Modeling of Data

CHAPTER 15

15.0 Introduction

Given a set of observations, one often wants to condense and summarize the data by fitting it to a *model* that depends on adjustable parameters. Sometimes the model is simply a convenient class of functions, such as polynomials or Gaussians, and the fit supplies the appropriate coefficients. Other times, the model's parameters come from some underlying theory that the data are supposed to satisfy; examples are rate coefficients in a complex network of chemical reactions or orbital elements of a binary star. Modeling can also be used as a kind of constrained interpolation, where you want to extend a few data points into a continuous function, but with some underlying idea of what that function should look like.

One very general approach has the following paradigm: You choose or design a *figure-of-merit function* (merit function, for short) that measures the agreement between the data and the model with a particular choice of parameters. In frequentist statistics, the merit function is conventionally arranged so that small values represent close agreement. Bayesians choose as their merit function the probability of the parameters given the data (or often its logarithm) so that larger values represent closer agreement.

In either case, the parameters of the model are then adjusted to find a happy extremum in the merit function, yielding *best-fit parameters*. The adjustment process is thus a problem in minimization in many dimensions. This optimization was the subject of Chapter 10; however, there exist special, more efficient, methods that are specific to modeling, and we will discuss these in this chapter.

There are important issues that go beyond the mere finding of best-fit parameters. Data are generally not exact. They are subject to *measurement errors* (called *noise* in the context of signal processing). Thus, typical data never exactly fit the model that is being used, even when that model is correct. We need the means to assess whether or not the model is appropriate, that is, we need to test the *goodness-of-fit* against some useful statistical standard.

We usually also need to know the accuracy with which parameters are determined by the data set. In frequentist terms, we need to know the standard errors of the best-fit parameters. Alternatively, in Bayesian language, we want to find not just the peak of the joint parameter probability distribution, but the whole distribution.

Or we at least want to be able to sample from that distribution, typically by Markov chain Monte Carlo, as we will discuss at length in §15.8.

It is not uncommon in fitting data to discover that the merit function is not unimodal, with a single minimum. In some cases, we may be interested in global rather than local questions. Not, "how good is this fit?" but rather, "how sure am I that there is not a *very much better* fit in some corner of parameter space?" As we have seen in Chapter 10, especially §10.12, this kind of problem is generally quite difficult to solve.

The important message is that fitting of parameters is not the end-all of model parameter estimation. To be genuinely useful, a fitting procedure should provide (i) parameters, (ii) error estimates on the parameters or a way to sample from their probability distribution, and (iii) a statistical measure of goodness-of-fit. When the third item suggests that the model is an unlikely match to the data, then items (i) and (ii) are probably worthless. Unfortunately, many practitioners of parameter estimation never proceed beyond item (i). They deem a fit acceptable if a graph of data and model "looks good." This approach is known as *chi-by-eye*. Luckily, its practitioners get what they deserve.

15.0.1 Basic Bayes

Because the discussion in this and subsequent chapters will move freely between frequentist and Bayesian methods, this is a good place to compare these two powerfully useful ways of thinking. In §14.0, when we discussed tail, or p-value, tests, we were adopting a frequentist viewpoint. The central frequentist idea is that, given the details of a null hypothesis, there is an implied population (that is, probability distribution) of possible data sets. If the assumed null hypothesis is correct, then the actual, measured, data set is drawn from that population. (We expand on this in §15.6.) It then makes sense to ask questions about how "frequently" some aspect of the measured data occurs in the population. If the answer is "very infrequently," then the hypothesis is rejected. The frequentist viewpoint avoids questions like, "what is the *probability* that this hypothesis is true?" because its focus is on the distribution of data sets, not hypotheses. Indeed, whether by dogma or merely benign neglect, it eschews the machinery needed to handle the concept of a probability distribution of hypotheses.

That machinery is Bayes' theorem, which follows from the standard axioms of probability. Bayes' theorem relates the conditional probabilities of two events, say A and B:

$$P(A|B) = P(A)\,\frac{P(B|A)}{P(B)} \tag{15.0.1}$$

Here $P(A|B)$ is the probability of A *given* that B has occurred, and similarly for $P(B|A)$, while $P(A)$ and $P(B)$ are unconditional probabilities.

Bayesians allow a broader set of uses for probabilities than frequentists. To a Bayesian, $P(A|B)$ is a measure of the degree of plausibility of A (given B) on a scale ranging from zero to one. In this broader view, A and B need not be repeatable events; they can indeed be propositions or hypotheses. In equation (15.0.1), A might be a hypothesis and B might be some data, so that $P(A|B)$ expresses the probability of a hypothesis, given the data. The equations of probability theory thus become a set of consistent rules for conducting inference [1,2]. Interestingly, this viewpoint was

universal before the 20th century. The Bernoullis (both of them), Laplace, Gauss, Legendre, and Poisson, among others, made little or no distinction between inference and probability. An opposing frequentist view, that these concepts should be kept separate, became explicit only with the work of Fisher, Box, Kendall, Neyman, and Pearson (among others), much later.

