# Fitting Vector Moving Averages

Tucker McElroy

September 14, 2015

**Abstract**

This research studies three methods of estimation of a Vector Moving Average (VMA) process, each of which is intended to be computationally fast and moderately accurate. These estimates could be used as initial values in the optimization routine for maximum likelihood estimators. The first technique uses Kullback Leibler discrepancy with inverse spectra, and yields a Yule-Walker system of equations in inverse autocovariances for the VMA coefficients. The second technique takes a truncated periodogram with a ridge modification (to ensure a positive definite sequence) to which spectral factorization is applied. The third technique simply takes the moving average expansion of a fitted high order Vector AutoRegression (VAR), and truncates to the desired VMA lag. The paper provides a heuristic analysis of each method, and assesses them in simulation studies for speed, bias, and precision.

## 1 Introduction

The Vector Moving Average (VMA) is an important model in econometrics and engineering, second only to the Vector AutoRegression (VAR) in terms of scope in applied time series problems. Despite its flexibility in modeling stationary vector time series, the VMA is hampered by the difficulty of estimation; the parameter constraints required to enforce causality and invertibility on the matrix moving average polynomial are complex, involving a determinental condition that is not practical for straight-forward parametrization. Apart from this issue, common objective functions – such as the Gaussian likelihood or the Whittle likelihood – involve nonlinear optimization, which is expensive and unwieldy for high dimensional data (because the dimension of the parameter space quickly become quite large). This is in contrast to VAR estimation via the Whittle likelihood, which becomes a quadratic problem with a unique (Yule-Walker) solution provided by matrix algebra (cf. McElroy and Findley (2015)). The objective of this paper is to introduce three new methods of fitting VMA models, and assess their accuracy against available methodologies.

The first method is essentially an inversion of the Yule-Walker method for VAR estimation: we reverse the role of process and model spectrum in the Kullback-Leibler (KL) discrepancy, which yields an objective function that is quadratic in the VMA coefficients, and depends on the process' inverse autocovariances. The resulting minimizers correspond to an invertible VMA process; this is called the inverse KL methd. The second technique is based on the spectral factorization of estimated autocovariances; this alone is not novel, but we here modify the resulting truncated periodogram so as to guarantee a positive definite spectral density – without which spectral factorization is not well-defined. The modification is related to the ridge alteration in regression analysis, so we refer to this as a ridged spectral factorization (SF henceforth). Finally, the third method first estimates the infinite moving average representation by inverting the fitted high-order VAR coefficients, and then truncates the result to the required VMA order; this is referred to as the Wold method.

Each of these methods yields an invertible fitted VMA model in an amount of time that is typically substantially less than that required to perform a single likelihood evaluation via the Durbin-Levinson algorithm. Using unconstrained maximum likelihood estimation (MLE) is much slower, is sensitive to initial conditions, and typically does not guarantee an invertible solution – although root-flipping can be easily done to obtain an equivalent invertible representation. However, MLEs have better theoretical statistical properties, and are preferable whenever they are practicable to compute. The three methods of this paper will be useful either as initial estimates for an MLE routine, or as final estimates in the case of a high-dimensional problem.

There is a substantial literature on initial estimation for VMA and VARMA (Hillmer and Tiao (1979), Tiao and Box (1981), Shea (1989), Mauricio (2002), Mélard, Roy, and Saidi (2002), Dufour and Jouini (2005)). While there are certain specification challenges for VARMA modeling – see the discussion in Dufour and Jouini (2005) and Lütkepohl (2007) – the case of a VMA is considerably simpler, although the Gaussan likelihood is highly non-linear in the parameters. When dimension is moderate (four or more series) the parameter space becomes large, which presents a challenge for numerical optimization; although evaluation of the Gaussian likelihood via Durbin-Levison is quite fast for samples of less than a thousand in length, the likelihood surface typically requires a long search. Dufour and Jouini (2005) say: "For example, in the Gaussian case, maximizing the likelihood function of a VARMA(p, q) model is typically a burdensome numerical exercise, as soon as the model includes a moving average part. Even numerical convergence may be problematic."

The presence of local maxima (heuristically corresponding to constrained MLEs) – leading to false MLEs – motivates the need for reasonably accurate initial values for likelihood optimization algorithms. Hannan and Rissanen (1982) proposed first estimating time series residuals via a long vector autoregression, and regressing the data on the lagged data and estimated residuals to get VARMA estimates. (The extension of the original univariate method to the VARMA case was first studied in Hannan and Kavalieris (1984, 1986).) By the term initialization, we refer to the

procedure of obtaining initial estimates – that are fast to compute, and fairly accurate – for a nonlinear optimization algorithm. Related literature, which focuses on the VARMA initialization problem, includes Hannan and Deistler (1988), Koreisha and Pukkila (1989), Huang and Guo (1990), Poskitt (1992), Lütkepohl and Poskitt (1996), Lütkepohl and Claessen (1997), and Flores de Frutos and Serrano (2002). We view the three methods of this paper – inverse KL, ridged SF, and Wold – as initialization methods.

The methods seem to have some advantages over the Hannan-Rissanen (HR) method when samples are small, but for larger samples perform similarly. The Wold method – which is very similar to the HR method in its heuristics – produces very similar results, though with improved estimation of the innovation variance matrix. The KL and ridged SF methods have very different heuristics, and can produce quite different estimates. One possible application is to use all three (or four) initializations to a nonlinear optimization procedure, and check that the same MLE is obtained. We next provide a brief discussion of the background to each method.

Kullback-Leibler (KL) discrepancy has its roots in information theory and entropy, and is treated in Taniguchi and Kakizawa (2012). Essentially, one uses the integrated ratio of process spectrum to model spectrum as an objective function; substituting the periodogram as an estimate of the process spectrum yields the Whittle likelihood. For multivariate time series, the ratio of spectrum is replaced by the trace of the product of process spectrum and inverse model spectrum; for a VAR model, the Whittle likelihood then becomes quadratic in the VAR parameters, and hence the minimum can be computed analytically (McElroy and Findley, 2015). By reversing the role of process and model spectrum for a VMA fit, we obtain a quadratic function in the VMA parameters, and only need to compute estimates of the process' inverse autocovariances. A nice feature of this inverse KL algorithm is that the estimated VMA process is always invertible.

