Forecasts in a Slightly Misspecified Finite Order VAR*

Ulrich K. Müller
Department of Economics
Princeton University

James H. Stock
Department of Economics
Harvard University
and NBER

First Version: July 2010
This Version: July 2011

Abstract

We propose a Bayesian procedure for exploiting small, possibly long-lag linear predictability in the innovations of a finite order autoregression. We model the innovations as having a log-spectral density that is a continuous mean-zero Gaussian process of order $1/\sqrt{T}$. This local embedding makes the problem asymptotically a normal-normal Bayes problem, resulting in closed-form solutions for the best forecast. When applied to data on 132 U.S. monthly macroeconomic time series, the method is found to improve upon autoregressive forecasts by an amount consistent with the theoretical and Monte Carlo calculations.

JEL classification: C53; C32; C11

Keywords: Spectral Domain Prior; Posterior Approximation; Information Criteria

*This research was funded in part by NSF grant SBR-0617811 (Stock). We thank participants of workshops at Columbia, Harvard/MIT, Montreal, the Greater New York Econometrics Colloquium and the Cowles Summer Conference for helpful comments, and Adam Clark-Joseph for excellent research assistance.

1 Introduction

Low-order autoregressions provide good benchmark forecasts for many economic time series, yet there is reason to suspect that some small amount of linear predictability remains beyond the initial autoregressive approximation. One indication of this residual predictability is that for macroeconomic data, the autoregressive lag length estimated by the Akaike Information Criterion (AIC) often exceeds the lag length estimated by the Bayes Information Criterion (BIC) by a considerably greater degree than it should if the process were in fact a finite order autoregression. These information criteria estimate the lag length by minimizing the penalized logarithm of the regression sum of squared residuals, where the penalty increases proportionately with the number of lags. In the data set analyzed in Section 4 (132 monthly macroeconomic time series for the United States from 1959:1 – 2003:12), if the series truly are finite-order autoregressions then the AIC lag length should exceed the BIC lag length by 3 or more in only 12.0% of series, whereas empirically AIC exceeds BIC by 3 or more in 61.4% of the series.[1] Table 1 reports the empirical distribution of the difference between the AIC- and BIC-selected lags (maximum lag length $p_{max} = 18$). For this sample size of $T = 510$, the regression R^2 must increase by approximately 1.2 percentage points for BIC to include an additional lag but must increase by only 0.4 percentage points for AIC to include an additional lag. Thus Table 1 shows that, for many series, adding lags increases the regression R^2 by a small but nonzero amount far more frequently than one would expect by random chance alone if the series were in fact finite-order autoregressions. One response to Table 1 might be to adopt the AIC when fitting autoregressions for forecasting; however, as we report in Section 4, for these data the mean squared out-of-sample forecast error of AIC exceeds BIC, at least at short horizons. In short, in these data there appears to be residual predictability beyond the BIC-selected autoregression, but the estimation error introduced by longer autoregressions overwhelms this small residual predictability and degrades the forecasts. Framed in terms of AIC/BIC asymptotics for stationary time series, this suggests the presence of residual predictability that lies in a $1/\sqrt{T}$ neighborhood of no predictability.[2]

[1] If the process is an AR(p_0) and the information criteria are used to choose a lag up to $p_{\max} \geq p_0$, then BIC estimates p_0 consistently and AIC overestimates p_0 by Δp; the overestimation Δp is asymptotically distributed as $\Delta p \sim \arg\max_{0 \leq d \leq p_{\max}-p_0} \sum_{j=1}^{d}(Z_j^2 - 2)$, where Z_j are independent standard normals. This distribution depends only weakly on $p_{\max} - p_0$ if $p_{\max} - p_0$ is large. Unreported results show the small sample distribution for $T = 510$ and Gaussian errors to closely match the asymptotic approximation, at least for an AR(0).

[2] For example, consider the autoregression $y_t = a + \beta_1 y_{t-1} + \cdots + \beta_{p_0} y_{t-p_0} + (b/\sqrt{T}) y_{t-p_0-1} + \varepsilon_t$, where ε_t is serially uncorrelated. Then the BIC-selected lag satisfies $\hat{p}_{BIC} \xrightarrow{p} p_0$ for all finite b, whereas for AIC, $P(\hat{p}_{AIC} > p_0)$ is increasing in b and approaches 1 for b large.

This paper proposes a structure for exploiting this small residual linear predictability. Consider the univariate Gaussian AR(p),

$$\begin{aligned} \beta(L)y_t &= \alpha + u_t \\ &= \alpha + \sigma e_t, \qquad t = 1, \ldots, T, \end{aligned} \tag{1}$$

where $\beta(z) = 1 - \beta_1 z - \cdots - \beta_p z^p$ has its roots outside the unit circle, L is the lagoperator, $\sigma > 0$ and e_t is a mean-zero stationary Gaussian process. We model e_t as having a small amount of residual predictability by supposing that the spectral density of e_t, $f_e(\omega)$, is local-to-flat, specifically,

$$f_e(\omega) = \frac{1}{2\pi} e^{G(\omega)/\sqrt{T}} \tag{2}$$

where $\int_{-\pi}^{\pi} G(\omega)d\omega = 0$. This setup captures the notion that after a prewhitening with a parametric model, there remains residual predictability of order $1/\sqrt{T}$. On the one hand, by Kolmogorov's formula (Brockwell and Davis (1991, p. 191)), the optimal one-step ahead linear predictor of e_t has variance $V[e_t|e_{t-1}, e_{t-2}, \ldots] = 2\pi \exp\left(\int_{-\pi}^{\pi} \ln f_e(\omega)d\omega/2\pi\right) = 1$. On the other hand, ignoring this local predictability yields an error variance of approximately $V[e_t] = \int_{-\pi}^{\pi} f_e(\omega)d\omega \approx \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(1 + G(\omega)/\sqrt{T} + \frac{1}{2}G(\omega)^2/T\right)d\omega = 1 + T^{-1}\frac{1}{4\pi}\int_{-\pi}^{\pi} G(\omega)^2 d\omega$, so that the increase in the mean square forecast error is of order $1/T$, the same order that arises from standard $1/\sqrt{T}$ parameter uncertainty. Also, the $1/\sqrt{T}$ rate in (2) ensures contiguity to the model with white noise u_t (Dzhaparidze (1986), p. 64); this implies in particular that any consistent lag-selection rule for p (such as BIC) remains asymptotically unaffected by the slight misspecification of the AR(p).

The local perturbation G cannot be consistently estimated. Instead we model G as an unobserved continuous Gaussian process on $[-\pi, \pi]$ with a known covariance kernel. The optimal forecast of y_{T+1} under quadratic loss thus becomes a Bayes problem of computing the posterior mean of y_{T+1} under such a prior in the spectral domain. We show that the local-to-flat spectral density assumption provides substantial computational simplifications. In particular, the posterior mean of u_{T+1} is approximately a linear function of the autocovariances of u_t, and the log-likelihood is approximately quadratic in these autocovariances.

Asymptotically, the posterior mean of u_{T+1} thus becomes a standard normal-normal Bayes problem with a straightforward closed form solution. In the special case where $p = 0$ and G is a constant c times demeaned Brownian motion on $[-\pi, \pi]$, the approximate posterior mean for the jth autocovariance is simply the jth sample autocovariance of u_t with a shrinkage factor of $c^2/(c^2 + 2\pi^2 j^2)$. When $p > 0$, $\beta(L)$ needs to be estimated, so only the OLS residuals $\hat{u}_t$ are directly observed. The appropriate shrinkage of the sample autocovariances of $\hat{u}_t$ then involves an additional linear regression step. Formally, we show that with a sample size independent, non-dogmatic prior over $\beta(L)$, and an independent Gaussian process prior on G, this approximation to the posterior mean of y_{T+1} is within $o_p(T^{-1/2})$ of the exact posterior mean.

This result is particularly useful because computation of the exact posterior mean is challenging. Carter and Kohn (1997), Liseo, Marinucci, and Petrella (2001), Choudhuri, Ghosal, and Roy (2004a), McCoy and Stephens (2004), and Rosen, Stoffer, and Wood (2009), among others, consider Bayesian inference in time series models with priors in the spectral domain. They resort to computationally intensive Markov Chain Monte Carlo techniques to obtain posteriors. What is more, their samplers are all based on the Whittle (1957, 1962) approximation of the likelihood. The pseudo-posteriors obtained in this fashion thus contain an approximation error that could be as large as $O_p(1)$.[3]

Alternatively, one might approximate the spectral density prior (2) by a corresponding prior on the coefficients of a long-lag AR(q) with $q \gg p$. Under the demeaned Wiener process specification for G, the resulting prior on the AR coefficients $j = p+1, p+2, \dots$ is approximately independent mean-zero Gaussian with variance proportional to $1/(Tj^2)$. This rate of decay corresponds to the rate of decay the 'Minnesota' prior (cf. Doan, Litterman, and Sims (1984)) imposes for $j = 1, 2, \dots$. For G a demeaned Wiener process, our forecast might thus be interpreted as the approximate posterior under a "local" Minnesota prior on prewhitened data, which is implemented without estimation of a long-lag AR(q).

If the baseline AR model (1) for y_t is locally misspecified for low frequencies, then the function G in (2) has most of its variation close to zero.[4] A large literature considers good forecasting rules in the presence of low frequency phenomena, such as structural breaks or time varying means (cf. Chernoff and Zacks (1964), Clements and Hendry (1998), Pesaran, Pettenuzzo, and Timmermann (2006), for example). Such concerns are seamlessly accommodated in our framework by picking a prior for G with more variation close to the origin. The shrinkage of the sample autocovariances of u_t is then performed in a way that the resulting forecast tracks low frequency movements in the mean of y_t.

From a decision theoretic perspective, a Bayesian approach to forecasting is entirely natural, as the resulting forecast is efficient relative to the prior and thus admissible by con-

[3] In the light of the results of Choudhuri, Ghosal, and Roy (2004b), one would expect that the Whittle approximation induces errors in posterior means of *autocovariances* no larger than $O_p(T^{-1/2})$ for sufficiently smooth spectral densities, but the approximation error for *forecasts* computed from the Whittle likelihood is $O_p(1)$ in general.

[4] The spectral density of y_t in (1) is $f_y(\omega) = \sigma^2 f_e(\omega)/|\beta(e^{\mathbf{i}\omega})|^2$, with $\mathbf{i} = \sqrt{-1}$. Thus, G in (2) equivalently represents a local misspecification in the log spectral density of y_t.

struction. In contrast, in the literature on optimal lag selection for forecasting (Shibata (1980), Ing and Wei (2005), Schorfheide (2005), among others) attention is restricted to the class of OLS forecasts, which might be dominated by another forecasting function.[5] From a more technical perspective, the quadratic approximations to the log-likelihood underlying our results are similar to those employed by Müller and Petalas (2010) in the context of the estimation of parameter time variation of order $1/\sqrt{T}$. Finally, there is an interesting connection to the recent paper by Golubev, Nussbaum, and Zhou (2010): These authors establish a general but non-constructive equivalence between spectral density estimation and estimation of a nonparametric function on the unit interval, observed with Gaussian noise. In the local-to-flat spectrum framework, this link becomes quite explicit: The asymptotically normal-normal Bayes problem for the autocovariances of e_t corresponds to a Gaussian functional estimation problem with Gaussian prior, and the optimal forecast of e_{T+1} in (2) under the belief that G is equal to $\hat{G}$ (where $\int_{-\pi}^{\pi} \hat{G}(\omega)d\omega = 0$) has variance $\int_{-\pi}^{\pi} f_e(\omega)e^{-\hat{G}(\omega)/\sqrt{T}}d\omega \approx 1 + T^{-1}\frac{1}{4\pi}\int_{-\pi}^{\pi}(G(\omega) - \hat{G}(\omega))^2 d\omega$, so that the impact on the mean squared forecast error of estimation error in G becomes asymptotically proportional to the L_2-norm of the estimation error of G.

A heuristic derivation, including the simplifying steps arising from the local-to-flat spectrum assumption, details on the suggested forecast and the formal theoretical result are given in Section 2. The result covers a general VAR model, with unknown intercept and error variance, and ℓ-step ahead forecasts of $y_{T+\ell}$. Under a mixture prior on G (for example, we consider the case that G is c times demeaned Brownian motion and multiple values of c are used), it is possible to combine forecasts by Bayesian model averaging (BMA), and we provide a simple expression for the model averaging weights. Our main result is that difference between the approximate posterior mean of $y_{T+\ell}$ and the exact posterior mean (which is the optimal forecast under quadratic loss) is $o_p(T^{-1/2})$ for all fixed horizons ℓ.

Section 3 reports a Monte Carlo simulation which confirms the main features of the theory. The simulations focus on the case that the process G is c times a demeaned Brownian motion. The cases that the prior is correctly specified are seen to map out an asymptotic efficiency envelope. The BMA forecast comes reasonably close to this envelope. Even if the prior is misspecified, for example if u_t is a finite-order moving average with coefficients of order $1/\sqrt{T}$, the forecasts based on the scaled Brownian motion prior capture much of the additional predictability in the process. Moreover, when there is no predictability, the cost of using the BMA forecasts is found to be small. These findings suggest that the demeaned

[5]For instance, Ing and Wei (2005, Section 4) provide evidence that AIC based OLS forecasts are not in general admissible under squared loss.

Brownian motion prior provides a flexible way to exploit small remaining predictability in autoregressive residuals for general unspecified forms of that predictability.

Section 4 assesses the performance of the smooth-spectrum forecasts via a pseudo out-of-sample forecasting experiment using the U.S. monthly macro data set examined in Table 1. Table 1 suggested that there is a small additional linear predictability, at least in some series. The final 6 columns of Table 1 report the asymptotic distribution of the difference between AIC- and BIC-selected lags in model (2) with G a demeaned Brownian motion of scale c, for a true lag length of $p_0 = 0$ and $p_0 = 3$.[6] At least qualitatively, these differences match the empirical results in the first column quite closely. Accordingly, the approximately optimal forecasts computed from this model with $c = 20$ have median improvements of the mean squared forecast error, relative to BIC, of 1.2% at the one month horizon. For some series, the improvement is substantially greater and for only a few series do we find that this procedure imposes much of a cost, relative to using BIC: the 10% and 90% percentiles of relative mean squared forecast errors, relative to BIC, are 0.964 and 1.014, respectively. The magnitudes of these improvements are also in line with the theoretical expressions in Section 2 and the Monte Carlo results of Section 3.

2 Theoretical Results

This section provides an informal derivation of the approximately best forecast, first for the case of no autoregressive component, then with an autoregressive component. We then provide a detailed description of the suggested forecast in a VAR with unknown mean and scale. The section concludes with a statement of the main theorem, which is proven in the appendix.

2.1 Heuristic Derivation

We begin by obtaining an expression for approximate posterior mean of y_{T+1} in the case that $\beta(L) = 1$, so that $y_t = u_t$ and u_t is observed. For simplicity, we assume $\alpha = 0$ and $\sigma = 1$ known (so that $u_t = e_t$), and that T is odd.

Given G, the jth autocovariance of u_t is

$$E[u_t u_{t-j}|G] = \int_{-\pi}^{\pi} e^{\mathbf{i}\omega j} f_e(\omega) d\omega = 2\int_0^{\pi} \cos(\omega j) f_e(\omega) d\omega, \qquad j = 0, \pm 1, \pm 2, \ldots \tag{3}$$

[6]The asymptotic distribution of Δp described in footnote 1 then changes to $\Delta p \sim \arg\max_{0 \leq d \leq p_{\max} - p_0} \sum_{j=1}^{d} (\kappa_{j+p_0} Z_j^2 - 2)$, where Z_j are independent standard normals and $\kappa_j = 1 + c^2/2\pi^2 j^2$.

with $\mathbf{i} = \sqrt{-1}$. Define the autocovariances, scaled by $\sqrt{T}$,

$$\gamma_j(G) = \begin{cases} \sqrt{T}(E[u_t^2|G] - 1) & \text{for } j = 0 \\ \sqrt{T}E[u_t u_{t-j}|G] & \text{for } j = \pm 1, \pm 2, \ldots \end{cases}$$

Let $\gamma(G) = (\gamma_1(G), \ldots, \gamma_T(G))'$, and

$$\Gamma(G) = \begin{pmatrix} \gamma_0(G) & \gamma_1(G) & \cdots & \gamma_{T-1}(G) \\ \gamma_1(G) & \gamma_0(G) & \cdots & \vdots \\ \vdots & \vdots & \ddots & \gamma_1(G) \\ \gamma_{T-1}(G) & \cdots & \gamma_1(G) & \gamma_0(G) \end{pmatrix}$$

so that with U_T the (reversed) vector of observations on u_t, $U_T = (u_T, \ldots, u_1)'$, we have

$$V(G) = E[U_T U_T'] = I + T^{-1/2}\Gamma(G).$$

The dependence of $\gamma(G)$, $\Gamma(G)$ and $V(G)$ on T is suppressed to simplify notation.

Given G, the optimal forecast of u_{T+1} is $u_{T+1|T}(G) = T^{-1/2}\gamma(G)'V(G)^{-1}U_T$. Thus $u^*_{T+1|T} = T^{-1/2}E[\gamma(G)'V(G)^{-1}U_T|U_T]$ is the exact posterior mean of u_{T+1}, where the expectation is taken over the posterior of G (or, equivalently, over the posterior of $(\gamma(G), V(G))$). We are interested in obtaining a simple approximate expression $u^p_{T+1|T} \approx u^*_{T+1|T}$.

The derivation proceeds in four steps. The first is to approximate $u_{T+1|T}(G)$. Under the local model (2),

$$f_e(\omega) = \frac{1}{2\pi}e^{T^{-1/2}G(\omega)} \approx \frac{1}{2\pi}\left(1 + T^{-1/2}G(\omega)\right)$$

so that

$$\gamma_j(G) \approx \tilde{\gamma}_j(G) = \frac{1}{\pi}\int_0^{\pi} \cos(\omega j)G(\omega)d\omega. \tag{4}$$

Thus $V(G) \approx I$, so $u_{T+1|T}(G) = T^{-1/2}\gamma(G)'V(G)^{-1}U_T \approx T^{-1/2}\gamma(G)'U_T$ and the posterior mean is approximately

$$u^*_{T+1|T} \approx T^{-1/2}\gamma^{*\prime}U_T \tag{5}$$

where γ^* is the posterior mean of $\gamma(G)$.

The second step is to approximate the exact Gaussian likelihood by the Whittle (1957, 1962) likelihood. Let z_j be the jth element of the periodogram,

$$z_j = \left| \frac{1}{\sqrt{2\pi T}} \sum_{t=1}^{T} u_t e^{\mathbf{i}\omega_j t} \right|^2, \quad \text{where } \omega_j = 2\pi j/T, \quad j = 1, \cdots, (T-1)/2.$$

The Whittle log-likelihood is

$$\ln p(Y_T|G) \approx -\sum_{j=1}^{(T-1)/2} \ln f_e(\omega_j) - \sum_{j=1}^{(T-1)/2} \frac{z_j}{f_e(\omega_j)} = \sum_{j=1}^{(T-1)/2} l_j(T^{-1/2}\delta_j(G)) \tag{6}$$

where $\delta_j(G) = \sqrt{T}(2\pi f_e(\omega_j) - 1)$ and $l_j(x) = -\ln(1+x) - \frac{2\pi z_j}{1+x} + \ln 2\pi$. As noted in the introduction, the Whittle approximation is often employed to facilitate computation of posteriors in models with priors in the spectral domain. At the same time, direct use of (6) for inference about $f_e(\omega_j)$ amounts to a non-parametric regression problem with log-chi squared distributed errors, which requires non-Gaussian posterior sampler techniques (cf. Carter and Kohn (1997)). What is more, the Whittle likelihood approximation potentially induces non-trivial errors in posterior means.

The third and key step of our approximation exploits the local embedding of $f_e(\omega_j)$ to address both these issues. On the one hand, for nearly flat $f_e(\omega)$, the approximation (6) becomes highly accurate. On the other hand, since $\delta_j(G) \approx G(\omega_j)$ is $O(1)$ for fixed G, we can further approximate l_j by a second order Taylor series

$$\begin{aligned}\sum_{j=1}^{(T-1)/2} l_j(\delta_j(G)) &\approx \sum_{j=1}^{(T-1)/2} \left(l_j(0) + T^{-1/2} l_j'(0)\delta_j(G) + \tfrac{1}{2}T^{-1} l_j''(0)\delta_j(G)^2\right) \\ &= \sum_{j=1}^{(T-1)/2} \left(l_j(0) + T_j^{-1/2} \bar{z}_j \delta_j(G) - \tfrac{1}{2}T^{-1}\left(1+2\bar{z}_j\right)\delta_j(G)^2\right) \end{aligned} \tag{7}$$

where $\bar{z}_j \equiv 2\pi z_j - 1$. Because $G(\omega)$ is continuous and $\frac{1}{T}\sum_{j=1}^{(T-1)/2} \bar{z}_j \approx 0$, $\frac{1}{T}\sum_{j=1}^{(T-1)/2} \bar{z}_j \delta_j(G)^2 \approx \frac{1}{T}\sum_{j=1}^{(T-1)/2} \bar{z}_j G(\omega_j)^2 \approx 0$ (see Müller and Petalas (2010)). Thus the term $\bar{z}_j\delta_j(G)^2$ in (7) vanishes and, upon completing the square, we have

$$\ln p(Y_T|G) \approx -\tfrac{1}{2} \sum_{j=1}^{(T-1)/2} \left(\bar{z}_j - T^{-1/2}\delta_j(G)\right)^2 + \text{function of } \{\bar{z}_j\}. \tag{8}$$

The sample information about the spectral density of u_t (or, equivalently, G) is thus captured asymptotically by the Gaussian pseudo-model

$$\bar{Z}|\delta(G) \sim \mathcal{N}(T^{-1/2}\delta(G), I), \tag{9}$$

where $\bar{Z} = (\bar{z}_1, \ldots, \bar{z}_{(T-1)/2})'$ and $\delta(G) = (\delta_1(G), \ldots, \delta_{(T-1)/2}(G))'$. Note, however, that (9) is only an accurate approximation to the *likelihood* as a function of $\delta(G)$—the *sampling distribution* of the centred periodogram ordinates $\bar{z}_j$ does not, of course, become approximately Gaussian, even asymptotically.

The fourth step in the approximation is to use the pseudo-model in the frequency domain to obtain a corresponding time domain model and hence an approximate posterior of $\gamma(G)$. From (3), $\gamma_j(G) \approx \frac{1}{\pi}\int_0^\pi \cos(\omega j)G(\omega)d\omega \approx \frac{2}{T}\sum_{l=1}^{(T-1)/2}\cos(\omega_l j)\delta_l(G) = T^{-1/2}q_j'\delta(G)$, where $q_j = 2T^{-1/2}(\cos(\omega_1 j), \cdots, \cos(\omega_{(T-1)/2}j))'$. Thus $\gamma(G) \approx T^{-1/2}Q'\delta(G)$, where $Q = (q_1, \ldots, q_{(T-1)/2})$. Similarly, let

$$\hat{s}_j = T^{-1/2}\sum_{t=j+1}^{T} u_t u_{t-j} = 2T^{-1/2}\sum_{l=1}^{(T-1)/2}\cos(\omega_l j)\bar{z}_l$$

so the vector of sample autocovariances of u_t, scaled by $\sqrt{T}$, is $\hat{s} = Q'\bar{Z}$. From (9) and $Q'Q = I$, we therefore have the approximate pseudo-model

$$\hat{s}|\gamma(G) \sim \mathcal{N}(\gamma(G), I). \tag{10}$$

The Gaussian prior for G implies via (4) an approximately Gaussian prior for $\gamma(G)$,

$$\gamma(G) \approx \mathcal{N}(0, \Sigma), \text{ where } \Sigma_{jl} = \frac{1}{\pi^2}\int_0^\pi\int_0^\pi \cos(rj)k_G(r,s)\cos(sl)drds \tag{11}$$

and $k_G(r,s) = E[G(r)G(s)]$ is the covariance kernel of the stochastic process G. Note that if

$$G(s) = J(s) - \frac{1}{\pi}\int_0^\pi J(r)dr, \tag{12}$$

that is G is the demeaned version of the stochastic process J, then Σ_{jl} in (11) can be alternatively computed with $k_G(r,s)$ replaced by $k_J(r,s) = E[J(r)J(s)]$, since $\int_0^\pi \cos(sj)ds = 0$ for $j = 1, 2, \ldots$.

