Robust Estimation of ARMA Models with Near Root Cancellation

Timothy Cogley

Richard Startz*

May 2012

Abstract

Standard estimation of ARMA models in which the AR and MA roots nearly cancel, so that individual coefficients are only weakly identified, often produces inferential ranges for individual coefficients that give a spurious appearance of accuracy. We remedy this problem with a model that mixes inferential ranges from the estimated model with those of a more parsimonious model. The mixing probability is derived using Bayesian methods, but we show that the method works well in both Bayesian and frequentist setups. In particular, we show that our mixture procedure weights standard results heavily when given data from a well-identified ARMA model (which does not exhibit near root cancellation) and weights heavily an uninformative inferential region when given data from a weakly-identified ARMA model (with near root cancellation). When our procedure is applied to a well-identified process the investigator gets the "usual results," so there is no important statistical cost to using our procedure. On the other hand, when our procedure is applied to a weakly-identified process, the investigator learns that the data tell us little about the parameters—and is thus protected against making spurious inferences. We recommend that mixture models be computed routinely when inference about ARMA coefficients is of interest.

* Tim Cogley: Department of Economics, New York University, 19 W. 4th St., 6FL, New York, NY 10012, email: tim.cogley@nyu.edu. Dick Startz: Department of Economics, 2127 North Hall, University of California, Santa Barbara, CA 93106, email: startz@econ.ucsb.edu. We acknowledge support from the Center for Scientific Computing at the CNSI and MRL: and NSF MRSEC (DMR-1121053) and NSF CNS-0960316.

Introduction

Near root cancellation can lead to misleading inference for ARMA models in both frequentist and Bayesian frameworks. The $\mathrm{ARMA}(p-1, q-1)$ model $\Phi_{p-1}(L)y_t = \Theta_{q-1}(L)\epsilon_t$ has an equivalent statement as the $\mathrm{ARMA}(p, q)$ model with an additional factor $(1-\gamma L)$ in both the AR and MA polynomials.

$$(1-\gamma L)\Phi_{p-1}(L)y_t = (1-\gamma L)\Theta_{q-1}(L)\epsilon_t \tag{1}$$

In equation (1), γ is not identified. Unfortunately, when roots cancel—or nearly cancel—maximum-likelihood estimation gives spuriously precise estimates of ARMA coefficients. For instance, Ansley and Newbold (1980, p. 181) write "…severe problems with these estimators arise in mixed models when, if there is anything approaching parameter redundancy, [the usual maximum likelihood interval estimators]…can be far too narrow." As an example, Nelson and Startz (2007, Table 1) report on a Monte Carlo experiment in which an ARMA(1,1) model is fit to data generated by an AR(1) process with autoregressive parameter $\phi_1 = 0.01$. Even in a large sample (1000 observations), the standard Wald-type *t*-test against a moving average coefficient of zero, $\theta_1 = 0$, had an actual size of 45.7 percent for nominal size of 5 percent.

The robust estimation procedure we set out in this paper works by mixing[1] an unrestricted $\mathrm{ARMA}(p, q)$ representations with a constrained $\mathrm{ARMA}(p-1, q-1)$ model that enforces an exact common factor $\phi_p = -\theta_q$ on higher-order lags. Since ϕ_p and θ_q are unidentified in the latter representation, their posteriors are the same as their prior, which

[1] To be clear, we don't want to do model selection. There are two reasons for this. First, doing model selection introduces a pre-test bias in the final results; as a practical matter this bias is rarely corrected. Second, in the case of near root cancellation the $\mathrm{ARMA}(p, q)$ actually is the "true" model but it has very bad sampling properties.

assigns unit mass to the unidentified ridge in $\mathrm{ARMA}(p,q)$ space. A Bayesian mixture therefore combines the posterior for the unrestricted $\mathrm{ARMA}(p,q)$ representation with the prior on $\gamma = \phi_p = -\theta_q$. When data are from a well-identified DGP, the posterior probability on the unrestricted model is close to 1, and the mixture inherits the good properties of standard ML estimators. When data are from a weakly-identified representation, however, the posterior probability on the unrestricted $\mathrm{ARMA}(p,q)$ representation is close to 0, and the posterior mixture resembles the prior on $\gamma = \phi_p = -\theta_q$. In this case, the mixture correctly conveys that the data are uninformative for ϕ_p or θ_q. We present both Bayesian and frequentist mixtures, and find that they work about equally well. In this context, whether you mix seems to be more important than how you mix.

As a motivating example, we study returns on the S&P 500 stock index. Asset pricing models typically imply that stock returns should be a martingale difference. In the first column of Table 1, we provide the maximum likelihood estimate of an $\mathrm{ARMA}(1,1)$ model fit to 658 monthly observations on the return on the S&P 500.[2,3] The parameters ϕ_1 and θ_1 are not identified under the null hypothesis that stock returns are a white-noise process. Despite that, the maximum likelihood estimate makes it appear that they are quite well identified. The point estimates are -0.67 and 0.75, respectively, with reported asymptotic standard errors of 0.18 and 0.16, and the likelihood ratio statistic against the $\mathrm{ARMA}(0,0)$ null is 7.6 with a *p*-value of

[2] Throughout we use conditional maximum likelihood, dropping the first p observations. Since T is large and the data is not close to having a unit root, the difference between conditional and exact maximum likelihood should be small. We ignore the considerable heteroskedasticity in the stock return data used for the illustration; both reported coefficients are significant at the 0.05 level using Huber-White standard errors.

[3] The return is defined as the demeaned value of $\log(P_t) - \log(P_{t-1})$, where P_t is the first observation in the month from the St. Louis Federal Reserve Economic Data (FRED) series SP500.

0.02, which seems rather convincingly to reject the white-noise hypothesis. Still, the point estimates make us suspect (near) root cancellation, and Ansley and Newbold (1980) warn us to distrust conventional asymptotic approximations in cases like this. ML estimates for the ARMA parameters are unsatisfactory because they suggest that we know ϕ_1 and θ_1 quite precisely—despite strong theoretical reasons to believe the parameters are not well-identified.

	Maximum Likelihood $\mathrm{ARMA}(1,1)$	Bayesian $\mathrm{ARMA}(1,1)$	Maximum Likelihood $\mathrm{ARMA}(0,0)$	Bayesian $\mathrm{ARMA}(0,0)$	Maximum Likelihood Mixture	Bayesian Mixture
ϕ_1	-0.67 (0.18)	-0.51 (0.20)			-0.05 (0.58)	-0.02 (0.58)
θ_1	0.75 (0.16)	0.60 (0.18)			0.04 (0.59)	0.04 (0.58)
$\lambda = \log \sigma_\epsilon^2$	-6.29 (0.06)	-6.28 (0.06)	-6.28 (0.06)	-6.28 (0.06)	-6.29 (0.06)	-6.29 (0.06)
log likelihood	1134.18		1130.38			
log marginal likelihood (Chib and Jeliazkov method)		1123.9		1126.8		
log marginal likelihood (Laplace approximation)		1114.0		1123.5		
log marginal likelihood (Schwarz approximation)		1124.4		1127.1		
95% HPD		(-0.86,0.07)			(-1.0,1.0)	(-1.0,1.0)

Note: Parameter estimates are maximum likelihood estimates and Bayesian posterior means, respectively. Standard errors in parentheses. For the mle, standard errors are taken from the last step of the Gauss-Newton conditional mle regression. For the Bayesian estimates, standard errors are numerical standard deviations from the sampled posterior. The 95% HPD gives the bounds of the 95% highest posterior density.

