Multivariate Normal Model

Assume that $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ is a random sample from a p-variate normal distribution with mean vector $\boldsymbol{\theta}$ and covariance matrix $\boldsymbol{\Sigma}$.

The density of $\boldsymbol{Y}_i$ is defined on p. 119N.

A conjugate family of priors is one in which

$$\boldsymbol{\Sigma} \sim \text{inverse-Wishart}(\nu_0, \boldsymbol{S}_0^{-1})$$

and

$$\boldsymbol{\theta}|\boldsymbol{\Sigma} \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}/\tau_0),$$

where $\nu_0 > 0$, $\boldsymbol{S}_0$ is a positive definite matrix, $\boldsymbol{\mu}_0$ is any p-vector and $\tau_0 > 0$.

To understand the *inverse-Wishart*, we first define the *Wishart*, which is a probability distribution for a *positive definite matrix*.

The tricky part of defining a distribution over positive-definite matrices is ensuring that each matrix considered is positive definite.

Let $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{\nu_0}$ be arbitrary p-vectors, and suppose that $\nu_0 > p$. *Then the $p \times p$ matrix*

$$\boldsymbol{U} = \sum_{i=1}^{\nu_0} \boldsymbol{z}_i \boldsymbol{z}_i^T$$

is positive definite.

This result is the basis for constructing the Wishart distribution, as follows:

- Let $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{\nu_0}$ be i.i.d. $N(\boldsymbol{0}, \boldsymbol{\Omega}_0)$.

- Then, if $\nu_0 > p$, we say that $\sum_{i=1}^{\nu_0} \boldsymbol{Z}_i \boldsymbol{Z}_i^T$ is distributed Wishart$(\nu_0, \boldsymbol{\Omega}_0)$.

Note that the expectation of a random matrix having Wishart$(\nu_0, \boldsymbol{\Omega}_0)$ distribution is $\nu_0 \boldsymbol{\Omega}_0$.

Let tr($\boldsymbol{A}$) denote the trace of matrix $\boldsymbol{A}$.

The random $p \times p$ matrix $\boldsymbol{X}$ has the Wishart($\nu_0, \boldsymbol{\Omega}_0$) distribution if and only if its density is

$$p(\boldsymbol{x}|\nu_0, \boldsymbol{\Omega}_0) \propto |\boldsymbol{x}|^{(\nu_0-p-1)/2} \exp\left(-\text{tr}(\boldsymbol{\Omega}_0^{-1}\boldsymbol{x})/2\right)$$

for $\boldsymbol{x}$ positive definite, and 0 otherwise.

The Wishart distribution may be regarded as a multivariate analog of the gamma distribution. Indeed, when $p = 1$, Wishart($\nu_0, \boldsymbol{\Omega}_0$) is gamma($\nu_0/2, (2\boldsymbol{\Omega}_0)^{-1}$).

If matrix $\boldsymbol{X}$ has the Wishart distribution, then $\boldsymbol{X}^{-1}$ has the inverse-Wishart distribution. See p. 257 of your text for a definition of the density.

Let $\boldsymbol{Y}$ be the $p \times n$ data matrix with ith column $\boldsymbol{y}_i$. The normal likelihood is

$$p(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{\Sigma}) \propto \frac{1}{|\boldsymbol{\Sigma}|^{n/2}} \exp\left(-\frac{1}{2} Q_n\right),$$

where

$$Q_n = \sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\theta})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_i - \boldsymbol{\theta}).$$

A very useful matrix result in deriving the form of the posterior is

$$\text{tr}(AB) = \text{tr}(BA).$$

Using this result we have

$$\begin{aligned} Q_n &= \sum_{i=1}^{n} \text{tr}\left((\boldsymbol{y}_i - \boldsymbol{\theta})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_i - \boldsymbol{\theta})\right) \\ &= \sum_{i=1}^{n} \text{tr}\left((\boldsymbol{y}_i - \boldsymbol{\theta})(\boldsymbol{y}_i - \boldsymbol{\theta})^T \boldsymbol{\Sigma}^{-1}\right) \\ &= \text{tr}\left[\left(\sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\theta})(\boldsymbol{y}_i - \boldsymbol{\theta})^T\right) \boldsymbol{\Sigma}^{-1}\right]. \end{aligned}$$

The prior distribution has the form

$$p(\boldsymbol{\theta}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(\nu_0+p+2)/2} \exp\left(-\frac{1}{2}\text{tr}(\boldsymbol{S}_0\boldsymbol{\Sigma}^{-1})\right)$$

$$\times \exp\left(-\frac{\tau_0}{2}(\boldsymbol{\theta}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu}_0)\right).$$

We can show that the posterior has exactly the same form with different parameters.

In place of ν_0, τ_0, μ_0 and S_0, we have, respectively,

$$\nu_n = \nu_0 + n, \quad \tau_n = \tau_0 + n,$$

$$\boldsymbol{\mu}_n = \left(\frac{\tau_0}{\tau_0+n}\right)\boldsymbol{\mu}_0 + \left(\frac{n}{\tau_0+n}\right)\bar{\boldsymbol{y}}$$

and

$$\boldsymbol{S}_n = \boldsymbol{S}_0 + n\boldsymbol{S}^2 + \left(\frac{\tau_0 n}{\tau_0+n}\right)(\bar{\boldsymbol{y}}-\boldsymbol{\mu}_0)(\bar{\boldsymbol{y}}-\boldsymbol{\mu}_0)^T,$$

where $\boldsymbol{S}^2$ is the usual *sample covariance matrix*, i.e.,

$$\boldsymbol{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})^T.$$

Finally, we have the following properties of component distributions of the posterior:

- The conditional distribution of $\boldsymbol{\theta}$ given $\boldsymbol{\Sigma}$ and the data is $N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}/\tau_n)$.

- The marginal posterior of $\boldsymbol{\theta}$ is multivariate t.

- The marginal posterior of $\boldsymbol{\Sigma}$ is

  $$\text{inverse-Wishart}(\nu_n, \boldsymbol{S}_n^{-1}).$$

The Jeffreys noninformative prior for the multivariate normal model is

$$p(\boldsymbol{\theta}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(p+1)/2},$$

which is a limiting form of the conjugate prior.

The resulting posterior is proper and given by

$$\boldsymbol{\Sigma}|\boldsymbol{Y} \sim \text{inverse-Wishart}(n-1, (n\boldsymbol{S}^2)^{-1})$$

and

$$\boldsymbol{\theta}|\boldsymbol{\Sigma}, \boldsymbol{Y} \sim N(\bar{\boldsymbol{y}}, \boldsymbol{\Sigma}/n).$$

Example 13 Analysis of blood work

The data in this example come from *Journal of Statistics Education*, Volume 13, Number 3 (November 2005).

(y_1, y_2, y_3) measured on each of $n = 174$ elderly persons.

y_1 = alkaline phospatase

y_2 = calcium

y_3 = inorganic phosphorous

We'll assume the data are a random sample from a trivariate normal distribution.

