RESEARCH REPORT SERIES
(Statistics #2008-5)

Signal Extraction Revision Variances as a Goodness-of-Fit Measure

Tucker McElroy
Marc Wildi[1]

[1]Institute of Data Analysis and Process Design

Statistical Research Division
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: July 2, 2008

Signal Extraction Revision Variances as a Goodness-of-Fit Measure

Tucker McElroy and Marc Wildi
U.S. Census Bureau and Institute of Data Analysis and Process Design

Abstract

Typically, model misspecification is addressed by statistics relying on model-residuals, i.e., on one-step ahead forecasting errors. In practice, however, users are often interested in problems involving (also) multi-step ahead forecasting performances, which are not explicitly addressed by traditional diagnostics. In this article, we consider the topic of misspecification from the perspective of signal extraction. More precisely, we emphasize the connection between models and real-time (concurrent) filter performances by analyzing revision errors instead of one-step ahead forecasting errors. In applications, real-time filters are important for computing trends, for performing seasonal adjustment or for inferring turning-points towards the current boundary of time series. Since revision errors of real-time filters generally rely on particular linear combinations of one- and multi-step ahead forecasts, we here address a generalization of traditional diagnostics. Formally, a hypothesis testing paradigm for the empirical revision measure is developed through theoretical calculations of the asymptotic distribution under the null hypothesis, and the method is assessed through real data studies as well as simulations. In particular, we analyze the effect of model misspecification with respect to unit roots, which are likely to determine multi-step ahead forecasting performances. We also show that this framework can be extended to general forecasting problems by defining suitable artificial signals.

Keywords. Model-diagnostics; Nonstationary time series; Real-time filtering; Seasonality; Signal extraction.

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.

1 Introduction

Generally speaking, time series models of economic data are misspecified, as models are essentially simplified portraits of the underlying stochastic dynamics. The task of model diagnostics is then to identify "relevant" mismatches so that faulty models can be refined accordingly. The predicate "relevant" means that diagnostic tools should account for the purpose of a particular application

by emphasizing model failures that are likely to affect results. Traditional diagnostics in time series analysis focus on one-step ahead forecasting errors. Typical examples are (partial) autocorrelation functions of model residuals, as well as Ljung-Box (Ljung and Box, 1978) and Box-Pierce statistics. If the purpose of a particular application is short term one-step ahead forecasting, then these tools are appropriate. But for many applications, the performance of a model over multiple forecast leads is more important than the modeling of short-term behavior. We now briefly discuss such an application field.

Signal extraction concerns the definition and the estimation of interesting components of a time series. In practice, signal estimates at the current boundary are important, because of the need for timely information (Findley, Bell, Monsell, Otto, and Chen, 1998). Unfortunately, symmetric filters cannot be used directly because future data hasn't been observed yet. Traditional methods overcome this difficulty by expanding series on both ends by backcasts and forecasts generated by a time series model – typically an ARIMA model – so that the symmetric filter can be used. If the coefficients of the symmetric filter decay slowly, then forecasts of longer horizons are emphasized. Therefore, a "good" forecasting model should perform well with respect to all forecasting horizons simultaneously. Unfortunately, as the following example illustrates, traditional diagnostics cannot account for "relevant" model failures in general.

Wildi (2008) compares real-time performances of various approaches in the context of leading indicators. The so-called KOF-Economic-Barometer (see *www.kof.ethz.ch*) is based on business survey data. The latter time series are bounded by construction in $[-100\%, 100\%]$. For a particular series seen in Figure 1 (solid line), TRAMO[1] selects the following airline-model

$$(1-B)(1-B^{12})X_t = (1-0.662B)(1-0.824B^{12})\epsilon_t \tag{1}$$

after adjustments for outliers and calendar effects. As can be seen from typical diagnostic plots in Figure 2 standard model assumptions are met; neither the autocorrelation nor the partial autocorrelation function nor the Ljung-Box statistics suggest significant departures from the null hypothesis[2]. However, a realization (simulation) of the process defined by (1) in Figure 1 (dotted line) shows obvious departures from the original path. A longer simulation of the same process in Figure 3 confirms that the artificial series is dominated by a strong trend component which is a "stylized fact" of $I(2)$-processes and therefore of the identified airline model. The level of the original time series is much more "stationary" because the series is *bounded*, as are many important economic time series (such as inflation rates, interest rates or unemployment rates, for example). In fact, a glance at the sample ACF plot of the series shows no indications of trend or seasonal nonstationarity (Figure 4).

[1]TSW-package (March 2006) which can be downloaded from the Bank of Spain (http://www.bde.es/servicio /software /econome).

[2]TRAMO as well as X-12-ARIMA provide additional diagnostic tools such as heteroscedasticity or model stability tests which did not lead to a rejection of the above model either.

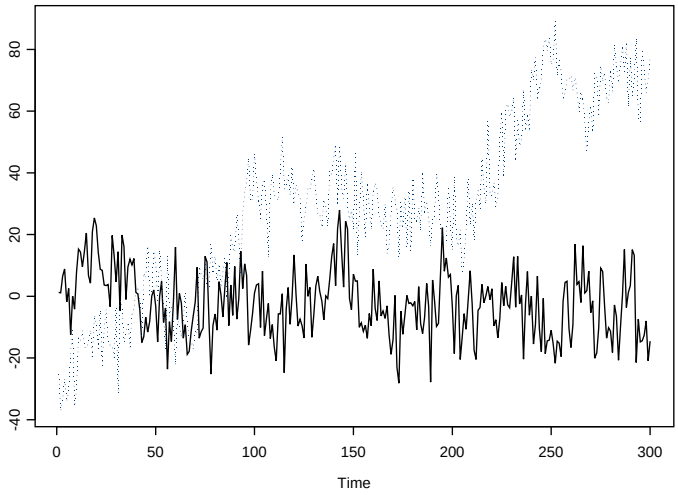

Figure 1: Series 31 (solid) and model simulation (dotted)

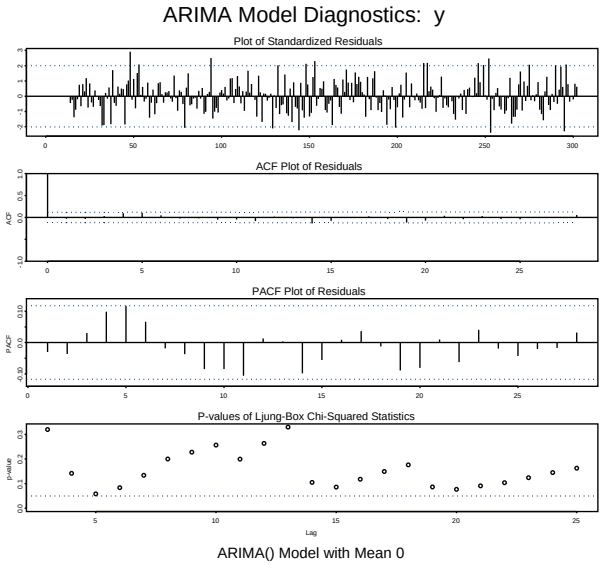

Figure 2: Diagnostics airline model for series 31

With respect to the limited one-step ahead forecasting perspective the above model performs well, thus confirming the usefulness of traditional diagnostics in this particular setting. Unfortunately, the "relevant" real-time estimates that are based on one- and multi-step ahead forecasts are inefficient – see Wildi (2004, 2008). Moreover, signal definitions based on misspecified models (for example canonical components) automatically inherit model failures and may be difficult to interpret in practice. Therefore, specific diagnostics are needed that match the signal extraction problem.

