JOURNAL OF TIME SERIES ANALYSIS
J. Time Ser. Anal. (2017)
Published online in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1111/jtsa.12267

SPECIAL ISSUE

RECURSIVE COMPUTATION FOR BLOCK-NESTED COVARIANCE MATRICES

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

Covariance matrices corresponding to samples of multivariate time series or spatial random fields have a block-Toeplitz structure that has a nested pattern. Also, non-lattice data samples yield nested covariance matrices, although they are no longer block-Toeplitz. The nested structure of such matrices facilitates the computation of their inverses, among other related quantities. Recursive algorithms, based upon this nested structure, are presented, yielding applications such as the simulation of vector time series, the evaluation of Gaussian likelihoods and Whittle likelihoods, the computation of spectral factorization, and the calculation of projections. Both multivariate time series applications and two-dimensional random fields applications are discussed, as well as applications to non-lattice data.

Received 20 December 2016; Accepted 30 August 2017

Keywords: Innovations algorithm; Levinson–Durbin algorithm; likelihood evaluation; spectral factorization; Whittle likelihood.

MOS subject classification: 62m10, 62m15, 62m30.

1. INTRODUCTION

This article develops recursive algorithms for the computation of block-nested covariance matrices that arise in statistical model-fitting. For many applications, such as when handling multivariate time series data or lattice random field data, the covariance matrix of the sample is block-Toeplitz, and this special structure can be utilized to yield fast algorithms for matrix inversion. In other types of applications, the covariance matrix may not be block-Toeplitz, though it will still be block-nested and block non-negative definite. The modified Cholesky decomposition (MCD) described in Newton and Pagano (1983) is essentially a block version of the well-known Cholesky–Banachiewicz algorithm, which is also described in Golub and Van Loan (1996). This article describes the MCD and makes several applications, including likelihood evaluation and projections.

The MCD is used in Newton and Pagano (1983) to obtain a faster way to compute predictors and likelihoods for stationary (univariate) processes, with a focus on ARMA processes. The basic concept is to write a block-nested matrix Γ as $L\,D\,L'$, with L unit block lower triangular and D block diagonal (from now on, we suspend the adjective 'block', and assume it throughout the paper). If we iterate on the dimension of the matrices, an elegant nested structure emerges from the nested form; this allows one to recursively compute the final row of L from the new entries (e.g. in the time series case these are the autocovariances) and previous iterates. This recursion follows easily from the derivations in Newton and Pagano (1983) and does not rely on an ARMA assumption. Here we develop this structure into so-called forward and backward decompositions, which allow rapid calculation of Γ^{-1}. Furthermore, the quantities arising from the MCD can be interpreted as partial covariances, suggesting a natural statistical parameterization.

* 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

These recursions are useful in statistical applications that involve vector time series or spatial data. This work supplies a few gaps in the literature: for time series not embeddable in state space (e.g. long memory processes), or for spatial data, it is important for practitioners to be able to compute likelihoods and projections in a stable and efficient manner. For such processes, the new recursions render applications such as forecasting and signal extraction feasible. As for processes that are embeddable in state space (see Durbin and Koopman 2001 for an overview), neither the Whittle likelihood nor spectral factorization can be computed from state space algorithms, and an alternative approach such as that advocated by this article is required. A potential advantage of this article's approach over state space algorithms occurs when there are many latent structures present in the process, such that the state vector becomes very long – the order of operations in the Kalman filter depends on the dimension of the state vector. In contrast, the projection approach of this article depends only on sample size and time series dimension, assuming that the process' autocovariances have been computed.

While the forward and backward decompositions of Section 2 are related to the finite forward and backward predictors encountered in time series literature (see Haykin 1996 Chapter 6), our treatment is slightly more general, allowing for non-Toeplitz matrices — this allows for broader applications to spatial data or non-lattice covariance structures. The basic recursions are discussed in Section 2, and they are applied in Section 3 to generate algorithms for simulation, likelihood evaluation, and spectral factorization for multivariate time series. Applications to spatial data are discussed at the end of Section 3, and Section 4 gives a few numerical illustrations. Appendix S1, Supporting Information contains proofs and an extensive treatment of the application of signal extraction – as well as R code for the Gaussian likelihood algorithm.

2. NESTED MATRIX RECURSIONS

We consider matrices that have a block-nested structure, and review the recursive algorithms for calculation of the MCD for both the matrix and its inverse. These recursions help us to illustrate the nested structure of the Cholesky factors, which is further developed through various applications in Section 3. There are two versions of this, the lower and upper versions, depending on whether we take the Cholesky factors in the MCD to be either lower triangular or upper triangular respectively. Early literature on such decompositions includes Bauer (1955), Akaike (1973), and Pagano (1976), followed by Newton and Pagano (1983). The innovations algorithm of Brockwell and Davis (1991) is related to the MCD approach – this is discussed in Section 3.2.

A lower nested matrix is defined to be a sequence of block matrices Γ_t with structure

$$\Gamma_{t+1} = \begin{bmatrix} \Gamma_t & \underline{\gamma}_t \\ \underline{\gamma}_t' & g_{t+1} \end{bmatrix}. \tag{1}$$

We suppose that Γ_t is a $tm \times tm$-dimensional block matrix, and the block matrix $\underline{\gamma}_t$ is $tm \times m$-dimensional, while g_{t+1} is $m \times m$-dimensional.

For example, for a block-Toeplitz matrix corresponding to a covariance stationary vector time series process $\{x_t\}$ $\underline{\gamma}_t' = [\gamma(t), \ldots, \gamma(1)]$ and $g_{t+1} = \gamma(0)$, where $\gamma(h)$ is the autocovariance function of the process defined via $\gamma(h) = \mathrm{Cov}(x_{t+h}, x_t)$. In this case, $\underline{\gamma}_{t-1}$ is identical with the latter $t-1$ block elements of $\underline{\gamma}_t$; however, there are cases of interest where Γ_t is not Toeplitz, so that $\underline{\gamma}_{t-1}$ and $\underline{\gamma}_t$ have no redundancy in their elements. This can happen if the time series is nonstationary or is irregularly sampled (see Section 3.6 for further discussion).

Throughout this section and the paper, 1 will denote an identity matrix of dimension corresponding to the context, and 0 will denote a matrix of zeroes of appropriate dimension. Usually, these matrices are $m \times m$-dimensional. A lower triangular matrix with block diagonals given by the identity matrix is said to be unit lower triangular. An upper nested matrix is defined similar to (1), but now with the formula

$$\Gamma_{t+1} = \begin{bmatrix} g_{t+1} & \overline{\gamma}_t' \\ \overline{\gamma}_t & \Gamma_t \end{bmatrix}. \tag{2}$$

DOI: 10.1111/jtsa.12267

This $\overline{\gamma}_t$ is also $tm \times m$-dimensional, and in the case of a stationary vector time series it would correspond to $\overline{\gamma}_t' = [\gamma(-1), \ldots, \gamma(-t)]$, in contrast to $\underline{\gamma}_t'$.

