Electronic Journal of Statistics
ISSN: 1935-7524

Supplement to Casting Vector Time Series: Algorithms for Forecasting, Imputation, and Signal Extraction

Tucker McElroy[1]

[1] *Research and Methodology Directorate U.S. Census Bureau, 4600 Silver Hill Road, Washington, D.C. 20233 e-mail:* tucker.s.mcelroy@census.gov

Disclaimer

This paper 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 those of the U.S. Census Bureau. The statistics in this paper were approved for dissemination by the Disclosure Review Board (release number: CBDRB-FY21-CED004-001).

1. Discussion of Signal Extraction Algorithms

Some further discussion of the motivating computational issues of signal extraction are discussed here. When the sample size T is small, say 500 or less (though cross-sectional dimension N also plays a role here), non-recursive (e.g., brute force) methods based upon explicit matrix formulas (McElroy and Trimbur, 2015) can be employed to obtain signal extraction estimates. Essentially, the determination of predictors and error covariances revolves upon the calculation of partial covariances, wherein a matrix inversion is required; straight Cholesky approaches are viable when the overall matrix dimension is small – as advocated in McElroy (2008). However, with high frequency data or high-dimensional time series the direct approaches tend to fail – either outright through memory allocation violations and/or numerical stability issues associated with ill-conditioned matrices (this can happen when the process' spectral density is non-invertible, yielding at least one eigenvalue that is approximately zero when the Toeplitz autocovariance matrix is large), or through impracticable computation times, e.g., a minute for a single likelihood evaluation indicates that days or weeks may be necessary to run numerical optimization. Moreover, the matrix approach is not easily extended to handle missing values.

In order to circumvent the calculation of exact matrix formulas, the author proceeded to determine the WK filter through frequency domain calculations and apply a suitably truncated sequence to the extended data; once the forecasts

and aftcasts were computed and appended, convolution methods rendered the filtering extremely speedy. However, the signal extraction error available from the WK error (the error variance arising from the ideal case of a bi-infinite sample) only represents part of the total extraction error – one must also account for the casting errors, and their covariances (aggregated appropriately by the truncated filter) offer a substantial contribution at the sample boundary. Given these motivations, the algorithms of this paper render feasible the efficient computation of this second contribution to signal extraction error. For an *ad hoc* filter, there is no first portion (because signal is essentially defined differently) and its error cannot be described at all without this second portion arising from error covariances.

TABLE 1
Comparison of Algorithmic Methods. T is sample size, and N is series dimension.

Capability	Recursive	Matrix	SSF
Large T, moderate N	yes	no	yes
Ragged Edge	yes	partial	partial
Growth Rate Signals	yes	yes	partial
Ad hoc Filters	yes	yes	partial
Correct Initialization	yes	yes	no
Models	any acf	any acf	markov

While the new algorithms here require a substantial exertion to implement, the framework is both more flexible and less strenuous than SSF. Table 1 provides a comparison of the new recursive algorithms of this paper to direct matrix approaches (discussed in McElroy and Trimbur (2015) and McElroy and Monsell (2015)) and SSF, summarizing the introductory discussion given above. Regarding the ragged edge problem, it seems that with some custom coding the SSF can be adapted to handle such situations, although the basic implementation (Gómez, 2016) treats values as missing (at some particular time) across all variables. Similarly, linear combinations of signals (such as growth rates) can be generated from SSF with little difficulty, but the uncertainty is not obtainable without a customized Kalman filter that iterates prediction error covariances across time lags. The problem of applying *ad hoc* filters can be viewed as a generalization of computing linear combinations of a signal. As for the problems with the diffuse intialization commonly used in SSF for non-stationary processes, this has been extensively discussed in Bell and Hillmer (1991) and Gómez (2016).

The paper is written to give a comprehensive description of recursive algorithms needed for casting and signal extraction, and therefore contains a mixture of known and novel results. We summarize the main contributions below, indicating where they can be found in the paper:

- Factorization of the Gaussian divergence for difference-stationary processes: extends univariate results of McElroy and Monsell (2015) to the multivariate case, allowing for ragged edge missing values and non-scalar differencing operators (Section 2.1).

- Recursive algorithms for one-step ahead and multi-step ahead predictors for difference-stationary processes: generalizes recursions for stationary predictors given in McElroy (2018) to difference-stationary and multi-step cases (Section 3.1).
- Predictors for the past via time reversal (Section 3.2).
- Recursive casting algorithm for ragged edge difference-stationary processes: new method with minimal storage, providing both casts and casting error covariances (Section 3.3).
- Algorithms for filter MSE: the algorithm for combining a given filter (which can be *ad hoc*) with casting error covariances is given, and the WK filter formula is extended from the scalar differencing polynomial case of McElroy and Trimbur (2015) to allow for matrix differencing polynomials (Section 3.4).
- Extensions to growth rates of signals: algorithms to yield estimates and MSE for linear combinations of a desired signal are provided (Section 3.4).

