Parameter Estimation with Out-of-Sample Objective

Peter Reinhard Hansen[a] and **Elena-Ivona Dumitrescu**[b]

[a]*European University Institute & Stanford University & CREATES*

[b]*University of Orléans & Maastricht University*

Very preliminary/June 5, 2012

Abstract

We discuss parameter estimation in a situation where the objective is good out-of-sample performance. A discrepancy between the out-of-sample objective and the criterion used for in-sample estimation can seriously degrade the performance. Using the same criterion for estimation and evaluation typically ensures that the estimator is consistent for the ideal parameter value, however this approach need not be optimal. In this paper, we show that the optimal out-of-sample performance is achieved through maximum likelihood estimation (MLE), and that MLE can can be vastly better than the criterion based estimation (QBE). This theoretical result is analogous to the well known Cramer-Rao bound for in-sample estimation. A drawback of MLE is that it suffers from misspecification in two ways. First, the MLE (now a quasi MLE) is inefficient under misspecification. Second, the MLE approach involves a transformation of likelihood parameters to criterion parameters, which depends on the truth. So that misspecification can result in inconsistent estimation causing MLE to be inferior to QBE. We illustrate the theoretical result in a context with an asymmetric (linex) loss function, where the CBE performs on par with MLE when the loss is close to being symmetric, while the MLE clearly dominates QBE the the loss is asymmetric. We also illustrate the theoretical result in an applicable to long-horizon forecasting.

Keywords: Forecasting, Out-of-Sample, Linex Loss, Long-horizon forecasting.
JEL Classification: C52

1 Introduction

Out-of-sample evaluation is nowadays considered an acid-test for a forecasting model (Clements and Hendry, 2005), and the "...ability to make useful ex-ante forecasts is the real test of a model" (Klein, 1992). However, even though much interest has been given to the choice of the model specification and to proposing new evaluation tests (for equal forecasting abilities; encompassing, adapted for nested models, allowing to compare several specifications, etc.), the loss-function considered in such an out-of-sample experiment is generally arbitrarily chosen, regardless of the one used in-sample to estimate the parameters of the model. The most obvious choice is the mean square error (MSE), for its simplicity, but other criteria like mean absolute deviation (MAD), likelihood (LIK), or even asymmetric criteria such as linex or asymmetric quadratic loss are considered.

Important links between the way in which the parameters are estimated and the measures of predictive ability have, nevertheless, been noted, in particular by (Granger, 1969), (Weiss and Andersen, 1984) and Weiss (1996). As (Granger, 1969) states, evaluation criteria are used twice in an econometric analysis: first, to find the 'best' estimator of parameters and then to assess the predictive abilities of this model. He then argues that asymptotically the same criterion should be used in estimation and evaluation. This possible solution has been also embraced by (Weiss and Andersen, 1984) and Weiss (1996), who suggest that relative to a given loss function, out-of-sample predictive ability is enhanced if the same loss function is used to estimate parameters rather than using some other means of estimating the parameters. More recently, Schorfheide (2005) considers the particular case of quadratic loss functions in a potentially misspecified VAR(p) framework. He proposes a modification to Shibata (1980)'s final prediction error criterion to jointly choose between the maximum likelihood predictor and the loss function predictor and to select the lag length of the forecasting model.

Besides, several theoretical studies (see Weiss (1996), West, 1996, 2006) analyze the asymptotical properties of the evaluation criteria when taking into account estimation risk. West (1996) actually shows that parameter estimation error, which is the main issue arising from estimating with one criterion and evaluation with another, vanishes asymptotically when the same loss function is used both in- and out-of-sample. This asymptotic irrelevance holds quite generally when the out-of-sample sample (P) is much smaller than the in-sample one (R). Simulation evidence in West (1996), 2001), McCracken (2004) and Clark and McCracken (2001) suggests that a ratio $P/R < 0.1$ more or less justifies using the asymptotic approximation that assumes asymptotic irrelevance.

Although the estimator for the ideal parameter value seems to be consistent in this framework, it might not be optimal. In this paper we hence scrutinize the issue of wether this classic approach relying on the use of the same criterion in- and out-of-sample dominates other forms of estimation. To this aim, we consider the case of M-estimators (Amemiya, 1985, Huber, 1981) and rely on Akaike (1974)'s framework, known as the fixed scheme in the forecasting literature. We use a second-order Taylor expansion of the evaluation criterion around the optimal parameter value for each of the estimators considered and hence derive the expression for the expected value of the difference between the values of the evaluation criterion corresponding to the two estimators.

We show that the optimal out-of-sample performance is achieved through maximum likelihood estimation (MLE), and that MLE can can be vastly better than the performance produced by the criterion based estimation (CBE) whatever the out-of-sample criterion considered. Our theoretical result is analogous to the well known Cramer-Rao bound for in-sample estimation. We also consider the case where the likelihood has more parameters than the evaluation criterion, and discuss the losses incurred by the misspecification of the likelihood.

We illustrate the theoretical result in a context with an asymmetric (linex) loss function. Criterion based estimation performs on par with maximum likelihood when the loss is near-symmetric, whereas the MLE clearly dominates QBE with asymmetric loss. Most importantly, not only the asymptotic but also the finite-sample findings support these conclusions. In contrast, if the likelihood has the same number of parameters as the criterion-based predictor CBP (the other parameter being set to its true value), the gains from using MLE in forecasting relatively to CBE increase.

A second application of our theoretical result pertains to long-horizon forecasting. We consider the case of a well-identified gaussian linear AR(1) process, where the maximum-likelihood predictor (MLP) and the CBP are labeled Iterated (or 'plug-in') and Direct forecasts. It results that MLP outperforms CBP both when the model is estimated with and without an intercept. Besides, the longer the forecasting horizon the better the MLP relatively to CBP. Another important finding is that the relative performance of MLP with respect to CBP plunges when the process is nearly integrated (the autoregressive coefficient is close to 1).

The rest of the paper is structured as follows. Section 2 unfolds our theoretical results, whereas section 3 presents the results of the two applications. Section 4 concludes and the appendix collects the mathematical proofs.

2 Theoretical Framework

Let the objective be given by a criterion function, $Q(\mathcal{Y}, \theta)$, where $\mathcal{Y}$ is a sample and θ is a vector of unknown parameters. These parameters are to be estimated from the observed sample, $\mathcal{X}$. We follow standard convention and refer to $\mathcal{X}$ and $\mathcal{Y}$ as the *in-sample* and the *out-of-sample*, respectively.

In this paper we discuss estimation of θ when the objective is to maximize the expected value of $Q(\mathcal{Y}, \theta)$. The specific objective is to determine the estimator, $\hat{\theta}(\mathcal{X})$, that maximizes

$$\mathrm{E}[Q(\mathcal{Y}, \hat{\theta}(\mathcal{X}))]. \tag{1}$$

A natural candidate is the *direct estimator* given by

$$\hat{\theta}_D = \hat{\theta}_D(\mathcal{X}) = \arg\max_{\theta} Q(\mathcal{X}, \theta),$$

which is deduced from the same criterion that defines the out-of-sample objective. Since all estimators in this paper will be functions of $\mathcal{X}$, we shall drop the $\mathcal{X}$-argument to simplify the exposition.

The direct estimator need not be optimal, in the sense of maximizing (1). So, we shall compare it to estimators that are based on other criteria than Q, and we will show that the optimal estimator is the one deduced from maximum likelihood estimation. Another reason for considering alternative estimators is the common practice of estimating parameters using conventional methods without regard to the out-of-sample objective. This practice has many pitfalls and can result in out-of-sample performance that is substantially worse than that of the direct estimator.

For instance it is important that the estimator is consistent for the "true" parameter, which can sometimes be defined by

$$\theta_0 = \arg\max_{\theta} \mathrm{E}[Q(\mathcal{Y}, \theta)],$$

but more generally it is define as the maximizer of $\lim_{n\to\infty} n^{-1} Q(\mathcal{Y}, \theta)$,(below we make regularity conditions that ensure that θ_0 is well-defined).

Because the direct estimator is intrinsic to the criterion Q, it will be consistent for θ_0 under standard regularity conditions, in the sense that $\hat{\theta}_D \xrightarrow{p} \theta_0$ as the in-sample size increases. This consistency need not be satisfied by estimators based on other criteria.

Consider an alternative estimator defined by

$$\tilde{\theta}(\mathcal{X}) = \arg\max_{\theta} \tilde{Q}(\mathcal{X}, \theta).$$

In this paper we will compare the merits of $\tilde{\theta}$ with those of the direct estimator $\hat{\theta}_D$. This is done within the theoretical framework of M-estimators (extremum estimators), see Huber, 1981, Amemiya, 1985, and White (1994). Our exposition and notation will largely follow that in Hansen (2010).

To simplify the exposition, we consider the case where the sample size is the same for both $\mathcal{X}$ and $\mathcal{Y}$, and denote this by n. The case where the sample sizes do not coincide will be discussed.

In the following we let D_n denote a normalizing diagonal matrix, whose diagonal elements go to ∞ as $n \to \infty$. In the canonical case where parameter estimator converges in distribution at rate $\sqrt{n}$, we simply have $D_n = n^{1/2} I$, where I denotes the identity matrix.

