Supplementary notes

Introduction to Computational Statistics

1 Introduction

Most practical statistics involves computational algorithms. Therefore, a course on computational statistics is, at least in part, just a course on statistics. Some basic interrelated questions include

1. What can you learn about something from partial and inaccurate data about it?
2. How accurate is your inference?
3. How do you predict behavior of something based on partial information about it?
4. How do errors in estimation impact predictions?
5. What measurement or data collection strategies help the most?
6. What actions (controls) should you take based on the data? This depends on the cost or value of outcomes that depend in part on parameters that you infer, with errors, from data.

The first question may be called *statistical inference*. *Parameter estimation* is a basic form of statistical inference. The second question addresses *uncertainty quantification*. Traditional error bars/confidence intervals and modern Bayesian are commonly used tools. The third and fourth questions are related to *uncertainty propagation*. Parameter uncertainties affect some predictions more than others. The fifth question is related to *experimental design*. Some designs are *adaptive* in that outcomes of early measurements influence future measurement choices. Different communities refer to the sixth issue as *decision theory* or *adaptive stochastic control*.

Traditional statistical methods often relied on unrealistic idealized models that lead to simple computations. Some of these simplifications are called "linear, quadratic, gaussian". This means linear models, quadratic costs for errors, and gaussian measurement errors. The resulting statistical estimation and control methods are linear regression and Kalman filtering. The computations involve linear algebra. If you look, for example, in a book on econometrics (statistical methods for economics), you will find dozens of models concocted ingeniously to have closed form or computationally simple solutions.

Modern computational statistics seeks general computational methods, typically calling for much more computing resources, that can handle more general models. The goal is to allow the statistical practitioner to use models that they think are realistic rather than ones that have tractable solution algorithms. The job of the computational statistician is to develop general computational algorithms that allow more realistic modeling. That's what this class is about.

There are important aspects of practical statistics not discussed here, including (but not limited to)

1. Data collection and analysis protocols. Data collection protocols include *double blinding* in drug trials in which neither the patient nor the healthcare worker knows whether the patient has received the drug being tested. Hypothesis testing protocols ask, for example, whether you can confirm ideas about the data you get after looking at the data.
2. Theoretical questions such as the (approximate) probability distribution of uncertain parameter estimates in limits such as large sample size.

Terminology

Look here if you see something in a formula you don't understand. You will find different ways of expressing the same thing. You don't have to read and digest all this now.

Random variables A random variable with n components will be written $X = (X_1, \cdots, X_n)$. Each of the components, X_k is a random number. We write $X \in \mathbb{R}^n$. If $n = 1$ (a one component random variable), we write X for X_1. We write use lower-case letters to represent generic values a random variable might take. For example, we might write "if $X_2 > x_2$" to ask whether the random number is larger than the value x_2.

Probability Probabilities or probability densities are usually denoted by p, P, Pr or π. A probability density is usually $p(x)$ or $\pi(x)$. We call p the PDF (probability density function) of X and write $X \sim p$. The probability of an event A is $\Pr(A)$ or $P(A)$. For example, if X is a one component random variable, then

$$\Pr(a \leq X \leq b) = \int_a^b p(x)\,dx \ .$$

If $X \in \mathbb{R}^n$ then

$$\Pr(\|X\| \leq 1) = \int_{\|x\| \leq 1} p(x)\,dx \ .$$

Marginal probabilities If $X = (X_1, X_2) \in \mathbb{R}^2$, the marginal probability density of X_1 is

$$p(x_1) = \int_{-\infty}^{\infty} p(x_1, x_2)\,dx_2 \ .$$

Be aware that the p in the left is a function of one variable while the p on the right is a function of two variables. The probability density that $p(x)$ represents can be different in different formulas or even in the same formula, as above.

Normalization A normalization factor (or normalization constant) is a factor that you apply to a vector or a function to make it satisfy some formula. In probability, we often know the PDF only "up to a normalization constant". This means that $p(x) = \frac{1}{Z} f(x)$ where there is a formula for f but the normalization constant Z is unknown. This might be written $p \propto f$. The normalization factor is given by the integral of f, because

$$\int p(x)\,dx = \frac{1}{Z} \int f(x)\,dx = 1 \ .$$

Therefore

$$Z = \int f(x)\,dx \ .$$

For example, the standard normal (one component, mean zero, variance 1, gaussian) random variable has a density

$$p(x) \propto f(x) = e^{-\frac{1}{2}x^2} \ .$$

The normalization factor is

$$Z = \int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}\,dx = \sqrt{2\pi} \ .$$

There may be no explicit formula for Z. An example is $p(x) \propto f(x) = e^{-x^4}$. There is no expression for the integral

$$Z = \int e^{-x^4}\,dx \ .$$

Writing $\frac{1}{Z} f(x)$ without the formula for Z may be convenient even when there is a formula for Z. For example, the Student T density with n degrees of freedom is

$$p(x) \propto f(x) = \frac{1}{\left(1 + \frac{x^2}{n}\right)^{\frac{n+1}{2}}} \quad .$$

