MODEL SELECTION: BAYESIAN AND FREQUENTIST METHODS

Carl Heiles, November 8, 2011

Before getting started let's make some definitions:

1. We distinguish between *probabilities*, denoted by the letter P, and *probability density functions*, denoted by the Greek letters ψ or ϕ. We also use what we call a "probability function" ρ, which is like a PDF except that it is not normalized, so that its integral is not equal to unity.

2. The subscript $*$ means the true values. The subscript d means the least-squares value derived from the data. (They would be equal if there were no noise).

3. We have M datapoints, d_m. The set of M datapoints is indicated by curly brackets, i.e. $\{d\}$.

4. To make things specific, we'll assume Gaussian noise for the data.

1. PAIR OF SAMPLE MODELS

We have a dataset with M data points d_m taken at times t_m. We want to compare two models, each of which has a single parameter:

1. Model A describes the data by a constant, μ_*. We estimate μ_* from the mean of the d_m, i.e. $\mu_d = \frac{\sum(d_m)}{M}$.

2. Model B describes the data by a slope λ_* with no zero offset. We estimate λ_* from the least squares solution, which gives $\lambda_d = \frac{\sum t_m d_m}{\sum t_m^2}$.

2. BAYESIAN APPROACH

This is excerpted from the beginning of chapter 4 in Sivia.

In principle, Bayes' theorem allows one to specify the probability that a model is correct. Specifically, Bayes' theorem says for model A and data d that

$$P(A \mid d) = \frac{P(d \mid A)}{P(d)} P(A) \tag{1}$$

If one has a model A, then one has to be able to calculate $P(d \mid A)$; this is central to fitting data to a model. Similarly, one ought to have some idea of the prior $P(A)$. So all one needs to know is

$P(d)$, which is called the "evidence". Unfortunately, as we discuss below in §6, it is unusual to be able to know $P(d)$. So we can hardly ever calculate $P(A \mid d)$.

Should we give up? NO!! We can still use Bayes' theorem to compare pairs of models by taking the ratio, because $P(d)$ cancels out and we get

$$\mathcal{R} = \frac{P(A \mid \{d\}, \sigma)}{P(B \mid \{d\}, \sigma)} = \frac{P(\{d\} \mid A, \sigma)}{P(\{d\} \mid B, \sigma)} \, \frac{P(A)}{P(B)} \tag{2}$$

Here we've included the noise level σ for the data as a specified parameter.

We need the likelihood probabilities for the data $[P(\{d\} \mid A, \sigma), P(\{d\} \mid B, \sigma)]$. However, when we ask ourselves what these symbols mean, we think: "The probability of obtaining a given piece of data d_m depends not only on its associated σ_m, but it also depends on the true value of what we're measuring." That is, our brain is telling us that we need to write $[P(\{d\} \mid A, \mu_*, \sigma), P(\{d\} \mid B, \lambda*, \sigma)]$ instead of $[P(\{d\} \mid A, \sigma), P(\{d\} \mid B, \sigma)]$. But doing this puts us in the following quandary: we want to which model (A or B) is better, *independent* of the numerical values of the true model parameters (μ_* or λ_*)), because *we don't know what these true values are*—we only have *estimates* for them in the form of the least-square fit results μ_d and λ_d.

What should we do? Let's consider Model A. The needed likelihood probability $P(\{d\} \mid A, \sigma)$ is for all possible values of $\mu*$, not just any assumed one or the most likely one from our data (i.e., μ_d). To obtain its probability $P(\{d\} \mid A, \sigma)$ we need to consider the PDF $\phi(\{d\} \mid A, \mu*, \sigma)$ and marginalize (integrate) over $\mu*$. But we can't marginalize over the *given* quantity μ_* (i.e., we can't marginalize over a quantity on the right-hand side of the " | "); rather, we can only marginalize over a variable in a joint PDF $\phi(\{d\}, \mu_* \mid A, \sigma)$.