Assumption 1. *(i) As $n \to \infty$ $n^{-1} Q(\mathcal{Y}, \theta) \xrightarrow{p} Q(\theta)$ uniformly in θ, where $Q(\theta)$ is a non-stochastic function with a unique global maximum, θ_0.*
(ii) For suitable functions, $S(\mathcal{Y}, \theta)$ and $H(\mathcal{Y}, \theta)$ the criterion function is such that

$$Q(\mathcal{Y}, \theta) = Q(\mathcal{Y}, \theta_0) + S(\mathcal{Y}, \theta_0)'(\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)' H(\mathcal{Y}, \dot{\theta})(\theta - \theta_0),$$

where $\| \dot{\theta} - \theta_0 \| = O(\| \theta - \theta_0 \|)$.
(iii) The maximizers $\hat{\theta} = \arg\max_{\theta \in \Theta} Q(\mathcal{X}, \theta)$ and $\tilde{\theta} = \arg\max_{\theta \in \Theta} \tilde{Q}(\mathcal{X}, \theta)$ converge in probability to θ_0 and $\tilde{\theta}_0$, respectively (both interior to Θ), and these estimators are uniquely given from

$$0 = S(\mathcal{X}, \hat{\theta}) = S(\mathcal{X}, \theta_0) + H(\mathcal{X}, \ddot{\theta})(\hat{\theta} - \theta_0),$$

with $\| \ddot{\theta} - \theta_0 \| = O(\| \hat{\theta} - \theta_0 \|)$. Moreover, for suitable functions $\tilde{S}(\mathcal{Y}, \theta)$ and $\tilde{H}(\mathcal{Y}, \theta)$,

$$0 = \tilde{S}(\mathcal{X}, \tilde{\theta}) = \tilde{S}(\mathcal{X}, \tilde{\theta}_0) + \tilde{H}(\mathcal{X}, \dddot{\theta})(\tilde{\theta} - \tilde{\theta}_0),$$

with $\| \dddot{\theta} - \tilde{\theta}_0 \| = O(\| \tilde{\theta} - \tilde{\theta}_0 \|)$;
(iv) $\Delta_n \xrightarrow{p} 0$ as $n \to \infty$ implies that $D_n^{-1} H(\mathcal{X}, \theta_0 + \Delta_n) D_n^{-1} \xrightarrow{p} -A$, $D_n^{-1} H(\mathcal{Y}, \theta_0 + \Delta_n) D_n^{-1} \xrightarrow{p}$

$-A$, *and* $D_n^{-1}\tilde{H}(\mathcal{X},\tilde{\theta}_0+\Delta_n)D_n^{-1})\xrightarrow{p}-\tilde{A}$, *and*

$$\begin{pmatrix} D_n^{-1/2}S(\mathcal{X},\theta_0) \\ D_n^{-1/2}\tilde{S}(\mathcal{X},\tilde{\theta}_0) \\ D_n^{-1/2}S(\mathcal{Y},\theta_0) \end{pmatrix} \xrightarrow{d} N(0, \begin{pmatrix} B & * & 0 \\ * & \tilde{B} & 0 \\ 0 & 0 & B \end{pmatrix}.$$

(iv) As $n\to\infty$ *we have* $\mathrm{E}[D_n^{-1/2}S(\mathcal{Y},\theta)|\mathcal{X}]\to 0$ *a.s.*

We leave part of the covariance matrix unspecified, because it does not play a role in our analysis.

Definition 1. Q and $\tilde{Q}$ are coherent if $\theta_0=\tilde{\theta}_0$, otherwise the criteria are said to be incoherent. Similarly, we refer to an estimator as coherent for the criterion Q if its probability limit is θ_0.

The effect of parameter estimation is given by the quantity $Q(\mathcal{Y},\hat{\theta})-Q(\mathcal{Y},\theta_0)$, and it follows from Hansen (2010) that

$$Q(\mathcal{Y},\hat{\theta})-Q(\mathcal{Y},\theta_0)\xrightarrow{d}-\frac{1}{2}Z_1'\Lambda Z_1+Z_1'\Lambda Z_2,$$

as $n\to\infty$, where $Z_1,Z_2\sim iidN(0,I)$ and Λ is a diagonal matrix with the eigenvalues of $A^{-1}B$. The expected loss that arises from parameter estimation using the direct estimator is (in an asymptotic sense) give by

$$\tfrac{1}{2}\mathrm{tr}\{A^{-1}B\}.$$

This result is related to Takeuchi (1976) who generalized the result by Akaike (1974) to the case with misspecified models.

Lemma 1. *Consider an alternative estimator,* $\tilde{\theta}$*, deduced from an incoherent criteria, so that* $\tilde{\theta}\xrightarrow{p}\tilde{\theta}_0\neq\theta_0$*. Then*

$$Q(\mathcal{Y},\theta_0)-Q(\mathcal{Y},\tilde{\theta})\to\infty,$$

in probability. In the canonical case the divergence is at rate n*.*

Proof. From Assumption 1 we have

$$Q(\mathcal{Y},\tilde{\theta})-Q(\mathcal{Y},\theta_0)=S(\mathcal{Y},\theta_0)(\tilde{\theta}-\theta_0)+\frac{1}{2}(\tilde{\theta}-\theta_0)'H(\mathcal{Y},\dot{\theta})(\tilde{\theta}-\theta_0),$$

where the first term is bounded in probability, $S(\mathcal{Y},\theta_0)(\tilde{\theta}-\theta_0)=O_p(1)$. Apply the decomposition, $(\tilde{\theta} - \theta_0) = (\tilde{\theta} - \tilde{\theta}_0 + \tilde{\theta}_0 - \theta_0)$ to the second term, reveals that the leading term is

$$-\Delta' D_n D_n^{-1} H(\mathcal{Y}, \theta_0) D_n^{-1} D_n \Delta = \Delta' D_n (A + o_p(1)) D_n \Delta,$$

where $\Delta = \tilde{\theta}_0 - \theta_0$. This term drifts to infinity $n \to \infty$ at rate $\| D_n \|^2$. □

So, the direct estimator strongly dominates all estimators that are based on incoherent criteria. This shows that consistency for θ_0 is a critical requirement.

2.1 Likelihood-Based Estimator

Next we consider the maximum likelihood estimator. Let $\{P_\vartheta\}_{\vartheta \in \Xi}$ be a statistical model, with P_{ϑ_0} the true probability measure, so that the expected value is defined by $\mathrm{E}_{\vartheta_0}(\cdot) = \int (\cdot) \mathrm{d}P_{\vartheta_0}$. In particular we have

$$\theta_0 = \arg\max_\theta \mathrm{E}_{\vartheta_0}[Q(\mathcal{Y}, \theta)],$$

which defines θ_0 as a function of ϑ_0, i.e. $\theta_0 = \theta(\vartheta_0)$.

Assumption 2. *There exists $\tau(\vartheta)$ so that $\vartheta \leftrightarrow (\theta, \tau)$ is continuous and one-to-one in an open neighborhood of $(\theta_0, \tau_0) = (\theta(\vartheta_0), \tau(\vartheta_0))$.*

Lemma 2. *Given Assumption 1 and 2, let $\tilde{\vartheta}$ be the MLE. Then $\tilde{\theta} = \theta(\tilde{\vartheta})$ is a coherent estimator.*

Proof. Let P denote the true distribution. Consider the parameterized model, $\{P_\vartheta : \vartheta \in \Xi\}$, which is correctly specified so that $P = P_{\vartheta_0}$ for some $\vartheta_0 \in \Xi$. Since θ_0 is defined to be the maximizer of

$$\mathrm{E}[Q(\mathcal{Y}, \theta)] = \mathrm{E}_{\vartheta_0}[Q(\mathcal{Y}, \theta)] = \int Q(\mathcal{Y}, \theta) \mathrm{d}P_{\vartheta_0},$$

it follows that θ_0 is just a function of ϑ_0, i.e., $\theta_0 = \theta(\vartheta_0)$. □

Remark. One challenge to using the MLE is that it may be complicated to determine the mapping $\theta(\cdot)$.

Theorem 1. *OPTIMALITY OF MLE. Let $\hat{\theta}_D = \arg\max_{\theta \in \Theta} Q(\mathcal{X}, \theta)$ denote the direct estimator and $\tilde{\theta}_{ML} = \theta(\tilde{\vartheta})$ denote the MLE based estimator, where $\tilde{\vartheta}$ denotes the maximum likelihood estimator. Then, as $n \to \infty$*

$$Q(\mathcal{Y}, \hat{\theta}_D) - Q(\mathcal{Y}, \tilde{\theta}_{ML}) \xrightarrow{d} \xi,$$

where $\mathrm{E}[\xi] \leq 0$.

Remark. We will, in most cases, have a strict inequality in which case the ML based estimator is superior to the direct estimator, in an asymptotic sense.

Proof. To simplify notation we write $Q_x(\theta)$ in place of $Q(\mathcal{X},\theta)$, and similarly $S_x(\theta) = S(\mathcal{X},\theta)$, $H_x(\theta) = H(\mathcal{X},\theta)$, $Q_y(\theta) = Q(\mathcal{Y},\theta)$, $\tilde{Q}_x(\theta) = \tilde{Q}(\mathcal{X},\theta)$, etc. Moreover, we denote the direct estimator by $\hat{\theta}$ and the MLE by $\tilde{\theta}$, where we recall that for a correctly specified likelihood we have the information matrix equality, $\tilde{A} = \tilde{B}$.

Out-of-sample we have that

$$\begin{aligned} Q_y(\hat{\theta}) - Q_y(\theta_0) &= S_y(\theta_0)'(\hat{\theta} - \theta_0) + \frac{1}{2}(\hat{\theta} - \theta_0)'H_y(\theta^*)(\hat{\theta} - \theta_0) + o_p(1) \\ &\simeq S_y(\theta_0)'[-H_x(\theta_0)]^{-1}S_x(\theta_0) - \frac{1}{2}S_x(\theta_0)'[-H_x(\theta_0)]^{-1}[-H_y(\theta_0)][-H_x(\theta_0)]^{-1}S_x(\theta_0) \end{aligned} \tag{2}$$

whereas for the MLE we find

$$\begin{aligned} Q_y(\tilde{\theta}) - Q_y(\theta_0) &= S_y(\theta_0)'(\tilde{\theta} - \theta_0) + \frac{1}{2}(\tilde{\theta} - \theta_0)'H_y(\theta_0)(\tilde{\theta} - \theta_0) + o_p(1) \\ &\simeq S_y(\theta_0)'[-\tilde{H}_x(\theta_0)]^{-1}\tilde{S}_x(\theta_0) - \frac{1}{2}\tilde{S}_x(\theta_0)'[-\tilde{H}_x(\theta_0)]^{-1}[-H_y(\theta_0)][-\tilde{H}_x(\theta_0)]^{-1}\tilde{S}_x(\theta_0) \end{aligned} \tag{3}$$

So that the difference in the criterion value for the two estimators is given by

$$\begin{aligned} Q_y(\hat{\theta}) - Q_y(\tilde{\theta}) &\simeq S_y(\theta_0)'[-H_x(\theta_0)]^{-1}S_x(\theta_0) - \frac{1}{2}S_x(\theta_0)'[-H_x(\theta_0)]^{-1}[-H_y(\theta_0)][-H_x(\theta_0)]^{-1}S_x(\theta_0) \\ &- S_y(\theta_0)'[-\tilde{H}_x(\theta_0)]^{-1}\tilde{S}_x(\theta_0) + \frac{1}{2}\tilde{S}_x(\theta_0)'[-\tilde{H}_x(\theta_0)]^{-1}[-H_y(\theta_0)][-\tilde{H}_x(\theta_0)]^{-1}\tilde{S}_x(\theta_0). \end{aligned}$$