Combining (10) and (11), the approximate posterior mean of γ is

$$\gamma^* \approx \gamma^p = \Sigma(I+\Sigma)^{-1}\hat{s} \tag{13}$$

which can be seen as a generalized shrinkage estimator of γ. With (5), the approximate posterior mean of u_{T+1} is thus

$$u^*_{T+1|T} \approx u^p_{T+1|T} = T^{-1/2}U_T'\gamma^p = T^{-1/2}\Sigma(I+\Sigma)^{-1}\hat{s}.$$

We now present explicit expressions for Σ and for the approximate posterior mean $\gamma^p = (\gamma_1^p, \cdots, \gamma_T^p)'$ in three cases.

Demeaned Brownian motion. Let $J(\omega) = cW(\omega)$, where W is Brownian motion on $[0,\pi]$ with $E[W(\pi)^2] = 1$, and via (12), G is distributed as a demeaned Brownian motion

$G(s) = cW^\mu(s) = cW(s) - \frac{c}{\pi}\int_0^\pi W(r)dr$. Then $k_J(r,s) = \frac{c^2}{\pi}\min(r,s)$ for $r, s \geq 0$, and direct evaluation of Σ_{jl} yields

$$\Sigma_{jl} = \begin{cases} 0, & j \neq l \\ \frac{c^2}{2\pi^2 j^2}, & j = l \end{cases}, \text{ and } \gamma_j^p = \frac{c^2}{c^2 + 2\pi^2 j^2}\hat{s}_j \quad \text{(Brownian motion prior).} \tag{14}$$

Thus the approximate posterior mean of u_{T+1}, $u^p_{T|T+1}$, is computed using shrinkage estimates of the autocovariances of u_t, with the shrinkage factor given in (14).

Demeaned integrated Brownian Bridge. Let $J(\omega) = cIB(\omega)$, where $IB(\omega) = \frac{1}{\pi}\int_0^\omega \left(B(s) - \frac{s}{\pi}B(\pi)\right) ds$, with W as above. Then

$$\Sigma_{jl} = \begin{cases} 0, & j \neq l \\ \frac{c^2}{2\pi^4 j^4}, & j = l \end{cases}, \text{ and } \gamma_j^p = \frac{c^2}{c^2 + 2\pi^4 j^4}\hat{s}_j \quad \text{(integrated Brownian Bridge prior).}$$

Demeaned Brownian motion restricted to a frequency band. A focus on deviations from the flat spectral density in a particular frequency band $[\underline{\omega}, \bar{\omega}]$, $0 \leq \underline{\omega} < \bar{\omega} \leq \pi$, is obtained by G that are constant for $\omega \notin [\underline{\omega}, \bar{\omega}]$. To be specific, suppose the baseline model is assumed to be misspecified for frequencies lower than $\bar{\omega}$ (i.e. $\underline{\omega} = 0$). One suitable process is then given by

$$J(\omega) = \begin{cases} \frac{c}{\sqrt{\bar{\omega}}} W(\omega) & \text{for } 0 \leq \omega < \bar{\omega} \\ \frac{c}{\sqrt{\bar{\omega}}} W(\bar{\omega}) & \text{otherwise} \end{cases}$$

where W is a standard Wiener process with $E[W(\pi)^2] = 1$. The scaling by $1/\sqrt{\bar{\omega}}$ ensures that the total variation in J is comparable to the demeaned Brownian motion case, since $E[J(1)^2] = c^2$. The covariance kernel of J is $k_J(r,s) = c^2 \min(r, s, \bar{\omega})/\bar{\omega}$ for $r, s \geq 0$, and Σ has j, lth element

$$\begin{cases} \frac{c^2}{\bar{\omega}\pi^2} \frac{j\cos(\bar{\omega}j)\sin(\bar{\omega}i) - l\cos(\bar{\omega}l)\sin(\bar{\omega}l)}{l^3 j - l j^3} & \text{for } j \neq l \\ \frac{c^2}{\bar{\omega}\pi^2} \frac{2\bar{\omega}j - \sin(2\bar{\omega}j)}{4j^3} & \text{for } j = j \end{cases} \quad \text{(Brownian motion prior on } [0, \bar{\omega}]\text{).} \tag{15}$$

We now discuss the extension to a baseline AR(p) model, but maintain $\alpha = 0$ and $\sigma = 1$ known. Suppose that $\beta(L)$ is estimated as $\hat{\beta}(L)$ by OLS for a given lag length p, which is asymptotically equivalent to the posterior mean of a Bayes estimation with a non-dogmatic prior on $\beta(L)$. The task then is to forecast $\hat{u}_{T+1}$, where $\hat{u}_t = \hat{\beta}(L)y_t$, so $\hat{u}_t$ is the AR(p) residual for $t = 1, \ldots, T$.

The approximate posterior mean of $\hat{u}_{T+1}$, $\hat{u}^p_{T+1|T}$, is obtained following the steps for the case $\beta(L) = 1$, except using the spectral density for the residuals. Because $\hat{u}_t = \hat{\beta}(L)y_t$, the spectral density of $\hat{u}_t$ is,

$$f_{\hat{u}}(\omega) = \frac{\left|\hat{\beta}(e^{\mathbf{i}\omega})\right|^2}{\left|\beta(e^{\mathbf{i}\omega})\right|^2} f_e(\omega) = \frac{1}{2\pi}\frac{\left|\hat{\beta}(e^{\mathbf{i}\omega})\right|^2}{\left|\beta(e^{\mathbf{i}\omega})\right|^2} e^{G(\omega)/\sqrt{T}}$$

$$\approx \frac{1}{2\pi}[1 + T^{-1/2}\mu(\omega)'b + T^{-1/2}G(\omega)] + O(T^{-1})$$

where $b = T^{1/2}(\beta - \hat{\beta})$ is $O_p(1)$ and

$$\mu(\omega) = \frac{\partial}{\partial\beta} \frac{\left|\hat{\beta}(e^{\mathbf{i}\omega})\right|^2}{|\beta(e^{\mathbf{i}\omega})|^2}\Bigg|_{\beta=\hat{\beta}}.$$

As before, set $\delta_j(G) = \sqrt{T}(2\pi f_{\hat{u}}(\omega_j) - 1)$, so $\delta_j(G) \approx \mu(\omega_j)'b + G(\omega_j)$. With this redefined notation for δ, the argument leading to (8) applies directly (because δ_j remains $O_p(1)$, so after prewhitening, the relevant spectral density remains local-to-flat). Thus $\hat{s}|\gamma_{\hat{u}}(G,b) \sim \mathcal{N}(\gamma_{\hat{u}}(G,b), I)$, where $\hat{s}$ now collects the autocovariances of $\hat{u}_t$ scaled by $\sqrt{T}$, $\gamma_{\hat{u}}(G,b) = T^{-1/2}Q'\delta(G) = mb + \gamma(G)$, $m = T^{-1/2}Q'\mu$ and $\mu = (\mu(\omega_1), \ldots, \mu(\omega_{(T-1)/2}))'$. The $T \times p$ matrix m has a simple form. Let $\Psi(L) = \beta(L)^{-1}$ and $\hat{\Psi}(L) = \hat{\beta}(L)^{-1}$, so $\hat{\Psi}_0 = 1$ and $\hat{\Psi}_j$ is the jth term in the MA representation of $\beta(L)$. Then

$$m = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \hat{\Psi}_1 & 1 & 0 & \cdots & 0 \\ \hat{\Psi}_2 & \hat{\Psi}_1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{\Psi}_{T-1} & \hat{\Psi}_{T-2} & \hat{\Psi}_{T-3} & \cdots & \hat{\Psi}_{T-p} \end{pmatrix}.$$

With an approximate $\mathcal{N}(0,\Sigma)$ prior on $\gamma(G)$, where Σ is given in (11), $\gamma_{\hat{u}}(G,b) = mb + \gamma(G)$ implies $\gamma_{\hat{u}}(G,b)|b \sim \mathcal{N}(mb, \Sigma)$. Thus $\hat{s}|\gamma_{\hat{u}}(G,b) \sim \mathcal{N}(\gamma_{\hat{u}}(G,b), I)$ yields

$$\hat{s}|b \sim \mathcal{N}(mb, I + \Sigma), \tag{16}$$

so $E[\gamma_{\hat{u}}(G,b)|\hat{s}, b] = mb + \Sigma(I+\Sigma)^{-1}(\hat{s} - mb)$. With a continuous prior on β, the prior on b is asymptotically flat, so that the posterior for b simply reflects the shape of the integrated likelihood (16), $E[b|\hat{s}] = (m'(I+\Sigma)^{-1}m)^{-1}m'(I+\Sigma)^{-1}\hat{s}$. Combining these expressions and using the identity $I - \Sigma(I+\Sigma)^{-1} = (I+\Sigma)^{-1}$, we obtain the approximate posterior mean,

$$\gamma_{\hat{u}}^* \approx \gamma_{\hat{u}}^p = \Sigma(I+\Sigma)^{-1}\hat{s} + (I+\Sigma)^{-1}m(m'(I+\Sigma)^{-1}m)^{-1}m'(I+\Sigma)^{-1}\hat{s}.$$

The approximate posterior mean $\hat{u}^p_{T+1|T}$ of $\hat{u}_{T+1}$may then be computed analogously to (13), that is, $\hat{u}^p_{T+1|T} = T^{-1/2}\hat{U}_T'\gamma^p_{\hat{u}}$, where $\hat{U}_T$ is the vector of AR(p) residuals, $\hat{U}_T = (\hat{u}_T, \ldots, \hat{u}_1)'$, and the approximate posterior mean of y_{T+1} is then given by $y^p_{T+1|T} = \hat{\beta}_1 y_T + \ldots + \hat{\beta}_p y_{T-p+1} + \hat{u}^p_{T+1|T}$.

Conditional heteroskedasticity

In the presence of conditional heteroskedasticity, the sampling distribution of $\hat{s}$ has an asymptotic covariance matrix that is no longer proportional to the identity matrix. In general, we have $T\,\mathrm{Var}[\hat{s}_j] = \mathrm{Var}[T^{-1/2}\sum_{t=j+1}^{T} u_t u_{t-j}] \approx E[u_t^2 u_{t-j}^2] = d_j$, so that

$$\hat{s}|\gamma \sim \mathcal{N}(\gamma, D) \tag{17}$$

with $D = \mathrm{diag}(d_1, d_2, \cdots, d_T)$. The sampling distribution of $\hat{s}$ thus differs from the pseudo-model (10). In analogy to results of Müller (2009) and Müller and Petalas (2010), one would therefore expect that one obtains better forecasts by employing the pseudo-model (17) that reflects the actual sample information about γ. Proceeding as above, in the case where $\beta(L) = 1$, we obtain with $\hat{s}_D = D^{-1/2}\hat{s}$, $\gamma_D(G) = D^{-1/2}\gamma(G)$ and $\Sigma_D = D^{-1/2}\Sigma D^{-1/2}$ that $\gamma_D^p = \Sigma_D(I+\Sigma_D)^{-1}\hat{s}_D$, so that now

$$\gamma^p = D^{1/2}\gamma_D^p = D^{1/2}\Sigma_D(I+\Sigma_D)^{-1}\hat{s}_D.$$

Similarly, with a baseline AR(p) model and $m_D = D^{-1/2}m$, we have instead of (16) that

$$\hat{s}_D|b \sim \mathcal{N}(T^{-1/2}m_D b, I+\Sigma_D),$$

so that now

$$\begin{aligned}\gamma_{\hat{u}}^p &= D^{1/2}\gamma_{D,\hat{u}}^p \\ &= D^{1/2}\Sigma_D(I+\Sigma_D)^{-1}\hat{s}_D + D^{1/2}(I+\Sigma_D)^{-1}m_D(m_D'(I+\Sigma_D)^{-1}m_D)^{-1}m_D'(I+\Sigma_D)^{-1}\hat{s}_D.\end{aligned}$$

The approximate posterior mean $\gamma_{\hat{u}}^p$ is thus $D^{1/2}\gamma_{D,\hat{u}}^p$, where $\gamma_{D,\hat{u}}^p$ is obtained by computing the posterior mean as before with $\hat{s}$, Σ and m replaced by $\hat{s}_D$, Σ_D and m_D. The matrix D is usually unknown, but can be estimated via $\hat{d}_j = T^{-1}\sum_{t=j+1}^{T} u_t^2 u_{t-j}^2$ and $\hat{D} = \mathrm{diag}(\hat{d}_1, \hat{d}_2, \cdots, \hat{d}_T)$.

2.2 Suggested Forecast and Formal Results

Consider the VAR generalization of (1)

$$\begin{aligned} y_t &= \alpha + \beta_1 y_{t-1} + \ldots + \beta_p y_{t-p} + u_t \\ &= \alpha + \beta_1 y_{t-1} + \ldots + \beta_p y_{t-p} + Pe_t, \quad t \geq 1 \end{aligned} \tag{18}$$

where $\{y_t\}_{t=-p+1}^{T}$ are observed $k \times 1$ vectors, $\{y_t\}_{t=-p+1}^{0}$ are fixed values independent of $\theta = (\alpha, \beta_1, \cdots, \beta_p)$, $\beta_j \in \mathbb{R}^{k\times k}$, $j = 1, \cdots, p$, P is a full rank lower-triangular $k \times k$ matrix, and $\{e_t\}$ is a zero mean stationary Gaussian process. Suppose the spectral density of $\{e_t\}$ is

$$f_e(\omega) = \frac{1}{2\pi}\exp[T^{-1/2}G(\omega)]$$

where G is a fixed Hermitian $k \times k$ matrix valued function on $[-\pi, \pi]$ with $G(-\omega) = G(\omega)'$ and $\int_{-\pi}^{\pi} G(\omega) d\omega = 0$, so that the jth autocovariance of e_t is given by $E[e_t e_{t-j}'] = \int_{-\pi}^{\pi} e^{\mathbf{i}j\omega} f_e(\omega) d\omega$.

Let A^* be the conjugate transpose of a complex matrix A. We consider the following class of priors for G.

Condition 1 *Under the prior measure,*

(a) $G(\omega)$ is a $k \times k$ Hermitian matrix for all $\omega \in [-\pi, \pi]$, and $G(-\omega) = G(\omega)'$ a.s.;

(b) $\int_{-\pi}^{\pi} G(\omega) d\omega = 0$ a.s.;

(c) $\mathrm{vec}(G(\omega))$ is a $k^2 \times 1$ mean-zero complex Gaussian process on $[-\pi, \pi]$, with a.s. bounded sample paths and covariance kernel $k_G(r, s) = E[\mathrm{vec}(G(s)) \mathrm{vec}(G(r))^*]$, $r, s \in [-\pi, \pi]$;

(d) $k_2(r, s) = \partial^2 k_G(r, s)/\partial r \partial s$ exists and is continuous for all $s \neq r$, $\int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \varphi(r)^* k_2(r, s) \varphi(s) ds dr < \infty$, $k_\Delta(s) = \lim_{\epsilon \to 0} [\partial k_G(r, s)/\partial r|_{r=s-\epsilon} - \partial k_G(r, s)/\partial r|_{r=s+\epsilon}]$ exists and $\int_{-\pi}^{\pi} \varphi(s)^* k_\Delta(s) \varphi(s) ds < \infty$, for all $k^2 \times 1$ complex valued functions φ with $\int_{-\pi}^{\pi} \varphi(s)^* \varphi(r) ds = 1$.

Parts (a) and (b) ensure that with prior probability one, e_t has a well defined spectral density, and $V[e_t | e_{t-1}, e_{t-2}, \ldots] = I_k$. Part (c) imposes a mean-zero Gaussian process prior for G. Loosely speaking, the differentiability assumptions on the covariance kernel k_G in part (d) ensure that under the prior, sample paths of G are at least as smooth as a Brownian motion. Note that a bounded modulus of the elements in $k_2(r, s)$ and $k_\Delta(s)$ over $r, s \in [-\pi, \pi]$ is sufficient (but not necessary) for the existence of the integrals in part (d).

All priors satisfying Condition 1 imply an eventual decay in $\tilde{\gamma}_j(G) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{\mathbf{i}js} G(s) ds \to 0$ as $j \to \infty$ almost surely. At the cost of a minor additional approximation error, one can therefore avoid large matrix inversions by explicitly setting $\tilde{\gamma}_j(G) = 0$ for $j \geq N$ for some large enough N.[7] Thus, define the $Nk^2 \times Nk^2$ matrix Σ with j,lth $k^2 \times k^2$ block equal to

$$\frac{1}{4\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} e^{\mathbf{i}js} k_G(r, s) e^{-\mathbf{i}lr} dr ds \tag{19}$$

so that Σ is the prior covariance matrix of $\mathrm{vec}(\tilde{\gamma}_1(G), \tilde{\gamma}_2(G), \cdots, \tilde{\gamma}_N(G))$ (cf. (4) of the heuristic discussion).

Condition 1 allows for a wide range of priors on the local-to-flat spectral density of e_t. Smoothness assumptions can be expressed continuously by letting G be a (demeaned) fractional Brownian with Hurst parameter $H \geq 1/2$ (with the Brownian motion case $H = 1/2$ the least smooth choice allowable under Condition 1 (d)), or integrated fractional Brownian motions. Condition 1 also covers the three priors for which Σ is worked out in Section 2.1.

[7] We stress that N is not a bandwidth choice, but simply a device that simplifies exposition and computations.

When y_t is a vector, one can choose to treat all k^2 elements of the autocovariance function of e_t symmetrically by letting the k^2 elements of G be i.i.d. copies of the real scalar Gaussian process $\bar{G}$ under the prior. With $k_{\bar{G}}(r,s) = E[\bar{G}(r)\bar{G}(s)]$, this leads to $k_G(r,s) = k_{\bar{G}}(r,s) \otimes I_{k^2}$, and Σ has a corresponding Kronecker structure $\Sigma = \bar{\Sigma} \otimes I_{k^2}$, where $\bar{\Sigma}$ is a $N \times N$ matrix with i,jth element constructed from $k_{\bar{G}}$ as in (11). For example, with a demeaned Wiener process prior for $\bar{G}$, one obtains the diagonal $Nk^2 \times Nk^2$ matrix $\Sigma = (c^2/2\pi^2)\,\mathrm{diag}(1/1^2, 1/2^2, \cdots, 1/N^2) \otimes I_{k^2}$.

Alternatively, one might want to allow for relatively greater non-flatness in the k diagonal elements of the spectrum, as residual cross-equation correlations might be expected to be relatively smaller than residual autocorrelations in the individual series in e_t. It might also make sense in some applications to treat the individual series in e_t asymmetrically, based on assumptions about frequencies and magnitudes of cross equation correlations, or different smoothness properties of individual and cross spectra.

The approximately best ℓ steps ahead forecast of y_t in the VAR system (18) under a Condition 1 prior is computed as follows: Let $\hat{\alpha}$ and $\hat{\beta}$ be the OLS estimates of α and β from a regression of y_t on a constant and p lagged values, and denote by $\hat{u}_t$ the OLS residuals. Let $\hat{P}$ the $k \times k$ lower diagonal matrix that satisfies $\hat{P}\hat{P}' = T^{-1}\sum_{t=1}^{T} \hat{u}_t\hat{u}_t'$, set $\widehat{e}_t = \hat{P}^{-1}\hat{u}_t$, and let $\hat{\Psi}_j$, $j = 1, 2, \ldots$ be $k \times k$ matrices satisfying $(I_k - \hat{\beta}_1 x - \cdots - \hat{\beta}_p x^p)^{-1} = I_k + \hat{\Psi}_1 x + \hat{\Psi}_2 x^2 + \ldots$. Define the $Nk \times pk$ matrix

$$\hat{\Psi} = \begin{pmatrix} I_k & 0 & 0 & \cdots & 0 \\ \hat{\Psi}_1' & I_k & 0 & \cdots & 0 \\ \hat{\Psi}_2' & \hat{\Psi}_1' & I_k & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{\Psi}_{N-1}' & \hat{\Psi}_{N-2}' & \hat{\Psi}_{N-3}' & \cdots & \hat{\Psi}_{N-k}' \end{pmatrix},$$

the $Nk^2 \times pk^2$ matrix $\hat{m} = \hat{\Psi} \otimes I_k$, and the $Nk^2 \times Nk^2$ block diagonal matrix $\hat{D}$ with jth $k^2 \times k^2$ block equal to $T^{-1}\sum_{t=j+1}^{T} \widehat{e}_{t-j}\widehat{e}_{t-j}' \otimes \widehat{e}_t\widehat{e}_t'$. Further, define the $Nk^2 \times 1$ vector $\widehat{s}$, which is the vec of the $k \times Nk$ matrix with jth $k \times k$ block equal to $T^{-1/2}\sum_{t=j+1}^{T} \widehat{e}_t\widehat{e}_{t-j}'$, $j = 1, \cdots, N$, and the $Nk^2 \times 1$ vector

$$\gamma^p = \Sigma(\hat{D} + \Sigma)^{-1}\widehat{s} + \hat{D}(\hat{D} + \Sigma)^{-1}\hat{m}(\hat{m}'(\hat{D} + \Sigma)^{-1}\hat{m})^{-1}\hat{m}'(\hat{D} + \Sigma)^{-1}\widehat{s}.$$

Let γ_j^p, $j = 1, \cdots, N$ be the $k \times k$ matrices such that $\mathrm{vec}(\gamma_1^p, \gamma_2^p, \cdots, \gamma_N^p) = \gamma^p$ and define

$$\hat{u}_{T+l|T}^p = T^{-1/2}\hat{P}\sum_{t=0}^{N-l} \gamma_{t+l}^p \widehat{e}_{T-t}, \quad l = 1, \cdots, \ell.$$

The approximately best forecast of $y_{T+\ell}$, $y^p_{T+\ell|T}$, then is obtained by iterating the VAR (18) forward by ℓ periods, using the OLS estimates for α and $\beta(L)$, with future disturbances set equal to $\hat{u}^p_{T+l|T}$.

These calculations are quite insensitive to N, as long as N is not chosen too small. A reasonable default is $N = \lfloor T^{3/4} \rfloor$. If conditional heteroskedasticity is not a concern, then $\hat{D}$ can be set equal to I_{Nk^2}. Also, if P is known (or partially known), then $\hat{P}$ can be replaced by the true P (or any other consistent estimator).

The main result of the paper is as follows.

Theorem 1 *Suppose that* $\{y_t\}_{t=1}^T$ *is generated from (18) with* G *any fixed and bounded function, and* $\theta = \theta^0$ *so that (18) is causal. Assume further that the prior on* G *satisfies Condition 1, and the prior on* $\theta = (\alpha, \beta_1, \cdots, \beta_p)$ *is a sample size independent, bounded Lebesgue probability density* w *satisfying* $\sup_{t \geq 1} \int E_\theta[||y_t||^2] w(\theta) d\theta < \infty$ *when* $G = 0$*. Let* $y^*_{T+\ell|T}$ *be the posterior mean for* $y_{T+\ell}$ *under this prior, assuming knowledge of* P*. If* $w(\theta^0) > 0$ *and* w *is continuous at* θ_0*, and* $N \to \infty$ *with* $N/T \to 0$*, then*

$$T^{1/2} ||y^*_{T+\ell|T} - y^p_{T+\ell|T}|| \xrightarrow{p} 0.$$

Moreover, if the prior is a mixture of n *Gaussian processes as above, with kernels*

$$k_{G(i)}(r,s) = E[(\operatorname{vec} G_{(i)}(r))(\operatorname{vec} G_{(i)}(s))^*],$$

$i = 1, \cdots, n$ *and mixture weights* $p_i \geq 0$, $\sum_{i=1}^n p_i = 1$*, then the same results holds with* $y^p_{T+\ell|T}$ *a corresponding convex combination of* n *versions of* $y^p_{T+\ell|T}$*, computed using* $\Sigma_{(i)}$ *(with* j, l*th block as in (19) with* k_G *replaced by* $k_{G(i)}$*) instead of* Σ*, with weights proportional to*

$$p_i \det(\hat{D} + \Sigma_{(i)})^{-1/2} \det(\hat{m}'(\hat{D} + \Sigma_{(i)})^{-1}\hat{m})^{-1/2}$$
$$\times \exp[-\tfrac{1}{2}\widehat{s}'(\hat{D} + \Sigma_{(i)})^{-1}\widehat{s} + \tfrac{1}{2}\widehat{s}'(\hat{D} + \Sigma_{(i)})^{-1}\hat{m}(\hat{m}'(\hat{D} + \Sigma_{(i)})^{-1}\hat{m})^{-1}\hat{m}'(\hat{D} + \Sigma_{(i)})^{-1}\widehat{s}].$$

3 Monte Carlo Results

This section reports selected results from a Monte Carlo study of the forecasting performance of the approximate posterior mean in the univariate case. We report the results of three experiments. In all three experiments, the results for multi-step ahead forecasts are for cumulative values, that is, forecasts of $y_{T+1} + \cdots + y_{T+\ell}$.

