Class notes: Monte Carlo methods
Week 5, Bayesian statistics

Jonathan Goodman

April 14, 2015

1 The philosophy

This section of notes is more expository and less technical than most. The practice of statistics, particularly Bayesian statistics, calls for lots of good judgement as well as technical, mathematical, and programming chops. The verbal material here is as important to the practice of Monte Carlo for Bayesian statistics as Hermite polynomials and the Perron Frobenius theorem. Pay attention and read carefully.

Statistics, as an activity, means gathering and analyzing data to learn about the world. In particular, data helps determine the values of parameters in models. The mathematics of statistics has much to say about what data says about parameters, in particular, precisely how strongly data constrain the parameters. The word *constrain* is not in the sense of optimization theory, which would be inequalities or equalities that cannot be violated. Here, it means that the data suggest certain combinations of parameters that are favored by the data and other combinations that are made very unlikely by the data. There are few absolute certainties, but there can be very high or low probabilities.

We denote the model parameters by $x = (x_1, \ldots, x_d)$. The data are $Y = Y_1, \ldots, Y_n)$. There will be a *likelihood function*, $L(x|y)$, that determines the goodness of fit – how well parameters x explain data y. Traditional *frequentist* statistics determines the *maximum likelihood* estimate of the parameters by finding the parameter combination that best fits the data in this sense:

$$\widehat{X}(y)_{\text{ML}} = \arg\max_x L(x|y) \; . \tag{1}$$

The modern Bayesian viewpoint (after Thomas Bayes, pioneering book on probability published 1763) criticizes this for not saying how strongly the data, y, constrain x.

The frequentist answers this criticism by constructing *confidence regions* in parameter space. A frequentist views the true parameters, x_*, as fixed, not random, and unknown. The data, Y, are drawn from a true probability distribution that depends on this fixed but unknown parameter set x_*. The maximum likelihood estimate (1) gives $\widehat{X}_{\text{ML}}$ that is a function of Y and therefore also random. A confidence region, with confidence level β, is a set in parameter space $C(Y) \subset \mathbb{R}^d$, so that

$$\text{P}(\, x_* \in C(Y)) \geq 1 - \beta \; . \tag{2}$$

It may happen that you are interested in a single parameter, say x_1. A confidence region for a single parameter may take the form of a *confidence interval* determined by a lower bound X_l and upper bound X_r:

$$\mathrm{P}(\, x_{1,*} \in [X_{1,l}(Y), X_{1,r}]\,) \geq 1 - \beta \; . \tag{3}$$

For example, it was recently announced that measurements at the CERN accelerator confirmed the predicted *Higgs boson* and gave estimates of its mass

$$a \leq m_H \leq b$$

The announcement said these were $5\,\sigma$ estimates, which meant that if $Z \sim \mathcal{N}(0,1)$, then

$$\mathrm{P}(\, m_H \notin [a,b]) \leq \mathrm{P}(\, Z \notin [-5,5]) \approx 5.7 \cdot 10^{-7} \; .$$

There are several Bayesian criticisms of the confidence region/confidence interval approach:

- Individual confidence intervals potentially leave out much information about the parameters that is contained in the data.

- Individual confidence intervals are not joint confidence intervals. If parameter estimates X_1 and X_2 are correlated, we do now know immediately what confidence intervals for X_1 and X_2 separately say about the pair (X_1, X_2) This is particularly important when there are many parameters.

- Confidence regions and confidence intervals are often found using approximations that are based on the central limit theorem and are valid when the sample size is large. These may be unreliable in practical situations with limited data. Many practical applications are not in the large data central limit regime, even when the data set is very large. For example, a year of stock trading *tick data*, recording every price change, may contain only a few significant economic events.

- Confidence regions $C(Y)$ may be hard to describe, given the curse of dimensionality. The best way to describe $C(Y)$ may be to give a collection of random samples $X \in C(Y)$. The question then arises, what probability distribution should the X be drawn from?

- It is somewhat arbitrary to insist on a hard boundary set $C(Y)$. It may be better to give a *soft region* confidence distribution $X \sim f(x|Y)$ to describe the constraint placed on parameter combinations by the data. If you have the data There does not seem to be a frequentist statement to be made about the “probability” that x_* comes from the probability density $f(x|Y)$.

The frequentist will answer these criticisms and the debate will go on. But it seems helpful to have a general way to understand how data constrain parameters in statistical models.

