Multi-Step Ahead Estimation of Time Series Models

Tucker McElroy[1] and Marc Wildi[2]
U.S. Census Bureau and Institute of Data Analysis and Process Design

Abstract

We study the fitting of time series models via 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 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 provably 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] Center for Statistical Research and Methodology, U.S. Census Bureau, 4600 Silver Hill Road, Washington, D.C. 20233-9100

[2] Institute of Data Analysis and Process Design

Keywords. ARIMA, Forecasting, Frequency Domain, Nonstationary, Signal Extraction.

1 Introduction

It is well-known that fitting models via the minimization of one-step ahead forecasting error 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 (1994), Taniguchi and Kakizawa (2000). But in reality models are never correctly specified, 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 one-step

ahead forecast error for a given process utilizing a certain misspecified model. Given that for some applications more interest focuses on forecasting performance at high leads, it is natural to consider the following questions: can we fit time series models such that multi-step ahead forecasting error is minimized? Is there an objective function analogous to KL, 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 directly derived from a multi-step ahead forecasting error criterion. This GKL can be used with very little programming effort to fit linear time series models; 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., the minimizers of the GKL discrepancy between model and truth) under standard conditions, and are also asymptotically normal (consistency results under quite mild conditions are established in Findley, Pötscher, and Wei (2004)). When the model is correctly specified, these estimates are inefficient, i.e., they perform worse than the 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 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 model specification and estimation should be oriented around a particular objective function associated with the application. McElroy and Findley (2010) addresses 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 truth and 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 equivalency of Gaussian likelihood and Whittle likelihood in the case that the model is correctly specified; when mis-specified, the proper reference is Dahlhaus and Wefelmeyer (1994).

Let us briefly discuss the econometric motivations for considering the multi-step ahead perspective. Since time series models in reality are always misspecified, the crucial thing is to find a model that performs well according to the prescribed task of interest to the practitioner; using GKL as an objective function means the practitioner is interested in a model that forecasts well at a particular lead. In econometric business cycle analysis there is little interest in mere one-step ahead performance of misspecified models, since the period of a typical cycle is 8 to 40 observations for quarterly

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

data. A model or collection of models that can forecast well at lead 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 implicitly or explicitly on long-horizon forecasts generated from the same misspecified model; see Dagum (1980), Findley et al. (1998), Wildi (2004), and McElroy (2008a). Real-time – or concurrent – signal extraction is discussed in Wildi (2004, 2008), where the nefarious impact of model mis-specification on long-term forecasting performance and signal extraction is highlighted through numerous empirical studies. This situation has ramifications for official seasonal adjustment procedures such as X-12-ARIMA[2] and TRAMO-SEATS[3]. Beyond these obvious applications, any data analysis that is contingent on long-run forecasts – such as 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 on this topic that deserves 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 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 multi-step ahead forecast mean squared error is minimized. Another treatment of the topic is 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. This 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 Haywood and Tunnicliffe-Wilson (1997), which provides an explicit formula for the objective function written in the frequency domain. A limitation of theirs is that the variables of the objective function do not in general correspond to ARMA parameters, as the paper essentially fits an unusual parametrization of moving average models.

There is also substantial interest among econometricians in multi-step ahead forecasting arising from autoregressive and difference autoregressive models. Marcellino, Stock, and Watson (2006) expounds a common approach involving ordinary least squares estimation of these models so as

[2]The seasonal adjustment software of the U.S. Census Bureau; see Findley et al. (1998) for a discussion of the methodology. The signal extraction method uses nonparametric symmetric filters applied to the data, which is forecast and backcast extended (implicitly) in order to obtain signal estimates at the sample boundary.

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

to minimize an empirical multi-step ahead forecast error. Proietti (2011) expands on this work, investigating the forecast performance of these multi-step ahead fitted parameters. However, what is lacking so far is a coherent general treatment of the subject that handles difference linear processes, i.e., nonstationary process that have a Wold decomposition when suitably differenced. The main objective of this paper is to summarize and generalize all the preceding literature, compactly expressing the appropriate objective functions in the frequency domain.

The reason for recourse to the frequency domain is for concision of formulas as well as computational efficiency. For example, certain formulas in Tiao and Xu (1993) for the multi-step ahead forecast mean square error involve infinite summations, that would only be calculated via truncation in practice. Using the frequency domain, exact expressions can be derived utilizing the calculus of residues, avoiding the need to truncate. Well-known Fourier transform algorithms can be used to speedily compute the asymptotic multi-step ahead forecast mean squared errors, 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 in principle could be used to generalize our GKL, the actual implementation of such is yet unsolved. However, it seems a fruitful direction for future work. Secondly, our method only optimizes over one forecast lead at a time – 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 asymptotic forecast mean squared error as a model fitting criterion in Section 2. A key contribution is the practical formula for its computation. 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 correctly specified; 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 empirical forecast error and GKL through a chemical time series, and display results from a forecasting exercise involving several time series. 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 associated Toeplitz covariance matrix $\Gamma(f)$, whose jkth entry is simply $\gamma_{j-k}(f)$. We also use the notation $< g >$ 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 $\Psi(z) = \sum_{j \geq 0} \psi_j z^j$. 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., $\sigma^2 = \exp\{< \log f >\}$. 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 discussed in 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 we choose to base our approach on the semi-infinite predictors, rather than finite sample predictors (see Newton and Pagano (1982) for a discussion of their computation for stationary processes), is that we obtain a single time-invariant filter. This is in contrast to managing a suite of time-varying forecast error filters, whose length each depends upon one's time location within the sample. The net effect of this approach is to create a computationally simpler objective function, which 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 Stock and Watson (1999), Marcellino, Stock, and Watson (2006), and Proietti (2011) 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 the derivations in McElroy and Findley (2010). Also see Findley, Pötscher, and Wei (2004) for an alternative formulation. The square brackets denote the truncation of an infinite power series to those coefficients with index 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

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

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 one-step ahead forecast error does not involve the unit root properties of the time series, whereas multi-step ahead forecast error does. In McElroy and Findley (2010) the formula (1) is utilized as the basis of a model goodness-of-fit diagnostic, and is 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 might substitute the periodogram I (see below) for $\widetilde{f}$ in (1).

