FINAL EXAM

Answer all questions. The exam has a total of 180 points and is a three hour exam.

(1) **40 points** Suppose

$$y_t = .98x_{t-1} + \varepsilon_t \tag{1}$$
$$x_t = .96y_{t-1} + \nu_t \tag{2}$$
$$\text{Var}\left(\begin{bmatrix}\varepsilon_t \\ \nu_t\end{bmatrix}\right) = I \tag{3}$$

and ε_t, ν_t form the innovation vector for the joint y, x process.

(a) Calculate the spectral density of y.
One can do this with matrix notation, or derive by substituting (2) into (1)

$$y_t = .98 \cdot .96 \cdot y_{t-1} + \varepsilon_t + .98 \cdot \nu_{t-1}\,.$$

Then

$$y_t = \frac{\varepsilon_t}{1 - .9408L^2} + \frac{.98\nu_{t-1}}{1 - .9408L^2}$$

Since ε_t and ν_t are serially uncorrelated and uncorrelated with each other at all leads and lags, this implies that the spectral density of y is

$$S_y(\omega) = \frac{1.9604^2}{|1 - .9408e^{-2i\omega}|^2} = \frac{3.8432}{1.8851 - 1.8816\cos(2\omega)}\,.$$

(b) Explain why, if this model has been estimated on quarterly data, it implies strong seasonality. [This can be answered based on the calculations in part 1a, but it might also be answerable directly, even if 1a stumps you.]
The expression for y's spectral density obviously peaks at frequencies where $\cos(2\omega) = 1$, i.e.at integer multiples of π. The ratio of the height there to the height of the spectral density at its minimal values (at the even integer multiples of π) is 1076, i.e. large. π is just one of the quarterly seasonal frequencies, but this process will have persistent oscillations of period 2. This will certainly appear as a seasonal pattern, even though not all seasonal frequencies are represented.

(2) **55 points** A simple new Keynesian model has this form

Phillips curve	$p_t = p_{t-1} + .7y_{t-1} + \varepsilon_{1t}$	(4)
IS	$\rho_{t-1} = y_t - y_{t-1} + \eta_{1t}$	(5)
Fisher relation	$p_t = p_{t-1} + r_{t-1} - \rho_{t-1} + \eta_{2t}$	(6)
Monetary policy	$r_t = .5(y_t - y_{t-1}) + 1.1(p_t - p_{t-1}) + \varepsilon_{2t}\,.$	(7)

The η's are expectational errors. The two ε's are independent of each other and across time.

Date: January 31, 2006.

When the model is solved, it takes the form

$$z_t = Gz_{t-1} + \xi_t \,, \tag{8}$$

where the z_t vector is p_t, y_t, ρ_t, r_t stacked up, and

$$G = \begin{bmatrix} 1.0 & 0.70 & 0 & 0 \\ 0 & 0.19 & 0 & 0 \\ 0 & 0.23 & 0 & 0 \\ 0 & 0.37 & 0 & 0 \end{bmatrix}, \qquad \xi_t = \begin{bmatrix} 1.0 & 0 \\ 0.79 & 0.72 \\ 0.94 & 0.86 \\ 1.49 & 1.36 \end{bmatrix} \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix} . \tag{9}$$

(a) Which variable or variables in this system are stationary, if any? Does the system show cointegration?
The second equation in the solution is $y_t = .19y_{t-1} + .79\varepsilon_{1t} + .72\varepsilon_{2t}$. Because the coefficient on lagged y is less than one in absolute value, y will be stationary. The third and fourth equations assert that r and ρ will be lagged y plus a serially uncorrelated disturbance. Thus those two variables are also stationary. The first equation asserts that the first difference of p is a constant times lagged y plus a shock. This implies p′s ARMA representation will show a unit root, and it is non-stationary. The system thus has one unit root and one stable root, with one non-stationary and three stationary variables. You could say it shows cointegration in a degenerate sense — there are three stationary "linear combinations" — but these stationary linear combinations are just y, r and ρ themselves. Since one does not need to take linear combinations to find stationary variables, according to common usage of the term we might not say this system shows cointegration.

