How should parameter estimation be tailored to the objective?

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

[a] *University of North Carolina, Chapel Hill & CREATES & Copenhagen Business School**

[b] *Paris Nanterre University*

September 14, 2020

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 when a model is estimated for the purpose of describing other data than those used for estimation, such as in forecasting problems. A natural candidate for solving $\max_{T\in\sigma(\mathcal{X})} \mathbb{E}Q(\mathcal{Y},T)$ is the *innate* estimator, $\hat{\theta} = \arg\max_\theta Q(\mathcal{X},\theta)$. While the innate estimator has certain advantages, we show 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 true model, causing an improper mapping to be used under misspecification.

Keywords: Estimation, Model Selection, LinEx Loss, Multistep forecasting

JEL Classification: C18,C13, C51, C52

*An earlier version of this paper was circulated under the title "Parameter estimation with out-of-sample objective". We thank Valentina Corradi, Nour Meddahi, Werner Ploberger, Barbara Rossi, Mark Watson, and seminar speakers at University of Pennsylvania, Penn State, and Cambridge University 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, $\mathbb{E}[L(T(\mathcal{X}), \theta)]$, where L is a loss function and $\mathcal{X}$ is the random sample available for estimation.

In this paper, we consider parameter estimation with a different objective that pertains to many empirical problems. This objective involves a second random sample, $\mathcal{Y}$, that is distinct from the data available for estimation, $\mathcal{X}$. We ask: How should we estimate the model from $\mathcal{X}$ when our true objective is to apply the estimated model to the sample $\mathcal{Y}$? This is the structure that emerges in forecasting problems where $\mathcal{Y}$ represents future data and $\mathcal{X}$ is the sample available for estimation. More generally, the sample $\mathcal{Y}$ can represent a random draw from the general population for which an estimated model is to be used. For instance, from a pilot study (based on $\mathcal{X}$), one may seek to optimize tuning parameters in a policy program before the program is implemented more widely (to $\mathcal{Y}$). We argue that this is the motivation for much empirical research as well as popular information criteria for model selection, see Akaike (1978).

To fix ideas: let the objective be $\max_{T \in \sigma(\mathcal{X})} \mathbb{E}Q(\mathcal{Y}, T)$, where Q is a criterion function and $T \in \sigma(\mathcal{X})$ refers to the requirement that T must be measurable with respect to the σ-field generated by $\mathcal{X}$. 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 that are characterized by $\tilde{\theta} = \arg\max_\theta \tilde{Q}(\mathcal{X}, \theta)$, where $\tilde{Q}$ is a different criterion. Some might find it bizarre to estimate parameters using a criterion, $\tilde{Q}$, that differs from that of the actual objective, Q, but this approach is quite common in practice. Sometimes this approach is taken for the sake of convenience such as simplifying the estimation problem. More interestingly, the most efficient estimator will often take this form. In fact, as we show in this paper, the asymptotically efficient estimator is one deduced from a carefully crafted $\tilde{Q}$, which we refer to as the likelihood-based estimator (LBE). However, the LBE also has drawbacks because the harmful effects of misspecification can offset any advantage the LBE has under correct specification. We analyze this issue, and show that the key problem with misspecification is not the lack of efficiency, but an inconsistency that arises from evaluating expectations with a misspecified model.

These results have implications for model selection in econometrics. The advantages of the likelihood-based estimation speak in favor of using a likelihood approach in the model construction. Meanwhile, the pitfalls of LBE emphasize the importance of paying attention to model diagnostics, with the caveat

that some forms of misspecification are relatively innocuous to the criterion, Q, while others can be severe yet difficult to detect empirically.[1]

Some key results in this paper can be illustrated with the following simple example: Let $X_1, \ldots, X_n$, $X_{n+1}, \ldots, X_{n+m}$ be iid random variables. Set $\mathcal{X} = (X_1, \ldots, X_n)$ and $\mathcal{Y} = (X_{n+1}, \ldots, X_{n+m})$, and suppose that the criterion is given by $Q(\mathcal{Y}, T(\mathcal{X})) = \sum_{i=1}^{m} q(Y_i - T(\mathcal{X}))$, where $q(u) = u - \exp(u)$. The innate estimator solves $\max_\theta \sum_{i=1}^{n} q(X_i - \theta)$, and is given by $\hat{\theta} = \log[\frac{1}{n}\sum_{i=1}^{n} \exp(X_i)]$. Under the assumption that $X_i \sim N(\mu, \sigma^2)$, one can instead use the estimator, $\tilde{\theta} = \tilde{\mu} + \frac{1}{2}\tilde{\sigma}^2$, where $\tilde{\mu}$ and $\tilde{\sigma}^2$ are the maximum likelihood estimates of μ and σ^2, respectively, both estimated from the sample $\mathcal{X}$.[2] We will show that $\tilde{\theta}$ dominates $\hat{\theta}$ by a substantial amount, provided that the Gaussian assumption is true. The flip-side is that the likelihood-based estimator can be vastly inferior to $\hat{\theta}$ under misspecification. But it is important to emphasize that the fragility does not stem from $(\tilde{\mu}, \tilde{\sigma}^2)$ having an asymptotic variance that exceeds the lower bound. Rather it is a consequence of the mapping, $\tilde{\theta} = \tilde{\mu} + \frac{1}{2}\tilde{\sigma}^2$, being invalid if the Gaussian assumption does not hold. For a broad class of criteria, we obtain similar results in an asymptotic framework, as $n, m \to \infty$. In this simple problem, which is a special case of a LinEx example used in Section 3, we can compare the relative merits of $\hat{\theta}$ and $\tilde{\theta}$ without the need for $n, m \to \infty$.

We establish results in an asymptotic framework that is based on standard assumptions in the context of M-estimation. While our framework and objective differ from the conventional approach used to analyze efficient parameter estimation, the classical structure emerges after manipulating the asymptotic expressions. This enables us to utilize the Cramer-Rao lower bound to establish a likelihood-based estimator as the asymptotically efficient estimator, albeit new and important issues arise in the case where the likelihood is misspecified. Under correct specification, the likelihood-based estimator dominates the innate estimator, in some cases by a wide margin. When the likelihood is misspecified, the asymptotic efficiency argument perishes but, more importantly, the likelihood approach requires a mapping of likelihood parameters to criterion parameters that hinges on the likelihood being correctly specified. Misspecification distorts this mapping, which causes $\tilde{\theta}$ to be inconsistent for the value of θ that maximizes the objective. So, our results cast light on the relative merits of likelihood-based estimation versus innate estimation. An advantage of the likelihood approach is that a single estimated model can, in principle, be customized to suit a broad range of objectives. In contrast, the innate estimator is intrinsically tied to the objective. So if the objective changes then the innate estimator must be redefined accordingly. Our limit results do not univocally point to one approach being preferred

[1] A likelihood approach to model selection with extensive use of model diagnostics is sometimes associated with the London School of Economics approach to econometric modeling.

[2] In this example, $\tilde{Q}(\mathcal{X}, \theta)$ is simply the Gaussian log-likelihood function, reparametrized by $(\mu, \sigma^2) \mapsto (\theta, \tau)$, with the log-likelihood function concentrated to eliminate the nuisance parameter, τ, see Section 2.1.

to the other. It ultimately rests on how confident one is about the specification. If the likelihood is correctly specified, the limit theory clearly favors the likelihood-based estimator, while the innate estimator is preferred asymptotically under a fixed degree of misspecification. At moderate levels of misspecification, as defined in a framework with local-to-correct specifications, the choice is less obvious. The misspecification threshold at which the innate estimator becomes superior to the likelihood-based estimator is context-specific and depends on many factors including the nature of misspecification and the criterion function, Q. Model diagnostics and misspecification tests can provide some assistance in the choice of estimator, but should be tailored for the specific problem at hand.

In the context of forecasting, many have argued that the estimation criterion should coincide with the actual objective, starting with Granger (1969), see also Weiss (1996). For empirical support of this approach, see e.g. Weiss and Andersen (1984) and Christoffersen and Jacobs (2004). Changing the forecasting horizon amounts to changing the objective even if the same loss function is applied to the forecasting errors. For this reason, one might expect the optimal estimation method to vary with the forecasting horizon. In the autoregressive setting with quadratic prediction loss Bhansali (1999) and Ing (2003) have established that the relative merits of the two estimation methods depend on the degree of misspecification. This led Schorfheide (2005) to propose a model selection criterion that accounts for the bias-variance trade-off, while Hansen (2010*a*) developed a leave-h-out cross-validation criterion that is well suited for the selection of h-step ahead forecasting models.

The existing literature has primarily focused on the case with a mean square error (MSE) loss function and likelihood functions based on 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). In this general framework we establish analytical results that are asymptotic in nature. Specifically, we will compare the relative merits of estimators in terms of the limit distributions that arise in this context. The asymptotic results are complemented and illustrated in an application with an asymmetric (LinEx) criterion. 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 performance of the LBE deteriorates and does so increasingly as the degree of misspecification increases.

We illustrate the relevance of the theoretical results in an empirical application to volatility forecasting. The direct forecast (innate estimation) is found to dominate the iterated forecast (likelihood-based estimation) when the underlying model is misspecified. Conversely, with a more flexible model speci-

fication, the likelihood-approach dominates direct forecasting – in particular with an asymmetric loss function where the inefficiency of the innate estimation is more pronounced.