Let us rewrite (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 (1) with a generic function g, we obtain

$$J([\theta], g) = < \frac{f^{(h)}_{[\theta]}}{f_{[\theta]}}\, g > . \tag{2}$$

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. But $J([\theta], I)$ 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 discussed in McElroy and Findley (2010) – with derivations in Findley (1991) – $J([\theta], I)$ approximately corresponds to an empirical sum of h-step ahead forecast errors calculated from finite-sample predictors.

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 in 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 (2) can be computed, although now $\nabla_{[\theta]}\sigma^2 \neq 0$. We henceforth suppose that $J([\theta], g)$ can be evaluated; this is easy for ARIMA models, as 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 smallest possible forecast error within the model $\mathcal{F}$. Likewise we can obtain an empirical estimate via 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, knowledge of $[\theta_g]$ is sufficient, because the forecast filter does not depend on the innovation variance. But if a knowledge of forecast precision is desired, we must also obtain σ^2. The true innovation variance is denoted by $\widetilde{\sigma}^2 = \exp\{< \log \widetilde{f} >\}$, 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 obtain the true innovation variance, and $g = I$ for 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. For either of the separable and non-separable cases, (3) holds.

It follows that σ_I^2 will be consistent for $\widetilde{\sigma}^2$, as shown in Section 3 below. Note that setting $g = \widetilde{f}$ in (3) affords an interpretation for 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.

So these equations together give us an algorithm: first minimize (2) with respect to $[\theta]$, and then compute the minimal σ^2 via (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 and Kakizawa (2000)), and these are equal to the true parameters when

the model is correctly specified. But when $g = I$, this method provides us with parameter estimates $([\theta_I], \sigma_I^2)$ that are consistent for the pseudo-true values (no matter whether the model is correctly or incorrectly specified).

We now make some further connections of J with 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 and 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

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

If we wish to fit a model to the data, we minimize $KL(f_\theta, I)$ with respect to θ, denoting 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 $\log \sigma^2 + \sigma^{-2} < I/f_{[\theta]} >$ so that the optimal σ_I^2 equals $< I/f_{[\theta_I]} >$ (this requires $\nabla_{[\theta]}\sigma^2 = 0$). In other words when the model is separable, minimization of KL is equivalent to the two-step minimization of (2) and (3) for $h = 1$.

So Generalized Kullback-Leibler (GKL) discrepancy is defined analogously for $h \geq 1$:

$$GKL_\delta^{(h)}(f, g) = < \log f > + \log < f^{(h)} > + \frac{< \frac{g}{f} f^{(h)} >}{< f^{(h)} >}. \tag{5}$$

Note that this reduces to (4) when $h = 1$, since then $f^{(h)} \equiv 1$. But for $h > 1$ we have the extra $\log < f^{(h)} >$ term, without which minimization of (5) would not be equivalent to optimization via (2) and (3). This relationship is described in Proposition 2 of Section 3.

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

$$< fI > = \frac{1}{n} W' \Gamma(f) W$$

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

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

This is quite easy to compute for ARIMA models, for which the autocovariances are readily obtained (see Appendix A.2). In particular, no matrices need be inverted (unlike with maximum likelihood estimation). Computation of multi-step forecasts and forecast error covariances from a finite past for stationary processes is discussed in Newton and Pagano (1982); our approach utilizes semi-infinite forecast error filters instead, and thereby avoids much of the complexity required for matrix inversion.

The formula also holds for the non-separable case – 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 (when they exist) of $GKL_\delta^{(h)}(f_\theta, \widetilde{f})$. The statistical properties of the parameter estimates are treated in the Section 3.

Thus, formula (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 optimal h-step ahead (asymptotic) mean square forecast error. Of course, these parameters will vary widely in practice, since there is no need that optimality 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 available multiple parameter fits of the same model is useful, since each fit is optimal with respect to its own h-step ahead forecasting objective[4].

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 (in-sample this is necessarily so). 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 be easily related to optimization of multi-step ahead forecasting error J. Then we state 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 (5) that

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

Note that σ^2 may well depend upon $[\theta]$ in the non-separable case, but this dependency will be suppressed in the notation. Now (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).

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

We next state the gradient and Hessian functions of G for the separable and non-separable cases. In the former case, $\nabla'_\theta = [\nabla'_{[\theta]}, \frac{\partial}{\partial \sigma^2}]$, 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*

$$\nabla_{[\theta]} G(\theta) = \frac{< \nabla_{[\theta]} f^{(h)}_{[\theta]} >}{< f^{(h)}_{[\theta]} >} + \frac{\nabla_{[\theta]} J([\theta], g)}{\sigma^2 < f^{(h)}_{[\theta]} >} - \frac{J([\theta], g) < \nabla_{[\theta]} f^{(h)}_{[\theta]} >}{\sigma^2 < f^{(h)}_{[\theta]} >^2}$$

$$\frac{\partial}{\partial \sigma^2} G(\theta) = \sigma^{-2} - \sigma^{-4} \frac{J([\theta], g)}{< f^{(h)}_{[\theta]} >}$$