(b) If this system described the true behavior of the economy, suppose we estimated a structural VAR, identifying the monetary policy equation as the first equation in a triangular identification scheme. (That is, the identification puts r_t on the left of the first equation and allows it to impact all other variables contemporaneously.) Would this approach correctly estimate the impulse response of the model to shocks in ε_{2t}, the policy shock? Justify your answer.
In a Cholesky scheme, the fact that r appears contemporaneously in all the other equations means that in the "r equation" there are no contemporaneous values of other variables on the right-hand side. This equation will therefore be just a least-square projections of r_t on past values of all variables in the system. That is just what the last equation in (9) is, so it is that equation that will be estimated. The residual in that equation is a linear combination of ε_{1t} and ε_{2t}, not ε_{2t} alone, so this approach will not give correct identification or (therefore) correct impulse responses.

(c) What if instead the policy equation were put last in the Cholesky ordering?
This would put contemporaneous values of all the other variables in the regression equation. It can be seen that the last three rows of G and ξ_t in (9) are just scalar multiples of each other (to the two-decimal precision that has been displayed). So with current y on the right-hand side, the fit of the "r equation" will be perfect, with no residual at all, except rounding error. So this approach also does not correctly recover the policy equation.

(d) What if the monetary policy equation had lagged output growth and inflation, instead of their current values, on the right-hand side?
Here we are considering a modification of the structural model itself, as opposed to a particular VAR identification scheme. With this change, because lagged z must be uncorrelated with ε_{2t}, the residual in the reduced-form VAR equation for r will in fact be the structural shock

ε_{2t}. A Cholesky identification scheme with r first in the ordering will therefore correctly estimate the impulse responses to a policy shock.
This answer has assumed (as you were told to verbally at the start of the exam) that lagging the right-hand-side variables in the monetary policy equation does not affect the existence or uniqueness of the solution to the original model. In fact, with coefficients set at the values given in the problem, existence and uniqueness are maintained.

(3) **55 points** Consider the following model, which might be used for house prices, for example.

$$y_t = \mu_t + \rho y_{t-1} + \varepsilon_t \tag{10}$$
$$P[\mu_t = \mu_{t-1}] = \theta\,, \text{ all } t \tag{11}$$
$$\tag{12}$$

We assume that, conditional on $\mu_t \neq \mu_{t-1}$, $\mu_t - \mu_{t-1} \sim N(0, \omega^2)$ for all t. Note that this means μ_t is a martingale. The evolution of the μ's is independent of the whole ε sequence. If we model μ_t as depending on the state, i.e. as $\mu(S_t)$, we can treat this as a hidden Markov chain model, at least over any finite sample.

(a) Show what the transition probability matrix for the state must look like in order to implement this model. Assume we normalize by insisting that S_1 is the state prevailing at time 1.
If we treat each new value of μ_t as a new state, then the states will (with probability one) never repeat. With the probability of persisting in the same state always θ, we need the transition matrix to be of the form

$$\begin{bmatrix} \theta & 0 & 0 & \dots & 0 & 0 \\ 1-\theta & \theta & 0 & \dots & 0 & 0 \\ 0 & 1-\theta & \theta & \dots & 0 & 0 \\ 0 & 0 & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 1-\theta & \theta & 0 \\ 0 & 0 & 0 & 0 & 1-\theta & 1 \end{bmatrix}.$$

Using a fixed finite number of states (i.e. a fixed dimension for the transition matrix) is at best an approximation, though, unless we allow a possibility of one state for each date in the sample, which is unreasonably many. In practice one would assume a reasonable finite number, then check whether using a larger number improves the fit.

