The multivariate normal model

Up until now all of our statistical models have been univariate models, that is, models for a single measurement on each member of a sample of individuals or each run of a repeated experiment. However, datasets are frequently multivariate, having multiple measurements for each individual or experiment. This chapter covers what is perhaps the most useful model for multivariate data, the multivariate normal model, which allows us to jointly estimate population means, variances and correlations of a collection of variables. After first calculating posterior distributions under semiconjugate prior distributions, we show how the multivariate normal model can be used to impute data that are missing at random.

7.1 The multivariate normal density

Example: Reading comprehension

A sample of twenty-two children are given reading comprehension tests before and after receiving a particular instructional method. Each student $i$ will then have two scores, $Y_{i,1}$ and $Y_{i,2}$ denoting the pre- and post-instructional scores respectively. We denote each student’s pair of scores as a $2 \times 1$ vector $Y_i$, so that

$$Y_i = \begin{pmatrix} Y_{i,1} \\ Y_{i,2} \end{pmatrix} = \begin{pmatrix} \text{score on first test} \\ \text{score on second test} \end{pmatrix}.$$ 

Things we might be interested in include the population mean $\theta$,

$$E[Y] = \begin{pmatrix} E[Y_{i,1}] \\ E[Y_{i,2}] \end{pmatrix} = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$$

and the population covariance matrix $\Sigma$,

$$\Sigma = \text{Cov}[Y] = \begin{pmatrix} E[Y_{1}^2] - E[Y_1]^2 & E[Y_1 Y_2] - E[Y_1] E[Y_2] \\ E[Y_1 Y_2] - E[Y_1] E[Y_2] & E[Y_2^2] - E[Y_2]^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2 \end{pmatrix},$$

where the expectations above represent the unknown population averages. Having information about $\theta$ and $\Sigma$ may help us in assessing the effectiveness of the teaching method, possibly evaluated with $\theta_2 - \theta_1$, or the consistency of the reading comprehension test, which could be evaluated with the correlation coefficient $\rho_{1,2} = \sigma_{1,2}/\sqrt{\sigma_1^2 \sigma_2^2}$.

The multivariate normal density

Notice that $\theta$ and $\Sigma$ are both functions of population moments, or population averages of powers of $Y_1$ and $Y_2$. In particular, $\theta$ and $\Sigma$ are functions of first- and second-order moments:

- first-order moments: $E[Y_1], E[Y_2]$
- second-order moments: $E[Y_1^2], E[Y_1 Y_2], E[Y_2^2]$

Recall from Chapter 5 that a univariate normal model describes a population in terms of its mean and variance ($\theta, \sigma^2$), or equivalently its first two moments ($E[Y] = \theta, E[Y^2] = \sigma^2 + \theta^2$). The analogous model for describing first- and second-order moments of multivariate data is the multivariate normal model. We say a $p$-dimensional data vector $Y$ has a multivariate normal distribution if its sampling density is given by

$$p(y|\theta, \Sigma) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\{- (y - \theta)^T \Sigma^{-1} (y - \theta)/2\}$$

where

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix}, \quad \theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \sigma_{1,p} \\ \sigma_{1,2} & \sigma_2^2 & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1,p} & \cdots & \cdots & \sigma_p^2 \end{pmatrix}.$$
This means that the marginal distributions for $Y_1$ from the three populations in Figure 7.1 are identical (the same holds for $Y_2$). The only thing that differs across the three populations is the relationship between $Y_1$ and $Y_2$, which is controlled by the covariance parameter $\sigma_{1,2}$.

![Multivariate normal samples and densities.](image)

**Fig. 7.1.** Multivariate normal samples and densities.

### 7.2 A semiconjugate prior distribution for the mean

Recall from Chapters 5 and 6 that if $Y_1, \ldots, Y_n$ are independent samples from a univariate normal population, then a convenient conjugate prior distribution for the population mean is also univariate normal. Similarly, a convenient prior distribution for the multivariate mean $\theta$ is a multivariate normal distribution, which we will parameterize as

$$p(\theta) = \text{multivariate normal}(\mu_0, \Lambda_0),$$

where $\mu_0$ and $\Lambda_0$ are the prior mean and variance of $\theta$, respectively. What is the full conditional distribution of $\theta$, given $y_1, \ldots, y_n$ and $\Sigma$? In the univariate case, having normal prior and sampling distributions resulted in a normal full conditional distribution for the population mean. Let’s see if this result holds for the multivariate case. We begin by examining the prior distribution as a function of $\theta$:

$$p(\theta) = (2\pi)^{-p/2}|\Lambda_0|^{-1/2} \exp\left\{-\frac{1}{2}(\theta - \mu_0)^T \Lambda_0^{-1}(\theta - \mu_0)\right\}$$

$$\propto \exp\left\{-\frac{1}{2} \theta^T \Lambda_0^{-1} \theta + \theta^T A_0^{-1} \mu_0 - \frac{1}{2} \mu_0^T A_0^{-1} \mu_0 \right\}$$

$$= \exp\left\{-\frac{1}{2} \theta^T A_1 \theta + \theta^T b_1 \right\}; \quad (7.1)$$
where \( A_0 = \Lambda_0^{-1} \) and \( b_0 = \Lambda_0^{-1} \mu_0 \). Conversely, Equation 7.1 says that if a random vector \( \theta \) has a density on \( \mathbb{R}^p \) that is proportional to \( \exp\{-\theta^T A \theta / 2 + \theta^T b\} \) for some matrix \( A \) and vector \( b \), then \( \theta \) must have a multivariate normal distribution with covariance \( A^{-1} \) and mean \( A^{-1} b \).

