AMS 207: Intermediate Bayesian Modeling

4: Multivariate Models (continued); Hierarchical Modeling

David Draper

Department of Applied Mathematics and Statistics
University of California, Santa Cruz

draper@ams.ucsc.edu
http://www.ams.ucsc.edu/~draper

© 2010 David Draper (all rights reserved)
Gaussian Modeling Review

You’ll recall from AMS 206 that the conjugate prior with the Gaussian sampling model

\[
(y_i|\theta) = (y_i|\mu, \sigma^2) \overset{IID}{\sim} N(\mu, \sigma^2), \quad i = 1, \ldots, n, \tag{1}
\]

with both components of \(\theta = (\mu, \sigma^2)\) unknown, can be expressed most intuitively in a hierarchical way, with a scaled \(\chi^{-2}\) prior on \(\sigma^2\) and then a conditional normal prior on \((\mu|\sigma^2)\):

\[
\begin{align*}
\sigma^2 & \sim \chi^{-2}(\nu_0, \sigma_0^2) \\
(\mu|\sigma^2) & \sim N\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right)
\end{align*}
\tag{2}
\]

\[
(y_i|\theta) = (y_i|\mu, \sigma^2) \overset{IID}{\sim} N(\mu, \sigma^2), \quad i = 1, \ldots, n.
\]

GCSR call \(\chi^{-2}\) the scaled inverse \(\chi^2\) distribution (written Inv-\(\chi^2\) in their book), but \(\chi^{-2}\) captures the inverse \(\chi^2\) idea well, and I’ll use that notation in this course.

As noted in their Appendix A, \(\chi^{-2}(\nu_0, \sigma_0^2)\) is a special case of the inverse Gamma distribution, namely \(\Gamma^{-1}\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right)\) (what I call \(\Gamma^{-1}(\alpha, \beta)\), GCSR call Inv-\(\Gamma(\alpha, \beta)\)); you’ll recall that \(\lambda \sim \Gamma^{-1}(\alpha, \beta)\) just means that \(\frac{1}{\lambda} \sim \Gamma(\alpha, \beta)\).

\(\sigma^2 \sim \chi^{-2}(\nu_0, \sigma_0^2)\) means that \(\sigma^2\) has density

\[
p(\sigma^2) = c \left(\sigma^2\right)^{-\left(\frac{\nu_0}{2}+1\right)} \exp\left(-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right), \quad \text{where}
\]

\[
c = \frac{\left(\frac{\nu_0}{2}\right)^{\frac{\nu_0}{2}} \left(\sigma_0^2\right)^{\frac{\nu_0}{2}}}{\Gamma\left(\frac{\nu_0}{2}\right)}. \tag{3}
\]
\[ \chi^{-2} \text{ Prior for } \sigma^2 \]

As GCSR’s Appendix A points out, if \( \sigma^2 \sim \chi^{-2}(\nu_0, \sigma_0^2) \) then

\[
E_B(\sigma^2) = \frac{\nu_0 \sigma_0^2}{\nu_0 - 2} \quad \text{for } \nu_0 > 2;
\]

\[
\text{mode}_B(\sigma^2) = \frac{\nu_0 \sigma_0^2}{\nu_0 + 2}, \quad \text{and}
\]

\[
V_B(\sigma^2) = \frac{2\nu_0^2 \sigma_0^4}{(\nu_0 - 2)^2(\nu_0 - 4)} \quad \text{for } \nu_0 > 4.
\]  

(4)

From this you can see that \( E_B(\sigma^2) \approx \text{mode}_B(\sigma^2) \approx \sigma_0^2 \), so with this prior \( \sigma_0^2 \) acts like a prior estimate of \( \sigma^2 \), and it turns out (see p. 50 of the book) that, if you use this prior in the simpler Gaussian model where \( \sigma^2 \) is unknown but \( \mu \) is known, the \( \chi^{-2}(\nu_0, \sigma_0^2) \) prior is conjugate and the posterior for \( \sigma^2 \) is

\[
(\sigma^2 | y) \sim \chi^{-2}\left(\nu_0 + n, \frac{\nu_0 \sigma_0^2 + n v}{\nu_0 + n}\right),
\]  

(5)

where \( v = \frac{1}{n} \sum_{i=1}^{n} (y_i - \mu)^2 \) is a sufficient statistic and as usual \( y = (y_1, \ldots, y_n) \).