Spectral factorization (SF) is not well-known outside the engineering community. It refers to the following question: given a sequence of matrix autocovariances for lags 0 through $q$, is there a VMA process corresponding to it? The problem is only well-posed if the spectral density corresponding to the autocovariances is positive semidefinite (psd), and in this case there exists a unique causal solution (proved in Hannan and Deistler (1988)). The problem has generated an immense literature, and many algorithms exist to produce the VMA coefficients (see Sayed and Kailath (2001) for an overview). One algorithm (Bauer, 1955), which we focus on in this paper, views these VMA coefficients as the limit of a certain sequence of partial covariances, which in turn are the entries of the lower triangular matrix in the Cholesky decomposition of the block Toeplitz covariance matrix for the process. (See McElroy (2015) for implementation.) An alternative SF algorithm – with application to VMA estimation – is discussed in Zadrozny (1998). A naïve approach to Whittle estimation for the VMA model would proceed by truncatation of the multivariate sample autocovariance sequence to lag $q$, followed by computation of the spectral factorization – this has merit, but the approach fails when the truncated autocovariances no longer form a psd spectral

density. We modify the method implicit in Zadrozny (1998) by a ridge modification of the truncated autocovariances, ensuring a psd spectrum.

The third method is easiest to explain: any invertible VMA process has an infinite order VAR representation. We propose first obtaining an estimate of this VAR($\infty$) by fitting a long autoregression (like the HR technique), and then computing the corresponding infinite order VMA representation, truncated to lag $q$. There is a fast recursive algorithm for obtaining these first $q$ VMA coefficients; higher order coefficients should be small in magnitude, if the VMA($q$) model is correctly specified. This part of the Wold method differs from HR, which utilizes the data twice – once the long autoregression's residuals are computed, the data is again regressed, this time on the residuals.

The paper's focus is on computational and empirical results. The methods are discussed in Section 2; Section 3 contains several worked examples and simulation studies, and Section 4 concludes.

# 2 Methodology

## 2.1 Background

We first set out some known results in order to establish notation; see Brockwell and Davis (2013) or Lütkepohl (2007) for more detail. An $m$-variate time series $\{X_t\}$ is a causal VMA(q) if there exists $m \times m$ matrix coefficients $\Theta_0, \Theta_1, \cdots, \Theta_q$ such that

$$X_t = \sum_{k=0}^{q} \Theta_k \epsilon_{t-k} \tag{1}$$

for a vector white noise process $\{\epsilon_t\}$. These are the innovations, and their covariance matrix $\Sigma$ is symmetric and positive semidefinite (psd). Invertibility of the process can fail when $\Sigma$ has reduced rank, and typically this generates an insoluble estimation problem; hence we assume that $\Sigma$ is full rank, and it can easily be parametrized to ensure the positive definite (pd) property. However, non-invertibility of the VMA process can still occur through the coefficients, and we allow for this in our estimation – such types of non-invertibility imply the spectral density is singular at some finite set of frequencies, as opposed to singularity of $\Sigma$, which implies the spectrum is singular at all frequencies.

For identifiability, it is typical to enforce that $\Theta_0$ is an identity matrix, denoted $1_m$. (The symbol $I$ is reserved for the periodogram.) However, an alternative VMA representation, which we also make use of below, allows the first coefficient to be non-trivial while imposing that $\Sigma = 1_m$. This is accomplished via a Cholesky factorization (there are many such factorizations possible, but we will utilize a lower triangular decomposition of the form $\Sigma = L\,D\,L'$, with $D$ a positive diagonal matrix and $L$ a unit lower triangular matrix – see Pinheiro and Bates (1998)) and the change of

variables $\Omega_k = \Theta_k L$, so that

$$X_t = \sum_{k=0} \Omega_k \eta_{t-k}$$

with $\eta_t = L^{-1}\epsilon_t$ the transformed white noise, which is now decorrelated in the sense that its covariance matrix is $D$.

With $B$ denoting the backshift operator, we define the matrix polynomial $\Theta(B) = \sum_{k=0}^{q} \Theta_k B^k$ so that (1) can be written compactly via $X_t = \Theta(B)\epsilon_t$. The necessary and sufficient conditions on invertibility of the VMA process are that the zeroes of the determinant of $\Theta(x)$, where $x$ is a complex number, lie outside the unit circle of the complex plane. Any such matrix polynomial $A(x)$ that satisfies this condition – namely, that the zeroes of the determinant of $A(x)$ lie outside the unit circle of the complex plane – is said to be *stable*; cf. discussion in Roy, McElroy, and Linton (2014).

Thus, a VAR process is stable if its VAR matrix polynomial is stable, and a VMA process is invertible if its VMA matrix polynomial is stable. In the case that $q = 1$, stability is equivalent to asserting that all $m$ eigenvalues of $\Theta_1$ have magnitude less than one. In practice, some zeroes of the determinant might actually lie on the unit circle, and this will not present insurmountable difficulties for estimation, unlike in the VAR case. However, for applications of the fitted VMA model, such as forecasting, stability of $\Theta(B)$ can be crucial, and some more discussion is warranted.

We can grapple with the stability question more directly using the frequency domain. Let $z = e^{-i\lambda}$ for $\lambda \in [-\pi, \pi]$, and define the spectral density to be the Fourier transform of the autocovariance sequence. For a stationary time series, $\Gamma(h) = \mathrm{Cov}(X_t, X_{t-h})$ and $f(\lambda) = \sum_{h=-\infty}^{\infty} \Gamma(h) z^h$ is the spectral density. This definition can be inverted via the formula $\Gamma(h) = \langle f z^{-h} \rangle$, where the angled brackets indicate integration over $[-\pi, \pi]$, divided by $2\pi$. Then it is known that the VMA spectral density has the formula

$$f(\lambda) = \Theta(z)\, \Sigma\, \Theta'(\overline{z}) \tag{2}$$

($\overline{z}$ denotes complex conjugation). A spectral density matrix has the Hermitian property that $f' = \overline{f}$, and its psd property is described in terms of complex vectors: we say $A$ is complex psd if and only if $a' A \overline{a} \geq 0$ for all complex $m$-vectors $a$. Spectral densities are complex psd for every frequency $\lambda$, and are said to be invertible if they are complex pd, namely that in addition whenever $a' f(\lambda) \overline{a} = 0$, it must be the case that $a = 0$. For any frequencies $\lambda$ such that a psd spectrum $f$ is also pd, it is true that $f(\lambda)$ is invertible. The connection to VMA processes is the following: if $\Theta(x)$ is stable and $\Sigma$ has full rank, then $f(\lambda)$ is nonsingular for all $\lambda$, i.e., it is invertible. In this case,

$$f^{-1}(\lambda) = \Theta^{\dagger}(\overline{z})\, \Sigma^{-1}\, \Theta^{-1}(z), \tag{3}$$

which can be rearranged to resemble a VAR process. Here $\dagger$ stands for inverse transpose. The formula (3) is crucial for defining the Whittle likelihood, which we discuss next – this is a useful

discussion, because it provides the context for inverse Kullback-Leibler and spectral factorization fitting.

