JOURNAL OF TIME SERIES ANALYSIS
J. Time Ser. Anal. **39**: 172–191 (2018)
Published online 7 December 2017 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1111/jtsa.12276

ORIGINAL ARTICLE

THE INVERSE KULLBACK–LEIBLER METHOD FOR FITTING VECTOR MOVING AVERAGES[1]

TUCKER MCELROY* AND ANINDYA ROY
Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, DC, USA

A new method for the estimation of a vector moving average (VMA) process is presented. The technique uses Kullback–Leibler discrepancy with inverse spectra, and yields a Yule–Walker system of equations in inverse autocovariances for the VMA coefficients. This provides a direct formula for the coefficients, which always results in a stable matrix polynomial. The paper provides asymptotic results, as well as an analysis of the method's performance, in terms of speed, bias, and precision. Applications to preliminary estimation of VMA models are discussed, and the method is illustrated on retail data.

Received 07 March 2017; Accepted 31 October 2017

Keywords: Spectral factorization; stability; vector autoregression

JEL: C18

MOS subject classification: 62M10.

1. INTRODUCTION

The vector moving average (VMA) is an important model in econometrics and engineering, second only to vector autoregression (VAR) in terms of scope in applied time series problems. The VMA model is superior to VAR when the process's autocorrelations truncate to zero at some lag; they can arise naturally for processes consisting of underlying random walk trend and white noise components (see Harvey 1989). Despite its flexibility in modeling stationary vector time series, 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 determinantal condition that is not practical for straightforward parameterization. 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 becomes 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. The objective of this paper is to introduce a new method of fitting VMA models and assess its accuracy against other available methodologies.

The inverse Kullback–Leibler (IKL) 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's inverse autocovariances. The resulting formulas are a multivariate generalization of the method of Durbin (1959), with minimizers that correspond to an invertible VMA process. The IKL method yields an invertible fitted VMA model in an amount of time that is typically substantially less than that required to perform a single Gaussian likelihood evaluation via the Durbin–Levinson algorithm. The use of unconstrained maximum likelihood estimation

* Correspondence to: Tucker McElroy, Center for Statistical Research and Methodology, U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233-9100, USA. E-mail: tucker.s.mcelroy@census.gov

[1] 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.

(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. Therefore, in cases where computation is challenging (such as in high-dimensional time series data), the IKL method of this paper will be useful, and even in cases where MLE is utilized, the IKL estimates can function as initial estimates.

There is a substantial literature on initial estimation for VMA and VARMA, including Hillmer and Tiao (1979), Tiao and Box (1981), Shea (1989), Mauricio (2002), Mélard *et al*. (2006), and 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 nonlinear in the parameters. When the 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) and saddle points leads to false MLEs, and motivates the need for accurate initial values for likelihood optimization algorithms. The general task of obtaining initial estimates that are accurate and fast to compute for a nonlinear optimization algorithm is referred to as the initialization problem. Hannan and Rissanen (1982) proposed first estimating time series residuals via a long VAR, and regressing the data upon 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).] Related literature that 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). Aside from the motivation of generating initial estimates for an MLE optimization, the estimates arising from IKL may be useful when there are departures from Gaussianity – MLEs, even under a correct time series model specification, are inefficient if the tri-spectrum is nontrivial (appendix to Holan *et al*., 2017), and hence the trade-off between computation speed and variance favors analytical methods such as IKL.

Recall that KL discrepancy has its roots in information theory and entropy – the topic is treated in Taniguchi and Kakizawa (2012). Essentially, one uses the integrated ratio of the process spectrum to the 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 the process spectrum and the 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 the process spectrum and the model spectrum for a VMA fit, we obtain a quadratic function in the VMA parameters and only need to compute estimates of the process's inverse autocovariances. A desirable feature of IKL is that the estimated VMA process is always invertible.

In the univariate case, the IKL method is easier to describe. If fitting an MA(q) to some process with spectral density f, the KL discrepancy is $(2\pi)^{-1} \int_{-\pi}^{\pi} f(\lambda)\, |\theta(\mathrm{e}^{-\mathrm{i}\lambda})|^{-2}\, \mathrm{d}\lambda/\sigma^2 + \log \sigma^2$, where $\theta(B) = 1 + \sum_{j=1}^{q} \theta_j B^j$ is the MA polynomial (with B denoting the backshift operator) and σ^2 is the innovation variance. The KL discrepancy is a nonlinear function of the MA coefficients; swapping the role of process and model spectral density yields the IKL

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} |\theta(\mathrm{e}^{-\mathrm{i}\lambda})|^2 / f(\lambda)\, \mathrm{d}\lambda\, \sigma^2 - \log \sigma^2,$$

where the first term is quadratic in the MA coefficients. Therefore, the minimum value of the IKL is given by negative one times the solution of the Yule–Walker equations involving inverse autocovariances (i.e., the

DOI: 10.1111/jtsa.12276

autocovariances corresponding to f^{-1}), and σ^2 equals the integral. For example, if $q = 1$, we obtain

$$\left(\xi_0 (1 + \theta_1^2) + 2\xi_1 \theta_1\right) \sigma^2 - \sigma^{-2},$$

where ξ_1 and ξ_0 are the lag one and lag zero inverse autocovariances. The minimizers are $\theta_1 = -\xi_1/\xi_0$ and $\sigma^2 = \xi_0 - \xi_1^2/\xi_0$. The IKL solutions correspond to a stable polynomial (i.e., all roots outside the unit circle), and, moreover, in the case of a mis-specified model, also provide an MA description that is close to the true process in the sense of KL. For example, if the true process were an AR(1) of parameter ϕ_1, then the IKL minimizer is $\theta_1 = \phi_1/(1 + \phi_1^2)$. This exactly parallels the result that fitting an AR(1) to an MA(1) with KL yields (via the Yule–Walker equations) the minimizer $\phi_1 = \theta_1/(1 + \theta_1^2)$. While these ideas are simple enough in the univariate case, their development for multivariate time series requires more care – but there is more potential gain, in terms of computational cost, as we demonstrate below.

2. BACKGROUND

Although a VMA(q) process can be approximated by a VAR(p) for p sufficiently large, this does not mean we should just use high-order VAR(p) models even though these are easier to estimate. We illustrate this with the following univariate example: suppose we want to forecast h steps ahead, and $h \geq q + 1$. Also suppose that the MA(q) is well approximated by an AR(1), because the MA polynomial is $\theta(B) = \sum_{j=0}^{q} \phi_1^j B^j$, which equals $(1 - \phi_1^{q+1} B^{q+1})/(1 - \phi_1 B)$. If we fitted both the AR(1) and MA(q) models, the former model would likely provide estimates close to ϕ_1 if ϕ_1^{q+1} is small (and sample size is large enough), and hence seems more parsimonious than the MA(q) model, which requires q parameters. However, the h-step-ahead forecast of the MA(q) process is zero, with forecast error γ_0, the lag zero autocovariance (see McElroy and Findley, 2010 for calculations). If we forecast with an approximating AR(1) model of parameter ϕ_1, the h-step-ahead forecast mean squared error is

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \sum_{k=0}^{h-1} \phi_1^k \mathrm{e}^{-\mathrm{i}\lambda k} (1 - \phi_1 \mathrm{e}^{-\mathrm{i}\lambda}) \right|^2 f(\lambda)\, \mathrm{d}\lambda = (1 + \phi_1^{2h})\, \gamma_0 - 2\phi_1^h\, \gamma_h.$$

Because the true process is an MA(q), the lag h autocovariance γ_h equals zero, and so the forecast mean squared error is strictly larger than γ_0 when $\phi_1 \neq 0$. Therefore, the slightly mis-specified AR(1) model, even though it seems to be more parsimonious, has inferior forecast performance when $h \geq q + 1$. The point is, it is indeed important to get the model form correct when applications such as forecasting or signal extraction are needed.

We next set out some known results in order to establish notation; see Brockwell and Davis (1991) or Lütkepohl (2007) for more details. An m-variate mean-zero time series $\{X_t\}$ is a causal VMA(q) if there exist $m \times m$ matrix coefficients $\Theta_0, \Theta_1, \ldots, \Theta_q$ 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 Σ is symmetric and positive semidefinite (psd). Invertibility of the process can fail when Σ has reduced rank, and typically this generates an insolvable estimation problem; hence we assume that Σ is full rank, and it can easily be parameterized to ensure the positive definite (pd) property. However, non-invertibility of the VMA process can still occur through the coefficients; such types of non-invertibility imply that the spectral density is singular at some finite set of frequencies, as opposed to singularity of Σ, which implies that the spectrum is singular at all frequencies. For identifiability, it is typical to enforce that Θ_0 is an identity matrix, denoted I_m.