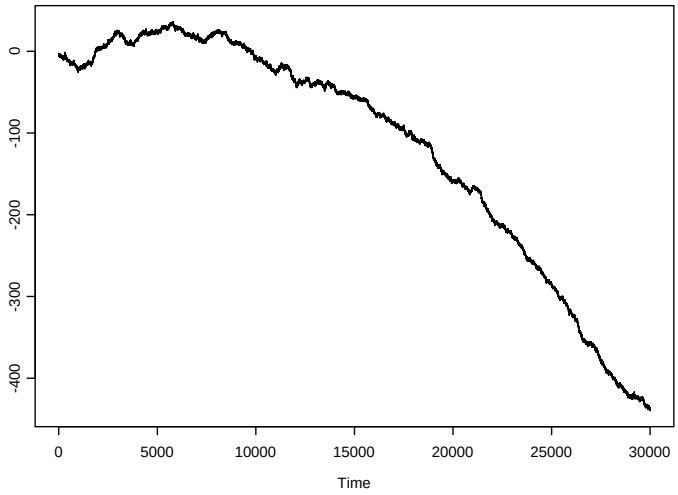

Figure 3: Long model simulation for series 31

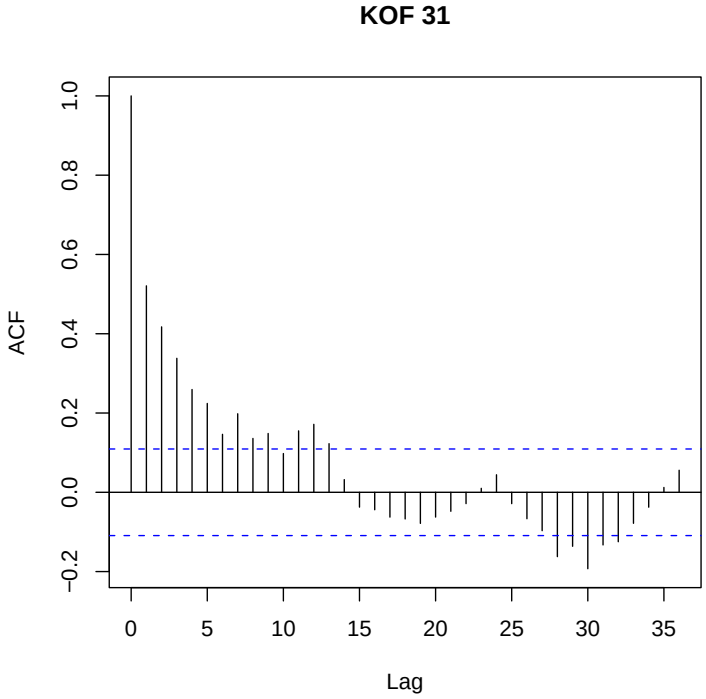

Figure 4: Sample ACF plot for series 31

A popular measure of the quality of signal estimates is the revision variance, because it conveys information as to what extent real-time (concurrent) estimates are subject to ex-post adjustments. Revision variances can be generated through model-based calculations, such as in SEATS (Mar-

avall and Caporello, 2004) and X13-ARIMA-SEATS. (Revision variance calculations are discussed in McElroy and Gagnon, 2006.) These revision variances should coincide asymptotically with empirical revision sample variances if the model is the "true" one. Therefore, a diagnostic test based on a comparison of revision variances accounts specifically for the signal extraction problem. Moreover, such a test involves one- and multi-step ahead model forecasts simultaneously. The main contribution of the paper is the proposal of a new test statistic that matches the signal extraction problem and a derivation of its distribution under the hypothesis that the model fits the DGP.

In Section 2 we discuss some of the background theory needed for a finite sample approach to signal extraction in a model-based context. We define the goodness-of-fit test statistic RV, and discuss its important finite sample and asymptotic properties under the null hypothesis that the given model is correct. Section 3 gives some of the details on implementing our testing procedure, with a discussion of the decomposition, structural, and direct approaches to defining a "signal." In Section 4 we apply these concepts to several real series where there is suspicion of model misspecification; the series are sectoral leading indices used in the KOF economic barometer. Section 5 concludes and mathematical proofs are in the appendix.

2 Theory

We begin with a background discussion on model-based signal extraction in a finite-sample context; then we discuss signal extraction revisions for such estimates, and their autocovariance structure is provided in Proposition 1. We then define our goodness-of-fit test statistic RV and determine its statistical properties.

2.1 Background on Signal Extraction

The following material can be found in an expanded form in McElroy (2008a). We consider the additive decomposition of our data vector $Y = (Y_1, Y_2, \cdots, Y_n)'$ into signal S and noise N, via $Y = S + N$. The signal might be the trend component, while the noise includes the seasonal and irregular components. Following Bell (1984), we let Y_t be an integrated process such that $W_t = \delta(B)Y_t$ is stationary, where B is the backshift operator and $\delta(z)$ is a polynomial with all roots located on the unit circle of the complex plane. (Also, $\delta(0) = 1$ by convention.) This $\delta(z)$ is referred to as the differencing operator of the series, and we assume it can be factored into relatively prime polynomials $\delta^S(z)$ and $\delta^N(z)$ (i.e., polynomials with no common zeroes), such that the series

$$U_t = \delta^S(B)S_t \qquad V_t = \delta^N(B)N_t \tag{2}$$

are mean zero stationary time series that are uncorrelated with one another. Note that $\delta^S = 1$ and/or $\delta^N = 1$ are included as special cases (in these cases either the signal or the noise or both

are stationary). We let d be the order of δ, and d_S and d_N are the orders of δ^S and δ^N; since the latter operators are relatively prime, $\delta = \delta^S \cdot \delta^N$ and $d = d_S + d_N$.

As in Bell and Hillmer (1988), we assume Assumption A of Bell (1984) holds for the component decomposition, and we treat the case of a finite sample with $t = 1, 2, \cdots, n$ with $n > d$. Assumption A states that the initial d values of Y_t, i.e., the variables $Y_* = (Y_1, Y_2, \cdots, Y_d)$, are independent of $\{U_t\}$ and $\{V_t\}$. For a discussion of the implications of this assumption, see Bell (1984) and Bell and Hillmer (1988).

Now we can write (2) in a matrix form, as follows. Let Δ be a $(n-d) \times n$ matrix with entries given by $\Delta_{ij} = \delta_{i-j+d}$ (the convention being that $\delta_k = 0$ if $k < 0$ or $k > d$).

$$\Delta = \begin{bmatrix} \delta_d & \cdots & \delta_1 & 1 & 0 & 0 & \cdots \\ 0 & \delta_d & \cdots & \delta_1 & 1 & 0 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \delta_d & \cdots & \delta_1 & 1 \end{bmatrix}$$

The matrices Δ_S and Δ_N have entries given by the coefficients of $\delta^S(z)$ and $\delta^N(z)$, but are $(n - d_S) \times n$ and $(n - d_N) \times n$ dimensional respectively. This means that each row of these matrices consists of the coefficients of the corresponding differencing polynomial, horizontally shifted in an appropriate fashion. Hence

$$W = \Delta Y \qquad U = \Delta_S S \qquad V = \Delta_N N \tag{3}$$

where W, U, V, S, and N are column vectors of appropriate dimension. Then it is possible to write the mean square linear optimal estimate $\hat{S}$ as a linear matrix operating on Y, i.e., $\hat{S} = FY$. The error covariance matrix, i.e., the covariance matrix of $\hat{S} - S$, is denoted by M; both F and M are given in McElroy (2008a). The formulas for F and M are given by:

$$F = \left(\Delta_S' \Sigma_U^{-1} \Delta_S + \Delta_N' \Sigma_V^{-1} \Delta_N\right)^{-1} \Delta_N' \Sigma_V^{-1} \Delta_N \tag{4}$$

$$M = \left(\Delta_S' \Sigma_U^{-1} \Delta_S + \Delta_N' \Sigma_V^{-1} \Delta_N\right)^{-1} \tag{5}$$

where Σ_X denote the covariance matrix for any random vector X.