If our sampling model is that \( \{Y_1, \ldots, Y_n | \theta, \Sigma\} \) are i.i.d. multivariate normal(\( \theta, \Sigma \)), then similar calculations show that the joint sampling density of the observed vectors \( y_1, \ldots, y_n \) is

\[
p(y_1, \ldots, y_n | \theta, \Sigma) = \prod_{i=1}^{n} \left(2\pi\right)^{-p/2} |\Sigma|^{-1/2} \exp\{-\frac{1}{2} \sum_{i=1}^{n} (y_i - \theta)^T \Sigma^{-1} (y_i - \theta)\}
\]

\[
\propto \exp\left\{-\frac{1}{2} \theta^T A_1 \theta + \theta^T b_1\right\}, \tag{7.2}
\]

where \( A_1 = n \Sigma^{-1} \), \( b_1 = n \Sigma^{-1} \bar{y} \) and \( \bar{y} \) is the vector of variable-specific averages \( y = (\frac{1}{n} \sum_{i=1}^{n} y_{i,1}, \ldots, \frac{1}{n} \sum_{i=1}^{n} y_{i,p})^T \). Combining Equations 7.1 and 7.2 gives

\[
p(\theta | y_1, \ldots, y_n, \Sigma) \propto \exp\left\{-\frac{1}{2} \theta^T A_0 \theta + \theta^T b_0\right\} \times \exp\left\{-\frac{1}{2} \theta^T A_1 \theta + \theta^T b_1\right\}
\]

\[
\propto \exp\left\{-\frac{1}{2} \theta^T A_n \theta + \theta^T b_n\right\}, \quad \text{where}
\]

\[
A_n = A_0 + A_1 = \Lambda_0^{-1} + n \Sigma^{-1} \quad \text{and} \quad b_n = b_0 + b_1 = \Lambda_0^{-1} \mu_0 + n \Sigma^{-1} \bar{y}. \tag{7.3}
\]

From the comments in the previous paragraph, Equation 7.3 implies that the conditional distribution of \( \theta \) therefore must be a multivariate normal distribution with covariance \( A_n^{-1} \) and mean \( A_n^{-1} b_n \), so

\[
\text{Cov}[\theta | y_1, \ldots, y_n, \Sigma] = A_n = (\Lambda_0^{-1} + n \Sigma^{-1})^{-1} \tag{7.4}
\]

\[
\text{E}[\theta | y_1, \ldots, y_n, \Sigma] = \mu_n = (\Lambda_0^{-1} + n \Sigma^{-1})^{-1} (\Lambda_0^{-1} \mu_0 + n \Sigma^{-1} \bar{y}) \tag{7.5}
\]

\[
p(\theta | y_1, \ldots, y_n, \Sigma) = \text{multivariate normal}(\mu_n, A_n). \tag{7.6}
\]

It looks a bit complicated, but can be made more understandable by analogy with the univariate normal case: Equation 7.4 says that posterior precision, or inverse variance, is the sum of the prior precision and the data precision, just as in the univariate normal case. Similarly, Equation 7.5 says that the posterior expectation is a weighted average of the prior expectation and the sample mean. Notice that, since the sample mean is consistent for the population mean, the posterior mean also will be consistent for the population mean even if the true distribution of the data is not multivariate normal.
7.3 The inverse-Wishart distribution

Just as a variance $\sigma^2$ must be positive, a variance-covariance matrix $\Sigma$ must be positive definite, meaning that $x' \Sigma x > 0$ for all vectors $x$.

Positive definiteness guarantees that $\sigma^2_j > 0$ for all $j$ and that all correlations are between -1 and 1. Another requirement of our covariance matrix is that it is symmetric, which means that $\sigma_{j,k} = \sigma_{k,j}$. Any valid prior distribution for $\Sigma$ must put all of its probability mass on this complicated set of symmetric, positive definite matrices. How can we formulate such a prior distribution?

Empirical covariance matrices

The sum of squares matrix of a collection of multivariate vectors $z_1, \ldots, z_n$ is given by

$$
\sum_{i=1}^{n} z_i z_i^T = Z^T Z,
$$

where $Z$ is the $n \times p$ matrix whose $i$th row is $z_i^T$. Recall from matrix algebra that since $z_i$ can be thought of as a $p \times 1$ matrix, $z_i z_i^T$ is the following $p \times p$ matrix:

$$
z_i z_i^T = \begin{pmatrix}
z_{i,1}^2 & z_{i,1} z_{i,2} & \cdots & z_{i,1} z_{i,p} \\
z_{i,2} z_{i,1} & z_{i,2}^2 & \cdots & z_{i,2} z_{i,p} \\
\vdots & \vdots & \ddots & \vdots \\
z_{i,p} z_{i,1} & z_{i,p} z_{i,2} & \cdots & z_{i,p}^2
\end{pmatrix}.
$$

If the $z_i$’s are samples from a population with zero mean, we can think of the matrix $z_i z_i^T / n$ as the contribution of vector $z_i$ to the estimate of the covariance matrix of all of the observations. In this mean-zero case, if we divide $Z^T Z$ by $n$, we get a sample covariance matrix, an unbiased estimator of the population covariance matrix:

$$
\frac{1}{n} [Z^T Z]_{j,j} = \frac{1}{n} \sum_{i=1}^{n} z_{i,j}^2 = s_{j,j} = s_j^2 \\
\frac{1}{n} [Z^T Z]_{j,k} = \frac{1}{n} \sum_{i=1}^{n} z_{i,j} z_{i,k} = s_{j,k}.
$$

If $n > p$ and the $z_i$’s are linearly independent, then $Z^T Z$ will be positive definite and symmetric. This suggests the following construction of a “random” covariance matrix: For a given positive integer $\nu_0$ and a $p \times p$ covariance matrix $\Phi_0$,

1. sample $z_1, \ldots, z_{\nu_0}$ i.i.d. multivariate normal($0, \Phi_0$);
2. calculate $Z_{\nu_0}^T Z = \sum_{i=1}^{\nu_0} z_i z_i^T$.