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

DOI: 10.1111/jtsa.12276

$\Theta(x)$, 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 *et al*. (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 $q = 1$, stability is equivalent to asserting that all m eigenvalues of Θ_1 have magnitude less than 1. In practice, some zeroes of the determinant might actually lie on the unit circle, but 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 address the stability question more directly using the frequency domain. Let $z = \mathrm{e}^{-\mathrm{i}\lambda}$ 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 2π. Then it is known (Brockwell and Davis, 1991) that the VMA spectral density has the formula

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

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

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

which can be rearranged to resemble the spectrum of a VAR process. Here, † stands for the 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 KL 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, θ is the parameter vector corresponding to the VMA coefficients, and σ corresponds to the innovation variance matrix, given by

$$\theta = \mathrm{vec}\left[\Theta_1, \Theta_2, \ldots, \Theta_q\right], \qquad \sigma = \mathrm{vech}\, \Sigma.$$

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

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

(Strictly speaking, an additional $\langle \log \det f^{-1} \rangle$ term is needed, but as this does not depend upon the parameters, it is left out of the definition.) 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 matrix is parameterized separately from the other parameters of the process). When f is the multivariate periodogram, (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 σ parameter, but it is more convenient for us to work with the unconcentrated Whittle likelihood.]

DOI: 10.1111/jtsa.12276

Minimization of $\mathcal{D}(\theta, \sigma)$ with respect to θ and σ (enforced to lie in the stable 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 $\tilde{f}$ to denote this true spectrum for the data process. When the model is correctly specified, the PTVs are identical to the true parameters; but otherwise they 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 (Taniguchi and Kakizawa, 2012, chapter 3); 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 $\widehat{f}$. Let the discrete Fourier transform (DFT) of a sample of length T from the time series, denoted $X_1, X_2, \ldots, X_T$, be given by

$$d(\lambda) = T^{-1/2} \sum_{t=1}^{T} X_t z^t.$$

This is a stochastic complex m-vector. The periodogram is the rank-1 matrix formed from the outer product of the DFT:

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

The second equation follows by rearranging terms, together with the definition of the sample autocovariances as $\widehat{\Gamma}(h) = T^{-1} \sum_{t=1}^{T-h} X_{t+h} X_t'$ for $h \geq 0$ (and $\widehat{\Gamma}(h) = \widehat{\Gamma}(-h)'$ for $h \leq 0$). While $\widehat{f}$ is clearly psd at all frequencies, it is evidently not invertible, being in fact rank-1.

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

$$(\widehat{\theta}, \widehat{\sigma}) = \arg\min \widehat{\mathcal{D}}(\theta, \sigma)$$

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

$$(\tilde{\theta}, \tilde{\sigma}) = \arg\min \widetilde{\mathcal{D}}(\theta, \sigma).$$

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.

3. METHODOLOGY

3.1. Inverse Kullback–Leibler

We begin with a discussion of KL discrepancy between the model and process, conceived in terms of multivariate spectral densities, and obtain a population (or process) version of IKL. Later in this section we discuss how plug-in estimators for the process's spectral density can then provide actual estimators. Equation (4) is an expression of the KL discrepancy

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

DOI: 10.1111/jtsa.12276

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), and utilizing the fact that the model spectral density is separable (recall the discussion following equation (4)), yields

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

and optimizing with respect to Σ (see McElroy and Findley, 2015 for the VAR case in the Whittle likelihood) yields

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

Concentrating refers to substituting Σ_θ back into the inverse KL and obtaining a function that depends only on θ (not on σ):

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

which up to a constant is just the negative log determinant of Σ_θ. Once θ has been obtained, the parameters σ are immediately obtained via (7). First note that, because f is Hermitian, Σ_θ 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

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

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

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

which is a quadratic equation in the matrix coefficients Θ_k. In McElroy and Findley (2015) it was shown that for such a matrix system each individual equation (i.e., each entry of the matrix Σ_θ^{-1}) is minimized with respect to θ by the expression

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

Hence, θ corresponding to (8) also minimizes the determinant of Σ_θ^{-1}, and thereby $\mathcal{H}(\theta)$ itself. We emphasize that (8) is a population equation, and cannot yield an actual estimator until we supply sample estimates for the inverse autocovariances. The value at this minimizer is

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

and the corresponding VMA will have a stable matrix polynomial $\Theta(B)$.

Proposition 1. The VMA polynomial $\Theta(B) = I_m + \sum_{k=1}^{q} \Theta_k B^k$ with coefficients given by (8), where $\Xi(h)$ is the inverse autocovariance function of a process with invertible spectral density f, is stable.

DOI: 10.1111/jtsa.12276

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 p of the VAR order. Let $\Psi(z)$ denote the infinite causal MA 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

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

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(p) produces estimates of the $\{\Pi_j\}$ and Σ, 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

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

This yields the approximate matrix system

$$\begin{bmatrix} \Gamma(0) & \dots & \Gamma(1-H) & \dots & \Gamma(2-2H) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \Gamma(H-1) & \dots & \Gamma(0) & \dots & \Gamma(1-H) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \Gamma(2H-2) & \dots & \Gamma(H-1) & \dots & \Gamma(0) \end{bmatrix} \begin{bmatrix} \Xi(-H+1)' \\ \vdots \\ \Xi(0)' \\ \vdots \\ \Xi(H-1)' \end{bmatrix} \approx \begin{bmatrix} 0 \\ \vdots \\ I_m \\ \vdots \\ 0 \end{bmatrix},$$

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

$$\begin{bmatrix} \widehat{\Xi}(-H+1)' \\ \vdots \\ \widehat{\Xi}(0)' \\ \vdots \\ \widehat{\Xi}(H-1)' \end{bmatrix} = \begin{bmatrix} \widehat{\Gamma}(0) & \dots & \widehat{\Gamma}(1-H) & \dots & \widehat{\Gamma}(2-2H) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \widehat{\Gamma}(H-1) & \dots & \widehat{\Gamma}(0) & \dots & \widehat{\Gamma}(1-H) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \widehat{\Gamma}(2H-2) & \dots & \widehat{\Gamma}(H-1) & \dots & \widehat{\Gamma}(0) \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ \vdots \\ I_m \\ \vdots \\ 0 \end{bmatrix}.$$

Yet another possibility is to use a tapered autocovariance spectral estimator or a windowed periodogram (Phillips *et al*., 2006). These nonparametric estimators involve a choice of bandwidth and kernel, but have the advantage of allowing direct computation of f and its inverse. (One can guarantee that $f(\lambda)$ is pd for each λ by appropriate choice of taper or kernel, and hence ensure invertibility of the estimate.) All these methods involve some choices by the user: either p in the VAR(p) estimator, H in the approximate inverse peridogram approach, or the size of bandwidth.

With any of these estimators for the inverse autocovariances, we then plug into (8) to fit the VMA. These derivations implicitly assume that the inverse of $\Psi(B)$ exists and is the convergent power series $\Pi(B)$. The coefficients of $\Pi(B)$ are utilized to compute multi-step-ahead predictors (from an infinite past); therefore for forecasting and signal extraction applications, the convergence of $\Pi(B)$ is crucial (see McElroy and McCracken, 2017 for details). In

DOI: 10.1111/jtsa.12276

the case of a VMA, $\Theta(B) = \Psi(B)$ and convergence of $\Pi(B)$ is equivalent to the stability of the VMA polynomial. Put another way, unstable VMA polynomials yield explosive (and nonsensical) forecasts.

3.2. Alternative Methodologies

We review the Hannan-Rissanen (HR) technique, which is based on viewing (1) as a multivariate regression of the dependent variable X_t on independent variables $\epsilon_{t-1}, \epsilon_{t-2}, \ldots, \epsilon_{t-q}$. The term ϵ_t has coefficient matrix I_m, and so the equation can be rewritten as

$$X_t - \epsilon_t = \sum_{k=1}^{q} \Theta_k \epsilon_{t-k}, \tag{9}$$

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 (9). This is the HR procedure, which can be generalized to fit VARMA models as well. [Although attributed to Hannan and Rissanen (1982), examination of Durbin (1960, equation (25)) shows the same methodology.] 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 Σ can be estimated via the VAR innovation covariance matrix. This will be called the WOLD procedure; it is very similar to the method of Galbraith *et al*. (2002). [In the method of Galbraith *et al*. (2002), $\Pi(B)\Theta(B)$ is approximately equal to the identity matrix, and hence the VMA coefficients can be recursively solved in terms of the coefficients of $\Pi(B)$.] 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 ($T = 50$).

The estimated VMA polynomial from the HR (or WOLD) procedure might not be stable, unlike the estimators arising from the IKL method. 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 first compute the acf – which is guaranteed to be a psd sequence – to lag q, and apply the Bauer (1955) spectral factorization algorithm. Spectral factorization refers to the calculation of an order-q matrix polynomial $\Theta(x)$, as well as Σ, from $\Gamma(0), \ldots, \Gamma(q)$ such that (2) holds. Many algorithms exist (Sayed and Kailath, 2001), but here we focus on the method of Bauer (1955), which has the property that the resulting $\Theta(x)$ is always stable.