ARMA Estimates of Monthly S&P500 Returns

Table 1

What would we *like* to see for a confidence set, whether frequentist or Bayesian, when an $\mathrm{ARMA}(p,q)$ model is fit to $\mathrm{ARMA}(p-1,q-1)$ data? Presumably, the confidence set should reflect ignorance of the highest order ARMA terms, except possibly a restriction that the additional roots lie in the stationary and invertible regions. For instance, if we have

distributions for $f_{\Phi(\mathrm{p}-1)}(\Phi_{p-1})$ and $f_{\Theta(q-1)}(\Theta_{q-1})$, we would like the higher order lag polynomials, $\tilde{f}_{\Phi(\mathrm{p})}(\Phi_p)$ and $\tilde{f}_{\Theta(\mathrm{q})}(\Theta_q)$, to reflect the factors $(1-\gamma L)\Phi_{p-1}$ and $(1-\gamma L)\Theta_{q-1}$, $\gamma \sim U(-1,1)$, where $U(\cdot)$ indicates a uniform prior distribution.[4] At the same time, if the AR and MA terms are strongly identified, we want the conventional estimates to be left alone.

We propose Bayesian and frequentist techniques that achieve these objectives. When a near common factor is present, our methods correctly signal that the data are uninformative for weakly-identified parameters. On the other hand, when a model is well-identified, our methods return the usual parameter distributions. Confidence sets and distributions of ARMA parameters therefore behave appropriately in both well-identified and poorly-identified situations.

Our methods are based on mixture models for $\mathrm{ARMA}(p,q)$ and $\mathrm{ARMA}(p-1,q-1)$ specifications. For a Bayesian implementation, the mixture probability is computed using a Bayes factor. For a frequentist approach, the models are weighted in accordance with the Schwartz information criterion (SIC) approximation to the Bayes factor. Both seem to work well; the Bayesian approach is slightly more costly to compute.

[4] More generally, if informative prior information on γ is available, the posterior on γ should resemble the prior.

Bayesian implementation

Our basic idea is to construct posteriors for both the $\text{ARMA}(p-1, q-1)$ and $\text{ARMA}(p, q)$ models and then mix the two using posterior model probabilities. We begin by using simulation to draw posteriors for the two specifications using a variant of the Chib and Greenberg (1994) algorithm.[6] Assuming Gaussian errors with $\log \text{var}(\epsilon) = \lambda$, we approximate the log likelihood function as in Harvey (1993, p. 62),

$$\ell(\Phi, \Theta, \lambda) = -\frac{T-p}{2}\log(2\pi) - \frac{T-p}{2}\lambda - \frac{1}{2\exp(\lambda)}\sum_{t=p+1}^{T} e_t^2 \tag{2}$$

$$e_t = y_t - \phi_1 y_{t-1} - \cdots - \phi_p y_{t-p} - \theta_1 e_{t-1} - \cdots - \theta_q e_{t-q}$$
$$t = p+1, \ldots, T \quad e_p = e_{p-1} = e_{p-q+1} = 0$$

We choose normal priors, $p(\psi)$, $\psi = \{\Phi, \Theta, \lambda\}$, that are fairly non-informative, but that keep most of the weight in the stationary and invertible regions. The prior means for ϕ and θ are 0_p and 0_q. The prior mean for λ is -6, which is roughly the log variance of the unconditional S&P 500 return. Prior variances for ϕ and θ are $4 \cdot p \cdot I_p$ and $4 \cdot q \cdot I_q$, which in this example leaves the $(-1,1)$ stationary and invertible regions within $\pm\frac{1}{2}$ standard deviation from the prior mean. The prior variance for λ is 2^2.

We simulate the posterior using an independence chain Metropolis-Hastings algorithm with a candidate generating density that is normal around the maximum-likelihood estimates, truncated to the stationary and invertible regions. Specifically we estimate $\psi'_{mle} = \{\phi'_{mle}, \theta'_{mle}, \lambda_{mle}\}$, obtaining ϕ_{mle} and θ_{mle} by grid search and letting

[6] We simplify Chib and Greenberg by dropping the mean function, conditioning on initial observations, and modeling $\text{var}(\epsilon)$ as log normal rather than inverse gamma.

$\lambda_{mle} = \log\left(\frac{1}{T-p}\sum_{t=p+1}^{T} e_t^2\right)$. We take the variance-covariance for the ARMA parameters, $V_{\phi,\theta}$, from the Jacobean, $V_{\phi,\theta} = \exp(\lambda)\left[\frac{\partial \ell}{\partial\{\phi,\theta\}}' \frac{\partial \ell}{\partial\{\phi,\theta\}}\right]^{-1}$, and then let $V_{\psi_{mle}} = \begin{bmatrix} V_{\phi,\theta} & 0 \\ 0 & \frac{2}{T-p} \end{bmatrix}$.

The Metropolis-Hastings (MH) algorithm proceeds in the following steps.

1. Draw a candidate $\psi^{(s)}$ from $N\left(\psi_{mle}, V_{\psi_{mle}}\right)$. If the draw lies outside the stationary or invertible region, reject and draw again.

2. Compute the MH log acceptance probability

$$\begin{aligned}\log \alpha(\psi^s) = &\left\{\ell\left(\psi^{(s)}\right) + \log p\left(\psi^{(s)}\right) - \log f_n\left(\psi^{(s)}; \psi_{mle}, V_{\psi_{mle}}\right)\right\} \\ &- \left\{\ell\left(\psi^{(s-1)}\right) + \log p\left(\psi^{(s-1)}\right) - \log f_n\left(\psi^{(s-1)}; \psi_{mle}, V_{\psi_{mle}}\right)\right\}\end{aligned} \tag{3}$$

   where the log-likelihood, $\ell(\cdot)$, is given in equation (2) and $f_n(\cdot)$ gives the normal pdf.[7]

3. Accept $\psi^{(s)}$ with probability $\min\left(\exp\left(\log \alpha\left(\psi^{(s)}\right)\right), 1\right)$, otherwise $\psi^{(s)} = \psi^{(s-1)}$.

For the stock return example in Table 1, the acceptance probability was 57 percent for an $\mathrm{ARMA}(1,1)$ specification and 98 percent for an ARMA(0,0) model. In each case 55,000 values of ψ were drawn, the first 5,000 of which were discarded.

We next compute the marginal likelihood following Chib and Jeliazkov (2001). We evaluate the basic marginal likelihood identity, equation (4), at the mean of the Metropolis-Hastings sample, $\bar{\psi}$,

[7] The acceptance ratio accounts for truncation to the stationary and invertible regions implicitly. See Appendix available from the authors.

$$\log p(y) = \ell(\bar{\psi}) + \log p(\bar{\psi}) - \log p(\bar{\psi}|y) \tag{4}$$

The first two terms in equation (4) are computed directly. The posterior density is computed as follows.

