Astronomy 8824: Statistics Notes 4
Fisher Matrix Forecasts and Linear Models

The most relevant section of Ivezic is §4.2.

For the case of uncorrelated errors, many of the points in this section are described well and practically in the "Modeling of Data" chapter of Numerical Recipes (ch. 15 in the 3rd edition).

However, NR doesn't really deal with correlated errors at all. Most of the key results (including some I won't get to) are summarized and derived in Gould (2003, `arXiv:astro-ph/0310577`). This paper is dense, but although it covers "standard" results, they are results that are hard to find in one place anywhere else.

I will present many results in the context of forecasting – predicting the precision of parameter determination from a future experiment.

Gould (2003) comes primarily from the point of view of analyzing data-in-hand, but with some comments about forecasting.

I phrase the discussion below in terms of likelihood, which is most common in the literature, but for a more properly Bayesian formulation one could substitute posterior probability for likelihood throughout.

Single-Parameter Warmup

Suppose we have an observable y_1 that we can predict given some model parameter θ_1, and that we measure y_1 with some observational error $\sigma(y_1)$.

Our best estimate of θ_1 is the value that gives the observed value of y_1.

In the neighborhood of this best fit value $\hat{\theta}_1$, a linear Taylor expansion implies

$$y_1(\theta_1) = y_1(\hat{\theta}_1) + \left(\frac{d\,y_1}{d\theta_1}\right)(\theta_1 - \hat{\theta}_1).$$

Simple "chain rule" error propagation then tells us that the error on θ_1 is

$$\sigma(\theta_1) = \left(\frac{d\,y_1}{d\theta_1}\right)^{-1} \sigma(y_1).$$

Often we are interested in the fractional error

$$\frac{\sigma(\theta_1)}{\theta_1} = \sigma(\ln\theta_1) = \left(\frac{d\ln y_1}{d\ln\theta_1}\right)^{-1} \sigma(\ln y_1).$$

For example, if $y_1 \propto \theta_1^3$, then the fractional error on θ_1 is only 1/3 the fractional error on y_1.

These results break down if the linear Taylor expansion becomes inaccurate over the observationally allowed range of y_1.

In general, the error on a parameter depends on the error on the observable and on the sensitivity of that observable to that parameter. A more sensitive observable gives greater leverage on the parameter.

Fisher Matrix Error Forecasting

Suppose we are considering some future experiment rather than data we have in hand. If we can *predict* what the measurement errors on the data will be, and we know how the data depend on model parameters, then we can forecast how accurately we will be able to constrain parameters.

If we have a parameter vector $\vec{\theta}$, the Fisher information matrix is defined by

$$F_{ij} = -\left\langle \frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j} \right\rangle .$$

The Fisher matrix is thus the expected value of the curvature/Hessian matrix.

To the extent that the likelihood is well described by a quadratic Taylor expansion, the expected error on parameter θ_i is

$$\sigma_i \equiv \sigma(\theta_i) = (F_{ii}^{-1})^{1/2}$$

if all of the parameters are being estimated from the data set and

$$\sigma_i \equiv \sigma(\theta_i) = (F_{ii})^{-1/2}$$

if all parameters other than θ_i are known.

Under more general conditions, the error of any unbiased estimator must be greater than or equal to these values, a result known as the *Cramér-Rao Bound*.

In *Fisher matrix forecasting*, we assume a fiducial model and properties of a data set to predict the Fisher matrix and thereby forecast the errors that will be obtained on model parameters.

There is a pretty good high-level discussion of this in section 2 of Tegmark, Taylor, & Heavens (1997, ApJ, 480, 22) and a valuable but dense presentation in Gould (2003).

Note that a Fisher matrix forecast only gives you accurate error forecasts if the 2nd-order expansion of the likelihood is accurate.

If you're worried this might not be true, then you can use MCMC instead, with your anticipated measurement errors and setting the data equal to the values expected for your fiducial model.

Parameter sensitivity and observational errors

We can decompose a Fisher matrix into a matrix product:

$$F_{ij} = -\left\langle \frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j} \right\rangle = -\left\langle \frac{\partial \Delta y_k}{\partial\theta_i} \cdot \frac{\partial^2 \ln L}{\partial \Delta y_k \partial \Delta y_l} \cdot \frac{\partial \Delta y_l}{\partial\theta_j} \right\rangle ,$$

where $\Delta y_k = y_{\rm mod}(x_k) - y_k$ is the difference between the model prediction for data point k and the observed value, and we are using the Einstein summation convention.