The rest of the paper is structured as follows. Section 2 presents the theoretical framework and the asymptotic results. We illustrate the results in Sections 3 with an application to asymmetric loss function and comment on the relation to direct and iterated forecasts in Section 4 through an empirical application on volatility prediction. Section 5 concludes and the three appendices contain proofs of the theoretical results in Section 2, auxiliary results related to the application, and details about the simulation studies, respectively.

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), and our notation will largely follow that in Hansen (2010*b*). We will first make the simple observation that a discrepancy between Q and $\tilde{Q}$ can seriously degrade the performance of the estimator, $\tilde{\theta}$. Then, 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 Cramer-Rao bound. We address the case where the likelihood function involves a parameter of higher dimension than θ, and finally we derive results for the situation where the likelihood is misspecified. Although our results are not specific to forecasting problems, we sometimes use standard forecasting terminology by referring to $\mathcal{X}$ and $\mathcal{Y}$ as *in-sample* and *out-of-sample*, respectively.

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 $\tilde{q}(\mathbf{x}_t,\theta)=-(X_t-\varphi X_{t-1})^2$. Here we might have $\theta=\varphi$, or some other transformation that is better suited for maximizing the true objective, defined $q(\mathbf{x}_t,\theta)$. For instance, we could have $q(\mathbf{x}_t,\theta)=-|X_t-\theta X_{t-1}|$. Additional examples will be given below.

Assumption 1. *Suppose that $\{X_t\}$ is stationary and ergodic.*

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 $\mathbb{E}[q(\mathbf{x}_t,\theta)]$ and $\mathbb{E}[\tilde{q}(\mathbf{x}_t,\theta)]$, respectively, where θ_* and θ_0 are interior to Θ; (iii) $\mathbb{E}[\sup_{\theta\in\Theta}|q(\mathbf{x}_t,\theta)|] < \infty$ and $\mathbb{E}[\sup_{\theta\in\Theta}|\tilde{q}(\mathbf{x}_t,\theta)|] < \infty$;*

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

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

Lemma 1. *Given Assumptions 1 and 2. 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 sample size, n, increases. This consistency need not hold for the alternative estimator, $\tilde{\theta}$, unless θ_* and θ_0 coincide.

Next we assume the following regularity conditions that enable us to derive the limit results that form 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, $n^{-1/2}\sum_{t=1}^n (s(\mathbf{x}_t,\theta_*)', \tilde{s}(\mathbf{x}_t,\theta_0)')' \xrightarrow{d} \mathrm{N}(0,\Sigma_S)$; (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 $A = -\mathbb{E}h(\mathbf{x}_t,\theta_*)$ and $\tilde{A} = -\mathbb{E}\tilde{h}(\mathbf{x}_t,\theta_0)$ 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 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. *Given Assumption 3. We have*

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

In this literature it 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 relates to $\mathcal{X}$ and the score that relates to

$\mathcal{Y}$. Lemma 2 shows that the assumed independence between $\mathcal{X}$ and $\mathcal{Y}$ can be dispensed with, because the asymptotic independence of the scores is a simple consequence of the central limit theorem being applicable. No additional assumption is needed, and the asymptotic independence holds regardless of the properties of the scores, which may be serially dependent, as long as the central limit theorem applies.

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 the criterion Q if its probability limit is θ_*.

To simplify the exposition, we set $m = n$, which is without loss of generality. The situation $m, n \to \infty$, possibly at different rates, is qualitatively identical but requires different normalizations for different terms.[3]

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*

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

in probability. The divergence is at rate n.

The result shows that any incoherent estimator is asymptotically inferior to the innate estimator, which serves as a reminder that mindless estimation without attention to the actual objective is undesirable. Lemma 3 shows that consistency for θ_* is a critical requirement, which limits the set of criteria, $\tilde{Q}$, that are suitable for estimation. It is, however, possible to craft a coherent criterion, $\tilde{Q}$, from a likelihood function, as we shall show below.

Theorem 1. *If Assumptions 1-3 hold and $\tilde{Q}$ is a coherent criterion, then the limit distribution of $Q(\mathcal{Y}, \tilde{\theta}) - Q(\mathcal{Y}, \theta_0)$, as $n \to \infty$ is given by*

$$Z_y' B^{1/2} \tilde{A}^{-1} \tilde{B}^{1/2} Z_x - \tfrac{1}{2} Z_x' \tilde{B}^{1/2} \tilde{A}^{-1} A \tilde{A}^{-1} \tilde{B}^{1/2} Z_x,$$

where $Z_x, Z_y \sim iid\mathrm{N}(0, I)$, and its expected value 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,

[3]The setup with $m = n$ is common in this literature, and it was used in the motivation of the Akaike information criterion (AIC), see Akaike (1978).

$-\frac{1}{2}\text{tr}\{A^{-1}B\}$, can be related to a result by Takeuchi (1976), who generalized the AIC to the case with misspecified models.

The expected value of the limit distribution motivates the following definition of criterion risk:

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

$$R(\tilde{\theta}) = \tfrac{1}{2}\text{tr}\{A\tilde{A}^{-1}\tilde{B}\tilde{A}^{-1}\}.$$

The finite sample analog is defined by $R_n(\tilde{\theta}) = \mathbb{E}[Q(\mathcal{Y}, \theta_0) - Q(\mathcal{Y}, \tilde{\theta})]$.

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

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

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

$$\frac{R(\tilde{\theta})}{R(\hat{\theta})} = \frac{\text{tr}\{A\tilde{A}^{-1}\tilde{B}\tilde{A}^{-1}\}}{\text{tr}\{A^{-1}B\}},$$

provided that $\tilde{\theta}$ is a coherent estimator. For an incoherent estimator it follows by Lemma 3 that RQE $\to \infty$, as $n \to \infty$. Coherency is therefore imperative to good asymptotic properties of an estimator.

2.1 Likelihood-Based Estimator

In this section we focus on estimators that are deduced from a likelihood criterion. In some cases, one can obtain $\tilde{\theta}$ directly as a maximum likelihood estimator, but a mapping of the vector of likelihood parameters, ϑ say, to the vector of criterion parameters, θ, is typically needed. This is obviously the case if the dimensions of θ and ϑ do not coincide.

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 $\mathbb{E}_{\vartheta_0}(\cdot) = \int(\cdot)\text{d}P_{\vartheta_0}$, and we therefore have that

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

which defines θ_0 as a function of ϑ_0. This mapping, $\theta_0 = \theta(\vartheta_0)$, is defined implicitly and requires no estimation. Closed-form expressions for $\theta(\vartheta_0)$ are available in many cases, but the mapping can be

difficult to determine in some problems, see (6) for an example.

For some analytical results, we need the mapping from ϑ to θ to satisfy the following regularity conditions.

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 distills θ from ϑ is invertible in a manner that does not degenerate in the limit distribution. The intuition is simply that this assumption ensures that $\tilde{\theta} = \theta(\tilde{\vartheta})$ is the extremum estimator that maximizes the concentrated log-likelihood function $\ell_c(\theta) = \ell(\theta, \tilde{\tau}(\theta))$, where $\tilde{\tau}(\theta) = \arg\max_\tau \ell(\theta, \tau)$.

The assumption is relatively innocuous, but we do consider a case in Section 3 where it is violated. This special case yields additional intuition about the structure of the problem.

Lemma 4. *Given Assumption 1–4, let $\tilde{\vartheta}$ be the MLE. Then $\tilde{\theta} = \theta(\tilde{\vartheta})$ is a coherent estimator.*

When $\tilde{\theta}$ is estimated from a correctly specified likelihood function, the information matrix identity, $\tilde{A} = \tilde{B}$, follows from standard regularity conditions. In terms of asymptotic criterion risk, the innate estimator relative to the likelihood-based estimator, amounts to 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)**.** *Suppose that Assumptions 1–4 hold, and let $\tilde{\vartheta}$ be the maximum likelihood estimator so that $\tilde{\theta} = \theta(\tilde{\vartheta})$ is the likelihood-based estimator. 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 $\mathbb{E}[\xi] = R(\hat{\theta}) - R(\tilde{\theta}) \geq 0$.

Theorem 2 shows that the likelihood-based approach is asymptotically superior to the criterion-based approach and the following Corollary shows that the likelihood-based estimator is in fact superior to any other M-estimator.

Corollary 1. *Let $\tilde{\theta} = \theta(\tilde{\vartheta})$ be the likelihood-based estimator and let $\breve{\theta}$ be any other coherent estimator. Suppose that Assumptions 1-4 hold with Assumptions 2 and 3 adapted to $\breve{\theta}$ in place of $\hat{\theta}$. Then*

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

where $\mathbb{E}[\xi] = R(\check{\theta}) - R(\tilde{\theta}) \geq 0$.

An inspection of the proof of Corollary 1 reveals that the inequality is strict, unless the alternative estimator, $\check{\theta}$, is asymptotically equivalent to $\tilde{\theta}$. So the likelihood-based estimator is (also) asymptotically efficient in the present framework with an out-of-sample objective. 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.

2.2 The Case with a Misspecified Likelihood

Misspecification is harmful to the likelihood-based estimator in two ways. 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 the mapping from ϑ to θ will be improper if the likelihood is misspecified. The reason is that the mapping is deducted from an expected value that is evaluated with the wrong distribution when the model is misspecified. The likelihood-based estimator $\tilde{\theta}$ may therefore be inconsistent for θ_* under misspecification, so that the (misspecified) likelihood-based criterion is incoherent in the sense of Definition 1.

