Take-Home Answers

1. The log posterior p.d.f. will be

$$-(T+2)\log\sigma - \frac{T}{2}\log\left(1-\rho^2\right) - \tfrac{1}{2}\sum_{t=1}^{T}\frac{\left(y(t)-\rho\, y(t-1)\right)^2}{\sigma^2\left(1-\rho^2\right)} - \frac{y_0^2}{2\sigma^2} \quad . \tag{1}$$

This is exactly the p.d.f. of the data vector $y(0),\ldots,y(T)$, except that the factor on $-\log\sigma$ is T+2 instead of T+1, reflecting the flat prior on $\log\sigma$. The parameter ρ appears twice in this expression. As $\rho \to 1$ the log term containing ρ tends to plus infinity, while the summation term tends to minus infinity (unless the sum of squared differences of y is zero, which we will assume it is not). However, a term increasing as a log grows slower in the limit than a term increasing linearly, so the summation term dominates. (This can be verified with l'Hôpital's rule if you like.) Thus the log likelihood goes to minus infinity as $\rho \to 1$, which implies that the likelihood itself goes to zero. But this does not necessarily mean that the marginal p.d.f. for ρ must go to zero. [Here's a simple example: If a,b are distributed on the region where $0<a<1,\ 0<b<(1-a)/a$ with p.d.f. $p(a,b)=a/(1-a)$, then the p.d.f. goes to zero as $a \to 0$ as a function of a for each fixed b, but the marginal p.d.f. of a is uniform on (0,1).] The most straightforward way to complete the argument is probably to integrate out the other parameter, σ, and then check directly. Despite the overall non-standard shape of the p.d.f., as a function of σ for fixed ρ it is in the form of a standard inverse-gamma p.d.f. on $1/\sigma^2$. Integrating it, we get

$$\left(\tfrac{1}{2}\sum_{t=1}^{T}\frac{\left(y(t)-\rho\, y(t-1)\right)^2}{\left(1-\rho^2\right)} + \frac{y_0^2}{2}\right)^{-\frac{T+1}{2}} \left(1-\rho^2\right)^{-\frac{T}{2}} \quad . \tag{2}$$

This expression goes to zero as $\rho \to 1$, because the term in $1-\rho^2$ raised to the negative power $T/2$ will be dominated by the term raised to the positive power $(T+1)/2$.

Note that if the question had asked about a flat prior on σ itself instead of on its log, the answer would have been different, with the marginal posterior on ρ not going to zero as $\rho \to 1$. This suggests a way to construct a prior that allows smoothly integrating analysis of $|\rho| \geq 1$ with analysis of the stationary case.

It is easy to see from (1) that the log-likelihood is not quadratic in ρ for fixed σ, and thus that the posterior is not Gaussian in ρ, even conditioning on σ, and even if we ignore the term contributed by the marginal distribution of the initial condition.

2. The problem statement should have said that θ is of lower dimension than $(n^2+n)/2$, not $(n^2-n)/2$, though this did not really affect anything. In either case the model is "overidentified".

One might begin the answer to this question by asking whether we really "have to" avoid the redundancy. Consider the simple two-by-two case where

$$A_0(\theta) = \begin{bmatrix} \theta_1 & 0 \\ 0 & \theta_2 \end{bmatrix}. \tag{3}$$

A natural normalization is to require that both θ_i's be positive. If we let the posterior p.d.f. on them spread over all four quadrants of R^2, the likelihood, and also the posterior if the prior has been consistent in giving equivalent models equal prior p.d.f. values, will have, corresponding to any local maximum in the positive quadrant, four peaks of the same shape in the four other quadrants. If we are directly interested in, say, θ_1, then it would be quite misleading to use MCMC methods to sample from this posterior and conclude (as we would have to, because the marginal posterior on θ_1 would be symmetric about zero) that θ_1 is "insignificantly different from zero". But we could easily compute the posterior p.d.f. of the absolute values of the θ's, even if sampling from a posterior that gave positive probability to negative values. Similarly, if we wanted p.d.f.'s of impulse responses, we could draw from a posterior on the *A*'s that allowed negative θ's, yet normalize the impulse responses that we compute from the draws of *A* so that, say, initial responses of the *i*'th variable to the *i*'th shock are always positive.

So it is possible to argue that no restriction on the parameter space is necessary, so long as we are careful not to compute distributions of functions of the parameters that have different values for different, but equivalent, *A*'s.

