Econometric Reviews

ISSN: 0747-4938 (Print) 1532-4168 (Online) Journal homepage: http://www.tandfonline.com/loi/lecr20

Multistep ahead forecasting of vector time series

Tucker McElroy & Michael W. McCracken

To cite this article: Tucker McElroy & Michael W. McCracken (2017) Multistep ahead forecasting of vector time series, Econometric Reviews, 36:5, 495-513, DOI: 10.1080/07474938.2014.977088

To link to this article: http://dx.doi.org/10.1080/07474938.2014.977088

Accepted author version posted online: 20 Oct 2014.
Published online: 20 Oct 2014.

ECONOMETRIC REVIEWS
2017, VOL. 36, NO. 5, 495–513
http://dx.doi.org/10.1080/07474938.2014.977088

Taylor & Francis
Taylor & Francis Group

Multistep ahead forecasting of vector time series

Tucker McElroy[a] and Michael W. McCracken[b]

[a]Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, D.C., USA; [b]Research Division, Federal Reserve Bank of St. Louis, St. Louis, Missouri, USA

ABSTRACT

This article develops the theory of multistep ahead forecasting for vector time series that exhibit temporal nonstationarity and co-integration. We treat the case of a semi-infinite past by developing the forecast filters and the forecast error filters explicitly. We also provide formulas for forecasting from a finite data sample. This latter application can be accomplished by using large matrices, which remains practicable when the total sample size is moderate. Expressions for the mean square error of forecasts are also derived and can be implemented readily. The flexibility and generality of these formulas are illustrated by four diverse applications: forecasting euro area macroeconomic aggregates; backcasting fertility rates by racial category; forecasting long memory inflation data; and forecasting regional housing starts using a seasonally co-integrated model.

KEYWORDS

Euro area; fertility rates; frequency domain; housing starts; multivariate time series; VAR models

JEL CLASSIFICATION

C01; C51

1. Introduction

This article develops the theory of multistep ahead forecasting for vector time series that exhibit temporal nonstationarity and may possibly be co-integrated. We begin by treating the case of an infinite past (i.e., all past observations of a time series are available) by developing the forecast filters and the forecast error filters explicitly. These filters are principally of theoretical interest, since they require an infinite span of data to be applied. They represent the long-term aspect of forecasting when the dataset is quite large and can be examined in the frequency domain. By considering the gain and phase delay of various components of the forecast filters, one can learn how various components are attenuated or advanced corresponding to the spectral structure of the vector process.

A secondary objective is to provide formulas for forecasting from a finite data sample. This can be accomplished by using large matrices, which remains practicable when the total sample size is moderate. Expressions for the mean square error (MSE) of forecasts are also derived and can be implemented readily. These expressions hold quite generally and are valid when finite-dimensional State Space (SS) approaches cannot be used (e.g., with long-memory processes, such as VARFIMA; see Sela and Hurvich, 2009).

Our formulation allows for fairly general temporal differencing, which is allowed to differ for each series. In particular, we suppose (i) a vector autoregressive (VAR) operator exists that reduces the unstable data vector process to stationarity and (ii) that some of the VAR roots are unit and others are outside the unit circle. This formulation includes the possibility of co-integrated components. Furthermore, we assume that the VAR-differenced process has a Wold decomposition, which should be invertible when forecasting from an infinite past. This includes co-integrated VAR and VARMA processes but is also more general. When working with a finite sample, we can allow the differenced process to be non-invertible at a finite number of frequencies, which makes our methods applicable to structural VARMA processes that may be collinear (e.g., the case of common trends).

CONTACT Tucker McElroy tucker.s.mcelroy@census.gov U.S. Census Bureau, Time Series Group, 4600 Silver Hill Road, Washington, D.C., 20233, USA.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lecr.

Given the importance of forecasting for econometric applications, a considerable literature on the topic already exists. The VARMA case is treated in Chapter 11 of Brockwell and Davis (1991), and Lütkepohl (2006) provides an overview of the (co-integrated) VAR and VARIMA cases. An older literature includes works by Whittle (1984) and Lütkepohl (1986, 1987), all of which focus on VARMA processes. Athanasopoulus and Vahid (2008) make a recent contribution. Our results generalize this prior literature in three ways: (i) for an infinite past, only an invertible Wold form is assumed for the differenced data; (ii) for a finite sample of data, only the autocovariances and the differencing operator need to be known; and (iii) for finite samples, we allow for missing values and more general prediction problems. For instance, the forecasting of VARFIMA (see Pai and Ravishanker, 2010) is not covered by the previous literature. (Sela and Hurvich, 2009, do handle the stationary case; our treatment is more general.)

Therefore, the stated formulas in this article are novel and useful at the level of generality needed for a multivariate time series analysis. Although vector time series are frequently modeled and forecasted using a SS formulation, there are cases not amenable to this approach such as vector long-memory models and vector cepstral models (Holan et al., 2014); moreover, the formulas in this article are very easy to encode, circumventing the need to purchase SS software. Section 2 considers the case of a semi-infinite past, while Section 3 describes the case of a finite sample of data. Section 4 provides applications and discussion. Section 5 concludes. Proofs are in the Appendix.

2. Multistep ahead forecasting from a semi-infinite past

Consider an n-dimensional nonstationary time series $\{\mathbf{X}_t\}$. We use bold notation for vector objects. In this section, we derive and state formulas for MSE optimal prediction of the time series h steps ahead, given a semi-infinite past and the assumption that $\{\mathbf{X}_t\}$ is difference stationary. Here MSE optimality means that the estimators have the best MSE among all linear functions of the data (from the semi-infinite past). If the data are Gaussian, they have the best MSE among all functions of the data. The semi-infinite past is useful for investigating the long-term impact of forecasting since the frequency response function (FRF) of the forecast filter can be calculated and studied. In addition, in some cases the semi-infinite predictors are easier to calculate than the finite-sample predictors and hence might be used for long samples. (Recall that the FRF is the discrete Fourier transform of the filter coefficients, expressed as a function of frequency $\lambda \in [-\pi, \pi]$.)

We consider difference stationary processes and generally follow the treatments of vector time series in Brockwell and Davis (1991), Taniguchi and Kakizawa (2000), and Lütkepohl (2006). Included in our framework are the popular co-integated VAR and VARIMA models used by econometricians, as well as structural VARIMA models. The formulas also cover more unconventional processes with long-range dependence. For notation, an underscore is used for every matrix, which generally are $n \times n$. The identity matrix is denoted by $\underline{I}_n$. In general, capital letters refer to composite objects and lowercase letters refer to components (such as coefficients). Latin letters refer to random variables/vectors, and Greek letters refer to deterministic quantities (e.g., parameters). Matrix polynomial and power series functions are defined as $\underline{A}(x) = \sum_{k=0}^{p} \underline{a}_j x^j$ with $p < \infty$ or $p = \infty$, as the case may be. We use B for the backshift operator, which sends a time series back in time: $B\mathbf{X}_t = \mathbf{X}_{t-1}$, working on all components of the vector at once. Then the action of $\underline{A}(B)$ on $\mathbf{X}_t$ is understood by linear extension.

The difference stationarity assumption means there exists a differencing polynomial whose application to the observed time series $\{\mathbf{X}_t\}$ yields a covariance stationary time series $\{\mathbf{W}_t\}$. In particular, suppose that $\underline{\Delta}(B)$ of order p exists such that

$$\underline{\Delta}(B)\mathbf{X}_t = \mathbf{W}_t, \tag{1}$$

and there exist complex variables z such that $|z| = 1$ and $\det \underline{\Delta}(z) = 0$. The operator $\underline{\Delta}(B) = \sum_{k=0}^{p} \underline{\delta}_j B^j$ is referred to as the VAR-differencing operator, which in general contains both stable and unstable elements that are not easily separated. We assume that $\underline{\delta}_0 = \underline{I}_n$. As discussed in Lütkepohl (2006),

the zeroes of det $\underline{\Delta}(z)$ include some on the unit circle of the complex plane, and the rest outside. Suppose that the number of unit roots is d. Then there exists an order d polynomial (in terms of the backshift operator) with only unit roots such that its application to each series individually produces stationarity. However, the net effect may produce singularities in the resulting spectral density matrix when co-integration is present. In fact, cointegration is present whenever $\underline{\Delta}(B)$ cannot be neatly decomposed into a pure unit root matrix polynomial times a purely stable VAR matrix polynomial. At any unit root z of det $\underline{\Delta}(z)$, the matrix $\underline{\Delta}(z)$ has rank less than n. For further background, see Granger (1981), Engle and Granger (1987), and Stock and Watson (1988).

The inverse of $\underline{\Delta}(B)$ exists only in a formal sense as $\underline{\Xi}(B) = \sum_{k\geq 0} \underline{\xi}_k B^k$, a formal power series. This expression is not well defined on the unit circle, but it is mathematically convenient and is used later. The series $\{\mathbf{W}_t\}$ is then stationary with mean vector $\mathbf{m}$, and we suppose that it is purely nondeterministic and invertible (i.e., the spectrum is always nonsingular). Note that in the co-integration formulation of Engle and Granger (1987) and Stock and Watson (1988), all variables have been fully differenced to stationarity, so the resulting stationary component processes have a singular spectral density. In contrast (as in Lütkepohl, 2006), we avoid this potential over-differencing, such that the resulting stationary vector process is invertible. Then by the Wold decomposition (Brockwell and Davis, 1991, or Reinsel, 1997) for multivariate time series we can write

$$\mathbf{W}_t = \mathbf{m} + \underline{\Psi}(B)\mathbf{A}_t, \tag{2}$$

where the series $\mathbf{A}_t$ is mean zero and uncorrelated (but possibly dependent) over time with covariance matrix $\underline{\sigma}$. Here $\underline{\Psi}(B)$ is a causal power series with coefficient matrices $\underline{\psi}_k$. By the invertibility assumption, det $\underline{\Psi}(z) \neq 0$ on the unit circle, so that $\underline{\Psi}^{-1}(z)$ is well defined for $|z| \leq 1$. More properly we should write $[\underline{\Psi}(z)]^{-1}$ for the inverse, since it is different for each value of z, but we use the first expression by abuse of notation (as is common). Note that allowing $\{\mathbf{A}_t\}$ to be dependent allows our results to be applied to nonlinear difference stationary processes (see Maïnassara and Francq, 2011, for a complete discussion).

