I confirm that this submitted assignment wholly represents my own work. I have neither received nor given assistance on the completion of this MidTerm assessment.

Signature: ______________________   Date: _____________
1. Revisit the SOI data for one last look at AR(1)-like analyses. We know there is a great deal of variation across the years in this data that our AR(1) model does not capture. To investigate whether this may be due to structural change we can simply fit the AR(1) model to shorter sections of the data and examine the estimated parameters to see if they seem to vary across time. Do this: Take lag = 100 months; for each month $t = 101 : 440$, repeatedly fit the AR(1) model to the 201 data points centered at month $t$, saving the estimates of $\phi$ which we can now regard as being estimated “locally” around each month using a “windowed” data set of width $\pm 100$ months. Plot these estimates of $\phi$ over time and comment on what you see in the plot. What does this suggest for more general models that might do a better job of imitating this data?

2. Observed measurements on mRNA expression levels of three related genes are denoted by $y$, $x$ and $z$; a sample of size $n = 150$ such measures from a set of cancer cell lines is displayed in Figure 1. The data are standardised to zero mean and unit variance, and there is clear evidence of high, positive correlations (sample correlations in the 0.7-0.9 range) between each pair of genes. The plots are apparently consistent with pairwise normality and a trivariate normal distribution, though of course may be consistent with other distributional models.

![Scatter plots](image)

Figure 1: Scatter plots 150 samples of gene expression measures on three related genes.

Biologically, there is some interest in the general question of whether the positive associations between $y$ and each of $x$ and $z$ represent separate, though related relationships, or whether perhaps $x$ and $z$ are representing one common underlying phenomenon.

Consider three potential distributions for this data: in each of these three models, $(y, x, z)' \sim N(0, \Sigma)$ and the models differ only through the values of the variances. They are:

- **Model A:**
  \[
  \Sigma_A = \begin{pmatrix}
  1 & 0.8 & 0.8 \\
  0.8 & 1 & 0.8 \\
  0.8 & 0.8 & 1
  \end{pmatrix}
  \]

- **Model B:**
  \[
  \Sigma_B = \begin{pmatrix}
  1 & 0.9 & 0.72 \\
  0.9 & 1 & 0.8 \\
  0.72 & 0.8 & 1
  \end{pmatrix}
  \]

- **Model C:**
  \[
  \Sigma_C = \begin{pmatrix}
  1 & 0.72 & 0.9 \\
  0.72 & 1 & 0.8 \\
  0.9 & 0.8 & 1
  \end{pmatrix}
  \]
(a) In each of these models, find the precision matrix of the trivariate distribution.
(b) What is the implied conditional distribution for \((y|x, z)\) in each case?
(c) Discuss the interpretation of each of these three models, and comment on how they reflect different potential associations between the three variables. You may find graphical representation of the distributions useful in explaining your reasoning.
(d) Comment briefly on how the data might be inspected at a simple exploratory level to explore which of these models might be most relevant for this data.

A fourth distributional model is now suggested:

Model D: \((y, x, z)' \sim N(0, \Sigma_B)\) with probability 0.5, and otherwise \((y, x, z)' \sim N(0, \Sigma_C)\). That is, 
p(y, x, z) is a discrete mixture of two normals,
\[
p(y, x, z) = \frac{1}{2} N(0, \Sigma_B) + \frac{1}{2} N(0, \Sigma_C).
\]

(e) Generate multiple repeat samples of size \(n = 150\) from this distribution and inspect scatter plots, sample variance matrices and resulting correlations. Based on this exploration, discuss how this model compares with the real data, and also compare it as a potential alternative to any of the Models A, B and C in how it reflects the appearance of the real data in the figure.

(f) What is \(p(x, z)\) in Model D?

(g) What is \(p(y|x, z)\) in Model D?

(h) Discuss the differences in interpretation of Model D compared to Model A, and the potential relevance of these differences in a biological context. As part of this, consider a context in which, based on the observed levels of activity of genes \(x\) and \(z\), you want to then predict \(y\).

(i) Briefly comment on how you might investigate which of the four models best represents the data from a more formal statistical viewpoint, and why this might be relevant in view of your comments in response to part 2h above.