Now these basic notions are generalized slightly for the development needed below. We will be considering samples of varying dimension; denote the signal extraction matrix of dimension m by $F^{(m)}$, and the MSE matrix by $M^{(m)}$. Also e_m denotes the mth unit vector in $\mathbb{R}^l$, where the dimension $l \geq m$ will be apparent from the context. We introduce a general notation for signal extraction estimates: $\hat{S}_{t|_s^m}$. This is an estimate of S_t, which is a linear function of the data $Y_s, Y_{s+1}, \cdots, Y_m$ such that the associated error $\hat{S}_{t|_s^m} - S_t$ is uncorrelated with the data $Y_s, Y_{s+1}, \cdots, Y_m$ under Assumption A. Such a signal extraction estimate has minimum Mean Squared Error (MSE) among all estimates that are linear in the data. Note that Assumption A has to do with the initial values

$Y_1, \cdots, Y_d$, which may not even be a part of the sample $Y_s, \cdots, Y_m$ (e.g., say $s > d$). The actual initial values in this sample are $Y_s, \cdots, Y_{s+d-1}$, but these can be expressed as a linear combination of the initial values Y_*. Therefore Assumption A does indeed guarantee the validity of all the signal extraction formulas for samples computed at subsequent time periods.

We make a final distinction. Any model-based signal extraction matrix will have the form F given by (4), though we allow that the model may be mis-specified. That is, any of δ^S, δ^N, Σ_U, or Σ_V may be in error. If we wish to denote the "true" specifications of these quantities, we place a tilde over it, e.g., $\tilde{\Sigma}_U$ is the true autocovariance matrix of U_t, whereas Σ_U denotes the matrix implied by our model. Misspecifying δ^S and δ^N is a worse error than the misspecification of Σ_U and Σ_V.

2.2 Revisions

The main concept in revision calculations is to consider a "window-sample" of size n; this is a sample $Y_{t+1}, Y_{t+2}, \cdots, Y_{t+n}$ for some $t = 0, 1, \cdots, N-1$, where N denotes the number of windows that we consider (not to be confused with the noise vector N). We focus on the concurrent signal extraction estimate, where we are interested in the signal at time $t+n$; simple extensions of our method can deal with the signal considered at other time points within the sample. Hence we consider signal extraction estimates $\hat{S}_{t+n|t+1}^{t+n}$, and are interested in the revision error that occurs if our sample was increased by a further $h > 0$ data points; the revised estimate would then be $\hat{S}_{t+n|t+1}^{t+n+h}$. Using the convention that the revision is "new minus old," the revision equals

$$\epsilon_t = \hat{S}_{t+n|t+1}^{t+n+h} - \hat{S}_{t+n|t+1}^{t+n}.$$

Of course the revision ϵ_t depends on n and h as well as t, but these will be held fixed throughout our analysis, so they don't enter the notation for the revision. If the nonstationary operators δ^S and δ^N have been correctly specified, then ϵ_t will be a stationary sequence; this is because $\hat{S}_{t+n|t+1}^{t+n+h}$ and $\hat{S}_{t+n|t+1}^{t+n}$ will have no noise nonstationarity, and will both contain signal nonstationarity in such a manner that their difference is in fact stationary. The following proposition describes some of the important statistical properties of these revisions. Let $\tilde{e}_n$ denote the nth unit vector in $\mathbb{R}^{n+h}$, whereas e_n denotes the nth unit vector in $\mathbb{R}^n$.

Proposition 1 *Assume that the signal extraction conditions of Section 2 hold, and in particular that δ^S and δ^N are correctly specified (though Σ_U and Σ_V need not be). Then the sequence of*

revisions ϵ_t is weakly stationary with mean zero and autocovariance sequence

$$\begin{aligned}\gamma_\epsilon(k) = &\left(\tilde{e}_n' M^{(n+h)} \Delta_S' \Sigma_U^{-1} \left[1_{n+h-d_S}\, 0_k\right] - e_n' M^{(n)} \Delta_S' \Sigma_U^{-1} \left[1_{n-d_S}\, 0_{k+h}\right]\right) \tilde{\Sigma}_U \\ &\left([0_k\, 1_{n+h-d_S}]' \Sigma_U^{-1} \Delta_S M^{(n+h)} \tilde{e}_n - [0_k\, 1_{n-d_S}\, 0_h]' \Sigma_U^{-1} \Delta_S M^{(n)} e_n\right) \\ + &\left(\tilde{e}_n' M^{(n+h)} \Delta_N' \Sigma_V^{-1} \left[1_{n+h-d_N}\, 0_k\right] - e_n' M^{(n)} \Delta_N' \Sigma_V^{-1} \left[1_{n-d_N}\, 0_{k+h}\right]\right) \tilde{\Sigma}_V \\ &\left([0_k\, 1_{n+h-d_N}]' \Sigma_V^{-1} \Delta_N M^{(n+h)} \tilde{e}_n - [0_k\, 1_{n-d_N}\, 0_h]' \Sigma_V^{-1} \Delta_N M^{(n)} e_n\right).\end{aligned}$$

The dimension of the M matrices is indicated by the superscript, and the 1 *refers to an identity matrix of indicated dimension. The subscript on the* 0 *then indicates the number of zero columns. The other matrices, such as Σ_U, Δ_S, etc., have dimensions implied by the other matrices that multiply them.*

Proposition 1 will be useful for determining the statistical properties of our goodness-of-fit statistic. Our null hypothesis (stated below) states that the model used actually describes the true process, so that $\Sigma_U = \tilde{\Sigma}_U$ and $\Sigma_V = \tilde{\Sigma}_V$. Hence for implementation, one needs to compute $\gamma_\epsilon(k)$ under this type of assumption, for a sufficient number of lags k. Below, we discuss the test statistic RV in more detail.

2.3 Goodness-of-Fit Test Statistic

Now we want to use the empirical within-sample revision error as a measure of goodness-of-fit; since the theoretical mean of the revisions is zero, we can compute an estimate of their variance via $\frac{1}{N}\sum_{t=0}^{N-1} \epsilon_t^2$. More generally, let our Revision Variance statistic be defined as

$$RV(B) = \frac{1}{N}\, \epsilon' B \epsilon,$$

where B is a square matrix and $\epsilon = (\epsilon_0, \epsilon_1, \cdots, \epsilon_{N-1})'$. Clearly, taking B equal to the identity matrix yields the sample second moment of the revisions, but other choices of B will grant better size and power properties. This $RV(B)$ has mean

$$\mathbb{E}RV(B) = \frac{1}{N}\, tr(B\, \widetilde{\Sigma}_\epsilon),$$

where $\widetilde{\Sigma}_\epsilon$ is the (true) covariance matrix of ϵ. Hence taking $B = \Sigma_\epsilon^{-1}$ based on our model specification (using Proposition 1), the mean of the revision statistic will be equal to 1 under the null hypothesis. Moreover, if the data is Gaussian, the variance will be equal to $2/N$. Now $RV(\Sigma_\epsilon^{-1})$ is the goodness-of-fit statistic considered in this paper; we will just use RV for short. The normalized test statistic is defined to be

$$\sqrt{N}\, \frac{RV - 1}{\sqrt{2}}. \tag{6}$$

Note that if the data is Gaussian, $\epsilon' \Sigma_\epsilon^{-1} \epsilon$ has a χ_N^2 distribution. The following result, which is essentially Theorem 1 of McElroy (2008b), gives the statistical properties of RV. Suppose that we

specify δ^S and δ^N correctly, so that by Proposition 1 the revision process is stationary; let f_ϵ be the spectral density corresponding to the given autocovariance sequence. If $\Sigma_U = \widetilde{\Sigma}_U$ and $\Sigma_V = \tilde{\Sigma}_V$, then the model is correctly specified with correct parameter values as well. The corresponding spectral density is the true spectrum for the revision process, and is denoted by $\tilde{f}_\epsilon$. Likewise, let Σ_ϵ and $\widetilde{\Sigma}_\epsilon$ be the associated covariance matrices.

