SUPPLEMENTAL MATERIAL: THE CEPSTRAL MODEL FOR MULTIVARIATE TIME SERIES: THE VECTOR EXPONENTIAL MODEL

Scott H. Holan, Tucker S. McElroy, and Guohui Wu

University of Missouri, U.S. Census Bureau, and SAS Institute Inc.

S1 Exponential Representation and Proofs

S1.1 Convergence of the Exponential Representation

Let $\Psi(z)$ be a causal power series with $\Psi_0 = \Psi(0) = I$. Let $\mathcal{P}_D$ be the set of all causal matrix power series that converge on $D = \{z \in \mathbb{C} : |z| \leq 1\}$. The matrix logarithm, when it converges, is given by

$$\log\Psi(z) = -\sum_{k\geq 1} k^{-1}(I - \Psi(z))^k = -\sum_{k\geq 1} k^{-1}\left(\sum_{j\geq 1} \Psi_j z^j\right)^k. \tag{S1.1}$$

With $\|\cdot\|$ denoting some matrix norm, a sufficient condition for $\log\Psi(z) \in \mathcal{P}_D$ is

$$\|I - \Psi(z)\| < 1 \quad \forall z \in D. \tag{S1.2}$$

The sufficiency of (S1.2) follows from the fact that $\|\log\Psi(z)\| < \infty$, using (S1.1).

Theorem 1. *If $\log\Psi(z) \in \mathcal{P}_D$, then there exists $\Omega(z) \in \mathcal{P}_D$ such that $\Omega(0) = 0$ and $\exp\{\Omega(z)\} \in \mathcal{P}_D$, and such that $\Psi(z) = \exp\{\Omega(z)\}$ for $z \in D$. If $\Omega(z) \in \mathcal{P}_D$ with $\Omega(0) = 0$, then $\Psi(z)$ defined as $\exp\{\Omega(z)\}$ exists in $\mathcal{P}_D$. Moreover, if $\|\Omega(z)\| < \log 2$ for all $z \in D$, then $\log\Psi(z) = \Omega(z)$.*

Proof of Theorem 1. First assuming $\log\Psi(z) \in \mathcal{P}_D$ and denoting this quantity by $\Omega(z)$, we have $\Omega(0) = \log\Psi(0) = \log I$, which equals the zero matrix by (S1.1). Hence $\Omega(z) = \sum_{k\geq 1} \Omega_k z^k$. Then $\exp\{\Omega(z)\} \in \mathcal{P}_D$ by Artin (1991) (the radius of convergence of the matrix exponential is all of $\mathbb{C}$), and equals $\Psi(z)$, which follows from Lemma 1 below.

Lemma 1. *Let A and B be matrices. If $\log A$ is defined, $\exp\{\log A\} = A$. If $\|\exp B - I\| < 1$, then $\log(\exp B) = B$.*

We only prove the second assertion of Lemma 1, the first being similar.

$$\log(\exp B) = -\sum_{k\geq 1} k^{-1}(I - \exp B)^k = -\sum_{k\geq 1} k^{-1}(-1)^k\left(\sum_{j\geq 1} B^j/j!\right)^k,$$

whenever the series converges. A sufficient condition for convergence, given the discussion preceding (S1.2), is that $\|I - \exp B\| < 1$. Then we see that $\log(\exp B) = \sum_{\ell \geq 1} \psi_\ell B^\ell$ for some scalar coefficients $\{\psi_\ell\}$, which are identical to the coefficients in the scalar expansion of $\log(\exp x)$ for $x \in \mathbb{R}$. Hence $\psi_\ell = 1_{\{\ell=1\}}$, and $\log(\exp B) = B$.

Returning to the theorem's proof, we now assume that $\Omega(z) \in \mathcal{P}_D$ with $\Omega(0) = 0$. Again, $\Psi(z) = \exp\{\Omega(z)\} \in \mathcal{P}_D$. By assumption $\|\Omega(z)\| < \log 2$, and

$$\|\exp \Omega(z) - I\| \leq \sum_{k \geq 1} \|\Omega(z)\|^k / k! = \exp \|\Omega(z)\| - 1 < 1$$

using the definition of the matrix exponential. Applying Lemma 1, we obtain $\log\Psi(z) = \Omega(z)$. □

Corollary 1. *Assume either: (i) $log\,\Psi(z) \in \mathcal{P}_D$ or (ii) $\Omega(z) \in \mathcal{P}_D$ such that $\|\Omega(z)\| < log\,2$ for all $z \in D$. Then $\Psi(B) = \exp \Omega(B)$ and $\Omega(B) = log\,\Psi(B)$ hold and are well-defined, and*

$$\det \Psi(z) = \exp\{tr\,\Omega(z)\} \tag{S1.3}$$

for all $z \in D$. Moreover, $\det \Psi(z) \neq 0$ for all $z \in D$, so that $\Psi(B)$ is invertible with $\Psi(B)^{-1} = \exp\{-\Omega(B)\}$.

Proof of Corollary 1. From Theorem 1, the relations of $\Psi(B)$ and $\Omega(B)$ follow, and (S1.3) follows from Artin (1991, p.286) whenever $\Psi(z) = \exp \Omega(z)$ holds. Invertibility also holds via Corollary 9.10 of Artin (1991). □

S1.2 Proofs of Propositions

Proof of Proposition 1. We note that this proof requires the main result of Proposition 3, but in terms of exposition it makes more sense to state Proposition 1 first. From $\|I - \Psi(z)\| < 1$ we know that $\log\Psi(z) \in \mathcal{P}_D$, and hence we can apply Corollary 1. Then

$$\sum_{k \geq 1} \Psi_k \Psi_k' = \frac{1}{2\pi} \int_{-\pi}^{\pi} \exp\{\Omega(e^{-i\lambda})\} \exp\{\Omega'(e^{i\lambda})\}\, d\lambda,$$

so that by applying the trace $\sum_{k \geq 1} \|\Psi_k\|^2 = (2\pi)^{-1} \int_{-\pi}^{\pi} \|\exp \Omega(e^{-i\lambda})\|^2\, d\lambda$, where $\|\cdot\|$ here denotes the Frobenius norm. Hence, $\{\Psi_k\}$ is square-summable with respect to the Frobenius norm, which indicates that the space of sequences described by condition (S1.2) is a subset of ℓ_2 with respect to Frobenius. This also shows that the concept of a mean square Cauchy sequence is well-defined for such a process. To prove the time series is mean square Cauchy, we take differences for $q = m + h$ and $q = m$, where m and h are large integers:

$$X_t^{(m+h)} - X_t^{(m)} = \left[\Psi^{(m+h)}(B) - \Psi^{(m)}(B)\right] \epsilon_t.$$

Here the time index t is immaterial, since we will compute the covariance matrix of the above vector difference. The covariance equals

$$\sum_{k \geq 0} \left(\Psi_k^{(m+h)} - \Psi_k^{(m)}\right) \underline{\Sigma} \left(\Psi_k^{(m+h)} - \Psi_k^{(m)}\right)'.$$

In order to show that this matrix tends to zero as m and h grow to infinity, it suffices to examine the sequence $\Psi_k^{(m+h)} - \Psi_k^{(m)}$, which by Proposition 3 can be written as follows. Let $\Upsilon(z) = \Omega(z)/z$, where $\Omega(z)$ corresponds to $\Psi^{(m)}(z)$; but the cepstral representation for the moving average series $\Psi^{(m+h)}(z)$ equals $\Omega(z)$ plus a second term $\Xi(z) = \sum_{j=1}^{h} \Omega_{j+m} z^{j+m}$. Thus $\{\Omega(z) + \Xi(z)\}/z = \Upsilon(z) + \Pi(z)$, say. With these notations, we have

$$\Psi_k^{(m+h)} - \Psi_k^{(m)} = \frac{1}{k!} \sum_{\ell=1}^{k} \binom{k}{\ell} \left(\left[\{\Upsilon(z) + \Pi(z)\}^\ell \right]^{(k-\ell)} - \left[\Upsilon(z)^\ell \right]^{(k-\ell)} \right) |_{z=0}.$$

To evaluate this expression further, we must expand the term $\Upsilon(z) + \Pi(z)$. Let us denote these matrices, for any fixed z, by A_0 and A_1 respectively. Then the ℓ-th power of $A_0 + A_1$ can be written as

$$\sum_{i_1, i_2, \ldots, i_\ell \in \{0,1\}^\ell} \prod_{j=1}^{\ell} A_{i_j}. \tag{S1.4}$$

Here $\{0,1\}^\ell$ denotes the space of binary strings of length ℓ, and we sum over all such strings. Note that we next subtract off $\Upsilon(z)^\ell$, which equals the single summand of (S1.4) that corresponds to the zero string, i.e., $i_j = 0$ for all $1 \le j \le \ell$. Using the linearity of differentiation, we have

$$\Psi_k^{(m+h)} - \Psi_k^{(m)} = \frac{1}{k!} \sum_{\ell=1}^{k} \binom{k}{\ell} \left(\sum_{i_1, i_2, \ldots, i_\ell \in \{0,1\}^\ell \setminus \{0\}^\ell} \prod_{j=1}^{\ell} A_{i_j} \right)^{(k-\ell)} |_{z=0}.$$

Note that $\Pi(z)$ occurs in at least one summand of every term of $\Psi_k^{(m+h)} - \Psi_k^{(m)}$. In taking the derivatives and evaluating zero, term by term we either produce zero or an expression involving a coefficient matrix of $\Pi(z)$. The smallest coefficient matrix is Ω_{m+1}, so taking matrix norms it will be sufficient to show that $\|\Omega_m\| \to 0$ as $m \to \infty$. Since the matrix logarithm of $\Psi(z)$ is well-defined, we find that

$$\sum_{k \ge 1} \Omega_k \Omega_k' = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log\Psi(e^{-i\lambda}) \log\Psi'(e^{i\lambda})\, d\lambda.$$

Taking the trace and again using the Frobenius norm $\|\cdot\|$, we obtain $\sum_{k \ge 1} \|\Omega_k\|^2 = (2\pi)^{-1} \int_{-\pi}^{\pi} \|\log\Psi(e^{-i\lambda})\|^2 \, d\lambda$; this is finite for all λ because $\|I - \Psi(z)\| < 1$. Hence $\|\Omega_k\|$ is square summable, and in particular the sequence tends to zero. This establishes that the sequence is Cauchy in mean square, and hence $X_t^{(q)} \to X_t^{(\infty)}$ in mean square, and $X_t = X_t^{(\infty)}$. □

Proof of Proposition 2. Because $q < \infty$, the convergence of $[\Omega(z)]_1^q$ for $z \in D$ is assured, and we apply Corollary 1, obtaining $\Psi(z) \in \mathcal{P}_D$. So the process is stable. It is also invertible, with inverse $\exp\{-[\Omega(z)]_1^q\}$. (Note that the stronger condition that $\|\Omega(z)\| < \log 2$ is not needed to establish invertibility.) Next, we establish identifiability.