3. Two stationary, univariate AR(1) processes are driven by correlated innovation sequences. That is, we observe two processes \(y_t\) and \(z_t\) where

\[
y_t = \phi y_{t-1} + \epsilon_t \quad \epsilon_t \sim N(0, v),
\]
\[
z_t = \gamma z_{t-1} + \eta_t \quad \eta_t \sim N(0, w),
\]

where, as usual, \(\epsilon_t \perp \perp \epsilon_s\) and \(\eta_t \perp \perp \eta_s\). However, the innovations of the two processes are cross-correlated at time \(t\) - the two series are subject to a common influence. In particular, the vector \((\epsilon_t, \eta_t)\)' has a bivariate normal distribution with correlation \(\rho > 0\).

Write \(L\) for the Cholesky component of the variance matrix of \((\epsilon_t, \eta_t)\)' and define the 2—vector \(x_t\) by

\[
x_t = L^{-1} \begin{pmatrix} y_t \\ z_t \end{pmatrix}.
\]

- Show that \(x_t = \Phi x_{t-1} + \alpha_t\) where \(\alpha_t \sim N(0, I)\) is a bivariate error sequence with \(\alpha_t \perp \perp \alpha_s\), and identify the \(2 \times 2\) matrix \(\Phi\) as a function of \(\{L, \phi, \gamma\}\).

\(nb.\) \(x_t\) is a bivariate AR(1) process.
4. **Importance sampling in the Dirichlet gene variant example.**

Return to the analysis of data in the gene variant example with the following specific prior for the three probabilities $\theta$: the prior expected rates are $E(\theta_0) = 0.92$, $E(\theta_1) = 0.07$ and $E(\theta_2) = 0.01$, and the precision of the Dirichlet prior is 10, so that $\theta \sim \text{Dir}(a)$ with $a' = (9.2, 0.7, 0.1)$.

Now, suppose we have a sample of $y$ of $n = 40$ independent individuals tested. For the first 36 individuals we see the actual test result, and learn that 35 are wild-type and 1 is a type A variant. The final four individuals are tested as variants, clearly not wild types, but variants of unknown type.

- What is the posterior for $\theta$ given only these first 36 fully observed test results? What are the posterior means for $\theta_j$, $(j = 0, 1, 2)$, based on this set of data only? What is the kernel of the corresponding posterior density?
- Write down the form of the kernel of the posterior density $p(\theta|y)$ based on all $n = 40$ observations, now including the four “uncertain” mutation cases. Explain why is this not a Dirichlet posterior.
- For importance sampling of this posterior $p(\theta|y)$, one possible importance sampling function is the prior Dirichlet density. Suggest an alternative importance density $g(\theta)$ that is also a Dirichlet density, but that is theoretically a better approximation to the posterior.
- Generate an importance sample and use it to generate MC estimates of $E(\theta_j|y)$, $(j = 0, 1, 2)$. Repeat this once or twice with varying choices of MC sample size. Comment, in particular, on the stability (or otherwise) of these MC estimates of posterior means, and on the question of how easy/difficult it is to accurately estimate small probabilities such as $\theta_2$ here.
5. Return to the linear, normal AR(1) plus noise model where we observe \( y_t \) with

- \( (y_t|x_t) \sim N(x_t, w) \) independently of \( x_{t-j}, j \geq 1 \),
- \( (x_t|x_{t-1}) \sim N(\phi x_{t-1}, v) \) independently of \( x_{t-j}, j > 1 \).

Gibbs sampling now allows us to consider a number of approaches to MCMC analysis; one key detail is that we can now properly treat initial values. We do this by augmenting the state space with the latent pre-initial value \( x_0 \). Gibbs sampling then aims to simulate the full posterior

\[
p(x_{0:n}, \phi, v, w | y_{1:n}).
\]

Assume the prior

\[
p(x_0, \phi, v, w) = p(x_0)p(\phi)p(v)p(w)
\]

where \( x_0 \sim N(0, u) \) for some (large?) \( u > 0 \), \( \phi \sim N(c, C) \), \( v^{-1} \sim Ga(a/2, av_0/2) \) and \( w^{-1} \sim Ga(b/2, bw_0/2) \). Here the prior parameters are specified; \((v_0, w_0)\) are prior point estimates (guesses) of the variances \((v, w)\), and we will typically choose the prior degrees of freedom \((a, b)\) to be relatively small. Note also that the process is not assumed to be necessarily stationary.