By the law of iterated expectations, two of the terms drop out when taking expectations, so the expected value of the limit distribution is given from the two quadratic forms. The limit distribution of these two terms are given by

$$\frac{1}{2}\left(\tilde{Z}'\tilde{B}^{1/2}\tilde{A}^{-1}A\tilde{A}^{-1}\tilde{B}^{1/2}\tilde{Z} - Z'B^{1/2}A^{-1}B^{1/2}Z\right), \tag{4}$$

where $Z, \tilde{Z} \sim N(0, I)$. The expected value of the first term is given from

$$\mathrm{tr}\left\{\tilde{B}^{1/2}\tilde{A}^{-1}A\tilde{A}^{-1}\tilde{B}^{1/2}\mathrm{E}(\tilde{Z}\tilde{Z}')\right\} = \mathrm{tr}\left\{A\tilde{B}^{-1}\right\},$$

where we used the information matrix equality. The expectation for the second term is found similarly, so that the expectation of (4) is given by

$$\frac{1}{2}\left(\mathrm{tr}\left\{A\tilde{B}^{-1}\right\} - \mathrm{tr}\left\{A^{-1}B\right\}\right) = \frac{1}{2}\left(\mathrm{tr}\left\{A^{1/2}(\tilde{B}^{-1} - A^{-1}BA^{-1})A^{1/2}\right\}\right) \leq 0.$$

The inequality follows from the fact that $\tilde{B}^{-1}$ is the asymptotic covariance matrix of the MLE whereas $A^{-1}BA^{-1}$ is the asymptotic covariance of the direct estimator, so that $A^{-1}BA^{-1} - \tilde{B}^{-1}$ is positive semi-definite by the Cramer-Rao bound. The line of arguments are valid whether θ has the same dimension as ϑ or not, because we can reparametrize model in $\vartheta \mapsto (\theta, \gamma)$ that results in block-diagonal information matrices. This is achieved with

$$\gamma(\vartheta) = \tau(\vartheta) - \Sigma_{\tau\theta}\Sigma_{\theta\theta}^{-1}\theta(\vartheta),$$

where

$$\begin{pmatrix} \Sigma_{\theta\theta} & \Sigma_{\theta\tau} \\ \Sigma_{\tau\theta} & \Sigma_{\tau\tau} \end{pmatrix},$$

denotes the asymptotic covariance of the MLE for the parametrization (θ, τ).

□

2.2 The Case with a Misspecified Likelihood

Misspecification deteriorates the performance of the likelihood based estimators through two channels. First, the resulting estimator is no longer efficient, eliminating the argument in favor of adopting the likelihood-based estimator. Second and more important, misspecification can impact the proper mapping from ϑ to θ. Thus the MLE-based estimator $\tilde{\theta}$ may become inconsistent under misspecification, making in very inferior to the direct estimator.

2.3 Quantifying the Relative Efficiency

In the next two sections we illustrate the theoretical results with applications to the case with an asymmetric linex loss function and the case with multistep-ahead forecasting. For this purpose we define the *relative criterion efficiency*

$$\mathrm{RQE}(\hat{\theta}, \tilde{\theta}) = \frac{\mathrm{E}[Q(\mathcal{Y}, \tilde{\theta}(\mathcal{X})) - Q(\mathcal{Y}, \theta_0)]}{\mathrm{E}[Q(\mathcal{Y}, \hat{\theta}(\mathcal{X})) - Q(\mathcal{Y}, \theta_0)]}. \tag{5}$$

We now illustrate the results obtained in the previous section. The first application studies the out-of-sample relative efficiency of the maximum-likelihood (MLP) and criterion-based predictors (CBP) in the framework of asymmetric loss functions. The second one looks at direct vs. iterated multi-period forecasts in the case of a stationary linear AR(1) process with gaussian innovations in the case where the model is estimated with and respectively without a constant term. Note that in this well-specified model the iterated estimator is equivalent to the maximum-likelihood one (Bhansali, 1999).

3 The Linex Loss Function

In this section we apply the theoretical results to the case where the criterion function is given by the linex loss function. Symmetric loss, such as that implied by the mean square error is inappropriate in many application, e.g. Granger, 1986, Christoffersen and Diebold, 1997, and Hwang et al., 2001. The linex loss function is a highly tractable asymmetric loss function that was introduced by Varian, 1974, that has found many applications in economics, see e.g. Weiss and Andersen, 1984, Zellner, 1986, Diebold and Mariano, 1995, and Christoffersen and Diebold, 1997).

Here we shall adopt the following parameterization of the linex loss function

$$L_c(e) = c^{-2}[\exp(ce) - ce - 1], \qquad c \in \mathbb{R}\backslash\{0\}, \tag{6}$$

which has minimum at $e = 0$. The parameter c determines the degree of asymmetry, in the sense that the sign of c determines whether the symmetry is skewed to the left or right. The asymmetry increases with the absolute value of c while quadratic loss arises as the limited case, $\lim_{c\to 0} L_c(e) = 3e^2$, see Figure 1.

The optimal linex predictor for x solves $x^* = \underset{\hat{x}}{\operatorname{argmin}} E[L_c(e)]$. Christoffersen and Diebold, 1997 showed that the optimal predictor is actually the population mean μ plus a function of the prediction-error variance σ^2 and the degree of asymmetry c,

$$x^* = \mu + \frac{c\sigma^2}{2}. \tag{7}$$

It immediately follows that the maximum-likelihood predictor is

$$\tilde{x} = \tilde{\mu} + \frac{c\tilde{\sigma}^2}{2}, \tag{8}$$

where $\tilde{\mu}$ is the sample mean and $\tilde{\sigma}^2$ is the sample variance. At the same time, the criterion-based predictor $\tilde{x}$ minimizes the sum of in-sample losses $\tilde{x} = \arg\min_{\breve{x}} \sum_{i=1}^{n} L_c(e)$. In fact, in the particular case of the linex loss this minimization problem has a closed-form solution

$$\hat{x} = \frac{1}{c}\log[\frac{1}{n}\sum_{i=1}^{n}\exp(cx_i)], \tag{9}$$

where n is the sample size. The tractability of the three predictors under linex loss reduces computational burden and makes this loss function very attractive for performing Monte-Carlo simulations. We hence compare the out-of-sample performance of the maximum-likelihood predictor and that of the criterion-based predictor by relying on the opposite of the linex loss as an evaluation criterion $Q(\breve{e}) = -\sum_{i=n+1}^{2n} c^{-2}[\exp(c\breve{e}_i) - c\breve{e}_i - 1]$, where $\breve{e}_i = x_i - \breve{x}$ is the prediction error for the i$^{\text{th}}$ out-of-sample observation and $\breve{x}$ represents each of the predictors at a time. Denote by $Q(e^*)$ the evaluation criterion for the optimal predictor, by $Q(\hat{e})$ the criterion for the maximum likelihood predictor and by $Q(\tilde{e})$ the one for the linex criterion-based predictor, respectively. To achieve our objective, we hence consider the following experiment.

Step 1. A sample of size $2n$ is drawn from the normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$. The first n observations (in-sample) are used to generate the prediction and the other n observations constitute the out-of-sample set with which the prediction is compared in order to calculate the loss.

Step 2. The three predictors are immediately obtained by applying eq.7, eq.8, eq.9 to the in-sample data.

Step 3. Compute the out-of-sample evaluation criterion for the three estimators.

Step 4. Repeat steps 1 to 3 a large number of times (100,000 and 500,000 simulations are considered here).

Step 5. We can now evaluate the out-of-sample performance of the predictors. Since x^* is the optimal predictor under the linex loss, the expected value of the evaluation criterion associated with x^* is always larger than the one corresponding to the two other predictors. It follows that both the numerator and denominator of eq. 5 are negative, so that a $RQE < 1$ indicates that the maximum likelihood predictor performs better than the criterion-based one under the linex out-of-sample evaluation criterion. Conversely, $RQE > 1$ would support the choice of the linex criterion-based predictor over the ML one.

The experiment is repeated for different sample sizes $n \in \{100; 1,000; 1,000,000\}$, so as to emphasize both the finite-sample and the asymptotic relative efficiency of the two predictors. Note also the equivalence between increasing (decreasing) the asymmetry parameter c and increasing (decreasing) the standard-deviation σ in the case where x_i is normally distributed with mean 0 and variance σ^2. Denote by L_1 the loss associated with the optimal predictor x^* for an asymmetry coefficient c. Therefore, $L^*_{c,i} = c^{-2}[\exp(c(x_i - x^*)) - c(x_i - x^*) - 1]$, which, by equation 7, is equivalent to $L^*_{c,i} = c^{-2}[\exp(cx_i - c^2\sigma^2/2)) - cx_i + c^2\sigma^2/2\text{-}1]$. Let z be a random variable such that $z_i = cx_i$. Then, $z_i \sim N(0, c^2\sigma^2)$, so that the loss incurred by its optimal predictor is given by $L_{2,i} = d^{-2}[\exp(dz_i - dz^*) - d(z_i - z^*) - 1]$, where d is the asymmetry coefficient and $z^* = d(c\sigma)^2/2$. It follows that there is a value of d for which $L_{1,i}$ and $L_{2,i}$ are equivalent $\forall i \in \{1, n\}$. It is then clear that decreasing (increasing) the asymmetry coefficient c times while increasing (decreasing) the standard-deviation c times does not change the output in terms of loss. In view of this result we decide to fix the standard deviation to 1 while considering several values of the asymmetry coefficient, i.e. $c \in \{0.01; 0.1; 1; 2; 3\}$.

The results are reported in table 1. Part *i* displays the asymptotic results ($n = 1,000,000$) whereas part *ii* reports the finite-sample findings ($n \in \{100; 1,000\}$). 100,000 simulations have been performed in the former case, whereas 500,000 repetitions were considered in the latter case.