The first experiment checks the results of the theorem in a special case by quantifying the discrepancy between the exact ($y^*_{T+1|T}$) and approximate ($y^p_{T+1|T}$) posterior mean of y_{T+1} in the case that the prior G is correctly specified demeaned Brownian motion. The data were generated as in (1) and (2), with $\beta(L) = 1$. The estimators were implemented including an intercept and a fixed number p of autoregressive lags, $p = 0, \ldots, 6$. In all the Monte Carlo work the approximate posterior mean $y^p_{T+1|T}$ was computed as described in Section 2.2 using the demeaned Brownian motion prior (so Σ is given by (14)), and $\hat{D} = I_N$.

The results of this first experiment are reported in Table 2. The entries are $T(\text{MSFE}_{\text{posterior mean}} - \text{MSFE}_{\text{AR}})/\text{MSFE}_{\text{AR}}$, that is, the scaled relative increase of the mean squared forecast error of the approximate and exact posterior mean, relative to the MSFE of the simple AR forecast with the same lag length. A value less than zero indicates an improvement upon the AR forecast. The scaling by T is such that the entries of Table 2 stabilize as $T \to \infty$. To get some sense for the magnitude of the entries, note that estimating an AR(p) by OLS for forecasting an exact AR(p_0) (i.e. $G = 0$) with $p_0 < p$ leads to an asymptotic scaled relative deterioration of $p - p_0$ over a forecast using the more parsimonious AR(p_0) model. Roughly speaking, the entries of Table 2 (and Figures 1-4 below) are thus in units of "unnecessarily estimated parameters".

As can be seen in Table 2, consistent with Theorem 1, the discrepancy between the scaled MSFEs for the exact and approximate posterior means tends to zero as T increases. In addition, the smaller c and the larger p and T, the more accurate the approximate posterior mean. Intuitively, estimating $\beta(L)$ soaks up some of the variability of G, making the quadratic approximation of the log-likelihood more accurate. The exact posterior mean is only a substantially better forecast for large c and very small p, which is unlikely to be a much of a concern in practice, as highly variable realizations of G would lead one to include some AR lags in small samples.

The second experiment examines the performance of the posterior mean forecast using the prior $G \sim cW^{\mu}$ and some lag-length selection for the AR(p) under a data generating process where $\beta(L) = 1$ and $G \sim c_0 W^{\mu}$. Specifically, for a given Monte Carlo draw, $y^p_{T+\ell|T}$ was computed with either no autoregressive component, or with an AR(p) with BIC-selected lag length (AR(BIC)), with $0 \leq p \leq 4$), for c on a unit grid of 0 to 20 and for $\ell = 1, 2, 3,$ and 4. In addition, the Bayes model averaging (BMA) forecast was computed using an equal weighted prior over the grid of c, weighted by the normalized Bayes factor given in Theorem 1. The unadjusted ℓ-step ahead iterated OLS AR(BIC) forecast was also retained. This exercise was repeated on a grid of c_0 and evaluated for $T = 200$, with 20,000 Monte Carlo repetitions.

Figures 1 and 2 plot $T(\text{MSFEf} - \text{MSFEBIC})/\text{MSFEBIC}$, where MSFEf is a candidate forecast, similar to the entries of Table 2. Each line in the plot corresponds to the relative

MSFE of a given forecast (a given value of c) as a function of the true DGP c_0. The lower envelope of these lines maps out the forecast gains possible using the Bayes procedure with correct c_0 (up to our asymptotic approximation), and each forecasting procedure achieves this envelope when $c = c_0$. Figure 1 shows the results when no AR component is estimated, Figure 2 shows the results when an AR(BIC) component is estimated. Evidently improvements over the benchmark linear forecast are obtained for a wide range of $c \neq c_0$. The main case in which $y^p_{T+\ell|T}$ is worse than BIC is when c_0 is very small (recall that $c_0 = 0$ corresponds to no residual predictability), especially when c is large. This is not surprising since larger values of c yield less shrinkage and introduce more estimation error which, when $c_0 = 0$, simply adds noise to the forecast. The BMA forecast is never very far from the envelope and is only worse than the AR forecast for $c_0 \leq 2.5$. These conclusions hold whether $p = 0$ (no AR term is estimated; Figure 1) or an AR(BIC) is estimated (Figure 2), although not surprisingly the magnitude of the gains is less when an AR term is estimated. The magnitude of the forecasting gains depend on the true amount of residual predictability; when an AR(BIC) term is estimated, a typical value for the relative scaled MSFE gains is –1.5, which corresponds to percentage MSFE improvements of 1.5% with $T = 100$ or 0.75% with $T = 200$.

The third experiment examines the performance using the prior $G = cW^\mu$ under a non-stochastic local-to-flat spectral density. Specifically, data are generated according to a MA(1), with MA coefficients on a grid from $-4/\sqrt{T}$ to $4/\sqrt{T}$. The AR(BIC), approximate $y^p_{T+\ell|T}$, BMA, and AR(AIC) forecasts ($\ell = 1$ and 4 are reported here) were computed, where the maximum AIC and BIC lag lengths were $0 \leq p \leq 4(T/100)^{1/3}$. This experiment was repeated for $T = 50, 100, 200$, and 400, with 20,000 Monte Carlo repetitions each.

The results are summarized in Figures 3 and 4, which (like Figures 1 and 2) plot T(MSFEf – MSFEBIC)/MSFEBIC, where MSFEf is the MSFE of the candidate forecast. Consistent with the theory, if there is no residual predictability then the Bayes procedures simply add noise. For moderate amounts of residual predictability, the Bayes procedures improve upon the AR(BIC) forecast. The BMA forecasts improve upon AR(BIC) for all values of the MA coefficient except for those very close to zero, in which case the BMA procedure produces only a very small deterioration of the forecast. AR(AIC) improves upon AR(BIC) for larger amounts of predictability, but BMA uniformly improves upon AR(AIC).

4 Empirical Analysis

This section reports the results of an empirical comparison of the pseudo out-of-sample forecasting performance of the univariate posterior mean forecasts $y^p_{T+\ell|T}$, relative to unadjusted

AR forecasts. The data set consists of monthly data on 132 U.S. monthly macroeconomic time series, including data on real output, employment, wages and prices, monetary variables, interest rates and exchange rates, and miscellaneous other indicators of monthly economic activity from 1959:1 – 2003:12. The data set and data transformations are taken from Stock and Watson (2005) and the reader is referred to that article for details. As in Stock and Watson (2005), for variables transformed to growth rates or changes, ℓ-step ahead forecasts are forecasts of cumulative changes, for example, of cumulative employment growth over the next ℓ months, or of the cumulative change in the rate of price inflation over the next ℓ months. Nominal series, such as prices, are modeled in changes of inflation, and the ℓ-step ahead forecasts are of average inflation over the next ℓ months.

Forecasts were computed recursively with the first forecast date being the earliest date that 198 observations on the transformed variable were observed (so for real variables the first forecast date is 1975:7). The final forecast date is 2003:12 $-$ ℓ, and forecasts were computed for horizons ℓ up to 12 months ahead. At each date, the AR forecasts and AR component of the posterior mean forecasts were computed using lag length estimated by BIC, where $0 \leq p \leq 18$. We consider two types of priors on G for the posterior mean forecasts: First, a demeaned Brownian motion of scale c (cf. (14)), and second, a sum of an independent demeaned Brownian motion of scale c, and a (demeaned) truncated Brownian motion that varies only over frequencies below $\bar{\omega} = 2\pi/96$ of fixed scale 20 (cf. (15)). The latter is motivated by an expectation that the baseline AR model exhibits relatively more pronounced misspecification below business cycle frequencies (=cycles with periods of 96 months or more). The posterior mean forecasts were computed for c fixed and equal to $c = 0, 10, 20, 30$, and two BMA forecasts for $0 \leq c \leq 20$ and for $0 \leq c \leq 40$.

The results for the full data set are summarized in Table 3. All posterior mean procedures provide improvements over the AR(BIC) forecasts, with mean and median relative MSFE over the 132 series below unity at all horizons. The improvements are rather insensitive to the choice of c and to the use of BMA averaging instead of a fixed c. Imposing additional misspecification below business cycle frequencies leads to further improvements in forecast performance at longer horizons. The mean relative reduction in the MSFE in the one month ahead BMA ($0 \leq c \leq 40$) forecast is 0.011. Because of the recursive design, the sample size varies, but the average sample size is approximately 350. The improvement of 0.011 thus corresponds roughly to a scaled improvement of $350 \times 0.011 = 3.8$ in the units of Table 2 and Figures 1-4, and is at the upper end of the range of improvements reported there.

For some series, the improvements are even greater and, in the context of the literature that has used this data set, are in fact quite substantial: 10% of the series have relative reductions in MSFEs of 6% or more for most of the posterior mean forecasting procedures at the six month horizon. At the same time, the cost from using the procedure is small, for example the 90% percentile of relative MSFEs is at most 1.011 at all horizons reported for all the demeaned Wiener process BMA procedures. At longer horizons, also AR(AIC) outperforms AR(BIC), with very large improvements at the 10% percentile. At the same time, the cost associated with using AIC is also large in the sense that for some series using the additional lags results in a marked deterioration of the AR(AIC) forecasts, with 90% percentiles exceeding 1.077 at the four reported horizons. Interestingly, in results not reported here, it appears that the residual predictability found in Table 3 is particularly pronounced in the nominal series.

5 Conclusion

This paper develops a framework to study slight misspecifications of finite order VARs by modeling driving disturbances with a local-to-flat spectral density. We focus on the impact of the misspecification on forecasts, and derive a computationally straightforward modification of standard VAR based forecasts that effectively exploits this residual predictability in large samples. Monte Carlo and empirical evidence suggests that the large sample results provide meaningful approximations for practically relevant sample sizes.

The suggested framework and some of our theoretical results could also be applied to study other issues involving slightly misspecified VARs. For instance, our likelihood approximations naturally lead to computationally straightforward tests of the null hypothesis of correct specification that maximize weighted average power in large samples. One could also consider the problem of the estimation of the spectral density, or the impact of the slight misspecification on standard large-sample inference about the VAR parameters, in analogy to Müller and Petalas (2010) and Li and Müller (2009). A potentially difficult but interesting extension could be a combination of locally varying VAR parameters as in Stock and Watson (1996) with the slight misspecification in the spectral domain considered here.

A Appendix: Proof of Theorem 1

Notation and Preliminaries: Let $|A| = \sqrt{\operatorname{tr} A^*A}$, and denote by $||A||^2$ the largest eigenvalue of A^*A, so that $||\cdot||$ is a submultiplicative matrix norm, and $||A|| \leq |A|$. Recall the following identities for conformable matrices A, B, C, D:

$$\begin{aligned}
(A \otimes B)(C \otimes D) &= AC \otimes BD & (20)\\
\operatorname{vec}(ABC) &= (C' \otimes A) \operatorname{vec} B & (21)\\
||A \otimes B|| &= ||A|| \cdot ||B|| & (22)\\
||A - B|| &\leq |A - B| = ||\operatorname{vec}(A - B)||. & (23)
\end{aligned}$$

Let $\Gamma(G)$ be the $kT \times kT$ symmetric block Toeplitz matrix with jth $k \times k$ block in the lower block diagonals equal to

$$\gamma_j(G) = \frac{T^{1/2}}{2\pi} \int_{-\pi}^{\pi} e^{\mathbf{i}\omega j}(\exp[T^{-1/2}G(\omega)] - I_k)d\omega, \quad j \geq 0 \tag{24}$$

and $\gamma(G)$ is the $k^2(T+\ell) \times 1$ vector that is the vec of the $k \times (T+\ell)k$ matrix with blocks $\gamma_j(G)$, $j = 1, \cdots, T+\ell$. Let $V(G) = I_{kT} + T^{-1/2}\Gamma(G)$, so that $e = (e_1', e_2', \cdots, e_T')' \sim \mathcal{N}(0, V(G))$, conditional on G.

Similarly, let $\tilde{\Gamma}(G)$ be the $kT \times kT$ symmetric block Toeplitz matrix with jth $k \times k$ block in the lower block diagonals equal to

$$\tilde{\gamma}_j(G) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{\mathbf{i}\omega j} G(\omega) d\omega, \; j \geq 0 \tag{25}$$

and $\tilde{\gamma}(G)$ is the $k^2N \times 1$ vector that is the vec of the $k \times Nk$ matrix with blocks $\tilde{\gamma}_j(G)$, $j = 1, \cdots, N$. Note that $\tilde{\gamma}_0(G) = 0$ from $\int_{-\pi}^{\pi} G(\omega)d\omega = 0$. Also, let $D_\Gamma(G) = \Gamma(G) - \tilde{\Gamma}(G)$.

Let Λ be the $Nk^2 \times Nk^2$ diagonal matrix $\operatorname{diag}(1, 1/2, 1/3, \cdots, 1/N) \otimes I_{k^2}$. Note that by (22) and Lemma 4 (ii), $||\Lambda^{-1}\Sigma\Lambda^{-1}|| = O(1)$. Define the $T \times (kp+1)$ matrix

$$\bar{X} = \begin{pmatrix} 1 & y_0' & y_{-1}' & \cdots & y_{-p+1}' \\ 1 & y_1' & y_0' & \cdots & y_{-p+2}' \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & y_{T-1}' & y_{T-2}' & \cdots & y_{T-p}' \end{pmatrix}$$

and note that with $\theta^0 = (\alpha^0, \beta_1^0, \beta_2^0, \cdots, \beta_p^0) = (\alpha^0, \beta^0)$,

$$\begin{aligned}
P^{-1}(y_1, \cdots, y_T) &= P^{-1}\theta^0 \bar{X}' + (e_1, \cdots, e_T)\\
(I_T \otimes P^{-1})y &= (\bar{X} \otimes P^{-1}) \operatorname{vec} \theta^0 + e\\
&= X \operatorname{vec} \theta^0 + e
\end{aligned}$$

where $y = \operatorname{vec}(y_1, \cdots, y_T)$ and $X = \bar{X} \otimes P^{-1}$.

Let $\mu(\theta) = \lim_{t\to\infty} E_\theta[y_t]$, and denote by $w_\mu : \mathbb{R}^{(k+1)\times p} \mapsto \mathbb{R}$ the prior density of $(\mu(\theta), \beta)$ induced by w. Let $\mu^0 = \mu(\theta^0)$. It is clear that the posterior mean of $y_{T+\ell}$ is equal to the sum of μ^0 and the posterior mean of $y_{T+\ell} - \mu^0$ given data $\{y_t - \mu^0\}_{t=-p+1}^T$, where the latter posterior is computed using the prior w^0 on θ such that the prior w_μ^0 on $(\mu(\theta), \beta)$ induced by w^0 satisfies $w_\mu^0((\mu, \beta)) = w_\mu((\mu + \mu^0, \beta))$, for all (μ, β). Also note that $y_{T+\ell|T}^p$ is equivariant to translations of $\{y_t\}_{t=-p+1}^T$. It thus suffices to show the result for $y_{T+\ell} - \mu^0$ given data $\{y_t - \mu^0\}_{t=-p+1}^T$ and using prior w^0. Noting that $w^0((0, \beta^0)) = w((\alpha^0, \beta^0))$, this amounts to showing the result for $\alpha^0 = 0$ and prior w^0. Thus, from now on, we will assume $\alpha^0 = 0$ and replace w with w^0. Note that $\theta \mapsto (\mu(\theta), \beta)$ is a continuous and invertible function at θ^0, so that $w^0(\theta)$ is continuous and positive at $(0, \beta^0)$, since w is continuous and positive at θ^0 by assumption.

All subsequent convergences and expectations with respect to the data are for the model where $(I_T \otimes P^{-1})y \sim \mathcal{N}(X \operatorname{vec} \theta^0, I_{Tk})$. Corresponding convergences in probability under the data generating process assumed in the Theorem follow from contiguity, which is shown for a fixed G in Dzhaparidze (1986), Theorem 4, page 64 for the univariate case. For brevity, we omit the conceptually straightforward extension of the contiguity proof.

Let $h = (a', b')' = T^{1/2} \operatorname{vec}(\theta - \theta^0)$, where a is $k \times 1$. The log-likelihood ratio of the model where $(I_T \otimes P^{-1})y \sim \mathcal{N}(X \operatorname{vec} \theta, V(G)^{1/2})$ to the model where $(I_T \otimes P^{-1})y \sim \mathcal{N}(X \operatorname{vec} \theta^0, I_{Tk})$, evaluated at $(I_T \otimes P^{-1})y = X \operatorname{vec} \theta^0 + e \sim \mathcal{N}(X \operatorname{vec} \theta^0, I_{Tk})$, is

$$\begin{aligned} \ln L_T(G, h) &= -\tfrac{1}{2} \ln \det V(G) - \tfrac{1}{2}(e - T^{-1/2} X h)' V(G)^{-1} (e - T^{-1/2} X h) + \tfrac{1}{2} e' e \\ &= -\tfrac{1}{2} \ln \det V(G) - \tfrac{1}{2} e'(V(G)^{-1} - I_{kT}) e + T^{-1/2} e' V(G)^{-1} X h - \tfrac{1}{2} T^{-1} h' X' V(G)^{-1} X h. \end{aligned}$$

Let $A(b)$ be the $pk \times pk$ companion matrix of the VAR with parameter β satisfying $\operatorname{vec} \beta = \operatorname{vec} \beta^0 + T^{-1/2} b$,

$$A(b) = \begin{pmatrix} \beta_1 & \beta_2 & \cdots & \beta_{p-1} & \beta_p \\ I_k & 0 & \cdots & 0 & 0 \\ 0 & I_k & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I_k & 0 \end{pmatrix}$$

define $Y_T = \operatorname{vec}(y_T, y_{T-1}, \cdots, y_{T-p+1})$ and J_1 as the first k columns of I_{pk}. Note that $y_{T+\ell}$ can be written as

$$y_{T+\ell} = \sum_{l=1}^{\ell} J_1' A(b)^{\ell - l} J_1 (P e_{T+l} + T^{-1/2} a) + J_1' A(b)^\ell Y_T.$$

Let R be the $Tk \times Tk$ matrix such that $\operatorname{vec}(e_T, e_{T-1}, \cdots, e_1) = Re$. Note that $E[Ree'R'] = RV(G)R' \neq V(G)$, but $R^{-1} = R' = R$, and $||R|| = 1$. Since

$$\begin{pmatrix} e_{T+l} \\ Re \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} I_k + T^{-1/2} \gamma_0(G) & T^{-1/2}(\gamma_l(G), \cdots, \gamma_{T+l-1}(G)) \\ T^{-1/2}(\gamma_l(G), \cdots, \gamma_{T+l-1}(G))' & RV(G)R \end{pmatrix}\right),$$

we have, using (21)

$$\begin{aligned} E[e_{T+l}|e] &= T^{-1/2}(\gamma_l(G), \cdots, \gamma_{T+l-1}(G))(RV(G)R)^{-1} Re \\ &= T^{-1/2}((RV(G)R)^{-1} Re \otimes I_k)'(\Delta_l \gamma(G)) \\ &= T^{-1/2} e_{T+l|T}(G) \end{aligned}$$

where Δ_l is the $Tk^2 \times (T+\ell)k^2$ matrix such that $\Delta_l \gamma(G) = \text{vec}(\gamma_l(G), \cdots, \gamma_{T+l-1}(G))$. Conditional on (h, G, e), the difference between the posterior mean of $y_{T+\ell}$ and $J_1' A(0)^\ell Y_T$, multiplied by $T^{1/2}$, is

$$\begin{aligned} f_T(G,h) &= T^{1/2} J_1'(A(b)^\ell - A(0)^\ell) Y_T + \sum_{l=1}^{\ell} J_1' A(b)^{\ell-l} J_1 (a + P e_{T+l|T}(G)) \\ &= T^{1/2} J_1'(A(b)^\ell - A(0)^\ell) Y_T \\ &\quad + \sum_{l=1}^{\ell} (J_1(a + P((RV(G)R)^{-1} Re \otimes I_k)'(\Delta_l \gamma(G))) \otimes J_1)' \,\text{vec}(A(b)^{\ell-l}) \end{aligned}$$

where the second equality uses (21). Let $\hat{e} = e - T^{-1/2} X \hat{h}$, where

$$\hat{h} = (\hat{a}', \hat{b}')' = (T^{-1} X'X)^{-1} T^{-1/2} X'e.$$

We will use the approximation

$$\begin{aligned} \tilde{f}_T(G,h) &= T^{1/2} J_1'(A(\hat{b})^\ell - A(0)^\ell) Y_T + \sum_{l=1}^{\ell} ((A(\hat{b})^{l-1} \tilde{Y}_T)' \otimes J_1' A(\hat{b})^{\ell-l} J_1)(b - \hat{b}) \\ &\quad + \sum_{l=1}^{\ell} J_1(a + P(\tilde{R}\hat{e} \otimes I_k)'(\tilde{\Delta}_l \tilde{\gamma}(G)) \otimes J_1)' \,\text{vec}(A(\hat{b})^{\ell-l}) \end{aligned}$$

where the $Nk \times kT$ matrix $\tilde{R}$ satisfies $\tilde{R}\hat{e} = \text{vec}(\hat{e}_T, \hat{e}_{T-1}, \cdots, \hat{e}_{T-N+1})$, $\tilde{\Delta}_l$ is the $Nk^2 \times Nk^2$ matrix such that $\tilde{\Delta}_l \tilde{\gamma}(G) = \text{vec}(\tilde{\gamma}_l(G), \cdots, \tilde{\gamma}_N(G), 0, \cdots, 0)$ (so that $\tilde{\Delta}_l = \bar{\Delta}_l \otimes I_{k^2}$ with $\bar{\Delta}_l$ a $N \times N$ matrix with ones on the $(l-1)$th upper diagonal, and zero elsewhere), and

$$\tilde{Y}_T = Y_T - (1, \cdots, 1)' \otimes (T^{-1/2} \sum_{t=0}^{N-1} \hat{\Psi}_t a) - A(\hat{b})^N Y_{T-N}$$

with $Y_{T-N} = \text{vec}(y_{T-N}, y_{T-N-1}, \cdots, y_{T-N-p+1})$. Note that $\tilde{f}_T(G,h)$ is linear in $\tilde{\gamma}(G)$ and h.