Let θ denote a parameter vector describing all the various entries of the cepstral matrices, in some order, and $f(\lambda; \theta)$ the associate spectral density. Let $\theta^{(1)}$ and $\theta^{(2)}$ denote two values of the parameter vector, but with $f(\cdot; \theta^{(1)}) = f(\cdot; \theta^{(2)})$. Writing the spectral density in the form (2), we can invert the moving average filters (because $\det\Psi(z) \ne 0$ holds for $z \in D$) to obtain

$$\exp\{\Omega_0^{(1)}\} = \exp\{-\Omega^{(1)}(z)\} \exp\{\Omega^{(2)}(z)\} \exp\{\Omega_0^{(2)}\} \exp\{[\Omega^{(2)}(z)]^*\} \exp\{-[\Omega^{(1)}(z)]^*\}.$$

The composition of the two causal power series $\exp\{-\Omega^{(1)}(z)\}$ and $\exp\{\Omega^{(2)}(z)\}$ is another causal power series with leading coefficient of I; because the spectral density has full rank (because $\mathrm{tr}(\Omega_0) > -\infty$), the spectral factorization is unique (see Hannan and Deistler (1988)). It follows that we must have $\exp\{\Omega^{(1)}\} = \exp\{\Omega^{(2)}\}$ and $\exp\{-\Omega^{(1)}(z)\}\exp\{\Omega^{(2)}(z)\} = I$ for all z. Hence the moving average filters are identically the same; applying the matrix logarithm reveals that $\Omega^{(1)}(z) = \Omega^{(2)}(z)$ for all $z \in D$. Now we use the uniqueness of the Fourier basis to learn that each of the coefficient matrices are the same, and hence $\theta^{(1)} = \theta^{(2)}$. This establishes that $\theta \mapsto f(\cdot;\theta)$ is injective. □

Proof of Proposition 3. Under condition (S1.2), $\log\Psi(B)$ is well-defined. To prove (2.4) we begin with the matrix exponential expansion, which is valid because the series is invertible:

$$\Psi(z) = \exp\{z\,\Upsilon(z)\} = 1 + \sum_{j\geq 1} \frac{z^j}{j!}\Upsilon(z)^j.$$

Differentiating k times with respect to the complex variable z yields

$$\Psi^{(k)}(z) = \sum_{j\geq 1}\sum_{\ell=1}^{k} \binom{k}{\ell} \frac{\partial^\ell}{\partial z^\ell}\frac{z^j}{j!} \cdot \frac{\partial^{k-\ell}}{\partial z^{k-\ell}}\Upsilon(z)^j,$$

where we can use the Abelian product rule because the scalar quantities commute with the matrix powers $\Upsilon(z)^j$. Interchanging the summations over j and ℓ, we see that if we evaluate at $z = 0$ – which is equivalent to coefficient matching – we have $k!\Psi_k$ on the left hand side, but the right hand side will be zero unless $j = \ell$. This produces the first formula of (2.4), and the second follows from algebra. The proof of (2.5) is similar, but using the expansion for the logarithm instead of the exponential. □

S2 Applications of the VEXP Model

In terms of modeling with a VEXP(q), in the frequentist context, we can proceed with a higher-order model – confident by Proposition 1 that we can get an arbitrarily accurate approximation to causal invertible processes – and then refine the model by replacing "small" parameter values with zeroes. In order to construct parameter estimates, one proceeds by computing the acf for any posited parameter values and evaluating the Gaussian or Whittle likelihood as desired (cf. Brockwell and Davis (1991) and Taniguchi and Kakizawa (2000)). Then numerical optima can be determined using BFGS (a quasi-Newton method) or other methods as desired. Appendix S4 has additional material on the asymptotic properties of parameter estimates.

Alternatively, using the exact Gaussian likelihood, we can proceed with estimation using a Bayesian approach. In this setting, the cepstral model order can be chosen using Bayes factor or by minimizing some previously selected criterion, such as deviance information criterion (DIC) (Spiegelhalter, Best, Carlin, and Van Der Linde (2002)) or out-of-sample mean squared prediction error. Within a given model order, estimation of cepstral matrix entries can proceed using stochastic search variable selection (SSVS) (George and McCulloch (1993, 1997)).

Having fitted a time series model, one may be interested in a variety of applications: forecasting, signal extraction, transfer function modeling, or spectral estimation/plotting, etc. The versatility of the VEXP model readily allows us these applications. Plotting the moving average filter $\Psi(z)$ for $z = \exp(-i\lambda)$ as a function of frequency allows us to visualize the transfer function of the process operating on white noise inputs. Evaluating (2.2) allows plotting of the fitted spectrum, which becomes arbitrarily accurate as q is increased.

For forecasting, it is necessary to know either the autocovariance structure or the moving average coefficients. McElroy and McCracken (2014) describes multi-step forecasting for non-stationary vector time series with a general moving average form, including integrated VARMA models as special cases. From that work, the forecast filter (from an infinite past) for h-step ahead forecasting of a stationary process with invertible moving average power series $\Psi(z)$ is

$$\Pi(z) = z^{-h}[\Psi]_h^\infty(z)\,\Psi^{-1}(z).$$

Noting that the VAR(1), VMA(1), and VEXP(1) all involve the same number of unknown coefficients, it is of interest to compare their h-step ahead forecast functions. As in the previous subsection, let the VAR(1) be written $\Psi(z) = (I - \Phi z)^{-1}$, whereas the VMA(1) is $\Psi(z) = I + \Psi_1 z$. Of course the VEXP(1) is $\Psi(z) = \exp(\Omega_1 z)$, which can be expanded into the moving average form with $\Psi_k = \Omega_1^k/k!$. Moreover, $z^{-h}[\Psi]_h^\infty(z) = \sum_{k\geq 0} \Omega_1^{k+h}/(k+h)!$, from which $\Pi(z)$ can be computed by a convolution (note that the matrices involved are just powers of Ω_1, and hence are Abelian). The forecast filters for the VAR(1), VMA(1), and VEXP(1) are then respectively given by

$$\begin{aligned}
\Pi(z) &= \Phi^h \\
\Pi(z) &= 1_{\{h=1\}} \sum_{k\geq 0} (-1)^k \Psi_1^{k+1} z^k \\
\Pi(z) &= \sum_{k\geq 0} \sum_{\ell=0}^{k} \frac{1}{\ell!(k+h-\ell)!}\, \Omega_1^{k+h} z^k.
\end{aligned}$$

Note that the VAR(1) forecast only relies on present data; the VMA(1) uses past data when $h = 1$, but otherwise offers the pathetic prediction of zero when $h > 1$. The VEXP(1) uses a geometrically decaying pattern of weights of past data, like the VMA(1). As h increases, all the forecast filters tend to the zero matrix, essentially dictating that long-run forecasts are given by the mean for a stationary process.

Generalizing to VAR(q), VMA(q), and VEXP(q), it is difficult to provide explicit formulas for $\Pi(z)$ (except in the VAR case), but we know that the VAR(q) filter utilizes the past q values of the series, whereas the VMA(q) uses all the data so long as $h \leq q$; when $h > q$ the filter is zero. For the VEXP(q), a weighted average of all past data is implied. The repercussions are that VAR forecasts tend to be based upon recent activity, even when h is large; VMA and VEXP forecasts can reach deeper into the past, which may be desirable when h is quite large. A VARMA forecast filter will have behavior more like that of a VEXP, but the VEXP can be estimated without concerns regarding identifiability.

S3 Simulated Examples

To illustrate the utility of the VEXP model we present a simulation study and two distinct simulated examples. The simulation we present uses the exact Gaussian likelihood and consists of 200 simulated datasets. The parameters are identical to those found in Simulated Example I (Section S3.1). As seen from Tables 1 and 2, the parameter estimates appear to be precise with the MSE decreasing as the sample size increases. In contrast, the first simulated example highlights estimation through SSVS, whereas the second simulated example considers estimation without SSVS. The two simulated examples presented here are designed to demonstrate various aspects associated with the analyses presented in Section 4.

Table 1: Simulation results for Appendix S3 using the exact Gaussian likelihood and parameter specification identical to Simulated Example I for $T = 100$ and $T = 150$.

	$T = 100$			$T = 150$			
	mean	sd	mse	mean	sd	mse	True
$\Omega_0(1,1)$	1.18811	0.16473	0.04066	1.23576	0.12450	0.02022	1.30500
$\Omega_0(2,1)$	-2.55051	0.24031	0.06658	-2.52714	0.15575	0.02934	-2.45500
$\Omega_0(2,2)$	0.03336	0.08533	0.00726	0.02508	0.09233	0.00851	0.03000
$\Omega_1(1,1)$	0.32113	0.10999	0.01204	0.31511	0.10348	0.01068	0.32000
$\Omega_1(1,2)$	-1.15171	0.70488	0.49471	-1.15778	0.57063	0.32414	-1.17000
$\Omega_1(2,1)$	0.00878	0.08384	0.00707	-0.00456	0.07691	0.00591	0.00000
$\Omega_1(2,2)$	0.23114	0.12710	0.01643	0.25343	0.10172	0.01031	0.25000
$\Omega_2(1,1)$	0.09975	0.13673	0.01901	0.11769	0.11135	0.01234	0.12000
$\Omega_2(1,2)$	1.49438	0.69472	0.48033	1.55049	0.46567	0.21783	1.50500
$\Omega_2(2,1)$	0.00702	0.08055	0.00650	-0.00617	0.07675	0.00590	0.00000
$\Omega_2(2,2)$	0.19337	0.11450	0.01332	0.20310	0.09350	0.00875	0.21000
$\Omega_3(1,1)$	0.13677	0.14469	0.02083	0.12342	0.11414	0.01310	0.13500
$\Omega_3(1,2)$	-0.11464	0.72906	0.52889	0.12342	0.11414	0.01310	-0.11000
$\Omega_3(2,1)$	0.00838	0.07984	0.00641	-0.00472	0.07675	0.00588	0.00000
$\Omega_3(2,2)$	0.02083	0.12324	0.01570	0.03866	0.09591	0.00919	0.04500
$\Omega_4(1,1)$	0.11893	0.13369	0.01791	0.12217	0.10724	0.01150	0.13000
$\Omega_4(1,2)$	-2.57316	0.74659	0.55478	-2.64134	0.60769	0.37406	-2.56000
$\Omega_4(2,1)$	0.01124	0.09455	0.00902	-0.00511	0.07680	0.00589	0.00000
$\Omega_4(2,2)$	-0.01461	0.12206	0.01504	-0.01449	0.09941	0.01004	0.00000
μ_1	0.03537	0.38324	0.14739	0.00960	0.33458	0.11147	0.00000
μ_2	-0.00488	0.05083	0.00259	-0.00058	0.04219	0.00177	0.00000

Table 2: Simulation results for Appendix S3 using the exact Gaussian likelihood and parameter specification identical to Simulated Example I for $T = 200$ and $T = 250$.