That is, we need to evaluate

$$P(\{d\} \mid A, \sigma) = \int \phi(\{d\}, \mu_* \mid A, \sigma) d\mu_* \tag{3}$$

which means that, first, we need to obtain the joint PDF inside the integral from the product rule:

$$\phi(\{d\}, \mu_* \mid A, \sigma) = \phi(\{d\} \mid A, \mu_*, \sigma) \times \psi(\mu_* \mid A) \tag{4}$$

The first term on the right-hand hand side contains μ_* as a given quantity, However, we don't know μ_*, i.e. we don't know $\phi(\{d\} \mid A, \mu_*, \sigma)$; we only know $\phi(\{d\} \mid A, \mu_d, \sigma)$. We relate these using the following, which is pretty obscure but will become clear when we do the Gaussian statistics example:

$$\phi(\{d\} \mid A, \mu_*, \sigma) = P(\{d\} \mid A, \mu_d, \sigma) \times \frac{\phi(\mu_* \mid A, \mu_d, \delta\mu_d)}{\phi(\mu_* \mid A, \mu_d, \delta\mu_d)|_{\mu_* = \mu_d}} \tag{5}$$

The term in the denominator, $\phi(\mu_* \mid A, \mu_d, \delta\mu_d)|_{\mu_*=\mu_d}$, is a PDF evaluated at the specific value $\mu_* = \mu_d$. Thus, it is a probability, not a PDF, and we could have written $P(\mu_* = \mu_d \mid A, \mu_d, \delta\mu_d)$ (but without this little remark this wouldn't be very clear). Let's define the "probability function" ρ:

$$\rho(\mu_* \mid A, \mu_d, \delta\mu_d) = \frac{\phi(\mu_* \mid A, \mu_d, \delta\mu_d)}{\phi(\mu_* \mid A, \mu_d, \delta\mu_d)|_{\mu_*=\mu_d}} \tag{6}$$

This is like a PDF because it expresses how things change with μ_*. However, its integral over μ_* isn't equal to unity, so it's not a PDF. In particular, its maximum value, which occurs at $\mu_* = \mu_d$, is unity. Plugging all this into equation 4, we get

$$\phi(\{d\}, \mu_* \mid A, \sigma) = P(\{d\} \mid A, \mu_d, \sigma) \times \rho(\mu_* \mid A, \mu_d, \delta\mu_d) \times \psi(\mu_* \mid A) \tag{7}$$

Now we can integrate over μ_*:

$$P(\{d\} \mid A, \sigma) = P(\{d\} \mid A, \mu_d, \sigma) \times \int \underbrace{\rho(\mu_* \mid A, \mu_d, \delta\mu_d)}_{first\ term} \times \underbrace{\psi(\mu_* \mid A)}_{prior}\ d\mu_* \tag{8}$$

and similarly for Model B

$$P(\{d\} \mid B, \sigma) = P(\{d\} \mid B, \lambda_d, \sigma) \times \int \rho(\lambda_* \mid B, \lambda_d, \delta\lambda_d) \times \psi(\lambda_* \mid B)\ d\lambda_* \tag{9}$$

We could now plug equations 8 and 9 into equation 2, which would result in complicated-looking thing, which could be evaluated numerically given any kind of statistics.

However, let's assume the usual Gaussian statistics with variance σ^2 for the data, which makes things lots easier:

$$P(d_m \mid A, \mu_d, \sigma_m) = \frac{1}{\sqrt{2\pi}\sigma_m} \exp - \left(\frac{(d_m - \mu_d)^2}{2\sigma_m^2} \right) \tag{10}$$

The probability for the set of data $\{d\}$ is the product, so this gives for the term in front of the integral in equation 8

$$P(\{d\} \mid A, \mu_d, \sigma) = \prod_m P(d_m \mid A, \mu_d, \sigma_m) \tag{11}$$

Given the Gaussian statistics for the data, the PDF for μ_* is also Gaussian. Having applied least squares for Model A, the data provide us the best fit for μ_*, which we denote μ_d; and its uncertainty, $\delta\mu_d$. These define the posterior PDF for μ_*:

$$\phi(\mu_* \mid A, \mu_d, \delta\mu_d) = \frac{1}{\sqrt{2\pi}\ \delta\mu_d} \exp - \left(\frac{(\mu_* - \mu_d)^2}{2\delta\mu_d^2} \right) \tag{12}$$

Dividing this by the function evaluated at $(\mu_* = \mu_d)$ gives

$$\rho(\mu_* \mid A, \mu_d, \delta\mu_d) = \exp - \left(\frac{(\mu_* - \mu_d)^2}{2\delta\mu_d^2} \right) \tag{13}$$

This is the "first term" under the integral in equation 8. For the "prior" term in equation 8, let's assume a uniform prior $\phi(\mu_* \mid A) =$ constant between (μ_{min}, μ_{max}) and zero elsewhere:

$$\phi(\mu_* \mid A) = \frac{1}{\mu_{max} - \mu_{min}} = \frac{1}{\Delta\mu} \tag{14}$$

We substitute these into equation 8 and obtain (finally!)