The beginning of a Bayesian approach is to ask not which parameter combination best fits the data (1), but what is the set of good fits. One mathematical formulation of this involves a family of probability distributions build from the likelihood function and depending on a parameter, β:

$$f(x|y,\beta) = \frac{1}{Z(Y,\beta)} L(x|Y)^\beta \; . \tag{4}$$

The parameter β determines how closely x is to the best fit $\widehat{X}_{\text{ML}}$. Large β constrains x to be close to $\widehat{X}_{\text{ML}}$, while smaller $\beta > 0$ allows more variation. This may be clearer if you use the negative of the *log likelihood* function

$$\phi(x|y) = -\log(L(x|y)) \; .$$

The maximum likelihood point minimizes ϕ, which measures the *badness of fit.*

$$\widehat{X}_{\text{ML}} = \arg \; \min_x \phi(x|y) \; .$$

The distributions (4) now take the form

$$f(x|y,\beta) = \frac{1}{Z(Y,\beta)} \, e^{-\beta\phi(x|y)} \; .$$

This represents the statistical uncertainty in the form we used earlier for the Gibbs Boltzmann distribution in thermodynamic equilibrium, with β being interpreted as the inverse temperature. Large β corresponds to low temperature, which keeps the parameters close to the most likely values. As $\beta \to 0$, we become less interested in constraints the likelihood function place on the data. *The simplest Bayesian approach to model based data analysis is simply to produce many samples of the distribution $f(x|Y,\beta)$.*

The research described in the talk

`http://www4.ncsu.edu/~ctk/TALKS/NLSQ.pdf`

illustrates the case for this primitive Bayesian approach. The researchers want to use blood flow measurements to determine whether some of the major blood vessels are partly blocked. The model parameters represent the state of different arteries. It turns out that the model is ill-conditioned: there are many parameter combinations that give nearly identical measurements, which makes the optimization problem (1) hard. The researchers show that the optimization problem becomes more tractable if they impose constraints on the parameters to select specific representatives from collections of parameters with nearly identical predictions. The details are interesting and clever.

But the approach has the drawback that it does not tell the end user, a physician in this case, what uncertainties remain after using the data. If there are several different explanations for the data, it may be important for the physician to know that. Rather than reporting a single best fit, or a single fit that is nearly optimal in the sense of (1), it may be better to provide the physician with a collection of parameter combinations that represent the different disease states that are consistent with the data. The physician should know if the measurements do not determine which artery is blocked. Giving a single likely parameter set, or even the single most likely parameter set, does not convey all the information the user (physician) wants or needs to know.

It makes sense to sample the distribution (4) only if it is possible to do so. Advances in Monte Carlo sampling technique make it possible to apply the Bayesian philosophy in more and more sophisticated situations.

2 Parametric statistics, likelihoods, probability modeling

Parametric statistics means using data to identify parameters in probability models. There are other aspects of statistics. For example, one may want to know the mean of a population without having a model of the population distribution. As an example, think about estimating the average income of a New York City resident without using a probability model governing the distribution New York City incomes. The average income is a well defined number, but it is not a parameter in an income probability density function. Often called *non-parametric* statistics would be estimating the PDF of New York City incomes, from a collection of random samples. Ironically, this is actually a problem involving a large number or an infinite number of parameters, as one seeks to estimate a whole function (the PDF), which is an element of an infinite dimensional function space. In practice, we would represent the estimated PDF by a large but finite collection of numbers.

A simple case is the model is that the data consists of n independent observations of a random variable from a probability density depending on parameters $x = (x_1, \ldots, x_d)$.

$$Y_k \sim g(y|x) \;\; , \quad Y_j \text{ independent of } Y_k \; , \;\; j \neq k \; .$$

In that case $Y = (Y_1, \ldots, Y_n)$ and joint PDF Y is

$$L(y|x) = \prod_{k=1}^{n} g(y_j|x) \; .$$

A standard example is estimating the mean and variance of a one dimensional normal: $g(y|\mu,\sigma) = \frac{1}{Z}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$. We do not always denote the parameters by generic x in specific applications. A fancier model in the same spirit would be estimating the mean, “variance parameter”, and power law decay of a Student t:

$$g(y|\mu,\sigma,p) = \frac{1}{Z}\left(\frac{1}{1+\frac{1}{2p}\left(\frac{(x-\mu)^2}{\sigma}\right)}\right)^p \; . \tag{5}$$

The actual variance of Y in this distribution is larger than σ^2. The parameter p here is within $\frac{1}{2}$ of the "degrees of freedom" parameter in the standard t distribution. This formula is simpler and represents the same family of distributions. This converges to the normal in the limit $p \to \infty$. The p parameter determines the rate at which $g \to 0$ as $y \to \infty$. For finite p, this is a power law rather than an exponential. These are *fat tailed* distributions. They are used to model situations where there are occasional "outliers" that are many σ's from the mean. Many random objects have fat tailed distributions. Examples include the sizes of files transmitted over the web (text messages to movies), incomes of people, and sizes of earthquakes.