Our main finding in large-samples is that the relative efficiency of MLP with respect to CBP increases as the degree of asymmetry rises (see column 2 of Table 1). To be more precise, as c increases, the ratio RQE gets closer to 0, since the expected value of the criterion for MLP does not diverge from the one corresponding to optimal predictor as rapidly as that of CBP. Besides, it seems that for near-symmetric loss, i.e. very low values of c such as 0.01 or 0.1, the two predictors fare similarly relative to the optimal one (RQE equals 1).

At the same time, columns 3 and 4 display the difference between the expected value of the criterion for the two predictors and that of the optimal predictor x^*. The expected values are accurately estimated, with standard deviations smaller than 10^{-2} for all sample sizes considered. [1] It is clear that the larger the asymmetry, the smaller the expected values, i.e. the larger the forecast loss. This indicates that the two predictors move away from the optimal one, CBP's speed of divergence being higher than that of MLP.

The last 3 columns of the table include the expected value of the optimal predictor, as well as the expected values of the biasses associated with MLP and CBP. Small values of standard deviations associated with the expected values, i.e. less than 10^{-7}, insure the desired level of accuracy of the results. As expected, we find that the bias asymptotically vanishes (see part *i*

[1] These results are available upon request.

of the table). Additionally, it appears that CBP exhibits a bias larger than MLP.

3.1 Finite-Sample Considerations

The small sample findings basically support our asymptotic results with the following caveats (see part *ii* of table 1). First, the relative efficiency of MLP with respect to CBP indeed increases as c increases. However, as expected, RQE decreases more slowly in small samples (e.g. it reaches 0.279 for $n = 100$ as opposed to 0.012 for $n = 1,000,000$ when $c = 3$). Second, to compare results across sample sizes, we emphasize the fact that $Q(\cdot)$ is the opposite of the sum of out-of-sample losses associated with a certain predictor. In other words, is necessary to properly rescale these results by dividing by the number of out-of-sample observations considered in a specific experiment. We hence obtain the expected value of per-observation out-of-sample criterion relative to the optimal predictor, which is independent of the sample size. Notice that the smaller the sample size, the smaller the evaluation criterion and implicitly the larger the loss associated with the out-of-sample forecast. For example, the per-observation criterion is -0.005 for $n = 100$, -0.000495 for $n = 1,000$ and -0.0000000495 for $n = 1,000,000$ when $c = 0.01$. Third, the predictors exhibit small-sample bias, which increases with the degree of asymmetry (see part *ii* of the table).

All in all, the MLP performs relatively better than CBP whatever the sample size as long as the degree of asymmetry considered is close to 1 or larger. Furthermore, according to RQE, CBP does not outperform MLP in any case. These results confirm our theoretical findings that MLP is the out-of-sample equivalent of Cramer-Rao lower bound.

3.2 Further results

As discussed in the theoretical section, in the case of the linex loss, the MLP relies on two MLE ($\hat{\mu}$ and $\hat{\sigma}$) whereas the CBP is obtained directly. It follows that less estimation risk should characterize the CBP relatively to the MLP. In this context, for comparability reasons, we consider here the case with only one parameter estimated by ML, the other one being considered known and equal to its true value. Table 2 reports the asymptotic results obtained through 100,000 simulations for different values of the asymmetry coefficient. To be more precise, panel *i* presents the case of estimated mean $\hat{\mu}$ and known variance σ, while panel *ii* includes the results obtained for known mean μ and estimated variance $\hat{\sigma}$. Notice that when only the mean is estimated by MLE, the loss associated with the MLP is constant (and roughly -0.5), the relative efficiency of this predictor increasing faster than in the case where both the mean and variance are estimated (see panel i in table 1). By contrast, if only the variance is estimated by MLE, the loss increases with the degree of asymmetry. It is important to note that the MLE loss in table 1 is actually the sum of the losses reported in panels i and ii in table 2. At the same time, the relative efficiency of the MLP with respect to the CBP soars when the variance is estimated, relatively to the case where both the mean and variance are estimated. In this case, MLP is relatively more efficient than CBP even for nearly symmetric loss (small values of the asymmetry coefficient). All in all, if the two predictors (MLP and CBP) rely on the same number of estimators, i.e. one in this case, and an asymmetric loss function is considered, the gains from using MLE in forecasting relatively to CBE increase (RQE is closer to 0).

3.3 Likelihood Misspecification

To study the effects of likelihood misspecification, we now consider that the sample is drawn from the normal-inverse gaussian (NIG) distribution. Two cases are considered. First, a NIG(0,1,0,3) is considered, where the numbers between parentheses represent the first four moments. Second, a NIG(-20.47,46.77,-1,1.67), i.e. with negative asymmetry, is used. Besides, the asymmetry coefficient in the linex function is set to 0.01 and 0.1, respectively. The experiment implemented is similar to the one presented at the beginning of this section, with the specification that the optimal predictor is now

$$x^*_{NIG} = \frac{kc + \delta\sqrt{\alpha^2 - \beta^2}) - \delta\sqrt{\alpha^2 - (\beta + c)^2})}{c}, \tag{10}$$

instead of that given by eq.7, and that the MLP in eq.8 is now called QMLP, since the gaussian distribution is no longer the true one. Note also that k, α, β and δ represent the location, tail heaviness, asymmetry and scale parameters of the NIG distribution respectively. The top part of table 3 shows that QMLP performs on par with CBP if a standard NIG distribution is employed. If, however, a different parametrization is considered, the QMLE becomes inconsistent, and its performance drops beneath that of the CBP (see the bottom part of table 3). Additionally, in this context, the more the loss function is asymmetric, the larger the gap between the relative efficiency of the two predictors.

4 Long-horizon forecasting

Our theoretical results are also applicable to multi-period ahead forecasting, specifically to the debate on the relative merits of direct forecast versus iterated forecast. There is a vast literature tackling this issue, and we do not have much to add to this particular setting beyond showing how this literature is related to our framework. Consider the case of a mean-square error (MSE) loss function and a true model given in the form of a stationary finite-order autoregressive model, for which the asymptotic theory has been established in Bhansali (1997) and Ing et al., 2003

$$y_t = \mu + \sum_{i=1}^{p} \varphi_i y_{t-i} + \varepsilon_t, \tag{11}$$

where $1 \leq p < \infty$ is the order of the autoregressive process, $\{\varphi_i\}_{i=1}^{p} \neq 0$ is the set of parameters and $\{\varepsilon_t\}$ is a sequence of (unobservable) independent and identically distributed (i.i.d.) random noises, each with mean 0 and variance σ^2 for $t \in \{1, T\}$. Recall that direct forecasts are obtained by regressing the multi-period ahead value of the variable on its present and past values for each forecast horizon. In contrast, iterated forecasts (also called "plug-in" forecasts) are obtained by considering the same fitted model across all forecast horizons and iterating forward.

Mapping this iterated vs. direct forecasting debate into our theoretical framework thus involves answering the question of wether using the same criterion for parameter estimation and forecast evaluation, namely the h-periods ahead MSE, i.e. direct forecasting, improves the forecasting abilities of the model compared to the use of the 1-period ahead MSE in the estimation step, i.e. iterated forecasting. Furthermore, to match perfectly our theoretical setup, the model is assumed correctly specified ($k \geq p$) and the disturbances are supposed to be normally distributed so that the least-squares estimators of the AR coefficients in the estimated AR(k) model are asymptotically equivalent to the MLE (Bhansali, 1999; Ing et al., 2003). At the same time, the estimator of the true parameter corresponding to the direct forecasts is the OLS estimator (Kabaila, 1981; Bhansali, 1997). It follows that the first estimator is asymptotically efficient whereas the second is asymptotically inefficient for a forecast horizon larger than 1. To keep the notation consistent with the rest of the paper we hereafter label the iterated estimator MLE and the direct estimator CBE.

For ease of exposition, without loss of generality, we now restrict our attention to the case of a first-order autoregressive model where the disturbances are normally distributed with mean 0 and variance 1. Two cases are considered. First, the mean of the autoregressive process, μ

is considered to be known and equal to 0. Equation 11 becomes $y_t = \varphi y_{t-1} + \varepsilon_t$. It follows that the optimal predictor is $y^*_{T+h} = \theta^* y_T$, where θ^* solves $\arg\min_\theta E[MSE(y_{T+h}, \theta)]$, where $MSE(y_{T+h}, \theta) = \mathrm{E}(y_{T+h} - y^*_{T+h})^2$. It can actually be shown that for a stationary process $\theta^* = \varphi^h$, so that the optimal predictor is

$$y^*_{T+h} = \varphi^h y_T. \tag{12}$$

At the same time, the iterated estimator (MLE) $\tilde{\theta} = \tilde{\varphi}^h$, where $\tilde{\varphi}$ minimizes the in-sample 1-step-ahead loss, $\tilde{\varphi} = \arg\min_\varphi \sum_{t=1}^T (y_t - \varphi y_{t-1})^2$. The MLE predictor is hence

$$\tilde{y}_{T+h} = \tilde{\varphi}^h y_T. \tag{13}$$

Last but not least, the direct estimator (CBE) solves $\hat{\theta} = \arg\min_\theta \sum_{t=1}^T (y_t - \theta y_{t-h})^2$, so that its associated predictor is

$$\hat{y}_{T+h} = \hat{\theta} y_T. \tag{14}$$

Second, the mean of the autoregressive process is considered unknown and hence it is estimated along with φ. Since the true mean is zero, the optimal predictor remains unchanged. By contrast, the MLE and CBE now solve $\{\tilde{\varphi}, \tilde{\mu}\} = \arg\min_{\varphi,\mu} \sum_{t=1}^T (y_t - \mu - \varphi y_{t-1})^2$ and $\{\hat{\theta}, \hat{\mu}\} = \arg\min_{\theta,\mu} \sum_{t=1}^T (y_t - \mu - \theta y_{t-h})^2$ respectively, so that the two predictors become

$$\tilde{y}_{T+h} = \tilde{\varphi}^h y_T + \tilde{\mu} \sum_{j=0}^{h-1} \tilde{\varphi}^j \tag{15}$$

and

$$\hat{y}_{T+h} = \hat{\mu} + \hat{\theta} y_T. \tag{16}$$

The forecasting abilities of the two predictors, i.e. MLP and CBP, can now be scrutinized in both cases (with known / unknown μ) by looking at the gap between the MSE associated with each of the two predictors (MLP and CBP) and that corresponding to the optimal predictor. For this, we rely on the relative efficiency criterion RQE as in the linex application

$$RQE = \frac{E[\tilde{Q} - Q^*]}{E[\hat{Q} - Q^*]}, \tag{17}$$

where Q^* represents the evaluation criterion for the optimal predictor, $\tilde{Q}$ corresponds to the iterated predictor (MLE) and $\hat{Q}$ is the one for the direct predictor (CBP), respectively. Besides, the criterion Q is the opposite of the out-of-sample MSE loss, $\breve{Q} = -\sum_{t=T+1}^{2T}(y_t - \breve{y}_t)^2$, where $\breve{}$ denotes each of the three predictors and associated value of the criterion at a time.

