(This is a sample cover image for this issue. The actual cover is not yet available at this time.)

International Journal of Forecasting 29 (2013) 378–394

Contents lists available at SciVerse ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Multi-step-ahead estimation of time series models

Tucker McElroy [a], Marc Wildi [b,*]

[a] Center for Statistical Research and Methodology, US Census Bureau, 4600 Silver Hill Road, Washington, DC 20233-9100, United States
[b] Institute of Data Analysis and Process Design, Switzerland

ARTICLE INFO

Keywords:
ARIMA
Forecasting
Frequency domain
Nonstationary
Signal extraction

ABSTRACT

We study the fitting of time series models via the minimization of a multi-step-ahead forecast error criterion that is based on the asymptotic average of squared forecast errors. Our objective function uses frequency domain concepts, but is formulated in the time domain, and allows the estimation of all linear processes (e.g., ARIMA and component ARIMA). By using an asymptotic form of the forecast mean squared error, we obtain a well-defined nonlinear function of the parameters that is proven to be minimized at the true parameter vector when the model is correctly specified. We derive the statistical properties of the parameter estimates, and study the asymptotic impact of model misspecification on multi-step-ahead forecasting. The method is illustrated through a forecasting exercise, applied to several time series.

1. Introduction

It is well-known that fitting models via the minimization of one-step-ahead forecasting errors is equivalent to maximum likelihood estimation of the Gaussian likelihood for a stationary time series, and thus provides efficient parameter estimation for correctly specified Gaussian time series models; see Hannan and Deistler (1988), Dahlhaus and Wefelmeyer (1996), and Taniguchi and Kakizawa (2000). However, in reality, models are never specified correctly, and thus the maximum likelihood estimates converge to so-called "pseudo-true" values under certain regularity conditions, and these pseudo-true values minimize the Kullback–Leibler (KL) discrepancy between the specified model spectral density and the true spectrum. This approach can be viewed as an attempt to minimize the one-step-ahead forecast error for a given process, utilizing a certain misspecified model. Given that the focus for some applications is more on the forecasting performance at high leads, it is natural to consider the following questions: can we fit time series models such that the multi-step-ahead forecasting error is minimized? Is there an objective function which is analogous to KL, and which generalizes it to the multi-step-ahead case? What are the statistical properties of the resulting parameter estimates? This paper provides answers to some of these questions.

We present a Generalized Kullback–Leibler (GKL) measure — which is really a multi-step version of KL — and demonstrate that this measure can be derived directly from a multi-step-ahead forecasting error criterion. This GKL can be used to fit linear time series models with very little programming effort; in this paper we focus on the univariate ARIMA class[1] of models. The resulting parameter estimates are consistent for the pseudo-true values (i.e., those that minimize the GKL discrepancy between model and truth) under standard conditions, and are also asymptotically normal (consistency results under quite mild conditions are established by Findley, Pötscher, & Wei, 2004). When the model is specified correctly, these estimates are inefficient, i.e., they perform worse than the

* Corresponding author.
E-mail address: marc.wildi@zhaw.ch (M. Wildi).

[1] Although the KL does not depend on unit root factors in the total autoregressive polynomial (i.e., the differencing polynomial) for one-step-ahead forecasting, this object participates directly in the GKL function in the multi-step-ahead case.

0169-2070/$ – see front matter © 2012 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.ijforecast.2012.08.003

classical one-step-ahead estimates; we discuss the reasons for this below. However, since GKL is derived from a multi-step-ahead forecasting error criterion, it is reasonable to hope that the forecasts generated from such a model (at that particular lead) will perform better than the classical forecasts. This reflects an application-driven modeling philosophy: both the model specification and estimation should be oriented around a particular objective function associated with the application. McElroy and Findley (2010) address the model specification problem from a multi-step-ahead forecasting perspective, and here we focus on the model estimation aspect.

The GKL can be used to investigate the behavior of multi-step pseudo-true values — the minimizers of the discrepancy between the truth and the misspecified model — and is also the basis for actual parameter estimates that generalize the (one-step-ahead) quasi-maximum likelihood estimates associated with the Whittle likelihood. We note in passing that Theorem 4.3.1 of Hannan and Deistler (1988) provides a discussion of the equivalence of the Gaussian likelihood and the Whittle likelihood when the model is specified correctly; when misspecified, the proper reference is Dahlhaus and Wefelmeyer (1996).

Let us briefly discuss the econometric motivations for considering the multi-step-ahead perspective. Since, in reality, time series models are always misspecified, the crucial thing is to find a model that performs well according to the particular task which is of interest to the practitioner; using GKL as an objective function means that the practitioner is interested in a model that forecasts well at a particular lead time. In econometric business cycle analysis there is little interest in the mere one-step-ahead performance of misspecified models, since the period of a typical cycle is 8 to 40 observations for quarterly data. A model or collection of models that can forecast well at a lead of h for $8 \leq h \leq 40$ is needed here. Another application is in the field of seasonal adjustment, and, more generally, the area of real-time signal extraction. All model-based asymmetric signal extraction filters rely, either implicitly or explicitly, on long-horizon forecasts generated from the same misspecified model; see Dagum (1980), Findley, Monsell, Bell, Otto, and Chen (1998), McElroy (2008a), and Wildi (2004). Real-time (or concurrent) signal extraction is discussed by Wildi (2004, 2008), and the nefarious impact of model mis-specification on both long-term forecasting performance and signal extraction has been highlighted by numerous empirical studies. Beyond these obvious applications, any data analysis that is contingent on long-run forecasts — such as those that occur in climatology (e.g., the hot topic of global warming) and demographics (e.g., forecasting long-term changes in human population) — should not rely solely upon one-step-ahead forecasting model fitting criteria.

In light of these important motivations, there has been substantial prior work done on this topic that deserves a mention. Cox (1961) describes how multi-step-ahead forecast filters can be constructed from exponentially weighted moving averages, by fitting the smoothing parameter such that the forecast mean squared error is minimized when the underlying process is autoregressive. Tiao and Xu (1993) later expanded this work, pointing out that the exponential weighted moving average is the forecast filter that arises from multi-step forecasts from an ARIMA(0, 1, 1) model, where the moving average parameter is the negative of the exponential smoothing parameter. Their focus is on estimating the parameters of the forecast filter such that the multi-step-ahead forecast mean squared error is minimized. Another treatment of the topic is that of Gersch and Kitagawa (1983); they estimate structural models using a heuristic 12-step-ahead form of the usual Gaussian likelihood, expressed in a state space form. Their innovative paper illustrates the impact of a multi-step-ahead model fitting criterion on forecasting and trend estimation; as expected, the trends resulting from the 12-step-ahead criterion are much smoother than those derived from the classical approach. A more recent contribution is that of Haywood and Tunnicliffe-Wilson (1997), who provide an explicit formula for the objective function written in the frequency domain. A limitation of their formula is that, in general, the variables of the objective function do not correspond to ARMA parameters, as the paper essentially fits an unusual parametrization of moving average models.

There is also a substantial level of interest among econometricians in multi-step-ahead forecasting arising from autoregressive and difference autoregressive models. Marcellino, Stock, and Watson (2006) expound a common approach involving ordinary least squares estimation of these models, so as to minimize an empirical multi-step-ahead forecast error. Proietti (2011) expands on this work, investigating the forecasting performances of these multi-step-ahead fitted parameters. However, what has been lacking so far is a coherent general treatment of the subject that handles difference linear processes, i.e., nonstationary processes that have a Wold decomposition when suitably differenced. The main objective of this paper is to summarize and generalize all of the preceding literature, expressing the appropriate objective functions compactly in the frequency domain.

The reason for this recourse to the frequency domain is for the sake of concision of formulas, as well as computational efficiency. For example, certain of Tiao and Xu's (1993) formulas for the multi-step-ahead forecast mean square error involve infinite summations, which would only be calculated via truncation in practice. Using the frequency domain, exact expressions can be derived, utilizing the calculus of residues, and thus avoiding the need to truncate. Well-known Fourier transform algorithms can be used to compute the asymptotic multi-step-ahead forecast mean squared errors speedily, and in turn fit models to data, as well as determine pseudo-true values.

It is appropriate to outline the limitations of our approach. We do not consider multivariate time series models here; although the forecast error filter in this context is known and could, in principle, be used to generalize our GKL, its actual implementation has not yet been solved. However, it seems a fruitful direction for future work. Secondly, our method only optimizes over one forecast lead at a time, and simultaneous optimization over many leads is not considered; a discussion of this

is provided in Section 2, where we discuss a composite forecasting rule. Finally, our methods are exposited only for ARIMA models, where the gradient of the spectral density with respect to the parameter vector has a particularly simple form.

This paper provides the development of the asymptotic forecast mean squared error as a model fitting criterion in Section 2. A key contribution is the practical formula for its computation. The statistical properties of this GKL function and its optima are discussed in Section 3. Our formulation of the problem provides well-defined objective functions that are optimized by the true parameters when the model is specified correctly; otherwise, the parameter estimates converge to the GKL pseudo-true values. Section 4 explores the GKL function through several illustrations, both analytically and numerically. Then, in Section 5 we explore the discrepancy between the empirical forecast error and GKL through a chemical time series, and display results from a forecasting exercise involving housing starts. Here we take models that may be mis-specifications for the data, and fit them according to a variety of forecast lead criteria, generating the resulting forecasts. The multi-step out-of-sample forecasts are then computed and compared across model fitting criteria. Section 6 provides our conclusions, and the Appendix contains proofs and implementation notes for ARIMA models.

2. Forecasting as model fitting criteria

In this section we formulate a discrepancy measure for model fitting, which generalizes the KL discrepancy. This is derived from the asymptotic mean square multi-step-ahead forecasting error for that model. We utilize $\gamma_k(f)$ for the lag k autocorrelation function (acf) corresponding to a given spectral density f — with the convention that $\gamma_k(f) = (2\pi)^{-1}\int_{-\pi}^{\pi} f(\lambda)e^{i\lambda k}\,d\lambda$ — and its associated Toeplitz covariance matrix $\Gamma(f)$, whose jkth entry is simply $\gamma_{j-k}(f)$. We also use the notation $\langle g\rangle$ for any function g, with domain $[-\pi,\pi]$, to denote $(2\pi)^{-1}\int_{-\pi}^{\pi} g(\lambda)\,d\lambda$.

We will speak of time series models in terms of their spectral densities, since we are primarily concerned with the second-order behavior of difference stationary time series. It will be convenient to restrict ourselves to the "linear class" of spectra $\mathcal{L}$, consisting of integrable functions f that can be written as $f(\lambda) = |\Psi(e^{-i\lambda})|^2\sigma^2$ for some causal power series $\Psi(z) = \sum_{j\geq 0}\psi_j z^j$ (this holds iff $|\langle \log f\rangle| < \infty$, see Hannan & Deistler, 1988). We will assume that this is an invertible representation, so that $1/\Psi(z)$ is well-defined on the unit circle. Here $\psi_0 = 1$, and σ^2 is the innovation variance of the associated time series, i.e., $\sigma^2 = \exp\{\langle \log f\rangle\}$. Then a linear model is some subset $\mathcal{F}$ of $\mathcal{L}$ parametrized by a vector θ, and we may write $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ for a parameter space Θ; we will refer to $\mathcal{F}$ as a model.

When σ^2 is a parameter of the model, it does not depend upon the other components of θ, and we can order things such that σ^2 is the last component. If there are $r+1$ parameters in total, then $\theta_{r+1} = \sigma^2$, and we refer to the first r components by the notation $[\theta]$, which omits the innovation variance. In this case we say that f_θ is "separable". Clearly, $\nabla_{[\theta]}\sigma^2 = 0$ for separable models; if this gradient is nonzero, then σ^2 is not a parameter of the model, but rather a function of the other model parameters. Then we have $[\theta] = \theta$, for a total of r parameters; this case is referred to as a non-separable model. For example, ARMA models are separable, but component ARMA models are not. For a separable model, $f_{[\theta]}$ can be defined via f_θ/σ^2, and clearly only depends on $[\theta]$. In the non-separable case we use the same definition of $f_{[\theta]}$, by a convenient abuse of notation.

As was discussed by McElroy and Findley (2010), there exist simple formulas for the h-step-ahead forecast error from a given model applied to a semi-infinite sample of a process. The reason that we choose to base our approach on semi-infinite predictors, rather than finite sample predictors (see Newton & Pagano, 1983, for a discussion of their computation for stationary processes), is that we obtain a single time-invariant forecast error filter. This is in contrast to managing a suite of time-varying forecast error filters, the length of each of which depends upon one's time location within the sample; see the Appendix for a brief treatment of such a fitting criterion, which we will refer to as "Least Squares" (LS). The net effect of using the semi-infinite predictors is to create a computationally simpler objective function that is more tractable for asymptotic analysis and faster for estimation (there are no matrix inversions involved). Also, note that we consider the direct forecasting problem; see Marcellino et al. (2006), Proietti (2011), and Stock and Watson (1999) for comparisons to iterative one-step-ahead forecasting filters.

