# Module 7

## Some multivariate normal examples

Prof. Stephen B. Vardeman
Statistics and IMSE
Iowa State University

March 5, 2008

# Multivariate Analyses

It is common to have measurements of multiple types on a given item. For example, a test might produce emission levels of multiple pollutants for diesel engines, or torques produces by hydraulic motors under multiple operating conditions. It is then useful to be able to treat those measurements jointly and do inference for their joint properties.

# Example 6

A company uses a torque gun (set during a laboratory calibration session to deliver an average of 370 ft lb of torque) to tighten some bolts on an assembly line. The sensor used to measure torque in the lab can be employed at considerable inconvenience and expense to measure peak torques delivered tightening bolts on the line. Torque wrench A or B can later be used to loosen bolts and record the torque required to do so. Data from 51 bolts follow.

| Tighten | A | B |
|---|---|---|
| 343 | 463 | |
| 332 | 478 | |
| 365 | 378 | |
| 332 | 448 | |
| 353 | 465 | |
| 357 | 546 | |
| 346 | 473 | |
| 343 | 461 | |
| 378 | | 381 |
| 345 | | 372 |
| 354 | | 365 |
| 347 | | 362 |
| 351 | | 349 |

| Tighten | A | B |
|---|---|---|
| 340 | | 330 |
| 365 | | 365 |
| 345 | | 338 |
| 360 | | 382 |
| 357 | | 383 |
| 345 | | 376 |
| 347 | | 370 |
| | 495 | |
| | 509 | |
| | 521 | |
| | 529 | |
| | 568 | |
| | 554 | |

| Tighten | A | B |
|---|---|---|
| | 494 | |
| | 573 | |
| | 523 | |
| | 496 | |
| | 535 | |
| | 509 | |
| | 442 | |
| | 481 | |
| | 481 | |
| | 468 | |
| | | 333 |
| | | 336 |
| | | 331 |

| Tighten | A | B |
|---|---|---|
| | | 320 |
| | | 361 |
| | | 370 |
| | | 387 |
| | | 385 |
| | | 342 |
| | | 364 |
| | | 315 |
| | | 430 |
| | | 365 |
| | | 338 |
| | | 382 |
| | | |

# Example 6 (cont.)

Here there are at most 2 measurements for each bolt, a tightening peak torque and a loosening torque taken with one or the other of wrenches A or B. The tightening torque is the best available indication of how well a bolt is tightened, but can not be routinely obtained. Relationships between the 3 kinds of measurement are of interest. What we will do here will treat on the previous panel as $n = 51$ *incomplete* data vectors, each of length $k = 3$.

# Multivariate Normal Models

What is first needed here is a model for complete data vectors

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_k \end{pmatrix}$$

The standard (essentially only mathematically tractable) choice is the so-called *multivariate normal* model. To say

$$\mathbf{Y} \sim \mathrm{MVN}_k\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$$

is to say that it has a (joint) probability density

$$f\left(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \left(\det\boldsymbol{\Sigma}\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\left(\mathbf{y}-\boldsymbol{\mu}\right)' \boldsymbol{\Sigma}^{-1}\left(\mathbf{y}-\boldsymbol{\mu}\right)\right)$$

# Multivariate Normal Models (cont.)

The vector

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix}$$

contains the means of the coordinates of $\mathbf{Y}$, and the matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1k}\sigma_1\sigma_k \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \cdots & \rho_{2k}\sigma_2\sigma_k \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1k}\sigma_1\sigma_k & \rho_{2k}\sigma_2\sigma_k & \cdots & \sigma_k^2 \end{pmatrix}$$

is called the *covariance matrix* for $\mathbf{Y}$ and conveniently summarizes the standard deviations and correlations between the coordinates of $\mathbf{Y}$. (In our example, these are the mean torque readings, the standard deviations of torque readings, and the correlations between torque readings of various kinds.)

# Example 6 (cont.)

Notice that in our data set, there is direct information on $\mu_1, \mu_2, \mu_3, \sigma_1, \sigma_2, \sigma_3, \rho_{12}$, and $\rho_{13}$. Anything that can be known about $\rho_{23}$ (the correlation between wrench A and wrench B readings) will have to be of the most indirect nature, available only through their relationships to the measured tightening torque. Notice also, that by most standards, the data set here is most incomplete, and most standard (non-Bayes) analysis methods will simply not be applicable to understanding what the data say about the 3 variables.