The Bauer spectral factorization (discussed in McElroy, 2017) provides a stabilization of the HR estimates. The algorithm proceeds by computing the unit lower triangular Cholesky decomposition of the block Toeplitz covariance matrix corresponding to a given f and taking the bottom row of the lower Cholesky factor to be the VMA coefficients written in reverse order. This computation can proceed recursively as the dimension of the block Topelitz matrix increases, with termination occurring when the extracted coefficients converge. Conceptually this approach is quite similar to the method of Mitchell and Brockwell (1997), which applies the innovation algorithm to sample autocovariances in order to efficiently compute VMA coefficient estimates. The Bauer procedure is typically fast, but can be time consuming if the original matrix polynomial has any roots close to unity. This means that stabilization of HR (or WOLD) estimates can be quite slow in some cases, whereas the IKL method is automatically stable (and much faster). In simulation we found that the stabilized estimates were not as competitive as the unstabilized estimates.

In fact, the Bauer algorithm suggests another method for fitting VMA that is implicit in Zadrozny (1998): simply apply Bauer spectral factorization to the sample autocovariance sequence computed to lag q. [Here we intend to utilize block Toeplitz covariance matrices of increasing dimension, based on the truncated autocovariance sequence, with termination of the algorithm when a convergence criterion is satisfied; this is slightly different from the application made in the linear process bootstrap – see McMurry and Politis (2010) and Jentsch and Politis

DOI: 10.1111/jtsa.12276

(2015) – where the dimension of the covariance matrix is fixed to be the sample size.] This truncation is equivalent to working with a modified periodogram, defined as

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

This has the important drawback that $[\widehat{f}]_q$ need not be psd – although we might increase q to some q' until the psd property holds, and discard the latter $q' - q$ VMA coefficients that are obtained. The resulting parameter estimates, assuming that $[\widehat{f}]_q$ is psd, are exactly the Whittle estimates obtained by fitting a VMA(q) to this truncated periodogram, i.e., minimization of $\mathcal{K}(f_{\theta,\sigma}, [\widehat{f}]_q)$ for KL given by (5) and spectral density $f_{\theta,\sigma}$ corresponding to a VMA(q) model.

We have experimented with a version of this procedure – wherein the truncated periodogram is 'ridge'-modified (by the addition of a multiple of the sample variance) to a psd version – but have found that the results (in terms of accuracy and computation speed) are inferior to those of IKL across a range of VMA processes. This method is very expensive when the roots are close to unity; moreover, the ridge modification causes some distortion to the empirical dynamics. Therefore, we have abandoned this class of estimators in favor of the higher performing IKL (and WOLD) methods.

3.3. Theoretical Properties of Estimators

This section describes the asymptotic theory for the VMA estimators in the case that an autoregressive estimator is used for f. The PTVs in this case are the minimizers of (7) with respect to $\Theta_1, \ldots, \Theta_q$, where f is the true spectral density, and the unique solution is given by (8). To emphasize that the inverse autocovariances are computed from some true spectral density f, we write $\tilde{\Xi}(j)$, and denote the corresponding PTVs by $\tilde{\Theta}_j$. In contrast, when a VAR(p) spectral estimator $\widehat{f}$ is utilized, the corresponding inverse autocovariances are denoted $\widehat{\Xi}(j)$ with IKL estimators $\widehat{\Theta}_j$. These estimators are consistent for the PTVs under broad regularity conditions and, moreover, are asymptotically normal. The following theorem provides the asymptotic normality result. Before stating the theorem, we first establish some notations. Let $[\Phi_1, \Phi_2, \ldots, \Phi_p]$ denote the coefficients of the VAR(p) approximation, and let Σ be the innovation variance. Let K denote the commutation matrix (Lütkepohl, 2007, A.12, p. 663) such that $\operatorname{vec} \Theta_j' = K \operatorname{vec} \Theta_j$, and let D be the $m^2 \times \frac{1}{2}m(m+1)$ duplication matrix such that for any symmetric $m \times m$ matrix G, $\operatorname{vec}(G) = D \operatorname{vech}(G)$; see Lütkepohl (2007, A.12, p. 662). Let D^+ be the Moore–Penrose inverse of D, and set

$$A^{(p)} = \begin{pmatrix} A_{00} & \cdots & A_{0p} \\ A_{10} & \cdots & A_{1p} \\ \vdots & \cdots & \vdots \\ A_{q0} & \cdots & A_{qp} \end{pmatrix} \equiv (A_0^{(p)} \,:\, A_1^{(p)}), \tag{10}$$

where $A_0^{(p)}$ consists of the first $\frac{1}{2}m(m+1)$ columns of $A^{(p)}$ with blocks $A_{j0} = \sum_{l=1}^{p-j} [\Phi_{l+j}' \Sigma^{-1} \otimes \Phi_l' \Sigma^{-1}] D$, $j = 0, \ldots, q$, and $A_1^{(p)}$ consists of the rest of the columns with the (j,k) block given by

$$A_{jk} = \begin{cases} \Phi_{k+j}' \Sigma^{-1} \otimes I_m, & k = 1, \ldots, j, \\ \Phi_{k+j}' \Sigma^{-1} \otimes I_m + I_m \otimes \Phi_{k-j}' \Sigma^{-1}, & k = (j+1), \ldots, (p-j), \\ I_m \otimes \Phi_{k-j}' \Sigma^{-1}, & k = (p-j+1), \ldots, p, \end{cases}$$

for $j = 0, \ldots, q$. Also let for each p

$$V_p = A^{(p)} \begin{bmatrix} 2\, D^+(\Sigma \otimes \Sigma) D^{+\prime} & 0 \\ 0 & \Gamma_p^{-1} \otimes \Sigma \end{bmatrix} A^{(p)\prime}, \tag{11}$$

DOI: 10.1111/jtsa.12276

where Γ_p is the covariance matrix of p consecutive observations $(X_t, X_{t+1}, \ldots, X_{t+p-1})$ from a VAR(p) with coefficients $[\Phi_1, \Phi_2, \ldots, \Phi_p]$ and error variance Σ.

Theorem 1. Let $\{X_t\}$ be the VMA(q) process defined in (1), where the $\{\epsilon_t\}$ are i.i.d. with components $\epsilon_{t,i}$. Let $(\Phi_1, \ldots, \Phi_p)$ be the coefficients of the VAR(p) approximation to $\{X_t\}$ with innovation variance Σ. Assume

(A1): $\exists\, C > 0$ such that $\mathrm{E}|\epsilon_{t,i}\epsilon_{t,j}\epsilon_{t,k}\epsilon_{t,l}| \le C < \infty$ for all $1 \le i, j, k, l \le m$.

(A2): The order of the VAR(p) approximation p is chosen as a function of T such that $p^3/T \to 0$ as $p, T \to \infty$.

(A3): The order p also satisfies $T^{1/2} \sum_{j=p+1}^{\infty} \|\Pi_j\| \to 0$ as $p, T \to \infty$, where $\{\Pi_j\}_0^\infty$ is the sequence of coefficient matrices for the VAR(∞) representation of $\{X_t\}$.

(A4): $V = \lim_{p\to\infty} V_p$ exists.

Then

$$T^{1/2}\left[\operatorname{vec}\widehat{\Theta}_{1:q} - \operatorname{vec}\tilde{\Theta}_{1:q}\right] \overset{\mathcal{L}}{\Longrightarrow} \mathcal{N}\left(0, B\,V\,B'\right),$$

where

$$B = \left[\left(\tilde{\Xi}_q^{\dagger} \otimes K^{-1}\,\tilde{\Xi}_{1:q}\,\tilde{\Xi}_q^{-1}\right)(J, 0 \otimes I_m^2) - \begin{pmatrix} 0 \otimes I_m^2 \\ \tilde{\Xi}_q^{\dagger} \otimes I_m \end{pmatrix}\right], \tag{12}$$

$\tilde{\Xi}_q, \tilde{\Xi}_{1:q}$ are functions of the PTV for the inverse autocovariances as defined in (8), and J is a transposition matrix such that $\operatorname{vec}\Xi_q = J \operatorname{vec}\Xi_{0:q}$.

In the case that the VAR(∞) coefficients decrease rapidly, e.g., $\|\Pi_j\| \le C\,\rho^j$ for some $C > 0$ and $\rho \in (0, 1)$, then (A3) is satisfied by any choice of p such that $\log T/p \to 0$; such a condition would be compatible with (A2). The existence of V in (A4) follows from the discussion given in Lewis and Reinsel (1985, p. 401).

As a practical issue, how is p to be selected? Implementation of both the proposed method and the WOLD method relies on modeling the original process by a higher order VAR(p) model, and the choice of p will impact the performance of the estimators. For the asymptotic results, p should satisfy assumptions (A2) and (A3), where (A2) puts an upper bound on the growth rate of p and (A3) puts a lower bound on the rate. Of course, in a finite sample we have to use other criteria to choose reasonable values of p. We investigated the issue of choosing p in such a manner that the resulting estimator has superior performance.