2. Monthly Housing Starts

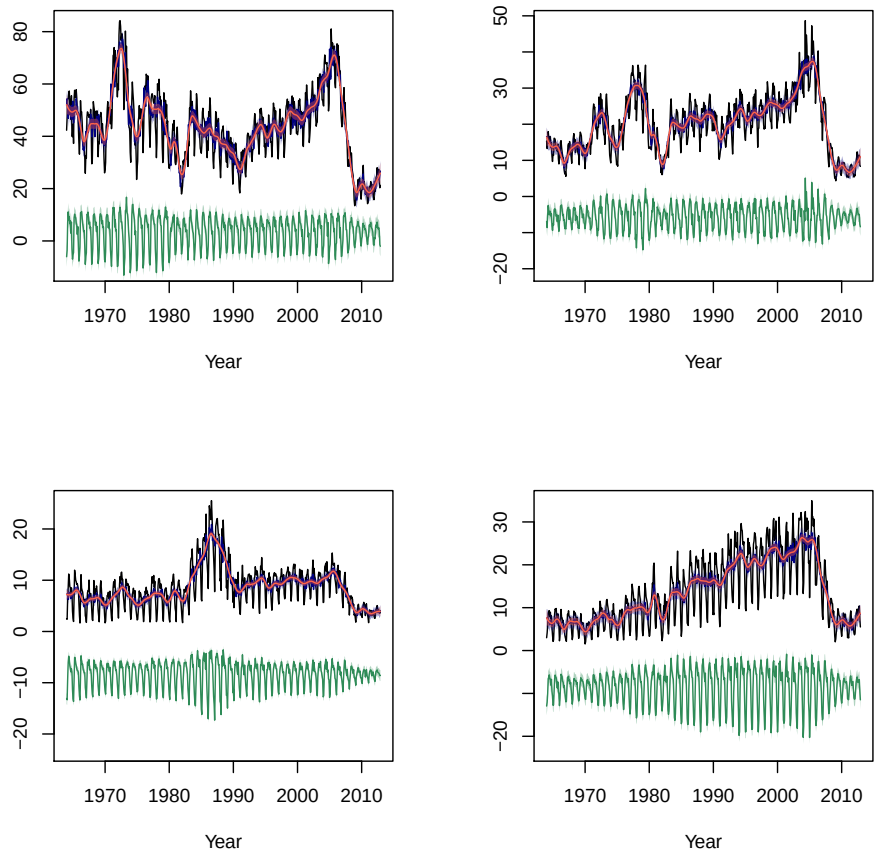

FIG 1. *Extracted components with (shaded) uncertainty bands of housing starts, for South (upper left panel), West (upper right panel), NorthEast (lower left panel), and MidWest (lower right panel). The data (black) is overlaid with trend (red), trend-irregular (blue), and seasonal (green) components; the seasonal component has been vertically displaced for easier visualization.*

Here we study "New Residential Construction (1964-2012), Housing Units Started, Single Family Units," or *housing starts* for short[1], corresponding to the four regions of South, West, NorthEast, and MidWest. Our objective with this illustration is to show that the direct matrix approach is identical to the

[1]The four series are from the Survey of Construction of the U.S. Census Bureau, available at http://www.census.gov/construction/nrc/how_the_data_are_collected/soc.html.

methods of Section 3.4, as we take the filter truncation level m larger and larger. Secondly, we show how growth rates can be generated, along with uncertainty, using either method. The direct matrix formulas for signal extraction (McElroy and Trimbur, 2015) were computed; this was feasible, because there are 49 full years of data (no missing values), covering 1964-2012, so that $T = 588$ is of moderate size.

We proceeded by fitting the structural model discussed in McElroy (2017) to the full data span (after some pre-processing for outliers), disallowing any co-integration constraints. We computed the WK filter for trend, trend-irregular (or seasonally adjusted), and seasonal components, with filter coefficients and frequency response functions computed in the manner described in Section 5 of the Supplement. With forecasts and aftcasts, we then generated the signal extractions along with uncertainty, using both the direct matrix approach and $m = 250$ casts (see Figure 1). The shading around each extraction indicates a confidence interval of plus or minus two times the square root MSE.

The time-varying MSE is increased at the sample boundaries, which is reflected by the width of the uncertainty shading being somewhat wider (though this is hard to discern in these plots). To better visualize the signal extraction standard error, we have plotted these for the seasonal adjustment extraction in Figures 2 and 3, corresponding to $m = 10$ and $m = 50$ number of aftcasts and forecasts, respectively. Whereas some discrepancies are apparent in Figure 2 between the direct matrix approach and the truncated filter approach, these discrepancies are largely reduced in Figure 3. There is virtually perfect agreement at $m = 250$ (not shown). Of course, the analyst has complete control over selecting m, the only limitations being on processing time and memory.

Finally, suppose we wish to know growth rates for the trend. According to Section 3.4, we select $\varphi(z) = I_N - I_N z$, as this corresponds to a first difference of each component time series. Utilizing *Ecce Signum*, the resulting trend growth rates (with corresponding uncertainty as shasing) are displayed in Figure 4.

