Chapter 1: Preliminaries

Bayesian inference in simple conjugate families

We start with a few of the simplest building blocks for complex multivariate statistical models: the beta/binomial, normal, and inverse-gamma conjugate families.

(A) Suppose that we take independent observations $x_1, \ldots, x_N$ from a Bernoulli sampling model with unknown probability w. That is, the x_i are the results of flipping a coin with unknown bias. Suppose that w is given a Beta(a,b) prior distribution:

$$p(w) = \frac{\Gamma(a+b)}{\Gamma(a) \cdot \Gamma(b)} \, w^{a-1}(1-w)^{b-1} \,,$$

where $\Gamma(\cdot)$ denotes the Gamma function. Derive the posterior distribution $p(w \mid x_1, \ldots, x_N)$.[1]

(B) The probability density function (PDF) of a gamma random variable, $x \sim \mathrm{Ga}(a,b)$, is

$$p(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx) \,.$$

Suppose that $x_1 \sim \mathrm{Ga}(a_1, 1)$ and that $x_2 \sim \mathrm{Ga}(a_2, 1)$. Define two new random variables $y_1 = x_1/(x_1 + x_2)$ and $y_2 = x_1 + x_2$. Find the joint density for (y_1, y_2) using a direct PDF transformation (and its Jacobian).[2] Use this to characterize the marginals $p(y_1)$ and $p(y_2)$, and propose a method that exploits this result to simulate beta random variables, assuming you have a source of gamma random variables.

(C) Suppose that we take independent observations $x_1, \ldots, x_N$ from a normal sampling model with unknown mean θ and *known* variance σ^2: $x_i \sim \mathrm{N}(\theta, \sigma^2)$. Suppose that θ is given a normal prior distribution with mean m and variance v. Derive the posterior distribution $p(\theta \mid x_1, \ldots, x_N)$.

(D) Suppose that we take independent observations $x_1, \ldots, x_N$ from a normal sampling model with *known* mean θ but *unknown* variance σ^2. (This seems even more artificial than the last, but is conceptually important.) To make this easier, we will re-express things in terms of the precision, or inverse variance $\omega = 1/\sigma^2$:

$$p(x_i \mid \theta, \omega) = \left(\frac{\omega}{2\pi}\right)^{1/2} \exp\left\{-\frac{\omega}{2}(x_i - \theta)^2\right\} \,.$$

[1] I offer two tips here that are quite general. (1) Your final expression will be cleaner if you reduce the data to a sufficient statistic. (2) Start off by ignoring normalization constants (that is, factors in the density function that do not depend upon the unknown parameter, and are only there to make the density integrate to 1.) At the end, re-instate these normalization constants based on the functional form of the density.

[2] Take care that you apply the important change-of-variable formula from basic probability. See, e.g., Section 1.2 of `http://www.stat.umn.edu/geyer/old/5102/n.pdf`.

Suppose that ω has a gamma prior with parameters a and b, implying that σ^2 has what is called an inverse-gamma prior.[3] Derive the posterior distribution $p(\omega \mid x_1, \ldots, x_N)$. Re-express this as a posterior for σ^2, the variance.

[3] Written $\sigma^2 \sim \text{IG}(a, b)$.

(E) Suppose that, as above, we take independent observations $x_1, \ldots, x_N$ from a normal sampling model with unknown, common mean θ. This time, however, each observation has its own idiosyncratic (but known) variance: $x_i \sim \text{N}(\theta, \sigma_i^2)$. Suppose that θ is given a normal prior distribution with mean m and variance v. Derive the posterior distribution $p(\theta \mid x_1, \ldots, x_N)$. Express the posterior mean in a form that is clearly interpretable as a weighted average of the observations and the prior mean.

(F) Suppose that $(x \mid \omega) \sim \text{N}(m, \omega^{-1})$, and that ω has a Gamma$(a/2, b/2)$ prior, with PDF defined as above. Show that the marginal distribution of x is Student's t with d degrees of freedom, center m, and scale parameter $(b/a)^{1/2}$. This is why the t distribution is often referred to as a *scale mixture of normals*.