Suppose that our data process $\{X_t\}$ is differenced to stationarity with differencing operator $\delta(B)$, which has all its roots on the unit circle, such that the resulting $W_t = \delta(B)X_t$ is mean zero and stationary. Suppose that $\{W_t\}$ follows a model $f_\theta \in \mathcal{F}$, so that we can write $f_\theta(\lambda) = |\Psi(e^{-i\lambda})|^2\sigma^2$. Note that each coefficient ψ_j potentially depends on each of the first r components of θ. Then the h-step-ahead forecast error (based on an infinite past) at time t is equal to

$$\frac{[\Psi/\delta]_0^{h-1}(B)}{\Psi(B)}\,W_t;$$

see McElroy and Findley (2010) for the derivations, and Findley et al. (2004) for an alternative formulation. The square brackets denote the truncation of an infinite power series to those coefficients with indexes lying between the lower and upper bounds. In other words, $[\Psi/\delta]_0^{h-1}(B)$ is given by computing a (nonconvergent) power series $\Psi(B)/\delta(B)$, and taking only the first h terms. We then designate the rational filter $[\Psi/\delta]_0^{h-1}(B)\Psi^{-1}(B)$ as the h-step-ahead forecast error filter.

If this forecast error filter is applied to a semi-infinite sample from $\{W_t\}$, then the mean square of the resulting forecast errors equals

$$\left\langle \widetilde{f}\,\frac{|[\Psi/\delta]_0^{h-1}(e^{-i\cdot})|^2}{|\Psi(e^{-i\cdot})|^2}\right\rangle, \tag{1}$$

where $\widetilde{f}$ is the true spectral density of the Data Generating Process (DGP) for the $\{W_t\}$ series. Observe that this quantity depends explicitly on $\delta(B)$ if and only if $h > 1$, which means that the one-step-ahead forecast error does

not involve the unit root properties of the time series, whereas the multi-step-ahead forecast error does. McElroy and Findley (2010) utilize Eq. (1) as the basis of a model goodness-of-fit diagnostic; it is also related to the popular statistic of Diebold and Mariano (1995). However, in this paper we are primarily interested in using it to fit time series models; in this case, one could substitute the periodogram I (see below) for $\widetilde{f}$ in Eq. (1).

Let us rewrite Eq. (1) as a function of the model parameters $[\theta]$. For any $f \in \mathcal{L}$ and a given δ, define $f^{(h)}(\lambda)$ via $|[\Psi/\delta]_0^{h-1}(e^{-i\lambda})|^2$ (its dependence on δ is suppressed in this notation). Then, replacing $\widetilde{f}$ in Eq. (1) with a generic function g, we obtain

$$J([\theta], g) = \left\langle \frac{f^{(h)}_{[\theta]}}{f_{[\theta]}}\, g \right\rangle. \tag{2}$$

That is, $J([\theta], \widetilde{f})$ is the asymptotic mean square h-step-ahead forecast error arising from model $f_{[\theta]}$; note that the model's innovation variance plays no role in the forecast error filter. However, $J([\theta], I)$ is an empirical estimate of the mean squared error, where $I(\lambda) = n^{-1}|\sum_{t=1}^{n} W_t e^{-i\lambda t}|^2$ is the periodogram computed on a sample of size n taken from the differenced series $\{W_t\}$. As was discussed by McElroy and Findley (2010), with derivations from Findley et al. (2004), $J([\theta], I)$ approximately corresponds to the empirical sum $S([\theta])$ of h-step-ahead forecast errors calculated from finite-sample predictors (see the Appendix for the definition of $S([\theta])$).

When the model spectrum is separable, one can compute $J([\theta], g)$ for any given g. If it is non-separable, e.g., it is an unobserved components model, then computing the Wold coefficients is laborious. For instance, if the model consists of an ARMA(2,1) cycle plus white noise irregular (say, using the basic structural models described by Harvey, 1989), then the parameters readily determine the spectral density, but its Wold form Ψ must be determined using spectral factorization techniques. Note that spectral factorization will produce a moving average where the leading coefficient need not be unity; this can be factored into the innovation variance. In this way, Eq. (2) can be computed, although now $\nabla_{[\theta]}\sigma^2 \neq 0$. We henceforth assume that $J([\theta], g)$ can be evaluated; this is easy for ARIMA models, as is explained in the Appendix.

Now consider the minimization of $J([\theta], \widetilde{f})$ with respect to $[\theta]$: the optimum $[\widetilde{\theta}]$ yields a fitted model $f_{[\widetilde{\theta}]}$ with the smallest possible forecast error within the model $\mathcal{F}$. Likewise, we can obtain an empirical estimate by minimizing $J([\theta], I)$. Denote a minimum of J via $[\theta_g]$, where g is alternatively I or $\widetilde{f}$, depending on our interest. Consistency of $[\theta_I]$ for $[\theta_{\widetilde{f}}]$ will then follow from asymptotic results for linear functionals of the periodogram (see Section 3 below).

For the purposes of forecasting, a knowledge of $[\theta_g]$ is sufficient, because the forecast filter does not depend on the innovation variance. However, if a knowledge of the forecast precision is desired, we must also obtain σ^2. The true innovation variance is denoted by $\widetilde{\sigma}^2 = \exp\{\langle \log \widetilde{f} \rangle\}$, and we can write $\widetilde{f} = f_{[\theta_{\widetilde{f}}]}\widetilde{\sigma}^2$ whenever the model is correctly specified. If the model is separable, then the innovation variance (either true or empirical) can be computed via

$$\sigma_g^2 = \frac{J\left([\theta_g], g\right)}{J\left([\theta_g], f_{[\theta_g]}\right)}. \tag{3}$$

As usual, take $g = \widetilde{f}$ to be the true innovation variance, and $g = I$ to be our estimate of it. However, if the model spectrum is non-separable, we would already have determined σ_g^2 during the process of finding the Wold decomposition of the aggregate spectrum. That is, we would already know both f_{θ_g} and $f_{[\theta_g]}$, whose ratio is σ_g^2. Thus, Eq. (3) holds for both the separable and non-separable cases.

It follows that σ_I^2 will be consistent for $\widetilde{\sigma}^2$, as is shown in Section 3 below. Note that setting $g = \widetilde{f}$ in Eq. (3) provides an interpretation of the pseudo-true value of the innovation variance, i.e., $\sigma_{\widetilde{f}}^2$; namely, it is equal to the h-step-ahead forecast MSE $J([\theta_{\widetilde{f}}], \widetilde{f})$ arising from using the specified model, *divided by* the normalization factor $J([\theta_{\widetilde{f}}], f_{[\theta_{\widetilde{f}}]})$. When $h = 1$, this latter term equals unity, and plays no role, but when $h > 1$ it has an impact. As a result, we have no reason to expect $\sigma_{\widetilde{f}}^2$ to be increasing in h, even though the h-step-ahead forecast MSE is indeed typically increasing in h.

Thus, these equations together give us an algorithm: first minimize Eq. (2) with respect to $[\theta]$, then compute the minimal σ^2 via Eq. (3). When $g = \widetilde{f}$, this provides us with the so-called pseudo-true values (which in turn are h-step generalizations of the classical pseudo-true values of the KL discrepancy, cf. Taniguchi & Kakizawa, 2000), and these are equal to the true parameters when the model is specified correctly. However, when $g = I$, this method provides us with parameter estimates $([\theta_I], \sigma_I^2)$ that are consistent for the set of pseudo-true values (regardless of whether the model is correctly or incorrectly specified).

We now make some further connections between J and the KL discrepancy. It is well-known that the log Gaussian likelihood for the differenced data $\{W_t\}$ is approximately proportional to the Whittle likelihood (Taniguchi & Kakizawa, 2000), which is simply the KL discrepancy between the periodogram I and the model f_θ. This KL discrepancy can be computed for any two positive bounded functions f, g via the formula

$$KL(f, g) = \langle \log f + g/f \rangle. \tag{4}$$

If we wish to fit a model to the data, we minimize $KL(f_\theta, I)$ with respect to θ, and denote the resulting estimate by θ_I. This can be done in two steps when f_θ is separable, since then the KL is rewritten as $\log\sigma^2 + \sigma^{-2}\langle I/f_{[\theta]}\rangle$, so that the optimal σ_I^2 equals $\langle I/f_{[\theta_I]}\rangle$ (this requires $\nabla_{[\theta]}\sigma^2 = 0$). In other words, when the model is separable, the minimization of KL is equivalent to the two-step minimization of Eqs. (2) and (3) for $h = 1$.

The Generalized Kullback–Leibler (GKL) discrepancy is therefore defined analogously for $h \geq 1$:

$$GKL_\delta^{(h)}(f, g) = \langle \log f \rangle + \log\langle f^{(h)} \rangle + \frac{\left\langle \frac{g}{f} f^{(h)} \right\rangle}{\langle f^{(h)} \rangle}. \tag{5}$$

Note that this reduces to Eq. (4) when $h = 1$, since then $f^{(h)} \equiv 1$. However, for $h > 1$ we have the extra $\log\langle f^{(h)}\rangle$ term, without which the minimization of Eq. (5) would

not be equivalent to optimization via Eqs. (2) and (3). This relationship is described in Proposition 2 of Section 3.

In practice, we can utilize the identities $\langle g \rangle = \gamma_0(g)$ and

$$\langle fI \rangle = \frac{1}{n} W' \Gamma(f) W$$

to compute $GKL_\delta^{(h)}(f_\theta, I)$, where $W = (W_1, W_2, \ldots, W_n)'$ is the available sample. Also, because $\langle \log f_{[\theta]} \rangle = 0$, we obtain

$$GKL_\delta^{(h)}(f_\theta, I) = \log\left(\sigma^2 \gamma_0(f_{[\theta]}^{(h)})\right) + \frac{W' \Gamma\left(f_{[\theta]}^{(h)}/f_{[\theta]}\right) W}{n\sigma^2 \gamma_0(f_{[\theta]}^{(h)})}. \quad (6)$$

This is quite easy to compute for ARIMA models, for which the autocovariances are readily obtained (see the Appendix). In particular, no matrices need to be inverted (unlike with maximum likelihood estimation). The computation of multi-step forecasts and forecast error covariances from a finite past for stationary processes is discussed by Newton and Pagano (1983); our approach utilizes semi-infinite forecast error filters instead, thereby avoiding much of the complexity required for matrix inversion.

The formula also holds for the non-separable case, but one must first determine the Wold decomposition for f_θ, as described above. The pseudo-true values, i.e., the values to which parameter estimates converge, are given by the minimizers of $GKL_\delta^{(h)}(f_\theta, \widetilde{f})$. The statistical properties of the parameter estimates are covered in Section 3.

Thus, Eq. (6) gives a unified method for fitting models to time series data W, which generalizes Whittle estimation from $h = 1$ to $h > 1$. If this procedure is repeated over a range of h, say $1 \leq h \leq H$ for some user-defined forecast threshold H, we obtain many different fits of the specified model, with each corresponding parameter estimate $\widehat{\theta}^{(h)}$ yielding the optimal h-step-ahead (asymptotic) mean square forecast error. Of course, these parameters will vary widely in practice, since there is no need for optimality to be achieved over a range of forecast leads for one single choice of parameters; this is illustrated in the numerical studies of Section 4. Having multiple parameter fits of the same model available is useful, since each fit is optimal with respect to its own h-step-ahead forecasting objective.[2]

Hence, a strategy for optimal multi-step-ahead forecasting is the following. For each h desired, utilize the forecast filters based on the model fitted according to the $GKL_\delta^{(h)}$ criterion. Over repeated forecasts, in an average sense, this procedure should prove to be advantageous (this is necessarily so in-sample). We refer to this process as the composite forecasting rule. It is explored further in Section 5 on a real time series.

3. Statistical properties of the estimates

In this section we develop the statistical properties of GKL. First we present gradient and Hessian expressions for the separable and non-separable cases. Optimization of GKL can then easily be related to the optimization of the multi-step-ahead forecasting error J. We then state the consistency and asymptotic normality results for the parameter estimates under standard regularity conditions.

