When are Direct Multi-Step and Iterative Forecasts Identical?

Tucker McElroy*
U.S. Census Bureau

Abstract

Although both direct multi-step ahead forecasting and iterated one-step ahead forecasting are two popular methods for predicting future values of a time series, it is not clear that the direct method is superior in practice, even though from a theoretical perspective it has lower Mean Squared Error (MSE). We first show that both methods are identical – if produced from the same fitted model – when the information set is semi-infinite. Then we show how discrepancies can arise when the sample is finite. Formulas for forecast error are derived, which are useful for determining the real MSE when the forecasting model is misspecified.

Keywords. ARIMA Models, Forecasting, Time Series Prediction.

1 Introduction

There is considerable interest among econometricians in forecasting time series, and both direct and iterated forecasting methods play a prominent role. Relevant literature includes Findley (1983, 1985), Weiss (1991), Tiao and Xu (1993), Lin and Granger (1994), Tiao and Tsay (1994), Clements and Hendry (1996), Bhansali (1996, 1997), Kang (2003), Chevillon and Hendry (2005), and Schorfheide (2005). A recent study by Marcellino, Stock, and Watson (2006) made comparisons between the direct and iterated methods, with the surprising conclusion that in practice the iterated method often performed better. Extensions of these results appear in Proietti (2011), which considers a particular class of ARIMA models as the basis of the forecast functions.

To frame our discussion, we must highlight that either method – direct or iterated – involves not only the use of forecast weights (or filters) peculiar to each method, but also model parameters fitted

*Center for Statistical Research and Methodology, U.S. Census Bureau, 4600 Silver Hill Road, Washington, D.C. 20233-9100, tucker.s.mcelroy@census.gov

accordingly. Thus, the model parameters used in the direct method and the iterated method could differ in practice. Hence, discrepancies in performance can arise from several sources: (1) different models being used; (2) different fitting methods being used; (3) different forecasting functions being used. By the third point, we refer to forecasting functions that differ even when the same models and same parameter estimates are plugged in. To focus our results initially, we focus on this scenario – namely that the models and fitting methods (Gaussian MLE in this case) are identical, so that only the forecast functions may differ.

Given this framework, we wish to answer the question: in general, when are iterated one-step ahead forecasts identical with multi-step ahead forecasts? To make the problem well-posed, we consider forecasting formulas arising from difference stationary time series models (including ARIMA models, for example) such that the resulting forecasts – under the assumption that the Data Generating Process (DGP) has been correctly identified – have minimal Mean Squared Error (MSE) given an information set not involving future values of the time series. For difference stationary time series we provide explicit forecasting formulas for either procedure, for both a semi-infinite information set and a finite information set. We also derive the forecast error processes and determine the MSEs for each case, allowing for model misspecification and parameter error (although uncertainty in parameter estimates is not quantified).

In the case of a semi-infinite past, the multi-step and iterative methods are identical – the semi-infinite concurrent filters in each case are algebraically the same. Essentially this is due to a property of nested conditional expectations. When a finite past is utilized, the forecast error is mean zero in both cases, and explicit expressions for it give insight into the MSEs for either method. We present the main mathematical results in Section 2. The numerical results of Section 3 suggest that the semi-infinite case is indicative of the general situation, in that the direct and iterated forecast functions differ very little in the finite past case when the sample size is above 10 or so. This indicates that the main discrepancies in the methods – assuming the same model is used, which is common enough – must arise due to the difference in fitting methods. We make some further comments on this aspect in Section 4.

2 Mathematics of Direct and Iterative Forecasting

Some of the following material can be found in a variety of time series references, but we assemble the mathematics here with a coherent notation. A related treatment of direct multi-step ahead forecasting can be found in McElroy and Findley (2010). We begin by focusing on the case of a semi-infinite past as the information set, and then treat a finite past in the following subsection.

2.1 Semi-infinite Past

Suppose that the time series $\{X_t\}$ is difference stationary with operator $\delta(B)$, such that $W_t = \delta(B)X_t$ is covariance stationary with mean zero, and hence has a causal Wold representation

$$W_t = \sum_{j\geq 0} \psi_j \epsilon_{t-j} = \Psi(B)\epsilon_t,$$

where the process $\{\epsilon_t\}$ is uncorrelated with variance σ^2. This type of causal difference linear process includes all ARIMA and SARIMA processes, and is fairly general. Suppose that at time t we are interested in generating h-step ahead forecasts based on present and past information, denoted by $X_{t:} = \{X_s : s \leq t\}$. The problem is to compute $\mathbb{E}[X_{t+h}|X_{t:}]$ for some $h > 0$ under a Gaussian assumption – or equivalently, to find the minimal MSE *linear* estimate of X_{t+h} given data up to time t. This optimal estimate – denoted by $\widehat{X}_{t+h|t}$ – can be expressed as a causal filter operating on the $\{X_t\}$ time series, called $\Upsilon_h(B) = \sum_{j\geq 0} \upsilon_j B^j$, namely $\widehat{X}_{t+h|t} = \Upsilon_h(B)X_t$. Because this filter works in an optimal fashion, it may be called the Direct multi-step ahead forecasting filter (cf. Proietti, 2011). In contrast, we might consider applying $\Upsilon_1(B)$ repeatedly, each time appending the previous forecasts to the end of the series, and thereby attaining an Iterated multi-step ahead forecasting filter. This will be denoted by $\Pi_h(B) = \sum_{j\geq 0} \pi_j^{(h)} B^j$, and is described below.