A fixed degree of misspecification is likely to result in an inconsistent estimator and Lemma 3 dictates that the LBE is asymptotically inferior to the innate estimator as $n \to \infty$. However, this does not imply that slight misspecification renders the LBE inferior. In finite samples, one would expect a small degree of misspecification to have a modest impact on the LBE. We therefore seek an asymptotic design that can accommodate the situation with "mild" misspecification. This is achieved with an asymptotic design with local-to-correct specifications that induce a discrepancy between θ_0 and θ_* that is of order $n^{-1/2}$. This asymptotic design is analogous to designs with local-to-unit roots and weak instrumental variables.

2.2.1 Local-to-Correct Specification

We consider situations 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. Let the true probability measure be P_n. Within the statistical model, $\{P_\vartheta\}_{\vartheta\in\Xi}$, that defines the log-likelihood function, we let the best approximating likelihood parameter be denoted by $\vartheta_0^{(n)}$, and the corresponding criterion parameter is denoted $\theta_0^{(n)} = \theta(\vartheta_0^{(n)})$. As n increases, P_n approaches 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 $b \in \mathbb{R}^k$. So, the vector b defines the degree of local misspecification, where correct specification corresponds to

the case: $b = 0$. Under the local-to-correct specification the limit distribution of $Q(\mathcal{Y}, \theta_*) - Q(\mathcal{Y}, \tilde{\theta})$ is as follows:

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

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

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

Interestingly, the likelihood-based estimator retains its efficiency in terms of asymptotic variance under local misspecification but is negatively affected by the asymptotic bias. Thus, under local misspecification the relative criterion efficiency is tug-of-war between the magnitude of the bias, bb', and $A^{-1}BA^{-1} - \tilde{B}^{-1}$. The former is the asymptotic penalty caused by local misspecification, whereas the latter is the advantage that the likelihood-based estimator has over the innate estimator in terms of asymptotic variance.

One can measure the degree of local misspecification by a measure of non-centrality. In the univariate case ($\theta \in \mathbb{R}$) this can be expressed in units of standard deviations, $d = b\sqrt{\tilde{B}}$, which can be interpreted as the expected value of the t-statistic, $(\tilde{\theta} - \theta_*)/\sqrt{\mathrm{avar}(\tilde{\theta})}$. In the multivariate case the non-centrality may be measured as $d = \sqrt{b'\tilde{B}b}$. The degree of non-centrality is a measure of the (average) statistical evidence of misspecification. When $k = 1$ the non-centrality parameter, d, translates into criterion risk trough the term $b^2 = d^2/\tilde{B}$. In higher dimensions ($k \geq 2$) different types of misspecification will amount to the same non-centrality, $d = \sqrt{b'\tilde{B}b}$, but will different impact on criterion risk, $\mathrm{tr}\{Abb'\} = b'Ab$, unless $\tilde{B} \propto A$. The following Theorem states upper and lower bounds on the criterion risk that can result from a given level of misspecification.

Theorem 4. *Let the local misspecification,* b*, 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* $|A - \lambda\tilde{B}| = 0$*.*

The degree of misspecification could be measured in other ways than by the non-centrality parameter d, for instance, in terms of Kullback-Leibler divergence. In one of our simulation designs below, where the likelihood-based estimator is deduced from a Gaussian likelihood, we translate the value of d into the expected value of the Jarque-Bera test statistic.

In the next section we consider an application with correct and local-to-correct specifications. The illustration is based on the asymmetric LinEx loss function, that facilitates closed-form expressions for

key asymptotic quantities. The application also enables us to consider a special case where Assumption 4 is violated.

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 asymmetric loss is appropriate, see e.g. Granger and Newbold (1977), Christoffersen and Diebold (1997), and Hwang et al. (2001). The LinEx loss function is a tractable asymmetric loss function that was introduced by Varian (1974), and it 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 a unique 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 $L_0(x)$. We adopt the LinEx loss function because it produces simple estimators in closed-form and analytical expressions of criterion risk that ease the computational burden.

The objective in this application is to estimate θ for the purpose of minimizing the expected loss, $\mathbb{E}L_c(Y_i - \theta)$. This problem maps into our theoretical framework by defining the criterion function with $q(X_i, \theta) = -L_c(X_i - \theta)$, and it is easy to show that $\theta_* = \arg\min \mathbb{E}L_c(X_i - \theta) = c^{-1}\log[\mathbb{E}\exp(cX_i)]$, provided that $\mathbb{E}\exp(cX_i) < \infty$. 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, which carries over to any function of X_i, including $\exp(cX_i)$, it follows that $\hat{\theta} \xrightarrow{as} \theta_*$. Next, we introduce a likelihood-based estimator that is deduced from the assumption that

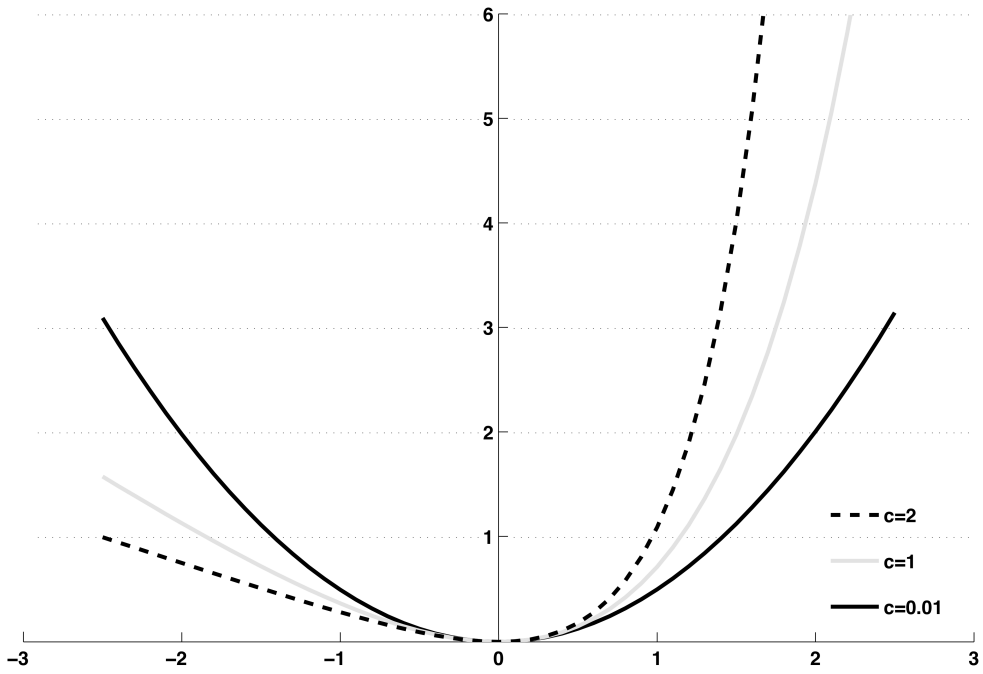

Figure 1: The LinEx loss function for three values of c.

$X_i \sim \text{iidN}(\mu_0, \sigma_0^2)$, 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 the likelihood parameter vector, $\vartheta = (\mu, \sigma^2)'$, 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 $\mathrm{N}(\mu_0, \sigma_0^2)$. 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 $X_i \sim iid\mathrm{N}(\mu, \sigma^2)$ it can be shown that

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

see Appendix B.1. Consequently, in this application we have

$$\mathrm{RQE} = \frac{\mathrm{tr}\{A\tilde{B}^{-1}\}}{\mathrm{tr}\{A^{-1}B\}} = \frac{1}{B\tilde{B}} = \frac{(c\sigma)^2 + (c\sigma)^4/2}{\exp(c\sigma)^2 - 1}, \qquad \text{for} \quad c\sigma \neq 0,$$

which is (unsurprisingly) less than one, and $\mathrm{RQE} = 1$ arises only in the limit as $c\sigma \to 0$, where the two estimators coincide.

The relative efficiency of $\hat{\theta}$ and $\tilde{\theta}$ is compared in Table 1 for the case with a correctly specified likelihood function. Panel A of Table 1 displays the asymptotic results based 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 C.1. The asymmetry parameter is given in the first column followed by the population value of θ_*, the RQE and the criterion losses resulting from estimation error.

Table 1 shows that the likelihood-based estimator dominates the innate estimator and does so increasingly as c increases in absolute value. The exception is the case $c = 0$, where the two estimators coincide. Our analytical results imply the superiority of the LBE in the asymptotic design of Panel A; however the LBE also dominates the innate estimator in finite samples, albeit to a less extent. The innate estimator appears to be somewhat less inferior in finite samples 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. To the contrary, the per observation criterion loss, $R_n(\hat{\theta})/n$,

[4] The standard deviations of all simulated quantities are smaller than 10^{-5}.

Table 1: Relative Efficiency under LinEx Loss

Panel A: Asymptotic Results

c	θ_*	RQE	$R(\hat{\theta})$	$R(\tilde{\theta})$	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	0.880	0.853	0.748	-0.001	0.000
1.5	0.750	0.600	1.777	1.068	-0.003	-0.001
2	1.000	0.350	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	0.250	0.991	0.567	0.562	-0.003	-0.003
1	0.500	0.900	0.839	0.753	-0.009	-0.005
1.5	0.750	0.720	1.493	1.075	-0.023	-0.008
2	1.000	0.560	2.745	1.526	-0.058	-0.010
2.5	1.250	0.418	5.273	2.203	-0.122	-0.012

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 former dominates the innate estimator 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.

is decreasing in n. Moreover, the innate estimator has a larger finite sample bias relative to that of the likelihood-based estimator in this application.

