Parameter Estimation with Out-of-Sample Objective

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

[a] *European University Institute & CREATES**

[b] *Paris-Ouest Nanterre la Défense University*

Preliminary/December 2, 2013

Abstract

We study parameter estimation from the sample $\mathcal{X}$, when the objective is to maximize the expected value of a criterion function, Q, for a distinct sample, $\mathcal{Y}$. This is the situation that arises in forecasting problems and whenever an estimated model is to be applied to a draw from the general population. A natural candidate for solving $\max_{T\in\sigma(\mathcal{X})} \mathrm{E}Q(\mathcal{Y},T)$, is the *innate* estimator, $\hat{\theta} = \arg\max_\theta Q(\mathcal{X},\theta)$. While the innate estimator has certain advantages, we show, under suitable regularity conditions, that the asymptotically efficient estimator takes the form $\tilde{\theta} = \arg\max_\theta \tilde{Q}(\mathcal{X},\theta)$, where $\tilde{Q}$ is defined from a likelihood function in conjunction with Q. The likelihood-based estimator is, however, fragile, as misspecification is harmful in two ways. First, the likelihood-based estimator may be inefficient under misspecification. Second, and more importantly, the likelihood approach requires a parameter transformation that depends on the truth, causing an improper mapping to be used under misspecification. The theoretical results are illustrated with two applications, one involving a Gaussian likelihood and an asymmetric loss function; and another is the problem of making multi-period ahead forecasts.

Keywords: Forecasting, Out-of-Sample, LinEx Loss, Long-horizon forecasting.

JEL Classification: C52

*We thank Valentina Corradi, Nour Meddahi, Barbara Rossi, and Mark Watson for helpful comments. The first author acknowledges support from CREATES - Center for Research in Econometric Analysis of Time Series (DNRF78), funded by the Danish National Research Foundation.

1 Introduction

Efficient parameter estimation is a well explored topic. For instance, an estimator $T(\mathcal{X})$ is said to be efficient for θ, if it minimizes the expected loss, $\mathrm{E}[L(T(\mathcal{X}),\theta)]$, where L is a loss function and $\mathcal{X}$ is the random sample that is available for estimation. In this paper, we consider parameter estimation with a different objective. Our objective is characterized by the intended use of the estimated "model" that involves a second random sample, $\mathcal{Y}$, which is distinct from that sample used for estimation, $\mathcal{X}$. This is the structure that emerges in forecasting problems where $\mathcal{Y}$ represents future data and $\mathcal{X}$ is the sample available for estimation. In the context of forecasting it is standard convention to refer to $\mathcal{X}$ and $\mathcal{Y}$ as *in-sample* and *out-of-sample*, respectively. Our framework is, however, not specific to forecasting problems. The sample, $\mathcal{Y}$, can also represent a random draw from the general population, for which an estimated model is to be used. For instance, based on a pilot study (based on $\mathcal{X}$), one may seek to optimize tuning parameters in a policy program (e.g. a job training program), before the program is implemented more widely (to $\mathcal{Y}$).

To fix ideas: Let the objective be $\max_{T\in\sigma(\mathcal{X})}\mathrm{E}Q(\mathcal{Y},T)$, where Q is a criterion function. A natural candidate is the extremum estimator, $\hat{\theta}=\arg\max_\theta Q(\mathcal{X},\theta)$, which we label the *innate estimator*, because it is deduced directly from Q. While the innate estimator seeks to maximize the objective, Q, it need not be efficient and a better estimator may be available. To study this problem we consider a class of extremum estimators, where a typical element is given by, $\tilde{\theta}=\arg\max_\theta \tilde{Q}(\mathcal{X},\theta)$, where $\tilde{Q}$ is another criterion.

While it may seem unnatural to estimate parameters using a criterion, $\tilde{Q}$, that differs from that of the actual objective, Q, this approach is quite common in practice. This is sometimes done out of convenience,[1] but, as we will show, a carefully crafted $\tilde{Q}$ will produce the asymptotically efficient estimator. The use of a different criterion, $\tilde{Q}$, for estimation has many pitfalls, and the asymptotic efficiency hinges on additional assumption.

We shall establish results in an asymptotic framework, that are based on conventional assumptions made in the context of M-estimation. While our framework and objective differs for that usually used to study efficient parameter estimation, the classical structure emerges after manipulating the asymptotic expressions. The enables us to make use of the Cramer-Rao lower bound to establish a likelihood-based estimator as the asymptotically efficient estimator, albeit new and important issues arises in the case where the likelihood is misspecified. Under correct specification, the likelihood-based estimator dominates the innate estimator, sometimes by a wide margin. When the likelihood is misspecified,

[1] For instance, estimation by simple regression analysis although the the objective may be predictions of Value-at-Risk.

the asymptotic efficiency argument perish but, more importantly, the likelihood approach requires a mapping of likelihood parameters to criterion parameters that hinges on the likelihood being correctly specified. Under misspecification, this mapping becomes improper and causes $\tilde{\theta}$ to be inconsistent for the value of θ that maximizes Q.

In the context of forecasting, many have argued for the estimation criterion to be synchronized with the actual objection, starting with Granger (1969), see also Weiss (1996). For empirically support of a synchronized approach, see Weiss and Andersen (1984) and Christoffersen et al. (2001), where the objective in the latter is option pricing. In the autoregressive setting with quadratic prediction loss Bhansali (1999) and Ing (2003) have established the relative merits of the estimation methods, direct and plug-in, depends on the degree of misspecification. This led Schorfheide (2005) to propose a model selection criterion that targets trade-off between variance and bias, that the choice of estimation method entails.

The existing literature has primarily focused on the case with a mean square error loss function and and likelihood functions based on a Gaussian specifications. In this paper, we establish results for the case where, both Q and $\tilde{Q}$, belong to a general class of criteria that are suitable for M-estimation, see e.g. Huber (1981) and Amemiya (1985). This generality comes at the expense of results being asymptotic in nature. Specifically, we will compare the relative merits of estimators in terms of the limit distributions that arise in this context. The theoretical result will be complemented by two applications that also considers finite sample properties and, interestingly, cases with local misspecification of various forms.

First we make the simple observation that a discrepancy between Q and $\tilde{Q}$ can seriously degrade the performance.

Second, we show that the asymptotically optimal estimator is an estimator that is deduced from the maximum likelihood estimator. This theoretical result is analogous to the well known Cramer-Rao bound for in-sample estimation. We address the case where the likelihood function involves a parameter of higher dimension than θ, 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. The innate estimator performs on par with the likelihood-based estimator (LBE) when the loss is near-symmetric, whereas the LBE clearly dominates the innate estimator under asymmetric loss. In contrast, when the likelihood is misspecified the LBE suffers and its performance drops considerably with the degree of misspecification.

A second application pertains to long-horizon forecasting, where two competing forecasting methods are known as the *direct* and the *iterated* forecasts. The latter is also known as the plug-in forecasts. The

direct and iterated forecasts can be related to the innate and likelihood-based estimators, respectively. Well known results for direct and iterated forecasts in the context with an autoregressive model and MSE loss, are emerges as special cases in our framework. We contribute to this literature by, considering a case with asymmetric loss and derive results for the case with correct specification and the case with local misspecification. The asymmetry exacerbates the advantages of iterated forecasts (the likelihood approach), so that it take a relatively higher degree of misspecification for the direct forecast to be competitive. This casts light on the two approaches to multi-period forecasting.

The rest of the paper is structured as follows. Section 2 presents the theoretical framework an asymptotic results. Sections 3 and 4 present the two applications to asymmetric loss function and multi-period forecasting. Section 4 concludes and the appendix collects the mathematical proofs.

2 Theoretical Framework

We will compare the merits of the innate estimator $\hat{\theta}$ to a generic alternative estimator $\tilde{\theta}$. This is done within the theoretical framework of M-estimators, see Huber (1981), Amemiya (1985), and White (1994). Our exposition and notation will largely follow that in Hansen (2010).

The criterion functions take the form

$$Q(\mathcal{X},\theta)=\sum_{t=1}^{n} q(\mathbf{x}_t,\theta) \qquad \text{and} \qquad \tilde{Q}(\mathcal{X},\theta)=\sum_{t=1}^{n} \tilde{q}(\mathbf{x}_t,\theta),$$

with $\mathbf{x}_t=(X_t,\ldots,X_{t-k})$ for some k. This framework includes criteria deduced from Markovian models. For instance, least squares estimation of an AR(1) model, $X_t=\varphi X_{t-1}+\varepsilon_t$, would translate into $\mathbf{x}_t=(X_t,X_{t-1})$ and $\tilde{q}(\mathbf{x}_t,\theta)=-(X_t-\varphi X_{t-1})^2$.

Assumption 1. *Suppose that* $\{X_t\}$ *is stationary and ergodic, and that* $\mathrm{E}|q(\mathbf{x}_t,\theta)|<\infty$ *and* $\mathrm{E}|\tilde{q}(\mathbf{x}_t,\theta)|<\infty$.

The assumed stationarity carries over to $q(\mathbf{x}_t,\theta)$ and $\tilde{q}(\mathbf{x}_t,\theta)$, and their derivatives that we introduce below. Next we make some regularity assumptions about the criteria functions.

Assumption 2. *(i) The criteria functions* $q(\mathbf{x}_t;\theta)$ *and* $\tilde{q}(\mathbf{x}_t;\theta)$ *are continuous in* θ *for all* $\mathbf{x}_t$ *and measurable for all* $\theta\in\Theta$*, where* Θ *is compact. (ii)* θ_* *and* θ_0 *are the unique maximizers of* $\mathrm{E}[q(\mathbf{x}_t,\theta)]$ *and* $\mathrm{E}[\tilde{q}(\mathbf{x}_t,\theta)]$*, respectively, where* θ_* *and* θ_0 *are interior to* Θ*; (iii)* $\mathrm{E}[\sup_{\theta\in\Theta}|q(\mathbf{x}_t,\theta)|]<\infty$ *and* $\mathrm{E}[\sup_{\theta\in\Theta}|\tilde{q}(\mathbf{x}_t,\theta)|]<\infty$*;*

The assumed stationarity and Assumption 2 ensure that the θ that maximizes $\mathrm{E}[Q(\mathcal{X},\theta)]$ is unique, invariant to the sample, and identical to $\theta_* = \arg\max_\theta \mathrm{E}[q(\mathbf{x}_t,\theta)]$. Similarly for $\tilde{Q}$ and θ_0.

The following consistency follows from the literature on M-estimators.

Lemma 1. *The extremum estimators* $\hat{\theta} = \arg\max_{\theta\in\Theta} \sum_{t=1}^n q(\mathbf{x}_t,\theta)$ *and* $\tilde{\theta} = \arg\max_{\theta\in\Theta} \sum_{t=1}^n \tilde{q}(\mathbf{x}_t,\theta)$ *converge in probability to* θ_* *and* θ_0*, respectively, as* $n\to\infty$*.*

Because the innate estimator (as its label suggests) is intrinsic to the criterion Q, it will be consistent for θ_* under standard regularity conditions, in the sense that $\hat{\theta} \xrightarrow{p} \theta_*$ as the in-sample size increases. This consistency need not be satisfied by alternative estimators, including $\tilde{\theta}$.

Next we assume the following regularity conditions that enable us to derive the limit results that will be the basis for our main results. These conditions are also standard in the literature on M-estimation.

Assumption 3. *The criteria,* q *and* $\tilde{q}$*, are twice continuously differentiable in* θ*, where (i) the first derivatives,* $s(\mathbf{x}_t,\theta)$ *and* $\tilde{s}(\mathbf{x}_t,\theta)$*, satisfy a central limit theorem,* $n^{1/2}\sum_{t=1}^n (s(\mathbf{x}_t,\theta_*)', \tilde{s}(\mathbf{x}_t,\theta_0)')' \xrightarrow{d} N(0,\Sigma_S)$*; (ii) the second derivatives,* $h(\mathbf{x}_t,\theta)$ *and* $\tilde{h}(\mathbf{x}_t,\theta)$*, are uniformly integrable in a neighborhood of* θ_* *and* θ_0*, respectively, where the matrices* $A = -\mathrm{E}h(\mathbf{x}_t,\theta_*)$ *and* $\tilde{A} = -\mathrm{E}\tilde{h}(\mathbf{x}_t,\theta_0)$ *are invertible.*

Let B and $\tilde{B}$ denote the long-run variances of $s(\mathbf{x}_t,\theta_*)$ and $\tilde{s}(\mathbf{x}_t,\theta_0)$, respectively. Then, Σ_S, will have a block diagonal structure, with B and $\tilde{B}$ as diagonal blocks. There is no need to introduce a notation for the off-diagonal blocks in Σ_S, as they are immaterial to subsequent results.

The following result establishes an asymptotic independence between the in-sample scores and out-of-sample scores, which is useful for the computation of conditional expectations in the limit distribution.