The normalization factor is

$$Z = \frac{\sqrt{\pi n}\,\Gamma(\frac{n}{2})}{\Gamma(\frac{n+1}{2})} \quad .$$

Conditional probability If A and B are two events, the conditional probability of A given B is $\Pr(A|B)$. The event that both A and B happen is written "A and B" or $A \cap B$. Bayes' rule for probabilities is

$$\Pr(A|B) = \frac{\Pr(A \text{ and } B)}{\Pr(B)} \; .$$

If $X = (X_1, X_2)$ has PDF $p(x_1, x_2)$, the conditional density of X_1 given a value of X_2 is given by the PDF form of Bayes' rule:

$$X_1 \sim p(x_1|X_2 = x_2) = p(x_1|x_2) = \frac{1}{Z(x_2)}\, p(x_1, x_2) \; .$$

In this, $p(x_1|x_2)$ is a shorter and less formal way to write $p(x_1|X_2 = x_2)$. The conditional probability is a PDF only as a function of x_1. Therefore, the normalization constant can depend on x_2. As before, the normalization constant is found by requiring that the integral is equal to 1. It can be different for each value of the parameter x_2:

$$1 = \int p(x_1|x_2)\, dx_1 = \int \frac{1}{Z(x_2)}\, p(x_1, x_2)\, dx_1 \implies Z(x_2) = \int p(x_1, x_2)\, dx_1 \; .$$

The formula for $Z(x_2)$ is the same as the marginal probability density for X_2, which can be expressed in various ways, including

$$Z(x_2) = p(x_2) \quad , \quad p(x_1|x_2) = \frac{p(x_1, x_2)}{p(x_2)} \; .$$

The probability densities p in the numerator and denominator on the right are different functions denoted by the same letter.

Parameters A probability distribution may depend on parameters. If there are d parameters, we call them $(\theta_1, \cdots, \theta_d) = \theta$. Parameters usually have specific names in specific applications. For example, a one dimensional Gaussian has parameters μ (the mean) and σ^2 (the variance), so we might write $\theta = (\mu, \sigma^2)$. The normalization constant can depend on the parameters. For example, the gamma distribution for a positive random variable X may be written as

$$p(x) \propto x^\beta e^{-\alpha x} \quad \text{or} \quad p(x) = \frac{1}{Z(\alpha, \beta)}\, x^\beta e^{-\alpha x} \; .$$

It is common to use (misuse) conditional probability notation or put the parameters as subscripts:

$$p(x|\theta) = p_\theta(x) \; .$$

Expected value The expectation value of a random number Y is $\mathrm{E}[Y]$. If the distribution of Y depends on parameters, we put the parameters in somewhere. Some possibilities are:

$$\mathrm{E}[Y|\theta] = \mathrm{E}_\theta[Y] = \mathrm{E}^\theta[Y]$$

The "observable" (or "quantity of interest", QOI), Y, may be a function of the possibly multivariate random variable X. We write this as $Y = u(X)$. In this case

$$\mathrm{E}_\theta[u(X)] = \int u(x)\, p(x|\theta)\, dx \ .$$

Data We will usually use Y to denote a dataset. A dataset may consist of m "observations", $Y = (Y_1, \cdots, Y_m)$. Each observation may have ℓ components, which is expressed by writing observation Y_j in terms of its components: $Y_j = (Y_{j,1} \cdots, Y_{j,\ell})$. Whole dataset then consists of $m{\cdot}\ell$ numbers. It is common to organize such a dataset into an $m \times \ell$ data matrix, also called Y. Row j of Y contains all the components of observation Y_j. Column k of Y contains component k from all of the observations.

$$Y = \begin{pmatrix} Y_{1,1} & Y_{1,2} & \cdots & Y_{1,\ell} \\ Y_{2,1} & Y_{2,2} & \cdots & Y_{2,\ell} \\ \vdots & \vdots & & \vdots \\ Y_{m,1} & Y_{m,2} & \cdots & Y_{m,\ell} \end{pmatrix} \ .$$

2 Maximum Likelihood

In statistics, *point estimation* means using a dataset to make a guess at the values of some unknown set of parameters. The d parameters are $\theta = (\theta_1, \cdots, \theta_d)$. The *estimator* is a function of the data

$$\widehat{\theta} = \widehat{\theta}(Y) = (\, \widehat{\theta}_1(Y), \cdots, \widehat{\theta}_d(Y)\,) \ .$$

The term "point estimate" emphasizes that the estimator function returns just one estimate, which is a "point" in parameter space: $\widehat{\theta}(Y) \in \mathbb{R}^d$. "Point estimate" also emphasizes that the estimator does not return other information, such as the accuracy or reliability of the estimate relative to the true parameters. We do not learn how close $\widehat{\theta}$ is likely to be to the true but unknown parameters θ^*. That important issue, which is sometimes called *uncertainty quantification* is addressed in one way below using the *Fisher information* matrix, I. It is addressed in a deeper way using Bayesian statistics later in the course.