We are interested in the following problem: how to compute the h-step ahead forecast of $\mathbf{X}_{t+h}$ from data $\{\mathbf{X}_s; s \leq t\}$. This is an important problem and has been studied in more limited forms by other authors.[1] The solution can be obtained using projections, given that our criterion function is the best linear MSE estimator. Denote this estimator as $\widehat{\mathbf{X}}_{t+h|t}$, which is the conditional expectation of $\mathbf{X}_{t+h}$ given the semi-infinite past when the series is Gaussian. As a first step, we consider the problem of determining $\widehat{\mathbf{W}}_{t+h|t}$, the best estimate of $\mathbf{W}_{t+h}$ given the semi-infinite past $\{\mathbf{W}_s; s \leq t\}$. We introduce the following convenient notation for any matrix power series $\underline{A}(x)$: $[\underline{A}]_0^j(x) = \sum_{k=0}^j \underline{a}_k x^k$. The optimal forecast and forecast filter are given below, and are quite similar to the univariate case discussed in McElroy and Findley (2010).

Proposition 1. *Suppose that* $\{\mathbf{W}_t\}$ *is strictly stationary with mean* $\mathbf{m}$*, finite variance, and is purely non-deterministic and invertible. The best linear MSE estimator of* $\mathbf{W}_{t+h}$ *given the semi-infinite past* $\{\mathbf{W}_s; s \leq t\}$ *is*

$$\widehat{\mathbf{W}}_{t+h|t} = \mathbf{m} + \sum_{k\geq 0} \underline{\psi}_{k+h} \underline{\Psi}^{-1}(B) \ \left(\mathbf{W}_{t-k} - \mathbf{m}\right). \tag{3}$$

With the forecast filter $\Phi(B)$ *defined via*

$$\widehat{\mathbf{W}}_{t+h|t} = \mathbf{m} + \Phi(B) \ (\mathbf{W}_t - \mathbf{m}) = \left(\underline{I}_n - \Phi(1)\right)\mathbf{m} + \Phi(B)\,\mathbf{W}_t,$$

[1] Chapter 9 of Whittle (1984) provides formulas for one-step ahead forecasting of a stationary process, given its Wold decomposition, but does not treat difference-stationary processes explicitly. Lütkepohl (1986, 1987) and Schorfheide (2005) focus on VAR or VARMA models, as does the treatment in Brockwell and Davis (1991). No published work to date handles the general case of a difference-stationary process; our Propositions 1 and 2 provide forecast formulas generally, allowing for vector long memory (e.g., VARFIMA processes). Another possible application is the multivariate extension of exponential (or cepstral) models as described in Bloomfield (1973).

its FRF has the formula

$$\Phi(z) = z^{-h}[\underline{\Psi}]_h^{\infty}(z)\underline{\Psi}^{-1}(z) \tag{4}$$

for $z = e^{-i\lambda}$. *The corresponding h-step ahead forecast error is*

$$\mathbf{E}_{t+h} = -[\underline{\Psi}]_0^{h-1}(B)\underline{\Psi}^{-1}(B)\ \left(\mathbf{W}_{t+h} - \mathbf{m}\right). \tag{5}$$

Remark 1. The forecast MSE is defined as the covariance matrix of $\mathbf{E}_{t+h}$. Eq. (5) can be written as

$$\mathbf{E}_{t+h} = -[\underline{\Psi}]_0^{h-1}(B)\ \mathbf{A}_{t+h},$$

and hence the MSE is $\sum_{j=0}^{h-1} \underline{\psi}_j \underline{\sigma}\, \underline{\psi}_j'$. The diagonal entries correspond to mean square forecast errors for each individual series.

The forecasting problem occurs when $h \geq 1$. However, it is convenient to extend the above formulas to the case $h \leq 0$, in which case $\Phi(B)$ should equal B^{-h}. Since $[\underline{\Psi}]_h^{\infty} = \underline{\Psi}$ when $h \leq 0$, the stated formula (4) for the forecast FRF immediately reduces to the appropriate quantity when $h \leq 0$. Also, in this case the forecast error $\mathbf{E}_{t+h} = 0$, since $[\underline{\Psi}]_0^{h-1} = 0$.

Now consider the calculation of $\widehat{\mathbf{X}}_{t+h|t}$. First, we need an elementary description for the difference-stationary time series (1) that differs slightly from the treatment in Bell (1984). It is immediate from (1) that

$$\mathbf{X}_t = \mathbf{W}_t - \sum_{j=1}^{p} \underline{\delta}_j \mathbf{X}_{t-j}$$

for all times t. Let $\underline{Q}(B) = \underline{I}_n - \underline{\Delta}(B)$, and note that it is a matrix polynomial in B such that B occurs to a power $k \geq 1$. Moreover, $B^{-1}\underline{Q}(B)$ is a matrix polynomial in B. Then the above equation is expressed as $\mathbf{X}_t = \mathbf{W}_t + \underline{Q}(B)\mathbf{X}_t$, which upon iteration h times in $\mathbf{X}_t$ yields

$$\mathbf{X}_{t+h} = \left(\underline{I}_n + \underline{Q}(B) + \cdots\ [\underline{Q}(B)]^{h-1}\right)\mathbf{W}_{t+h} + \ [B^{-1}\underline{Q}(B)]^h \mathbf{X}_t. \tag{6}$$

Equation (6) provides a decomposition in terms of linear combinations of various $\mathbf{W}_s$ for $s \leq t+h$ and various $\mathbf{X}_s$ for $s \leq t$. This latter aspect is emphasized here: $[B^{-1}\underline{Q}(B)]^h$ is a polynomial in B, and so the latter term on the right-hand side of (6) depends only on present and past values of the data. The matrix operating on $\mathbf{W}_{t+h}$ in (6) will be denoted $\underline{P}^{(h)}(B) = \sum_{j\geq 0} \underline{p}_j B^j$ and has maximum order $p(h-1)$. Recall that $\widehat{\mathbf{W}}_{s|t}$ is given by (3) with $h = s - t$ if $s > t$, but just $\mathbf{W}_s$ if $s \leq t$. With this representation of the data process, we can state our forecasting result.

Proposition 2. *Suppose that* $\{\mathbf{W}_t\}$ *is strictly stationary with mean* $\mathbf{m}$, *finite variance, and is purely non-deterministic and invertible. Also assume that the future innovations* $\mathbf{A}_{t+j}$ *are uncorrelated with all past values of the process* $\mathbf{X}_t$, *for any* $j \geq 1$. *Then the best linear MSE estimator of* $\mathbf{X}_{t+h}$ *given the semi-infinite past* $\{\mathbf{X}_s; s \leq t\}$ *is*

$$\widehat{\mathbf{X}}_{t+h|t} = \underline{P}^{(h)}(B)\widehat{\mathbf{W}}_{t+h|t} + \ [B^{-1}\underline{Q}(B)]^h \mathbf{X}_t, \tag{7}$$

where $\underline{P}^{(h)}(B)\widehat{\mathbf{W}}_{t+h|t}$ *means* $\sum_{j\geq 0} \underline{p}_j \widehat{\mathbf{W}}_{t+h-j|t}$. *Also (7) can be rewritten as*

$$\widehat{\mathbf{X}}_{t+h|t} = \underline{P}^{(h)}(1)\ \left(\underline{I}_n - \Phi(1)\right)\mathbf{m} + \Phi(B)\mathbf{X}_t,$$

with the forecast filter $\Phi(B)$ *defined via the FRF as*

$$\Phi(z) = \underline{P}^{(h)}(z)z^{-h}[\underline{\Psi}]_h^{\infty}(z)\underline{\Psi}^{-1}(z)\underline{\Delta}(z) + \ [z^{-1}\underline{Q}(z)]^h \tag{8}$$

for $z = e^{-i\lambda}$. *The corresponding h-step ahead forecast error is*

$$\mathbf{E}_{t+h} = -\sum_{j=0}^{h-1} \underline{\xi}_j [\underline{\Psi}]_0^{h-1-j}(B)\underline{\Psi}^{-1}(B)\ \left(\mathbf{W}_{t+h-j} - \mathbf{m}\right) = -[\underline{\Delta}^{-1}(B)\underline{\Psi}(B)]_0^{h-1}\underline{\Psi}^{-1}(B)\ (\mathbf{W}_t - \mathbf{m})\,. \quad (9)$$

3. Projection from a finite past

Now consider the data process at times $t = 1, 2, \ldots, T$. When working with a finite sample, we do not require full knowledge of the Wold filter $\underline{\Psi}(B)$, but rather the autocovariances of the process. We write $\underline{\gamma}(h) = \mathbb{E}[\mathbf{W}_t\mathbf{W}'_{t-h}]$ for the hth covariance matrix, which has individual entries $\underline{\gamma}_{jk}(h)$. We first consider general projection formulas, discuss computational issues, and then work through the examples of forecasting and imputation.

Although finite-sample results are currently available for VAR and VARMA models (see Lütkepohl, 2006), more general formulas for vector long-memory, vector cepstral, or vector structural processes (see Harvey, 1989, or Section 4 below for elaboration) have not been published. Our Theorem 1 below provides the general solution, allowing for missing values (of an arbitrary pattern). While forecasting and imputation for VARMA models can be done using SS methods, long-memory and cepstral processes cannot be embedded in a finite-dimensional SS formulation. Chan and Palma (1998) discuss an infinite-dimensional SS representation for long-memory proceesses, with a finite-dimensional approximation. However, as discussed in McElroy and Holan (2012), a severe degree of distortion to the autocovariance structure can result from such truncations when the long-memory exponent is quite high. Vector long-memory models are discussed in Sela and Hurvich (2009), while vector cepstral processes are formulated in Holan et al. (2014).