The sufficient statistics are

$$(\bar{y}_1, \bar{y}_2, \bar{y}_3) = (91.87, 2.36, 1.10)$$

and

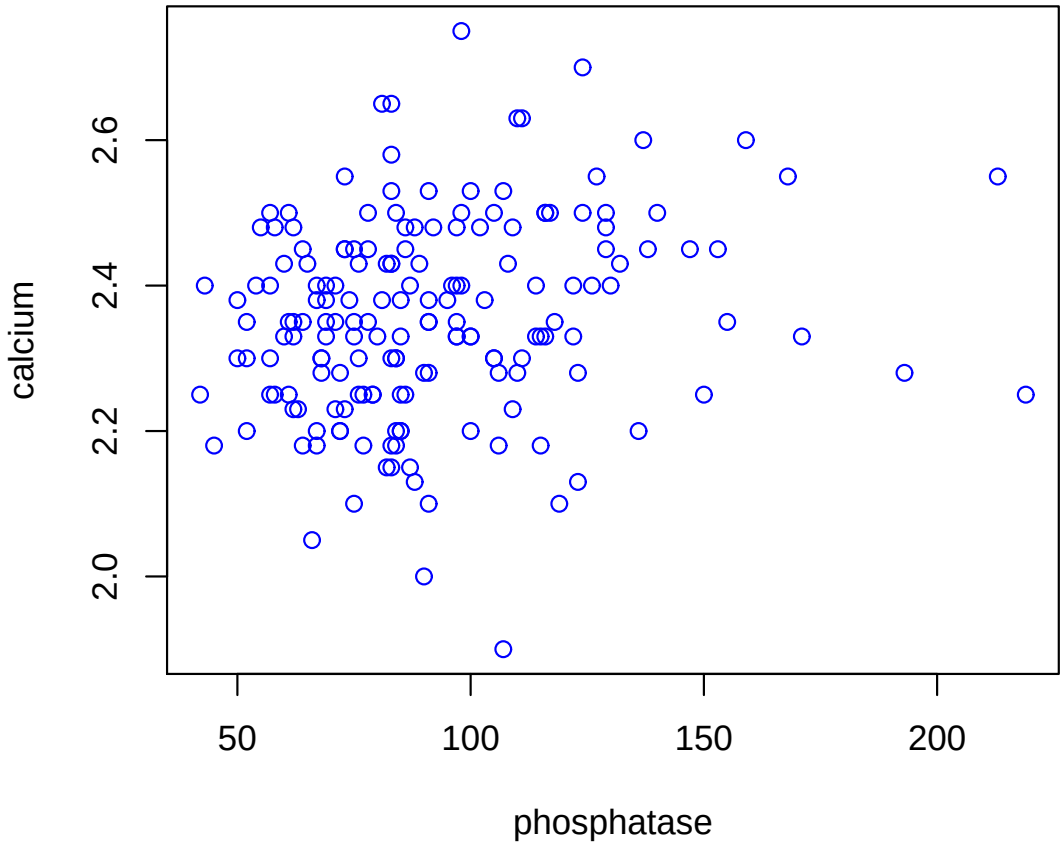
$$\boldsymbol{S}^2 = \begin{bmatrix} 923.97 & 0.79952 & 0.60863 \\ 0.79952 & 0.018728 & 0.0026920 \\ 0.60863 & 0.0026920 & 0.031118 \end{bmatrix}.$$

The sample correlation matrix is

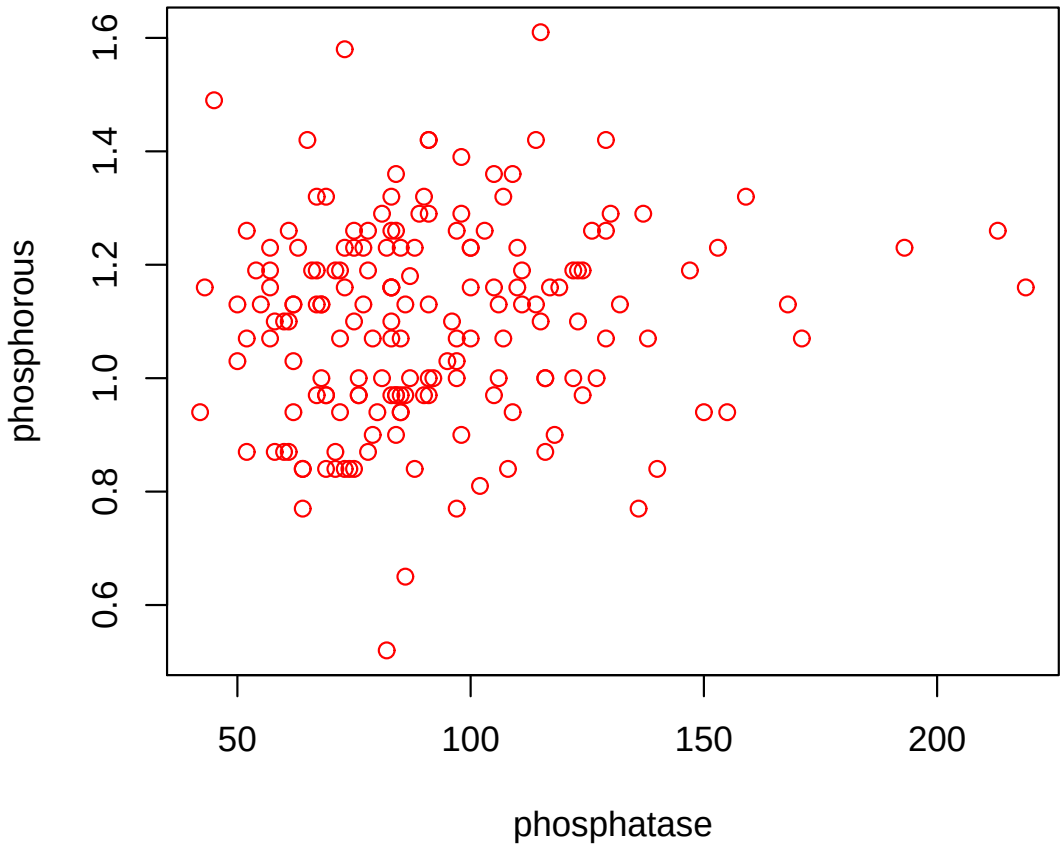
$$\widehat{\boldsymbol{\rho}} = \begin{bmatrix} 1 & 0.1922 & 0.1135 \\ 0.1922 & 1 & 0.1115 \\ 0.1135 & 0.1115 & 1 \end{bmatrix}.$$

Data plots are shown on the next six pages.

