Review Problem Answers

1. Saying the residual at each *t* is drawn from a mixture of normals means that the p.d.f. of $\varepsilon(t)$ at each *t* is the sum of three normal p.d.f.'s with weights π_i summing to one. But this can be modeled as involving the unobservable random variable i_t, which takes on the values 1, 2, 3 with probabilities π_i, *i*=1, 2, 3, and with $\varepsilon(t)\big| i_t \sim N\left(0, \Sigma_{i_t}\right)$. Collecting lagged *y*'s into a single row vector $X(t)$ and the coefficients in *B* into a single parameter matrix β, we can rewrite (1) from the problem set as

$$y(t)' = X(t)\beta + \varepsilon(t)' \ . \qquad [1]$$

Then the log of the joint p.d.f. of $\{i_t\}$ and $\{y(t)\}$ is

$$\sum_{t=1}^{T}\left\{-\tfrac{1}{2}\log\left|\Sigma_{i_t}\right| - \tfrac{1}{2}\left(y(t)' - X(t)\beta\right)\Sigma_{it}^{-1}\left(y(t)' - X(t)\beta\right)' + \log\pi_{i_t}\right\} . \qquad [2]$$

To form the likelihood in terms just of the parameters (not including $\{i_t\}$), we would have to integrate this p.d.f. over the $\{i_t\}$ sequences, which would give it an inconvenient, non-Gaussian form. It is therefore convenient not to integrate out the $\{i_t\}$'s, but to leave them in and use Monte Carlo methods that draw jointly from a p.d.f. proportional to [2] (exponentiated) in the parameters and $\{i_t\}$ jointly. We are not directly interested in the $\{i_t\}$'s, but drawing from the flat-prior posterior (the likelihood) is much easier if we draw them together with the parameters.

Conditional on a sequence $\{i_t\}_{t=1}^{T}$, each disturbance is normal, and this p.d.f. has a Gaussian shape in β. It is not a very nice Gaussian shape, because the occurrence of different Σ_i's at different observations, if they are not diagonal, prevents application of any of the usual SUR conditions that allow equation-by-equation estimation of the peak of the likelihood. Nonetheless, it is a big Gaussian distribution. Its shape can be found by stacking up the data for the system, treating it as a big GLS problem with *nT* observations, where *n* is the number of variables in *y* and *T* is the number of observations. This will give a mean and covariance matrix for the elements of *B* , and it is then straightforward to draw a random *B* from this distribution.

Conditional on all the other parameters and the data, the p.d.f. as a function of the i_t's is straightforward. The $\varepsilon(t)$ sequence is a known function of the conditioning variables. A particular i_t enters the log likelihood only through a single element in the sum in [2], the term indexed by *t*. This means that the p.d.f. as a function of $\{i_t\}_{t=1}^{T}$ has the form of independent draws for the elements of the sequence. There are three possible values of this term, corresponding to the three possible values of i_t, so to draw from this conditional distribution for i_t requires just a simple draw from a three-point discrete distribution at each *t*.

Conditional on all the other parameters, the data, and the $\{i_t\}$ sequence, the three covariance matrices Σ_i, i=1, 2, 3 enter the joint p.d.f. through three separate components of the sum in [2]. For Σ_1, for example, the corresponding component of [2] is

$$-\frac{T_1}{2}\log|\Sigma_1| - \tfrac{1}{2}\sum_{t\in I_1}\left(y(t)' - X(t)\beta\right)\Sigma_1^{-1}\left(y(t)' - X(t)\beta\right)' - T_1\log\alpha_1 \qquad [3]$$

where T_1 is the number of observations in I_1, the set of t values for which $i_t = 1$. Note that as a function of Σ_1 this is in the form of a standard inverse-Wishart p.d.f. One forms the residual covariance matrix for each of the three subsamples I_j, then draws a value for the corresponding Σ_j from the associated inverse-Wishart. (See Gelman, Carlin, Stern and Rubin, p.474-5 and 481, for description of the inverse-Wishart).