To compare the relative efficiency of the two predictors (MLP and CBP), the following setup is considered for the Monte-Carlo simulations.

Step 1. We draw a vector of disturbances $\{\varepsilon\}_{t=1}^{2T}$ from a normal distribution with mean 0 and variance 1. Then we generate the AR(1) vector $y_t = \varphi y_{t-1} + \varepsilon_t$, where the initial value y_0 has been set to 0 and the autoregressive parameter $\varphi \in (-1, 1)$ to ensure the stationarity of the process. The first T observations constitute the in-sample data and are used to estimate the parameters of the models, whereas the other T observations serve for the out-of-sample forecasting exercise.

Step 2. The MLE, CBE and optimal estimator can now be determined by relying on the in-sample dataset. Recall that we consider the fixed forecasting scheme, so that the parameters are estimated only once, independent of the number of out-of-sample periods to forecast. We next compute the three predictors for each out-of-sample period by relying on eq.12 - eq.14 in the case where μ is known and on eq.12 and eq.15 - eq.16 if μ is estimated.

Step 3. Subsequently, the out-of-sample evaluation criterion is computed for each of the predictors (optimal, MLP and CBP).

Step 4. Repeat steps 1 to 3 a large number of times (100,000 and 500,000 simulations are run).

Step 5. We can now evaluate the out-of-sample performance of the predictors by relying on the relative criteria efficiency (RQE) indicator (eq. 17). Since y^* is the optimal predictor under the squared loss, the expected value of the evaluation criterion associated with y^* is always larger than the one corresponding to the two other predictors. It follows that both the numerator and denominator of are negative, so that a $RQE < 1$ indicates that the MLP, i.e. iterated predictor, performs better than the CBP, i.e. direct one, under the h-periods-ahead quadratic evaluation criterion. Conversely, $RQE > 1$ would support the choice of the criterion-based predictor over the ML one. Note also that several values have been considered for φ, so as to study the change in efficiency when the process approaches unit-root. Besides, we set the forecast horizon h to 2; 4 and 12 respectively.

4.1 Asymptotic Findings

Part i) of table 4 displays the forecast evaluation results for $n = 100,000$ and $\varphi \in \{0; 0.3; 0.8; 0.99\}$ both when the mean μ is known (Panel A.) and when it is estimated (Panel B.). The forecasting superiority of the iterated method with respect to the direct one in this setup has been emphasized theoretically in the literature (Bhansali, 1999; Ing et al., 2003). Nevertheless, the role of the autoregressive parameter in the evaluation has not been explicitly tackled, even though it deserves attention. First, notice that the larger φ, i.e. the higher the persistence of the process, the more the relative efficiency of the MLP with respect to CBP diminishes. As expected, when the constant must be estimated, the relative efficiency of the iterated predictor is lower than for known μ since the variance of the evaluation criterion rises. Actually two particular cases can be distinguished. First, recall that multi-period prediction errors have a moving average component in them. Hence, in the special case where $\varphi = 0$, this moving average component has a unit root which (in the model without a constant to be estimated, i.e. $\mu = 0$) causes $\tilde{Q} - Q^*$ to degenerate to zero at a fast rate. This explains why $\mathrm{E}(\tilde{Q} - Q^*) \approxeq 0$ in our simulation design. Naturally, $\varphi = 0$, is not an interesting case for multi-period ahead prediction in practice. Second, when the autoregressive parameter approaches near unit-root, i.e. $\varphi = 0.99$, the RQE advantages from using the iterated approach fade almost entirely. One intuition behind this is that when φ is near-integrated the iterated estimator losses in efficiency since its variance is approaching at a fast rate the variance of the direct estimator.

Moreover, an increase in the forecast loss (shrinkage in the evaluation criterion) adds to the reduction in relative efficiency as φ rises, the loss being more important when the constant is estimated (see columns 3-4 and 10-11). To put it another way, persistent processes seem to be more difficult to forecast accurately, especially if the number of estimated parameters increases.

At the same time, asymptotically the MLP and CBP are unbiased. We also note the high precision of the simulation results, with standard deviations are less than 10^{-4} for the evaluation criteria, and less than 10^{-6} for the bias of the estimators.

Now let us compare the short-run forecasting abilities of the models ($h = 2$) with the ones associated with longer horizons $h = 4$ (table 5) and 12 (table 6) respectively. First, it appears that the larger the forecast horizon, the more MLP is efficient relatively to the CBP, as already noted in the literature. Still, our simulation framework allows us to note several interesting facts. To be more precise, RQE gets closer and closer to 0 as h increases independent of the values of the autoregressive parameter, as long as φ does not approaches 1. In this particular near unit-root case, RQE always remains close to 1, emphasizing the fact that the behavior of the evaluation criterion changes in such circumstances. The improvement in RQE is nevertheless less significant when the constant is estimated (see Panel B in tables 5 and 6). Second, the forecast loss increases exponentially as h grows, underlying the difficulty to correctly forecast at long-horizons independent of the underlying model considered.

4.2 Finite-Sample Results

As aforementioned, large-sample properties of the two predictors have been the object of numerous studies. By contrast, to our knowledge only Bhansali (1997) presents small-sample results (in the particular case of AR(2) and ARMA(2,2) models) by relying on only 500 simulations. Besides, his framework is different than ours, as he studies the impact of selecting the order of the process for different models on the MSE associated with the direct and iterated forecasts, respectively.

In part *ii*) of table 4 we hence report the results for $n = 1,000$ and $n = 100$. One of our main findings is that the small sample results are consistent with the asymptotic findings, which means that matching estimation and evaluation criteria does not improve forecasting abilities in a setting where the other estimator is the MLE. Notice that the RQE seems to slightly improve when the sample-size is reduced, even though the per-observation evaluation criterion decreases (as the forecast loss rises). For this, recall that to compare results across the different sample-sizes the values must be rescaled by dividing by the number of simulations (as in the linex application) so as to obtain the per-observation evaluation criterion.

At the same time, the MLE and CBE exhibit small-sample bias. We stress the fact that the larger φ, i.e. the more persistent the process, the larger the bias. Besides, the bias increases with the shrinkage of the sample size and rises when the constant μ is estimated as opposed to the case where μ is known.

Tables 5 and 6 present the finite-sample results for $h = 4$ and 12. As in the case $h = 2$, RQE seems to improve with respect to large-samples. At the same time, the value of the evaluation criterion exponentially drops as the forecast horizon increases, while the estimation bias enlarges. These findings are particularly true when μ is estimated.

All in all, the MLP is proven to be relatively more efficient than the CBP with respect to the optimal predictor, even in small samples and when additional parameters are estimated. Most importantly, by acknowledging the fact that for persistent processes the relative gain of

MLE decreases while the bias increases, we recommend to pay more attention to the estimated autoregressive parameters in empirical applications that look at multi-step-ahead forecasting. Furthermore, gain in relative predictive ability in small-samples could result from using bias-corrected estimators (e.g. Roy-Fuller estimator (Roy and Fuller, 2001), bootstrap mean bias-corrected estimator (Kim, 2003), grid-bootstrap (Hansen, 1999), Andrews' estimator (Andrews, 1993; Andrews and Chen, 1994 for AR(p) processes). Indeed, more accurate forecasts seem to be obtained when comparing such estimators with the traditional ones by relying on the cumulated root-mean square error. It is particularly the case of persistent processes, i.e. near-unit-root, for which we have shown that the direct and iterated predictors perform on par (Kim, 2003; Kim and Durmaz, 2009). Further investigation into this issue would be interesting.

5 Conclusion

In this paper we address the question of whether the use of the same criterion in- and out-of-sample dominates other forms of estimation. Taking the case of M-estimators and using a second-order Taylor expansion, we show that the optimal out-of-sample performance is achieved through MLE. Most importantly, MLP can can be vastly better than CBP, whatever the out-of-sample criterion considered. Our theoretical result is analogous to the well known Cramer-Rao bound for in-sample estimation. We also discuss the case where the likelihood is misspecified, in particular the optimal transformation of the likelihood parameters into evaluation-criterion parameters.

In a context with an asymmetric (linex) loss function we show that the criterion based estimation performs on par with maximum likelihood when the loss is near-symmetric, whereas the MLE clearly dominates QBE with asymmetric loss. Most importantly, not only the asymptotic but also the finite-sample findings support these conclusions. In contrast, if the likelihood has the same number of parameters as the criterion-based predictor CBP (the other parameter being set to its true value), the gains from using MLE in forecasting relatively to CBE increase. Second, in the case of a well-identified gaussian linear AR(1) process it appears that MLP outperforms CBP both when the model is estimated with and without an intercept. The longer the forecasting horizon the better the MLP relatively to CBP. Still, the relative performance of MLP with respect to CBP plunges when the process is nearly integrated (the autoregressive coefficient is close to 1).

