Astronomy 8824: Statistics Notes 3
Correlated errors, χ^2, Gaussian Likelihood, MCMC

Bivariate and Multivariate Gaussians

(Ivezic §§3.5.2-3.5.4)

Suppose we have two independent variables x and y drawn from Gaussian distributions of width σ_x and σ_y. The joint distribution $p(x,y) = p(x)p(y)$ is a bivariate Gaussian, and the values of x and y are uncorrelated:

$$\langle (x-\mu_x)(y-\mu_y)\rangle = 0.$$

If we now consider

$$x' = x\cos\alpha - y\mathrm{sin}\alpha$$
$$y' = x\mathrm{sin}\alpha + y\cos\alpha$$

then we "rotate" the distribution by angle α. The distribution $p(x',y')$ is still a bivariate Gaussian, but now the values of x' and y' are correlated.

If we have a number of random variables y_i, $i = 1...M$, which we combine into a vector $\mathbf{y}$, then the covariance matrix is

$$C_{ij} = \langle (y_i - \langle y_i\rangle)(y_j - \langle y_j\rangle)\rangle\ .$$

If the distribution $p(\mathbf{y})$ is a multivariate Gaussian then

$$p(\mathbf{y}) = \frac{1}{(2\pi)^{M/2}\sqrt{\det(\mathbf{C})}}\exp\left(-\frac{1}{2}\Delta y_i C^{-1}_{ij}\Delta y_j\right)\ ,$$

where $\Delta y_i = y_i - \langle y_i\rangle$, C^{-1}_{ij} is the inverse covariance matrix and I have used the Einstein summation convention: repeated indices (i,j in this case) are automatically summed over.

This can also be written in vector/matrix notation.

Correlated Errors: Observables and Parameters

Sometimes the errors on data points are correlated.

For example, there may be a calibration uncertainty that affects many data points in the same way. For galaxy clustering statistics, measurement errors at different scales are usually correlated.

Even if the errors on data points are uncalibrated, the errors on *parameters* derived from a multi-parameter fit to the data (e.g., the slope and amplitude of a line) are often correlated, unless one has deliberately constructed parameters that have uncorrelated errors.

It is also possible to have correlated errors on data and uncorrelated errors on parameters, though this is less generic than the reverse case.

Gaussian Likelihoods and χ^2

If we have uncorrelated, Gaussian errors on observables y and a model that predicts $y_k = y_{\rm mod}(x_k)$ then the likelihood is $L \propto e^{-\chi^2/2}$ where

$$\chi^2 = \sum_k \frac{(\Delta y_k)^2}{\sigma_k^2}$$

with $\Delta y_k = y_k - y_{\rm mod}(x_k)$.

However, if the errors are correlated then we instead have

$$\chi^2 = \Delta y_k C_{kl}^{-1} \Delta y_l.$$

The two definitions coincide for a diagonal covariance matrix $C_{kl} = \sigma_k^2 \delta_{kl}$, in which case $C_{kl}^{-1} = \delta_{kl}\sigma_k^{-2}$.

Parameter Errors in a Maximum Likelihood (or MAP) Estimate

(Ivezic section 4.2.5.)

For a Gaussian probability distribution $p(x) = (2\pi\sigma^2)^{-1/2} e^{-(x-\mu)^2/2\sigma^2}$,

$$\ln p = -\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2} + \text{const.}$$

Suppose we have estimated a parameter θ by maximizing either the likelihood L or the posterior probabiliity L_p. The first derivative vanishes at the maximum, so a Taylor expansion gives

$$\ln L \approx \ln L_0 + \frac{1}{2}\left(\frac{\partial^2 \ln L}{\partial \theta^2}\right)(\theta - \theta_0)^2 \ ,$$

where θ_0 is the location of the maximum.

Identifying the two equations, we infer that if $L(\theta)$ is adequately described by this Taylor expansion, the 1σ error on θ is

$$\sigma_\theta = \left(-\frac{\partial^2 \ln L}{\partial \theta^2}\right)^{-1/2} ,$$

where the derivative is evaluated at the maximum.

For the more general case of a vector of parameters θ_i, we can define the second-derivative matrix

$$H_{jk} = -\frac{\partial^2 \ln L}{\partial \theta_j \partial \theta_k},$$

which is sometimes called the Hessian matrix or curvature matrix (though terminology and notation are not standard).

One can approximate the log-likelihood as a multi-dimensional paraboloid near its maximum, to find that the likelihood itself is a multi-dimensional Gaussian with covariance matrix

$$\text{Cov}(\theta_j, \theta_k) = \sigma_{jk} = H^{-1}_{jk}$$

Here $(\sigma_{ii})^{1/2}$ is the error on parameter θ_i marginalized over uncertainties in other parameters.

If $\sigma_{jk} \neq 0$ for some $j \neq k$ then the uncertainties on parameters θ_j and θ_k are correlated.

The expression in Ivezic equation (4.6) is different (it takes the reciprocal of matrix elements instead of the inverse of the matrix), but I think it is incorrect for the case where H_{jk} is not diagonal (and I think I even confirmed this with Ivezic). When H_{jk} is diagonal, the two definitions are the same.