$$P(\{d\} \mid A, \sigma) = \frac{P(\{d\} \mid A, \mu_d, \sigma)}{\Delta\mu} \int_{\mu_{min}}^{\mu_{max}} \exp - \left(\frac{(\mu_* - \mu_d)^2}{2\delta\mu_d^2} \right)\ d\mu_* \tag{15}$$

or

$$P(\{d\} \mid A, \sigma) = P(\{d\} \mid A, \mu_d, \sigma)\ \sqrt{2\pi}\ \frac{\delta\mu}{\Delta\mu} \tag{16}$$

Here we have assumed $\Delta\mu \gg \delta\mu$, so we can take the limits of the integral as $\pm\infty$. For the ratio of equation 2, we have

$$\mathcal{R} = \frac{P(A \mid d)}{P(B \mid d)} = \frac{P(A)}{P(B)}\ \frac{P(\{d\} \mid A, \mu_d, \sigma)}{P(\{d\} \mid B, \lambda_d, \sigma)}\ \frac{\delta\mu}{\Delta\mu}\ \frac{\Delta\lambda}{\delta\lambda} \tag{17}$$

Some comments:

1. The ratio is scale-invariant, as it must be for it to make sense.

2. In equation 15 we assumed $\Delta\mu \gg \delta\mu_d$, i.e. that the experimental errors are small compared to the range over which the prior is nonzero. If this weren't the case, the integral in equation 15 would be an error function and would depend on the limits. This implicit assumption is the useful case, because if the experimental errors are large then the experiment cannot provide any additional information to what the prior already provides.

3. The first ratio on the right hand side of equation 17 is the ratio of priors for the two models. Typically, you don't prefer one over the other so this ratio is unity. However, you might have

prior information that favors one over the other (like your girlfriend invented one and your ex-girlfriend the other)...

4. The second ratio on the right hand side of equation 17 is the ratio of the likelihoods. Each likelihood is the product of M terms, where M is the number of datapoints d_m. Because of Gaussian statistics, each datapoint d_m has a probability $P(d_m \mid A, \mu_d, \sigma)$ given by equation 10. To obtain the likelihood, we multiply all the $P(d_m \mid A, \mu_d, \sigma)$, as in equation 11. In practice, and also in the analytical development of maximum likelihood, we take the log of each term, add, and then exponentiate; the result is, of course, identical. The likelihood is the sum of the log of the $P(d_m \mid A, \sigma_m, \mu_d)$, so apart from constants like $\sqrt{2\pi}$ we have

$$\ln P(\{d\} \mid A, \mu_d, \sigma) \propto \sum_m \frac{(d_m - \mu_d)^2}{2\sigma_m^2} = \frac{\chi^2(A, \mu_d)}{2} \tag{18}$$

5. Let's plug equation 18, evaluated for both Models A and B, into equation 17:

$$\mathcal{R}_{\mathcal{A}/\mathcal{B}} = \frac{P(A \mid d, \sigma)}{P(B \mid d, \sigma)} = \frac{P(A)}{P(B)} \exp\left(\frac{-\chi^2(A, \mu_d) + \chi^2(B, \lambda_d)}{2}\right) \frac{\delta\mu}{\Delta\mu} \frac{\Delta\lambda}{\delta\lambda} \tag{19}$$

To summarize, we can write equation 17 in words:

$$\mathcal{R} = [ratio\ of\ priors] \times [ratio\ of\ \exp(-\frac{\chi^2}{2})] \times [ratio\ of\ (\delta params/\Delta params)] \tag{20}$$

3. NUMERICAL EXAMPLE

Figure 1 shows our numerical experiment for the above. We chose model B for the example. We tried two numerical experiments, distinguished only by the different values of $\sigma_{data} = [1.0, 10.0]$. We used 64 data points. The reduced chi-squares for Model A are about 1.4 and 10 for these two experiments, while those for Model B are both 1.4. The chi-squares are these numbers multiplied by 63, so they are large! Calculating $\mathcal{R}$ from equation 19 involves exponentiating these large numbers, which exceed single precision float ranges. So we can only calculate using logarithms.

For $\sigma_d = 1.0$, we find

$$\ln(\mathcal{R}_{\mathcal{A}/\mathcal{B}})_{\sigma_{data}=1} = -263 \ \ ; \ \ \mathcal{R}_{\mathcal{A}/\mathcal{B}} \sim 10^{-114} \tag{21}$$

This means that the probability of model A, the constant fit, being correct is totally negligible, as it should be for this visually-obvious example!