(b) Describe an algorithm for sampling from the posterior of this model. Note that the restrictions on the transition matrix may make the posterior fail to be in Dirichlet form as a function of the transition probabilities.
Actually the likelihood, as a function of θ, is in the form of a Beta *distribution, which is a special case of the Dirichlet. I neglected to make explicit the standard assumption that $\varepsilon_t \sim N(0, \sigma^2)$ and i.i.d. across t. The full log likelihood, before invoking any priors and treating the sequence of state values (with k the number of states and n_j, $j = 1, \dots, k$ the*

number of observations in state j) as part of the parameter vector, is

$$\left(\sum_{j=1}^{k} n_j - k\right)\log\theta + k\log(1-\theta) - T\log\sigma - \frac{\sum_{t=1}^{T}(y_t - \rho y_{t-1} - \mu(S_t))^2}{2\sigma^2} - \sum_{t=1}^{T}\mathcal{I}_{\{\mu_t \neq \mu_{t-1}\}}(\mu_t, \mu_{t-1})\log\omega - \frac{\sum_{t=1}^{T}(\mu_t - \mu_{t-1})^2}{2\omega^2}.$$

Notice the last part of the likelihood function, which reflects the assumption that the μ_j's form a random walk and is not a standard part of a time series regression model with Markov-switching coefficients.
To keep things convenient, we will want a conjugate prior, i.e. one that is Beta in θ, inverse-Gamma in σ^2 and ω^2, and jointly Gaussian in the k values of μ and the coefficient ρ, conditional on the sequence of states and on θ. Note that the extra terms in μ are quadratic, so the overall posterior on $\{\mu_t\}$ is Gaussian. However it is not a simple OLS framework. An ideal answer would display the form of the covariance matrix and mean of the joint posterior for $\rho, \{\mu(j)\}$.. A good answer would note that the log likelihood is a quadratic form and describe how in general one extracts the implied posterior mean and variance.
The algorithm, then, would on each iteration

(i) *Draw θ from a* $\text{Beta}(\sum_{j=1}^{k} n_j - k + n_0 + 1, k + m_0 + 1)$, *where n_0 and m_0 are the parameters of the prior on θ. For the prior to imply persistent states, one would want m_0/n_0 small.*

(ii) *Draw $\{S_t\}$ using the usual hidden Markov chain forward-backward recursions.*

(iii) *Draw σ^2 and ω^2 from their respective mutually independent inverse-gamma distributions. For σ^2 the distribution is* $\text{Gamma}(N, u'u/2)$, *where N is the total number of observations, including both actual and dummy observations on ρ and $\{\mu_j\}$ and u is the vector of residuals, both actual and dummy, from the model equations. For ω^2 the distribution is* $\text{Gamma}(\hat{k} + k_0, \sum((\Delta\mu_t)^2 + \ell_0)/2$, *where k_0 and $\ell_0/2$ are the degrees of freedom and scale parameters from an inverse-gamma prior on ω^2 and $\hat{k}$ is the number of state changes in the $\{S_t\}$ sequence.*

(iv) *Draw ρ and $\{\mu_j\}$ from their joint-normal conditional posterior.*

It should be possible to integrate $\rho, \{\mu_j\}$ out of the conditional joint distribution of them and σ^2, ω^2, obtaining a marginal (conditional on $\theta, \{S_t\}$) on the variance parameters. Drawing from this for the variance parameters instead of from their conditional distribution would make the iteration more efficient.

(c) **30 points**

(i) Explain why, in a univariate first-order autoregressive model, confidence intervals generated from a naive bootstrap without bias correction will be misleading from both frequentist and Bayesian perspectives.
The model is $y_t = \rho y_{t-1} + \varepsilon_t$. (You could have included a constant, or not. The general form of the answer is the same,though the constant introduces subtleties if you wanted a chance to show off having read and understood Sims, Stock and Watson.)
It is well known that estimates of ρ by OLS, which is MLE conditional on initial conditions, are biased, as a matter of their distribution in repeated samples, toward zero. The naïve parametric bootstrap first estimates ρ, say by MLE conditional on the initial observation, then generates a large number of artificial samples from the distribution of the data by either drawing from the theoretical normal distribution