Lemma 2. *We have*

$$n^{1/2}\left(\sum_{t=1}^n s(\mathbf{x}_t,\theta_*)', \sum_{t=1}^n \tilde{s}(\mathbf{x}_t,\theta_0)', \sum_{t=n+1}^{2n} s(\mathbf{x}_t,\theta_*)'\right)' \xrightarrow{d} N(0, \begin{pmatrix} \Sigma_S & 0 \\ 0 & B \end{pmatrix}).$$

Proof. For simplicity write $s_t = s(\mathbf{x}_t,\theta_*)$ and similarly for $\tilde{s}_t$. By Assumption 3, the asymptotic variance of $(2n)^{1/2}\left(\sum_{t=1}^{2n} s_t', \sum_{t=1}^{2n} \tilde{s}_t'\right)'$ is Σ_S. Now use the simple identity for the variance of a sum to deduce that the asymptotic covariance of $n^{1/2}\sum_{t=1}^n s_t$ and $n^{1/2}\sum_{t=n+1}^{2n} s_t$ is zero. The same argument can be applied to establish the (zero) asymptotic covariance between $n^{1/2}\sum_{t=1}^n \tilde{s}_t$ and $n^{1/2}\sum_{t=n+1}^{2n} s_t$. □

In this literature is is often assumed that $\mathcal{X}$ and $\mathcal{Y}$ are independent, see e.g. Schorfheide (2005, assumption 4), which implies independence of the score that relate to $\mathcal{X}$ and the score that relate to $\mathcal{Y}$. The Lemma shows that the asymptotic independence of the scores is a simple consequence of the central

limit theorem being applicable, and no further assumption is needed. The asymptotic independence holds whether the scores are martingale difference sequences or, more generally, serially dependent as is the case in some cases. To simplify the exposition we focus on the case where the sample size of $\mathcal{Y} = (\mathbf{x}_{n+1}, \ldots, \mathbf{x}_{2n})$, coincides with that of $\mathcal{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$.[2]

Definition 1. Two criteria, Q and $\tilde{Q}$, are said to be coherent if $\theta_* = \theta_0$, otherwise the criteria are said to be incoherent. Similarly, we refer to an estimator as being coherent for Q if its probability limit is θ_*.

Next, we state the fairly obvious result that an incoherent criterion will lead to inferior performance.

Lemma 3. *Consider an alternative estimator, $\tilde{\theta}$, deduced from an incoherent criterion, so that $\tilde{\theta} \xrightarrow{p} \theta_0 \neq \theta_*$. Then*

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

in probability. The divergence is at rate n.

Proof. Since $\tilde{\theta} \xrightarrow{p} \theta_0$ it follows by Assumptions 1 and 2 that $n^{-1} \sum_{t=1}^{n} q(\mathbf{x}_t, \tilde{\theta}) \xrightarrow{p} \mathrm{E}[q(\mathbf{x}_t, \theta_0)]$, which is strictly smaller than $\mathrm{E}[q(\mathbf{x}_t, \theta_*)]$, as a consequence of Assumption 2.*ii*. □

The results shows that any incoherent estimator will be inferior to the innate estimator. This shows that consistency for θ_* is a critical requirement, which limits the choice of criteria, $\tilde{Q}$, to be used for estimation. It is, however, possible to craft a coherent criterion, $\tilde{Q}$, from a likelihood function, as we shall show below.

Theorem 1. *Let Assumptions 1-3 be satisfied and suppose that $\tilde{Q}$ is a coherent criterion. Then*

$$Q(\mathcal{Y}, \tilde{\theta}) - Q(\mathcal{Y}, \theta_0) \xrightarrow{d} Z_y' B^{1/2} \tilde{A}^{-1} \tilde{B}^{1/2} Z_x - \frac{1}{2} Z_x' \tilde{B}^{1/2} \tilde{A}^{-1} A \tilde{A}^{-1} \tilde{B}^{1/2} Z_x,$$

where $Z_x, Z_y \sim iidN(0, I)$, and the expected value of the limit distribution is:

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

Interestingly, for the case with the innate estimator, the expected value of the limit distribution

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

[2]This setup is quite common in this literature, and was, for instance, used in Akaike (1974) to derive the Akaike's Information Criterion.

can be related to a result by Takeuchi (1976), who generalized the result by Akaike (1974) to the case with misspecified models.

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. Since $\tilde{Q}$ is coherent, we have $\tilde{\theta} \xrightarrow{p} \theta_0 = \theta_*$, and by a Taylor expansion we have

$$Q(\mathcal{Y},\tilde{\theta}) - Q(\mathcal{Y},\theta_0) = S_y(\theta_0)'(\tilde{\theta}-\theta_0) + \frac{1}{2}(\tilde{\theta}-\theta_0)'H_y(\theta_0)(\tilde{\theta}-\theta_0) + o_p(1).$$

By Assumption 3 and Lemma 2 we have that $n^{-1}H(\mathcal{Y},\theta_*) \xrightarrow{p} -A$, $n^{-1}\tilde{H}(\mathcal{X},\theta_*) \xrightarrow{p} -\tilde{A}$, and $\{\tilde{S}(\mathcal{X},\theta), S(\mathcal{Y},\theta)\} \xrightarrow{d} \{\tilde{B}^{1/2}Z_x, B^{1/2}Z_y\}$ where Z_x and Z_y are independent and both distributed $N(0,I)$. The result $Q(\mathcal{Y},\tilde{\theta}) - Q(\mathcal{Y},\theta) \xrightarrow{d} Z_y'B^{1/2}\tilde{A}\tilde{B}^{1/2}Z_x + \frac{1}{2}Z_x'\tilde{B}^{1/2}\tilde{A}^{-1}[-A]\tilde{A}^{-1}\tilde{B}^{1/2}Z_x$ now follows. The expectation of the first term is zero, and the final result follows by

$$\mathrm{tr}\{\mathrm{E}Z_x'\tilde{B}^{1/2}\tilde{A}^{-1}A\tilde{A}^{-1}\tilde{B}^{1/2}Z_x\} = \mathrm{tr}\{\tilde{A}^{-1}A\tilde{A}^{-1}\tilde{B}^{1/2}\mathrm{E}Z_xZ_x'\tilde{B}^{1/2}\},$$

and using that $\mathrm{E}Z_xZ_x' = I$. □

This result motivates the following definition of criterion risk

Definition 2. The asymptotic criterion risk, induced by estimation error of $\tilde{\theta}$, is defined by

$$R_\infty(\tilde{\theta}) = \tfrac{1}{2}\mathrm{tr}\{A\tilde{A}^{-1}\tilde{B}\tilde{A}^{-1}\}.$$

The finite sample equivalent is defined by

$$R_n(\tilde{\theta}) = \mathrm{E}[Q(\mathcal{Y},\theta_0) - Q(\mathcal{Y},\tilde{\theta})].$$

For the innate estimator we have $R_\infty(\hat{\theta}) = \frac{1}{2}\mathrm{tr}\{A^{-1}B\}$ and its magnitude relative to $\frac{1}{2}\mathrm{tr}\{A\tilde{A}^{-1}\tilde{B}\tilde{A}^{-1}\}$ defines which of the two coherent estimators is most efficient. We formulate this by defining the *relative criterion efficiency*

$$\mathrm{RQE}(\hat{\theta},\tilde{\theta}) = \frac{\mathrm{E}[Q(\mathcal{Y},\theta_0) - Q(\mathcal{Y},\tilde{\theta}(\mathcal{X}))]}{\mathrm{E}[Q(\mathcal{Y},\theta_0) - Q(\mathcal{Y},\hat{\theta}(\mathcal{X}))]} = \frac{R_n(\tilde{\theta})}{R_n(\hat{\theta})}. \tag{1}$$

Note that an RQE < 1 defines the case where $\tilde{\theta}$ outperforms the innate estimator, $\hat{\theta}$. The asymptotic expression for the RQE is

$$\frac{R_\infty(\tilde{\theta})}{R_\infty(\hat{\theta})} = \frac{\mathrm{tr}\left\{A\tilde{A}^{-1}\tilde{B}\tilde{A}^{-1}\right\}}{\mathrm{tr}\left\{A^{-1}B\right\}},$$

provided that $\tilde{\theta}$ is a coherent estimator. For an incoherent estimator it follows by Lemma 3 that RQE $\to \infty$. as $n \to \infty$.

2.1 Likelihood-Based Estimator

In this section we will consider estimators that are deduced from a likelihood criterion. In some cases, one can obtain $\tilde{\theta}$ directly as a maximum likelihood estimator. However, more generally, there will be a need to map the likelihood parameters, ϑ say, into those of the criterion function, θ. This is for instance needed if the dimensions of the two do no coincide.

So consider a statistical model, $\{P_\vartheta\}_{\vartheta\in\Xi}$, and suppose that P_{ϑ_0} is the true probability measure, with $\vartheta_0 \in \Xi$. The implication is 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 4. *There exists $\tau(\vartheta)$ so that $\vartheta \mapsto (\theta,\tau)$ is continuously differentiable, with $\frac{\partial}{\partial\vartheta}(\theta(\vartheta)',\tau(\vartheta)')'$ having non-zero determinant at ϑ_0.*

The assumption ensures that the reparameterization (that isolates θ) is invertible in a way that does not degenerate the limit distribution. While the assumption is relatively innocuous, we will present a special case in our first application where the assumption is violated.[3]

Lemma 4. *Given Assumption 1 to 4, 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_* 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., $\theta_0 = \theta(\vartheta_0)$. □

One potential challenge to using the likelihood-based estimator is that the mapping from ϑ to θ may be difficult to obtain.

When $\tilde{\theta}$ is estimated from a correctly specified likelihood function, one has $\tilde{A} = \tilde{B}$. In terms of asymptotic criterion risk the comparison of the innate estimator to the likelihood-based estimator,

[3]The assumption also allows us to interpret $\tilde{\theta} = \theta(\tilde{\vartheta})$ as an extremum estimator, that maximizes the reparameterized and concentrated log-likelihood function $\ell_c(\theta) = \ell(\theta,\tilde{\tau}(\theta))$, where $\tilde{\tau}(\theta) = \arg\max_\tau \ell(\theta,\tau)$.

becomes a comparison of the quantities $\frac{1}{2}\text{tr}\{A^{-1}B\}$ and $\frac{1}{2}\text{tr}\{A\tilde{B}^{-1}\}$. The following Theorem shows that the latter is smaller.

Theorem 2 (Optimality of likelihood-based estimator)**.** *Let $\tilde{\vartheta}$ be the maximum likelihood estimator so that $\tilde{\theta} = \theta(\tilde{\vartheta})$ is the LBE of the criterion parameters. If the likelihood function is correctly specified, then, as $n \to \infty$*

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

where $\text{E}[\xi] = R_\infty(\hat{\theta}) - R_\infty(\tilde{\theta}) \leq 0$. The same result holds with $\check{\theta}$ in place of $\hat{\theta}$, provided that $\check{\theta}$ satisfies Assumptions 2 and 3.

Theorem 2 shows that the likelihood-based approach is superior to the criterion-based approach (and any other M-estimator for that matter). An inspection of the proof reveals that the inequality is strict, unless the estimator is asymptotically equivalent to the MLE. So the likelihood-based estimator can be said to be asymptotically efficient. The proof also reveals that manipulation of the asymptotic expression simplifies the comparison to one that is well known from the asymptotic analysis of estimation.

Proof. Consider first the case where $\vartheta = \theta$. From Theorem 1 and a slight variation of its proof it follows that

$$\begin{aligned} Q(\mathcal{Y}, \hat{\theta}) - Q(\mathcal{Y}, \tilde{\theta}) \quad \xrightarrow{d} \quad & +Z_y' B^{1/2} A^{-1} B^{1/2} Z_x - \frac{1}{2} Z_x' B^{1/2} A^{-1} B^{1/2} Z_x \\ & -Z_y' B^{1/2} \tilde{A}^{-1} \tilde{B}^{1/2} \tilde{Z}_x + \frac{1}{2} \tilde{Z}_x' \tilde{B}^{1/2} \tilde{A}^{-1} A \tilde{A}^{-1} \tilde{B}^{1/2} \tilde{Z}_x, \end{aligned}$$

where Z_y, Z_x, and $\tilde{Z}_x$ are all distributed as $N(0, I)$, with Z_y independent of $(Z_x, \tilde{Z}_x)$. Two of the terms vanish after taking the expected value, which yields

$$-\frac{1}{2}\text{tr}\{A^{-1}B\} + \frac{1}{2}\text{tr}\{\tilde{A}^{-1}A\tilde{A}^{-1}\tilde{B}\} = \frac{1}{2}\text{tr}\{A\tilde{A}^{-1} - A^{-1}B\},$$

where we have used the information matrix equality, $\tilde{A} = \tilde{B}$. Manipulating this expression, leads to

$$\frac{1}{2}\text{tr}\left\{A^{1/2}(\tilde{A}^{-1} - A^{-1}BA^{-1})A^{1/2}\right\} \leq 0,$$

where the inequality follows from the fact that $\tilde{A}^{-1} = \tilde{B}^{-1}$ is the asymptotic covariance matrix of the MLE whereas $A^{-1}BA^{-1}$ is the asymptotic covariance of the innate estimator, so that $A^{-1}BA^{-1} - \tilde{B}^{-1}$ is positive semi-definite by the Cramer-Rao bound. These arguments are valid whether θ has the

same dimension as ϑ or not, because we can reparametrize the model in $\vartheta \mapsto (\theta, \gamma)$, which 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 is harmful to the likelihood-based estimator for two reasons. First, the resulting estimator is no longer efficient, which eliminates the argument in favor of adopting the likelihood-based estimator. Second, and more importantly, the mapping from ϑ to θ depends on the true probability measure, so that a misspecified likelihood will result in an improper mapping from ϑ to θ. The likelihood-based estimator $\tilde{\theta}$ may therefore be inconsistent under misspecification, i.e. incoherent.

An incoherent likelihood-based estimator, as the result of a fixed degree of misspecification, will be greatly inferior to the innate estimator in the sense that RQE $\to \infty$ as $n \to \infty$. Such an asymptotic design will in many cases by misleading for the relative performance of competing estimators in finite samples. For this reason we turn our attention to the case with a slightly misspecified model. This can be achieved with an asymptotic design where the likelihood is misspecified, albeit local to correct, in the sense that the likelihood gets closer and closer to being correctly specified as $n \to \infty$, where the rate of convergence is such that $\theta_0 - \theta_* \propto n^{-1/2}$. This form of misspecification may be labelled as local-to-correct.