1. The posterior kernel for draw (s) is $\log p_k(\psi^{(s)}) = \ell\left(\psi^{(s)}\right) + \log p\left(\psi^{(s)}\right)$. The posterior kernel for $\bar{\psi}$ is $\log p_k(\bar{\psi}) = \ell(\bar{\psi}) + \log p(\bar{\psi})$.

2. Compute the numerator acceptance probability

$$\begin{aligned} \log a_N^{(s)} = \min\big(\{&\log p_k(\bar{\psi}) - \log f_n\left(\bar{\psi}; \psi_{mle}, V_{\psi_{mle}}\right)\} \\ &- \left\{\log p_k\left(\psi^{(s)}\right) - \log f_n\left(\psi^{(s)}; \psi_{mle}, V_{\psi_{mle}}\right)\right\}, 0\big) \end{aligned} \tag{5}$$

3. Draw $\psi^{(s_2)}$, $s_2 = 1, \ldots, S$ draws from the candidate density $N\left(\psi_{mle}, V_{\psi_{mle}}\right)$.

4. The posterior kernel for draw (s_2) is $\log p_k(\psi^{(s_2)}) = \ell\left(\psi^{(s_2)}\right) + \log p\left(\psi^{(s_2)}\right)$.

5. Compute the denominator acceptance probability

$$\begin{aligned} \log a_D^{(s_2)} = \min\big(\{&\log p_k\left(\psi^{(s_2)}\right) - \log f_n\left(\psi^{(s_2)}; \psi_{mle}, V_{\psi_{mle}}\right)\} \\ &- \left\{\log p_k(\bar{\psi}) + - \log f_n\left(\bar{\psi}; \psi_{mle}, V_{\psi_{mle}}\right)\right\}, 0\big) \end{aligned} \tag{6}$$

6. Compute the posterior density

$$p(\bar{\psi}|y) = \frac{\frac{1}{S}\sum\left(a_N^{(s)} \times f_n\left(\bar{\psi}; \psi_{mle}, V_{\psi_{mle}}\right)\right)}{\frac{1}{S}\sum a_D^{(s_2)}} \tag{7}$$

As a computationally convenient alternative to equation (7), the marginal likelihood can also be approximated in a large sample by a Laplace approximation,[8]

[8] See Kass and Raftery (1995) for discussion and references.

$$\log \tilde{p}(y) = \ell(\psi_{mle}) + \log p(\psi_{mle}) + (p+q+1)\log 2\pi + \frac{1}{2}\log\left|V_{\psi_{mle}}\right| - \frac{p+q+1}{2}\log T \tag{8}$$

However, $\tilde{p}(y)$ may be a poor approximation for a weakly-identified model. Since the asymptotic variance matrix is nearly singular, $\log\left|V_{\psi_{mle}}\right|$ is likely to be inaccurate.

Assuming even prior odds on the two representations, the mixing probability in favor of the more parsimonious model, p_0, ($\mathrm{ARMA}(0,0)$ in preference to $\mathrm{ARMA}(1,1)$ in our example) is given by[9]

$$p_0 = \frac{BF_{01}}{1+BF_{01}}, \qquad BF_{01} = \frac{p(y|M_0)}{p(y|M_1)}. \tag{9}$$

Having run MCMC samplers on both $\mathrm{ARMA}(p-1,q-1)$ and $\mathrm{ARMA}(p,q)$ models, we numerically mix the posterior distributions.

1. With probability $1-p_1$, draw with replacement $\{\Phi_p, \Theta_q\}$ from the $\mathrm{ARMA}(p,q)$ sample.

2. With probability p_1, draw with replacement $\{\Phi_{p-1}, \Theta_{q-1}\}$ from the $\mathrm{ARMA}(p-1,q-1)$ sample. Draw $\gamma \sim U(-1,1)$.[10] Generate the draw the coefficients of the AR polynomial $\tilde{\phi}_1 = \phi_1 + \gamma, \tilde{\phi}_i = \phi_i - \phi_{i-1}\gamma, i = 2, \ldots, p$. Generate the coefficients of the MA polynomial $\tilde{\theta}_i = \theta_i - \theta_{i-1}\gamma, i = 1, \ldots, q$. (For $p = 0, \phi_1 = 0$ and for $q = 0, \theta_1 = 0$.)

[9] One could assign uneven prior odds if desired.

[10] Here we assume that p(γ) is $U(-1,1)$ and independent a priori from the other parameters of the ARMA(p-1,q-1) model. It follows that the posterior for the ARMA(p-1,q-1) model is $\mathrm{p}(\Phi_p, \Theta_q|Y) = \mathrm{p}(\Phi_{p-1}, \Theta_{q-1}|Y)p(\gamma)$. We adopt a uniform prior in part to maintain consistency with the frequentist results that follow and in part for simplicity.

In the example in Table 1, the ARMA(0,0) specification has a posterior probability of 0.95. The Bayes factor computed using the Laplace approximation differs considerably from the Chib-Jeliazkov calculation (which is Monte Carlo consistent and thus preferred aside from computational costs), primarily because of differences in the marginal likelihood calculation for the $\mathrm{ARMA}(1,1)$ model. In this particular application the difference is of little consequence, as p_0 would be estimated to be 0.9999 rather than 0.95.

Figure 1 shows three distributions for ϕ_1 for the $\mathrm{ARMA}(1,1)$ model for stock returns. The maximum likelihood estimate concentrates spuriously around the negative point estimate. The Bayesian estimate, although influenced by mildly informative priors pulling the distribution toward zero, is also quite concentrated around negative values. In contrast, the mixture model shows that ϕ_1 is effectively unidentified. Because the weight on the ARMA(0,0) model is close to 1, the posterior mixture on $\phi_1$is essentially the same as the prior on γ.

