Statistics and nonlinear fits

2

In this chapter, we provide a small window into the field of *statistics*, the mathematical study of data.[1] We naturally focus on the statistics of nonlinear models: how these models are fit to data, and how to then extract predictions and estimate their reliability. For those familiar with statistics, our purpose here is to define nomenclature and notation. More generally, for those (like our previous selves) who remain unfamiliar with the vocabulary and definitions of statistics, we discuss standard definitions and methods from the traditional viewpoint. We provide here only brief hints in the margins on how these definitions and methods relate to our reformulation of the subject in the later portions of this book.

[1] The mathematical study of statistics has surprisingly little overlap in vocabulary with the physical science of statistical mechanics, despite the similarity in subject matter.

2.1 Fits: Data, errors, models, and parameters

We are interested in the study of models $\vec{y}_{\boldsymbol{\tau}}(\boldsymbol{\theta})$ depending on parameters $\boldsymbol{\theta}$ and control variables $\boldsymbol{\tau}$, as they are fit to data $\vec{d}$ with errors $\vec{\sigma}$. Let us introduce each in turn.

Let M be the number of data points, and $\vec{d}$ be the data as a vector of length M, so[2] $\vec{d} = \{d_i\} = \{d_0, d_1, \ldots, d_{M-1}\}$. These might be the number of clicks in a Geiger counter at times $\{t_i\}$, or the amount of rain at (time, position) coordinates $\{(t_i, \mathbf{x}_i)\} = \{(t_0, \mathbf{x}_0), (t_1, \mathbf{x}_1), \ldots\}$, or the chemical concentration of protein species s_i at time t_i under experimental condition n_i. In these cases, the control variables[3] $\boldsymbol{\tau}$ consist of the time t, or the time and position $(t, \mathbf{x})$. The data $\vec{d}$ is a vector in *data space* $\mathbb{R}^M$; we will consistently denote data–space vectors with an arrow above them, and will use Latin letters $i, j, k \ldots$ as indices.

Let $\vec{y}(\boldsymbol{\theta}) = \{y_{\boldsymbol{\tau}_i}(\theta_\alpha)\}$ be the theoretical model being fit to the data, depending upon N parameters $\boldsymbol{\theta} = \{\theta_0, \theta_1, \ldots, \theta_{M-1}\}$; these might be the amounts and decay rates of radioactive elements, cloud droplet nucleation rates, or reaction rates and binding constants for models of Geiger counters/climate/cellular chemistry. The parameters $\boldsymbol{\theta}$ are a point in *parameter space*, which we denote with a boldface font and index with Greek letters $\alpha, \beta, \mu, \nu, \ldots$.

Together with each data point is an error estimate σ_i, the standard deviation of the data point d_i. In statistics, σ_i is usually considered part of the model $\vec{y}$. For example, in radioactive decay with a relatively weak signal one expects that the number of clicks d_i in the time window

[2] In this book, we will use the convention that vector indices start at zero, common to modern programming languages (C, C++, and Python, the last of which we use in *SloppyCell*). Standard mathematics notation starts vectors with the index one (as does Fortran, Matlab, and Mathematica). This will arise rarely.

[3] Later in the text, we will study the implications of *interpolating* between different values of these control variables to explain features of sloppy models. Hence we focus on continuous variables like time and position, even though chemical species also varies between data points. We will often drop the $\boldsymbol{\tau}$ subscript for $\vec{y}$

centered at t_i will fluctuate about some mean, and that $y_i(\boldsymbol{\theta})$ will be an estimate of this average decay rate. If the decays happen independently (an excellent approximation), then the probability of getting the precise number d_i will be given by a Poisson distribution. Since a Poisson distribution of mean y has standard deviation $\sqrt{y}$, one can use $\sigma_i = \sqrt{y_i(\boldsymbol{\theta})}$ as a theoretical estimate of the expected error in the measured data point d_i. This is not the perspective we will take in this book.[4]