	$T = 200$			$T = 250$			
	mean	sd	mse	mean	sd	mse	True
$\Omega_0(1,1)$	1.24610	0.11664	0.01701	1.23931	0.12357	0.01951	1.30500
$\Omega_0(2,1)$	-2.49927	0.21655	0.04862	-2.47532	0.22081	0.04893	-2.45500
$\Omega_0(2,2)$	0.02592	0.05605	0.00314	0.03476	0.05712	0.00327	0.03000
$\Omega_1(1,1)$	0.32507	0.07191	0.00517	0.32286	0.06320	0.00398	0.32000
$\Omega_1(1,2)$	-1.23435	0.47634	0.22990	-1.15178	0.43278	0.18670	-1.17000
$\Omega_1(2,1)$	-0.00133	0.04190	0.00175	-0.00370	0.04717	0.00223	0.00000
$\Omega_1(2,2)$	0.25518	0.06863	0.00471	0.24909	0.06890	0.00472	0.25000
$\Omega_2(1,1)$	0.11750	0.09078	0.00821	0.11304	0.07443	0.00556	0.12000
$\Omega_2(1,2)$	1.45744	0.49902	0.25003	1.52588	0.45581	0.20716	1.50500
$\Omega_2(2,1)$	-0.00064	0.04914	0.00240	0.00070	0.04758	0.00225	0.00000
$\Omega_2(2,2)$	0.20385	0.07563	0.00573	0.20655	0.06638	0.00440	0.21000
$\Omega_3(1,1)$	0.13520	0.08617	0.00739	0.13792	0.07191	0.00515	0.13500
$\Omega_3(1,2)$	-0.09788	0.48411	0.23334	-0.11043	0.43414	0.18753	-0.11000
$\Omega_3(2,1)$	-0.00382	0.05515	0.00304	-0.00012	0.04742	0.00224	0.00000
$\Omega_3(2,2)$	0.04136	0.07395	0.00545	0.03111	0.06052	0.00384	0.04500
$\Omega_4(1,1)$	0.12468	0.08306	0.00689	0.13045	0.07660	0.00584	0.13000
$\Omega_4(1,2)$	-2.57213	0.52039	0.26960	-2.56157	0.48179	0.23097	-2.56000
$\Omega_4(2,1)$	0.00369	0.05056	0.00256	0.00066	0.04560	0.00207	0.00000
$\Omega_4(2,2)$	-0.00651	0.07729	0.00599	-0.00747	0.07170	0.00517	0.00000
μ_1	-0.01936	0.28118	0.07904	-0.00423	0.23463	0.05479	0.00000
μ_2	0.00224	0.03112	0.00097	0.00043	0.03094	0.00095	0.00000

S3.1 Simulated Example I

The goal of this example is to illustrate that, given an underlying dependence structure, the modeling approach using SSVS is able to provide shrinkage toward the simulated dependence structure with high probability. This is especially useful in the context of multi-step ahead forecasting, as presented in Section 4.1, where our approach averages over several candidate models with the expectation of improved long-term forecasts.

For illustration, we simulate data based on estimates from a VEXP(4) model, with $T = 192$, based on the forecasting example presented in Section 4.1. In particular, the elements of the cepstral matrices are based on estimated values obtained from a VEXP(4) model applied to the bivariate retail sales forecasting example. Recalling that $V_j = \text{vec}(\Omega_j)$, the exact model used for data generation is given by

$$\begin{aligned} V_0 &= (1.305, 0.030, 0.030, -2.455)', \\ V_1 &= (0.320, -1.170, 0.000, 0.250)', \\ V_2 &= (0.120, 1.505, 0.000, 0.210)', \\ V_3 &= (0.135, -0.110, 0.000, 0.045)', \\ V_4 &= (0.130, -2.560, 0.000, 0.000)', \end{aligned}$$

where the mean of the two time series is set equal to zero (i.e., $\delta = (0, 0)'$). In terms of prior distributions we assume that $\delta \sim N(\overline{x}, \text{diag}(\sigma^2_{\mu_1}, \sigma^2_{\mu_2}))$, $\text{diag}(\Omega_0) \sim N(0, \text{diag}(\sigma^2_1, \sigma^2_2))$, and $\sigma^2_{\mu_1}, \sigma^2_{\mu_2}, \sigma^2_1, \sigma^2_2 \sim \text{IG}(A, B)$, where the elements of $\overline{x}$ constitute the estimated sample means for the bivariate time series. In addition, we choose $A = 2.1$ and $B = 1.1$; i.e., we assume an inverse-gamma distribution with mean and variance both being 1. The prior specification for μ follows from the fact that, for independent and identically distributed (iid) data, $\overline{x}$ is the maximum likelihood estimate (as well as the asymptotic mean). Finally, based on a sensitivity analysis (see Section 4.1), the hyperparameters for the SSVS were specified as $\tau_i \equiv \tau = .10$ and $c_i \equiv c = 10$. The MCMC sampling algorithm was run for 60,000 iterations with the first 40,000 discarded for burn-in. Convergence was assessed through visual inspection of the sample chains with no evidence of lack of convergence detected.

Table 3 displays the frequency that a particular cepstral matrix specification appeared in the model throughout the 20,000 post burn-in MCMC iterations. This table clearly illustrates that the SSVS prior is selecting the data generating model specification with high probability. Additionally, in cases where competing cepstral matrix specifications are chosen, typically the additional elements selected have parameters estimated relatively close to zero.

In contrast, Table 4 presents posterior summaries of the estimated mean and cepstral matrix elements. Importantly, in all cases, the 95% credible intervals (CIs) capture the true values, with most intervals relatively narrow. Although the SSVS is implicitly averaging over several model specifications, the fact that the 95% CIs capture the true values reinforces the fact that the SSVS is able to recover the correct dependence structure with high probability.

Table 3: SSVS results for the VEXP(4) model from Simulated Example I. Note that only cepstral matrices appearing in the model more than 200 times are detailed in the table and the column labeled SSVS corresponds to an indicator function specifying the elements of $\text{vec}(\Omega_j)$ $(j = 1, \ldots, 4)$ appearing in the model. Finally, note that the bolded entry represents the model structure used to generate the data.

Parameters	SSVS	Freq (out of 20,000)
$\Omega_0(1,2)$	**1**	11,387
	0	8,613
Ω_1	**1101**	15,358
	1111	4,528
Ω_2	**1101**	11,405
	1100	4,500
	1111	2,040
	1110	960
	0101	685
	0100	234
Ω_3	**1101**	5,005
	1100	4,756
	0101	2,597
	0100	2,668
	1000	739
	1111	733
	1110	725
	0000	503
	0001	471
	0111	444
	0110	397
Ω_4	1101	7,733
	1100	4,476
	0101	3,641
	0100	1,648
	1111	1,107
	1110	627
	0111	533
	0110	235

Table 4: Posterior summary of the VEXP(4) parameters using SSVS for Simulated Example I. Here, "mean," "sd," and "Q"denote the posterior mean, posterior standard deviation, and quantile of the posterior distribution, respectively.

Parameters	mean	sd	$Q_{.025}$	$Q_{.5}$	$Q_{.975}$	True
$\Omega_0(1,1)$	1.24532	0.11567	1.02450	1.24352	1.47568	1.30500
$\Omega_0(2,1)$	-2.47678	0.11247	-2.69252	-2.47981	-2.25332	-2.45500
$\Omega_0(2,2)$	0.03862	0.04419	-0.01551	0.02115	0.13369	0.03000
$\Omega_1(1,1)$	0.29931	0.07165	0.15771	0.29951	0.43823	0.32000
$\Omega_1(2,1)$	-1.28023	0.47853	-2.23111	-1.28870	-0.33292	-1.17000
$\Omega_1(1,2)$	0.01052	0.00875	-0.00508	0.00994	0.02985	0.00000
$\Omega_1(2,2)$	0.24884	0.06896	0.11437	0.24841	0.38286	0.25000
$\Omega_2(1,1)$	0.16605	0.08028	-0.00144	0.16904	0.31906	0.12000
$\Omega_2(2,1)$	1.75680	0.45747	0.88372	1.75858	2.66037	1.50500
$\Omega_2(1,2)$	0.00590	0.00813	-0.00980	0.00581	0.02224	0.00000
$\Omega_2(2,2)$	0.08299	0.08090	-0.02128	0.07803	0.24665	0.21000
$\Omega_3(1,1)$	0.06619	0.07615	-0.02600	0.04505	0.22631	0.13500
$\Omega_3(2,1)$	-0.38706	0.48649	-1.43690	-0.34742	0.46667	-0.11000
$\Omega_3(1,2)$	0.00384	0.00791	-0.01112	0.00348	0.01947	0.00000
$\Omega_3(2,2)$	-0.02885	0.06065	-0.17981	-0.00751	0.07065	0.04500
$\Omega_4(1,1)$	0.08438	0.08466	-0.01970	0.07374	0.25861	0.13000
$\Omega_4(2,1)$	-2.79556	0.48298	-3.78661	-2.79562	-1.88095	-2.56000
$\Omega_4(1,2)$	-0.00115	0.00773	-0.01658	-0.00134	0.01391	0.00000
$\Omega_4(2,2)$	-0.06388	0.07357	-0.22113	-0.04491	0.03247	0.00000
μ_1	-0.10955	0.24152	-0.58511	-0.11213	0.36615	0.00000
μ_2	0.02241	0.02709	-0.03138	0.02249	0.07616	0.00000
σ_1^2	1.16845	1.42409	0.28159	0.82241	4.10003	NA
σ_2^2	2.64336	6.08424	0.62769	1.81727	8.99748	NA
$\sigma_{\mu_1}^2$	0.70184	0.80633	0.16888	0.49273	2.54956	NA
$\sigma_{\mu_2}^2$	0.69542	0.90139	0.16607	0.48470	2.44190	NA

S3.2 Simulated Example II

The second simulated example considers bivariate spectral estimation and, in particular, estimation of squared coherence, where squared coherence is defined as

$$\rho^2_{X_1 \cdot X_2}(\lambda) = \frac{|f_{X_1 X_2}(\lambda)|^2}{f_{X_1 X_1}(\lambda) f_{X_2 X_2}(\lambda)}. \tag{S3.1}$$

This simulation is designed to behave similar to the bivariate critical radio frequency – sunspots example presented in Section 4.2 and uses a VEXP(4), with $T = 240$, for illustration. Additionally, this example does not use SSVS; instead, it demonstrates the VEXP framework in situations where model averaging is not necessarily desired.