3.1.1 Likelihoods with One-Dimensional Parameter

With a likelihood deduced from $X_t \sim \mathrm{N}(\mu, \sigma^2)$ 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 the one-dimensional θ. This could be favorable to the innate estimator in finite samples, but the dimension of ϑ is actually immaterial to the asymptotic comparison. This follows from the fact 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 point, we consider two nested models that both have a one-dimensional ϑ. The first nested model has σ_0^2 to be known, so that only μ is to be estimated, and the second nested model has μ_0 to be known so that $\vartheta = \sigma^2$. The latter design is of separate interest because it constitutes a case where Assumption 4 is violated. The asymptotic variance of $\tilde{\theta}$ that arises in the various models is the following:

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

When σ_0^2 is known, the asymptotic variance of $\tilde{\theta}$ is smaller than when σ_0^2 is estimated, which gives $\tilde{\theta}$ an even greater advantage over the innate estimator, $\hat{\theta}$. In this design, the "stakes are raised" for the likelihood-based estimator, because misspecification can also result from an incorrect assumed value for σ_0^2 (in addition to the other forms of misspecification). An interesting situation arises when μ_0 is known. In this case the mapping from ϑ to θ is simply $\theta = \theta(\sigma^2) = \mu_0 + c\frac{\sigma^2}{2}$, which does not depend on σ^2 when $c = 0$. This is the case where Assumption 4 is violated because $\partial\theta(\sigma^2)/\partial\sigma^2 = 0$ if $c = 0$. Consequently, the asymptotic results need not apply in this case. It turns out that this particular violation of Assumption 4 is advantageous to the likelihood-based estimator because with μ_0 known and $c = 0$ the optimal estimator is known without any need for estimation. If c is small, then the LBE benefits from having a small asymptotic variance, because it is proportional to c^2.

Asymptotic criterion risks and their corresponding RQEs in the situation where either μ_0 or σ_0^2

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

		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	$R(\hat{\theta})$	$R(\tilde{\theta})$	RQE	$R(\hat{\theta})$	$R(\tilde{\theta})$
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

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 former dominates the innate estimator and the performance gap increases with the degree of asymmetry. Panel A corresponds to the case where the variance σ_0^2 is known, whereas panel B presents the case where the mean μ_0 is known. The results are based on 500,000 simulations.

is known are reported in Table 2. As expected, the likelihood-based estimator performs even better when the dimension of ϑ is smaller. Panel A has the case $\vartheta = \mu$ (and σ_0^2 is known) and Panel B has the case where $\vartheta = \sigma^2$ (and μ_0 is known). In Panel A the asymptotic criterion risk, $R(\tilde{\theta})$, 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, $R(\tilde{\theta})$ is increasing in c starting from zero at $c = 0$. The theoretical explanation 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 $R(\hat{\theta}) = \frac{1}{2}\text{tr}\{A^{-1}B\} = \frac{1}{2}[\exp(c^2)-1]/c^2$, whereas the expressions are different for the likelihood-based estimators. 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, misspecification distorts the likelihood-based estimator through two channels. First, it erodes the efficiency argument that favors the likelihood-based estimator and second, the transformation of the likelihood parameter into the criterion parameter is distorted and will be improper for most types of misspecification. To study the impact of misspecification we now consider the case where the truth is defined by a normal inverse Gaussian (NIG) distribution, that was introduced by Barndorff-Nielsen (1977, 1978).

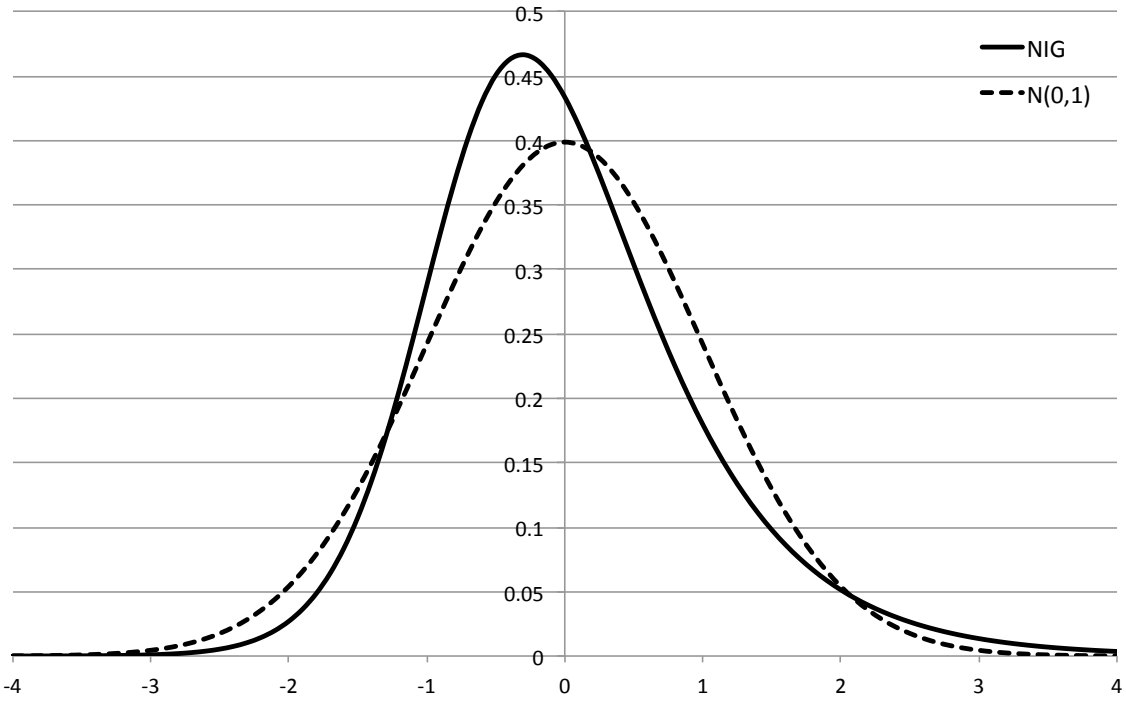

Figure 2: The density of a standardized NIG distribution (mean zero and unit variance) with $\xi = 0.5$ and $\chi = 0.25$ and the standard Gaussian density.

A NIG distribution is characterized by four parameters, λ, δ, α, and β, that represent the location, scale, tail heaviness, and asymmetry, respectively. The density of a NIG distribution is presented in Figure 2. The NIG-distribution is flexible and well suited for the present problem, because the Gaussian distribution, $\mathrm{N}(\mu, \sigma^2)$, 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 expressions for the quantities that are relevant for our analysis of the LinEx loss function.

The mean and variance of $\mathrm{NIG}(\lambda, \delta, \alpha, \beta)$ 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 (based on the Gaussian likelihood) 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

$$\theta_* = \begin{cases} \lambda + \frac{\delta}{c}\left[\sqrt{\alpha^2 - \beta^2} - \sqrt{\alpha^2 - (\beta + c)^2}\right] & \text{if} \quad c \neq 0, \\ \lambda + \frac{\delta\beta}{\gamma} & \text{if} \quad c = 0, \end{cases} \tag{6}$$

see Lemma B.1 in the Appendix. So θ_0 and θ_* generally do not coincide and the (misspecified) likelihood-based estimator is therefore incoherent, with the exception of the following two special cases. The

estimators are identical under symmetric loss, $c = 0$, and in the limited case: $\delta = \sigma^2\alpha$, $\alpha \to \infty$, and $\beta = O(\alpha^{1-a})$ with $a \in (\frac{1}{2}, 1]$, see Theorem B.1, because this causes the NIG to converge to the Gaussian distribution.

To make our misspecified design comparable to our previous design (where $X_i \sim iid\mathrm{N}(0,1)$) we consider the standard NIG distribution with mean zero and unit variance. The zero mean and unit variance are achieved by setting $\lambda = -\frac{\delta\beta}{\gamma}$ and $\delta\frac{\alpha^2}{\gamma^3} = 1$. This family of standardized 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 $0 \leq |\chi| < \xi < 1$, see Barndorff-Nielsen et al. (1985). Figure 2 displays the density for the case with $\xi = 0.5$ and $\chi = 0.25$. The original parameter values can be obtained using the expressions

$$\alpha = \xi\frac{\sqrt{1-\xi^2}}{\xi^2-\chi^2} \qquad \text{and} \qquad \beta = \chi\frac{\sqrt{1-\xi^2}}{\xi^2-\chi^2}, \tag{7}$$

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

We now construct a local-to-correct specification by letting $(\theta_0 - \theta_*) = b/\sqrt{n}$ in an asymmetric design where $\chi = -\xi^{3/2} \to 0$. To ease the interpretation of the simulation results, we use that $\xi \propto n^{-1/3}$ to set ξ to $d \times n^{-1/3}$ with d the degree of local misspecification. The parameters for the NIG are deduced from ξ and the optimal predictor is computed using (6). In particular, with $\xi \propto n^{-1/3}$ so that $\chi \propto -n^{-1/2}$, it can be shown that $\alpha \propto n^{1/3}$ and $\beta \propto -n^{1/6}$, see Appendix C.2 where the simulation design is fully detailed. Besides, the results can be easily mapped in terms of the asymptotic bias b.

To assess the extent to which the misspecification is statistically detectable, we have computed the expected value of the Jarque-Bera test statistic and the power of the Jarque-Bera test for various values of d and three sample sizes. These numbers are reported in Table 3. The large sample size, $n = 10^6$, is intended to emulate the asymptotic results, while the sample sizes $n = 1000$ and $n = 200$ provide a finite-sample analogy. The second column of the table reports the mapping of the level of misspecification to the asymptotic bias in the case $c = 1$. The Table shows that the statistical evidence for misspecification is strong once $d \geq 2$.