To fit a VMA model, we consider a target spectral density $f$, which is either the periodogram (for the empirical problem) or the true spectrum (for the theoretical problem), and we propose a VMA spectrum $f_{\theta,\sigma}$ as an approximation to $f$. Here $\theta$ is the parameter vector corresponding to the VMA coefficients, and $\sigma$ corresponds to the innovation variance matrix, given by

$$\theta = \operatorname{vec}[\Theta_1, \Theta_2, \cdots, \Theta_q] \qquad \sigma = \operatorname{vech} \Sigma.$$

So $f_{\theta,\sigma}$ is our notation for a spectral density of the form (2). The Kullback-Leibler discrepancy between model and truth (cf. Taniguchi and Kakizawa (2012)), viewed as a function of model parameters $\theta$ and $\sigma$, is given by

$$\mathcal{D}(\theta, \sigma) = \langle \operatorname{tr}\left(f_{\theta,\sigma}^{-1} f\right)\rangle + \langle \log\det f_{\theta,\sigma}\rangle. \tag{4}$$

Also see McElroy and Findley (2015) for more discussion (and analysis of the VAR case), and the connection to the Gaussian likelihood and one-step ahead forecast error. The second term in (4) can be simplified to $\log\det\Sigma$, as is well-known for separable spectra (i.e., the innovation variance martrix is parametrized separately from the other parameters of the process). When $f$ is the multivariate peridogram, (4) is referred to as the Whittle likelihood. (There is also a concentrated form of the Whittle likelihood, as described in McElroy and Findley (2015), which removes the presence of the $\sigma$ parameter, but it is more convenient for us to work with the unconcentrated Whittle likelihood.)

Minimization of $\mathcal{D}(\theta, \sigma)$ with respect to $\theta$ and $\sigma$ (enforced to lie in the causal VMA parameter space described above) yields the quasi-maximum likelihood estimators (QMLEs) in the case that $f$ is the periodogram, but in the case that $f$ is a true (but unknown) spectral density we obtain the pseudo-true values (PTVs). The latter are useful for studying the impact of mis-specification, e.g., fitting a VMA(1) to a VAR(1) process, and we use the symbol $\widetilde{f}$ to denote this true spectrum for the data process. When the model is correctly specified, the PTVs are identical with the true parameters, but otherwise can be quite different. It is known that the QMLEs are consistent for the PTVs, and moreover satisfy a Central Limit Theorem under regularity conditions involving higher order cumulants of the process (Chapter 3 of Taniguchi and Kakizawa (2012)); when the model is correctly specified and there is no kurtosis, the QMLEs are also efficient.

In order to construct a QMLE, we first define the multivariate periodogram $I$. Let the discrete Fourier transform (DFT) of a sample of length $T$ from the time series, denoted $X_1, X_2, \cdots, X_T$, be given by

$$d(\lambda) = \sum_{t=1}^{T} X_t z^t.$$

This is a stochastic complex $m$-vector. The periodogram is the rank one matrix formed from the outer product of the DFT:

$$I(\lambda) = T^{-1} d(\lambda) d'(-\lambda) = \sum_{|h|<T} \widehat{\Gamma}(h) z^h.$$

The second equation follows by rearranging terms, together with the definition of the sample autocovariances as $\widehat{\Gamma}(h) = T^{-1} \sum_{t=1}^{T-|h|} X_{t+h} X_t'$. While $I$ is clearly psd at all frequencies, it is evidently not invertible, being in fact rank one.

Now to distinguish the empirical and theoretical estimation problems, which involve respectively the choice of $f = I$ and $f = \widetilde{f}$ in (4), we write $\widehat{\mathcal{D}}$ and $\widetilde{\mathcal{D}}$ for the respective Kullback-Leibler discrepancies. A QMLE is then

$$(\widehat{\theta}, \widehat{\sigma}) = \overset{-1}{\min} \widehat{\mathcal{D}}(\theta, \sigma),$$

when it exists (and solutions need not be unique), whereas the PTV is analogously defined as

$$(\widetilde{\theta}, \widetilde{\sigma}) = \overset{-1}{\min} \widetilde{\mathcal{D}}(\theta, \sigma).$$

Typically, the QMLEs (and PTVs, when desired for theoretical work) are calculated via nonlinear minimization of $\widehat{\mathcal{D}}$, e.g., via a conjugate gradient method. These methods are time-consuming, and need not even converge. The next subsections describe alternative estimation procedures that avoid nonlinear optimization, and therefore are substantially faster.

### 2.2 Inverse Kullback-Leibler

Equation (4) is an expression of the Kullback-Leibler discrepancy

$$\mathcal{K}(g, h) = \langle \operatorname{tr}\left(g^{-1} h\right) \rangle + \langle \log \det g \rangle \tag{5}$$

with $g = f_{\theta,\sigma}$ and $h = f$. Consider instead – assuming that $f$ is invertible – the KL discrepancy of the *inverse* spectra, setting $g = f_{\theta,\sigma}^{-1}$ and $h = f^{-1}$. This results in

$$\mathcal{H}(\theta, \sigma) = \langle \operatorname{tr}\left(f_{\theta,\sigma} f^{-1}\right) \rangle - \langle \log \det f_{\theta,\sigma} \rangle. \tag{6}$$

Substituting the specification (2) for $f_{\theta,\sigma}$ in (6) yields

$$\langle \operatorname{tr}\left(\Sigma \Theta'(\overline{z}) f^{-1} \Theta(z)\right) \rangle - \log \det \Sigma,$$

and optimizing with respect to $\Sigma$ (cf. McElroy and Findley (2015) for the VAR case in the Whittle likelihood) yields

$$\Sigma_\theta = \langle \Theta'(\overline{z}) f^{-1} \Theta(z) \rangle^{-1}. \tag{7}$$