We can repeat this procedure over and over again, generating matrices $Z_1^T Z_1, \ldots, Z_{\nu_s}^T Z_{\nu_s}$. The population distribution of these sum of squares matrices is called a Wishart distribution with parameters $(\nu_0, \Phi_0)$, which has the following properties:
• If $\nu > p$, then $Z^T Z$ is positive definite with probability 1.
• $Z^T Z$ is symmetric with probability 1.
• $E[Z^T Z] = \nu \Phi_0$.

The Wishart distribution is a multivariate analogue of the gamma distribution (recall that if $z$ is a mean-zero univariate normal random variable, then $z^2$ is a gamma random variable). In the univariate normal model, our prior distribution for the precision $1/\sigma^2$ is a gamma distribution, and our full conditional distribution for the variance is an inverse-gamma distribution. Similarly, it turns out that the Wishart distribution is a semi-conjugate prior distribution for the precision matrix $\Sigma^{-1}$, and so the inverse-Wishart distribution is our semi-conjugate prior distribution for the covariance matrix $\Sigma$. With a slight reparameterization, to sample a covariance matrix $\Sigma$ from an inverse-Wishart distribution we perform the following steps:

1. sample $z_1, \ldots, z_{\nu_0} \sim \text{i.i.d. multivariate normal}(0, S_0^{-1})$;
2. calculate $Z^T Z = \sum_{i=1}^{\nu_0} z_i z_i^T$;
3. set $\Sigma = (Z^T Z)^{-1}$.

Under this simulation scheme, the precision matrix $\Sigma^{-1}$ has a Wishart($\nu_0, S_0^{-1}$) distribution and the covariance matrix $\Sigma$ has an inverse-Wishart($\nu_0, S_0^{-1}$) distribution. The expectations of $\Sigma^{-1}$ and $\Sigma$ are

$$E[\Sigma^{-1}] = \nu_0 S_0^{-1}$$
$$E[\Sigma] = \frac{1}{\nu_0 - p - 1} (S_0^{-1})^{-1} = \frac{1}{\nu_0 - p - 1} S_0.$$

If we are confident that the true covariance matrix is near some covariance matrix $\Sigma_0$, then we might choose $\nu_0$ to be large and set $S_0 = (\nu_0 - p - 1)\Sigma_0$, making the distribution of $\Sigma$ concentrated around $\Sigma_0$. On the other hand, choosing $\nu_0 = p + 2$ and $S_0 = \Sigma_0$ makes $\Sigma$ only loosely centered around $\Sigma_0$.

**Full conditional distribution of the covariance matrix**

The inverse-Wishart($\nu_0, S_0^{-1}$) density is given by

$$p(\Sigma) = \left[ 2^{\nu_0 p/2} \pi^{(p^2)/2} |S_0|^{-\nu_0/2} \prod_{j=1}^{p} \Gamma([\nu_0 + 1 - j]/2) \right]^{-1} \times$$

$$|\Sigma|^{-(\nu_0 + p + 1)/2} \times \exp\{-\text{tr}(S_0 \Sigma^{-1})/2\}. \quad (7.7)$$

The normalizing constant is quite intimidating. Fortunately we will only have to work with the second line of the equation. The expression “tr” stands for trace and for a square $p \times p$ matrix $A$, tr($A$) = $\sum_{j=1}^{p} a_{j,j}$, the sum of the diagonal elements.

We now need to combine the above prior distribution with the sampling distribution for $Y_1, \ldots, Y_n$:
The inverse-Wishart distribution

\[ p(y_1, \ldots, y_n | \theta, \Sigma) = (2\pi)^{-np/2} |\Sigma|^{-n/2} \exp\left\{ -\sum_{i=1}^{n} (y_i - \theta)^T \Sigma^{-1} (y_i - \theta)/2 \right\}. \]

(7.8)

An interesting result from matrix algebra is that the sum \( \sum_{k=1}^{K} b_k^T A b_k = \text{tr}(B^T BA) \), where \( B \) is the matrix whose \( k \)th row is \( b_k^T \). This means that the term in the exponent of Equation 7.8 can be expressed as

\[ \sum_{i=1}^{n} (y_i - \theta)^T \Sigma^{-1} (y_i - \theta) = \text{tr}(S_\theta \Sigma^{-1}), \]

where \( S_\theta = \sum_{i=1}^{n} (y_i - \theta)(y_i - \theta)^T \).