However, the fact that every peak in one quadrant in our simple example is mirrored by three other equivalent peaks in the unrestricted parameter space can create severe problems with convergence, and with assessing when convergence has occurred, with MCMC methods. If there is a single, sharp peak in the positive quadrant, MCMC methods are likely to generate samples that stay entirely in the positive quadrant for many draws. Eventually, though, if there is non-zero posterior probability in the neighborhood of $\theta_i = 0$, the MCMC sample will cross into another quadrant, and then will very likely start generating observations in the neighborhood of that quadrant's image of the sharp peak in the first quadrant. Mechanical approaches to assessing whether the MCMC algorithm is converged will conclude that, since the latter part of the artificial sample looks very different from the first part, it has not converged, when actually, taking account of the equivalence of the quadrants, it may be well converged. Assessing convergence, or generating samples of functions of the parameters that don't violate the symmetry, will in any case involve mapping points in other quadrants back in to the first quadrant. So why not just do this from the start? Every time there is a MCMC draw that goes outside the first quadrant, just map the draw back in to the first quadrant. In general, this means simply flipping the signs of all the coefficients in any row of any draw of $\left[A_0 \; A^+\right]$ that violates our normalizing restrictions. In the case of our simple 2-variable example, this means simply changing the sign of coefficients in equation *i* whenever a draw of θ_i comes out negative.

However, for Metropolis sampling, this apparently violates the rule that the jump distribution must be symmetric. Suppose we start from the point θ_1^0, θ_2^0 and jump by the Metropolis algorithm to $\left(\theta_1^0 + \varepsilon_1, \theta_2^0 + \varepsilon_2\right) = \left(\theta_1^1, \theta_2^1\right)$. If it then turns out that, say, $\theta_1^1 < 0$, we would flip

signs and make our new point instead $\left(-\theta_1^1, \theta_2^1\right)$. If our jump distribution was independent of the initial θ and had a p.d.f. J satisfying $J\left(\theta^1 \middle| \theta^0\right) = J\left(\theta^1 - \theta^0\right) = J\left(\theta^0 - \theta^1\right) = J\left(\theta^0 \middle| \theta^1\right)$, then a jump back to θ^1 from θ^0 could occur either by directly drawing a new $\varepsilon^* = \theta^0 - \theta^1$, which would have the same p.d.f. value as directly drawing $-\varepsilon^*$, or by drawing $\varepsilon^* = (\varepsilon_1, -\varepsilon_2)$. The latter draw would reverse the sign of θ_1^1, and after the sign flip would bring us back to the original $\theta^0 = \left(-\left(\theta_1^1 + \varepsilon_1\right), \theta_2^1 - \varepsilon_2\right)$. But the usual requirement on J for standard Metropolis sampling is $J(\varepsilon) = J(-\varepsilon)$, not $J(\varepsilon_1, \varepsilon_2) = J\left(\varepsilon_1, -\varepsilon_2\right)$. Thus to make Metropolis sampling work in this context, we would have to strengthen the requirement on the jump p.d.f., making it symmetric about zero in each of its arguments separately, not just in the two arguments jointly. In more general models the requirement would be that the jump p.d.f. be symmetric in each equation's coefficient vector separately.

One could also apply Metropolis sampling under the usual conditions, with the posterior p.d.f interpreted as dropping to zero outside the quadrant of normalization. This is the same kind of thing you had to do on one of our exercises, where there was an upper bound of 1 on a parameter. At the boundary, the symmetric jump distribution will produce many draws that are rejected, resulting in some inefficiency. Also, a sharp peak near a boundary can result in the algorithm getting stuck for a long time. If the algorithm starts in the quadrant of normalization, but closer to the peak across the boundary than to the peak that is in the quadrant of normalization, the Metropolis algorithm may tend to "bounce" repeatedly against the boundary for a long time, with many rejected jumps because likelihood drops off as one moves away from the boundary within the quadrant of normalization. Only when the algorithm gets near the peak that is in the quadrant of normalization will it start to show "convergent" behavior.

All of this discussion presumed the case where constraints are zero restrictions, so that flipping signs of coefficients in equations never affects whether they satisfy the restrictions. Also it is important that in this simple case it is coefficients in *A* themselves that we are jumping among with the Metropolis algorithm. If $A_0(\theta)$ is a nonlinear mapping, there will not be any simple way to restrict the jump p.d.f. so that $J\left(\theta^1 \middle| \theta^0\right) = J\left(\theta^0 \middle| \theta^1\right)$ even though jumps out of the quadrant of normalization are "sign-flipped". Of course the direct solution, of treating the posterior p.d.f. as zero outside the quadrant of normalization, is still available, and still can create the same possible inefficiencies.