Here is a more complex model involving observations of a dynamical system at various times. Suppose $y(t)$ is the amount of a given chemical in an experiment at time t. Suppose $T(t)$ is the temperature at time t given, say, in degrees Kelvin. Suppose y changes because it is consumed by a temperature dependent chemical reaction. In particular, we take $\frac{d}{dt}y = \dot{y} \sim -y$, with the proportionality constant given by a standard chemical kinetics model

$$\dot{y} = -Ae^{\frac{-E_a}{RT(t)}}y(t) \ , \tag{6}$$

$$\dot{T} = -C\dot{y} = CAe^{\frac{-E_a}{RT(t)}}y(t) \ . \tag{7}$$

Finally, suppose the data consist of noisy observations of y at observation times t_k:

$$Y_k = y(t_k) + \sigma Z_k \ \ , \quad Z_k \sim \mathcal{N}(0,1) \ , \ \text{i.i.d.} \ . \tag{8}$$

The parameters to be identified are

- A, the *prefactor*

- E_a, the *activation energy*. You can think that the molecule of substance y has to get energy E_a to do its reaction. The probability to get this energy at temperature T has the Gibbs Boltzmann probability $E^{\frac{-E_a}{kT}}$, except that in chemistry the conversion factor between temperature and energy is called R (the *ideal gas constant*) rather than k (the *Boltzmann constant*).

- C, the *heat release*. When one molecule reacts, it releases energy, which in turn raises the temperature. The reaction starts slowly, but accelerates as the temperature starts to rise.

- σ, the *observation noise standard deviation*. This may be an example of a *nuisance parameter*, which is a parameter whose value we are not interested in, but which must be "estimated" (i.e., sampled) along with the others because it is a part of the model and is unknown.

The ideal gas constant, R, is not a parameter because its value is known. Other known quantities include the initial concentration, $y(0)$, the initial temperature, $T(0)$, and the observation times t_k.

There is not a simple closed form expression for the likelihood function as a function of the parameters and observations. The probability density to get observations $Y = \{Y_k\}$ is

$$L(y|A, E_a, C, \sigma) = \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_k - y(t_k|A, E_a, C))^2}{2\sigma^2}} \ . \tag{9}$$

The values $y(t_k|A, E_a, C)$ are computed from (6) and (7) using a ODE solver and the known initial conditions.

The *log likelihood*, which is the log of the likelihood, is often easier to work with in practice. The likelihood itself can easily get to be smaller than the smallest positive floating point number in double precision, which is $e^{-1022} \approx 1.4 \cdot 10^{-444}$. This can happen, for example, if there are 1022 data points each contributing a factor $\frac{1}{e}$.

It can be very hard, from a technical point of view, to sample the distribution

$$X \sim f(x) = \frac{1}{Z(Y)} L(Y|x) \ .$$

Among the difficulties you will meet in practice:

- It can be expensive to evaluate the likelihood function. The formula (9) asks you to solve a differential equation and cycle through a data set. Other problems ask you to solve a partial differential equation and/or cycle through a much larger data set. If it takes minutes or hours to compute $f(x)$ for a single proposal x, then we should seek more sophisticated ways make good proposals.

- There can be more than one locally best fit. For example, if your data has frequency ω, then a fit with frequency 2ω may fit better than $2\omega \pm \epsilon$, even though none of these fits as well as frequencies nearer to ω. Local moves typical of Metropolis samplers are slow to move from one local "potential well". The potential, ϕ, is the negative log of the likelihood, L. ϕ has a local minimum where the L has a local maximum. The problem can be so severe that a sampler can spend all its time in a local potential well without ever finding a deeper well with better fits. The theorem (Perron Frobenius) says the sampler will eventually find the better fits, but "eventually" may be longer than a practical run time.

- The burn-in phase can be very slow. Ironically, this often happens when the data strongly constrain the parameters. Your starting guess may be a bad fit, and nearby parameter combinations may be little better. For example, if the frequency ω is very clear in the data and you start with $23 \cdot \omega$, then Metropolis perturbations to $22 \cdot \omega$ may be nearly equally bad fits to the data. This means the samples will drift only slowly toward good fits.

- The probability distribution may be ill conditioned. A simple form of ill conditioning is that the different variables have different units and, as numbers, vastly different ranges. For example, the activation energy E_a may be on the order of one (if measured in electron volts), while the prefactor A may be of the order of 10^8, if measured in inverse seconds. Consider a simple Gaussian proposal move $X \to X + \rho Z$ (where $Z \sim \mathcal{N}(0, I_d)$). This proposes to move E_a by an amount of order ρ, and A by order ρ. For E_a, we should adjust ρ to be order 1. But if we do that, the perturbation of A will be, relative to A, so small that it takes 10^8 such steps to make an order 1 relative change to A. More troubling is that the data may imply strong relations among the parameters without constraining the individual parameters so tightly. For example, a proposal to raise A without also raising C a proportionate amount may take you from good to bad fits, and be rejected.