# Bayes One-Sample Multivariate Normal Analyses

A likelihood based on $n$ data vectors $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ is

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{n} f(\mathbf{y}_i | \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

and in order to do a Bayes analysis, one must invent a (joint) prior distribution for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ (a mean vector and a *covariance matrix*)!

Standard choices for a prior for the mean vector are

$$\boldsymbol{\mu} \sim \text{MVN}_k(\mathbf{0}, \mathbf{W})$$

for some large covariance matrix $\mathbf{W}$ or

$$\mu_1, \mu_2, \ldots, \mu_k \text{ independent "Uniform}(-\infty, \infty)\text{" (i.e. "flat")}$$

The first of these provides a diffuse prior specification centered at $\mathbf{0}$ (actually, any mean $\mathbf{m}$ could be used). The second is the multivariate version of the flat prior for a univariate mean used in the previous module.

# Bayes One-Sample Multivariate Normal Analyses (cont.)

The choice of a prior for the covariance matrix $\mathbf{\Sigma}$ is far more subtle, as covariance matrices are much more special than one might naively expect (not just every choice of correlations between $-1$ and 1 will produce a valid covariance matrix). What is commonly done is based on the so-called Wishart distributions, and to get to the Bayes analysis here we must know some things about these.

# Wishart Distributions

If $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ are $n$ independent (complete) random vectors from the $\text{MVN}_k(\boldsymbol{\theta}, \boldsymbol{\Delta})$ distribution and $\mathbf{S}$ is their ($n-k$ divisor) sample covariance matrix

$$\begin{aligned}
\mathbf{S} &= \frac{1}{n-k} \sum_{i=1}^{n} (\mathbf{X}_i - \bar{\mathbf{X}}) (\mathbf{X}_i - \bar{\mathbf{X}})' \\
&= \frac{1}{n-k} \left( \sum_{i=1}^{n} (X_{il} - \bar{X}_l) (X_{im} - \bar{X}_m) \right)_{\substack{l=1,2,\ldots,k \\ m=1,2,\ldots,k}}
\end{aligned}$$

then $(n-k)\,\mathbf{S}$ has a so-called Wishart distribution with "degrees of freedom" $(n-k)$ and scale matrix $\boldsymbol{\Delta}$. (The mean of a Wishart distribution with degrees of freedom $\nu$ and scale matrix $\boldsymbol{\Delta}$ is $\nu\boldsymbol{\Delta}$.)

# Inverse-Wishart Distributions and Priors for Covariance Matrices

The inverse of a Wishart$(\nu, \mathbf{\Delta})$-distributed matrix is said to have an "Inv-Wishart$(\nu, \mathbf{\Delta})$" distribution, and the most convenient type of prior distribution in a Bayes analysis of multivariate normal data is to set *a priori*

$$\begin{aligned} \mathbf{\Sigma}^{-1} &\sim \text{Wishart}\left(\nu, \mathbf{\Delta}\right) \quad \text{i.e.} \\ \mathbf{\Sigma} &\sim \text{Inv-Wishart}\left(\nu, \mathbf{\Delta}\right) \end{aligned}$$

for appropriately chosen $\nu$ and $\mathbf{\Delta}$.

The parameter $\nu$ functions as a kind of "prior sample size." In some sense an inverse Wishart prior with parameter $\nu$ weights a prior mean for a covariance matrix about as strongly as the information provided by an additional $\nu$ observations. So, smaller $\nu$'s correspond to "flatter"/"less informative" priors. But there are technical problems related to the ability to invert matrices and existence of means if $\nu$ is too small, so it is common to take $\nu \geq k + 1$.