The MCD for the lower nested matrices is $\Gamma_t = L_t D_t L_t'$ for unit block lower triangular matrices L_t and block diagonal matrices D_t, whereas the MCD for the upper nested matrices is $\Gamma_t = U_t C_t U_t'$ for unit upper triangular matrices U_t and block diagonal matrices C_t. First we describe the nested structure of these MCDs and how the lower and upper nested matrices can be computed recursively.

Proposition 1. A lower nested matrix has MCD given by $\Gamma_t = L_t D_t L_t'$, where the unit block lower triangular matrix L_t and block diagonal matrix D_t have the form (for $t \geq 1$)

$$L_{t+1} = \begin{bmatrix} L_t & 0 \\ \ell_{t+1}' & 1 \end{bmatrix} \qquad D_{t+1} = \begin{bmatrix} D_t & 0 \\ 0 & d_{t+1} \end{bmatrix}.$$

The new entries are ℓ_{t+1} and d_{t+1}, and they are updated via the equations

$$\ell_{t+1} = D_t^{-1} L_t^{-1} \underline{\gamma}_t \qquad d_{t+1} = g_{t+1} - \ell_{t+1}' D_t \ell_{t+1}.$$

An upper nested matrix has MCD given by $\Gamma_t = U_t C_t U_t'$, where the unit block upper triangular matrix U_t and block diagonal matrix C_t have the form (for $t \geq 1$)

$$U_{t+1} = \begin{bmatrix} 1 & u_{t+1}' \\ 0 & U_t \end{bmatrix} \qquad C_{t+1} = \begin{bmatrix} c_{t+1} & 0 \\ 0 & C_t \end{bmatrix}.$$

The new entries are u_{t+1} and c_{t+1}, and they are updated via the equations

$$u_{t+1} = C_t^{-1} U_t^{-1} \overline{\gamma}_t \qquad c_{t+1} = g_{t+1} - u_{t+1}' C_t u_{t+1}.$$

The proof of this result follows directly from writing out the block matrices and matching appropriate expressions. The initialization for the lower MCD is $L_1 = 1$ and $D_1 = g_1 = \Gamma_1$, whereas the initialization for the upper MCD is $U_1 = 1$ and $C_1 = g_1 = \Gamma_1$. We can also modify these MCDs such that the triangular matrices are no longer unit, with the square root of the diagonal entries of the block diagonal matrices featured on the diagonals.

These recursions have application to spectral factorization, to the computation of the covariance matrix of a vector autoregressive moving average (VARMA) process (Lütkepohl 2007), and to the calculation of signal extraction filters. However, it is also useful to have recursions to compute the inverses of nested matrices, as these are used to calculate Gaussian likelihoods and projections (such as forecasts).

Introduce the notation $A^\dagger = A'^{-1}$ for the inverse transpose of A. If a nested matrix has a lower MCD $\Gamma_t = L_t D_t L_t'$, then its inverse has a lower MCD as well, given by $\Gamma_t^{-1} = L_t^\dagger D_t^{-1} L_t^{-1}$, although the inverse matrix is no longer nested. Similarly, the upper MCD for the upper nested matrix is $\Gamma_t^{-1} = U_t^\dagger C_t^{-1} U_t^{-1}$. The next result shows that, if we have calculated the MCD of Γ_t recursively, we can obtain Γ_t^{-1} fairly easily.

Proposition 2. A lower nested matrix has MCD for its inverse given by $\Gamma_t^{-1} = L_t^\dagger D_t^{-1} L_t^{-1}$, where the unit block lower triangular matrix L_t^{-1} and block diagonal matrix D_t^{-1} have the form (for $t \geq 1$)

$$L_{t+1}^{-1} = \begin{bmatrix} L_t^{-1} & 0 \\ -\tilde{\ell}_{t+1}' & 1 \end{bmatrix} \qquad D_{t+1}^{-1} = \begin{bmatrix} D_t^{-1} & 0 \\ 0 & d_{t+1}^{-1} \end{bmatrix}.$$

The new entries are $\tilde{\ell}_{t+1}$ and d_{t+1}^{-1}, and they are updated via

$$\tilde{\ell}_{t+1} = L_t^\dagger \ell_{t+1} = \Gamma_t^{-1} \underline{\gamma}_t \qquad d_{t+1} = g_{t+1} - \underline{\gamma}_t' \tilde{\ell}_{t+1},$$

DOI: 10.1111/jtsa.12267

so long as d_{t+1} is invertible. The inverse matrix does not have a lower nested form, but can be updated via (for $t \geq 1$)

$$\Gamma_{t+1}^{-1} = \begin{bmatrix} \Gamma_t^{-1} + \tilde{\ell}_{t+1}\, d_{t+1}^{-1}\, \tilde{\ell}'_{t+1} & -\tilde{\ell}_{t+1}\, d_{t+1}^{-1} \\ -d_{t+1}^{-1}\, \tilde{\ell}'_{t+1} & d_{t+1}^{-1} \end{bmatrix}. \tag{3}$$

An upper nested matrix has MCD for its inverse given by $\Gamma_t^{-1} = U_t^{\dagger}\, C_t^{-1}\, U_t^{-1}$, where the unit block upper triangular matrix U_t^{-1} and block diagonal matrix C_t^{-1} have the form (for $t \geq 1$)

$$U_{t+1}^{-1} = \begin{bmatrix} 1 & -\tilde{u}'_{t+1} \\ 0 & U_t^{-1} \end{bmatrix} \qquad C_{t+1}^{-1} = \begin{bmatrix} c_{t+1}^{-1} & 0 \\ 0 & C_t^{-1} \end{bmatrix}.$$

The new entries are $\tilde{u}_{t+1}$ and c_{t+1}^{-1}, and they are updated via (for $t \geq 1$)

$$\tilde{u}_{t+1} = U_t^{\dagger}\, u_{t+1} = \Gamma_t^{-1}\, \overline{\gamma}_t \qquad c_{t+1} = g_{t+1} - \overline{\gamma}'_t\, \tilde{u}_{t+1},$$

so long as c_{t+1} is invertible. The inverse matrix does not have an upper nested form, but can be updated via

$$\Gamma_{t+1}^{-1} = \begin{bmatrix} c_{t+1}^{-1} & -c_{t+1}^{-1}\, \tilde{u}'_{t+1} \\ -\tilde{u}_{t+1}\, c_{t+1}^{-1} & \Gamma_t^{-1} + \tilde{u}_{t+1}\, c_{t+1}^{-1}\, \tilde{u}'_{t+1} \end{bmatrix}. \tag{4}$$

Remark 1. The quantities d_{t+1} and c_{t+1} are Schur complements (of Γ_{t+1} with respect to Γ_t – see Axelsson 1996) in the lower and upper nested cases respectively. For block-Toeplitz matrices that are block positive definite, such Schur complements will be positive definite themselves, and hence invertible. More generally, the block positive definite property of covariance matrices follows from Mercer's Theorem (Tanaka 1996); see further discussion in Section 3.6.