3 Priors and posteriors

So far, Bayesian statistics has been a heuristic. You learn what parameter sets are consistent with the data by sampling from that set. But there is a more formal version.

The formal version is that first the parameters are chosen by sampling a *prior* probability distribution $\pi(x)$. The prior distribution represents our knowledge or beliefs about x before ("prior to") looking at the data. The *data model* is that the data, Y, are generated from the likelihood function with the selected parameters. Therefore the joint distribution of random parameters $X \sim \pi(x)$, and conditional data $Y \sim L(y|X)$ is given by Bayes' rule"

$$(X, Y) \sim L(y|x)\pi(x) \; .$$

Once X and Y have been chosen as described, we look at Y. The remaining uncertainty over X is given by the conditional distribution of X given Y, which is

$$X \sim f(x|Y) = \frac{1}{Z(Y)} L(y|x)\pi(x) \; . \tag{10}$$

This is the *posterior* distribution of X, which is the distribution after ("posterior to") looking at the data. Modern Bayesian statistics consists of producing samples of the posterior.

The prior distribution π is the weak link in the Bayesian chain of ideas. It is not likely that we know enough about x before taking data to believe very strongly in a specific prior. This is particularly true, for example, when the parameters are physical constants like activation energies. Often the best we can do is to estimate bounds. For example, the activation energy must be positive, and it presumably is less that three electron volts, which is enough energy to break chemical bonds. There are similar physical arguments for bounds on the reaction rate prefactor and heat release parameter. The person building the detector may have a reasonable idea (say, within a factor of two) of the observation error parameter, σ. In the prior, the parameters may be taken to be independent and uniformly distributed within their bounds. There are cases where our prior bounds span many orders of magnitude. The masses of planets in our solar system are an example. Jupiter mass is more than 500 times earth mass. This makes it hard to put tighter priors on the masses that planets in other systems might have. If we assume the log of the parameter is uniformly distributed in a range, that is called (in astronomy) a *Jeffries prior.* We may take $\pi(x) = 1$ for all x. That is not *normalizable* in the sense that there is no Z so that $\frac{1}{Z}\pi(x)$ is a probability density. But the posterior distribution (10) may still be normalizable. That is called a *flat prior.* Slight less flat, but still not normalizable, are priors that enforce, say, positivity of a parameter (such as all the parameters in our chemical reaction model).

Priors are often used to "regularize" ill conditioned models. An example is given below.

An important weakness of this Bayesian approach is the need to specify a prior distribution for the parameters. In a majority of cases I have been involved in, the prior is chosen for its qualitative properties, rather than some theory or deep prior knowledge. An expression often used in turbulence modeling is appropriate here: *replacing ignorance with fiction.* We do not know exactly what the prior should be, so we make one up as best we can.

4 Gaussian examples

Example 1. Suppose the parameter X is a scalar. Before taking data, the statistician believes that $X \sim \mathcal{N}(\mu, \sigma^2)$. There are n independent observations of X, which have observation noise with standard deviation ρ. The observations may be written

$$Y_k \sim \mathcal{N}(\mathcal{X}, \rho^{\in}) \ ,$$

or as

$$Y_k = X + \rho Z_k \ \ , \quad Z_k \sim \mathcal{N}(0,1) \ \ , \quad \text{i.i.d.}$$

The prior is

$$\pi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \ .$$

The likelihood function is

$$L(y|x) = \frac{1}{(2\pi)^{n/2}} \frac{1}{\rho^n} \, e^{-\frac{1}{2\rho^2}\sum_{k=1}^{n}(y_k - x)^2} \ .$$

The posterior distribution is

$$\begin{aligned} f(x|Y) &= \frac{1}{Z}\pi(x)L(Y|x) \\ &= \frac{1}{Z} \ \exp\!\left(\frac{-1}{2}\left[\frac{(x-\mu)^2}{\sigma^2} + \sum_{k=1}^{n} \frac{(x-y_k)^2}{\rho^2} \right] \right) \ . \end{aligned}$$

This is a Gaussian, so it is determined by its mean and variance. The *posterior variance* is the coefficient of x^2 in the exponent, which is

$$\sigma_p^2 = \frac{\rho^2}{n + \frac{\sigma^2}{\rho^2}} \ .$$

The variance of the sample mean is

$$\sigma_{\overline{X}}^2 = \text{var}\left(\frac{1}{n} \sum_{k=1}^{n} Y_k \right) = \frac{\rho^2}{n} \ .$$