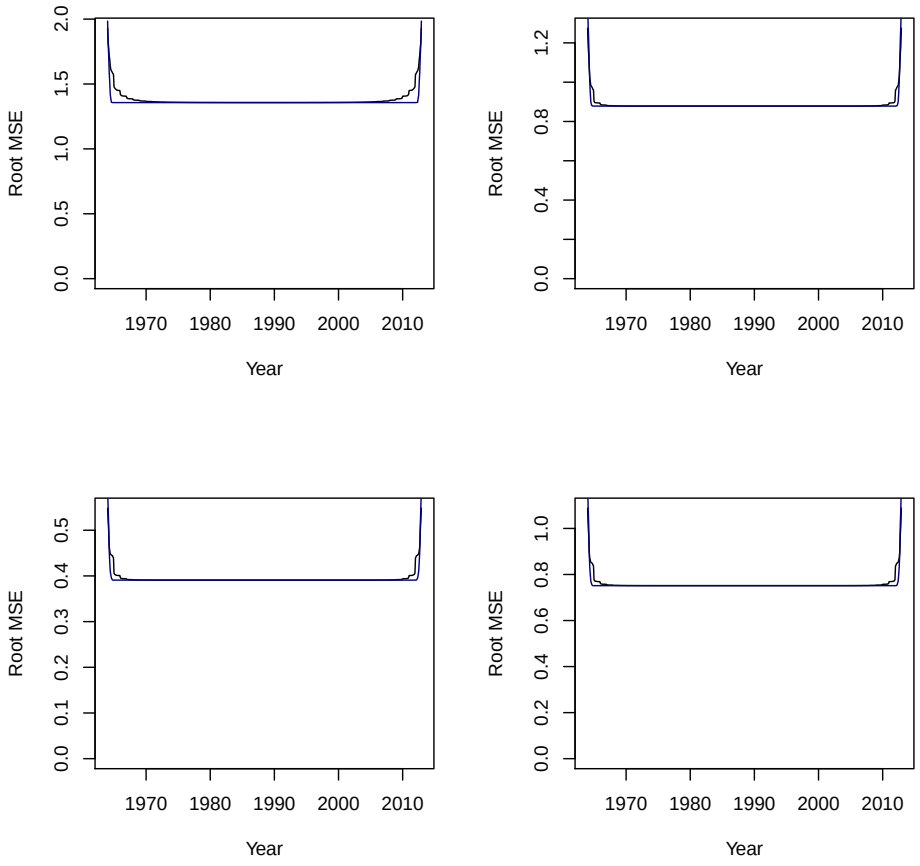

FIG 2. *Square root MSE over time for the trend-irregular component of housing starts, for South (upper left panel), West (upper right panel), NorthEast (lower left panel), and MidWest (lower right panel). The exact square root MSE (black) is compared to the approximation (blue) based on truncating the WK filter using* $m = 10$ *casts.*

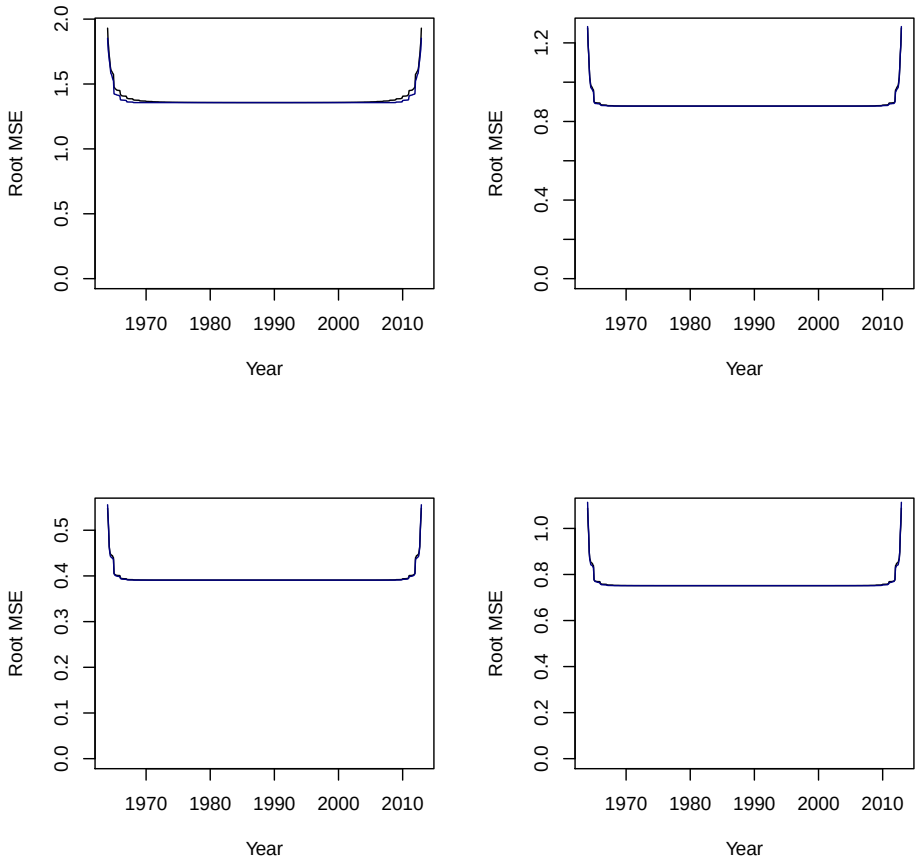

FIG 3. *Square root MSE over time for the trend-irregular component of housing starts, for South (upper left panel), West (upper right panel), NorthEast (lower left panel), and MidWest (lower right panel). The exact square root MSE (black) is compared to the approximation (blue) based on truncating the WK filter using* $m = 50$ *casts.*