We begin by studying $GKL_\delta^{(h)}(f_\theta, g)$ as a function of θ, abbreviated as $G(\theta)$. It follows from the definition in Eq. (5) that

$$G(\theta) = \log \sigma^2 + \log\langle f_{[\theta]}^{(h)} \rangle + \frac{J([\theta], g)}{\sigma^2 \langle f_{[\theta]}^{(h)} \rangle}. \quad (7)$$

Note that σ^2 may well depend upon $[\theta]$ in the non-separable case, but this dependency will be suppressed in the notation. Now Eq. (7) is convenient because it involves the function J. We begin our treatment by noting that $J([\theta], \widetilde{f})$ has a global minimum at $[\theta_{\widetilde{f}}]$ when the model is correctly specified; this follows from MSE optimality of the h-step-ahead forecast filter. In this case, $\widetilde{f} \in \mathcal{F}$, and there exists $\widetilde{\theta}$ such that $\widetilde{f} = f_{\widetilde{\theta}}$, so that $[\theta_{\widetilde{f}}] = [\widetilde{\theta}]$ when the minimum is unique (this is really a property of the parametrization of the model).

We next state the gradient and Hessian functions of G for the separable and non-separable cases. In the former case, $\nabla_\theta' = \left[\nabla_{[\theta]}', \frac{\partial}{\partial \sigma^2}\right]$, whereas in the latter case we have $\nabla_\theta = \nabla_{[\theta]}$, since there is no differentiation with respect to innovation variance (σ^2 is not a parameter).

Proposition 1. *For a separable model, the gradient and Hessian functions of GKL are given by*

$$\nabla_{[\theta]} G(\theta) = \frac{\langle \nabla_{[\theta]} f_{[\theta]}^{(h)} \rangle}{\langle f_{[\theta]}^{(h)} \rangle} + \frac{\nabla_{[\theta]} J([\theta], g)}{\sigma^2 \langle f_{[\theta]}^{(h)} \rangle} - \frac{J([\theta], g) \langle \nabla_{[\theta]} f_{[\theta]}^{(h)} \rangle}{\sigma^2 \langle f_{[\theta]}^{(h)} \rangle^2}$$

$$\frac{\partial}{\partial \sigma^2} G(\theta) = \sigma^{-2} - \sigma^{-4} \frac{J([\theta], g)}{\langle f_{[\theta]}^{(h)} \rangle}$$

$$\nabla_{[\theta]} \nabla_{[\theta]}' G(\theta) = \frac{\langle \nabla_{[\theta]} \nabla_{[\theta]}' f_{[\theta]}^{(h)} \rangle}{\langle f_{[\theta]}^{(h)} \rangle} - \frac{\langle \nabla_{[\theta]} f_{[\theta]}^{(h)} \rangle \langle \nabla_{[\theta]}' f_{[\theta]}^{(h)} \rangle}{\langle f_{[\theta]}^{(h)} \rangle^2} - \frac{\nabla_{[\theta]} J([\theta], g) \langle \nabla_{[\theta]}' f_{[\theta]}^{(h)} \rangle + \langle \nabla_{[\theta]} f_{[\theta]}^{(h)} \rangle \nabla_{[\theta]}' J([\theta], g)}{\sigma^2 \langle f_{[\theta]}^{(h)} \rangle^2} + \frac{\nabla_{[\theta]} \nabla_{[\theta]}' J([\theta], g)}{\sigma^2 \langle f_{[\theta]}^{(h)} \rangle} - \frac{\langle \nabla_{[\theta]} \nabla_{[\theta]}' f_{[\theta]}^{(h)} \rangle J([\theta], g)}{\sigma^2 \langle f_{[\theta]}^{(h)} \rangle^2} + 2 \frac{\langle \nabla_{[\theta]} f_{[\theta]}^{(h)} \rangle \langle \nabla_{[\theta]}' f_{[\theta]}^{(h)} \rangle J([\theta], g)}{\sigma^2 \langle f_{[\theta]}^{(h)} \rangle^3}$$

$$\frac{\partial}{\partial \sigma^2} \nabla_{[\theta]} G(\theta) = \sigma^{-4} J([\theta], g) \frac{\langle \nabla_{[\theta]} f_{[\theta]}^{(h)} \rangle}{\langle f_{[\theta]}^{(h)} \rangle^2} - \sigma^{-4} \frac{\nabla_{[\theta]} J([\theta], g)}{\langle f_{[\theta]}^{(h)} \rangle}$$

$$\frac{\partial^2}{\partial^2 \sigma^2} G(\theta) = -\sigma^{-4} + 2\sigma^{-6} \frac{J([\theta], g)}{\langle f_{[\theta]}^{(h)} \rangle}.$$

[2] The first author thanks Donald Gaver for this insightful perspective.

For a non-separable model, the gradient and Hessian functions of GKL are given by

$$\nabla_\theta G(\theta) = \left(\frac{\nabla_\theta \sigma^2}{\sigma^2} + \frac{\langle \nabla_\theta f^{(h)}_{[\theta]}\rangle}{\langle f^{(h)}_{[\theta]}\rangle}\right)\left(1 - \frac{J([\theta], g)}{\sigma^2 \langle f^{(h)}_{[\theta]}\rangle}\right) + \frac{\nabla_\theta J([\theta], g)}{\sigma^2 \langle f^{(h)}_{[\theta]}\rangle}$$

$$\begin{aligned}\nabla_\theta \nabla'_\theta G(\theta) = {} & \left(\frac{\nabla_\theta \nabla'_\theta \sigma^2}{\sigma^2} - \frac{\nabla_\theta \sigma^2 \nabla'_\theta \sigma^2}{\sigma^4} + \frac{\nabla_\theta \nabla'_\theta f^{(h)}_{[\theta]}}{\langle f^{(h)}_{[\theta]}\rangle} - \frac{\nabla_\theta f^{(h)}_{[\theta]} \nabla'_\theta f^{(h)}_{[\theta]}}{\langle f^{(h)}_{[\theta]}\rangle^2}\right) \cdot \left(1 - \frac{J([\theta], g)}{\sigma^2 \langle f^{(h)}_{[\theta]}\rangle}\right) \\ & + \frac{J([\theta], g)}{\sigma^2 \langle f^{(h)}_{[\theta]}\rangle}\left(\frac{\nabla_\theta \sigma^2}{\sigma^2} + \frac{\nabla_\theta f^{(h)}_{[\theta]}}{\langle f^{(h)}_{[\theta]}\rangle}\right) \\ & \times \left(\frac{\nabla'_\theta \sigma^2}{\sigma^2} + \frac{\nabla'_\theta f^{(h)}_{[\theta]}}{\langle f^{(h)}_{[\theta]}\rangle}\right) \\ & - \frac{\nabla_\theta J([\theta], g)}{\sigma^2 \langle f^{(h)}_{[\theta]}\rangle}\left(\frac{\nabla'_\theta \sigma^2}{\sigma^2} + \frac{\nabla'_\theta f^{(h)}_{[\theta]}}{\langle f^{(h)}_{[\theta]}\rangle}\right) \\ & - \left(\frac{\nabla_\theta \sigma^2}{\sigma^2} + \frac{\nabla_\theta f^{(h)}_{[\theta]}}{\langle f^{(h)}_{[\theta]}\rangle}\right)\frac{\nabla'_\theta J([\theta], g)}{\sigma^2 \langle f^{(h)}_{[\theta]}\rangle} \\ & + \frac{\nabla_\theta \nabla'_\theta J([\theta], g)}{\sigma^2 \langle f^{(h)}_{[\theta]}\rangle}.\end{aligned}$$

The proof follows from calculus, and is omitted. These expressions are written in terms of J, and its gradient and Hessian, which can also be expanded further algebraically. The resulting expressions could be used in the numerical optimization of GKL, though the implementation would be quite burdensome: it would require the calculation of the various derivatives of spectral densities and their associated inverse Fourier Transforms (FTs). Thus, for many models, these formulas are of no practical use, although they do serve the purpose of establishing that the minimization of GKL coincides with the minimization of Eq. (2), together with the computation of Eq. (3).

Proposition 2. *Suppose that the model is separable. If* $[\theta_g]$ *is a minimum of Eq. (2) and* σ_g^2 *is computed via Eq. (3), then* $([\theta_g], \sigma_g^2)$ *is a global minimum of* $G(\theta)$. *Conversely, for any minimizer* θ_g *of* $G(\theta)$, $[\theta_g]$ *minimizes* $J([\theta], g)$. *The minimal value of* G *is* $1 + \log J([\theta_g], g)$. *When the model is non-separable, the minima of* $J([\theta], g)$ *are also minimizers of* $G(\theta)$.

Thus, GKL really corresponds to the multi-step-ahead forecast error minimization problem. As a practical matter, minimization of Eq. (2) is more convenient than minimization of GKL, as it involves one parameter fewer (in the separable case). However, GKL is more convenient as a discrepancy measure between spectra, and for establishing asymptotic results for parameter estimates.

Proposition 2 can be adapted to data fitting (let $g = I$) or the computation of pseudo-true values (let $g = \widetilde{f}$). We always assume that the order of integration d has been specified correctly, and that appropriately differenced data are passed into the routines.

Recall that when the model is correctly specified, $\theta_{\widetilde{f}}$ corresponds to the true parameter vector $\widetilde{\theta}$, and we can expect that θ_I will converge to this value. However, when the model is misspecified, θ_I converges to $\theta_{\widetilde{f}}$ under fairly classical regularity conditions. A first treatment of consistency has been given by Findley et al. (2004), but here we extend the result to asymptotic normality under some more stringent conditions. Our central limit theorem shows that multi-step estimation has asymptotic variance that, in general, is not equal to the inverse of the Fisher information matrix, when the model is correctly specified. This implies that the estimates are inefficient, but when the model is misspecified, we can no longer say what types of estimates have the minimal variance, except on a case by case basis.

We shall assume that our pseudo-true parameters are not on the boundary of the parameter set, because the limit theory is non-standard in this case (cf. Self & Liang, 1987). If the pseudo-true parameter is unique, the Hessian of GKL should be positive definite at that value, and hence invertible. The so-called Hosoya-Taniguchi (HT) conditions (Hosoya & Taniguchi, 1982; Taniguchi & Kakizawa, 2000) impose sufficient regularity on the process $\{W_t\}$ to ensure a central limit theorem; these conditions require that the process is a causal filter of a higher-order martingale difference. Finally, we suppose that the fourth order cumulant function of the process is identically zero, which says that the process looks Gaussian in terms of the second and fourth order structure. This condition is not strictly necessary, but facilitates a simple expression for the asymptotic variance of the parameter estimates. Let the Hessian of $G(\theta)$ with $g = \widetilde{f}$ be denoted $H(\theta)$.

Theorem 1. *Suppose that* $\theta_{\widetilde{f}}$ *exists uniquely in the interior of* Θ *and that* $H(\theta_{\widetilde{f}})$ *is invertible. Suppose that the process* $\{W_t\}$ *has finite fourth moments, conditions* (HT1)–(HT6) *of* Taniguchi and Kakizawa (2000, pp. 55–56) *hold, and that the fourth order cumulant function of* $\{W_t\}$ *is zero. Then, as* $n \to \infty$,

$$\sqrt{n}\left(\theta_I - \theta_{\widetilde{f}}\right) \stackrel{\mathcal{L}}{\Longrightarrow} \mathcal{N}\left(0, H^{-1}(\theta_{\widetilde{f}}) V(\theta_{\widetilde{f}}) H^{-1}(\theta_{\widetilde{f}})\right). \tag{8}$$

Here, the matrix $V(\theta, f)$ *is defined via*

$$V(\theta) = 2\left\langle \nabla_\theta \frac{f^{(h)}_{[\theta]}}{f_\theta \langle f^{(h)}_{[\theta]}\rangle} \nabla'_\theta \frac{f^{(h)}_{[\theta]}}{f_\theta \langle f^{(h)}_{[\theta]}\rangle} \widetilde{f}^2 \right\rangle.$$

Remark 1. In order to produce estimated standard errors for parameter estimates, it is best to proceed as if the model was misspecified (since otherwise we will mis-state the uncertainty); the quantities in $H^{-1}VH^{-1}$ are computed by substituting parameter estimates for pseudo-true values, while plugging in I for $\widetilde{f}$ and $I^2/2$ for $\widetilde{f}^2$ (cf. Chiu, 1988; and McElroy & Holan, 2009). With these substitutions, the matrices can be computed using quadratic forms in the data vector W, as well as its sample autocovariance vector. Of course, if the exact gradient and Hessian are already used in the numerical optimization procedure, then these quantities can be used to find H.

4. Illustrations

Although in general it is difficult to compute $J([\theta], g)$ explicitly, it is possible in some special cases. We first provide several analytical examples involving stationary and nonstationary DGPs, then consider several numerical illustrations of the GKL objective functions.