Concentrating refers to substituting $\Sigma_\theta$ back into the inverse KL, and obtaining a function that only depends on $\theta$ (not on $\sigma$):

$$\mathcal{H}(\theta) = m + \log \det \langle \Theta'(\overline{z}) f^{-1} \Theta(z) \rangle,$$

which up to a constant is just the negative log determinant of $\Sigma_\theta$. Once $\theta$ has been obtained, the parameters $\sigma$ are immediately obtained via (7). To compute the minimization of $\mathcal{H}(\theta)$, it is sufficient to minimize each entry of $\Sigma_\theta^{-1}$. First note that because $f$ is Hermitian, $\Sigma_\theta$ is symmetric. Therefore, we can work with the transpose; letting $\Xi(h)$ denote the sequence of inverse autocovariances (i.e., $\Xi(h) = \langle f^\dagger z^{-h} \rangle$), we obtain

$$\Sigma_\theta^{-1} = \langle \Theta'(z) f^\dagger \Theta(\overline{z}) \rangle = \Xi(0) + \sum_{j=1}^{q} \Xi(j)\,\Theta_j + \sum_{k=1}^{q} \Theta_k'\,\Xi(-k) + \sum_{j,k=1}^{q} \Theta_j'\,\Xi(k-j)\,\Theta_k.$$

Let us write $\Xi_q$ for the block Toeplitz matrix with $jk$th block entry given by $\Xi(k-j)$, and $\Xi_{1:q} = [\Xi(1), \cdots, \Xi(q)]$. Then

$$\Sigma_\theta^{-1} = \Xi(0) + \Xi_{1:q} \begin{bmatrix} \Theta_1 \\ \vdots \\ \Theta_q \end{bmatrix} + \begin{bmatrix} \Theta_1', \cdots, \Theta_q' \end{bmatrix} \Xi_{1:q}' + \begin{bmatrix} \Theta_1', \cdots, \Theta_q' \end{bmatrix} \Xi_q \begin{bmatrix} \Theta_1 \\ \vdots \\ \Theta_q \end{bmatrix},$$

and the minimizers of the negative log determinant (it can also be shown, as in McElroy and Findley (2015), that the solution minimizes each entry of $\Sigma_\theta^{-1}$) are

$$\begin{bmatrix} \Theta_1', \cdots, \Theta_q' \end{bmatrix} = -\Xi_{1:q}\,\Xi_q^{-1}. \tag{8}$$

The corresponding VMA will have a stable matrix polynomial $\Theta(B)$, via the same proof as for the Yule-Walker case. The value at this minimizer is

$$\Xi(0) - \Xi_{1:q}\,\Xi_q^{-1}\,\Xi_{1:q}' = \Sigma_\theta^{-1}.$$

In order to construct empirical estimators, it is necessary to obtain an estimate of $f^\dagger$. One possibility is to use an autoregressive estimator of $f$, and compute the inverse transpose; however, this requires some choice of the VAR order. Essentially, one wants to take the order as large as possible, because there is no issue of overfitting here – one is concerned only with a finite collection $q$ of inverse autocovariances. Let $\Psi(z)$ denote the infinite causal moving average representation of the true process, so that

$$f(\lambda) = \Psi(z)\,\Sigma\,\Psi'(\overline{z}).$$

Assuming that $f$ is invertible, and letting $\Pi(z) = \Psi(z)^{-1}$, we have

$$f^\dagger(\lambda) = \Pi'(z)\,\Sigma^{-1}\,\Pi(\overline{z}),$$

and hence

$$\Gamma(h) = \sum_{j \geq 0} \Psi_{j+h}\,\Sigma\,\Psi_j' \qquad \Xi(h) = \sum_{j \geq 0} \Pi_{j+h}'\,\Sigma^{-1}\,\Pi_j.$$

Fitting a high order VAR produces estimates of the $\{\Pi_j\}$ and $\Sigma$, from which $\Xi(h)$ can be computed.

Another estimator of $f^\dagger$ is based on the inverse transpose of the periodogram, although this matrix is singular. Using the fact that $\langle f\, f^{-1} z^{-h} \rangle$ equals zero unless $h = 0$, we obtain a system of equations

$$\sum_{k \in \mathbb{Z}} \Gamma(h-k)\, \Xi'(k) = \begin{cases} 1_m & \text{if } h = 0 \\ 0 & \text{if } h \neq 0 \end{cases} .$$

This yields the approximate matrix system

$$\begin{bmatrix} \Gamma(0) & \cdots & \Gamma(1-T) & \cdots & \Gamma(2-2T) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \Gamma(T-1) & \cdots & \Gamma(0) & \cdots & \Gamma(1-T) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \Gamma(2T-2) & \cdots & \Gamma(T-1) & \cdots & \Gamma(0) \end{bmatrix} \begin{bmatrix} \Xi'(-T+1) \\ \vdots \\ \Xi'(0) \\ \vdots \\ \Xi'(T-1) \end{bmatrix} \approx \begin{bmatrix} 0 \\ \vdots \\ 1_m \\ \vdots \\ 0 \end{bmatrix},$$

where the approximation improves as $T \to \infty$. At this point, the matrix can be inverted, and sample autocovariance substituted for the true autocovariances, to yield the estimator

$$\begin{bmatrix} \widehat{\Xi}'(-T+1) \\ \vdots \\ \widehat{\Xi}'(0) \\ \vdots \\ \widehat{\Xi}'(T-1) \end{bmatrix} = \begin{bmatrix} \widehat{\Gamma}(0) & \cdots & \widehat{\Gamma}(1-T) & \cdots & \widehat{\Gamma}(2-2T) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \widehat{\Gamma}(T-1) & \cdots & \widehat{\Gamma}(0) & \cdots & \widehat{\Gamma}(1-T) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \widehat{\Gamma}(2T-2) & \cdots & \widehat{\Gamma}(T-1) & \cdots & \widehat{\Gamma}(0) \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ \vdots \\ 1_m \\ \vdots \\ 0 \end{bmatrix}.$$

With either type of estimator for the inverse autocovariances, we then plug into (8) to fit the VMA.

### 2.3 Ridged Spectral Factorization

If the true spectrum $\widetilde{f}$ corresponded to a $q$-dependent process, so that one knows $\Gamma(h) = 0$ for $|h| > q$, then we would know that $\widetilde{f}$ is psd and we could immediately compute its spectral factorization, thereby obtaining a stable $\Theta(B)$. If we now substitute the periodogram for $\widetilde{f}$, and do the spectral factorization, we may hope to obtain estimates of $\Theta(B)$ – however, the periodogram need not have sample autocovariances that truncate at lag $q$. We might enforce that $I$ looks like a VMA(q) spectral density, and compute estimates via spectral factorization; we know that the resulting error should tend to zero in probability, because integrals of the periodogram converge to integrals of $\widetilde{f}$ (cf. Lemma 3.1.1 of Taniguchi and Kakizwa (2012)).