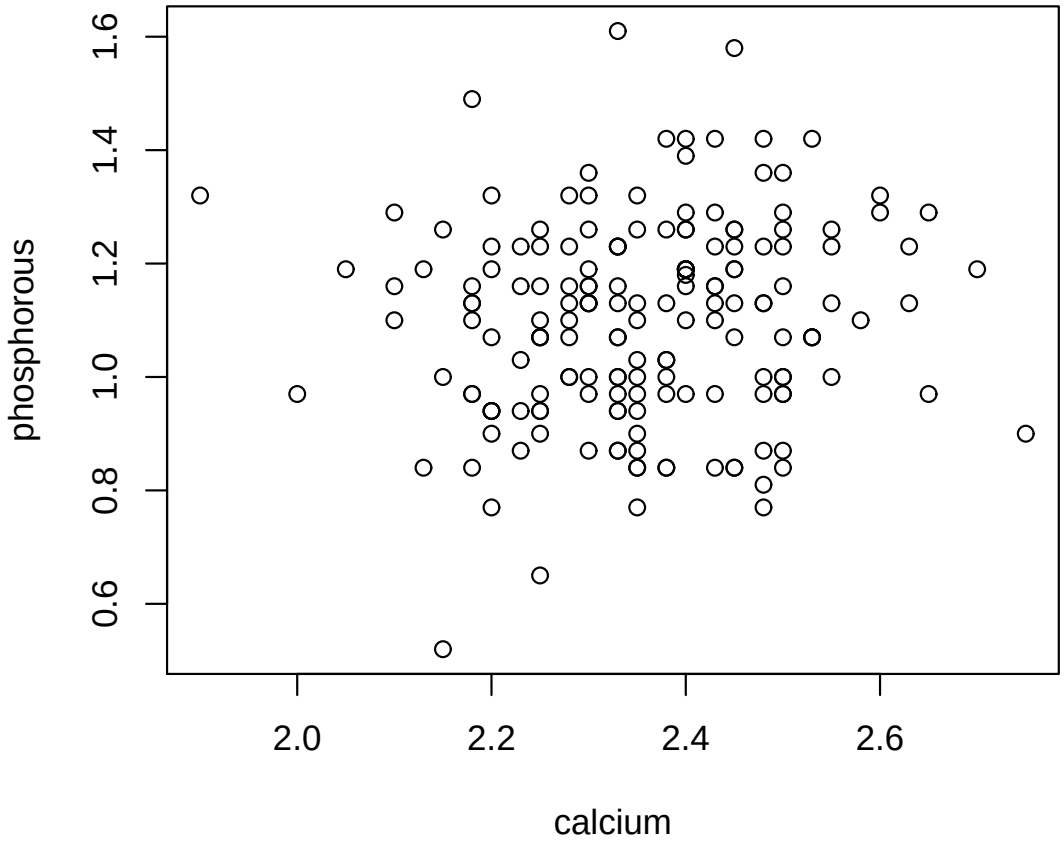
Calcium versus alkaline phosphatase



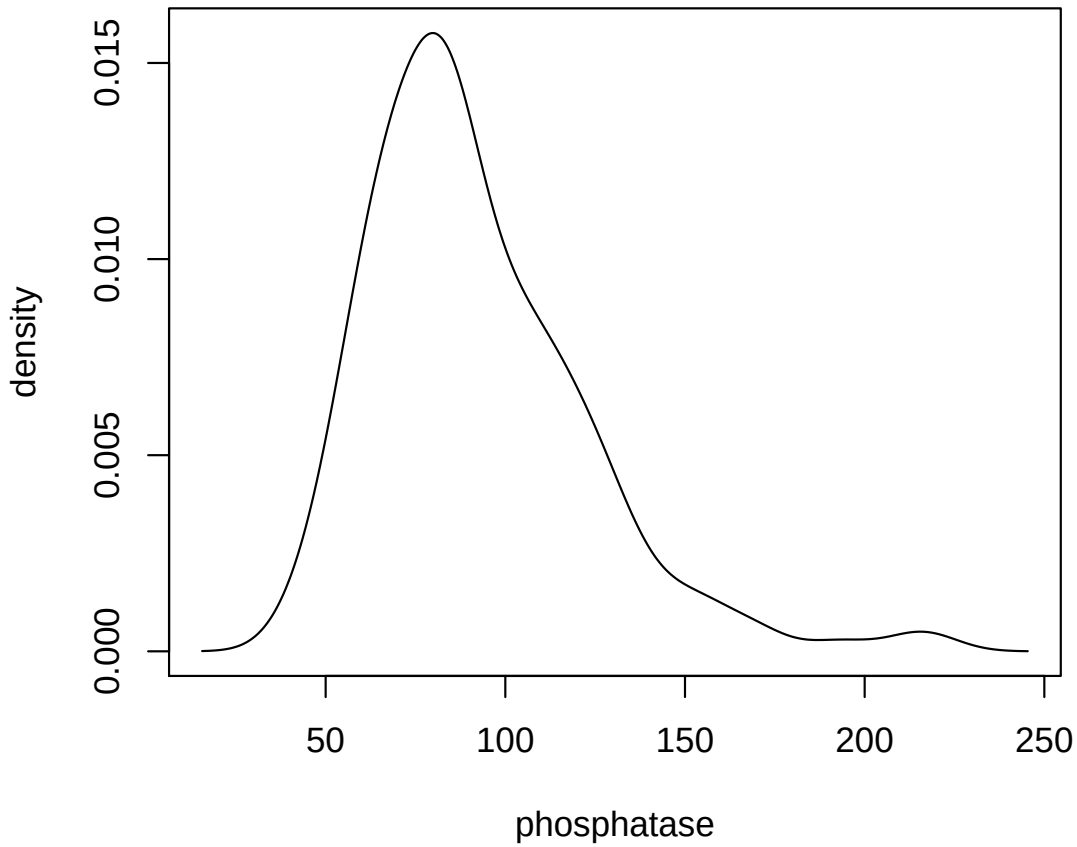
Phosphorous versus alkaline phosphatase



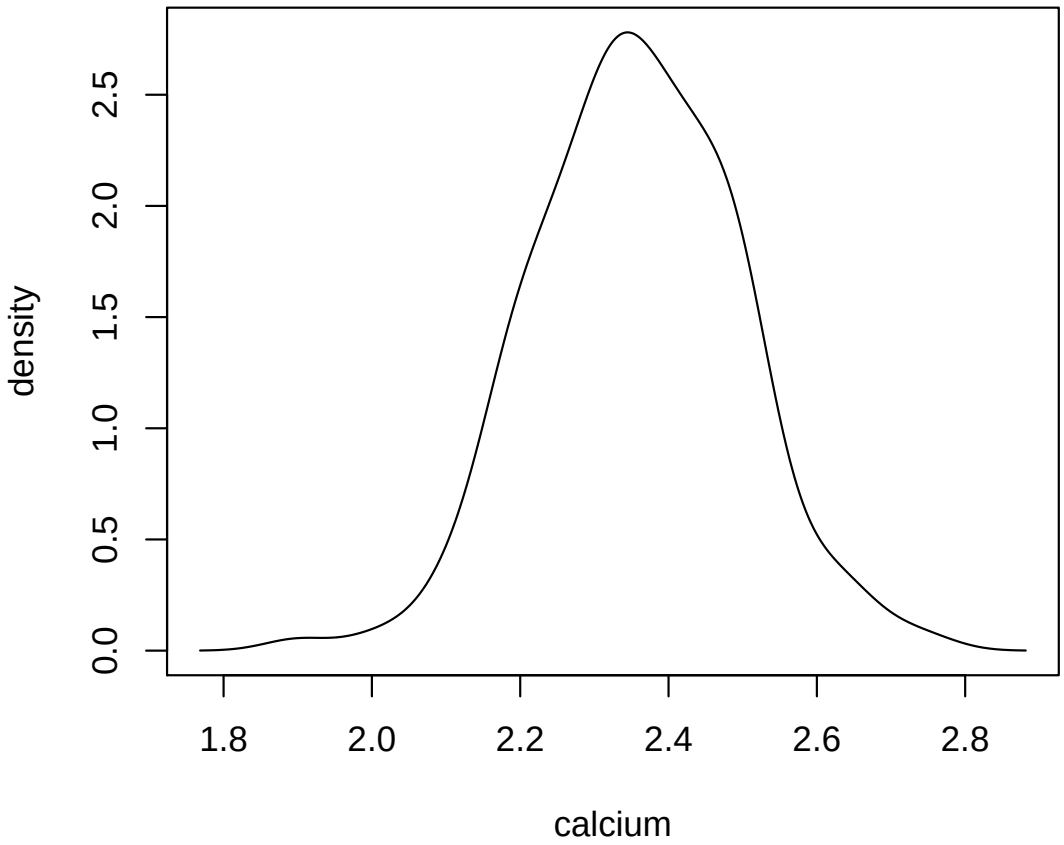
Phosphorous versus calcium



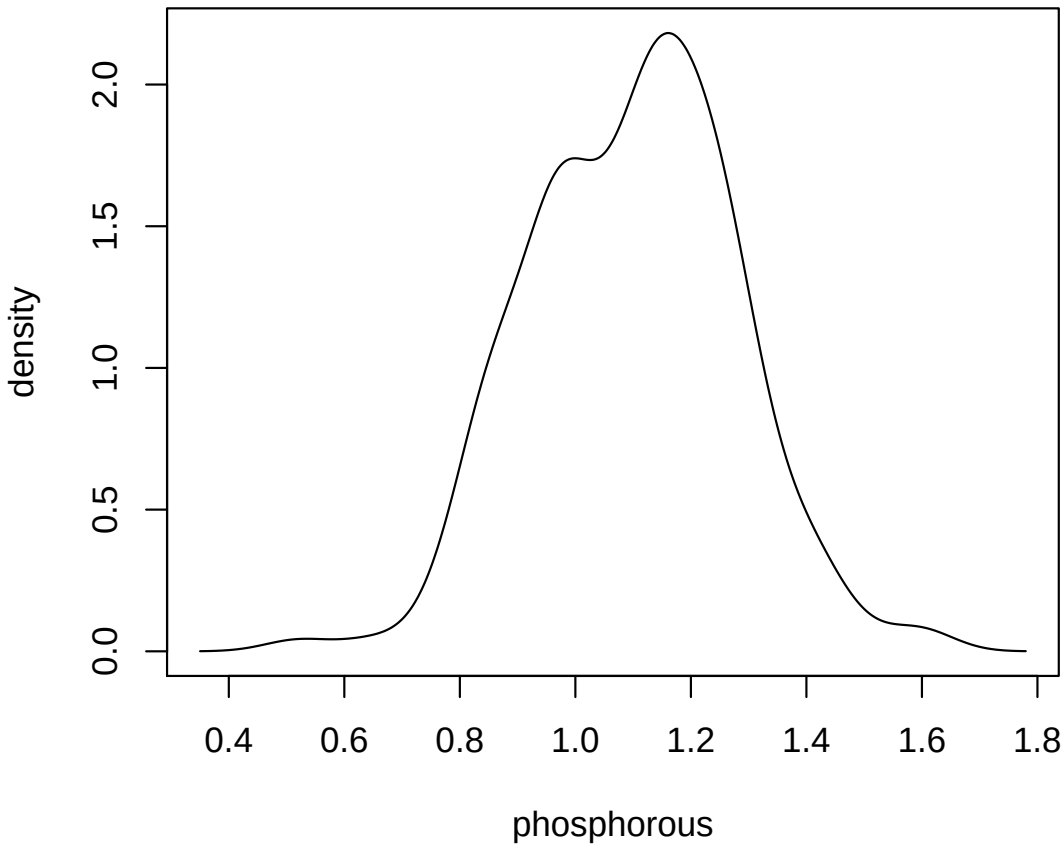
Kernel estimate: alkaline phosphatase



Kernel estimate: calcium



Kernel estimate: phosphorous



We'll use the Jeffreys noninformative prior, and so

$$\boldsymbol{\Sigma}|\boldsymbol{Y} \sim \text{inverse-Wishart}(173, (174\boldsymbol{S}^2)^{-1})$$

and

$$\theta|\boldsymbol{\Sigma}, \boldsymbol{Y} \sim N\left(\bar{\boldsymbol{y}}, \frac{\boldsymbol{\Sigma}}{174}\right).$$

R has a function, `rWishart`, that will generate matrices from a Wishart distribution. One may then invert each of the generated matrices to get a sample from the inverse-Wishart.

A number of different R packages have functions to generate observations from a multivariate normal. The function `rmvnorm` in the package `mixtools` will do so.

I generated 10,000 independent observations from the posterior using these functions.

Using these 10,000 observations I can:

1. Obtain point estimates of all 9 parameters.

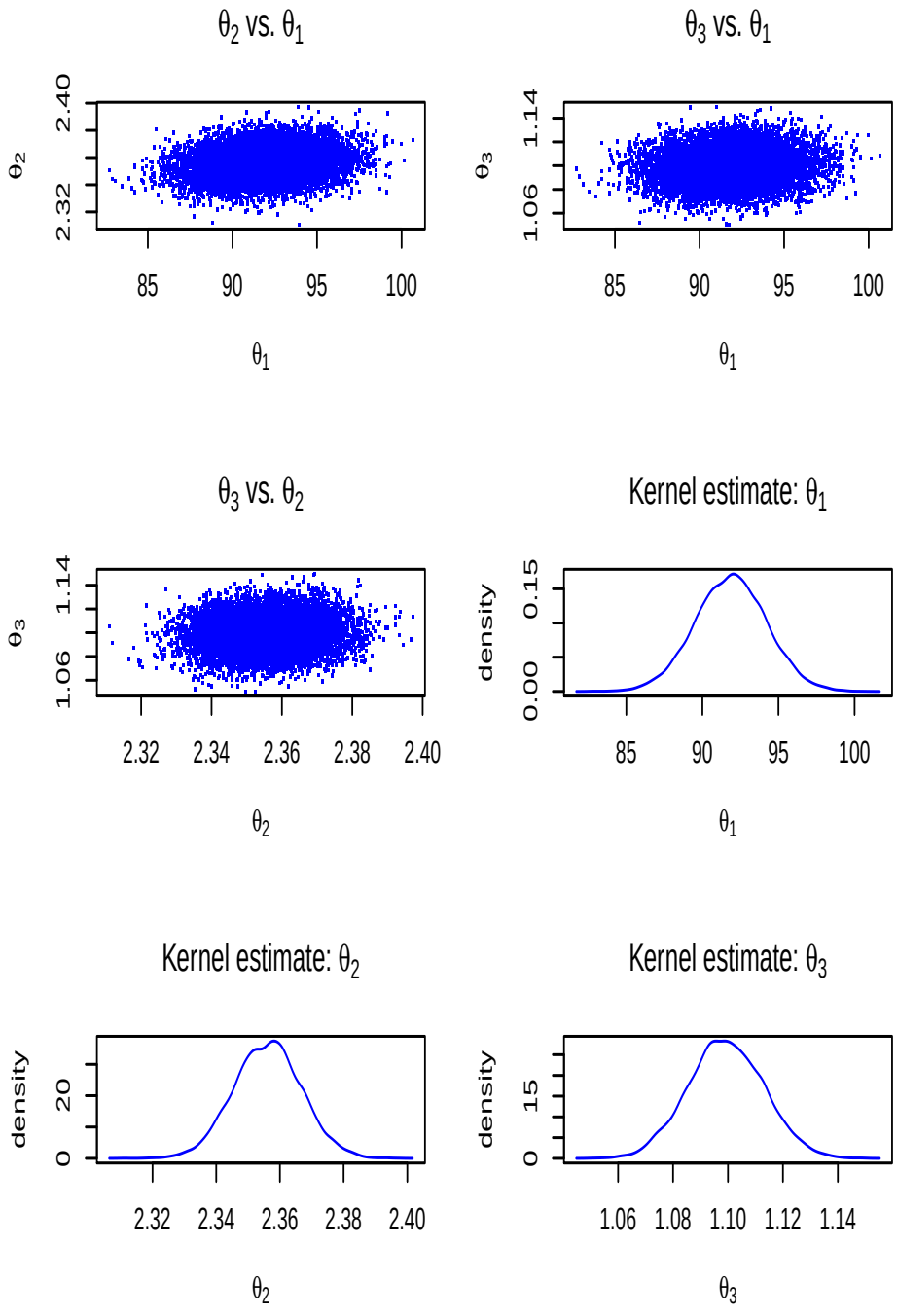
2. Investigate whether, given the data, there are correlations among any of the parameters.

3. Construct kernel density approximations of each of the nine marginal posteriors.

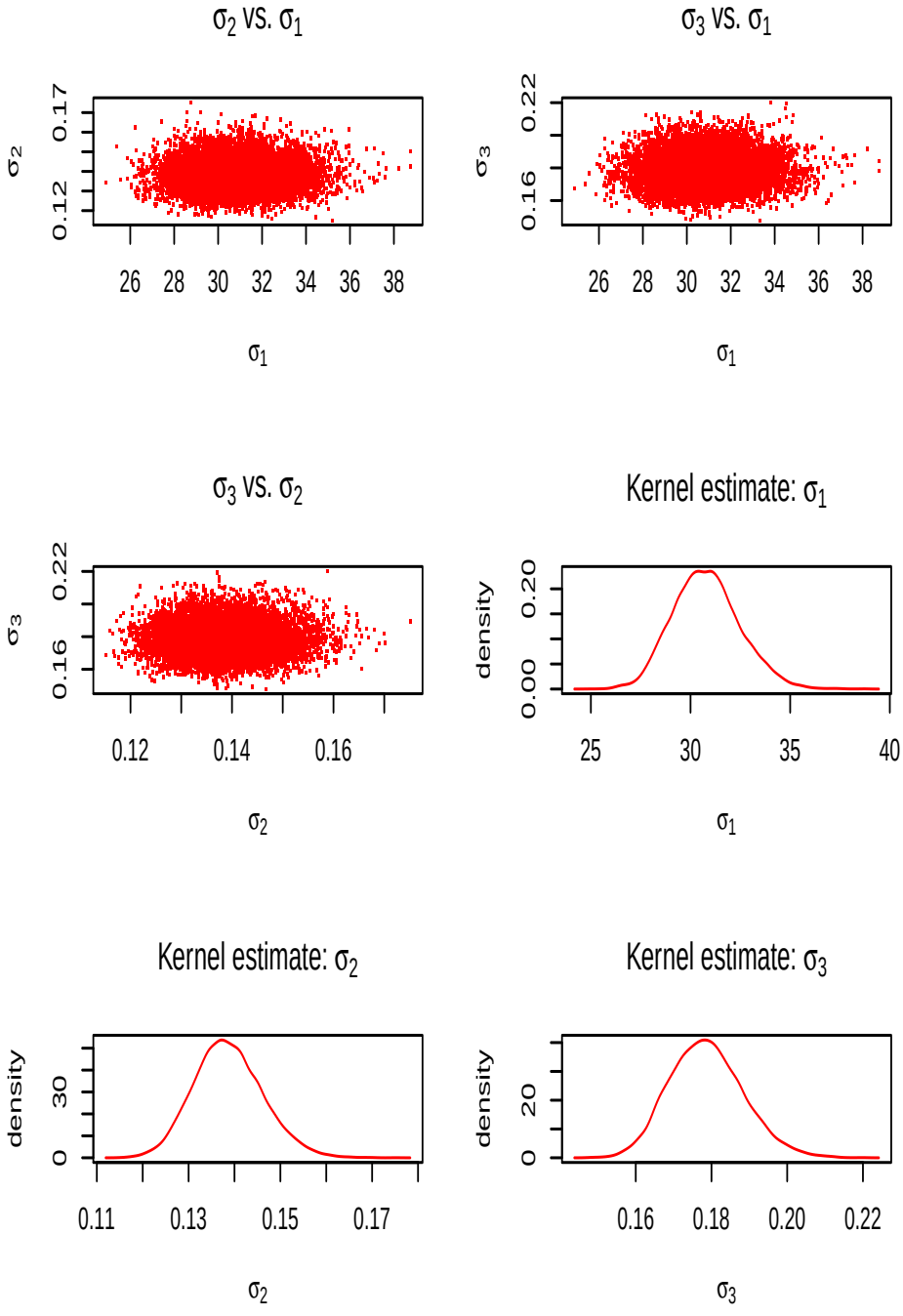
4. Find approximate HPD or credible regions for the parameters.