We begin the treatment with some results from Bell (1984) on nonstationary stochastic processes. Let $\delta(z) = 1 - \sum_{j=1}^{d} \delta_j z^j$, and its reciprocal power series is $\xi(z) = 1/\delta(z) = \sum_{j\geq 0} \xi_j z^j$. One can recursively solve for the $\{\xi_j\}$ via $\xi_0 = 1$ and $\xi_j = \sum_{k=1}^{\min(d,j)} \delta_k \xi_{j-k}$ for $j \geq 1$. Moreover, certain time-dependent coefficient functions $A_{j,t}$ lying in the null space of $\delta(B)$ are defined via

$$A_{j,t} = \xi_{t-j} - \sum_{k=1}^{d-j} \delta_k \xi_{t-j-k}$$

for $j = 1, 2, \cdots, d$ and $t \geq 1$. Then the process $\{X_t\}$ can be represented at time $t+h$ for any $h \geq 0$ via

$$X_{t+h} = \sum_{j=1}^{d} A_{j,d+h} X_{t+j-d} + \sum_{j=0}^{h-1} \xi_j W_{t+h-j}.$$

Then the direct forecast filter is given by

$$\Upsilon_h(B) = \sum_{j=1}^{d} A_{j,d+h} B^{d-j} + \sum_{k=1}^{h} \xi_{h-k} [\Psi]_k^{\infty}(B) F^k \delta(B) \Psi^{-1}(B). \tag{1}$$

Here the bracket notation is used to refer to that portion of the power series that is retained, namely $[\Psi]_a^b(B) = \sum_{j=a}^{b} \psi_j B^j$ for integers a and b. The derivation is sketched in McElroy and Findley (2010), but to prove its optimality it suffices to show that the error process is orthogonal to $X_{t:}$.

The forecast error is

$$\begin{aligned}\varepsilon_t = X_{t+h} - \Upsilon_h(B)X_t &= \sum_{j=0}^{h-1} \xi_j \left(F^{h-j} - [\Psi]_{h-j}^{\infty}(B) F^{h-j} \Psi^{-1}(B) \right) W_t \\ &= F^h \sum_{j=0}^{h-1} \xi_j B^j [\Psi]_0^{h-j-1}(B) \Psi^{-1}(B) W_t.\end{aligned}$$

It can be shown that $\sum_{j=0}^{h-1} \xi_j B^j [\Psi]_0^{h-j-1}(B) = [\Psi/\delta]_0^{h-1}(B)$ using simple algebra. Now when the filter exactly matches the DGP, we have $\Psi^{-1}(B)W_t = \epsilon_t$, so that the error process is $\varepsilon_t = [\Psi/\delta]_0^{h-1}(B)\epsilon_{t+h}$, which only depends on future innovations; hence the error process is orthogonal to $X_{t:}$. More generally, our forecasting model may be misspecified such that

$$Var(\varepsilon_t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\left| [\Psi/\delta]_0^{h-1}(z) \right|^2}{\left| \Psi(z) \right|^2} \widetilde{f}(\lambda) d\lambda, \tag{2}$$

where $z = e^{-i\lambda}$ and $\widetilde{f}$ is the true spectral density of $\{W_t\}$.

Now evaluating (1) for $h = 1$ produces the one-step ahead direct forecast filter. Applying its iterative definition yields

$$\Pi_{p+1}(B) = \upsilon_0 \Pi_p(B) + \upsilon_1 \Pi_{p-1}(B) + \cdots + \upsilon_{p-1} \Pi_1(B) + F^p [\Upsilon_1]_p^{\infty}(B).$$