$$\begin{aligned}
\nabla_{[\theta]} \nabla'_{[\theta]} G(\theta) = & \frac{< \nabla_{[\theta]} \nabla'_{[\theta]} f^{(h)}_{[\theta]} >}{< f^{(h)}_{[\theta]} >} - \frac{< \nabla_{[\theta]} f^{(h)}_{[\theta]} >< \nabla'_{[\theta]} f^{(h)}_{[\theta]} >}{< f^{(h)}_{[\theta]} >^2} \\
& - \frac{\nabla_{[\theta]} J([\theta], g) < \nabla'_{[\theta]} f^{(h)}_{[\theta]} > + < \nabla_{[\theta]} f^{(h)}_{[\theta]} > \nabla'_{[\theta]} J([\theta], g)}{\sigma^2 < f^{(h)}_{[\theta]} >^2} \\
& + \frac{\nabla_{[\theta]} \nabla'_{[\theta]} J([\theta], g)}{\sigma^2 < f^{(h)}_{[\theta]} >} - \frac{< \nabla_{[\theta]} \nabla'_{[\theta]} f^{(h)}_{[\theta]} > J([\theta], g)}{\sigma^2 < f^{(h)}_{[\theta]} >^2} + 2 \frac{< \nabla_{[\theta]} f^{(h)}_{[\theta]} >< \nabla'_{[\theta]} f^{(h)}_{[\theta]} > J([\theta], g)}{\sigma^2 < f^{(h)}_{[\theta]} >^3}
\end{aligned}$$

$$\frac{\partial}{\partial \sigma^2} \nabla_{[\theta]} G(\theta) = \sigma^{-4} J([\theta], g) \frac{< \nabla_{[\theta]} f^{(h)}_{[\theta]} >}{< f^{(h)}_{[\theta]} >^2} - \sigma^{-4} \frac{\nabla_{[\theta]} J([\theta], g)}{< f^{(h)}_{[\theta]} >}$$

$$\frac{\partial^2}{\partial^2 \sigma^2} G(\theta) = -\sigma^{-4} + 2\sigma^{-6} \frac{J([\theta], g)}{< f^{(h)}_{[\theta]} >}.$$

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

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

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

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 calculation of the various derivatives of spectral densities and their associated inverse Fourier Transforms (FTs). So for many models these formulas are not practically useful, although they serve the purpose of establishing that minimization of GKL coincides with minimization of (2) together with computation of (3).

Proposition 2 *Suppose that the model is separable. If* $[\theta_g]$ *is a minimum of (2) and* σ_g^2 *is computed via (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* $G(\theta)$*.*

So GKL really corresponds to the multi-step ahead forecast error minimization problem. As a practical matter, minimization of (2) as opposed to GKL is more convenient, as it involves one less parameter (in the separable case). But 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 computation of pseudo-true values (let $g = \widetilde{f}$). We typically assume that the order of integration d has been correctly specified, and that appropriately differenced data is 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. But 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 in Findley, Pötscher, and Wei (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 is not in general equal to the inverse of the Fisher information matrix, when the model is correctly specified. This implies that estimates are inefficient. But when the model is mis-specified, we can no longer say what types of estimates have 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 and 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 and Taniguchi (1982) and Taniguchi and 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 in terms of second and fourth order structure the process looks Gaussian. 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

$$V(\theta) = 2 < \nabla_\theta \frac{f_{[\theta]}^{(h)}}{f_\theta < f_{[\theta]}^{(h)} >} \nabla_\theta' \frac{f_{[\theta]}^{(h)}}{f_\theta < f_{[\theta]}^{(h)} >} \widetilde{f}^2 > .$$

Remark 1 In order to produce estimated standard errors for parameter estimates, it is best to proceed as if the model was mis-specified (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 and 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 it is difficult in general to compute $J([\theta], g)$ explicitly, in some special cases this is possible. We first provide several analytical examples involving stationary and nonstationary DGPs. Then we consider several numerical illustrations of the GKL objective functions.

4.1 Analytical Derivations of Optima

When forecasting stationary processes long-term, 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 have no impact on forecasting performance. This situation is dramatically different in the presence of non-stationarity, because the large h behavior of GKL instead tends to infinity, rather than a constant. Our results below are computed in terms of generic g, which can be taken as either I or $\widetilde{f}$ as context dictates.

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

$$\begin{aligned} f_{[\theta]}(\lambda) &= |1-\phi z|^{-2} \\ f^{(h)}_{[\theta]}(\lambda) &= \left|\sum_{j=0}^{h-1} \phi^j z^j\right|^2 \\ J([\theta], g) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} g(\lambda)|1-\phi^h z^h|^2 \, d\lambda = (1+\phi^{2h})\gamma_0(g) - 2\phi^h \gamma_h(g). \end{aligned}$$

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 $J([\theta_g], g)$ illustrate 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 and Jenkins (1976) – $\rho_h(g) = \widetilde{\varphi}^{h-1}(\widetilde{\varphi}+\widetilde{\omega})(1+\widetilde{\varphi}\widetilde{\omega})(1+2\widetilde{\omega}\widetilde{\varphi}+\widetilde{\omega}^2)^{-1}$. So ϕ_g is either equal to zero, when this correlation $\rho_h(g)$ is negative and h is even, or is equal to $\rho_h^{1/h}(g)$ otherwise. So 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 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 equal to $\widetilde{\varphi}^{h/2}$ when the lag h is even. Then $\phi_g = 0$ 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 fairly complicated in general, but when $h > q$ we have $f_{[\theta]}^{(h)} = f_{[\theta]}$, 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 facet is no longer present 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 in Tiao and Xu (1993), and is easily adapted into our framework; write the MA polynomial as $1 - \theta B$ and consider 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

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

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

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

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

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

which is trivially done 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, 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 $1 - \theta B$ 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 don't 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 MA polynomial Ω and AR polynomial Φ. Also the innovation variance $\sigma^2 = 1$ in all cases.