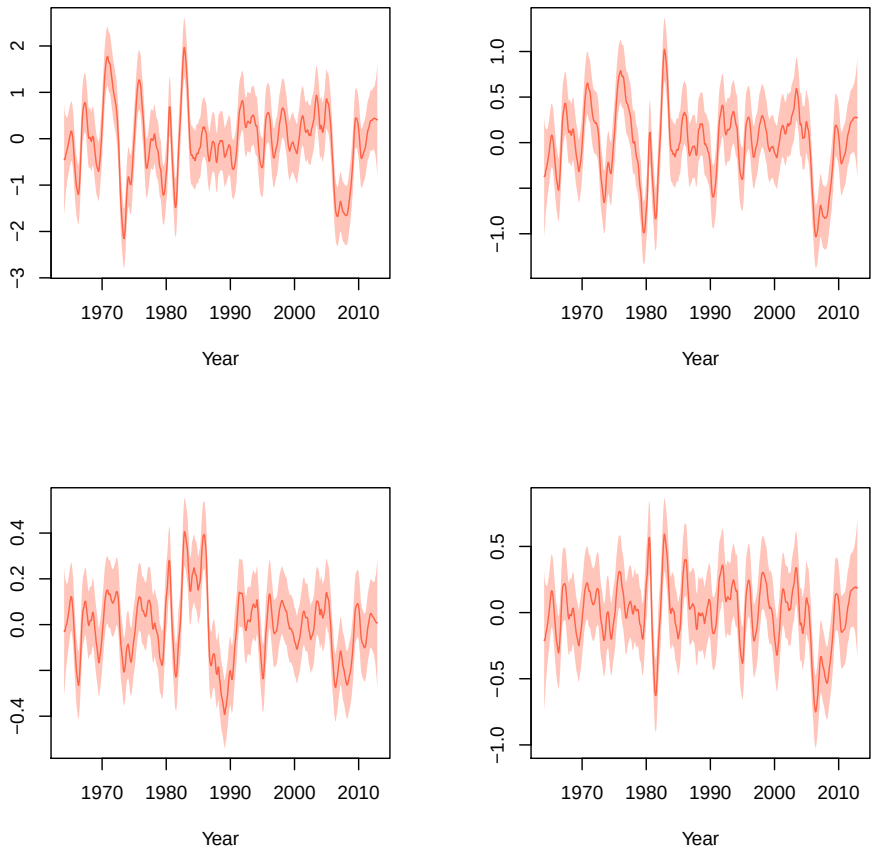

FIG 4. *Trend growth rates (red) with (shaded) uncertainty bands of housing starts, for South (upper left panel), West (upper right panel), NorthEast (lower left panel), and MidWest (lower right panel).*

3. Weekly Retail Products

The next example consists of weekly time series obtained from Dominick's Database, published by the Kilts Center for Marketing (University of Chicago School of Business). This database contains weekly store-level sales from Dominick's Finer Foods from 1989 through 1994 around the Chicago, Illinois area in the U.S.A. For one of the stores we have extracted the sales data for bathroom tissues (tti) and paper towels (ptw), and refer to the bivariate series as *products* for short. The purpose of this example is to illustrate the solution to the ragged edge problem in a low-dimensional case where it is feasible to compare to matrix-based solutions.