3.1. The general treatment

As in Section 2, we allow the differencing matrix $\underline{\Delta}(B)$ to have both stable and unstable elements, such that the resulting $\{\mathbf{W}_t\}$ is stationary, although we can now allow this differenced process to be noninvertible. An equation involving matrices and random vectors can be obtained from the basic relation (1) simply by stacking over values of t as follows:

$$\begin{bmatrix} \underline{I}_n & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \underline{I}_n & 0 & 0 & \cdots & 0 \\ \underline{\delta}_p & \cdots & \underline{\delta}_1 & \underline{\delta}_0 & 0 & \cdots & 0 \\ 0 & \underline{\delta}_p & \ddots & \underline{\delta}_1 & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & \underline{\delta}_p & \cdots & \underline{\delta}_1 & \underline{\delta}_0 \end{bmatrix} \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_p \\ \mathbf{X}_{p+1} \\ \vdots \\ \mathbf{X}_T \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_p \\ \mathbf{W}_{p+1} \\ \vdots \\ \mathbf{W}_T \end{bmatrix}.$$

This relation can be expressed compactly by writing $\Delta\mathbf{X} = [\mathbf{X}'_*, \mathbf{W}']'$, where $\mathbf{X}_* = \mathrm{vec}[\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p]$, and these are called the initial values of the process. Also $\mathbf{W} = \mathrm{vec}[\mathbf{X}_{p+1}, \mathbf{X}_{p+2}, \ldots, \mathbf{X}_T]$. In time-series analysis, the initial values are typically assumed to be independent of the differenced series $\{\mathbf{W}_t\}$, though it is sufficient for our purposes that the initial values be uncorrelated with $\mathbf{W}$. The covariance matrix of the random vector $\mathbf{W}$ is the $n(T-p)$-dimensional matrix $\underline{\Gamma}$, which is block Toeplitz with the jkth block given by $\underline{\gamma}(j-k)$. The mean vector of $\mathbf{W}$ is written as $\iota \otimes \mathbf{m}$, where ι is a column vector of ones of length $T-p$.

We can relax the invertibility assumption on the $\{\mathbf{W}_t\}$ process used in Section 2, provided the spectral density is only singular on a set of frequencies of Lebesgue measure zero. In this case (discussed in McElroy and Trimbur, 2012), any covariance matrix obtained by sampling from $\{\mathbf{W}_t\}$ will be positive

definite. For instance, for any block vector $\underline{a}' = [\underline{a}_1', \ldots, \underline{a}_{T-p}']$,

$$\underline{a}' \, \underline{\Gamma} \, \underline{a} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \underline{a}'(\lambda) \mathbf{F}_{\mathbf{W}}(\lambda) \underline{a}(\lambda) \, d\lambda$$

holds for any vector $\underline{a}$, and with $\underline{a}(\lambda) = \sum_{t=1}^{T-p} \underline{a}_t e^{-i\lambda t}$ and $\mathbf{F}_{\mathbf{W}}$ the Hermitian spectral density matrix of $\{\mathbf{W}_t\}$. At worst (assuming $\underline{a}$ is not the zero vector), there is a set of frequencies of Lebesgue measure zero such that the integrand is zero, but otherwise the integrand is positive; hence the entire expression is positive. This shows that $\underline{\Gamma}$ is positive definite, and a similar argument applies to the covariance matrix of any subvector of $\mathbf{W}$. It is necessary that such covariance matrices be invertible—not only for our forecasting and projection results below—but also so that the Gaussian likelihood function is well defined, and we can obtain model parameter estimates in the first place.

The case where $\mathbf{F}_{\mathbf{W}}$ is singular on a set of frequencies of Lebesgue measure zero can arise in the following way. We might consider potentially over-differencing in (1), such that in the resulting Wold form $\Psi(z)$ has less than full rank at some frequencies; these frequencies would correspond to the co-integrating frequencies, typically just $\lambda = 0$ for most econometric applications. Then it follows that $\mathbf{F}_{\mathbf{W}}$ is singular at those same frequencies but is invertible at all other values. See Stock and Watson (1988) or the discussion in McElroy and Trimbur (2012).

A different case arises when the observed process $\{\mathbf{X}_t\}$ is a sum of unobserved signal and noise processes, as described in McElroy and Trimbur (2012). Then when the signal process has collinear innovations (i.e., the innovations that drive its Wold representation have a singular covariance matrix), the resulting $\mathbf{F}_{\mathbf{W}}$ can be singular at a finite number of co-integrating frequencies. In fact, as discussed in McElroy and Trimbur (2012), such a latent signal, together with an invertible latent noise, gives rise to a co-integrated process, where the co-integrating frequencies correspond to the roots of the signal processes' differencing operator. For example, if the signal is a common trend process (see Stock and Watson, 1988, and Nyblom and Harvey, 2000) and the noise is white, then the differenced data process has a spectrum that is singular only at $\lambda = 0$, corresponding to the signal unit root.

The forecasting problem is now a special case of a more general prediction problem. As in the previous section, optimality refers to the minimum MSE linear predictor, which is also the conditional expectation in the case of a Gaussian process. Suppose that the observed data are actually given by $J\mathbf{X}$, where J has a number of rows less than or equal to its number of columns nT. If J consists of ones and zeroes, we can allow for any type of missing data, backcasting, or forecasting problem. The target quantities of interest are given by $K\mathbf{X}$, where K is another selection matrix. For forecasting and imputation problems, row permutations of the juxtaposition of J and K typically equals the identity matrix.

Apart from trivial applications (such as one-step ahead forecasting), careful thought must be given to the construction of $\mathbf{X}$, J, and K. We must (i) consider all times t of the process $\{\mathbf{X}_t\}$ that enter into the observed data, as well as the target quantities, and (ii) extend this index set if necessary to be a contiguous subset of the integers. Label this contiguous set as $\{1, 2, \ldots, T\}$, and define $\mathbf{X}$ accordingly. At this point, the definition of J and K such that $J\mathbf{X}$ is our observed data and $K\mathbf{X}$ is our target becomes immediate. Examples are given below.

The general solution for optimal prediction of $K\mathbf{X}$ given data $J\mathbf{X}$ is

$$K \, \mathbb{E}[\mathbf{X}] + K \, \mathrm{Cov}(\mathbf{X}, \mathbf{X}) \, J' \left[J \mathrm{Cov}(\mathbf{X}, \mathbf{X}) J' \right]^{-1} J \, (\mathbf{X} - \mathbb{E}\mathbf{X}) \, .$$

Computing these covariances in terms of $\underline{\Gamma}$ and Δ requires some additional analysis. Observe that Δ is block lower triangular and invertible, and $J\mathbf{X} = J\Delta^{-1}[\mathbf{X}_*', \mathbf{W}']'$. Suppose that we can find an invertible matrix R such that

$$RJ\Delta^{-1} = \begin{bmatrix} \underline{I}_{np} & 0 \\ 0 & \underline{B} \end{bmatrix}, \tag{10}$$

where $\underline{B}$ has fewer rows than columns. Such factorizations often exist and are easy to demonstrate in cases of forecasting and backcasting. In addition, condition (10) can be checked by calculating $J\Delta^{-1}$

and then attempting to row reduce it to an upper triangular (but potentially rectangular) matrix; in this case, R is the composition of all row operations. We provide specific examples of the factorization (10) in the next section; a detailed treatment of the factorization problem for mixed-frequency univariate series is given in McElroy and Monsell (2012). The general formula for the projection is given below.

Theorem 1. *Suppose that* $\{\mathbf{W}_t\}$ *is covariance stationary with mean* $\mathbf{m}$*, and has finite variance with spectral density* $\mathbf{F_W}$ *that is singular only at a set of frequencies of Lebesgue measure zero. Also suppose that an invertible matrix R exists such that the factorization (10) holds for some* $\underline{B}$*, and suppose that* $\mathbf{X}_*$ *is uncorrelated with* $\{\mathbf{W}_t\}$*. Then the optimal estimate of* $K\mathbf{X}$ *given* $J\mathbf{X}$ *is*

$$K\,\Delta^{-1}\begin{bmatrix} \underline{I}_{np} & 0 \\ 0 & \underline{\Gamma B}'\left(\underline{B\Gamma B}'\right)^{-1}\end{bmatrix} RJ\mathbf{X} + K\,\Delta^{-1}\begin{bmatrix} 0 \\ \left(\underline{I}_{n(T-p)} - \underline{\Gamma B}'\left(\underline{B\Gamma B}'\right)^{-1}\underline{B}\right)\ [\iota \otimes \mathbf{m}]\end{bmatrix}.$$

The covariance of the error process is

$$K\,\Delta^{-1}\begin{bmatrix} 0 & 0 \\ 0 & \underline{\Gamma} - \underline{\Gamma B}'\left(\underline{B\Gamma B}'\right)^{-1}\underline{B\Gamma}\end{bmatrix}\Delta'^{-1}\,K'.$$

This general projection formulation is useful for a range of applications. For example, the series are allowed to have different lengths (some series go further into the past) or perhaps to have missing values. Regardless, these situations can be handled by adapting J and K appropriately. If back-casting is desired, it is sometimes easier to time reverse all the data and then apply Theorem 1. We discuss additional aspects of the result in the following subsections.

3.2. Computation

In the special case when there are no missing values, forecasts of $\mathbf{X}_{T+h}$ could be easily computed using only recursions and forecasts of the $\{\mathbf{W}_t\}$ process. For example, if $\underline{\delta}(B) = \underline{I}_n - B \cdot \underline{I}_n$ is the differencing operator, then $\widehat{\mathbf{X}}_{T+1} = \mathbf{X}_T + \widehat{\mathbf{W}}_{T+1}$, and $\widehat{\mathbf{W}}_{T+1} = [\underline{\gamma}(T), \ldots, \underline{\gamma}(1)]\,\underline{\Gamma}^{-1}\,\Delta\,\mathbf{X}$. But if some of the components of X_T are missing, these must first be estimated, and the value of a recursive scheme is diminished. For general and complicated patterns of omission (e.g., values missing from different constituent series at different times), the formula in Theorem 1 should be directly computed instead. Here we discuss how these quantities can be obtained.

Given J and Δ, we first compute Δ^{-1}, which is discussed below. Then we calculate $J\Delta^{-1}$ and use row operations to reduce it to the form given in (10) if possible; R is then the composition of these row operations (and hence is invertible). Now that R and $\underline{B}$ are available, we only need to compute the autocovariances and generate $\underline{\Gamma}$ and $\underline{\Gamma}^{-1}$. (There are numerically stable methods for inverting block Toeplitz matrices, such as the Durbin–Levinson algorithm, discussed in Brockwell and Davis, 1991.)

When a time-series model is fitted, we typically are able to compute the Wold coefficients, and from these we can obtain the autocovariances

$$\underline{\gamma}(h) = \sum_{j\geq 0} \underline{\psi}_{j+h}\,\underline{\sigma}\,\underline{\psi}_j'. \tag{11}$$

For a long-memory model, we might use (11) directly to compute autocovariances (this requires truncation of the summation) or perhaps use a direct formula (see Brockwell and Davis, 1991, and Sela and Hurvich, 2009). For VAR and VARMA models, the autocovariances can be computed from the parameters directly, and the Wold coefficients are not needed (see discussion in Mittnik, 1987, 1990, 1993).