4.1. Analytical derivations of optima

When forecasting stationary processes long-term, the forecasts tend to revert to the mean independent of the parameter values (this can also be seen in the large h behavior of GKL when $\delta(z) = 1$), and as a result the objective function will be flat on the majority of its domain, i.e., changes in parameter values will have no impact on the forecasting performance. This situation is dramatically different in the presence of non-stationarity, because the large h behavior of GKL then tends to infinity, rather than a constant. Our results below are computed in terms of a generic g, which can be taken as either I or $\widetilde{f}$, as the context dictates.

First consider fitting an AR(1) model, and denote the AR parameter by ϕ. Then

$$f_{[\theta]}(\lambda) = |1 - \phi z|^{-2}$$

$$f^{(h)}_{[\theta]}(\lambda) = \left| \sum_{j=0}^{h-1} \phi^j z^j \right|^2$$

$$\begin{aligned} J([\theta], g) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} g(\lambda) |1 - \phi^h z^h|^2 \, d\lambda \\ &= (1 + \phi^{2h}) \gamma_0(g) - 2\phi^h \gamma_h(g). \end{aligned}$$

Thus, the concentrated objective function is equal to $\gamma_0(g)$ times $1 - \rho_h(g) + (\phi^{2h} - \rho_h(g))^2$. Unless the correlation is negative and h is even, this is minimized by ϕ_g satisfying $\phi_g^h = \rho_h(g) = \gamma_h(g)/\gamma_0(g)$ (otherwise the minimizer is $\phi_g = 0$). When $\phi_g = \rho_h^{1/h}(g)$, then $\sigma_g^2 = \gamma_0(g)(1 - \phi_g^2)$, and the minimal h-step forecast error is $J([\theta_g], g) = \gamma_0(g)(1 - \rho_h^2(g))$. A glance at the formulas for σ_g^2 and $J([\theta_g], g)$ illustrates a point made in Section 2: although the latter is increasing in h (note that $\rho_h^2(g) \to 0$ as $h \to \infty$ for processes with summable autocovariance functions), the former need not be, as in the ARMA example below.

Let us further suppose that g is the spectrum of an AR(1), so that $\rho_h(g) = \widetilde{\phi}^h$. Then $\phi_g = \widetilde{\phi}$, the case of consistency in the presence of correct model specification. It is easy to check that σ_g^2 equals the innovation variance of g as well. The minimal forecast error function is proportional to $(1 - \widetilde{\phi}^{2h})/(1 - \widetilde{\phi}^2)$, an increasing function in h. If instead g is the periodogram of the above AR(1), our estimate is the hth root of $\gamma_h(I)/\gamma_0(I)$, the lag h (biased) sample autocorrelation. There is an efficiency loss, in general, in using this estimate versus just $\gamma_1(I)/\gamma_0(I)$.

Next, suppose that g is the spectrum of an ARMA(1,1) of the form

$$f_{[\widetilde{\theta}]}(\lambda) = \frac{|1 + \widetilde{\omega} z|^2}{|1 - \widetilde{\varphi} z|^2}.$$

The MA(∞) representation of the process has coefficients $\psi_j = \widetilde{\varphi}^{j-1}(\widetilde{\varphi} + \widetilde{\omega})$ for $j \geq 1$ and $\psi_0 = 1$. We then obtain the autocorrelation sequence (cf. Box & Jenkins, 1976) $\rho_h(g) = \widetilde{\varphi}^{h-1}(\widetilde{\varphi} + \widetilde{\omega})(1 + \widetilde{\varphi}\widetilde{\omega})(1 + 2\widetilde{\omega}\widetilde{\varphi} + \widetilde{\omega}^2)^{-1}$. Thus, ϕ_g is equal to either zero, in which case this correlation $\rho_h(g)$ is negative and h is even, or $\rho_h^{1/h}(g)$. Therefore, as $h \to \infty$, $\rho_h^{1/h}(g) \to \widetilde{\varphi}$ and the MA parameter has no impact on the minima, which is interesting; this is because the AR parameter governs the long-term serial correlation. Also, $\sigma_g^2 = \widetilde{\sigma}^2 \cdot (1 - \widetilde{\phi}^2)$, which shows that the pseudo-true value of the innovation variance is less than the actual true $\widetilde{\sigma}^2$. Moreover, $\widetilde{\phi}$ is an increasing function of h, so that σ_g^2 is *decreasing* in h.

Finally, suppose that the process is a gap AR(2) with spectrum

$$f_{[\widetilde{\theta}]}(\lambda) = |1 - \widetilde{\varphi} z^2|^{-2}.$$

The autocorrelations are zero at odd lags, and are equal to $\widetilde{\varphi}^{h/2}$ when the lag h is even. Then, ϕ_g equals 0 whenever h is odd, and equals $\sqrt{|\widetilde{\varphi}|}$ unless $\widetilde{\varphi} < 0$ and $h \equiv 2 \bmod 4$, in which case $\phi_g = 0$ as well.

Now suppose that we fit an MA(q) model, which has spectral density

$$f_{[\theta]}(\lambda) = \left| 1 + \sum_{j=1}^{q} \omega_j z^j \right|^2.$$

The resulting expression for J is generally fairly complicated, but when $h > q$ we have $f^{(h)}_{[\theta]} = f_{[\theta]}$, so that $J([\theta], g) = \gamma_0(g)$. Thus, the concentrated objective function is completely flat with respect to the parameters. This reflects the fact that an MA(q) model has no serial information by which to forecast at leads exceeding q. However, this aspect is no longer relevant when non-stationary differencing is present.

In particular, suppose $q = 1$ and $h = 1$, so that

$$J(\omega, g) = \frac{\gamma_0(g) + 2 \sum_{k \geq 1} \gamma_k(g)(-\omega)^k}{1 - \omega^2}.$$

This poses a highly non-linear optimization problem, unless g has a special form.

The ARIMA(0, 1, 1) model was studied by Tiao and Xu (1993), and can easily be adapted to fit our framework; write the MA polynomial as $1 - \theta B$ and consider an arbitrary h. Then,

$$[\Psi/\delta]_0^{h-1}(B) = 1 + (1 - \theta) \sum_{j=1}^{h-1} B^j$$

when $h > 1$. The full forecast error filter works out to be

$$\begin{aligned} \frac{[\Psi/\delta]_0^{h-1}(B)}{\Psi(B)} &= \frac{1 - B^h}{1 - B} + \frac{\theta B^h}{1 - \theta B} \\ &= \frac{1 + (1 - \theta) \sum_{j=1}^{h-1} B^j}{1 - \theta B}. \end{aligned}$$

Note that this filter corresponds to the transfer function of an ARMA(1, h), and its Wold coefficients have the curious pattern of being equal to unity up to index $h - 1$, and equal

to θ^{k-h+1} at index k when $k \geq h$. Then the autocovariance sequence satisfies

$$\gamma_k(f_\theta) = \begin{cases} h - k + \theta\,\dfrac{1-\theta^k}{1-\theta} + \theta^k\,\dfrac{\theta^2}{1-\theta^2} & k < h \\ \theta^{k-h+1}\,\dfrac{1-\theta^k}{1-\theta} + \theta^k\,\dfrac{\theta^2}{1-\theta^2} & k \geq h. \end{cases}$$

Then $J(\theta, g) = \sum_k \gamma_k(f_\theta)\gamma_k(g)$, and substituting our expressions yields Eq. (2.3) of Tiao and Xu (1993). Numerical minimization with $g = I$ essentially truncates the infinite summations to the sample size, because $\gamma_k(I) = 0$ for $|k| \geq n$. It is hard to say anything about pseudo-true values analytically, as the optimization problem is highly nonlinear.

Finally, consider the example of an ARIMA(1, 1, 0), which was fitted for multi-step-ahead forecasting via ordinary least squares by Marcellino et al. (2006). Denote the AR polynomial by the usual $1 - \phi B$. Then the forecast error filter is

$$\frac{[\Psi/\delta]_0^{h-1}(B)}{\Psi(B)} = \frac{1-B^h}{1-B} - \phi\,\frac{1-\phi^h}{1-\phi}\,B^h.$$

This corresponds to an MA(h) with all unit entries except the last coefficient, which is equal to $-\phi(1-\phi^h)(1-\phi)^{-1}$; call this $\zeta(\phi)$ for short. Then the autocovariances have a simple structure: $\gamma_0(f_\phi) = h + \zeta^2(\phi)$ and $\gamma_k(f_\phi) = h - k + \zeta(\phi)$ for $k \leq h$, and zero otherwise. Then $J(\phi, g)$ will still be nonlinear in ϕ, but it is interesting that only a finite number of autocovariances of g are involved. In particular,

$$J(\phi, g) = \gamma_0(g)[h + \zeta^2(\phi)] + 2\sum_{k=1}^{h} \gamma_k(g)[h - k + \zeta(\phi)].$$

Taking the derivative with respect to ϕ provides two solutions: either $\dot\zeta(\phi) = 0$, or we must have $\zeta(\phi) = -\sum_{k=1}^{h} \rho_k(g)$. The first case demands a solution to

$$0 = 1 + 2\phi + 2\phi^2 + \cdots + h\phi^{h-1}$$

and in no way depends on the properties of g. The second case requires the polynomial equation

$$\phi + \phi^2 + \cdots + \phi^h = \sum_{k=1}^{h} \rho_k(g),$$

to be solved, which is done trivially by root-finding. Note that when $h = 1$ we recover the familiar $\phi_g = \rho_1(g)$; recall that the differencing operator has no impact on parameter estimates when $h = 1$, so we should just be fitting the AR(1) to the differenced data, as indicated by the Whittle likelihood. When $h > 1$, a different solution is called for; in this particular case it is very fast to compute.

4.2. Numerical calculation of pseudo-true optima

We look at experimental results by determining pseudo-true values for a range of DGPs and models. By examining the resulting concentrated GKL objective functions and the pseudo-true values, we can get a sense of how each model is fitted to the respective DGPs. We will consider the Local Level Model (LLM) of Harvey (1989), which is defined as consisting of a random walk trend plus independent white noise. Such a process can be re-written as an ARIMA(0, 1, 1), where the MA polynomial is $1 - \theta B$ (with $\theta \geq 0$), as in the previous subsection. If the signal-to-noise ratio (SNR) is $q > 0$, i.e., the innovation variance of the random walk component is $q\sigma^2$ and the white noise variance is σ^2, then it is known that

$$\theta = \frac{q + 2 - \sqrt{q^2 + 4q}}{2},$$

by solving the spectral factorization problem. Note that as $q \to 0$ we obtain $\theta \to 1$, or in other words the process becomes more like a pure white noise as the SNR decreases. We also consider the Smooth Trend Model (STM) of Harvey (1989), which is like the LLM except with two differencings. Then the aggregate process is an ARIMA(0, 2, 2), and the coefficients ω_1, ω_2 of the MA(2) are complicated functions of the signal-to-noise ratio (see McElroy, 2008b, for a spectral factorization of the STM).

For our numerical studies, our DGPs are selected from the following list, where $d = 1, 2$; we do not consider $d = 0$, for the reasons discussed in the previous subsection. In general, we use the notation $\Omega(z) = 1 + \omega_1 z + \omega_2 z^2$ and $\Phi(z) = 1 - \phi_1 z - \phi_2 z^2$ for the ARMA process with the MA polynomial Ω and AR polynomial Φ; also, the innovation variance $\sigma^2 = 1$ in all cases.

- D1: $d = 1, \omega_1 = -0.1, \omega_2 = 0 = \phi_1 = \phi_2$.
- D2: $d = 1, \omega_1 = -0.8, \omega_2 = 0 = \phi_1 = \phi_2$.
- D3: $d = 1, \omega_1 = 0.7, \phi_1 = 0.2$, and $\omega_2 = 0 = \phi_2$.
- D4: $d = 1, \phi_1 = 0.9\cos(\pi/60), \phi_2 = -0.81$, and $\omega_1 = \omega_2 = 0$.
- D5: $d = 2, \phi_1 = 0 = \phi_2, \omega_1, \omega_2$ corresponding to SNR = 0.1 in STM.
- D6: $d = 2, \phi_1 = 0 = \phi_2, \omega_1, \omega_2$ corresponding to SNR = 10 in STM.

This provides an interesting collection of DGPs. The first two processes correspond to the LLM with a high (D1) and a low (D2) trend-to-noise ratio, respectively. The STM is explored for different values of the SNR through D5 and D6. Process D3 follows a mixed ARMA model, while D4 generates a cyclical effect with a period of 60 observations. The models considered are ARIMA(p, d, q) with $p = 1, q = 0$ (AR), $p = 0, q = 1$ (MA), and $p = 0 = q$ (WN), with $d = 1, 2$ corresponding to the DGP.