Comparing these expressions shows that the posterior variance is a little smaller. The mean of the posterior, because it is Gaussian, if the minimizer of the exponent. The result is

$$\widehat{X}_p = \frac{1}{n + \frac{\rho^2}{\sigma^2}} \left(\sum_{k=1}^{n} Y_k + \frac{\rho^2}{\sigma^2} \mu \right) \ .$$

This is an average of the data numbers Y_k, each taken with weight $\frac{1}{n}$, and the prior mean μ, taken with weight $\frac{\rho^2}{\sigma^2}$. This may be rewritten as

$$\begin{aligned} \widehat{X}_p &= \frac{1}{1 + \frac{\rho^2}{n\sigma^2}} \left(\frac{1}{n} \sum_{k=1}^{n} Y_k \right) \ + \ \frac{\frac{\rho^2}{n\sigma^2}}{1 + \frac{\rho^2}{n\sigma^2}} \ \mu \\ &= w_1 \overline{Y} + w_2 \mu \ , \quad w_1 + w_2 = 1 \ . \end{aligned} \tag{11}$$

This shows that the posterior mean is a weighted average of the sample mean and the prior. At least in the Gaussian case, Bayesian statistics means averaging the data together with your prior beliefs. The posterior variance in the Bayesian formulation is smaller than the variance of the posterior mean because the prior beliefs are considered to be data too. The relative weights of the prior belief and the data are determined by the ratio between the prior variance and the observation variance. The prior variance is smaller than the observation variance if $\sigma^2 < n\rho^2$. In that case, the prior mean μ gets more weight than the sample mean $\overline{Y}$ in sum (11) that gives the posterior mean.

Example 2. This example illustrates using the prior to regularize the problem. A linear least squares problem is to minimize

$$\| Ax - y \|_{l^2}^2 \ . \tag{12}$$

Here A is a coefficient matrix with n rows and d columns, x is a d component vector of parameters (or fitting coefficients), and y is an n component observation vector. We assume $d < n$. A perfect fit would be $Ax = y$, but this is probably impossible because there are more equations than unknowns, x. The *residual* from an imperfect fit is $r = Ax - y$. The linear least squares problem is to find coefficients x_j that minimize the sum of the squares of the residuals r_k.

As an example, consider fitting n data points y_k to a polynomial of degree $d-1$ in t. The fitting polynomial is $p(t) = x_0 + x_1 t + \cdots + x_{d-1}t^{d-1}$. The residuals are $r_k = p(t_k) - y_k$. In matrix form this is

$$\begin{pmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^{d-1} \\ 1 & t_2 & t_2^2 & \cdots & t_2^{d-1} \\ \vdots & \vdots & & & \vdots \\ 1 & t_n & t_n^2 & \cdots & t_n^{d-1} \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{d-1} \end{pmatrix} - \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{pmatrix} .$$

Many other fitting problems can be formulated in a similar way.

The linear least squares problem is equivalent to a maximum likelihood estimator of x in the probability model that the observations are given by

$$Y_k = p(t_k) + \rho Z_k \ , \quad Z_k \sim \mathcal{N}(0,1) \ , \quad \text{i.i.d.}$$

More generally, if a_k is row k of the matrix A, the probability model would be

$$Y_k = a_k x + \rho Z_k \ , \quad Z_k \sim \mathcal{N}(0,1) \ , \quad \text{i.i.d.}$$

Here, a_k is a d component row vector and x is a d component column vector. The likelihood function is

$$L(r|x) = \frac{1}{(2\pi\rho^2)^{n/2}} \exp\left(-\sum_{k=1}^{n} \frac{r_k^2}{2\rho^2} \right) .$$

Minimizing this likelihood is the same as minimizing the $\sum r_k^2$, which is the linear least squares problem.

The *singular value decomposition* of A (also called *principal component analysis*) is a way to solve the linear least squares problem, and to understand some difficulties that may come up. The singular value decomposition of A is the factorization

$$A = U\Sigma V^t \ ,$$

where V is a $d \times d$ orthogonal matrix (*orthogonal* means $VV^t = I$), and Σ is $d \times d$ diagonal matrix with *singular values* $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d > 0$ on the diagonal, and U is an $n \times d$ matrix with orthogonal unit norm columns, which means $U^t U = I$. The expression for the residual becomes

$$U\Sigma V^t x - y = r \ .$$

We multiply this from the left by U^t and use the fact that $U^t U = I$. The result is

$$\Sigma V^t x - U^t y = U^t r \ .$$

The columns of V are a convenient basis for x space and the columns of U are a convenient set of ortho-normal vectors in data space. We use tildes for the coefficients with respect to these vectors, which means $\widetilde{x} = V^t x$, $\widetilde{y} = U^t y$, and $\widetilde{r} = U^t r$. The equations become

$$\Sigma \widetilde{x} - \widetilde{y} = \widetilde{r} \ .$$

In components, this is

$$\sigma_j \widetilde{x}_j - \widetilde{y}_j = \widetilde{r}_j \ .$$

Note that by our conventions, $\widetilde{y}$ and $\widetilde{r}$ have $d < n$ components. Assuming (as we did above) that all the singular values are positive, we may set all components of $\widetilde{r}$ to zero, which gives

$$\widetilde{x}_j = \frac{1}{\sigma_j} \, \widetilde{y}_j \ . \tag{13}$$

This is the solution to the original linear least squares problem. We have chosen x that leaves a residual vector r that is orthogonal to the columns of A. It is impossible to make $\|r\|_{l^2}$ smaller than this.