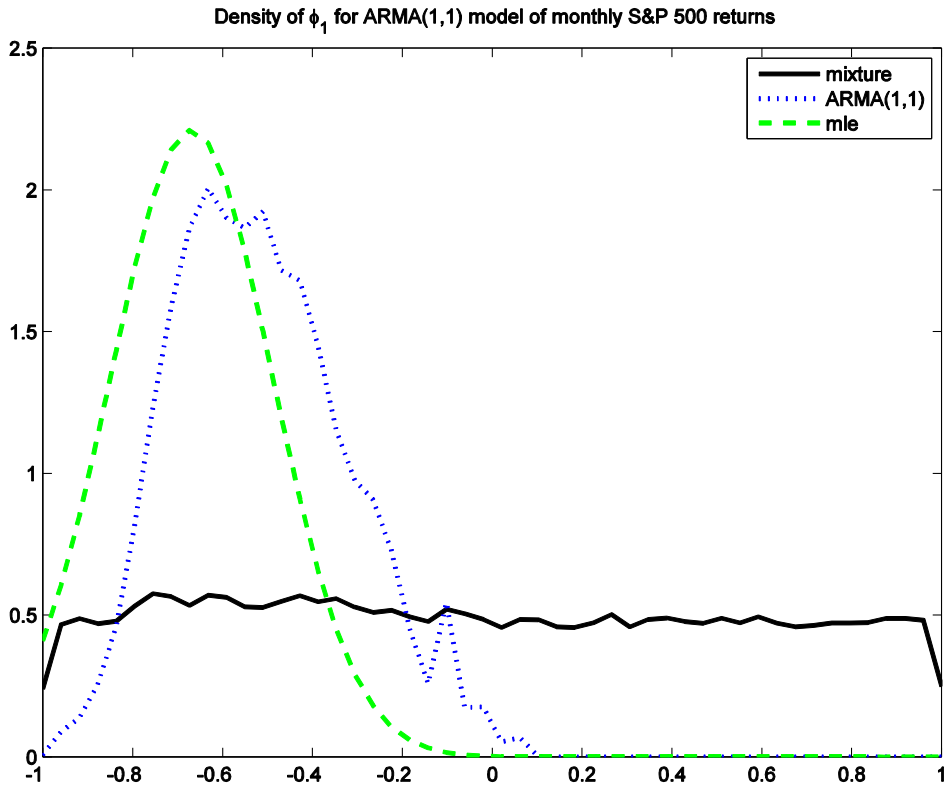

Figure 1

Frequentist implementation

The same procedure can be deployed in a maximum-likelihood framework either to economize on the need for MCMC runs or simply because the investigator prefers a frequentist approach. In a sufficiently large sample, Bayesian and frequentist estimators coincide. More interestingly, we find that for the size samples often used in ARMA estimates and with relatively diffuse priors, the Bayesian and frequent results of our procedure are quite similar.

The need for priors and MCMC simulations to estimate the marginal likelihood can be eliminated by use of the Schwarz criterion

$$\log p(y) \approx S = \ell(\psi_{mle}) - \frac{p+q+1}{2}\log T \tag{10}$$

The Schwarz information criterion (also called the Bayesian information criterion (BIC)), which equals minus twice the difference in the Schwarz criterion, is often used as a frequentist model selection technique. Rather than selecting a model, we use S to compute the frequentist mixture probability.[11] Writing the log likelihood from the $\mathrm{ARMA}(p-1, q-1)$ and $\mathrm{ARMA}(p, q)$ as ℓ_0 and ℓ_1 respectively, we can write the frequentist approximations to the Bayes factor and mixing probability as

$$BF_{01} = \exp(\ell_0 - \ell_1 + \log T), \quad p_0 = \frac{BF_{01}}{1 + BF_{01}} \tag{11}$$

The mixing procedure is the same as given above for the Bayesian case, except that draws are taken from the asymptotic distribution for the maximum likelihood estimators, $N\left(\psi_{mle}, V_{\psi_{mle}}\right)$ instead of from the posterior draws.

For the S&P example, the frequentist mixture probability is 0.94, and the mixture distribution shown in Figure 2 correctly reflects the lack of identification. The rightmost two columns of Table 1 show that frequentist and Bayesian mixture distributions are essentially identical for this example, both being close to $U(-1,1)$.[12]

[11] Hannan (1980) shows that selecting with the BIC gives a consistent estimate of the true order of an ARMA model.

[12] The HPD intervals reported in Table 1 are computed by binning draws in a histogram, collecting the bins that contain the highest 95 percent of the draws, and then recording the edges of the highest and lowest bins in the set. In principle, the HPD need not be continuous so looking at these bounds somewhat exaggerates the width of the HPD. If the posterior were literally $U(-1,1)$, then the HPD would not be unique. If the posterior is close to uniform, then the exact location and width of the HPD somewhat random with respect to the simulation results. Thus our reported HPDs are slightly wider than expected.

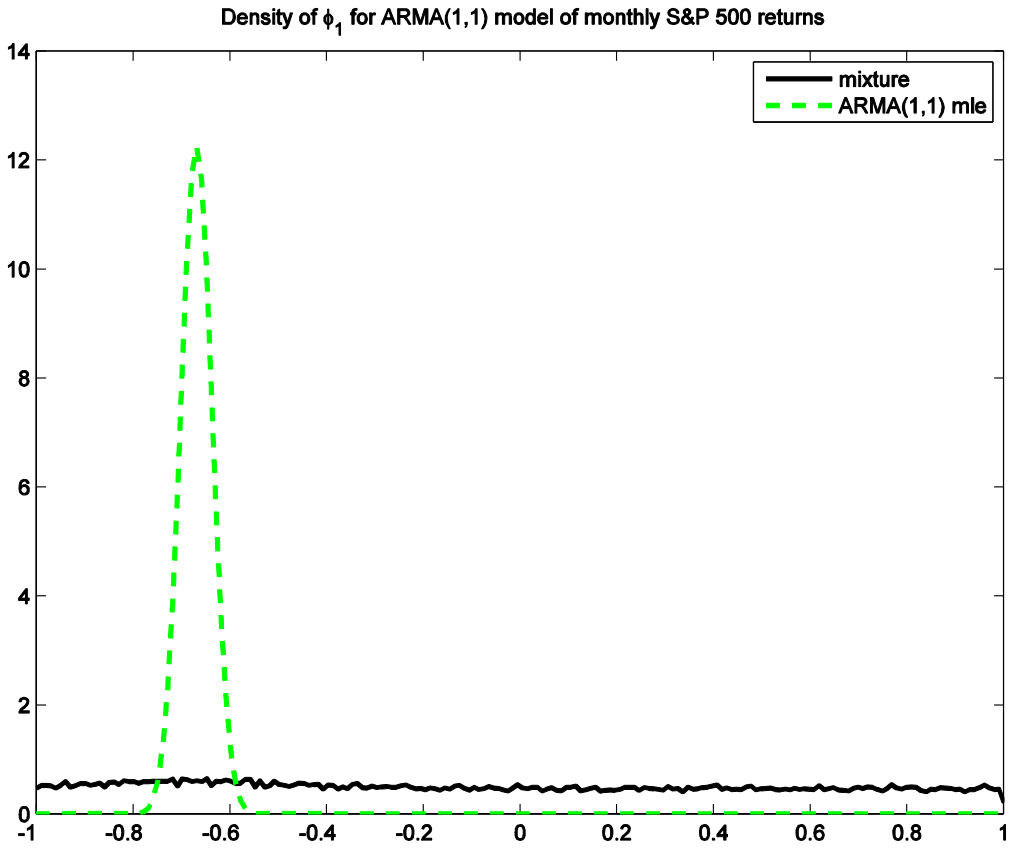

Figure 2

Monte Carlo results

For a well-identified problem, our procedure should closely mimic standard results. For a weakly-identified problem, our procedure should indicate that we are largely ignorant about the location of parameters. For the S&P 500 example, the results indicate the latter.

In this section we present Monte Carlo results for three data generating processes, one well-identified, one weakly-identified, and one twixt well- and weakly-identified. The three DGPs are $\mathrm{ARMA}(1,1)$ models. In all cases, the moving average coefficient θ is set to 0. The autoregressive parameter ϕ is set equal to 0.9, 0.2, and 0.01, respectively, for strongly-, twixt,

and weakly-identified DGPs. Asymptotically, the models are all identified—the issue is whether the conventional asymptotic approximation for strongly-identified representations is reliable in finite samples.