2.2.1 Local-to-Correct Specification

We consider a case where the true probability measure does not coincide with P_{ϑ_0}. To make matter interesting, we consider a case with a locally misspecified model, where the degree of misspecification is balanced with the sample size. Thus, let the true probability measure be P_n, and let the corresponding (best approximating) likelihood parameter be denoted by $\vartheta_0^{(n)}$, and let $\theta_0^{(n)} = \theta(\vartheta_0^{(n)})$. As n increases, P_n approaches an element, P_{ϑ_0} for some $\vartheta_0 \in \Xi$, and this occurs at a rate so that $\theta_0^{(n)} - \theta_* = n^{-1/2}b$, for some $b \in \mathbb{R}^k$ that defines the degree of local misspecification – correct specification being the case where

$b = 0$. In this scenario, the limit distribution of $Q(\mathcal{Y}, \theta_*) - Q(\mathcal{Y}, \tilde{\theta})$ is given by the following Theorem.

Theorem 3. *Suppose that* $P_n \to P_{\vartheta_0}$ as $n \to \infty$, so that $\theta(\vartheta_0^{(n)}) - \theta(\vartheta_0) = n^{-1/2}b$, then

$$R_\infty(\tilde{\theta}) = \tfrac{1}{2}\text{tr}\{A(\tilde{B}^{-1} + bb')\},$$

where $\tilde{B} = \text{E}[-\tilde{h}(\text{x}_i, \theta_*)]$.

Proof. With $P_n \to P_{\vartheta_0}$ we have $\theta(\vartheta_0) = \theta_*$. Thus with $\theta_0^{(n)} = \theta(\vartheta_0^{(n)})$ and consider the Taylor expansion

[] Since $\theta(\vartheta_0^{(n)}) - \theta(\vartheta_0) = n^{-1/2}b$, and $\tilde{\theta} \xrightarrow{p} \theta(\vartheta_0^{(n)}) = \theta_0$, the Taylor expansion in the proof of Theorem 1 becomes

$$Q(\mathcal{Y}, \tilde{\theta}) - Q(\mathcal{Y}, \theta_*) = S_y(\theta_*)'(\tilde{\theta} - \theta_0^{(n)} + \theta_0^{(n)} - \theta_*) + \frac{1}{2}(\tilde{\theta} - \theta_0^{(n)} + \theta_0^{(n)} - \theta_*)'H_y(\bar{\theta})(\tilde{\theta} - \theta_0^{(n)} + \theta_0^{(n)} - \theta_*).$$

By Assumption 3 and Lemma 2 we have that $n^{-1}H(\mathcal{Y}, \bar{\theta}) \xrightarrow{p} -A$, $n^{-1}\tilde{H}(\mathcal{X}, \theta_0^{(n)}) \xrightarrow{p} -\tilde{A}$, and $\{\tilde{S}(\mathcal{X}, \theta), S(\mathcal{Y}, \theta)\} \xrightarrow{d} \{\tilde{B}^{1/2}Z_x, B^{1/2}Z_y\}$ where Z_x and Z_y are independent and both distributed $N(0, I)$. The result $Q(\mathcal{Y}, \tilde{\theta}) - Q(\mathcal{Y}, \theta_*) \xrightarrow{d} Z_y'B^{1/2}\tilde{A}\tilde{B}^{1/2}Z_x + \frac{1}{2}Z_x'\tilde{B}^{1/2}\tilde{A}^{-1}[-A]\tilde{A}^{-1}\tilde{B}^{1/2}Z_x + \frac{1}{2}b'[-A]b$ follows since $\tilde{\theta} - \theta_0 = \tilde{S}(\mathcal{X}, \theta_0)\tilde{H}(\mathcal{X}, \theta_0)$. The expectation of the first term is zero, and the final result follows from

$$\frac{1}{2}\text{tr}\{\text{E}Z_x'\tilde{B}^{1/2}\tilde{A}^{-1}A\tilde{A}^{-1}\tilde{B}^{1/2}Z_x\} + \frac{1}{2}tr\{b'Ab\} = \frac{1}{2}\{\text{tr}\{A\tilde{B}^{-1}\} + tr\{Abb'\}\},$$

by using that $\text{E}Z_xZ_x' = I$ and the information matrix equality. □

So the likelihood-based estimator retains it efficiency variance under local misspecification, but involves an asymptotic bias term. The implication is that under local misspecification the asymptotic RQE becomes a question of the relative magnitude of the bias, bb' and the relative advantages that the likelihood-based estimator has under correct specification, as measured by $A^{-1}BA^{-1} - \tilde{B}^{-1}$.

One can measure the degree of local misspecification, in terms of the measure on non-centrality. In the univariate case ($\theta \in \mathbb{R}$) this can be expressed as $d = b\sqrt{\tilde{B}}$, which can be interpreted as the expected value of the t-statistic, $(\tilde{\theta} - \theta_*)/\sqrt{\text{avar}(\tilde{\theta})}$. In the multivariate case the non-centrality may be measured as $\sqrt{b'\tilde{B}b}$. While the degree of non-centrality is, in some sense, a measure of the (average) statistical evidence of misspecification, it does not (unless $k = 1$) directly map into a particular value of criterion risk, because different vectors of b can translate into the same non-centrality, $d = \sqrt{b'\tilde{B}b}$, but different values of $\text{tr}\{Abb'\} = b'Ab$. The following Theorem puts upper and lower bounds on the criterion risk that results from a given level of misspecification.

Theorem 4. *Let the local misspecification be such that* $d = \sqrt{b'\tilde{B}b}$. *Then the asymptotic criterion risk*

resulting from this misspecification, $b'Ab$ is bounded by $\lambda_{\min} d^2 \leq b'Ab \leq \lambda_{\max} d^2$ where $\lambda_{\min}$ and $\lambda_{\max}$ are the smallest and largest solutions (eigenvalues) to $|A - \lambda \tilde{B}| = 0$.

Proof. With $y = \tilde{B}^{1/2} b$ we have $b'Ab/b'\tilde{B}b = y'\tilde{B}^{-1/2} A \tilde{B}^{-1/2} y / y'y$ which is bounded by the smallest and largest eigenvalues of $\tilde{B}^{-1/2} A \tilde{B}^{-1/2}$. If λ is a solution to $|\tilde{B}^{-1/2} A \tilde{B}^{-1/2} - \lambda I| = 0$ then λ also solves $|A - \lambda \tilde{B}| = 0$, and the result follows. □

A more general way of measuring the misspecification is in terms of the KLIC, $\mathrm{E}_{P_n}[\log \frac{f_{\vartheta_0^{(n)}}(\mathcal{X})}{g_n(\mathcal{X})}]$, or one can measure the discrepancy in terms on the non-centrality of a suitable test statistic. For instance, the misspecification of a Gaussian distribution may intuitively be expressed in terms of the non-centrality parameter in a Jarque–Bera test.

We shall study designs with local misspecification in the following two sections. The first section is an application to a criterion function defined by the asymmetric LinEx loss function and the second section is an application of our framework to the problem of making multistep-ahead forecasting.

3 The Case with Asymmetric Loss and a Gaussian Likelihood

In this section we apply the theoretical results to the case where the criterion function is given by the LinEx loss function. In forecasting problems, there are many applications where asymmetry is thought to be appropriate, see 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), and 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

$$L_c(x) = \begin{cases} c^{-2}[\exp(cx) - cx - 1] & \text{for } c \in \mathbb{R}\backslash\{0\}, \\ \frac{1}{2}x^2 & \text{for } c = 0 \end{cases} \tag{2}$$

which has minimum at $x = 0$. The absolute value of the parameter c determines the degree of asymmetry and its sign defines whether the asymmetry is left-skewed or right-skewed, see Figure 1. The quadratic loss arises as the limited case, $\lim_{c \to 0} L_c(x) = \frac{1}{2}x^2$, which motivates the definition of $L_0(x)$.

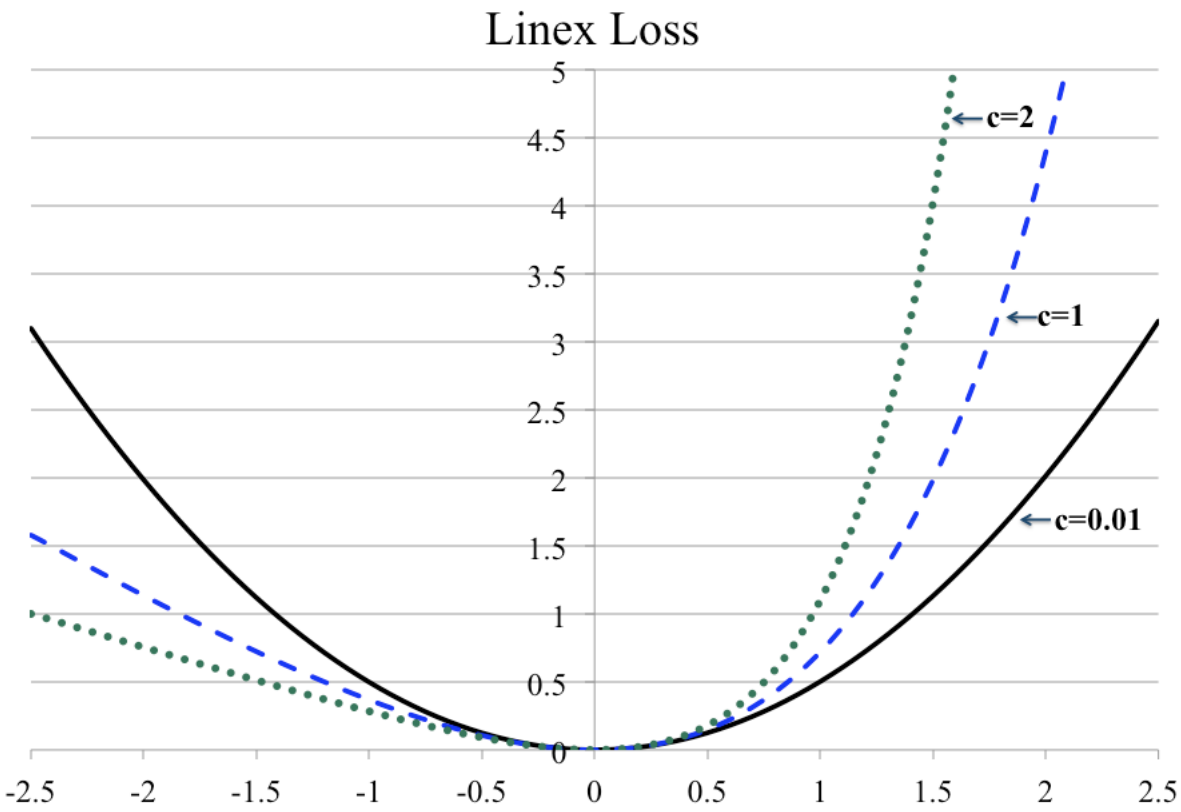

Figure 1: The LinEx loss function for three values of c.

We adopt the LinEx loss function because it produces simple estimators in closed-form that ease the computational burden.

The objective in this application is to estimate θ for the purpose of minimizing the expected loss, $\mathrm{E}L_c(Y_i-\theta)$. This problem maps in to our theoretical framework by setting $q(X_i,\theta)=-L_c(X_i-\theta)$, and it is easy to show that $\theta_* = \arg\min \mathrm{E}L_c(X_i-\theta) = c^{-1}\log[\mathrm{E}\exp(cX_i)]$, provided that $\mathrm{E}\exp(cX_i)<\infty$. Similarly, it can be shown that the innate estimator, which is given as the solution to $\min_\theta \sum_{i=1}^n L_c(X_i-\theta)$, can be written in closed-form as

$$\hat{\theta} = \frac{1}{c}\log[\frac{1}{n}\sum_{i=1}^{n}\exp(cX_i)], \tag{3}$$

and by the ergodicity of X_i, hence $\exp(cX_i)$, it follows that $\hat{\theta} \xrightarrow{as} \theta_*$. Next, we introduce a likelihood-based estimator that is deduced from the assumption that $X_i \sim \mathrm{iid}N(\mu_0,\sigma_0^2)$, for which it can be shown that

$$\theta_0 = \mu_0 + \frac{c\sigma_0^2}{2}, \tag{4}$$

see Christoffersen and Diebold (1997). The likelihood-based estimator is therefore given by

$$\tilde{\theta} = \tilde{\mu} + \frac{c\tilde{\sigma}^2}{2}, \tag{5}$$

where $\tilde{\mu} = n^{-1}\sum_{t=1}^{n} X_i$ and $\tilde{\sigma}^2 = n^{-1}\sum_{t=1}^{n}(X_i - \tilde{\mu})^2$ are the maximum likelihood estimators of μ_0 and σ_0^2, respectively.

Equation (4) illustrates the need to map likelihood parameters, $\vartheta = (\mu, \sigma^2)'$, into criterion parameter, θ. The likelihood-based estimator will be consistent for θ_0, which coincides with θ_* if the Gaussian assumption is correct. Under misspecification the two need not coincide.

We shall compare the two estimators in terms of the LinEx criterion

$$Q(\mathcal{Y}; \theta) = -\sum_{i=1}^{n} c^{-2}[\exp\{c(Y_i - \theta)\} - c(Y_i - \theta) - 1],$$

where Y_i are iid and independent of $(X_1, \ldots, X_n)$. First we consider the case with correct specification, i.e. the case where $(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$ are iid with marginal distribution $N(\mu_0, \sigma_0^2)$. Subsequently we turn to the case where the marginal distribution is a normal inverse Gaussian (NIG) distribution, which causes the Gaussian likelihood to be misspecified.