The VEXP(4) model used to generate data for this example was based on estimates obtained from the critical radio frequency - sunspots data discussed in Section 4.2. Specifically, the model is given by

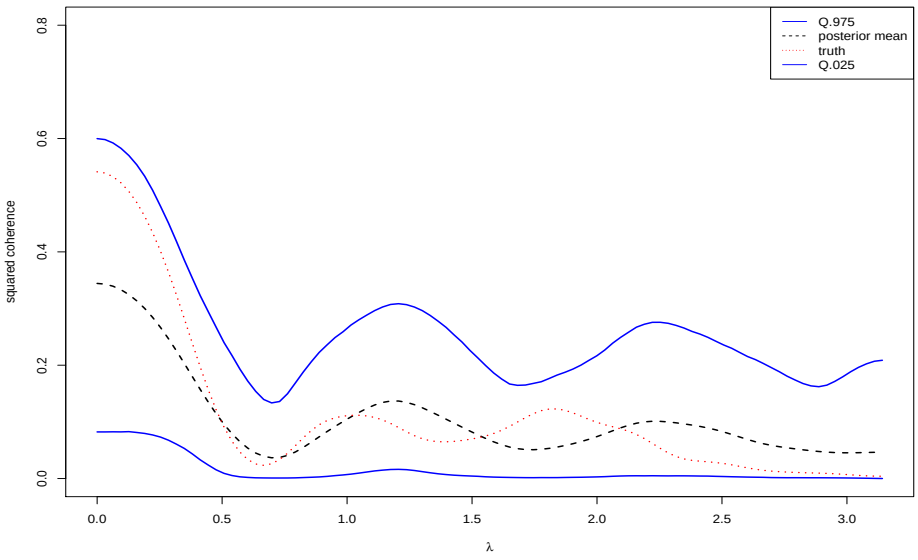
$$\begin{aligned} V_0 &= (-0.249, 0.211, 0.211, -0.023)', \\ V_1 &= (1.343, 0.081, 0.073, 0.803)', \\ V_2 &= (0.261, 0.169, -0.109, 0.432)', \\ V_3 &= (-0.108, 0.160, 0.138, 0.234)', \\ V_4 &= (0.127, 0.080, 0.114, 0.244)', \end{aligned}$$

where the mean of the bivariate time series is set equal to zero (i.e., $\delta = (0, 0)'$). In terms of prior distributions we assume that $\delta \sim N(\overline{x}, \text{diag}(\sigma^2_{\mu_1}, \sigma^2_{\mu_2}))$, $\text{diag}(\Omega_0) \sim N(0, \text{diag}(\sigma^2_1, \sigma^2_2))$, $\sigma^2_{\mu_1}, \sigma^2_{\mu_2}, \sigma^2_1, \sigma^2_2 \sim \text{IG}(A, B)$, where $\overline{x}$ is the estimated sample mean for the bivariate time series. Again, we choose $A = 2.1$ and $B = 1.1$; i.e., we assume an inverse-gamma distribution with mean and variance are both one. For the off-diagonal element of Ω_0, $\Omega_0(1, 2)$, and all elements in Ω_j for $j = 1, 2, 3, 4$, we assumed a $N(0, 10^2)$ prior distribution.

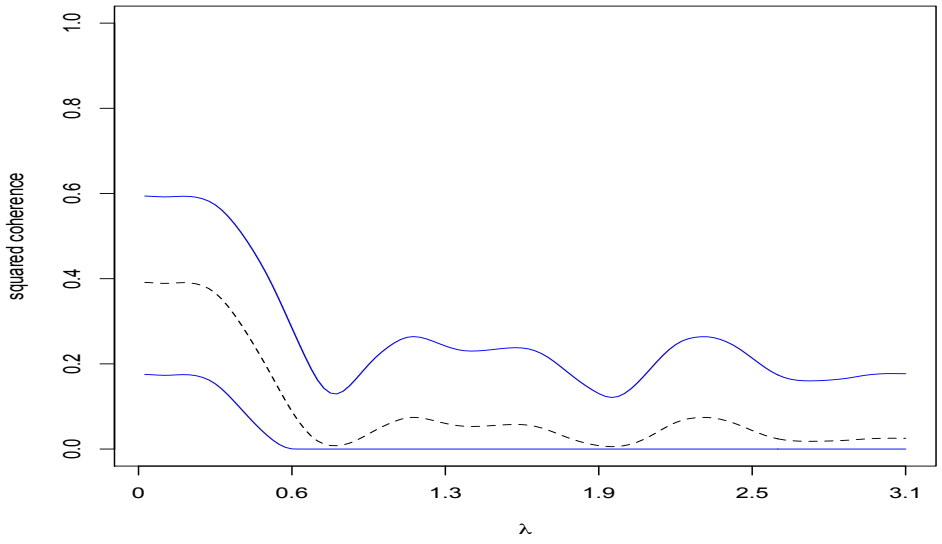
As shown in Table 5, this example clearly demonstrates the ability for our Bayesian estimation procedure to produce reliable results. In particular, all of the 95% CIs capture the true values and, in most cases, the intervals are relatively narrow. Additionally, as depicted in Figure 1a, the posterior mean squared coherence (obtained as the pointwise mean from the posterior distribution of squared coherence functions) and true squared coherence, as defined by (S3.1), are in close agreement, with the pointwise 95% CIs relatively narrow away from frequency zero and capturing the true squared coherence. It is important to note that the deviations between the true and estimated squared coherence in this example are due to the fact that the estimated squared coherence is based on one stochastic realization of the truth, with the Bayesian VEXP(4) estimate agreeing with an empirical estimate obtained through smoothing the multivariate discrete Fourier transform using a modified Daniell window in R (using kernel("modified.daniell", c(8,8,8)) with taper=.2 in the function spec.pgram).

Table 5: Posterior summary of the VEXP(4) parameters without SSVS for Simulated Example II. Here, "mean," "sd," and "Q"denote the posterior mean, posterior standard deviation, and quantile of the posterior distribution, respectively.

Parameters	mean	sd	$Q_{.025}$	$Q_{.5}$	$Q_{.975}$	True
$\Omega_0(1,1)$	-0.14897	0.09878	-0.33487	-0.15222	0.05479	-0.24900
$\Omega_0(2,1)$	0.11234	0.06381	-0.01192	0.11252	0.23769	0.21100
$\Omega_0(2,2)$	0.10710	0.10031	-0.07976	0.10539	0.31661	-0.02300
$\Omega_1(1,1)$	1.31882	0.06530	1.19197	1.31802	1.44882	1.34300
$\Omega_1(2,1)$	0.07137	0.05760	-0.04718	0.07241	0.18201	0.08100
$\Omega_1(1,2)$	0.04971	0.07239	-0.08775	0.04805	0.19345	0.07300
$\Omega_1(2,2)$	0.78013	0.06338	0.65332	0.78029	0.90351	0.80300
$\Omega_2(1,1)$	0.22706	0.06532	0.10014	0.22675	0.35491	0.26100
$\Omega_2(2,1)$	0.14242	0.05574	0.03353	0.14276	0.25154	0.16900
$\Omega_2(1,2)$	-0.05140	0.07554	-0.20083	-0.05122	0.09722	-0.10900
$\Omega_2(2,2)$	0.43926	0.06320	0.31438	0.43998	0.56269	0.43200
$\Omega_3(1,1)$	-0.10610	0.06468	-0.23236	-0.10613	0.02079	-0.10800
$\Omega_3(2,1)$	0.08094	0.06024	-0.03730	0.08142	0.20182	0.16000
$\Omega_3(1,2)$	0.21579	0.07655	0.06273	0.21588	0.36484	0.13800
$\Omega_3(2,2)$	0.24103	0.06452	0.11220	0.24061	0.36782	0.23400
$\Omega_4(1,1)$	0.15975	0.06788	0.02469	0.15914	0.29192	0.12700
$\Omega_4(2,1)$	-0.00110	0.05933	-0.11442	-0.00269	0.11792	0.08000
$\Omega_4(1,2)$	0.03664	0.07427	-0.10839	0.03684	0.18223	0.11400
$\Omega_4(2,2)$	0.20952	0.06442	0.08318	0.20893	0.33763	0.24400
μ_1	0.48925	0.29485	-0.09672	0.48887	1.06666	0.00000
μ_2	0.14156	0.33476	-0.52073	0.14177	0.79180	0.00000
σ_1^2	0.69913	0.81959	0.16847	0.48853	2.50540	NA
σ_2^2	0.69215	0.83041	0.16946	0.48956	2.41251	NA
$\sigma_{\mu_1}^2$	0.71532	0.81970	0.17109	0.50502	2.53710	NA
$\sigma_{\mu_2}^2$	0.72942	0.81193	0.17450	0.51109	2.64821	NA



(a)



(b)

Figure 1: (a) Comparison of true, pointwise posterior mean, and pointwise posterior 95% credible intervals of squared coherence for the VEXP(4) model presented in Simulated Example II. Note that the red dotted line and black dashed line denote the truth and posterior mean, respectively. (b) Empirical squared coherence and pointwise 95% confidence intervals using modified Daniell window.

S4 Results on Parameter Estimates

Given the discussion of MLE and WLE (Whittle Likelihood Estimates) in Section 3.1, we here summarize the asymptotic theory for parameter estimates. We also develop the theory of estimation for Quasi-Maximum Likelihood Estimates (QMLEs), and provide results applicable to large classes of short memory vector time series. Most texts on multivariate time series – Hannan and Deistler (1988), Lütkepohl (2007), Tsay (2013), Shumway and Stoffer (2010), Brockwell and Davis (1991), Taniguchi and Kakizawa (2000) – either establish asymptotic efficiency (i.e., asymptotic normality with asymptotic covariance equal to the inverse Fisher information matrix) of QMLEs, or establish asymptotic efficiency of MLEs in the univariate case. Although many papers claim that the asymptotic efficiency of MLEs in the multivariate case has been established, in fact the above books do not contain such a result, and neither do published papers on this topic. Moreover, there is much confusion over this issue; we attempt to provide a rigorous resolution for the Gaussian case.

The quasi-likelihood for a linear process of known marginal distribution is an approximation to the exact likelihood that is more convenient for computation. Moreover, in the case of a Gaussian marginal distribution the exact likelihood is computable, being given by the innovations algorithm, and the approximation of the exact Gaussian likelihood by the quasi-likelihood can be studied more precisely. We provide an analysis of the difference of these objective functions, such that the difference between Gaussian MLE and QMLE is asymptotically negligible; then asymptotic efficiency of the Gaussian MLE is inherited from that of the QMLE. We develop these results for a fairly broad class of linear short memory Gaussian processes, which include VEXP and VARMA processes.

As a preliminary step, we summarize the asymptotic properties of WLEs, and then discuss the quasi-likelihood. Among the above cited references, Taniguchi and Kakizawa (2000) – henceforth TK – provides the most rigorous and thorough treatment of vector linear processes. TK's result on the asymptotic efficiency of QMLEs presumes the marginal density is parameter-free, which amounts to assuming that the innovation covariance matrix is known ahead of time. Below, we extend the TK results to the more general case of an unknown innovation covariance matrix that is separately parametrized. Both this extension of QMLE asymptotic efficiency, as well as Gaussian MLE asymptotic efficiency, are apparently (and surprisingly, given the importance and centrality of this statistical issue) novel results.