The multivariate normal distribution

Basics

We all know the univariate normal distribution, whose long history began with de Moivre's 18th-century work on approximating the (analytically inconvenient) binomial distribution. This led to the probability density function

$$p(x) = \frac{1}{\sqrt{2\pi v}} \exp\left\{-\frac{(x-m)^2}{2v}\right\}$$

for the normal random variable with mean m and variance v, written $x \sim \text{N}(m, v)$.

Here's an alternative characterization of the univariate normal distribution in terms of moment-generating functions:[4] a random variable x has a normal distribution if and only if $E\{\exp(tx)\} = \exp(mt + vt^2/2)$ for some real m and positive real v. Remember that $E(\cdot)$ denotes the expected value of its argument under the given probability distribution. We will generalize this definition to the multivariate normal.

[4] Laplace transforms to everybody but statisticians.

(A) First, some simple moment identities. The covariance matrix $\text{cov}(x)$ of a vector-valued random variable x is defined as the matrix whose (i, j) entry is the covariance between x_i and x_j. In matrix notation, $\text{cov}(x) = E\{(x-\mu)(x-\mu)^T\}$, where μ is the mean

vector whose ith component is $E(x_i)$. Prove the following: (1) $\text{cov}(x) = E(xx^T) - \mu\mu^T$; and (2) $\text{cov}(Ax + b) = A\text{cov}(x)A^T$ for matrix A and vector b.

(B) Consider the random vector $z = (z_1, \ldots, z_p)^T$, with each entry having an independent standard normal distribution (that is, mean 0 and variance 1). Derive the probability density function (PDF) and moment-generating function (MGF) of z, expressed in vector notation.[5] We say that z has a standard multivariate normal distribution.

[5] Remember that the MGF of a vector-valued random variable x is the expected value of the quantity $\exp(t^T x)$, as a function of the vector argument t.

(C) A vector-valued random variable $x = (x_1, \ldots, x_p)^T$ has a *multivariate normal distribution* if and only if every linear combination of its components is univariate normal. That is, for all vectors a not identically zero, the scalar quantity $z = a^T x$ is normally distributed. From this definition, prove that x is multivariate normal, written $x \sim \text{N}(\mu, \Sigma)$, if and only if its moment-generating function is of the form $E(\exp\{t^T x\}) = \exp(t^T\mu + t^T\Sigma t/2)$. Hint: what are the mean, variance, and moment-generating function of z, expressed in terms of moments of x?

(D) Another basic theorem is that a random vector is multivariate normal if and only if it is an affine transformation of independent univariate normals. You will first prove the "if" statement. Let z have a standard multivariate normal distribution, and define the random vector $x = Lz + \mu$ for some $p \times p$ matrix L of full column rank.[6] Prove that x is multivariate normal. In addition, use the moment identities you proved above to compute the expected value and covariance matrix of x.

[6] The full rank restriction turns out to be unnecessary; relaxing it leads to what is called the *singular normal distribution.*

(E) Now for the "only if." Suppose that x has a multivariate normal distribution. Prove that x can be written as an affine transformation of standard normal random variables. (Note: a good way to prove that something can be done is to do it!) Use this insight to propose an algorithm for simulating multivariate normal random variables with a specified mean and covariance matrix.

(F) Use this last result, together with the PDF of a standard multivariate normal, to show that the PDF of a multivariate normal $x \sim \text{N}(\mu, \Sigma)$ takes the form $p(x) = C\exp\{-Q(x-\mu)/2\}$ for some constant C and quadratic form $Q(x - \mu)$.[7]

[7] A useful fact is that the Jacobian matrix of the linear map $x \to Ax$ is simply A.

(G) Let $x_1 \sim N(\mu_1, \Sigma_1)$ and $x_2 \sim N(\mu_2, \Sigma_2)$, where x_1 and x_2 are independent of each other. Let $y = Ax_1 + Bx_2$ for matrices A, B of full column rank and appropriate dimension. Note that x_1 and x_2

need not have the same dimension, as long as Ax_1 and Bx_2 do. Use your previous results to characterize the distribution of y.

