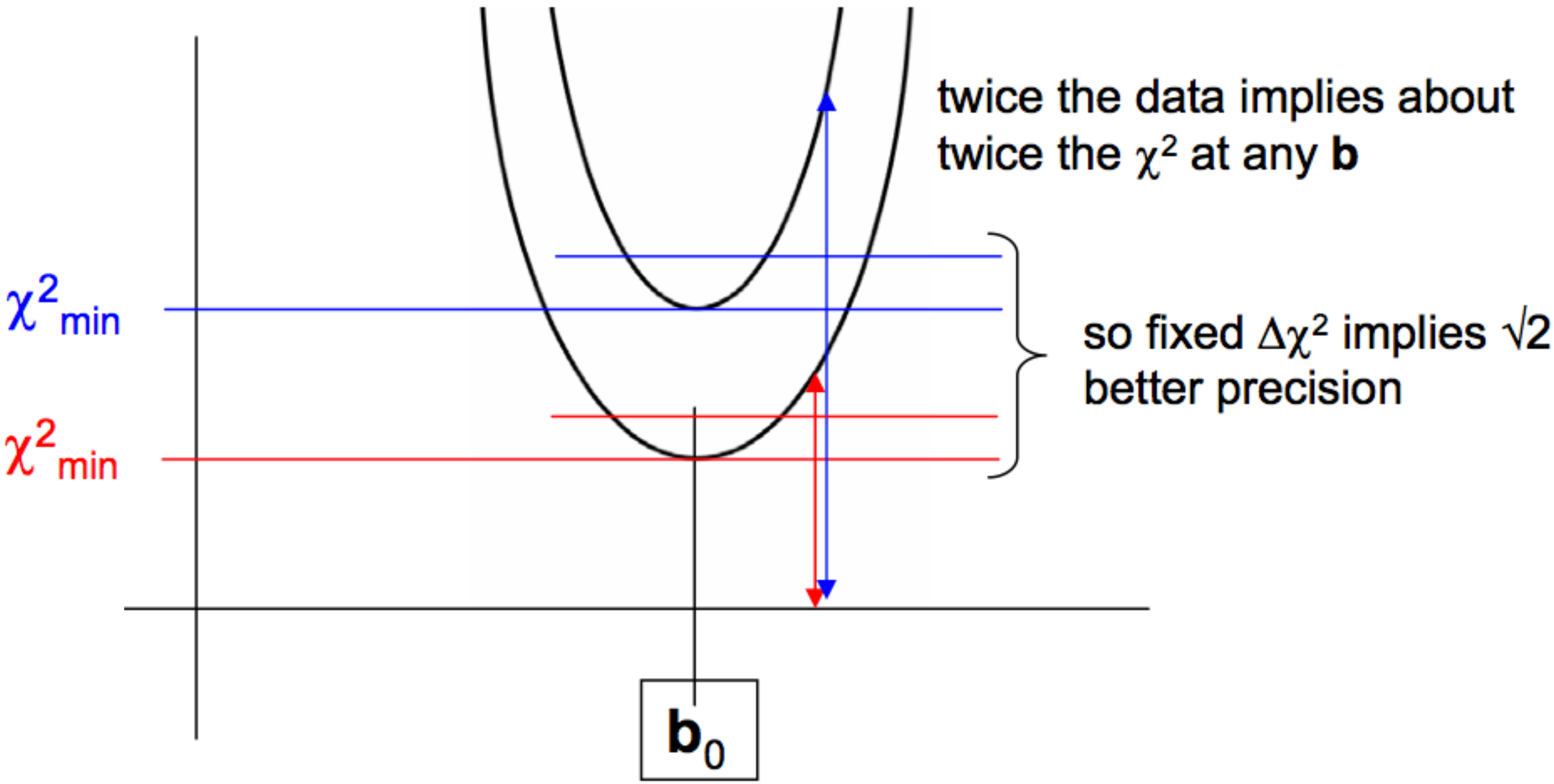
It is *conventional* to use contours of probability density as the shapes (= contours of $\Delta\chi^2$) since these are maximally compact.