From (3), it is evident that a knowledge of d_{t+1} and $\tilde{\ell}_{t+1}$ is sufficient to generate the new Γ_{t+1}^{-1} from the previous Γ_t^{-1}. Similarly, from (4) we can compute Γ_{t+1}^{-1} if we know c_{t+1} and $\tilde{u}_{t+1}$. The next section considers an application that intertwines both the lower and upper MCDs to yield an algorithm for computing the Gaussian likelihood, which in the case of block-Toeplitz covariance matrices is identical with the Levinson–Durbin algorithm.

We can also provide a statistical interpretation of the quantities L and D in the MCD; this is similar to the known fact about the entries of Γ^{-1} being partial covariances, but is actually a different result. The proposition is stated for a lower nested MCD, but a similar result also holds for the upper nested MCD. We suppose that Γ_t is the covariance matrix of some random vector x, and the individual random vectors are denoted x_j for $1 \leq j \leq t$. Let $\mathrm{Cov}_Z(x, y) = \mathrm{Cov}(x - P_Z(x), y - P_Z(y))$ denote the partial covariance, given conditioning variables Z, where $P_Z(x)$ denotes the linear projection of x onto Z, i.e. $P_Z(x) = \mathrm{Cov}(x, Z)\, \mathrm{Var}(Z)^{-1}\, Z$.

Proposition 3. In the lower nested MCD, the Schur complement d_j is equal to $\mathrm{Cov}_{x_1,\ldots,x_{j-1}}(x_j, x_j)$. Also, when d_j^{-1} exists, the kjth block entry of the lower Cholesky factor L is given by the transpose of $d_j^{-1}\, \mathrm{Cov}_{x_1,\ldots,x_{j-1}}(x_j, x_k)$.

Because the MCD is a block decomposition, each partial covariance is actually a matrix. When $k = j$, we obtain the identity matrix as the block diagonal of L, and when $k < j$ the partial covariance is zero (because in this case x_k is included in the conditioning set). This new result can be constrasted with the known result that the precision matrix Γ_t^{-1} has block entries given by partial covariances, where the conditioning is on *all* the variables; Proposition 3 describes the entries of L as (scaled) sequential partial covariances.

This interpretation can be useful when modeling multivariate data, and one wishes to impose some sparsity on Γ_t – in this case, placing a zero in L_t corresponds to imposing that a partial covariance is zero. Moreover, it suggests a parameterization of Γ_t that is useful for model fitting. Allowing the lower block entries of L to be

DOI: 10.1111/jtsa.12267

generic real matrices, and parameterizing the Schur complements d_j as symmetric positive definite matrices, we find by Proposition 3 that the product of such is interpretable as a partial covariance. McElroy (2017) provides an application of this parameterization to model seasonally co-integrated economic data.

3. APPLICATIONS OF THE RECURSIONS

3.1. Simulation of Multivariate Time Series

Here we present a known algorithm for generating simulations of a Gaussian m-variate time series, with a sample of length T. If the autocovariance function for the generating model is known, we can compute the entries of Γ_T. If Z_T is a Tm-vector of i.i.d. normal random variables, then $\Gamma_T^{1/2} Z_T$ will be a normal random vector with covariance matrix Γ_T, and hence is a simulation of the time series. Here, our interest focuses on quickly computing the matrix square root $\Gamma_T^{1/2}$.

Let Z_t be the tm-vector, for all $1 \leq t \leq T$, corresponding to the first t block-components of Z_T, and let z_j denote an individual element for $1 \leq j \leq T$. So $z_1, \ldots, z_T$ are i.i.d. normal random vectors with covariance matrix 1_m. Similarly, let $X_T = \Gamma_T^{1/2} Z_T$, which consists of individual elements x_j, and sub-vectors X_t. Then

$$Z_{t+1} = \begin{bmatrix} Z_t \\ z_{t+1} \end{bmatrix} \qquad X_{t+1} = \begin{bmatrix} X_t \\ x_{t+1} \end{bmatrix},$$

and utilizing the formula for L_{t+1} in Proposition 1 prompts the definition $X_t = L_t D_t^{1/2} Z_t$ – it is immediate that the covariance matrix of X_t for each $1 \leq t \leq T$ is Γ_t. It then follows that

$$X_{t+1} = L_{t+1} D_{t+1}^{1/2} Z_{t+1} = \begin{bmatrix} X_t \\ \tilde{\ell}'_{t+1} X_t + d_{t+1}^{1/2} z_{t+1} \end{bmatrix},$$

so that the sequence of simulations at time t, namely X_t, gets appended with the new value $\tilde{\ell}'_{t+1} X_t + d_{t+1}^{1/2} z_{t+1}$. This derivation uses $\ell'_{t+1} Z_t = \tilde{\ell}'_{t+1} L_t D_t^{1/2} Z_t = \tilde{\ell}'_{t+1} X_t$. The step requires order of t calculations (ignoring the impact of m) in general, and the computation of $\tilde{\ell}'_{t+1} X_t$ is discussed in the next section .

3.2. Calculation of Gaussian Likelihoods

Here we present a Gaussian likelihood algorithm, which is essentially the same as the Levinson–Durbin algorithm and shares the form of the innovations algorithm in Brockwell and Davis (1991) – where one computes a Gaussian likelihood by subtracting from an observation x_{t+1} its best linear predictor $\tilde{\ell}'_{t+1} X_t$ and square, normalizing by the prediction error variance d_{t+1}. This topic has been well studied (e.g. see Whittle 1963), and illustrates a key application of the MCD.

Given a sample $\{x_1, x_2, \ldots, x_T\}$ from a stationary m-variate time series, a time series model can be fitted via maximizing the Gaussian likelihood. Let $X_t = \mathrm{vec}\{x_1, x_2, \ldots, x_t\}$. Once the values for the parameters have been selected, the model's autocovariance sequence $\{\gamma(h)\}$ can be calculated for $-T < h < T$, and the Toeplitz matrix Γ_T can be computed, where by definition the jkth $m \times m$ block of Γ_T is $\gamma(j-k)$. The log determinant we write as l_T, while the quadratic form is denoted Q_T, so that the log Gaussian likelihood is defined to be

$$l_T + Q_T = \log |\Gamma_T| + X_T' \Gamma_T^{-1} X_T. \tag{5}$$

Brute-force inversion of Γ_T involves order T^3 flops, whereas a straight Cholesky decomposition is much faster. When a block-Toeplitz structure is available, as in the case of a stationary time series, an even faster recursive method is available. Consider viewing Q_T in (5) as the sum of squared model residuals $Z_t = \Gamma_t^{-1/2} X_t$ for each t.