Theorem 1 *(Theorem 1 of McElroy (2008b)) The mean of RV is $tr(\Sigma_\epsilon^{-1}\tilde{\Sigma}_\epsilon)/N$, and if the third and fourth cumulants are zero the variance is $2tr([\Sigma_\epsilon^{-1}\tilde{\Sigma}_\epsilon]^2)/N^2$. If $\tilde{f}_\epsilon$ and $1/f_\epsilon$ are continuously differentiable, then*

$$\mathbb{E}RV \to \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\tilde{f}_\epsilon(\lambda)}{f_\epsilon(\lambda)}\, d\lambda$$
$$N\,VarRV \to \frac{2}{2\pi}\int_{-\pi}^{\pi} \frac{\tilde{f}_\epsilon^2(\lambda)}{f_\epsilon^2(\lambda)}\, d\lambda$$

as $N \to \infty$. Also if the revision process satisfies either condition (B) or (HT) referenced below, then as $N \to \infty$

$$\frac{RV - \mathbb{E}RV}{\sqrt{VarRV}} \stackrel{\mathcal{L}}{\Longrightarrow} \mathcal{N}(0,1).$$

Remark 1 Some mild conditions on the data are required for the asymptotic theory; we follow the material in Taniguchi and Kakizawa (2000, Section 3.1.1). Condition (B), due to Brillinger (1981), states that the process is strictly stationary and condition (B1) of Taniguchi and Kakizawa (2000, p. 55) holds. Condition (HT), due to Hosoya and Taniguchi (1982), states that the process has a linear representation, and conditions (H1) through (H6) of Taniguchi and Kakizawa (2000, pp. 55 – 56) hold.

Remark 2 The computations required for the variance of the empirical revision measure RV are considerable, since we must consider up to $N + h$ different MSE matrices of various dimensions. There is no straight-forward way to obtain the required quantities using a State Space smoother – one must use the direct matrix approach of McElroy (2008a).

Our null hypothesis is that the model is correctly specified with correct covariance structure for the components as well, i.e.,

$$H_0 : \delta^N = \tilde{\delta}^N,\ \delta^S = \tilde{\delta}^S,\ \Sigma_U = \tilde{\Sigma}_U,\ \Sigma_V = \tilde{\Sigma}_V.$$

The alternative hypothesis is that the model is incorrectly specified, which includes not only the case that the proposed differencing operators may be incorrect, but also that the models for U_t and/or V_t may be incorrect. Not only may the parameter values be faulty, but the model specifications for these components may be wrong as well. In general we may speak of over- and under-specification of differencing operators. This refers to assigning too many or too few unit root differencing factors

in δ (which are then allocated among the signal and the noise). For example, if the true process is $I(1)$ and we use an $I(2)$ model, this corresponds to over-specification, whereas using an $I(0)$ model is under-specification. Generally speaking, our test is much more powerful for detection of under-specification, because in this case the revision process is nonstationary and the RV statistic explodes asymptotically. But with over-specification, the revision process will still be stationary; only now the variance normalization will be incorrect, leading us to reject H_0. There are many other interesting cases that arise, for example: $\tilde{\delta}(z) = 1 - z^{12}$ but our model specifies $\delta(z) = 1 - z$; this is under-specification, because the operator $1 + z + \cdots + z^{11}$ has been omitted.

3 Implementation

The previous section discussed the theoretical properties of the revision diagnostic RV, given that we compute signal extraction estimates using (4). We now discuss the details of implementing these ideas. In order to construct the signal extraction matrix F, we must specify the matrices Σ_U and Σ_V (as well as δ^S and δ^N) – or equivalently, their spectral densities f_U and f_V. Our null hypothesis states that our model is correct, and thus we have a true knowledge of the spectral density f_W under H_0. Generally, f_U and f_V are in turn determined from f_W in a variety of ways. If we are dealing with ARIMA processes, it may be possible to mathematically solve for f_U and f_V using the canonical decomposition approach of Hillmer and Tiao (1982). The structural approach (Harvey, 1989) is to specify models for f_U and f_V ahead of time and determine their parameters implicitly through f_W during estimation, noting that

$$f_W(\lambda) = |\delta^N(e^{-i\lambda})|^2 f_U(\lambda) + |\delta^S(e^{-i\lambda})|^2 f_V(\lambda). \tag{7}$$

A third approach is to set up equations relating the signal and noise pseudo-spectra to that of the original process, i.e.,

$$f_S(\lambda) = g(\lambda) f_Y(\lambda) \qquad f_N(\lambda) = (1 - g(\lambda)) f_Y(\lambda),$$

where $g : [-\pi, \pi] \to [0, 1]$ ia a user-defined function, and $f_S(\lambda) = f_U(\lambda)|\delta^S(e^{-i\lambda})|^{-2}$, $f_N(\lambda) = f_V(\lambda)|\delta^N(e^{-i\lambda})|^{-2}$, $f_Y(\lambda) = f_W(\lambda)|\delta(e^{-i\lambda})|^{-2}$. We call this the direct approach (see Kaiser and Maravall, 2005).

The canonical decomposition approach is quite popular, and is the method of the widely-used seasonal adjustment program TRAMO-SEATS. The structural approach is also widely used, being implemented in the the program STAMP. Both of these methods use the fitted model f_W to determine the component models. The direct approach depends on f_W as well, but by choosing g one can control what sorts of signals one is interested in. The basic conditions on g are that $g|\delta^N(e^{-i\cdot})|^{-2}$ and $(1-g)|\delta^S(e^{-i\lambda})|^{-2}$ are bounded functions. (These conditions ensure that f_S and f_N only have poles at the appropriate signal and noise frequencies.) Since there is less literature on

the direct approach, we illustrate it through a particular choice of g. Suppose that $\delta^S(z) = 1-z$ and $\delta^N(z) = 1+z+\cdots+z^{11}$, which correspond to trend signal and seasonal noise processes respectively. Let $g(\lambda) = |\delta^N(e^{-i\lambda})|^2/144$; it then follows that

$$\frac{1-g(\lambda)}{|\delta^S(e^{-i\lambda})|^2} = |h(e^{-i\lambda})|^2/144,$$

where $h(z) = 10.787+8.570z+6.672z^2+5.070z^3+3.738z^4+2.652z^5+1.788z^6+1.123z^7+.634z^8+.297z^9+.093z^{10}$. Then we obtain

$$f_U(\lambda) = f_W(\lambda)/144 \qquad f_V(\lambda) = |h(e^{-i\lambda})|^2 f_W(\lambda)/144.$$

These equations represent a very direct and clear relationship between f_W and f_U, f_V. In contrast, (7) does not (in general) provide a unique f_U, f_V for each specification of f_W. From these definitions of f_U and f_V, it is straightforward to obtain Σ_U and Σ_V, their associated covariance matrices. The general procedure for computing the revision goodness-of-fit measure is the following:

1. Begin with a proposed model f_Y, which consists of: signal and noise differencing operators δ^S and δ^N, and spectrum of the differenced process f_W.

2. Obtain f_U and f_V from f_W. In the structural approach, f_U and f_V are actually determined at the model estimation stage, whereas for the decomposition and direct approaches there are algorithms for computing f_U and f_V from f_W.

3. Construct the filter matrix F and the revision process ϵ by applying the appropriate rows of F to the data.

4. Obtain the covariance matrix of ϵ under the null hypothesis (by using Proposition 1). Compute the normalized RV via (6) and get the p-value using the χ^2_N distribution.