In applications, it is usually more natural to associate the error estimates with the data $\vec{d}$. This is particularly natural with data that is itself an average over several measurements:[5] if data point d_i is an average over L_i independent measurements $d_i^{(\ell)}$ with estimated standard deviation $\sigma_i^{\text{sample}} = \sqrt{\sum_{\ell=0}^{L}(d_i^{(\ell)} - d_i)^2/(L-1)}$, then the error estimate for d_i is the standard deviation of the mean $\sigma_i = \sigma_i^{\text{mean}} = \sigma_i^{\text{sample}}/\sqrt{L}$.[6]

In many applications, the data points are not independent of one another. In our avalanche model, for example (Fig. **??**) runs with a very large avalanche will tend to have fewer avalanches of all sizes; the small and large avalanche populations will be anticorrelated, and the small avalanche populations will tend to be correlated with one another.[7] For simplicity, in this book we will assume that the data is uncorrelated, and in much of it we will also assume that the errors are all of the same magnitude.[8][9]

2.2 Quality of fit: Residuals, likelihood, and the cost function

The deviation of the best fit from the data is described by the *residuals* $r_i = (y_i(\boldsymbol{\theta}) - d_i)/\sigma_i$. Finding the best fit involves finding the parameter set $\boldsymbol{\theta}$ that zeros the residual vector as best as possible.

The definition of 'best' usually involves the *likelihood function* $P(\vec{d}\,|\,\boldsymbol{\theta})$,[10] the probability density that the data $\vec{d}$ would have been generated by the model using parameters $\boldsymbol{\theta}$, added to errors given by $\vec{\sigma}$. In general, this probability depends on the distribution function for the errors; knowing σ_i only tells the root–mean–square average of the errors, not the shape of the distribution. For simplicity again, we will assume in this book that our probabilities are 'normally' distributed, so the deviation x of d_i from the true value has probability density proportional to $\exp(-x^2/(2\sigma_i^2))$.[11] Since the deviation of theory from data point d_i is the residual r_i, this implies that our likelihood function is

$$P(\vec{d}\,|\,\boldsymbol{\theta}) \propto \exp(-\sum_m r_i^2/2) = \exp\left(-\sum_m (y_i(\boldsymbol{\theta}) - d_i)^2/(2\sigma_i^2)\right). \quad (2.1)$$

The distance[12] between the data and the model prediction is measured by the argument of the exponential in eqn 2.1, which we will call the

[4] XXX Say something about how one might incorporate $\vec{\sigma} = \vec{\sigma}(\boldsymbol{\theta})$ into our formalism, or why it is hard?

[5] In our example, one could position several Geiger counters at the same distance from the sample, or repeat the experiment several times with identically prepared radioactive samples.

[6] The distinction between the spread of the data σ^{sample} and the estimated spread of the mean (which is smaller by the square root $\sqrt{L}$ of the number of independent measurements) is commonly missed by new researchers. If your system is expected to have a smooth behavior plus noise, and your data is much smoother than your error estimates, you should check for this, and also for the possibility of correlations between your measured data.

[7] These data correlations are nicely described by a covariance matrix, closely related to the covariance matrix we introduce in Section 2.4 for the model predictions.

[8] Correlated errors would be straightforward to include, at the cost of complicating our notation. For example, in Section 4.2 the inverse of the covariance matrix of the data would become the metric in data space.

[9] XXX Say something about non-Gaussian errors, and how they make the embedding metric incompatible with the Fisher Information metric.

[10] This formula would be pronounced 'pee of dee given theta'.

[11] This is also known as a Gaussian; properly normalized, it is $1/\sqrt{2\pi\sigma^2}\exp -x^2/(2\sigma^2)$. For averages of many independent measurements of finite variance, the *central limit theorem* guarantees convergence to a Gaussian.

[12] This will indeed become literally true in Section 4.2, where we use this as the metric in data space.

cost $C(\boldsymbol{\theta})$:

$$C(\boldsymbol{\theta}) = \sum_i r_i^2/2 = \sum_{i=0}^{M-1} (y_i(\boldsymbol{\theta}) - d_i)^2/(2\sigma_i^2). \tag{2.2}$$

This (up to an additive constant) is the *negative log-likelihood* $C(\boldsymbol{\theta}) = -\log P(\vec{d}\,|\,\boldsymbol{\theta})$.[13] It is half of the measure $\xi^2 = \sum r_i^2$ (pronounced 'khi squared'), commonly used by statisticians.