We report the results of 1,000 Monte Carlo trials. Data series of length $100 + T$ were generated, and the 100 initial burn-ins were discarded, for samples of size $T \in \{50, 100, 200, 500, 1000, 5000, 10000\}$. As a first step, we computed standard maximum-likelihood estimates to reconfirm the Ansley and Newbold's (1980) statement.[13] Figure 3 shows 90 percentile bands for the reported standard error of ϕ_{mle}, as well as the asymptotic standard error. For the well-identified, $\phi = 0.9$, model in the bottom panel, asymptotic approximations work well even for small ($T = 50$) samples, and asymptotic and the median reported standard errors are essentially the same (0.068 and 0.078 respectively). For the weakly-identified model, the asymptotic approximation is off by a factor of three even at $T = 10{,}000$—at $T = 200$ the asymptotic standard error is 56 times the reported median. As expected, $\phi = 0.2$ gives results in-between the other two models.

[13] We computed the likelihood function on a coarse 40,000 point grid for $\phi \in \{-0.99, 0.99\}$ and $\theta \in \{-0.99, 0.99\}$ with a resolution of 0.01 to obtain initial estimates ϕ^C and θ^C. We then searched on a finer grid $\phi^C \pm .01$ and $\theta^C \pm .0099$ with a resolution of 0.001 to obtain our final estimates.

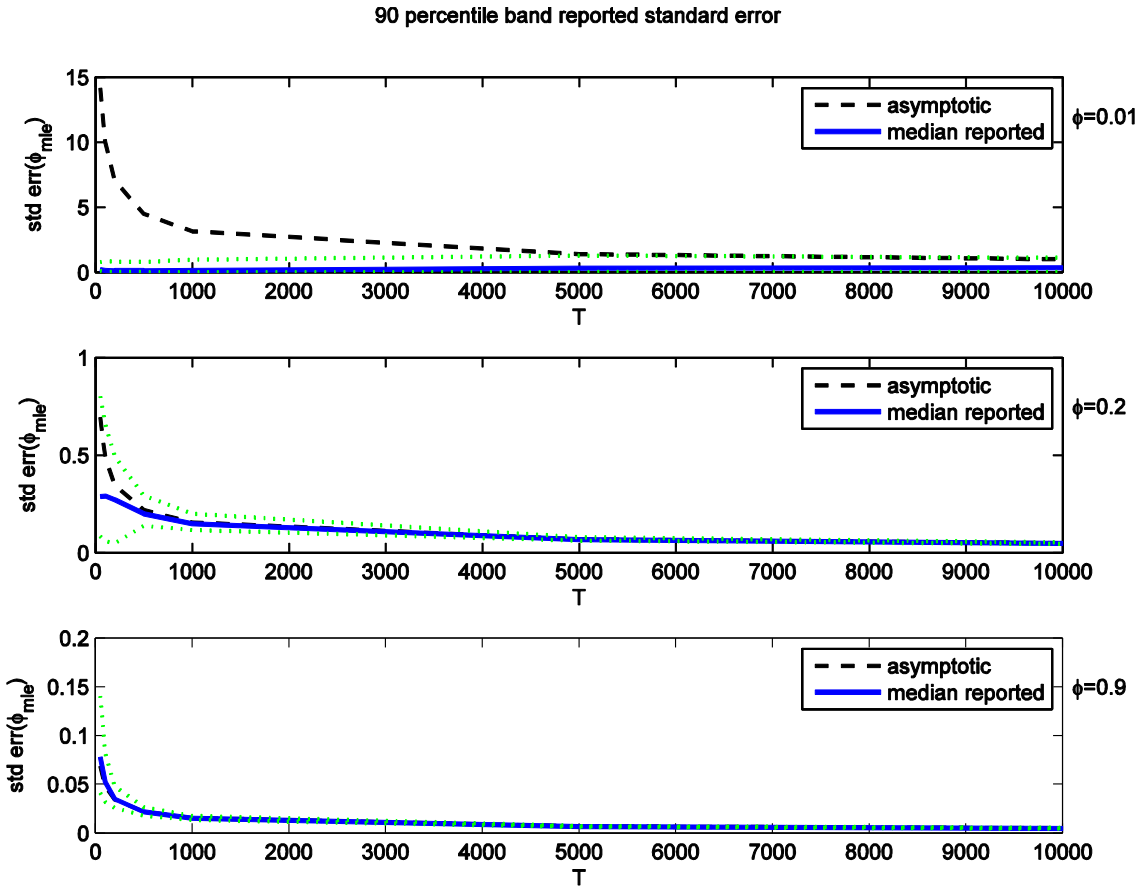

Figure 3

Figure 4 gives the empirical size of nominal five percent Wald tests on ϕ for each data generating process. In the well-identified model, nominal and actual size are approximately equal for all our sample lengths. In contrast, the weakly identified model rejects the true value 66 percent of the time at $T = 50$ and 41 percent of the time even at $T = 10{,}000$. The middle model performs quite badly at modest sample sizes, rejecting 37 percent of the time at $T = 100$, but by $T = 10{,}000$ achieves a 7 percent size.

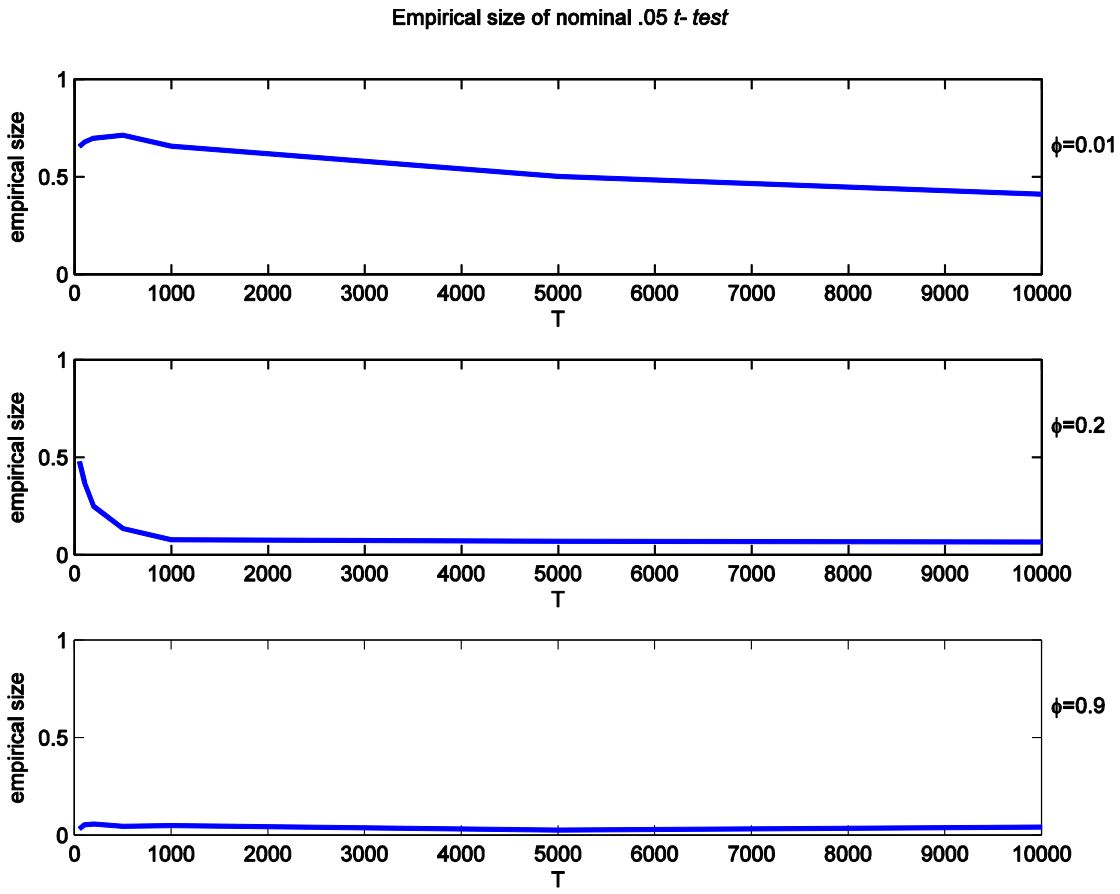

Figure 4

The size results suggest that we would like largely to ignore the estimated results for the weakly-identified model. At the same time, the estimated results for the well-identified model should be left alone. In betwixt, we should down-weight estimation results at small sample sizes but lean heavily on the estimated results with large samples.