Maximum likelihood estimation chooses the $\widehat{\theta}$ that maximizes the "likelihood" of getting data Y given parameters θ. A *probability data model* is a PDF $p(y|\theta)$ that gives the probability density of the dataset y given parameters θ. The function $p(y|\theta)$ is called a probability density when it is thought of as a function of y. If you fix y to be the dataset you actually have and look at $p(Y|\theta)$ as a function of θ, that is called a *likelihood* function. The term "likelihood" as opposed to "probability" is supposed to remind you that θ is not a random variable and $p(Y|\theta)$ is not a probability density as a function of θ.

The *frequentist* point of view for parameter estimation is that the true parameter set θ^* is unknown but not random. Suppose you're trying to estimate, say, the speed of a car from police radar measurements. The estimate, $\widehat{\theta}$, will be different from the true speed θ^*. The difference between $\widehat{\theta}(Y)$ and θ^* is random only because Y is random. The probability model says the dataset has distribution $Y \sim p(\cdot|\widehat{\theta}^*)$. What makes $\widehat{\theta}$ random, is that $\widehat{\theta}(Y)$ is a function of the random variable Y. There is also a *Bayesian* point of view in which θ^* is also random and given by some *prior* distribution $\theta^* \sim \pi(\cdot)$. Some advantages and disadvantages of the Bayesian approach are given later.

The maximum likelihood estimate of θ^* is the parameter combination that maximizes the likelihood function

$$\widehat{\theta}(Y) = \arg\max_\theta \ p(Y|\theta) \ . \tag{1}$$

This chapter discusses three aspects of maximum likelihood estimation:

1. What functions p are likely to arise as probability data models?

2. Given a dataset, Y, and a probability model, p, how do you solve the optimization problem (1) that determines $\widehat{\theta}$?

3. What can you learn about the probability distribution of $\widehat{\theta}$ from that dataset, Y?

Points 2 and 3 are useful contexts for discussing many aspects of computational statistics. For the optimization problem (1), there are generic optimization algorithms related to gradient descent and Newton's method. There are also specialized methods for problems that often arise in statistical applications. One is *expectation maximization*, which is used when some of the components $Y_{j,k}$ are unknown (the "missing data") problem. Another is *stochastic gradient descent*, which is used then the dataset Y is so large that you don't want to use all of Y every iteration.

There is corollary of the central limit theorem saying that if there are enough observations in the dataset then the distribution of $\widehat{\theta}$ is approximately a multivariate normal whose covariance matrix is the inverse of the *Fisher information* matrix I. The entries of I are expected values (integrals) over the m dimensional datapoint space of possible values of an observation. For dimensions higher than $m > 4$ (say), Monte Carlo is the most practical way to estimate this integral.

The central tool of maximum likelihood estimation is numerical optimization. This involves classical optimization ideas such as gradient descent, Newton methods, line search, etc. More recently, there are variants on these ideas for cases in which the likelihood function is expensive to evaluate because of complex models or large datasets. For complex models there are methods such as model reduction that derive simplified approximations to help guide the optimization more efficiently. For large datasets there are methods derived from stochastic gradient descent, which make approximate models using small randomly chosen pieces of the data.

3 Bayesian methods, sampling

The modern Bayesian point of view is that the data "constrain" the model rather than specifying it exactly as a point estimate would. In some sense (explained in detail later), the Bayesian approach defines a probability distribution in parameter space, with more likely parameter combinations being ones that fit the data better. This is the *posterior* distribution, "posterior" meaning a distribution that is determined only *after* the data are collected. *Sampling* methods are used to explore the posterior. Using sampling methods from statistical physics, particularly Markov chain Monte Carlo (MCMC), has made all this possible. The most effective modern sampling algorithms often require the "user" (the computational statistician) to design methods tailored to specific applications. It also is challenging to determine how effective an MCMC sampler has been.

4 Inference and discovery

Frequentist and Bayesian are two approaches to inference, which often means learning about parameters in models that are specified up to parameters. *Discovery* means looking for (finding, hopefully) patterns in data that were not necessarily expected. A related issue is making predictions that you hope are accurate but that are heuristic in that they are not based on statistical models. Much machine learning is like this. As a simple example, a modern neural net can be trained to distinguish cat from dog images reasonably well. But the net itself has millions of tunable parameters and nobody knows exactly how the net tells cats from dogs.

This is not a course on machine learning but we will cover some ingredients, including feature selection, regularization (which seems necessary when a prediction model has many parameters), and stochastic optimization algorithms related to stochastic gradient descsent.