Let $\hat{s}$ be the $Nk^2 \times 1$ vector which is the vec of the $k \times Nk$ matrix with blocks $\hat{s}_j = T^{-1} \sum_t \hat{e}_t \hat{e}_{t-j}'$, $j = 1, \cdots, N$. Define m to be the $Nk^2 \times pk^2$ matrix with $k^2 \times pk^2$ blocks equal to

$$m_j = (P'\Psi_{j-1}', P'\Psi_{j-2}', \cdots, P'\Psi_{j-p}') \otimes P^{-1}$$

for $j = 1, \cdots, N$, where $(I_k - \beta_1^0 x - \cdots - \beta_p^0 x^p)^{-1} = \sum_{j=-\infty}^{\infty} \Psi_j x^j$ (so that $\Psi_j = 0$ for $j < 0$, and $\Psi_0 = I_k$). Note that

$$m'm = \left(\sum_{j=1}^{N} \begin{pmatrix} \Psi_{j-1}P \\ \Psi_{j-2}P \\ \vdots \\ \Psi_{j-p}P \end{pmatrix} \begin{pmatrix} \Psi_{j-1}P \\ \Psi_{j-2}P \\ \vdots \\ \Psi_{j-p}P \end{pmatrix}' \right) \otimes (P^{-1\prime} P^{-1}) \qquad (26)$$

$$\overset{N\to\infty}{\to} \lim_{t\to\infty} \begin{pmatrix} E[y_t y_t'] & E[y_t y_{t-1}'] & \cdots & E[y_t y_{t-p+1}'] \\ E[y_{t-1} y_t'] & E[y_t y_t'] & \cdots & E[y_t y_{t-p+2}'] \\ \vdots & \vdots & & \vdots \\ E[y_{t-p+1} y_t'] & E[y_{t-p+2} y_t'] & \cdots & E[y_t y_t'] \end{pmatrix} \otimes \Omega^{-1}.$$

where $\Omega = PP'$, and it is not difficult to see that $\hat{m}'\hat{m} - m'm \xrightarrow{p} 0$. Also, by Lemma 8 (ii), $\hat{m}'\hat{s} \xrightarrow{p} 0$. Define $\hat{M} = (I+\Sigma)^{-1}\hat{m}$ and $\hat{S} = (I+\Sigma)^{-1}\hat{s}$.

Note that by Lemma 5 (i),

$$\tilde{f}_T(G,h) = T^{1/2} J_1'(A(\hat{b})^\ell - A(0)^\ell)Y_T + \sum_{l=1}^{\ell} J_1' A(\hat{b})^{\ell-l} J_1[a + P(\tilde{R}\hat{e} \otimes I_k)'\tilde{\Delta}_l(\tilde{\gamma}(G) + \hat{m}(b-\hat{b}))].$$

The approximation of the likelihood proceeds as follows

$$\begin{aligned} \ln L_T(G,h) &= -\tfrac{1}{2}\ln\det V(G) - \tfrac{1}{2}e'(V(G)^{-1} - I)e + T^{-1/2}e'V(G)^{-1}Xh - \tfrac{1}{2}T^{-1}h'X'V(G)^{-1}Xh \\ &\approx \tfrac{1}{2}T^{-1/2}e'\tilde{\Gamma}(G)e - \tfrac{1}{2}T^{-1/2}\operatorname{tr}\tilde{\Gamma}(G) - \tfrac{1}{4}T^{-1}\operatorname{tr}\tilde{\Gamma}(G)^2 + T^{-1/2}e'Xh \\ &\qquad -T^{-1}e'\tilde{\Gamma}(G)Xh - \tfrac{1}{2}T^{-1}h'X'Xh \\ &\approx \tfrac{1}{2}T^{-1/2}\hat{e}'\tilde{\Gamma}(G)\hat{e} - \tfrac{1}{2}T^{-1/2}\operatorname{tr}\tilde{\Gamma}(G) - \tfrac{1}{4}T^{-1}\operatorname{tr}\tilde{\Gamma}(G)^2 - T^{-1}(h-\hat{h})'X'\tilde{\Gamma}(G)e \\ &\qquad +T^{-1}\hat{h}'X'Xh - \tfrac{1}{2}T^{-1}h'X'Xh \\ &\approx \hat{s}'\tilde{\gamma}(G) - \tfrac{1}{2}\tilde{\gamma}(G)'\tilde{\gamma}(G) - (b-\hat{b})'\hat{m}'\tilde{\gamma}(G) + \hat{b}'\hat{m}'\hat{m}b - \tfrac{1}{2}b'\hat{m}'\hat{m}b + \hat{a}'\Omega^{-1}a - \tfrac{1}{2}a'\Omega^{-1}a \\ &\approx -\tfrac{1}{2}||\hat{s} - \tilde{\gamma}(G) - \hat{m}(b-\hat{b})||^2 - \tfrac{1}{2}(a-\hat{a})'\Omega^{-1}(a-\hat{a}) + \tfrac{1}{2}||\hat{s}||^2 + \tfrac{1}{2}\hat{b}'\hat{m}'\hat{m}\hat{b} + \tfrac{1}{2}\hat{a}'\Omega^{-1}\hat{a} \\ &= \ln\tilde{L}_T(G,h). \end{aligned}$$

so that the approximate log-likelihood $\ln\tilde{L}_T(G,h)$ is quadratic in $(h, \tilde{\gamma}(G))$.

Define $\mathcal{S}_T(G) = \mathbf{1}[\sup_{-\pi\le\omega\le\pi}||G(\omega)|| \le T^\kappa]$ for some $0 < \kappa < 1/6$. Let E_y stand for integration over the sampling distribution of e (and y, X, etc.) in the model where $(I_T \otimes P^{-1})y \sim \mathcal{N}(X\operatorname{vec}\theta^0, I_{Tk})$ and $e \sim \mathcal{N}(0, I_{Tk})$, E_G integration over G, E_h integration over $\mathcal{N}(\hat{h}, C_T^{-1}I_{pk^2+k})$, where C_T is defined in Lemma 7. Also, let $w_c^0(h) = w^0(\theta)$, where $\operatorname{vec}\theta = \operatorname{vec}(\theta^0) + T^{-1/2}h$.

Additional Lemmata: The proof of Theorem 1 below makes use of the following Lemmata. Their proofs are collected in the supplementary material section.

Lemma 1 *(i) Let x_1 and $\tilde{x}_1$ be sequences (in T) of random vectors, and let x_j and $\tilde{x}_j$, $j = 2, \cdots, l$ be sequences (in T) of scalar random variables, and denote by E_c a conditional expectation.*

(i) If $E_c||\tilde{x}_1||^2 = O_p(1)$, $E_c||x_1 - \tilde{x}_1||^2 \xrightarrow{p} 0$, and for all $K \in \mathbb{N}$ and $j \ge 2$, $E_c|\tilde{x}_j|^K = O_p(1)$ and $E_c|x_j - \tilde{x}_j|^K \xrightarrow{p} 0$, then $E_c||\prod_{j=1}^l x_j - \prod_{j=1}^l \tilde{x}_j||^{4/3} \xrightarrow{p} 0$.

(ii) For all $j \ge 2$, if $E_c[\tilde{x}_j{}^K] \xrightarrow{p} 1$ for all $K \in \mathbb{N}$, then $E_c|\tilde{x}_j - 1|^K \xrightarrow{p} 0$ for all $K \in \mathbb{N}$.

Lemma 2 *(i) $T^K E_G(1 - \mathcal{S}_T(G)) \to 0$ for all finite K.*

(ii) $E_G \int (1 - \mathcal{S}_T(G)) w_c^0(h)||f_T(G,h)||L_T(G,h)dh \xrightarrow{p} 0$.

(iii) $\mathcal{S}_T(G)||D_\Gamma(G)|| \le 2T^{-1+\kappa}$ a.s. and $\mathcal{S}_T(G)T^{-1}\operatorname{tr} D_\Gamma(G)^2 \le 4kT^{-2+4\kappa}$ a.s.

(iv) $\mathcal{S}_T(G)||\tilde{\Gamma}(G)|| \le T^\kappa$ a.s. and $\mathcal{S}_T(G)||\Gamma(G)|| \le 2T^\kappa$ a.s.

(v) $\mathcal{S}_T(G)||V(G)^{-1}|| \le 1 + 4T^{\kappa-1/2}$ *a.s. for* $T > 64$.
(vi) $\mathcal{S}_T(G)||\gamma(G) - \tilde{\gamma}(G)|| \le 2kT^{-1+2\kappa}$ *a.s.*
(vii) $\mathcal{S}_T(G)||\gamma(G)|| \le 2kT^{\kappa}$ *a.s.*
(viii) $E_G\mathcal{S}_T(G)\sum_{j=N+1}^{T}|\gamma_j(G)|^2 \to 0$.

Lemma 3 *For any fixed* $K > 0$,

$$E_hE_G\mathcal{S}_T(G)\exp[K\xi_T(G,h)] \xrightarrow{p} 1$$

where $\xi_T(G,h)$ *is one of the following expressions:*
(i) $\ln\det V(G) - T^{-1/2}\operatorname{tr}\Gamma(G) + \frac{1}{2}T^{-1}\operatorname{tr}\Gamma(G)^2$
(ii) $e'(V(G)^{-1} - I)e - e'(-T^{-1/2}\Gamma(G) + T^{-1}\Gamma(G)^2)e$
(iii) $T^{-1/2}e'V(G)^{-1}Xh - (T^{-1/2}e'Xh - T^{-1}e'\Gamma(G)Xh)$
(iv) $T^{-1}h'X'V(G)^{-1}Xh - T^{-1}h'X'Xh$
(v) $T^{-1}e'\Gamma(G)^2e - T^{-1}\operatorname{tr}\Gamma(G)^2$
(vi) $T^{-1/2}e'D_\Gamma(G)e - T^{-1/2}\operatorname{tr}D_\Gamma(G)$
(vii) $T^{-1}\operatorname{tr}\Gamma(G)^2 - T^{-1}\operatorname{tr}\tilde{\Gamma}(G)^2$
(viii) $T^{-1}e'D_\Gamma(G)Xh$
(ix) $T^{-3/2}\hat{h}'X'\tilde{\Gamma}(G)Xh$
(x) $T^{-3/2}\hat{h}'X'\tilde{\Gamma}(G)X\hat{h}$
(xi) $T^{-1}\frac{1}{2}\operatorname{tr}\tilde{\Gamma}(G)^2 - \tilde{\gamma}(G)'\tilde{\gamma}(G)$
(xii) $T^{-1}h'X'Xh - a'\Omega^{-1}a - b'\hat{m}'\hat{m}b$
(xiii) $T^{-1}\hat{h}'X'Xh - \hat{a}'\Omega^{-1}a - \hat{b}'\hat{m}'\hat{m}b$
(xiv) $T^{-1}(h-\hat{h})'X'\tilde{\Gamma}(G)e - (b-\hat{b})'\hat{m}'\tilde{\gamma}(G)$
(xv) $\hat{s}'\hat{m}(b-\hat{b})$
(xvi) $\ln w_c^0(h) - \ln w_c^0(0)$

Lemma 4 *(i)* $\operatorname{tr}\Sigma = O(1)$.
(ii) $||\Lambda^{-1}\Sigma\Lambda^{-1}|| = O(1)$.

Lemma 5 *(i)* $\sum_{l=1}^{\ell}((A(\hat{b})^{l-1}\tilde{Y}_T)' \otimes J_1'A(\hat{b})^{\ell-l}J_1)(b-\hat{b}) = \sum_{l=1}^{\ell}J_1'A(\hat{b})^{\ell-l}J_1P(\tilde{R}\hat{e}\otimes I)'\tilde{\Delta}_l\hat{m}(b-\hat{b})$.
(ii)$P(\tilde{R}\hat{e}\otimes I)'\tilde{\Delta}_l(\hat{s} - \hat{S} + \hat{M}(\hat{m}'\hat{M})^{-1}\hat{m}'\hat{S}) = P(\tilde{R}\hat{e}\otimes I_k)'\tilde{\Delta}_l[(I_N\otimes P'\hat{P}^{-1\prime}\otimes P^{-1}\hat{P})][\Sigma(\hat{D}+\Sigma)^{-1}\widehat{\hat{s}} + \hat{D}\widehat{\hat{M}}(\hat{m}'\widehat{\hat{M}})^{-1}\hat{m}'\hat{S}] + o_p(1)$ *for* $l = 1,\cdots,\ell$, *where* $\widehat{\hat{M}} = (\hat{D}+\Sigma)^{-1}\hat{m}$ *and* $\widehat{\hat{S}} = (\hat{D}+\Sigma)^{-1}\widehat{\hat{s}}$.

Lemma 6 *(i)* $E_hE_G||\tilde{f}_T(G,h)||^2 = O_p(1)$.
(ii) $E_hE_G\mathcal{S}_T(G)||f_T(G,h) - \tilde{f}_T(G,h)||^2 \xrightarrow{p} 0$.

Lemma 7 *(i)* $E_G\exp[K\hat{s}'\tilde{\gamma}(G)] = O_p(1)$ *for any fixed* $K \in \mathbb{R}$.
(ii) $E_G\tilde{L}_T(G,h)^4 \le \tilde{\zeta}_T\exp[-\frac{1}{2}4C_T||h-\hat{h}||^2]$ *uniformly in h, where* C_T,C_T^{-1} *and* $\tilde{\zeta}_T$ *are* $O_p(1)$ *and do not depend on h.*

Lemma 8 *(i)* $\hat{s}'\Lambda^2\hat{s} = O_p(1)$.
(ii) $\hat{m}'\hat{s} \xrightarrow{p} 0$.

Lemma 9

$$
\begin{aligned}
b^p &= \frac{E_G \int b\tilde{L}_T(G,h)dh}{E_G \int \tilde{L}_T(G,h)dh} = (\hat{m}'\hat{M})^{-1}\hat{m}'\hat{S} \\
\gamma^p &= \frac{E_G \int \tilde{\gamma}(G)\tilde{L}_T(G,h)dh}{E_G \int \tilde{L}_T(G,h)dh} = \hat{s} - \hat{S} + (\hat{M} - \hat{m})b^p \\
a^p &= \frac{E_G \int a\tilde{L}_T(G,h)dh}{E_G \int \tilde{L}_T(G,h)dh} = \hat{a} \\
\frac{E_G \int \tilde{L}_T(G,h)dh}{\int \tilde{L}_T(0,h)dh} &= \det(I+\Sigma)^{-1/2}\det(\hat{m}'\hat{M})^{-1/2}\exp[-\tfrac{1}{2}\hat{s}'\hat{S} + \tfrac{1}{2}\hat{S}'\hat{m}(\hat{m}'\hat{M})^{-1}\hat{m}'\hat{S}]
\end{aligned}
$$

where $\hat{S} = (I+\Sigma)^{-1}\hat{s}$ *and* $\hat{M} = (I+\Sigma)^{-1}\hat{m}$.

Proof of Theorem 1:

We start with the first claim. We show

$$
\frac{E_G \int w_c^0(h) f_T(G,h) L_T(G,h)dh}{E_G \int w_c^0(h) L_T(G,h)dh} - \frac{E_G \int w_c^0(0)\tilde{f}_T(G,h)\tilde{L}_T(G,h)dh}{E_G \int w_c^0(0)\tilde{L}_T(G,h)dh} \stackrel{p}{\to} 0 \tag{27}
$$

via

$$
E_G \int ||w_c^0(h) f_T(G,h) L_T(G,h) - w_c^0(0)\tilde{f}_T(G,h)\tilde{L}_T(G,h)||dh \stackrel{p}{\to} 0 \tag{28}
$$

$$
E_G \int |w_c^0(h) L_T(G,h) - w_c^0(0)\tilde{L}_T(G,h)|dh \stackrel{p}{\to} 0. \tag{29}
$$

These imply (27) if we can show that $\tilde{d}_T^{-1} = O_p(1)$ with $\tilde{d}_T = E_G \int \tilde{L}_T(G,h)dh$. Write E_h^* for the integration over the ball in $\mathbb{R}^{pk^2+k}$ of volume 1 and centered at zero, so that $\tilde{d}_T \geq E_h^* E_G \tilde{L}_T(G,h)$. Then

$$
\tilde{d}_T^{-1} \leq \frac{1}{E_h^* E_G \tilde{L}_T(G,h)} \leq \exp[-E_h^* E_G \ln \tilde{L}_T(G,h)]
$$

where the second inequality follows from Jensen, and

$$
\begin{aligned}
-2E_h^* E_G \ln \tilde{L}_T(G,h) &= E_G \tilde{\gamma}(G)'\tilde{\gamma}(G) + E_h^*(b'\hat{m}'\hat{m}b) + E_h^*(a'\Omega^{-1}a) \\
&= \operatorname{tr}\Sigma + \operatorname{tr}(\hat{m}'\hat{m}E_h^*[hh']) + \operatorname{tr}(\Omega^{-1}E_h^*(aa')) = O_p(1)
\end{aligned}
$$

using Lemma 4 (i). We are thus left to show (28) and (29). We focus on (28), which is strictly more difficult.

Now

$$
\begin{aligned}
&\int E_G ||w_c^0(h) f_T(G,h) L_T(G,h) - w_c^0(0)\tilde{f}_T(G,h)\tilde{L}_T(G,h)||dh \\
\leq &\int E_G ||\mathcal{S}_T(G) w_c^0(h) f_T(G,h) L_T(G,h) - w_c^0(0)\tilde{f}_T(G,h)\tilde{L}_T(G,h)||dh
\end{aligned}
$$

$$+\int E_G(1-\mathcal{S}_T(G))w_c^0(h)||f_T(G,h)||L_T(G,h)dh.$$

The last term converges in probability to zero by Lemma 2 (ii). By Hoelder, Lemma 7 (ii), Jensen and Minkowksi, we obtain

$$\begin{aligned}
& \int E_G||\mathcal{S}_T(G)w_c^0(h)f_T(G,h)L_T(G,h)-w_c^0(0)\tilde{f}_T(G,h)\tilde{L}_T(G,h)||dh \\
\leq & \int (E_G(\tilde{L}_T(G,h)^4))^{1/4}\cdot(E_G||\mathcal{S}_T(G)w_c^0(h)f_T(G,h)\frac{L_T(G,h)}{\tilde{L}_T(G,h)}-w_c^0(0)\tilde{f}_T(G,h)||^{4/3})^{3/4}dh \\
= & \zeta_T E_h(E_G||\mathcal{S}_T(G)w_c^0(h)f_T(G,h)\frac{L_T(G,h)}{\tilde{L}_T(G,h)}-w_c^0(0)\tilde{f}_T(G,h)||^{4/3})^{3/4} \\
\leq & \zeta_T (E_hE_G||\mathcal{S}_T(G)w_c^0(h)f_T(G,h)\frac{L_T(G,h)}{\tilde{L}_T(G,h)}-w_c^0(0)\tilde{f}_T(G,h)||^{4/3})^{3/4} \\
\leq & \zeta_T (E_hE_G||\mathcal{S}_T(G)w_c^0(h)f_T(G,h)\frac{L_T(G,h)}{\tilde{L}_T(G,h)}-w_c^0(0)\mathcal{S}_T(G)\tilde{f}_T(G,h)||^{4/3})^{3/4} \\
& +\zeta_T w_c^0(0)(E_GE_h||(1-\mathcal{S}_T(G))\tilde{f}_T(G,h)||^{4/3})^{3/4}
\end{aligned}$$

where $\zeta_T = O_p(1)$, and by Hoelder and Lemmas 2 (i) and 6 (i),

$$(E_GE_h||(1-\mathcal{S}_T(G))\tilde{f}_T(G,h)||^{4/3})^{3/4} \leq (E_G(1-\mathcal{S}_T(G)))^{1/4}\cdot(E_hE_G||\tilde{f}_T(G,h)||^2)^{1/2} \xrightarrow{p} 0.$$

Now apply Lemma 1 with $E_c = E_hE_G$, $x_1 = \mathcal{S}_T(G)\mathcal{U}_T(h)f_T(G,h)$, $\tilde{x}_1 = \mathcal{S}_T(G)\mathcal{U}_T(h)\tilde{f}_T(G,h)$, $x_2 = w_c^0(0)$, $\tilde{x}_2 = w_c^0(h)$, $\tilde{x}_j = \mathcal{S}_T(G)$ for $j > 2$ and x_j various products in $\mathcal{S}_T(G)L_T(G,h)/\tilde{L}_T(G,h)$ and $\tilde{x}_j = 1$, so that Lemmas 3 and 6 imply

$$E_hE_G\mathcal{S}_T(G)||w_c^0(h)f_T(G,h)\frac{L_T(G,h)}{\tilde{L}_T(G,h)}-w_c^0(0)\tilde{f}_T(G,h)||^{4/3} \xrightarrow{p} 0$$

as was left to be shown. We therefore established (27).

Now by Lemma 9,

$$\begin{aligned}
& \frac{E_G\int w_c^0(0)\tilde{f}_T(G,h)\tilde{L}_T(G,h)dh}{E_G\int w_c^0(0)\tilde{L}_T(G,h)dh} - T^{1/2}J_1'(A(\hat{b})^\ell - A(0)^\ell)Y_T \\
= & \sum_{l=1}^{\ell} J_1'A(\hat{b})^{\ell-l}J_1[\hat{a}+P(\tilde{R}\hat{e}\otimes I)'\tilde{\Delta}_l(\gamma^p+\hat{m}b^p)] \\
= & \sum_{l=1}^{\ell} J_1'A(\hat{b})^{\ell-l}J_1[\hat{a}+P(\tilde{R}\hat{e}\otimes I)'\tilde{\Delta}_l(\hat{s}-\hat{S}+\hat{M}(\hat{m}'\hat{M})^{-1}\hat{m}'\hat{S})]
\end{aligned}$$

and the first claim follows from Lemma 5 (ii).

The second claim of Theorem 1 follows from the same arguments as in the proof of Theorem 4 of Müller and Petalas (2010), using (29) and the last claim of Lemma 9.

References

Brockwell, P. J., and R. A. Davis (1991): *Time Series: Theory and Methods.* Springer, New York, second edn.

Carter, C. K., and R. Kohn (1997): "Semiparametric Bayesian Inference for Time Series with Mixed Spectra," *Journal of the Royal Statistical Society Series B*, 59, 255–268.

Chernoff, H., and S. Zacks (1964): "Estimating the Current Mean of a Normal Distribution Which is Subject to Changes in Time," *Annals of Mathematical Statistics*, 35, 999–1028.

Choudhuri, N., S. Ghosal, and A. Roy (2004a): "Bayesian Estimation of the Spectral Density of a Time Series," *Journal of the American Statistical Association*, 99, 1050–1059.

——— (2004b): "Contiguity of the Whittle Measure for a Gaussian Time Series," *Biometrika*, 91, 211–218.

Clements, M. P., and D. F. Hendry (1998): *Forecasting Economic Time Series.* Cambridge University Press, Cambridge, UK.

Davies, R. B. (1973): "Asymptotic Inference in Stationary Gaussian Time-Series," *Advances in Applied Probability*, 5, 469–497.

Doan, T., R. Litterman, and C. Sims (1984): "Forecasting and Conditional Projection Using Realistic Prior Distributions," *Econometric Reviews*, 3, 1–100.

Dzhaparidze, K. (1986): *Parameter Estimation and Hypothesis Testing in Spectral Analysis of Stationary Time Series.* Springer, New York.

Golubev, G. K., M. Nussbaum, and H. H. Zhou (2010): "Asymptotic Equivalence of Spectral Density Estimation and Gaussian White Noise," *The Annals of Statistics*, 38, 181–214.

Ing, C., and C. Wei (2005): "Order Selection for Same-Realization Predictions in Autoregressive Processes," *The Annals of Statistics*, 33, 2423–2474.

Li, H., and U. K. Müller (2009): "Valid Inference in Partially Unstable General Method of Moment Models," *Review of Economic Studies*, 76, 343–365.

Liseo, B., D. Marinucci, and L. Petrella (2001): "Bayesian Semiparametric Inference on Long-Range Dependence," *Biometrika*, 88, 1089–1104.

Lütkepohl, H. (2005): *New Introduction to Multiple Time Series Analysis*. Springer, Berlin.

McCoy, E. J., and D. A. Stephens (2004): "Bayesian Time Series Analysis of Periodic Behaviour and Spectral Structure," *International Journal of Forecasting*, 20, 713–730.

Müller, U. K. (2009): "Risk of Bayesian Inference in Misspecified Models, and the Sandwich Covariance Matrix," *Working paper, Princeton University.*

Müller, U. K., and P. Petalas (2010): "Efficient Estimation of the Parameter Path in Unstable Time Series Models," *Review of Economic Studies*, 77, 1508–1539.

Pesaran, M. H., D. Pettenuzzo, and A. Timmermann (2006): "Forecasting Time Series Subject to Multiple Structural Breaks," *Review of Economic Studies*, 73, 1057–1084.