3.1 Results for the Case with Correct Specification

With $q_i(X_i, \theta) = -L_c(Y_i - \theta)$ we have $s_i(X_i, \theta) = c^{-1}[\exp\{c(Y_i - \theta)\} - 1]$ and $h_i(X_i, \theta) = -[\exp\{c(X_i - \theta)\}$. With $X_i \sim iidN(\mu, \sigma^2)$ it can be verified that

$$\begin{aligned} A &= \mathrm{E}[-h_i(X_i, \theta_0)] = 1 \\ B &= \mathrm{var}[s_i(X_i, \theta_0)] = \frac{\exp(c^2\sigma^2) - 1}{c^2}, \quad (= \sigma^2 \text{ if } c = 0), \\ \tilde{A} = \tilde{B} &= 1/\mathrm{avar}(\tilde{\theta}) \qquad = 1/(\sigma^2 + c^2\sigma^4/2), \end{aligned}$$

see Appendix A. Consequently, in this application we have

$$\mathrm{RQE} = \frac{\mathrm{tr}\{A\tilde{B}^{-1}\}}{\mathrm{tr}\{A^{-1}B\}} = \frac{1}{B\tilde{B}} = \frac{(c\sigma)^2 + (c\sigma)^4/2}{\exp(c\sigma)^2 - 1},$$

which is (unsurprisingly) less than or equal to one for all combinations of c and σ, and $\mathrm{RQE} = 1$ if and only if $c\sigma = 0$.

The relative efficiency of $\hat{\theta}$ and $\tilde{\theta}$ is compared in Table 1 for the case with a correctly specified

likelihood function.

Table 1: Relative Efficiency under LinEx Loss

Panel A: Asymptotic Results

c	θ_*	RQE	$R_\infty(\hat{\theta})$	$R_\infty(\tilde{\theta})$	bias$(\hat{\theta})$	bias$(\tilde{\theta})$
0	0	1	0.5	0.5	0.000	0.000
0.25	0.125	0.999	0.516	0.516	0.000	0.000
0.5	0.250	0.990	0.568	0.563	0.000	0.000
1	0.500	0.873	0.859	0.750	0.000	0.000
1.5	0.750	0.563	1.886	1.063	0.000	0.000
2	1.000	0.224	6.700	1.500	0.000	0.000
2.5	1.250	0.050	41.36	2.063	0.000	0.000

Panel B: Finite Sample Results: $n = 1,000$

c	θ_*	RQE	$R_n(\hat{\theta})$	$R_n(\tilde{\theta})$	bias$(\hat{\theta})$	bias$(\tilde{\theta})$
0	0	1	0.499	0.499	0.000	0.000
0.25	0.125	0.999	0.518	0.518	0.000	0.000
0.5	0.250	0.991	0.569	0.563	0.000	0.000
1	0.500	0.88	0.853	0.748	-0.001	0.000
1.5	0.750	0.60	1.777	1.068	-0.003	-0.001
2	1.000	0.35	4.341	1.513	-0.010	-0.001
2.5	1.250	0.217	9.670	2.100	-0.030	-0.001

Panel C: Finite Sample Results: $n = 100$

c	θ_*	RQE	$R_n(\hat{\theta})$	$R_n(\tilde{\theta})$	bias$(\hat{\theta})$	bias$(\tilde{\theta})$
0	0	1	0.498	0.498	0.000	0.000
0.25	0.125	0.999	0.515	0.514	-0.001	-0.001
0.5	0.750	0.991	0.567	0.562	-0.003	-0.003
1	0.500	0.90	0.839	0.753	-0.009	-0.005
1.5	0.750	0.72	1.493	1.075	-0.023	-0.008
2	1.000	0.56	2.745	1.526	-0.058	-0.010
2.5	1.250	0.418	5.273	2.203	-0.122	-0.012

Note: The likelihood-based estimator $\tilde{\theta}$ is compared to the innate estimator, $\hat{\theta}$, in terms of the relative criterion efficiency in the case with LinEx loss and iid Gaussian observations with zero mean and unit variance. The LBE based predictor dominates the innate criterion based predictor, and does so increasingly as the asymmetry increases. The upper panel is intended to match the asymptotic results, whereas the next two panels present the corresponding results in finite samples, $n = 100$ and $n = 1,000$. The results are based on 500,000 simulations.

Panel A of Table 1, displays the asymptotic results base on our analytical expressions, whereas Panels B and C present finite sample results based on simulations with $n = 1,000$ and $n = 100$, respectively.

500,000 replications were used to compute all statistics.[4] The simulation design is detailed in Appendix B.1. The asymmetry parameter is given in the first column followed by the population value of θ_*, the RQE, the criterion losses resulting from estimation error, and the bias of the two estimators.

Table 1 shows that (for $c \neq 0$) the likelihood-based estimator dominates the innate estimation, and increasingly so, as c increases. In the asymptotic design this simply reflects the effect that c has on the A and B matrices. The superiority of the LBE is (as dictated by our analytical results) found in our asymptotic design, however the LBE also dominates the innate estimator in finite samples, albeit to a less extent. The main reason why the innate estimator appears to be relatively better in finite samples, is because its criterion loss tends to be relatively smaller in finite samples. However this does not imply that the innate estimator performs better with a smaller sample size, because the per observation criterion loss, $R_n(\hat{\theta})/n$, is decreasing in n. Moreover, the innate estimator has a larger finite sample bias relative to that of the likelihood-based estimator.

3.1.1 Likelihoods with One-Dimensional Parameter

With a likelihood deduced from $X_t \sim N(\mu, \sigma^2)$, we have in some sense stacked the results against the likelihood-based estimators. The likelihood approach involves a two-dimensional estimator, $(\tilde{\mu}, \tilde{\sigma}^2)$, whereas the innate estimator only estimates a one-dimensional object, and this might be viewed as being favorable to the innate estimator. While this may be true in finite samples, the dimension of ϑ is immaterial to the asymptotic comparison, in the sense that the asymptotic RQE for the likelihood-based estimator is always bounded by one, regardless of the dimension of ϑ. However, the asymptotic variance of $\tilde{\theta}$ could be influenced by the complexity of the underlying likelihood function, so that a simpler likelihood (one with fewer degrees of freedom) may be even better in terms of RQE. To illustrate this, we considered a restricted model, in which σ_0^2 is known, so that only μ is to be estimated. We also consider the case where μ_0 is known so that σ_0^2 is the only parameter to be estimated. This design is of separate interest because it constitutes a case where Assumption 4 is violated when $c = 0$.

When σ_0^2 is known, the asymptotic variance of $\tilde{\theta}$ is smaller than in our first design. This results in a more efficient estimator. The design corresponds to a case where the "stakes" in using the likelihood approach are raised, because misspecification can now result from an incorrect assumed value for σ_0^2 (in addition to the previous forms of misspecification).

[4] These quantities are estimated by simulations with a high accuracy. Standard deviations are smaller than 10^{-5} in all cases, and smaller that 10^{-7} in the case of the estimated biases.

$$\text{avar}(\tilde{\theta}) = \begin{cases} \sigma_0^2 + \frac{c^2}{2}\sigma_0^4 & \\ \sigma_0^2 & \text{if } \sigma_0^2 \text{ is known} \\ \frac{c^2}{2}\sigma_0^4 & \text{if } \mu_0 \text{ is known} \end{cases}$$

When μ_0 is known, the mapping from ϑ to θ is simply one from σ^2 to θ, which does not depend on σ^2 when $c = 0$. This constitutes a case where Assumption 4 is violated, because $\partial\theta(\sigma^2)/\partial\sigma^2 = 0$ when $c = 0$. Consequently the asymptotic results need not apply in this case. This particular violation of Assumption 4, as it turns out, is to the advantage of the likelihood-based estimator, because with μ_0 known and $c = 0$, the optimal estimator is known without any need for estimation. For c close to zero, the LBE benefits from having a very small asymptotic variance (which is proportional to c^2).

Table 2: Relative Criterion Efficiency: 1-dimensional likelihood parameters

		Panel A: σ_0^2 known $\tilde{\theta} = \tilde{\mu} + c\sigma_0^2/2$			Panel B: μ_0 known $\tilde{\theta} = \mu_0 + c\tilde{\sigma}^2/2$		
c	θ_*	RQE	$R_\infty(\hat{\theta})$	$R_\infty(\tilde{\theta})$	RQE	$R_\infty(\hat{\theta})$	$R_\infty(\tilde{\theta})$
0	0	1	0.5	0.5	0	0.5	0
0.25	0.125	0.969	0.516	0.5	0.030	0.516	0.016
0.5	0.25	0.880	0.568	0.5	0.110	0.568	0.063
1.0	0.50	0.582	0.859	0.5	0.291	0.859	0.250
1.5	0.75	0.265	1.886	0.5	0.298	1.886	0.563
2.0	1.00	0.075	6.700	0.5	0.149	6.700	1.000
2.5	1.25	0.012	41.36	0.5	0.038	41.36	1.563

Note: ....

Table 2 reports the results for the two cases where ϑ is one-dimensional. Panel A has the case $\vartheta = \mu$ (and σ^2 known) and Panel B has the case where $\vartheta = \sigma^2$ (and μ known). As expected, the likelihood-based estimator performs even better, in these cases where the dimension of ϑ is smaller. In Panel A, where σ_0^2 is known, the asymptotic criterion risk, $R_\infty(\tilde{\theta})$, for the likelihood estimator does not depend on c, while the corresponding criterion loss for the innate estimator is increasing in c. In Panel B, $R_\infty(\tilde{\theta})$ is increasing in c starting from zero at $c = 0$. The theoretical explanations for this follows from the underlying information matrices. Because the innate estimator is unaffected by the choice of specification for the likelihood, we continue to have $A = 1$ and $B = [\exp(c^2\sigma^2) - 1]/c^2$ in both cases. Consequently, we have the same expression for $R_\infty(\hat{\theta}) = \frac{1}{2}\text{tr}\{A^{-1}B\} = \frac{1}{2}[\exp(c^2) - 1]/c^2$. For

the likelihood-based estimators the expressions are different. For the specification in Panel A we have $\tilde{B} = 1/\sigma_0^2 = 1$, so that $\frac{1}{2}\text{tr}\{A\tilde{B}^{-1}\} = \frac{1}{2}$. Similarly, for the specification in Panel B we have $\tilde{B}^{-1} = c^2/2$, so that $\frac{1}{2}\text{tr}\{A\tilde{B}^{-1}\} = c^2/4$.

3.2 Local Misspecification

As previously discussed, likelihood misspecification entails problems in two directions: the efficiency argument for the likelihood-based estimator perishes and the transformation of likelihood parameters into criterion parameters is likely to be wrong. To study the impact of misspecification we now consider the case where the truth is defined by a normal inverse Gaussian (NIG) distribution, so that the Gaussian likelihood is misspecified.

A NIG distribution is characterized by four parameters, λ, δ, α, and β, that represent location, scale, tail heaviness, and asymmetry, respectively, see Figure 2. The NIG-distribution is flexible and well suited for the present problem, because the Gaussian distribution, $N(\mu, \sigma^2)$, can be obtained as the limited case where $\lambda = \mu$, $\delta = \sigma^2\alpha$, $\beta = 0$, and $\alpha \to \infty$, and because the distribution yields tractable analytical expression for the quantities that are relevant for our analysis of the LinEx loss function.

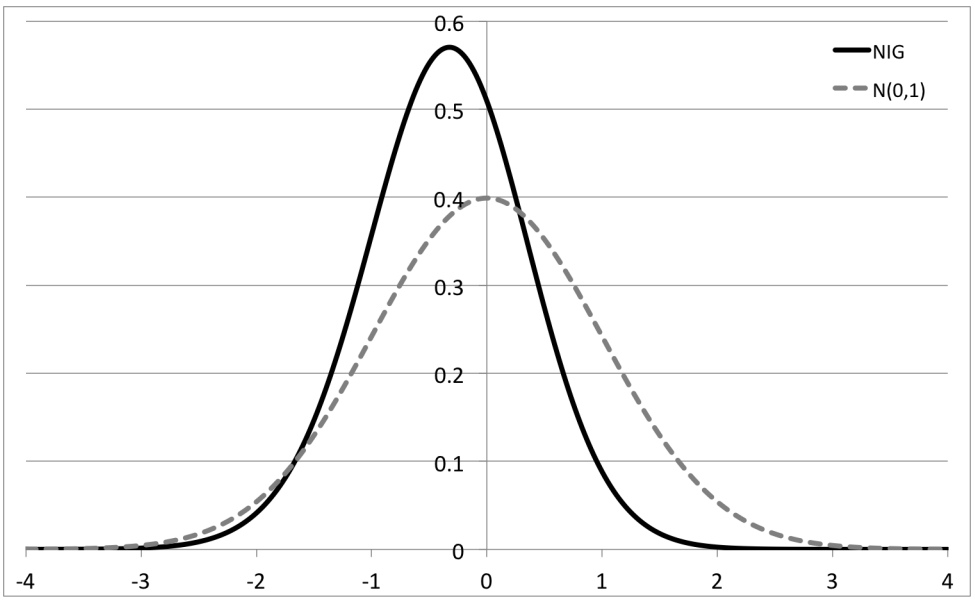

Figure 2: The density of the NIG distribution for a particular parameterization (with mean zero and unit variance) and the standard Gaussian density.