The matrix \( S_\theta \) is the residual sum of squares matrix for the vectors \( y_1, \ldots, y_n \) if the population mean is presumed to be \( \theta \). Conditional on \( \theta \), \( \frac{1}{n} S_\theta \) provides an unbiased estimate of the true covariance matrix \( \text{Cov}[Y] \) (more generally, when \( \theta \) is not conditioned on the sample covariance matrix is \( \sum (y_i - \bar{y})(y_i - \bar{y})^T/(n - 1) \) and is an unbiased estimate of \( \Sigma \)). Using the above result to combine Equations 7.7 and 7.8 gives the conditional distribution of \( \Sigma \):

\[ p(\Sigma | y_1, \ldots, y_n, \theta) \propto p(\Sigma) \times p(y_1, \ldots, y_n | \theta, \Sigma) \]
\[ \propto \left( |\Sigma|^{-(\nu_0+p+1)/2} \exp\{-\text{tr}(S_0 \Sigma^{-1})/2\} \right) \times \left( |\Sigma|^{-n/2} \exp\{-\text{tr}(S_\theta \Sigma^{-1})/2\} \right) \]
\[ = |\Sigma|^{-(\nu_0+n+p+1)/2} \exp\{-\text{tr}([S_0 + S_\theta] \Sigma^{-1})/2\}. \]

Thus we have

\[ \{ \Sigma | y_1, \ldots, y_n, \theta \} \sim \text{inverse-Wishart}(\nu_0 + n, [S_0 + S_\theta]^{-1}). \]

(7.9)

Hopefully this result seems somewhat intuitive: We can think of \( \nu_0 + n \) as the “posterior sample size,” being the sum of the “prior sample size” \( \nu_0 \) and the data sample size. Similarly, \( S_0 + S_\theta \) can be thought of as the “prior” residual sum of squares plus the residual sum of squares from the data. Additionally, the conditional expectation of the population covariance matrix is

\[ \mathbb{E}[\Sigma | y_1, \ldots, y_n, \theta] = \frac{1}{\nu_0 + n - p - 1} (S_0 + S_\theta) \]
\[ = \frac{\nu_0 - p - 1}{\nu_0 + n - p - 1} \frac{1}{\nu_0 - p - 1} S_0 + \frac{n}{\nu_0 + n - p - 1} \frac{1}{n} S_\theta \]

and so the conditional expectation can be seen as a weighted average of the prior expectation and the unbiased estimator. Because it can be shown that \( S_\theta \) converges to the true population covariance matrix, the posterior expectation of \( \Sigma \) is a consistent estimator of the population covariance, even if the true population distribution is not multivariate normal.
7.4 Gibbs sampling of the mean and covariance

In the last two sections we showed that

\[
\{\theta | y_1, \ldots, y_n, \Sigma\} \sim \text{multivariate normal}(\mu_n, \Lambda_n)
\]

\[
\{\Sigma | y_1, \ldots, y_n, \theta\} \sim \text{inverse-Wishart}(\nu_n, S_n^{-1}),
\]

where \{\Lambda_n, \mu_n\} are defined in Equations 7.4 and 7.5, \(\nu_n = \nu_0 + n\) and \(S_n = S_0 + \Sigma\). These full conditional distributions can be used to construct a Gibbs sampler, providing us with an MCMC approximation to the joint posterior distribution \(p(\theta, \Sigma | y_1, \ldots, y_n)\). Given a starting value \(\Sigma^{(0)}\), the Gibbs sampler generates \(\{\theta^{(s+1)}, \Sigma^{(s+1)}\}\) from \(\{\theta^{(s)}, \Sigma^{(s)}\}\) via the following two steps:

1. Sample \(\theta^{(s+1)}\) from its full conditional distribution:
   a) compute \(\mu_n\) and \(\Lambda_n\) from \(y_1, \ldots, y_n\) and \(\Sigma^{(s)}\);
   b) sample \(\theta^{(s+1)} \sim \text{multivariate normal}(\mu_n, \Lambda_n)\).

2. Sample \(\Sigma^{(s+1)}\) from its full conditional distribution:
   a) compute \(S_n\) from \(y_1, \ldots, y_n\) and \(\theta^{(s+1)}\);
   b) sample \(\Sigma^{(s+1)} \sim \text{inverse-Wishart}(\nu_0 + n, S_n^{-1})\).

Steps 1.a and 2.a highlight the fact that \{\mu_n, \Lambda_n\} depend on the value of \(\Sigma\), and that \(S_n\) depends on the value of \(\theta\), and so these quantities need to be recalculated at every iteration of the sampler.

Example: Reading comprehension

Let’s return to the example from the beginning of the chapter in which each of 22 children were given two reading comprehension exams, one before a certain type of instruction and one after. We’ll model these 22 pairs of scores as i.i.d. samples from a multivariate normal distribution. The exam was designed to give average scores of around 50 out of 100, so \(\mu_0 = (50, 50)^T\) would be a good choice for our prior expectation. Since the true mean cannot be below 0 or above 100, it is desirable to use a prior variance for \(\theta\) that puts little probability outside of this range. We’ll take the prior variances on \(\theta_1\) and \(\theta_2\) to be \(\lambda_{0,1}^2 = \lambda_{0,2}^2 = (50/2)^2 = 625\), so that the prior probability \(\Pr(\theta_j \notin [0, 100])\) is only 0.05. Finally, since the two exams are measuring similar things, whatever the true values of \(\theta_1\) and \(\theta_2\) are it is probable that they are close. We can reflect this with a prior correlation of 0.5, so that \(\lambda_{1,2} = 312.5\). As for the prior distribution on \(\Sigma\), some of the same logic about the range of exam scores applies. We’ll take \(S_0\) to be the same as \(\Lambda_0\), but only loosely center \(\Sigma\) around this value by taking \(\nu_0 = p + 2 = 4\).