A Appendix: Joint variance-covariance matrix of transformed estimators

Recall that we want to prove that the joint variance-covariance matrix of the transformed estimators is given by

$$J_\gamma = \text{Avar}\begin{pmatrix} n^{1/2}(\hat{\gamma}_D - \gamma^*) \\ n^{1/2}(\tilde{\gamma}_{ML} - \gamma^*) \end{pmatrix} = \begin{pmatrix} V & I \\ I & I \end{pmatrix}.$$

The two blocks on the diagonal of J_γ are the variance-covariance matrices of the two estimators. We thus have to show that $\text{E}(\hat{\gamma}_D - \gamma)(\tilde{\gamma}_{ML} - \gamma)' = I$ to prove that the joint variance-covariance matrix of these estimators is indeed J_γ. To this aim, we rely on the fact that $-n^{-1}\tilde{H}(\gamma^*) = I$, so that $cov(\hat{\gamma}_D - \gamma^*, \tilde{\gamma}_{ML} - \gamma^*) \overset{a}{=} cov(\hat{\gamma}_D - \gamma^*, \tilde{S}(\gamma^*)')$. Moreover, since $\tilde{S}(\gamma^*)$ is the first derivative of the maximum-likelihood criterion, it can be written as $\frac{\partial logL}{\partial \gamma^*}$. Note also that by the definition of the MLE, $\text{E}(\frac{\partial logL}{\partial \gamma^*}) = 0$. Consequently,

$$\begin{aligned}
cov(\hat{\gamma}_D - \gamma^*, \tilde{S}(\gamma^*)') &= \text{E}((\hat{\gamma}_D - \gamma^*)\tilde{S}(\gamma^*)') \\
&= \text{E}(\hat{\gamma}_D \frac{\partial logL}{\partial \gamma^{*\prime}}) - \text{E}(\gamma^* \frac{\partial logL}{\partial \gamma^{*\prime}}) \\
&= \text{E}(\hat{\gamma}_D \frac{\partial L}{\partial \gamma^{*\prime}} \frac{1}{L}) \\
&= \int (\hat{\gamma}_D \frac{\partial L}{\partial \gamma^{*\prime}} \frac{1}{L} L \, d\mathcal{X}) \\
&= \int (\hat{\gamma}_D \frac{\partial L}{\partial \gamma^{*\prime}} \, d\mathcal{X}) \\
&= \frac{\partial}{\partial \gamma^{*\prime}} \int \hat{\gamma}_D L \, d\mathcal{X} \\
&= \frac{\partial}{\partial \gamma^{*\prime}} \text{E}(\hat{\gamma}_D) = I \text{ by interchanging the operations of integration and differentiation.}
\end{aligned}$$

For a theorem giving the sufficient conditions for the interchange of operations to hold see Amemiya (1985) page 17.

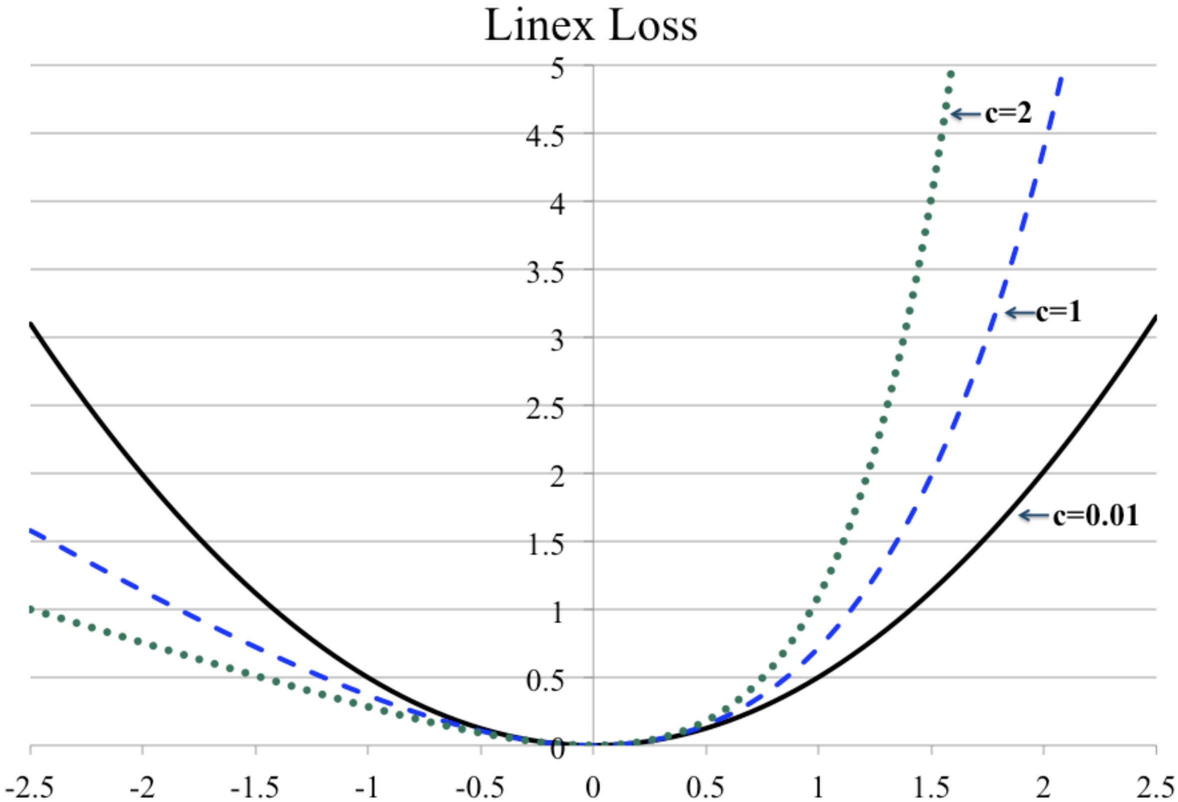

Figure 1: Linex

Table 1: Linex Loss

i) Asymptotic results

c	RQE	$\mathrm{E}[\tilde{Q}-Q^*]$	$\mathrm{E}[\hat{Q}-Q^*]$	$\mathrm{E}(x^*)$	$\mathrm{E}(\tilde{x}_{ML}-x^*)$	$\mathrm{E}(\hat{x}-x^*)$
0.01	1.00	-0.50	-0.50	0.005	0.000	0.000
0.1	1.00	-0.50	-0.50	0.050	0.000	0.000
1	0.87	-0.75	-0.86	0.500	0.000	0.000
1.5	0.56	-1.06	-1.88	0.750	0.000	0.000
2	0.23	-1.50	-6.64	1.000	0.000	0.000
3	0.01	-2.72	-224	1.500	0.000	-0.001

ii) Finite sample results: n=1,000

c	RQE	$\mathrm{E}[\tilde{Q}-Q^*]$	$\mathrm{E}[\hat{Q}-Q^*]$	$\mathrm{E}(x^*)$	$\mathrm{E}(\tilde{x}_{ML}-x^*)$	$\mathrm{E}(\hat{x}-x^*)$
0.01	1.00	-0.50	-0.50	0.005	0.000	0.000
0.1	1.00	-0.50	-0.50	0.050	0.000	0.000
1	0.88	-0.75	-0.85	0.500	0.000	-0.001
1.5	0.60	-1.07	-1.78	0.750	-0.001	-0.003
2	0.35	-1.51	-4.34	1.000	-0.001	-0.010
3	0.14	-2.76	-19.9	1.500	-0.001	-0.073

iii) Finite sample results: n=100

c	RQE	$\mathrm{E}[\tilde{Q}-Q^*]$	$\mathrm{E}[\hat{Q}-Q^*]$	$\mathrm{E}(x^*)$	$\mathrm{E}(\tilde{x}_{ML}-x^*)$	$\mathrm{E}(\hat{x}-x^*)$
0.01	1.00	-0.50	-0.50	0.005	0.000	0.000
0.1	1.00	-0.50	-0.50	0.050	0.000	0.000
1	0.90	-0.75	-0.84	0.500	-0.005	-0.009
1.5	0.72	-1.08	-1.49	0.750	-0.008	-0.023
2	0.56	-1.53	-2.75	1.000	-0.010	-0.058
3	0.28	-3.05	-10.9	1.500	-0.015	-0.215

Note: We compare the out-of-sample performance of the maximum-likelihood estimator (MLE) $\tilde{x}_{ML}$ with respect to that of the criterion-based estimator (CBE) $\hat{x}$ by looking at the ratio of the expected values of the gaps between the evaluation criterion for each of the two estimators ($\tilde{Q}$ and $\hat{Q}$) and that corresponding to the optimal estimator (Q^*) under linex loss, i.e. $RQE = \mathrm{E}[\tilde{Q}-Q^*]/\mathrm{E}[\hat{Q}-Q^*]$. The evaluation criterion is set to the opposite of the loss function. When $RQE < 1$ the MLE outperforms the CBE. The expected value of the optimal estimator $\mathrm{E}(x^*)$ as well as the expected bias of MLE and CBE, $\mathrm{E}(\tilde{x}_{ML}-x^*)$, and $\mathrm{E}(\hat{x}-x^*)$, are also included. Besides, the fixed forecasting scheme is used for estimation, where the estimation and evaluation samples have the same size, n. The results are presented for several levels of asymmetry c, different out-of-sample sizes n and have been obtained by performing 500,000 simulations in finite-samples and 100,000 in large samples.

Table 2: Linex Loss - Asymptotic results (with only one parameter estimated by both metods - ML and CB -)

i) Estimated mean ($\hat{\mu}$); Known variance (σ^2)

c	RQE	$\mathrm{E}[\tilde{Q}-Q^*]$	$\mathrm{E}[\hat{Q}-Q^*]$	$\mathrm{E}(x^*)$	$\mathrm{E}(\tilde{x}_{ML}-x^*)$	$\mathrm{E}(\hat{x}-x^*)$
0.01	1.00	-0.50	-0.50	0.005	0.000	0.000
0.1	0.99	-0.50	-0.50	0.050	0.000	0.000
1	0.58	-0.50	-0.87	0.500	0.000	0.000
1.5	0.26	-0.50	-1.88	0.750	0.000	0.000
2	0.08	-0.50	-6.67	1.000	0.000	0.000
3	0.00	-0.52	-219	1.500	0.000	-0.001

ii) Known mean (μ); Estimated variance ($\hat{\sigma}^2$)

c	RQE	$\mathrm{E}[\tilde{Q}-Q^*]$	$\mathrm{E}[\hat{Q}-Q^*]$	$\mathrm{E}(x^*)$	$\mathrm{E}(\tilde{x}_{ML}-x^*)$	$\mathrm{E}(\hat{x}-x^*)$
0.01	0.00	0.00	-0.50	0.005	0.000	0.000
0.1	0.00	0.00	-0.50	0.050	0.000	0.000
1	0.29	-0.25	-0.87	0.500	0.000	0.000
1.5	0.30	-0.56	-1.89	0.750	0.000	0.000
2	0.15	-0.97	-6.63	1.000	0.000	0.000
3	0.01	-2.44	-221	1.500	0.000	-0.001

Note: See note to table 1.