DOI: 10.1111/jtsa.12267

The notation is the same as in the previous subsection, but now we imagine that Γ_t corresponds to the covariance structure of a fitted model. Utilizing Proposition 2 with $Z_t = L_t^{-1} X_t$, we easily obtain that

$$Z_{t+1} = L_{t+1}^{-1} X_{t+1} = \begin{bmatrix} Z_t \\ d_{t+1}^{-1/2}[x_{t+1} - \tilde{\ell}'_{t+1} X_t] \end{bmatrix}.$$

This recursion is initialized with $Z_1 = \Gamma_1^{-1/2} x_1$. In order to compute the next entry of the residual, we must compute $\tilde{\ell}_{t+1} = \Gamma_t^{-1} \underline{\gamma}_t$ and $d_{t+1} = g_{t+1} - \underline{\gamma}_t' \tilde{\ell}_{t+1}$. This $\tilde{\ell}_{t+1}$ is the one-step-ahead predictor, i.e. the best forecast of x_{t+1} from X_t is $\tilde{\ell}'_{t+1} X_t$. We proceed to discuss the efficient calculation of $\tilde{\ell}_{t+1}$.

Proposition 4. A recursive algorithm to compute (5) is given by the initializations $\tilde{u}_2 = \gamma^{-1}(0)\gamma(1)$, $\tilde{\ell}_2 = \gamma^{-1}(0)\gamma(-1)$, $c_2 = \gamma(0) - \overline{\gamma}_1' \tilde{u}_2$, $d_2 = \gamma(0) - \underline{\gamma}_1' \tilde{\ell}_2$, $Q_2 = x_1' \gamma^{-1}(0) x_1 + (x_2 - \tilde{\ell}'_2 x_1)' d_2^{-1} (x_2 - \tilde{\ell}'_2 x_1)$, and $l_2 = \log|\gamma(0)| + \log|c_2|$, followed by the loop for $2 \leq t \leq T-1$ of

$$\xi_t = \gamma(t) - \underline{\gamma}_{t-1}' \tilde{u}_t \tag{6}$$

$$\tilde{\ell}'_{t+1} = \left[\xi_t c_t^{-1},\ \tilde{\ell}'_t - \xi_t c_t^{-1} \tilde{u}_t'\right] \tag{7}$$

$$\tilde{u}'_{t+1} = \left[\tilde{u}'_t - \xi_t' d_t^{-1} \tilde{\ell}'_t,\ \xi_t' d_t^{-1}\right] \tag{8}$$

$$c_{t+1} = \gamma(0) - \overline{\gamma}_t' \tilde{u}_{t+1} \tag{9}$$

$$d_{t+1} = \gamma(0) - \underline{\gamma}_t' \tilde{\ell}_{t+1} \tag{10}$$

$$Q_{t+1} = Q_t + (x_{t+1} - \tilde{\ell}'_{t+1} X_t)' d_{t+1}^{-1} (x_{t+1} - \tilde{\ell}'_{t+1} X_t) \tag{11}$$

$$l_{t+1} = l_t + \log|c_{t+1}|. \tag{12}$$

The total computation cost is of order T^2.

When a Markovian structure on the covariances exists, other algorithms can be entertained – this is the foundation for the efficient state space approach to likelihood evaluation, called the innovations algorithm of the Kalman filter (Durbin and Koopman 2001). A Markov structure implies that the partial autocovariances are eventually zero for high lags. Such a property could be used in the definition of $\tilde{\ell}'_{t+1}$, in order to come up with an alternative recursive approach to its computation; the quantity ξ_t equals the partial autocovariance $\mathrm{Cov}[x_t, x_0 | X_{t-1}]$, and hence is zero whenever t exceeds the Markov order p – indicating by (7) that $\tilde{\ell}_{t+1}$ and $\tilde{u}_{t+1}$ require no additional updating (neither will c_{t+1} or d_{t+1}) when $t > p$. Thus, (6)– (10) are accomplished in p steps. This shows that dramatic computational gains are possible when an additional structure is present.

In the case of a univariate time series, analysts sometimes work with the concentrated likelihood – in the case where the innovation variance is a separate parameter in the time series model, one can algebraically solve for this parameter's minimizer in terms of the others, and then re-substitute. The result is called the concentrated likelihood, being an objective function that depends on one less parameter. If Γ_T above corresponds to the 'innovation-free' covariance matrix, i.e. the calculation of the covariance matrix pretending that the innovation variance equals unity, then Q_T/T is the maximum likelihood estimator of the innovation variance, and the concentrated likelihood then equals $T + T\log(Q_T/T) + l_T$ (here we would likewise compute $l_T = \log|\Gamma_T|$ under the assumption that Γ_T is innovation-free). These quantities are generated by the above algorithm, so it follows that the concentrated likelihood can also be efficiently computed.

3.3. Calculation of Exact Whittle Likelihoods

Given a sample $\{x_1, x_2, \ldots, x_T\}$ from a stationary m-variate time series, a time series model can be fitted via minimizing the exact Whittle likelihood. We use the adjective 'exact' to delineate the Whittle likelihood described

DOI: 10.1111/jtsa.12267

in Taniguchi and Kakizawa (2000), which considers the periodogram at all possible frequencies, which is distinct from the 'approximate' Whittle likelihood, which considers the periodogram only at Fourier frequencies. The exact Whittle likelihood has the advantage over the approximate Whittle likelihood in that the latter is a step further away from the asymptotic objective function, i.e. the exact Whittle likelihood is a closer approximation to the Kullback–Leibler divergence between the true spectral density and the postulated model; see Taniguchi and Kakizawa (2000).

Once values for the parameters have been selected, the model's inverse autocovariance sequence $\{\omega(h)\}$ can be calculated for $-T < h < T$, and the Toeplitz matrix Ω_T can be computed, where, by definition, the jkth $m \times m$ block of Ω_T is $\omega(j-k)$. Recall that inverse autocovariances have the defining property that, when convolved with the autocovariances, one obtains the Kronecker delta function. Put another way, if F is the spectral density corresponding to an invertible postulated model, then the inverse autocovariance sequence is the inverse Fourier transform of F^{-1}. In Taniguchi and Kakizawa (2000), the Whittle likelihood is defined via

$$\log\det\Sigma + \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{tr}\{F^{-1}\, I\}\, d\lambda,$$

where I is the multivariate periodogram (the Fourier transform of the sample autocovariance sequence) and Σ is the innovation variance matrix of the model. Noting that the periodgram can be expressed via

$$I(\lambda) = T^{-1}\left(\sum_{j=1}^{T} x_j z^j\right)\left(\sum_{k=1}^{T} x_k' z^{-k}\right)$$

for $z = e^{-i\lambda}$, it follows that

$$\begin{aligned}
\frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{tr}\{F^{-1}\, I\}\, d\lambda &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{tr}\{F^{-1}\, T^{-1} \sum_{j,k=1}^{T} x_j x_k' z^{j-k}\}\, d\lambda \\
&= T^{-1}\sum_{j,k=1}^{T} \frac{1}{2\pi}\int_{-\pi}^{\pi} x_k' F^{-1} x_j\, z^{j-k}\, d\lambda \\
&= T^{-1}\sum_{j,k=1}^{T} x_k'\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} F^{-1}\, z^{j-k}\, d\lambda\right) x_j \\
&= T^{-1} X_T'\, \Omega_T\, X_T,
\end{aligned}$$

with Ω_T the block Toeplitz matrix corresponding to the inverse spectral density F^{-1}. Therefore the exact Whittle likelihood is defined to be

$$\log|\Sigma| + T^{-1}\, X_T'\, \Omega_T\, X_T, \tag{13}$$

where $X_t = \mathrm{vec}\{x_1, x_2, \ldots, x_t\}$. The log determinant does not depend on T, and so can be calculated once. The quadratic form is denoted $Q_T = X_T'\, \Omega_T\, X_T$, but differs from the Gaussian likelihood's quadratic form in that no matrix inversion is required. The algorithm is given by expressing Q_{t+1} in terms of previously computed Q_t scalars. (Here, $\underline{\omega}_t'$ is the bottom block row of Ω_t, excepting the final entry $\omega(0)$.)