In the following, you can identify the conditional prior/posteriors discussed by the conditioning arguments. Each question has an answer that will be either a normal or gamma distribution and you need to specify the parameters of each such distribution.

**Prior to observing the data \( y \):**

(a) Write down the compositional form of \( p(x_{0:n} | \phi, v) \).
(b) What is \( p(\phi | x_{0:n}, v) \)?
(c) What is \( p(v^{-1} | x_{0:n}, \phi) \)?
(d) What is \( p(x_0 | x_{1:n}, \phi, v) \)?

Then, using the above priors and now updating to the **posteriors on observing the data \( y \):**

(e) What is \( p(\phi | x_{0:n}, v, w, y_{1:n}) \)?
(f) What is \( p(v^{-1} | x_{0:n}, \phi, w, y_{1:n}) \)?
(g) What is \( p(x_0 | x_{1:n}, \phi, v, w, y_{1:n}) \)?
(h) What is \( p(w^{-1} | x_{0:n}, \phi, v, y_{1:n}) \)?

- Describe how the conditional posteriors (e)-(h) are four of the five required distributions for a Gibbs sampler to generate samples from the full posterior distribution \( p(x_{0:n}, \phi, v, w | y_{1:n}) \).
- What is the fifth required conditional posterior?

**BONUS:** NOT part of the exam assessment, but of interest and of use later: Write Matlab functions to implement the four conditional simulations above in (e)-(h). Begin to explore this approach to MCMC in this model fitted to the SOI time series. Think a little about the specification of the priors for \( v, w \) and \( \phi \). A modest bonus grade will be available for any cogent and documented attempts submitted with the Midterm.
6. $x = (x_1, x_2, \ldots, x_5)'$ has a 5-variate normal distribution with precision matrix

$$K = \begin{pmatrix}
  K_{1,1} & K_{1,2} & K_{1,3} & 0 & 0 \\
  K_{2,1} & K_{2,2} & K_{2,3} & K_{2,4} & K_{2,5} \\
  K_{3,1} & K_{3,2} & K_{3,3} & 0 & 0 \\
  0 & K_{4,2} & 0 & K_{4,4} & K_{4,5} \\
  0 & K_{5,2} & 0 & K_{5,4} & K_{5,5}
\end{pmatrix}$$

where the $K_{i,j}$ elements indicated are all non-zero.

Show that:

(a) $p(x_1, x_3 | x_2, x_4, x_5)$ does not depend on $x_4, x_5$; deduce that $(x_1, x_3) \independent (x_4, x_5) \mid x_2$.

(b) The joint density $p(x)$ can be written as

$$p(x) = \frac{p(x_1, x_2, x_3)p(x_2, x_4, x_5)}{p(x_2)}.$$

(c) Comment on how this density decomposition relates to the block structure of the precision matrix $K$, and highlights the role of $\{x_2\}$ as a “separating” variable in the joint distribution.

nb. This is a simple example of decomposition of a multivariate distribution in terms of simpler marginal components that exploits the conditional independencies in the joint distribution – a key idea in graphical models.

7. A Gibbs Sampling Linear Regression Exercise. Refer to the discussion of reference Bayesian analysis of linear regression models with T distributed errors.

A gene expression study in breast cancer investigates the predictive association between a response gene $Y$ and two others - a transcription factor $H_1$ that is thought to positively promote the RNA levels of $Y$, and a gene $H_2$ whose action is thought to inhibit that of $Y$. $n = 50$ breast tumors generate DNA microarray data from which the $n$ observations $\{Y_i, H_{1,i}, H_{2,i}\}$ are made on RNA levels of the three genes, representing biological variability tumor-to-tumor plus experimental noise. A linear regression is assumed for the log base 2 expression levels $\{y_i, h_{1,i}, h_{2,i}\}$, namely

$$y_i = h_i' \beta + \epsilon_i$$

with $h_i' = (1, h_{1,i}, h_{2,i})$ and $\beta' = (\beta_0, \beta_1, \beta_2)$, and with $y_i = \log_2(Y_i)$, $h_{1,i} = \log_2(H_{1,i})$ and $h_{2,i} = \log_2(H_{2,i})$.