The mean and variance of NIG$(\lambda, \delta, \alpha, \beta)$ are given by $\mu = \lambda + \frac{\delta\beta}{\gamma}$ and $\sigma^2 = \delta\frac{\alpha^2}{\gamma^3}$, respectively, where $\gamma = \sqrt{\alpha^2 - \beta^2}$. So it follows that the likelihood-based estimator converges in probability to,

$$\theta_0 = (\lambda + \tfrac{\delta\beta}{\gamma}) + \tfrac{c}{2}\delta\frac{\alpha^2}{\gamma^3}.$$

The ideal value for θ is, however, equal to

$$\theta_* = \lambda + \frac{\delta}{c}\left[\sqrt{\alpha^2 - \beta^2} - \sqrt{\alpha^2 - (\beta + c)^2}\right], \tag{6}$$

see Appendix A, and the two values do not coincide except for some special cases that can be obtained as various limits. So the (misspecified) likelihood-based estimator is incoherent, unless $c = 0$. The latter follows because both θ_* and θ_0 converge to $\lambda + \frac{\delta\beta}{\gamma}$ in probability as $c \to 0$, see (9) in Appendix A. Moreover, if we set $\delta = \sigma^2\alpha$ and $\beta = 0$ then $\theta_* - \theta_0 \to 0$ as $\alpha \to \infty$, for any value of c.

To make our misspecified design comparable to our previous design (where $X_i \sim iidN(0, 1)$) we consider the standard NIG distribution. The zero mean and unit variance is achieved by setting $\lambda = -\frac{\delta\beta}{\gamma}$ and $\delta\frac{\alpha^2}{\gamma^3} = 1$. This family of standard NIG distributions can, conveniently, be characterized by the two parameters

$$\xi = \frac{1}{\sqrt{1 + \delta\gamma}} \qquad \text{and} \qquad \chi = \xi\frac{\beta}{\alpha},$$

that will be such that $0 \leq |\chi| < \xi < 1$. The original parameter values can be backed out using

$$\alpha = \xi\frac{\sqrt{1 - \xi^2}}{\xi^2 - \chi^2} \qquad \text{and} \qquad \beta = \chi\frac{\sqrt{1 - \xi^2}}{\xi^2 - \chi^2},$$

that implies $\gamma = \sqrt{1 - \xi^2}$. The limited case where $\xi = 0$ (and hence $\chi = 0$) corresponds to the standard Gaussian distribution.

We now construct a local-to-correct specified model by making the truth on in which

$$\xi_n = -\chi_n^{2/3} = b/\sqrt{n}$$

Table 3: Local Misspecification

	d	θ_*	RQE	$R_\infty(\hat{\theta})$	$R_\infty(\tilde{\theta})$	bias$(\tilde{\theta})$	$R_\infty^{\text{bias}}(\tilde{\theta})$
$c = 0.0$	0	0.000	1.00	0.497	0.497	0.000	0.000
	0.3	0.000	1.00	0.499	0.499	0.000	0.000
	0.5	0.000	1.00	0.500	0.500	0.000	0.000
	1	0.000	1.00	0.489	0.489	0.000	0.000
	4	0.000	1.00	0.500	0.500	0.000	0.000
	7	0.000	1.00	0.491	0.491	0.000	0.000
	10	0.000	1.00	0.503	0.503	0.000	0.000

	d	θ_*	RQE	$R_\infty(\hat{\theta})$	$R_\infty(\tilde{\theta})$	bias$(\tilde{\theta})$	$R_\infty^{\text{bias}}(\tilde{\theta})$
$c = 0.25$	0	0.125	1.00	0.513	0.51	0.000	0.000
	0.3	0.125	1.00	0.512	0.51	0.000	0.000
	0.5	0.125	1.00	0.518	0.52	0.000	0.001
	1	0.125	1.00	0.513	0.51	0.000	0.001
	4	0.125	1.06	0.515	0.54	0.000	0.030
	7	0.125	1.18	0.516	0.61	0.000	0.094
	10	0.125	1.39	0.510	0.71	0.001	0.200

	d	θ_*	RQE	$R_\infty(\hat{\theta})$	$R_\infty(\tilde{\theta})$	bias$(\tilde{\theta})$	$R_\infty^{\text{bias}}(\tilde{\theta})$
$c = 0.5$	0	0.250	0.99	0.574	0.57	0.000	0.000
	0.3	0.250	1.00	0.567	0.56	0.000	0.003
	0.5	0.250	1.00	0.569	0.57	0.000	0.007
	1	0.250	1.04	0.560	0.59	0.000	0.030
	4	0.249	1.85	0.556	1.03	0.001	0.474
	7	0.248	3.62	0.560	2.03	0.002	1.471
	10	0.248	6.62	0.546	3.62	0.002	3.061

	d	θ_*	RQE	$R_\infty(\hat{\theta})]$	$R_\infty(\tilde{\theta})$	bias$(\tilde{\theta})$	$R_\infty^{\text{bias}}(\tilde{\theta})$
$c = 1$	0	0.500	0.87	0.859	0.75	0.000	0.000
	0.3	0.500	0.92	0.861	0.79	0.000	0.039
	0.5	0.500	1.00	0.860	0.86	0.000	0.110
	1	0.499	1.42	0.834	1.18	0.001	0.445
	4	0.496	9.49	0.828	7.86	0.004	7.109
	7	0.493	27.5	0.815	22.5	0.007	21.73
	10	0.491	57.0	0.790	45.1	0.009	44.32

	d	θ_*	RQE	$R_\infty(\hat{\theta})]$	$R_\infty(\tilde{\theta})$	bias$(\tilde{\theta})$	$R_\infty^{\text{bias}}(\tilde{\theta})$
$c = 2$	0	1.000	0.22	6.592	1.48	0.000	0.000
	0.3	0.999	0.32	6.612	2.11	0.001	0.582
	0.5	0.998	0.49	6.564	3.22	0.002	1.697
	1	0.996	1.24	6.461	8.02	0.004	6.576
	4	0.986	17.6	5.739	101	0.014	99.44
	7	0.976	55.7	5.264	293	0.024	291.7
	10	0.966	119	4.829	575	0.034	573.2

Note: The likelihood-based estimator $\tilde{\theta}$ is compared with the innate estimator, $\hat{\theta}$, in the case where the Gaussian likelihood is local-to-correctly specified, for different levels of asymmetry. $R_\infty^{\text{bias}}(\tilde{\theta})$ captures the bias component of the risk, $b'Ab/2$ The data generating process is a standard NIG distribution, where the degree of local misspecification is determined by d. The "asymptotic" results are based on 100,000 replication with $n = 10^6$.

Table 3 displays the behavior of the two forecasts in the case when we depart from the normality assumption.

The first panel, where $c = 0$, corresponds to the case of the MSE loss function. Here, the likelihood-based and innate estimators are equivalent and equal to the sample average, a consistent estimator of the conditional mean, such that RQE is constant at 1. The estimation loss increases with the degree of misspecification...

The RQE in column 3 shows how the performance of the (quasi) likelihood-based estimator is linked to the degree of misspecification and asymmetry level. For low levels of misspecification, the LBE dominates the innate estimator, and its performance improves as c increases. By contrast, for large departures from normality its performance worsens and it becomes much inferior to the innate one. This can be explained by the fact that the mapping from the MLE ϑ to the criterion parameters θ becomes improper as the misspecification level d increases. Figure 3 provides a clearer insight into the impact of local misspecification on the RQE under LinEx loss. For values of d up to about 1 RQE is always lower than or equal to 1, suggesting that the QMLE is robust to small levels of misspecification. In this case, the larger the asymmetry c the better LBE is. In contrast, once the bias parameter exceeds the threshold at 1 the relative performance results are completely reversed. The innate predictor becomes preferable, with an exponential drop in the LBE's performance as the level of asymmetry in the LinEx loss increases. These results are supported by the fact that the LBE risk surges with d because the asymptotic bias of the estimator increases. Meanwhile, the risk of the innate estimator is due to the variance of the estimator (the innate estimator is consistent even in the case of misspecified models).

The design of the simulations is detailed in appendix B.2.

4 Multi-Period Forecasting

Our theoretical framework can be applied to the problem of making multi-period forecasts, where forecasts based on the innate and likelihood-based estimators are known as the direct forecast and iterated forecast, respectively, see e.g. Marcellino et al. (2006). The iterated forecasts are also known as the plug-in forecasts. There is a vast literature on this issue, see e.g. Cox (1961), Tiao and Tsay (1994), Clements and Hendry (1996), Bhansali (1997), Ing (2003), Chevillon (2007), and references therein. This literature has mainly focused on the case with MSE loss with or without misspecification. We make ancillary contributions to this literature by showing that the merits of direct versus iterated forecasts can be analyze in the theoretical setting of Section 2. We also contribute to the literature by

establishing results beyond the quadratic loss function. We shall see that the asymmetric LinEx loss function exacerbates the inefficiency of the direct estimator.

Consider an autoregressive process of order p. In this context, the direct forecast of Y_{T+h} at time T is obtained by regressing Y_t on $(Y_{t-h}, \ldots, Y_{t-p-h+1})$ and a constant for $t = 1, \ldots, T$, whereas the iterated forecast are obtained by estimating the an AR(p) model, that yields a forecast of Y_{T+1}, which is subsequently used to construct a forecast of Y_{T+2}, and so forth until the forecast of Y_{T+h} is obtained, by repeated use of the estimated autoregressive model.

For ease of exposition, we restrict our attention to the case of a simple first-order autoregressive model

$$Y_t = \mu + \varphi Y_{t-1} + \varepsilon_t, \qquad t = 1, 2, \ldots$$

where $\varepsilon_t \sim iidN(0, \sigma^2)$. It follows that the conditional distribution of Y_{t+h} given Y_t is $N(\varphi^h Y_t + \frac{1-\varphi^h}{1-\varphi}\mu, \frac{1-\varphi^{2h}}{1-\varphi^2}\sigma^2)$, so that the optimal predictor under LinEx loss is given by

$$Y^0_{t+h,t} = \varphi^h Y_t + \tfrac{1-\varphi^h}{1-\varphi}\mu + \frac{c}{2}\tfrac{1-\varphi^{2h}}{1-\varphi^2}\sigma^2. \tag{7}$$

The iterated (likelihood-based) predictor, $\tilde{Y}_{t+h,t}$, is given by plugging the maximum likelihood estimators, $\tilde{\mu}$, $\tilde{\varphi}$, and $\tilde{\sigma}^2_\varepsilon$ into this expression. In the notation of Section 2, we have $\vartheta = (\mu, \varphi, \sigma^2)'$ and

$$\theta(\vartheta) = \left(\tfrac{1-\varphi^h}{1-\varphi}\mu + \frac{c}{2}\tfrac{1-\varphi^{2h}}{1-\varphi^2}\sigma^2, \varphi^h\right)',$$

and for simplicity we use the notation $\theta = (\alpha, \beta)'$ for the two elements of θ, so that the iterated forecast can be expressed in terms of the likelihood-based estimators, $\tilde{Y}_{t+h,t} = \tilde{\alpha} + \tilde{\beta} Y_t$, with $\tilde{\alpha} = \frac{1-\tilde{\varphi}^h}{1-\tilde{\varphi}}\tilde{\mu} + \frac{c}{2}\frac{1-\tilde{\varphi}^{2h}}{1-\tilde{\varphi}^2}\tilde{\sigma}^2$ and $\tilde{\beta} = \tilde{\varphi}^h$.

The direct forecast is based on the innate estimators, $\hat{\alpha}$ and $\hat{\beta}$, that are obtained by solving $\min_{\alpha,\beta} \sum_{t=1}^{T} L_c(Y_t - \alpha - \beta Y_{t-h})$. The resulting forecast is simply $\hat{Y}_{t+h,t} = \hat{\alpha} + \hat{\beta} Y_t$.

Panel A. in Table 4 displays the asymptotic forecast evaluation results ($n = 100,000$) for several levels of asymmetry of the LinEx loss $c \in \{0.1; 0.5; 1; 2\}$ and for different levels of persistence of the autoregressive process $\varphi \in \{0.3; 0.8; 0.99\}$. The forecast horizon is fixed to $h = 2$. First, note that the results for $c = 0.1$ (see Part i) of the table) roughly mimic the behavior of the two estimators in the MSE case. The forecasting superiority in this setup of the iterated method with respect to the direct one has been emphasized theoretically in the literature (Bhansali, 1999; Ing, 2003). Nevertheless, the role of the autoregressive parameter φ in the evaluation has not been explicitly tackled, even though it

deserves attention. Observe that the larger φ, i.e. the higher the persistence of the process, the more the relative efficiency of the likelihood-based predictor with respect to the innate one diminishes such that 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 likelihood-based estimator losses in efficiency since its variance is approaching at a fast rate the variance of the innate estimator. Moreover, an increase in the forecast loss (shrinkage in the evaluation criterion) adds to the reduction in relative efficiency as φ rises.

Second, parts ii) to iv) display the relative behavior of the two estimators when the asymmetry in the evaluation criterion increases. The larger the asymmetry, the smaller RQE and the better the LBE, iterated predictor with respect to the innate, direct one. The improvement is notable especially for highly persistent processes ($\varphi \in \{0.8; 0.99\}$). At the same time, as expected since the model is correctly specified, in all cases the likelihood-based and the innate predictors are asymptotically unbiased (see columns 6 and 7). We also note the high precision of the simulation results, with standard deviations less than 10^{-4} for the evaluation criteria, and less than 10^{-6} for the bias of the estimators.

4.1 Finite-Sample Results

As aforementioned, the large-sample properties of the two predictors under MSE loss 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. In panel B. of Table 4 we hence report the results under the LinEx loss for $n = 1,000$ and a more realistic $n = 200$ sample size, by performing $500,000$ simulations. 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 alternative estimator is the likelihood-based one.