Table 4 displays the performance of the two estimators under local misspecification, as defined by d, for different levels of asymmetry, c. The table also reports the optimal predictor, RQE, the criterion risk induced by estimation, the bias of the likelihood-based predictor and the asymptotic criterion risk

Table 3: Local Misspecification

		"Asymptotic"		$n = 1,000$		$n = 200$	
d	b	Jarque-Bera	Power	Jarque-Bera	Power	Jarque-Bera	Power
0	0.00	2.000	5.00%	1.985	4.89%	1.912	4.51%
0.1	0.02	2.001	5.13%	1.981	4.89%	1.910	4.52%
0.2	0.04	2.012	5.14%	2.000	5.00%	1.941	4.70%
0.3	0.08	2.042	5.32%	2.030	5.25%	1.985	5.02%
0.5	0.17	2.184	6.43%	2.212	6.63%	2.216	6.49%
0.6	0.23	2.325	7.50%	2.373	7.83%	2.417	7.74%
0.8	0.35	3.098	11.3%	3.276	12.0%	3.624	12.0%
1	0.49	3.609	19.0%	3.765	19.0%	4.276	18.7%
1.5	0.89	7.184	52.1%	8.211	51.6%	10.91	48.0%
2	1.36	14.17	88.7%	17.90	86.8%	28.29	81.1%
2.5	1.89	25.61	99.5%	36.33	98.9%	70.14	97.0%
3	2.47	42.95	100%	69.52	100%	169.2	99.8%

The misspecification level, d, is measured in terms of the non-centrality level in the Jarque-Bera test. The "asymptotic" results are based on $100,000$ replications with $n = 10^6$, and the finite-sample ones rely on $500,000$ simulations for $n = 200$ and $n = 1000$, respectively. The normality is rejected at the 5% level when the Jarque-Bera statistic is larger than the $\chi^2(2)$, i.e. 5.99. The misspecification level d is also mapped into the asymptotic bias b for the case where $c = 1$.

induced by the misspecification.

The first panel, where $c = 0$, corresponds to the case of the MSE loss function. Here, the likelihood-based and innate estimators coincide, as they are both equal to the sample average. The LBE is therefore not affected by misspecification if $c = 0$. The subsequent panels display the results with increasing levels of asymmetry, c. The RQE in the last column of Table 4 shows how the performance of the (quasi) likelihood-based estimator is affected by misspecification for different levels of asymmetry. For low levels of misspecification, the LBE dominates the innate estimator and its performance improves as c increases. In contrast, with larger deviations from normality its performance worsens and it becomes increasingly inferior to the innate estimator. This is explained by the fact that the mapping from $\vartheta \mapsto \theta$ is distorted under misspecification, and the distortion increases as the degree of misspecification increases. The likelihood-based estimator still outperforms the innate estimator in terms of variance. This is evident from the LBE's variance component in column 5, which can be compared to column 3 (because the bias is asymptotically negligible for the innate estimator). But the improper mapping of the LBE estimator induces a bias-related risk component, which is exponentially increasing in d (see column 6), and this bias is responsible for the degradation of the LBE's performance.

To complement the asymptotic results in Table 4, we present results based on a sample size with $n = 200$ in Table 5. The results are similar, although for large values of c we observe that the innate

Table 4: Local Misspecification

			Innate	Likelihood-Based			
	d	θ_*	$R(\hat{\theta})$	$R(\tilde{\theta})$	$R^{\text{var}}(\tilde{\theta})$	$R^{\text{bias}}(\tilde{\theta})$	RQE
$c = 0.0$	0	0.000	0.493	0.493	0.493	0.000	1.000
	0.5	0.000	0.502	0.502	0.502	0.000	1.000
	1	0.000	0.493	0.493	0.493	0.000	1.000
	1.5	0.000	0.501	0.501	0.501	0.000	1.000
	2	0.000	0.498	0.498	0.498	0.000	1.000
	2.5	0.000	0.500	0.500	0.500	0.000	1.000
	3	0.000	0.494	0.494	0.494	0.000	1.000
	d	θ_*	$R(\hat{\theta})$	$R(\tilde{\theta})$	$R^{\text{var}}(\tilde{\theta})$	$R^{\text{bias}}(\tilde{\theta})$	RQE
$c = 0.25$	0	0.125	0.514	0.514	0.514	0.000	1.000
	0.5	0.125	0.507	0.507	0.507	0.000	1.000
	1	0.125	0.522	0.522	0.521	0.001	1.000
	1.5	0.125	0.516	0.517	0.515	0.002	1.003
	2	0.125	0.507	0.511	0.507	0.004	1.008
	2.5	0.125	0.520	0.528	0.520	0.008	1.014
	3	0.125	0.513	0.526	0.514	0.012	1.025
	d	θ_*	$R(\hat{\theta})$	$R(\tilde{\theta})$	$R^{\text{var}}(\tilde{\theta})$	$R^{\text{bias}}(\tilde{\theta})$	RQE
$c = 0.5$	0	0.250	0.566	0.561	0.561	0.000	0.991
	0.5	0.250	0.560	0.557	0.555	0.001	0.994
	1	0.250	0.576	0.579	0.571	0.008	1.004
	1.5	0.250	0.567	0.588	0.562	0.026	1.036
	2	0.250	0.559	0.615	0.554	0.061	1.101
	2.5	0.250	0.573	0.686	0.569	0.118	1.197
	3	0.249	0.564	0.761	0.563	0.198	1.350
	d	θ_*	$R(\hat{\theta})$	$R(\tilde{\theta})$	$R^{\text{var}}(\tilde{\theta})$	$R^{\text{bias}}(\tilde{\theta})$	RQE
$c = 1$	0	0.500	0.857	0.750	0.750	0.000	0.876
	0.5	0.500	0.852	0.762	0.746	0.015	0.895
	1	0.500	0.873	0.885	0.767	0.119	1.015
	1.5	0.499	0.855	1.145	0.750	0.395	1.340
	2	0.499	0.844	1.670	0.741	0.929	1.978
	2.5	0.498	0.862	2.552	0.763	1.789	2.961
	3	0.498	0.843	3.806	0.759	3.047	4.516
	d	θ_*	$R(\hat{\theta})$	$R(\tilde{\theta})$	$R^{\text{var}}(\tilde{\theta})$	$R^{\text{bias}}(\tilde{\theta})$	RQE
$c = 2$	0	1.000	6.723	1.535	1.535	0.000	0.228
	0.5	0.999	6.678	1.736	1.500	0.236	0.260
	1	0.998	6.560	3.355	1.562	1.793	0.511
	1.5	0.997	6.440	7.382	1.509	5.873	1.146
	2	0.995	6.370	15.15	1.476	13.67	2.378
	2.5	0.993	6.210	27.58	1.548	26.04	4.442
	3	0.991	6.067	45.42	1.521	43.90	7.486

The likelihood-based estimator, $\tilde{\theta}$, is compared with the innate estimator, $\hat{\theta}$, under local misspecification. The data generating process is a NIG distribution whose discrepancy from the Gaussian specification is characterized by d. $R^{\text{bias}}(\tilde{\theta})$ measures the bias component of the risk, $b'Ab/2$. The simulation results are based on 100,000 replications with $n = 10^6$.

Table 5: Local Misspecification (finite samples $n = 200$)