S4.1 Whittle Estimation

The theory for Whittle estimation is thoroughly covered in TK, and we merely summarize those results with application to the VEXP process. The Whittle theory allows for model misspecification, unlike the case of QMLE and MLE. The WLE is defined as the minimizer (when it exists and is unique) of either criterion (3.1) or (3.2). The Kullback-Leibler (KL) discrepancy between a model spectral density f_ω (e.g., given by a VEXP(q), or by some other multivariate

model) and the true process' spectrum $\widetilde{f}$ is given by

$$\mathrm{KL}(f_\omega, \widetilde{f}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{tr}\{f_\omega^{-1}(\lambda)\, \widetilde{f}(\lambda)\}\, d\lambda + \log\det \underline{\Sigma}, \tag{S4.1}$$

where $\underline{\Sigma}$ is the innovation variance matrix of the model (determined by spectral factorization). In the case of a VEXP(q) model, we can write

$$\log\det \underline{\Sigma} = \mathrm{tr}\Omega_0.$$

Substituting the periodogram I_T for $\widetilde{f}$ in equation (S4.1) yields the Whittle objective function $\mathcal{W}$ of equation (3.2). The minimizer of KL with respect to ω is denoted $\omega(\widetilde{f})$, and is called the Pseudo-True Value (PTV). In the case that the model is correctly specified, $\widetilde{f}$ is a VEXP(q) process with some parameter vector $\widetilde{\omega}$, and $\omega(\widetilde{f}) = \widetilde{\omega}$. Otherwise, the model is mis-specified and the WLE will be consistent for the PTV.

We explicate asymptotic efficiency results for the WLE, when the true marginal structure is non-Gaussian. Following TK, we will assume the cumulant conditions of Brillinger (1981). Set

$$c^X_{a_1,\ldots,a_k}(t_1,\ldots,t_{k-1}) = \mathrm{cum}\left(X_0^{(a_1)}, X_{t_1}^{(a_2)}, \ldots, X_{t_{k-1}}^{(a_k)}\right)$$

for $a_1,\ldots,a_k \in \{1,\ldots,m\}$ and $X_t^{(a)}$ denoting the ath component of X_t. Then consider the following conditions described in TK:

(**B**): for each $j = 1, 2, \ldots, k-1$ and any k-tuple $a_1, a_2, \ldots, a_k$ we have

$$\sum_{t_1,t_2,\ldots,t_{k-1}\in\mathbb{Z}} (1 + |t_j|)\, |c^X_{a_1,\ldots,a_k}(t_1,\ldots,t_{k-1})| < \infty$$

for $k \geq 2$.

(**A1**): f_ω is twice continuously differentiable, and ∇f_ω is continuous

(**A2**): the PTV exists uniquely in $\mathbb{R}^{\binom{m+1}{2}+m^2q}$

(**A3**): the Hessian $H(\omega) = \nabla\nabla' \mathrm{KL}(f_\omega, \widetilde{f})$ is invertible at the PTV

These assumptions allow for a broad class of short memory non-Gaussian vector processes. As can be seen from Proposition 3, the moving average representation for a VEXP(q) is continuously differentiable with respect to ω, and hence (**A1**) holds. Conditions (**A2**) and (**A3**) amount to assuming that a unique minimizer exists, and similar assumptions are made in the analysis of VARMA models. Finally, let

$$V(\omega) = \frac{1}{\pi} \int_{-\pi}^{\pi} \mathrm{tr}\{\nabla f_\omega^{-1}(\lambda)\, \widetilde{f}(\lambda)\, \nabla' f_\omega^{-1}(\lambda)\, \widetilde{f}(\lambda)\}\, d\lambda,$$

which generalizes H, in the sense that it equals $2H(\omega)$ when $\widetilde{f} = f_\omega$. Recall that the WLE is denoted $\widehat{\omega}_{WLE}$. The following theorem establishes consistency and asymptotic efficiency of WLEs for the non-Gaussian vector processes, and hence can be applied in the case of VEXP modeling.

Theorem 2. *Suppose that $\{X_t\}$ is a strictly stationary process satisfying (**B**), and that (**A1**), (**A2**), and (**A3**) hold. Then the WLE is consistent for the PTV and*

$$\sqrt{T}\left(\widehat{\omega}_{WLE} - \omega(\widetilde{f})\right) \stackrel{\mathcal{L}}{\Longrightarrow} \mathcal{N}\left(0, H(\omega(\widetilde{f}))^{-1} V(\omega(\widetilde{f})) H(\omega(\widetilde{f}))^{-1}\right)$$

if the fourth order cumulants are zero.

Proof of Theorem 2. This is Theorem 3.1.2 of TK, which follows from Lemma 3.1.1 of TK utilizing condition (**B**). □

Remark 1. A more complicated variance expression is available when the fourth order cumulants are nonzero. In typical applications, one supposes that the model is correctly specified, as well as the conditions of Theorem 2. Then the asymptotic covariance is $4V(\omega(\widetilde{f}))^{-1}$, i.e., the inverse Fisher information matrix. This would be consistently estimated by $4V(\widehat{\omega}_{WLE})^{-1}$. Below we explicitly derive $V(\omega)$ for a separable linear process, including the VEXP and VARMA cases.

Nested model hypotheses can be tested by considering the log ratio of Whittle likelihoods. For a full VEXP(q) model, we can restrict any of the $\binom{m+1}{2} + m^2 q$ real parameters to fixed values, such as zero. The nested model is obtained by fitting the Whittle likelihood with the restrictions in play. The Generalized Likelihood Ratio (GLR) test statistic is the difference between the restricted Whittle and the unrestricted likelihood, and the asymptotic theory then follows from the discussion on pp. 59 – 61 of TK.

Corollary 2. *Suppose that $\{X_t\}$ is a strictly stationary process satisfying (**B**) with fourth order cumulants equal to zero, and that (**A1**), (**A2**), and (**A3**) hold. If the restricted VEXP(q) model is correctly specified and we impose r restrictions in fitting the restricted model, then*

$$GLR = T\left(\mathcal{W}(\widehat{\omega}_{WLE^*}; X) - \mathcal{W}(\widehat{\omega}_{WLE}; X)\right) \stackrel{\mathcal{L}}{\Longrightarrow} \chi^2_r,$$

where $\widehat{\omega}_{WLE^*}$ is the Whittle estimate from the restricted model.

We next proceed to derive $V(\omega)$ in the special case of a linear process, where $\widetilde{f} = f_\omega$. Suppose we have a causal linear process

$$X_t = \sum_{j \geq 0} \Psi_j \epsilon_{t-j} \tag{S4.2}$$

with $\{\epsilon_t\}$ an iid sequence of covariance matrix $\underline{\Sigma}$, and Ψ_0 equal to the identity by assumption. The causal moving average filter $\Psi(B)$ depends on the parameter vector ω in a smooth way in the case of a VEXP process, or a VARMA process. Also, these processes are examples of separable processes, which means that the innovation covariance $\underline{\Sigma}$ is separately parametrized from $\Psi(B)$. Supposing that the moving average filter is invertible, we write $\Pi(B) = \Psi^{-1}(B)$, which we denote as the inverse moving average filter. It follows that Π_0 is the identity matrix. Also, as discussed in Section 2.1, the spectral density is $f_\omega(\lambda) = \Psi(z)\underline{\Sigma}\Psi'(\overline{z})$ and the inverse spectrum is

$$f_\omega^{-1}(\lambda) = \Pi'(\overline{z})\underline{\Sigma}^{-1}\Pi(z).$$

Suppose ω_j corresponds to a parameter of $\Psi(B)$ (and not of $\underline{\Sigma}$); then

$$\frac{\partial}{\partial\omega_j} f_\omega^{-1}(\lambda) = \frac{\partial}{\partial\omega_j}\Pi'(\overline{z})\cdot\underline{\Sigma}^{-1}\Pi(z) + \Pi'(\overline{z})\underline{\Sigma}^{-1}\frac{\partial}{\partial\omega_j}\Pi(z).$$

It follows that

$$\frac{\partial}{\partial\omega_j} f_\omega^{-1}(\lambda)\cdot f_\omega(\lambda) = \frac{\partial}{\partial\omega_j}\Pi'(\overline{z})\cdot\Psi'(\overline{z}) + \Pi'(\overline{z})\underline{\Sigma}^{-1}\frac{\partial}{\partial\omega_j}\Pi(z)\cdot\Psi(z)\underline{\Sigma}\Psi'(\overline{z}).$$

On the other hand, if ω_k corresponds to a parameter of $\underline{\Sigma}$, then

$$\frac{\partial}{\partial\omega_k} f_\omega^{-1}(\lambda) = -\Pi'(\overline{z})\underline{\Sigma}^{-1}\cdot\left[\frac{\partial}{\partial\omega_k}\underline{\Sigma}\right]\cdot\underline{\Sigma}^{-1}\Pi(z)$$
$$\frac{\partial}{\partial\omega_k} f_\omega^{-1}(\lambda)\cdot f_\omega(\lambda) = -\Pi'(\overline{z})\underline{\Sigma}^{-1}\cdot\left[\frac{\partial}{\partial\omega_k}\underline{\Sigma}\right]\cdot\Psi'(\overline{z}).$$

Then $V(\omega)$ has a block structure. The upper left block corresponds to parameters of the process: suppose j,k pertain to this block. Then

$$\begin{aligned}
\operatorname{tr}\left(\frac{\partial}{\partial\omega_j} f_\omega^{-1}(\lambda)\cdot f_\omega(\lambda)\frac{\partial}{\partial\omega_k} f_\omega^{-1}(\lambda)\cdot f_\omega(\lambda)\right) &= \operatorname{tr}\left(\frac{\partial}{\partial\omega_j}\Pi'(\overline{z})\cdot\Psi'(\overline{z})\frac{\partial}{\partial\omega_k}\Pi'(\overline{z})\cdot\Psi'(\overline{z})\right)\\
&+ \operatorname{tr}\left(\frac{\partial}{\partial\omega_j}\Pi'(\overline{z})\cdot\underline{\Sigma}^{-1}\cdot\frac{\partial}{\partial\omega_k}\Pi(z)\cdot f_\omega(\lambda)\right)\\
&+ \operatorname{tr}\left(\frac{\partial}{\partial\omega_k}\Pi'(\overline{z})\cdot\underline{\Sigma}^{-1}\cdot\frac{\partial}{\partial\omega_j}\Pi(z)\cdot f_\omega(\lambda)\right)\\
&+ \operatorname{tr}\left(\frac{\partial}{\partial\omega_j}\Pi(z)\cdot\Psi(z)\frac{\partial}{\partial\omega_k}\Pi(z)\cdot\Psi(z)\right)
\end{aligned}$$

after simplification. Note that the derivatives of $\Pi(z)$ only depend on positive powers of z, because Π_0 does not depend on ω; thus the first summand above only depends on positive powers of $\overline{z}$, and the fourth summand depends on positive powers of z. Therefore these both have integral zero, and drop out of the expression for $V_{jk}(\omega)$. Introducing the shorthand $\langle\cdot\rangle$ for average integration over $[-\pi,\pi]$, we obtain

$$\begin{aligned}
V_{jk}(\omega) &= 2\,\langle\operatorname{tr}\left(\frac{\partial}{\partial\omega_j} f_\omega^{-1}\cdot f_\omega\frac{\partial}{\partial\omega_k} f_\omega^{-1}\cdot f_\omega\right)\rangle\\
&= 2\,\operatorname{tr}\underline{\Sigma}^{-1}\langle\frac{\partial}{\partial\omega_k}\Pi(z)\cdot f_\omega(\lambda)\cdot\frac{\partial}{\partial\omega_j}\Pi'(\overline{z})\rangle\\
&+ 2\,\operatorname{tr}\underline{\Sigma}^{-1}\langle\frac{\partial}{\partial\omega_j}\Pi(z)\cdot f_\omega(\lambda)\cdot\frac{\partial}{\partial\omega_k}\Pi'(\overline{z})\rangle\\
&= 2\sum_{\ell_1,\ell_2\geq 0}\operatorname{tr}\underline{\Sigma}^{-1}\left(\frac{\partial}{\partial\omega_k}\Pi_{\ell_1}\,\Gamma_{\ell_2-\ell_1}\,\frac{\partial}{\partial\omega_j}\Pi'_{\ell_2} + \frac{\partial}{\partial\omega_j}\Pi_{\ell_1}\,\Gamma_{\ell_2-\ell_1}\,\frac{\partial}{\partial\omega_k}\Pi'_{\ell_2}\right).
\end{aligned}$$