\[
\text{mu0<-c}(50,50) \\
\text{L0<-matrix(c(625,312.5,312.5,625),nrow=2,ncol=2)}
\]

\[
\text{nu0<-4} \\
\text{S0<-matrix(c(625,312.5,312.5,625),nrow=2,ncol=2)}
\]
The observed values $y_1, \ldots, y_{22}$ are plotted as dots in the second panel of Figure 7.2. The sample mean is $\bar{y} = (47.18, 53.86)^T$, the sample variances are $s_1^2 = 182.16$ and $s_2^2 = 243.65$, and the sample correlation is $s_{1,2}/(s_1 s_2) = 0.70$. Let’s use the Gibbs sampler described above to combine this sample information with our prior distributions to obtain estimates and confidence intervals for the population parameters. We begin by setting $\Sigma^{(0)}$ equal to the sample covariance matrix, and iterating from there. In the R-code below, $Y$ is the $22 \times 2$ data matrix of the observed values.

```r
data(chapter7) ; Y<-Y.reading
n<-dim(Y)[1] ; ybar<-apply(Y,2,mean)
Sigma<-cov(Y) ; THETA<-SIGMA<-NULL

set.seed(1)
for(s in 1:5000)
{
  ###update theta
  Ln<-solve( solve(L0) + n*solve(Sigma) )
  mum<-Ln%*%( solve(L0)%*%mu0 + n*solve(Sigma)%*%ybar )
  theta<-rmvnorm(1,mun,Ln)
  ###update Sigma
  Sn<- S0 + ( t(Y)-c(theta) )%*%t( t(Y)-c(theta) )
  Sigma<-solve( rwish(1, nu0+n, solve(Sn)) )
  ### save results
  THETA<-rbind(THETA,theta) ; SIGMA<-rbind(SIGMA,c(Sigma))
}
```

The above code generates 5,000 values ($\{\boldsymbol{\theta}^{(1)}, \Sigma^{(1)}\}, \ldots, \{\boldsymbol{\theta}^{(5000)}, \Sigma^{(5000)}\}$) whose empirical distribution approximates $p(\boldsymbol{\theta}, \Sigma | y_1, \ldots, y_n)$. It is left as an exercise to assess the convergence and autocorrelation of this Markov chain. From these samples we can approximate posterior probabilities and confidence regions of interest.

```r
> quantile( THETA[,2] - THETA[,1] , prob=c(.025,.5,.975) )
2.5%  97.5% 6.668097 11.794824

> mean( THETA[,2] > THETA[,1] )
[1] 0.9942
```

The posterior probability $\Pr(\theta_2 > \theta_1 | y_1, \ldots, y_n) = 0.99$ indicates strong evidence that, if we were to give exams and instruction to a large population
of children, then the average score on the second exam would be higher than that on the first. This evidence is displayed graphically in the first panel of Figure 7.2, which shows 97.5%, 75%, 50%, 25% and 2.5% highest posterior density contours for the joint posterior distribution of $\theta = (\theta_1, \theta_2)^T$. A highest posterior density contour is a two-dimensional analogue of a confidence interval. The contours for the posterior distribution of $\theta$ are all mostly above the 45-degree line $\theta_1 = \theta_2$.

![Fig. 7.2. Reading comprehension data and posterior distributions](image)

Now let’s ask a slightly different question - what is the probability that a randomly selected child will score higher on the second exam than on the first? The answer to this question is a function of the posterior predictive distribution of a new sample $(Y_1, Y_2)^T$, given the observed values. The second panel of Figure 7.2 shows highest posterior density contours of the posterior predictive distribution, which, while mostly being above the line $y_2 = y_1$, still has substantial overlap with the region below this line, and in fact $\Pr(Y_2 > Y_1 | y_1, \ldots, y_n) = 0.71$. How should we evaluate the effectiveness of the between-exam instruction? On one hand, the fact that $\Pr(\theta_2 > \theta_1 | y_1, \ldots, y_n) = 0.99$ seems to suggest that there is a “highly significant difference” in exam scores before and after the instruction, yet $\Pr(Y_2 > Y_1 | y_1, \ldots, y_n) = 0.71$ says that almost a third of the students will get a lower score on the second exam. The difference between these two probabilities is that the first is measuring the evidence that $\theta_2$ is larger than $\theta_1$ without regard to whether or not the magnitude of the difference $\theta_2 - \theta_1$ is large compared to the sampling variability of the data. Confusion over these two different ways of comparing populations is common in the reporting of results from experiments or surveys: studies with very large values of $n$ often result in values of $\Pr(\theta_2 > \theta_1 | y_1, \ldots, y_n)$ that are very close to 1 (or $p$-values
that are very close to zero), suggesting a “significant effect,” even though such results say nothing about how large of an effect we expect to see for a randomly sampled individual.

7.5 Missing data and imputation

Figure 7.3 displays univariate histograms and bivariate scatterplots for four variables taken from a dataset involving health-related measurements on 200 women of Pima Indian heritage living near Phoenix, Arizona (Smith et al., 1988). The four variables are glu (blood plasma glucose concentration), bp (diastolic blood pressure), skin (skin fold thickness) and bmi (body mass index). The first ten subjects in this dataset have the following entries:
The NA’s stand for “not available,” and so some data for some individuals are “missing.” Missing data are fairly common in survey data: Sometimes people accidentally miss a page of a survey, sometimes a doctor forgets to write down a piece of medical data, sometimes the response is unreadable, and so on. Many surveys (such as the General Social Survey) have multiple versions with certain questions appearing in only a subset of the versions. As a result, all the subjects may have missing data.

In such situations it is not immediately clear how to do parameter estimation. The posterior distribution for $\boldsymbol{\theta}$ and $\Sigma$ depends on $\prod_{i=1}^{n} p(y_i|\boldsymbol{\theta},\Sigma)$, but $p(y_i|\boldsymbol{\theta},\Sigma)$ cannot be computed if components of $y_i$ are missing. What can we do? Unfortunately, many software packages either throw away all subjects with incomplete data, or impute missing values with a population mean or some other fixed value, then proceed with the analysis. The first approach is bad because we are throwing away a potentially large amount of useful information. The second is statistically incorrect, as it says we are certain about the values of the missing data when in fact we have not observed them.