[] RQE registers very similar values to Panel A. regardless of the level of asymmetry c, even though the per-observation 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 observations (as in the LinEx application) so as to obtain the per-observation loss in the evaluation criterion due to estimation. At the same time, the LBE and innate estimator exhibit small-sample bias. We stress the fact that the larger φ, i.e. the more persistent the process, the larger the bias.

All in all, the likelihood-based, iterated, predictor is proven to be relatively more efficient than the innate, direct, predictor even in small samples. Most importantly, by acknowledging the fact that

for persistent processes the relative gain of LBE decreases while the bias increases, we recommend to pay more attention to the estimated autoregressive parameters and the choice of evaluation criterion 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 could be obtained when comparing such estimators with the traditional ones. Further investigation into this issue under the LinEx loss would be interesting.

4.2 Local Misspecification

To study the effect of local misspecification on the relative efficiency of the likelihood-based predictor with respect to the innate one, we adopt an asymptotic design from Schorfheide (2005). Specifically, we keep the complexity of our prediction models fixed, and introduce local misspecification in the conditional mean. Unlike our previous application, where we deviated from the Gaussian distribution, we maintain the Gaussian distribution for the innovations, ε_t, but define the true data generating process to be an AR(2) model, $Y_t = \mu + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \varepsilon_t$. The local-to-correct specification is achieved by letting $\varphi_2 = O(n^{-1/2})$. Specifically, we set $\varphi_2 = n^{-1/2}\frac{d}{\sigma_{\varphi_2}}$ where $\sigma^2_{\varphi_2} =$ is the asymptotic variance of $\tilde{\varphi}_2$ when estimated by maximum likelihood when $\varphi_2 = 0$. The constant, d, defines the degree of misspecification and can be interpreted as the non-centrality of the t-statistics associated with testing $\varphi_2 = 0$. In this local-to-correct design we hold the the first order autocorrelation, $\rho_1 = \text{corr}(Y_t, Y_{t-1})$, constant, which is achieved by setting $\varphi_1 = \rho_1(1 - \varphi_2)$. We focus on the two cases $\rho_1 = 0.8$ and $\rho_1 = 0.99$.

[new specification] The degree of misspecification is defined in terms of the expected value of the t-statistic for the second autoregressive parameter, $\varphi_2 = \frac{d}{\sqrt{Avar(\varphi_2)}}$, where $\hat{Avar}(\hat{\varphi}_2) = \frac{(1+\hat{\varphi}_2)(1-\hat{\varphi}_1-\hat{\varphi}_2)(1+\hat{\varphi}_1-\hat{\varphi}_2)}{(1-\hat{\varphi}_1)\times n}$ since $Var(Y_T) = \frac{(1-\hat{\varphi}_2)\sigma}{(1+\hat{\varphi}_2)(1-\hat{\varphi}_1-\hat{\varphi}_2)(1+\hat{\varphi}_1-\hat{\varphi}_2)}$, with $\hat{\varphi}_1$ and $\hat{\varphi}_2$ estimated from a correctly specified AR(1) model.

The optimal predictor is given by

$$Y^*_{T+2} = (\varphi_1^2 + \varphi_2)Y_T + \varphi_1\varphi_2 Y_{T-1} + \frac{c}{2}(1 + \varphi_1^2), \tag{8}$$

but neither the direct nor the iterated estimator make use of two lags of Y_t. Using the probability limit

for the iterated estimator we have

$$Y_{T+2}(\tilde{\theta}_0) = \varphi_1^2 Y_T + \frac{c}{2}(1+\varphi_1^2),$$

whereas the direct estimator takes the same form as in the AR(1) case... based on the minimization of the two-periods-ahead LinEx loss. The direct method will correctly estimate β as the second order autocorrelation ρ_2, which, however, is no longer equal to ρ_1^2 as in the AR(1) case. It follows that the larger the local misspecification parameter d, the larger the bias in the quasi-MLE which approximates ρ_2 by ρ_1^2. The simulations design is similar to the one described in Appendix B.3 except that we generate an AR(2) process with the particularities described above instead of an AR(1).

The results obtained for $\rho_1 = 0.8$ and different levels of asymmetry $c \in \{0.1; 0.5; 1; 2\}$ are presented in Table 5. First, we notice that the larger the asymmetry c, the more efficient the quasi-likelihood-based estimator is relatively to the innate one and the more resilient it is, i.e. the larger the level of local-misspecification up to which the quasi-LBE is preferable to the innate estimator. Then, as expected, the asymptotic bias of the QLBE (column 6) increases with d, i.e. it becomes inconsistent. The loss associated with the quasi-likelihood-based predictor hence increases faster than the one of the innate one (see columns 3 and 4), leading to a progressive increase in RQE. The loss for the plim (quasi-LBE), θ_0, (column 8) exhibits a similar behavior.

The top panel in figure 4 provides a clearer insight into the impact of local misspecification on the RQE criterion under LinEx loss. For values of d up to 4, or equivalently, a φ_2 of up to 0.01, the quasi-LBE performs better than the innate estimator. It proves to be robust to even larger levels of misspecification, d=15 ($\varphi_2 = 0.05$) if the asymmetry in the LinEx evaluation criterion is set to 2. Similar results have been obtained for a first order autocorrelation coefficient ρ_1 equal to 0.99[5]. The bottom panel in Figure 4 shows that the local misspecification impacts the relative efficiency of the two estimators in a similar way. Note that in this case the symmetric quadratic loss leads to a constant RQE, equal to 1 regardless of the level of misspecification. In contrast, the RQE for a largely asymmetric LinEx loss ($c = 2$) exhibits a more non-linear trend than for a $\rho_1 = 0.8$.

The larger the asymmetry c the better the quasi-LBE is. In contrast, once the bias parameter exceeds the threshold at 1 the relative performance results are completely reversed. The innate predictor becomes preferable, with an exponential drop in the performance of the quasi-likelihood-based predictor as the level of asymmetry in the LinEx loss increases. Note also that for a larger forecast horizon ($h = 4$) similar results have been obtained, which are available upon request.

[5]The table of results is available upon request.

5 Conclusion

In this paper we have studied parameter estimation in the situation where the objective is to obtain a good description of future data (or data different from those used for estimation), in terms of a given criterion function. We have studied a broad family of m-estimators, and compared these in terms of the limit distributions that arise. A natural estimator is the innate estimator that used the same criterion for estimation as defines the objective. Estimators based on other criteria can be considered, but the notation of coherency between the criteria is essential. One alternative estimator is the likelihood-based estimator that is deduced from the maximum likelihood parameter of a statistical model. We have established that the likelihood-based estimator is asymptotically efficient, and our applications have shown that this estimator can be vastly better than the innate estimator is some circumstances. These advantages, however, require the likelihood function to be correctly specified. When the likelihood function is misspecified, the asymptotic efficiency that is inherited from the underlying maximum likelihood estimators, perish. However, the most damaging consequence of misspecification is that a required mapping of likelihood parameters to criterion parameters hinges on the specification, causing a likelihood-based estimator, deduced from at misspecified likelihood function to be incoherent.

Our results cast some light on how one ought to estimate parameters in the present context. Two competing approaches map into what we have labelled the innate estimator and the likelihood-based estimator. The latter corresponds to the case where a statistical model is formulated and estimated, without attention to the ultimate use of the estimated model. One the model is estimated it can, in principle, be tailored to suit any purpose, including one defined by a criterion such as Q. The innate estimator is directly tied to to objective, so if the objective is modified, so must the estimator be. Our limit results do not univocally point to one approach being preferred to the other. If the likelihood is correctly specified, the limit theory clearly favors the likelihood-based estimator, while the innate estimator is preferred under a fixed degree of misspecification. However, our results based on slightly misspecified likelihood functions (local-to-correct specification) showed that the likelihood-based estimator continues to dominate the innate estimator when the misspecification is "small". The degree of misspecification at which the innate estimator begins to be superior depends on the context, such as the criterion. For instance, in our applications based on the LinEx loss function we saw the superiority of likelihood-based estimator increases with the degree of asymmetry of the objective. For this reason, it takes a relatively high degree of (local) misspecification before the innate estimator outperforms the likelihood-based estimator when the asymmetry is large, but relatively little misspecification when the loss function is symmetric.

Our results may be viewed as an argument in favor of conducting a thorough model diagnostic in the present context, diagnostics that arguably should be targeted toward the form of misspecification that distorts the mapping of likelihood-parameters to criterion parameters.

A Appendix: Proof of Auxiliary Results

The expression for A follows by

$$\begin{aligned} A = \mathrm{E}[-h_i(X_i,\theta_0)] &= \mathrm{E}\exp\{c(X_i-\theta_0)\} \\ &= \mathrm{E}\exp\{c(X_i-\mu) - \frac{c^2\sigma^2}{2}\} \\ &= \exp\{-\frac{c^2\sigma^2}{2} + \frac{1}{2}c^2\sigma^2\} = 1, \end{aligned}$$

where the second last equality follows by using that the moment generating function for $V \sim N(\lambda, \tau^2)$ is $\mathrm{mgf}(t) = \mathrm{E}(\exp\{tV\}) = \exp\{\lambda t + \frac{1}{2}\tau^2 t^2\}$, and setting $\lambda = -\frac{c^2\sigma^2}{2}$, $\tau^2 = c^2\sigma^2$, and $t = 1$.

For B we note that

$$\begin{aligned} \mathrm{E}[s_i(X_i,\theta_0)]^2 &= c^{-2}\mathrm{E}[\exp\{2c(X_i-\theta_0)\} - 2\exp\{c(X_i-\theta_0)\} + 1] \\ &= c^{-2}\mathrm{E}[\exp\{2c(X_i-\mu) - c^2\sigma^2)\} - 2\exp\{c(X_i-\mu) - \frac{c^2\sigma^2}{2}\} + 1] \\ &= c^{-2}[\exp\{-c^2\sigma^2 + 2c^2\sigma^2\} - 2\exp\{-\tfrac{c^2\sigma^2}{2} + \tfrac{c^2\sigma^2}{2}\} + 1] \\ &= c^{-2}[\exp\{c^2\sigma^2\} - 1]. \end{aligned}$$

Here we have used the expression for the moment generating function for a Gaussian random variable twice.

Proof of (6). We seek the solution to

$$\min_\theta \mathrm{E}L_c(X-\theta) = \min_\theta \mathrm{E}[\exp\{c(X-\theta)\} - c(X-\theta) - 1],$$

when $X \sim \mathrm{NIG}(\lambda,\delta,\alpha,\beta)$. Using the moment generating function for the NIG-distribution the problem becomes to minimize

$$\exp\{-c\theta\}\exp\{c\lambda + \delta(\gamma - \sqrt{\alpha^2-(\beta+c)^2})\} - c(\lambda + \tfrac{\delta\beta}{\gamma} - \theta),$$

with respect to θ. The first order conditions are therefore

$$-c\exp\{-c\theta\}\exp\{c\lambda + \delta(\gamma - \sqrt{\alpha^2-(\beta+c)^2})\} + c = 0,$$

hence by rearranging and taking the logarithm, we have

$$-c\theta + c\lambda + \delta(\gamma - \sqrt{\alpha^2 - (\beta + c)^2}) = 0,$$

and rearranging yields the expression (6).

As $c \to 0$ we can use l'Hospital rule to establish that $\lim_{c\to 0} \lambda + \frac{\delta}{c}\left[\sqrt{\alpha^2 - \beta^2} - \sqrt{\alpha^2 - (\beta + c)^2}\right]$ equals

$$\lambda + \lim_{c\to 0} \delta \frac{2(\beta + c)\frac{1}{2}[\alpha^2 - (\beta + c)^2]^{-1/2}}{1} = \lambda + \delta\beta/\gamma. \tag{9}$$

If we set $\lambda = \mu$, $\delta = \sigma^2\alpha$, and $\beta = 0$, we have $\theta_* = \mu + \frac{\sigma^2\alpha}{c}\left[\sqrt{\alpha^2} - \sqrt{\alpha^2 - c^2}\right] = \mu + \frac{\sigma^2}{c}\left[\frac{1-\sqrt{1-xc^2}}{x}\right]$, if we set $x = \alpha^{-2}$. To obtain the limit as $x \to 0$ ($\alpha \to \infty$) we apply l'Hospital rule again

$$\lim_{\alpha\to\infty} \theta_0 = \mu + \lim_{x\to 0} \frac{\sigma^2}{c} \frac{\frac{1}{2}c^2(1 - xc^2)^{-1/2}}{1} = \mu + c\frac{\sigma^2}{2}.$$

Theorem 5. *Consider the* NIG$(\lambda, \delta, \alpha, \beta)$*, where* $\lambda = \mu - \delta\beta/\gamma$*,* $\delta = \sigma^2\gamma^3/\alpha^2$ *and* $\beta = b\alpha^{1-a}$ *for* $a \in (\frac{1}{3}, 1]$ *and* $b \in \mathbb{R}$*. Then*

$$\text{NIG}(\lambda, \delta, \alpha, \beta) \to N(\mu, \sigma^2), \qquad as \quad \alpha \to \infty$$

Proof. Define $x = \alpha^{-2}$ so that $\alpha = x^{-1/2}$ and $\beta = bx^{-(1-a)/2}$ and note that $\beta/\alpha = b\alpha^{-a} = bx^{a/2}$ so that

$$\frac{\gamma}{\alpha} = \sqrt{1 - (\beta/\alpha)^2} = \sqrt{1 - b^2x^a}.$$

Now consider the characteristic function for the NIG$(\lambda, \delta, \alpha, \beta)$ which is given by

$$\exp\{i\lambda t + \delta(\gamma - \sqrt{\alpha^2 - (\beta + it)^2}\}.$$

With $\delta = \sigma^2\gamma^3/\alpha^2$ and $\lambda = \mu - \delta\beta/\gamma = \mu - \sigma^2(\gamma/\alpha)^2\beta$, the first part of the characteristic function is given by

$$\lambda = \mu - \sigma^2(1 - b^2x^a)bx^{-\frac{1-a}{2}} = \mu - \sigma^2 bx^{-\frac{1-a}{2}} + \sigma^2 b^3 x^{\frac{3a-1}{2}}.$$

We observe that the last term vanish as $x \to 0$ provided that $a > \frac{1}{3}$, while the second term, $it\sigma^2 bx^{-\frac{1-a}{2}} = it\sigma^2 bx^{\frac{1+a}{2}}/x$, will be accounted for below.