Conditionals and marginals

Suppose that $x \sim \mathrm{N}(\mu, \Sigma)$ has a multivariate normal distribution. Let x_1 and x_2 denote an arbitrary partition of x into two sets of components. Because we can relabel the components of x without changing their distribution, we can safely assume that x_1 comprises the first k elements of x, and x_2 the last $p - k$. We will also assume that μ and Σ have been partitioned conformably with x:

$$\mu = (\mu_1, \mu_2)^T \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} .$$

Clearly $\Sigma_{21} = \Sigma_{12}^T$, as Σ is a symmetric matrix.

(A) Derive the marginal distribution of x_1. (Remember your result about affine transformations.)

(B) Let $\Omega = \Sigma^{-1}$ be the inverse covariance matrix, or precision matrix, of x, and partition Ω just as you did Σ:

$$\Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{12}^T & \Omega_{22} \end{pmatrix} .$$

Using (or deriving!) identities for the inverse of a partitioned matrix, express each block of Ω in terms of blocks of Σ.

(C) Derive the conditional distribution for x_1, given x_2, in terms of the partitioned elements of x, μ, and Σ. There are several keys to inner peace: work with densities on a log scale, ignore constants that don't affect x_1, and remember the cute trick of completing the square from basic algebra.[8] Explain briefly how one may interpret this conditional distribution as a linear regression on x_2, where the regression matrix can be read off the precision matrix.

[8] In scalar form:

$$\begin{aligned} x^2 - 2bx + c &= x^2 - 2bx + b^2 - b^2 + c \\ &= (x-b)^2 - b^2 + c\,. \end{aligned}$$

Multiple regression: three classical principles for inference

Suppose we observe data that we believe to follow a linear model, where $y_i = x_i^T \beta + \epsilon_i$ for $i = 1, \ldots, n$. To fix notation: y_i is a scalar response; x_i is a p-vector of predictors or features; and the ϵ_i are errors. By convention we write vectors as column vectors. Thus $x_i^T \beta$ will be our typical way of writing the inner product between the vectors x_i and β.

Notice we have no explicit intercept. For now you can imagine that all the variables have had their sample means subtracted, making an intercept superfluous. Or you can just assume that the leading entry in every x_i is equal to 1, in which case β_1 will be an intercept term.

Consider three classic inferential principles that are widely used to estimate β, the vector of regression coefficients. In this context we will let $\hat{\beta}$ denote an estimate of β—think, it wears a hat because it's masquerading as the true value.[9]

[9] This metaphor once came back to me in somewhat garbled fashion on an undergraduate's midterm: "$\hat{\beta}$ wears a hat because he is an impostor."

Least squares: make the sum of squared errors as small as possible.

$$\hat{\beta} = \arg\min_{\beta \in \mathcal{R}^p} \left\{ \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 \right\} .$$

Maximum likelihood under Gaussianity: assume that the errors are independent, mean-zero normal random variables with common variance σ^2. Choose $\hat{\beta}$ to maximize the likelihood:

$$\hat{\beta} = \arg\max_{\beta \in \mathcal{R}^p} \left\{ \prod_{i=1}^{n} p(y_i \mid \beta, \sigma^2) \right\} .$$

Here $p_i(y_i \mid \sigma^2)$ is the conditional probability density function of y_i, given the model parameters β and σ^2.

Method of moments: Choose $\hat{\beta}$ so that the sample covariance between the errors and each of the p predictors is exactly zero. This gives you a system of p equations and p unknowns.