In the context of model-based seasonal adjustment or trend estimation of economic data, typically steps 1 and 2 (and part of 3) are already performed by the analyst. The implementation challenge lies in the correct construction of Σ_ϵ based on Proposition 1, which takes some care. Also, as noted in the previous section, the computation of RV requires a choice of revision lead h and window size n. We have written our implementation (of the decomposition and direct approaches) in Ox, utilizing SsfPack routines (Koopman, Shephard, and Doornik, 1999).

4 Empirical Studies

In this section, we focus on the finite-sample statistical properties of the empirical revision measure RV, considering both the decomposition and direct approaches (for the direct approach, we take g as defined in Section 3). In 4.1 we summarize various size and power studies, and in 4.2 we examine the method on several series from the KOF-Economic-Barometer.

4.1 Simulations

We wish for our power studies to correspond with the empirical studies of section 4.2. The series we consider are of length 322, so we take three window sizes $n = 120, 150, 180$ – hence the number of windows is $N = 202 - h, 172 - h, 142 - h$, where h is the revision lead. We consider several values of h, up to five years out (the data is monthly): $h = 12, 24, 36, 48, 60$. For our first study, we employ the decomposition approach applied to the popular Box-Jenkins airline model. Our second study employs the direct approach, but in this case the model is only $I(1)$ plus seasonal. Details on these two implementations are provided below.

In the decomposition study, there are three components: trend, seasonal, and irregular. The airline model is given by the ARIMA equation

$$(1 - B)(1 - B^{12})X_t = (1 - \theta B)(1 - \Theta B^{12})\epsilon_t.$$

Both the trend and seasonal are typically nonstationary in economic data, and thus are the components of greatest interest for our purposes. Here the trend differencing operator is $(1 - B)^2$, whereas the seasonal differencing operator is $U(B) = 1 + B + \cdots + B^{11}$. Hence we will consider either the trend or the seasonal as the signal of interest – note that the revision process for the associated noise is that of the signal multiplied by -1. So the RV for the seasonal component and the seasonally adjusted component will be identical. We consider a null hypothesis of a Box-Jenkins airline model with various specifications of the parameters θ, Θ. Given the specification of a null model via a choice θ, Θ, we can determine RV for either the trend or seasonal components as discussed in Section 3.

In the direct study, there are two components: the seasonal and the nonseasonal. The spectra of these components are defined through $g(\lambda) = |U(e^{-i\lambda})|^2/144$, as discussed in Section 3. In that section, S_t is nonseasonal and N_t is seasonal; note that if we swap roles and let the seasonal be the signal instead, the revision measure RV will yield identical results (again, since the revision process for noise is related to the revision process of signal via multiplication by -1). So, we only report results for the nonseasonal. The model for the data process is

$$(1 - B^{12})X_t = (1 - \Theta B^{12})\epsilon_t, \tag{8}$$

which can be viewed as a subset model of the airline model when $\theta = 1$ (after cancelation).

These clearly do not reflect a comprehensive study, but nevertheless will reveal some useful observations. First, airline models form a fairly basic trend-seasonal model, and thus are a good starting place. The window sizes were chosen to reflect common data lengths – typically monthly seasonal time series at many statistical agencies may be between 10 and 15 years long. Of course, the number of revisions N is much larger than it would be in practice, though in our case the length of the KOF series facilitates a large N. The asymptotics of Theorem 1 are with respect to

increasing N, so decreasing n and h should provide a RV that is more normally distributed. The revision leads h are fairly typical – in practice the revisions from model-based seasonal adjustments (using SEATS or X-12-ARIMA) are generally negligible after 5 years.

In order to investigate the power of the diagnostic tests in both studies, we consider the following alternative models:

$$(1-\phi B)(1-\Phi B^{12})X_t = (1-.6B)(1-.6B^{12})\epsilon_t$$

with $\phi, \Phi = .6, .9, 1$. Therefore, taking all possible combinations and making cancelations where appropriate, we obtain the following 9 models: $\phi = \Phi = 1$ (Model 0); $\phi = .9$, $\Phi = 1$ (Model 1); $\phi = .6$, $\Phi = 1$ (Model 2); $\phi = 1$, $\Phi = .9$ (Model 3); $\phi = 1$, $\Phi = .6$ (Model 4); $\phi = .9 = \Phi$ (Model 5); $\phi = .9$, $\Phi = .6$ (Model 6); $\phi = .6$, $\Phi = .9$ (Model 7); $\phi = .6 = \Phi$ (Model 8).

These alternative models have varying degrees of nonstationarity (and Model 8 corresponds to white noise after cancelation of factors). In computing the power, we simulated Gaussian data from the models but applied the signal extraction filters associated with the null model, which for the first study was a $.6, .6$ airline model (or Model 0); we consider both the trend and seasonal signals. For the second study (direct approach), the null model corresponds to the choice $\Theta = .6$ in the data process (8), which is actually Model 2. In the first study all 8 alternative models (i.e., Model 1 through 8) correspond to over-specification of the order of nonstationarity. However, in the second study Model 0 corresponds to under-specification and Models 3 through 8 correspond to over-specification (i.e., the null model over-differences these processes). (A note on the simulation of nonstationary stochastic processes: in the over-specification case, the application of the revision filter in the calculation of RV always annihilates all initial values, but there will be initial values leftover in the under-specification case. Thus, power in the under-specification case actually depends upon the choice of starting values in the stochastic process. For the simulation of Model 0 in study 2, we initialize with 13 zeroes and discard the first 500 observations, which amounts to a random initialization of the process.)

The results are reported in Table 1 below. In the first study, the power is quite low for Models 1, 3, and 5, which are fairly close to the null model. Models 2 and 4 are further from the null model, and the power is unexceptional, breaking past 50% for the larger sample sizes (i.e., smaller h and n). Models 6 and 7 are quite different from the null, and the power is decent, approaching 75% in larger samples. Finally Model 8 is very different from the null, and the power is quite high. Thus the power results are intuitive, increasing both with sample size and the discrepancy between null and alternative model. Generally the power for trend and seasonal are similar, though the former was usually greater than the latter. In the second study, Model 0 produces perfect power (in this case the normalized RV statistic was explosive, taking on values in the thousands), as to be expected. Model 2 just provides the size; power was surprisingly high for Models 3 and 4, which have no nonstationary seasonality. Models 1, 5 and 6 provide very good power as well, even though

the latter two correspond to the over-specification case. Power is low for Model 7, which perhaps is closest to the null model in some sense, and power is moderate for Model 8.

In order to provide a reference frame for these results, we computed Ljung-Box (LB) statistics on the same data generating processes, when fitting an airline model. As with the RV study, we kept the parameters in the fitted model fixed at .6, .6. We do this, rather than using the MLEs in each simulation, in order that comparisons with the RV statistic – which uses fixed parameters – will be meaningful. Table 2 summarizes the results; we consider the LB at lags 12, 24, and 36, since these are multiples of the seasonal lag. Although there are some problems with the size (Model 0 for Study 1, and Model 2 for Study 2), the power is generally quite good – especially in the under-specification cases. In terms of comparing the RV and LB methods, we note that for Study 1 our RV statistic is marginally more powerful for Models 4, 5, and 6 (note that we can maximize power by taking h and n smaller, but there is no *a priori* reason to consider one of the lags 12, 24, or 36 as preferable to the others in the LB statistics), but the LB statistics are superior in the other cases. In Study 2, only Model 8 provides more power than the LB. Therefore, according to the simulation studies the RV statistic is not competitive with LB.