The second part of the characteristic function equals

$$\delta(\gamma - \sqrt{\alpha^2 - (\beta + it)^2} = \sigma^2 \frac{(\frac{\gamma}{\alpha})^4 - (\frac{\gamma}{\alpha})^3 \sqrt{1 - (\beta/\alpha + it/\alpha)^2}}{\alpha^{-2}},$$

which, in terms of x, is expressed as

$$\sigma^2 \frac{(1 - b^2 x^a)^2 - (1 - b^2 x^a)^{3/2} \sqrt{1 - (bx^{a/2} + itx^{1/2})^2}}{x}.$$

Including the second term from the first part of the CF, we arrive at,

$$\sigma^2 \frac{-itbx^{\frac{1+a}{2}} + (1 - b^2 x^a)^2 - (1 - b^2 x^a)^{3/2} \sqrt{1 - (bx^{a/2} + itx^{1/2})^2}}{x},$$

and applying l'Hospital's rule as $x \to 0$, we find (apart for the scale σ^2)

$$-itb\tfrac{1+a}{2} x^{\frac{a-1}{2}} - 2ab^2 x^{a-1} + \tfrac{3}{2} ab^2 x^{a-1} - \tfrac{1}{2}(-b^2 x^{a-1} - 2itb\tfrac{1+a}{2} x^{\frac{a-1}{2}} + t^2) = -\tfrac{1}{2} t^2.$$

So the CF for the NIG converges to $\exp\{i\mu t - \frac{\sigma^2}{2} t^2\}$ as $x \to 0$, which is the CF for $N(\mu, \sigma^2)$. □

Corollary 1. NIG$(\mu, \sigma^2\alpha, \alpha, 0) \to N(\mu, \sigma^2)$, *as* $\alpha \to \infty$.

Proof. The results follows from Theorem , or directly by observing that the CF for NIG$(\mu, \sigma^2\alpha, \alpha, 0)$ is

$$\exp\{i\mu t + \sigma^2\alpha^2(1 - \sqrt{1 + \alpha^{-2}t^2}\}.$$

Now by l'Hospital's rule ote that $\partial\sqrt{1 + xt^2}/\partial x = \frac{1}{2} t^2 (1 + xt^2)^{-1/2}$, so by setting $x = \alpha^{-2}$ and applying L'Hospital rule we find

$$\lim_{x \to 0} \frac{\sigma^2(1 - \sqrt{1 + xt^2})}{x} = \frac{\lim_{x \to 0}[-\frac{1}{2} t^2 (1 + xt^2)^{-1/2}]}{1} = -\frac{1}{2}\sigma^2 t^2.$$

□

B Details Concerning the Simulations Designs

B.1 LinEx Loss

Our simulate design is based on random variables with mean zero and unit variance. This is without loss of generality because a simulation design based on random variables X_i with mean μ, variance σ^2 and

asymmetry parameter c, is equivalent to a design based on $Z_i = (X_i - \mu)/\sigma$ with asymmetry parameter $d = \sigma c$. To establish this result, suppose that an estimator, $\check{\theta}$, under LinEx loss, L_c, $(\mathcal{X})$, is such that

$$\check{\theta}_c(\mathcal{X}) = \mu + \sigma\check{\theta}_d(\mathcal{Z}). \tag{10}$$

Then

$$c\{Y_i - \check{\theta}_c(\mathcal{X})\} = c\{\sigma\tfrac{Y_i-\mu}{\sigma} + \mu - \mu - \sigma\check{\theta}_d(\mathcal{Z})\} = d\{\tfrac{Y_i-\mu}{\sigma} - \check{\theta}_d(\mathcal{Z})\}.$$

Thus,

$$L_c(Y_i - \check{\theta}_c(\mathcal{X})) = \sigma^2 L_d(\tfrac{Y_i-\mu}{\sigma} - \check{\theta}_d(\mathcal{Z})).$$

Because the scale, σ^2, is shared by all estimators that satisfy (10) the relative performance of such estimators are unaffected.

Both the innate estimator and the likelihood-based estimator takes the form stated in (10). This follows from:

$$\begin{aligned}\hat{\theta}_c(\mathcal{X}) &= \frac{1}{c}\log\{\tfrac{1}{n}\sum\exp(cX_i)\} = \frac{1}{c}\log\{\tfrac{1}{n}\sum\exp(c\sigma Z_i)\exp(c\mu)\} \\ &= \mu + \frac{1}{c}\log\{\tfrac{1}{n}\sum\exp(dZ_i)\} = \mu + \sigma\hat{\theta}_d(\mathcal{Z}),\end{aligned}$$

and similarly for the likelihood-based estimator:

$$\tilde{\theta}_c(\mathcal{X}) = \bar{X} + \frac{c}{2}\frac{1}{n}\sum_i (X_i - \bar{X})^2 = \mu + \sigma\bar{Z} + \frac{c}{2}\frac{1}{n}\sum_i \sigma^2(Z_i - \bar{Z})^2 = \mu + \sigma\tilde{\theta}_c(\mathcal{Z}).$$

The experiment considered under the gaussian distribution and LinEx loss function consists in the following 5 steps.

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 ML and CB predictors and the other n observations constitute the out-of-sample set, that of realizations, with which the predictors are compared in order to calculate the losses.

Step 2. The three predictors are immediately obtained from (4) and by applying (5), (3) to the in-sample data, respectively.

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

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 θ^* is the optimal predictor under the LinEx loss, the expected value of the evaluation criterion associated with θ^* is always larger than the one corresponding to the two other predictors. It follows that both the numerator and denominator of (1) 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.

In view of this result we decide to fix the standard deviation to 1 while considering several values of the asymmetry coefficient, i.e. $c \in \{0.01; 0.1; 1; 2; 3\}$.

B.2 LinEx under NIG distribution

This set of simulations tackles the case of local-misspecification by considering that the data follows the normal inverse gaussian distribution. We normalize the NIG$(\lambda, \delta, \alpha, \beta)$ distribution to have mean zero and unit variance by setting $\delta = \gamma^3/\alpha^2$ and $\lambda = -\delta\beta/\gamma$ where $\gamma = \sqrt{\alpha^2 - \beta^2}$. Actually, with the parameterization

$$\xi = \frac{1}{\sqrt{1+\delta\gamma}} \qquad \chi = \xi\frac{\beta}{\alpha},$$

one has $0 \leq |\chi| < \xi < 1$, with $\xi = 0$ corresponding to the Gaussian case.

The distributional parameters can be easily computed in our zero-mean unit-variance design since we have that $\xi = (1 + \gamma^4/\alpha^2)^{-1/2}$. Hence $\frac{\beta}{\alpha} = \frac{\chi}{\xi}$, and

$$\frac{\sqrt{1-\xi^2}}{\xi} = \frac{\gamma^2}{\alpha} = \frac{\alpha^2 - \beta^2}{\alpha} = \alpha(1 - \tfrac{\chi^2}{\xi^2}).$$

Finally, we see that

$$\alpha = \xi\frac{\sqrt{1-\xi^2}}{\xi^2 - \chi^2} \qquad \beta = \chi\frac{\sqrt{1-\xi^2}}{\xi^2 - \chi^2}.$$

We choose the asymmetric negative design by fixing χ to $-\xi^{3/2}$, such that the optimal predictor under NIG is a function only of ξ. This facilitates the setup of the local-misspecification experiments where an increase in the level of the bias will be immediately mirrored by a modification of ξ. We hence make sure that the NIG optimal predictor is always computable, regardless of the changes in the distributional parameters entailed by the larger level of bias.

The experiments are organized in several steps.

Step 1. We generate the data by accounting for the local-misspecification. We hence define the bias parameter $d = bA^{1/2}$ such that the bias engendered by the local-misspecification is given by $d'd$. To be more precise, d takes values from 0 (correctly specified model) to 10. Then, for each value of d we compute the distributional parameter ξ by setting the LinEx assymmetry coefficient to 1. For this, we rely on the fact that the asymptotic variance of the QMLE predictor can be computed as $Avar = var(\mu + c\frac{\sigma^2}{2}) = 1 + 1 \times \frac{2}{4} = 1.5$. For each value of ξ we draw a sample of size $2n$ from the associated standard NIG distribution.

Step 2. The three predictors are then computed from the in-sample data. Note that the optimal predictor is now associated with the NIG distribution and it is given by (6).

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

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

Step 5. We evaluate the out-of-sample performance of the predictors by relying on the RQE criterion.

The sample size is set to $1,000,000$ and that several levels of asymmetry are considered in the evaluation step, i.e. $c \in \{0.1; 0.5; 1; 2\}$. The analysis performed is asymptotic in the sense that a lower sample size, $n = 100,000$ leads to very similar results.

B.3 Long-horizon forecasting

To compare the relative efficiency of the two predictors in the context of multi-period forecasting, 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, innate estimator and optimal estimator can now be determined by relying on the in-sample dataset and theoretical distribution respectively. 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 (7) - (**??**).

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

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

Step 5. Evaluate the out-of-sample performance of the predictors by relying on the relative criteria efficiency (RQE) indicator in (**??**). Note also that several levels of persistence in the process have been considered so as to study the change in efficiency when the process approaches unit-root. Besides, we set the forecast horizon h to 2.

