Opinionated Lessons in Statistics

by Bill Press

#21 Marginalize vs. Condition Uninteresting Fitted Parameters

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

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

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

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

Special case of one variable at a time: Just take diagonal components in $\boldsymbol{\Sigma}_b$

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.

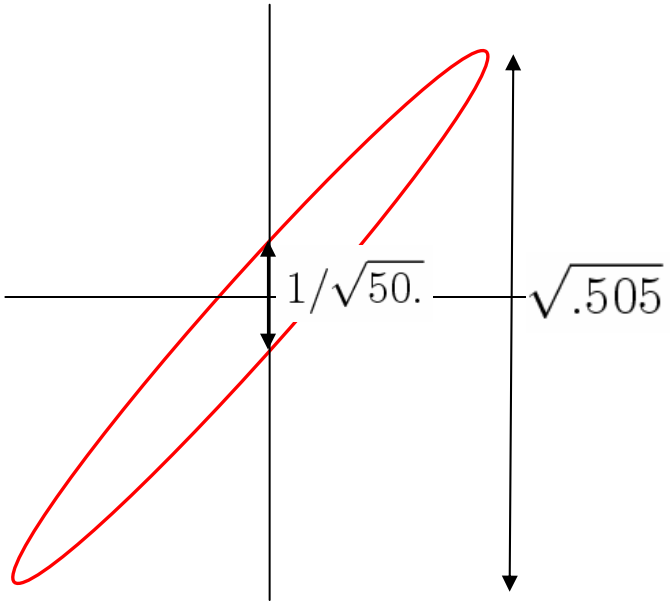
In $\boldsymbol{\Sigma}_b^{-1} = \left[\frac{1}{2}\frac{\partial^2\chi^2}{\partial\mathbf{b}\partial\mathbf{b}}\right]$ submatrix of *interesting* rows and columns is new $\boldsymbol{\Sigma}_b^{-1}$

Take matrix inverse if you want their covariance $\boldsymbol{\Sigma}_b$

(If you fix uninteresting parameters at any value other than $\mathbf{b_0}$, the mean also shifts – exercise for reader to calculate, or see Wikipedia "Multivariate Normal Distribution".)

Example of 2 dimensions marginalizing or conditioning to 1 dimension:

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



$$\boldsymbol{\Sigma}_b^{-1} = \left[\tfrac{1}{2}\frac{\partial^2\chi^2}{\partial\mathbf{b}\partial\mathbf{b}}\right] = \begin{pmatrix} 50. & -49. \\ -49. & 50. \end{pmatrix}$$

$$\boldsymbol{\Sigma}_b = \begin{pmatrix} .505 & .495 \\ .495 & .505 \end{pmatrix}$$

By the way, don't confuse the "covariance matrix of the fitted parameters" with the "covariance matrix of the data". For example, the data covariance is often diagonal (uncorrelated σ_i's), while the parameters covariance is essentially never diagonal!

If the data has correlated errors, then the starting point for $\chi^2(\mathbf{b})$ is (recall):

$$\chi^2 = [\mathbf{y}_{\{\mathbf{i}\}} - \mathbf{y}(\mathbf{x}_{\{i\}}|\mathbf{b})]^T\boldsymbol{\Sigma}^{-1}[\mathbf{y}_{\{\mathbf{i}\}} - \mathbf{y}(\mathbf{x}_{\{i\}}|\mathbf{b})] \quad \text{instead of} \quad \sum_i \left(\frac{y_i - y(\mathbf{x}_i|\mathbf{b})}{\sigma_i}\right)^2$$

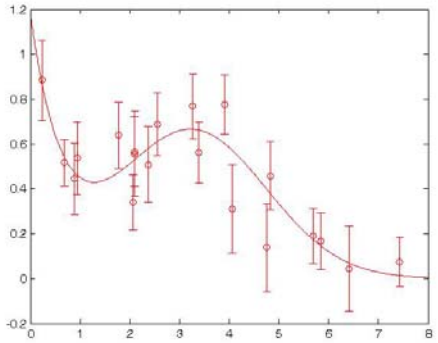
For our example, we are conditioning or marginalizing from 5 to 2 dims:

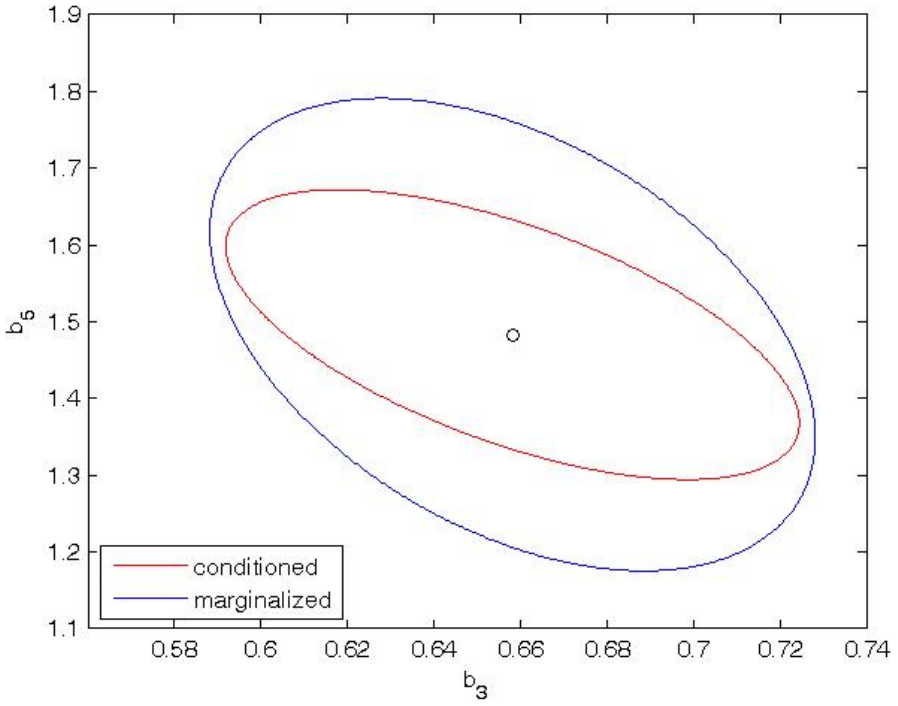
$$y(x|\mathbf{b}) = b_1 \exp(-b_2 x) + b_3 \exp\left(-\frac{1}{2}\frac{(x-b_4)^2}{b_5^2}\right)$$

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 other ways to measure (accurately) the parameters that you want to condition on.