					Innate			Likelihood-Based			
	d	θ_*	bias$(\hat{\theta})$	bias$(\tilde{\theta})$	$R_n(\hat{\theta})$	$R_n^{\text{var}}(\hat{\theta})$	$R_n^{\text{bias}}(\hat{\theta})$	$R_n(\tilde{\theta})$	$R_n^{\text{var}}(\tilde{\theta})$	$R_n^{\text{bias}}(\tilde{\theta})$	RQE
$c = 0.0$	0	0.000	0.000	0.000	0.500	0.500	0.000	0.500	0.500	0.000	1.000
	0.5	0.000	0.000	0.000	0.503	0.503	0.000	0.503	0.503	0.000	1.000
	1	0.000	0.000	0.000	0.501	0.501	0.000	0.501	0.501	0.000	1.000
	1.5	0.000	0.000	0.000	0.499	0.499	0.000	0.499	0.499	0.000	1.000
	2	0.000	0.000	0.000	0.498	0.498	0.000	0.498	0.498	0.000	1.000
	2.5	0.000	0.000	0.000	0.499	0.499	0.000	0.499	0.499	0.000	1.000
	3	0.000	0.000	0.000	0.500	0.500	0.000	0.500	0.500	0.000	1.000
	d	θ_*	bias$(\hat{\theta})$	bias$(\tilde{\theta})$	$R_n(\hat{\theta})$	$R_n^{\text{var}}(\hat{\theta})$	$R_n^{\text{bias}}(\hat{\theta})$	$R_n(\tilde{\theta})$	$R_n^{\text{var}}(\tilde{\theta})$	$R_n^{\text{bias}}(\tilde{\theta})$	RQE
$c = 0.25$	0	0.125	0.001	0.001	0.516	0.516	0.000	0.516	0.516	0.000	0.999
	0.5	0.124	0.001	0.000	0.507	0.507	0.000	0.507	0.507	0.000	1.000
	1	0.122	0.001	0.002	0.487	0.487	0.000	0.487	0.487	0.000	1.000
	1.5	0.118	0.001	0.003	0.470	0.470	0.000	0.470	0.469	0.001	0.999
	2	0.113	0.001	0.005	0.446	0.446	0.000	0.446	0.443	0.003	0.998
	2.5	0.107	0.000	0.008	0.424	0.424	0.000	0.423	0.416	0.006	0.997
	3	0.100	0.001	0.011	0.395	0.395	0.000	0.395	0.382	0.012	0.999
	d	θ_*	bias$(\hat{\theta})$	bias$(\tilde{\theta})$	$R_n(\hat{\theta})$	$R_n^{\text{var}}(\hat{\theta})$	$R_n^{\text{bias}}(\hat{\theta})$	$R_n(\tilde{\theta})$	$R_n^{\text{var}}(\tilde{\theta})$	$R_n^{\text{bias}}(\tilde{\theta})$	RQE
$c = 0.5$	0	0.250	0.001	0.001	0.568	0.565	0.000	0.563	0.560	0.000	0.991
	0.5	0.247	0.001	0.002	0.548	0.548	0.000	0.545	0.545	0.000	0.995
	1	0.242	0.001	0.007	0.509	0.509	0.000	0.513	0.509	0.005	1.009
	1.5	0.235	0.001	0.014	0.473	0.473	0.000	0.491	0.473	0.018	1.038
	2	0.228	0.001	0.021	0.430	0.430	0.000	0.473	0.428	0.045	1.101
	2.5	0.219	0.001	0.030	0.389	0.389	0.000	0.473	0.384	0.089	1.215
	3	0.209	0.001	0.040	0.345	0.345	0.000	0.496	0.339	0.157	1.438
	d	θ_*	bias$(\hat{\theta})$	bias$(\tilde{\theta})$	$R_n(\hat{\theta})$	$R_n^{\text{var}}(\hat{\theta})$	$R_n^{\text{bias}}(\hat{\theta})$	$R_n(\tilde{\theta})$	$R_n^{\text{var}}(\tilde{\theta})$	$R_n^{\text{bias}}(\tilde{\theta})$	RQE
$c = 1$	0	0.500	0.004	0.003	0.847	0.828	0.002	0.752	0.741	0.000	0.887
	0.5	0.489	0.004	0.009	0.775	0.774	0.001	0.719	0.710	0.008	0.927
	1	0.470	0.003	0.028	0.664	0.663	0.001	0.716	0.639	0.077	1.078
	1.5	0.447	0.003	0.051	0.568	0.568	0.001	0.825	0.574	0.251	1.451
	2	0.422	0.002	0.076	0.472	0.472	0.001	1.066	0.508	0.558	2.258
	2.5	0.395	0.002	0.102	0.391	0.390	0.000	1.472	0.462	1.009	3.766
	3	0.367	0.002	0.130	0.317	0.317	0.000	2.085	0.450	1.635	6.579
	d	θ_*	bias$(\hat{\theta})$	bias$(\tilde{\theta})$	$R_n(\hat{\theta})$	$R_n^{\text{var}}(\hat{\theta})$	$R_n^{\text{bias}}(\hat{\theta})$	$R_n(\tilde{\theta})$	$R_n^{\text{var}}(\tilde{\theta})$	$R_n^{\text{bias}}(\tilde{\theta})$	RQE
$c = 2$	0	1.000	0.035	0.005	3.139	2.869	0.119	1.523	1.470	0.003	0.485
	0.5	0.959	0.028	0.036	2.580	2.503	0.077	1.475	1.342	0.133	0.572
	1	0.896	0.021	0.099	1.903	1.861	0.042	2.038	1.110	0.928	1.071
	1.5	0.826	0.014	0.169	1.377	1.355	0.021	3.499	0.939	2.560	2.542
	2	0.754	0.010	0.241	0.961	0.951	0.010	5.792	0.817	4.975	6.030
	2.5	0.682	0.007	0.313	0.658	0.653	0.005	8.794	0.773	8.021	13.38
	3	0.610	0.005	0.385	0.443	0.442	0.002	12.45	0.819	11.63	28.08

The likelihood-based estimator $\tilde{\theta}$ is compared with the innate estimator, $\hat{\theta}$, in the case where the Gaussian likelihood is (locally) misspecified for different levels of asymmetry. The data generating process is a standard NIG distribution where the degree of local misspecification is determined by d. The finite-sample results are based on 500,000 replications with $n = 200$.

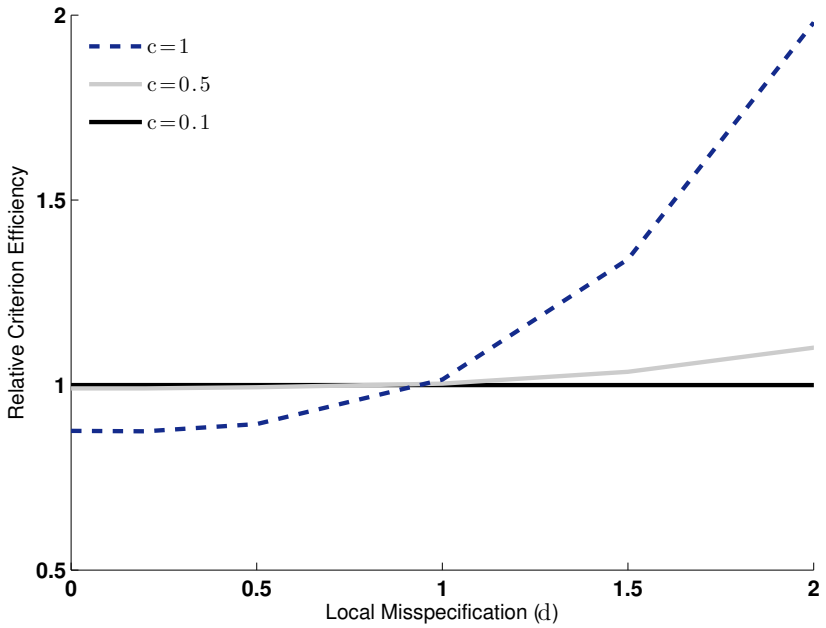

Figure 3: RQE: Local Misspecification.

estimator begins to dominate the LBE once the misspecification parameter reaches $d = 1$. By comparing the values of d for which the LBE dominates the innate estimator with the power reported in Table 3, it is obvious that diagnostic tests can be beneficial in the selection of an estimator. However, diagnostic tests are unlikely to provide perfect guidance in this respect for a number of reasons. One reason is that there are going to be levels of misspecification that are difficult to detect, yet detrimental to the likelihood-based estimator. This is illustrated in this application for medium levels of misspecification, $d \in [1, 2]$ and $c \geq 1$. In this design one would prefer to use the innate estimator over the LBE. However, for this level of misspecification one cannot rely on the JB test for the selection of the estimator, because it only has moderate power in this range for d. If more powerful tests were adopted, we would face the opposite problem of having higher rejection rates for $d < 1$ where the LBE still dominates the innate estimator.

Figure 3 illustrates the impact of local misspecification on the RQE under LinEx loss.[5] It shows that the LBE dominates the innate estimator when the misspecification is small ($d \leq 1$). At some point (at about $d \simeq 1$) the local misspecification more than offsets the advantages that the likelihood approach offers in terms of a reduced asymptotic variance. The larger is c the more advantageous it is to use the LBE over the innate estimator for low levels misspecification. The ranking is reversed under severe misspecification, where a high degree of asymmetry magnifies the distortions caused by misspecification.

[5]These results are based on $n = 1,000,000$ and $100,000$ replications.

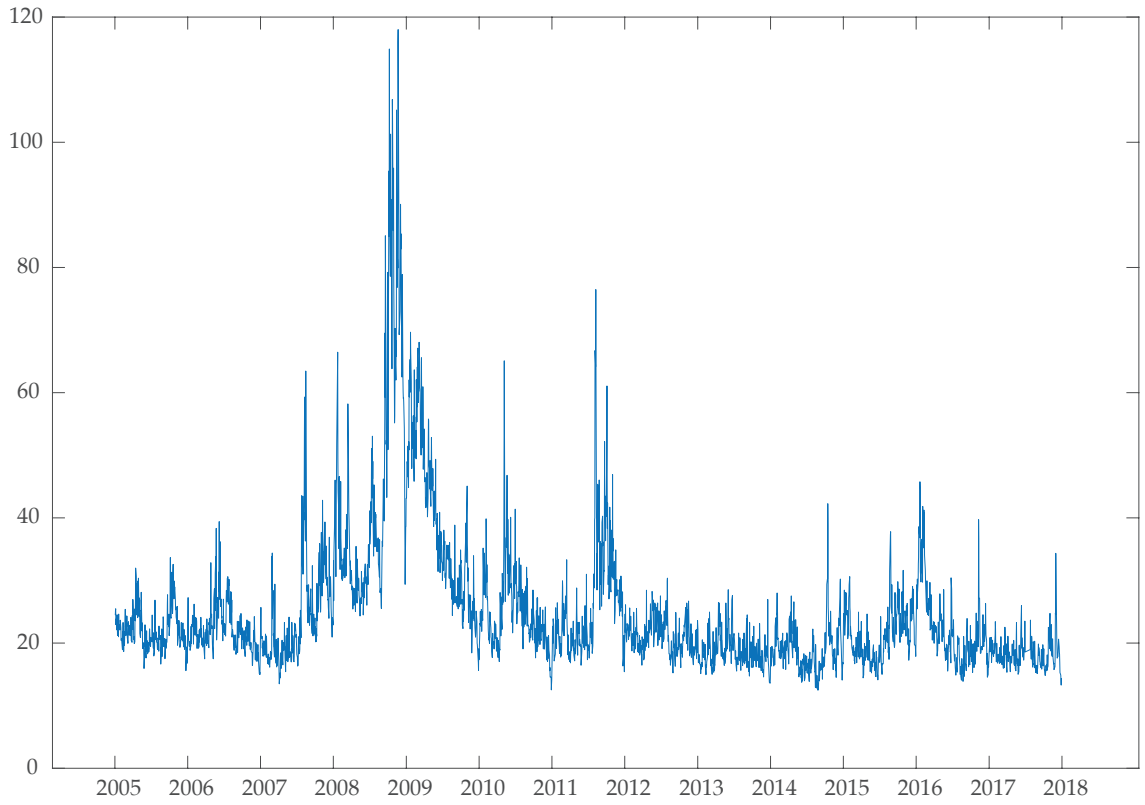

Figure 4: Annualized volatility computed from the geometric average of the daily realized kernel estimator for companies that were included in the S&P 500 index during the period January 3rd 2005 to December 31st 2017.