Introduce the following notation: for any spectral density $f$, let $[f]_q$ denote its truncation to the first $q$ lags:

$$[f]_q(\lambda) = \sum_{|h| \leq q} \Gamma(h) z^h,$$

with residual $[f]_r = f - [f]_q = \sum_{|h|>q} \Gamma(h) z^h$. Note that $[f]_q$ need not be psd. Focusing on the periodogram, certainly $[I]_q$ need not be psd, although with increasing probability this will be the

case as sample size $T \to \infty$, because by assumption $[\widetilde{f}]_r = 0$. In order to alter $[I]_q$ to become psd, we make a ridge modification: we add some white noise to the underlying process, which is equivalent to considering $[I]_q + \alpha\,\widehat{\Gamma}(0)$ for some $\alpha \in \mathbb{R}$.

We here allow negative $\alpha$ in the case that $[I]_q$ is already psd and some advantage is obtained by subtracting white noise, although conceptually the opposite case where $\alpha$ is positive makes more intuitive sense. Also note that we consider scalar multiples of the matrix $\widehat{\Gamma}(0) = \langle I \rangle$, the sample variance of the sample. This is not the only choice available, but utilizes a known psd matrix; certainly for $\alpha$ sufficiently large, $[I]_q + \alpha\,\langle I \rangle$ will be psd.

Although we have in mind utilizing this decomposition with the periodogram, we proceed to describe the algorithm for a generic $f$. We propose the decomposition, for any scalar $\alpha$, as follows:

$$f = \Big([f]_q + \alpha\langle f\rangle\Big) + \left([f]_r - \alpha\langle f\rangle\right). \tag{9}$$

We suppose that $\alpha$ is restricted to a range of values such that the first term in (9) is psd. Plugging (9) into (4), and using the linearity of the trace, produces the following decomposition:

$$\mathcal{D}(\theta,\sigma) = \langle \operatorname{tr}\Big(f_{\theta,\sigma}^{-1}\Big([f]_q + \alpha\langle f\rangle\Big)\Big)\rangle + \langle \log\det f_{\theta,\sigma}\rangle + \langle \operatorname{tr}\Big(f_{\theta,\sigma}^{-1}\left([f]_r - \alpha\langle f\rangle\right)\Big)\rangle. \tag{10}$$

Let us decompose this Kullback-Leibler discrepancy into two terms, with the first two summands of (10) denoted by $\mathcal{D}_\alpha(\theta,\sigma)$, and the last term denoted by $\mathcal{R}_\alpha(\theta,\sigma)$. We propose to minimize $\mathcal{D}_\alpha$, for appropriately chosen $\alpha$, and ignore $\mathcal{R}_\alpha$. Since $[f]_q + \alpha\langle f\rangle$ is psd and $q$-dependent, a spectral factorization exists, which means we can find causal $\Theta_\alpha(x)$ and $\Sigma_\alpha$ such that

$$[f]_q + \alpha\langle f\rangle = \Theta_\alpha(z)\,\Sigma_\alpha\,\Theta_\alpha'(\overline{z}). \tag{11}$$

Noting that $\mathcal{D}_\alpha(\theta,\sigma)$ itself has the form of a Whittle likelihood, whose minimal possible value (see McElroy and Findley (2015)) is $m + \langle \log\det\Big([f]_q + \alpha\langle f\rangle\Big)\rangle$, or more simply $m + \log\det\Sigma_\alpha$, we see that setting $\Theta(x) = \Theta_\alpha(x)$ and $\Sigma = \Sigma_\alpha$ yields this minimial value, i.e., the minimizer of $\mathcal{D}_\alpha(\theta,\sigma)$ is obtained with $\theta$ and $\sigma$ corresponding to the spectral factorization (11).

However, such a solution minimizes $\mathcal{D}_\alpha(\theta,\sigma)$, but not necessarily $\mathcal{D}(\theta,\sigma)$, which is our ultimate goal. When $[f]_r = 0$ (so that $[f]_q$ is psd), then $\Theta_0(x)$ will be a minimizer of the Whittle likelihood, but otherwise it may be the case that other values of $\Theta_\alpha$ and $\Sigma_\alpha$ yield the minimizer of the Whittle likelihood. Consider evaluating the Whittle likelihood at the class of $\theta$ and $\sigma$ corresponding to the spectral factorization (11) – denoted by $\theta(\alpha)$ and $\sigma(\alpha)$ – such that the resulting concentrated function only depends upon $\alpha$; call this $\mathcal{Q}(\alpha)$. From the preceding discussion, we have

$$\mathcal{Q}(\alpha) = \mathcal{D}(\theta(\alpha),\sigma(\alpha)) = m + \log\det\Sigma_\alpha + \langle \operatorname{tr}\Big(f_{\theta(\alpha),\sigma(\alpha)}^{-1}\left([f]_r - \alpha\langle f\rangle\right)\Big)\rangle.$$

This provides a profiled Whittle likelihoood, utilizing a one-parameter sub-class of solutions. One possibility is to minimize $\mathcal{Q}$ with respect to $\alpha$, over the admissible range of values (i.e., such

that $[f]_q + \alpha\langle f\rangle$ is psd), say $\alpha_* = \min^{-1}\mathcal{Q}$, and taking the resulting $\theta(\alpha_*)$ and $\sigma(\alpha_*)$ as our final estimates. When working with the periodogram, we place a hat over $\mathcal{Q}$ and $\alpha_*$, but for the theoretical optimization problem we use a tilde instead.

Another approach, which is sub-optimal but avoids numerical optimization, is to choose

$$\alpha_* = \min\{-\frac{\min_{\lambda\in[-\pi,\pi]}\text{eigen}_1([f]_q(\lambda))}{\text{eigen}_1\langle f\rangle}, 0\},$$

where $\text{eigen}_1$ denotes the smallest eigen value of the matrix (it will be real, because $f$ is Hermitian). It easily follows that $[f]_q + \alpha_*\langle f\rangle$ is psd. This approach has the advantage of speed, and if the model is correctly specified then $\alpha_*$ should approach zero in probability as the sample size increases.

### 2.4 Hannan-Rissanen and Wold Estimation

The HR technique is based on viewing (1) as a multivariate regression of the dependent variable $X_t$ on independent variables $\epsilon_{t-1}, \epsilon_{t-2}, \cdots, \epsilon_{t-q}$. The term $\epsilon_t$ has coefficient matrix $1_m$, and so the equation can be rewritten

$$X_t - \epsilon_t = \sum_{k=1}^{q}\Theta_k\epsilon_{t-k}, \tag{12}$$

so that $X_t - \epsilon_t$ becomes the dependent variable. As a preliminary step, one estimates the infinite VAR representation $\epsilon_t = \Pi(B)X_t$ – say via fitting a high order VAR and retaining the residuals – and inserts the estimated residuals into the regression (12). This is the HR procedure, which can be generalized to fit VARMA models as well. It is very fast to compute, only requiring two ordinary least squares (OLS) multivariate regressions.