From this the conjugate updating rule is evidently

\[
\nu_0 \rightarrow \nu_0 + n, \quad \sigma_0^2 \rightarrow \frac{\nu_0 \sigma_0^2 + n \nu}{\nu_0 + n},
\]

(6)

and it’s clear that, with this sampling model, the \( \chi^{-2}(\nu_0, \sigma_0^2) \) prior for \( \sigma^2 \) is like a data set with prior sample size \( \nu_0 \) and prior estimate \( \sigma_0^2 \).

We can readily verify these conclusions graphically:
\chi^2 \text{ Prior for } \sigma^2 \text{ (continued)}

sauternes 333> maple

\begin{verbatim}
> p := (sigma_2, nu_0, sigma_0) -> sigma_2^(- (1 + nu_0/2)) * 
    (nu_0/2)^(-nu_0/2) * (sigma_0^2)^(-nu_0/2) * 
    exp(-nu_0*sigma_0^2 / (2 * sigma_2)) / GAMMA(nu_0/2);

(-1 - 1/2 nu_0)

p := (sigma_2, nu_0, sigma_0) -> sigma_2

(1/2 nu_0) 2 (1/2 nu_0) nu_0 sigma_0
(1/2 nu_0) (sigma_0) exp(-1/2  

GAMMA(1/2 nu_0)

> assume( nu_0 > 0, sigma_0 > 0 );

> simplify( integrate( p( sigma_2, nu_0, sigma_0 ), 
    sigma_2 = 0 .. infinity ) );

1

> simplify( integrate( sigma_2 * p( sigma_2, nu_0, sigma_0 ), 
    sigma_2 = 0 .. infinity ) );

nu_0^\sigma_0^2

---------------------------------

nu_0^\sigma_0^2 - 2

> plotsetup( x11 );
\end{verbatim}
\( \chi^{-2} \) Prior for \( \sigma^2 \) (continued)

> plot( \{ p( \text{sigma}_2, 1, 1 ), p( \text{sigma}_2, 10, 1 ), p( \text{sigma}_2, 100, 1 ) \}, \text{sigma}_2 = 0 .. 3, \text{color} = \text{black} );

(becoming more **peaked** around the **same location** as \( \nu_0 \) increases (check))

> plot( \{ p( \text{sigma}_2, 1, 1 ), p( \text{sigma}_2, 1, 2 ), p( \text{sigma}_2, 1, 3 ) \}, \text{sigma}_2 = 0 .. 10, \text{color} = \text{black} );

(location **moving to the right** as \( \sigma_0^2 \) increases (check))
Conditional Normal Prior for $\mu$

What about the interpretation of $\mu_0$ and $\kappa_0$ in the conditional normal prior for $\mu$: $(\mu|\sigma^2) \sim N\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right)$?

In part 2 of these lecture notes we agreed that if the sampling model was $(y_i|\mu) \sim^{\text{IID}} N(\mu, \sigma^2)$ ($i = 1, \ldots, n$) with $\sigma$ known, the conjugate prior is $N(\mu_0, \sigma_0^2)$ and the resulting posterior is Gaussian with mean

$$E_B(\mu|y) = \frac{\frac{1}{\sigma_0^2} \mu_0 + \frac{n}{\sigma^2} \bar{y}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} = \frac{n_0\mu_0 + n\bar{y}}{n_0 + n} \quad (7)$$

and variance

$$V_B(\mu|y) = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1} = \frac{\sigma^2}{n_0 + n}, \quad (8)$$

where the prior sample size is $n_0 = \frac{\sigma_0^2}{\kappa_0}$.

Set $\sigma_0^2 = \frac{\sigma^2}{\kappa_0}$ to get that, in the conditional normal prior, $n_0 = \frac{\sigma^2}{\kappa_0} = \kappa_0$; thus in the Gaussian sampling model with both $\mu$ and $\sigma^2$ unknown, the hierarchical prior

$$\sigma^2 \sim \chi^{-2}(\nu_0, \sigma_0^2)$$

$$(\mu|\sigma^2) \sim N\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right) \quad (9)$$

acts like

(a) a data set for $\sigma^2$ with prior sample size $\nu_0$ and prior estimate $\sigma_0^2$, and

(b) a data set for $\mu$ with prior sample size $\kappa_0$ and prior estimate $\mu_0$.  