# Choice of a Scale Matrix for an Inverse-Wishart Prior

For a target *a priori* mean of $\mathbf{V}$ for $\mathbf{\Sigma}$, let

$$\mathbf{\Delta} = \frac{1}{\nu - k - 1}\mathbf{V}^{-1}$$

and take

$$\begin{aligned} \mathbf{\Sigma} &\sim \text{Inv-Wishart}\left(\nu, \mathbf{\Delta}\right) \text{ i.e.} \\ \mathbf{\Sigma}^{-1} &\sim \text{Wishart}\left(\nu, \mathbf{\Delta}\right) \end{aligned}$$

Because `WinBUGS` parameterizes its distributions with precisions instead of variances, to implement this in `WinBUGS` one must set

$$\begin{aligned} \mathbf{\Sigma}^{-1} &\sim \texttt{WinBUGS-Wishart}\left(\nu, \mathbf{\Delta}^{-1}\right) \text{ i.e.} \\ \mathbf{\Sigma}^{-1} &\sim \texttt{WinBUGS-Wishart}\left(\nu, \left(\nu - k - 1\right)\mathbf{V}\right) \end{aligned}$$

# Example 6 (cont.)

So we can finally return to Example 6 and the bolt torques. The file

`BayesASQEx6.odc`

contains `WinBUGS` code for doing a Bayes analysis for the torques. The choice of prior

$$\boldsymbol{\mu} \sim \mathrm{MVN}_3\left(\begin{pmatrix} 370 \\ 370 \\ 370 \end{pmatrix}, \begin{pmatrix} 10^6 & 0 & 0 \\ 0 & 10^6 & 0 \\ 0 & 0 & 10^6 \end{pmatrix}\right)$$

was made. This makes the means *a priori* independent normal with means equal to the lab setting of the torque gun and standard deviations of 1000 $\mathrm{ft\,lb}$. This should be a fairly non-informative specification.

# Example 6 (cont.)

Then, the choice was made to use an inverse Wishart prior for $\mathbf{\Sigma}$ with $\nu = 5$ and target *a priori* mean

$$\mathbf{V} = \begin{pmatrix} 2000 & 0 & 0 \\ 0 & 2000 & 0 \\ 0 & 0 & 2000 \end{pmatrix}$$

Note that if $\mathbf{\Sigma}$ were to be equal to its prior mean, torque measurements from the 3 devices would be independent with standard deviation about 45 $\mathrm{ft\,lb}$. One could argue against independence of the measurements. If anything really varies, they should be positively correlated. But this kind of mean covariance matrix reflects a prior belief that in fact all that one is seeing in the data is independent measurement noise. The 45 is "about the right order of magnitude" and really came about from some "playing around to find something sensible." The analysis is fairly sensitive to this choice and diagonal elements of $\mathbf{V}$ too small or too large seem to overwhelm the sample standard deviations from the data.

# Example 6 Code

```
model
    {
        for(i in 1:51)
        {
            Y[i, 1:3] ~dmnorm(mu[], R[ , ])
        }
        mu[1:3] ~dmnorm(alpha[],Tau[ , ])
        #remember that WinBUGS uses precisions
        #for specifying its normal distributions
        R[1:3 , 1:3] ~dwish(Lambda[ , ], nu)
        D[1:3, 1:3]<-inverse(R[1:3, 1:3])
        sig1<-sqrt(D[1,1])
        sig2<-sqrt(D[2,2])
        sig3<-sqrt(D[3,3])
        rho12<-D[1,2]/(sig1*sig2)
        rho13<-D[1,3]/(sig1*sig3)
        rho23<-D[2,3]/(sig2*sig3)
```

# Example 6 Code (cont.)

```
        diff21<-mu[2]-mu[1]
        diff31<-mu[3]-mu[1]
    }

list(nu=5, alpha=c(370,370,370),
Tau = structure(.Data = c(0.000001, 0, 0, 0, 0.000001,
0, 0, 0, 0.000001), .Dim = c(3, 3)),
Lambda = structure(.Data = c(2000, 0, 0, 0, 2000,
0, 0, 0, 2000), .Dim = c(3, 3)),

 Y = structure(.Data = c(
                         343,463,NA,
                         332,478,NA,
                         365,378,NA,
                         332,448,NA,
```

# Example 6 Code (cont.)

```
                    353,465,NA,
                    357,546,NA
.
.
.
                    NA,NA,430,
                    NA,NA,365,
                    NA,NA,338,
                    NA,NA,382), .Dim = c(51, 3)))
```

Notice that `WinBUGS` was quite willing to let me treat the missing parts of the 51 data vectors as exactly that. I ran 3 parallel chains, using starting values generated by `WinBUGS` and found the chains had apparently burned in by the end of 1000 iterations. The following figures summarize the results of a large number of iterations after burn in.