Actually, in the case of a VMA model this procedure can be simplified, because clearly $\Theta(B)$ corresponds to the first $q$ terms of the power series $\Psi(B) = \Pi(B)^{-1}$. Hence, only one OLS regression needs to be computed; instead of computing residuals, we can just determine the infinite VMA representation (or Wold form) corresponding to the fitted high order VAR, and take the first $q$ coefficients to estimate $\Theta(B)$. The covariance matrix $\Sigma$ can be estimated via the VAR innovation covariance matrix. This is the Wold estimation procedure.

Unsurprisingly, the HR and Wold procedures perform very similarly on data, although in our simulations the Wold procedure appears to have better performance for small samples ($T = 50$). But we can also describe the Wold estimates as the minimizers of an objective function, just like the KL and SF methods. Suppose the true process has Wold representation $X_t = \Psi(B)\epsilon_t$, but we fit the VMA model (1). The true innovations are $\epsilon_t = \Pi(B)X_t$, and plugging these into the model yields a pseudo-process $Y_t = \Theta(B)\epsilon_t$, whereas the true process is $X_t = \Psi(B)\epsilon_t$. The variance of the difference between these two processes should be small if the model is not badly mis-specified, so we adopt this as our objective function:

$$\mathcal{W}(\theta,\sigma) = \text{Var}[X_t - Y_t] = \langle(\Psi(z) - \Theta(z))\ \Sigma\ \left(\Psi'(\overline{z}) - \Theta'(\overline{z})\right)\rangle. \tag{13}$$

Given the VMA model of order $q$, the criterion (13) becomes

$$\sum_{k=1}^{q} (\Psi_k - \Theta_k)\ \Sigma\ \left(\Psi_k' - \Theta_k'\right) + \sum_{k>q} \Psi_k\, \Sigma\, \Psi_k',$$

with solutions $\Theta_k = \Psi_k$ for $1 \leq k \leq q$. Note that in this approach, $\Psi(B)$ and $\Sigma$ are already given – perhaps they've been estimated by a high order VAR, as described in section 2.2.

The estimated VMA polynomial might not be stable, unlike the estimators arising from the KL and ridged-SF methods above. (Also, the HR estimate need not be stable.) By the term *stabilization*, we refer to a procedure that calculates a stable matrix polynomial with the same VMA acf as the original. To do this, we only need to compute the acf to lag $q$ – which is guaranteed to be a psd sequence – and apply spectral factorization. This procedure is typically fast, but can be time-consuming if the original matrix polynomial has any roots close to unity.

# 3 Numerical Results

We next evaluate these methods on two bivariate VMA(1) processes. We measure performance in terms of bias and variance, even though the various objective functions used – $\mathcal{H}$, $\mathcal{Q}$, and $\mathcal{W}$ – are not directly tied to parameter mean squared error. We also compare the speed of each method (KL, SF, Wold, and HR) to a single Gaussian likelihood evaluation. We first discuss the processes, and then summarize the results.

## 3.1 Simulation Processes

The first process is a bivariate VMA(1) with a stable moving average polynomial:

$$\Theta_1 = \begin{bmatrix} -.088 & -.325 \\ .655 & -.705 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 3.612 & 1.631 \\ 1.631 & 4.410 \end{bmatrix}. \tag{14}$$

The eigenvalues have modulus .52.

The second process is a bivariate VMA(1) with eigenvalues close to unity ($-.97$ and .69):

$$\Theta_1 = \begin{bmatrix} .407 & .875 \\ .445 & -.692 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 2.016 & .301 \\ .301 & 1.146 \end{bmatrix}. \tag{15}$$

We generated 1000 Gaussian time series of length $T = 50, 100, 200, 300, 400, 500$, and evaluated each of the four methods, recording bias, variance, and average runtime. The results for HR are not reported here; they were fairly similar to the Wold method's results, although with substantial error for the $T = 50$ case (using an initial VAR estimation by Yule-Walker, as opposed to OLS, appears to improve the results).

### 3.2 Simulation Performance

Tables 1, 2, and 3 have reference to the first VMA(1) process (14). In the rows and columns corresponding to $\Theta_1$ or $\Sigma$ are the bias and variance for the corresponding matrix entry. For the sample sizes under consideration, the average run time for evaluation of the Gaussian likelihood was .063, .126, .273, .426, .621, and .741 respectively.

| Sample Size | $\Theta_1$ Bias | | $\Theta_1$ Var | | $\Sigma$ Bias | | $\Sigma$ Var | | Run-time |
|---|---|---|---|---|---|---|---|---|---|
| 50 | -.031 | -.004 | .037 | .045 | -.076 | .015 | .790 | .547 | .221 |
| | .007 | -.015 | .060 | .060 | .015 | .026 | .547 | 2.08 | |
| 100 | -.009 | -.025 | .017 | .026 | -.164 | -.082 | .351 | .261 | .136 |
| | .025 | -.052 | .028 | .046 | -.082 | -.193 | .261 | .766 | |
| 200 | -.010 | -.017 | .007 | .014 | -.095 | -.051 | .182 | .165 | .106 |
| | .008 | -.034 | .014 | .031 | -.051 | -.139 | .165 | .470 | |
| 300 | -.009 | -.012 | .006 | .009 | -.079 | -.055 | .123 | .105 | .108 |
| | .012 | -.033 | .009 | .021 | -.055 | -.154 | .105 | .306 | |
| 400 | -.004 | -.012 | .003 | .006 | -.054 | -.039 | .088 | .074 | .107 |
| | .003 | -.017 | .007 | .014 | -.039 | -.087 | .073 | .208 | |
| 500 | -.004 | -.008 | .003 | .004 | -.049 | -.030 | .064 | .056 | .100 |
| | .005 | -.016 | .005 | .010 | -.030 | -.071 | .056 | .162 | |

Table 1: SF Performance for Process 1.

Tables 4, 5, and 6 have reference to the second VMA(1) process (15). For the sample sizes under consideration, the average run time for evaluation of the Gaussian likelihood was .061, .121, .251, .415, .558, and .807.

Both bias and variance are decreasing, for all the methods, as sample size increases. In general, the KL method appears to be superior to SF in terms of bias and variance, but is slightly inferior to the Wold method. All the methods suffer from a downward bias in the estimation of $\Sigma$.