[13] In cases where our errors are estimated from the parameters, we indeed would use the negative log-likelihood as our cost.

2.3 Bayesians versus the Frequentists

The next section (Section 2.4, discussing the estimation of parameter errors from data fits) approaches a touchy subject in the field of statistics: the competing Bayesian and Frequentist interpretations. As we find it convenient to adopt the language and methods of the minority (Bayesian) viewpoint, it behooves us to explain here the distinctions. The results and methods of the rest of the book do not depend on a Bayesian framework, and should be of interest and relevance to both groups.[14]

[14] To our physics readers, we point out the startlingly complete parallel with the Copenhagen and many–worlds interpretation of quantum mechanics.

The majority viewpoint is that of the Frequentists. Frequentists believe in one true model, with a true parameter set $\boldsymbol{\theta}$. They define their job as weeding out implausible models and implausible parameter sets. Models with low likelihood are implausible. The best parameters $\boldsymbol{\theta}^*$ are those which maximize the likelihood $P(\boldsymbol{\theta}^*\,|\,\vec{d}) = max_{\boldsymbol{\theta}} P(\boldsymbol{\theta}\,|\,\vec{d})$. Parameter uncertainties are given by the range of parameters that are *not* ruled out by the data.

The Bayesians believe in an *ensemble* of models, compatible with our current knowledge. They are interested in the relative probability of different models. In estimating the parameter uncertainties, they are therefore interested in $P(\boldsymbol{\theta}\,|\,\vec{d})$, the probability density of a given set of model parameters, given the data $\vec{d}$. Since $P(A\,|\,B)P(B) = P(A \text{ and } B)/P(B)$ (the probability of A given B times the probability of B is the probability that both A and B happen), we see that $P(\boldsymbol{\theta}\,|\,\vec{d})P(\vec{d}) = P(\vec{d}\,|\,\boldsymbol{\theta})P(\boldsymbol{\theta})$, and hence derive Bayes' theorem:

$$P(\boldsymbol{\theta}\,|\,\vec{d}) = P(\vec{d}\,|\,\boldsymbol{\theta})P(\boldsymbol{\theta})/P(\vec{d}). \tag{2.3}$$

The left–hand side is the result we want: $P(\boldsymbol{\theta}\,|\,\vec{d})$ is the relative probabilities of different sets of parameters. The first term on the right, $P(\vec{d}\,|\,\boldsymbol{\theta})$, is the likelihood function given in the last section (Section 2.2). The last term $P(\vec{d})$ (the 'probability of the data') can be viewed as a normalization constant—making the total probability of having some set of parameters integrate to one.

The remaining term, $P(\boldsymbol{\theta})$ is called the *prior*. The term 'prior' is used to suggest that it is the relative probability of the different models before (prior) to the current set of data d_i. For example, if one has pre-existing measurements of some or all of the parameter values, $P(\boldsymbol{\theta})$ would represent the probabilities of those parameters. More generally,

if you have no previous measurements,[15] but have a general experience of the rough scale of certain measurements (protein binding energies are a few k_BT) you are required to incorporate that into the analysis. The prior can be viewed as a measure on parameter space.[16]

A Frequentist might argue that the Bayesians, by using priors, prejudice their models. Indeed, the maximum in the Bayesian probability distribution (where $\nabla_{\boldsymbol{\theta}} P(\boldsymbol{\theta} \,|\, \vec{d}) = \mathbf{0}$) is shifted away from the maximum likelihood (where $\nabla_{\boldsymbol{\theta}} P(\vec{d} \,|\, \boldsymbol{\theta}) = \mathbf{0}$):

$$\nabla_{\boldsymbol{\theta}} P(\boldsymbol{\theta} \,|\, \vec{d}) = P(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} P(\vec{d} \,|\, \boldsymbol{\theta}) + P(\vec{d} \,|\, \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} P(\boldsymbol{\theta}) \tag{2.4}$$

with the last term representing a pressure due to the 'bias' introduced by the prior. Should a new measure of the gravitational constant, for example, incorporate previously known measurements into the analysis?[17]