6
Joint Posterior for $\mu$ and $\sigma^2$

Hierarchical models always describe mixtures: in this case $p(\mu, \sigma^2) = p(\sigma^2) p(\mu|\sigma^2)$, and to simulate from this prior you can just

(a) make a $\sigma^2$ draw from $\chi^2(\nu_0, \sigma_0^2)$, obtaining $\sigma_*^2$, and then

(b) make a $\mu$ draw from $N\left(\mu_0, \frac{\sigma_*^2}{\kappa_0}\right)$.

In a small departure from GCSR’s name for this joint conjugate prior, let’s call it the $N-\chi^2\left(\mu_0, \frac{\sigma^2}{\kappa_0}; \nu_0, \sigma_0^2\right)$ distribution.

By conjugacy the posterior is another member of this family, which can be written $N-\chi^2\left(\mu_n, \frac{\sigma_n^2}{\kappa_n}; \nu_n, \sigma_n^2\right)$, in which the updating rules are given by

\[
\begin{align*}
\nu_n &= \nu_0 + n \\
\kappa_n &= \kappa_0 + n \\
\mu_n &= \frac{\kappa_0 \mu_0 + n \bar{y}}{\kappa_0 + n} \\
\nu_n \sigma_n^2 &= \nu_0 \sigma_0^2 + (n - 1)s^2 + \frac{\kappa_0 n}{\kappa_0 + n} (\bar{y} - \mu_0)^2,
\end{align*}
\]

in which $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ and $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$ are sufficient for $\mu$ and $\sigma^2$.

This updating is simple and beautifully intuitive, and it also indicates how to achieve diffuseness in the conjugate prior, if that’s appropriate scientifically (choose $\nu_0$ and $\kappa_0$ positive but small relative to $n$).

The joint posterior for $\mu$ and $\sigma^2$ can be examined graphically, but it’s probably even more important to look at the marginals $p(\sigma^2|y)$ and $p(\mu|y)$. 
Multivariate Generalization

Since $\sigma^2$ is at the top of the hierarchy in the prior, we can just read the form of the posterior for it directly from the updating rules:

$$ (\sigma^2|y) \sim \chi^{-2}(\nu_n, \sigma_n^2). \quad (11) $$

The marginal for $\mu$ requires a bit more work:

$$ (\mu|y) = \int_0^\infty p(\mu, \sigma^2|y) \, d\sigma^2. \quad (12) $$

This is an unpleasant but analytically tractable integral; when the dust settles you get

$$ (\mu|y) \sim t_{\nu_n}(\mu_n, \sigma_n^2), \quad (13) $$

in which $t_{\nu}(\mu, \sigma^2)$ is the scaled t distribution with location $\mu$, scale $\sigma$ and degrees of freedom $\nu$ (this just means that if $\theta \sim t_{\nu}(\mu, \sigma^2)$ then $\frac{\theta - \mu}{\sigma} \sim t_{\nu}$, where $t_{\nu}$ is the standard t distribution in the back of any frequentist textbook).

As GCSR note in their Appendix A, if $\theta \sim t_{\nu}(\mu, \sigma^2)$ then

$$ p(\theta|\mu, \sigma^2, \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sigma \Gamma\left(\frac{\nu}{2}\right) \sqrt{\pi \nu}} \left[ 1 + \frac{1}{\nu} \left(\frac{\theta - \mu}{\sigma}\right)^2 \right]^{-\frac{\nu+1}{2}} $$

mode($\theta$) = $\mu$

$E_B(\theta) = \mu$ if $\nu > 1$

$V_B(\theta) = \frac{\nu}{\nu - 2} \sigma^2$ if $\nu > 2$. \quad (14)

How does all of this generalize when each entry $y_i$ in $y = (y_1, \ldots, y_n)$ is a vector of length $d > 1$?