This is initialized with $\Pi_1(B) = \Upsilon_1(B)$. Iterative forecasting produces an h-step ahead estimate $\widetilde{X}_{t+h|t} = \Pi_h(B)X_t$, and it is not initially obvious whether this performs as well as the Direct forecast filter. Now $\Pi_h(B)$ can be expressed compactly in terms of $\Upsilon_1(B)$ as follows. Define the degree k polynomials $p_k(B)$ recursively via $p_0(F) = 1$ and

$$p_{k+1}(F) = \sum_{j=0}^{k} \upsilon_j p_{k-j}(F) + F^{k+1}.$$

For example, $p_1(F) = \upsilon_0 + F$ and $p_2(F) = \upsilon_0^2 + \upsilon_1 + \upsilon_0 F + F^2$, etc. Then

$$\Pi_h(B) = F^h + p_{h-1}(F) \left[\Upsilon_1(B) - F \right]. \tag{3}$$

The proof of (3) is by induction. But this formula is also convenient, because it allows quick calculation of the forecast error process. The iterated forecast error is

$$\eta_t = X_{t+h} - \Pi_p(B)X_t = \left(F^h - \Pi_p(B) \right) X_t = p_{h-1}(F) F \Psi^{-1}(B) W_t.$$

In the case that the filter exactly matches the DGP, the error process $\eta_t = p_{h-1}(F)\epsilon_{t+1}$, which is orthogonal to $X_{t:}$. Hence the iterated forecasts are also optimal, and by uniqueness of the Gaussian conditional expectation (i.e., the MSE optimal linear estimate) we must have $\Upsilon_h(B) = \Pi_h(B)$. In fact, we have

$$\varepsilon_t = \eta_t + \left(\Upsilon_h(B) - \Pi_h(B) \right) X_t$$

with the two quantities on the RHS orthogonal (this is because η_t is orthogonal to all linear functions of $X_{t:}$). Thus the optimal MSE is equal to $Var(\eta_t)$ plus a non-negative quantity; by optimality, this quantity must be zero, and it follows that $\Upsilon_h(z) = \Pi_h(z)$ almost everywhere.

A second derivation of the result stems from probability theory alone. From Theorem 1.2 of Durrett (1996, p.226) within a conditional expectation we can always additionally condition on a larger information set, since "the smaller σ-field always wins." Thus for $h > 1$

$$\begin{aligned}\mathbb{E}[X_{t+h}|X_{t:}] &= \mathbb{E}[\mathbb{E}[X_{t+h}|X_{t+h-1:}]|X_{t:}] \\ &= \mathbb{E}[\Upsilon_1(B)X_{t+h-1}|X_{t:}] \\ &= \sum_{j=0}^{h-2} \upsilon_j \mathbb{E}[X_{t+h-1-j}|X_{t:}] + [\Upsilon_1]_{h-1}^{\infty}(B)X_t.\end{aligned}$$

From here we use induction on h to prove that $\Pi_h(B)X_t = \mathbb{E}[X_{t+h}|X_{t:}]$.

Comparing the alternative expressions for the error processes ε_t and η_t yields

$$[\Psi/\delta]_0^{h-1}(B)F^{h-1} = p_{h-1}(F),$$

which is not easy to show algebraically (and is not obvious). In applications, we may want to compute (2) for a given model and arbitrary DGP. In this case the LHS formula $[\Psi/\delta]_0^{h-1}(B)$ is more convenient to work with, since the power series representation of $\Psi/\delta(z)$ up to a finite number of terms is easily computed in R. The recursive polynomials p_{h-1} require knowledge of the first $h-2$ coefficients of $\Upsilon_1(B)$, which requires a separate calculation.

2.2 Finite Past

Here we maintain the same basic assumptions on the process, but suppose that we are interested in forecasts based on a finite information set $X_{1:n}$, where n denotes the present observation time, as well as the sample size. The conditional expectation $\mathbb{E}[X_{n+h}|X_{1:n}]$ is the target of the direct approach, and we begin by presenting matrix formulas for forecasting and the covariance of the forecast error process. Although the treatment is standard (and some of the results can be found in McElroy (2008) and other literature), we review all derivations for a cohesive treatment.

In the finite-sample treatment it is not necessary to utilize a causal Wold representation for the differenced data process; we only require that the covariance function γ_h of the $\{W_t\}$ process be well-defined. We require the following notation. Let Δ_m be the (square) differencing matrix of dimension m such that the upper left $d \times d$ block is an identity matrix and the lower $m-d$ rows are given by the coefficients of $\delta(z)$ appropriately shifted. The jkth entry of Δ_m for $j > d$ is given by the $j-k$th coefficient of $\delta(z)$ (by convention, a coefficient of $\delta(z)$ with index less than zero or greater than d is just equal to zero). This differencing matrix is unit lower triangular, as is its inverse. Likewise, the Toeplitz covariance matrix of dimension m for $\{W_t\}$ is denoted Σ_m, i.e., $[\Sigma_m]_{jk} = \gamma_{j-k}$.

