Statistics 335 – Assignment 2
Due: Thursday, November 6, 2003

1. This first question will investigate two possible ways of simulating random vectors from a multivariate normal distribution. The multivariate normal distribution is defined by two parameters, mu, a vector of length p, and Σ, the $p \times p$, variance-covariance matrix. The density of this distribution is

$$f(\mathbf{x}|\mu, \sigma) = (2\pi)^{-p/2}|\Sigma|^{-p/2} \exp(-0.5(\mathbf{x} - \mu)^t \Sigma^{-1}(\mathbf{x} - \mu))$$

While one of the most important distributions in statistics, there is no random number generator built into S-Plus or R for this distribution. What is desired is a function that will generate realizations from this distribution where the call to the function `mrnorm(n, mu, sigma)` will return a $n \times p$ matrix, where each row is a realization from the p dimensional multivariate normal.

(a) Matrix approach

Suppose that $\mathbf{z} = [z_1 z_2 \ldots z_p]^t$ is a vector of p independent standard normal random variables. Then $\mu + R^t\mathbf{z}$ is a realization from $N(\mu, \Sigma)$ where R is a matrix satisfying $R^t R = \Sigma$. The are number of ways of producing R with the most common based on the Choleski decomposition or the eigenvalue/eigenvector decomposition. R for the Choleski decomposition can gotten with `R = chol(Sigma)`. Note that the form of the matrix R doesn't matter for the distributional result to hold.

Write the function for generating from the multivariate normal based on this Choleski decomposition idea.

(b) Gibbs sampler

The Gibbs sampler is an approach that allows one to sample from complicated distributions. The Gibbs sampler, and Markov Chain Monte Carlo (MCMC) methods in general have opened up many areas of statistics, for example Bayesian statistics, and have made them tractible.

Suppose you wanted to generate samples from the joint density $f(x, y, z)$, but f is complicated. A scheme that will generate (dependent) samples (asymptotically) is

```
initialize x, y, and z as x(0), y(0), and z(0)
for i = 1 to n {
  draw x(i) from f(x|y(i-1), z(i-1))
  draw y(i) from f(y|x(i), z(i-1))
  draw z(i) from f(z|x(i), y(i))
}
```

The realizations `(x(i), y(i), z(i))` form a Markov Chain with a stationary distribution having density $f(x, y, z)$. Note that the realizations from the chain are not initially from the desired distribution, as the result is asymptotic. Due to this fact, the initial part of the chain is usually thrown away (known as burn-in). Also the realizations are dependent, as the the realization at step i depends on the previous realizations. The Matrix approach mentioned above does not have this problem.

Write a function for implementing the Gibbs sampler to draw from a bivariate normal (assume p is 2). As part of the function, allow for a burn-in figure to be given, and for the user to be able to specify the starting state of the chain. However for the starting state, have the mean of the distribution be used as the default. The conditional distributions you need to implement this sampler are

$$X|Y = y \sim N(\mu_x + \frac{\sigma_{xy}}{\sigma_y^2}(y - \mu_y), \sigma_x^2 - \frac{\sigma_{xy}^2}{\sigma_y^2})$$

$$Y|X = x \sim N(\mu_y + \frac{\sigma_{xy}}{\sigma_x^2}(x - \mu_x), \sigma_y^2 - \frac{\sigma_{xy}^2}{\sigma_x^2})$$

where σ_x^2 and σ_y^2 are the variances of the two components and σ_{xy} is the covariance.

(c) Generate 1000 realizations with both functions, with `mu = c(10,5)` and `Sigma = matrix(c(10,5,5,5),ncol=2)` (this corresponds to the correlation ρ being 0.707. For the Gibbs sampler sample, use a burnin of 100 iterations.

(d) Calculate the mean and variance-covariance matrices for both samples to see if they are close to the desired values.

(e) Calulate the lag one autocorrations for both variables for both samples. You can do this along the lines of `cor(mvn[-1,1], mvn[-1000,1])`, where `mvn` is a matrix containing the draws from one of the functions. Also for both samples, plot x_i against x_{i-1} and y_i against y_{i-1}.

(f) Generate histograms for both variables and both samplers, superimposing the normal density curves with the matching mean and variance and a kernel density estimate. Is there any evidence or non-normality in any of the plots.

2. Data Analysis

Wisconsin Power and Light studied the effectiveness of two devices for improving the efficiency of gas home-heating systems. The electric vent damper (EVD) reduces heat loss through the chimney when the furnace is in its off cycle by closing off the vent. It is controlled electrically. The thermally activated vent damper (TVD) is the same as the EVD, except that it is controlled by the thermal properties of a set of bimetal fins set in the vent. Ninety test houses were used, 40 with TVDs and 50 with EVDs. For each house, energy consumption was measured for a period of several weeks with the vent damper active and for a period with the damper not active. This should help show how effective the vent damper is in each house. For further information on the data set and an electronic version of the data file, see Assignments web page.

(a) Plot the relationship between BTUIn and BTUOut for the different combinations of CHShape and Damper. Is there any evidence from the plot that relationship between BTUIn and BTUOut is different for some combinations of CHShape and Damper?

(b) Generate the scatterplot matrix for all of the continuous variables in the dataset (Omit the variables that should be treated as factors). Are there any interesting relationships suggested by the plot?

(c) Fit the linear model predicting BTUIn by BTUOut, CHShape, and Damper where all two-way interactions are included in the model and give a summary of the fitting results. Does this output agree with the conclusions you made in part (a)?

(d) Fit the additive model predicting BTUIn by BTUOut and Damper. Is there any evidence that the EVD type of damper reduces energy consumption after accounting for BTUOut? What is the estimated difference between the different dampers?