(A) Show that all three of these principles lead to the same estimator.[10]

[10] You will end up tearing your hair out if you try to deal with sums of scalar quantities. Thus convert everything to matrix-vector notation. Remember your basic results on moments of linear combinations of random variables, and the following two identities on derivatives of linear and quadratic forms:

$$\begin{aligned} \frac{\partial (a^T x)}{\partial x} &= a \\ \frac{\partial (x^T A x)}{\partial x} &= (A + A^T) x . \end{aligned}$$

(B) Now suppose you trust some observations more than others, and will estimate β by minimizing the weighted sum of squared errors,

$$\hat{\beta} = \arg\min_{\beta \in \mathcal{R}^p} \left\{ \sum_{i=1}^{n} w_i (y_i - x_i^T \beta)^2 \right\} ,$$

where the w_i are weights. (Trustworthy observations have large weights.) Derive this estimator, and show that it corresponds to the maximum-likelihood solution under *heteroscedastic* Gaussian error:

$$\hat{\beta} = \arg\max_{\beta \in \mathcal{R}^p} \left\{ \prod_{i=1}^{n} p(y_i \mid \beta, \sigma_i^2) \right\} .$$

Make sure you explicitly connect the weights w_i and the idiosyncratic variances σ_i^2.

Quantifying uncertainty: some basic frequentist ideas

In linear regression

In frequentist inference, inferential uncertainty is usually characterized by the sampling distribution, which expresses how one's estimate is

likely to change under repeated sampling. The idea is simple: unstable estimators shouldn't be trusted, and should therefore come with large error bars. This should be a familiar concept, but in case it isn't, consult the tutorial on sampling distributions in this chapter's references.

Suppose, as in the previous section, that we observe data from a linear regression model with Gaussian error:

$$y = X\beta + \epsilon \,, \quad \epsilon \sim \mathrm{N}(0, \sigma^2 I) \,.$$

(A) Derive the sampling distribution of your estimator for β from the previous problem.

(B) This sampling distribution depends on σ^2, yet this is unknown. Suppose that you still wanted to quantify your uncertainty about the individual regression coefficients. Propose a strategy for calculating standard errors for each β_j. Then consult the data set on ozone concentration in Los Angeles, where the goal is to regress daily ozone concentration on a set of other atmospheric variables. This is available from the R package "mlbench," with my R script "ozone.R" giving you a head start on processing things.

Calculate standard errors using your method, and then using the pre-packaged lm function in R. Note: you may have an essentially correct strategy for calculating standard errors that yields something slightly different from the lm function. If so, that's OK—can you explain the discrepancy?

Propagating uncertainty

Suppose you have taken data and estimated some parameters $\theta_1, \ldots, \theta_P$ of a multivariate statistical model—for example, the regression model of the previous problem. Call your estimate $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_P)^T$. Suppose that you also have an estimate of the covariance matrix of the sampling distribution of $\hat{\theta}$:

$$\hat{\Sigma} \approx \mathrm{cov}(\hat{\theta}) = E\left\{(\hat{\theta} - \bar{\theta})(\hat{\theta} - \bar{\theta})^T\right\} \,,$$

where the expectation is under the sampling distribution for the data, given the true parameter θ. Here $\bar{\theta}$ denotes the mean of the sampling distribution.

If you want to report uncertainty about the $\hat{\theta}_j$'s, you can do so by peeling off the diagonal of the estimated covariance matrix: $\hat{\Sigma}_{jj} = \hat{\sigma}_j^2$ is the square of the ordinary standard error of $\hat{\theta}_j$. But what if you want to report uncertainty about some function involving multiple components of the estimate $\hat{\theta}$?

(A) Start with the trivial case where you want to estimate

$$f(\theta) = \theta_1 + \theta_2 \,.$$

Calculate the standard error of $f(\hat{\theta})$, and generalize this to the case where f is the sum of all p components of $\hat{\theta}$.

(B) What now if f is a nonlinear function of the $\hat{\theta}_j$'s? Propose an approximation for $\text{var}\{f(\hat{\theta})\}$, where f is any sufficiently smooth function. (As above, the variance is under the sampling distribution of the data, given the true parameter.)

There are obviously many potential strategies that might work, but here's one you might find fruitful: try a first-order Taylor approximation of $f(\hat{\theta})$ around the unknown true value θ. Try to bound the size of the likely error of the approximation, or at least talk generally about what kinds of assumptions or features of f or $p(\hat{\theta} \mid \theta)$ might be relevant. You should also reflect on some of the potential caveats of this approach.