For a structural model, autocovariances must be computed for each component, and then the result is summed. Suppose that $\mathbf{X}_t = \mathbf{S}_t + \mathbf{N}_t$, where $\{\mathbf{S}_t\}$ and $\{\mathbf{N}_t\}$ are two latent processes that are stationary and uncorrelated with one another. Hence, the respective autocovariance functions $\{\underline{\gamma}_S(h)\}$ and $\{\underline{\gamma}_N(h)\}$

satisfy

$$\underline{\gamma}_X(h) = \underline{\gamma}_S(h) + \underline{\gamma}_N(h)$$

for all integers h. The formulas of Theorem 1 are given in terms of $\{\underline{\gamma}_X(h)\}$, but when unobserved components models are fitted (Harvey, 1989), we typically obtain fitted models for both $\{\mathbf{S}_t\}$ and $\{\mathbf{N}_t\}$ (the discussion generalizes to more than two latent components). If these latent processes follow VARMA models, we can compute the individual autocovariance functions $\{\underline{\gamma}_S(h)\}$ and $\{\underline{\gamma}_N(h)\}$, which we then sum to get $\{\underline{\gamma}_X(h)\}$. This only allows us to forecast the aggregate process; forecasting a latent component itself requires signal extraction formulas, as discussed in McElroy and Trimbur (2012).

Next we discuss speedy computation of differencing matrices. Let L_T denote a $T \times T$-dimensional lag matrix, which takes the value 1 on the first subdiagonal but is zero elsewhere. Matrix powers of this lag matrix shift the unit subdiagonal further down, whereas the zeroth power is the identity matrix. We then define the matrix polynomial $\underline{A}(L_T)$ analogously to the univariate construction. If $\underline{A}(x)$ is a degree p polynomial as described in Section 2, then

$$\underline{A}(L_T) = \sum_{k=0}^{p} L_T^k \otimes \underline{a}_k,$$

which is $nT \times nT$ dimensional. The rule for multiplying two such expressions is

$$\underline{A}(L_T) \cdot \underline{C}(L_T) = (\underline{A} * \underline{C})(L_T), \tag{12}$$

where $*$ denotes the convolution of functions. In other words, we compute the product of polynomials $\underline{A}(x)$ and $\underline{C}(x)$ and evaluate at $x = L_T$ (this is derived in the appendix). As a corollary, one can show that

$$[\underline{A}(L_T)]^{-1} = \underline{A}^{-1}(L_T),$$

where $\underline{A}^{-1}(x)$ is the inverse of the matrix polynomial $\underline{a}(x)$. Such an inverse need not always exist, but if $\underline{a}_0$ is invertible, then the coefficients $\underline{b}_k$ of $\underline{A}^{-1}(x)$ can be recursively computed via

$$\underline{b}_k = -\underline{a}_0^{-1} \sum_{j=1}^{k} \underline{a}_j \underline{b}_{k-j}.$$

This treatment can be applied to the case of the differencing matrix polynomial $\underline{\Delta}(x)$, which has inverse matrix power series $\underline{\Xi}(x)$, with coefficients $\underline{\xi}_k$ given by the above type of recursion. This works because $\underline{\delta}_0$ is invertible, being the identity matrix. Hence, we can quickly compute $\underline{\Xi}(x)$ and set

$$[\underline{\Delta}(L_T)]^{-1} = \underline{\Xi}(L_T).$$

Now it is easy to see that

$$\Delta = \begin{bmatrix} \underline{I}_{np} & 0 \\ \begin{bmatrix} 0 & \underline{I}_{n(T-p)} \end{bmatrix} & \underline{\Delta}(L_T) \end{bmatrix} = \begin{bmatrix} D^{-1} & 0 \\ 0 & \underline{I}_{n(T-p)} \end{bmatrix} \underline{\Delta}(L_T),$$

where D is the upper-left $np \times np$ block of $\underline{\Delta}(L_T)$. Thus

$$\Delta^{-1} = \underline{\Xi}(L_T) \begin{bmatrix} D & 0 \\ 0 & \underline{I}_{n(T-p)} \end{bmatrix} \tag{13}$$

gives a fast method for computing the inverse of Δ.

3.3. Forecasting

Suppose that the data go up to time N and we want to forecast over the subsequent $T - N$ periods. Then $J = [\underline{I}_{nN}\ 0]$ and $K = [0\ \underline{I}_{n(T-N)}]$ and hence stacking J on top of K yields $\underline{I}_{nT}$. Also, it is elementary that $J\Delta^{-1} = [\Delta_N^{-1}\ 0]$ due to the block lower triangular shape of Δ^{-1}, where Δ_N is the upper nN rows and

columns of Δ. Then we can take $R = \Delta_N$ in our projection results, and $RJ\Delta^{-1} = [\underline{I}_{nN}\ 0]$, implying that $\underline{B} = [\underline{I}_{n(N-p)}\ 0]$, a matrix with $n(T-p)$ columns. Then the forecast formula is

$$K\Delta^{-1}\begin{bmatrix}\mathbf{X}_* \\ \underline{\Gamma}\,[\underline{I}_{n(N-p)}\ 0]'\left\{[\underline{I}_{n(N-p)}\ 0]\,\underline{\Gamma}\,[\underline{I}_{n(N-p)}\ 0]'\right\}^{-1}\mathbf{W}\end{bmatrix}$$

$$+K\,\Delta^{-1}\begin{bmatrix}0 \\ \left(\underline{I}_{n(T-p)} - \underline{\Gamma}\,[\underline{I}_{n(N-p)}\ 0]'\left\{[\underline{I}_{n(N-p)}\ 0]\,\underline{\Gamma}\,[\underline{I}_{n(N-p)}\ 0]'\right\}^{-1}\right)\ [\iota\otimes\mathbf{m}]\end{bmatrix}$$

where ι is a column vector of ones of length $T-p$. This type of formula is trivial to encode, and the matrix inversions are fast even when nT is as large as 1,000, which is sufficient for many macroeconomic applications of interest.

Two examples of the application of this forecasting formula are given in the next section, although we also provide an application of informed backcasting. But we should also point out that the above formulas are simple extensions of familiar ordinary least squares (OLS) forecasting formulas. In the special case of a VAR(p), the model can be easily fit using OLS (Lütkepohl, 2006) and forecast by taking linear combinations of the past p observations. If the data are also differenced, the differencing operator can be merged with the VAR polynomial. Alternatively, an unstable VAR may be fitted to the data; in this case, no differencing is done beforehand, but unit roots are found in the estimated polynomial. Our formulas include all these scenarios as special cases. Suppose that forecasts from a stable VAR(p) are desired H steps ahead. Writing

$$\mathbf{Y}_t = [\mathbf{X}'_{t-N+1},\ldots,\mathbf{X}'_{t-1},\mathbf{X}'_t]',$$

the VAR(p) becomes a VAR(1) in terms of $\{\mathbf{Y}_t\}$, with transition matrix

$$A = \begin{bmatrix}0 & \underline{I}_n & 0 & \cdots & \cdots & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -\underline{a}_p & \cdots & -\underline{a}_1\end{bmatrix}.$$

The full VAR(p) polynomial is $\underline{A}(x) = \sum_{j=0}^{p}\underline{a}_j x^j$; note that our sign convention differs from that of Lütkepohl (2006) here. Ignoring mean effects for simplicity and omitting K, the forecast formula above can be written as

$$\underline{\Gamma}\,[\underline{I}_{n(N-p)}\ 0]'\left\{[\underline{I}_{n(N-p)}\ 0]\,\underline{\Gamma}\,[\underline{I}_{n(N-p)}\ 0]'\right\}^{-1}\mathbf{X} = \begin{bmatrix}\mathrm{Cov}(\mathbf{Y}_N,\mathbf{Y}_N) \\ P\,\mathrm{Cov}(\mathbf{Y}_{N+1},\mathbf{Y}_N) \\ \vdots \\ P\,\mathrm{Cov}(\mathbf{Y}_{N+H},\mathbf{Y}_N)\end{bmatrix}[\mathrm{Cov}(\mathbf{Y}_N,\mathbf{Y}_N)]^{-1}\,\mathbf{X},$$

where $P = [0\ 0\ \cdots\ \underline{I}_n]$. Since $\mathrm{Cov}(\mathbf{Y}_{N+h},\mathbf{Y}_N) = A^h\,\mathrm{Cov}(\mathbf{Y}_N,\mathbf{Y}_N)$, the above simplifies to

$$\begin{bmatrix}\underline{I}_{nN} \\ PA \\ \vdots \\ PA^H\end{bmatrix}\mathbf{X}; \tag{14}$$

this is the familiar formula given in Lütkepohl (2006). If instead we have an unstable VAR(p), we can apply the formulas with $\underline{\Delta}(B)$ equal to the full VAR polynomial $\underline{A}(B)$ and $\{\mathbf{W}_t\}$ given as multivariate white noise. Then our forecasting formula becomes (again, omitting K and the mean $\mathbf{m}$ for simplicity)

$$\Delta^{-1}\begin{bmatrix}\mathbf{X}_* \\ \mathbf{W} \\ 0\end{bmatrix} = \underline{\Xi}(L_T)\begin{bmatrix}\underline{\Delta}(L_N) \\ 0\end{bmatrix}\mathbf{X}, \tag{15}$$

using (13). Here the zero matrix above has nH rows and nN columns. The appendix shows that (15) reduces to the familiar (14) formula.

Our general formulas clearly encompass those associated with stable/unstable VAR(p) representations. The next section provides an illustration for forecasting a co-integrated process that does not have a VAR(p) representation and hence requires us to use the general formulas.

4. Application and discussion

The formulas of the previous section are quite general and allow for a diverse set of applications. We begin with a somewhat common forecasting application. We then discuss three more interesting cases: backcasting disaggregated data, forecasting when long-memory is present, and forecasting series that are co-integrated, where the co-integration is formulated in terms of reduced rank latent processes (such as trends and seasonals). Except for the long-memory example, each of these applications could be handled via SS methods. The chief advantage of matrix formulas in such cases lies in the ease of implementation of the formulas of Theorem 1.

4.1. Forecasting euro area series

Our first application uses a VAR to model quarterly euro area real gross domestic product (RGDP), the Harmonised Index of Consumer Prices (HICP), and the unemployment rate (UR), from 1995:Q1 through 2011:Q4. Whereas the first two series exhibit strong trend growth and plausibly are $I(1)$, the UR should be stationary (in the long term). Hence, we use a single differencing for RGDP and the HICP (after a log transform for both), but no differencing for the UR. These specifications were carefully checked by exploratory analysis (regression of the HICP onto a line was not sufficient to remove all nonstationarity). The resulting stationary series were modeled as a VAR(4) based on using AIC to choose the lag length with a maximum lag set to 5.