I have phrased this discussion in terms of likelihood, but it could equally well be phrased in terms of posterior probability: the log of the posterior probability can also be approximated as a paraboloid about its maximum, and one would just substitute $P_{\text{posterior}}$ for L in the expressions.

Notational caution: Whenever I write A^{-1}_{jk} I mean the jk element of the inverse of matrix A, not the reciprocal of the jk element of A, which I would write $(A_{jk})^{-1}$.

Monte Carlo Markov Chains

A fairly common statistical problem is estimating the probability distribution of parameters in a high-dimensional parameter space.

If the 2nd-order expansion described above is adequate, then one "just" needs to find the maximum likelihood solution and compute the second-derivatives of the likelihood with respect to the parameters.

But sometimes this approximation isn't adequate – a rule-of-thumb that doesn't always work is that the parabolic approximation is good when the fractional errors on *all* of the parameters are small.

One option is to grid the parameter space finely and compute the posterior probability at all grid locations within it. Marginal distributions can be computed by summing over axes.

This approach is robust and therefore shouldn't be ignored, but it is often computationally impractical.

For example, we might be trying to determine the constraints from a CMB data set D on the set of cosmological parameters $\vec{\theta} = (\Omega_m, h, \Omega_b, A, n, \tau)$ that determines the CMB spectrum in the simplest current cosmological scenario.

There are tools for calculating $p(D|\vec{\theta}I)$, but this calculation might take a few seconds, or minutes, for each model in the parameter space.

Since the parameter space is six-dimensional, even a relatively coarse grid with 10 points along each parameter direction over the plausible range requires 10^6 evaluations of $p(D|\vec{\theta}I)$, and if we add another two parameters then 10^6 becomes 10^8.

Thus, a naive grid-based evaluation of the likelihood to find best-fit parameters and error bars may be prohibitively expensive.

Monte Carlo Markov Chains (MCMC) are a useful tool for this kind of problem, and this approach has taken rapid hold in the cosmology literature.

In effect, MCMC is doing the necessary integrals for marginalization by Monte Carlo integration.

For details, see the references listed below and the things that they in turn refer to, but in brief the idea is as follows.

The goal is to map the posterior probability distribution $p(\vec{\theta}|DI) \propto p(\vec{\theta}|I)p(D|\vec{\theta}I)$, in the neighborhood of its maximum value.

If the prior $p(\vec{\theta}|I)$ is flat, then we just have $p(\vec{\theta}|DI) \propto L$.

Procedure:

1. Start from a randomly chosen point in the parameter space, $\vec{\theta} = \vec{\alpha}_1$.

2. Take a random step to a new position $\vec{\alpha}_2$.

3. If $p(\vec{\alpha}_2|DI) \geq p(\vec{\alpha}_1|DI)$, "accept" the step: add $\vec{\alpha}_2$ to the chain, and substitute $\vec{\alpha}_2 \rightarrow \vec{\alpha}_1$. Return to step 2.

4. If $p(\vec{\alpha}_2|DI) < p(\vec{\alpha}_1|DI)$, draw a random number x from a uniform distribution from 0 to 1. If $x < p(\vec{\alpha}_2|DI)/p(\vec{\alpha}_1|DI)$, "accept" the step and proceed as in 3. If $x \geq p(\vec{\alpha}_2|DI)/p(\vec{\alpha}_1|DI)$, reject the step. Save $\vec{\alpha}_1$ as another (repeated) link on the chain, and go back to 2.

The chain takes some time to "burn in," i.e., to reach the neighborhood of the most likely solutions.

However, once this happens, a "long enough" chain will have a density of points that is proportional to $p(\vec{\theta}|DI)$.

To get, for example, the joint pdf of a pair of parameters, one can just make contours of the density of points in the space of those two parameters. Other "nuisance" parameters are marginalized over automatically, because the points sample the full space.

If you want to calculate the posterior distribution of some *function* of the parameters (e.g., the age of the Universe, given parameter estimates from the CMB), you can just calculate that function for all points in the chain, then plot the pdf of the result.

There are numerous technical issues related to determining whether a chain has "converged" (i.e., is fairly sampling the probability distribution), and to choosing steps in a way that produces fast convergence and good "mixing" (sampling the distribution fairly with a relatively small number of points).

There is an increasingly extensive literature on MCMC methods. Some starting points are: Sections 5.8.1 and 5.8.2 of Ivezic et al., and section 15.8 of the 3rd edition of Numerical Recipes, though this topic wasn't in the 1st or 2nd edition.

An exceedingly useful and enjoyably written reference is Hogg & Foreman-Mackey (2017, arXiv:1710.06068). Another that goes a bit further in introducing more advanced methods is Sharma (2017, ARAA, 55, 213).

Properly implemented, MCMC should sample tails or multiple modes of a distribution that are not well described by the Gaussian approximation.

However, if the Gaussian approximation is adequate, then MCMC is not a computationally efficient way to find the parameter PDF.