As a preliminary, we have a representation of $X_{1:n}$ in terms of initial values given as follows. Interpreting $X_{1:n}$ as a column vector, we have $\Delta_n X_{1:n} = [X'_{1:d}, W'_{d+1:n}]'$. Here subindices refer to the collection of corresponding random variables, collected into a column vector. The first d values $X_{1:d}$ are referred to as initial values in the forecasting and signal extraction literature. The distribution of these initial values is typically not included in time series models, being an unknown quantity, and it is common to assume the initial values are uncorrelated with the increment process $\{W_t\}$. This assumption is ubiquitous in the time series forecasting literature, and is implicit in all State Space Smoothing algorithms for forecasting and signal extraction. When deriving forecasting results under this assumption, the filters do not depend on the distribution of the initial values, and moreover the forecast error process is independent of the initial values themselves, which is a nice feature.

Let $\widehat{X}^{(D)}_{n+h|1:n} = \mathbb{E}[X_{n+h}|X_{1:n}]$ be the optimal direct h-step ahead forecast. We first derive this formula and its properties, and then discuss the iterated forecast formulas. Letting e_{n+h} denote a vector of length $n+h$ that is zero except for a unit in the last entry, we claim that

$$\widehat{X}^{(D)}_{n+h|1:n} = e'_{n+h}\Delta^{-1}_{n+h}\begin{bmatrix} 1_d & 0_{d\times n-d} \\ 0_{n-d\times d} & \Sigma_{n+h-d}\begin{bmatrix} 1_n \\ 0_{h\times n}\end{bmatrix}\Sigma^{-1}_{n-d}\end{bmatrix}\Delta_n X_{1:n}. \tag{4}$$

Optimality is defined as having the property that the forecast error is uncorrelated with the sample, which for Gaussian processes implies MSE optimality. We proceed to compute the forecast error $\widehat{X}^{(D)}_{n+h|1:n} - X_{n+h}$. We first note that due to the special structure of Δ_{n+h}, we have $\Delta_n[1_n\ 0_{n\times h}] = [1_n\ 0_{n\times h}]\Delta_{n+h}$. Therefore the direct forecast error $\widehat{X}^{(D)}_{n+h|1:n} - X_{n+h}$ equals

$$\begin{aligned}
&e'_{n+h}\Delta^{-1}_{n+h}\begin{bmatrix} 1_d & 0_{d\times n-d} \\ 0_{n+h-d\times d} & \Sigma_{n+h-d}\begin{bmatrix} 1_n \\ 0_{h\times n}\end{bmatrix}\Sigma^{-1}_{n-d}\end{bmatrix}[\Delta_n\ 0_{n\times h}]X_{1:n+h} - e'_{n+h}X_{1:n+h} \\
&= e'_{n+h}\Delta^{-1}_{n+h}\left(\begin{bmatrix} 1_d & 0_{d\times n-d} \\ 0_{n+h-d\times d} & \Sigma_{n+h-d}\begin{bmatrix} 1_n \\ 0_{h\times n}\end{bmatrix}\Sigma^{-1}_{n-d}\end{bmatrix}[1_n\ 0_{n\times h}] - 1_{n+h}\right)\begin{bmatrix} X_{1:d} \\ W_{d+1:n+h}\end{bmatrix} \\
&= e'_{n+h}\Delta^{-1}_{n+h}\begin{bmatrix} 0_d & 0_{d\times n+h-d} \\ 0_{n+h-d\times d} & \Sigma_{n+h-d}\begin{bmatrix} 1_n \\ 0_{h\times n}\end{bmatrix}\Sigma^{-1}_{n-d}[1_n\ 0_{n\times h}] - 1_{n+h-d}\end{bmatrix}\begin{bmatrix} X_{1:d} \\ W_{d+1:n+h}\end{bmatrix} \\
&= e'_{n+h}\Delta^{-1}_{n+h}\begin{bmatrix} 0_{1:d} \\ \left(\Sigma_{n+h-d}\begin{bmatrix} 1_n \\ 0_{h\times n}\end{bmatrix}\Sigma^{-1}_{n-d}[1_n\ 0_{n\times h}] - 1_{n+h-d}\right)W_{d+1:n+h}\end{bmatrix},
\end{aligned}$$

which shows that the initial values are not present in the forecast error. To show optimality,

consider the covariance of this forecast error with the data $\Delta_n^{-1}[X'_{1:d}, W'_{d+1:n}]'$; we obtain

$$e'_{n+h}\Delta_{n+h}^{-1}\begin{bmatrix} 0_d & 0_{d\times n+h-d} \\ 0_{n+h-d\times d} & \left(\Sigma_{n+h-d}\begin{bmatrix} 1_n \\ 0_{h\times n}\end{bmatrix}\Sigma_{n-d}^{-1}[1_n\ 0_{n\times h}] - 1_{n+h-d}\right)\Sigma_{n+h-d}\begin{bmatrix} 1_n \\ 0_{h\times n}\end{bmatrix}\end{bmatrix}\Delta_n^{\dagger}$$