A Bayesian might argue that the Frequentists are implicitly using a 'flat' prior. From their point of view, the maximum of eqn 2.4 of $P(\boldsymbol{\theta} \,|\, \vec{d})$ that shifts due to the prior is not particularly meaningful. Indeed, if we do a nonlinear change of variables $\boldsymbol{\phi}(\boldsymbol{\theta})$ (say, from decay rates γ_α to lifetimes $\tau_\alpha = 1/\gamma_\alpha$) the probability density is multiplied by the Jacobian determinant $|\partial\phi/\partial\theta|$ of the transformation: $P(\boldsymbol{\phi} \,|\, \vec{d}) = |\partial\phi/\partial\theta| P(\boldsymbol{\theta} \,|\, \vec{d})$. The 'most likely' rate is not the inverse of the 'most likely' lifetime, because what we are using is a likelihood density and the density of rates and lifetimes differ.[18] A Frequentist, of course, would point out that by maximizing the likelihood $P(\vec{d} \,|\, \boldsymbol{\theta})$ that the model would generate the data, their probability densities are in data space (where they are invariant under parameter changes).

2.4 Parameter error estimates: Hessians and Jacobians

Fitting a model to data is usually a precursor to drawing conclusions and making predictions. Whether predictions are possible depends on how well the existing data constrain the behavior of the model. In most cases, this is quantified by calculating the parameters uncertainties—how far the parameters $\boldsymbol{\theta}$ can vary away from the best fit $\boldsymbol{\theta}^*$ before the fit becomes unacceptable. Here we outline the traditional approaches to quantifying the parameter ranges consistent with the data.

We shall see in Chapter 3 and Section 8.1 that in typical multiparameter nonlinear fits the parameters can vary over enormous ranges while still describing the data and still allowing useful predictions.[19] This makes error bars for individual parameters almost useless, and suggests more broadly that measuring the range of predictions of a property of interest directly might be more useful than quoting parameter distributions as an intermediate step. We proceed nonetheless, because the machinery we need to explain traditional error estimation (Hessians, approximate Hessians, and covariance matrices) will be useful in our larger discussion.[20]

[15] What if you have
Bayesian holy grail
'objective' prior. V
sensible choice for t
ter 7.

[16] In mechanics and statistical physics, Liouville's theorerm (Sethna, 2006) tells us that the natural prior in phase space is uniform, $\prod_n \mathrm{d}p_n \mathrm{d}q_n$. There is no analogy to Liouville's theorem for parameter space.

[17] A Bayesian approach here would suppress some of the spurious scientific announcements of statistically significant, scientifically implausible measurements.

[18] Indeed, under a nonlinear parameter transformation the prior must be multiplied by this Jacobian.

[19] That is, they are *sloppy* models.

[20] Their eigenvalues will be used to characterize sloppiness in Chapter 3, and the approximate Hessian will become the metric tensor in Section 4.2.

In the limit that one has large amounts of high–accuracy data and a model that correctly describes the data, one can generally prove that the best fit parameters $\boldsymbol{\theta}^*$ converge to the true parameters, and that the acceptable fits converge to a small region around the best fit.[21] The traditional error estimates assume that we are in this limit, and thus expand about the best fit in a (multidimensional) Taylor series.

[21] Indeed, the best fits in this limit are distributed according to eqn **??** below.

From the Frequentist point of view (Section 2.3), parameters become unlikely when they are unlikely to generate the observed data. Thus the best fit $\boldsymbol{\theta}^*$ maximizes the likelihood $P(\vec{d}\,|\,\boldsymbol{\theta}) \propto \exp(-C(\boldsymbol{\theta}))$, which is equivalent to maximizing the cost $C(\theta)$. If the parameters are changed by $\boldsymbol{\delta}$, so $\boldsymbol{\theta} = \boldsymbol{\theta}^* + \boldsymbol{\delta}$, we can expand the cost in powers of $\boldsymbol{\delta}$:[22]

$$\begin{aligned} C(\boldsymbol{\theta}) &\approx C(\boldsymbol{\theta}^*) + \delta_\mu \partial_\mu C + {}^1\!/_2\, \delta_\mu \delta_\nu \partial_\mu \partial_\nu C \\ &= C(\boldsymbol{\theta}^*) + \delta_\mu \delta_\nu \mathcal{H}^{\rm cost}_{\mu\nu}/2 \end{aligned} \tag{2.5}$$