Since plausibility is itself always conditioned on some, perhaps unarticulated, set of assumptions, all Bayesian probabilities are viewed as conditional on some collective background information I. Suppose H is some hypothesis. Even before there exist any explicit data, a Bayesian can assign to H some degree of plausibility $P(H|I)$, called the "Bayesian prior." Now, when some data D_1 comes along, Bayes theorem tells how to reassess the plausibility of H,

$$P(H|D_1 I) = P(H|I)\frac{P(D_1|HI)}{P(D_1|I)} \tag{15.0.2}$$

The factor in the numerator on the right of equation (15.0.2) is calculable as the probability of a data set *given* the hypothesis (comparable to "likelihood" as we will define it in §15.1). The denominator, called the "prior predictive probability" of the data, is in this case merely a normalization constant that can be calculated by the requirement that the probability of all hypotheses should sum to unity. (In other Bayesian contexts, the prior predictive probabilities of two qualitatively different models can be used to assess their relative plausibility.)

If some additional data D_2 come along tomorrow, we can further refine our estimate of H's probability, as

$$P(H|D_2 D_1 I) = P(H|D_1 I)\frac{P(D_2|HD_1 I)}{P(D_2|D_1 I)} \tag{15.0.3}$$

Using the product rule for probabilities, $P(AB|C) = P(A|C)P(B|AC)$, we find that equations (15.0.2) and (15.0.3) imply

$$P(H|D_2 D_1 I) = P(H|I)\frac{P(D_2 D_1|HI)}{P(D_2 D_1|I)} \tag{15.0.4}$$

which shows that we would have gotten the same answer if all the data $D_1 D_2$ had been taken together.

We might wonder, before we adopt the laws of probability as our calculus of inference and thus become Bayesians, whether there are any other alternatives. The answer is, basically, no. Cox [3] showed that making a small number of very reasonable assumptions about "degree of belief" leads inevitably to the axioms of probability, and thus the application of Bayes theorem to the evaluation of hypotheses, given data. Either you become a Bayesian or else you must live in a world with no general calculus of inference.

CITED REFERENCES AND FURTHER READING:

Bevington, P.R., and Robinson, D.K. 2002, *Data Reduction and Error Analysis for the Physical Sciences*, 3rd ed. (New York: McGraw-Hill), Chapters 6–11.

Devore, J.L. 2003, *Probability and Statistics for Engineering and the Sciences*, 6th ed. (Belmont, CA: Duxbury Press), Chapters 12–13.

Brownlee, K.A. 1965, *Statistical Theory and Methodology*, 2nd ed. (New York: Wiley).

Martin, B.R. 1971, *Statistics for Physicists* (New York: Academic Press).

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. 2004, *Bayesian Data Analysis*, 2nd ed. (Boca Raton, FL: Chapman & Hall/CRC).

Sivia, D.S. 1996, *Data Analysis: A Bayesian Tutorial* (Oxford, UK: Oxford University Press).

Jaynes, E.T. 1976, in *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, W.L. Harper and C.A. Hooker, eds. (Dordrecht: Reidel).[1]

Jaynes, E.T. 1985, in *Maximum-Entropy and Bayesian Methods in Inverse Problems*, C.R. Smith and W.T. Grandy, Jr., eds. (Dordrecht: Reidel).[2]

Cox, R.T. 1946, "Probability, Frequency, and Reasonable Expectation," *American Journal of Physics*, vol. 14, pp. 1–13.[3]

15.1 Least Squares as a Maximum Likelihood Estimator

Suppose that we are fitting N data points $(x_i, y_i), i = 0, \ldots, N-1$, to a model that has M adjustable parameters a_j, $j = 0, \ldots, M-1$. The model predicts a functional relationship between the measured independent and dependent variables,

$$y(x) = y(x|a_0 \ldots a_{M-1}) \tag{15.1.1}$$

where the notation indicates dependence on the parameters explicitly on the right-hand side, following the vertical bar.

What, exactly, do we want to minimize to get fitted values for the a_j's? The first thing that comes to mind is the familiar least-squares fit,

$$\text{minimize over } a_0 \ldots a_{M-1} : \quad \sum_{i=0}^{N-1} [y_i - y(x_i|a_0 \ldots a_{M-1})]^2 \tag{15.1.2}$$

But where does this come from? What general principles is it based on?

To answer these questions, let us start by asking, "*Given a particular set of parameters*, what is the probability that the observed data set should have occurred?" If the y_i's take on continuous values, the probability will always be zero unless we add the phrase, "...plus or minus some small, fixed Δy on each data point." So let's always take this phrase as understood. If the probability of obtaining the data set is too small, then we can conclude that the parameters under consideration are "unlikely" to be right. Conversely, our intuition tells us that the data set should not be too improbable for the correct choice of parameters.

To be more quantitative, suppose that each data point y_i has a measurement error that is independently random and distributed as a normal (Gaussian) distribution around the "true" model $y(x)$. And suppose that the standard deviations σ of these normal distributions are the same for all points. Then the probability of the data set is the product of the probabilities of each point:

$$P(\text{data} \mid \text{model}) \propto \prod_{i=0}^{N-1} \left\{ \exp\left[-\frac{1}{2} \left(\frac{y_i - y(x_i)}{\sigma} \right)^2 \right] \Delta y \right\} \tag{15.1.3}$$