$$= e'_{n+h}\Delta_{n+h}^{-1}\begin{bmatrix} 0_d & 0_{d\times n+h-d} \\ 0_{n+h-d\times d} & \Sigma_{n+h-d}\begin{bmatrix} 1_n \\ 0_{h\times n}\end{bmatrix}\Sigma_{n-d}^{-1}\left([1_n\ 0_{n\times h}]\,\Sigma_{n+h-d}\begin{bmatrix} 1_n \\ 0_{h\times n}\end{bmatrix} - \Sigma_{n-d}\right)\end{bmatrix}\Delta_n^{\dagger},$$

which is identically zero. Similarly, the variance of the forecast error (i.e., the forecast MSE) is equal to

$$e'_{n+h}\Delta_{n+h}^{-1}\begin{bmatrix} 0_d & 0_{d\times n+h-d} \\ 0_{n+h-d\times d} & \Sigma_{n+h-d} - \Sigma_{n+h-d}\,[1_n\ 0_{n\times h}]\,\Sigma_{n-d}^{-1}\begin{bmatrix} 1_n \\ 0_{h\times n}\end{bmatrix}\Sigma_{n+h-d}\end{bmatrix}\Delta_{n+h}^{\dagger}e_{n+h}. \quad (5)$$

This assumes that the model is correctly specified and precisely known (no parameter error); otherwise the formula is slightly more complex, because the covariance matrix of $W_{d+1:n+h}$ is no longer equal to Σ_{n+h-d}.

Both (4) and (5) are very easy to program. We now proceed to discuss iterative forecasting. Let $\widehat{X}^{(I)}_{n+h|1:n}$ be the iterative forecast obtained by inductively applying the one-step ahead direct forecast filter, as described above. Let η' denote the row vector of coefficients yielding $\widehat{X}^{(D)}_{n+1|1:n}$ from the data, i.e.,

$$\eta' = e'_{n+1}\Delta_{n+1}^{-1}\begin{bmatrix} 1_d & 0_{d\times n-d} \\ 0_{n-d\times d} & \Sigma_{n+h-d}\begin{bmatrix} 1_n \\ 0_{h\times n}\end{bmatrix}\Sigma_{n-d}^{-1}\end{bmatrix}\Delta_n.$$

The iterative procedure amounts to appending the most recent forecast to past data and forecasts, which may be formalized in matrix notation as follows. Define

$$J = \begin{bmatrix} 0_{1:n-1}\ 1_{n-1} \\ \eta'\end{bmatrix},$$

so that $JX_{1:n}$ consists of data values $X_{2:n}$ with the one-step ahead forecast appended. Then $\widehat{X}^{(I)}_{n+h|1:n} = e'_n J^h X_{1:n}$. Note the similarity to results in Proietti (2011), although the matrix power here involves matrices of full dimension n rather than just the model order. It is possible to prove that the iterative forecasts result in a forecast error process that does not depend on initial values, and moreover explicit formulas for the forecast MSE can be derived. The forecast error can be

written

$$\begin{aligned}\widehat{X}^{(I)}_{n+h|1:n} - X_{n+h} = \eta' J^{h-2} \left(J \left[1_n \, 0_{1:n}\right] - \left[0_{1:n} \, 1_n\right]\right) X_{1:n+1} \\ + \cdots + \eta' \left(J \left[1_n \, 0_{1:n}\right] - \left[0_{1:n} \, 1_n\right]\right) X_{h-1:n+h-1} \\ + \left(\left[\eta' \, 0_{1:n}\right] - e'_{n+1}\right) X_{h:n+h}.\end{aligned}$$

This expansion is established by telescoping the sum. Now $J \left[1_n \, 0_{1:n}\right] - \left[0_{1:n} \, 1_n\right]$ has the first $n-1$ rows identically zero and the final row equal to $\left[\eta' \, 0\right] - e'_{n+1}$. But this is the forecast error "filter" for 1-step ahead direct forecasting. Let

$$K = \Sigma_{n+1-d} \begin{bmatrix} 1_n \\ 0_{1\times n} \end{bmatrix} \Sigma^{-1}_{n-d} \left[1_n \; 0_{n\times 1}\right] - 1_{n+1-d}$$

by definition, so that

$$\left(\left[\eta' \, 0_{1:n}\right] - e'_{n+1}\right) X_{1:n+1} = e'_{n+1} \Delta^{-1}_{n+1} \begin{bmatrix} 0_d & 0_{d\times n+1-d} \\ 0_{n+1-d\times d} & K - 1_{n+1-d} \end{bmatrix} \begin{bmatrix} X_{1:d} \\ W_{d+1:n+1} \end{bmatrix}.$$