It is very common, unfortunately, to have a linear least squares problem with a matrix A that has some very small singular values. Polynomial fitting with even moderate d is an example. The solution formula (14) implies that this gives a very large x. A proposed solution is *Tikhonov regularization*, which is to minimize

$$\|r\|_{l^2}^2 + \varepsilon^2 \, \|x\|_{l^2}^2 \ .$$

The point is to look for good fits while not having such large fitting coefficients. Minimizing over x using the SVD as above gives

$$\widetilde{x}_j = \frac{1}{\sqrt{\sigma_j^2 + \varepsilon^2}} \, \widetilde{y}_j \ . \tag{14}$$

If ε is smaller than σ_j, then the regularized solution is almost the same as without. But regularization essentially replaces the singular value σ_j with the *regularization parameter* ε if $\sigma_j \ll \varepsilon$. If done right, this can have a small effect on the quality of the fit, as measured by the residual norm $\|r\|_{l^2}^2$, but have a big impact on the size of the fitting parameters, as measured by the norm $\|x\|_{l^2}^2$.

The Bayesian interpretation Tikhonov regularization is simple, just introduce a prior with variance ε^{-2}

$$\pi(x) = \frac{1}{Z} e^{\frac{-\varepsilon^2}{2} \|x\|_{l^2}^2} \ .$$

Then the posterior mean is given by (14).

Using a regularizing prior, you can do Bayesian “estimation” with more parameters than data.

5 Bernstein theorem and uncertainty quantification

Uncertainty quantification (UQ) is the process of learning how precisely the parameters are constrained by the data. This is different from a *point estimate*, which means giving the "user" your best guess at the parameter values. It is better for the statistician also to communicate a level of confidence in the supplied numbers: how close are the numbers to being right? The modern Bayesian statistician will use the data to determine a posterior distribution of the parameters. This distribution has information on parameter uncertainties, and also on relations between parameters. There are posterior correlations.

How significant the posterior uncertainties and posterior correlations are depends on the application. In science and engineering, parameter estimates are used to predict what the system will do in situations where you do not have data yet. A more sophisticated form of UQ is finding the uncertainty of these predictions. It may be that the application model is insensitive to certain parameter variability in the same way the fitting model is. On the other hand, if a physician wants to know which artery in a patient is damaged, it may matter more if the answer is wrong.

The pre-Bayesian approach, the *frequentist* approach, to uncertainty quantification is confidence intervals and confidence regions. *Bernstein theorems* relate frequentist and Bayesian estimates of uncertainty, but always in the limit of large data sets. We describe a simple theorem of this type, which applies to the case of n independent samples from a PDF that depends on parameters. The frequentist point estimate is the maximum likelihood point, which is a random point that depends on the (random) data. In the limit of large n, the maximum likelihood point is approximately Gaussian with mean close to the true value and covariance determined by the *Fisher information.* The Bayesian posterior, for large n, is also approximately Gaussian, with (approximately) the same mean and covariance.

5.1 Maximum likelihood distribution for large n

We write $f(y|x)$ for the PDF for random variable Y with parameters x. Let Y_k, for $k = 1, \ldots, n$ be iid. samples of f with parameter set x_*. The likelihood function is

$$L(\{Y_k\}\,|x) = \prod_{k=1}^{n} f(Y_k|x) \; . \tag{15}$$

The maximum likelihood point estimate of x is

$$\widehat{X} = \arg\ \max_{x} L(\{Y_k\}\,|x) \; .$$

For large n, and with some non-degeneracy hypotheses on f, there is a central limit theorem analysis of the distribution of $\widehat{X}$.

The analysis starts with the log likelihood represention

$$g(y|x) = \log(f(y|x)) \; ,$$

The maximum likelihood point also maximizes $\log(L)$, so

$$\nabla_x \log[\, L(\{Y_k\}\,|x)] = 0 \; .$$

In terms of g, this is

$$\sum_{k=1}^{n} \nabla_x g(Y_k|\widehat{X}) = 0 \ . \qquad (16)$$

We analyze $\widehat{X} = x_*$ in two steps. First we use the central limit theorem to quantify the *residual* in equation (16) when $\widehat{X}$ is replaced by x_*. Then a Taylor series analysis relates this residual to the error, $\widehat{X} = x_*$. Two matrices arise. It seems to be a coincidence that these are equal.