In order to explain these results and, in particular, the seemingly low power of RV for Models 1, 3 and 5 in study 1 (see Table 1) it is useful to recall that our statistic is designed with real-time signal-extraction in mind; misspecification is directly related to performances of real-time filters. With that particular design-aspect in mind we propose to compare the performances of the misspecified filters (based on Model 0) for Models 1, 3 and 5 with the performance obtained for Model 0 (no misspecification). For that purpose, we computed empirical revision error variances ($N^{-1}\epsilon'\epsilon$) by simulations based on 1000 replications of Models 0, 1, 3 and 5. The results are reported in Table 3. As can be seen, the revision error variances of the misspecified filters (the last three columns) are close to the performance obtained for the true model (Model 0). Since misspecification does not have a dramatic impact with regard to the interesting real-time signal extraction performances, it is not surprising that our RV statistic does not lead to rejection of the false model in these cases. However, increased power gains are to be expected whenever real-time performances are affected by the misspecification, as shown in the analysis of the KOF series below.

In summary, we note that the RV procedure is quite flexible, as any combination of unit roots can be specified in the null hypothesis, and tested against an alternative where some or all of the roots no longer lie on the unit circle. The size is good and the power results are reasonable in finite sample (though not as good as LB, generally speaking). We observe that our statistic emphasizes signal extraction problems so that it cannot detect misspecifications that do not affect real-time filter performances. In terms of a recommendation for the choice of h and n, it is noted that smaller values effectively increase the sample size N used in the RV statistic, and thus increase the power; therefore, these should be taken as small as practicable.

4.2 Revisions of the KOF Data

We next applied these diagnostic tests to the KOF series mentioned in the Introduction. To focus the discussion, we concentrate on four series that were all identified by X-12-ARIMA as having seasonality and an $I(2)$ trend. Due to the bounded nature of these series, the $I(2)$ trend seems to be misspecified; so we expect our diagnostic tests to reject these models. We applied both revision diagnostic tests discussed above; the first is used to show that the $I(2)$ is over-specified, and the second shows that $I(1)$ is also over-specified. The series KOF9, KOF25, KOF27, and KOF29 were specified as airline models by X-12-ARIMA. Values of the standardized RV statistic are reported in Table 4.

All of the RV statistics were computed with the null model given by the maximum likelihood parameter estimates, for each given model specification, when fitted to a subset of the data (given by the window size). Recall from the introduction that LB statistics were generally not significant for all of the KOF series; in particular, at lag 12 the LB statistics are above the 5% level for all four series, though KOF9 and KOF29 have a few significant LB statistics at other lags. (Some rejections due to pure chance are to be expected due to multiple testing.) Yet even the smallest of these RVs is significant (p-value $< .025$) as a two-sided test, using the Gaussian distribution. All the values were negative, indicating over-specification, and in fact, there is a remarkable pattern in the test statistics evident from Table 4. Letting the effective sample size $N = 322 - h - n$, each value in Table 4 (independent of the type of study, or the series) is approximately equal to $-\sqrt{N/2}$, which can be easily verified. This is because the (un-standardized) RV statistic is approximately zero in every case, hence by (6) we obtain $-\sqrt{N/2}$.

Why do we obtain these results? By glancing at ACF plots (such as Figure 4), we know that these KOF series are close to stationarity, even though conventional model identification software classifies them as $I(2)$[3]. So the computed trends (and trend-irregulars) derived from such an over-specified model will tend to be very smooth and fairly flat. In this case, there will be little improvement to the trend estimate obtained by moving from a concurrent filter to a less asymmetric filter, and therefore the revisions will tend to be small with no particular directionality. In fact, the empirical revision process will have zero mean, little discernible autocorrelation pattern, and a small variance. Hence the RV will be quite small (values of .001 typically), resulting in the standardized RV being approximately $-\sqrt{N/2}$. It is primarily for this reason that the RV is so powerful at detecting model misspecification for the KOF series.

Another explanation can be based on results obtained in Wildi (2008) where the author shows that substantial performance gains (reduction of revision error variances) can be obtained in real-time by selecting filters that are not misspecified in the unit-root frequencies. Specifically, the author

[3]This is because models in differences are parsimonious (only two parameters for the airline model) and because one-step ahead forecasts are quite good. Of course, multi-step ahead forecasts are of poor quality because the series are bounded.

obtained reductions of 16%, 41%, 22% and 23% of the revision error variances for series 9, 25, 27 and 29 by relaxing the misspecified I(2)-filter-constraints (see Table 10.9 in the cited literature). This result confirms that the power of the RV-statistic increases as real-time performances are affected by the model misspecification.

5 Conclusion

It is well-known that models that pass traditional one-step ahead diagnostic tests may perform rather poorly in a multi-step ahead perspective – recall the discussion in section 1. It is therefore necessary to account for the purpose of a particular application when selecting and checking model performances. We have proposed a test for model misspecification which fits a general class of forecasting problems.

Although we restricted attention towards real-time signal-extraction problems, the scope of the proposed approach is more general because we allowed for arbitrary signals. Therefore, revision errors can be "designed" by choosing suitable (artificial) signal definitions. As an example, assume that a signal is defined by a symmetric MA(3)-filter with coefficients $\gamma_{-1}, \gamma_0, \gamma_1$ where $\gamma_{-1} = \gamma_1$. If $\gamma_1 = 1$, then the revision error would correspond to the one-step ahead forecasting error. Thus, traditional (one-step ahead) diagnostics can be replicated in our framework by choosing the above artificial filter. More generally, revision errors relying on arbitrary linear combinations of one- and multi-step ahead forecasts can be derived by specifying a corresponding symmetric MA-filter. (Note that the central weight γ_0 is not important here.) Therefore, a diagnostic test can be set-up which accounts for performances involving any linear combination of forecasts. As a consequence, the proposed diagnostic test can fit a variety of practically relevant estimation problems whose precise structures can be accounted for explicitly.

Our simulation results confirm a good concordance between asymptotic and finite sample test distributions; although in our Monte Carlo simulations the LB statistics generally out-performed RV, the latter is very successful in real-data studies, particularly if the misspecification directly affects filter performances. Results in the context of the KOF economic barometer suggest stronger rejection of false unit roots hypotheses, in both seasonal and trend roots. This is due to the fact that the above series exhibit mid-term trend reversion which cannot be detected by statistics relying on short-term forecasting performances exclusively.

Traditional model-fitting diagnostics are based on computing model residuals and testing them for whiteness, i.e., whether or not they are serially uncorrelated. The signal extraction revision process is similar in many ways to model residuals, although under the null hypothesis of correct model and covariance specification they do not behave as white noise, but rather have another covariance structure (as given in Proposition 1). Both model residuals and signal extraction revisions can be used to assess poorness of model fit, but each examines different aspects of the data's

dynamics. For the KOF series, model residuals appear to be white and hence no problems with the over-specified model are indicated, whereas the signal extraction revisions also appear to be white and generate a small value of RV, indicating a strong rejection of the fitted model. Hence, model residuals and signal extraction revisions present different information about a series. Essentially, signal extraction revisions allow the practitioner to focus on particular aspects or sections of the data's pseudo-spectrum, whereas model residuals look at the spectrum as a whole.

Given these findings, we present the RV statistic as a useful tool to be used in addition to standard goodness-of-fit statistics, such as LB and unit root tests. One drawback of the RV statistic is that it takes some time and effort to encode the formulas of Proposition 1, and some thought must also be given to how the models for signal and noise are related to the data process. We have found Ox to be a convenient language for implementation. A second caution is the finite-sample power of the RV statistic, as shown in Table 1, which will tend to be lower than desired in the over-specification case. However, we feel that these results are more than offset by the tremendous "empirical power" of the method on typical business-survey data where real-time performances are often substantially affected by model misspecification.