Let’s carefully think about the information that is available from subjects with missing data. Let $O_i = (O_{i,1}, \ldots, O_{i,p})^T$ be a binary vector of zeros and ones such that $O_{i,j} = 1$ implies that $Y_{i,j}$ is observed and not missing, whereas $O_{i,j} = 0$ implies $Y_{i,j}$ is missing. Our observed information about subject $i$ is therefore $O_i = o_i$ and $Y_{i,j} = y_{i,j}$ for variables $j$ such that $o_{i,j} = 1$. For now, we’ll assume that missing data are missing at random, meaning that $O_i$ and $Y_i$ are statistically independent and that the distribution of $O_i$ does not depend on $\boldsymbol{\theta}$ or $\Sigma$. In cases where the data are missing but not at random, then sometimes inference can be made by modeling the relationship between $O_i$, $Y_i$ and the parameters (see Chapter 21 of Gelman et al (2004)).

In the case where data are missing at random, the sampling probability for the data from subject $i$ is

\[
p(\omega_i, \{y_{i,j} : o_{i,j} = 1\}|\boldsymbol{\theta},\Sigma) = p(\omega_i) \times p(\{y_{i,j} : o_{i,j} = 1\}|\boldsymbol{\theta},\Sigma)
\]

\[
= p(\omega_i) \times \int \left\{ p(y_{i,1}, \ldots, y_{i,p}|\boldsymbol{\theta},\Sigma) \prod_{y_{i,j}:o_{i,j}=0} dy_{i,j} \right\}.
\]
In words, our sampling probability for data from subject \( i \) is \( p(o_i) \) multiplied by the marginal probability of the observed variables, after integrating out the missing variables. To make this more concrete, suppose \( y_i = (y_{i,1}, \text{NA}, y_{i,3}, \text{NA})^T \), so \( o_i = (1, 0, 1, 0)^T \). Then

\[
p(o_i, y_{i,1}, y_{i,3}|\theta, \Sigma) = p(o_i) \times p(y_{i,1}, y_{i,3}|\theta, \Sigma) = p(o_i) \times \int p(y_j|\theta, \Sigma) \, dy_2 \, dy_4.
\]

So the correct thing to do when data are missing at random is to integrate over the missing data to obtain the marginal probability of the observed data. In this particular case of the multivariate normal model, this marginal probability is easily obtained: \( p(y_{i,1}, y_{i,3}|\theta, \Sigma) \) is simply a bivariate normal density with mean \( (\theta_1, \theta_3)^T \) and a covariance matrix made up of \( (\sigma_1^2, \sigma_{1,3}, \sigma_3^2) \). But combining marginal densities from subjects having different amounts of information can be notationally awkward. Fortunately, our integration can alternatively be done quite easily using Gibbs sampling.

**Gibbs sampling with missing data**

In Bayesian inference we use probability distributions to describe our information about unknown quantities. What are the unknown quantities for our multivariate normal model with missing data? The parameters \( \theta \) and \( \Sigma \) are unknown as usual, but the missing data are also an unknown but key component of our model. Treating it as such allows us to use Gibbs sampling to make inference on \( \theta, \Sigma \), as well as to make predictions for the missing values.

Let \( Y \) be the \( n \times p \) matrix of all the potential data, observed and unobserved, and let \( O \) be the \( n \times p \) matrix in which \( o_{i,j} = 1 \) if \( Y_{i,j} \) is observed and \( o_{i,j} = 0 \) if \( Y_{i,j} \) is missing. The matrix \( Y \) can then be thought of as consisting of two parts:

- \( Y_{\text{obs}} = \{ y_{i,j} : o_{i,j} = 1 \} \), the data that we do observe, and
- \( Y_{\text{miss}} = \{ y_{i,j} : o_{i,j} = 0 \} \), the data that we do not observe.

From our observed data we want to obtain \( p(\theta, \Sigma, Y_{\text{miss}}|Y_{\text{obs}}) \), the posterior distribution of unknown and unobserved quantities. A Gibbs sampling scheme for approximating this posterior distribution can be constructed by simply adding one step to the Gibbs sampler presented in the previous section: Given starting values \( \{ \Sigma^{(0)}, Y_{\text{miss}}^{(0)} \} \), we generate \( \{ \theta^{(s+1)}, \Sigma^{(s+1)}, Y_{\text{miss}}^{(s+1)} \} \) from \( \{ \theta^{(s)}, \Sigma^{(s)}, Y_{\text{miss}}^{(s)} \} \) by

1. sampling \( \theta^{(s+1)} \) from \( p(\theta|Y_{\text{obs}}, Y_{\text{miss}}^{(s)}, \Sigma^{(s)}) \);
2. sampling \( \Sigma^{(s+1)} \) from \( p(\Sigma|Y_{\text{obs}}, Y_{\text{miss}}^{(s)}, \theta^{(s+1)}) \);
3. sampling \( Y_{\text{miss}}^{(s+1)} \) from \( p(Y_{\text{miss}}|Y_{\text{obs}}, \theta^{(s+1)}, \Sigma^{(s+1)}) \).