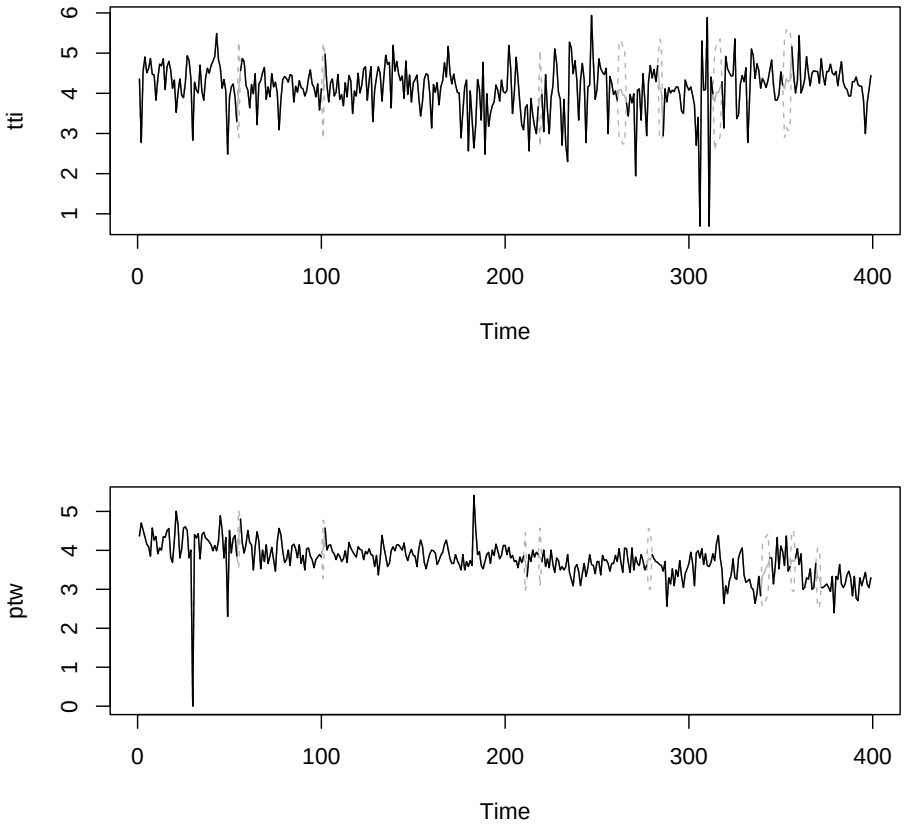

FIG 5. *Plots of bathroom tissues (tti) and paper towels (ptw) by week (black), with imputed values (solid grey) and confidence intervals (dashed grey) for NAs.*

In addition to some missing values (which are missing for both tti and ptw), there are meager values corresponding to virtually no sales activity. Following McElroy and Penny (2019), we replace any non-positive values with an NA, treating them as missing so that they can be imputed – this is akin to treating such meager values as (small) extremes, and facilitates taking log transforms of the data. However, these meager values occur in a ragged pattern, so that there are times for which only one variable is an NA. In particular these times are

$$t = 55, 101, 219, 262, 263, 264, 265, 284, 285, 314, 315, 316, 317, 352, 353, 354, 355$$

for tti and

$$t = 55, 101, 211, 219, 278, 279, 340, 341, 342, 343, 356, 357, 370, 371$$

for ptw. Based on exploratory analysis of this data, a VAR model was identified and various orders of fit were attempted. The likelihood evaluation uses the exact methods of this article, implicitly casting and evaluating throughout the optimization routine. The stable VAR parameterization of Roy, McElroy, and Linton (2019) was utilized, because the data does not display non-stationarity. The final model was of order 3 with a mean and no other regressors. The final value of the likelihood was -317.8978, which was verified by a direct matrix approach based on Section 2 – this is fairly easy to implement because there are no differencing operators. The diagnostics for the residuals indicate that the model is adequate for illustrative purposes. Finally, casts (with casting MSEs) were generated for all the NAs, and plotted in Figure 5. Available data is in black, whereas imputed NAs are solid grey, with dashed grey lines denoting a confidence interval of plus or minus two times the square root MSE.

4. Figures for Retail Shoe Data

The signal extractions for daily retail series 4482 are displayed in Figure 6. The data is in black and the seasonally adjusted component is in red. If we only suppress the weekly effect, we obtain the blue line. Uncertainty bands are also plotted, with shading, but are quite narrow and hard to discern. Figure 7 displayes sample autocorrelations of the residuals, showing that there is little serial correlation remaining in the residuals.

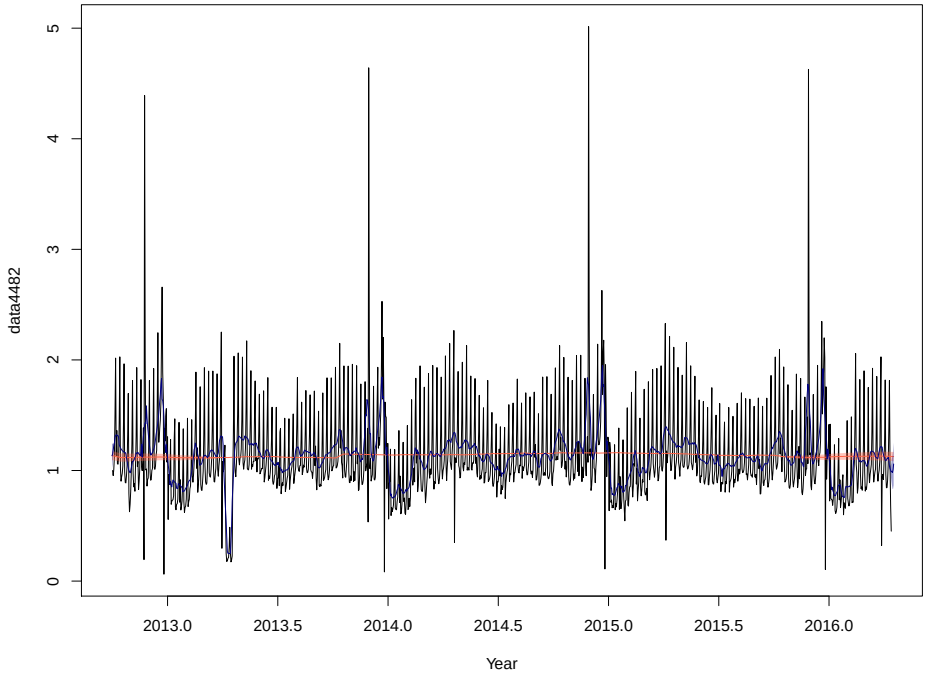

FIG 6. *Time series plot of daily retail series 4482 (Shoe stores), with seasonal adjustment (red) and non-weekly effect (blue); uncertainty bands are shaded.*