The second point is important because if two parameters are correlated, *a joint credible region for the two might be smaller than the rectangular region implied by the intersection of two credible intervals.*

Mean vector summary

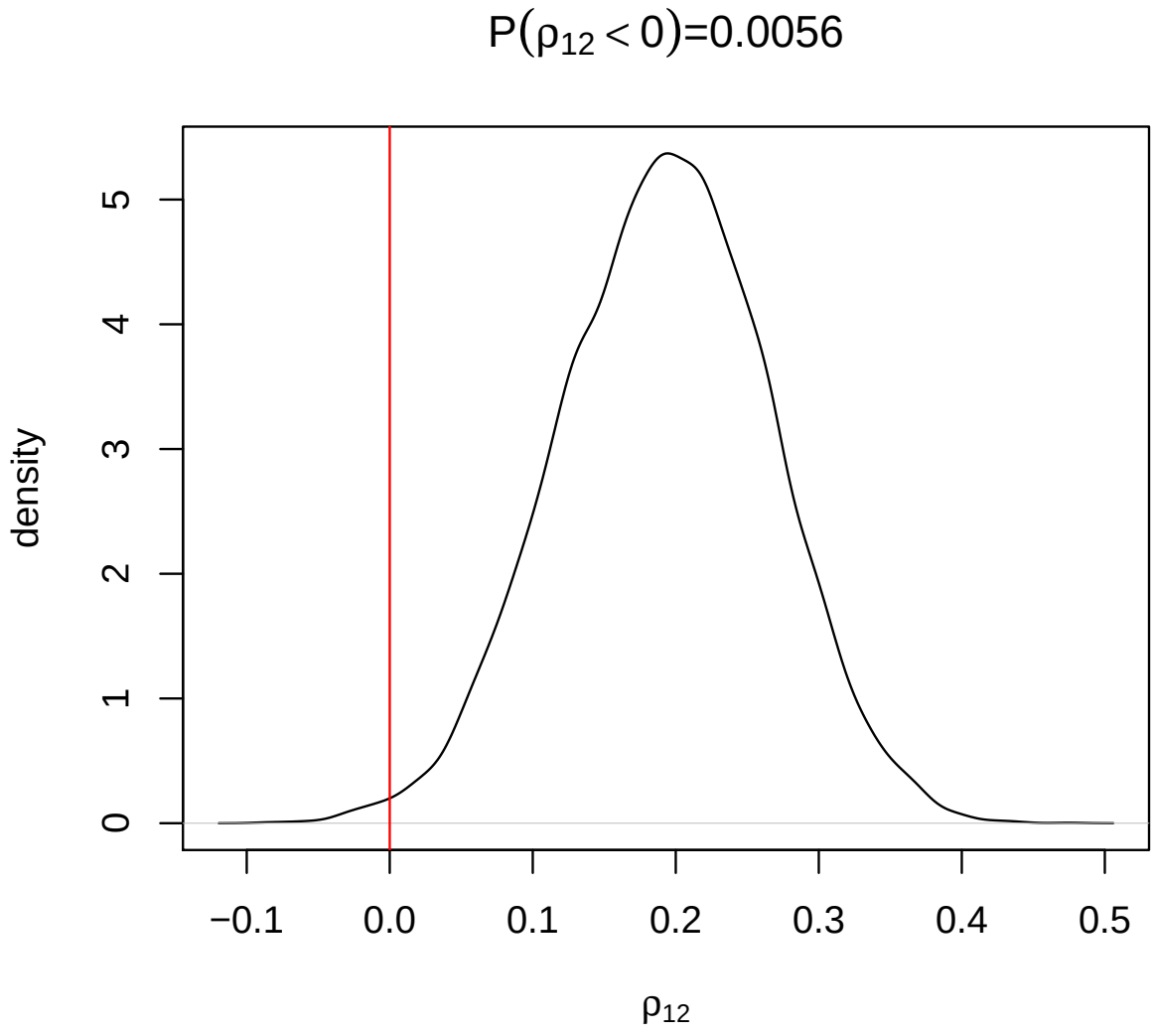


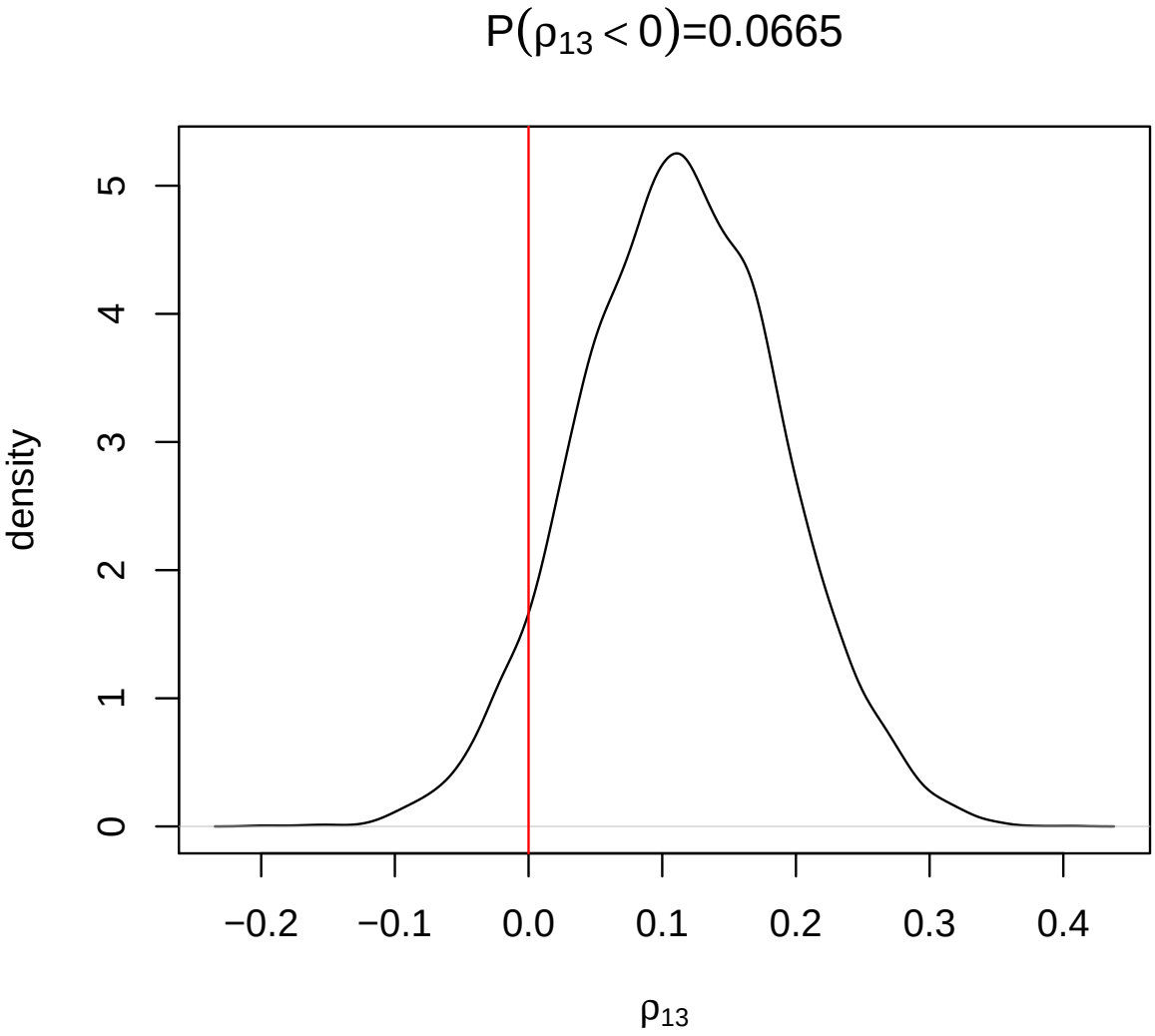
Summary of standard deviations

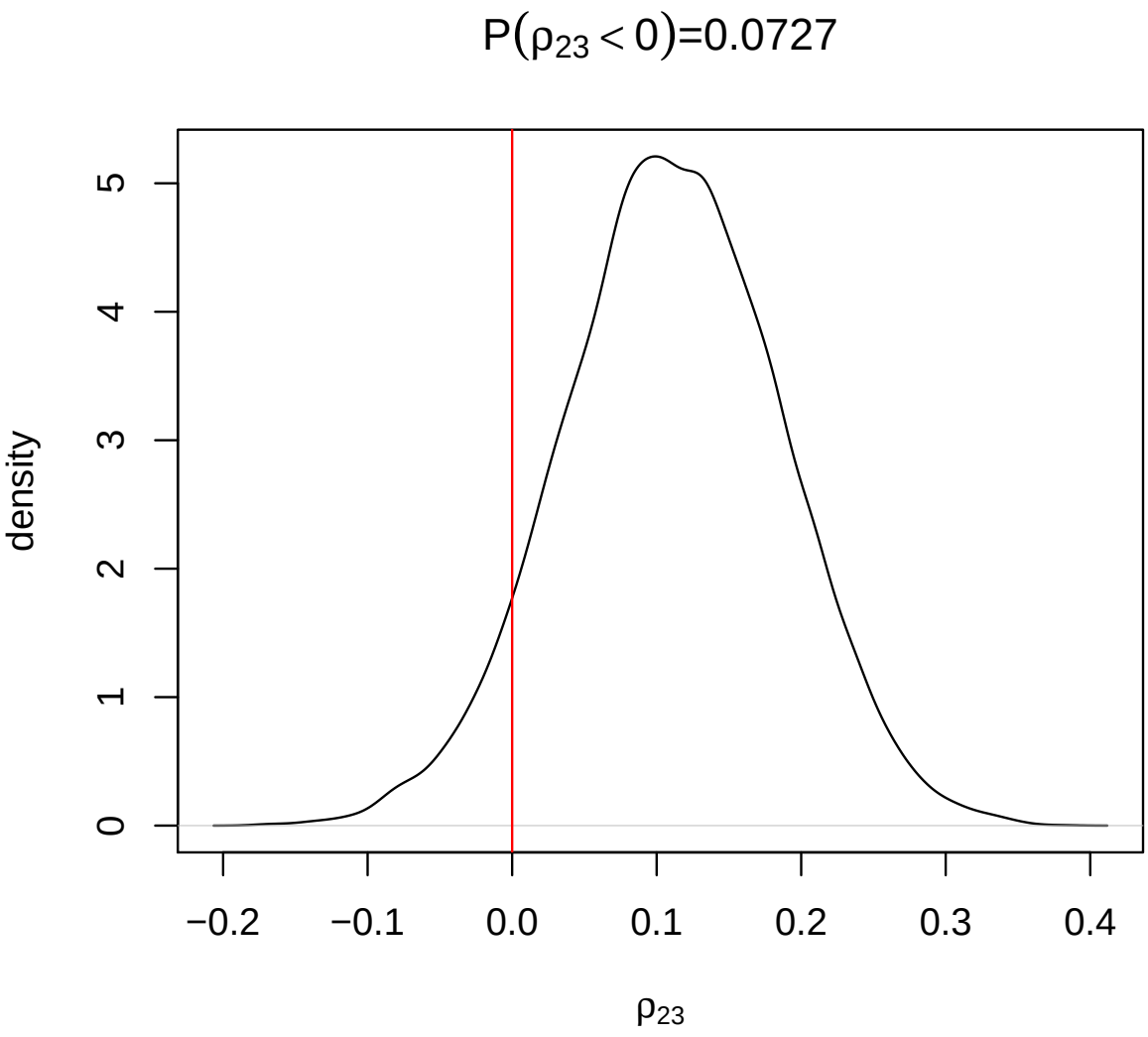


Are any of the variables actually correlated?

We may investigate this question by looking at the marginal posteriors of correlations. Let $\rho_{ij} = \text{Corr}(Y_i, Y_j)$.




P(ρ13 < 0)=0.0665
density
0
1
2
3
4
5
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
ρ13


P(ρ23 < 0)=0.0727
density
0
1
2
3
4
5
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
ρ23

Missing data

Missing data are a common phenomenon in practice. We may start out with n subjects, with the intention of measuring, say, p quantities on each subject.

However, some subjects may drop out of the study, and/or we may have subjects for whom we have some but not all of the p measurements.

We'll consider a fairly simple situation and see how the likelihood can be affected by missing data.

Suppose that potential p-variate observations $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ are i.i.d., *but not all components of every $\boldsymbol{Y}_i$ are observed.*

Define the indicator vector $\boldsymbol{O}_i = (O_{i1}, \ldots, O_{ip})$ by

$$O_{ij} = \begin{cases} 1, & \text{if } Y_{ij} \text{ is not missing,} \\ 0, & \text{if } Y_{ij} \text{ is missing.} \end{cases}$$