[22] Here we are using an incredibly convenient set of conventions called Einstein notation. ∂_μ is a shorthand for $\partial X/\partial\theta_\mu$, and repeated indices are automatically summed over, so $\delta_\mu \partial_\mu C = \sum_\mu \delta_\mu \partial C/\partial\theta_\mu = \boldsymbol{\delta}\cdot\nabla C$ and $\delta_\mu\delta_\nu\mathcal{H}^{\rm cost}_{\mu\nu} = \boldsymbol{\delta}^T\mathcal{H}^{\rm cost}\boldsymbol{\delta}$.

where $\mathcal{H}^{\rm cost} = \partial^2 C/\partial\theta_\mu\partial\theta_\nu$ is the Hessian of the cost at the best fit, and the linear term disappears because θ^* is a local maximum of the cost, and hence the likelihood of $\boldsymbol{\theta}^* + \boldsymbol{\delta}$ is given by

$$P(\vec{d}\,|\,\boldsymbol{\theta}^* + \boldsymbol{\delta}) = \exp(-C(\boldsymbol{\theta}^*) - \delta_\mu\delta_\nu\mathcal{H}^{\rm cost}_{\mu\nu}/2) \tag{2.6}$$

and is thus less likely than the best fit by a factor of

$$P(\vec{d}\,|\,\boldsymbol{\theta}^* + \boldsymbol{\delta})/P(\vec{d}\,|\,\boldsymbol{\theta}^*) = \exp(-\delta_\mu\delta_\nu\mathcal{H}^{\rm cost}_{\mu\nu}/2). \tag{2.7}$$

For a Bayesian, the relative probability densities of different parameters are given by $P(\boldsymbol{\theta}\,|\,\vec{d}) \propto P(\boldsymbol{\theta})P(\vec{d}\,|\,\boldsymbol{\theta})$. In the limit of lots of high–quality data, one can show that the prior $P(\boldsymbol{\theta})$ becomes unimportant.[23] More generally, one can often incorporate the prior in terms of extra effective residuals (see Section 8.4), literally adding data points representing the previous experimental knowledge. In either case, the probability density is reduced from that of the best fit by the same factor

[23] XXX Formulas here, showing as $M \to \infty$ the prior quadratic form gets swamped, and the linear term is not important.

$$P(\boldsymbol{\theta}\,|\,\vec{d})/P(\boldsymbol{\theta}^*\,|\,\vec{d}) \approx P(\vec{d}\,|\,\boldsymbol{\theta})/P(\vec{d}^*\,|\,\boldsymbol{\theta}^*) \approx \exp(-\delta_\mu\delta_\nu\mathcal{H}^{\rm cost}_{\mu\nu}/2). \tag{2.8}$$

Thus, for both Frequentists and Bayesians, we find the ranges of acceptable parameters are given by a multidimensional normal distribution characterized by the Hessian $\mathcal{H}^{\rm cost}$ of the cost. This distribution incorporates not only the error ranges of the various parameters, but also correlations between the uncertainties in parameter combinations. In statistics, what is usually quoted is the *covariance matrix*, which is precisely the inverse of the Hessian ${\mathcal{H}^{\rm cost}}^{-1}$. We shall see in Section 3.2 that the diagonal entries of the covariance matrix give the ranges of individual parameters, but that the off–diagonal entries are crucial to understanding the uncertainties in the model predictions.[24]

[24] That is, predictions will be possible because of strongly constrained 'stiff' combinations of parameters, none of which are by themselves constrained well by the data.