In other words, suppose you had 2 or more outcome variables, each of which you thought could be modeled marginally as (univariate) Gaussian; how does the modeling go then?
Multivariate Normal Sampling Model

Evidently we first need a multivariate generalization of the univariate normal distribution.

This is not as easy as it may sound, because — in the world of multivariate sampling models — you and I might have different ideas about what properties we want the multivariate generalization to have, and we might therefore end up with different multivariate distributions.

We’re trying to “invent” a multivariate normal sampling model, in which the data set would be \( n \) vectors, each of length \( d \); let’s arrange that in an \((n \times d)\) matrix \( Y \) whose typical row \( y_i \) has elements \((y_{i1}, \ldots, y_{id})\).

So, OK, what do you like about the univariate Gaussian distribution that you’d want to preserve in its multivariate generalization?

- It’s nice that in the Gaussian family if you know \( \mu \) and \( \sigma^2 \) you’ve put your finger on a unique Gaussian density; with \( d \) outcome variables the analogue of \( \mu \) would be a mean vector \( \mu = (\mu_1, \ldots, \mu_d) \), the analogue of \( \sigma^2 \) would be a (symmetric) \((d \times d)\) covariance matrix \( \Sigma \) — whose \((j, j')\) entry is the covariance \( \text{Cov}_B(y_{ij}, y_{ij'}) \) (assumed the same for all \( i \)) — and the analogue of the uniqueness result would be that the \( d \)-dimensional multivariate normal distribution \( N_d(\mu, \Sigma) \) would be uniquely identified in the Gaussian family by \( \mu \) and \( \Sigma \).

- The analogue of the standard normal distribution \( N(0, 1) \) would be \( N_d(0, I) \) where \( 0 \) is a vector of zeros of length \( d \) and \( I \) is the \((d \times d)\) identity matrix.

- The analogue of converting to standard units — \( y \sim N(\mu, \sigma^2) \rightarrow \frac{y - \mu}{\sigma} = (y - \mu)(\sigma^2)^{-\frac{1}{2}} \sim N(0, 1) \) — looks like it would involve raising the covariance matrix \( \Sigma \) to the power \(-\frac{1}{2}\); to get the dimensions right in the matrix multiplication, this would have to be
Multivariate Normal Model

\[ y \sim N_d(\mu, \Sigma) \rightarrow \Sigma^{-\frac{1}{2}}(y - \mu) \sim N_d(0, I); \] it turns out that this requires that the symmetric matrix \( \Sigma \) be invertible (= of full rank \( d = \) non-singular), in which case all of its eigenvalues \( \lambda_1 \geq \ldots \geq \lambda_d \) are positive, in which case by the Spectral Decomposition Theorem we can write \( \Sigma = \Gamma \Lambda \Gamma' \), where \( \Lambda = \text{diag}(\lambda_i) \) and \( \Gamma \) is an orthogonal matrix whose \( i \)th column is the standardized eigenvector of \( \Sigma \) corresponding to \( \lambda_i \), in which case \( \Sigma^{-\frac{1}{2}} \) can be defined to be \( \Gamma \Lambda^{-\frac{1}{2}} \Gamma' \).

- The univariate normal density is

\[
\begin{align*}
(y|\mu, \sigma^2) & \sim N(\mu, \sigma^2) \rightarrow \\
p(y|\mu, \sigma^2) & = (2\pi\sigma^2)^{-\frac{1}{2}} \exp \left[ -\frac{1}{2} \left( \frac{y - \mu}{\sigma} \right)^2 \right] \\
& = (2\pi\sigma^2)^{-\frac{1}{2}} \exp \left[ -\frac{1}{2}(y - \mu)(\sigma^2)^{-1}(y - \mu) \right];
\end{align*}
\] (15)

the multivariate generalization of this turns out to be

\[
\begin{align*}
(y|\mu, \Sigma) & \sim N_d(\mu, \Sigma) \rightarrow \\
p(y|\mu, \Sigma) & = |2\pi\Sigma|^{-\frac{1}{2}} \exp \left[ -\frac{1}{2}(y - \mu)'\Sigma^{-1}(y - \mu) \right],
\end{align*}
\] (16)

where \( |A| \) is the determinant of the square matrix \( A \) (a measure of its size).