Proposition 5. A recursive algorithm to compute (13) is given by the initialization

$$Q_1 = x_1'\, \omega(0)\, x_1 \qquad \rho_1 = x_2'\, \omega(1)\, x_1,$$

DOI: 10.1111/jtsa.12267

followed by the loop for $1 \le t \le T-1$ of

$$\rho_t = x'_{t+1}\,\underline{\omega}'_t\,X_t \tag{14}$$

$$Q_{t+1} = Q_t + \rho_t + \rho'_t + x'_{t+1}\,\omega(0)\,x_{t+1}. \tag{15}$$

The total computation cost is of order T^2.

It is also possible to concentrate out the Whittle likelihood when the innovation covariance matrix is separately parameterized in the model – see the discussion in McElroy and Findley (2015). In this case, the time series model is represented in causal Wold form via $x_t = \Psi(B)\epsilon_t$ for a vector white noise sequence $\{\epsilon_t\}$ of covariance Σ, and the causal multivariate filter $\Psi(z)$. Then, concentration of the Whittle likelihood amounts to computation of the empirical forecast error variance (FEV) matrix, defined as

$$\widehat{\Sigma} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \Psi^{-1}(e^{-i\lambda})\,I(\lambda)\,\Psi^{\dagger}(e^{i\lambda})\,d\lambda.$$

The concentrated Whittle likelihood is actually equal to $m + \log\det\widehat{\Sigma}$ (McElroy and Findley 2015). Expanding $\Psi^{-1}(z)$ as $\sum_{j\ge 0} \upsilon_j z^j$ results in the empirical FEV being written as

$$\widehat{\Sigma} = \sum_{j,k\ge 0} \upsilon_j \widehat{\gamma}(j-k)\upsilon'_k = [\upsilon_0, \upsilon_1, \ldots, \upsilon_{T-1}]\;\widehat{\Gamma}_T\;[\upsilon_0, \upsilon_1, \ldots, \upsilon_{T-1}]',$$

where $\widehat{\Gamma}_T$ is the block-Toeplitz matrix consisting of the sample autocovariance matrices. Hence, we see that the algorithms for the (unconcentrated) Whittle likelihood can be applied, but now with the coefficients υ_t playing the role of x_t and the inverse autocovariances of the model $\omega(h)$ replaced by the sample autocovariances $\widehat{\gamma}(h)$.

3.4. Spectral Factorization

The multivariate spectral factorization problem is the following: given a positive definite sequence of autocovariances $\gamma(0), \ldots, \gamma(q-1), \gamma(q)$ (i.e. the matrix function $\sum_{|h|\le q}\gamma(h)e^{-i\lambda h}$ must be positive definite for all $\lambda \in [-\pi, \pi]$), determine a vector moving average (VMA) representation of order q that has this covariance structure. For a discussion of VMA and vector autoregressive (VAR) processes, see Brockwell and Davis (1991) or Lütkepohl (2007). This problem is important in engineering applications; see Sayed and Kailath (2001) for an overview. Spectral factorization also provides a way to compute inverse autocovariances, which are needed in the Whittle likelihood.

The basis for our spectral factorization relies on the following known results: If we write down the lower nested MCD of Γ_{t+1} via Proposition 1, it is shown below that the bottom block-row of L_t (i.e. ℓ'_{t+1}) asymptotically corresponds to the coefficients of the VMA. Specifically, if F is the spectral density matrix given by

$$F(\lambda) = \sum_{h=-q}^{q} \Gamma(h)e^{-ih\lambda}$$

for $\lambda \in [-\pi, \pi]$, then the spectral factorization is

$$F(\lambda) = \Theta(z)\,\Sigma\,\Theta'(\overline{z}) \tag{16}$$

for $z = e^{-i\lambda}$. Here, Θ is a degree q polynomial with nonzero leading coefficient Θ_0, which is taken to be the identity matrix as an identifiability restriction ((Lütkepohl, 2007)). Thus $\Theta(z) = \sum_{k=0}^{q}\Theta_k z^k$ is a matrix polynomial. There

DOI: 10.1111/jtsa.12267

are many spectral factorization algorithms available in the literature (c.f., Sayed and Kailath 2001); we here present a method (Bauer 1955) with proof for matrix power series, namely the case where $q = \infty$.

Proposition 6. Suppose that the lower MCD of Γ_t corresponds to a positive definite sequence $\{\gamma(h)\}$, where the kjth block entry of L_{t+1} is written as $[L_{t+1}]_{k,j}$. Then for any $h \geq 0$

$$\lim_{t\to\infty} [L_{t+1}]_{t+1,t+1-h} = \Theta_h.$$

In the case where $q < \infty$, corresponding to a VMA process, Proposition 6 says that the bottom block row of L_t in the non-unit lower MCD is approximately given as a sequence of zero matrices followed by the VMA coefficients. Hence, an iterative algorithm can be constructed to compute ℓ'_t, looping until convergence. Using Propositions 1 and 2, we iteratively compute

$$\begin{aligned}
\ell_{t+1} &= D_t^{-1} L_t^{-1} \underline{\gamma}_t \\
d_{t+1} &= \gamma(0) - \ell'_{t+1} D_t \ell_{t+1} \\
L_{t+1}^{-1} &= \begin{bmatrix} L_t^{-1} & 0 \\ -\ell'_{t+1} L_t^{-1} & 1 \end{bmatrix}.
\end{aligned}$$

Then the last $q+1$ matrices of ℓ'_{t+1} yield the VMA coefficients. It is straightforward to see from this recursion that, when $t > q$, due to the lower triangular structure of L_t^{-1} and the fact that $\gamma(h) = 0$ for $h > q$, the first $t - q$ columns of $-\ell'_{t+1} L_t^{-1}$ are necessarily zero – this is what we desire, because only the first $q+1$ coefficients of $\Theta(z)$ should be nonzero. Efficient computation of ℓ_{t+1} and d_{t+1} follow from the ideas discussed in Section 3.1.

We remark that the Bauer spectral factorization algorithm yields a stable factorization asymptotically. This last result uses the notion of convergence of matrix polynomials: by the expression $\Pi^{(t)}(B) \to \Pi^{(\infty)}(B)$ as $t \to \infty$ we mean that each coefficient matrix of the polynomial $\Pi^{(t)}(B)$ tends to the corresponding coefficient matrix of the power series $\Pi^{(\infty)}(B)$, in the sense that the matrix norm of their difference tends to zero. Moreover, a matrix polynomial $\Pi(B)$ is stable if it has the property that $\det \Pi(z) = 0$ implies $|z| > 1$.