Appendix

Proof of Proposition 1. For the first assertion, we write out ϵ_t in vector form.

$$\begin{aligned}
\epsilon_t &= \tilde{e}_n' F^{(n+h)} \begin{bmatrix} Y_{t+1} \\ \vdots \\ Y_{t+n+h} \end{bmatrix} - e_n' F^{(n)} \begin{bmatrix} Y_{t+1} \\ \vdots \\ Y_{t+n} \end{bmatrix} \\
&= \tilde{e}_n' \left(F^{(n+h)} \begin{bmatrix} Y_{t+1} \\ \vdots \\ Y_{t+n+h} \end{bmatrix} - \begin{bmatrix} S_{t+1} \\ \vdots \\ S_{t+n+h} \end{bmatrix} \right) - e_n' \left(F^{(n)} \begin{bmatrix} Y_{t+1} \\ \vdots \\ Y_{t+n} \end{bmatrix} - \begin{bmatrix} S_{t+1} \\ \vdots \\ S_{t+n} \end{bmatrix} \right) \\
&= \tilde{e}_n' E_t^{(n+h)} - e_n' E_t^{(n)},
\end{aligned}$$

where $E_t^{(n)}$ denotes the error process at time t based on the sample from time $t+1$ to $t+n$. Such an error process is simply a linear combination of U_s and V_s – the differenced signal and noise processes – at times $t+1 \leq s \leq t+n$. The same goes for $E_t^{(n+h)}$, so ϵ_t is a linear combination of $\{U_s\}$ and $\{V_s\}$, which are weakly stationary and uncorrelated with one another. Thus, the revisions are weakly stationary, too (and if the $\{U_s\}$ and $\{V_s\}$ processes are strictly stationary, then so is the revision process). Since these error processes have mean zero, so does the revision process.

Finally, we consider the autocovariance at lag k; considering $k \geq 0$, we have

$$\begin{aligned}
\epsilon_t \epsilon_{t+k} = \tilde{e}_n' E_t^{(n+h)} {E_{t+k}^{(n+h)}}' \tilde{e}_n - \tilde{e}_n' E_t^{(n+h)} {E_{t+k}^{(n)}}' e_n \\
- e_n' E_t^{(n)} {E_{t+k}^{(n+h)}}' \tilde{e}_n + e_n' E_t^{(n)} {E_{t+k}^{(n)}}' e_n.
\end{aligned}$$

Next, we compute each of the error processes:

$$\begin{aligned}
E_t^{(n)} &= -M^{(n)}\Delta_S'\Sigma_U^{-1}U_{t+1+d_S:t+n} + M^{(n)}\Delta_N'\Sigma_V^{-1}V_{t+1+d_N:t+n} \\
E_{t+k}^{(n)} &= -M^{(n)}\Delta_S'\Sigma_U^{-1}U_{t+k+1+d_S:t+k+n} + M^{(n)}\Delta_N'\Sigma_V^{-1}V_{t+k+1+d_N:t+k+n} \\
E_t^{(n+h)} &= -M^{(n+h)}\Delta_S'\Sigma_U^{-1}U_{t+1+d_S:t+n+h} + M^{(n+h)}\Delta_N'\Sigma_V^{-1}V_{t+1+d_N:t+n+h} \\
E_{t+k}^{(n+h)} &= -M^{(n+h)}\Delta_S'\Sigma_U^{-1}U_{t+k+1+d_S:t+k+n+h} + M^{(n+h)}\Delta_N'\Sigma_V^{-1}V_{t+k+1+d_N:t+k+n+h}
\end{aligned}$$

We can conceive of a vector U of dimension $k+n+h-d_S$, which contains the U_j for $t+1+d_S \leq j \leq t+k+n+h$. Then we can substitute selection matrices into the above expressions, such as $[1_{n+h-d_S}\ 0]U$, and so forth. Similarly, we can do the same with the vector V. These expressions may be substituted into the formula for $\epsilon_t\epsilon_{t+k}$ above, and the expectation of UU' is Σ_U of appropriate dimension. The same holds for V, though note that $\mathbb{E}UV'$ is a zero matrix due to our assumptions on the components. Then by rearranging terms, we arrive at the stated formula. □

References

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

[2] Bell, W. and Hillmer, S. (1988) A matrix approach to likelihood evaluation and signal extraction for ARIMA component time series models. *SRD Research Report No. RR−88/22*, Bureau of the Census. http://www.census.gov/srd/papers/pdf/rr88-22.pdf

[3] Brillinger, D. (1981) *Time Series Data Analysis and Theory,* San Francisco: Holden-Day.

[4] Findley, D. F., Monsell, B. C., Bell, W. R., Otto, M. C. and Chen, B. C. (1998) New capabilities and methods of the X-12-ARIMA seasonal adjustment program. *Journal of Business and Economic Statistics* **16**, 127–177 (with discussion).

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

[6] Hillmer, S. and Tiao, G. (1982) An ARIMA-model-based approach to seasonal adjustment. *Journal of the American Statistical Association* **77**, 63–70.

[7] Hosoya, Y., and Taniguchi, M. (1982) A central limit theorem for stationary processes and the parameter estimation of linear processes. *Annals of Statistics* **10**, 132–153.

[8] Kaiser, R. and Maravall, A. (2005) Combining Filter Design with Model-based Filtering: An Application to Business-cycle Estimation. *International Journal of Forecasting* **21**, 691–710.

[9] Koopman, S., Shephard, N., and Doornik, J. (1999) Statistical algorithms for models in state space using SsfPack 2.2. *Econometrics Journal* **2**, 113 – 166.

[10] Ljung, G. and Box, G. (1978) On a measure of lack of fit in time series models. *Biometrika* **65**, 297–303.

[11] Maravall, A. and Caporello, G. (2004) Program TSW: Revised Reference Manual. *Working Paper 2004, Research Department, Bank of Spain.* http://www.bde.es

[12] McElroy, T. (2008a) Matrix Formulas for Nonstationary ARIMA Signal Extraction. *Econometric Theory* **24**, 1-22.

[13] McElroy, T. (2008b) Statistical Properties of Model-Based Signal Extraction Diagnostic Tests. *Communications in Statistics, Theory and Methods* **37**, 591–616.

[14] McElroy, T., and Gagnon, R. (2006) Finite Sample Revision Variances for ARIMA Model-Based Signal Extraction. *SRD Research Report No. RRS*2006 − 05, U.S. Census Bureau.

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

[16] Wildi, M. (2004) Signal Extraction: How (In)efficient Are Model-Based Approaches? An Empirical Study Based on TRAMO/SEATS and Census X-12-ARIMA. *KOF-Working Paper Nr. 96*, ETH-Zurich.

[17] Wildi, M. (2008) *Real-Time Signal-Extraction: Beyond Maximum Likelihood Principles.* http://www.idp.zhaw.ch/de/engineering/idp/forschung/finance-risk-management-and-econometrics/signal-extraction-and-forecasting/signal-extraction.html