Rosen, O., D. Stoffer, and S. Wood (2009): "Local Spectral Analysis Via a Bayesian Mixture of Smoothing Splines," *Journal of the American Statistical Association*, 104, 249–262.

Schorfheide, F. (2005): "VAR Forecasting under Misspecification," *Journal of Econometrics*, 128, 99–136.

Shibata, R. (1980): "Asymptotically Efficient Selection of the Order of the Model for Estimating Parameters of a Linear Process," *The Annals of Statistics*, 8, 147–164.

Stock, J. H., and M. W. Watson (1996): "Evidence on Structural Instability in Macroeconomic Time Series Relations," *Journal of Business and Economic Statistics*, 14, 11–30.

——— (2005): "Implications of Dynamic Factor Models for VAR Analysis," *manuscript, Harvard University.*

Whittle, P. (1957): "Curve and Periodogram Smoothing," *Journal of the Royal Statistical Society Series B*, 19, 38–63.

——— (1962): "Gaussian Estimation in Stationary Time Series," *Bulletin of the International Statistical Institute*, 39, 105–129.

Table 1

Distributions of Estimated Lag Differences, AIC – BIC:

Empirical (132 monthly U.S. macroeconomic time series) and Asymptotic

$\hat{p}_{AIC} - \hat{p}_{BIC}$	Empirical	Asymptotic						
			p_0=0			p_0=3		
		AR(0)	c=10	c=20	c=40	c=10	c=20	c=40
0	0.205	0.714	0.462	0.185	0.021	0.646	0.482	0.169
1	0.098	0.114	0.196	0.175	0.043	0.132	0.157	0.116
2	0.083	0.056	0.114	0.157	0.070	0.071	0.101	0.112
3	0.098	0.034	0.071	0.122	0.089	0.044	0.070	0.103
4	0.068	0.023	0.045	0.093	0.100	0.030	0.049	0.089
5	0.030	0.016	0.031	0.069	0.100	0.021	0.035	0.079
6	0.038	0.011	0.022	0.051	0.094	0.015	0.027	0.067
7	0.045	0.008	0.016	0.037	0.086	0.011	0.020	0.057
8	0.045	0.006	0.011	0.028	0.075	0.008	0.015	0.046
9	0.030	0.005	0.008	0.021	0.064	0.006	0.012	0.039
10	0.015	0.003	0.006	0.016	0.054	0.005	0.009	0.032
11	0.053	0.003	0.005	0.012	0.046	0.004	0.008	0.026
≥12	0.189	0.008	0.014	0.035	0.157	0.008	0.016	0.065

Notes: Entries are the probability of observing the indicated difference between AIC and BIC. The empirical distribution is computed from the data set described in Section 4 consisting of 132 U.S. monthly macroeconomic time series, 1959:1 – 2003:12, with the regressions all run on a balanced panel with $T = 510$ observations, and $0 \le p \le p_{\max}=18$ for both AIC and BIC. The "Asymptotic AR(0)" reports the asymptotic distribution of $\hat{p}_{AIC} - \hat{p}_{BIC}$ in model (1) with white noise u_t, and the remaining columns report the asymptotic distribution in model (1) and (2) with G distributed as c times a demeaned Brownian motion.

Table 2
Mean Square Forecast Errors of
Approximate and Small Sample Exact Bayes Procedures

	p=0	p=1	p=2	p=3	p=6
	T=100, c=10				
Exact Bayes $y^*_{T+1\|T}$	-4.6	-1.0	-0.4	-0.4	-0.2
Approx. $y^p_{T+1\|T}$	-4.1	-1.0	-0.3	-0.2	-0.1
Difference	-0.6	0.0	-0.1	-0.2	-0.2
	T=100, c=20				
Exact Bayes $y^*_{T+1\|T}$	-22.2	-8.9	-5.0	-2.3	-0.4
Approx. $y^p_{T+1\|T}$	-7.5	-7.2	-4.2	-2.0	-0.3
Difference	-14.7	-1.7	-0.7	-0.3	-0.2
	T=100, c=30				
Exact Bayes $y^*_{T+1\|T}$	-49.1	-22.6	-10.1	-6.6	-2.0
Approx. $y^p_{T+1\|T}$	31.9	-8.6	-7.0	-4.9	-1.5
Difference	-81.1	-13.9	-3.1	-1.7	-0.5
	T=200, c=10				
Exact Bayes $y^*_{T+1\|T}$	-5.4	-0.9	-0.3	-0.4	-0.2
Approx. $y^p_{T+1\|T}$	-5.2	-0.9	-0.3	-0.3	-0.1
Difference	-0.2	0.0	0.0	0.0	-0.1
	T=200, c=20				
Exact Bayes $y^*_{T+1\|T}$	-25.5	-7.7	-2.9	-2.1	-0.5
Approx. $y^p_{T+1\|T}$	-18.6	-6.8	-2.8	-2.1	-0.4
Difference	-6.9	-1.0	-0.1	0.0	-0.1
	T=200, c=30				
Exact Bayes $y^*_{T+1\|T}$	-55.7	-21.0	-10.6	-6.8	-1.8
Approx. $y^p_{T+1\|T}$	-14.6	-14.2	-8.6	-6.0	-1.7
Difference	-41.1	-6.8	-2.1	-0.8	-0.1
	T=400, c=10				
Exact Bayes $y^*_{T+1\|T}$	-6.2	-1.0	-0.4	0.0	0.0
Approx. $y^p_{T+1\|T}$	-6.1	-1.0	-0.4	0.0	0.0
Difference	-0.1	0.0	0.0	0.0	0.0
	T=400, c=20				
Exact Bayes $y^*_{T+1\|T}$	-29.0	-7.1	-2.8	-1.4	-0.3
Approx. $y^p_{T+1\|T}$	-24.6	-7.0	-3.0	-1.4	-0.3
Difference	-4.4	-0.1	0.2	0.0	0.0
	T=400, c=30				
Exact Bayes $y^*_{T+1\|T}$	-64.8	-20.9	-7.1	-6.4	-1.6
Approx. $y^p_{T+1\|T}$	-39.9	-17.5	-7.1	-6.3	-1.7
Difference	-24.9	-3.3	0.0	-0.2	0.1

Notes to Table 2: Entries are Monte Carlo estimates of $T(\text{MSFE}_{posterior\ mean} - \text{MSFE}_{AR})/\text{MSFE}_{AR}$, where MSFE_{AR} is the one-step ahead mean square forecast error of the AR(p) model estimated by OLS (where p is given in the column heading), and $\text{MSFE}_{posterior\ mean}$ is the small sample exact ($y^*_{T+1|T}$) and approximate ($y^p_{T+1|T}$) posterior mean of $y_{T+1|T}$, respectively (computed using AR(p) residuals, for p given in the column heading), along with the difference of these two numbers. For the exact Bayes posterior, the prior on the AR(p) coefficients and the constant is N(0, I_{p+1}), independent of G. Based on 1000 Monte Carlo draws, with analytical integration over y_{T+1} - $y^*_{T+1|T}$. The exact Bayes procedure is implemented via importance sampling over *G* (discretely approximated with 2,000 steps) using the exact (not Whittle) likelihood, with the prior as proposal, using up to 1,000,000 draws.

Table 3

Pseudo Out-of-Sample Mean Square Forecast Errors of Univariate Approximate Posterior Mean Forecasts, Relative to BIC, for 132 Monthly U.S. Macro Time Series, 1959:1 – 2003:12

	AIC	A. Demeaned Brownian motion prior					B. Additional Below Business Cycle Frequency Misspecification			
ℓ		**c = 10**	**c = 20**	**c = 30**	**BMA, 0≤*c*≤20**	**BMA, 0≤*c*≤40**	**c = 0**	**c = 20**	**c = 30**	**BMA, 0≤*c*≤40**
Mean										
1	1.025	0.992	0.988	0.988	0.991	0.989	0.998	0.987	0.988	0.988
3	1.010	0.987	0.977	0.975	0.984	0.978	0.996	0.975	0.973	0.976
6	0.982	0.983	0.968	0.963	0.978	0.970	0.992	0.963	0.958	0.966
12	0.977	0.983	0.968	0.962	0.978	0.969	0.987	0.959	0.954	0.962
Median										
1	1.002	0.992	0.988	0.988	0.991	0.989	1.000	0.987	0.987	0.988
3	1.002	0.991	0.987	0.985	0.992	0.988	0.999	0.984	0.983	0.986
6	0.997	0.988	0.980	0.976	0.989	0.982	0.998	0.974	0.969	0.976
12	0.995	0.990	0.983	0.982	0.990	0.987	0.997	0.973	0.971	0.975
10% Percentile										
1	0.956	0.981	0.964	0.953	0.976	0.964	0.992	0.960	0.951	0.962
3	0.892	0.963	0.933	0.914	0.955	0.934	0.981	0.925	0.912	0.933
6	0.858	0.955	0.909	0.888	0.940	0.925	0.964	0.900	0.880	0.913
12	0.836	0.956	0.905	0.873	0.938	0.911	0.933	0.869	0.849	0.876
90% Percentile										
1	1.135	1.006	1.014	1.023	1.006	1.011	1.004	1.012	1.020	1.012
3	1.150	1.004	1.012	1.020	1.005	1.008	1.011	1.012	1.018	1.011
6	1.091	1.003	1.008	1.015	1.003	1.006	1.018	1.015	1.023	1.016
12	1.077	1.003	1.009	1.019	1.003	1.008	1.029	1.026	1.034	1.023

Notes: Entries are the relative mean square forecast error of cumulative ℓ-step ahead approximate posterior mean forecasts ($y^p_{T+1|T}$ + … + $y^p_{T+\ell|T}$) for real series, and of average inflation over the next ℓ months for nominal series, relative to the BIC forecast with the same lag length. The posterior mean forecast in Panel A was computed using the value of c (or BMA-weighted average over c) given in the column heading, using the demeaned Brownian motion prior; Panel B uses the prior corresponding to the sum of demeaned Brownian motion of scale c plus a demeaned truncated Brownian motion with variation below frequencies corresponding to cycles of 96 months with fixed scale 20. Lag lengths were chosen by BIC, with $0 \le p \le 18$. MSFEs were computed using recursive forecasts, with the first forecast made when there were 198 observations on the transformed variable (for real series,1975:7) and the final forecast made in 2003:12 – ℓ.

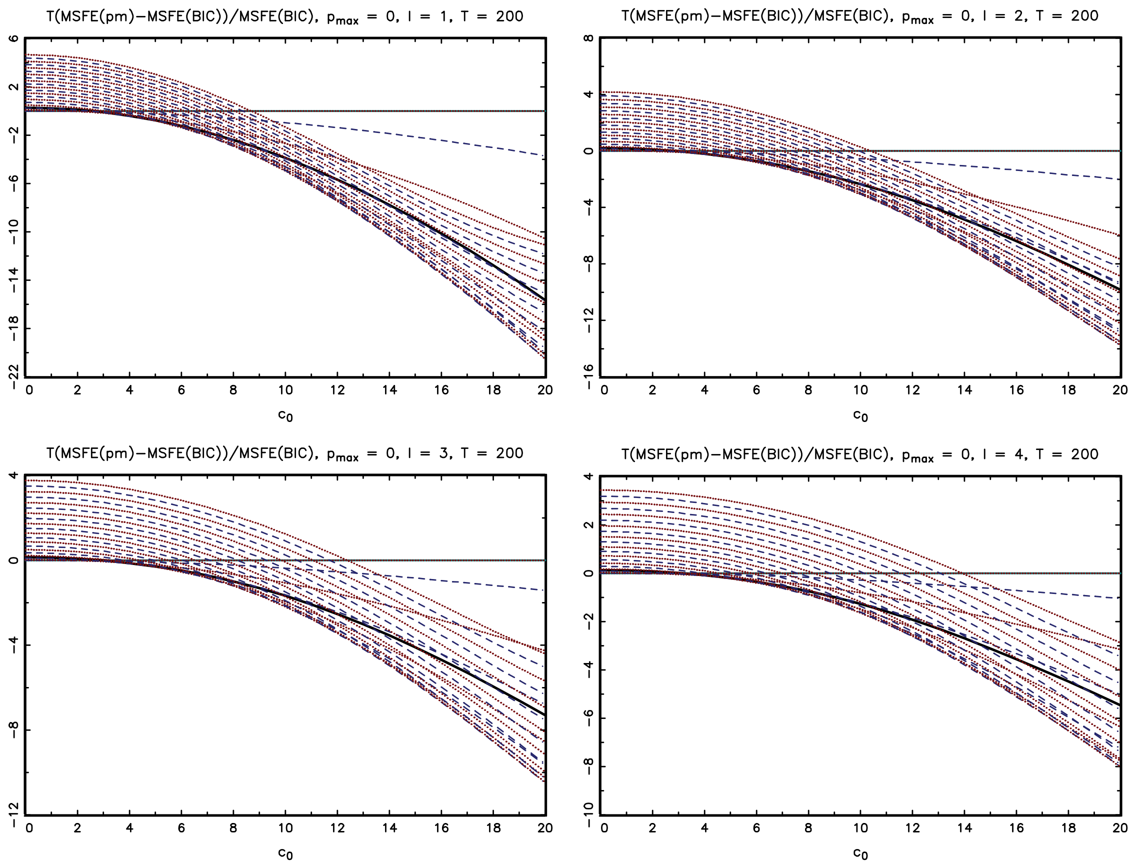

Figure 1

Monte Carlo estimates of T (MSFE$_{posterior\ mean}$ – MSFE$_{BIC}$)/MSFE$_{BIC}$ for cumulative forecasts, $\ell = 1,\ldots, 4$

Experiment 2: Estimator prior cW^{μ}, DGP prior c_0W^{μ} for c, $c_0 = 0,\ldots, 20$, $T = 200$

No estimated AR component

Dashed lines: MSFEs for fixed-c estimators; Solid line: BMA forecast

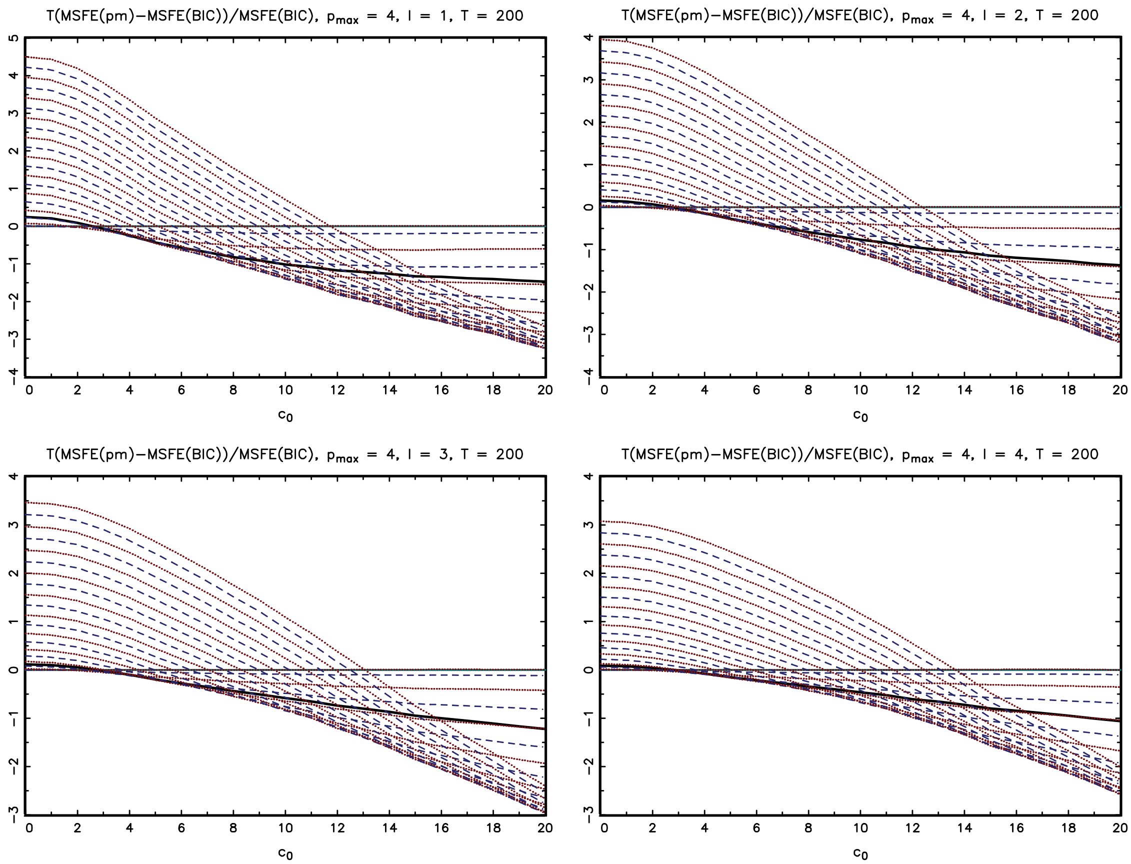

Figure 2

Monte Carlo estimates of $T(\text{MSFE}_{posterior\ mean} - \text{MSFE}_{BIC})/\text{MSFE}_{BIC}$ for cumulative forecasts, $\ell = 1,\ldots, 4$

Experiment 2: Estimator prior cW^{μ}, DGP prior c_0W^{μ} for c, $c_0 = 0,\ldots, 20$, $T = 200$

AR(BIC) component, $0 \le p \le 4$

Dashed lines: MSFEs for fixed-c estimators; Solid line: BMA forecast

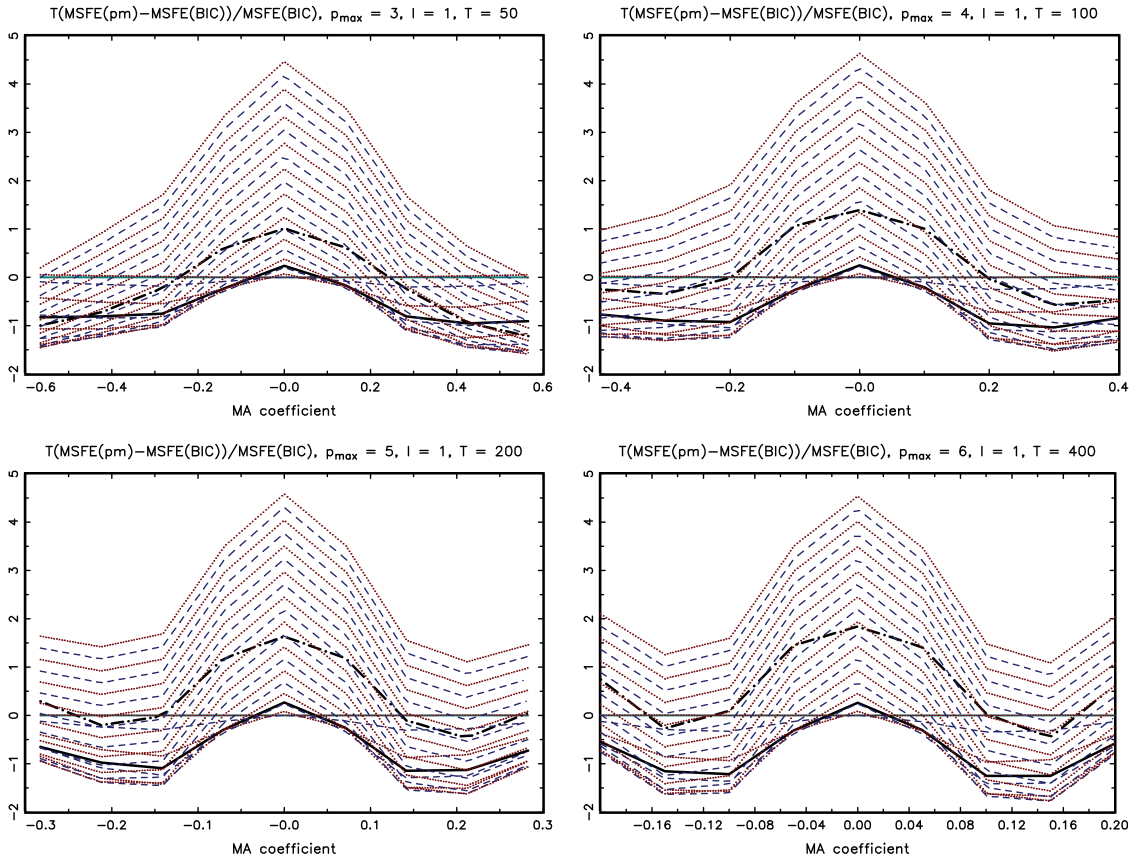

Figure 3

Monte Carlo estimates of $T(\text{MSFE}_{posterior\ mean} - \text{MSFE}_{BIC})/\text{MSFE}_{BIC}$ for $\ell = 1$

Experiment 3: MA(1) DGP; estimator prior cW^{μ}, $c = 0,\ldots, 20$

AR(BIC) component, $0 \le p \le 3$, 4, 5, 6 for $T = 50$, 100, 200, 400

Dashed lines: fixed-c estimators; Solid line: BMA; Heavy dash-dot: AIC

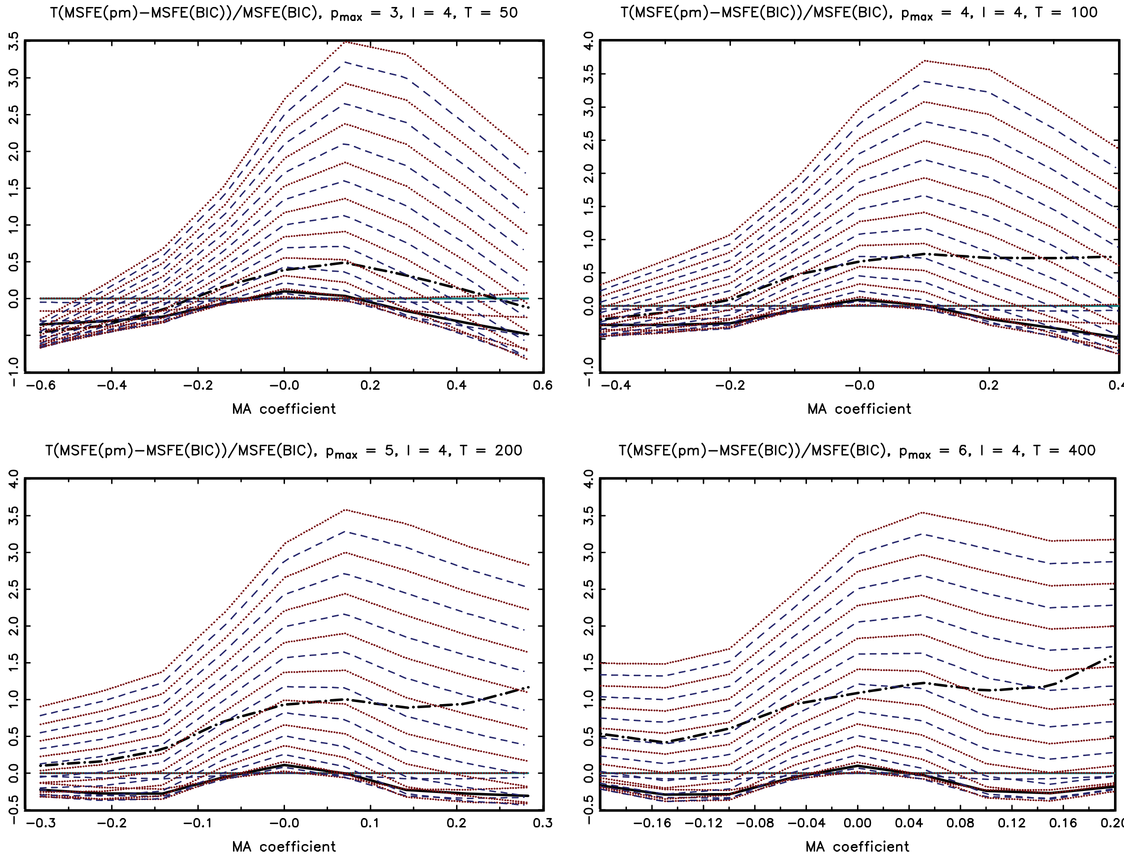

Figure 4

Monte Carlo estimates of $T(\text{MSFE}_{posterior\ mean} - \text{MSFE}_{BIC})/\text{MSFE}_{BIC}$ for cumulative forecasts, $\ell = 4$

Experiment 3: MA(1) DGP; estimator prior cW^{μ}, $c = 0,\ldots, 20$

AR(BIC) component, $0 \leq p \leq 3, 4, 5, 6$ for $T = 50, 100, 200, 400$

Dashed lines: fixed-c estimators; Solid line: BMA; Heavy dash-dot: AIC

Supplementary Material to
Forecasts in a Slightly Misspecified Finite Order VAR

Proof of Lemma 1: (i) Define $z_j(0) = ||\tilde{x}_j||$ and $z_j(1) = ||x_j - \tilde{x}_j||$. Then

$$||\prod_{j=1}^{l} x_j - \prod_{j=1}^{l} \tilde{x}_j|| \leq \sum_{v \in S_v} \prod_{j=1}^{l} z_j(v_j)$$

almost surely, where S_v is the set of all vectors $v = (v_1, \cdots, v_l)' \in \mathbb{R}^l$ with elements v_j equal to zero or one, and $v'v \geq 1$. Let $\#S_v$ be the number of elements in S_v. Then, for any $p > 1$, by convexity

$$(\frac{1}{\#S_v} \sum_{v \in S_v} \prod_{j=1}^{l} z_j(v_j))^p \leq \frac{1}{\#S_v} \sum_{v \in S_v} \prod_{j=1}^{l} z_j(v_j)^p$$

Furthermore, by Hoelder,

$$E_c|| \prod_{j=1}^{l} x_j - \prod_{j=1}^{l} \tilde{x}_j||^{4/3} \leq (\#S_v)^{4/3} \sum_{v \in S_v} \prod_{j=1}^{l} \left(E_c z_j(v_j)^{4p_j/3} \right)^{1/p_j}$$

for any $p_j > 1$,$j = 1, \cdots, l$ such that $\sum_{j=1}^{l} 1/p_j = 1$. The result now follow from setting $p_1 = 3/2$.