# Example 6 (cont.)

Node statistics

| node | mean | sd | MC error | 2.5% | median | 97.5% | start | sample |
|---|---|---|---|---|---|---|---|---|
| D[1,1] | 228.5 | 78.9 | 0.4617 | 122.0 | 213.2 | 423.0 | 61001 | 150000 |
| D[1,2] | 6.835 | 181.7 | 1.788 | -363.6 | 9.305 | 365.5 | 61001 | 150000 |
| D[1,3] | 110.5 | 113.6 | 0.8264 | -105.9 | 106.4 | 348.7 | 61001 | 150000 |
| D[2,1] | 6.835 | 181.7 | 1.788 | -363.6 | 9.305 | 365.5 | 61001 | 150000 |
| D[2,2] | 1943.0 | 572.3 | 2.95 | 1126.0 | 1843.0 | 3333.0 | 61001 | 150000 |
| D[2,3] | 4.861 | 620.5 | 11.17 | -1097.0 | 5.921 | 1100.0 | 61001 | 150000 |
| D[3,1] | 110.5 | 113.6 | 0.8264 | -105.9 | 106.4 | 348.7 | 61001 | 150000 |
| D[3,2] | 4.861 | 620.5 | 11.17 | -1097.0 | 5.921 | 1100.0 | 61001 | 150000 |
| D[3,3] | 702.7 | 197.9 | 1.063 | 415.9 | 669.6 | 1180.0 | 61001 | 150000 |
| diff21 | 145.1 | 9.56 | 0.0567 | 126.3 | 145.1 | 163.9 | 61001 | 150000 |
| diff31 | 10.01 | 5.632 | 0.0341 | -1.056 | 9.987 | 21.2 | 61001 | 150000 |
| mu[1] | 350.0 | 3.484 | 0.02393 | 343.0 | 350.0 | 356.8 | 61001 | 150000 |
| mu[2] | 495.0 | 9.058 | 0.05606 | 477.2 | 495.0 | 512.9 | 61001 | 150000 |
| mu[3] | 360.0 | 5.079 | 0.02996 | 349.9 | 360.0 | 370.0 | 61001 | 150000 |
| rho12 | 0.01092 | 0.2529 | 0.002574 | -0.4813 | 0.0153 | 0.4821 | 61001 | 150000 |
| rho13 | 0.2723 | 0.249 | 0.001868 | -0.262 | 0.2956 | 0.6862 | 61001 | 150000 |
| rho23 | 0.003574 | 0.5105 | 0.009519 | -0.8102 | 0.00554 | 0.809 | 61001 | 150000 |
| sig1 | 14.92 | 2.452 | 0.01418 | 11.04 | 14.6 | 20.57 | 61001 | 150000 |
| sig2 | 43.65 | 6.192 | 0.03181 | 33.56 | 42.94 | 57.73 | 61001 | 150000 |
| sig3 | 26.27 | 3.579 | 0.01917 | 20.39 | 25.88 | 34.35 | 61001 | 150000 |