This gives 18 combinations of models and DGPs. For the AR and MA models, the objective function J in Eq. (2) can be computed, and is displayed in Figs. 1 and 2 for $1 \leq h \leq 10$ as a function of the single parameter (the individual objective functions are not labeled with regard to h, to avoid cluttering the picture). In some cases the minima are fairly obvious and change smoothly with respect to h, but in other cases the objective functions can be either flat (resulting in less reliable estimates of the optima) or criss-crossing (resulting in oscillatory patterns in the optima as h changes). Tables 1 to 6 summarize the numerical minima, and also present the pseudo-true innovation variances.

Table 1
Pseudo-true values for models fitted to DGP D1.

Minima											
Models		Leads									
		1	2	3	4	5	6	7	8	9	10
AR(1)	$\widetilde{\phi}$	−0.0998	−0.1118	−0.1098	−0.1098	−0.1098	−0.1098	−0.1098	−0.1098	−0.1098	−0.1098
AR(1)	$\widetilde{\sigma}^2$	1.0001	1.0118	1.0052	1.0035	1.0023	1.0016	1.0010	1.0006	1.0003	1.0000
MA(1)	$\widetilde{\omega}$	−0.0998	−0.0998	−0.0998	−0.0998	−0.0998	−0.0998	−0.0998	−0.0998	−0.0998	−0.0998
MA(1)	$\widetilde{\sigma}^2$	1	0.9998	0.9997	0.9997	0.9997	0.9996	0.9996	0.9996	0.9996	0.9996
WN	$\widetilde{\sigma}^2$	1.010	0.910	0.877	0.860	0.850	0.843	0.839	0.835	0.832	0.830

Table 2
Pseudo-true values for models fitted to DGP D2.

Minima											
Models		Leads									
		1	2	3	4	5	6	7	8	9	10
AR(1)	$\widetilde{\phi}$	−0.4870	−0.5010	−0.6347	−0.6068	−0.7066	−0.6707	−0.7525	−0.7146	−0.7824	−0.7465
AR(1)	$\widetilde{\sigma}^2$	1.2498	1.1069	0.7716	0.6979	0.5916	0.5458	0.4943	0.4630	0.4324	0.4099
MA(1)	$\widetilde{\omega}$	−0.8004	−0.8004	−0.8004	−0.8004	−0.8004	−0.8004	−0.8004	−0.8004	−0.8004	−0.8004
MA(1)	$\widetilde{\sigma}^2$	1.0000	1.0002	1.0003	1.0004	1.0006	1.0007	1.0008	1.0009	1.0010	1.0011
WN	$\widetilde{\sigma}^2$	1.6400	0.8400	0.5733	0.4400	0.3600	0.3067	0.2686	0.2400	0.2178	0.200

Table 3
Pseudo-true values for models fitted to DGP D3.

Minima											
Models		Leads									
		1	2	3	4	5	6	7	8	9	10
AR(1)	$\widetilde{\phi}$	0.5788	0.4731	0.4391	0.4271	0.4232	0.4212	0.4212	0.4212	0.4212	0.4212
AR(1)	$\widetilde{\sigma}^2$	1.2242	1.5562	1.6185	1.6239	1.6124	1.6011	1.5871	1.5766	1.5686	1.5623
MA(1)	$\widetilde{\omega}$	0.7804	0.8164	0.8224	0.8244	0.8244	0.8244	0.8244	0.8244	0.8244	0.8244
MA(1)	$\widetilde{\sigma}^2$	1.0253	1.0959	1.1856	1.2328	1.2612	1.2791	1.2914	1.3004	1.3072	1.3125
WN	$\widetilde{\sigma}^2$	1.8438	2.9125	3.4113	3.6820	3.8479	3.9590	4.0385	4.0981	4.1445	4.1816

Table 4
Pseudo-true values for models fitted to DGP D4.

Minima											
Models		Leads									
		1	2	3	4	5	6	7	8	9	10
AR(1)	$\widetilde{\phi}$	0.4970	0.1178	−0.7385	−0.6068	−0.8663	−0.1896	0.0938	−0.1018	−0.7745	−0.7465
AR(1)	$\widetilde{\sigma}^2$	2.9078	5.1052	8.7612	6.7663	4.1471	2.3597	1.6041	2.6589	5.5698	5.0134
MA(1)	$\widetilde{\omega}$	0.7964	−0.7385	−0.7565	−0.6926	−0.5469	−0.2934	0.7305	−0.6966	−0.7166	−0.6727
MA(1)	$\widetilde{\sigma}^2$	1.9458	9.0817	9.6785	8.2420	5.4644	2.9958	0.6728	9.8307	10.5664	9.0148
WN	$\widetilde{\sigma}^2$	3.8594	5.7759	5.4789	3.9234	2.4333	1.7826	1.9040	2.2479	2.3479	2.1288

Table 5
Pseudo-true values for models fitted to DGP D5.

Minima											
Models		Leads									
		1	2	3	4	5	6	7	8	9	10
AR(1)	$\widetilde{\phi}$	−0.6567	−0.9980	−0.8044	−0.9980	−0.8583	−0.9980	−0.8862	−0.9980	−0.9042	−0.9980
AR(1)	$\widetilde{\sigma}^2$	1.5540	1.4719	0.7471	0.7761	0.5472	0.5735	0.4538	0.4758	0.3991	0.4177
MA(1)	$\widetilde{\omega}$	−0.8543	−0.8144	−0.7924	−0.7804	−0.7725	−0.7665	−0.7625	−0.7585	−0.7545	−0.7525
MA(1)	$\widetilde{\sigma}^2$	1.1997	0.7769	0.6609	0.6185	0.6011	0.5934	0.5916	0.5897	0.5870	0.5886
WN	$\widetilde{\sigma}^2$	2.7263	1.2961	0.8747	0.6704	0.5485	0.4671	0.4086	0.3646	0.3301	0.3024

Table 6
Pseudo-true values for models fitted to DGP D6.

Minima											
Models		Leads									
		1	2	3	4	5	6	7	8	9	10
AR(1)	$\widetilde{\phi}$	−0.2495	−0.2495	−0.2435	−0.2395	−0.2375	−0.2375	−0.2355	−0.2355	−0.2355	−0.2355
AR(1)	$\widetilde{\sigma}^2$	1.0003	0.9999	0.9955	0.9961	0.9975	1.0006	1.0003	1.0021	1.0037	1.0049
MA(1)	$\widetilde{\omega}$	−0.2335	−0.2056	−0.1956	−0.1916	−0.1896	−0.1876	−0.1856	−0.1836	−0.1836	−0.1836
MA(1)	$\widetilde{\sigma}^2$	1.0044	0.9660	0.9655	0.9693	0.9732	0.9752	0.9756	0.9750	0.9774	0.9795
WN	$\widetilde{\sigma}^2$	1.0670	0.8536	0.7907	0.7602	0.7420	0.7299	0.7212	0.7146	0.7094	0.7053

Firstly, DGP D1 (Table 1) shows the MA(1) parameter equal to truth (up to numerical error), as this model is correctly specified; however, the misspecified ARIMA(1, 1, 0) model exhibits an h-step pseudo-true value for ϕ that varies slightly for small values of h and then stabilizes as h increases. The first two panels of Fig. 1 confirm this behavior. More or less the same behavior is evident for DGP D2 in Table 2, with only the true parameter value having been changed. The fact that the innovation variance for the WN fit decreases as h increases should cause no confusion, in light of the comments made previously about the proper interpretation of this parameter.

For DGP D3 we see that the fitted parameters also seem to stabilize for increasing values of h (Table 3), and the objective functions for this case (bottom row of Fig. 1) look qualitatively quite similar to those for D2 and D1. DGP D4 is much more interesting, with the objective functions overlapping one another for different values of h (top row of Fig. 2). As a result, the pseudo-true values for the AR and MA parameters change quite a bit, and seem not to stabilize in h (Table 4). This is no surprise, given the strong spectral peak in the data process that is captured badly by the grossly misspecified models. As h increases, a different snap-shot of this cyclical process is obtained, and the h-step-ahead forecast error is optimized accordingly.

Finally, we have DGPs D5 and D6 (Tables 5 and 6), which exhibit distinct behavior in the objective functions from the other cases (middle and bottom rows of Fig. 2). Unfortunately, portions of these likelihoods (especially in the AR model case) are extremely flat, resulting in numerical imprecisions in the optima shown. The ARIMA(0,2,1) performs slightly better, since, in a sense, it is less badly misspecified, the true model being an ARIMA(0, 2, 2). Also, the increased SNR in D6 makes the trend in the STM more dominant, which presumably facilitates forecasting (compared to a noisier process), and this may be the reason that the optima are better behaved.

5. Empirical results

We first study a time series of chemical data from an in-sample forecasting perspective, in order to show the correspondence between GKL and LS. We then examine housing starts, and demonstrate the superior long-term forecasting performance of the 12-step GKL over the conventional MLE.

5.1. Chemical data

We consider chemical process concentration readings (Chem for short).[3] The sample has 197 observations. This series was studied by McElroy and Findley (2010), who argued that an ARIMA(0, 1, 1) model was the most appropriate of several contenders, according to multi-step-ahead forecasting criteria (based on parameter estimates obtained using MLEs). The same model was identified for the series by Box and Jenkins (1976), and was also studied by Tiao and Xu (1993). Fitting Chem using $GKL_\delta^{(h)}$ yields the MA(1) polynomials $1-0.698B$, $1-0.798B$, and $1-0.841B$ for $h = 1, 2, 3$ respectively.

We noted earlier (Section 2) that the GKL objective function given in Eq. (2) is an asymptotic form of the forecast mean squared error. In contrast, the LS method is based on empirical forecasts generated from a finite sample of data. As was discussed by McElroy and Findley (2010), $J([\theta], I)$ differs from the empirical forecast mean squared error $S([\theta])$ by $O_P(n^{-1/2})$, where n is the number of h-step-ahead forecasts. We also fitted the ARIMA(0, 1, 1) model using the LS method, to see whether there were any substantial discrepancies in the parameter estimates. We obtained the MA(1) polynomials $1 - 0.694B$, $1 - 0.773B$, and $1 - 0.809B$ for $h = 1, 2, 3$ respectively, with values of the objective function given by 0.101, 0.114, and 0.121. In contrast, the GKL objective function had minimum values of 0.102, 0.115, and 0.124, respectively; it was also substantially faster to compute.

If we generate forecasts of Chem using a moving window of sub-samples, and average the squared forecast errors, the resulting behavior should mimic that of S (and J) as the window size increases. In particular, let us consider the forecast h steps ahead, for $h = 1, 2, 3$, from a sample consisting of time indices $t = 1, 2, \ldots, 197 - n - h + s$, repeated for $s = 1, 2, \ldots, n$. Moreover, let us generate these forecasts from each of the three GKL and LS objective functions, for $h = 1, 2, 3$. Then, the within-sample forecast errors are calculated, squared, and averaged over s. The results can be summarized in a 3×3 table, where the row j corresponds to the $GKL^{(j)}$ or $LS^{(j)}$ parameter used, and column k corresponds to the forecast horizon. Note that the diagonal entries of the forecast error matrix correspond to forecasts generated from the composite forecasting procedure described in the last paragraph of Section 2.

[3] Available from http://www.stat.wisc.edu/~reinsel/bjr-data/index.html.