In Galbraith *et al*. (2002), the authors chose p using the Akaike information criterion (AIC). However, given that we have additional information about the true model, i.e., it is a VMA(q), it may be possible to improve upon the model choice. To this end, we considered choosing p such that the fitted value of the criterion (6), where f^{-1} has been estimated as the inverse spectrum of a VAR(p) model, is the smallest over different choices of p. We compared the performance of the proposed estimator obtained using this choice of p with the one where the initial VAR(p) fit is based on AIC. No one choice performed significantly better than the other, and hence we advocate using AIC, and this choice is reflected in our simulations and data analysis below.

4. NUMERICAL RESULTS

We next compare the finite sample performance of the HR, WOLD, and IKL methods based on several VMA processes. We measure performance in terms of the root mean squared error (RMSE), even though the objective function $\mathcal{H}$ is not directly associated with parameter MSE. We also compare the speed of each method and tabulate

DOI: 10.1111/jtsa.12276

the percentage of time the HR and WOLD methods (without stabilization) produce non-invertible estimates. We first discuss the processes, and then summarize the results.

4.1. Simulation Processes

4.1.1. Bivariate VMA(1)
We consider the simplest model in the VMA(q) class, namely a bivariate VMA(1) process:

$$X_t = \epsilon_t + \begin{pmatrix} \vartheta & 0 \\ 1 & .8 \end{pmatrix} \epsilon_{t-1}, \tag{13}$$

where $\epsilon_t \overset{\text{iid}}{\sim} \text{N}(0, I_2)$. The parameter that we vary is the upper left entry of Θ_1, which has eigenvalues ϑ and 0.8. The values of ϑ are chosen from the set $\{-0.99, -0.95, -0.8, -0.5, 0, 0.5, 0.8, 0.95, 0.99\}$, which ensures that the true process is invertible.

4.1.2. Bivariate VMA(2)
Let $X_{t,k}$ denote the kth component of the vector X_t. The bivariate VMA(2) process is defined via

$$\begin{pmatrix} X_{t,1} \\ X_{t,2} \end{pmatrix} = \begin{pmatrix} \epsilon_{t,1} \\ \epsilon_{t,2} \end{pmatrix} + \begin{pmatrix} -\vartheta & 0 \\ -1 & -0.4 \end{pmatrix} \begin{pmatrix} \epsilon_{t-1,1} \\ \epsilon_{t-1,2} \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & -0.45 \end{pmatrix} \begin{pmatrix} \epsilon_{t-2,1} \\ \epsilon_{t-2,2} \end{pmatrix}, \tag{14}$$

where the errors ϵ_t are $\overset{\text{iid}}{\sim} \text{N}(0, I_2)$. This particular parameterization provides a one-dimensional parameterization in terms of one of the roots, ϑ, of the VMA(2) process. This is convenient for illustrating the performance of the estimators as the process changes from invertible to nearly non-invertible. The other roots of the VMA polynomial are 0.9, 0 and -0.5. The scenarios considered are $\vartheta \in \{-0.99, -0.95, -0.8, -0.5, 0, 0.5, 0.8, 0.95, 0.99\}$.

4.1.3. Trivariate VMA(1)
The trivariate VMA(1) process is

$$X_t = \epsilon_t + \begin{pmatrix} \vartheta & 0 & 0 \\ .1 & .5 & 0 \\ 1 & .4 & .8 \end{pmatrix} \epsilon_{t-1}, \tag{15}$$

where $\epsilon_t \overset{\text{iid}}{\sim} \text{N}(0, I_3)$. The parameter that we vary is the upper left entry of Θ_1, which has eigenvalues ϑ, 0.5 and 0.8. The values of ϑ are chosen from the set $\{-0.99, -0.95, -0.8, -0.5, 0, 0.5, 0.8, 0.95, 0.99\}$, which ensures that the true process is invertible.

4.1.4. Higher Dimensional VMA
We also simulated processes with a higher number of parameters. In particular, we chose $m = 15$ dimensional VMA(2) processes and $m = 25$ dimensional VMA(1) processes. Given that there are a large number (570 and 925, respectively) of parameters in either of these settings, we used only larger sample sizes ($T = 200, 500$) for the simulation. For each setting we generated nine random scenarios. The distribution of the roots for the nine cases in each of these settings are given in Figures 1 and 2.

4.2. Simulation Performance

We generated $M = 1000$ Gaussian time series of length $T = 50, 200$, and 500 for each of the simulation models, and evaluated each of the three methods, recording the RMSE of the coefficient matrices and the innovation variance. We also recorded the average runtime for each of the methods. We did not employ the stabilization algorithm

DOI: 10.1111/jtsa.12276

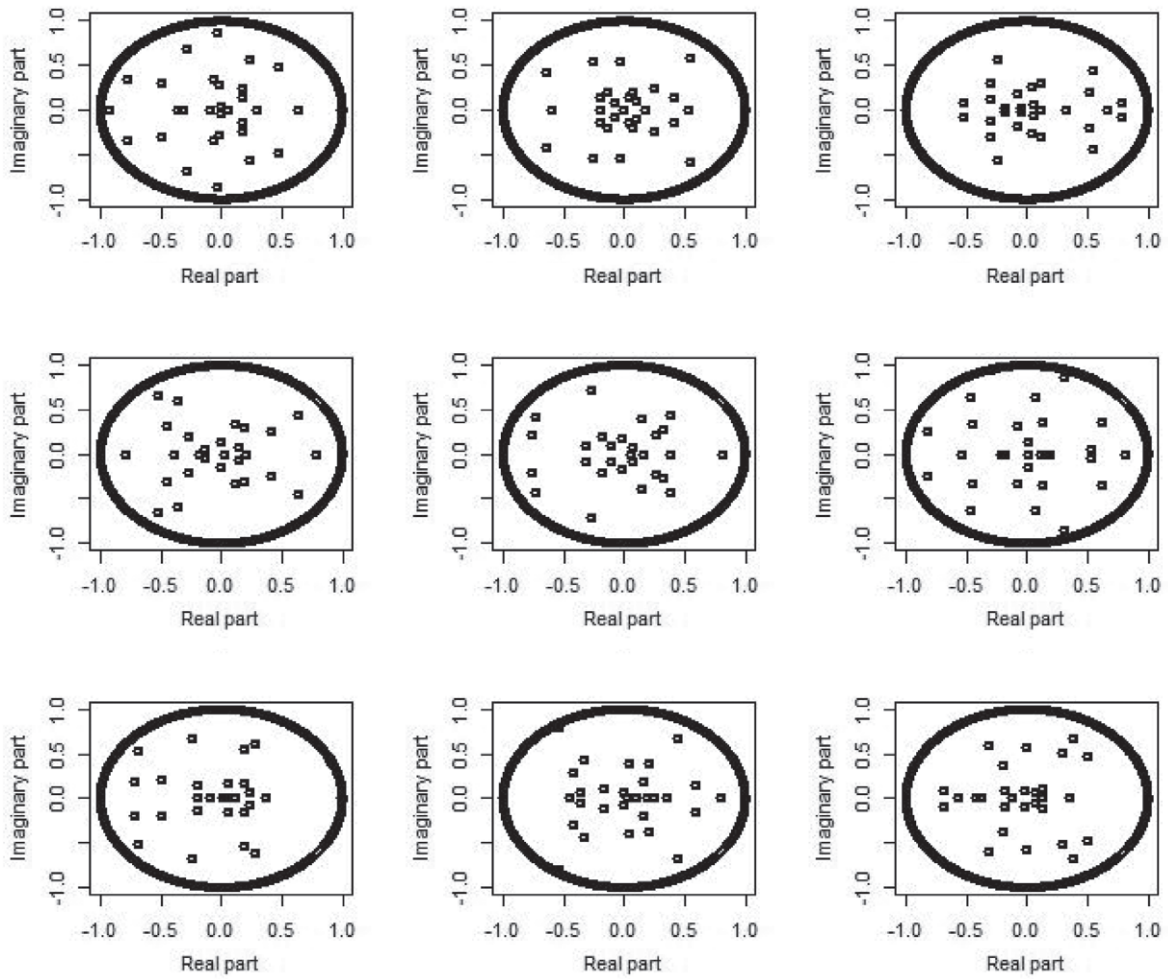

Figure 1. Distribution of the roots associated with the VMA polynomial for nine randomly generated cases for invertible 15-dimensional VMA(2) processes. The roots are shown as complex conjugate pairs with the unit circle

for the HR and WOLD estimates, as this would distort the straight assessment of their accuracy and also add to their runtime (quite significantly, in some cases where a simulation has near-unit roots). IKL enforces stability automatically, with no sacrifice to runtime, while maintaining a competitive estimation accuracy. Let the Monte Carlo MSE for the coefficient matrices of the VMA(q) polynomial be

$$M^{-1}\sum_{j=1}^{M}\sum_{k=1}^{q}\|\widehat{\Theta}_k^{(j)}-\Theta_k\|^2,$$

where $\|\cdot\|$ is the Frobenius norm of a matrix, and $\widehat{\Theta}_k^{(j)}$ is the jth Monte Carlo estimate of Θ_k. For a single entry in a coefficient matrix, the Monte Carlo MSE is defined as the average square distance of the estimated value to the true value, where the average is over M Monte Carlo replications. We examined the RMSE as a function of ϑ.