We considered fitting the models via both OLS and the Yule–Walker (YW) methods, with fairly similar results for both methods. Since the mean parameter is a drift for the first two series, it has a tremendous impact on long-term forecasts. For the YW method, we compute the sample mean (Lütkepohl, 2006), whereas for OLS we include regression onto a constant to estimate the mean, as is commonly done. Because we feel confident that the differenced series are stationary, we choose the YW method as it guarantees a stable fitted VAR(k) (Lütkepohl, 2006). The estimated YW means were 0.00397, 0.00522, and 9.31493 for RDGP (in differences), the HICP (in differences), and UR, respectively.

The data length was 68, and the forecast horizon was set to 50. To compute the autocovariances needed in the forecasting formulas, the VAR(4) is imbedded as a 12-dimensional VAR(1) as described in Lütkepohl (2006), and the YW recursions are used to compute exact expressions. The R code for the forecasts (along with their MSEs) finished in less than a second (in less time than it took to fit the models), so for series this short, computation time is not an issue. All 50 forecasts are generated at once, along with the full error covariance matrix, whose diagonal entries are the forecast MSEs. Taking the $\pm$ twice the square root of these MSEs forms an envelope (pointwise confidence intervals) around the forecasts, depicted in shaded colors in the following plots. Black denotes the original series with forecasts appended (where the colored lines begin denotes the start of the forecast period; see Fig. 1).

To contrast with purely univariate techniques, we also fitted high-order AR models to the differenced series and then applied the same forecast formulas. More lags seemed necessary to explain the series when the cross-series information was not used. We arrived at an AR(10) model for each series (a lower-order AR model would also be satisfactory, but for illustrative purposes we keep the same order for all three series). The univariate forecasts in Fig. 1 are shown as dashed lines with a slightly lighter shade to depict the pointwise confidence intervals. While there is no discernible difference for the HICP, there is very different behavior for the RGDP and UR series. Also, for RGDP the univariate confidence intervals happen to be wider. Note that while the RGDP and HICP series exhibit fairly typical forecasts for $I(1)$ series—with expanding confidence intervals—the UR forecasts are interesting in their initial curvature

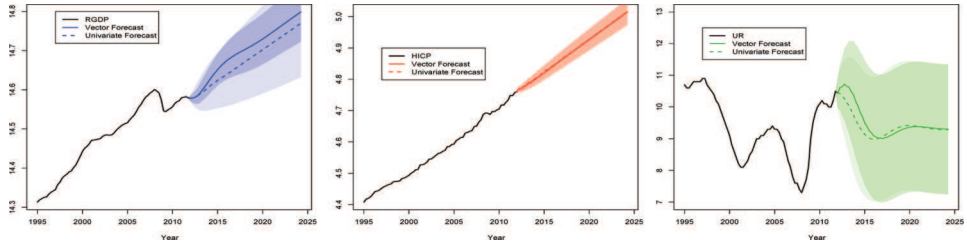

Figure 1. Euro area series RGDP (left), HICP (middle), and UR (right). Series values are in black and forecasts are colored (blue, red, green), with confidence intervals shaded. Dashed lines depict forecasts from a univariate AR(10) model, while the solid colored lines depict forecasts from the fitted VAR(4).

and stabilization (due to being $I(0)$), with confidence intervals that have essentially converged to their asymptotic value by the 50th forecast.

4.2. Backcasting fertility series

The second application involves fertility rate data (total births divided by population, by year) from the U.S. Census Bureau stratified into different racial categories. From 1980 to 2003 the racial categories were Hispanic (H), Non-Hispanic Black (NHB), and Non-Hispanic Non-Black (NHNB). Starting in 1989, birth data were collected based on five categories and emanate from a slightly different source. For 1989 through 2009, the categories are Non-Hispanic White (NHW), NHB, Hispanic (H), Non-Hispanic American Indian or Alaskan Native (NHI), and Non-Hispanic Asian or Pacific Islander (NHA). Intuitively it seems that the old NHNB category should include the new categories of NHW, NHI, and NHA. However, for the 1989–2003 overlap period the sum of the latter three race categories does *not* sum to the NHNB category because the data consist of rates. Because the NHI and NHA categories are not available for the 1980–1988 span, our objective is to backcast these series to this time period.

It seems sensible to use older-vintage fertility series if they have a clear correlation with the NHI and NHA data. After some initial exploratory analysis, it is apparent that the NHB series in its older (1980–2003) and newer (1989–2009) vintages shares some coherence with the NHI and NHA series; see the top left panel of Fig. 2. Note the imperfect matchup of the older and newer vintages of NHB in their overlap period (1989–2003), which is due to differences in data collection. As previously, we used AIC to choose the lag length with a maximum of 5 lags. The final model for forecasting was a trivariate VAR(1), with mean parameters 2.2106, 2.2349, and 1.8487.

We use the NHI, NHA, and old NHB categories for the projection calculations. Although the model was fitted with the new NHB category, we backcast using the old NHB. (We also could have fitted the model using the NHI, NHA, and old NHB categories by constructing a likelihood with missing values. Given the short length of the series, we avoided this approach.) For the series $\{\mathbf{X}_t\}$ with its components defined in this order, and with $t = 1980$ through $t = 2009$, the observation matrix J has the following structure:

$$J = \begin{bmatrix} \underline{I}_9 \otimes [1\ 0\ 0] & 0 & 0 \\ 0 & \underline{I}_{15} \otimes \underline{I}_3 & 0 \\ 0 & 0 & \underline{I}_6 \otimes \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \end{bmatrix}.$$

In other words, the first nine rows feature only old NHB data, followed by 15 groups of three rows of all series (the common period), and completed by 6 pairs of two rows for just the NHI and NHA series. Our interest focuses on the early period, so

$$K = \left[\underline{I}_9 \otimes \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \right].$$

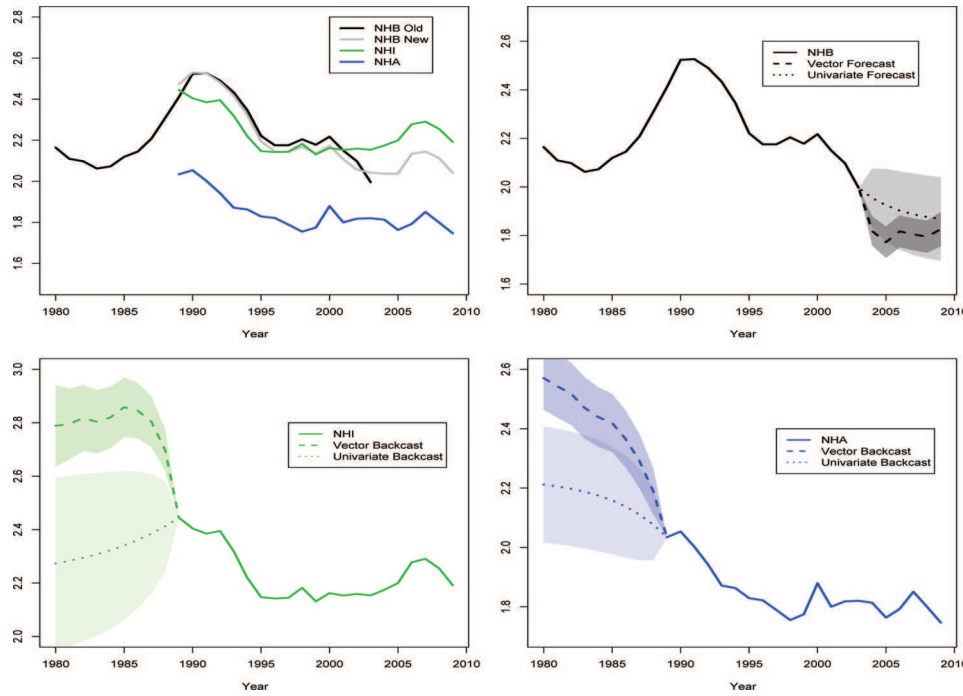

Figure 2. Fertility rates backcasted by racial group. Top left panel has NHB, NHI, and NHA series, with both an older and newer vintage for the NHB overlaid. Top right panel contains forecasts of the older NHB vintage, with error bands. Bottom panels have backcasts of NHI (left) and NHA (right) with error intervals. Dashed lines are trivariate forecasts, dotted lines are univariate forecasts.

However, we also produce forecasts of the old NHB in the top right panel of Fig. 2 for comparison with the new NHB data for that period; the K for this objective is just $\underline{I}_6 \otimes [0\,0\,1]$. Because there is no nonstationarity, the formulas for backcasting simplify dramatically. The projection is given by

$$K\,\underline{\Gamma}\,J'\left(J\,\underline{\Gamma}\,J'\right)^{-1}J\mathbf{X} + K\left[\underline{I}_{90} - \underline{\Gamma}\,J'\left(J\,\underline{\Gamma}\,J'\right)^{-1}J\right][\iota \otimes \mathbf{m}],$$

and the covariance of the error process is

$$K\left[\underline{\Gamma} - \underline{\Gamma}\,J'\left(J\,\underline{\Gamma}\,J'\right)^{-1}J\underline{\Gamma}\right]K'.$$

In the bottom panels of Fig. 2 we plot the backcasts (dashed lines) of NHI (left) and NHA (right), along with error bands derived from using ± 2 standard errors (the square root of the diagonals of the error covariance matrix). Both backcasts have an initial increase that tapers in the more distant past, which is due to the downward effect of the past NHB data. However, the forecasts for these two series do not closely track the NHB shape in the early 1980s because the coherence between the three series is not close to full. We also fitted univariate AR(1) models to the series; the resulting forecasts are plotted in Fig. 2 as dotted lines. The discrepancy between the univariate and trivariate forecasts is quite dramatic in this case.