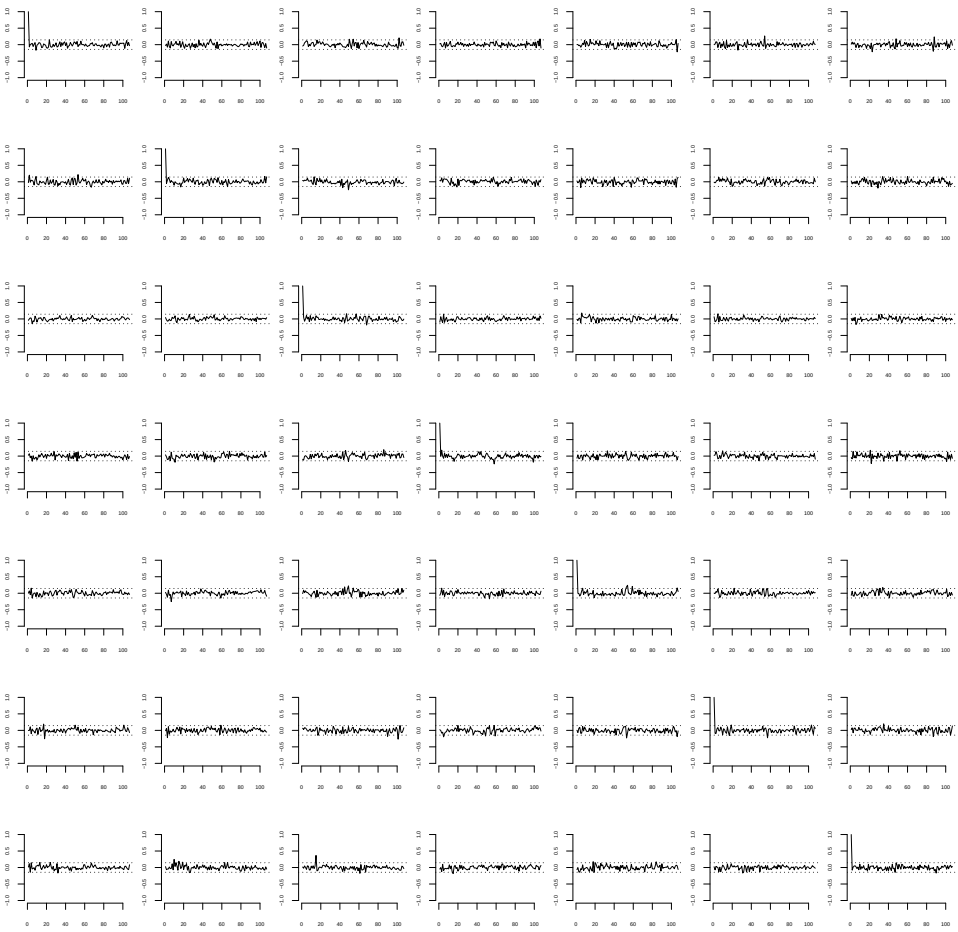

FIG 7. *Sample autocorrelation of residuals for daily retail series 4482 (Shoe stores) embedded as a weekly series. Plots give cross-autocorrelations for Sunday through Saturday (from left to right and top to bottom), with horizontal bands denoting the white noise hypothesis test.*

5. WK Computations for Structural Time Series

This section describes the calculation of frequency response functions for WK signal extraction filters. Adopting the general framework of McElroy (2017), we suppose the data process $\{x_t\}$ can be decomposed in terms of $J+2$ latent processes, where J of these processes are difference stationary with scalar differencing operators. The remaining two components correspond to a cycle $\{\rho_t\}$ and an irregular $\{\iota_t\}$, both of which are stationary. We denote the non-stationary components by $\{\chi_t^{(j)}\}$ for $1 \leq j \leq J$.

$$x_t = \sum_{j=1}^{J} \chi_t^{(j)} + \rho_t + \iota_t.$$

Let each scalar differencing operator be denoted $\delta^{(j)}(B)$; these have distinct unit roots by assumption. When differenced to stationarity, a non-stationary component is denoted $\{\underline{\chi}_t^{(j)}\}$, where $\underline{\chi}_t^{(j)} = \delta^{(j)}(B)\chi_t^{(j)}$. In order to reduce the data process $\{x_t\}$ to stationarity – necessary to evaluate the Gaussian likelihood via Durbin-Levinson algorithm – we must apply the differencing operator $\delta(B) = \prod_{j=1}^{J} \delta^{(j)}(B)$, and this is the minimal degree operator with this property. Setting $\delta^{(-j)}(B) = \prod_{k \neq j} \delta^{(k)}(B)$ (if $J = 1$, this is equal to one), we obtain

$$\underline{x}_t = \delta(B)x_t = \sum_{j=1}^{J} \delta^{(-j)}(B)\, \underline{\chi}_t^{(j)} + \delta(B)\rho_t + \delta(B)\iota_t.$$