- The level curves of this sampling distribution (curves of constant density in \( y \)) are evidently of the form 

\( (y - \mu)'\Sigma^{-1}(y - \mu) = c \), which are ellipsoids; this is consistent with the elliptical shape for bivariate-normal scatterplots.

- As GCSR note in their Appendix A, the simplest way to draw a random vector \( y \) from \( N_d(\mu, \Sigma) \) is to (a) compute the Cholesky decomposition of \( \Sigma = AA' \) (named after André-Louis Cholesky (1875–1918), a French mathematician and military officer), where \( A \) is lower triangular, (b) generate \( d \) IID values \( z \sim N(0, 1) \) and (c) set \( y = \mu + Az \).
Conjugate Prior for $N_d(\mu, \Sigma)$

With a sample $(y_i|\mu, \Sigma) \sim \mathcal{N}_d(\mu, \Sigma)$ ($i = 1, \ldots, n$), the likelihood function is

$$l(\mu, \Sigma|y) = c p(y|\mu, \Sigma)$$

$$= c \prod_{i=1}^{n} |2\pi \Sigma|^{-\frac{1}{2}} \exp \left[ -\frac{1}{2} (y_i - \mu)' \Sigma^{-1} (y_i - \mu) \right]$$

$$= c |\Sigma|^{-\frac{n}{2}} \exp \left[ -\frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)' \Sigma^{-1} (y_i - \mu) \right],$$

and this can be shown to equal

$$l(\mu, \Sigma|y) = c |\Sigma|^{-\frac{n}{2}} \exp \left\{ -\frac{1}{2} \text{tr} [\Sigma^{-1} S_0(\mu)] \right\}, \quad \text{where}$$

$$S_0(\mu) = \sum_{i=1}^{n} (y_i - \mu)(y_i - \mu)'.$$  \hspace{1cm} (17)

If there’s a conjugate prior for this likelihood, what is it?

Earlier we saw that with the univariate Gaussian sampling model the conjugate prior, expressed hierarchically, is

$$\sigma^2 \sim \chi^{-2}(\nu_0, \sigma_0^2)$$  \hspace{1cm} (18)

$$(\mu|\sigma^2) \sim \mathcal{N} \left( \mu_0, \frac{\sigma^2}{\kappa_0} \right).$$

If this is going to generalize, the second line should evidently become

$$(\mu|\Sigma) \sim \mathcal{N}_d \left( \mu_0, \frac{1}{\kappa_0} \Sigma \right), \quad \text{ (19)}$$

but how does the prior on $\sigma^2$ generalize?

It looks like it should be of the form

$$\Sigma \sim D(\nu_0, \Sigma_0),$$  \hspace{1cm} (20)

for some distribution $D$ on the $(d \times d)$ covariance matrix $\Sigma$ that’s a multivariate generalization of the $\chi^{-2}$ distribution; how to proceed?
Conjugate Prior for $N_d(\mu, \Sigma)$

This sounds like a tall order: $\Sigma$ is supposed to be not only invertible but also positive definite, meaning that $y'\Sigma y > 0$ for all $(d \times 1)$ vectors $y$; how can you create a density that's guaranteed to generate samples that are positive-definite invertible matrices?

The guy who solved this problem was John Wishart (1898–1956), a British statistician who worked at Rothamsted and Cambridge; he worked in the frequentist paradigm, and he asked the following question: if my sampling model is $(y_i|\mu, \Sigma) \overset{\text{IID}}{\sim} N_d(\mu, \Sigma)$ ($i = 1, \ldots, n$), what are good frequentist estimates of $\mu$ and $\Sigma$, and what are their repeated-sampling distributions?

The usual answers with $d = 1$ are of course

$$\hat{\mu} = \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \quad \text{and} \quad s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2;$$

[Q: how do these generalize?]