Figure: Summary of a Bayes analysis of the torque data

# Example 6 (cont.)

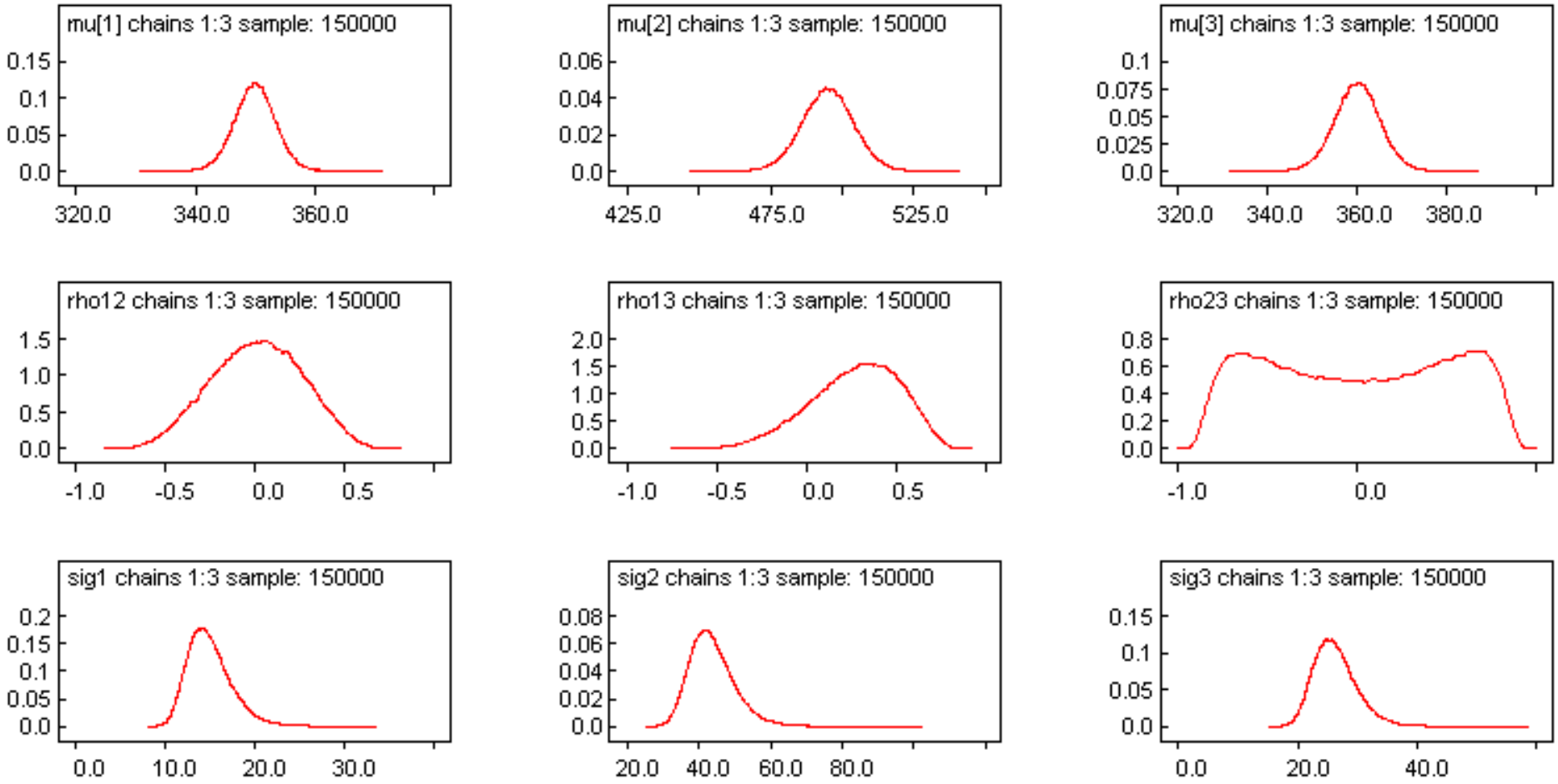



Figure: Approximate posteiors of means, standard deviations and correlations for the torque problem