But which $\Delta\chi^2$ contour contains 95% of the probability?



χ^2 distribution

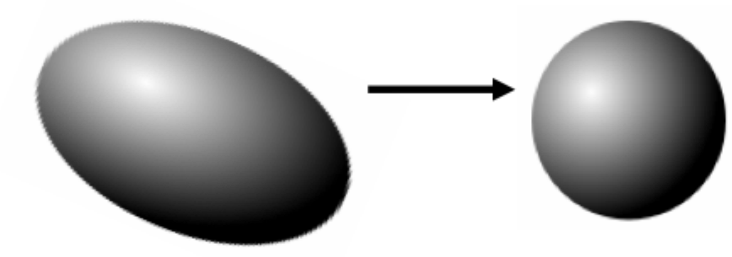
Measurement precision improves with the amount of data *N* as *N*$^{-1/2}$



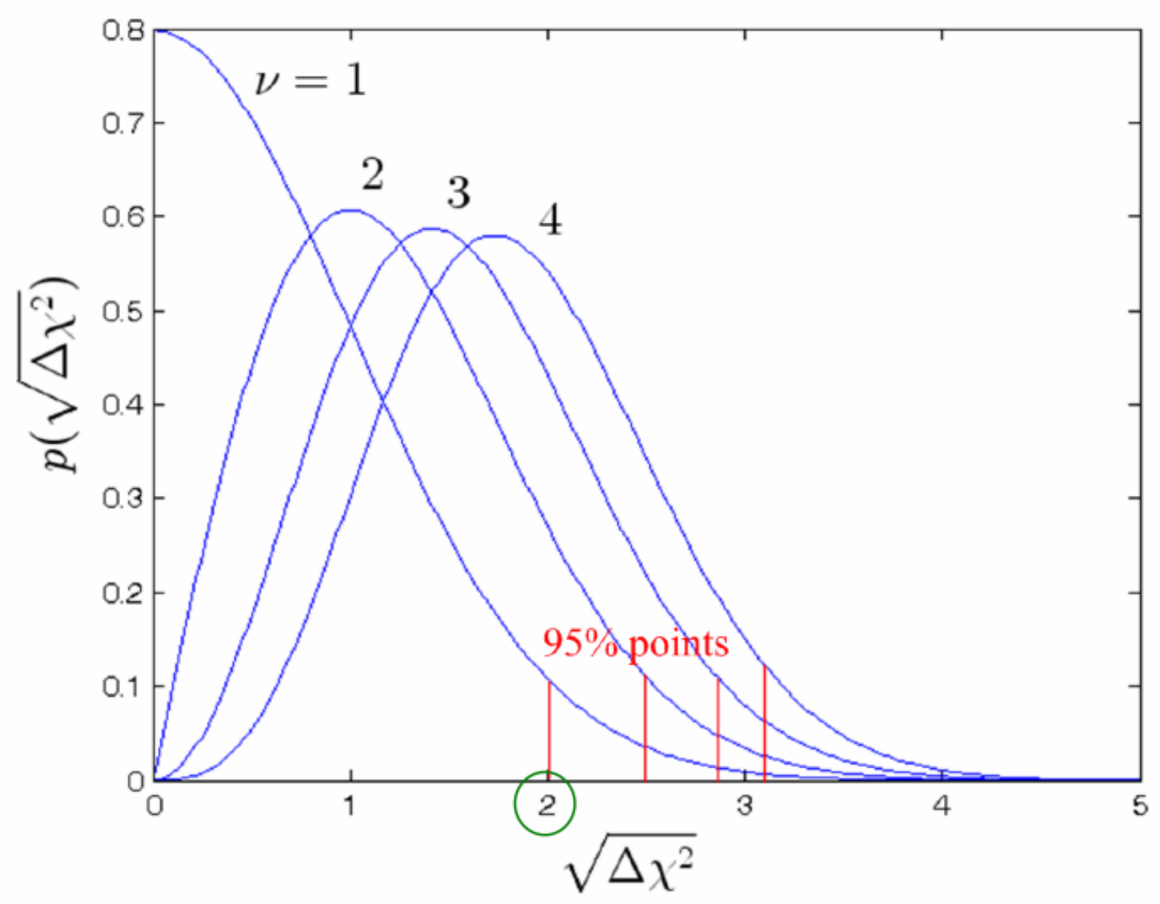
confidence intervals

What $\Delta\chi^2$ contour in ν dimensions contains some percentile probability?