For $\sigma_d = 10.0$, we find

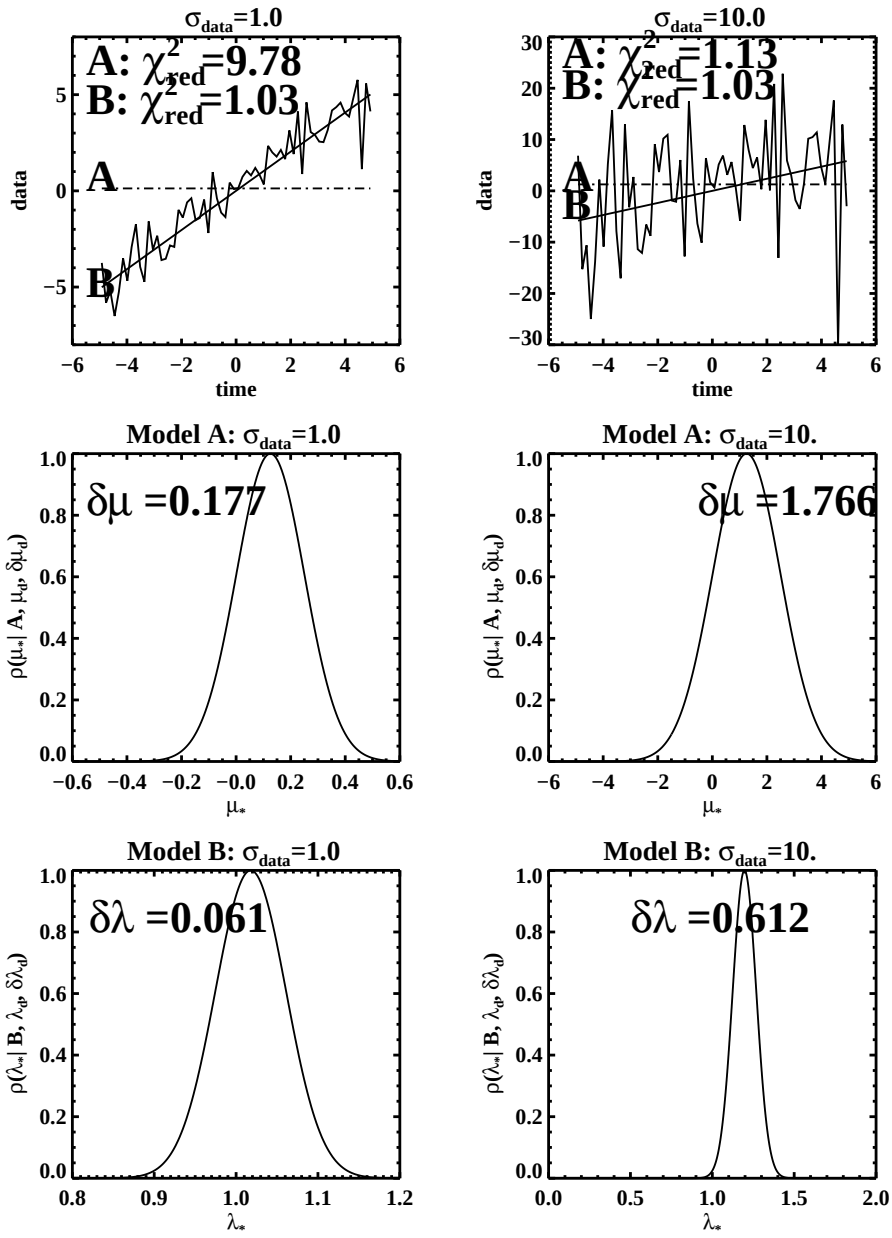

Fig. 1.— *Top panels:* Data versus time for two values of σ_{data} whose ratio is 10. *Columns:* The low $\sigma_{data} = 1$ case is the left column of panels. We consider two model fits, A (a constant μ—the dash-dot line in the top panels) and B (a slope λ—the solid line in the top panels). The reduced χ^2 for the two fits and the two cases of σ_{data} are noted in the top panels. *Middle row of panels:* The probability function $\rho(\mu_* \mid A, \mu_d, \delta\mu_d)$ for the two σ_{data} cases. Because of the Gaussian statistics, each probability function is a Gaussian with dispersion $\delta\mu$, which is noted on the plots. *Bottom row:* As for the middle row, but for Model B and λ_*.

$$\ln(\mathcal{R}_{\mathcal{A}/\mathcal{B}})_{\sigma_{data}=10} = -1.2 \;\; ; \;\; \mathcal{R}_{\mathcal{A}/\mathcal{B}} \sim 0.31 \tag{22}$$

So with low signal/noise, we find that the probability of model A, the constant fit, being correct is 0.31 for this particular trial. Other trials, with different noise, typically find smaller probabilities.