Table B.1 of Appendix B (Supporting Information) shows the RMSE of the three estimators for the coefficient matrix Θ_1 and the innovation variance Σ for $T = 50$, 200, and 500. It also shows for each method the percentage of Monte Carlo estimates that were invertible. Since the IKL method is theoretically guaranteed to be invertible, its

DOI: 10.1111/jtsa.12276

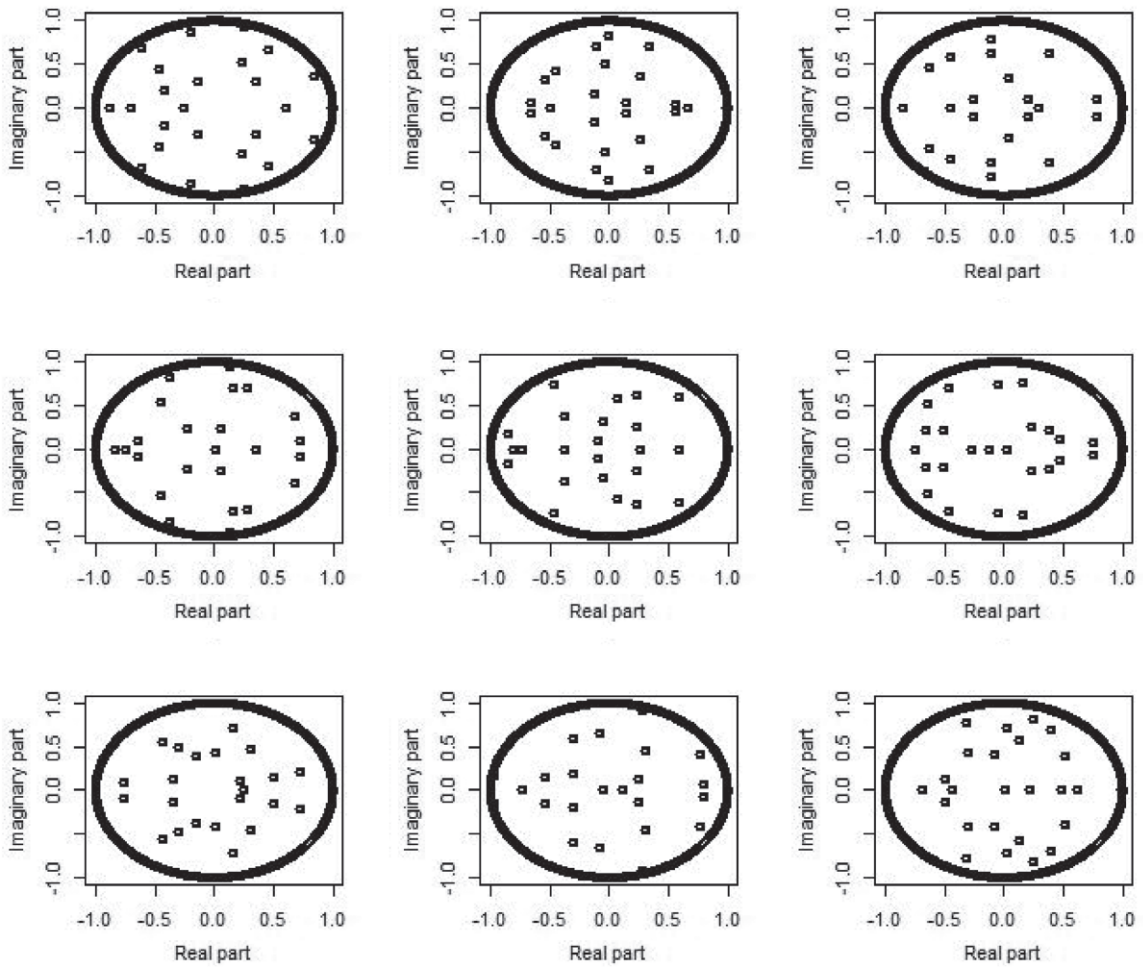

Figure 2. Distribution of the roots associated with the VMA polynomial for nine randomly generated cases for invertible 25-dimensional VMA(1) processes. The roots are shown as complex conjugate pairs with the unit circle

percentage is not reported. The final three columns of the table show the average runtime of each of the methods for the cases considered.

The overall MSE (aggregation over the estimated coefficients and the estimated error variance) of the estimators are comparable, with the HR method for the coefficient matrices having slightly better efficiency than the WOLD and IKL methods, particularly near the unit-root boundary. However, this gain in efficiency is obtained at the cost of frequently estimating outside the invertible space. This is evident from the column containing the percentage of simulations where HR or WOLD estimated a non-invertible process. Even for the simple case of a two-dimensional VMA(1), there is significant probability, depending on the sample size and the parameter configuration, that the unconstrained methods will estimate processes with roots outside the unit disk. The relative comparison of the methods are similar for the other two low-dimensional cases as well, and they are reported in Tables B.2 and B.3. The runtimes of all three methods are comparable.

Tables B.4 and B.5 report the efficiency, percentage of non-invertible estimates, and runtime for the two moderately large dimensional cases considered. When the processes have roots close to unity, the IKL method gains in efficiency over the other methods. This is due to the large number of roots estimated outside the unit circle by the other two methods. Moreover, the percentage of non-invertible estimates reaches 100% in certain cases

DOI: 10.1111/jtsa.12276

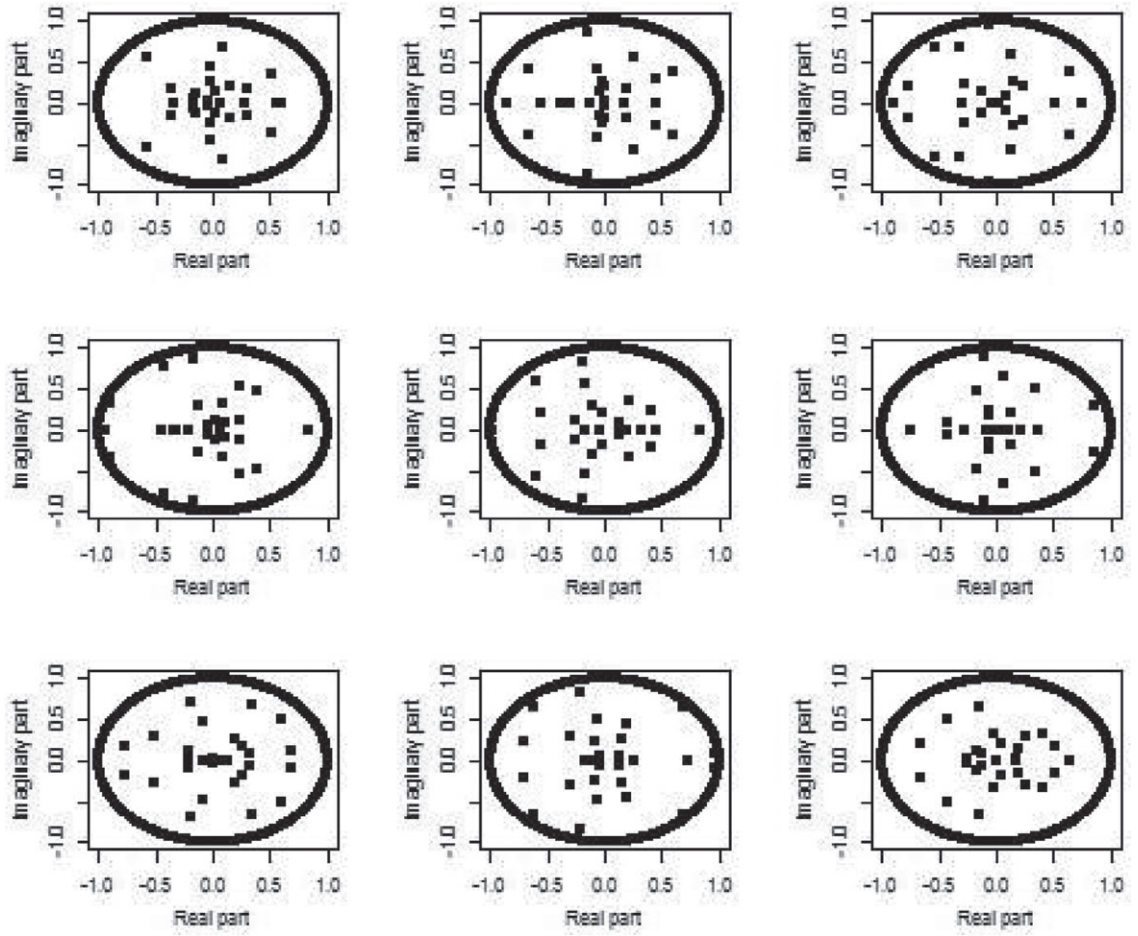

Figure 3. Distribution of the roots associated with the VMA polynomial for nine randomly generated cases for invertible 15-dimensional VMA(2) processes with multivariate skew-*t* errors. The roots are shown as complex conjugate pairs with the unit circle

for the HR and WOLD methods. Therefore, the case for using the IKL method over competitors in the case of high-dimensional time series is fairly strong.