4 Empirical Application to Multistep Forecasting

The estimation problem in multistep forecasting emerges as an important special case in the present framework. In multistep forecasting there are estimation methods known as the *direct* forecast and the *iterated* forecast, see e.g. Marcellino et al. (2006).[6] The focus of this literature has been on the case with mean square error loss function, see e.g. Cox (1961), Tiao and Tsay (1994), Clements and Hendry (1996), Bhansali (1997), Ing (2003), Chevillon (2007), and references therein. The direct and iterated forecasts are closely related to the innate and the likelihood-based estimators, respectively. However, the crux of this problem is not multistep forecasting, but the fact that different criteria are used for estimation. This nuance can be overlooked in the standard setting with regression-based forecasts, and MSE loss function, because both the direct forecast and the iterated forecast are deduced from least squares estimation problems in this context.

We consider the problem of multistep forecasting, with an empirical application to volatility prediction. Figure 4 displays the daily geometric average of realized volatility for 646 assets from January 3rd,

[6]The former is also known as the prediction error estimation and the latter as the plug-in forecast.

2005 to December 31st, 2017.[7] We included ticker symbols for which there were at least ten transactions every day during the sample period.[8]

We analyze multi-step forecasts of $Y_t = \frac{1}{n}\sum_i \log(\mathrm{RK}_{i,t})$ where the objective is given by Linex loss function with the mean squared error loss function being a special case. We compare innate estimation and likelihood-based estimation for different forecasting horizons using three model specifications: The first specification is a restricted AR(22) model, known as the HAR model, and the other specifications are standard AR(5) and AR(3) models. The three specifications will be used to illustrate how different levels of misspecification impact the forecasting performance. We will follow the standard convention and use *direct forecast* and *iterated forecast* to refer to forecasts based on innate estimation and likelihood-based estimation, respectively.

The direct HAR forecast is obtained by estimating the parameters by

$$\min_{\mu,\beta_d,\beta_w,\beta_m} \sum_{t=1}^{T} L_c(Y_t - \mu - \beta_d Y_{t-h} - \beta_w Y^w_{t-h} + \beta_m Y^m_{t-h}),$$

where h is the forecasting horizon and $Y^w_t = \frac{1}{5}\sum_{j=0}^{4} Y_{t-j}$ and $Y^m_t = \frac{1}{22}\sum_{j=0}^{21} Y_{t-j}$ are the weekly and monthly averages, respectively. When $c = 0$ we have the case with mean square error loss function, where the parameters can be estimated by simple regression,

$$Y_t = \mu + \beta_d Y_{t-h} + \beta_w Y^w_{t-h} + \beta_m Y^m_{t-h} + e_{t-h,t}, \qquad t = 1, \ldots, T. \tag{8}$$

The likelihood-based approach is (for all c and h) deduced from the regression model

$$Y_t = \mu + \beta_d Y_{t-1} + \beta_w Y^w_{t-1} + \beta_m Y^m_{t-1} + \varepsilon_t, \qquad t = 1, \ldots, T, \tag{9}$$

where the likelihood function is based on the assumption $\varepsilon_t \sim iidN(0, \sigma^2)$. This model is equivalent to an AR(22) model, $Y_t = \mu + \varphi_1 Y_{t-1} + \cdots + \varphi_{22} Y_{t-22} + \varepsilon_t$, subject to the restrictions, $\varphi_2 = \cdots = \varphi_5$ and $\varphi_6 = \cdots = \varphi_{22}$, and the model can be estimated with regression (8) by setting $h = 1$. From the estimated model, we obtain the estimate of the conditional mean, $\hat{Y}^e_{t,t+h} = \hat{\mu} + \hat{\varphi}_1 \hat{Y}^e_{t,t+h-1} + \cdots + \hat{\varphi}_{h-1}\hat{Y}^e_{t,t+1} + \hat{\varphi}_h Y_t + \cdots + \hat{\varphi}_{22} Y_{t+h-22}$, and the corresponding forecast error variance, $\hat{\sigma}^2_h$, see (B.1), and

[7] The underlying volatility time series were previously used in Archakov (2016). Our time series is extended with four years of data (2014-2017), and the data were kindly provided to us by Ilya Archakov.

[8] The annualized measure of volatility for asset i on day t is given by $\sqrt{250 \cdot \mathrm{RK}_{i,t}}$, where $\mathrm{RK}_{i,t}$ is the realized kernel estimator computed from high-frequency returns on asset i (scaled by 100 to express it as a percentage), see Barndorff-Nielsen et al. (2008). Our analysis is based on the time series, $Y_t = \frac{1}{n}\sum_i \log(\mathrm{RK}_{i,t})$, which we converted to the geometric average, $\left(\prod_i \sqrt{250 \cdot \mathrm{RK}_{i,t}}\right)^{\frac{1}{n}} = \exp\{\frac{1}{2}[\log(250) + Y_t]\}$, that is displayed in Figure 4.

the resulting iterated (likelihood-based) forecast is $\hat{Y}_{t,t+h} = \hat{Y}^e_{t,t+h} + c\frac{\hat{\sigma}^2_{t,t+h}}{2}$. The forecasts based on the AR(5) and AR(3) models are derived similarly.

We compare the direct forecast with the iterated forecast in Table 6. Each of the models is estimated once using 8 years of data, 2005 to 2012, and the remaining five years of data, 2013 to 2017, are used for out-of-sample evaluation and comparisons. We consider forecasting horizons ranging from $h = 1$ day to $h = 22$ days, and five levels of asymmetry with the LinEx loss function, $c \in \{0, 0.25, 0.5, 1, 1.5\}$, where $c = 0$ corresponds to the mean square error loss function. We report the average loss for the direct forecast (scaled by 100), and below each of the average losses we report the percentage improvement offered by the iterated forecasts.[9] Positive percentages (shaded green) correspond to cases where the likelihood-based iterated forecasts dominate the direct forecast, and negative percentages (shaded blue) are cases where the direct forecast had the smallest sample loss. Cases where the sample losses are significantly different in a pairwise comparison using a Diebold-Mariano test and a 10% significance level are indicated with asterisks.

The sample loss tends to be monotonically increasing in the forecasting horizon, h, as one would expect. For the iterated forecast the loss is strictly increasing in h for all loss functions and all model specifications. This also holds for the direct forecast with few minor exceptions. For instance the average MSE loss for the direct HAR forecast with $h = 14$ is slightly smaller than for $h = 13$.

In the upper panel of Table 6 we present results based on the HAR specification. Direct and iterated forecasts are identical if $c = 0$ and $h = 1$, but different for any other combination of c and h, and the relative performance of the two types of forecast clearly depend on h and c. Both h and c induce discrepancies between the criterion that defines the forecasting objective and the objective used to estimate parameters for the iterated forecast. The fact that the direct forecast with the HAR model significantly outperforms the likelihood-based forecast strongly suggests that the HAR model (9) with Gaussian errors is misspecified. Misspecification takes many forms, including dynamic misspecification and distributional misspecification, see White (1994). In the present context, dynamic misspecification refers to the conditional mean implied by autoregressive models being incorrect, while distributional misspecification would refer to ε_t not being normally distributed. The two are related because dynamic misspecification would typically imply distributional misspecification. The results in the upper panel of Table 6 suggest that the model is dynamically misspecified because the relative performance of the direct forecasts increases as h increases. The likelihood-based forecasts rely on the Gaussian assumption

[9] The reported number is defined as $(\bar{L}_{\text{direct}} - \bar{L}_{\text{iterated}})/[(\bar{L}_{\text{direct}} + \bar{L}_{\text{iterated}})/2]$, where $\bar{L}_{\text{direct}}$ is the out-of-sample loss by the direct forecast (reported in the table) and $\bar{L}_{\text{iterated}}$ is the corresponding out-of-sample loss by the iterated forecasts.