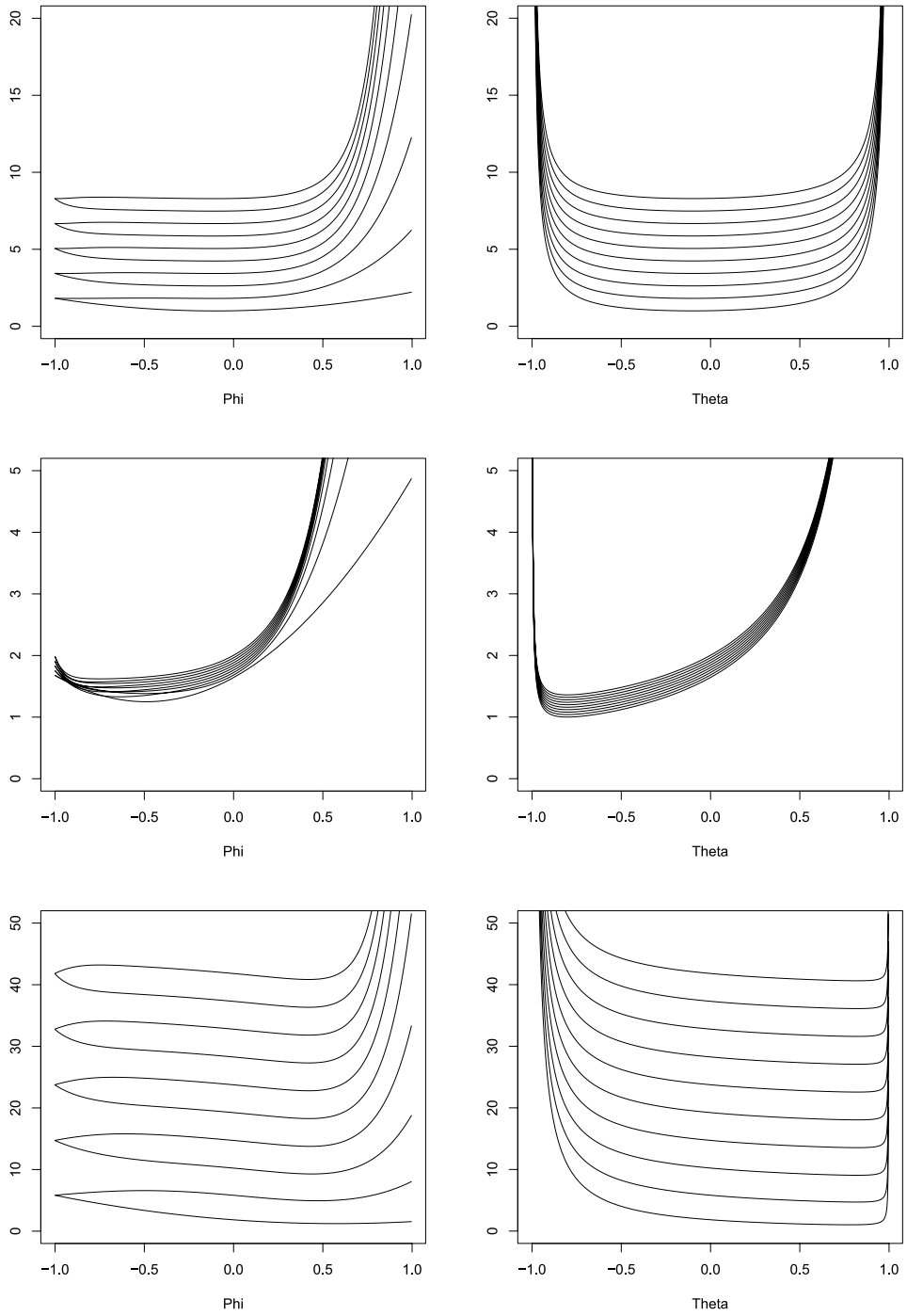

Fig. 1. The left panels display the function J for the AR model, while the right panels display the function J for the MA model. The upper panels correspond to DGP D1, the middle panels to DGP D2, and the lower panels to DGP D3. Overlaid objective functions correspond to h-step-ahead forecast MSEs, for $1 \leq h \leq 10$. Higher curves correspond to greater values of h.

Referring to this forecast error matrix via $F(n)$, we can expect the column minima to occur on the diagonals, as $n \to \infty$. That is, $\min_{\{j\}} F_{jk}(n) = F_{kk}(n)$ for each $k = 1, 2, 3$, for large values of n.

This is heuristic, because as we increase n we reduce the length of the filters used to generate forecasts; nevertheless, Table 7 displays the pattern of $F(n)$ for $n = 50, 75, 100, 125, 150$, and the expected property holds starting at $n = 125$ (for GKL, and to a limited extent for LS as well). The 2-step GKL does well for 3-step-ahead forecasting for smaller values of n, presumably due to the close values (-0.798 and -0.841) of their respective MA parameters.

5.2. Housing starts

Here we study the series of "Total new privately owned housing units started" in the U.S., from 1959 to 2004. We omit the period of the Great Recession and some antecedent years for illustrative purposes (no model fitted

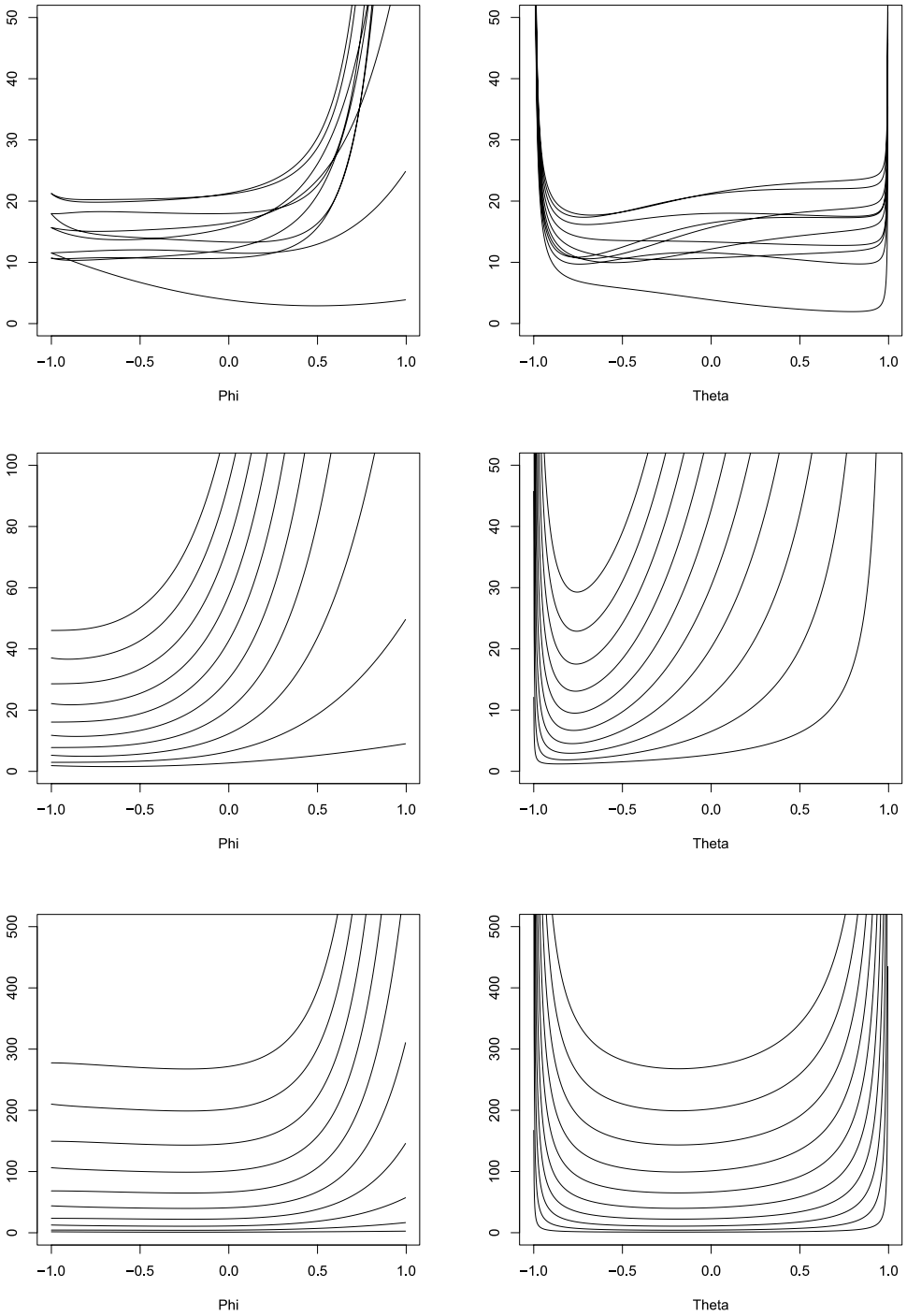

Fig. 2. The left panels display the function J for the AR model, while the right panels display the function J for the MA model. The upper panels correspond to DGP D4, the middle panels to DGP D5, and the lower panels to DGP D6. Overlaid objective functions correspond to h-step-ahead forecast MSEs, for $1 \leq h \leq 10$. Higher curves correspond to greater values of h.

to the pre-recession period has a reasonable forecasting performance during the slump). The series was adjusted for outliers and other regression effects by X-12-ARIMA, and the log-transformed data were fitted by an airline model using GKL and LS for $h = 1, 12$. In order to perform an out-of-sample forecasting exercise, we only fit to the first 40 years of data. The parameter estimates are given in Table 8, along with the values of the objective function. There were some differences between the LS and GKL estimates in this case. While the $h = 1$ GKL parameters were 0.279 (nonseasonal MA) and 0.827 (seasonal MA), the LS parameters were 0.284 and 0.936. The minimum values of the objective functions were 0.00814 (GKL) and 0.00793 (LS); however, whereas a single evaluation required 86 s for LS, the time was less than a second for GKL. For $h = 12$, the GKL parameters were 0.157 (nonseasonal MA) and 0.906 (seasonal MA), while the LS parameters were 0.176 and 0.999. The minimum values of the objective functions were 0.06147 (GKL) and 0.05864 (LS), and a single evaluation required 83 s for LS, versus less than

Table 7
Empirical mean square forecast error grids $F(n)$ by window size $n = 50, 75, 100, 125, 150$, for the Chem series, utilizing $GKL_\delta^{(j)}$ and $LS^{(j)}$ optima $j = 1, 2, 3$ (by row) and forecast horizons $k = 1, 2, 3$ (by column). Values in bold denote the lowest values in each column, for each n. On the left are results for using GKL, and on the right are results for using LS.

$F(n)$ from GKL				$F(n)$ from LS			
Fit lead	Forecast lead			Fit lead	Forecast lead		
$n = 50$	1	2	3	$n = 50$	1	2	3
1	**0.09832**	0.13084	0.14282	1	**0.09828**	0.13119	0.14335
2	0.09974	0.12207	0.12997	2	0.09928	0.12417	0.13300
3	0.10094	**0.11896**	**0.12540**	3	0.09997	**0.12123**	**0.12875**
$n = 75$	1	2	3	$n = 75$	1	2	3
1	**0.08627**	0.11267	0.12505	1	**0.08615**	0.11281	0.12530
2	0.09018	0.10961	0.11919	2	0.08901	0.11026	0.12057
3	0.09266	**0.10886**	**0.11711**	3	0.09073	**0.10938**	**0.11863**
$n = 100$	1	2	3	$n = 100$	1	2	3
1	**0.08673**	0.10831	0.12017	1	**0.08668**	0.10847	0.12044
2	0.08882	0.10472	0.11365	2	0.08814	0.10553	0.11523
3	0.09035	**0.10357**	**0.11110**	3	0.08915	**0.10441**	**0.11300**
$n = 125$	1	2	3	$n = 125$	1	2	3
1	**0.08527**	0.10725	0.12125	1	**0.08516**	0.10733	0.12145
2	0.08934	**0.10621**	0.11717	2	0.08802	**0.10624**	0.11798
3	0.09232	0.10675	**0.11632**	3	0.08998	0.10626	**0.11689**
$n = 150$	1	2	3	$n = 150$	1	2	3
1	**0.09537**	0.11315	0.12473	1	**0.09534**	0.11326	0.12494
2	0.09764	**0.11150**	0.12081	2	0.09675	0.11163	0.12151
3	0.09996	0.11198	**0.12026**	3	0.09811	**0.11153**	**0.12059**

Table 8
Results for the fitting of the housing starts data via using a one-step and a twelve-step GKL criterion, using the first 20, 30, or 40 years of data. Values for the GKL minimum are given, along with parameter estimates for the fitted airline model, and the average of squared forecast errors based on forecasting until 2004.

	GKL one-step			GKL 12-step		
	20 years	30 years	40 years	20 years	30 years	40 years
GKL minimum	0.00893	0.00910	0.008137	0.06881	0.07427	0.061470
Nonseasonal MA	0.273	0.252	0.279	0.032	0.116	0.157
Seasonal MA	0.661	0.814	0.827	0.948	0.909	0.906
Forecast error	0.44170	0.01279	0.00316	0.07550	0.01186	0.00266

a second for GKL. Although these discrepancies are of some interest, in the rest of our discussion we focus on the GKL estimates. The key point is that the seasonal MA parameter is larger for $h = 12$, implying a more stable form of seasonality. When generating forecasts 5 years ahead, the empirical mean square forecast error over this period for $h = 12$ GKL is 84% of the result from the $h = 1$ GKL. See Table 8 and Fig. 3. We repeat the analysis with different amounts of data withheld in order to see a longer span of forecasts. Parameter estimates from using the first 20 or 30 years of data are given in Table 8, with a dramatic reduction in forecast error in the former case (see Fig. 3); in this case, the discrepancies between seasonal MA parameter estimates are the greatest. It seems that the 1-step GKL forecasts use an older trajectory of the data, perhaps due to an increased dependence upon the past, merited by a more chaotic seasonal pattern; the 12-step GKL forecasts seem to utilize a more nascent trajectory, consistent with presuming a more stable seasonal pattern. Generalizing these observations, it seems that a seasonal MA parameter which is closer to unity generates more "conservative" forecasts that are based on the very recent past, whereas a smaller seasonal MA parameter put a greater weight on the distant past observations. In the extreme case that a seasonal MA parameter is equal to one, the airline model reduces to an ARIMA(0, 1, 1) with a fixed seasonal regressor, which is no longer $I(2)$, betokening a much less ambitious forecast pattern. For long-term performance, the housing starts data seem to prefer this conservative approach, and the performance gains can be substantial.