# Example 6 (cont.)

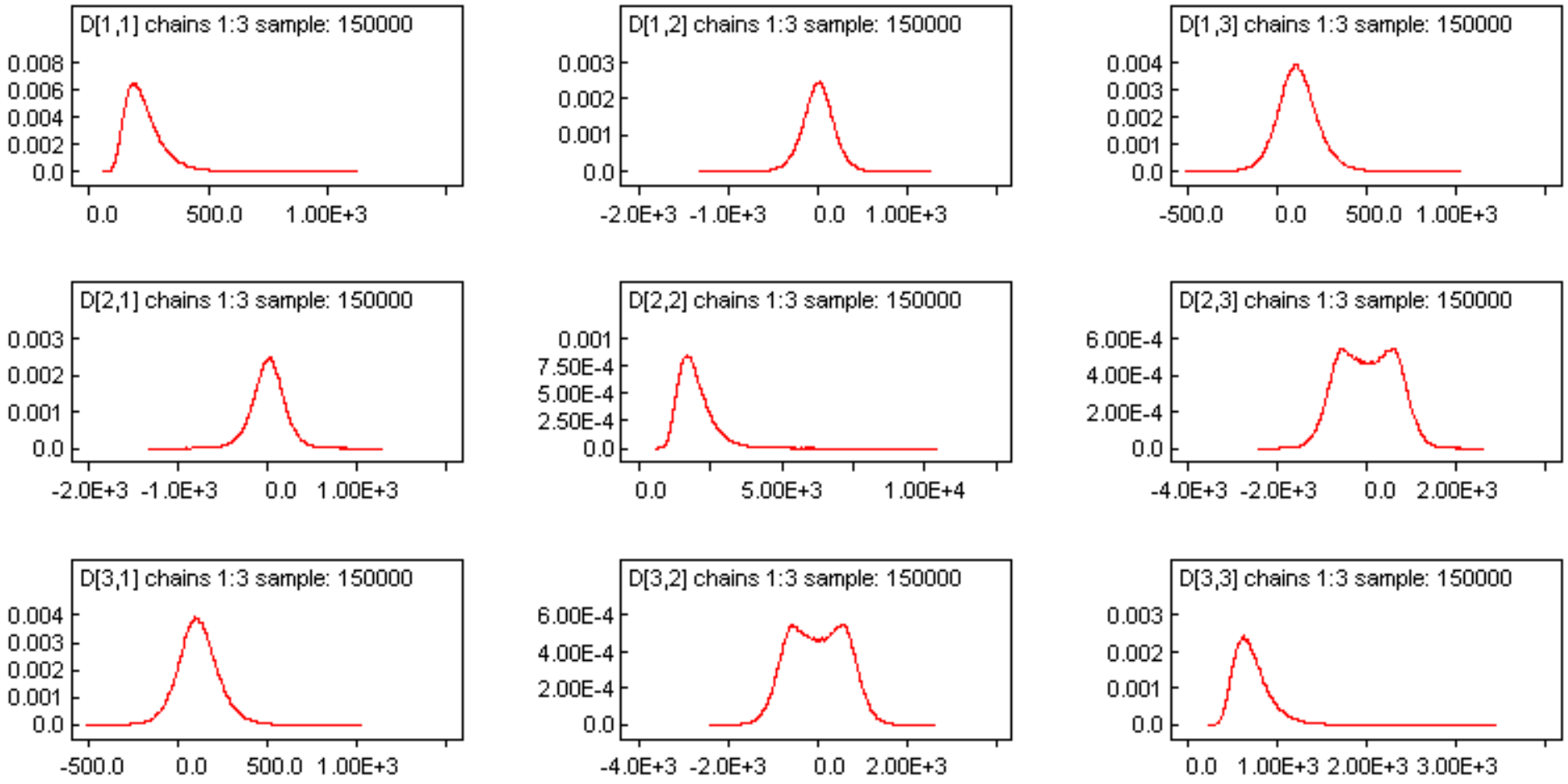



Figure: Approximate posterior distributions of the elements of the covariance matrix for the torque problem

# Example 6 (cont.)

Here is a further summary of some 95% credible intervals from this analysis.

**Table 1** 95% Credible Intervals for 3 Types of Torque Measurement

| | $\mu$ | $\sigma$ |
|---|---|---|
| Tightening Torque | $(343, 357)$ | $(11.04, 20.57)$ |
| Wrench A Loosening Torque | $(477, 513)$ | $(33.56, 57.73)$ |
| Wrench B Loosening Torque | $(350, 371)$ | $(20.39, 34.35)$ |

And here are the 95% credible intervals for the 3 correlations.

**Table 2** 95% Credible Intervals for the Correlations Between Measurement Types

| | Wrench A | Wrench B |
|---|---|---|
| Tightening | $(-.48, .48)$ | $(-.26, .69)$ |
| Wrench A | | $(-.81, .81)$ |

# Example 6 (cont.)

Why measured peak tightening torque on bolts at assembly is roughly 20 ft lb less than what the gun was set for in the lab is unknown. The measured loosening torques are pretty clearly systematically different from the measured tightening torque and each other. Measured tightening torque is clearly more consistent than either loosening torque. None of the correlations are definitively non-zero, and not surprisingly, essentially nothing is known about the unobservable correlation between wrench A and wrench B measurements. This "no definitive correlations" picture would be a very unhappy one if there was reason to believe that in fact there were real process differences across the 51 bolts (since it would say that the loosening torques are no reliable indicator of what the varying tightening torques were). But if in fact there are none, this correlation picture is consistent with "nothing but pure measurement noise" view of the data.