Table 1. RV Size and Power

	Models								
St1 (T)	0	1	2	3	4	5	6	7	8
Ld 12	.05 .05 .05	.09 .08 .07	.59 .53 .45	.08 .08 .07	.60 .53 .46	.19 .17 .15	.78 .71 .63	.75 .68 .60	.98 .96 .92
Ld 24	.05 .05 .05	.08 .08 .07	.56 .50 .42	.08 .08 .07	.57 .50 .42	.18 .16 .14	.75 .68 .59	.73 .65 .56	.97 .95 .90
Ld 36	.05 .05 .05	.08 .08 .07	.54 .47 .39	.08 .08 .07	.54 .47 .39	.18 .15 .13	.72 .65 .55	.70 .62 .52	.96 .93 .87
Ld 48	.05 .04 .05	.08 .08 .07	.51 .44 .35	.08 .07 .07	.52 .43 .36	.16 .14 .12	.69 .61 .50	.67 .58 .48	.96 .91 .83
Ld 60	.05 .04 .05	.08 .07 .07	.48 .40 .32	.07 .07 .07	.49 .41 .32	.15 .13 .11	.66 .57 .46	.64 .55 .43	.94 .89 .79
St1 (S)	0	1	2	3	4	5	6	7	8
Ld 12	.05 .05 .05	.09 .08 .08	.59 .53 .46	.08 .08 .07	.59 .53 .45	.20 .17 .16	.78 .71 .62	.75 .68 .61	.98 .96 .92
Ld 24	.05 .05 .05	.08 .08 .08	.57 .50 .43	.08 .08 .07	.52 .45 .37	.18 .16 .13	.71 .63 .53	.72 .64 .56	.97 .93 .87
Ld 36	.05 .05 .05	.08 .08 .07	.54 .47 .40	.07 .07 .07	.47 .39 .31	.17 .14 .12	.66 .56 .45	.68 .60 .51	.95 .90 .81
Ld 48	.05 .05 .05	.08 .08 .07	.51 .44 .37	.07 .07 .06	.43 .34 .26	.15 .13 .11	.61 .51 .38	.65 .57 .46	.93 .86 .75
Ld 60	.05 .05 .05	.08 .07 .07	.48 .40 .33	.07 .07 .06	.39 .30 .21	.14 .12 .10	.57 .45 .32	.62 .52 .41	.91 .82 .67
St2	0	1	2	3	4	5	6	7	8
Ld 12	1.0 1.0 1.0	.87 .83 .77	.05 .05 .05	1.0 1.0 1.0	1.0 1.0 1.0	.91 .87 .83	.98 .97 .94	.09 .09 .08	.62 .57 .49
Ld 24	1.0 1.0 1.0	.86 .80 .75	.05 .05 .05	1.0 1.0 1.0	1.0 1.0 1.0	.90 .85 .80	.98 .96 .93	.09 .09 .08	.60 .54 .46
Ld 36	1.0 1.0 1.0	.84 .78 .72	.05 .05 .05	1.0 1.0 1.0	1.0 1.0 1.0	.88 .83 .78	.97 .95 .91	.09 .09 .08	.58 .51 .43
Ld 48	1.0 1.0 1.0	.82 .76 .70	.05 .05 .05	1.0 1.0 .99	1.0 1.0 .99	.87 .81 .74	.97 .94 .89	.09 .08 .08	.56 .48 .40
Ld 60	1.0 1.0 1.0	.80 .73 .66	.05 .05 .05	1.0 1.0 .99	1.0 1.0 .99	.85 .79 .71	.96 .92 .86	.09 .09 .08	.53 .45 .37

Table 1: Entries indicate empirical size and power as a percentage, computed via 10,000 Monte Carlo simulations, of the RV statistic. The Models 0 through 8 indicates the data generating process that was simulated, with the Lead (Ld) and type of Study (St1, St2) on the left. St1 (S) refers to revision statistics based on the seasonal signal, whereas St1 (T) refers to the trend. For these studies Model 0 corresponds to the null hypothesis, so this column gives size, whereas the other columns give power. For St2 the null hypothesis corresponds to Model 2, so this column gives size and the other columns give power. The three numbers in each cell are size/power for window sizes 120, 150, and 180 respectively, from left to right.

Table 2. LB Size and Power

	Models								
	0	1	2	3	4	5	6	7	8
St1	.05 .06 .07	.15 .14 .13	.97 .83 .72	.07 .08 .10	.57 .60 .56	.18 .18 .18	.72 .72 .68	.97 .83 .77	1.0 .99 .98
St2	1.0 1.0 1.0	1.0 1.0 1.0	.06 .06 .07	1.0 1.0 1.0	1.0 1.0 1.0	1.0 1.0 1.0	1.0 1.0 1.0	.08 .09 .10	.56 .60 .56

Table 2: Entries indicate empirical size and power as a percentage, computed via 10,000 Monte Carlo simulations, of the LB statistic (computed using fixed parameters). The Models 0 through 8 indicates the data generating process that was simulated, with the type of Study (St1, St2) on the left. Model 0 corresponds to the null hypothesis for St1, so this column gives size, whereas the other columns give power. For St2 the null hypothesis corresponds to Model 2, so this column gives size and the other columns give power. The three numbers in each cell are size/power for LB lags 12, 24, and 36 respectively, from left to right.

Table 3. Revision Error Variances: misspecification versus true model

	Models			
	0	1	3	5
St1	.192 (0.033)	0.207 (0.037)	0.209 (0.037)	0.227 (0.041)

Table 3: Empirical revision variances based on 1000 simulations from Models 0, 1, 3, and 5 under the null Model 0. Numbers in parentheses are empirical standard deviations of the revision variance estimates.

Table 4. Standardized RV statistics for the KOF series.

	Series			
St1 (T)	KOF9	KOF25	KOF27	KOF29
Ld 12	-9.73 -8.93 -8.05	-9.74 -8.94 -8.06	-9.74 -8.94 -8.06	-9.74 -8.94 -8.06
Ld 24	-9.42 -8.59 -7.67	-9.43 -8.60 -7.68	-9.43 -8.60 -7.68	-9.43 -8.60 -7.68
Ld 36	-9.10 -8.24 -7.27	-9.10 -8.24 -7.27	-9.10 -8.24 -7.28	-9.11 -8.24 -7.27
Ld 48	-8.76 -7.87 -6.85	-8.77 -7.87 -6.85	-8.77 -7.87 -6.85	-8.77 -7.87 -6.85
Ld 60	-8.42 -7.48 -6.40	-8.42 -7.48 -6.40	-8.42 -7.48 -6.40	-8.42 -7.48 -6.40
St1 (S)	KOF9	KOF25	KOF27	KOF29
Ld 12	-9.73 -8.92 -8.03	-9.74 -8.94 -8.06	-9.70 -8.90 -7.99	-9.74 -8.94 -8.06
Ld 24	-9.40 -8.57 -7.64	-9.43 -8.59 -7.67	-9.32 -8.50 -7.57	-9.43 -8.60 -7.68
Ld 36	-9.08 -8.22 -7.25	-9.10 -8.24 -7.27	-9.05 -8.19 -7.21	-9.11 -8.24 -7.28
Ld 48	-8.75 -7.87 -6.83	-8.77 -7.87 -6.85	-8.71 -7.82 -6.80	-8.77 -7.87 -6.85
Ld 60	-8.40 -7.46 -6.39	-8.42 -7.48 -6.40	-8.36 -7.43 -6.34	-8.42 -7.48 -6.40
St2	KOF9	KOF25	KOF27	KOF29
Ld 12	-9.74 -8.94 -8.06	-9.75 -8.94 -8.06	-9.75 -8.94 -8.06	-9.75 -8.94 -8.06
Ld 24	-9.43 -8.60 -7.68	-9.43 -8.60 -7.68	-9.43 -8.60 -7.68	-9.43 -8.60 -7.68
Ld 36	-9.11 -8.24 -7.28	-9.11 -8.25 -7.28	-9.11 -8.24 -7.28	-9.11 -8.25 -7.28
Ld 48	-8.77 -7.87 -6.85	-8.77 -7.87 -6.86	-8.77 -7.87 -6.85	-8.77 -7.87 -6.85
Ld 60	-8.42 -7.48 -6.40	-8.43 -7.48 -6.40	-8.43 -7.48 -6.40	-8.43 -7.48 -6.40

Table 4: Normalized RV test statistics for KOF series 9, 25, 27, and 29. An airline model was fitted to a subset of the data (corresponding to the window size) using Maximum Likelihood Estimation, and the corresponding parameter values were used to determine the null hypothesis. The type of Study (St1, St2) is on the left, along with revision lead. The three numbers in each cell are normalized RV at window sizes 120, 150, and 180 respectively, from left to right.