It can be shown that the trace of both summands is equal. Next, the lower right block of $V(\omega)$ corresponds to parameters of the innovation covariance matrix: supposing j,k pertain to this block, we have

$$\operatorname{tr}\left(\frac{\partial}{\partial\omega_j} f_\omega^{-1}(\lambda)\cdot f_\omega(\lambda)\frac{\partial}{\partial\omega_k} f_\omega^{-1}(\lambda)\cdot f_\omega(\lambda)\right) = \operatorname{tr}\left(\underline{\Sigma}^{-1}\,\frac{\partial}{\partial\omega_j}\underline{\Sigma}\cdot\underline{\Sigma}^{-1}\,\frac{\partial}{\partial\omega_k}\underline{\Sigma}\right).$$

Because this term is constant with respect to λ, $V_{jk}(\omega)$ is equal to twice this quantity. Finally, the lower left and upper right blocks of $V(\omega)$ are zero; suppose that ω_j corresponds to the

process and ω_k to the innovation variance matrix. Then

$$\begin{aligned}
&\mathrm{tr}\left(\frac{\partial}{\partial\omega_j}f_\omega^{-1}(\lambda)\cdot f_\omega(\lambda)\frac{\partial}{\partial\omega_k}f_\omega^{-1}(\lambda)\cdot f_\omega(\lambda)\right)\\
&=\mathrm{tr}\left(\frac{\partial}{\partial\omega_j}\Pi'(\overline{z})\cdot\Psi'(\overline{z})+\Pi'(\overline{z})\underline{\Sigma}^{-1}\frac{\partial}{\partial\omega_j}\Pi(z)\cdot\Psi(z)\underline{\Sigma}\Psi'(\overline{z})\right)\cdot\left(-\Pi'(\overline{z})\underline{\Sigma}^{-1}\cdot\left[\frac{\partial}{\partial\omega_k}\underline{\Sigma}\right]\cdot\Psi'(\overline{z})\right)\\
&=-\mathrm{tr}\left(\frac{\partial}{\partial\omega_j}\Pi'(\overline{z})\cdot\underline{\Sigma}^{-1}\cdot\left[\frac{\partial}{\partial\omega_k}\underline{\Sigma}\right]\cdot\Psi'(\overline{z})+\underline{\Sigma}^{-1}\frac{\partial}{\partial\omega_j}\Pi(z)\cdot\Psi(z)\left[\frac{\partial}{\partial\omega_k}\underline{\Sigma}\right]\right),
\end{aligned}$$

and the first summand involves only positive powers of $\overline{z}$, whereas the second summand involves only positive powers of z. Therefore this term integrates to zero. The Fisher information is half of $H(\omega)$, or one quarter of $V(\omega)$:

$$\frac{1}{4}V(\omega)=\begin{bmatrix}\mathrm{tr}\underline{\Sigma}^{-1}\langle\nabla\Pi(z)\cdot f_\omega(\lambda)\cdot\nabla'\Pi'(\overline{z})\rangle & 0\\ 0 & \frac{1}{2}\,\mathrm{tr}\left(\underline{\Sigma}^{-1}\,\nabla\underline{\Sigma}\cdot\underline{\Sigma}^{-1}\,\nabla'\underline{\Sigma}\right)\end{bmatrix}.$$

This agrees with the expression in the proof of Theorem 3.1.12 of TK (which considers the case of constant innovation variance).

S4.2 Quasi-Maximum Likelihood Estimation

Whereas Whittle estimation does not presume a linear structure for the data process, the Quasi-Likelihood does, i.e., (S4.2) is assumed to hold. Although in this paper we are principally interested in Gaussian processes, we begin our treatment somewhat more generally, following TK closely. Hence, we denote by p_ω the multivariate pdf of the innovation process $\{\epsilon_t\}$, allowing the pdf to depend on the parameter vector. We also assume that our linear process is invertible, so that

$$X_t=\Psi(B)\epsilon_t\quad\text{and}\quad\epsilon_t=\Pi(B)X_t.$$

Given a model, it produces an approximation to the true $\Psi(B)$ and $\Pi(B)$, which we denote via $\Psi_\omega(B)$ and $\Pi_\omega(B)$ – we presume a correctly specified model (unlike the Whittle case), so we only are concerned with obtaining the correct value of ω. The true ω is denoted $\widetilde{\omega}$, and hence $\Psi_{\widetilde{\omega}}(B)=\Psi(B)$, and so forth.

For any guess ω, we can compute $\Pi_\omega(B)$ to any desired truncation level, and proceed to compute estimated innovations, or residuals. (For a VARMA model, one needs only invert the VMA polynomial, whereas for the VEXP we generate the moving average form corresponding to cepstral matrices multiplied by negative one.) However, $\Pi_\omega(B)X_t$ involves an infinite past of the data, and our sample begins at time point $t=1$. Therefore we have a truncation of our "ideal" residuals $e_t=\Pi_\omega(B)X_t$ given by

$$\widehat{e}_t=\sum_{j=0}^{t-1}\Pi_j(\omega)X_{t-j}.$$

This is tantamount to setting $X_s=0$ for all $s\le 0$ in the ideal residual calculation. Clearly, this is an unwarranted approximation, but if the $\Pi_j(\omega)$ decay rapidly in matrix norm as j increases, the impact will hopefully be insubstantial. The ideal, or exact, likelihood is written in log scale

as

$$\mathcal{L} = -2 \sum_{t=1}^{T} \log p_\omega(e_t),$$

which we seek to minimize with respect to ω. This is typically not computable, though in the case of a Gaussian pdf one may replace the missing X_s for $s \leq 0$ in the ideal residuals by their backcasts from available data and obtain the exact likelihood. This differs somewhat from replacing the past data by zeroes, which results in the Quasi-Likelihood. In log scale, this is written

$$\mathcal{Q} = -2 \sum_{t=1}^{T} \log p_\omega(\widehat{e}_t).$$

The minimizer of $\mathcal{Q}$ is the QMLE, denoted $\widehat{\omega}_{QMLE}$, when it exists uniquely. Asymptotic efficiency of the QMLE is addressed by many authors; the treatment of TK, which summarizes much of the current literature, considers the case that p does not depend on ω (Theorem 3.1.12 of TK). However, the Local Asymptotic Normality (LAN) theory upon which their results rely are established for the case of p_ω – see Theorem 2.2.4 of TK. Therefore, it seems possible to extend Theorem 3.1.12 to the case that p depends on ω. We state this result as a theorem, which is a straightforward extension of Theorem 3.1.12 of TK. We begin by summarizing the sufficient conditions from TK. We define $|\cdot|$ to be the matrix norm defined by the sum of absolute values of all entries, and say a matrix power series $\{A_j\}$ has "geometric decay" if there exists $0 < \rho < 1$ such that $|A_j| = O(\rho^j)$ for $j \geq 1$. Also, let $\|\cdot\|$ denote the Euclidean norm.

(**C1**): (i) Uniformly in ω, $\{\Psi_j(\omega)\}$ has geometric decay.
(ii) Each matrix entry of $\Psi_j(\cdot)$ is twice continuously differentiable, and the mixed partial derivatives of order two of $\{\Psi_j(\omega)\}$ have geometric decay, uniformly in ω.
(iii) The mixed partial derivatives of order two of Ψ_j is Lipschitz in ω.
(iv) $\Psi(B)$ is invertible with inverse $\Pi(B)$, which has geometric decay.
(v) Each matrix entry of $\Pi_j(\cdot)$ is twice continuously differentiable, and the mixed partial derivatives of order two of $\{\Pi_j(\omega)\}$ have geometric decay, uniformly in ω.
(vi) The mixed partial derivatives of order two of Π_j is Lipschitz in ω.

(**C2**): (i) The distribution corresponding to p_ω has finite second moments, with mean zero and variance $\underline{\Sigma}_\omega$, a p.d. matrix. Also $\lim_{\|e\|\to\infty} p_\omega(e) = 0$.
(ii) The first and second derivatives with respect to ω and e exist, and are Lipschitz functions.
(iii) Letting $\phi_\omega(e) = \nabla_e \log p_\omega(e)$ and $\eta_\omega(e) = \nabla_\omega \log p_\omega(e)$, we assume

$$\int |\phi_\omega(e)|^4 \, p_\omega(e)\, de < \infty \qquad \int |\eta_\omega(e)|^4 \, p_\omega(e)\, de < \infty$$

uniformly in ω. Moreover,

$$\int \nabla_e \nabla_e' p_\omega(e)\, de = 0 \qquad \text{and} \qquad \int \nabla_\omega \nabla_\omega' p_\omega(e)\, de = 0 \tag{S4.3}$$

for all ω.

Assumption (**C1**) includes short memory vector time series, such as VARMA and VEXP – Section 2.2 shows that the moving average representation of a VEXP(q) process (with $q < \infty$) has geometric decay, and trivially the same is true of its inverse moving average representation. The conditions on p_ω are mild extensions of those used to prove Theorem 3.1.12 of TK, and are clearly satisfied by a Gaussian pdf. The first part of the Hessian mean condition (S4.3) is equivalent to assuming that

$$-\mathbb{E}[\nabla_e \nabla_e' \log p_\omega(e)] = \mathbb{E}[\nabla_e \log p_\omega(e) \cdot \nabla_e' \log p_\omega(e)],$$

where the expectation is with respect to p_ω. We denote this quantity by $\mathcal{F}(p_\omega)$; it is a sort of Jacobian term in the Fisher information. In the Gaussian case, this quantity is equal to $\underline{\Sigma}_\omega^{-1}$. Similarly, under the second part of (S4.3) we have the quantity $\mathcal{G}(p_\omega)$, whose jkth entry is the matrix

$$-\mathbb{E}[\nabla_{\omega_j} \nabla_{\omega_k}' \log p_\omega(e)] = \mathbb{E}[\nabla_{\omega_j} \log p_\omega(e) \cdot \nabla_{\omega_k}' \log p_\omega(e)],$$

which is the Fisher information of the marginal distribution. The following result establishes consistency and asymptotic efficiency for the non-Gaussian Quasi-Likelihood, assuming that the model (and marginal pdf) is correctly specified. In particular, the QMLEs for a VEXP are asymptotically normal.