- D1: $d = 1$, $\omega_1 = -.1$, $\omega_2 = 0 = \phi_1 = \phi_2$.
- D2: $d = 1$, $\omega_1 = -.8$, $\omega_2 = 0 = \phi_1 = \phi_2$.
- D3: $d = 1$, $\omega_1 = .7$, $\phi_1 = .2$, and $\omega_2 = 0 = \phi_2$.
- D4: $d = 1$, $\phi_1 = .9\cos(\pi/60)$, $\phi_2 = -.81$, and $\omega_1 = \omega_2 = 0$.
- D5: $d = 2$, $\phi_1 = 0 = \phi_2$, ω_1, ω_2 corresponding to SNR = .1 in STM
- D6: $d = 2$, $\phi_1 = 0 = \phi_2$, ω_1, ω_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 through D5 and D6 for different values of the SNR. Process D3 follows a mixed ARMA model, while D4 generates a cyclical effect with a period of 60 observations. The Models considered are – with $d = 1, 2$ corresponding to the DGP – ARIMA(p,d,q) with $p = 1, q = 0$ (AR), $p = 0, q = 1$ (MA), and $p = 0 = q$ (WN).

This gives 18 combinations of models and DGPs. For the AR and MA models, the objective function J of (2) can be computed, and is displayed in Figures 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 cris-crossing (resulting in oscillatory patterns in the optima as h changes). Tables 1 through 6 summarize the numerical minima, also presenting the pseudo-true innovation variances.

Firstly, DGP D1 (Table 1) shows the MA(1) parameter equal to truth (up to numerical error), as this model is correctly specified; but the misspecified ARIMA(1,1,0) model exhibits a h-step pseudo-true value for ϕ that varies slightly for small h and then stabilizes as h increases. The first two panels of Figure 1 confirm this behavior. More or less the same behavior is evident for DGP D2 in Table 2, 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 seem to stabilize for increasing h as well (Table 3), and qualitatively the objective functions for this case (bottom row of Figure 1) look quite similar to those for D2 and D1. DGP D4 is much more interesting, with the objective functions overlapping

one another for different h (top row of Figure 2). As a result, 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 ill-captured 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 Figure 2). Unfortunately, portions of these likelihoods (especially in the AR model case) are extremely flat, resulting in numerical imprecisions in the statement of the optima. 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 (as 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, J, and empirical forecast error. Then we study several seasonal time series originally featured in the NN3 forecasting competition[5], with the interesting finding that GKL with $h = 12$ performs competitively with the classical $h = 1$ criterion.

5.1 Chemical Data

We consider Chemical Process Concentration Readings (Chem for short)[6]. The sample has 197 observations. This series was studied in McElroy and Findley (2010), where it was argued that an ARIMA$(0, 1, 1)$ model was most appropriate, given several contenders, according to multi-step ahead forecasting criteria. The same model was identified for the series by Box and Jenkins (1976), and was also studied in Tiao and Xu (1993). Fitting Chem using $GKL_\delta^{(h)}$ yields the MA(1) polynomials $1 - .7B$, $1 - .8B$, and $1 - .84B$ for $h = 1, 2, 3$ respectively.

We noted earlier (Section 2) that the objective function J given in (2) is an asymptotic form of the forecast mean squared error. Empirical forecasts are generated from a finite sample of data, so that the actual forecast error filter is an approximate truncation of $[\Psi/\delta]_0^{h-1}(B)/\Psi(B)$. As discussed in McElroy and Findley (2010), $J([\theta], I)$ differs from the empirical forecast mean squared error by $O_P(n^{-1/2})$, where n is the number of h-step ahead forecasts.

So if we generate forecasts of Chem using a windowed sub-sample, and average the squared forecast errors, the resulting behavior should mimic that of J as n increases. In particular, let

[5]See http://www.neural-forecasting-competition.com/NN3/.

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

us consider the forecast h steps ahead, for $h = 1, 2, 3$, from a sample consisting of time indices $t = 1, 2, \cdots, 197 - n - h + s$, repeated for $s = 1, 2, \cdots, n$. Moreover, let us generate these forecasts from each of the three GKL 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 $GKL^{(j)}$ parameter used (so $j = 1$ corresponds to MA parameter $-.7$, $j = 2$ corresponds to $-.8$, and $j = 3$ corresponds to $-.84$) 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. 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 n large.

This is heuristic, because as we increase n we reduce the length of the set of weights 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 $n = 125$. It is also interesting that the 2-step GKL does well at 3-step ahead forecasting for smaller n, in light of the close values ($-.8$ and $-.84$) for their respective MA parameters.

5.2 NN3 Data

Our goal here is to fit a common set of models to the various time series, generate out-of-sample forecasts, and compare performance across the various GKL criteria utilized. A realistic assessment of the composite forecasting rule (see the last paragraph of Section 2), as was done for the Chem data, is not really possible for the NN3 Data due to the short length of the series (most of the series are seasonal with less than 12 years in the sample, so that a windowing technique – such as was utilized with the Chem data – is not compatible with having enough remaining data to get reasonable parameter estimates). Instead, we examine a separate question: are there any NN3 series for which the forecasting performance "in the competition" generated by a particular $GKL_{\delta}^{(h)}$ was superior to the one-step ahead forecasts arising from $GKL_{\delta}^{(1)}$? We describe the results of this query below.

The original NN3 competition utilized 111 monthly time series of varying lengths and starting dates, for the most part exhibiting seasonal and trend dynamics, and from each a final span of 18 observations was with-held. We have obtained the full span of each time series from the contest's designers, so that we can assess performance.

In our study we attempt to mimic fairly closely the conditions of the competition, but restrict to a common set of models to fix comparisons and facilitate didactic purposes. Therefore, we used automatic SARIMA model identification software (X-12-ARIMA version 0.3) to determine the Box-Cox transform, the preferred SARIMA model, and regression effects (outliers, trading day, and Easter effects). Out of 111 series, 27 of them required a log transformation and were best

fitted by an Airline model, according to X-12-ARIMA. This was the largest subclass of identified models, so we restricted our viewpoint to these 27 (this also includes some (010)(011) SARIMA models, which of course are nested in the Airline model). Broadening our comparisons to include other models would seem to cloud the picture – especially as plenty of series and models are non-seasonal. We also work with the regression-adjusted series in order to allow us to focus upon estimation of SARIMA parameters.