6. Conclusion

Classical model-based approaches typically emphasize a short-term one-step-ahead forecasting perspective for estimating unknown model parameters. This procedure could be justified by assuming that the "true" model has been identified or that it is known to the analyst *a priori*. In contrast, we have emphasized the importance of inferences based on multi-step-ahead forecasting performances in the practically more relevant context of misspecified models. For this purpose, we have proposed a generalization of the well-known Kullback–Leibler discrepancy, and have derived an asymptotic distribution

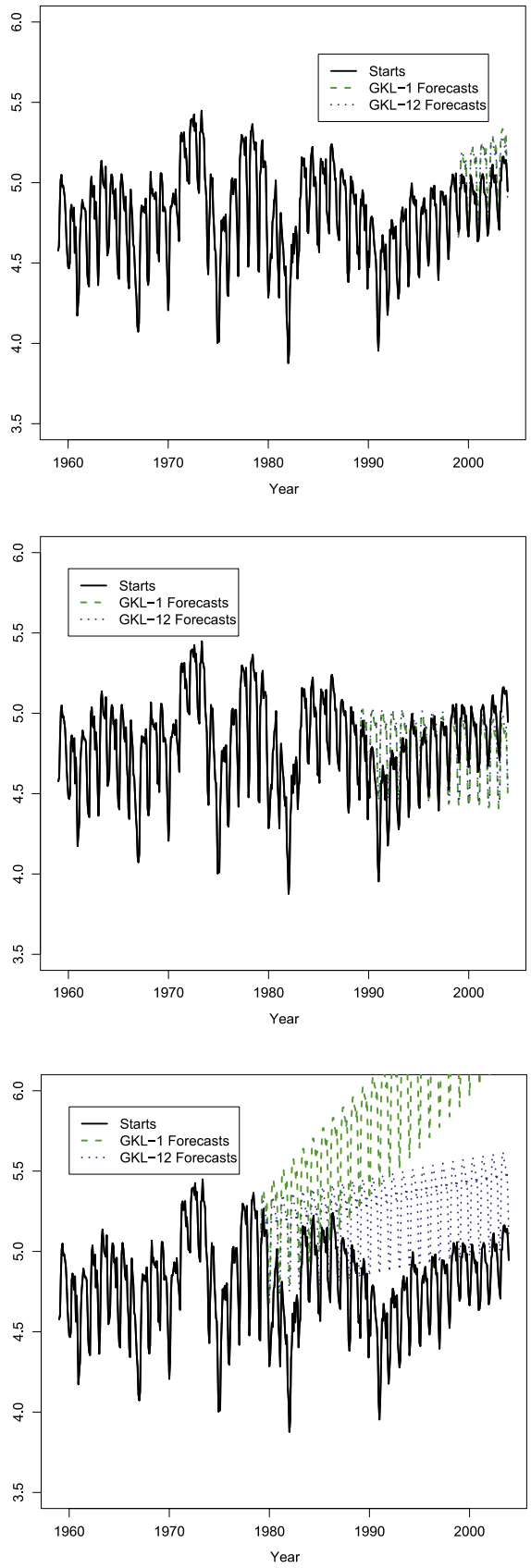

Fig. 3. Housing starts, along with forecasts from 40 years out (top), 30 years out (middle), and 20 years out (bottom). Forecasts are generated from an airline model fitted using either the 1-step GKL or the 12-step GKL.

theory for estimates that converge to "pseudo-true" values, expanding the consistency results of Findley et al. (2004) to central limit theorems. In contrast to earlier approaches (e.g., Tiao & Xu, 1993; or Haywood & Tunnicliffe-Wilson, 1997), our development is fairly general, covering all difference-stationary processes with a causal Wold decomposition.

We have illustrated the appeal of our approach by deriving closed-form solutions for a selection of simple processes, such as the popular ARIMA(1, 1, 0) model used in econometric forecasting. We then compared the performances of classical (one-step-ahead) and generalized (h-step-ahead) estimates in a controlled experimental design based on a selection of both simulated and practical time series. Our empirical findings confirm the asymptotic theory, i.e., that the smallest forecast errors for a given forecast lead arise from the corresponding criterion function for that lead (cf. the discussion of the Chem series in Section 5.1). We find evidence in the housing starts that unit root over-specification (i.e., specifying a differencing operator of too high an order) can be mitigated, to some extent, by longer-term forecasting criteria.

In this paper we have focused on univariate multi-step-ahead forecasting over one forecast lead at a time. In terms of future work, we are interested in addressing more complex forecasting problems, such as simultaneous optimization over many leads or real-time signal extraction (computation of concurrent trend or seasonal-adjustment filters) in univariate and multivariate frameworks. The real-time signal extraction problem can be parsed as an attempt to minimize revisions, which in turn depends on the forecast performance at a variety of leads. There is already substantial interest in the minimization of revisions at statistical agencies (such as the U.S. Census Bureau and the Australian Bureau of Statistics), and designing a model-fitting criterion to minimize the revision variance seems appealing. Such a procedure would have ramifications for official seasonal adjustment procedures such as X-12-ARIMA[4] and TRAMO-SEATS.[5] We expect the frequency-domain approach underlying GKL to offer some promising perspectives on these future topics.

Acknowledgments

The authors thank a referee and the associate editor for many helpful comments and references that improved our results.

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, not necessarily those of the U.S. Census Bureau.

[4] See Findley et al. (1998) for a discussion of the methodology. The signal extraction method uses nonparametric symmetric filters applied to the data, which are then forecast and backcast extended (implicitly) in order to obtain signal estimates at the sample boundary.

[5] The seasonal adjustment software of the Bank of Spain; see Maravall and Caporello (2004) for a discussion. Model-based signal extraction filters are obtained from component models deduced via the method of canonical decomposition (Burman, 1980; Hillmer & Tiao, 1982).

Appendix

A.1. The least squares criterion

Here, we develop the mathematics for the minimization of the h-step-ahead forecast MSE with respect to given parameters, based upon finite-sample predictors. This is referred to here as the least squares (LS) method, in contrast to the GKL approach of Section 2, which relies on semi-infinite predictors. The LS method is popular when the fitted model is an AR(p), because then parameter estimates can be obtained simply via regression (ordinary least squares). However, the forecast formulas are a bit more complicated for the general ARIMA(p, d, q), and we develop our treatment for difference stationary processes.

Let $X_{1:n}$ denote our data sample, consisting of observations up to time n, and the matrix version of $\delta(B)$ is Δ_n, an $n \times n$ dimensional matrix given by $\delta(L_n)$, where L_n is a lag matrix. That is, the jkth entry of L_n is one if $j = k + 1$ and zero otherwise, so that

$$\Delta_n = \delta(L_n) = \sum_{k=0}^{d} \delta_k L_n^k.$$

This matrix differences all but the first d entries of $X_{1:n}$ to be stationary $\{W_t\}$, and is always invertible. Consider the covariance matrices computed from f_θ with dimensions indicated by subscripts, denoted $\Sigma_{1:m,1:p}$ for an $m \times p$ matrix corresponding to the covariance of $W_{1:m}$ with $W_{1:p}$. Then the estimate of present and future values given $X_{1:n}$ is

$$\widehat{X}_{1:n+h} = \Delta_{n+h}^{-1} \begin{bmatrix} 1_d & 0 \\ 0 & \Sigma_{1:n+h-d,1:n-d}\, \Sigma_{1:n-d,1:n-d}^{-1} \end{bmatrix} \times \Delta_n\, X_{1:n}.$$

Here 1_d denotes a d-dimensional identity matrix. The final entry produces the forecast h steps ahead:

$$\varepsilon_n = \widehat{X}_{n+h|1:n} = [0, \ldots, 0, 1]\, \widehat{X}_{1:n+h}.$$

These formulas are derived in Eq. (11) of McElroy (2008a),[6] and require the assumption that the first d values $X_{1:d}$ are uncorrelated with $\{W_t\}$. Now the forecast error ε_n is a function of the model parameters, and does not depend on the innovation variance. Hence, we can write $\varepsilon_n([\theta])$. The average sum of squares of such forecast errors is then

$$S([\theta]) = \frac{1}{N - h - d} \sum_{n=d+1}^{N-h} \varepsilon_n^2([\theta]),$$

which, by definition, is the h-step-ahead LS criterion. Here, N is the total number of observations available.

This LS criterion is more expensive to compute than the GKL criterion function $J([\theta], I)$, requiring a loop over roughly N calculations, each of which requires matrix inversion. Although the formula can be greatly simplified and the algorithm quickened accordingly for certain models, such as the AR(p), in general repeated calculations must be made. (Newton & Pagano, 1983, give helpful results for stationary processes, and there are tricks for the fast calculation of Δ_{n+h}^{-1}.) In contrast, GKL involves no matrix inversion, because of the use of semi-infinite predictors. Now the finite sample predictors tend to be approximated well by semi-infinite predictors, even for very small sample sizes, and in practice the performance of $J([\theta], I)$ is quite close to that of $S([\theta])$, as Section 5 demonstrates. Also, $J([\theta], I)$ is much easier to analyze from a theoretical and numerical perspective than $S([\theta])$.

[6] Note an error in the final matrix on the right in the expression for D; the block matrices of the upper block row should be interchanged.

A.2. Proofs

Proof of Proposition 2. Plugging $[\theta] = [\theta_g]$ and $\sigma^2 = \sigma_g^2$ from Eq. (3) into the gradient formulas in the separable case of Proposition 1 shows that θ_g is a critical point of GKL, since $\nabla_{[\theta]} J([\theta], g)$ evaluated at $[\theta] = [\theta_g]$ equals zero. Plugging into the Hessian formula yields (after simplification):

$$\nabla_{[\theta]} \nabla'_{[\theta]} G(\theta)|_{\theta=\theta_g} = \frac{\nabla_{[\theta]} \nabla'_{[\theta]} J([\theta], g)|_{\theta=\theta_g}}{J([\theta_g], g)} + \left. \frac{\langle \nabla_{[\theta]} f_{[\theta]}^{(h)} \rangle \langle \nabla'_{[\theta]} f_{[\theta]}^{(h)} \rangle J([\theta], g)}{\langle f_{[\theta]}^{(h)} \rangle^2} \right|_{\theta=\theta_g}$$

$$\frac{\partial}{\partial \sigma^2} \nabla_{[\theta]} G(\theta)|_{\theta=\theta_g} = \frac{\nabla_{[\theta]} \int f_{[\theta]}^{(h)}|_{\theta=\theta_g}}{J([\theta_g], g)}$$

$$\frac{\partial^2}{\partial^2 \sigma^2} G(\theta)|_{\theta=\theta_g} = \sigma_g^{-4}.$$

This fills out a matrix $H(\theta_g)$, partitioned as

$$\begin{bmatrix} \sigma_g^4\, c\, c' + B & c \\ c' & \sigma_g^{-4} \end{bmatrix},$$

for $c = \nabla_{[\theta]} \int f_{[\theta]}^{(h)}|_{\theta=\theta_g} / J([\theta_g], g)$ and B equal to the Hessian of $J([\theta], g)$ evaluated at $[\theta_g]$, divided by $J([\theta_g], g)$. Then consider any vector a partitioned into the first r components $[a]$ and the final component b:

$$a' H(\theta_g) a = \left(b\sigma_g^{-2} + \sigma_g^2 [a]' c\right)^2 + [a]' B [a],$$

by completing the square. Now, since the Hessian of J is positive definite at $[\theta_g]$ by assumption, and $J([\theta_g], g) > 0$, we conclude that $H(\theta_g)$ is positive definite. For the converse, suppose that θ_g minimizes $G(\theta)$. Then, by the gradient expression in Proposition 1, Eq. (3) must hold, and in turn we must have $\nabla_{[\theta]} J([\theta], g)$ equal to zero at $[\theta] = [\theta_g]$.