Theorem 3. *Suppose that $\{X_t\}$ is a linear process (S4.2) with marginal pdf p_ω satisfying (**C1**) and (**C2**), and that the model is correctly specified. Suppose that $\Psi_\omega(B)$ and p_ω are separately parametrized. Then the QMLE is consistent and*

$$\sqrt{T}\,(\widehat{\omega}_{QMLE} - \widetilde{\omega}) \stackrel{\mathcal{L}}{\Longrightarrow} \mathcal{N}\left(0, F(\widetilde{\omega})^{-1}\right),$$

with the Fisher information matrix $F(\omega)$ given by

$$F_{jk}(\omega) = tr\{\mathcal{F}(p_\omega)\,\langle \nabla_j \Pi(z) \cdot f_\omega(\lambda) \cdot \nabla_k' \Pi'(\overline{z})\rangle\} + tr\,\mathcal{G}_{jk}(p_\omega).$$

Proof of Theorem 3. The proof mimics that of Theorem 3.1.12 of TK, but with extensions for a pdf depending on ω. Note that our Quasi-Likelihood is -2 times the objective function used in TK. Recall that ϕ and η are the gradients of $\log p_\omega(e)$ with respect to e and ω. Let the full Hessian of $\log p_\omega(e)$ be denoted M, with upper left block $M^{(ee)}$ corresponding to the Hessian with respect to e only, and the lower right block $M^{(\omega\omega)}$ the Hessian with respect to ω only. Using the chain rule for the derivative of $\log p_\omega(\widehat{e}_t(\omega))$, the scaled gradient of $\mathcal{Q}$ is

$$\begin{aligned} T^{-1/2}\,\nabla_{\omega_j}\mathcal{Q} &= -2T^{-1/2}\sum_{t=1}^{T}\left(\frac{\partial}{\partial\omega_j}\widehat{e}_t'\,\phi(\widehat{e}_t) + \eta_j(\widehat{e}_t)\right) \\ &= -2T^{-1/2}\sum_{t=1}^{T}\left(\frac{\partial}{\partial\omega_j}\widehat{e}_t'\,\phi(e_t) + \eta_j(e_t)\right) + O_P(T^{-1/2}), \end{aligned}$$

using (**C1**) and (**C2**) – the error $e_t - \widehat{e}_t = O_P(\rho^t)$ for some $0 < \rho < 1$. The scaled Hessian is

$$\begin{aligned} T^{-1}\,\nabla_{\omega_j}\nabla'_{\omega_k}\mathcal{Q} &= -2T^{-1}\sum_{t=1}^{T}\Big(\frac{\partial^2}{\partial\omega_j\partial\omega_k}\widehat{e}'_t\,\phi(\widehat{e}_t) + \frac{\partial}{\partial\omega_j}\widehat{e}'_t M^{(ee)}(\widehat{e}_t)\frac{\partial}{\partial\omega_k}\widehat{e}_t \\ &\quad + \frac{\partial}{\partial\omega_j}\widehat{e}'_t M^{(e\omega)}_{\cdot,k}(\widehat{e}_t) + \frac{\partial}{\partial\omega_k}\widehat{e}'_t M^{(e\omega)}_{\cdot,j}(\widehat{e}_t) + M^{(\omega\omega)}(\widehat{e}_t)\Big) \\ &= 2T^{-1}\sum_{t=1}^{T}\left(\frac{\partial}{\partial\omega_j}\widehat{e}'_t\phi(e_t)\phi'(e_t)\frac{\partial}{\partial\omega_k}\widehat{e}_t + \eta_j(e_t)\eta_k(e_t)\right) + O_P(T^{-1/2}). \end{aligned}$$

This approximation is achieved through several steps, utilizing the theorem's assumptions. As with the gradient term, we can replace $\widehat{e}_t$ by e_t inside of ϕ and the entries of M, at the cost of stochastic error terms. Furthermore, we have for any $t \geq 1$

$$\mathbb{E}\left[\frac{\partial^2}{\partial\omega_j\partial\omega_k}\widehat{e}'_t\,\phi(e_t)\right] = \mathbb{E}\left[\sum_{\ell=0}^{t-1} X'_{t-\ell}\frac{\partial^2}{\partial\omega_j\partial\omega_k}\Pi'_\ell\,\phi(e_t)\right] = 0,$$

because the coefficient of X_t in the sum is zero (because Π_0 is constant with respect to ω), and all other X_s with $s < t$ are independent of e_t and have mean zero. It follows that the average of such stochastic terms is $O_P(T^{-1/2})$; this handles the first term in the Hessian of $\mathcal{Q}$. For the second term, we have

$$M^{(ee)} = \nabla_e\nabla'_e p_\omega(e)\cdot p_\omega^{-1}(e) - \nabla_e p_\omega(e)\cdot\nabla'_e p_\omega(e)\cdot p_\omega^{-2}(e).$$

The second term is the negative of $\phi\phi'$, but the first term we denote by the matrix $J_\omega(e)$; its expectation with respect to p_ω is zero by (S4.3). Therefore for any $t \geq 1$,

$$\mathbb{E}\left[\frac{\partial}{\partial\omega_j}\widehat{e}'_t\,J(e_t)\frac{\partial}{\partial\omega_k}\widehat{e}_t\right] = \mathbb{E}\left[\sum_{\ell_1,\ell_2=0}^{t-1} X'_{t-\ell_1}\frac{\partial}{\partial\omega_j}\Pi'_{\ell_1}\,J(e_t)\,\frac{\partial}{\partial\omega_k}\Pi_{\ell_2}X_{t-\ell_2}\right],$$

which is again zero due to independence of $J(e_t)$ from X_s with $s < t$. As a result, the second term in the Hessian contributes $\frac{\partial}{\partial\omega_j}\widehat{e}'_t\phi(e_t)\phi'(e_t)\frac{\partial}{\partial\omega_k}\widehat{e}_t$ to the sum. The third and fourth terms in the Hessian of $\mathcal{Q}$ are handled similarly to the first term, and hence are negligible. The fifth term is analyzed along the lines of the second term, with

$$M^{(\omega\omega)} = \nabla_\omega\nabla'_\omega p_\omega(e)\cdot p_\omega^{-1}(e) - \nabla_\omega p_\omega(e)\cdot\nabla'_\omega p_\omega(e)\cdot p_\omega^{-2}(e);$$

the second term is $-\eta\eta'$. At this point, we have proved the asymptotic expressions for the gradient and Hessian of $\mathcal{Q}$, and we imitate the remaining argument of Theorem 3.1.12 of TK. In particular, we obtain the convergence

$$T^{-1}\,\nabla_{\omega_j}\nabla'_{\omega_k}\mathcal{Q} \xrightarrow{P} 2\,\mathrm{tr}\,\{\mathcal{F}(p_\omega)\sum_{\ell_1,\ell_2\geq 1}\frac{\partial}{\partial\omega_j}\Pi_{\ell_1}\Gamma_{\ell_2-\ell_1}\frac{\partial}{\partial\omega_k}\Pi'_{\ell_2}\} + 2\,\mathrm{tr}\,\mathcal{G}_{jk}(p_\omega). \tag{S4.4}$$

The quantity $\widehat{\omega}_{QMLE} - \widetilde{\omega}$ is equal, up to asymptotically negligible terms, to the negative of the inverse Hessian times the gradient; hence the factor of two cancels out. It follows that the asymptotic precision matrix for $\widehat{\omega}_{QMLE} - \widetilde{\omega}$ is one half the above limit, i.e., it is $F_{jk}(\widetilde{\omega})$, the Fisher information matrix. □

Remark 2. A Gaussian VEXP process satisfies the assumptions of Theorem 3 (as does a VARMA). In this case – with p_ω the pdf of the $\mathcal{N}(0, \underline{\Sigma}_\omega)$ distribution – we obtain

$$\mathcal{F}(p_\omega) = \underline{\Sigma}_\omega^{-1} \qquad \mathcal{G}_{jk}(p_\omega) = \frac{1}{2} \operatorname{tr}\left(\underline{\Sigma}^{-1} \nabla_j \underline{\Sigma} \cdot \underline{\Sigma}^{-1} \nabla_k' \underline{\Sigma}\right),$$

and hence (using the fact that the parametrization is separable) $F(\omega)$ equals $V(\omega)/4$, with V discussed in the previous section on Whittle estimation.

Finally, nested model hypotheses can be tested in the same way as with the Whittle likelihood, using a GLR statistic based upon the Quasi-Likelihood.

Corollary 3. *Suppose that $\{X_t\}$ is a linear process (S4.2) with marginal pdf p_ω satisfying (**C1**) and (**C2**), and that the model is correctly specified. Suppose that $\Psi_\omega(B)$ and p_ω are separately parametrized. Then*

$$GLR = (\mathcal{Q}(\widehat{\omega}_{QMLE^*}) - \mathcal{Q}(\widehat{\omega}_{QMLE})) \stackrel{\mathcal{L}}{\Longrightarrow} \chi^2_r,$$

where $\widehat{\omega}_{QMLE^*}$ is the QMLE from the restricted model.

Proof of Corollary 3. In proving Theorem 3, we establish the property

$$\sqrt{T}\,(\widehat{\omega}_{QMLE} - \widetilde{\omega}) = F(\widetilde{\omega})^{-1}\,\Delta_T + o_P(1),$$

where Δ_T is a sequence of random variables converging weakly to $\mathcal{N}(0, F(\widetilde{\omega}))$. Then we can mimic the arguments on pp. 60 – 61 of TK, using the convergence results (S4.4), to obtain the χ^2 limit. □

S4.3 Exact Maximum Likelihood Estimation

In this section we focus on Gaussian processes, and are interested in relating the quasi-likelihood $\mathcal{Q}$ to the exact likelihood $\mathcal{L}$. Ignoring constants, and following the treatment in Brockwell and Davis (1991) for the innovations algorithm, the Gaussian quadratic form in $\mathcal{L}$ can be decomposed by a block Cholesky factorization, yielding

$$\mathcal{L} = \sum_{t=1}^{T} (X_t - \widehat{X}_t)' V_t^{-1} (X_t - \widehat{X}_t) + \sum_{t=1}^{T} \log\det V_t,$$

where $\widehat{X}_t = \mathbb{E}[X_t | X_1, \ldots, X_{t-1}]$ is the projection of an observation onto past data, and V_t is the prediction error variance matrix, given by $V_t = \operatorname{Var}(X_t - \widehat{X}_t)$. The parameters of the model enter into the formulas used to computed the projections, and they also enter into each V_t. Minimizing $\mathcal{L}$ (assuming existence and uniqueness) yields the Gaussian MLE $\widehat{\omega}_{MLE}$. Under the same conditions as Theorem 3, the difference between MLE and QMLE is asymptotically negligible.