(ii) By Jensen, it suffices to consider even K. The result then follows from applying the premise to $(\tilde{x}_j - 1)^K = \sum_{j=0}^{K} b_{K,j} \tilde{x}_j^j (-1)^{K-j}$, where $b_{K,j}$ are the binomial coefficients.

Proof of Lemma 2: Define $\mathcal{S}_T^c(G) = 1 - \mathcal{S}_T(G)$, and $d_T(G, \omega) = T^{1/2} \exp[T^{-1/2} G(\omega)] - T^{1/2} I_k - G(\omega)$.

(i) By the Isoperimetric Inequality, the tail of the random variable $\sup_{-\pi \leq \omega \leq \pi} ||G(\omega)||$ is dominated by a normal with sufficiently large mean and variance, so that $E_G \mathcal{S}_T^c(G) \to 0$ exponentially fast.

(ii) For any $\varepsilon > 0$, by Markov

$$\begin{aligned} P(E_G \int \mathcal{S}_T^c(G) w_c^0(h) ||f_T(G,h)|| L_T(G,h) dh &> \varepsilon) \leq \varepsilon^{-1} E_y E_G \int \mathcal{S}_T^c(G) w_c^0(h) ||f_T(G,h)|| L_T(G,h) dh \\ &= E_G \int \mathcal{S}_T^c(G) w_c^0(h) E_{y|G,h} ||f_T(G,h)|| dh \end{aligned}$$

where $E_{y|G,h}$ denotes integration over (e, y, X) in the model $(I_T \otimes P^{-1}) y \sim \mathcal{N}(X \operatorname{vec} \theta, V(G))$, where $\operatorname{vec} \theta = \operatorname{vec} \theta^0 + T^{-1/2} h$.

Define $F_T(G,h) = T^{-1/2} f_T(G,h) + J_1' A(0)^\ell Y_T$, which is the conditional mean of $y_{T+\ell}$ given (y, h, G) by definition of $f_T(G,h)$. Therefore $E_{y|G,h}[||F_T(G,h)||^2] \leq E_{y|G,h}[||y_{T+\ell}||^2] \leq \sup_{t \geq 1} E_{y|G,h}[||y_t||^2]$. Furthermore, $E_{y|G,h}[||J_1' A(0)^\ell Y_T||^2]$ can also be be bounded in terms of $\sup_{t \geq 1} E_{y|G,h}[||y_t||^2]$.

Now in model (18), $\sup_{t\geq 1} E_{y|G,h}[||y_t||^2] \leq \sup_{t\geq 1} E_\theta[||y_t||^2] \cdot \sup_{-\pi\leq\omega\leq\pi} ||\exp[T^{-1/2}G(\omega)]||$. Furthermore, $\sup_{-\pi\leq\omega\leq\pi} ||\exp[T^{-1/2}G(\omega)]|| \leq \sup_{-\pi\leq\omega\leq\pi} \exp[T^{-1/2}||G(\omega)||]$ and by Cauchy-Schwarz

$$E_G \int \mathcal{S}_T^c(G) \sup_{-\pi\leq\omega\leq\pi} ||\exp[T^{-1/2}G(\omega)]||dh \leq (E_G \mathcal{S}_T^c(G)) \cdot \left(E_G \exp[2T^{-1/2} \sup_{-\pi\leq\omega\leq\pi} ||G(\omega)||] \right)^{1/2}.$$

But $E_G \exp[2T^{-1/2} \sup_{-\pi\leq\omega\leq\pi} ||G(\omega)||] = O(1)$, so that

$$E_G \int \mathcal{S}_T^c(G) w_c^0(h) E_{y|G,h}||f_T(G,h)||dh \leq (C_1 + C_2 T^{k/2+1/2} \int \sup_{t\geq 1} E_\theta[||y_t||^2] w(\theta)d\theta) \cdot E_G \mathcal{S}_T^c(G)$$

for some large enough C_1 and C_2, and the result follows from the result of part (i), since $\varepsilon > 0$ was arbitrary.

(iii) By definition, $\exp[T^{-1/2}G(\omega)] = \sum_{j=0}^\infty (T^{-1/2}G(\omega))^j/j!$. Thus,

$$\sup_{-\pi\leq\omega\leq\pi} \mathcal{S}_T(G)||d_T(G,\omega)|| \leq \sum_{j=2}^\infty \frac{(T^{\kappa-1/2})^j}{j!} \leq 2T^{-1+2\kappa} \tag{30}$$

almost surely, and the first result follows from Lemma 3.1 (i) of Davis (1973). Furthermore, by Lemma 3.1 (iii) of Davies (1973), $\mathcal{S}_T(G)T^{-1} \operatorname{tr} D_\Gamma(G)^2 \leq \frac{1}{2\pi}\int_{-\pi}^{\pi} |\mathcal{S}_T(G)d_T(G,\omega)|^2 d\omega \leq 4kT^{-2+4\kappa}$.

(iv) Follows from Lemma 3.1 (i) of Davis (1973), the norm inequality, and part (iii).

(v) Follows from $||V(G)^{-1}|| \leq 1/(1 - T^{-1/2}||\Gamma(G)||)$ and part (iv).

(vi) By construction, $\gamma_j(G) - \tilde{\gamma}(G)_j = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{\mathbf{i}j\omega} d_T(G,\omega)d\omega$. Thus, by Lemma 3.1 (iii) of Davies (1973), $\mathcal{S}_T(G)||\gamma(G) - \tilde{\gamma}(G)||^2 \leq \frac{1}{2\pi}\int_{-\pi}^{\pi} |\mathcal{S}_T(G)d_T(G,\omega)|^2 d\omega \leq 4kT^{-2+4\kappa}$ a.s., as in the proof of part (iii).

(vii) Follows from the norm inequality and part (vi).

(viii) $E_G \mathcal{S}_T(G) \sum_{j=N+1}^T |\gamma_j(G)|^2 \leq 4kT^{-2+4\kappa} + E_G \sum_{j=N+1}^T |\tilde{\gamma}_j(G)|^2 = 4kT^{-2+4\kappa} + \operatorname{tr}\Sigma_{NT} \to 0$, where Σ_{NT} is defined in the proof of Lemma 3 (xi).

Proof of Lemma 3: Note that by Lemma 2 (i)

$$E_h E_G \exp[K\mathcal{S}_T(G)\xi_T(G,h)] - E_h E_G \mathcal{S}_T(G) \exp[K\xi_T(G,h)] = E_G(1 - \mathcal{S}_T(G)) \to 0$$

so it suffices to show the claim for $E_h E_G \exp[K\mathcal{S}_T(G)\xi_T(G,h)]$. We repeatedly rely on $||T^{-1/2}e|| = O_p(1)$, $||T^{-1/2}X|| = O_p(1)$ and $||\hat{h}|| = O_p(1)$. Also, note that it suffices to show that

$$E_y E_h E_G \exp[K\mathcal{S}_T(G)\xi_T(G,h)] \to 1 \tag{31}$$

for any K, since by Jensen, $E_y(E_h E_G \exp[K'\mathcal{S}_T(G)\xi_T(G,h)])^2 \leq E_y E_h E_G \exp[2K'\mathcal{S}_T(G)\xi_T(G,h)]$, so that (31) with $K = K', 2K'$ implies $E_h E_G \exp[K\mathcal{S}_T(G)\xi_T(G,h)] \to 1$ in quadratic mean.

(i) By Lemma A1.1 of Dzhaparidze (1986), for any matrix A with $||A|| < 1$,

$$|\ln\det(I + A) - \operatorname{tr} A + \tfrac{1}{2}\operatorname{tr} A^2| \leq \tfrac{1}{3}\frac{||A|| \cdot (\operatorname{tr} A^2)}{(1 - ||A||)^3} \leq \tfrac{1}{3}\frac{T||A||^3}{(1 - ||A||)^3}$$

where the second inequality follows from $\operatorname{tr} A^2 \leq T||A||^2$. Since by Lemma 2 (iv), $\mathcal{S}_T(G)||KT^{-1/2}\Gamma(G)|| \to 0$, for large enough T,

$$\begin{aligned} & \mathcal{S}_T(G)K|\ln\det(I + T^{-1/2}\Gamma(G)) - T^{-1/2}\operatorname{tr}\Gamma(G) + \tfrac{1}{2}T^{-1}\operatorname{tr}\Gamma(G)^2| \\ \leq\ & K\tfrac{1}{3}\mathcal{S}_T(G)T^{-1/2}\frac{||\Gamma(G)||^3}{(1 - T^{-1/2}||\Gamma(G)||)^3} \to 0 \end{aligned}$$

uniformly in G, where the convergence follows from Lemma 2 (iv).

(ii) With $V(G)^{-1} = I - T^{-1/2}\Gamma(G) + T^{-1}\Gamma(G)^2V(G)^{-1}$, we find

$$\begin{aligned} & |\mathcal{S}_T(G)Ke'(V(G)^{-1} - I + T^{-1/2}\Gamma(G) - T^{-1}\Gamma(G)^2)e| \\ =\ & |\mathcal{S}_T(G)T^{-3/2}Ke'\Gamma(G)^3V(G)^{-1}e| \leq T^{-1/2}||T^{-1/2}e||^2\mathcal{S}_T(G)K||V(G)^{-1}|| \cdot ||\Gamma(G)||^3 \xrightarrow{p} 0 \end{aligned}$$

uniformly in G, where the convergence follows from Lemma 2 (iv) and (v).

(iii)

$$\begin{aligned} & E_h\exp[\mathcal{S}_T(G)KT^{-1/2}e'(V(G)^{-1} - I + T^{-1/2}\Gamma(G))Xh] \\ =\ & E_h\exp[\mathcal{S}_T(G)KT^{-3/2}e'\Gamma(G)^2V(G)^{-1}Xh] \\ =\ & \exp[\tfrac{1}{2}\mathcal{S}_T(G)T^{-3}C_T^{-1}K^2||e'\Gamma(G)^2V(G)^{-1}X||^2 + \mathcal{S}_T(G)KT^{-3/2}e'\Gamma(G)^2V(G)^{-1}X\hat{h}] \\ \leq\ & \exp[\tfrac{1}{2}\mathcal{S}_T(G)C_T^{-1}K^2||T^{-1/2}e||^2||T^{-1/2}X||^2||V(G)^{-1}||^2T^{-1}||\Gamma(G)^2|| \\ & +\mathcal{S}_T(G)KT^{-1/2}||T^{-1/2}e|| \cdot ||\Gamma(G)^2|| \cdot ||V(G)^{-1}|| \cdot ||T^{-1/2}X|| \cdot ||\hat{h}||] \xrightarrow{p} 1 \end{aligned}$$

uniformly in G, where the convergence follows from Lemma 2 (iv) and (v).

(iv)

$$\begin{aligned} & E_h\mathcal{S}_T(G)\exp[KT^{-1}h'X'(V(G)^{-1} - I)Xh] \\ \leq\ & \mathcal{S}_T(G)E_h\exp[||V(G)^{-1} - I||KT^{-1}h'X'Xh] \\ =\ & \mathcal{S}_T(G)\det(I_{pk^2+k} - C_T||V(G)^{-1} - I||KT^{-1}X'X)^{-1/2} \xrightarrow{p} 1 \end{aligned}$$

uniformly in G, since $\mathcal{S}_T(G)||V(G)^{-1} - I|| \leq 1/(1 - T^{-1/2}\mathcal{S}_T(G)||\Gamma(G)||)^{-1} - 1 \to 0$ uniformly in G from Lemma 2 (iv).

(v)

$$\begin{aligned} & E_y\exp[\tfrac{1}{2}\mathcal{S}_T(G)KT^{-1}e'\Gamma(G)^2e - \tfrac{1}{2}\mathcal{S}_T(G)KT^{-1}\operatorname{tr}\Gamma(G)^2] \\ =\ & \exp[-\tfrac{1}{2}\ln\det(I - \mathcal{S}_T(G)KT^{-1}\Gamma(G)^2) - \tfrac{1}{2}\mathcal{S}_T(G)KT^{-1}\operatorname{tr}\Gamma(G)^2] \end{aligned}$$

and for large enough T,

$$\mathcal{S}_T(G)KT^{-1}\operatorname{tr}\Gamma(G)^2 - \mathcal{S}_T(G)K^2T^{-2}\operatorname{tr}\Gamma(G)^4 \leq \ln\det(I - \mathcal{S}_T(G)KT^{-1}\Gamma(G)^2) \leq \mathcal{S}_T(G)KT^{-1}\operatorname{tr}\Gamma(G)^2$$

uniformly in G, because $||\mathcal{S}_T(G)KT^{-1}\Gamma(G)^2|| \to 0$ by Lemma 2 (iv), and for $|x| < 1/2$, $-x - x^2 \leq \ln(1 - x) \leq -x$. With $\operatorname{tr} A^4 \leq ||A||^2\operatorname{tr} A^2 \leq T||A||^4$, $T^{-2}\mathcal{S}_T(G)\operatorname{tr}\Gamma(G)^4 \leq \mathcal{S}_T(G)T^{-1}||\Gamma(G)||^4 \to 0$ uniformly in G by Lemma 2 (iv).

(vi)

$$E_y\exp[\mathcal{S}_T(G)KT^{-1/2}(e'D_\Gamma(G)e - \operatorname{tr} D_\Gamma(G))]$$

$$
\begin{aligned}
&= \exp[-\mathcal{S}_T(G)KT^{-1/2}\operatorname{tr} D_\Gamma(G)]\det(I-\mathcal{S}_T(G)2KT^{-1/2}D_\Gamma(G))^{-1/2} \\
&\le \exp[-\mathcal{S}_T(G)KT^{-1/2}\operatorname{tr} D_\Gamma(G)]\exp[\tfrac{1}{2}(\mathcal{S}_T(G)2KT^{-1/2}\operatorname{tr} D_\Gamma(G)+\mathcal{S}_T(G)4K^2T^{-1}\operatorname{tr} D_\Gamma(G)^2] \\
&= \exp[\mathcal{S}_T(G)4K^2T^{-1}\operatorname{tr} D_\Gamma(G)^2] \to 1
\end{aligned}
$$

uniformly in G by Lemma 2 (iii), where the inequality follows for sufficiently large T, because $||\mathcal{S}_T(G)2KT^{-1/2}D_\Gamma(G)|| \to 0$, and for $|x|<1/2$, $\ln(1-x)\ge -x-x^2$.

(vii)

$$
\begin{aligned}
\mathcal{S}_T(G)T^{-1}|\operatorname{tr}\Gamma(G)^2-\operatorname{tr}\tilde{\Gamma}(G)^2| &\le \mathcal{S}_T(G)T^{-1}\operatorname{tr} D_\Gamma(G)^2+2\mathcal{S}_T(G)T^{-1}|\operatorname{tr} D_\Gamma(G)\tilde{\Gamma}(G)| \\
&\le \mathcal{S}_T(G)T^{-1}\operatorname{tr} D_\Gamma(G)^2+2\mathcal{S}_T(G)(T^{-1}\operatorname{tr} D_\Gamma(G)^2)^{1/2}||T^{-1/2}\tilde{\Gamma}(G)|| \to 0
\end{aligned}
$$

uniformly in G by Lemma 2 (iii) and (iv).

(viii)

$$
\begin{aligned}
&E_h\exp[\mathcal{S}_T(G)KT^{-1}e'D_\Gamma(G)Xh] \\
&= \exp[\tfrac{1}{2}\mathcal{S}_T(G)C_T^{-1}K^2T^{-2}e'D_\Gamma(G)XX'D_\Gamma(G)e+\mathcal{S}_T(G)KT^{-1}e'D_\Gamma(G)X\hat{h}] \\
&\le \exp[\tfrac{1}{2}C_T^{-1}K^2||T^{-1/2}e||^2||T^{-1/2}X||^2\mathcal{S}_T(G)||D_\Gamma(G)||^2] \\
&\qquad\times\exp[K||T^{-1/2}e||\cdot\mathcal{S}_T(G)||D_\Gamma(G)||\cdot||T^{-1/2}X||\cdot||\hat{h}||] \xrightarrow{p} 1
\end{aligned}
$$

uniformly in G by Lemma 2 (iii).

(ix)

$$
\begin{aligned}
&E_h\exp[\mathcal{S}_T(G)KT^{-3/2}\hat{h}'X'\tilde{\Gamma}(G)h] \\
&= \exp[\tfrac{1}{2}\mathcal{S}_T(G)C_T^{-1}K^2T^{-3}\hat{h}'X'\tilde{\Gamma}(G)^2X\hat{h}+\mathcal{S}_T(G)KT^{-3/2}\hat{h}'X'\tilde{\Gamma}(G)\hat{h}] \\
&\le \exp[C_T^{-1}K^2\mathcal{S}_T(G)||T^{-1/2}\tilde{\Gamma}(G)||^2||T^{-1/2}X||^2||\hat{h}||^2+\mathcal{S}_T(G)KT^{-1}||\hat{h}||^2||T^{-1/2}X||\cdot||\tilde{\Gamma}(G)||] \xrightarrow{p} 1
\end{aligned}
$$

uniformly in G by Lemma 2 (iv).

(x)

$$
\exp[\mathcal{S}_T(G)KT^{-3/2}\hat{h}'X'\tilde{\Gamma}(G)X\hat{h}] \le \exp[\mathcal{S}_T(G)||T^{-1/2}\tilde{\Gamma}(G)||\cdot||T^{-1/2}X||^2||\hat{h}||^2] \xrightarrow{p} 1
$$

uniformly in G by Lemma 2.

(xi) By Lemma 3.1 of Davies (1973), $\sum_{j=-\infty}^{\infty}|\tilde{\gamma}_j(G)|^2=\frac{1}{2\pi}\int_{-\pi}^{\pi}|G(\omega)|^2d\omega<\infty$ a.s. Thus

$$
\begin{aligned}
|\tilde{\gamma}(G)'\tilde{\gamma}(G)-\tfrac{1}{2}T^{-1}\operatorname{tr}\tilde{\Gamma}(G)^2| &\le T^{-1}\sum_{j=1}^{N}j|\tilde{\gamma}_j(G)|^2+\sum_{j=N+1}^{T}|\tilde{\gamma}_j(G)|^2 \\
&\le \frac{N}{T}\sum_{j=1}^{N}|\tilde{\gamma}_j(G)|^2+\sum_{j=N+1}^{T}|\tilde{\gamma}_j(G)|^2 =: \varepsilon_\gamma(G)
\end{aligned}
$$

and a straightforward argument shows $\varepsilon_\gamma(G)\to 0$ for all G with square summable $|\tilde{\gamma}_j(G)|$. Furthermore,

$$
E_G\exp[2K\varepsilon_\gamma(G)] \le (E_G\exp[4K\frac{N}{T}\sum_{j=1}^{N}|\tilde{\gamma}_j(G)|^2])^{1/2}\cdot(E_G\exp[4K\sum_{j=N+1}^{T}|\tilde{\gamma}_j(G)|^2])^{1/2}.
$$

It is easy to see that $E_G \exp[4K\frac{N}{T}\sum_{j=1}^{N}|\tilde{\gamma}_j(G)|^2] = O(1)$. Since $\operatorname{tr}\Sigma = O(1)$ by Lemma 4 (i), $\operatorname{tr}\Sigma_{NT} \to 0$, where Σ_{NT} is the prior covariance matrix of $\tilde{\gamma}_{NT}(G) = \operatorname{vec}(\tilde{\gamma}_{N+1}(G), \cdots, \tilde{\gamma}_T(G))$. Therefore also $||\Sigma_{NT}|| \to 0$, and for large enough T,

$$
\begin{aligned}
E_G \exp[4K\sum_{j=N+1}^{T}|\tilde{\gamma}_j(G)|^2] &= \det(I - 8K\Sigma_{NT})^{-1/2} \\
&= \exp[-\tfrac{1}{2}\ln\det(I - 8K\Sigma_{NT})] \\
&\leq \exp[\tfrac{1}{2}16K\operatorname{tr}\Sigma_{NT}] \to 1
\end{aligned}
$$

where $\ln\det(I - 8K\Sigma_{NT}) \geq 16K\operatorname{tr}\Sigma_{NT}$ because for $0 \leq x < 1/2$, $\ln(1-x) \geq -2x$. Thus $\exp[K\varepsilon_\gamma(G)]$ has uniformly bounded second moments, so that $\varepsilon_\gamma(G) \to 0$ implies $E_G \exp[K\varepsilon_\gamma(G)] \to 1$.

(xii) Clearly, $\hat{m}'\hat{m} - m'm \xrightarrow{p} 0$, so that from a law of large numbers, with $\hat{D}_X = \operatorname{diag}(\Omega^{-1}, \hat{m}'\hat{m})$,

$$T^{-1}X'X - \hat{D}_X \xrightarrow{p} 0.$$

Thus,

$$E_h \exp[\tfrac{1}{2}Kh'(T^{-1}X'X - \hat{D}_X)h] \leq \det(I - C_T K(T^{-1}X'X - \hat{D}_X))^{-1/2} \xrightarrow{p} 1.$$

(xiii) In the notation of the proof of part (xii)

$$E_h \exp[Kh'(T^{-1}X'X - \hat{D}_X)\hat{h}] = \exp[\tfrac{1}{2}C_T^{-1}K^2\hat{h}'(T^{-1}X'X - \hat{D}_X)^2\hat{h} + K\hat{h}'(T^{-1}X'X - \hat{D}_X)\hat{h}] \xrightarrow{p} 1.$$

(xiv) Note that for any K',

$$
\begin{aligned}
E_h E_G \exp[K'(b-\hat{b})'(\hat{m}-m)'\tilde{\gamma}(G)] &= E_h \exp[K'^2(b-\hat{b})'(\hat{m}-m)'\Sigma(\hat{m}-m)(b-\hat{b})] \\
&\leq E_h \exp[K'^2||\Lambda^{-1}\Sigma\Lambda^{-1}|| \cdot ||(\hat{m}-m)'\Lambda^2(\hat{m}-m)||]
\end{aligned}
$$

and $||(\hat{m}-m)'\Lambda^2(\hat{m}-m)|| \xrightarrow{p} 0$ follows from $\hat{\theta} \xrightarrow{p} \theta^0$, and $||\Lambda^{-1}\Sigma\Lambda^{-1}| = O(1)$ by Lemma 4 (ii). The convergence of $E_h \exp[K'^2||\Lambda^{-1}\Sigma\Lambda^{-1}||\cdot||(\hat{m}-m)'\Lambda^2(\hat{m}-m)||] \xrightarrow{p} 1$ now follows as in the proof of part (xii). By Cauchy-Schwarz, it thus suffices to show the claim for $T^{-1}(h-\hat{h})'X'\tilde{\Gamma}(G)e - (b-\hat{b})'m'\tilde{\gamma}(G)$.