So to each of the 27 regression-adjusted series, the last 18 observations were withheld, and the log Airline model was fitted to the remaining data (so we could do the out-of-sample forecasting exercise), using the GKL objective function with $1 \leq h \leq 18$. The choice of leads was natural, given the original competition was to forecast each series up to 18 periods ahead. For each of the 27 series, we computed an 18 by 18 grid of absolute forecast errors, with each column corresponding to a forecast lead k and each row corresponding to a $GKL_\delta^{(j)}$ objective function. If we can sensibly combine results across leads (along each row), then we obtain an overall assessment of each $GKL_\delta^{(j)}$ as a forecasting procedure, for each of the 27 series.

The overall performance of submitted forecasts in the NN3 competition were judged according to Symmetric Mean Absolute Percentage Error (SMAPE), which is described in Armstrong (1985), so we adopt this as our method of synthesizing results. Letting A_k denote the target value (k steps ahead) and $\widehat{A}_k$ its forecast (from one of the $GKL_\delta^{(j)}$ forecasting methods), the formula is

$$SMAPE = \frac{1}{18} \sum_{k=1}^{18} \frac{2|A_k - \widehat{A}_k|}{A_k + \widehat{A}_k}$$

whenever A_k and $\widehat{A}_k$ are positive. These quantities were computed for each $1 \leq j \leq 18$, and for all 27 series. The resulting values are presented in Table 8. Also presented there is the best $GKL_\delta^{(j)}$ for each given series (i.e., the j that yielded lowest SMAPE), as well as the ratio of its SMAPE to that of $GKL_\delta^{(1)}$. In some cases there was substantial improvement over the $h = 1$ criterion – namely quasi-maximum likelihood estimation – though results were variable over all. In the best cases, there was close to a 20 percent improvement over classical estimation.

So if someone used a particular $GKL_\delta^{(j)}$ to produce all forecasts, and submitted the results to the competition, which criteria would be successful relative to $GKL_\delta^{(1)}$ (for those 27 series)? The two most successful such leads were $j = 1$ and $j = 12$, each of which performed best for 7 out of the 27 series. Given the seasonal nature of these series, it is not surprising that 12-step ahead forecasting is important to get right, and that a model optimized with respect to 12-step ahead forecasting may perform well at shorter forecast leads too.

We examine this idea further with Series 110. In this case $j = 12$ was the overall winner, and there is quite a bit of improvement over $j = 1$ – a 16 percent reduction to SMAPE. The parameters for the latter case, i.e., the conventional estimates, were $.08, .71$ for the nonseasonal and seasonal parameters of the Airline model. But when fitting with $GKL_\delta^{(12)}$, these became $.99, -.25$. This

is a radical, and meaningful, alteration to the parameters – from the long-term perspective, the process is not really $I(2)$, noting that the MA factor $(1-.99B)$ can essentially be canceled with one of the nonseasonal differencing operators of the model, reducing it to $I(1)$ (once a compensating mean regression effect is added). There is also a substantial change in the seasonal moving average parameter. For this series, it seems likely that the airline model is a misspecification[7] – it is in such scenarios that our multi-step criterion can be expected to offer some improvements to forecasting performance.

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 *a priori* to the analyst. 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 we have derived an asymptotic distribution theory for estimates that converge to "pseudo-true" values, expanding the consistency results of Findley, Pötscher, and Wei (2004) to central limit theorems. In contrast to earlier approaches (e.g., Tiao and Xu (1993) or Haywood and 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 performances of classical (one-step ahead) and generalized (h-step ahead) estimates in a controlled experimental design based on a selection of simulated as well as 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 Series 110 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. Specifically, we found that for $h = 12$ one of the MA roots approaches the unit circle, resulting in near cancelation of misspecified AR-roots.

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. We expect the frequency-domain approach underlying GKL to offer some promising perspectives on

[7]The software X-12-ARIMA identifies a SARIMA model by inserting a dubious level shift regressor, whereas the raw data shows little evidence of dynamic seasonality in its autocorrelation plot.

these future topics.

Appendix

A.1 Proofs

Proof of Proposition 2. Plugging $[\theta] = [\theta_g]$ and $\sigma^2 = \sigma_g^2$ (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:

$$\nabla_{[\theta]} \nabla'_{[\theta]} G(\theta)|_{\theta=\theta_g} = \frac{\nabla_{[\theta]} \nabla'_{[\theta]} J([\theta], g)|_{\theta=\theta_g}}{J([\theta_g], g)} + \frac{< \nabla_{[\theta]} f^{(h)}_{[\theta]} >< \nabla'_{[\theta]} f^{(h)}_{[\theta]} > J([\theta], g)}{< f^{(h)}_{[\theta]} >^2}|_{\theta=\theta_g}$$

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

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

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 $c = \nabla_{[\theta]} \int f^{(h)}_{[\theta]}|_{\theta=\theta_g} / J([\theta_g], g)$ 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, (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 (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

$$\left(\frac{\nabla_\theta \sigma^2}{\sigma^2} + \frac{\nabla_\theta f^{(h)}_{[\theta]}}{< f^{(h)}_{[\theta]} >} \right) \left(\frac{\nabla'_\theta \sigma^2}{\sigma^2} + \frac{\nabla'_\theta f^{(h)}_{[\theta]}}{< f^{(h)}_{[\theta]} >} \right) + \frac{\nabla_\theta \nabla'_\theta J([\theta], g)}{J([\theta_g], g)}|_{\theta=\theta_g},$$

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 and Kakizawa (2000))

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

where $r_\theta = \nabla_\theta f_{[\theta]}^{(h)} f_\theta^{-1} < f_{[\theta]}^{(h)} >^{-1}$. 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.2 Implementation for ARIMA models

In order to compute parameter estimates and pseudo-true values for a fitted ARIMA model, it is necessary to carefully set up an optimization algorithm. In the case that the DGP is a known ARIMA process and one seeks to obtain pseudo-true values, the integrand of J (2) can always be written as the spectral density of a composite ARMA process, 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 has 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

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

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 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 (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 (6).