Each latent process is driven by white noise innovations, with a covariance matrix of possibly reduced rank – though we stipulate that the irregular has full rank. These matrices are denoted by $\Sigma^{(j)}$ for the non-stationary processes, and by Σ^{ρ} and Σ^{ι} for the cycle and irregular:

$$\underline{\chi}_t^{(j)} \sim \mathrm{WN}(0, \Sigma^{(j)}) \qquad \iota_t \sim \mathrm{WN}(0, \Sigma^{\iota}).$$

The Generalized Cholesky Decomposition (GCD) for each white noise covariance matrix produces lower triangular matrices L and diagonal matrices D of column dimension r, where $r \leq N$ is the rank, such that $\Sigma = L\,D\,L'$. (See McElroy (2017) for more detail.) These matrices will be super-scripted in correspondence with each latent component. Let f_ρ denote the scalar spectral density of the autoregressive portion of the cycle, so that when multiplied by Σ^{ρ} we obtain that process' multivariate spectral density.

Given these preliminaries, we can write down formulas for the WK signal extraction frequency response functions, and examine their behavior at so-called co-integrating frequencies. An $\omega \in [-\pi, \pi]$ is a co-integrating frequency for the jth latent process if $e^{-i\omega}$ is a root of $\delta^{(j)}(B)$. The reason for the terminology is the following: latent process j is co-integrated of rank $n - r$ if and only if the space of left co-integrating vectors has dimension $n - r$ (which means that application of a co-integrating vector reduces the order of non-stationarity, up

to fixed effects, from $\delta(B)$ to $\delta^{(-j)}(B)$), which is true if and only if $\Sigma^{(j)}$ has rank r. The spectral density of $\{\underline{x}_t\}$ at a co-integrating frequency is equal to $\Sigma^{(j)}$, and hence has rank r. Hence, there are co-integrating vectors computable from the GCD of $\Sigma^{(j)}$, such that their application to the data process removes non-stationarity associated with frequency ω.

These claims are apparent once we compute the spectral density of the differenced data process:

$$f_{\underline{x}}(\lambda) = \sum_{j=1}^{J} |\delta^{(-j)}(e^{-i\lambda})|^2 \, \Sigma^{(j)} + |\delta(e^{-i\lambda})|^2 \, f_\rho(\lambda) \, \Sigma^\rho + |\delta(e^{-i\lambda})|^2 \, \Sigma^\iota.$$

Then if $\lambda^{(j)}$ is the co-integrating frequency for the jth latent process, we have $\delta^{(j)}(e^{-i\lambda^{(j)}}) = 0$, but $\delta^{(k)}(e^{-i\lambda^{(j)}}) \neq 0$ for $k \neq j$. Therefore

$$f_{\underline{x}}(\lambda^{(j)}) = |\delta^{(-j)}(e^{-i\lambda^{(j)}})|^2 \, \Sigma^{(j)},$$

which has rank $r_j \leq N$. Note that if the roots are close to one another, it is possible for $\delta^{(-j)}(e^{-i\lambda^{(j)}})$ to be close to zero.

The formulas for the WK frf in each case are obtained by application of Theorem 3.5; we denote the various filter frfs with a superscript corresponding to each signal, i.e., $\Psi^{(j)}$, Ψ^ρ, Ψ^ι. Away from co-integrating frequencies, they are explicitly given by

$$\begin{aligned}
\Psi^{(j)}(e^{-i\lambda}) &= |\delta^{(-j)}(e^{-i\lambda})|^2 \, \Sigma^{(j)} \, f_{\underline{x}}(\lambda)^{-1} \\
\Psi^\rho(e^{-i\lambda}) &= |\delta(e^{-i\lambda})|^2 \, f_\rho(\lambda) \, \Sigma^\rho \, f_{\underline{x}}(\lambda)^{-1} \\
\Psi^\iota(e^{-i\lambda}) &= |\delta(e^{-i\lambda})|^2 \, \Sigma^\iota \, f_{\underline{x}}(\lambda)^{-1}.
\end{aligned}$$