4.3. *Forecasting vector long-memory inflation data*

We now discuss a bivariate data analysis of long-memory inflation series studied by Sela and Hurvich (2009). In their article, using data from February 1956–January 2008, inflation rates associated with Goods and Services were fitted with both a FIVAR and VARFI model (see below). In the FIVAR

model (of order one) for demeaned growth rate data $\{\mathbf{X}_t\}$, $(\underline{I}_n - \underline{a}_1 B)\,\underline{D}(B)\mathbf{X}_t$ is vector white noise with covariance $\underline{\sigma}$, where $\underline{D}(B)$ is a diagonal matrix power series consisting of the scalar power series $(1-B)^{d_j}$ for $j = 1, 2$. The parameters d_1 and d_2 are the long-memory fractional exponents for each series (although their effects become commingled in the autocovariance functions, as described by Sela and Hurvich (2009). Algebraic expressions for the power series coefficients are available (see Brockwell and Davis, 1991). The matrix polynomial $\underline{A}(B) = \underline{I}_n - \underline{a}_1 B$ corresponds to the VAR(1) portion of the model. For the VARFI model, in contrast, $\underline{D}(B)\,(\underline{I}_n - \underline{a}_1 B)\mathbf{X}_t$ is vector white noise with covariance $\underline{\sigma}$. The only difference is the order of the long-memory and VAR operators. We use the maximum likelihood estimates in Sela and Hurvich (2009), which yield 0.228, 0.477 for the long-memory FIVAR parameters and 0, 0.484 for the VARFI parameters.

Here we derive the semi-infinite forecast filters given by (4) explicitly. One advantage of these formulas is fast computation since we only need to determine the Wold coefficients $\underline{\psi}_j$ rather than the autocovariances (as would be needed for finite-sample predictors), which require considerably more programming and computation time (Sela and Hurvich, 2009). Because the sample is long, we compute the prediction filter coefficients up to a large index, at which point they become negligible (about index 200 in this case). Write $D(B) = \sum_{j\geq 0} \underline{b}_j(d)B^j$ with $\underline{b}_j(d)$ a diagonal matrix with entries $\Gamma(j-d_k)/[\Gamma(j+1)\Gamma(-d_k)]$ for the kth diagonal, where Γ is the gamma function. Also we have $(1-\underline{a}_1 B)^{-1} = \sum_{j\geq 0} \underline{a}_1^j B^j$. Then for the FIVAR model, $\underline{\Psi}(B) = \underline{D}^{-1}(B)\,\underline{A}^{-1}(B)$ has coefficients $\underline{\psi}_k = \sum_{j=0}^{k} \underline{b}_j(-d)\,\underline{a}_1^{k-j}$. Moreover, the jth coefficient of $\underline{\Psi}^{-1}(B)$ is $\underline{b}_j(d) - \underline{a}_1\,\underline{b}_{j-1}(d)$, with the convention that $\underline{b}_{-1}(d)$ is set equal to the zero matrix. Finally, the h-step ahead forecast filter for the FIVAR has the ℓth coefficient

$$\sum_{k=0}^{\ell} \underline{\psi}_{k+h}\,(\underline{b}_{\ell-k}(d) - \underline{a}_1\,\underline{b}_{\ell-1-k}).$$

For computation of the coefficients $\underline{\psi}_{k+h}$, we can use the special structure of the VAR(1) to obtain the recursion $\underline{\psi}_{k+h} = \underline{b}_{k+h}(-d) + \underline{\psi}_{k+h-1}\,\underline{a}_1$. For any $h \geq 1$, this would be initialized with $\underline{\psi}_h = \sum_{j=0}^{h} \underline{b}_j(-d)\underline{a}_1^{h-j}$. For the VARFI model $\underline{\Psi}(B) = \underline{A}^{-1}(B)\,\underline{D}^{-1}(B)$ has coefficients $\underline{\psi}_k = \sum_{j=0}^{k} \underline{a}_1^{k-j}\,\underline{b}_j(-d)$. Moreover, the jth coefficient of $\underline{\Psi}^{-1}(B)$ is $\underline{b}_j(d) - \underline{b}_{j-1}(d)\,\underline{a}_1$, and the h-step ahead forecast filter has the ℓth coefficient

$$\sum_{k=0}^{\ell} \underline{\psi}_{k+h}\,(\underline{b}_{\ell-k}(d) - \underline{b}_{\ell-1-k}\,\underline{a}_1).$$

The recursion $\underline{\psi}_{k+h} = \underline{b}_{k+h}(-d) + \underline{a}_1\,\underline{\psi}_{k+h-1}$ can be used.

In this manner, we computed the predictors for leads $h = 1, 2, \ldots, 76$; these leads correspond to data, obtained from the FRED database, that have become available since publication of Sela and Hurvich (2009). The data have also been revised, which slightly modifies some values in the pre-2008 period. We apply the forecasts to the original data (and the parameter estimates are from the original, unrevised data) and check performance against the available figures starting in February 2008. This out-of-sample exercise replicates a practical scenario closely (i.e., we pretend only data at the time of January 2008 are available). The resulting forecasts (adding the estimated mean) are plotted against the observed values, with confidence intervals constructed using two standard errors (Fig. 3). The forecast MSE matrix is computed using Remark 1 immediately after Proposition 1.

We display results for only the recent portion of the series for better visualization of the forecast paths. The dip in services-based inflation, as a result of the Great Recession was not anticipated by the forecasts, which gradually converge to the sample mean. The high uncertainty in the forecasts seems appropriate given the subsequent shifts in the series after 2008.

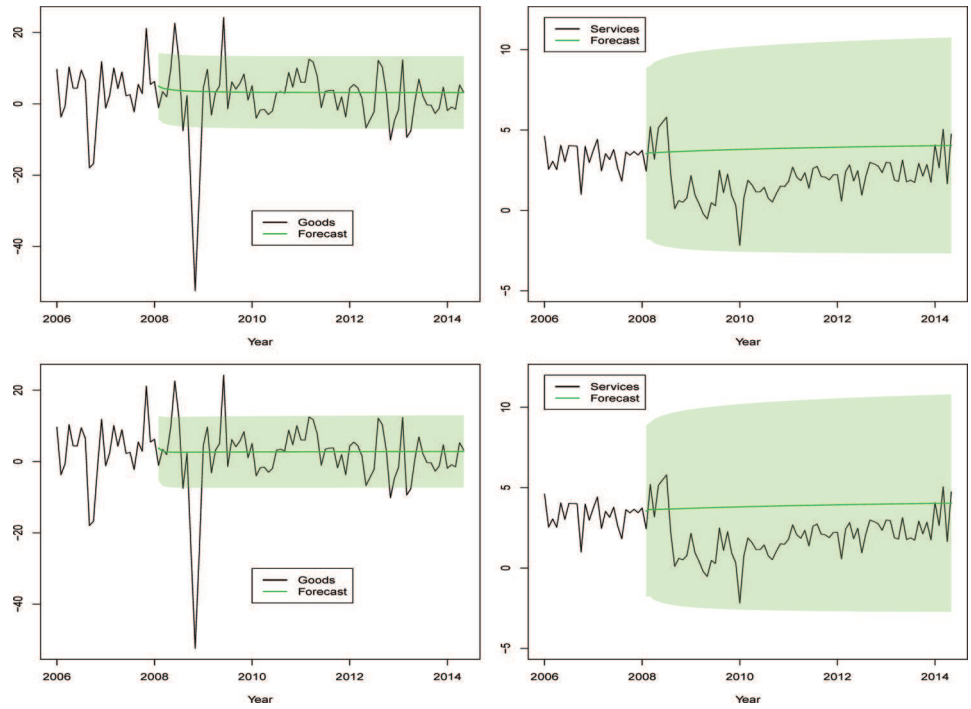

Figure 3. Forecast of inflation growth rates, 76 steps ahead. Upper panels correspond to the FIVAR model, lower panels correspond to the VARFI model. Left panels correspond to Goods, and right panels correspond to Services.

4.4. Forecasting seasonally co-integrated housing starts

In this section, we discuss forecasting regional housing starts using a seasonally cointegrated model. Classically, the concept of co-integration can be formulated as follows: The spectral density matrix of the differenced (stationary) process is singular at frequency zero. Stock and Watson (1988) showed how this can arise when the data process is viewed as the sum of a latent integrated process with collinear innovations, plus an independent stationary component. More generally, co-integration at nonzero frequencies can arise from latent nonstationary processes with collinear innovations. We propose to investigate seasonal and non-seasonal co-integration in housing series stratified by geographical region. The aggregate process will have a non-invertible Wold decomposition, so we cannot apply the forecasting techniques of Section 2, but we are free to apply the finite-sample methods of Section 3. In this case, the requisite autocovariance matrices for the aggregate data process are obtained from the component model (described in Section 3.2). The novelty here is in demonstrating forecasting from seasonally co-integrated models. (This can also be done in SS, but here we use exact formulas.)

Our data consists of monthly housing starts for both the South and West regions of the United States from 1992 through 2006 (with fixed effect removed by prior analysis in X-12-ARIMA). The housing slump of the Great Recession is avoided (although some downturn is already evident in 2006). Initial exploratory analysis of the data shows that similar seasonal and trend patterns exist in both series, so we consider unobserved component model specifications that allow for cross-correlation. To capture a slow-moving trend we specify the smooth trend model for the unobserved trend component $\{\boldsymbol{\tau}_t\}$, and a basic structural seasonal model for the seasonal $\{\boldsymbol{\xi}_t\}$, plus a white noise irregular $\{\boldsymbol{\iota}_t\}$. The monthly seasonal latent process has differencing operator $U(B) = 1 + B + \cdots + B^{11}$, and the trend latent process has differencing operator $(1 - B)^2$. (There is some evidence from the data that $1 - B$ suffices as well, but we adopt the $I(2)$ specification to make the resulting forecasts more visually apparent, purely for illustrative purposes.) The innovations for trend and seasonal are denoted $\{\boldsymbol{\eta}_t\}$ and $\{\boldsymbol{\nu}_t\}$, respectively.

We do not explicitly include a cycle component. Our model for the differenced series is

$$(1-B)^2U(B)\mathbf{X}_t = U(B)\boldsymbol{\eta}_t + (1-B)^2\boldsymbol{\nu}_t + (1-B)^2U(B)\boldsymbol{\iota}_t,$$

where $\{\boldsymbol{\eta}_t, \boldsymbol{\nu}_t, \boldsymbol{\iota}_t\}$ are each bivariate white noise sequences uncorrelated with each other, but with lag zero covariance matrices $\Sigma_{\boldsymbol{\eta}}$, $\Sigma_{\boldsymbol{\nu}}$, and $\Sigma_{\boldsymbol{\iota}}$. Generalizing the approach of Stock and Watson (1988), we obtain a related component (either of trend, seasonal, or irregular) when both eigenvalues of Σ are positive. However, if one eigenvalue is zero, then collinearity of the corresponding disturbances exists and the component is called common. In such a case, the covariance matrix can be rewritten as the outer product $[1, \vartheta]'[1, \vartheta]\sigma^2$, and only the two parameters ϑ, σ^2 need to be estimated. We say a component is unrelated if its disturbance covariance matrix Σ is diagonal.