The lesson here is that as we reduce σ_d by accumulating more data, or obtaining higher-quality data, the exponential term in equation 19 gets rapidly smaller, favoring model B.

Apart from the χ^2, the favoring of a model is proportional to its ratio $\frac{\delta\mu}{\Delta\mu}$. If the model's prior allows a wide possible range for the parameter, it is less favored; this makes sense, because it means that the model isn't very specific regarding its predictions.

4. CONVENTIONAL (FREQUENTIST) STATISTICS: THE F-TEST

With conventional statistics, models are compared using the F-test. The F-test uses $\mathcal{R}_{\widehat{\chi^2}}$, the ratio of the two *reduced* chi-squares for the best fit and, also, the degree of freedom ($\sim$ number of datapoints) for each fit. It does not marginalize over the fitted parameters as is done in the Bayesian approach.

The relevant statistic f is the probability (fraction of cases) for which the ratio is smaller than the specified value; f ranges from 0 to 1. If the f statistic is small, then the probability of the ratio being smaller than the observed ratio is itself small, meaning it's hard to do better, so that the numerator of the ratio is a better fit than the denominator.

For our case $\sigma_{data} = 1.0$, $\mathcal{R}_{\widehat{\chi^2}} \sim 7.2$. We find $f = 1.0$; it can't get any larger! This means it's hard to do worse than we observed, meaning model A is terrible with respect to model B. In contrast, for our case $\sigma_{data} = 10.0$, $\mathcal{R}_{\widehat{\chi^2}} \sim 1.05$, which yields $f = 0.58$. So 58% of the time we would find a larger ratio, meaning B is favored over A, but not by much..

To do the f-test In IDL, you use the function `f_pdf`. Also see Numerical Recipes and Bevington.

5. COMPARING PAIRS OF MODELS WITH MULTIPLE PARAMETERS

5.1. Bayesian Approach

The above refers to models having a single parameter. For models with multiple parameters, you must integrate over all of them in equation 8 above. If the statistics are Gaussian, things are simpler: first use least squares to find the combination of parameters that maximize the likelihood; above, this is equivalent to finding (μ_d, λ_d). Then do the multidimensional integral as in equation 8. In a multiparameter problem, it helps to choose parameters that are orthogonal over the interval of the data because then you have a series of one-parameter integrals instead of a multiparameter

integral.

5.2. Frequentist Approach

You use the best fit results—no marginalization over the fitted parameters. This is easier.

6. THE EVIDENCE $P(d)$

See Sivia, page 88. In the introduction, we mentioned the importance and difficulty of the factor $P(d)$ in Bayes' theorem. By treating pairs of models, we managed to ignore it. What is this factor? It expresses the probability of obtaining the given data, regardless of the model. Suppose we have N models; for concreteness in the discussion, let's take $N = 3$, the models being A,B,C.

The models A,B,C are mutually exclusive, and by assumption there are no other possibilities. One, or some combination, must be correct, so we must have

$$P(A) + P(B) + P(C) = 1. \tag{23}$$

or, since we have the data,

$$P(A \mid d) + P(B \mid d) + P(C \mid d) = 1 \tag{24}$$

From Bayes' theorem,

$$P(A \mid d) = \frac{P(d \mid A)}{P(d)} P(A) \tag{25}$$

etc., so we can write

$$\frac{P(d \mid A)P(A) + P(d \mid B)P(B) + P(d \mid C)P(C)}{P(d)} = 1 \tag{26}$$

We can calculate all ratios such as $\frac{P(A \mid d)}{P(B \mid d)}$, so we have enough information to calculate the actual values $P(A)$ (same as $P(A \mid d)$). So, if we know the priors $P(A)$, etc., we can calculate $P(d)$. Eureka—we know what's going on in absolute terms!

But wait: suppose there's another model, D, which we haven't considered and don't know about. Then our calculations are all for naught. We can never be sure that D doesn't exist, and that D isn't, in fact, the correct model. So obtaining actual probabilities like $P(A)$ is essentially impossible. The one exception is when you are comparing two models, A and $\bar{A}$, and when those are the only two possibilities. This is the "disease test" illustration that forms the beginning of every text on Bayesian statistics (except for Sivia's!).