of the shocks or drawing from the sample distribution of the shocks, then generating data recursively from the model equation with $\rho = \hat{\rho}_{OLS}$. The estimates of ρ from the artificial sample are of course also biased toward zero, so if the bootstrap draws are treated as if they were the conditional pdf for ρ given the data (which of course would be a mistake from a frequentist point of view) they are very misleading, because they in effect have doubled the original bias. Also, these simulations will show a long left tail in the simulated distribution which cannot be justified by looking at the shape of the likelihood. These objections are from the Bayesian viewpoint. From the frequentist viewpoint, the point about bias is also valid. A confidence interval generated this way will tend to be too close to zero and hence not to have close to the assumed coverage probability. Better results can be obtained by using a first round bootstrap to estimate the bias, then adjusting the original $\hat{\rho}$ to eliminate this bias, then repeat the bootstrapping. Killian showed that in realistic sized samples for small VAR models this procedure gives reasonable (by frequentist standards) results, even though it has no firm foundation in even asymptotic theory.

(ii) What does it mean for a MCMC sampling scheme to converge? Describe at least two ways to check whether it seems to have converged.

There are two senses of convergence that need to be checked. One is whether the Markov chain has run long enough so that it has settled in to ergodic behavior — in other words the effects of initial conditions have died away. The other is whether the sample is large enough to provide accurate estimates of whatever function of the distribution one might want to estimate from the sample. One technique that applies to both is the trace plot. If we are sampling the random vector X and we are interested in $E[f(X)]$, then we can plot $f(X_j)$ against draw number J. We should be able to visually cut off a first segment of the sample, after which the rest appears to behave in a stationary, repetitive way. This checks whether initial conditions are no longer important. One can also look for patterns indicating high serial dependence — volatility or level of the plotted serious showing just a few slow waves. These patterns would indicate that a much longer artificial sample is needed before an accurate estimate of the mean is possible.

A more formal check is to estimate effective sample size. Effective sample size computations involve two steps. First, an estimate of how much of the MCMC sample needs to be thrown away as "burn-in". Sometimes this is done by eye from trace plots. Sometimes it is done more formally, for example by checking whether when the sample is cut into n components, the first last $n-1$ components have means and variances that could easily have been drawn from the same distribution, which can be tested. After having thrown out the burn-in segment, one then fits a model of serial correlation to the rest. This can be done with frequency-domain methods, by fitting an AR model, or by other time-series modeling techniques. The fitted model is then used to construct an estimate of the the variance of the sample mean of $f(X_j)$, and this is compared to what the variance would be if the sample had been i.i.d. The actual MCMC sample size divided by this ratio is the effective sample size. Generally if it is less than 200, the MCMC sample is regarded as unreliable. If one has in mind a target standard error, based on substantive considerations, for the standard error of the estimate of $E[f(X)]$, one can decide whether to sample further depending on whether this target has been reached or not, so long as effective sample size is at least around 200.

Finally, one should check for fat tails. It is quite possible for a target distribution to fail to have second moments, even though its first moments are well defined. In this case MCMC will converge, but very slowly, and methods of deciding on burn-in period or estimating effective sample size that are based on second moments will be useless. Fat tails can be diagnosed by eye if the trace plots show recurring large outliers, with the largest outlier increasing fairly rapidly as sample size increases.

(iii) Suppose we are modeling a collection of 20 time series of logged prices of eggs in US cities. We believe that there is a common, non-stationary component in these series due to inflation in the general price level, but that this is the only source of non-stationarity. Can we say a priori what is the form of the cointegrating vectors for this system? If so, what are they? If not, why not?
The relative prices, in logs, are the differences among the log price (call them p_{it}) time series. A complete set of stationary linear combinations is then $p_{it} - p_{1t}$, $i = 2, \ldots, 20$, and the cointegrating vectors are the coefficients forming these linear combinations (1 in position i, -1 in position 1, zero elsewhere). There are then 19 cointegrating vectors. The vectors themselves are not unique. Any linearly independent set of 19 coefficient vectors all of which are orthogonal to the vector of 20 repeated 1's spans the same space, which is all that is uniquely defined.