5. Search for the next value of θ using BFGS or other numerical recipe.

Explicit formulas for the quantity in step 2 can be found in 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 the 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 of 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 bounds on the polynomial roots is automatically handled through the *flipIt* routine, so only the innovation variance needs to be constrained – this is most naturally handled through optimizing over $\log \sigma^2$ instead, which can take as value any real number. Then a conjugate gradient method such as BFGS (Golub and 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 of Section 4 in principle allow one to calculate these exact quantities, the programming effort is considerable and it is unclear whether there is any advantage to be gained, since the resulting formulas depend on multiple calls to *ARMAacvf* and the like.

References

[1] Armstrong, S. (1985) *Long-range forecasting*. New York: Wiley.

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

[3] Chiu, S. (1988) Weighted Least Squares Estimators on the Frequency Domain for the Parameters of a Time Series. *Ann. Statist.* **16**, 1315–1326.

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

[5] Dagum, E. (1980) *The X-11-ARIMA Seasonal Adjustment Method*. Ottawa: Statistics Canada.

[6] Dahlhaus, R., and Wefelmeyer, W. (1996) Asymptotically optimal estimation in misspecified time series models. *Ann. Statist.* **16**, 952–974.

[7] Diebold, F. and Mariano, R. (1995) Comparing predictive accuracy. *Journal of Business and Economics Statistics* **13**, 253–263.

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

[9] Findley, D., Pötscher, B., and Wei, C-Z. (2004) Modeling of time series arrays by multistep prediction or likelihood methods. *Journal of Econometrics* **118**, 151–187.

[10] Gersch, W. and Kitagawa, G. (1983) The prediction of time series with trends and seasonalities. *Journal of Business and Economics Statistics* **1**, 253–264.

[11] Golub, G. and Van Loan, C. (1996) *Matrix Computations*. Baltimore: Johns Hopkins University Press.

[12] Hannan, E. and Deistler, M. (1988) *The Statistical Theory of Linear Systems*. New York: Wiley.

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

[14] Haywood, J. and Tunnicliffe-Wilson, G. (1997) Fitting time series models by minimizing multistep-ahead errors: a frequency domain approach. *J.R. Statist. Soc. B* **59**, 237–254.

[15] Hillmer, S. and Tiao, G. (1982). An ARIMA-model-based approach to seasonal adjustment. *Journal of the American Statistical Association* **77**, 377, 63 – 70.

[16] Hosoya, Y. and Taniguchi, M. (1982) A Central Limit Theorem for Stationary Processes and the Parameter Estimation of Linear Processes. *Ann. Statist.* **10**, 132–153.

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

[18] Marcellino, M., Stock, J., and Watson, M. (2006) A Comparison of Direct and Iterated Multistep AR Methods for Forecasting Macroeconomic Time Series. *Journal of Econometrics* **135**, 499–526.

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

[20] McElroy, T. (2008b). Exact Formulas for the Hodrick-Prescott Filter. *Econometrics Journal* **11**, 1–9.

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

[22] McElroy, T. and Holan, S. (2009) A Local Spectral Approach for Assessing Time Series Model Misspecification. *Journal of Multivariate Analysis* **100**, 604–621.

[23] Newton, H. and Pagano, M. (1982) The finite memory prediction of covariance stationary time series. *SIAM J. Sci. Stat. Comput.* **4**, 330–339.

[24] Proietti, T. (2011) Direct and Iterated multistep AR Methods for Difference Stationary Processes. *International Journal of Forecasting* **27**, 266–280.

[25] Self, S. and 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.

[26] Stock, J. and Watson, M. (1999) Forecasting Inflation. *Journal of Monetary Economics* **44**, 293–335.

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

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

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

[30] 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

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 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)	$\widetilde{\phi}$	-0.4870	-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 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)	$\widetilde{\phi}$	0.5788	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 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)	$\widetilde{\phi}$	0.4970	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 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)	$\widetilde{\phi}$	-0.6567	-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 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)	$\widetilde{\phi}$	-0.2495	-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

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

$F(n)$			
GKL	Forecast Lead		
$n = 50$	1	2	3
1	**.0983**	.1307	.1426
2	.0998	.1219	.1298
3	.1009	**.1190**	**.1255**
$n = 75$	1	2	3
1	**.0863**	.1126	.1250
2	.0903	.1096	.1191
3	.0926	**.1089**	**.1171**
$n = 100$	1	2	3
1	**.0868**	.1082	.1201
2	.0889	.1047	.1136
3	.0903	**.1036**	**.1111**
$n = 125$	1	2	3
1	**.0853**	.1072	.1212
2	.0894	**.1062**	.1171
3	.0923	.1067	**.1163**
$n = 150$	1	2	3
1	**.0954**	.1131	.1247
2	.0977	**.1115**	.1208
3	.0999	.1119	**.1203**