Frequentists love MLE estimates (and not just the case with a Normal error model) because they have provably nice properties asymptotically as the size of the data set becomes large

- Consistency: converges to true value of the parameters
- Equivariance: estimate of function of parameter = function of estimate of parameter
- asymptotically Normal
- asymptotically efficient (optimal): among estimators with the above properties, it has the smallest variance

The "Fisher Information Matrix" is another name for the Hessian of the log probability (or, rather, log likelihood):

$$\mathbf{I}(\mathbf{b}) = -\left\langle \frac{\partial^2 \log P(\{y_i\}|\mathbf{b})}{\partial \mathbf{b} \partial \mathbf{b}} \right\rangle \approx 2\boldsymbol{\Sigma}_b^{-1}$$

except that, strictly speaking, it is an expectation over the population

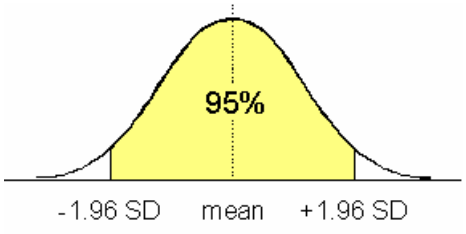
Bayesians tolerate MLE estimates because they are almost Bayesian – even better if you put the prior back into the minimization.

But Bayesians know that we live in a non-asymptotic world: none of the above properties are exactly true for finite data sets!

Small digression:

You can give confidence intervals or regions, instead of (co-)variances

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

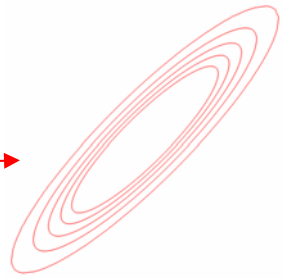


(Remember to take the square root of the variances to get the standard deviations!)

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!

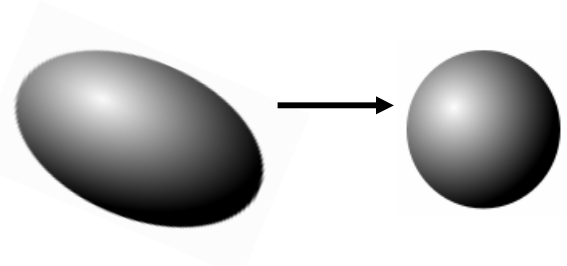
It is *conventional* to use contours of probability density as the shapes (= contours of $\Delta\chi^2$) since these are maximally compact.

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

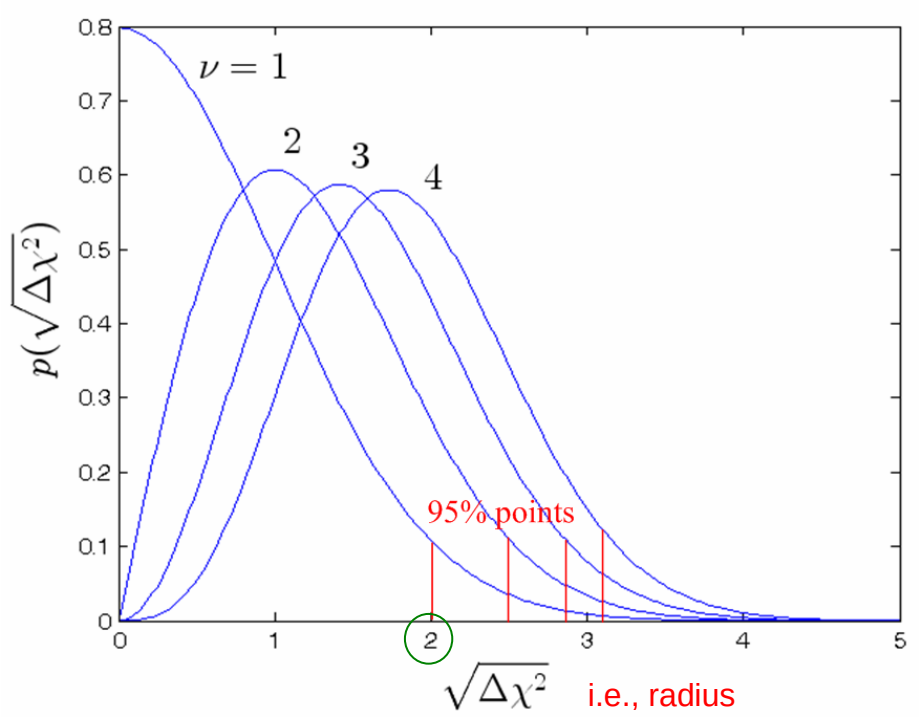


What $\Delta\chi^2$ contour in ν dimensions contains some percentile probability?

Rotate and scale the covariance to make it spherical. Contours still contain same probability. (In equations, this would be another "Cholesky thing".)



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						
	ν					
p	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.