The large sample approximation to the mixing probability based on the Schwarz criteria is easily calculated for maximum likelihood models. Figure 5 provides 90 percentile bands from our Monte Carlo. In the weakly-identified model, the median value of p_0 (the weight on the $\mathrm{ARMA}(0,0)$ representation) ranges from 94 percent in the smallest sample to 99.9 percent at $T = 10{,}000$. In contrast, for $\phi = 0.9$ p_0 is always zero for all practical purposes.

The median mixing probability for the middle model is above 85 percent through $T = 100$, although there is considerable variation around the median. At $T = 500$, median p_0 is close to zero. These middle results suggest that our proposed mixture greatly improves frequentist inference, although not completely achieving the asymptotic size.

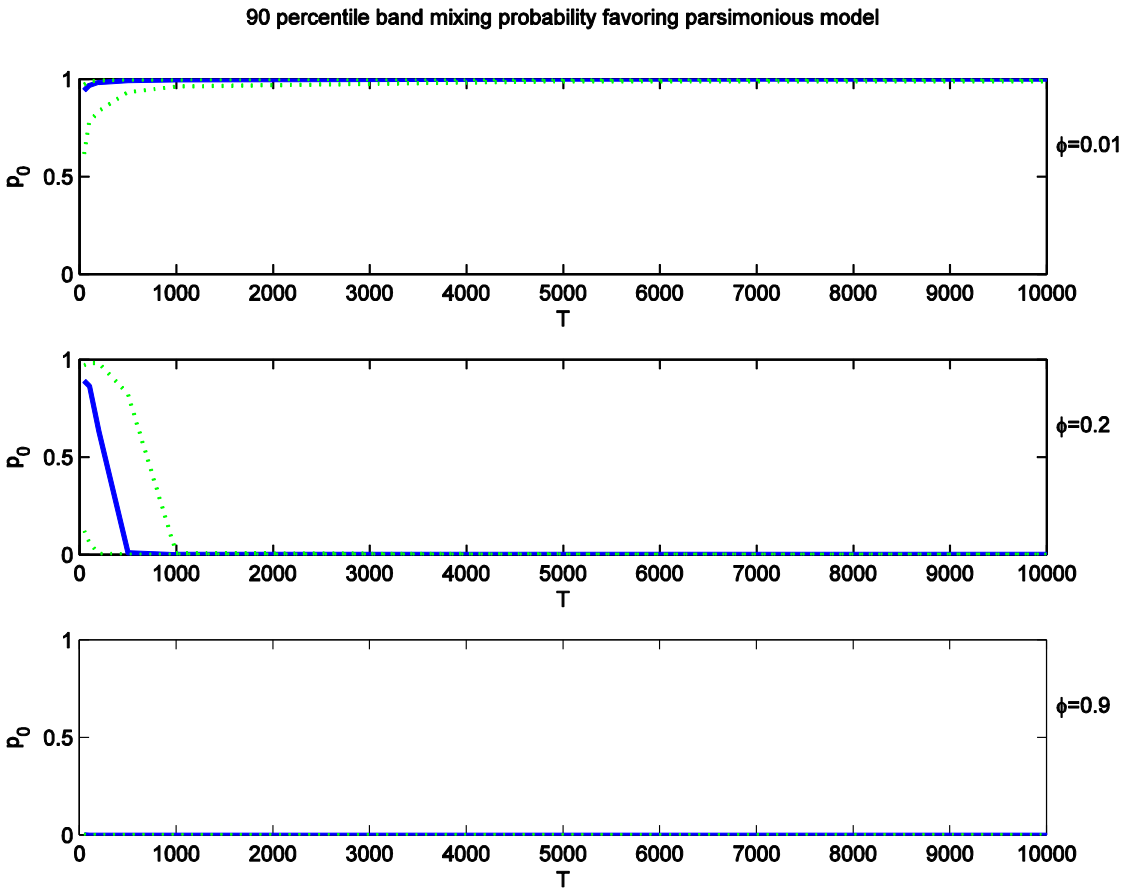

Figure 5

Turn now from examination of the frequentist approximation at varying sample sizes to a more detailed Monte Carlo examination of both Bayesian and frequentist approaches at $T = 200$. Note that since the asymptotic variance for ϕ_{mle} is $\frac{(1-\phi^2)}{T\phi^2}$, the asymptotic standard errors are 0.034,0.346, and 7.07, for $\phi_1 = 0.9, \phi_1 = 0.2$, and $\phi_1 = 0.01$, respectively.

We begin with the well-identified, $\phi = 0.9$, DGP. The results are summarized simply by saying that everything is well-behaved. Figure 6 shows the density of the Monte Carlo draws for ϕ_{mle}. The density is what one would expect from asymptotic distribution theory. The upper panel of Table 2 gives more detailed results. Note in particular that according to the MLE results given in the first column, the asymptotic standard deviation agrees with both the Monte Carlo standard deviation and the median reported standard deviation. Furthermore, the actual size in the Monte Carlo for a *Z*-score test against $\phi = 0.9$ matches the nominal size.