The data $\{Y_i, H_{1,i}, H_{2,i}\}$ are given in the $50 \times 3$ data matrix on the course web site: go to the Support page, then the Examples, Matlab code, Data link, to find the data listed as 50 samples of gene expression data on 3 breast cancer genes.

Implement the Gibbs sampler for this model, in general. Aim to produce Matlab code that is concise and parallelized as much as possible (this can be done in just a few lines). The main Gibbs sampling iteration will be coded as a for loop, but there should be no other looping.

Structure the code for the MCMC on this model as follows: Starting with $\lambda_i^{(0)} = 1$ for each $i = 1, \ldots, n$, run the Gibbs sampler for an initial 200 iterations before saving the simulated parameters. Then save $m = 2500$ samples from there on. That is, discard the initial 200 “burn-in” samples, and assume that the Gibbs sampler has converged at that point so that the 2500 saved samples represent, approximately, posterior samples from $p(\beta, \phi, \Lambda | y)$.

Choose $\nu = 4$ degrees of freedom for the model T distribution, and perform the following analysis:
(a) Run the Gibbs sampler as described. Explore lag-1 scatter plots for the MC series of values for \( \beta_0 \) (i.e., \texttt{scatter(b0[1 : (m - 1)], b0[2 : m])} or similar). This will let you assess, visually, evidence of dependence within the Gibbs parameter sequence. Comment on the plot and this question of Gibbs-induced dependence in posterior samples for \( \beta_0 \).

Repeat this for \( \beta_1 \), then \( \beta_2 \) and \( \phi \). Is there much evidence of serial dependence in any of these cases?

(b) Summarize the approximate posterior inferences for \( \beta_1 \) and \( \beta_2 \) in terms of posterior histograms, approximate posterior means and 95% intervals.

(c) Interest lies in inference on the relative absolute effects of the two transcription factors, \( \mu = |\beta_1/\beta_2| \). From the Monte Carlo sample, compute an approximate posterior mean and 95% interval for \( \mu \).

(d) From examining boxplots of posterior samples, and/or posterior means, and/or posterior quantiles (all at your choice and discretion) for the \( n \) weights \( \Lambda \), discuss whether there are any particular observations that might be considered as particularly interesting - potential “outliers”? Which ones? How “outlying” are they?

(e) Prediction: Scientific interest lies in the question of how gene \( Y \) will respond - in terms of predicted levels of RNA expression - in a circumstance when the regulatory gene \( H_1 \) is at average levels, say fixed at \( H_{1,n+1} = 190 \) on the original expression scale, but in which the gene \( H_2 \) that inhibits/controls \( Y \) is at extremely low levels and almost inactive, say fixed at \( H_{2,n+1} = 50 \) on the original expression scale.

Address this question by summarizing the posterior predictive distribution for \( Y_{n+1} \) based on predicting \( y_{n+1} \) at an appropriate specified value of \( h_{n+1} \). Armed with the full posterior sample for \{\( \beta, \phi \)\} from the T analysis with \( \nu = 4 \), we are now in a position to do this directly via simulation.

- Generate a Monte Carlo sample from the predictive distribution for \( Y_{n+1} \) at the values of \( H_{1,n+1}, H_{2,n+1} \) described above, and summarize it in terms of MC approximations to the predicted mean expression, a 95% predicted interval, and a histogram of sampled values.

(f) Refit the model with \( \nu = 200 \) (so that the T distribution is essentially normal). Compare the approximate posterior means and 95% intervals for \( \beta_1, \beta_2 \) and \( \mu \) in this refitted model to that with \( \nu = 4 \). Comment on how the T model has modified these inferences relative to the “normal” analysis.