Table 3: Likelihood Misspecification

A. Normal Inverse Gaussian (0,1,0,3)

i) Asymptotic results

c	RQE	$\mathrm{E}[\tilde{Q}-Q^*]$	$\mathrm{E}[\hat{Q}-Q^*]$	$\mathrm{E}(x^*)$	$\mathrm{E}(\tilde{x}_{QML}-x^*)$	$\mathrm{E}(\hat{x}-x^*)$
0.01	1.000	-0.496	-0.496	0.005	0.000	0.000
0.1	1.005	-0.508	-0.505	0.050	0.000	0.000

ii) Finite sample results: n=1,000

c	RQE	$\mathrm{E}[\tilde{Q}-Q^*]$	$\mathrm{E}[\hat{Q}-Q^*]$	$\mathrm{E}(x^*)$	$\mathrm{E}(\tilde{x}_{QML}-x^*)$	$\mathrm{E}(\hat{x}-x^*)$
0.01	1.000	-0.497	-0.497	0.005	0.000	0.000
0.1	0.990	-0.508	-0.514	0.050	0.000	0.000

iii) Finite sample results: n=100

c	RQE	$\mathrm{E}[\tilde{Q}-Q^*]$	$\mathrm{E}[\hat{Q}-Q^*]$	$\mathrm{E}(x^*)$	$\mathrm{E}(\tilde{x}_{QML}-x^*)$	$\mathrm{E}(\hat{x}-x^*)$
0.01	1.000	-0.499	-0.500	0.005	0.000	0.000
0.1	0.990	-0.506	-0.511	0.050	-0.001	-0.001

B. Normal Inverse Gaussian (-20.47,46.78,-1,1.67)

i) Asymptotic results

c	RQE	$\mathrm{E}[\tilde{Q}-Q^*]$	$\mathrm{E}[\hat{Q}-Q^*]$	$\mathrm{E}(x^*)$	$\mathrm{E}(\tilde{x}_{QML}-x^*)$	$\mathrm{E}(\hat{x}-x^*)$
0.01	1.573	-34.803	-22.122	-20.24	0.005	0.000
0.1	5663	-85970	-15.179	-18.547	0.417	0.000

ii) Finite sample results: n=1,000

c	RQE	$\mathrm{E}[\tilde{Q}-Q^*]$	$\mathrm{E}[\hat{Q}-Q^*]$	$\mathrm{E}(x^*)$	$\mathrm{E}(\tilde{x}_{QML}-x^*)$	$\mathrm{E}(\hat{x}-x^*)$
0.01	0.998	-21.847	-21.885	-20.240	0.005	0.000
0.1	6.398	-101.744	-15.903	-18.547	0.415	-0.002

iii) Finite sample results: n=100

c	RQE	$\mathrm{E}[\tilde{Q}-Q^*]$	$\mathrm{E}[\hat{Q}-Q^*]$	$\mathrm{E}(x^*)$	$\mathrm{E}(\tilde{x}_{QML}-x^*)$	$\mathrm{E}(\hat{x}-x^*)$
0.01	0.998	-21.920	-21.967	-20.240	0.001	-0.004
0.1	1.531	-24.465	-15.980	-18.547	0.395	-0.015

Note: We compare the out-of-sample performance of the quasi-maximum-likelihood estimator (QMLE) $\tilde{x}_{QML}$ with respect to that of the criterion-based estimator (CBE) $\hat{x}$ when the true distribution is normal inverse gaussian with the first four moment mentioned between parentheses. We thus look at the ratio of the expected values of the gaps between the evaluation criterion for each of the two estimators ($\tilde{Q}$ and $\hat{Q}$) and that corresponding to the optimal estimator (Q^*) under linex loss, i.e. $RQE = \mathrm{E}[\tilde{Q}-Q^*]/\mathrm{E}[\hat{Q}-Q^*]$. For further details, see note to table1.

Table 4: Long-horizon Forecasting: AR(1) Model, horizon $h = 2$

	Panel A. AR(1) without mean						Panel B. AR(1) with estimated mean					
i) Asymptotic results												
φ	RQE	$E[\tilde{Q}-Q^*]$	$E[\hat{Q}-Q^*]$	$E(\varphi^{*h})$	$E(\tilde{\varphi}^h-\varphi^{*h})$	$E(\hat{\varphi}^h-\varphi^{*h})$	RQE	$E[\tilde{Q}-Q^*]$	$E[\hat{Q}-Q^*]$	$E(\varphi^{*h})$	$E(\tilde{\varphi}^h-\varphi^{*h})$	$E(\hat{\varphi}^h-\varphi^{*h})$
0.00	0.00	0.00	-1.00	0.000	0.000	0.000	0.50	-1.00	-1.99	0.000	0.000	0.000
0.30	0.28	-0.36	-1.27	0.090	0.000	0.000	0.69	-2.05	-2.95	0.090	0.000	0.000
0.80	0.88	-2.57	-2.93	0.640	0.000	0.000	0.94	-5.81	-6.17	0.640	0.000	0.000
0.99	0.99	-3.95	-3.97	0.980	0.000	0.000	1.00	-8.04	-8.06	0.980	0.000	0.000
ii) Finite sample results: n=1,000												
φ	RQE	$E[\tilde{Q}-Q^*]$	$E[\hat{Q}-Q^*]$	$E(\varphi^{*h})$	$E(\tilde{\varphi}^h-\varphi^{*h})$	$E(\hat{\varphi}^h-\varphi^{*h})$	RQE	$E[\tilde{Q}-Q^*]$	$E[\hat{Q}-Q^*]$	$E(\varphi^{*h})$	$E(\tilde{\varphi}^h-\varphi^{*h})$	$E(\hat{\varphi}^h-\varphi^{*h})$
0.00	0.00	0.00	-1.00	0.000	0.001	0.000	0.50	-1.01	-2.00	0.000	0.001	-0.001
0.30	0.28	-0.36	-1.27	0.090	0.001	0.000	0.69	-2.05	-2.97	0.090	0.000	-0.002
0.80	0.87	-2.58	-2.97	0.640	-0.002	-0.003	0.93	-6.01	-6.45	0.640	-0.005	-0.006
0.99	0.99	-6.25	-6.31	0.980	-0.004	-0.004	0.99	-19.62	-19.82	0.980	-0.009	-0.009
iii) Finite sample results: n=100												
φ	RQE	$E[\tilde{Q}-Q^*]$	$E[\hat{Q}-Q^*]$	$E(\varphi^{*h})$	$E(\tilde{\varphi}^h-\varphi^{*h})$	$E(\hat{\varphi}^h-\varphi^{*h})$	RQE	$E[\tilde{Q}-Q^*]$	$E[\hat{Q}-Q^*]$	$E(\varphi^{*h})$	$E(\tilde{\varphi}^h-\varphi^{*h})$	$E(\hat{\varphi}^h-\varphi^{*h})$
0.00	0.03	-0.03	-1.00	0.000	0.010	0.000	0.49	-1.02	-2.06	0.000	0.010	-0.010
0.30	0.27	-0.35	-1.26	0.090	0.006	-0.003	0.65	-2.03	-3.11	0.090	-0.002	-0.021
0.80	0.83	-2.75	-3.30	0.640	-0.021	-0.025	0.87	-7.91	-9.10	0.640	-0.051	-0.060
0.99	0.95	-16.98	-17.85	0.980	-0.034	-0.035	0.93	-61.83	-66.33	0.980	-0.101	-0.105

Note: We compare the h-step-ahead out-of-sample performance of the maximum-likelihood estimator (MLE) $\tilde{\varphi}$ with respect to that of the criterion-based estimator (CBE) $\hat{\varphi}$ by looking at the ratio of the expected values of the gaps between the evaluation criterion for each of the two estimators ($\tilde{Q}$ and $\hat{Q}$) and that corresponding to the optimal estimator (Q^*) for a correctly specified AR(1) model with normal disturbances, i.e. $RQE = E[\tilde{Q}-Q^*]/E[\hat{Q}-Q^*]$. The evaluation criterion is set to the opposite of the MSE loss function. When $RQE < 1$ the MLE outperforms the CBE. The expected value of the optimal estimator $E(\varphi^{*h})$ as well as the expected bias of MLE and CBE, $E(\tilde{\varphi}^h-\varphi^{*h})$ and $E(\hat{\varphi}^{\mathrm{h}}-\varphi^{*\mathrm{h}})$, are also included. Besides, the fixed forecasting scheme is used for estimation, where the estimation and evaluation samples have the same size, n. Two cases are considered: an AR(1) process whose mean is known to be equal to 0 (see Panel A.) and an AR(1) model with unknown (estimated) mean (see Panel B.). Moreover, the results are presented for several levels of persistence of the autoregressive process φ, different out-of-sample sizes n and have been obtained by performing 500,000 simulations in finite-samples and 100,000 in large samples.