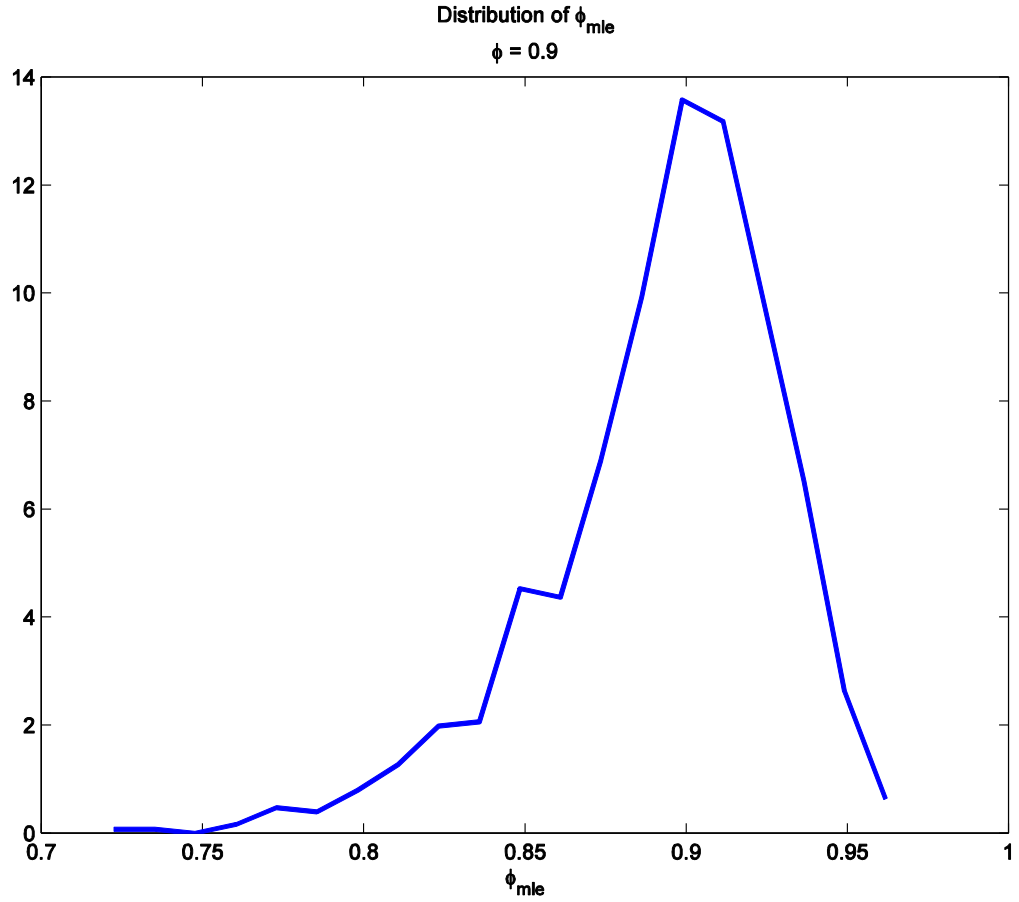

Figure 6

The second column of Table 2 provides results for the Bayesian estimates, which are essentially the same as the maximum likelihood estimates, as would be expected from a well-

identified model with a relatively noninformative prior. The 95 percent HPD is very slightly wider for the Bayesian estimate, presumably reflecting the priors having been centered at zero.[14] A 'Bayesian' Z-score[15] test of $\phi = 0.9$ has the correct frequentist size, rejecting the null hypothesis at a nominal 5 percent level in 5 percent of the samples.

The bottom three rows of the upper panel of Table 2 give the median values of the probability weight on the ARMA(0,0) specification, as well as the outer 2.5 percent tail values. The weight is always zero, so our proposed mixture model simply returns the traditional standard maximum likelihood and Bayesian results. The choice of Bayes factor calculation by Chib and Jeliazkov versus either of the approximations in this well-identified model makes no difference. The minor differences between the original and mixture HPD intervals simply reflect sampling with replacement from the original distributions.

To summarize the Monte Carlo results for the well-identified model, the standard procedures work well, and since our proposed modification replicates the standard procedure, our mixture model also works well. Robustness costs the investigator nothing other than a few seconds of computer time.

[14] The maximum likelihood "HPD" is calculated as $2 \times 1.96 \times median(std.dev.\phi_{mle})$.

[15] The 'Bayesian' Z-score is a conventional Z-score computed using the posterior mean and standard deviation.

	MLE	Bayes	MLE Mixture	Bayes Mixture
$\boldsymbol{\phi = 0.9}$ $asymp.std.dev.(\phi_{mle}) = 0.034$				
Monte Carlo std. dev.	0.037	0.037	0.037	0.037
Mean reported std. dev.	0.036	0.036	0.036	0.036
Median reported std. dev.	0.035	0.035	0.035	0.035
empirical size, 5 percent nominal test	0.048	0.050	0.050	0.050
Median 95% HPD width	0.138	0.181	0.150	0.185
{0.025, Median, 0.975} percentile mixture probability (Chib and Jeliazkov method)	$\{3.32 \times 10^{-92}, 3.55 \times 10^{-68}, 5.25 \times 10^{-45}\}$			
{0.025, Median, 0.975} percentile mixture probability (Laplace approximation)	$\{6.91 \times 10^{-90}, 7.18 \times 10^{-66}, 1.18 \times 10^{-42}\}$			
{0.025, Median, 0.975} percentile mixture probability (Schwarz approximation)	$\{5.01 \times 10^{-93}, 5.63 \times 10^{-69}, 1.16 \times 10^{-45}\}$			
$\boldsymbol{\phi = 0.2}$ $asymp.std.dev.(\phi_{mle}) = 0.346$				
Monte Carlo std. dev.	0.455	0.367	0.158	0.158
Mean reported std. dev.	0.277	0.257	0.464	0.464
Median reported std. dev.	0.265	0.257	0.526	0.526
empirical size, 5 percent nominal test	0.256	0.179	0.015	0.015
Median 95% HPD width	1.04	1.31	2.00	2.00
{0.025, Median, 0.975} percentile mixture probability (Chib and Jeliazkov method)	{0.002,0.702,0.999}			
{0.025, Median, 0.975} percentile mixture probability (Laplace approximation)	{0.295,0.998,1.0}			
{0.025, Median, 0.975} percentile mixture probability (Schwarz approximation)	{0.001,0.631,0.989}			
$\boldsymbol{\phi = 0.01}$ $asymp.std.dev.(\phi_{mle}) = 7.07$				
Monte Carlo std. dev.	0.801	0.663	0.024	0.024
Mean reported std. dev.	0.264	0.209	0.577	0.576
Median reported std. dev.	0.134	0.158	0.577	0.577
empirical size, 5 percent nominal test	0.703	0.611	0	0
Median 95% HPD width	0.525	0.652	2.00	2.00
{0.025, Median, 0.975} percentile mixture probability (Chib and Jeliazkov method)	{0.89,0.99,1.00}			
{0.025, Median, 0.975} percentile mixture probability (Laplace approximation)	{1.00,1.00,1.00}			
{0.025, Median, 0.975} percentile mixture probability (Schwarz approximation)	{0.74,0.98,0.99}			

Monte Carlo Results

Table 2

Turn now to the Monte Carlo results for the weakly-identified, $\phi = 0.01$, model in the bottom panel of Table 2. Unfortunately, as in the previously cited literature, the maximum likelihood estimator often understates the uncertainty associated with the estimate of ϕ. The median reported standard error is only 0.13, while the Monte Carlo standard deviation is 0.80. The bottom line is that a nominal 5 percent test has an actual size equal to 70 percent.

Results for the Bayesian estimator—shown in the second column of the lower panel of Table 2—are similarly unsatisfactory. The actual size is 61.1 percent.[16] Weak identification occurs when ϕ is close to $-\theta$. The prior for the $\mathrm{ARMA}(p, q)$ model assigns very little probability to this ridge in the parameter space, essentially limiting the $\mathrm{ARMA}(p, q)$ posterior to the well-identified region in which ϕ is not close to $-\theta$.[17] Hence the $\mathrm{ARMA}(p, q)$ posterior inherits the defects of mle. Ironically, a seemingly uninformative prior on the parameters of the $\mathrm{ARMA}(p, q)$ specification encodes a strong prior belief that the model is well identified.