C Appendix: Figures and Tables

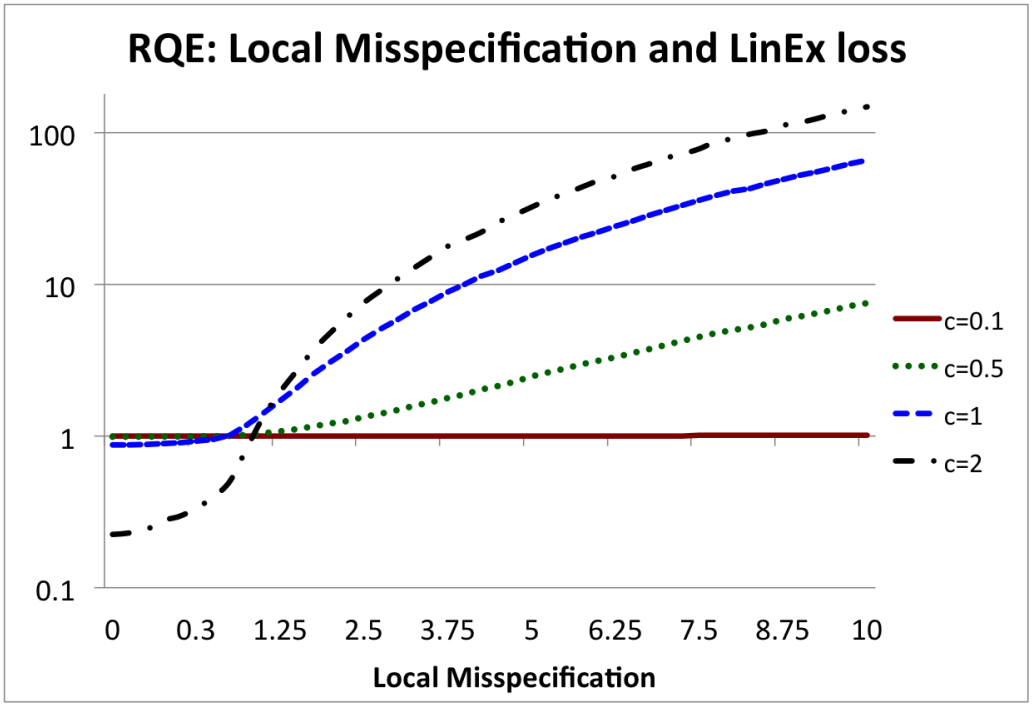

Figure 3: RQE: Local Misspecification

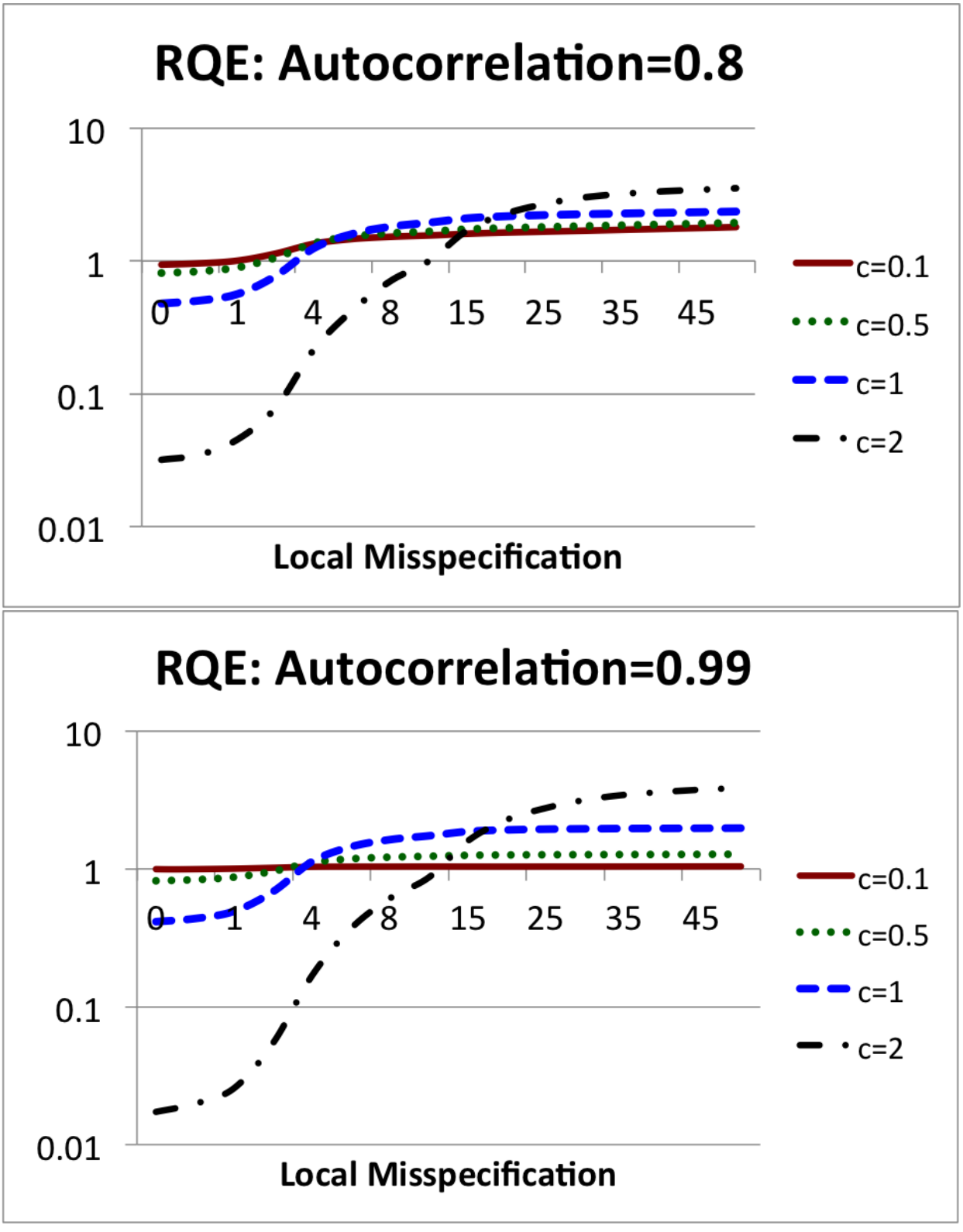

Figure 4: Locally Misspecified AR(1) Model

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

		Asymptotic Results			Finite Sample, $n = 1000$					Finite Sample, $n = 200$				
	φ	RQE	$R_\infty(\hat{\theta})$	$R_\infty(\tilde{\theta})$	RQE	$R_n(\hat{\theta})$	$R_n(\tilde{\theta})$	bias$(\hat{\beta})$	bias$(\tilde{\beta})$	RQE	$R_n(\hat{\theta})$	$R_n(\tilde{\theta})$	bias$(\hat{\beta})$	bias$(\tilde{\beta})$
$c = 0.1$	0.3	0.69	1.48	1.02	0.69	1.49	1.02	-0.002	0.000	0.67	1.52	1.03	-0.010	-0.001
	0.8	0.94	3.10	2.90	0.93	3.26	3.03	-0.006	-0.005	0.90	3.85	3.47	-0.030	-0.026
	0.99	0.99	4.06	4.02	0.99	10.00	9.87	-0.009	-0.009	0.96	27.4	26.4	-0.052	-0.051
	φ	RQE	$R_\infty(\hat{\theta})$	$R_\infty(\tilde{\theta})$	RQE	$R_n(\hat{\theta})$	$R_n(\tilde{\theta})$	bias$(\hat{\beta})$	bias$(\tilde{\beta})$	RQE	$R_n(\hat{\theta})$	$R_n(\tilde{\theta})$	bias$(\hat{\beta})$	bias$(\tilde{\beta})$
$c = .25$	0.3	0.69	1.52	1.05	0.68	1.54	1.05	-0.002	0.000	0.66	1.56	1.04	-0.011	-0.001
	0.8	0.92	3.22	2.96	0.91	3.38	3.09	-0.006	-0.005	0.89	4.00	3.56	-0.030	-0.026
	0.99	0.97	4.19	4.06	0.97	10.05	9.78	-0.009	-0.009	0.95	28.28	26.95	-0.052	-0.051
	φ	RQE	$R_\infty(\hat{\theta})$	$R_\infty(\tilde{\theta})$	RQE	$R_n(\hat{\theta})$	$R_n(\tilde{\theta})$	bias$(\hat{\beta})$	bias$(\tilde{\beta})$	RQE	$R_n(\hat{\theta})$	$R_n(\tilde{\theta})$	bias$(\hat{\beta})$	bias$(\tilde{\beta})$
$c = 0.5$	0.3	0.67	1.65	1.11	0.67	1.66	1.11	-0.002	0.000	0.65	1.69	1.11	-0.010	-0.001
	0.8	0.85	3.63	3.10	0.85	3.79	3.23	-0.006	-0.005	0.84	4.44	3.73	-0.030	-0.026
	0.99	0.88	4.85	4.25	0.91	11.09	10.12	-0.009	-0.009	0.91	31.62	28.75	-0.052	-0.051
	φ	RQE	$R_\infty(\hat{\theta})$	$R_\infty(\tilde{\theta})$	RQE	$R_n(\hat{\theta})$	$R_n(\tilde{\theta})$	bias$(\hat{\beta})$	bias$(\tilde{\beta})$	RQE	$R_n(\hat{\theta})$	$R_n(\tilde{\theta})$	bias$(\hat{\beta})$	bias$(\tilde{\beta})$
$c = 1$	0.3	0.56	2.43	1.36	0.57	2.39	1.36	-0.002	0.000	0.57	2.39	1.37	-0.011	-0.001
	0.8	0.58	6.36	3.67	0.61	6.33	3.85	-0.006	-0.005	0.66	6.87	4.51	-0.031	-0.026
	0.99	0.52	9.73	5.03	0.65	17.05	11.16	-0.009	-0.009	0.68	57.26	39.12	-0.052	-0.051
	φ	RQE	$R_\infty(\hat{\theta})$	$R_\infty(\tilde{\theta})$	RQE	$R_n(\hat{\theta})$	$R_n(\tilde{\theta})$	bias$(\hat{\beta})$	bias$(\tilde{\beta})$	RQE	$R_n(\hat{\theta})$	$R_n(\tilde{\theta})$	bias$(\hat{\beta})$	bias$(\tilde{\beta})$
$c = 1.5$	0.3	0.35	5.24	-1.81	0.39	-4.55	1.80	-0.002	0.000	0.45	4.07	1.82	-0.010	-0.001
	0.8	0.23	20.45	-4.64	0.35	-14.20	4.94	-0.008	-0.005	0.45	13.28	5.96	-0.033	-0.025
	0.99	0.15	41.15	-6.09	0.34	-36.39	12.53	-0.009	-0.009	0.13	596.7	75.89	-0.052	-0.051
	φ	RQE	$R_\infty(\hat{\theta})$	$R_\infty(\tilde{\theta})$	RQE	$R_n(\hat{\theta})$	$R_n(\tilde{\theta})$	bias$(\hat{\beta})$	bias$(\tilde{\beta})$	RQE	$R_n(\hat{\theta})$	$R_n(\tilde{\theta})$	bias$(\hat{\beta})$	bias$(\tilde{\beta})$
$c = 2$	0.3	0.13	18.11	2.40	0.25	9.73	2.39	-0.002	0.000	0.33	7.40	2.41	-0.011	-0.001
	0.8	0.06	105.9	6.13	0.19	33.36	6.48	-0.009	-0.005	0.27	30.51	8.32	-0.034	-0.026
	0.99	0.03	243.9	7.35	0.13	116.8	15.60	-0.009	-0.009	0.00	$> 10^6$	308.4	-0.051	-0.051

Note: We compare the h-step-ahead out-of-sample performance of the maximum-likelihood estimator (MLE) $\tilde{\varphi}$ with that of the direct estimator $\hat{\varphi}$ in terms of RQE based on the LinEx loss with different asymmetry levels. When $RQE < 1$ the MLE outperforms the direct estimator. The expected value of the optimal estimator $\mathrm{E}(\varphi^{*h})$ as well as the expected bias of the others are also included. The fixed forecasting scheme is used for estimation, where the estimation and evaluation samples have the same size, n. The results are obtained for several levels of persistence of the autoregressive process φ, different 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: Locally-misspecified AR(1) Model

A) LinEx asymmetry c=0.1

d	RQE	$R_\infty(\tilde{\theta})$	$R_\infty(\hat{\theta})$	ρ_2	$\mathrm{E}(\tilde{\rho}_1^2-\rho_2)$	$\mathrm{E}(\hat{\rho}_2-\rho_2)$	$\mathrm{E}[\tilde{Q}_0-Q^*]$
0	0.94	2.92	3.11	0.640	0.000	0.000	0.000
1	1.01	3.46	3.44	0.641	-0.001	0.000	-4.042
2	1.14	4.96	4.36	0.642	-0.002	0.000	-1.998
6	1.46	20.8	14.2	0.647	-0.007	0.000	-17.65
10	1.55	51.5	33.1	0.651	-0.011	0.000	-48.18
30	1.69	407	241	0.674	-0.034	0.000	-402.9
50	1.80	1052	585	0.697	-0.057	0.000	-1046

B) LinEx asymmetry c=0.5

d	RQE	$R_\infty(\tilde{\theta})$	$R_\infty(\hat{\theta})$	ρ_2	$\mathrm{E}(\tilde{\rho}_1^2-\rho_2)$	$\mathrm{E}(\hat{\rho}_2-\rho_2)$	$\mathrm{E}[\tilde{Q}_0-Q^*]$
0	0.81	2.95	3.63	0.640	0.000	0.000	0.000
1	0.89	3.53	3.96	0.641	-0.001	0.000	-4.044
2	1.06	5.17	4.87	0.642	-0.002	0.000	-2.186
6	1.53	22.5	14.7	0.647	-0.007	0.000	-19.32
10	1.67	56.0	33.6	0.651	-0.011	0.000	-52.73
30	1.83	442	242	0.674	-0.034	0.000	-437.5
50	1.94	1133	585	0.697	-0.057	0.000	-1128

C) LinEx asymmetry c=1

d	RQE	$R_\infty(\tilde{\theta})$	$R_\infty(\hat{\theta})$	ρ_2	$\mathrm{E}(\tilde{\rho}_1^2-\rho_2)$	$\mathrm{E}(\hat{\rho}_2-\rho_2)$	$\mathrm{E}[\tilde{Q}_0-Q^*]$
0	0.48	3.03	6.35	0.640	0.000	0.000	0.000
1	0.56	3.77	6.70	0.641	-0.001	0.000	-4.047
2	0.77	5.83	7.54	0.642	-0.002	0.000	-2.779
6	1.60	27.7	17.4	0.647	-0.007	0.000	-24.52
10	1.94	70.1	36.1	0.651	-0.011	0.000	-66.84
30	2.25	548	244	0.674	-0.034	0.000	-543.9
50	2.35	1381	587	0.697	-0.057	0.000	-1376

D) LinEx asymmetry $c = 2$

d	RQE	$R_\infty(\tilde{\theta})$	$R_\infty(\hat{\theta})$	ρ_2	$\mathrm{E}(\tilde{\rho}_1^2-\rho_2)$	$\mathrm{E}(\hat{\rho}_2-\rho_2)$	$\mathrm{E}[\tilde{Q}_0-Q^*]$
0	0.03	3.32	104.43	0.640	0.000	0.000	0.000
1	0.05	4.77	105.41	0.641	-0.001	0.000	-4.065
2	0.08	8.46	107.07	0.642	-0.002	0.000	-5.239
6	0.43	47.6	111.92	0.647	-0.007	0.000	-44.09
10	1.00	124	124.00	0.651	-0.011	0.000	-121.1
30	3.00	944	314.88	0.674	-0.034	0.000	-940.2
50	3.53	2275	645.41	0.697	-0.057	0.000	-2272

Note: The QML (iterated) predictor is compared with the direct one in termes of RQE based on the LinEx loss with different asymmetry levels. The locally misspecified alternative takes the form of an AR(2) model where the level of the misspecification is given by the d parameter. The larger d the more important the second autoregressive parameter and the further we are from the underlying AR(1) model. These asymptotic results (100,000 observations) are based on 100,000 simulations.

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.

Chevillon, G. (2007), 'Direct multi-step estimation and forecasting', *Journal of Economic Surveys* **21**(4), 746–785.

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

Christoffersen, P., Jacobs, K. and CIRANO. (2001), *The importance of the loss function in option pricing*, CIRANO.

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

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

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.-K. (2003), 'Multistep prediction in autoregressive processes', *Econometric Theory* **19**(2), 254–279.

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

Marcellino, M., Stock, J. H. and Watson, M. W. (2006), 'A comparison of direct and iterated multistep ar methods for forecasting macroeconomic time series', *Journal of Econometrics* **135**, 499–526.

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.

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

Tiao, G. C. and Tsay, R. S. (1994), 'Some advances in non-linear and adaptive modelling in time-series', *Journal of forecasting* **13**(2), 109–131.

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.

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.