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

SMAPE by Series and GKL																			
	GKL																		
Series	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	Ratio
013	.009	.010	.010	.010	.010	.011	.012	.011	.011	.012	.011	.011	.010	.010	.010	**.009**	.010	.009	0.986
021	**.010**	.011	.011	.011	.012	.012	.011	.012	.012	.011	.011	.013	.012	.012	.012	.012	.012	.012	1.000
024	.019	.018	.019	.018	**.017**	.019	.018	.019	.018	.019	.021	.020	.019	.019	.019	.019	.018	.018	0.942
030	**.005**	.005	.005	.005	.005	.005	.005	.005	.005	.005	.005	.005	.005	.005	.005	.005	.005	.005	1.000
036	**.010**	.010	.010	.010	.010	.010	.011	.011	.010	.010	.010	.011	.010	.011	.011	.011	.011	.011	1.000
050	.016	.016	.016	.016	.016	.016	.017	.017	.016	.017	.015	**.014**	.015	.015	.015	.015	.015	.015	0.864
051	**.018**	.019	.021	.022	.021	.020	.020	.020	.020	.020	.019	.019	.018	.019	.019	.019	.019	.019	1.000
057	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	**.002**	.002	.002	.002	.002	.002	.002	.002	0.757
059	.005	.005	.005	.005	.005	.005	.005	.005	.005	.005	.005	**.005**	.005	.005	.005	.005	.005	.005	0.925
061	.003	**.003**	.004	.004	.004	.004	.004	.004	.004	.004	.004	.007	.005	.005	.005	.005	.005	.005	0.994
062	.037	.039	.042	.039	.039	.036	.036	.039	.036	**.035**	.038	.044	.038	.037	.039	.038	.037	.039	0.935
066	.006	.005	**.005**	.005	.005	.005	.006	.006	.006	.006	.006	.007	.006	.005	.005	.005	.005	.005	0.807
074	**.016**	.018	.017	.017	.018	.019	.019	.019	.019	.019	.020	.020	.018	.018	.018	.017	.017	.018	1.000
076	.018	**.018**	.018	.018	.018	.019	.019	.019	.019	.019	.019	.019	.019	.019	.019	.019	.019	.020	0.998
079	.013	.013	**.013**	.013	.013	.013	.013	.013	.013	.013	.013	.014	.013	.013	.013	.013	.013	.013	0.986
081	**.018**	.019	.019	.019	.019	.019	.019	.020	.020	.020	.020	.020	.020	.020	.019	.020	.019	.019	1.000
082	.020	.021	.021	.021	.020	.020	.020	.021	.019	.018	.017	**.017**	.019	.019	.020	.020	.019	.020	0.830
083	.015	.015	.014	**.014**	.014	.015	.016	.017	.017	.016	.016	.016	.017	.017	.017	.016	.017	.017	0.910
084	**.003**	.003	.004	.005	.005	.005	.005	.004	.004	.004	.004	.004	.003	.003	.004	.005	.005	.005	1.000
087	.015	.015	.015	.015	.015	.015	.015	.016	.016	.016	.016	.021	.016	.016	.016	.016	.016	**.013**	0.899
098	.041	.040	.041	.041	.040	.039	.041	.041	.042	.042	.041	**.038**	.041	.040	.040	.040	.040	.040	0.940
100	.005	**.005**	.007	.010	.010	.010	.010	.010	.009	.009	.006	.006	.010	.009	.009	.009	.009	.009	0.897
101	.002	.002	.002	.002	.002	.002	.002	.002	.002	.003	.002	**.002**	.002	.002	.002	.002	.002	.002	0.996
103	.070	.070	.070	.067	.069	**.066**	.071	.070	.071	.073	.073	.072	.076	.075	.075	.078	.077	.076	0.949
105	.003	.004	.005	.005	.005	.005	.005	.005	.005	.005	.005	**.003**	.004	.005	.005	.005	.005	.005	0.982
106	.010	.009	.010	.009	.009	.008	**.008**	.009	.009	.008	.009	.009	.009	.009	.009	.008	.008	.008	0.831
110	.061	.060	.061	.062	.061	.060	.060	.059	.058	.057	.055	**.051**	.053	.054	.055	.055	.055	.055	0.837

Table 8: Values of SMAPE, averaged over various forecast leads, by series (row) and GKL fitting criteria (column). Bold entries indicate the row-minima, i.e., smallest SMAPE for each series, over various GKL criteria. The column marked "Best" refers to the ratio of this lowest bolded SMAPE to the SMAPE for the classical $k = 1$ GKL.