Because the data values do not depend on the model parameters (they are just observed),

$$\frac{\partial \Delta y_k}{\partial \theta_i} = \frac{\partial y_{\rm mod}(x_k)}{\partial \theta_i} \ .$$

As we will show below, *if the errors on the observables are Gaussian and independent of the model parameters,* then

$$-\frac{\partial^2 \ln L}{\partial \Delta y_k \partial \Delta y_l} = C^{-1}_{kl},$$

the inverse covariance matrix.

Thus, the Fisher matrix has an "outer" piece $\partial \vec{y}_{\rm mod}/\partial\vec{\theta}$ that represents the sensitivity of the observables to the parameters and an "inner" piece that represents the errors on the observables themselves.

If we consider the 1-parameter, 1-observable case, we get

$$F_{11} = \frac{dy_{\rm mod}}{d\theta} \cdot \frac{1}{\sigma_y^2} \cdot \frac{dy_{\rm mod}}{d\theta},$$

implying

$$\sigma^2(\theta) = 1/F_{11} = \sigma_y^2 \cdot \left(\frac{dy_{\rm mod}}{d\theta}\right)^{-2},$$

in agreement with our earlier chain rule result.

For a Fisher matrix forecast of parameter errors, we compute the parameter sensitivity from our model, and we take the expected values of the observable errors (and their covariances).

As far as I know, $\partial \vec{y}_{\rm mod}/\partial\vec{\theta}$ doesn't have a special name, but we can think of it as an "influence matrix" or "sensitivity matrix."

While computing the Fisher matrix requires assumptions about the data set, the sensitivity matrix requires only knowledge of the model, and it can be an interesting quantity to compute even if one doesn't have a specific data set in mind.

Fisher matrix for Gaussian Likelihoods

Suppose we have a Gaussian likelihood function for N data points

$$-\ln L = \frac{N}{2}\ln(2\pi) + \frac{1}{2}\ln[\det(\mathbf{C})] + \frac{1}{2}\Delta y_m C^{-1}_{mn}\Delta y_n.$$

(My reason for changing kl indices to mn indices will become evident shortly.)

For the case of a diagonal covariance matrix, $C_{mn} = \sigma_m^2\delta_{mn}$, this expression becomes

$$-\ln L = \frac{N}{2}\ln(2\pi) + \frac{1}{2}\sum \ln\sigma_m^2 + \frac{1}{2}\sum \frac{\Delta y_m^2}{\sigma_m^2},$$

but we will consider the full covariant case.

Assume that we can ignore any dependence of the covariance matrix on the model parameters. *This is a non-trivial assumption that will not always hold.*

For example, in cosmological applications we sometimes have "cosmic variance" errors that depend on the amplitude of matter or galaxy clustering, and the expected size of these errors depends on the cosmological parameters.

However, if we have a data set that provides tight constraints on parameters, then the allowed model dependence of the covariance matrix usually cannot be large.

If we do make this assumption, then the derivative of det$\mathbf{C}$ with respect to parameters vanishes, and C^{-1}_{mn} is independent of Δy_k and Δy_l, allowing us to rearrange the "inner" piece of the Fisher matrix:

$$\begin{aligned}
\frac{1}{2}\frac{\partial^2(\Delta y_m C^{-1}_{mn}\Delta y_n)}{\partial\Delta y_k\partial\Delta y_l} &= \frac{1}{2}\sum_{mn} C^{-1}_{mn}\frac{\partial^2(\Delta y_m\Delta y_n)}{\partial\Delta y_k\partial\Delta y_l} \\
&= \frac{1}{2}\sum_{mn} C^{-1}_{mn}\frac{\partial}{\partial\Delta y_k}\left(\frac{\partial(\Delta y_m\Delta y_n)}{\Delta y_l}\right) \\
&= \frac{1}{2}\sum_{mn} C^{-1}_{mn}\frac{\partial}{\partial\Delta y_k}\left(\Delta y_m\frac{\partial\Delta y_n}{\partial\Delta y_l} + \Delta y_n\frac{\partial\Delta y_m}{\partial\Delta y_l}\right) \\
&= \frac{1}{2}\sum_{mn} C^{-1}_{mn}\frac{\partial}{\partial\Delta y_k}\left(\Delta y_m\delta_{nl} + \Delta y_n\delta_{ml}\right) \\
&= \frac{1}{2}\sum_{mn} C^{-1}_{mn}(\delta_{km}\delta_{nl} + \delta_{kn}\delta_{ml}) \\
&= C^{-1}_{kl}\ .
\end{aligned}$$

On the right-hand sides I have written out sums explicitly for clarity and interchanged sums and derivatives.

This derivation is a bit mathematically loose, but I think it is correct.