The *residual* is the amount by which an equation fails to be satisfied. This term is used in numerical analysis and in statistics. Define the residual in (16) with x_* as

$$R = \sum_{k=1}^{n} \nabla_x g(Y_k|x_*) \ .$$

The central limit applies because this is the sum of a large number of independent terms. The expected value of a term is

$$\begin{aligned} \mathrm{E}[\,\nabla_x g(Y|x_*)] &= \int \nabla_x g(y|x_*) f(y|x_*)\, dy \\ &= \int \frac{\nabla_x f(y|x_*)}{f(y|x_*)} f(y|x_*)\, dy \\ &= \int \nabla_x f(y|x_*)\, dy \\ &= 0 \ . \end{aligned}$$

The last line is because f is a probability density as a function of y for each x. Therefore,

$$\int \nabla_x f(y|x_*)\, dy = \nabla_x \int f(y|x_*)\, dy = \nabla_x 1 = 0 \ .$$

The cancellation in the second line, with $f(y|x_*)$ in the denominator and numerator, is where $x = x_*$ is used.

Now that $\mathrm{E}[R] = 0$, the nature of R is determined by the covariance matrix of the summands. If V is a column vector with mean zero, the covariance matrix is

$$\mathrm{cov}(V) = \mathrm{E}\big[\,VV^t\big] \ .$$

The covariance of $\nabla_x g(Y|x_*)$ is a matrix called I, the *Fisher information*

$$I = \mathrm{cov}(\nabla_x g(Y|x_*)) = \mathrm{E}\bigg[\Big(\nabla_x g(Y|x_*)\Big)\Big(\nabla_x g(Y|x_*)\Big)^t\bigg] \ . \qquad (17)$$

For future reference, note that the entries of I are

$$\begin{aligned} I_{ij} &= \mathrm{E}\big[\partial_{x_i} g(Y|x_*)\partial_{x_j} g(Y|x_*)\big] \\ &= \mathrm{E}\bigg[\frac{\partial_{x_i} f(Y|x_*)\,\partial_{x_j} f(Y|x_*)}{f^2(Y,x_*)}\bigg] \\ &= \int \frac{\partial_{x_i} f(y|x_*)\,\partial_{x_j} f(y|x_*)}{f(y,x_*)}\, dy \ . \end{aligned}$$

We can see that I has something to do with parameter uncertainty by asking about the case $I = 0$. In that case $\nabla_x g(y|x_*) = 0$, which implies that $\nabla_x f(y|x_*) = 0$. If f does not depend on x, it will not be possible to use f to determine x. The definition (17) implies that I is positive semi-definite. We assume, as is usually true in applications, that I is non-degenerate and positive definite. In that case, the central limit theorem says that if n is large, then R is approximately normal with mean zero and covariance nI, with I being the Fisher information (17).

The second step in characterizing $\widehat{X} - x_*$ uses a Taylor approximation in (16). Let H be the Hessian matrix of g evaluated at x_*

$$H_{ij} = \partial_{x_i}\partial_{x_j} g(y|x_*) \ .$$

This H is a random variable because it depends on the random $Y \sim f(y|x_*)$. The Hessian matrices for the data values Y_k will be written $H_k = H(Y_k)$. The Taylor approximation is

$$\nabla_x g(y|\widehat{X}) \approx \nabla_x g(y|x_*) + H(y)\Big(\widehat{X} - x_*\Big) \ .$$

With this approximation, (16) becomes

$$R + M\Big(\widehat{X} - x_*\Big) \approx 0 \ , \quad \widehat{X} - x_* \approx -M^{-1}R \tag{18}$$

with

$$M = \sum_{k=1}^{n} H_k \ . \tag{19}$$

The relation between $M \approx n\mathrm{E}[H]$ and I is the coincidence that makes I so important.

The calculation that led to I can be continued:

$$\begin{aligned}\mathrm{E}\big[\partial_{x_i}\partial_{x_j} g(Y|x_*)\big] &= \mathrm{E}\left[\partial_{x_i}\left(\frac{\partial_{x_j} f(Y|x_*)}{f(Y,x_*)}\right)\right] \\ &= \mathrm{E}\left[\frac{\partial_{x_i}\partial_{x_j} f(Y|x_*)}{f(Y,x_*)}\right] - \mathrm{E}\left[\frac{\partial_{x_i} f(Y|x_*)\,\partial_{x_j} f(Y|x_*)}{f^2(Y,x_*)}\right] \ .\end{aligned}$$