Because there are distinct unit roots (at trend and seasonal frequencies), the topic of co-integration becomes more complex. Any common component (i.e., a component with collinear innovations represented as a dynamic factor model) has a nontrivial left null space, so there exist row vectors such that multiplication on the left yields zero for that component. Although we would prefer to call such row vectors co-integrating relations (corresponding to appropriate unit root frequencies), they only serve to reduce the order of nonstationarity by a degree. For example, a trend co-integrating row vector β has the property that $\beta\boldsymbol{\tau}_t = 0$ for all t. But since $\beta\boldsymbol{\xi}_t \neq 0$ with probability 1, seasonal nonstationarity remains: Application of the trend co-integrating removes the trend nonstationarity, but seasonality remains. (Moreover, only stochastic trend behavior is removed by β; deterministic trends will still remain.) When the left null space for trend and seasonal components have nontrivial intersection, we can find co-integrating vectors that remove both trend and seasonal nonstationarity.

Our data analysis began with the most general model, where all three components are allowed to be related (so the two components of each of the three disturbances can have any correlation between -1 and 1). The log likelihood times -2 was 700.4995 in this case, and the trend and seasonal disturbance

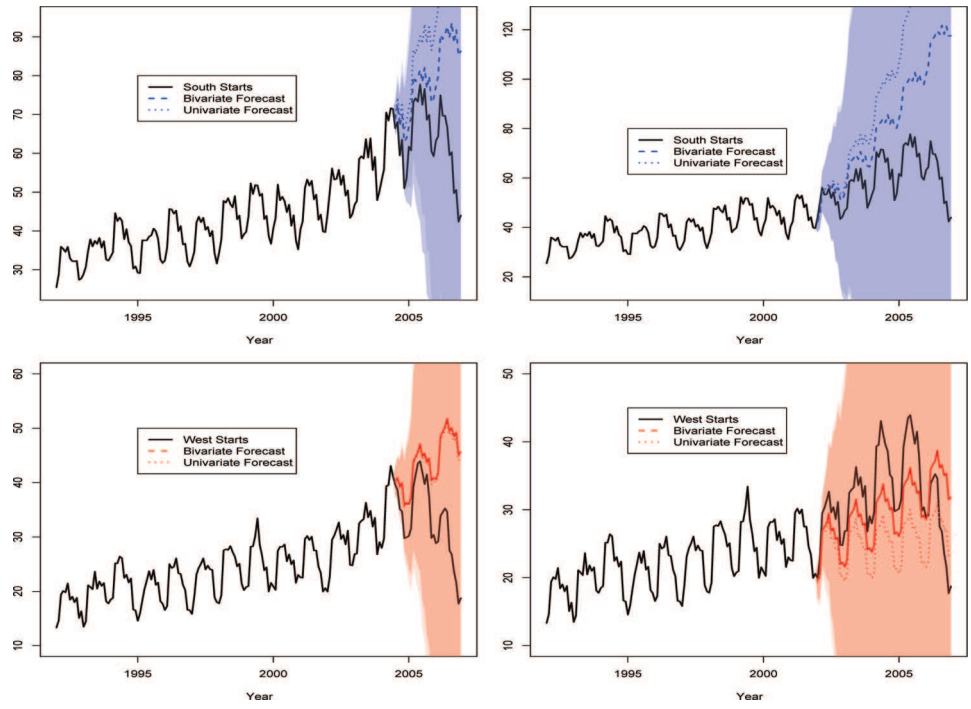

Figure 4. Housing starts series for the South (top) and West (bottom). Forecasts based on withholding the last 30 observations are on the left, and withholding 60 observations on the right. Series values are in black and forecasts are in colored (blue or red), with confidence intervals shaded. Dashed lines are bivariate forecasts, dotted lines are univariate forecasts.

correlations were 0.981 and 0.987. As these values are fairly close to unity, it was plausible to obtain a better fit using a common trend or common seasonal model. The former yielded 777.9362 for scaled likelihood, and the latter yielded 675.8092. With the superior common seasonal model, the trend disturbance correlation was now only 0.899. The unrelated components model (which corresponds to using univariate fits of each series) yielded a scaled likelihood of 752.0634. For the common seasonal, the estimated parameters were $\vartheta = -0.0168$ and $\sigma^2 = 0.0717$. The trend disturbance variances were 0.4669 and 0.1672 (for the South and West regions, respectively), while the irregular disturbance variances were 1.0874 and 0.6821 with a correlation of 0.1527. The estimated means were 0.16698 and 0.01357.

Application of the forecasting formulas is then straightforward, having identified the differencing operator and constructed the covariance matrix directly from the parameter estimates (forecasts of components could be obtained by applying signal extraction formulas to the forecast-extended data). As an in-sample exercise, we withhold the last 30 observations of both series and use the parameters fitted to the entire data span to forecast these values (dashed lines in Fig. 4). The same exercise is repeated by withholding 60 observations. MSEs were also calculated, and the resulting confidence intervals are quite wide due to the nonstationarity. For comparison, the forecasts resulting from the univariate models are presented in Fig. 4 as dotted lines. The bivariate forecasts appear to be more conservative, while the univariate forecasts are more extreme (and less accurate) in their projections.

5. Summary

These four diverse applications demonstrate the flexibility of the theoretical results to different forecasting and projection problems. While three of the applications can also be embedded and handled in a SS formulation, our approach uses explicit formulas and easily generates the full error covariance matrix. Because all the algorithms are transparent and easily encoded, the applied statistician need not be encumbered with the costs and limitations of the SS approach. All of our work was encoded in R and runs quite quickly. For other types of problems (e.g., the long-memory inflation example) it is more important to use the formulas herein, as SS approaches become less appealing.

When time-series practitioners use models similar to those described in our examples—particularly the VAR and structural VARMA models—for small samples there is no appreciable loss in computation speed compared with SS methods when the direct formulas of Theorem 1 are used.

One limitation of this work is that we do not assess the impact of model selection on forecasting. This article is concerned primarily with the derivations of general forecast formulas and their calculations. Modeling has a tremendous impact on forecasting results but is not the focus of our research here. Future studies could examine a particular set of models using the formulas in this paper to study the impact of model misspecification on performance.

Appendix

Proof of Proposition 1. The inverse of $\underline{\Psi}(z)$ exists by assumption. First, note that

$$\begin{aligned}\widehat{\mathbf{W}}_{t+h} - \mathbf{W}_{t+h} &= \sum_{k\geq h} \underline{\psi}_k B^{k-h} \underline{\Psi}^{-1}(B)\ (\mathbf{W}_t - \mathbf{m}) - \sum_{k\geq 0} \underline{\psi}_k B^{k-h} \underline{\Psi}^{-1}(B)\mathbf{W}_t \\ &= -\sum_{k=0}^{h-1} \underline{\psi}_k B^{k-h} \underline{\Psi}^{-1}(B)\ (\mathbf{W}_t - \mathbf{m}) = \mathbf{e}_{t+h},\end{aligned}$$

which proves (5). We only need to check that this error is uncorrelated with $\mathbf{W}_s$ for all $s \leq t$. But the errors can be further written as

$$\mathbf{E}_{t+h} = -\sum_{k=0}^{h-1} \underline{\psi}_k B^k \mathbf{A}_{t+h},$$

which is a linear combination of variables $\mathbf{A}_{t+1}, \cdots, \mathbf{A}_{t+h}$. By the causal representation of $\mathbf{W}_s$, it must be uncorrelated with each of these $\mathbf{A}_{t+j}$ for $1 \leq j \leq h$, since $s \leq t$. Expression (4) for the forecast filter FRF is immediate from (3). □

Proof of Proposition 2. Consider the formula (7) and subtract $\mathbf{X}_{t+h}$ given by (6), which yields the forecast errors as follows:

$$\mathbf{E}_{t+h} = \underline{P}^{(h)}(B)\left(\widehat{\mathbf{W}}_{t+h|t} - \mathbf{W}_{t+h}\right).$$

The right-hand side is of the form $\sum_{j\geq 0} \underline{p}_j \left(\widehat{\mathbf{W}}_{t+h-j|t} - \mathbf{W}_{t+h-j}\right)$. So, whenever j is such that $t+h-j \leq t$, the corresponding term in the sum is identically zero; therefore, we only need to consider coefficient matrices such that $0 \leq j < h$. Note that

$$\underline{P}^{(h)} = \underline{\Xi}(B) - \sum_{j\geq h} [\underline{P}(B)]^j \tag{A.1}$$

is a formal expression, since the geometric series for $\underline{P}(B)$ must formally equal the inverse of $\underline{I}_n - \underline{P}(B) = \underline{\Delta}(B) = \underline{\Xi}^{-1}(B)$. Outside the unit circle this is actually valid. Now in (A.1), consider taking only coefficient matrices such that the corresponding power of B is between 0 and h, this operation annihilates the right-hand term of (A.1), since the smallest power of B would be h. This leaves the definition of $[\underline{\Xi}]_0^{h-1}(B)$. This proves that the forecast error is

$$\mathbf{E}_{t+h} = \sum_{j=0}^{h-1} \underline{\xi}_j \left(\widehat{\mathbf{W}}_{t+h-j|t} - \mathbf{W}_{t+h-j}\right).$$

Now plugging in (5) yields the first equality in (9). Logically, the stated formula for the forecast (7) is MSE optimal if this error process is uncorrelated with the available data. But this is now trivial: We see from (9) and the fact that the stationary error process (5) depends only on innovations $\mathbf{A}_{t+1}$ through $\mathbf{A}_{t+h}$, that the same holds for the nonstationary case (i.e., (9) is a linear function in the innovations $\mathbf{A}_{t+1}$ through $\mathbf{A}_{t+h}$). Each observation $\mathbf{X}_s$ for $s \leq t$ can be written in the form of (6) with $t+h$ replaced by s. The $\mathbf{W}_s$ term is then a direct function of innovations $\mathbf{A}_j$ with $j \leq s \leq t$, and hence is uncorrelated with $\mathbf{E}_{t+h}$. Likewise, only values of $\mathbf{X}_j$ with $j \leq s \leq t$ occur, which are uncorrelated with the innovations by assumption. The second equality in (9) follows from algebra:

$$\sum_{j=0}^{h-1} \underline{\xi}_j B^j [\underline{\Psi}]_0^{h-1-j}(B) = \sum_{j=0}^{h-1} \underline{\xi}_j \sum_{k=0}^{h-1-j} \underline{\psi}_k B^{k+j} = \sum_{j=0}^{h-1} \left(\sum_{k=0}^{j} \underline{\xi}_k \underline{\psi}_{j-k} \right) B^j = [\underline{\Delta}^{-1}(B)\underline{\Psi}(B)]_0^{h-1}.$$