Proposition 7. Let $\Theta^{(t)}(B)$ be the VMA(t) polynomial with coefficients satisfying

$$[\Theta_t, \ldots, \Theta_1, \Theta_0] = \left[\ell'_{t+1}, 1_m\right],$$

the final block row of L_{t+1} in the lower MCD of Γ_{t+1}. Then $\Theta^{(t)}(B) \to \Theta^{(\infty)}(B)$, where $\Theta^{(\infty)}(B)$ is a stable order-q matrix polynomial satisfying (16) (for some Σ) with the leading coefficient equal to the identity matrix.

3.5. Lattice Random Fields

Here we discuss the case of a two-dimensional lattice random field, and how the recursive algorithms can be applied to compute the likelihood for a lattice model. Let us suppose that $\mathbb{X}$ is the $N_1 \times N_2$-dimensional data matrix, which consists of $N_1 \cdot N_2$ values of random variables, each at a corresponding location within a rectangle. For more details on inference for such lattice fields, see McElroy and Holan (2014). The underlying lattice field is denoted x_{t_1,t_2}, with $1 \leq t_1 \leq N_1$ and $1 \leq t_2 \leq N_2$. Supposing that the field is covariance-stationary (and mean zero for simplicity), we have

$$\mathrm{Cov}\left[x_{t_1,t_2}, x_{t_1+h_1,t_2+h_2}\right] = \gamma_{h_1,h_2}$$

for $h_1, h_2 \in \mathbb{Z}$. It is known that $\gamma_{-h_1,-h_2} = \gamma_{h_1,h_2}$. It is necessary to put the data matrix $\mathbb{X}$ into a vectorized form, and lexicographic ordering corresponds to the rule $X = \mathrm{vec}(\mathbb{X}')$. The covariance matrix of the column vector X

DOI: 10.1111/jtsa.12267

consists of $N_1 \times N_1$ blocks, where the jkth block matrix is denoted $\Sigma^{(j-k)}$ with $1 \leq j,k \leq N_1$. Each such block matrix is $N_2 \times N_2$-dimensional, being Toeplitz with ℓ, nth entry given by $\gamma_{j-k,\ell-n}$. Hence, $\mathrm{Cov}(X)$ is block-Toeplitz, denoted Γ_{N_1}, having the structure of (1), where $\underline{\gamma}_t' = \left[\Sigma^{(t)}, \ldots, \Sigma^{(1)}\right]$, and $g_{t+1} = \Sigma^{(0)}$. Note that the dimension of recursion in (1) is in the number of blocks N_1, each of size $N_2 \times N_2$. In order to verify that the structure of (1) is valid, we proceed as follows:

$$[\Sigma^{(h)}]'_{jk} = \Sigma^{(h)}_{kj} = \gamma_{h,k-j} = \gamma_{-h,j-k} = \Sigma^{(-h)}_{jk},$$

so that $\Sigma^{(-h)} = \Sigma^{(h)\prime}$. This shows that the upper right portion of Γ_{t+1} is indeed equal to the transpose of $\underline{\gamma}_t'$. As a result, we can utilize the method of Section 3.2 to calculate the Gaussian likelihood (given some model for the autocovariances); in the algorithm there we can replace each $\gamma(h)$ by $\Sigma^{(h)}$. Whereas in the vector time series case each block matrix $\gamma(h)$ is not Toeplitz, in the case of a lattice random field the block matrices $\Sigma^{(h)}$ are also Toeplitz – in a sense the covariance matrix Γ_{N_1} is doubly Toeplitz. Also note that the observation x_t in the time series case is replaced here by the transpose of the tth row of $\mathbb{X}$.

Now in many applications N_1 and N_2 have similar sizes, so that each of the N_1 steps of the likelihood algorithm requires matrix computations of size N_2; in the case of vector time series, we instead have matrices of fixed size m. Asymptotic results for random fields often stipulate that N_1 and N_2 grow at the same rate, and, due to the parity of these dimensions, computations for random fields can be expensive. While order m^2T^2 operations are required in the vector case, for lattice fields we have order $N_1^2N_2^2$ operations.

The other applications also follow, such as simulation and Whittle evaluation. In the case of the lattice Whittle, one can establish

$$X'_{N_1}\,\Gamma_{N_1}\,X_{N_1} = \sum_{j,k=1}^{N_1} x_j'\tilde{\Sigma}^{(j-k)}x_k,$$

where x_j' is the jth row of $\mathbb{X}$, and $\tilde{\Sigma}^{(h)}$ has ℓ, nth entry given by the autocovariance $\gamma_{h,\ell-n}$ of the inverse spectral density. From here it is straightforward to see how Whittle estimation of spatial data can be efficiently carried out. As a final note, it seems possible to consider dimension greater than two by continuing the block-Toeplitz nesting algorithms recursively in dimension.

3.6. Non-Lattice Data

The algorithms for likelihood evaluation (and others) are also valid for non-lattice data – or data resulting in non-Toeplitz covariance matrices Γ_t, such as time series samples with missing values – so long as the Schur complements (c.f., Remark 1) are still positive definite. This is because the results of Section 2 are purely algebraic. We now sketch how such non-Toeplitz structures can commonly arise.

Besides the obvious scenario of nonstationary time series data, which results in a non-Toeplitz covariance matrix, one might obtain a non-Toeplitz matrix from irregular sampling, i.e. sampling at non-lattice observation points. Consider a covariance stationary time series or spatial m-variate process $x(s)$ with s belonging to some subset of Euclidean space $\mathbb{R}^d$, such that the process has autocovariance function $R(h) = \mathrm{Cov}(x(s+h), x(s))$ for $h \in \mathbb{R}^d$. Sampling locations $s_1, s_2, \ldots, s_T$ are given, which may be times or spatial locations that are non-lattice, i.e. $\|s_j - s_k\|$ is not uniform between neighboring sampling locations. The covariance matrix of the sample $x(s_1), x(s_2), \ldots, x(s_t)$ for any $t \leq T$ is given by the matrix Γ_t, which has jkth block entry $R(s_j - s_k)$, for $1 \leq j,k \leq t$.

Hence the nonuniformity of the displacements $s_j - s_k$ implies there is no Toeplitz structure to Γ_t. However, so long as the autocovariances are computed in a correct and coherent manner (e.g. as derived from a covariance stationary model), the Schur complements will still be invertible, and the likelihood can be evaluated via an MCD algorithm – this follows from Mercer's Theorem (Tanaka 1996). Below, we study covariance stationary multivariate processes, and explicitly show that the resulting covariance matrices of samples must have invertible Schur complements.