Note that in steps 1 and 2, the fixed value of \( Y_{\text{obs}} \) combines with the current value of \( Y_{\text{miss}}^{(s)} \) to form a current version of a complete data matrix \( Y^{(s)} \) having
no missing values. The $n$ rows of the matrix of $Y^{(s)}$ can then be plugged into formulae 7.6 and 7.9 to obtain the full conditional distributions of $\theta$ and $\Sigma$. Step 3 is a bit more complicated:

$$p(Y_{\text{miss}} | Y_{\text{obs}}, \theta, \Sigma) \propto p(Y_{\text{miss}}, Y_{\text{obs}} | \theta, \Sigma)$$

$$= \prod_{i=1}^{n} p(y_{i,\text{miss}}, y_{i,\text{obs}} | \theta, \Sigma)$$

$$\propto \prod_{i=1}^{n} p(y_{i,\text{miss}} | y_{i,\text{obs}}, \theta, \Sigma),$$

so for each $i$ we need to sample the missing elements of the data vector conditional on the observed elements. This is made possible via the following result about multivariate normal distributions: Let $y \sim \text{multivariate normal}(\theta, \Sigma)$, let $a$ be a subset of variable indices $\{1, \ldots, p\}$ and let $b$ be the complement of $a$. For example, if $p = 4$ then perhaps $a = \{1, 2\}$ and $b = \{3, 4\}$. If you know about inverses of partitioned matrices you can show that

$$\{y_{[b]} | y_{[a]}, \theta, \Sigma\} \sim \text{multivariate normal}(\theta_{b|a}, \Sigma_{b|a}),$$

where

$$\theta_{b|a} = \theta_{[b]} + \Sigma_{[b,a]}(\Sigma_{[a,a]})^{-1}(y_{[a]} - \theta_{[a]})$$

(7.10)

$$\Sigma_{b|a} = \Sigma_{[b,b]} - \Sigma_{[b,a]}(\Sigma_{[a,a]})^{-1}\Sigma_{[a,b]}.$$  

(7.11)

In the above formulae, $\theta_{[b]}$ refers to the elements of $\theta$ corresponding to the indices in $b$, and $\Sigma_{[a,b]}$ refers to the matrix made up of the elements that are in rows $a$ and columns $b$ of $\Sigma$.

Let’s try to gain a little bit of intuition about what is going on in Equations 7.10 and 7.11. Suppose $y$ is a sample from our population of four variables glu, bp, skin and bmi. If we have glu and bp data for someone ($a = \{1, 2\}$) but are missing skin and bmi measurements ($b = \{3, 4\}$), then we would be interested in the conditional distribution of these missing measurements $y_{[b]}$ given the observed information $y_{[a]}$. Equation 7.10 says that the conditional mean of skin and bmi start off at their unconditional mean $\theta_{[b]}$, but then are modified by $(y_{[a]} - \theta_{[a]})$. For example, if a person had higher than average values of glu and bp, then $(y_{[a]} - \theta_{[a]})$ would be a $2 \times 1$ vector of positive numbers. For our data the $2 \times 2$ matrix $\Sigma_{[b,a]}(\Sigma_{[a,a]})^{-1}$ has all positive entries, and so $\theta_{b|a} > \theta_{[b]}$. This makes sense: If all four variables are positively correlated, then if we observe higher than average values of glu and bp, we should also expect higher than average values of skin and bmi. Also note that $\Sigma_{b|a}$ is equal to the unconditional variance $\Sigma_{[b,b]}$ but with something subtracted off, suggesting that the conditional variance is less than the unconditional variance. Again, this makes sense: having information about some variables should decrease, or at least not increase, our uncertainty about the others.

The R code below implements the Gibbs sampling scheme for missing data described in steps 1, 2 and 3 above:
```r
data(chapter7); Y<-Y.pima.miss
### prior parameters
n<-dim(Y)[1]; p<-dim(Y)[2]
mu0<-c(120,64,26,26)
sd0<-(mu0/2)
L0<-matrix(1,p,p); diag(L0)<-1; L0<-L0*outer(sd0,sd0)
u0<-p+2; S0<-L0
###
### starting values
Sigma<-S0
Y.full<-Y
O<-!is.na(Y)
for(j in 1:p)
{
  Y.full[is.na(Y.full[,j]),j]<-mean(Y.full[,j],na.rm=TRUE)
}
###
### Gibbs sampler
THETA<-SIGMA<-Y.MISS<-NULL
set.seed(1)
for(s in 1:1000)
{
  ###update theta
  ybar<-apply(Y.full,2,mean)
  Ln<-solve(solve(L0)+n*solve(Sigma))
  mnu<-Ln%*%(solve(L0)%*%mu0+n*solve(Sigma)%*%ybar)
  theta<-rmvnorm(1,mnu,Ln)
  ###
  ###update Sigma
  Sn<-S0+(t(Y.full)-c(theta))%*%t(t(Y.full)-c(theta))
  Sigma<-solve(rwish(1,u0+n,solve(Sn)))
  ###
  ###update missing data
  for(i in 1:n)
  {
    b <- ( O[i,]==0 )
    a <- ( O[i,]==1 )
    iSa<-solve(Sigma[a,a])
    beta.j <- Sigma[b,a]%*%iSa
    Sigma.j <- Sigma[b,b] - Sigma[b,a]%*%iSa%*%Sigma[a,b]
    theta.j<-theta[b] + beta.j%*(t(Y.full[i,a]) - theta[a])
    Y.full[i,b]<-rmvnorm(1,theta.j,Sigma.j)
  }
```
The prior mean of $\mu_0 = (120, 64, 26, 26)^T$ was obtained from national averages, and the prior variances were based primarily on keeping most of the prior mass on values that are above zero. These prior distributions are likely much more diffuse than more informed prior distributions that could be provided by someone who is familiar with this population or these variables.