Each $\Psi^{(j)}(B)$ is a type of signal-pass filter for extracting the jth non-stationary signal, whereas Ψ^ρ is the band-pass filter, and Ψ^ι is the high-pass filter. To express the jth signal extraction frf at signal frequency $\lambda^{(j)}$, let

$$G^{(j)}(\lambda) = \sum_{\ell \neq j}^{J} |\delta^{(-j,-\ell)}(e^{-i\lambda})|^2 \, \Sigma^{(\ell)} + |\delta^{(-j)}(e^{-i\lambda})|^2 \, f_\rho(\lambda) \, \Sigma^\rho + |\delta^{(-j)}(e^{-i\lambda})|^2 \, \Sigma^\iota$$

for $1 \leq j \leq J$, where $\delta^{(-j,-k)}(B) = \prod_{\ell \neq j,k} \delta^{(\ell)}(B)$; if $J = 1$, this polynomial is interpreted as zero, and equals one in the case $J = 2$. Then it follows that

$$f_{\underline{x}}(\lambda) = |\delta^{(-j)}(e^{-i\lambda})|^2 \, \Sigma^{(j)} + |\delta^{(j)}(e^{-i\lambda})|^2 \, G^{(j)}(\lambda).$$

Because $\delta^{(-j)}(e^{-i\lambda^{(j)}}) \neq 0$ and Σ^ι has full rank, $G^{(j)}(\lambda^{(j)})$ has full rank, and is invertible. Using the GCD for $\Sigma^{(j)}$ for λ in a neighborhood of $\lambda^{(j)}$ the following matrix is well-defined:

$$H^{(j)}(\lambda) = I_N - L^{(j)} \left(\frac{|\delta^{(j)}(e^{-i\lambda})|^2}{|\delta^{(-j)}(e^{-i\lambda})|^2} \, I_N + L^{(j)\prime} \, G^{(j)}(\lambda)^{-1} \, L^{(j)} \right)^{-1} L^{(j)\prime} \, G^{(j)}(\lambda)^{-1}.$$

Note that $L^{(j)\prime} G^{(j)}(\lambda)^{-1} L^{(j)}$ is well-defined for λ in a neighborhood of $\lambda^{(j)}$, and this matrix moreover is invertible (it has rank r_j, which is full). Also,

$$\lim_{\lambda \to \lambda^{(j)}} H^{(j)}(\lambda) = \left\{ I_N - L^{(j)} \left(L^{(j)\prime} G^{(j)}(\lambda^{(j)})^{-1} L^{(j)} \right)^{-1} L^{(j)\prime} G^{(j)}(\lambda^{(j)})^{-1} \right\},$$

which is taken as the definition of $H^{(j)}(\lambda^{(j)})$. Then

$$\Psi^{(j)}(e^{-i\lambda}) = \begin{cases} |\delta^{(-j,-k)}(e^{-i\lambda})|^2 \, \Sigma^{(j)} \, G^{(j)}(\lambda)^{-1} \, H^{(j)}(\lambda) & \text{if} \quad \lambda = \lambda^{(k)}, k \neq j \\ I_N - H^{(j)}(\lambda) & \text{if} \quad \lambda = \lambda^{(j)}, \end{cases}$$

which is proved in the first case by the Sherman-Woodbury identity

$$f_{\underline{x}}(\lambda)^{-1} = |\delta^{(j)}(e^{-i\lambda})|^{-2} \, G^{(j)}(\lambda)^{-1} \, H^{(j)}(\lambda).$$

The second case follows from

$$\Psi^{(j)}(e^{-i\lambda}) = I_N - |\delta^{(j)}(e^{-i\lambda})|^2 \, G^{(j)}(\lambda) \, f_{\underline{x}}(\lambda)^{-1}.$$

As for the band-pass and high-pass filters, it follows that at $\lambda = \lambda^{(j)}$ they can be expressed as

$$\begin{aligned} \Psi^{\rho}(e^{-i\lambda}) &= |\delta^{(-j)}(e^{-i\lambda})|^2 \, f_{\rho}(\lambda) \, \Sigma^{\rho} \, G^{(j)}(\lambda)^{-1} \, H^{(j)}(\lambda) \\ \Psi^{\iota}(e^{-i\lambda}) &= |\delta^{(-j)}(e^{-i\lambda})|^2 \, \Sigma^{\iota} \, G^{(j)}(\lambda)^{-1} \, H^{(j)}(\lambda). \end{aligned}$$

The error spectral densities for each signal can be computed in terms of the same quantities:

$$f_{\eta}^{(j)}(\lambda) = \begin{cases} |\delta^{(-j)}(e^{-i\lambda})|^{-2} \left(\Psi^{(j)}(e^{-i\lambda}) \, G^{(j)}(\lambda) + [I_N - \Psi^{(j)}(e^{-i\lambda})] \, \Sigma^{(j)} \right) & \text{if} \quad \lambda = \lambda^{(j)} \\ |\delta^{(j)}(e^{-i\lambda})|^{-2} \left(G^{(j)}(\lambda) \, \Psi^{(j)}(e^{-i\lambda})' + [I_N - \Psi^{(j)}(e^{-i\lambda})] \, \Sigma^{(j)} \right) & \text{if} \quad \lambda = \lambda^{(k)}, k \neq j \\ \Sigma^{(j)} \, f_{\underline{x}}(\lambda)^{-1} \, G^{(j)}(\lambda) & \text{else.} \end{cases}$$

For the band-pass and high-pass filters similar expressions can be computed.