DOI: 10.1111/jtsa.12267

The spectral representation of the autocovariance function, assuming this exists, is given by

$$R(h) = \int_{\mathbb{R}^d} F(\lambda)\, \exp\{i\lambda' h\}\, d\lambda \tag{17}$$

with spectral density F, which is a matrix-valued function of frequencies $\lambda \in \mathbb{R}^d$. The inverse Fourier transform yields the relationship

$$F(\lambda) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} R(s) \exp\{-i\lambda' s\}\, ds$$

for a given F. Because $R(-h) = R'(h)$ (this follows from the stationarity assumption), we must have $F'(\lambda) = \overline{F}(\lambda)$. The essential condition for stationarity is that $F(\lambda)$ be nonnegative definite for every λ, but because the spectral matrices are possibly complex-valued, the condition is slightly different from the real case. For any complex vector ω, the scalar $\omega' F(\lambda)\overline{\omega}$ must equal its own transpose, which is

$$\overline{\omega}' F'(\lambda)\omega = \overline{\omega' F(\lambda)\overline{\omega}},$$

and hence $\omega' F(\lambda)\overline{\omega}$ equals its own conjugate. Therefore it is real-valued. Then the positive definite condition on F is that for every λ, for any complex vector ω,

$$\omega' F(\lambda)\overline{\omega} \geq 0$$

and equals zero if and only if ω is the zero vector. We write $F > 0$ for short; c.f., Brockwell and Davis (1991). Below, we show that the resulting Γ_t is block positive definite.

Proposition 8. The block matrix Γ_t with jkth block entry equal to $R(s_j - s_k)$ for R given by (17) has MCD with Schur complements d_j $(1 \leq j \leq t)$ that are positive definite.

As a consequence of Proposition 8, the algorithms of Section 2 can be applied. However, the likelihood algorithm of Section 3.2 assumes a block-Toeplitz structure, and hence is not valid in this scenario; the cruder methods of Propositions 1 and 2 can be used instead.

4. ILLUSTRATIONS

When fitting a time series model, it is common for hundreds of likelihood evaluations to occur, so it is important to avoid crashes due to numerical instability; also it is important that each evaluation is rapidly computed.

4.1. Simulations

To illustrate the potential gains, we consider the application of Gaussian likelihood evaluation for a VARMA process. We consider a trivariate VAR process of order 1 given by $X_t = \Phi X_{t-1} + \epsilon_t$ for $\{\epsilon_t\}$ a vector white noise of covariance matrix Σ, where

$$\Phi = \begin{bmatrix} .5 & 0 & 0 \\ .1 & .1 & .3 \\ 0 & .2 & .3 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 2.25 & 0 & 0 \\ 0 & 1 & .5 \\ 0 & .5 & .74 \end{bmatrix}.$$

This corresponds to the VAR(1) process discussed in Example 2.1.14 of Lütkepohl (2007). We also consider a bivariate VARMA(1,1) process given by $X_t = \Phi X_{t-1} + \epsilon_t + \Theta\epsilon_{t-1}$ for $\{\epsilon_t\}$ a vector white noise of covariance matrix

DOI: 10.1111/jtsa.12267

Table I. Runtimes in hundredths of seconds for the MCD and Default algorithms for calculation of the Gaussian likelihood, as a function of sample size T. The covariance structures correspond to a three-dimensional VAR(1) process and to a a two-dimensional VARMA(1,1) process.

	VAR(1)		VARMA(1,1)	
T	MCD	Default	MCD	Default
50	1.20	3.76	1.03	3.18
100	2.56	17.39	2.18	13.72
200	5.77	117.25	4.67	65.88
300	9.71	373.97	7.44	215.76
400	13.93	834.53	10.45	483.97
500	20.61	1701.09	13.70	884.88

Σ, where

$$\Phi = \begin{bmatrix} .5 & .1 \\ .4 & .5 \end{bmatrix} \qquad \Theta = \begin{bmatrix} .6 & .2 \\ 0 & .3 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} .09 & 0 \\ 0 & .04 \end{bmatrix}.$$

This is similar to the VARMA(2,1) process described in Exercise 11.3 of Lütkepohl (2007). We simulate time series of length T from these specifications, and investigate the runtime to evaluate the Gaussian likelihood at the correct parameter values; i.e. we compute the autocovariance sequence corresponding to the above specification of Φ and Σ, and apply the algorithms of Section 3.2. We report for sample sizes $T = 50, 100, 200, 300, 400, 500$ the runtimes for evaluation of the Gaussian likelihood, comparing the MCD method of Section 3.2 and the Default method (direct Cholesky factorization). Both methods were coded in R (version 3.0.2, 32-bit) and implemented on the same machine (Intel Xeon CPU W3570 3.20 GHz, with 4 GB RAM, 64-bit OS and four cores) for a fair comparison. Table I contains the runtimes in hundredths of a second (obtained as averages over a hundred runs). While the runtimes for the MCD method increase roughly linearly in sample size, the Default method's runtime increases at worse than a quadratic rate.

4.2. Immigration Data

We consider the fitting of a structural model to New Zealand immigration data, displayed in Figure 1. There are six daily time series, recorded between September 1, 1997, and July 31, 2012. The first, third, and fifth series pertain to arrivals, while the second, fourth, and sixth series measure departures. The first and second series count the category of a temporary visit, while the third and fourth series count visitors (short term), and the last two series measure permanent and long-term visits.

Letting $\{x_t\}$ represent the logged data, suppose that there are latent components $\{\mu_t\}$ (a long-term component), $\{\xi_t\}$ (a weekly seasonal component), and $\{\iota_t\}$ (an idiosyncratic white noise), with $x_t = \mu_t + \xi_t + \iota_t$. The three component processes are each multivariate difference stationary processes, as described in McElroy (2017), which satisfy

$$(1 - B)\,\mu_t \sim \mathrm{WN}(0, \Sigma^{\mu})$$
$$(1 + B + B^2 + B^3 + B^4 + B^5 + B^6)\,\xi_t \sim \mathrm{WN}(0, \Sigma^{\xi}),$$

where B is the backshift operator. Here, WN denotes a multivariate white noise process, with the given covariance matrix; if Σ^{μ} (or Σ^{ξ}) is reduced rank, then the data process will have common trends (respectively, common weekly seasonals) and is co-integrated in a generalized sense. In these data, annual seasonality (i.e. stochastic effects with a period of 365.25 days) and trend are difficult to disentangle, and are together featured in the component μ_t. More nuanced modeling of the weekly seasonality can be accomplished by further decomposing ξ_t in terms of the weekly harmonics at frequencies $2\pi/7$, $4\pi/7$, and $6\pi/7$.