In Bayesian analysis, the main "trick" to deal with missing data is to modify the likelihood correctly to reflect the missingness. Then we just proceed as usual.

To determine the likelihood, we need the joint distribution of $(\boldsymbol{Y}_i, \boldsymbol{O}_i)$. In obtaining this distribution, we assume that the "missingness" of $\boldsymbol{Y}_i$ is unrelated to parameters of the underlying model, *and that it does not depend on* $\boldsymbol{Y}_i$.

This kind of data is called *missing at random*.

Let $\boldsymbol{o}_i$ be the observed value of $\boldsymbol{O}_i$ and let $\boldsymbol{y}_i^o$ be the components of the data vector $\boldsymbol{Y}_i$ that aren't missing.

The likelihood for $(\boldsymbol{y}_i^o, \boldsymbol{o}_i)$ is

$$P(\boldsymbol{O}_i = \boldsymbol{o}_i | \boldsymbol{y}_i^o) f(y_i^o | \boldsymbol{\theta}) =$$

$$P(\boldsymbol{O}_i = \boldsymbol{o}_i) f(\boldsymbol{y}_i^o | \boldsymbol{\theta}).$$

The quantity $f(\boldsymbol{y}_i^o|\boldsymbol{\theta})$ is just the marginal density of the nonmissing components of $\boldsymbol{Y}_i$ evaluated at $\boldsymbol{y}_i^o$.

This, of course, will depend upon the particular model being considered.

In the multivariate normal case, the joint distribution of $(Y_{ij_1}, \ldots, Y_{ij_k})$ is multivariate normal with mean vector $(\theta_{j_1}, \ldots, \theta_{j_k})$ and covariance matrix with rs element equal to $\text{Cov}(Y_{ij_r}, Y_{ij_s})$.