# Example 6 (cont.): Tomorrow's Prior

Note that should one decide to in the future collect and analyze more data of the present type, the posteriors from this analysis suggest suitable parameters for future prior distributions. A future prior mean for $\boldsymbol{\mu}$ of

$$\begin{pmatrix} 350 \\ 495 \\ 360 \end{pmatrix}$$

and future prior covariance matrix for $\boldsymbol{\mu}$ like

$$\begin{pmatrix} (3.5)^2 & 0 & 0 \\ 0 & (9.1)^2 & 0 \\ 0 & 0 & (5.1)^2 \end{pmatrix}$$

would be informative and consistent with the current posterior.

# Example 6 (cont.): Tomorrow's Prior

Further, since the posterior mean of $\mathbf{\Sigma}$ here is (using the means of the values "$D[i,j]$" from the report)

$$\mathbf{D} = \begin{pmatrix} 228.5 & 6.835 & 110.5 \\ 6.835 & 1943.0 & 4.861 \\ 110.5 & 4.861 & 702.7 \end{pmatrix}$$

upon choosing some degrees of freedom $\nu$ to represent the weight of the evidence provided by these data about $\mathbf{\Sigma}$ (because of the very fragmented/sparse nature of the data, I would not be inclined to use $\nu$ more than 4 or 5) an appropriate prior assignment for a future analysis might be

$$\mathbf{\Sigma}^{-1} \sim \texttt{WinBUGS-Wishart}\left(\nu, (\nu - 3 - 1)\,\mathbf{D}\right)$$

as this will make $\mathbf{D}$ the prior mean of $\mathbf{\Sigma}$.

# Example 6B

As a final variation on this example, suppose that the company decides that

- a tightening torque of at least 320 is essential, and
- since it is impractical to measure tightening torque, wrench B will be used to spot check bolts and any bolt whose wrench B loosening torque is at least 320 will be considered "OK."

An interesting function of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is then

$$p(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \text{the fraction of the current measurements that have wrench B loosening torque above 320 and tightening torque below 320}$$

This is not a simple function of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, but can be approximated inside a `WinBUGS` simulation by at each $(\boldsymbol{\mu},\boldsymbol{\Sigma})$ iteration simulating a fairly large number of fictitious observations from that set of parameters and simply counting the fraction that satisfy the above condition.

# Example 6B (cont.)

The file

BayesASQEx6B.odc

contains a slight revision of the earlier WinBUGS code that adds this feature. The new part of the code is

```
#below is some new code for approximating the
#fraction of the Y distribution for which Y[1]<320
#while Y[3]>320

    for (j in 1:5000) {
    Z[j,1:3] ~dmnorm(mu[], R[ , ])
    bad[j]<-step(320 - Z[j,1])*step(Z[j,3]-320)
    }
    prob<-sum(bad[1:5000])/5000
```

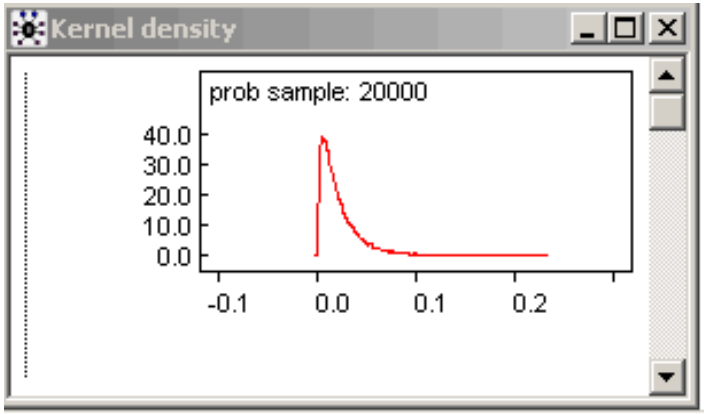
.

.

.

# Example 6B (cont.)

Below is a summary of a use of this code. It indicates that one can be fairly sure that the fraction of the current distribution of measurements with wrench B loosening torque above 320 and tightening torque below 320 is small.




Node statistics

| node | mean | sd | MC error | 2.5% | median | 97.5% | start | sample |
|---|---|---|---|---|---|---|---|---|
| prob | 0.02158 | 0.02092 | 3.396E-4 | 0.0014 | 0.0152 | 0.0784 | 1001 | 20000 |

In other contexts, this kind of trick could be used to make estimates of the fraction of a joint distribution that has all dimensions of a part inside their respective engineering specifications, or all emission levels of an engine below their respective EPA guidelines.