Looking again at [3] we can see that the p.d.f. as a function of the α's is proportional simply to

$$\alpha_1^{T_1}\alpha_2^{T_2}\left(1-\alpha_1-\alpha_2\right)^{T_3} . \qquad [4]$$

This is a three-dimensional Dirichlet p.d.f. (GCSR p. 476-7, 481), which is straightforward to draw from.

Thus we have the possibility of an exact Gibbs sampling scheme, drawing in turn from the conditional posteriors for β, $\{i_t\}_{t=1}^{T}$, $\{\Sigma_i\}_{i=1}^{3}$, $\{\alpha_i\}_{i=1}^{3}$.

All of this discussion has assumed we would use the conditional p.d.f. of the y's given initial observations. If we assume stationarity, there would be an implied marginal distribution for the initial y's we could use. However, this marginal distribution is extremely messy analytically. Since Markov-chain Monte Carlo methods assume that the p.d.f. of the data can be calculated for given parameter values at given data points, they will not help with this situation.

2. Asymptotically, for the stationary case ($|\rho|<1$), the usual OLS formulas for mean and variance of the OLS estimators and the usual Gaussian distribution for the estimators are justified. So when the model is true, we can write asymptotically

$$\sqrt{T}\left(\hat{\rho}_T - \rho\right) \sim N\left(0, 1-\rho^2\right) . \qquad [5]$$

Also, for the stationary case the error in $\hat{\rho}_T$ is negligibly correlated with the last observation in the sample, so the distribution of $\hat{\rho}_T^2 y(T)$ conditional on $y(T)$ is asymptotically just the marginal asymptotic distribution of $\hat{\rho}_T^2$ scaled by $y(T)$. Note that the mean-squared error of two-step-ahead forecast in this "true model" case is dominated by the variance of the forecast based on knowledge of the true ρ, which is $\left(1+\rho^2\right)\sigma_\varepsilon^2$. The component due to error in estimation of ρ declines at the rate $T^{-\frac{1}{2}}$. The asymptotic distribution of $\hat{\rho}_T^2$ is centered at ρ^2, with variance (after scaling by $\sqrt{T}$)

$$(2\rho)^2 \cdot \mathrm{var}\left(\sqrt{T}(\hat{\rho}_T - \rho)\right) = 4\rho^2 \cdot (1-\rho^2). \qquad [6]$$

If we estimate using OLS on (3) from the problem statement, the error in the coefficient estimate is

$$\frac{\sum (\varepsilon(t) + \rho\varepsilon(t-1)) y(t-2)}{\sum y(t-1)^2}. \qquad [7]$$

The denominator of this expression, when normalized by T^{-1}, converges in probability to the constant $\sigma^2/(1-\rho^2)$, while the numerator, being a sum of zero-mean, finite-variance, stationary random variables, converges, when normalized by $T^{-\frac{1}{2}}$, to a zero-mean normal distribution. There is a standard result, which I realized as I prepared this answer sheet might not have been covered in the prerequisites for this course, that time averages of a zero-mean, finite-variance stationary process x with autocovariance function $\mathrm{cov}(x(t), x(t-s)) = R_x(s)$ converge in distribution, when normalized by $T^{-\frac{1}{2}}$, to $N\left(0, \sum_t R_x(t)\right)$, assuming that R_x is absolutely summable. The numerator of [7] has ACF $R(t) = 0$ for $t > 1$, since the ε's are zero-mean and i.i.d. It has variance $R(0) = \sigma_\varepsilon^4 \cdot (1+\rho^2)/(1-\rho^2)$ and $R(1) = \sigma_\varepsilon^4 \rho^2/(1-\rho^2)$. Therefore the ratio in [7], normalized by $\sqrt{T}$, converges to a $N(0, 1+3\rho^2)$. (To see why, be sure you remember that one sums R over both positive and negative values of its argument.) It is not hard to check then that the asymptotic variance of this estimator, $1+3\rho^2$, must exceed that of the previous estimator based on a one-step-ahead regression, as given in [6]. In fact, the difference always exceeds 15/16, for any value of ρ. But remember that the unnormalized difference is shrinking at rate $T^{-\frac{1}{2}}$ and becomes a negligible part of the overall MSE of forecast as $T \to \infty$.