Hence the iterative forecast error does not depend on initial values, and the forecast error can be written

$$\begin{aligned}\widehat{X}^{(I)}_{n+h|1:n} - X_{n+h} = \eta' J^{h-2} \begin{bmatrix} 0_{n-1\times n+h-d} \\ \mu' \left(K - 1_{n+1-d}\right) \left[1_{n+1-d} \, 0_{n+1-d\times h-1}\right] \end{bmatrix} W_{d+1:n+h} \\ + \cdots + \eta' \begin{bmatrix} 0_{n-1\times n+h-d} \\ \mu' \left(K - 1_{n+1-d}\right) \left[0_{h-2\times n+1-d} \; 1_{n+1-d} \, 0_{n+1-d\times 1}\right] \end{bmatrix} W_{d+1:n+h} \\ + \mu' \left(K - 1_{n+1-d}\right) \left[0_{h-1\times n+1-d} \; 1_{n+1-d}\right] W_{d+1:n+h},\end{aligned}$$

where μ' corresponds to the final row of Δ^{-1}_{n+1}, with the first d entries omitted, i.e., $\mu' = e'_{n+1} \Delta^{-1}_{n+1} [1_d \, 0]'$. Thus this forecast error is equal to some $a' W_{d+1:n+h}$, with

$$\begin{aligned}a' = \eta' J^{h-2} \begin{bmatrix} 0_{n-1\times n+h-d} \\ \mu' \left(K - 1_{n+1-d}\right) \left[1_{n+1-d} \, 0_{n+1-d\times h-1}\right] \end{bmatrix} \\ + \cdots + \eta' \begin{bmatrix} 0_{n-1\times n+h-d} \\ \mu' \left(K - 1_{n+1-d}\right) \left[0_{h-2\times n+1-d} \; 1_{n+1-d} \, 0_{n+1-d\times 1}\right] \end{bmatrix} \\ + \mu' \left(K - 1_{n+1-d}\right) \left[0_{h-1\times n+1-d} \; 1_{n+1-d}\right].\end{aligned}$$

Then the MSE is $a' \Sigma_{n+h-d} a$, which must be greater than (5). When the model is incorrectly specified, substitute the true DGP covariance matrix for Σ_{n+h-d} in the above quadratic form.

In general, direct and iterative forecasting utilize different formulas that are only identical in special cases. With explicit formulas, it is simple to compare forecast filters numerically. The direct filters are optimal when the DGP is correctly specified and parameters are estimated perfectly. However, in practice there is parameter estimation error and model misspecification. Thus it can happen that the iterated forecasts are actually superior.

3 Numerical Studies

We then examined forecast coefficients for various ARIMA models, as a function of h and n, for both the direct and iterated methods. First note that in simple cases like an ARIMA(1,0,0) or ARIMA(1,1,0) the forecast coefficients are the same, but more generally can be different. For ARMA and MA processes (with $d = 1, 2$) we observed that filter coefficients can be quite different when n and h are quite low. Of course as n increases, we expect movement towards the results for the semi-infinite past, and hence the discrepancies will disappear. This behavior manifested with n as low as 6 in many cases. Also, increasing h eventually makes the two methods coincide numerically. Figures 1, 2, and 3 display results for an ARIMA(0,1,2) model with (noninvertible) MA polynomial $1 + .8B + .2B^2$.

4 Conclusions

The work of Marcellino, Stock, and Watson (2006) provides the interesting conclusion that iterative forecasting actually works better out-of-sample for many time series. However, we caution the reader that their analysis is confined to the use of ARIMA(p,d,0) models fitted using OLS, according to either a one-step or multi-step ahead criterion, for iterated and direct methods respectively. The actual forecast functions used correspond to those of Section 2.1 (their equation (2.3) corresponds to our (1) when $h = 1$), which in this special case (because the model is autoregressive) does not involve an infinite span of data. Therefore the results of this paper indicate that the discrepancies in performance they observed were mainly due to the difference in parameter estimates between the two methods.

The work of Proietti (2011) formalized a way to compare forecast MSE for competing methods. This suggests a test statistic, computed as the difference of forecast MSEs for each method, with their respective parameter estimates plugged in. In order to isolate the unimportant finite-sample effects on the forecast filters, one could proceed by considering the semi-infinite past forecast MSEs, where the differences chiefly arise from the different parameter values used. Intriguingly, there is a further connection that can be made between these forecast MSEs and parameter estimation, discussed below.

If fitting via Gaussian MLE, this is equivalent to minimizing the one-step ahead forecast MSE function – take (2) with $h = 1$ and the periodogram substituted for the unknown spectrum $\widetilde{f}$ to get the Whittle likelihood, up to a term involving the logged innovation variance. Likewise, one might define a multi-step ahead fitting criterion based on the $h > 1$ case; this is developed at length in McElroy and Wildi (2011). Hence, if we obtain parameters for the direct and iterated methods in this way – not via the regression formulas of Marcellino, Stock, and Watson (2006) – then they automatically form zeroes of the gradients of the forecast MSE functions, indicating that the

resulting test statistic (for significant differences between direct and iterated forecast performance) may have some sort of χ^2 distribution. Future work should focus on this development.