DOI: 10.1111/jtsa.12267

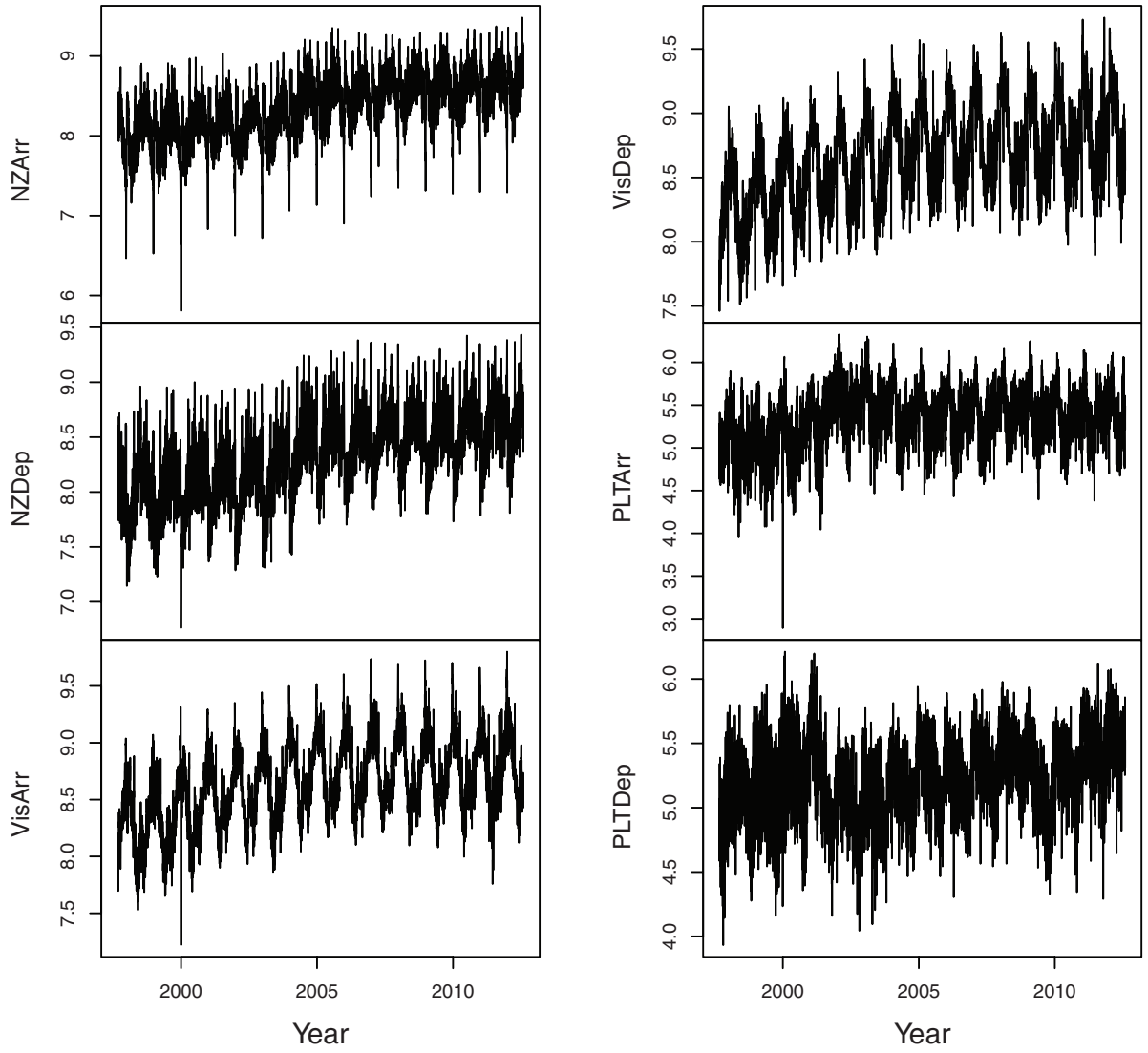

Figure 1. Six-variate (log-transformed) New Zealand immigration data.

One application of interest is to remove the weekly seasonality, so that the long-term features are more clearly discerned; the methods of Section A.1 of Appendix S1 could be applied to accomplish the extraction. Moreover, long-range forecasts of the data may be of interest to a government agency such as Statistics New Zealand, in order to understand and anticipate future immigration and travel patterns.

The model indicates that weekly differencing $1 - B^7$ should suffice to render the data stationary. With $m = 6$ and $T = 5448$, direct inversion of the sample's covariance matrix is impossible due to memory limitations, and recursive methods are necessary for forecasting and likelihood evaluation. The average runtime over a hundred repetitions (same machine as used for simulations) was 1.23 minutes for a single evaluation of the likelihood. Whereas multistep-ahead forecasting (which is based on computing $\tilde{\ell}_T$) is feasible in such a case, maximum likelihood estimation can take days or weeks (because hundreds of evaluations are necessary) – this is more of an issue with the high-dimensional parameter space (111 parameters) than with likelihood evaluation *per se*.

DOI: 10.1111/jtsa.12267

ACKNOWLEDGEMENTS

This report is released to inform interested parties of research and to encourage discussion. The views expressed on statistical issues are those of the author and not necessarily those of the U.S. Census Bureau.

SUPPORTING INFORMATION

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

REFERENCES

Akaike H. 1973. Block Toeplitz matrix inversion. *SIAM Journal on Applied Mathematics* **24**: 234–241.
Axelsson O. 1996. *Iterative Solution Methods*. New York: Cambridge University Press.
Bauer F. 1955. Ein direktes Iterationsverfahren zur Hurwitz-Zerlegung eines Polynoms. *Archiv der elektrischen Übertragung* **9**: 285–290.
Brockwell P, Davis R. 1991. *Time Series: Theory and Methods*. New York: Springer.
Durbin J, Koopman S. 2001. *Time Series Analysis by State Space Methods*. Oxford: Oxford University Press.
Golub G, Van Loan C. 1996. *Matrix Computations*. Baltimore: Johns Hopkins University Press.
Haykin S. 1996. *Adaptive Filter Theory*. Upper Saddle River: Prentice Hall.
Lütkepohl H. 2007. *New Introduction to Multiple Time Series Analysis*. Berlin: Springer-Verlag.
McElroy T. 2017. Multivariate seasonal adjustment, economic identities, and seasonal taxonomy. *Journal of Business and Economics Statistics* **35**(4): 511–525.
McElroy T, Findley D. 2015. Fitting constrained vector autoregression models. In *Empirical Economic and Financial Research. Advanced Studies in Theoretical and Applied Econometrics*, Vol. 48, Beran J, Feng Y, Hebbel H (eds.) Cham: Springer; 451–470.
McElroy T, Holan S. 2014. Asymptotic Theory of Cepstral Random Fields. *Annals of Statistics* **42**: 64–86.
Newton H, Pagano M. 1983. The finite memory prediction of covariance stationary time series. *SIAM Journal on Scientific and Statistical Computing* **4**: 330–339.
Pagano M. 1976. On the linear convergence of a covariance factorization algorithm. *Journal Association for Computing Machinery* **23**: 310–316.
Sayed A, Kailath T. 2001. A survey of spectral factorization methods. *Numerical Linear Algebra with Applications* **8**: 467–496.
Tanaka K. 1996. *Time Series Analysis: Nonstationary and Noninvertible Distribution Theory*. New York: John Wiley and Sons.
Taniguchi M, Kakizawa Y. 2000. *Asymptotic Theory of Statistical Inference for Time Series*. New York: Springer-Verlag.
Whittle P. 1963. On the fitting of multivariate autoregressions and the approximate canonical factorization of spectral density matrix. *Biometrika* **50**: 129–134.

DOI: 10.1111/jtsa.12267