**A:** Directly: $\hat{\mu} = \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ requires no change at all; with $y_i$ a $(d \times 1)$ vector, $\bar{y}$ is the $(d \times 1)$ vector whose $j$th entry is the mean of variable $j$; and the obvious thing to try to estimate the covariance matrix that generalizes $s^2$ is the $(d \times d)$ matrix

$$\hat{\Sigma} = S = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})(y_i - \bar{y})'.$$  \hspace{1cm} (21)

As for the repeated-sampling distributions, with $d = 1$ the answers are

$$\bar{y} \sim_{\text{RS}} N\left(\mu, \frac{1}{n} \sigma^2\right) \quad \text{and} \quad (n - 1)s^2 \sim_{\text{RS}} \sigma^2 \chi^2_{n-1};$$  \hspace{1cm} (22)

the obvious analogue for $\bar{y}$ with $d > 1$ is $\bar{y} \sim_{\text{RS}} N\left(\mu, \frac{1}{n} \Sigma\right)$, and this turns out to be correct, but what about $s^2$?
Wishart Distribution

Wishart knew that if \((y_i|\mu, \sigma^2) \overset{\text{IID}}{\sim} N(\mu, \sigma^2)\) \((i = 1, \ldots, n)\) then

\[(n - 1) s^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 \sim_{\text{RS}} \sigma^2 \chi^2_{n-1}; \quad (23)\]

to make this even simpler he got rid of \(\mu\) and \(\bar{y}\) it's also true that if \((y_i|\sigma^2) \overset{\text{IID}}{\sim} N(0, \sigma^2)\) \((i = 1, \ldots, n)\) then

\[\sum_{i=1}^{n} y_i^2 = y' y \sim_{\text{RS}} \sigma^2 \chi^2_n. \quad (24)\]

In our multivariate situation the univariate data vector \(y\) is replaced by the \((n \times d)\) data matrix \(Y\), and (as Wishart was the first to notice) the obvious thing to try is the matrix product \(Y'Y\), leading to the following

**Definition:** If \(Y (n \times d)\) with \(n \geq d\) has rows \(y_i\) with sampling distribution \((y_i|\Sigma) \overset{\text{IID}}{\sim} N_d(0, \Sigma)\) \((i = 1, \ldots, n)\), where \(\Sigma\) is positive definite (and therefore automatically invertible), then \(W = Y'Y\) has a Wishart distribution with scale matrix \(\Sigma\) and degrees of freedom \(n\): \(W \sim W_d(n, \Sigma)\) (GCSR call this the Wishart_{\(n\)}(\(\Sigma\)) distribution) (this construction ensures that draws from \(W_d(n, \Sigma)\) are invertible, positive-definite matrices).

This distribution has a hideous density that's never used for sampling (e.g., you can use the above definition to draw samples): if \(W \sim W_d(\nu, \Sigma)\) then

\[p(W) = \left|\Sigma\right|^{-\nu/2} |W|^{-\nu/d-1/2} \exp \left[ -\frac{1}{2} \text{tr}(\Sigma^{-1}W) \right] \frac{\nu d^{d/2} \pi^{d(d-1)/4}}{2^{\frac{3d}{2}} \pi \prod_{i=1}^{d} \Gamma \left( \frac{\nu + d - i}{2} \right)} ; \quad (25)\]

this distribution has mean

\[W \sim W_d(\nu, \Sigma) \rightarrow E_{RS}(W) = E_{B}(W) = \nu \Sigma. \quad (26)\]
Inverse Wishart Distribution

So now we have a **multivariate (matrix) generalization** of the $\chi^2$ distribution; for the **conjugate prior** for the $N_d(\mu, \Sigma)$ **sampling model** we wanted a multivariate generalization of the $\chi^{-2}$ distribution, which is provided by the

**Definition:** If $Y \ (n \times d)$ with $n \geq d$ has rows $y_i$ with sampling distribution $(y_i | \Sigma) \sim \text{IID } N_d(0, \Sigma) \ (i = 1, \ldots, n)$, where $\Sigma$ is **positive definite** (and therefore automatically invertible), then $W = Y'Y$ has a **Wishart** distribution with scale matrix $\Sigma$ and **degrees of freedom** $n$ — $W \sim W_d(n, \Sigma)$ — and $W^{-1}$ has an inverse Wishart distribution — $M = W^{-1} \sim W_d^{-1}(n, \Sigma^{-1})$ (GCSR call this the Inv-Wishart$_n(\Sigma^{-1})$ distribution).