The first expectation on the last line vanishes for a reason we just saw. Since $Y \sim f(y|x_*)$,

$$\mathrm{E}\left[\left(\frac{\partial_{x_i}\partial_{x_j} f(Y|x_*)}{f(Y|x_*)}\right)\right] = \int \partial_{x_i}\partial_{x_j} f(y|x_*)\,dy = \partial_{x_i}\partial_{x_j}\,1 = 0 \ .$$

Therefore

$$\mathrm{E}[H(Y|x_*)] = -I \ .$$

Since I is positive definite, the law of large numbers applies to the sum (19) up to statistical fluctuations:

$$M = -nI + O(\sqrt{n}) \ .$$

The approximations for R and M now characterize $\widehat{X} - x_*$, asymptotically for large n. We saw that R is approximately normal with mean zero and $\mathrm{cov}(R) \approx nI$. We also saw that M is approximately deterministic with $M \approx nI$. The formula (18) then expresses (approximately) $\widehat{X} - x_*$ as a linear function of a mean zero Gaussian. This makes $\widehat{X} - x_*$ itself mean zero Gaussian, with covariance (M is symmetric, so $\left(M^{-1}\right)^t = M^{-t} = M^{-1}$)

$$\begin{aligned} \mathrm{cov}\left[\widehat{X} - x_*\right] &\approx \mathrm{cov}\left[M^{-1}\left(\widehat{X} - x_*\right)\right] \\ &\approx M^{-1}\mathrm{cov}\left[\widehat{X} - x_*\right] M^{-t} \\ &\approx \left(\frac{1}{n}I^{-1}\right)(nI)\left(\frac{1}{n}I^{-1}\right) \\ \mathrm{cov}\left[\widehat{X} - x_*\right] &\approx \frac{1}{n}I^{-1} \; . \end{aligned} \tag{20}$$

This is the famous and very useful asymptotic characterization of the error in maximum likelihood estimation. The covariance decays like $1/n$, which means that the actual estimation error decays like $1/\sqrt{n}$. The inverse of the information matrix indicates that the more "information" there is in the dependence of f on x, the more accurately x can be determined.

5.2 Large n behavior of the Bayesian posterior

The Bernstein theorem for this case (independent samples, non-degeneracy hypotheses) is that the posterior is (approximately, for large n) normal with mean (approximately) the true parameter set and covariance (20). This result is useful even if you're a frequentist. If you are a frequentist wanting the covariance matrix of $\widehat{X}$, you can get it by sampling the posterior. If you're a Bayesian, this gives you some confidence that the Bayesian posterior has correct information about the remaining parameter uncertainty.

We study the "posterior" without a prior, which is the likelihood function. If x is the parameter set, generated at random if you believe the official Bayesian story, then the posterior of x, given n independent samples $Y_k \sim f(y|x)$, is just the normalized likelihood function (15),

$$X \sim F(x) = \frac{1}{Z(\{Y_k\})} L(\{Y_k\}\,|x) \; .$$

(Apologies for the notation, $f(y|x)$ is the PDF of samples, $F(x) = F(x|\,\{Y_k\})$ is the posterior.) We assume that $F(x) = e^{-\phi(x)}$, and that $\phi(x)$ is well approximated by a quadratic, at least for likely values of x. This is the situation of the Laplace method used in Assignment 1. The picture (quadratic or unlikely) can be verified using large deviation theory. That theory shows that x values far from the posterior mean, far enough so that the quadratic approximation is invalid, are very unlikely.

The calculations that give us a picture of ϕ are similar to the calculations above related to Fisher information. Here, we write (still using the notation $g = \log(f)$)

$$\begin{aligned}\phi(x) &= -\log(F(x)) \\ &= -\sum_{k=1}^{n} g(Y_k|x) \ .\end{aligned}$$

We find the minimum of ϕ by setting the derivative with respect to x equal to zero. This is the maximum likelihood point (16). The maximum likelihood point is the *mode* of the posterior, if you use a "flat prior". If you have a non-trivial prior, the maximum of the posterior is called *MAP*, or *maximum a posteriori*. We will see that the MAP point and the maximum likelihood point are close (with high probability) if the prior allows it. The Hessian of ϕ is the negative of M in (19). The analysis above shows that this is approximately nI, in terms of the Fisher information, I.

Putting in a prior usually does not change this picture much. If x is the true value that generates the data $\{Y_k\}$, and if $\pi(x)$ is continuous, and if $\pi(x) \neq 0$, then the "true" posterior $F = \frac{1}{Z}\pi L$ has the same asymptotic behavior as the posterior with a flat prior. To be specific, the MAP point, the maximizer of F, converges to the actual value as $n \to \infty$. Also, the posterior distribution about the MAP point converges to a Gaussian with covariance $\frac{1}{n}I^{-1}$. These facts are easy to verify.