To summarize, one may fix model and model estimation technique to be the same between the forecasting methods, and then their performance is extremely similar, even in finite sample. However, it is more common to use different fitting criteria (and the same model) in econometric practice, in which case the forecast weights can differ substantially (and iterative often performs better). Now, it could happen that – even though the parameter-fitting methods differ – the forecast weights come out pretty similar. We would like to know if their is a significant discrepancy in their performance, ahead of time. A test statistic involving forecast MSEs would cater to this need.

References

[1] Bhansali, R. (1996) Asymptotically efficient autoregressive model selection for multistep prediction problems. *Ann. Inst. Statist. Math.* **48**, 577–602.

[2] Bhansali, R. (1997) Direct autoregressive predictions for multistep prediction: order selection and performance relative to the plug in predictors." *Statistica Sinica* **7**, 425–449.

[3] Chevillon, G. and Hendry, D. (2005) Non-parametric direct multi-step estimation for forecasting economic processes. *International Journal of Forecasting* **21**, 201–218.

[4] Clements, M. and Hendry, D. (1996) Multi-step estimation for forecasting. *Oxford Bulletin of Economics and Statistics* **58**, 657–684.

[5] Durrett, R. (1996) *Probability: Theory and Examples*. New York: Duxbury Press.

[6] Findley, F. (1983) On the use of multiple models for multi-period forecasting. *Proceedings of the Business and Economics Statistics Section, Americal Statistical Association*, 528–531.

[7] Findley, F. (1985) Model selection for multi-step-ahead forecasting. *Proceedings of the Seventh Symposium on Identification and System Parameter Estimation* (H.A. Baker and P.C. Young, eds.) Pergamon, Oxford, 1039–1044.

[8] Kang, I. (2003) Multi-period forecasting using different models for different horizons: an application to U.S. economic time series data. *International Journal of Forecasting* **19**, 387–400.

[9] Lin, J. and Granger, C. (1994) Forecasting from non-linear models in practice. *Journal of Forecasting* **13**, 1–9.

[10] Marcellino, M., Stock, J., and Watson, M. (2006) A comparison of direct and iterated multistep AR methods for forecaasting microeconomic time series. *Journal of Econometrics* **135**, 499–526.

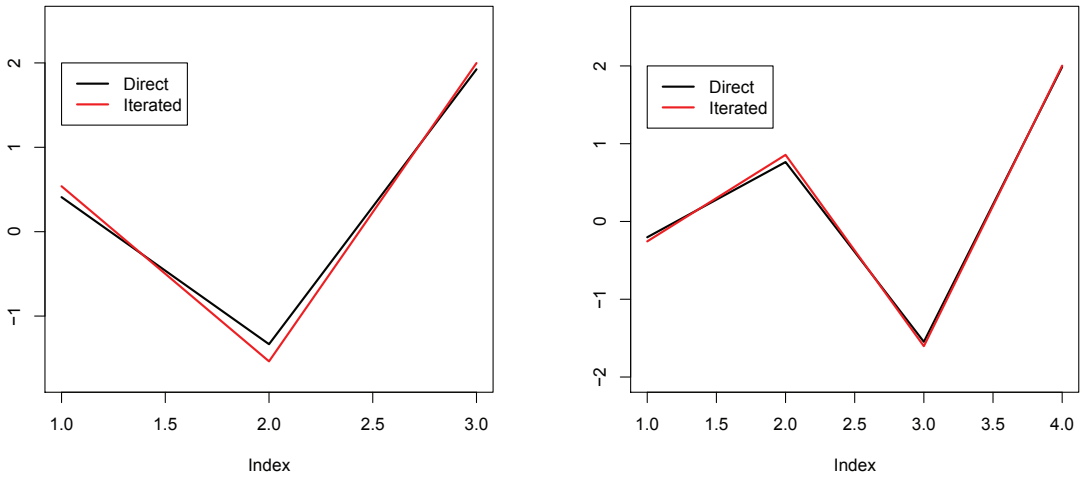

Figure 1: Forecast weights for the direct and iterated methods, for an ARIMA(0,1,2) model with forecast horizon $h = 2$ and $n = 3$ (left panel) or $n = 4$ (right panel) data points.

[11] McElroy, T. (2008) Matrix formulas for nonstationary ARIMA signal extraction. *Econometric Theory* **24**, 1–22.

[12] McElroy, T. and Findley, D. (2010) Discerning Between Models Through Multi-Step Ahead Forecasting Errors. *Journal of Statistical Planning and Inference* **140**, 3655–3675.