3. The question should have stated explicitly that it was making the conventional assumption that the i.i.d. $N(0, I)$ distribution for ε is conditional on A_0, A_1 and past *y*'s. If we use *a* to stand for $\text{vec}\left(\left[A_0 \; A_1\right]\right)$ and $\bar{a}$ to stand for the 8-dimensional mean vector for *a* given in the problem, we can write the joint p.d.f. of *a* and $y(6)$ conditional on $y(5)$ (and all past *y*'s) as the product of the conditional p.d.f. of *a* given *y*'s dated 5 and earlier with the conditional p.d.f for *y*(6) given *a* and *y*'s dated 5 and earlier. This is just

$$|A_0|\exp\left(-\tfrac{1}{2}\left(A_0 y(6)-A_1 y(5)\right)'\left(A_0 y(6)-A_1 y(5)\right)\right)\cdot .25^{-4}\exp\left(-2(a-\bar{a})'(a-\bar{a})\right). \qquad (4)$$

This is most of the answer. The conditional p.d.f. for *a* is then just (4) treated as a function of *a* with the *y*'s fixed. A more complete answer could have substituted in for $y(5)$, $y(6)$, and $\bar{a}$ the numerical values given in the problem and collected terms. This p.d.f., because of the determinant on the left, is not in any standard (and thus easily integrated) form. It is therefore not possible (I think) to give an analytic expression for the p.d.f. normalized to integrate to one.

4.

a) It is one of the handy, but sometimes unreasonable, features of a normal prior that normal observations, even observations wildly in conflict with the prior, always reduce the posterior variance below the prior variance. More information always shrinks variance. The Kalman smoothing formula makes this clear, as the smoothed variance differs from the filtered variance for the same date by a negative semi-definite matrix. [Multivariate *t* or mixed-normal priors can make observations in the tails raise posterior variance.]

b) Ordinary rejection sampling draws from a p.d.f. j, trying to get a sample from a p.d.f. *p* that can be scaled by some positive constant q such that $\mathsf{q}p(x)<\mathsf{j}(x)$, all *x*. It rejects each draw of *x* with probability $\mathsf{q}p(x)/\mathsf{j}(x)$. Though this is similar in many respects to Metropolis-Hastings, it is different. Every time a random variable is drawn in Metropolis-Hastings, it results either in a new point for the artificial sample, or a repetition of the previous point. In standard rejection sampling, "rejected" draws do not correspond to any repetition of a previous draw.

c) This problem aimed at our class discussion of fit criteria for false models. If the data are i.i.d. and have finite variance, then the sample mean of X_t^2 provides a consistent estimate of the variance of *X*, which is in turn the minimum mean-squared-error estimate of X_{T+1}^2. The sample mean of the X_t^2 is not generally the best possible estimate of X_{T+1}^2, even under a MSE criterion, but it will in large samples at least converge to producing the minimum possible MSE, which is what would be obtained if the second moment of *X* were known exactly. Twice the squared sample mean of $\{X_t\}$ will generally converge to something other than the second moment of *X*, and hence will have an error as a predictor of X_{T+1}^2 that does not become negligible in large samples.

The harder aspect of the question might have been figuring out what conditions make either estimator optimal. If *X* is Gaussian, then with a flat prior on the variance s^2 of *X*, the posterior distribution for s^2 is inverse-gamma, with mean equal to the sample mean of $\{X_t^2\}$. Thus this assumption makes the sample mean of $\{X_t^2\}$ optimal. If instead each X_t is exponentially distributed on $(0,\infty)$, with p.d.f. $e^{-x/\mathsf{b}}/\mathsf{b}$, then the second moment of *X* is $2\mathsf{b}^2$. The posterior mean of b^2 from a sample of size *T* with a flat prior, which is the second moment of an inverse-gamma with *T*-1 degrees of freedom, is

$$\frac{2\cdot\left(\sum X_t\right)^2}{(T-2)(T-3)}\ . \tag{5}$$

This is close to twice the sample mean squared, for large *T*. It can be shown, by tedious algebra, that the variance of $2\cdot\left(\sum X_t\right)^2 \Big/ T^2$ is smaller, asymptotically, than the variance of $\sum X_t^2/T$, by about 10%, so (5) is asymptotically better than the sample mean of $\left\{X_t^2\right\}$ for this particular assumption (gamma-distributed *X*) on the distribution of *X*. But obviously for many distributions, e.g. for Gaussian *X*, the inequality goes the other way, and by a large amount.

Note, though, that it is always true that use of false models can produce very bad results. The notion that by "using our loss function to generate the fit criterion" we always get robust conclusions is dangerous. Here we rely on the assumption of i.i.d. *X*, which might be incorrect. Also, we implicitly assume that the model is not "too far" from the Gaussian one that makes the sample mean of $\left\{X_t^2\right\}$ optimal. One can describe models for which the sample mean of $\left\{X_t^2\right\}$ is very bad in modest-sized samples. For example, *X* might be a mixture of a .999 probability on a uniform distribution over $[-a,a]$ and .0005 probabilities on $\pm 1/a$. For $a \ll 1$, use of ML based on the true model will in small samples be many times more accurate than use of the sample mean of $\left\{X_t^2\right\}$.