Including the "outer" piece, the Fisher matrix is

$$F_{ij} = \frac{\partial\Delta y_k}{\partial\theta_i}C^{-1}_{kl}\frac{\partial\Delta y_l}{\partial\theta_j} = \frac{\partial y_{\rm mod}(x_k)}{\partial\theta_i}C^{-1}_{kl}\frac{\partial y_{\rm mod}(x_l)}{\partial\theta_j}.$$

Though notationally different, I think this is equivalent to equation (15) of Tegmark et al. (1997), except that the term $\mathbf{A}_i\mathbf{A}_j$ in that equation has vanished because we have assumed that the dependence of C_{ij} on the parameters can be neglected.

Straight Line Model

Now consider a model $y_{\rm mod}(x) = \theta_1 + \theta_2 x$.

The derivatives are

$$\frac{\partial \Delta y_k}{\partial \theta_1} = 1, \qquad \frac{\partial \Delta y_k}{\partial \theta_2} = x_k,$$

making the 2×2 Fisher matrix

$$F_{ij} = \begin{pmatrix} \sum C_{kl}^{-1} & \sum C_{kl}^{-1} x_k \\ \sum C_{kl}^{-1} x_k & \sum C_{kl}^{-1} x_k x_l \end{pmatrix}.$$

This matrix can be inverted recalling that for

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \qquad A^{-1} = \frac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

The errors on the intercept and slope are, respectively, $(F_{11}^{-1})^{1/2}$ and $(F_{22}^{-1})^{1/2}$.

For a diagonal covariance matrix $C_{kl} = \delta_{kl}\sigma_k^2$,

$$F_{ij} = \begin{pmatrix} \sum \sigma_k^{-2} & \sum x_k \sigma_k^{-2} \\ \sum x_k \sigma_k^{-2} & \sum x_k^2 \sigma_k^{-2} \end{pmatrix} = \frac{N}{\sigma^2} \begin{pmatrix} 1 & \langle x \rangle \\ \langle x \rangle & \langle x^2 \rangle \end{pmatrix},$$

where the second equality is for homoscedastic errors $\sigma_k = \sigma$.

Inverting the last case yields

$$F_{ij}^{-1} = \frac{\sigma^2}{N} (\langle x^2 \rangle - \langle x \rangle^2)^{-1} \begin{pmatrix} \langle x^2 \rangle & -\langle x \rangle \\ -\langle x \rangle & 1 \end{pmatrix}.$$

This is the same as the top equation on p. 5 of Gould (2003). Here F_{ij}^{-1} refers to the forecast covariance matrix of the parameter errors, and $\langle x \rangle$ and $\langle x^2 \rangle$ refer to expected properties of the data set. In Gould (2003), the $\langle ... \rangle$ averages are over the actual data points obtained, and the result is the actual covariance matrix of the parameter errors.

χ^2-minimization for a general linear model

My discussion here follows that of Gould (2003) but with different notation.

Our analysis of the straight-line model can be generalized to the fit of a model that is linear in the parameters θ_i,

$$y_{\rm mod}(x) \equiv \sum_{i=1}^{n} \theta_i f_i(x),$$

where the $f_i(x)$ are specified functions.

Note that the $f_i(x)$ do not need to be linear, e.g., we could have

$$y_{\rm mod}(x) = \theta_1 + \theta_2 x + \theta_3 x^2 + \theta_4 \sin(2\pi x).$$

Again defining $\Delta y_k \equiv y_k - y_{\rm mod}(x_k)$, we now have

$$\frac{\partial \Delta y_k}{\partial \theta_i} = f_i(x_k).$$

Therefore, for a Gaussian likelihood function,

$$F_{ij} = f_i(x_k) C^{-1}_{kl} f_j(x_l).$$

As before,

$$\sigma_{ij} = F^{-1}_{ij}$$

is the expected covariance matrix of the parameter errors, with $(F^{-1}_{ii})^{1/2}$ the expected error on parameter θ_i if all parameters must be estimated from the data.

This is equivalent to the result on pp. 3 and 4 of Gould (2003), with the notational translations

$$\mathcal{B}_{kl} = C^{-1}_{kl} \qquad b_{ij} = F_{ij} \qquad c_{ij} = F^{-1}_{ij}.$$

Importantly, Gould also derives the solution for the minimum χ^2 (maximum likelihood) values of the parameters by requiring $\partial\chi^2/\partial\theta_i = 0$.

The result is

$$\hat{\theta}_i = F^{-1}_{ij} \left[y_k C^{-1}_{kl} f_j(x_l) \right] \quad (= c_{ij} d_j \text{ in Gould's notation}),$$

where there is an implicit sum over k, l inside the [..], and a sum over j.

This is a general result for χ^2 fitting of a model that is linear in the parameters θ_i, with F_{ij} defined as $f_i(x_k) C^{-1}_{kl} f_j(x_l)$.