The Monte Carlo approximation of $E[\theta|y_1, \ldots, y_n]$ is $(123.46, 71.03, 29.35, 32.18)$, obtained by averaging the 1,000 $\theta$-values generated by the Gibbs sampler. Posterior confidence intervals and other quantities can additionally be obtained in the usual way from the Gibbs samples. We can also average the 1,000 values of $\Sigma$ to obtain $E[\Sigma|y_1, \ldots, y_n]$, the posterior expectation of $\Sigma$. However, when looking at associations among a set of variables, it is often the correlations that are of interest and not the covariances. To each covariance matrix $\Sigma$ there corresponds a correlation matrix $C$, given by

$$ C = \left\{ c_{j,k} : c_{j,k} = \frac{\Sigma[j,k]}{\sqrt{\Sigma[j,j] \Sigma[k,k]}} \right\}. $$

We can convert our 1,000 posterior samples of $\Sigma$ into 1,000 posterior samples of $C$ using the following R-code:

```r
COR <- array(dim=c(p, p, 1000))
for (s in 1:1000)
{
  Sig <- matrix(SIGMA[s,], nrow=p, ncol=p)
  COR[, , s] <- Sig/sqrt(outer(diag(Sig), diag(Sig))
}
```

This code generates a $4 \times 4 \times 1000$ array, where each “slice” is a $4 \times 4$ correlation matrix generated from the posterior distribution. The posterior expectation of $C$ is

$$ E[C|y_1, \ldots, y_n] = \begin{pmatrix}
1.00 & 0.23 & 0.25 & 0.19 \\
0.23 & 1.00 & 0.25 & 0.24 \\
0.25 & 0.25 & 1.00 & 0.65 \\
0.19 & 0.24 & 0.65 & 1.00
\end{pmatrix} $$

and marginal posterior 95% quantile-based confidence intervals can be obtained with the command `apply(COR, c(1,2), quantile, prob=c(.025,.975))`.

These are displayed graphically in the left panel of Figure 7.4.

**Prediction and regression**

Multivariate models are often used to predict one or more variables given the others. Consider, for example, a predictive model of $glu$ based on measurements of $bp$, $skin$ and $bmi$. Using $a = \{2, 3, 4\}$ and $b = \{1\}$ in Equation 7.10,
the conditional mean of $y_{[b]} = \text{glu}$, given numerical values of $y_{[a]} = \{\text{bp, skin, bmi}\}$, is given by

$$E[y_{[b]} | \theta, \Sigma, y_{[a]}] = \theta_{[b]} + \beta_{[b|a]}^T (y_{[a]} - \theta_{[a]})$$

where $\beta_{[b|a]}^T = \Sigma_{[b,a]} (\Sigma_{[a,a]})^{-1}$. Since this takes the form of a linear regression model, we call the value of $\beta_{[b|a]}$ the regression coefficient for $y_{[b]}$ given $y_{[a]}$ based on $\Sigma$. Values of $\beta_{[b|a]}$ can be computed for each posterior sample of $\Sigma$, allowing us to obtain posterior expectations and confidence intervals for these regression coefficients. Quantile-based 95% confidence intervals for each of $\{\beta_{1|234}, \beta_{2|134}, \beta_{3|124}, \beta_{4|123}\}$ are shown graphically in the second column of Figure 7.4. The regression coefficients often tell a different story than the correlations: The bottom row of plots, for example, shows that while there

Fig. 7.4. Ninety-five percent posterior confidence intervals for correlations (left) and regression coefficients (right).
is strong evidence that the correlations between \textit{bmi} and each of the other variables are all positive, the plots on the right-hand side suggest that \textit{bmi} is nearly conditionally independent of \textit{glu} and \textit{bp} given \textit{skin}.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figs/missing_data.png}
\caption{True values of the missing data versus their posterior expectations.}
\end{figure}

\textit{Out-of-sample validation}

Actually, the dataset we just analyzed was created by taking a complete data matrix with no missing values and randomly replacing 10\% of the entries with \texttt{NA}'s. Since the original dataset is available, we can compare values predicted by the model to the actual sample values. This comparison is made graphically in Figure 7.5, which plots the true value of $y_{i,j}$ against its posterior mean for each $\{i,j\}$ such that $o_{i,j} = 0$. It looks like we are able to do a better job
predicting missing values of skin and bmi than the other two variables. This makes sense, as these two variables have the highest correlation. If skin is missing, we can make a good prediction for it based on the observed value of bmi, and vice-versa. Such a procedure, where we evaluate how well a model does at predicting data that were not used to estimate the parameters, is called out-of-sample validation, and is often used to quantify the predictive performance of a model.

7.6 Discussion and further references

The multivariate normal model can be justified as a sampling model for reasons analogous to those for the univariate normal model (see Section 5.7): It is characterized by independence between the sample mean and sample variance (Rao, 1958), it is a maximum entropy distribution and it provides consistent estimation of the population mean and variance, even if the population is not multivariate normal.

The multivariate normal and Wishart distributions form the foundation of multivariate data analysis. A classic text on the subject is Mardia et al (1979), and one with more coverage of Bayesian approaches is Press (1982). An area of much current Bayesian research involving the multivariate normal distribution is the study of graphical models (Lauritzen, 1996; Jordan, 1998). A graphical model allows elements of the precision matrix to be exactly equal to zero, implying some variables are conditionally independent of each other. A generalization of the Wishart distribution, known as the hyper-inverse-Wishart distribution, has been developed for such models (Dawid and Lauritzen, 1993; Letac and Massam, 2007).