4.3. Effect of Non-Gaussianity and Non-diagonal Σ

To investigate the performance of the proposed method when the time series are non-Gaussian and/or the error covariance matrix is non-diagonal, we generated errors from heavy-tailed asymmetric distributions and compared the performance. To generate heavy-tailed asymmetric marginal distributions we used a multivariate skew-*t* family (Azzalini and Capitanio, 2003). The VMA(2) polynomials were generated as before. The root distribution for the nine cases are given in Figure 3.

To generate errors that mimic the errors in the Gaussian case up to the second moment, we constrained the parameters of the skew-*t* process such that marginals have zero mean and identity covariance matrix. The key parameters (in the notation of Azzalini and Capitanio, 2003) are

$$\bar{\Omega} = (1-\rho)\, I_m + \rho\, 1_m\, 1_m', \qquad \omega = \sigma^2\, I_m, \qquad \xi = 0, \qquad \alpha = \kappa\, 1_m, \tag{16}$$

DOI: 10.1111/jtsa.12276

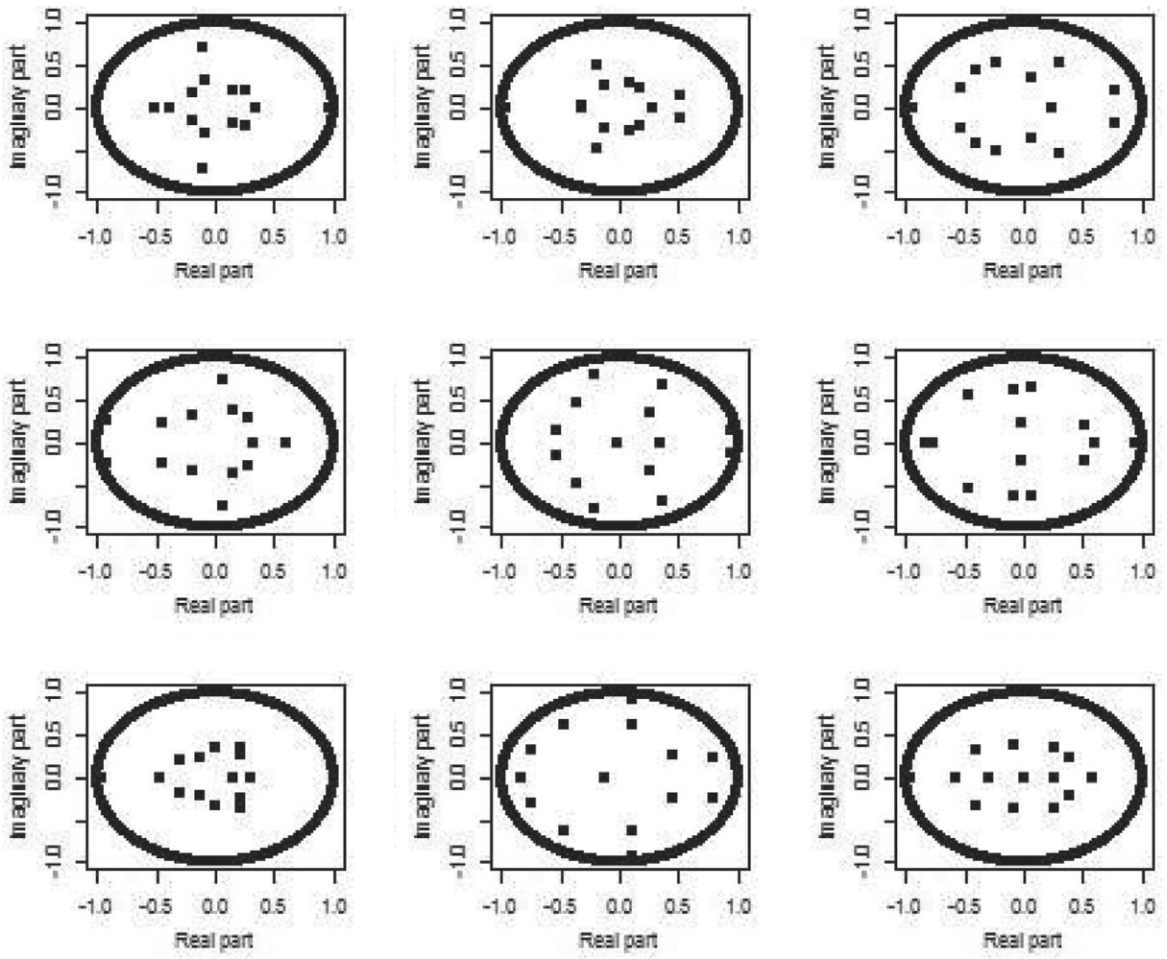

Figure 4. Distribution of the roots associated with the two-dimensional VMA(7) polynomials used in simulation to check the effect of estimation of model order q on the performance of the IKL estimator

where the parameters are constrained to make the variance, $(\frac{\nu}{\nu-2})\Omega - \omega\mu\mu'\omega$, equal to the identity matrix. The constrained values corresponding to the identity give $\sigma^2 = \frac{\nu-2}{\nu(1-\rho)}$, and ρ as the positive root of the quadratic $Ax^2 + Bx + C = 0$ with

$$\begin{aligned} A &= \kappa^2(m-1)[m - c_\nu(m-1)], \\ B &= \kappa^2 m + 1 - 2\kappa^2 c_\nu(m-1), \\ C &= -\kappa^2 c_\nu, \end{aligned}$$

and $c_\nu = \frac{(\nu-2)(\Gamma(\frac{\nu-1}{2}))^2}{\pi(\Gamma(\frac{\nu}{2}))^2}$. For simulation we chose $\nu = 3$ and $\kappa = 5$, resulting in highly skewed heavy-tailed marginals for the errors ϵ_t.

Table B.6 shows the relative performance of the different methods in this non-Gaussian case. From the table we see that the estimation error for the coefficient matrices is not affected a great deal, but the estimation error for the error covariance matrix is generally higher. The relative performance of the different methods remains similar to that obtained in the Gaussian case – with the proposed IKL method performing better than the existing methods.

DOI: 10.1111/jtsa.12276

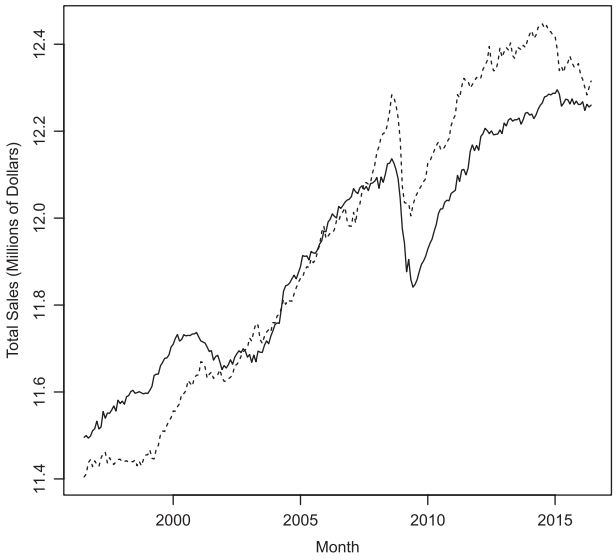

Figure 5. Monthly U.S. total sales (seasonally adjusted) for durable (solid) and nondurable (dashed) goods from July 1996 to June 2016

To generate errors with non-Gaussian and correlated marginals, we simulated using the above parameterization, except that ρ and σ^2 were adjusted ($\rho = 0.8099765$ and $\sigma^2 = 1/\sqrt{2}$) to obtain heavy-tailed skewed marginal distributions with an intraclass correlation structure with correlation ≈ 0.71. Thus, the errors $(\epsilon_{t,1}, \ldots, \epsilon_{t,m})$ were considerably correlated. We used a sample size of $T = 200$; the results are given in Table B.7. From the table we see that proposed method performs much better than the existing methods in this correlated non-Gaussian case.