The first $k \times 1$ block of $T^{-1}X'\tilde{\Gamma}(G)e$ is equal to

$$
\begin{aligned}
& \sum_{j=1}^{T-1} T^{-1} \sum_{t=j+1}^{T} P^{-1\prime}\tilde{\gamma}_j(G)e_{t-j} + \sum_{j=1}^{T-1} T^{-1} \sum_{t=j+1}^{T} P^{-1\prime}\tilde{\gamma}_j(G)'e_t \\
=& \sum_{j=1}^{T-1} \left(T^{-1} \sum_{t=j+1}^{T} e_{t-j}' \otimes P^{-1\prime} \right) \operatorname{vec}(\tilde{\gamma}_j(G)) + \sum_{j=1}^{T-1} \left(T^{-1} \sum_{t=j+1}^{T} e_t' \otimes P^{-1\prime} \right) \operatorname{vec}\left(\tilde{\gamma}_j(G)'\right) \\
=& \sum_{j=1}^{T-1} \left(T^{-1} \sum_{t=j+1}^{T} (e_{t-j} \otimes P^{-1} + C_{kk}(e_t \otimes P^{-1})) \right)' \operatorname{vec}(\tilde{\gamma}_j(G)) \\
=& \bar{\xi}(j)' \operatorname{vec}(\tilde{\gamma}_j(G))
\end{aligned}
$$

where C_{kk} is the commutation matrix that satisfies $C'_{kk}\operatorname{vec}(\tilde\gamma_j(G)) = \operatorname{vec}(\tilde\gamma_j(G)')$, and the following $k^2\times 1$ blocks of $T^{-1}X'\tilde\Gamma(G)e$ are given by, $i=1,\cdots,p$

$$\begin{aligned}
&\sum_{j=1}^{T-1} T^{-1}\sum_{t=j+1}^{T}[y_{t-i}\otimes P^{-1\prime}]\tilde\gamma_j(G)e_{t-j} + \sum_{j=1}^{T-1}T^{-1}\sum_{t=j+1}^{T}[y_{t-j-i}\otimes P^{-1\prime}]\tilde\gamma_j(G)'e_t\\
=\;& \sum_{j=1}^{T-1}\operatorname{vec}\left(P^{-1\prime}\tilde\gamma_j(G)T^{-1}\sum_{t=j+1}^{T}e_{t-j}y'_{t-i}\right) + \sum_{j=1}^{T-1}\operatorname{vec}\left(P^{-1\prime}\tilde\gamma_j(G)'T^{-1}\sum_{t=j+1}^{T}e_t y'_{t-j-i}\right)\\
=\;& \sum_{j=1}^{T-1}\left((T^{-1}\sum_{t=j+1}^{T}e_{t-j}y'_{t-i})'\otimes P^{-1\prime}\right)\operatorname{vec}(\tilde\gamma_j(G)) + \sum_{j=1}^{T-1}\left((T^{-1}\sum_{t=j+1}^{T}e_t y'_{t-j-i})'\otimes P^{-1\prime}\right)\operatorname{vec}\left(\tilde\gamma_j(G)'\right)\\
=\;& \sum_{j=1}^{T-1}\left(T^{-1}\sum_{t=j+1}^{T}(e_{t-j}y'_{t-i}\otimes P^{-1} + C_{kk}(e_t y'_{t-j-i}\otimes P^{-1}))\right)'\operatorname{vec}(\tilde\gamma_j(G))\\
=\;& \bar m(i)'_j\operatorname{vec}(\tilde\gamma_j(G)).
\end{aligned}$$

Thus

$$T^{-1}X'\tilde\Gamma(G)e = \begin{pmatrix} \bar\xi'\tilde\gamma(G) + \bar\xi'_{NT}\tilde\gamma_{NT}(G)\\ \bar m'\tilde\gamma(G) + \bar m'_{NT}\tilde\gamma_{NT}(G)\end{pmatrix}$$

where $\tilde\gamma_{NT}(G) = \operatorname{vec}(\tilde\gamma_{N+1}(G)',\cdots,\tilde\gamma_T(G)')$, the jth $k^2\times k$ block of $\bar\xi$ and $\bar\xi_{NT}$ equals $\bar\xi(j)$ and $\bar\xi(j+1)$, respectively, and the jth $k^2\times k^2p$ block of $\bar m$ and $\bar m_{NT}$ equals $\bar m_j = (\bar m(1)'_j,\cdots,\bar m(p)'_j)'$ and $\bar m_{N+j} = (\bar m(1)'_{N+j},\cdots,\bar m(p)'_{N+j})'$, respectively.

Thus

$$\begin{aligned}
T^{-1}(h-\hat h)'X'\tilde\Gamma(G)e - (b-\hat b)'m'\tilde\gamma(G) \;&=\; (a-\hat a)'\bar\xi'\tilde\gamma(G) + (a-\hat a)'\bar\xi'_{NT}\tilde\gamma_{NT}(G)\\
&\quad + (b-\hat b)'\bar m'_{NT}\tilde\gamma_{NT}(G) + (b-\hat b)'(\bar m - m)'\tilde\gamma(G)
\end{aligned}$$

and by repeated applications of Cauchy-Schwarz and $E_h\exp[x_T|||\hat h - h|||] = O_p(1)$ for all $x_T = O_p(1)$, it suffices to show that $\bar\xi'\Lambda^2\bar\xi \xrightarrow{p} 0$, $\bar\xi'_{TN}\Lambda^2_{TN}\bar\xi_{TN}\xrightarrow{p} 0$, $\bar m'_{NT}\Lambda^2_{TN}\bar m_{NT}\xrightarrow{p} 0$, and $(\bar m - m)'\Lambda^2(\bar m - m)\xrightarrow{p} 0$, where $\Lambda_{TN} = \operatorname{diag}((N+1)^{-1},\cdots,T^{-1})\otimes I_{k^2}$. These claims follow from Markov's inequality after noting that $E_y|\bar\xi(j)|^2\to 0$ and $E_y|\bar m_j - m_j|^2\to 0$ for all j, and $\sup_{j\ge N}|m_j|\to 0$.

(xv) $E_h\exp[K\hat s'\hat m(b-\hat b)] \le \exp[\frac{1}{2}K^2C_T\hat s'\hat m\hat m'\hat s]\xrightarrow{p} 1$ by Lemma 8 (ii).

(xvi) Immediate from dominated convergence.

Proof of Lemma 4: (i) Let $k_{G,l}(r,s)$ be the lth diagonal element of $k_G(r,s)$. Note that $k_{G,l}(r,s)$ is a non-negative definite kernel, so that by by Mercer's Theorem, there exist orthonormal functions $\varphi_{l,j}\in L^2[-\pi,\pi]$ and nonnegative numbers $\nu_{l,j}$ such that $k_{G,l}(r,s) = \sum_{j=1}^{\infty}\nu_{l,j}\varphi_{l,j}(r)\varphi_{l,j}(s)^*$ (where $\nu_{l,j}$ are the eigenvalues and $\varphi_{l,j}$ are the eigenfunctions of $k_{G,l}(r,s)$), and $\int_{-\pi}^{\pi}k_{G,l}(s,s)ds = \sum_{j=1}^{\infty}\nu_{l,j}$. Note that $\int_{-\pi}^{\pi}\varphi_{l,j}(s)ds = 0$, since $\int_{-\pi}^{\pi}G(s)ds = 0$ a.s.. Since the functions $\{\phi_j(\omega)\}_{j=0}^{\infty} = \{e^{\mathbf{i}j\omega}/\sqrt{2\pi}\}_{j=0}^{\infty}$ form a basis of $L^2[-\pi,\pi]$, we have

$$2\pi\operatorname{tr}\Sigma \;\le\; 2\pi\lim_{N\to\infty}\operatorname{tr}\Sigma$$

$$
\begin{aligned}
&= \sum_{l=1}^{k^2}\sum_{j=1}^{\infty}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\phi_j(s)k_{G,l}(s,r)\phi_j(r)dsdr \\
&= \sum_{l=1}^{k^2}\sum_{j=1}^{\infty}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\phi_j(s)\left(\sum_{i=1}^{\infty}\nu_{l,i}\varphi_{l,i}(r)\varphi_{l,i}(s)\right)\phi_j(r)dsdr \\
&= \sum_{l=1}^{k^2}\sum_{j=1}^{\infty}\sum_{i=1}^{\infty}\nu_{l,i}(\int_{-\pi}^{\pi}\phi_j(s)\varphi_{l,i}(s)ds)^2 \\
&= \sum_{l=1}^{k^2}\sum_{i=1}^{\infty}\nu_{l,i} = \mathrm{tr}\int_{-\pi}^{\pi}k_G(s,s)ds.
\end{aligned}
$$

(ii) The j,lth $k\times k$ block of $\Lambda^{-1}\Sigma\Lambda^{-1}$ is equal to

$$
\frac{jl}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}e^{\mathbf{i}rj}e^{-\mathbf{i}sl}k_G(r,s)drds.
$$

Let $k_1(r,s) = \partial k_G(r,s)/\partial r$. By integration by parts,

$$
j\int_{-\pi}^{\pi}e^{\mathbf{i}rj}k_G(r,s)dr = -\mathbf{i}\int_{-\pi}^{\pi}e^{\mathbf{i}rj}k_1(r,s)dr =: g_j(s)
$$

where

$$
\begin{aligned}
\frac{dg_j(s)}{ds} &= -\mathbf{i}\frac{d}{ds}\left(\int_{-\pi}^{s}e^{\mathbf{i}rj}k_1(r,s)dr + \int_{s}^{\pi}e^{\mathbf{i}rj}k_1(r,s)dr\right) \\
&= -\mathbf{i}\int_{-\pi}^{\pi}e^{\mathbf{i}rj}k_2(r,s)dr - \mathbf{i}e^{\mathbf{i}sj}k_\Delta(s).
\end{aligned}
$$

Thus, by another application of integration by parts,

$$
\begin{aligned}
l\int_{-\pi}^{\pi}e^{-\mathbf{i}sl}g_j(s)ds &= \mathbf{i}\int_{-\pi}^{\pi}e^{-\mathbf{i}sl}\frac{dg_j(s)}{ds}ds \\
&= \int_{-\pi}^{\pi}\int_{-\pi}^{\pi}e^{\mathbf{i}rj}e^{-\mathbf{i}sl}k_2(r,s)drds + \int_{-\pi}^{\pi}e^{\mathbf{i}sj}e^{-\mathbf{i}sl}k_\Delta(s)ds.
\end{aligned}
$$

Now let $v = (v_1', \cdots, v_N')'$ be such that $v'v = 1$, where v_j is $k^2\times 1$. Then

$$
\begin{aligned}
2\pi v'\Lambda^{-1}\Sigma\Lambda^{-1}v &= \frac{1}{2\pi}\sum_{j,l=1}^{N}v_j'\left(\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}e^{\mathbf{i}rj}e^{-\mathbf{i}sl}k_2(r,s)drds + \int_{-\pi}^{\pi}e^{\mathbf{i}sj}e^{-\mathbf{i}sl}k_\Delta(s)ds\right)v_l \\
&= \int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\left(\frac{1}{\sqrt{2\pi}}\sum_{j=1}^{N}v_je^{\mathbf{i}rj}\right)'k_2(r,s)\left(\frac{1}{\sqrt{2\pi}}\sum_{j=1}^{N}v_je^{-\mathbf{i}sj}\right)drds \\
&\quad + \int_{-\pi}^{\pi}\left(\frac{1}{\sqrt{2\pi}}\sum_{j=1}^{N}v_je^{\mathbf{i}sj}\right)'k_\Delta(s)\left(\frac{1}{\sqrt{2\pi}}\sum_{j=1}^{N}v_je^{-\mathbf{i}sj}\right)ds
\end{aligned}
$$

$$= \int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\varphi(r)^* k_2(r,s)\varphi(s)drds + \int_{-\pi}^{\pi}\varphi(s)^* k_\Delta(s)\varphi(s)ds$$

with $\varphi(s) = \frac{1}{\sqrt{2\pi}}\sum_{j=1}^N v_j e^{-\mathbf{i}sj}$, and the result follows.

Proof of Lemma 5: (i) Define $\hat\mu_j = (P'\hat\Psi'_{j-1}, P'\hat\Psi'_{j-2}, \cdots, P'\hat\Psi'_{j-p})$, and $\hat\mu$ the $Nk \times pk$ matrix with block rows equal to $\hat\mu_j$, $j = 1, \cdots, N$. Note that $\hat m = \hat\mu \otimes P^{-1}$. From $y_{T-l} = \sum_{t=0}^{N-1}\hat\Psi_t(P\hat e_{T-t-l} + T^{-1/2}\hat a) + J_1' A(\hat b)^N Y_{T-N-l}$, we obtain

$$\tilde Y_T = \begin{pmatrix} \sum_{t=0}^N \hat\Psi_t P\hat e_{T-t} \\ \sum_{t=0}^{N+1}\hat\Psi_{t-1}P\hat e_{T-t} \\ \vdots \\ \sum_{t=0}^{N+p}\hat\Psi_{t-p+1}P\hat e_{T-t} \end{pmatrix} = (\bar\Delta_1\hat\mu)'\tilde R\hat e$$

so that $A(\hat b)^{l-1}\tilde Y_T = (\bar\Delta_l\hat\mu)'\tilde R\hat e$, where $\bar\Delta_l$ is the $Nk \times Nk$ matrix such that $\bar\Delta_l\hat\mu$ has tth block row equal to $(P'\hat\Psi'_{t+l-2}, P'\hat\Psi'_{t+l-3}, \cdots, P'\hat\Psi_{t+l-p-1})$, $t = 1, \cdots, T$. Now for $l = 1, \cdots, \ell$, using (21),

$$\begin{aligned} & ((A(\hat b)^{l-1}\tilde Y_T)' \otimes J_1' A(\hat b)^{\ell-l}J_1)(b-\hat b) \\ = & (((\bar\Delta_l\hat\mu)'\tilde R\hat e)' \otimes J_1'A(\hat b)^{\ell-l}J_1)(b-\hat b) \\ = & J_1'A(\hat b)^{\ell-l}J_1PP^{-1}(b-\hat b)(\bar\Delta_l\hat\mu)'\tilde R\hat e \\ = & J_1'A(\hat b)^{\ell-l}J_1P((\bar\Delta_l\hat\mu)'\tilde R\hat e \otimes P^{-1\prime})'\operatorname{vec}(b-\hat b) \\ = & J_1'A(\hat b)^{\ell-l}J_1P[(\tilde\Delta_l\hat m)'(\tilde R\hat e\otimes I)]'(b-\hat b) \\ = & J_1'A(\hat b)^{\ell-l}J_1P(\tilde R\hat e\otimes I)'\tilde\Delta_l\hat m(b-\hat b). \end{aligned}$$

(ii) Note that $\hat s - \hat S = \Sigma(I+\Sigma)^{-1}\hat s$, $\hat s - \hat D\widehat{\hat S} = \Sigma(\hat D + \Sigma)^{-1}\widehat{\hat s}$ and $(I_N\otimes P\otimes P)\hat s = (I_N \otimes \hat P\otimes\hat P)\widehat{\hat s}$, so that $\widehat{\hat s} = (I_N\otimes \hat P^{-1}P\otimes\hat P^{-1}P)\hat s$. Furthermore, $||\hat D - I|| \le |\hat D - I|$, and $|\hat D - I|^2 = \sum_{i=1}^N |\hat d_j - I_{k^2}|^2 = \sum_{i=1}^N \operatorname{vec}(\hat d_j - I_{k^2})'\operatorname{vec}(\hat d_j - I_{k^2})$. By standard arguments, $\max_j E|\hat d_j - I_{k^2}|^2 = O(T^{-1})$, so that $E[|\hat D - I|^2] = O(N/T) = o(1)$, and thus $||\hat D - I|| = o_p(1)$. Therefore, also $||\hat D|| = O_p(1)$ and $||\hat D^{-1}|| \le 1/(1-||\hat D - I||) \xrightarrow{p} 1$.

Further, note that

$$\begin{aligned} ||\Lambda\hat s|| &= O_p(1) && (32)\\ ||\Lambda\tilde R\hat e||^2 &= O_p(1) && (33)\\ ||\Lambda^{-1}\Sigma(I+\Sigma)^{-1}\Lambda^{-1}|| &= O(1) && (34)\\ ||\Lambda^{-1}\hat m|| &= O_p(1) && (35)\\ ||(\hat m'(I+\Sigma)^{-1}\hat m)^{-1}|| &= O_p(1) && (36) \end{aligned}$$

where (32) follows from $||\Lambda\hat s||^2 = \hat s'\Lambda^2\hat s = O_p(1)$ by Lemma 8 (i); with $\hat e = e - T^{-1/2}X\hat h$, (33) follows from $(a+b)'A(a+b) \le a'Aa + b'Ab + 2[(a'Aa)(b'Ab)]^{1/2}$, $E_y e'\tilde R'\Lambda^2\tilde R' e = \operatorname{tr}[\tilde R(E_y ee')\tilde R'\Lambda^2] = O(1)$ and $T^{-1}\hat hX\Lambda^2X\hat h \le ||\hat h||^2T^{-1}||X||^2 = O_p(1)$; (34) follows from $||\Lambda^{-1}(I_N - (I_N+\Sigma)^{-1})\Lambda^{-1}|| \le ||\Lambda^{-1}\Sigma\Lambda^{-1}|| = O(1)$ by Lemma 4 (ii), since for any eigenvalue λ_i of Σ, the corresponding eigenvalue of $I_N - (I_N+\Sigma)^{-1}$ is given by $\lambda_i/(1+\lambda_i) \le \lambda_i$; (35) follows, since with probability converging to one, $\hat{\theta}$ is such that $||\hat{\Psi}_j||$ decays exponentially fast in j; (36) follows from $v'\hat{m}'(I+\Sigma)^{-1}\hat{m}v \geq v'\hat{m}'\hat{m}v/(1+||\Sigma||) \geq ||(\hat{m}'\hat{m})^{-1}||^{-1}/(1+||\Sigma||)$ for any any $pk^2 \times 1$ vector v with $||v|| = 1$, and $||(\hat{m}'\hat{m})^{-1}||^{-1} \xrightarrow{p} ||(m'm)^{-1}||^{-1} > 0$ via $\hat{m}'\hat{m} \xrightarrow{p} m'm$.

Moreover, also

$$\begin{aligned}\Sigma(\hat{D}+\Sigma)^{-1} - \Sigma(I+\Sigma)^{-1} &= (I+\Sigma)^{-1} - \hat{D}(\hat{D}+\Sigma)^{-1} \\ &= (I+\Sigma)^{-1}[\hat{D}+\Sigma-(I+\Sigma)\hat{D}](\hat{D}+\Sigma)^{-1} \\ &= (I+\Sigma)^{-1}\Sigma(I-\hat{D})(\hat{D}+\Sigma)^{-1}\end{aligned}$$

so that

$$\Lambda^{-1}[\Sigma(\hat{D}+\Sigma)^{-1} - \Sigma(I+\Sigma)^{-1}]\Lambda^{-1} = [\Lambda^{-1}(I+\Sigma)^{-1}\Sigma\Lambda^{-1}][I-\hat{D}][\Lambda(\hat{D}+\Sigma)^{-1}\Lambda^{-1}]$$

and

$$\begin{aligned}\Lambda(\hat{D}+\Sigma)^{-1}\Lambda^{-1} &= \Lambda(\hat{D}+\Sigma)^{-1}\hat{D}\Lambda^{-1}\hat{D}^{-1} \\ &= \hat{D}^{-1} - [\Lambda(\hat{D}+\Sigma)^{-1}\Lambda][\Lambda^{-1}\Sigma\Lambda^{-1}]\hat{D}^{-1}\end{aligned}$$

so that by $||(\hat{D}+\Sigma)^{-1}|| \leq ||\hat{D}^{-1}|| \xrightarrow{p} 1$, $||\Lambda|| = O(1)$, $||I-\hat{D}|| \xrightarrow{p} 0$, $||\hat{D}^{-1}|| = O_p(1)$ and by Lemma 4 (ii),

$$\begin{aligned}||\Lambda(\hat{D}+\Sigma)^{-1}\Lambda^{-1}|| &= O_p(1) \\ ||\Lambda^{-1}(\Sigma(\hat{D}+\Sigma)^{-1} - \Sigma(I+\Sigma)^{-1})\Lambda^{-1}|| &= o_p(1).\end{aligned}$$

The result now follows from (22), $||P\hat{P}^{-1} - I_k|| \xrightarrow{p} 0$, $||P^{-1}\hat{P} - I_k|| \xrightarrow{p} 0$ and repeated applications of the norm inequality.

Proof of Lemma 6: By definition and Lemma 5 (i)

$$\begin{aligned}\tilde{f}_T(G,h) &= T^{1/2}J_1'(A(\hat{b})^\ell - A(0)^\ell)Y_T + \sum_{l=1}^{\ell}(J_1[a+P(\tilde{R}e\otimes I)'(\tilde{\Delta}_l\tilde{\gamma}(G))]\otimes J_1)'\operatorname{vec}(A(\hat{b})^{\ell-l}) \\ &\quad + \sum_{l=1}^{\ell}((A(\hat{b})^{l-1}\tilde{Y}_T)'\otimes J_1'A(\hat{b})^{\ell-l}J_1)(b-\hat{b}) \\ &= T^{1/2}J_1'(A(\hat{b})^\ell - A(0)^\ell)Y_T + \sum_{l=1}^{\ell}(J_1[a+P(\tilde{R}e\otimes I)'(\tilde{\Delta}_l\tilde{\gamma}(G))]\otimes J_1)'\operatorname{vec}(A(\hat{b})^{\ell-l}) \\ &\quad + \sum_{l=1}^{\ell} J_1'A(\hat{b})^{\ell-l}J_1(b-\hat{b})A(\hat{b})^{l-1}\tilde{Y}_T \\ &= T^{1/2}J_1'(A(\hat{b})^\ell - A(0)^\ell)Y_T + \sum_{l=1}^{\ell} J_1'A(\hat{b})^{\ell-l}J_1[a+P(\tilde{R}\hat{e}\otimes I)'\tilde{\Delta}_l[\tilde{\gamma}(G)+\hat{m}(b-\hat{b})]].\end{aligned}$$

By Minkowki's inequality, it suffices to consider each summand in $\tilde{f}_T(G,h)$ and $\mathcal{S}_T(G)f_T(G,h) - \mathcal{S}_T(G)\tilde{f}_T(G,h)$ separately.