For the other case, where the AR model is false, using the one-step-ahead OLS estimate produces an inconsistent estimator of the coefficient in the best linear predictor of $y(T+2)$ based on $y(T)$. Thus the difference between the two estimators does not dwindle as $T \to \infty$. This is the main point of this question, and had been discussed in class. On an exam, recognizing this point and giving a quick argument as to why it's true in this case would have given you most of the credit for the question. But here we proceed with the details.

If y is drawn from a first-order MA of the form

$$y(t) = \eta(t) + \alpha\eta(t-1), \qquad [8]$$

then the best linear predictor of $y(t)$ based on $y(t-2)$ is just zero, since y is uncorrelated with values of itself lagged by more than one period, and its MSE is $\sigma_\eta^2 \cdot (1+\alpha^2)$. The best one-step ahead linear predictor for $y(t)$ based on $y(t-1)$ alone is $\theta y(t-1)$, with

$$\theta = R_y(1)/R_y(0) = \alpha/(1+\alpha^2). \qquad [9]$$

If we estimate a regression of $y(t)$ on $y(t-1)$ by OLS, we will obtain a consistent estimate of θ, and the resulting two-step ahead prediction error will be, in large samples

$$\sigma_e^2 \cdot (1+\alpha^2) + \theta^2 \sigma_y^2 = \sigma_e^2 \frac{1+3\alpha^2+\alpha^4}{1+\alpha^2} , \qquad [10]$$

which is clearly larger than what is obtained with the OLS regression on $y(t-2)$, except when $\alpha = 0$. The ratio of MSE's is worst when α^2 is at its upper limit of 1, when the 1-step-ahead regression gives an MSE 1.25 times as large as the 2-step-ahead regression.

3. i) The money neutrality proposition, in a log-linear system, is a single system-wide dummy observation. It will be a scalar multiple of an observation in which all nominal variables are set to one and all real variables are set to zero. When multiplied by the coefficients, this artificial data vector implies exactly the neutrality constraint. Since the likelihood itself has the desired Kronecker structure, adding this pseudo-observation will preserve that structure. Note that we will have to accept that the variance of the disturbance is the same in all equations for this dummy observation, so we cannot be more sure of the neutrality restriction in some equations than in others.

ii) A restriction that the sum of coefficients on variable *j* in equation *i* be zero is not in the form of a system-wide dummy observation, as it involves only coefficients in one equation. However, as the Sims-Zha article points out, differences across equations in the right-hand-side variable matrix, while destroying the strict Kronecker structure, do still allow equation-by-equation estimation when the model is formulated to imply a diagonal covariance matrix of disturbances. The requirement is just that block diagonality of the covariance matrix in the posterior for a^+ be preserved. Since the long-run restriction we are discussing here is a linear combination of coefficients in a single equation, it can be expressed by adding a dummy observation to the data matrix for that equation, and this will preserve block-diagonality.

Note that most existing empirical work using "long run restrictions" on VAR's has used 2-variable models, where the restriction that an off-diagonal element of the AR operator have coefficients summing to zero is equivalent to the restriction that the corresponding element of the MA operator have coefficients summing to zero. But in systems of dimension larger than 2 this correspondence no longer holds. Often a long-run restriction is more reasonably formulated as, say, "monetary policy shocks have no long run effects on *Y*" rather than as "a permanent change in *M* will have no long run effect on *Y*". The latter is a restriction on the AR, the former a restriction on the MA. Linear restrictions on the MA are in general complicated nonlinear restrictions on the AR, involving all equations even when the MA restriction involves only one *i* and *j*.