Theorem 4. *Suppose that $\{X_t\}$ is a linear process (S4.2) satisfying (**C1**) with Gaussian marginal that is $\mathcal{N}(0, \underline{\Sigma}_\omega)$, and that the model is correctly specified. Suppose that $\Psi_\omega(B)$ and $\underline{\Sigma}_\omega$ are separately parametrized. Then*

$$\sqrt{T}\,(\widehat{\omega}_{QMLE} - \widehat{\omega}_{MLE}) \xrightarrow{P} 0.$$

From this we conclude the asymptotic normality of the MLE, and the limit distribution of the associated GLR statistic.

Corollary 4. *Suppose that $\{X_t\}$ is a linear process (S4.2) satisfying (**C1**) with Gaussian marginal that is $\mathcal{N}(0, \underline{\Sigma}_\omega)$, and that the model is correctly specified. Suppose that $\Psi_\omega(B)$ and $\underline{\Sigma}_\omega$ are separately parametrized. Then the MLE is consistent and*

$$\sqrt{T}\ (\widehat{\omega}_{MLE} - \widetilde{\omega}) \stackrel{\mathcal{L}}{\Longrightarrow} \mathcal{N}\left(0, F(\widetilde{\omega})^{-1}\right),$$

with the Fisher information matrix given by $F(\omega) = V(\omega)/4$. Also,

$$GLR = (\mathcal{L}(\widehat{\omega}_{MLE^*}) - \mathcal{L}(\widehat{\omega}_{MLE})) \stackrel{\mathcal{L}}{\Longrightarrow} \chi^2_r,$$

where $\widehat{\omega}_{MLE^*}$ is the MLE from the restricted model.

Proof of Theorem 4. Let $\mathcal{E} = \mathcal{L} - \mathcal{Q}$. For any ω in a neighborhood of $\widehat{\omega}_{QMLE}$, we have

$$\nabla\mathcal{Q}(\omega) = \nabla\mathcal{Q}(\widehat{\omega}_{QMLE}) + \nabla\nabla'\mathcal{Q}(\widehat{\omega}_{QMLE})\,(\omega - \widehat{\omega}_{QMLE}) + \cdots$$

by Taylor series, where the first term on the right hand side is zero. Moreover, $\nabla\mathcal{Q} = \nabla\mathcal{L} - \nabla\mathcal{E}$, and hence

$$\widehat{\omega}_{MLE} - \widehat{\omega}_{QMLE} \approx \left[T^{-1}\nabla\nabla'\mathcal{Q}(\widehat{\omega}_{QMLE})\right]^{-1} T^{-1}\nabla\mathcal{E}(\widehat{\omega}_{MLE}),$$

where the approximation involves lower order stochastic terms. Given that T^{-1} times the Hessian of $\mathcal{Q}$ is known to converge to a p.d. matrix, as shown in the proof of Theorem 3, the assertion of this theorem follows from showing $T^{-1/2}\nabla\mathcal{E}(\omega) \xrightarrow{P} 0$ for any ω.

Because $e_t = \widehat{e}_t + \sum_{j\geq t}\Pi_j(\omega)X_{t-j}$ and $X_t = e_t - \sum_{j\geq 1}\Pi_j(\omega)X_{t-j}$, we have

$$\widehat{X}_t = -\sum_{j=1}^{t-1}\Pi_j(\omega)X_{t-j} - r_t = X_t - \widehat{e}_t - r_t,$$

with $r_t = \sum_{j\geq t}\Pi_j(\omega)\,\mathbb{E}[X_{t-j}|X_1,\ldots,X_{t-1}]$. Therefore we have the decomposition

$$X_t - \widehat{X}_t = \widehat{e}_t + r_t = e_t \oplus \sum_{j\geq t}\Pi_j(\omega)\,(\mathbb{E}[X_{t-j}|X_1,\ldots,X_{t-1}] - X_{t-j}),$$

where $\oplus$ denotes a sum where the summands are uncorrelated with each other. (This is because the second term is a linear function of X_s for $s < t$, and e_t is independent of these variables.) Therefore

$$V_t = \underline{\Sigma}_\omega + \mathrm{Var}\left[\sum_{j\geq t}\Pi_j(\omega)\,(\mathbb{E}[X_{t-j}|X_1,\ldots,X_{t-1}] - X_{t-j})\right],$$

and the second summand we denote by $M_t = V_t - \underline{\Sigma}_\omega$. Using the matrix inversion lemma,

$$\begin{aligned}(X_t - \widehat{X}_t)'\,V_t^{-1}\,(X_t - \widehat{X}_t) &= \widehat{e}_t'\,\underline{\Sigma}_\omega^{-1}\,\widehat{e}_t + 2r_t'\,\underline{\Sigma}_\omega^{-1}\,\widehat{e}_t + r_t'\,\underline{\Sigma}_\omega^{-1}\,r_t \\ &- (\widehat{e}_t + r_t)'\,\underline{\Sigma}_\omega^{-1}M_t[I + \underline{\Sigma}_\omega^{-1}M_t]^{-1}\underline{\Sigma}_\omega^{-1}\,(\widehat{e}_t + r_t).\end{aligned}$$

Here I is the T-dimensional identity matrix. Moreover, $\det V_t = \det\underline{\Sigma}_\omega \cdot \det[I + \underline{\Sigma}_\omega^{-1}M_t]$, so that

$$\begin{aligned}\mathcal{E} &= \sum_{t=1}^{T}\left(2r_t'\,\underline{\Sigma}_\omega^{-1}\,\widehat{e}_t + r_t'\,\underline{\Sigma}_\omega^{-1}\,r_t\right) - \sum_{t=1}^{T}(\widehat{e}_t + r_t)'\,\underline{\Sigma}_\omega^{-1}M_t[I + \underline{\Sigma}_\omega^{-1}M_t]^{-1}\,(\widehat{e}_t + r_t) \\ &+ \sum_{t=1}^{T}\log\det[I + \underline{\Sigma}_\omega^{-1}M_t].\end{aligned}$$

Denote these three terms respectively by $\mathcal{E}^{(1)}$, $\mathcal{E}^{(2)}$, and $\mathcal{E}^{(3)}$. For $\mathcal{E}^{(1)}$, consider the term $\sum_{t=1}^{T} r_t' \underline{\Sigma}_\omega^{-1} r_t$ (the other term follows a similar analysis). There exist matrix coefficients $D_{\ell,k}$ depending on ω such that $r_t = \sum_{\ell \leq 0} \Pi_{t-\ell}(\omega) \sum_{k=1}^{t-1} D_{\ell,k} X_k$. Hence any derivative with respect to process parameters of r_t will involve a linear combination of $X_1, \ldots X_{t-1}$; on the other hand the norm of r_t is bounded in probability of order ρ^t, because $\Pi_t(\omega)$ is the leading coefficient – here ρ governs the geometric decay of the $\{\Pi_j\}$. The derivative of $r_t' \underline{\Sigma}_\omega^{-1} r_t$ with respect to parameters in $\underline{\Sigma}_\omega$ is again a quadratic form in r_t, and hence has norm $O_P(\rho^t)$. Then the same argument used to analyze $\nabla \mathcal{Q}$ in the proof of Theorem 3 shows that $\nabla \sum_{t=1}^{T} r_t' \underline{\Sigma}_\omega^{-1} r_t$ is $O_P(1)$.

For $\mathcal{E}^{(2)}$ and $\mathcal{E}^{(3)}$, we analyze M_t, which is asymptotically negligible:

$$M_t = \sum_{j_1, j_2 \geq t} \Pi_{j_1}(\omega) \operatorname{Cov}\left[\mathbb{E}[X_{t-j_1}|X_1, \ldots, X_{t-1}] - X_{t-j_1}, \mathbb{E}[X_{t-j_2}|X_1, \ldots, X_{t-1}] - X_{t-j_2}\right] \Pi_{j_2}(\omega)'.$$

The error process $\mathbb{E}[X_{1-h}|X_1, \ldots, X_{t-1}] - X_{1-h}$ tends as $t \to \infty$ to a VMA(h) process, where $h \geq 1$. Hence the matrix norm is order ρ^{2t}, and this bound also applies to derivatives of M_t. In conjunction with the previous analysis of r_t and $\widehat{e}_t$, we obtain $\nabla \mathcal{E}^{(2)} = o_P(T^{1/2})$. Finally, the derivatives of the log determinant term with respect to process and innovation variance parameters, respectively, are given by

$$\sum_{t=1}^{T} \operatorname{tr}\{\frac{\partial}{\partial \omega_k} M_t \left(\underline{\Sigma}_\omega + M_t\right)^{-1}\}$$

$$\sum_{t=1}^{T} \operatorname{tr}\{\frac{\partial}{\partial \omega_k} \underline{\Sigma}_\omega \cdot \underline{\Sigma}_\omega^{-1} M_t \left(\underline{\Sigma}_\omega + M_t\right)^{-1}\}.$$

Therefore, the norm bound on M_t shows that $\nabla \mathcal{E}^{(3)}$ is also negligible. This completes the proof. □

Proof of Corollary 4. The conditions establish both Theorems 3 and 4; note that the Gaussian pdf satisfies assumption (**C2**). Then asymptotic normality for the MLE follows, and the Fisher information matrix in this special case is given by $V(\omega)/4$, as shown previously. The proof for the GLR results uses Corollary 3 together with the result of Theorem 4. □

Acknowledgements

We thank the Editor, associate editor, and two anonymous referees for providing valuable comments that have helped strengthen this manuscript. This research was partially supported by the U.S. National Science Foundation (NSF) and the U.S. Census Bureau under NSF grant SES-1132031, funded through the NSF-Census Research Network (NCRN) program. This report is released to inform interested parties of research and to encourage discussion. The views expressed on statistical issues are those of the authors and not necessarily those of the U.S. Census Bureau.

References

Artin, M. (1991). *Algebra*. Prentice Hall, Englewood Cliffs, New Jersey.

Brillinger, D. (1981). *Time Series: Data Analysis and Theory*. SIAM, Philadelphia, PA.

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

George, E.I. and McCulloch R.E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**, 881–889.

George, E.I. and McCulloch R.E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica* **7**, 339–373.

Hannan, E.J. and Deistler, M. (1988). *The Statistical Theory of Linear Systems*. SIAM, Pennsylvania.

Lütkepohl, H. (2007). *New Introduction to Multiple Time Series Analysis*. Springer-Verlag, Berlin.

McElroy, T.S. and Holan, S.H. (2012). On the computation of autocovariances for generalized Gegenbauer processes. *Statistica Sinica* **22**, 1661–1687.

McElroy, T.S. and McCracken, M.W. (2014). Multi-step ahead forecasting of vector time series. *Econometric Reviews* Published Online.

Shumway, R. H. and Stoffer, D. S. (2010). *Time Series Analysis and its Applications: With R Examples*. Springer, New York.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B* **64**, 583–639.

Taniguchi, M. and Kakizawa, Y. (2000). *Asymptotic Theory of Statistical Inference for Time Series*. Springer, New York.

Tsay, R.S. (2013). *Multivariate Time Series Analysis: With R and Financial Applications*. John Wiley & Sons, Hoboken, New Jersey.