Note that for any positive integer q, by a direct calculation, (cf. Lütkepohl (2005), page 96)

$$\frac{\partial \operatorname{vec} A(b)^q}{\partial b'} = \sum_{i=0}^{q-1} (A(b)')^{q-1-i} \otimes (A(b)^i J_1),$$

and with $g_A : \mathbb{R} \mapsto \mathbb{R}^{k^2 p^2}$ defined as $g_A(\lambda) = \operatorname{vec}(A(\hat{b} + \lambda(b - \hat{b}))^q)$ for some positive integer q, we have

$$\begin{aligned}
& \operatorname{vec}(A(b)^q) - \operatorname{vec}(A(\hat{b})^q) \\
= & \int_0^1 \frac{g_A(\lambda)}{d\lambda} d\lambda \\
= & T^{-1/2} \int_0^1 \sum_{i=0}^{q-1} (A(\hat{b} + \lambda(b - \hat{b}))')^{q-1-i} \otimes A(\hat{b} + \lambda(b - \hat{b}))^i J_1) d\lambda \cdot (b - \hat{b}). \qquad (37)
\end{aligned}$$

(i) Let $\bar{A}_0 = \sup_{0 \leq \lambda \leq 1} ||A(\lambda \hat{b})|| = O_p(1)$. Then

$$\begin{aligned}
||T^{1/2} J_1' (A(\hat{b})^\ell - A(0)^\ell) Y_T||^2 &= ||(Y_T \otimes J_1)' \int_0^1 \sum_{i=0}^{\ell-1} (A(\lambda \hat{b})')^{\ell-1-i} \otimes A(\lambda \hat{b})^i J_1) d\lambda \cdot \hat{b}||^2 \\
&\leq ||Y_T||^2 ||J_1||^4 ||\hat{b}||^2 \sum_{i=0}^{\ell-1} \bar{A}_0^{2\ell-2} = O_p(1) \\
E_h ||J_1' A(\hat{b})^{\ell-l} J_1 (a - \hat{a})||^2 &\leq \bar{A}_0^{2\ell-2l} ||J_1||^4 E_h ||a - \hat{a}||^4 \\
E_h ||J_1' A(\hat{b})^{\ell-l} J_1 (b - \hat{b}) A(\hat{b})^{l-1} Y_T||^2 &\leq \bar{A}_0^{2\ell-2l} ||J_1||^4 ||Y_T||^2 E_h ||b - \hat{b}||^2 = O_p(1)
\end{aligned}$$

Furthermore, $||J_1' A(\hat{b})^{\ell-l} J_1 P|| \leq ||J_1||^2 ||P|| \bar{A}_0^{\ell-l} = O_p(1)$, $\hat{e} = e - T^{-1/2} X \hat{h}$, and

$$\begin{aligned}
E_G T^{-1} ||(\tilde{R} X \hat{b} \otimes I)' \tilde{\Delta}_l \tilde{\gamma}(G)||^2 &\leq ||T^{-1/2} X||^2 ||\hat{b}||^2 ||\tilde{\Delta}_l||^2 E_G ||\tilde{\gamma}(G)||^2 = O_p(1) \\
E_y E_G ||(\tilde{R} e \otimes I)' \tilde{\Delta}_l \tilde{\gamma}(G)||^2 &\leq ||\tilde{\Delta}_l||^2 E_G ||\tilde{\gamma}(G)||^2 = O(1). \qquad (38) \\
E_h T^{-1} ||(\tilde{R} X \hat{b} \otimes I)' \tilde{\Delta}_l \hat{m} (b - \hat{b})||^2 &\leq ||T^{-1/2} X||^2 ||\hat{b}||^2 E_h ||b - \hat{b}||^2 ||\tilde{\Delta}_l||^2 ||\hat{m}||^2 = O_p(1) \\
E_y E_h ||(\tilde{R} e \otimes I)' \tilde{\Delta}_l \hat{m} (b - \hat{b})||^2 &\leq E_h ||b - \hat{b}||^2 ||\tilde{\Delta}_l||^2 ||\Lambda^{-1} \hat{m}||^2 E_y ||\Lambda(\tilde{R} e \otimes I)|| = O(1).
\end{aligned}$$

(ii) Define the $Nk \times (T + \ell)k$ matrix Δ_l^0 such that $\Delta_l^0 \gamma(G) = \operatorname{vec}(\gamma_l(G), \cdots, \gamma_N(G), 0, \cdots, 0)$. By Minkowski, it suffices to show that $E_h E_G \mathcal{S}_T(G) ||\xi_T(G, h)||^2 \xrightarrow{p} 0$ for $\xi_T(G, h)$ equal to any of

$$\begin{aligned}
& J_1' [A(b)^{\ell-l} - A(\hat{b})^{\ell-l}] J_1 a, \\
& J_1' A(b)^{\ell-l} J_1 P (((RV(G)R)^{-1} - I) Re \otimes I)' (\Delta_l \gamma(G)), \\
& J_1' A(b)^{\ell-l} J_1 P (e \otimes I)' [(\tilde{R} \otimes I)' \Delta_l^0 - (R \otimes I)' \Delta_l] \gamma(G), \\
& J_1' A(b)^{\ell-l} J_1 P (\tilde{R} e \otimes I)' [\Delta_l^0 \gamma(G) - \tilde{\Delta}_l \tilde{\gamma}(G)], \\
& J_1' [A(b)^{\ell-l} - A(\hat{b})^{\ell-l}] J_1 P (\tilde{R} e \otimes I)' (\tilde{\Delta}_l \tilde{\gamma}(G)), \\
& J_1' A(\hat{b})^{\ell-l} J_1 P (\tilde{R}(e - \hat{e}) \otimes I)' (\tilde{\Delta}_l \tilde{\gamma}(G)), \\
& J_1' A(\hat{b})^{\ell-l} J_1 (b - \hat{b}) A(\hat{b})^{l-1} (\tilde{Y}_T - Y_T),
\end{aligned}$$

for $l = 1, \cdots, \ell$, as well as for $\xi_T(G,h)$ equal to

$$T^{1/2}(Y_T \otimes J_1)'(\mathrm{vec}(A(b)^\ell) - \mathrm{vec}(A(\hat{b})^\ell)) - \sum_{l=1}^{\ell}((A(\hat{b})^{l-1}Y_T)' \otimes J_1' A(\hat{b})^{\ell-l}J_1)(b-\hat{b}). \tag{39}$$

Let $\bar{A}_{1,q} = E_h[||A(b)||^q]$, and $\bar{A}_{2,q} = \sup_{\lambda\in[0,1]} E_h[||A(\hat{b}+\lambda(b-\hat{b}))||^q]$, which are $O_p(1)$ for any fixed q, since $||A|| \leq |A|$, and multivariate normal distributions posses all moments. By Cauchy-Schwarz

$$E_h[||J_1'[A(b)^q - A(\hat{b})^q]J_1 a||^2]^2 \leq ||J_1||^4 E_h[||A(b)^q - A(\hat{b})^q||^4]E_h[||a||^4]$$

where clearly $E_h[||a||^4] = O_p(1)$, and using (23), (37), Minkowski and Cauchy-Schwarz, we find

$$\begin{aligned}
E_h||A(b)^q - A(\hat{b})^q||^4 &\leq E_h||\mathrm{vec}(A(b)^q - A(\hat{b})^q)||^4 \qquad (40)\\
&\leq T^{-1}E_h||\int_0^1 \sum_{i=0}^{q-1}(A(\hat{b}+\lambda(b-\hat{b}))')^{q-1-i} \otimes A(\hat{b}+\lambda(b-\hat{b}))^i J_1)d\lambda||^4||b-\hat{b}||^4\\
&\leq T^{-1}||J_1||^4(E_h||b-\hat{b}||^8)^{1/2}\sum_{i=0}^{q-1}(\bar{A}_{2,16q-16-16i}\bar{A}_{2,16i})^{1/4} \xrightarrow{p} 0.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&\mathcal{S}_T(G)E_y(\Delta_l\gamma(G))'[(RV(G)R)^{-1} - I)Ree'R(RV(G)R)^{-1} - I) \otimes I](\Delta_l\gamma(G))\\
=\ &\mathcal{S}_T(G)(\Delta_l\gamma(G))'[(RV(G)R)^{-1} - I)(RV(G)R)^{-1} - I) \otimes I](\Delta_l\gamma(G))\\
=\ &\mathcal{S}_T(G)(\Delta_l\gamma(G))'[R(V(G)^{-1} - I)^2 R \otimes I](\Delta_l\gamma(G))\\
=\ &\mathcal{S}_T(G)T^{-1}(\Delta_l\gamma(G))'[R\Gamma(G)V(G)^{-2}\Gamma(G)'R \otimes I](\Delta_l\gamma(G))\\
\leq\ &\mathcal{S}_T(G)T^{-1}||\Gamma(G)||^2||V(G)^{-2}|| \cdot ||\gamma(G)||^2 \to 0
\end{aligned}$$

uniformly in G by Lemma 2 (iv), (v) and (vii). Also

$$\begin{aligned}
&E_G\mathcal{S}_T(G)E_y\gamma(G)[(\tilde{R}\otimes I)'\Delta_l^0 - (R\otimes I)'\Delta_l](ee'\otimes I)[(\tilde{R}\otimes I)'\Delta_l^0 - (R\otimes I)'\Delta_l]\gamma(G)\\
\leq\ &E_G\mathcal{S}_T(G)\sum_{j=N+1}^{T}|\gamma_j(G)|^2 \to 0
\end{aligned}$$

by Lemma 2 (viii). Also

$$\mathcal{S}_T(G)E_y[\Delta_l^0\gamma(G) - \tilde{\Delta}_l\tilde{\gamma}(G)]'(\tilde{R}ee\tilde{R}'\otimes I)'[\Delta_l^0\gamma(G) - \tilde{\Delta}_l\tilde{\gamma}(G)] \leq \mathcal{S}_T(G)||\gamma(G) - \tilde{\gamma}(G)|| \to 0$$

uniformly in G by Lemma 2 (vi), and

$$\begin{aligned}
&E_hE_G\mathcal{S}_T(G)||J_1'[A(b)^{\ell-l} - A(\hat{b})^{\ell-l}]J_1P(\tilde{R}e\otimes I)'(\tilde{\Delta}_l\tilde{\gamma}(G))||^2\\
\leq\ &||J_1||^4||P||^2(E_G||(\tilde{R}e\otimes I)'\tilde{\Delta}_l\tilde{\gamma}(G)||^2)\cdot(E_h||A(b)^{\ell-l} - A(\hat{b})^{\ell-l}||^2) \xrightarrow{p} 0
\end{aligned}$$

since by (38), $E_G||(\tilde{R}e\otimes I)'\tilde{\Delta}_l\tilde{\gamma}(G)||^2 = O_p(1)$ as above. Furthermore,

$$E_G E_h ||J_1'[A(b)^{\ell-l} - A(\hat{b})^{\ell-l}]J_1 P(\tilde{R}e \otimes I)'(\tilde{\Delta}_l \tilde{\gamma}(G))||^2 \leq$$
$$||J_1||^3 E_h ||A(b)^{\ell-l} - A(\hat{b})^{\ell-l}||^2 E_G ||P(\tilde{R}e \otimes I)'(\tilde{\Delta}_l \tilde{\gamma}(G))||^2 \xrightarrow{p} 0$$

since $E_h||A(b)^q - A(\hat{b})^q||^2 \xrightarrow{p} 0$ from (40) and Jensen, and again $E_G||(\tilde{R}e \otimes I)'\tilde{\Delta}_l\tilde{\gamma}(G)||^2 = O_p(1)$.

Furthermore,

$$E_G||J_1' A(\hat{b})^{\ell-l} J_1 P(\tilde{R}(e-\hat{e}) \otimes I)'(\tilde{\Delta}_l \tilde{\gamma}(G))||^2 \leq ||J_1' A(\hat{b})^{\ell-l} J_1 P||^2 T^{-1} E_G ||(\tilde{R}X\hat{h} \otimes I)'(\tilde{\Delta}_l \tilde{\gamma}(G))||^2$$

and

$$\begin{aligned} E_G||(\tilde{R}X\hat{h} \otimes I)'(\tilde{\Delta}_l \tilde{\gamma}(G))||^2 &\leq \operatorname{tr}(\tilde{R}X\hat{h}\hat{h}'X'\tilde{R}' \otimes I)\tilde{\Delta}_l \Sigma \tilde{\Delta}_l' \\ &\leq ||\hat{h}||^2 \operatorname{tr}(\tilde{R}XX'\tilde{R}' \otimes I)\tilde{\Delta}_l \Sigma \tilde{\Delta}_l' = O_p(1) \end{aligned}$$

since, using Lemma 4 (ii),

$$E_y \operatorname{tr}(\tilde{R}XX'\tilde{R}' \otimes I)\tilde{\Delta}_l \Sigma \tilde{\Delta}_l' \leq ||\Lambda^{-1}\Sigma\tilde{\Delta}_l'\Lambda^{-1}|| \operatorname{tr}\Lambda(\tilde{R}E_y[XX']\tilde{R}' \otimes I)\Lambda = O(1).$$

Moreover

$$E_h||J_1' A(\hat{b})^{\ell-l} J_1 (b-\hat{b}) A(\hat{b})^{l-1}(\tilde{Y}_T - Y_T)||^2 \leq \bar{A}_0^{2\ell-2}||J_1||^4 ||Y_T - \tilde{Y}_T||^2 E_h ||b-\hat{b}||^2 \xrightarrow{p} 0$$

since $||Y_T - \tilde{Y}_T|| \xrightarrow{p} 0$ from $\sum_{t=0}^{\infty} ||\hat{\Psi}_{l+t-1}|| = O_p(1)$, $T^{-1/2}||\hat{a}|| \xrightarrow{p} 0$ and $||A(\hat{b})^N|| \cdot ||Y_{T-N}|| \xrightarrow{p} 0$.

Finally, using (37), the spectral norm of (39) is bounded above by

$$||Y_T|| \cdot ||J_1||^2 ||b-\hat{b}|| \cdot || \int_0^1 \sum_{l=1}^{\ell} [(A(\hat{b}+\lambda_0(b-\hat{b}))')^{\ell-l} \otimes A(\hat{b}+\lambda_0(b-\hat{b}))^{l-1} - (A(\hat{b})')^{\ell-l} \otimes A(\hat{b})^{l-1}] d\lambda_0||.$$

For any fixed integer q and $\lambda_0 \in [0,1]$, from (23) and (37),

$$||A(\hat{b}+\lambda_0(b-\hat{b}))^q - A(\hat{b})^q|| \leq T^{-1/2}|| \int_0^1 \sum_{i=0}^{q-1} (A(\hat{b}+\lambda_0\lambda(b-\hat{b}))')^{q-1-i} \otimes A(\hat{b}+\lambda_0\lambda(b-\hat{b}))^i J_1) d\lambda|| \cdot ||b-\hat{b}||$$

so that from the same argument as employed for (40), $E_h \sup_{\lambda_0 \in [0,1]} ||A(\hat{b}+\lambda_0(b-\hat{b}))^q - A(\hat{b})^q||^K \xrightarrow{p} 0$ for any integer K. But for conformable matrices A, B, C, D, $||(A \otimes B) - (C \otimes D)|| \leq ||A - C|| \cdot ||B|| + ||A|| \cdot ||B - D||$, so that this result, along with Minkowski and Cauchy-Schwarz, establishes that also $E_h||\xi_T(G,h)||^2 \xrightarrow{p} 0$ with $\xi_T(G,h)$ equal to (39).

Proof of Lemma 7: (i) $E_G \exp[K\hat{s}'\tilde{\gamma}(G)] = \exp[K^2\hat{s}'\Sigma\hat{s}] \leq \exp[K^2||\Lambda^{-1}\Sigma\Lambda^{-1}|| \cdot \hat{s}'\Lambda^2\hat{s}]$, so the result follows from Lemmata 4 (ii) and 8 (i).

(ii) By Cauchy-Schwarz

$$E_G \tilde{L}_T(G,h)^4 \leq (E_G \exp[8\hat{s}'\tilde{\gamma}(G)])^{1/2}$$
$$\cdot(E_G \exp[-8\tfrac{1}{2}\tilde{\gamma}(G)'\tilde{\gamma}(G) - 8(h-\hat{h})'\hat{m}'\tilde{\gamma}(G) - \tfrac{1}{2}8(h-\hat{h})'\hat{m}'\hat{m}(h-\hat{h}) + \tfrac{1}{2}8\hat{h}'\hat{m}'\hat{m}\hat{h}])^{1/2}$$

and a direct calculation

$$\begin{aligned}
& E_G \exp[-8(h-\hat{h})'\hat{m}'\tilde{\gamma}(G) - 8\tfrac{1}{2}\tilde{\gamma}(G)'\tilde{\gamma}(G)] \\
= \; & \det(I+8\Sigma)^{-1/2}\exp[\tfrac{1}{2}8(h-\hat{h})'\hat{m}'(I+(8\Sigma)^{-1})^{-1}\hat{m}(h-\hat{h})] \\
\leq \; & \exp[\tfrac{1}{2}8(h-\hat{h})'\hat{m}'(I+(8\Sigma)^{-1})^{-1}\hat{m}(h-\hat{h})].
\end{aligned}$$

Since $I-(I+(8\Sigma)^{-1})^{-1} = (I+8\Sigma)^{-1}$, we obtain

$$\begin{aligned}
& E_G \exp[-8\tfrac{1}{2}\tilde{\gamma}(G)'\tilde{\gamma}(G) - 8(h-\hat{h})'\hat{m}'\tilde{\gamma}(G) - \tfrac{1}{2}8(h-\hat{h})'\hat{m}'\hat{m}(h-\hat{h}) + \tfrac{1}{2}8\hat{h}'\hat{m}'\hat{m}\hat{h}] \\
\leq \; & \exp[-\tfrac{1}{2}8(h-\hat{h})'\hat{m}'(I+8\Sigma)^{-1}\hat{m}(h-\hat{h})] \leq \exp[-\tfrac{1}{2}4C_T||h-\hat{h}||^2]
\end{aligned}$$

where C_T is the smallest eigenvalue of $2\hat{m}'(I+8\Sigma)^{-1}\hat{m}$, and

$$\begin{aligned}
C_T &= 2||(\hat{m}'(I+8\Sigma)^{-1}\hat{m})^{-1}||^{-1} \geq \frac{2||(\hat{m}'\hat{m})^{-1}||^{-1}}{1+8||\Sigma||} \\
C_T &\leq 2||\hat{m}||^2.
\end{aligned}$$

From $\hat{m}'\hat{m} - m'm \xrightarrow{p} 0$ and (26), we have $C_T = O_p(1)$ and $C_T^{-1} = O_p(1)$. Also, by part (i), $E_G \exp[8\hat{s}'\tilde{\gamma}(G)] = O_p(1)$. Thus

$$E_G \tilde{L}_T(G,h)^4 \leq \tilde{\zeta}_T \exp[-\tfrac{1}{2}4C_T||h-\hat{h}||^2]$$

where C_T and $\tilde{\zeta}_T = O_p(1)$.

Proof of Lemma 8: (i) Define s just as $\hat{s}$, but with $\hat{e}_t$ replaced by e_t. Then $E_y s'\Lambda^2 s = \mathrm{tr}[(E_y ss')\Lambda^2] = O(1)$, because the elements of $E_y ss'$ are uniformly bounded. Furthermore, recall that $\hat{e}_t = e_t - T^{-1/2}X_t'\hat{h}$, where $X_t' = (1, y_{t-1}', \cdots, y_{t-p}') \otimes P^{-1}$, so that the jth $k^2 \times 1$ block of $s - \hat{s}$ is given by

$$\begin{aligned}
& T^{-1}\,\mathrm{vec}\textstyle\sum_t [e_t\hat{h}'X_{t+j} + X_t'\hat{h}e_{t+j} + T^{-1/2}X_t'\hat{h}\hat{h}'X_{t+j}] = \\
& \qquad T^{-1}\textstyle\sum_t [(X_{t+j}'\otimes e_t)\hat{h} + (e_{t+j}'\otimes X_t)\hat{h} + T^{-1/2}(X_{t+j}'\otimes X_t')\,\mathrm{vec}(\hat{h}\hat{h}')].
\end{aligned} \tag{41}$$

Thus, letting d_{Xe}, d_{eX} and d_{XX} be $Nk^2 \times (p+1)k^2$, $Nk^2 \times (p+1)k^2$ and $Nk^2 \times (p+1)^2k^4$ matrices with jth k^2 block of rows equal to $T^{-1}\sum(X_{t+j}'\otimes e_t)$, $T^{-1}\sum(e_{t+j}'\otimes X_t)$ and $T^{-3/2}\sum(X_{t+j}'\otimes X_t')$, respectively, we have $\hat{s} = s + d_{Xe}\hat{h} + d_{eX}\hat{h} + d_{XX}\,\mathrm{vec}(\hat{h}\hat{h}')$. It thus suffices to show $s'\Lambda^2 s = \mathrm{tr}\, ss'\Lambda^2 = O_p(1)$, $\hat{h}'d_{Xe}'\Lambda^2 d_{Xe}\hat{h} \leq ||\hat{h}||^2\,\mathrm{tr}\, d_{Xe}d_{Xe}'\Lambda^2 = O_p(1)$, $\hat{h}'d_{eX}'\Lambda^2 d_{eX}\hat{h} \leq ||\hat{h}||\,\mathrm{tr}\, d_{eX}d_{eX}'\Lambda^2 = O_p(1)$ and $\mathrm{vec}(\hat{h}\hat{h}')'d_{XX}'\Lambda^2 d_{XX}\,\mathrm{vec}(\hat{h}\hat{h}') \leq ||\,\mathrm{vec}(\hat{h}\hat{h}')||^2\,\mathrm{tr}\, d_{XX}d_{XX}'\Lambda^2 = O_p(1)$. These follow from Markov's inequality, since the diagonal elements of $E_y ss'$, $E_y d_{Xe}d_{Xe}'$, $E_y d_{eX}d_{eX}'$ and $E_y d_{XX}d_{XX}'$ are readily seen to be uniformly bounded, and $\mathrm{tr}\,\Lambda^2 = O(1)$.

(ii) Note that with $\hat{u}_t$ defined as $\hat{u}_t = 0$ for $t \leq 0$,

$$\hat{m}'\hat{s} = \sum_{j=1}^{N}\begin{pmatrix} \mathrm{vec}(P^{-1\prime}\hat{s}_j'P'\hat{\Psi}_j') \\ \vdots \\ \mathrm{vec}(P^{-1\prime}\hat{s}_j'P'\hat{\Psi}_{j-p+1}') \end{pmatrix} = \sum_{j=1}^{N}\begin{pmatrix} \mathrm{vec}(\Omega^{-1}T^{-1/2}\sum_{t=1}^{T}\hat{u}_t\hat{u}_{t-j}'\hat{\Psi}_{j-1}') \\ \vdots \\ \mathrm{vec}(\Omega^{-1}T^{-1/2}\sum_{t=1}^{T}\hat{u}_t\hat{u}_{t-j}'\hat{\Psi}_{j-p}') \end{pmatrix}$$

$$= \begin{pmatrix} \mathrm{vec}(\Omega^{-1}T^{-1/2}\sum_{t=1}^{T}\hat{u}_t\sum_{j=1}^{N}\hat{u}'_{t-j}\hat{\Psi}'_{j-1}) \\ \vdots \\ \mathrm{vec}(\Omega^{-1}T^{-1/2}\sum_{t=1}^{T}\hat{u}_t\sum_{j=1}^{N}\hat{u}'_{t-j}\hat{\Psi}'_{j-p}) \end{pmatrix}.$$

Furthermore, with $Y_{t-1} = A(\hat{b})^{-1}(Y_t - T^{-1/2}(1, \cdots, 1)' \otimes \hat{a})$ for $t = 0, -1, -2, \ldots$,

$$y_{t-l} = \sum_{j=1}^{N} \hat{\Psi}_{j-l}\hat{u}_{t-j} + T^{-1/2}\sum_{j=0}^{N-l}\hat{\Psi}_j\hat{a} + J_1'A(\hat{b})^{N-l+1}Y_{t-N+l-1}$$

we have

$$T^{-1/2}\sum_{t=1}^{T}\hat{u}_t\sum_{j=1}^{N}\hat{u}'_{t-j}\hat{\Psi}'_{j-l} = T^{-1/2}\sum_{t=1}^{T}\hat{u}_t(y_{t-l} - T^{-1/2}\sum_{j=0}^{N-l}\hat{\Psi}_j\hat{a} - J_1'A(\hat{b})^{N-l+1}Y_{t-N+l-1})' \quad (42)$$

Now $\sum_{t=1}^{T}\hat{u}_t y'_{t-l} = 0$ and $\sum_{t=1}^{T}\hat{u}_t = 0$ by the OLS first order condition, and $T^{-1/2}||A(\hat{b})^{N-l+1}|| \xrightarrow{p} 0$, since with probability converging to one, $||A(\hat{b})|| < 1 - \varepsilon_A/2 < 1$, where $\varepsilon_A = 1 - ||A(0)||$. Thus (42) is $o_p(1)$, and the result follows.

Proof of Lemma 9: Follows from "completing the squares", as $\ln \tilde{L}_T(G, h)$ is quadratic in h and $\tilde{\gamma}(G)$.