[13] McElroy, T. and Wildi, M. (2011) Multi-Step Ahead Estimation of Time Series Models. *Mimeo.*

[14] Proietti, T. (2011) Direct and iterated multistep AR methods for difference stationary processes. *International Journal of Forecasting* **27**, 266–280.

[15] Schorfheide, F. (2005) VAR forecasting under misspecfication. *Journal of Econometrics* **128**, 99–136.

[16] Tiao, G. and Tsay, R. (1994) Some advances in non-linear and adaptive modeling in time seies. *Journal of Forecasting* **13**, 109–131.

[17] Tiao, G. and Xu, D. (1993) Robustness of MLE for multi-step predictions: the exponential smoothing case. *Biometrika* **80**, 623–641.

[18] Weiss, A. (1991) Multi-step estimation and forecasting in dynamic models. *Journal of Econometrics* **48**, 135–149.

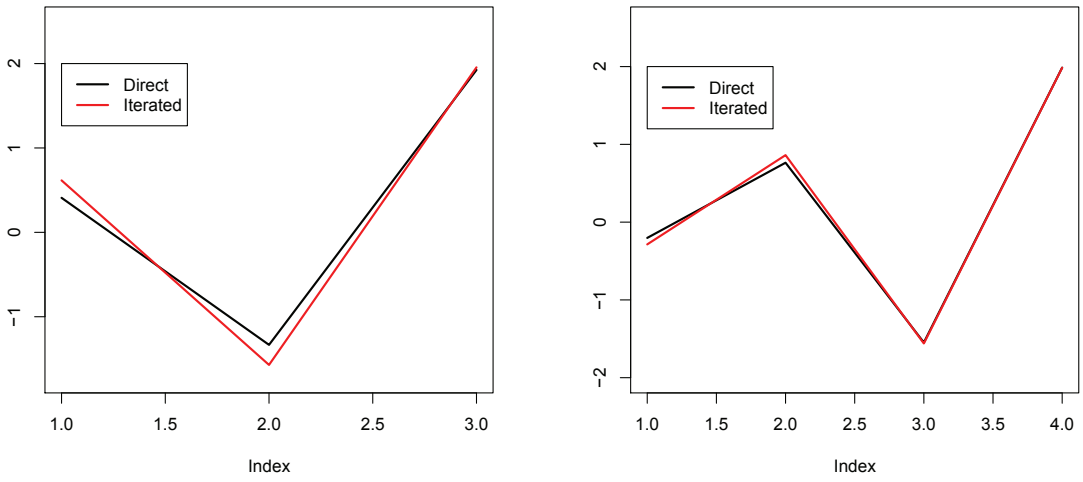

Figure 2: Forecast weights for the direct and iterated methods, for an ARIMA(0,1,2) model with forecast horizon $h = 3$ and $n = 3$ (left panel) or $n = 4$ (right panel) data points.

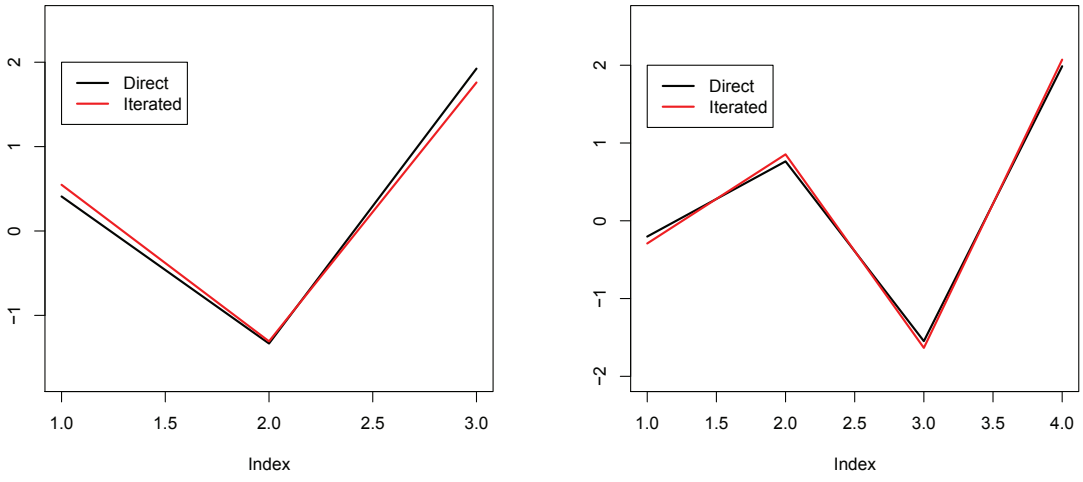

Figure 3: Forecast weights for the direct and iterated methods, for an ARIMA(0,1,2) model with forecast horizon $h = 6$ and $n = 3$ (left panel) or $n = 4$ (right panel) data points.