Next, suppose that the model is non-separable. Recall that ∇_θ is the same thing as $\nabla_{[\theta]}$. The expression for the gradient of $G(\theta)$ in Proposition 1 shows that when σ_g^2 satisfies Eq. (3) and $[\theta_g]$ is a critical point of $J([\theta], g)$, then θ_g is a critical point of $G(\theta)$. Plugging into the Hessian expression yields

$$\left(\frac{\nabla_\theta \sigma^2}{\sigma^2} + \frac{\nabla_\theta f_{[\theta]}^{(h)}}{\langle f_{[\theta]}^{(h)} \rangle} \right) \left(\frac{\nabla'_\theta \sigma^2}{\sigma^2} + \frac{\nabla'_\theta f_{[\theta]}^{(h)}}{\langle f_{[\theta]}^{(h)} \rangle} \right) + \left. \frac{\nabla_\theta \nabla'_\theta J([\theta], g)}{J([\theta_g], g)} \right|_{\theta=\theta_g},$$

which is positive definite. This completes the proof. □

Proof of Theorem 1. Note that θ_g is a zero of $G(\theta)$ with the function g, so we can do a Taylor series expansion of the gradient at θ_I and $\theta_{\widetilde{f}}$. This yields the asymptotic expression (cf. Taniguchi & Kakizawa, 2000)

$$\sqrt{n}\left(\theta_I - \theta_{\widetilde{f}}\right) = o_P(1) - H^{-1}(\theta_{\widetilde{f}})\ \sqrt{n}\left\langle \int r_{\theta_{\widetilde{f}}}\left(I - \widetilde{f}\right)\right\rangle,$$

where $r_\theta = \nabla_\theta f_{[\theta]}^{(h)} f_\theta^{-1} \langle f_{[\theta]}^{(h)} \rangle^{-1}$. Our assumptions allow us to apply Lemma 3.1.1 of Taniguchi and Kakizawa (2000) to the right hand expression above, and the stated central limit theorem is obtained. □

A.3. Implementation for ARIMA models

In order to compute parameter estimates and pseudo-true values for a fitted ARIMA model, it is necessary to set up an optimization algorithm carefully. In the case that the DGP is a known ARIMA process and one seeks to obtain pseudo-true values, the integrand of J in Eq. (2) can always be written as the spectral density of a composite ARMA process, with its AR and MA factors being determined by both the DGP and the fitted model. An exact formula for the integral is given as follows.

Say that the AR polynomial of degree p has the form $\Pi_j(1-\zeta_j^{-1}z)^{r_j}$ for roots ζ_j of multiplicity r_j. Similarly, let the MA polynomial of degree q have the form $\Pi_\ell(1-\xi_\ell^{-1}z)^{s_\ell}$ for roots ξ_ℓ of multiplicity s_ℓ. Then the variance of the ARMA spectrum is

$$\frac{1}{2\pi i}\int_C \frac{\Pi_\ell(1-\xi_\ell^{-1}z)^{s_\ell}(z-\xi_\ell^{-1})^{s_\ell}}{\Pi_j(1-\zeta_j^{-1}z)^{r_j}(z-\zeta_j^{-1})^{r_j}}\, z^{p-q-1}\, dz,$$

where C denotes the unit circle of the complex plane. The poles at ζ_j have multiplicity r_j, and the pole at zero has multiplicity $q+1-p$ when this is positive. When $q+1-p > 0$, the variance simplifies to

$$\sum_j \frac{\partial^{r_j-1}}{\partial z^{r_j-1}} \times \left[\frac{\Pi_\ell(1-\xi_\ell^{-1}z)^{s_\ell}(z-\xi_\ell^{-1})^{s_\ell} z^{p-q-1}(-\zeta_j)^{r_j}}{\Pi_{k\neq j}(1-\zeta_k^{-1}z)^{r_k}(z-\zeta_k^{-1})^{r_k}(z-\zeta_j^{-1})}\right]\Bigg|_{z=\zeta_j} + \frac{\partial^{q-p}}{\partial z^{q-p}}\left[\frac{\Pi_\ell(1-\xi_\ell^{-1}z)^{s_\ell}(z-\xi_\ell^{-1})^{s_\ell}}{\Pi_j(1-\zeta_j^{-1}z)^{r_j}(z-\zeta_j^{-1})^{r_j}}\right]\Bigg|_{z=0}.$$

In practice, this formula does not provide the fastest method of computation except in special cases. We now describe a method that works for both parameter estimation and the calculation of pseudo-true values. Let $\Psi(B) = \Omega(B)/\Phi(B)$, where $\Omega(z) = 1 + \omega_1 z + \cdots + \omega_q z^q$ and $\Phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ with $r = p + q$. First, the data should be differenced using $\delta(B)$. The main computational issue is the calculation of the autocovariances in Eq. (6); this is detailed in the following algorithm. The user fixes a given forecast lead $h \geq 1$.

1. Given: current value of θ.
2. Compute the first h coefficients of the moving average representation of $\Omega(B)/(\Phi(B)\delta(B))$ (e.g., in R use the function *ARMAtoMA*); the resulting polynomial is $[\Omega/(\Phi\delta)]_0^{h-1}(B)$.
3. Compute the autocovariances of $f_{[\theta]}^{(h)}(\lambda) = |[\Omega/(\Phi\delta)]_0^{h-1}(e^{-i\lambda})|^2$ and $f_{[\theta]}^{(h)}(\lambda)/f_{[\theta]}(\lambda)$, which both have the form of ARMA spectral densities (e.g., in R use *ARMAacf*).
4. Form the Toeplitz matrix and plug into Eq. (6).
5. Search for the next value of θ using BFGS or any other numerical recipe.

Explicit formulas for the quantity in step 2 are given by McElroy and Findley (2010). Our implementation is written in R, and utilizes the *ARMAtoMA* routine. Although one could find the autocovariances of $f_{[\theta]}^{(h)}(\lambda)/f_{[\theta]}(\lambda)$ directly through the *ARMAacf* routine, one still needs the integral of $f_{[\theta]}^{(h)}(\lambda)$, which is the sum of the square of the coefficients of its moving average representation. Moreover, finding the MA representation first happens to be more numerically stable. Also note that in step 3 the R routine *ARMAacf* has the defect of computing autocorrelations rather than autocovariances. We have adapted the routine to our own *ARMAacvf*, which rectifies this deficiency.

When mapping ARMA parameter values into the objective function, it is important to have an invertible representation. In particular, the roots of both the AR and MA polynomials must lie outside the unit circle. To achieve this, we utilize our routine *flipIt*, which computes the roots, flips those lying on or inside the unit circle (by taking the reciprocal of the magnitude), compensates the innovation variance (scale factor) appropriately, and passes the new polynomials back to the objective function. Step 4 is implemented using the *toeplitz* routine in R.

Step 5 requires a choice of optimizer. The R routine *optim* is reliable and versatile, as one can specify several different techniques. The implicit bound on the polynomial roots is handled automatically through the *flipIt* routine, so only the innovation variance needs to be constrained, and this is handled most naturally through optimizing over $\log\sigma^2$ instead, which can take any real number as its value. Then, a conjugate gradient method such as BFGS (Golub & Van Loan, 1996) can be used to compute the gradient and Hessian via a numerical approximation; some routines allow for the use of an exact gradient and Hessian. While the formulas in Section 4 allow one to calculate these exact quantities in principle, the programming effort is considerable and it is not clear whether there is any advantage to be gained, since the resulting formulas depend on multiple calls to *ARMAacvf* and the like.

References

Box, G. E. P., & Jenkins, G. M. (1976). *Time series analysis: forecasting and control*. Holden-Day.

Burman, P. (1980). Seasonal adjustment by signal extraction. *Journal of the Royal Statistical Society, Series A*, *143*, 321–337.

Chiu, S. (1988). Weighted least squares estimators on the frequency domain for the parameters of a time series. *The Annals of Statistics*, *16*, 1315–1326.

Cox, D. (1961). Prediction of exponentially weighted moving averages and related methods. *Journal of the Royal Statistical Society, Series B*, *23*, 414–422.

Dagum, E. (1980). *The X-11-ARIMA seasonal adjustment method*. Ottawa, Statistics Canada.

Dahlhaus, R., & Wefelmeyer, W. (1996). Asymptotically optimal estimation in misspecified time series models. *The Annals of Statistics*, *16*, 952–974.

Diebold, F., & Mariano, R. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, *13*, 253–263.

Findley, D. F., Monsell, B. C., Bell, W. R., Otto, M. C., & Chen, B. C. (1998). New capabilities and methods of the X-12-ARIMA seasonal adjustment program (with discussion). *Journal of Business and Economic Statistics*, *16*, 127–177.
Findley, D., Pötscher, B., & Wei, C.-Z. (2004). Modeling of time series arrays by multistep prediction or likelihood methods. *Journal of Econometrics*, *118*, 151–187.
Gersch, W., & Kitagawa, G. (1983). The prediction of time series with trends and seasonalities. *Journal of Business and Economic Statistics*, *1*, 253–264.
Golub, G., & Van Loan, C. (1996). *Matrix computations*. Baltimore: Johns Hopkins University Press.
Hannan, E., & Deistler, M. (1988). *The statistical theory of linear systems*. New York: Wiley.
Harvey, A. (1989). *Forecasting, structural time series models and the Kalman filter*. Cambridge: Cambridge University Press.
Haywood, J., & Tunnicliffe-Wilson, G. (1997). Fitting time series models by minimizing multistep-ahead errors: a frequency domain approach. *Journal of the Royal Statistical Society, Series B*, *59*, 237–254.
Hillmer, S., & Tiao, G. (1982). An ARIMA-model-based approach to seasonal adjustment. *Journal of the American Statistical Association*, *77*(377), 63–70.
Hosoya, Y., & Taniguchi, M. (1982). A central limit theorem for stationary processes and the parameter estimation of linear processes. *The Annals of Statistics*, *10*, 132–153.
Maravall, A., & Caporello, G. (2004). *Program TSW: revised reference manual*. Working Paper 2004, Research Department, Bank of Spain. http://www.bde.es.
Marcellino, M., Stock, J., & Watson, M. (2006). A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics*, *135*, 499–526.
McElroy, T. (2008a). Matrix formulas for nonstationary ARIMA signal extraction. *Econometric Theory*, *24*, 1–22.
McElroy, T. (2008b). Exact formulas for the Hodrick–Prescott filter. *Econometrics Journal*, *11*, 1–9.
McElroy, T., & Findley, D. (2010). Discerning between models through multi-step ahead forecasting errors. *Journal of Statistical Planning and Inference*, *140*, 3655–3675.
McElroy, T., & Holan, S. (2009). A local spectral approach for assessing time series model misspecification. *Journal of Multivariate Analysis*, *100*, 604–621.
Newton, H., & Pagano, M. (1983). The finite memory prediction of covariance stationary time series. *SIAM Journal on Scientific and Statistical Computing*, *4*, 330–339.
Proietti, T. (2011). Direct and iterated multistep AR methods for difference stationary processes. *International Journal of Forecasting*, *27*, 266–280.
Self, S., & Liang, K. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, *82*, 605–610.
Stock, J., & Watson, M. (1999). Forecasting inflation. *Journal of Monetary Economics*, *44*, 293–335.
Taniguchi, M., & Kakizawa, Y. (2000). *Asymptotic theory of statistical inference for time series*. New York: Springer-Verlag.
Tiao, G., & Xu, D. (1993). Robustness of maximum likelihood estimates for multi-step predictions: the exponential smoothing case. *Biometrika*, *80*, 623–641.
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 No. 96, ETH-Zurich.
Wildi, M. (2008). *Real-time signal-extraction: beyond maximum likelihood principles*. Berlin: Springer, http://www.idp.zhaw.ch/de/engineering/idp/forschung/finance-risk-management-and-econometrics/signal-extraction-and-forecasting/signal-extraction.html.

Tucker McElroy received a B.A. from Columbia College in 1996, cum laude in mathematics. From University of California, San Diego he received a Ph.D. in pure mathematics in 2001, and he also completed a postdoctoral fellowship at the same institution in 2003. He currently works as a principal researcher in time series analysis at the U.S. Census Bureau. His research interests include time series modeling and signal extraction, and he has published on the topics of goodness-of-fit diagnostics, seasonal adjustment, and frequency domain time series analysis.

Marc Wildi studied mathematics at ETH, in Zurich, and economics at the University of St Gallen. He is professor of econometrics at the University of Applied Sciences, Zurich. His research interests are time series analysis, signal extraction, robust statistics, time series analysis, and filtering.