Rotate and scale the covariance to make it spherical.
(Linear, so contours still contain same probability.)



Now, each dimension is an independent Normal, and contours are labeled by radius squared (sum of ν individual t^2 values), so $\Delta\chi^2 \sim$ Chisquare(ν)



$\Delta\chi^2$ as a Function of Confidence Level p and Number of Parameters of Interest ν

p	ν = 1	2	3	4	5	6
68.27%	1.00	2.30	3.53	4.72	5.89	7.04
90%	2.71	4.61	6.25	7.78	9.24	10.6
95.45%	4.00	6.18	8.02	9.72	11.3	12.8
99%	6.63	9.21	11.3	13.3	15.1	16.8
99.73%	9.00	11.8	14.2	16.3	18.2	20.1
99.99%	15.1	18.4	21.1	23.5	25.7	27.9

You sometimes learn "facts" like: "delta chi-square of 1 is the 68% confidence level". We now see that this is true only for one parameter at a time.

χ^2 distribution Maximum Likelihood marginalized parameters

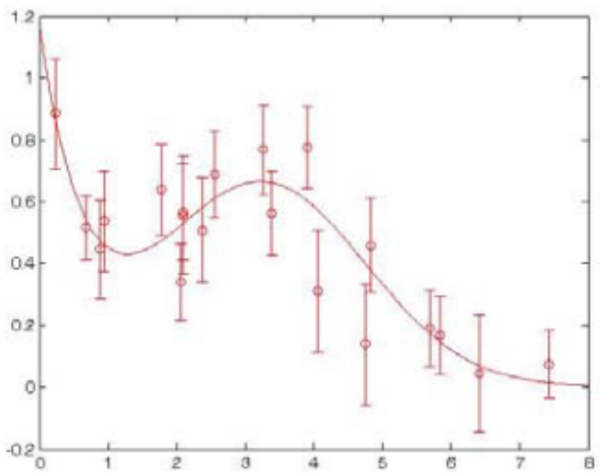
For our example, we are conditioning or marginalizing from 5 to 2 dims:

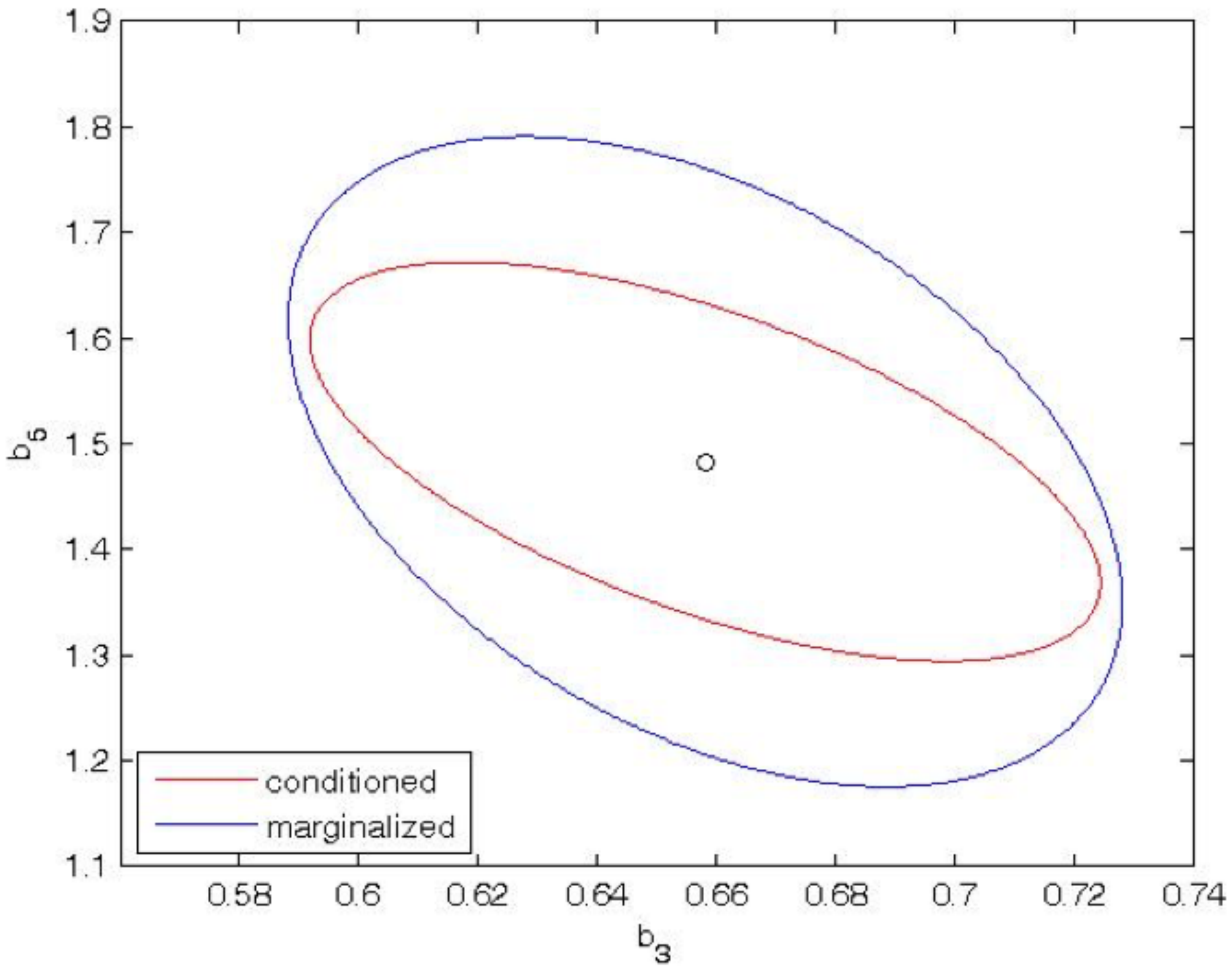
$$y(x|\mathbf{b}) = b_1 \exp(-b_2 x) + b_3 \exp\left(-\frac{1}{2}\frac{(x-b_4)^2}{b_5^2}\right)$$

the uncertainties on b_3 and b_5 jointly (as error ellipses) are

```
sigcond =
    0.0044   -0.0076
   -0.0076    0.0357

sigmarg =
    0.0049   -0.0094
   -0.0094    0.0948
```





Conditioned errors are always smaller, but are useful only if you can find <u>other ways</u> to measure (accurately) the parameters that you want to condition on.

correlated data - first glimpse

Multivariate Normal Distributions

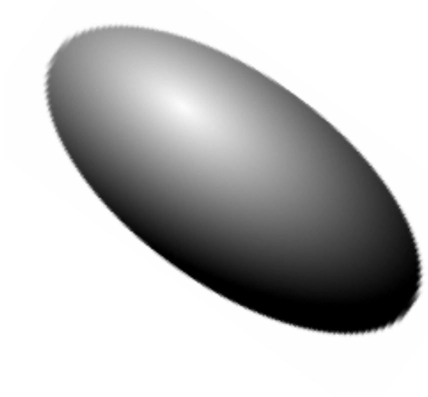
Generalizes Normal (Gaussian) to M-dimensions
Like 1-d Gaussian, completely defined by its mean and (co-)variance
Mean is a M-vector, covariance is a M x M matrix

$$N(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{M/2}\det(\boldsymbol{\Sigma})^{1/2}}\exp[-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})]$$

The mean and covariance of r.v.'s from this distribution are*

$$\boldsymbol{\mu} = \langle \mathbf{x} \rangle \qquad \boldsymbol{\Sigma} = \langle (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T \rangle$$

In the one-dimensional case σ is the standard deviation, which can be visualized as "error bars" around the mean.