One way to counterbalance this strong-identification prior is by mixing the $\mathrm{ARMA}(p, q)$ model with another that assigns a heavier weight of prior probability to the $\phi = -\theta$ ridge. That is what the $\mathrm{ARMA}(p-1, q-1)$ model does. Priors for the $p-1$ AR and $q-1$ MA terms are weakly informative, but the prior on γ enforces exact equality between ϕ_p and $-\theta_q$. Hence the prior for the $\mathrm{ARMA}(p-1, q-1)$ specification assigns a prior mass of 1 to the unidentified ridge in $\mathrm{ARMA}(p, q)$ space. By mixing the two models, we strike a balance between strong and weak identification.

[16] Since our prior is centered at zero and has some influence on the posterior, the slight improvement over the mle doesn't really reflect information from the data.

[17] Hannan (1980) points to this ridge as the source of inconsistency of MLE estimates.

For the example in Table 2, the median mixture probability places 99 percent of the weight on the augmented $\mathrm{ARMA}(0,0)$ model, resulting in a posterior for ϕ which is essentially $U(-1,1)$ in a given Monte Carlo run. The Bayesian and maximum likelihood mixture models give the same results, telling us that the data are uninformative about ϕ. This contrasts with both of the traditional estimators, which are distributed across $(-1,1)$ across Monte Carlo runs but are spuriously tightly distributed within a run.

All three methods of calculating the marginal likelihood do an excellent job of picking the $\mathrm{ARMA}(0,0)$ model (even though the true model is $\mathrm{ARMA}(1,1)$). If we take the Chib and Jeliazkov calculation as the standard, the Schwarz approximation occasionally understates p_0 while the Laplace approximation overstates p_0. The latter reflects the underestimate of $V_{\psi_{mle}}$ in equation (8).

Our twixt results are given in the middle panel of Table 2. As expected from the Monte Carlo run reported earlier, the maximum likelihood estimator performs poorly but not disastrously. Empirical size is one-fourth and the median standard error is too small. The poor size results despite the fact that the width of the median confidence interval covers is relatively large, 1.04. Standard Bayesian results aren't much different. The median mixture probabilities put more of the weight on the $\mathrm{ARMA}(1,1)$ model, but 30 percent of the time draws from the parsimonious model. As a result, the mixture models rarely reject the null and have somewhat larger median HPD widths than the standard estimators. Note that the mixing probabilities from the Schwarz approximation are close to those from Chib and Jeliazkov, but that the Laplace approximation again puts somewhat too much weight on the parsimonious model.

The Monte Carlo results indicate that our proposed mixture returns the standard results for inference for a well-identified data generating process. In contrast, when faced with a weakly identified data generating process the standard procedure indicate spurious precision while the mixture correctly reports our inability to infer the true parameter from the data.

Conclusion

It has long been known that standard estimation of ARMA models in the presence of near root cancellation produces spuriously tight confidence intervals for the estimated coefficients. Our mixture procedure avoids such spurious inference without any significant cost for well-identified models. While our procedure is derived with a Bayesian justification, it seems to work equally well in the maximum-likelihood context. Computation of the Schwarz approximation to the mixing probability works well. For maximum-likelihood the only extra computation is estimation of the $\mathrm{ARMA}(p-1, q-1)$ model and numerical mixing of two normals. We recommend, at a minimum, that mle mixture models be computed when inference about ARMA coefficients is of interest.

References

Ansley, C.F., Newbold, P., 1980. Finite sample properties of estimators for auto-regressive moving average processes, *Journal of Econometrics,* 13, 159–184.

Chib, S. and E. Greenberg, 1994. Bayes Inference in regression models with $\mathrm{ARMA}(p,q)$ errors, *Journal of Econometrics,* 64, 138-206.

_____ and I. Jeliazkov, 2001. Marginal Likelihood From the Metropolis-Hastings Output, *Journal of the American Statistical Association* 96, 270-281.

Hannan, E.J., 1980, The Estimation of the Order of an ARMA Process, *The Annals of Statistics,* 8, 1071-1081.

Harvey, A.C., 1993, *Time Series Models,* 2nd edition, M.I.T. Press, Cambridge, MA.

Kass, R.E. and A.E. Raftery, 1995, Bayes Factors, *Journal of the American Statistical Association,* 90, 773-795.

Nelson, C.R. and R. Startz, 2007. The zero-information-limit condition and spurious inference in weakly-identified models, *Journal of Econometrics* 138, 47-62.

Appendix – Not for publication

In the Metropolis-Hastings algorithm employed in the body of the paper we use a truncated normal proposal density. The truncation is not explicit in the calculation of the proposal density in equation (3). In this appendix we show why the calculations are correct. Our proposal density has probability

$$f_{tn}\left(\psi^{(s)};\psi_{mle},V_{\psi_{mle}}\right)=\frac{I\left(\psi^{(s)}\right)f_n\left(\psi^{(s)};\psi_{mle},V_{\psi_{mle}}\right)}{\int I(\psi)f_n\left(\psi;\psi_{mle},V_{\psi_{mle}}\right)d\psi} \tag{12}$$

where $f_n(\cdot)$ and $f_{tn}(\cdot)$ are normal and truncated normal densities, respectively, and $I(\psi)$ is an indicator function

$$I(\psi)=1 \text{ for admissible draws}$$

$$=0 \text{ otherwise}$$

The denominator of equation (12) is the probability of an acceptable draw. Step 1 in in the Metropolis-Hastings algorithm generates a proposal from this density.

Note further that the denominator, $\int I(\psi)f_n\left(\psi;\psi_{mle},V_{\psi_{mle}}\right)d\psi$, does not depend on the draw, and that since we are using an independence chain Metropolis-Hastings this value is the same for all draws so that

$$\frac{f_{tn}\left(\psi^{(s)};\psi_{mle},V_{\psi_{mle}}\right)}{f_{tn}\left(\psi^{(s-1)};\psi_{mle},V_{\psi_{mle}}\right)}=\frac{\left[\frac{I\left(\psi^{(s)}\right)f_n\left(\psi^{(s)};\psi_{mle},V_{\psi_{mle}}\right)}{\int I(\psi)f_n\left(\psi;\psi_{mle},V_{\psi_{mle}}\right)d\psi}\right]}{\left[\frac{I\left(\psi^{(s-1)}\right)f_n\left(\psi^{(s-1)};\psi_{mle},V_{\psi_{mle}}\right)}{\int I(\psi)f_n\left(\psi;\psi_{mle},V_{\psi_{mle}}\right)d\psi}\right]} \tag{13}$$

$$=\frac{I\left(\psi^{(s)}\right)}{I\left(\psi^{(s-1)}\right)}\times\frac{f_n\left(\psi^{(s)};\psi_{mle},V_{\psi_{mle}}\right)}{f_n\left(\psi^{(s-1)};\psi_{mle},V_{\psi_{mle}}\right)}$$

Since all draws that pass through step 1 are in the acceptance region, $I\left(\psi^{(s)}\right) = I\left(\psi^{(s-1)}\right) = 1$. Therefore in equations (3) and (13) evaluating the candidate density using the normal distribution gives the same value as using the truncated normal.

The analogous argument holds as well for evaluation of the prior densities.