Since we have an formula for the cost $C = \sum r_i^2/2$ (eqn 2.2) in terms

of the residuals $r_i = (y_i(\boldsymbol{\theta}) - d_i)/\sigma_i$ we can write the cost Hessian:

$$\begin{aligned}\mathcal{H}^{\text{cost}}_{\alpha\beta} &= \partial_\alpha \partial_\beta C = \sum_i \tfrac{1}{2}\, \partial_\alpha \partial_\beta r_i^2 = \sum_i \partial_\alpha (r_i \partial_\beta r_i) \\ &= \sum_i (\partial_\alpha r_i)(\partial_\beta r_i) + \sum_i r_i \partial_\alpha \partial_\beta r_i \\ &= \sum_i (\partial_\alpha y_i)(\partial_\beta y_i)\,/\sigma_i^2 + \sum_i (y_i - d_i)\partial_\alpha \partial_\beta y_i\,/\sigma_i^2.\end{aligned} \tag{2.9}$$

Note that the first term in the last line of eqn 2.9 only involves first derivatives, and is *independent of the data*. It is both useful and common to approximate the cost Hessian by this first term:

$$\mathcal{H}^{\text{cost}}_{\alpha\beta} \approx \mathcal{H}^{\text{aprx}}_{\alpha\beta} = \sum_i (\partial_\alpha y_i)(\partial_\beta y_i)\,/\sigma_i^2. \tag{2.10}$$

In statistics, the approximate Hessian is known as the *Fisher information matrix*.[25]

This approximate Hessian is convenient for three reasons. First, $\mathcal{H}^{\text{aprx}}$ only involves first derivatives. Usually, the user does not provide a routine for taking second derivatives of their residuals. Numerical derivatives introduce discretization errors.[26] Also, with N parameters the second derivative will cost $\sim N^2$ operations (or $\sim N$ residual gradient calculations), which for large numbers of parameters can be prohibitive.

Second, the approximate Hessian is *independent of the data*, depending only on the parameters we evaluate it at. Thus $\mathcal{H}^{\text{aprx}}(\boldsymbol{\theta})$ is a property of the model, not of the particular data set we use to fit it to.[27]

Third, the approximate Hessian is often an excellent approximation to the true Hessian.[28] The other term in the last line of eqn 2.9, $\mathcal{H}^{\text{cost}}_{\alpha\beta} - \mathcal{H}^{\text{aprx}}_{\alpha\beta} = +\sum_i (y_i - d_i)\partial_\alpha \partial_\beta y_i\,/\sigma_i^2$ is zero if the model agrees with the data $y_i = d_i$. Thus it is small when the model is doing a good job, and is zero when one generates artificial data using the model (Brown and Sethna, 2003, Brown *et al.*, 2004, Gutenkunst *et al.*, 2007). This last term also is usually equally often positive and negative, so for large amounts of data it tends to cancel out.

Finally, we introduce the important Jacobian

$$J_{i\alpha} = (1/\sigma_i)\partial_\alpha y_i = \partial y_i / \partial \theta_\alpha\,/\sigma_i, \tag{2.11}$$

an $M \times N$ matrix describing the dependence of the M predictions of the model $y_i(\boldsymbol{\theta})$ on the N parameters θ_α.[29] Now the first term in the last line of eqn 2.9 can be written as

$$\mathcal{H}^{\text{aprx}}_{\alpha\beta} = J_{i\alpha} J_{i\beta} = (J^T J)_{\alpha\beta}. \tag{2.12}$$

[25] More precisely, the Fisher information matrix is the expectation of $J^T J$ over the noise in the data. One can show that the expectation value of $\mathcal{H}^{\text{cost}}$ is *also* the Fisher information matrix: when averaging over all possible noise, the positive and negative residual cancellation becomes perfect. Finally, for the normally–distributed (Gaussian) noise we assume in this book, the expectation of $J^T J$ and $\mathcal{H}^{\text{cost}}$ over the noise are both equal to $J^T J$.

[26] We have found that this is particularly problematic for sloppy models, where small discretization errors can cause eigenvalues to go negative.

[27] We will indeed find that it provides the natural metric tensor $g_{\mu\nu} = \mathcal{H}^{\text{aprx}}_{\mu\nu}$ in Section 4.2.

[28] In particular, the remarkable sloppy properties of the Hessian that we introduce in Chapter 3 are seen both in the exact and the approximate Hessians.

[29] The Jacobian thus characterizes how the model stretches and rotates parameter space into data space (Section 4.2).