This distribution has another hideous density (see GCSR Appendix A) that's also **never used for sampling** (again, just use the definition); in this case

$$W \sim W_d^{-1}(\nu, \Sigma^{-1}) \rightarrow E_{RS}(W) = E_B(W) = \frac{1}{\nu - d + 1} \Sigma. \quad (27)$$

**Warning:** Various books and software packages use various parameterizations for the Wishart and inverse Wishart distributions, and it's easy to get confused; often you can sort things out by keeping track of the **mean relations** (26) and (27).

**NB:** You can readily **write your own code** to draw from the Wishart and inverse Wishart distributions, but they're also available in the MCMCpack library in R.

We're now finally ready to (hierarchically) specify the conjugate prior in the $N_d(\mu, \Sigma)$ sampling model:

$$\Sigma \sim W_d^{-1}(\nu_0, \Lambda_0^{-1}) \quad (28)$$

$$(\mu | \Sigma) \sim N_d\left(\mu_0, \frac{1}{\kappa_0} \Sigma\right).$$
Conjugate Updating with $N_d(\mu, \Sigma)$

GCSR call this the Normal-Inverse-Wishart$(\mu_0, \frac{1}{\kappa_0} \Lambda_0; \nu_0, \Lambda_0)$ distribution; in direct analogy with the univariate case, this prior acts like

(a) a data set for $\Sigma$ with prior sample size $\nu_0$ and prior estimate $\nu_0 \Lambda_0$, and

(b) a data set for $\mu$ with prior sample size $\kappa_0$ and prior estimate $\mu_0$.

With this prior, having observed $(y_i|\mu, \Sigma)^{\text{IID}} N_d(\mu, \Sigma)$ $(i = 1, \ldots, n)$, the posterior is Normal-Inverse-Wishart$(\mu_n, \frac{1}{\kappa_n} \Lambda_n; \nu_n, \Lambda_n)$; the updating rules are

\[
\begin{align*}
\nu_n &= \nu_0 + n \\
\kappa_n &= \kappa_0 + n \\
\mu_n &= \frac{\kappa_0 \mu_0 + n \bar{y}}{\kappa_0 + n} \\
\Lambda_n &= \Lambda_0 + S^* + \frac{\kappa_0 n}{\kappa_0 + n} (\bar{y} - \mu_0)(\bar{y} - \mu_0)',
\end{align*}
\]  

(29)
in which the sample mean vector $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ and the sums of squares and cross-products matrix $S^* = \sum_{i=1}^{n} (y_i - \bar{y})(y_i - \bar{y})'$ are sufficient for $\mu$ and $\Sigma$.

In this model, generalizing the earlier univariate findings, the marginal posterior for $\Sigma$ is $W_d^{-1}(\nu_n, \Lambda_n^{-1})$, and the marginal posterior for $\mu$ is

\[
(\mu|Y) \sim t_{\nu_n - d + 1} \left[ \mu_n, \frac{1}{\kappa_n(\nu_n - d + 1)} \Lambda_n \right],
\]

(30)

where $t_{\nu}(\mu, \Sigma)$ is the multivariate $t$ distribution with location vector $\mu$, scale matrix $\Sigma$, and degrees of freedom $\nu$ (see GCSR's Appendix A for details about this distribution).
Diffuse Priors

If a diffuse prior is scientifically indicated, evidently to achieve this in the conjugate family you can choose $\kappa_0$ small and positive and $\nu_0$ not much bigger than $d$ (this ensures that the prior is proper).

Another popular diffuse choice is the Jeffreys prior

$$ p(\mu, \Sigma) = c|\Sigma|^{-\frac{d+1}{2}}, \quad (31) $$

which arises as a formal limit of the conjugate prior as $\kappa_0 \to 0, \nu_0 \to -1,$ and $|\Lambda_0| \to 0.$

You can see this is an improper prior, but it will lead to the proper posterior

$$ (\Sigma|Y) \sim W_d^{-1}(n-1, S^*) \quad (32) $$

$$ (\mu|Y, \Sigma) \sim N_d\left(\bar{y}, \frac{1}{n} \Sigma\right) $$

as long as $n \geq d$; with this prior the marginal posterior for $\mu$ is

$$ (\mu|Y) \sim t_{n-d}\left[\bar{y}, \frac{1}{n(n-d)S^*}\right]. \quad (33) $$