The likelihood function may be expressed as

$$\prod_{i=1}^{n} P(\boldsymbol{O}_i = \boldsymbol{o}_i) f(y_i^o|\boldsymbol{\theta}) =$$

$$\prod_{i=1}^{n} P(\boldsymbol{O}_i = \boldsymbol{o}_i) \prod_{i=1}^{n} f(y_i^o|\boldsymbol{\theta}).$$

An important point about the likelihood is that the term

$$\prod_{i=1}^{n} P(\boldsymbol{O}_i = \boldsymbol{o}_i)$$

does not depend on $\boldsymbol{\theta}$ and hence is just a constant of proportionality. *It has no effect whatsoever on the posterior.*

In the Bayesian setting, a practical problem caused by missing data is that a conjugate prior may no longer be conjugate.

To see why, consider the multivariate normal case where only one of the n data vectors has any missing components. Suppose we only observe the first component of $\boldsymbol{Y}_1$.

If we use our conjugate prior, then

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\Sigma}|\boldsymbol{Y}) \;\propto\; & \text{(normal-inverse Wishart)} \\ & \times \frac{1}{\sigma_1} \exp\left(-\frac{(y_{11} - \theta_1)^2}{2\sigma_1^2}\right). \end{aligned}$$

So we can no longer sample from a normal-inverse Wishart to obtain samples from the posterior.

A solution to this problem is to treat the missing data like additional unknown parameters, and then use Gibbs sampling.

Let $\boldsymbol{Y}_{\text{obs}}$ represent the observed data and $\boldsymbol{Y}_{\text{miss}}$ the missing data. In general, if we can draw samples $(\boldsymbol{\theta}, \boldsymbol{Y}_{\text{miss}})$ from

$$p(\boldsymbol{\theta}, \boldsymbol{Y}_{\text{miss}} | \boldsymbol{Y}_{\text{obs}}),$$

then we may use all the values of $\boldsymbol{\theta}$ drawn as our sample from the posterior $p(\boldsymbol{\theta}|\boldsymbol{Y}_{\text{obs}})$.

Consider the multivariate normal case. If we use Gibbs sampling, we need each of the following full conditionals:

- $p(\boldsymbol{\theta}|\boldsymbol{\Sigma}, \boldsymbol{Y}_{\text{miss}}, \boldsymbol{Y}_{\text{obs}})$
- $p(\boldsymbol{\Sigma}|\boldsymbol{\theta}, \boldsymbol{Y}_{\text{miss}}, \boldsymbol{Y}_{\text{obs}})$
- $p(\boldsymbol{Y}_{\text{miss}}|\boldsymbol{\theta}, \boldsymbol{\Sigma}, \boldsymbol{Y}_{\text{obs}})$

Each of these distributions is known if we use the normal-inverse Wishart prior.

Example 14 Missing blood work data

In the full blood work data set there are $n = 178$ cases, but two cases are missing one of y_1, y_2 or y_3. Case 42 is missing the phosphorous reading and case 85 the calcium reading.

We'll assume that these observations are missing at random.

If we just deleted these two cases we'd be throwing away some information.

We'll continue to use the Jeffreys' prior. Using the result on p. 157N, we know that

$$p(\boldsymbol{\theta}|\boldsymbol{\Sigma}, \boldsymbol{Y}_{\text{miss}}, \boldsymbol{Y}_{\text{obs}})$$

is $N(\bar{\boldsymbol{y}}_*, \boldsymbol{\Sigma}/n)$, where $\bar{\boldsymbol{y}}_*$ is the sample mean vector of the $n \times 3$ data matrix that we get by augmenting $\boldsymbol{Y}_{\text{obs}}$ with $\boldsymbol{Y}_{\text{miss}}$.

When $\boldsymbol{\Sigma} \sim$ inverse Wishart$(\nu, \boldsymbol{M})$ and $\theta|\boldsymbol{\Sigma} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}/n_0)$, it follows that $\boldsymbol{\Sigma}|\boldsymbol{\theta}$ is distributed inverse Wishart with parameters $\nu + 1$ and

$$\left(\boldsymbol{M}^{-1} + n_0(\boldsymbol{\theta} - \boldsymbol{\mu})(\boldsymbol{\theta} - \boldsymbol{\mu})^T\right)^{-1}.$$

Using this result, the distribution of $\boldsymbol{\Sigma}$ given $(\boldsymbol{\theta}, \boldsymbol{Y}_{\text{miss}}, \boldsymbol{Y}_{\text{obs}})$ is inverse Wishart with parameters n and

$$\frac{1}{n}\left(\boldsymbol{S}_*^2 + (\boldsymbol{\theta} - \bar{\boldsymbol{y}}_*)(\boldsymbol{\theta} - \bar{\boldsymbol{y}}_*)^T\right)^{-1},$$

where $\boldsymbol{S}_*^2$ is the sample covariance matrix of the $n \times 3$ data matrix gotten by augmenting $\boldsymbol{Y}_{\text{obs}}$ with $\boldsymbol{Y}_{\text{miss}}$.

Finally, we need $p(\boldsymbol{Y}_{\text{miss}}|\boldsymbol{\theta}, \boldsymbol{\Sigma}, \boldsymbol{Y}_{\text{obs}})$, which in our example is

$$p(y_{42,3}|y_{42,1}, y_{42,2}, \boldsymbol{\theta}, \boldsymbol{\Sigma})$$

$$\times p(y_{85,2}|y_{85,1}, y_{85,3}, \boldsymbol{\theta}, \boldsymbol{\Sigma}).$$

Why?

As in Example 12, the last two conditional densities are normal.

We can now write down a Gibbs sampling algorithm to sample from the distribution of $(\boldsymbol{\theta}, \boldsymbol{\Sigma}, \boldsymbol{Y}_{\text{miss}})$ given $\boldsymbol{Y}_{\text{obs}}$.

Suppose the current observation is

$$(\boldsymbol{\theta}_t, \boldsymbol{\Sigma}_t, \boldsymbol{Y}_{\text{miss},t}).$$

1. Generate $\boldsymbol{\theta}_{t+1}$ from $p(\boldsymbol{\theta}|\boldsymbol{\Sigma}_t, \boldsymbol{Y}_{\text{miss},t}, \boldsymbol{Y}_{\text{obs}})$.

2. Generate $\boldsymbol{\Sigma}_{t+1}$ from $p(\boldsymbol{\Sigma}|\boldsymbol{\theta}_{t+1}, \boldsymbol{Y}_{\text{miss},t}, \boldsymbol{Y}_{\text{obs}})$.

3. Generate $\boldsymbol{Y}_{\text{miss},t+1}$ from $p(\boldsymbol{Y}_{\text{miss}}|\boldsymbol{\theta}_{t+1}, \boldsymbol{\Sigma}_{t+1}, \boldsymbol{Y}_{\text{obs}})$.

4. Iterate steps 1-3.