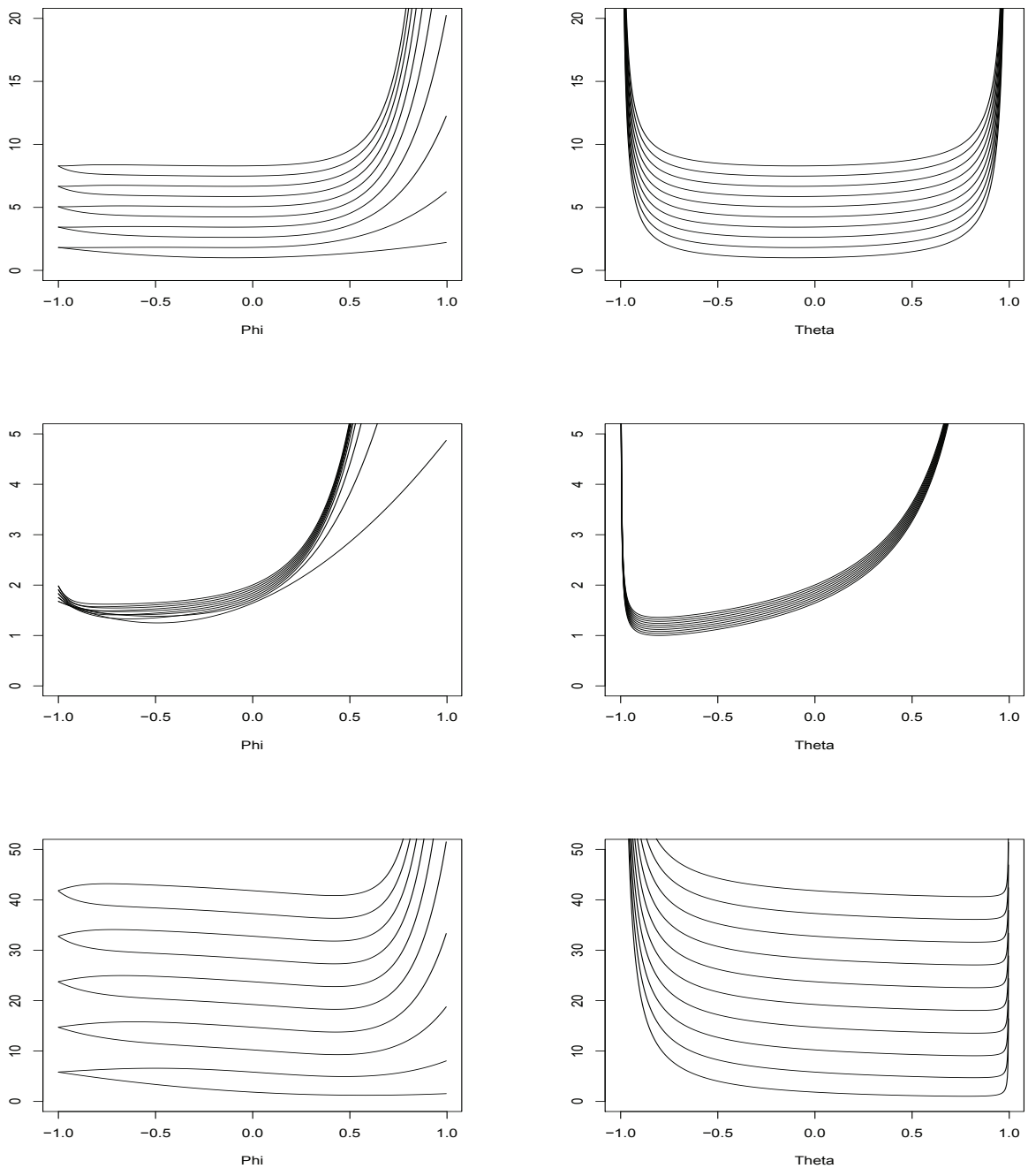

Figure 1: The left panels display the function J for the AR model, while the right panels displays the function J for the MA model. The upper panels correspond to DGP D1, the middle panels to DGP D2, and the lower panels to DGP D3. Overlaid objective functions correspond to h-step ahead forecast MSE,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

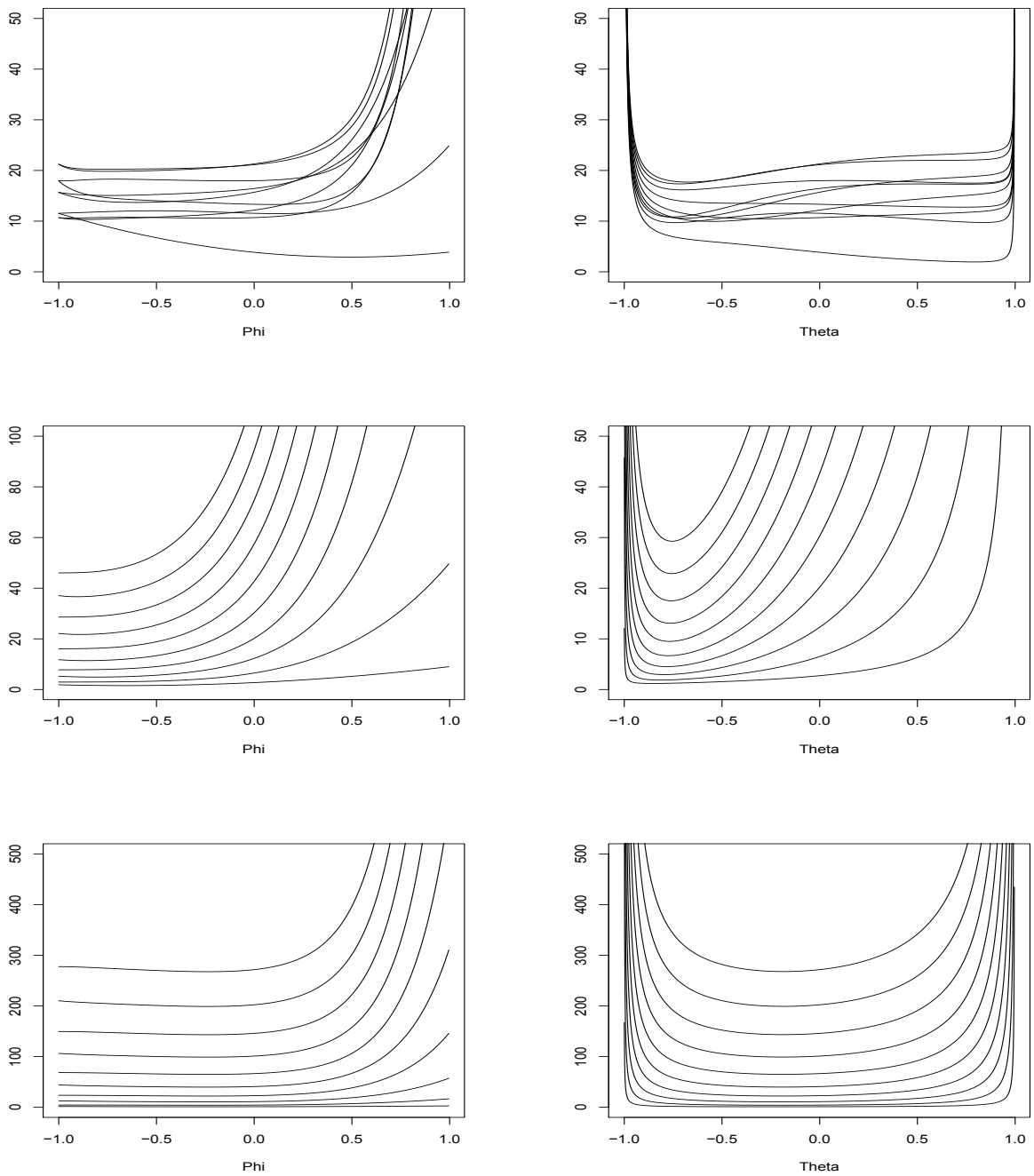

Figure 2: The left panels display the function J for the AR model, while the right panels displays the function J for the MA model. The upper panels correspond to DGP D4, the middle panels to DGP D5, and the lower panels to DGP D6. Overlaid objective functions correspond to h-step ahead forecast MSE,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