In more than one dimension Σ can be visualized as an error ellipsoid around the mean in a similar way.

$$1 = (\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

correlated data - first glimpse

Question: What is the generalization of

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

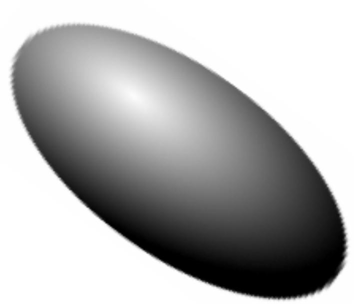
to the case where the x_i's are normal, **but not independent**?
I.e., **x** comes from a multivariate Normal distribution?

$$N(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{M/2} \det(\boldsymbol{\Sigma})^{1/2}} \exp[-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})]$$

The mean and covariance of r.v.'s from this distribution are*

$$\boldsymbol{\mu} = \langle \mathbf{x} \rangle \qquad \boldsymbol{\Sigma} = \left\langle (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T \right\rangle$$

In the one-dimensional case σ is the standard deviation, which can be visualized as "error bars" around the mean.



In more than one dimension Σ can be visualized as an error ellipsoid around the mean in a similar way.

$$1 = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

generic idea of covariance matrix

The covariance matrix is a more general idea than just for multivariate Normal.
You can compute the covariances of any set of random variables.
It's the generalizaton to M-dimensions of the (centered) second moment Var.

$$\mathrm{Cov}\,(x,y) = \langle (x-\overline{x})(y-\overline{y}) \rangle$$

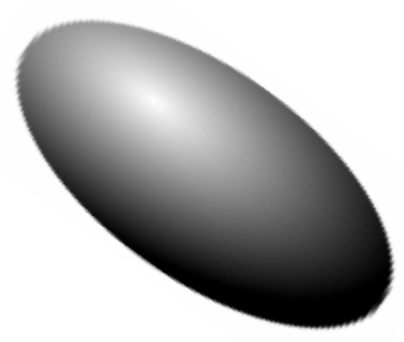
For multiple r.v.'s, all the possible covariances form a (symmetric) matrix:

$$\mathbf{C} = C_{ij} = \mathrm{Cov}\,(x_i, x_j) = \langle (x_i - \overline{x_i})(x_j - \overline{x_j}) \rangle$$

Notice that the diagonal elements are the variances of the individual variables.

The variance of any linear combination of r.v.'s is a quadratic form in $\mathbf{C}$:

$$\begin{aligned}
\mathrm{Var}\,\left(\sum \alpha_i x_i\right) &= \left\langle \sum_i \alpha_i (x_i - \overline{x_i}) \sum_j \alpha_j (x_j - \overline{x_j}) \right\rangle \\
&= \sum_{ij} \alpha_i \left\langle (x_i - \overline{x_i})(x_j - \overline{x_j}) \right\rangle \alpha_j \\
&= \boldsymbol{\alpha}^T \mathbf{C}\, \boldsymbol{\alpha}
\end{aligned}$$



This also shows that $\mathbf{C}$ is positive definite, so it can still be visualized as an ellipsoid in the space of the r.v.'s., where the directions are the different linear combinations.

generic idea of covariance matrix

The covariance matrix is closely related to the linear correlation matrix.

$$r_{ij} = \frac{C_{ij}}{\sqrt{C_{ii}C_{jj}}}$$

more often seen written out as

$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2}\sqrt{\sum_i (y_i - \overline{y})^2}}$$

When the null hypothesis is that X and Y are independent r.v.'s, then r is useful as a p-value statistic ("test for correlation"), because

1. For large numbers of data points *N*, it is normally distributed,

   $r \sim \mathrm{N}(0, N^{-1/2})$

   so $r\sqrt{N}$ is a normal t-value

2. Even with small numbers of data points, if the underlying distribution is multivariate normal, there is a simple form for the p-value (comes from a Student t distribution).