The expression (8) for the forecast filter FRF is immediate from (7). □

Proof of Theorem 1. To compute the optimal estimate, we proceed to calculate the Gaussian conditional expectation. First, let $A = \mathrm{Cov}(\mathbf{X}_*, \mathbf{X}_*)$, and observe that

$$J\mathbf{X} = R^{-1} \begin{bmatrix} \mathbf{X}_* \\ \underline{B}\mathbf{W} \end{bmatrix}.$$

Because R is invertible, the information in $J\mathbf{X}$ is the same as the information in $\mathbf{X}_*$ and $\underline{B}\mathbf{W}$. Therefore,

$$\begin{aligned} \mathbb{E}[K\mathbf{X}|J\mathbf{X}] &= \mathbb{E}[K\mathbf{X}|\mathbf{X}_*, \underline{B}\mathbf{W}] = K\Delta^{-1}\mathbb{E}\left\{ \begin{bmatrix} \mathbf{X}_* \\ \mathbf{W} \end{bmatrix} | \mathbf{X}_*, \underline{B}\mathbf{W} \right\} \\ &= K\Delta^{-1} \left\{ \begin{bmatrix} \mathbb{E}\mathbf{X}_* \\ \mathbb{E}\mathbf{W} \end{bmatrix} + \begin{bmatrix} A & 0 \\ 0 & \underline{\Gamma B}' \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & \underline{B\Gamma B}' \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_* - \mathbb{E}\mathbf{X}_* \\ \underline{B}\mathbf{W} - \underline{B}\mathbb{E}\mathbf{W} \end{bmatrix} \right\} \end{aligned}$$

$$= K\Delta^{-1}\begin{bmatrix} \underline{I}_{np} & 0 \\ 0 & \underline{\Gamma B}'[\underline{B\Gamma B}']^{-1} \end{bmatrix}\begin{bmatrix} \mathbf{X}_* \\ \underline{B}\mathbf{W} \end{bmatrix}$$
$$+K\Delta^{-1}\begin{bmatrix} 0 \\ \left(\underline{I}_{n(T-p)} - \underline{\Gamma B}'[\underline{B\Gamma B}']^{-1}\underline{B}\right)\ [\iota \otimes \mathbf{m}] \end{bmatrix},$$

which yields the stated formula. With $C = \underline{I}_{n(T-p)} - \underline{\Gamma B}'[\underline{B\Gamma B}']^{-1}\underline{B}$, the error process is

$$\widehat{K\mathbf{X}} - K\mathbf{X} = K\Delta^{-1}\begin{bmatrix} 0 & 0 \\ 0 & -C \end{bmatrix}\begin{bmatrix} \mathbf{X}_* \\ \underline{B}\mathbf{W} \end{bmatrix} + K\Delta^{-1}\begin{bmatrix} 0 \\ C\,[\iota \otimes \mathbf{m}] \end{bmatrix},$$

which has the stated covariance matrix. □

Derivation of (12) The product expands as

$$\sum_{k=0}^{p}\sum_{\ell=0}^{q}(L_T^k \otimes \underline{a}_k)(L_T^\ell \otimes \underline{b}_\ell) = \sum_{k=0}^{p}\sum_{\ell=0}^{q}(L_T^{k+\ell} \otimes \underline{a}_k\underline{b}_\ell) = \sum_{j=0}^{p+q} L_T^j \otimes \left[\sum_i \underline{a}_i\underline{b}_{j-i}\right]$$

Equivalency of (15) and (14). First, decompose $\underline{\Xi}(L_T)$ so (15) is

$$\begin{bmatrix} \underline{\Xi}(L_N) & & 0 \\ \begin{bmatrix} \underline{\xi}_N & \cdots & \underline{\xi}_1 \\ \underline{\xi}_{N+1} & \cdots & \underline{\xi}_2 \\ \vdots & \cdots & \vdots \end{bmatrix} & & \underline{\Xi}(L_H) \end{bmatrix}\begin{bmatrix} \underline{\Delta}(L_N) \\ 0 \end{bmatrix}\mathbf{X} = \begin{bmatrix} \underline{I}_{nN} & \\ \begin{bmatrix} \underline{\xi}_N & \cdots & \underline{\xi}_1 \\ \underline{\xi}_{N+1} & \cdots & \underline{\xi}_2 \\ \vdots & \cdots & \vdots \end{bmatrix} & \underline{\Delta}(L_N) \end{bmatrix}\mathbf{X}.$$

The lower-left block matrix has nH rows and nN columns, and its hth block row is given by $[\underline{\xi}_{N-1+h}, \cdots, \underline{\xi}_h]\,\underline{\Delta}(L_N)$. We claim that $A^h\,\underline{\Xi}(L_N)$ is a matrix with jkth block entry $\underline{\xi}_{j-k+h}$, from which it follows that

$$[\underline{\xi}_{N-1+h}, \cdots, \underline{\xi}_h]\,\underline{\Delta}(L_N) = PA^h,$$

which will then match up with (14). To prove the claim, first consider $h = 1$:

$$A\,\underline{\Xi}(L_N) = \begin{bmatrix} \underline{\xi}_1 & \underline{\xi}_0 & 0 & \cdots & 0 \\ \underline{\xi}_2 & \underline{\xi}_1 & \underline{\xi}_0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \underline{\xi}_{N-1} & \underline{\xi}_{N-2} & \cdots & \cdots & \underline{\xi}_0 \\ \underline{\xi}_N & \underline{\xi}_{N-1} & \cdots & \cdots & \underline{\xi}_1 \end{bmatrix},$$

where the last block row is given by

$$\sum_{\ell=1}^{N}(-\underline{\delta}_{N-\ell+1})\,\underline{\xi}_{\ell-k} = -[\underline{\delta} * \underline{\xi}]_{N-k+1} + \underline{\delta}_0\,\underline{\xi}_{N+1-k} = \underline{\xi}_{N+1-k}$$

for the kth column because $\underline{\delta}(B)\underline{\xi}(B)$ is the identity matrix polynomial. This handles the base case $h = 1$, and the rest of the claim follows from induction by repeated application of A to the left-hand side along with the following calculation: If the claim holds for h, then the bottom row of $A^{h+1}\,\underline{\Xi}(L_N)$ is

$$\sum_{\ell=1}^{N}(-\underline{\delta}_{N-\ell+1})\,\underline{\xi}_{\ell-k+h} = -[\underline{\delta} * \underline{\xi}]_{N-k+h+1} + \underline{\delta}_0\,\underline{\xi}_{N+h+1-k} = \underline{\xi}_{N+h+1-k}.$$

This concludes the derivation.

Disclaimer

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, the Federal Reserve Bank of St. Louis, the Federal Reserve System, or the Federal Reserve Board of Governors.

References

Athanasopoulus, G., Vahid, F. (2008). VARMA versus VAR for macroeconomic forecasting. *Journal of Business and Economics Statistics* 26:237–252.

Bell, W. (1984). Signal extraction for nonstationary time series. *Annals of Statistics* 12:646–664.

Bloomfield, P. (1973). An exponential model for the spectrum of a scalar time series. *Biometrika* 60:217–226.

Brockwell, P., Davis, R. (1991). *Time Series: Theory and Methods*. 2nd ed. New York: Springer-Verlag.

Chan, N., Palma, W. (1998). State space modeling of long-memory processes. *Annals of Statistics* 26:719–740.

Engle, R., Granger, C. (1987). Cointegration and error correction: representation, estimation, and testing. *Econometrica* 55:251–276.

Granger, C. (1981). Some properties of time series data and their use in econometric model specification. *Journal of Econometrics* 16:121–130.

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

Holan, S. H., McElroy, T. S., Wu, G. (2014). The cepstral model for multivariate time series: The vector exponential model. arXiv preprint arXiv:1406.0801.

Lütkepohl, H. (1986). Forecasting vector ARMA processes with systematically missing observations. *Journal of Business and Economics Statistics* 4:375–390.

Lütkepohl, H. (1987). *Forecasting Aggregated Vector ARMA Processes*. Berlin: Springer-Verlag.

Lütkepohl, H. (2006). *New Introduction to Multiple Time Series Analysis*. Berlin: Springer-Verlag.

Maïnassara, B., Francq, C. (2011). Estimating structural VARMA models with uncorrelated but non-independent error terms. *Journal of Multivariate Analysis* 102:496–505.

McElroy, T., Findley, D. (2010). Discerning between models through multi-step ahead forecasting errors. *Journal of Statistical Planning and Inference* 140:3655–3675.

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

McElroy, T., Monsell, B. (2012). Model estimation, prediction, and signal extraction for nonstationary stock and flow time series observed at mixed frequencies. U.S. Census Bureau Research Report, RRS2012/09.

McElroy, T., Trimbur, T. (2012). Signal extraction for nonstationary multivariate time series with illustrations for trend inflation. Finance and Economics Discussion Series 2012–45, Federal Reserve Board.

Mittnik, S. (1987). Non-recursive methods for computing the coefficients of the autoregressive and the moving-average representation of mixed ARMA processes. *Economics Letters* 23:279–284.

Mittnik, S. (1990). Computation of theoretical autocovariance matrices of multivariate autoregressive moving average time series. *Journal of the Royal Statistical Society, Series B* 52:151–155.

Mittnik, S. (1993). Computing theoretical autocovariances of multivariate autoregressive moving average models by using a block Levinson method. *Journal of the Royal Statistical Society, Series B* 55:435–440.

Nyblom, J., Harvey, A. (2000). Tests of common stochastic trends. *Econometric Theory* 16:176–199.

Pai, J., Ravishanker, N. (2010). Fast Bayesian estimation for VARFIMA processes with stable errors. *Journal of Statistical Theory and Practice* 4:663–677.

Reinsel, G. (1997). *Elements of Multivariate Time Series Analysis*. New York: Springer-Verlag.

Schorfheide, F. (2005). VAR forecasting under misspecification. *Journal of Econometrics* 128:99–136.

Sela, R., Hurvich, C. (2009). Computationally efficient methods for two multivariate fractionally integrated models. *Journal of Time Series Analysis* 30:631–651.

Stock, J., Watson, M. (1988). Testing for common trends. *Journal of the American Statistical Association* 83:1097–1107.

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

Whittle, P. (1984). *Prediction and Regulation by Linear Least-Squares Methods*. 2nd ed. Oxford: Blackwell.