Run-time can actually decrease as $T$ increases, in some cases, because spectral factorization becomes less computationally expensive when there is better separation of the VMA eigenvalues from unity. That is, for larger $T$ the algorithm will have an easier time discriminating between the VMA roots and unity, with the result that spectral factorization takes less iterations to converge. In the KL procedure, no such factorization algorithm is features, so its run-time increases virtually monotonically. The KL method was fastest (for both processes), and the SF and Wold methods are quite a bit slower when there the process is close to instability. However, none of the methods are slow; the average run-time is comparable to a likelihood evaluation.

| Sample Size | $\Theta_1$ Bias | | $\Theta_1$ Var | | $\Sigma$ Bias | | $\Sigma$ Var | | Run-time |
|---|---|---|---|---|---|---|---|---|---|
| 50 | .031 | .068 | .023 | .015 | -.533 | -.141 | .544 | .345 | .045 |
| | -.086 | .148 | .036 | .026 | -.141 | -.436 | .345 | 1.016 | |
| 100 | .033 | .040 | .010 | .008 | -.372 | -.114 | .289 | .186 | .056 |
| | -.038 | .080 | .016 | .013 | -.114 | -.300 | .186 | .463 | |
| 200 | .016 | .026 | .005 | .004 | -.218 | -.066 | .138 | .092 | .070 |
| | -.023 | .045 | .007 | .006 | -.066 | -.197 | .092 | .216 | |
| 300 | .012 | .023 | .003 | .002 | -.165 | -.057 | .093 | .061 | .074 |
| | -.014 | .033 | .004 | .004 | -.057 | -.166 | .061 | .138 | |
| 400 | .010 | .014 | .002 | .002 | -.125 | -.051 | .068 | .047 | .086 |
| | -.014 | .024 | .003 | .003 | -.051 | -.128 | .047 | .101 | |
| 500 | .008 | .015 | .002 | .002 | -.111 | -.042 | .050 | .036 | .079 |
| | -.008 | .018 | .002 | .002 | -.042 | -.112 | .036 | .076 | |

Table 2: KL Performance for Process 1.

# 4 Conclusion

This paper introduces and studies three methods for fitting vector moving averages. Each method is justified through a different objective function, and hence can be expected to generate estimates with somewhat different properties. The SF, KL, and Wold methods are fast to compute, and competitive with the HR technique in limited simulation studies. Future research will focus on the harder problem of estimating a VARMA process, and extending the Wold and KL techniques to this class of models.

# References


[1] Bauer, F. (1955) Ein direktes Iterationsverfahren zur Hurwitz-Zerlegung eines Polynoms. *Archiv. Elekt. Ubertr.* **9**, 285–290.

[2] Brockwell, P. J., and Davis, R. A. (2013) *Time series: theory and methods*. Springer Science & Business Media.

[3] Dufour, J.-M. and Jouini, T. (2005) Asymptotic distribution of a simple linear estimmator for VARMA models in echelon form. In *Statistical Modeling and Analysis for Complex Data Problems*, eds. Duchesne, P. and Rémillard, B. Kluwer/Springer-Verlag, Canada, chapter 11, 209–240.

[4] Dufour, J.-M. and Jouini, T. (2014) Asymptotic distributions for quasi-efficient estimators in echelon VARMA models. *Computational Statistics and Data Analysis* **73**, 69–86.

| Sample Size | $\Theta_1$ Bias | | $\Theta_1$ Var | | $\Sigma$ Bias | | $\Sigma$ Var | | Run-time |
|---|---|---|---|---|---|---|---|---|---|
| 50 | -.011 | .054 | .028 | .021 | -.294 | -.065 | .553 | .372 | .063 |
| | -.048 | .101 | .038 | .029 | -.065 | -.225 | .372 | .946 | |
| 100 | .008 | .029 | .013 | .011 | -.206 | -.056 | .281 | .193 | .075 |
| | -.014 | .047 | .017 | .014 | -.056 | -.145 | .193 | .442 | |
| 200 | .001 | .017 | .006 | .005 | -.108 | -.025 | .137 | .095 | .088 |
| | -.011 | .025 | .008 | .007 | -.025 | -.090 | .095 | .210 | |
| 300 | .000 | .015 | .004 | .004 | -.083 | -.024 | .093 | .062 | .092 |
| | -.005 | .016 | .005 | .004 | -.024 | -.082 | .062 | .133 | |
| 400 | .002 | .005 | .003 | .003 | -.061 | -.025 | .068 | .049 | .104 |
| | -.007 | .011 | .004 | .003 | -.025 | -.061 | .049 | .100 | |
| 500 | .001 | .007 | .002 | .002 | -.056 | -.019 | .050 | .036 | .094 |
| | -.003 | .007 | .003 | .003 | -.019 | -.053 | .036 | .075 | |

Table 3: Wold Performance for Process 1.


[5] Flores de Frutos, R. and Serrano, G. R. (2002) A generalized least squares estimation method for VARMA models. *Statistics* **36**, 303–316.

[6] Hannan, E. J. and Deistler, M. (1988) *The Statistical Theory of Linear Systems*. John Wiley & Sons, New York.

[7] Hannan, E. J. and Rissanen, J. (1982) Recursive estimation of mixed autoregressive-movingaverage order. *Biometrika* **69**, 81–94. Errata 70 (1983), 303.

[8] Hannan, E. J. and Kavalieris, L. (1984) A method for autoregressive-moving average estimation. *Biometrika* **71**, 273–280.

[9] Hannan, E. J. and Kavalieris, L. (1986) Regression, autoregression models. *Journal of Time Series Analysis* **7**, 27–49.

[10] Hillmer, S. C. and Tiao, G. C. (1979) Likelihood function of stationary multiple autoregressive moving average models. *Journal of the American Statistical Association* **74**, 652–660.

[11] Huang, D. and Guo, L. (1990) Estimation of nonstationary ARMAX models based on the Hannan- Rissanen method. *The Annals of Statistics* **18**, 1729–1756.

[12] Kascha, C. (2012) A comparison of estimation methods for vector autoregressive moving-average models. *Econometric Reviews* **31**, 297–324.