Table 5: Long-horizon Forecasting: AR(1) Model, horizon $h = 4$

	Panel A. AR(1) without mean						Panel B. AR(1) with estimated mean					
i) Asymptotic results												
φ	RQE	$E[\tilde{Q}-Q^*]$	$E[\hat{Q}-Q^*]$	$E(\varphi^{*h})$	$E(\tilde{\varphi}^h-\varphi^{*h})$	$E(\hat{\varphi}^h-\varphi^{*h})$	RQE	$E[\tilde{Q}-Q^*]$	$E[\hat{Q}-Q^*]$	$E(\varphi^{*h})$	$E(\tilde{\varphi}^h-\varphi^{*h})$	$E(\hat{\varphi}^h-\varphi^{*h})$
0	0.00	0.00	-0.98	0.000	0.000	0.000	0.50	-1.00	-2.00	0.000	0.000	0.000
0.3	0.01	-0.01	-1.31	0.008	0.000	0.000	0.61	-2.02	-3.34	0.008	0.000	0.000
0.8	0.62	-4.19	-6.78	0.410	0.000	0.000	0.83	-13.0	-15.6	0.410	0.000	0.000
0.99	0.98	-15.2	-15.5	0.961	0.000	0.000	0.99	-30.8	-31.1	0.961	0.000	0.000
ii) Finite sample results: n=1,000												
φ	RQE	$E[\tilde{Q}-Q^*]$	$E[\hat{Q}-Q^*]$	$E(\varphi^{*h})$	$E(\tilde{\varphi}^h-\varphi^{*h})$	$E(\hat{\varphi}^h-\varphi^{*h})$	RQE	$E[\tilde{Q}-Q^*]$	$E[\hat{Q}-Q^*]$	$E(\varphi^{*h})$	$E(\tilde{\varphi}^h-\varphi^{*h})$	$E(\hat{\varphi}^h-\varphi^{*h})$
0	0.00	0.00	-1.00	0.000	0.000	0.000	0.50	-1.01	-2.01	0.000	0.000	-0.001
0.3	0.01	-0.01	-1.31	0.008	0.000	0.000	0.61	-2.02	-3.34	0.008	0.000	-0.002
0.8	0.61	-4.15	-6.84	0.410	-0.002	-0.003	0.82	-13.2	-16.2	0.410	-0.006	-0.009
0.99	0.97	-23.6	-24.4	0.961	-0.007	-0.008	0.97	-74.3	-76.7	0.961	-0.017	-0.018
iii) Finite sample results: n=100												
φ	RQE	$E[\tilde{Q}-Q^*]$	$E[\hat{Q}-Q^*]$	$E(\varphi^{*h})$	$E(\tilde{\varphi}^h-\varphi^{*h})$	$E(\hat{\varphi}^h-\varphi^{*h})$	RQE	$E[\tilde{Q}-Q^*]$	$E[\hat{Q}-Q^*]$	$E(\varphi^{*h})$	$E(\tilde{\varphi}^h-\varphi^{*h})$	$E(\hat{\varphi}^h-\varphi^{*h})$
0	0.00	0.00	-1.02	0.000	0.000	0.000	0.48	-1.01	-2.11	0.000	0.000	-0.010
0.3	0.02	-0.02	-1.32	0.008	0.004	-0.001	0.58	-2.02	-3.49	0.008	0.003	-0.019
0.8	0.55	-3.81	-6.97	0.410	-0.017	-0.031	0.70	-14.7	-21.2	0.410	-0.053	-0.088
0.99	0.86	-54.9	-63.6	0.961	-0.063	-0.067	0.82	-187	-229	0.961	-0.181	-0.198

Note: See note to 4.

Table 6: Long-horizon Forecasting: AR(1) Model, horizon $h = 12$

	Panel A. AR(1) without mean						Panel B. AR(1) with estimated mean					
i) Asymptotic results												
φ	RQE	$E[\tilde{Q} - Q^*]$	$E[\hat{Q} - Q^*]$	$E(\varphi^{*h})$	$E(\tilde{\varphi}^h - \varphi^{*h})$	$E(\hat{\varphi}^h - \varphi^{*h})$	RQE	$E[\tilde{Q} - Q^*]$	$E[\hat{Q} - Q^*]$	$E(\varphi^{*h})$	$E(\tilde{\varphi}^h - \varphi^{*h})$	$E(\hat{\varphi}^h - \varphi^{*h})$
0	0.00	0.00	-1.01	0.000	0.000	0.000	0.50	-1.00	-2.01	0.000	0.000	0.000
0.3	0.00	0.00	-1.31	0.000	0.000	0.000	0.61	-2.06	-3.38	0.000	0.000	0.000
0.8	0.08	-1.03	-12.3	0.069	0.000	0.000	0.67	-22.7	-33.9	0.069	0.000	0.000
0.99	0.93	-115	-124	0.886	0.000	0.000	0.96	-247	-256	0.886	0.000	0.000
ii) Finite sample results: n=1,000												
φ	RQE	$E[\tilde{Q} - Q^*]$	$E[\hat{Q} - Q^*]$	$E(\varphi^{*h})$	$E(\tilde{\varphi}^h - \varphi^{*h})$	$E(\hat{\varphi}^h - \varphi^{*h})$	RQE	$E[\tilde{Q} - Q^*]$	$E[\hat{Q} - Q^*]$	$E(\varphi^{*h})$	$E(\tilde{\varphi}^h - \varphi^{*h})$	$E(\hat{\varphi}^h - \varphi^{*h})$
0	0.00	0.00	-1.01	0.000	0.000	0.000	0.49	-1.00	-2.02	0.000	0.000	-0.001
0.3	0.00	0.00	-1.32	0.000	0.000	0.000	0.60	-2.03	-3.39	0.000	0.000	-0.002
0.8	0.09	-1.06	-12.0	0.069	0.001	-0.002	0.65	-22.7	-34.7	0.069	-0.001	-0.010
0.99	0.88	-164	-186	0.886	-0.019	-0.021	0.89	-531	-598	0.886	-0.045	-0.049
iii) Finite sample results: n=100												
φ	RQE	$E[\tilde{Q} - Q^*]$	$E[\hat{Q} - Q^*]$	$E(\varphi^{*h})$	$E(\tilde{\varphi}^h - \varphi^{*h})$	$E(\hat{\varphi}^h - \varphi^{*h})$	RQE	$E[\tilde{Q} - Q^*]$	$E[\hat{Q} - Q^*]$	$E(\varphi^{*h})$	$E(\tilde{\varphi}^h - \varphi^{*h})$	$E(\hat{\varphi}^h - \varphi^{*h})$
0	0.00	0.00	-1.11	0.000	0.000	0.000	0.44	-1.01	-2.31	0.000	0.000	-0.010
0.3	0.00	0.00	-1.45	0.000	0.000	0.000	0.53	-2.05	-3.86	0.000	0.000	-0.018
0.8	0.09	-1.04	-11.4	0.069	0.008	-0.013	0.53	-21.6	-40.6	0.069	-0.009	-0.098
0.99	0.63	-242	-383	0.886	-0.131	-0.156	0.55	-679	-1234	0.886	-0.371	-0.487

Note: See note to 4.

References

Akaike, H. (1974), ‘A new look at the statistical model identification’, *IEEE transactions on automatic control* **19**, 716–723.

Amemiya, T. (1985), *Advanced Econometrics*, Cambridge: Harvard University Press.

Andrews, D. (1993), ‘Exactly median-unbiased estimation of first order autoregressive/unit root models’, *Econometrica: Journal of the Econometric Society* pp. 139–165.

Andrews, D. and Chen, H. (1994), ‘Approximately median-unbiased estimation of autoregressive models’, *Journal of Business & Economic Statistics* pp. 187–204.

Bhansali, R. (1997), ‘Direct autoregressive predictors for multistep prediction: Order selection and performance relative to the plug in predictors’, *Statistica Sinica* **7**, 425–450.

Bhansali, R. (1999), *Parameter estimation and model selection for multistep prediction of time series: a review.*, 1 edn, CRC Press, pp. 201–225.

Christoffersen, P. and Diebold, F. (1997), ‘Optimal prediction under asymmetric loss’, *Econometric Theory* **13**, 808–817.

Clark, T. E. and McCracken, M. W. (2001), ‘Tests of equal forecast accuracy and encompassing for nested models’, *Journal of Econometrics* **105**, 85–110.

Clements, M. and Hendry, D. (2005), ‘Evaluating a model by forecast performance*’, *Oxford Bulletin of Economics and Statistics* **67**, 931–956.

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

Granger, C. (1969), ‘Prediction with a generalized cost of error function’, *OR* pp. 199–207.

Granger, C. (1986), *Forecasting Economic Time Series*, Academic Press.

Hansen, B. (1999), ‘The grid bootstrap and the autoregressive model’, *Review of Economics and Statistics* **81**(4), 594–607.

Hansen, P. R. (2010), ‘A winnerís curse for econometric models: On the joint distribution of in-sample fit and out-of-sample fit and its implications for model selection’, *working paper* .

Huber, P. (1981), *Robust Statistics*, Wiley.

Hwang, S., Knight, J. and Satchell, S. (2001), ‘Forecasting nonlinear functions of returns using linex loss functions’, *annals of economics and finance* **2**(1), 187–213.

Ing, C. et al. (2003), ‘Multistep prediction in autoregressive processes’, *Econometric Theory* **19**(2), 254–279.

Kabaila, P. (1981), 'Estimation based on one step ahead prediction versus estimation based on multi-step ahead prediction', *Stochastics: An International Journal of Probability and Stochastic Processes* **6**(1), 43–55.

Kim, H. and Durmaz, N. (2009), Bias correction and out-of-sample forecast accuracy. Manuscript, Auburn University.

Kim, J. (2003), 'Forecasting autoregressive time series with bias-corrected parameter estimators', *International Journal of Forecasting* **19**(3), 493–502.

Klein, L. (1992), The test of a model is its ability to predict. Manuscript, University of Pennsylvania.

McCracken, M. (2004), Asymptotics for out-of-sample tests of causality. Manuscript, Louisiana State University.

Roy, A. and Fuller, W. (2001), 'Estimation for autoregressive time series with a root near 1', *Journal of Business and Economic Statistics* **19**(4), 482–493.

Schorfheide, F. (2005), 'Var forecasting under misspecification', *Journal of Econometrics* **128**(1), 99–136.

Shibata, R. (1980), 'Asymptotically efficient selection of the order of the model for estimating parameters of a linear process', *The Annals of Statistics* pp. 147–164.

Takeuchi, K. (1976), 'Distribution of informational statistics and a criterion of model fitting', *Suri-Kagaku (Mathematical Sciences)* **153**, 12–18. (In Japanese).

Varian, H. (1974), *A Bayesian Approach to Real Estate Assessment*, North-Holland, pp. 195–208.

Weiss, A. (1996), 'Estimating time series models using the relevant cost function', *Journal of Applied Econometrics* **11**(5), 539–560.

Weiss, A. and Andersen, A. (1984), 'Estimating time series models using the relevant forecast evaluation criterion', *Journal of the Royal Statistical Society. Series A (General)* pp. 484–487.

West, K. (2001), 'Tests for forecasts encompassing when forecasts depend on estimated regression parameters', *Journal of Business and Economic Statistics* **19**(1), 29–33.

West, K. (2006), 'Forecast evaluation', *Handbook of economic forecasting* **1**, 99–134.

West, K. D. (1996), 'Asymptotic inference about predictive ability', *Econometrica* **64**, 1067–1084.

White, H. (1994), *Estimation, Inference and Specification Analysis*, Cambridge University Press, Cambridge.

Zellner, A. (1986), 'Bayesian estimation and prediction using asymmetric loss functions', *Journal of the American Statistical Association* pp. 446–451.