4.4. Effect of Choosing the Model Order

In real applications, one would not know the correct value of the model order q, and it must be estimated from the data. To investigate the possible loss of efficiency in the proposed method from estimating the model order, we performed a simulation experiment with estimated model order and compared the results to the case where the order is assumed to be known. For each VMA series, we used the function *auto.arima* in the library *forecast* in R to select the MA order for individual components and chose the VMA order for the vector series as the maximum of the chosen order for the individual component series. For each simulation replication, we generated data from two-dimensional VMA(7) models and fitted the IKL method using both the true order ($q = 7$) and estimated order $\hat{q} = \max\{\hat{q}_1, \hat{q}_2\}$, where $\hat{q}_j$ was the chosen order for the jth component series using the *auto.arima* function. We generated nine different VMA(7) models; the distribution of the 14 roots of the VMA polynomials are given in Figure 4. The error covariance matrix was the identity matrix. Since $\hat{q}$ is random and changes with each replication, we compared the one-step-ahead forecast performance of the IKL method with and without the assumption of known model order. Specifically, for each model we looked at the average one-step-ahead forecast error covariance matrix

$$\widehat{\Sigma}_{T+1|1:T} = L^{-1} \sum_{l=1}^{L} (Y_{T+1}^{(l)} - \widehat{Y}_{T+1|1:T}^{(l)})(Y_{T+1}^{(l)} - \widehat{Y}_{T+1|1:T}^{(l)})',$$

where $Y_{T+1}^{(l)}$ and $\widehat{Y}_{T+1|1:T}^{(l)}$ denote the $(T+1)$th observation and the forecast value of the $(T+1)$th observation, based upon observations 1 through T from the lth Monte Carlo sample; L was the total number of Monte Carlo replications, which we set to $L = 5000$.

As a measure of forecast accuracy, we examined the trace and the determinant of $\widehat{\Sigma}_{T+1|1:T}$. Table B.8 gives the values of the trace and determinant of the one-step-ahead forecast error covariance for both known and estimated

DOI: 10.1111/jtsa.12276

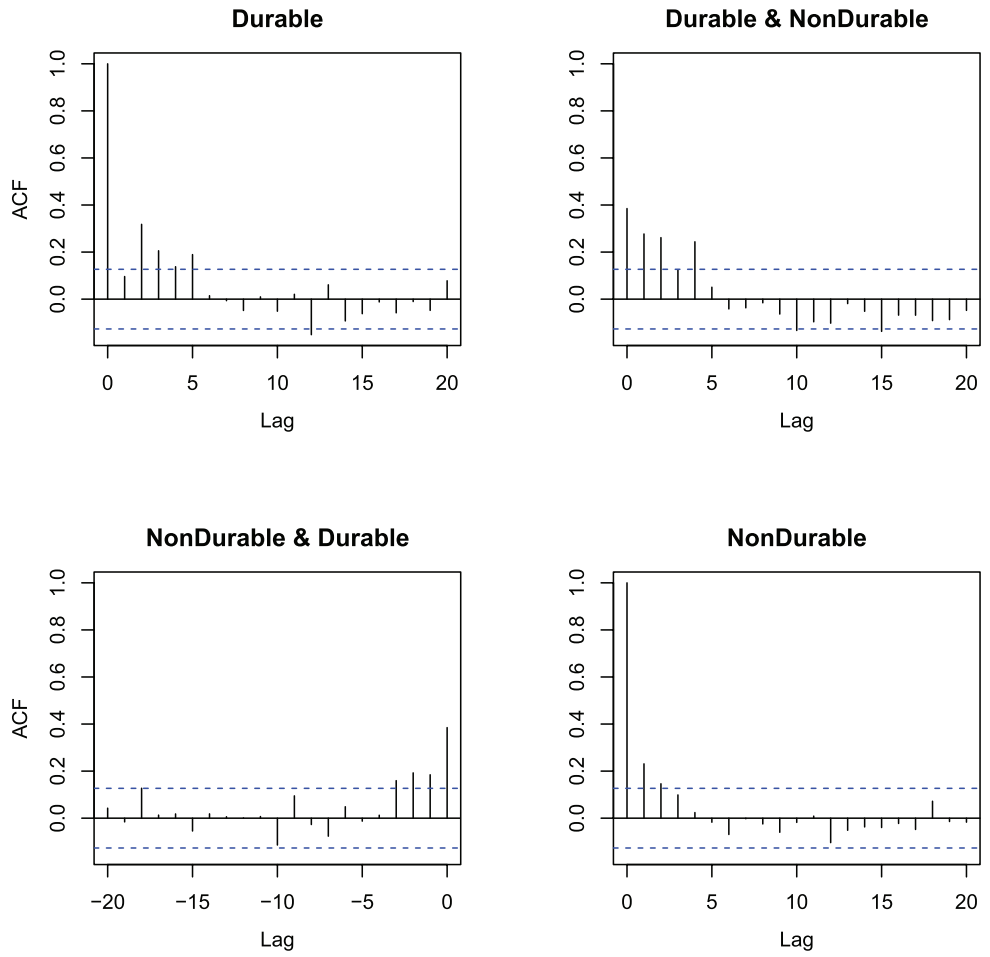

Figure 6. Autocorrelation and cross-correlation plots of first difference of logarithm of durable and nondurable goods sales series (seasonally adjusted). [Color figure can be viewed at wileyonlinelibrary.com]

q for two different sample sizes. As can be seen from the table, the effect of estimation of model order is minimal compared to the effect of estimation of the parameters. For reference, the true value of the trace and the determinant are 2 and 1, respectively. Thus, estimation of the parameters of the model seems to have a larger contribution toward the forecast error as compared to the estimation of the model order.

5. DATA ANALYSIS

We examined the monthly U.S. total sales series (seasonally adjusted) for durable and nondurable goods. The current series are available from the Monthly Wholesale Trade: Sales and Inventories section in http://www.census.gov/econ/currentdata. Galbraith *et al*. (2002) used a similar series (total inventory) to illustrate the WOLD method. We used the last 20 years of the August 9, 2016 vintage for our analysis, yielding start and end months of July 1996 and June 2016. The plot of the series in the original scale is shown in Figure 5. The series show increasing variability over time, and a log transformation was employed to stabilize the variability. Because of the increasing trend over time, we modeled the first differences of the series. The autocorrelation plot of the first difference of the log-transformed series are shown in Figure 6.

DOI: 10.1111/jtsa.12276

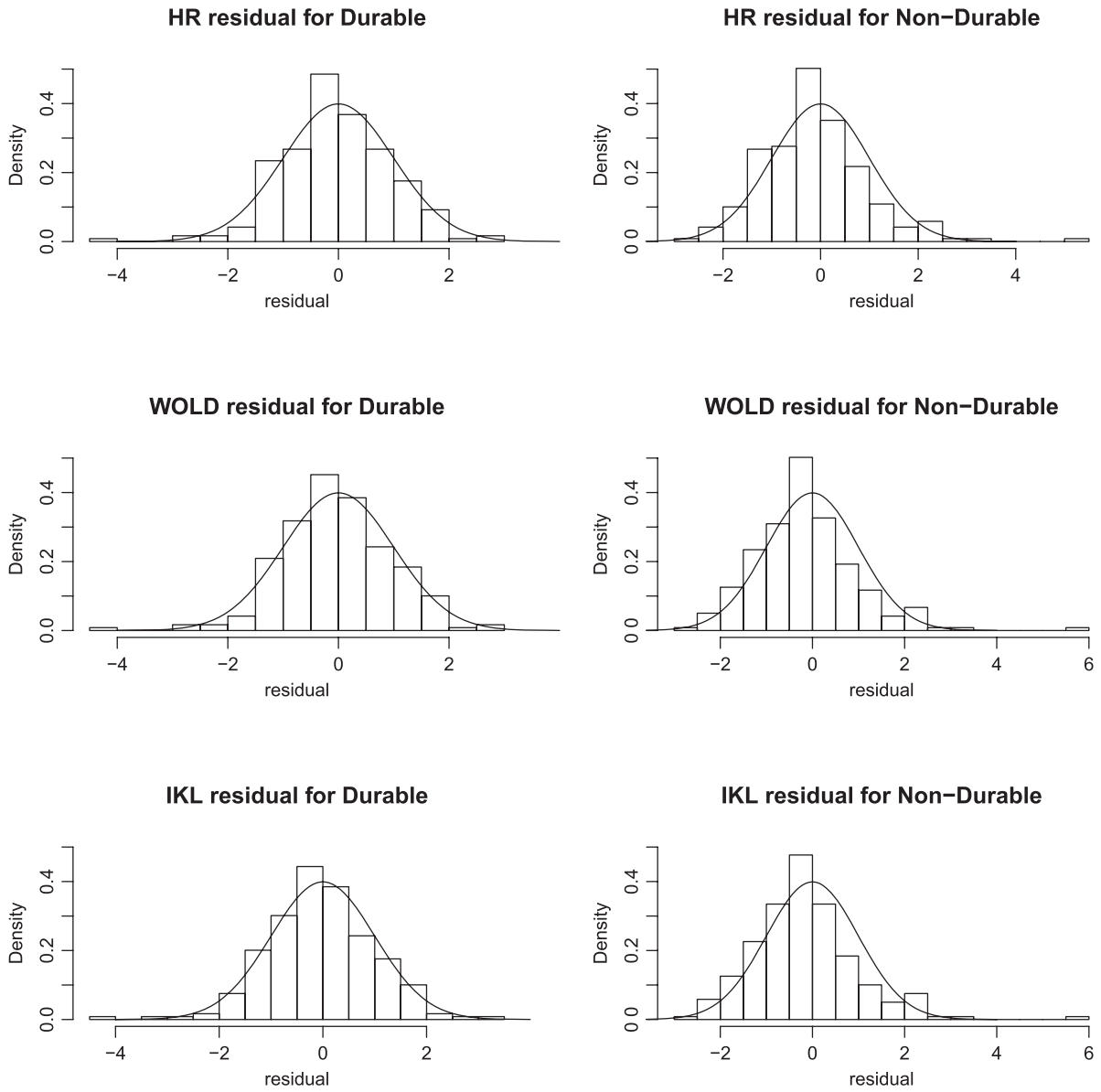

Figure 7. Histogram of residuals from the three different fitting methods for the durable and nondurable goods sales series