For a diagonal covariance matrix $C^{-1}_{kl} = \sigma_k^{-2} \delta_{kl}$,

$$F_{ij} = \sum_{k=1}^{n} \frac{f_i(x_k) f_j(x_k)}{\sigma_k^2}$$

and

$$y_k C_{kl}^{-1} f_j(x_l) = \sum_{k=1}^{n} \frac{y_k f_j(x_k)}{\sigma_k^2}.$$

As previously emphasized, a diagonal covariance matrix does *not* imply a diagonal Fisher matrix. One can have independent data points but still have correlated parameter errors, and vice versa.

For χ^2-minimization of a general linear model, one can find best-fit parameter values and the covariance matrix of parameter errors "analytically" (numerical matrix inversions may be required).

Gould (2003) also gives solutions for cases where one imposes constraints on the parameters (e.g., $\theta_1 + 2\theta_2 = 0$).

Illustration for a straight line

If we adopt a diagonal covariance matrix and further specify a straight-line model, $f_1(x) = 1$, $f_2(x) = x$, we obtain our earlier result for the Fisher matrix, but I will now adopt notation from Numerical Recipes section 15.2:

$$F_{ij} = \begin{pmatrix} S & S_x \\ S_x & S_{xx} \end{pmatrix},$$

with the inverse-variance weighted sums

$$S \equiv \sum \sigma_k^{-2}, \quad S_x \equiv \sum x_k \sigma_k^{-2}, \quad S_{xx} = \sum x_k^2 \sigma_k^{-2}.$$

The inverse Fisher matrix is

$$F_{ij}^{-1} = \frac{1}{S\,S_{xx} - S_x^2} \begin{pmatrix} S_{xx} & -S_x \\ -S_x & S \end{pmatrix}.$$

The vector $d_j \equiv y_k C_{kl}^{-1} f_j(x_l)$ is (d_1, d_2) with

$$d_1 = \sum y_k \sigma_k^{-2} \equiv S_y, \quad d_2 = \sum x_k y_k \sigma_k^{-2} \equiv S_{xy}.$$

The minimum-χ^2 solution is then

$$\theta_1 = F_{11}^{-1} d_1 + F_{12}^{-1} d_2 = \frac{S_{xx} S_y - S_x S_{xy}}{S S_{xx} - S_x^2}$$
$$\theta_2 = F_{21}^{-1} d_1 + F_{22}^{-1} d_2 = \frac{S S_{xy} - S_x S_y}{S S_{xx} - S_x^2},$$

in agreement with equation 15.2.6 of NR.

Expanding a non-linear model

Suppose that our model is a *non-linear* function of our parameters, but we know that the correct parameters are small perturbations about a fiducial model with parameters $\theta_{i,\rm fid}$.

In this case, we can make a Taylor expansion

$$y_{\rm mod}(x_k) = y_{\rm mod,fid}(x_k) + \Delta\theta_i \frac{\partial y_{\rm mod}(x_k)}{\partial \theta_i},$$

where $\Delta\theta_i = \theta_i - \theta_{i,\rm fid}$ and the derivative is evaluated for the fiducial values of all parameters.

This is now a linear model with parameters $\Delta\theta_i$ instead of θ_i and

$$f_i(x_k) = \frac{\partial y_{\rm mod}(x_k)}{\partial \theta_i}.$$

We can use this expansion to fit parameter values or compute parameter errors or forecast errors provided that the errors are small enough that the linear Taylor expansion remains accurate.

By definition, this expansion holds exactly for a true linear model.

Most Fisher matrix forecasts implicitly assume this kind of linear expansion around a fiducial model, so they will give accurate forecasts of parameter errors only to the extent that the linear expansion is accurate over the range of parameters allowed by the data.

An MCMC forecast does not rely on this linear approximation.

Adding Fisher matrices

Suppose we have two data sets that are statistically independent.

In this case, the joint likelihood (or posterior probability) is just the product of the individual likelihoods (or posterior probabilities), since $p(x,y) = p(x)p(y)$ for independent variables.

Therefore, one obtains $\langle \ln L \rangle$ for the two data sets by adding the two individual values of $\langle \ln L \rangle$, and the Fisher matrix for the two data sets is just the sum of the Fisher matrices for the individual data sets.

This still holds even if the data sets are quite different in character provided they constrain the same underlying parameters.

For example, one can forecast cosmological parameter errors that will be obtained by joint fits to CMB data, supernova data, and a direct measurement of H_0 by adding the Fisher matrices for the three data sets.

This is a powerful technique.

Note that Fisher information scales like an inverse variance, $F \propto \sigma^{-2}$, and information from independent data sets adds linearly.