[13] Koreisha, S. G. and Pukkila, T. M. (1989) Fast linear estimation methods for vector autoregressive moving-average models. *Journal of Time Series Analysis* **10**, 325–339.

| Sample Size | $\Theta_1$ Bias | | $\Theta_1$ Var | | $\Sigma$ Bias | | $\Sigma$ Var | | Run-time |
|---|---|---|---|---|---|---|---|---|---|
| 50 | -.083 | -.222 | .037 | .030 | .661 | -.117 | .842 | .123 | .335 |
| | -.159 | -.226 | .013 | .034 | -.117 | .464 | .123 | .260 | |
| 100 | -.055 | -.165 | .027 | .017 | .442 | -.089 | .318 | .054 | .373 |
| | -.126 | .189 | .008 | .024 | -.089 | .332 | .054 | .098 | |
| 200 | -.019 | -.130 | .019 | .012 | .315 | -.078 | .178 | .029 | .294 |
| | -.096 | .158 | .005 | .018 | -.078 | .258 | .029 | .054 | |
| 300 | -.012 | -.117 | .017 | .008 | .269 | -.070 | .126 | .023 | .326 |
| | -.083 | .137 | .004 | .014 | -.070 | .222 | .023 | .037 | |
| 400 | .005 | -.099 | .016 | .006 | .214 | -.068 | .104 | .018 | .354 |
| | -.069 | .126 | .003 | .012 | -.068 | .197 | .018 | .026 | |
| 500 | .003 | -.095 | .014 | .006 | .200 | -.070 | .073 | .014 | .392 |
| | -.071 | .122 | .003 | .011 | -.070 | .176 | .014 | .021 | |

Table 4: SF Performance for Process 2.


[14] Lütkepohl, H. (2007) *New introduction to multiple time series analysis*. Springer Science & Business Media.

[15] Lütkepohl, H. and Claessen, H. (1997) Analysis of cointegrated VARMA processes. *Journal of Econometrics* **80**, 223–239.

[16] Lütkepohl, H. and Poskitt, D. S. (1996) Specification of echelon-form VARMA models. *Journal of Business and Economic Statistics* **14**, 69–79.

[17] Mauricio, J. A. (2002) An algorithm for the exact likelihood of a stationary vector autoregressive-moving average model. *Journal of Time Series Analysis* **23**, 473–486.

[18] McElroy, T.S. (2015) Recursive computation for block nested covariance matrices. *Mimeo.*

[19] McElroy, T., and Findley, D. (2015) Fitting constrained vector autoregression models. In *Empirical Economic and Financial Research*, pp. 451–470. Springer International Publishing.

[20] Mélard, G., Roy, R., and Saidi, A. (2006) Exact maximum likelihood estimation of structured or unit root multivariate time series models. *Computational Statistics and Data Analysis* **50**, 2958-2986.

[21] Pinheiro, J. C., and Bates, D. M. (1996) Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing*, **6**(3), 289–296.

[22] Poskitt, D. S. (1992) Identification of echelon canonical forms for vector linear processes using least squares. *The Annals of Statistics* **20**, 195–215.

| Sample Size | $\Theta_1$ Bias | | $\Theta_1$ Var | | $\Sigma$ Bias | | $\Sigma$ Var | | Run-time |
|---|---|---|---|---|---|---|---|---|---|
| 50 | -.058 | -.224 | .012 | .028 | -.298 | -.155 | .226 | .044 | .044 |
| | -.175 | .269 | .010 | .011 | -.155 | -.093 | .044 | .094 | |
| 100 | -.019 | -.149 | .006 | .012 | -.252 | -.115 | .104 | .021 | .057 |
| | -.124 | .208 | .004 | .005 | -.115 | -.084 | .021 | .043 | |
| 200 | .010 | -.100 | .003 | .006 | -.190 | -.084 | .046 | .012 | .069 |
| | -.082 | .150 | .001 | .003 | -.084 | -.067 | .012 | .018 | |
| 300 | .013 | -.084 | .002 | .004 | -.160 | -.066 | .032 | .008 | .077 |
| | -.066 | .125 | .001 | .002 | -.066 | -.053 | .008 | .013 | |
| 400 | .015 | -.069 | .001 | .003 | -.140 | -.055 | .024 | .006 | .084 |
| | -.055 | .107 | .001 | .001 | -.055 | -.048 | .006 | .010 | |
| 500 | .017 | -.061 | .001 | .002 | -.121 | -.052 | .018 | .005 | .096 |
| | -.049 | .095 | .000 | .001 | -.052 | -.046 | .005 | .007 | |

Table 5: KL Performance for Process 2.


[23] Roy, A., McElroy, T., and Linton, P. (2014) Estimation of causal invertible VARMA models. arXiv:1406.4584 [math.ST]

[24] Sayed, A. and Kailath, T. (2001) A survey of spectral factorization methods. *Numerical Linear Algebra with Applications* **8**, 467–496.

[25] Shea, B. L. (1989) The exact likelihood of a vector autoregressive moving average model. *Journal of the Royal Statistical Society, Series C Applied Statistics* **38**, 161–184.

[26] Taniguchi, M., and Kakizawa, Y. (2012) *Asymptotic theory of statistical inference for time series*. Springer Science & Business Media.

[27] Tiao, G. C. and Box, G. E. P. (1981) Modeling multiple time series with applications. *Journal of the American Statistical Association* **76**, 802–816.

[28] Zadrozny, P. (1998) An eigenvalue method of undetermined coefficients for solving linear rational expectations models. *Journal of Economic Dynamics and Control* **22**, 1353–1373.

| Sample Size | $\Theta_1$ Bias | | $\Theta_1$ Var | | $\Sigma$ Bias | | $\Sigma$ Var | | Run-time |
|---|---|---|---|---|---|---|---|---|---|
| 50 | -.024 | -.168 | .022 | .039 | -.105 | -.142 | .203 | .053 | .573 |
| | -.128 | .200 | .016 | .022 | -.142 | .049 | .053 | .089 | |
| 100 | -.003 | -.111 | .011 | .018 | -.099 | -.108 | .095 | .024 | .512 |
| | -.085 | .147 | .008 | .012 | -.108 | .027 | .024 | .040 | |
| 200 | .016 | -.076 | .005 | .010 | -.074 | -.082 | .042 | .013 | .495 |
| | -.054 | .106 | .003 | .006 | -.082 | .020 | .013 | .017 | |
| 300 | .013 | -.061 | .004 | .006 | -.064 | -.065 | .029 | .009 | .680 |
| | -.043 | .085 | .002 | .004 | -.065 | .020 | .009 | .012 | |
| 400 | .015 | -.050 | .003 | .005 | -.053 | -.054 | .022 | .006 | .787 |
| | -.033 | .070 | .002 | .003 | -.054 | .018 | .006 | .009 | |
| 500 | .014 | -.043 | .002 | .004 | -.045 | -.052 | .017 | .005 | .794 |
| | -.033 | .064 | .001 | .002 | -.052 | .011 | .005 | .007 | |

Table 6: Wold Performance for Process 2.