The autocorrelation plots (Figure 6) indicate that a low-order MA may be suitable for both series. In addition, the cross-correlation plots show significant cross-correlation up to lag four. Based on these plots, as well as the partial autocorrelation plot (not shown), a VMA(5) model was selected for the bivariate time series of first differences of the logarithm of monthly total sales of durable and nondurable goods. There is some evidence in the literature to support a VMA model for sales series. Dekimpe and Hanssens (1995) used such a model, based upon the argument that the marketing incentives provided by traders reset the correlation frequently, thereby limiting the length of lagged dependence structure. We fit the fifth-order VMA model to the data using the three different methods discussed above (HR, WOLD, IKL). We report the modulus of the roots of the determinental equations for the estimated MA polynomial for each methods (values are in decreasing order and rounded to three decimal

DOI: 10.1111/jtsa.12276

places), with bold values corresponding to roots outside the unit circle.

$$\begin{aligned} \text{HR} &: (\mathbf{1.086}, 0.764, 0.764, 0.685, 0.685, 0.450, 0.450, 0.395, 0.366, 0.366) \\ \text{WOLD} &: (\mathbf{1.023}, 0.788, 0.788, 0.698, 0.698, 0.476, 0.476, 0.382, 0.382, 0.333) \\ \text{IKL} &: (0.807, 0.807, 0.741, 0.741, 0.633, 0.508, 0.508, 0.376, 0.321, 0.321) \end{aligned}$$

All three methods produce an adequate fit to the data. The histograms of the residuals obtained from the Durbin–Levinson algorithm using the estimated model for the three methods are given in Figure 7. The marginal distributions appear to be normal, excepting for two extreme values associated with the big dip in the sales values observed around the start of the Great Recession. All six residual series pass the Box–Ljung test for autocorrelation with p-values ranging from 0.78 to 0.96. However, both the HR and the WOLD methods estimate non-invertible VMA processes. As discussed earlier, the forecast formula based on an infinite past is no longer valid for unstable processes; these applications therefore strongly motivate IKL, or at a minimum applying a stabilization algorithm, to the raw HR (or WOLD) results.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

REFERENCES

Azzalini A, Capitanio A. 2003. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution. *Journal of the Royal Statistical Society. Series B* **65**: 367–389.

Bauer F. 1955. Ein direktes Iterationsverfahren zur Hurwitz-Zerlegung eines Polynoms. *Archiv für Elektronik und Übertragungstechnik* **9**: 285–290.

Brockwell PJ, Davis RA. 1991. *Time Series: Theory and Methods*. New York: Springer Science & Business Media.

Dekimpe MG, Hanssens DM. 1995. The persistence of marketing effects on sales. *Marketing Science* **14**: 1–21.

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

Durbin J. 1959. Efficient estimation of parameters in moving-average models. *Biometrika* **46**: 306–316.

Durbin J. 1960. The fitting of time-series models. *Review of the International Statistical Institute* **28**: 233–244.

Flores de Frutos R, Serrano GR. 2002. A generalized least squares estimation method for VARMA models. *Statistics* **36**: 303–316.

Galbraith JW, Ullah A, Zinde-Walsh V. 2002. Estimation of the vector moving average model by vector autoregression. *Economic Reviews* **21**: 205–219.

Hannan EJ, Deistler M. 1988. *The Statistical Theory of Linear Systems*. New York: John Wiley & Sons.

Hannan EJ, Kavalieris L. 1984. A method for autoregressive-moving average estimation. *Biometrika* **71**: 273–280.

Hannan EJ, Kavalieris L. 1986. Regression, autoregression models. *Journal of Time Series Analysis* **7**: 27–49.

Hannan EJ, Rissanen J. 1982. Recursive estimation of mixed autoregressive-moving average order. *Biometrika* **69**: 81–94. Errata 70 1983, 303.

Harvey A. 1989. *Forecasting, Structural Time Series Models, and the Kalman Filter*. Cambridge: Cambridge University Press.

Hillmer SC, Tiao GC. 1979. Likelihood function of stationary multiple autoregressive moving average models. *Journal of the American Statistical Association* **74**: 652–660.

Holan S, McElroy T, Wu G. 2017. The cepstral model for multivariate time series: the vector exponential model. *Statistica Sinica* **27**: 23–42.

Huang D, Guo L. 1990. Estimation of nonstationary ARMAX models based on the Hannan–Rissanen method. *The Annals of Statistics* **18**: 1729–1756.

Jentsch C, Politis DN. 2015. Covariance matrix estimation and linear process bootstrap for multivariate time series of possibly increasing dimension. *The Annals of Statistics* **43**(3): 1117–1140.

Koreisha SG, Pukkila TM. 1989. Fast linear estimation methods for vector autoregressive moving-average models. *Journal of Time Series Analysis* **10**: 325–339.

Lewis R, Reinsel GC. 1985. Prediction of multivariate time series by autoregressive model fitting. *Journal of Multivariate Analysis* **16**(3): 393–411.

DOI: 10.1111/jtsa.12276

Lütkepohl H. 2007. *New Introduction to Multiple Time Series Analysis*. Berlin-Heidelberg: Springer Science & Business Media.
Lütkepohl H, Claessen H. 1997. Analysis of cointegrated VARMA processes. *Journal of Econometrics* **80**: 223–239.
Lütkepohl H, Poskitt DS. 1996. Specification of echelon-form VARMA models. *Journal of Business and Economic Statistics* **14**: 69–79.
Mauricio JA. 2002. An algorithm for the exact likelihood of a stationary vector autoregressive-moving average model. *Journal of Time Series Analysis* **23**: 473–486.
McElroy TS. 2017. Recursive computation for block-nested covariance matrices. *Journal of Time Series Analysis*. DOI: 10.1111/jtsa.12267.
McElroy TS, Findley D. 2010. Discerning between models through multi-step ahead forecasting errors. *Journal of Statistical Planning and Inference* **140**: 3655–3675.
McElroy T, Findley D. 2015. *Fitting constrained vector autoregression models*. *Empirical Economic and Financial Research*. Cham: Springer International Publishing; 451–470.
McElroy T, McCracken M. 2017. Multi-step ahead forecasting of vector time series. *Econometric Reviews* **36**(5): 495–513.
McMurry TL, Politis DN. 2010. Banded and tapered estimates for autocovariance matrices and the linear process bootstrap. *Journal of Time Series Analysis* **31**(6): 471–482.
Mélard G, Roy R, Saidi A. 2006. Exact maximum likelihood estimation of structured or unit root multivariate time series models. *Computational Statistics and Data Analysis* **50**: 2958–2986.
Mitchell H, Brockwell P. 1997. Estimation of the coefficients of a multivariate linear filter using the innovations algorithm. *Journal of Time Series Analysis* **18**(2): 157–179.
Phillips PC, Sun Y, Jin S. 2006. Spectral density estimation and robust hypothesis testing using steep origin kernels without truncation. *International Economic Review* **47**(3): 837–894.
Poskitt DS. 1992. Identification of echelon canonical forms for vector linear processes using least squares. *The Annals of Statistics* **20**: 195–215.
Roy A, McElroy T, Linton P. 2014. Estimation of causal invertible VARMA models. arXiv:1406.4584 [math.ST].
Sayed A, Kailath T. 2001. A survey of spectral factorization methods. *Numerical Linear Algebra with Applications* **8**: 467–496.
Shea BL. 1989. The exact likelihood of a vector autoregressive moving average model. *Journal of the Royal Statistical Society, Series C Applied Statistics* **38**: 161–184.
Taniguchi M, Kakizawa Y. 2012. *Asymptotic Theory of Statistical Inference for Time Series*. New York: Springer Science & Business Media.
Tiao GC, Box GEP. 1981. Modeling multiple time series with applications. *Journal of the American Statistical Association* **76**: 802–816.
Zadrozny P. 1998. An eigenvalue method of undetermined coefficients for solving linear rational expectations models. *Journal of Economic Dynamics and Control* **22**: 1353–1373.

DOI: 10.1111/jtsa.12276