Table 6: Out-of-sample Linex loss for log-realized kernel

HAR

	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$	$h=9$	$h=10$	$h=11$	$h=12$	$h=13$	$h=14$	$h=15$	$h=16$	$h=17$	$h=18$	$h=19$	$h=20$	$h=21$	$h=22$
$c=0$	2.62	3.79	4.51*	4.95	5.34	5.77*	6.17*	6.26	6.41	6.47	6.62	6.88	6.99	6.98	7.03	7.13*	7.26*	7.27*	7.32*	7.45*	7.54*	7.70*
	0.0%	0.1%	0.9%	0.2%	-0.6%	-0.9%	-0.8%	-0.3%	0.4%	-0.2%	-1.2%	-1.3%	-1.5%	-2.1%	-2.6%	-3.1%	-3.5%	-3.7%	-3.7%	-4.1%	-4.7%	-5.5%
$c=\frac{1}{4}$	2.67	3.86	4.60*	5.06	5.46	5.89	6.31	6.42	6.57	6.64	6.82	7.10	7.23	7.22	7.27	7.39*	7.52*	7.54*	7.60*	7.74*	7.85*	8.02*
	0.1%	0.3%	1.4%	0.7%	-0.1%	-0.4%	-0.2%	0.1%	0.7%	0.0%	-1.0%	-1.1%	-1.3%	-1.8%	-2.4%	-2.9%	-3.3%	-3.6%	-3.8%	-4.2%	-4.9%	-5.7%
$c=\frac{1}{2}$	2.71	3.94	4.70*	5.18*	5.59*	6.02	6.47	6.59	6.76*	6.84	7.04	7.36	7.51	7.50	7.56	7.69	7.83*	7.86*	7.93*	8.10*	8.21*	8.41*
	0.2%	0.5%	1.8%	1.3%	0.6%	0.3%	0.6%	0.8%	1.2%	0.3%	-0.6%	-0.6%	-0.7%	-1.2%	-1.8%	-2.1%	-2.6%	-2.9%	-3.2%	-3.7%	-4.4%	-5.1%
$c=1$	2.81	4.10*	4.93*	5.45*	5.89*	6.32*	6.85*	7.00*	7.21*	7.34*	7.61*	8.01*	8.23*	8.24*	8.32	8.51	8.68	8.74	8.86	9.07	9.22	9.49
	0.4%	1.1%	3.0%	2.7%	2.3%	2.0%	2.7%	2.8%	3.1%	2.2%	1.4%	1.7%	1.9%	1.5%	1.1%	1.2%	0.9%	0.7%	0.4%	0.1%	-0.7%	-1.1%
$c=1.5$	2.93	4.28*	5.18*	5.76*	6.24*	6.71*	7.33*	7.57*	7.85*	8.08*	8.45*	8.99*	9.29*	9.33*	9.49*	9.74*	9.98*	10.14*	10.33*	10.62*	10.82*	11.23*
	0.8%	1.7%	4.4%	4.5%	4.6%	4.4%	6.0%	6.5%	7.1%	6.6%	6.1%	7.1%	7.4%	7.5%	7.6%	8.1%	8.4%	8.9%	8.8%	9.0%	8.3%	8.4%

AR(5)

	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$	$h=9$	$h=10$	$h=11$	$h=12$	$h=13$	$h=14$	$h=15$	$h=16$	$h=17$	$h=18$	$h=19$	$h=20$	$h=21$	$h=22$
$c=0$	2.68	3.90	4.66	5.20	5.64*	6.07*	6.49*	6.63*	6.80*	6.91*	7.12*	7.43*	7.58*	7.57*	7.56*	7.67*	7.78*	7.79*	7.86*	8.03*	8.19*	8.41*
	0.0%	0.0%	0.1%	-0.3%	-1.4%	-2.9%	-3.5%	-2.8%	-2.2%	-2.2%	-3.1%	-3.6%	-3.8%	-4.2%	-5.5%	-6.4%	-7.4%	-7.7%	-7.7%	-8.0%	-8.3%	-8.9%
$c=\frac{1}{4}$	2.72	3.95	4.72	5.27	5.72*	6.16*	6.60*	6.75*	6.94*	7.06*	7.29*	7.63*	7.79*	7.78*	7.78*	7.90*	8.02*	8.05*	8.13*	8.33*	8.50*	8.74*
	0.0%	-0.1%	0.0%	-0.3%	-1.3%	-2.8%	-3.4%	-3.0%	-2.6%	-2.9%	-3.8%	-4.4%	-4.7%	-5.4%	-6.8%	-7.8%	-8.9%	-9.4%	-9.6%	-10.0%	-10.5%	-11.2%
$c=\frac{1}{2}$	2.76	4.01	4.80	5.36	5.81*	6.25*	6.72*	6.88*	7.09*	7.23*	7.49*	7.85*	8.04*	8.03*	8.04*	8.18*	8.31*	8.35*	8.46*	8.68*	8.87*	9.14*
	-0.1%	-0.1%	0.0%	-0.2%	-1.1%	-2.5%	-3.1%	-3.0%	-2.8%	-3.3%	-4.3%	-4.9%	-5.3%	-6.2%	-7.7%	-8.7%	-9.8%	-10.5%	-10.9%	-11.4%	-11.9%	-12.6%
$c=1$	2.85	4.14	4.95	5.56	6.04	6.49*	7.00*	7.19*	7.44*	7.65*	8.00*	8.44*	8.70*	8.69*	8.74*	8.91*	9.09*	9.18*	9.36*	9.64*	9.89*	10.25*
	-0.1%	0.0%	0.0%	0.4%	-0.2%	-1.4%	-1.8%	-2.2%	-2.4%	-3.2%	-4.0%	-4.4%	-4.9%	-6.1%	-7.5%	-8.5%	-9.4%	-10%	-10.5%	-10.9%	-11.4%	-11.7%
$c=1.5$	2.96	4.28	5.13	5.81*	6.34*	6.83	7.39	7.63	7.95	8.27	8.76	9.35	9.68	9.68*	9.81*	10.02*	10.29*	10.5*	10.78*	11.16*	11.54*	12.09*
	0%	0.2%	0.3%	1.6%	1.7%	0.8%	0.6%	0.1%	0.0%	-0.5%	-0.8%	-0.7%	-1.3%	-2.5%	-3.5%	-4.4%	-4.8%	-4.8%	-5.0%	-4.8%	-4.7%	-4.2%

AR(3)

	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$	$h=9$	$h=10$	$h=11$	$h=12$	$h=13$	$h=14$	$h=15$	$h=16$	$h=17$	$h=18$	$h=19$	$h=20$	$h=21$	$h=22$
$c=0$	2.75	3.99*	4.72*	5.27*	5.87*	6.44*	6.87*	6.85*	6.95*	7.13*	7.45*	7.75*	7.82*	7.81*	7.95*	8.08*	8.14*	8.12*	8.10*	8.24*	8.39*	8.64*
	0.0%	-2.1%	-3.0%	-2.4%	-1.8%	-3.4%	-6.0%	-8.3%	-8.8%	-8.6%	-9.6%	-11.6%	-13.9%	-15.3%	-16.3%	-18.2%	-21.1%	-23.0%	-24.9%	-26.2%	-27.1%	-27.8%
$c=\frac{1}{4}$	2.79	4.05*	4.78*	5.34*	5.94*	6.51*	6.96*	6.96*	7.07*	7.28*	7.62*	7.95*	8.03*	8.02*	8.16*	8.30*	8.37*	8.35*	8.36*	8.52*	8.70*	8.97*
	-0.1%	-2.1%	-3.1%	-3.0%	-2.8%	-4.5%	-7.3%	-10.1%	-11.2%	-11.4%	-12.6%	-14.7%	-17.3%	-19.2%	-20.5%	-22.6%	-25.5%	-27.7%	-29.8%	-31.1%	-32.0%	-32.6%
$c=\frac{1}{2}$	2.83	4.11*	4.86*	5.41*	6.01*	6.59*	7.06*	7.07*	7.21*	7.44*	7.82*	8.18*	8.29*	8.27*	8.42*	8.57*	8.65*	8.64*	8.67*	8.86*	9.07*	9.37*
	-0.1%	-2.1%	-3.2%	-3.6%	-3.7%	-5.5%	-8.4%	-11.7%	-13.2%	-13.9%	-15.1%	-17.3%	-20.0%	-22.4%	-24.0%	-26.1%	-29.1%	-31.5%	-33.7%	-34.9%	-35.7%	-36.2%
$c=1$	2.92	4.24*	5.02*	5.59*	6.19*	6.80*	7.31*	7.35*	7.54*	7.84*	8.32*	8.78*	8.96*	8.94*	9.11*	9.29*	9.40*	9.43*	9.52*	9.79*	10.07*	10.50*
	-0.1%	-1.8%	-3.1%	-4.4%	-5.2%	-7.1%	-10.0%	-14.0%	-16.1%	-17.3%	-18.5%	-20.3%	-23.1%	-26.1%	-28.0%	-30.2%	-32.9%	-35.4%	-37.5%	-38.3%	-38.8%	-38.5%
$c=1.5$	3.15	4.55	5.43	6.07*	6.68*	7.41*	8.09*	8.39*	8.83*	9.56*	10.36*	11.15*	11.52*	11.57*	11.90*	12.08*	12.36*	12.56*	12.92*	13.46*	14.18*	15.20*
	0.4%	-0.7%	-1.7%	-4.4%	-6.3%	-7.2%	-9.1%	-11.6%	-13%	-12.1%	-12.1%	-12.6%	-15.2%	-18.2%	-19.6%	-22.3%	-24.0%	-25.6%	-26.2%	-25.6%	-23.7%	-20.5%

Note: Out-of-sample loss average of the direct forecast (innate estimation). The relative performance of the iterated forecast (likelihood-based estimation) is reported below. Cases where the direct forecast had the smallest sample loss are highlighted with green color, and cases where the iterated forecast faired better are highlighted with blue. Asterix denote cases where the performance is significantly different according to a pairwise Diebold-Mariano test.

because the adjustment $c\sigma_h^2/2$ is deduced from the normal distribution. The likelihood-based forecasts dominate the direct forecasts at shorter horizons when c is large, which suggests that the distributional misspecification is not the main issue in this application. Imposing the Gaussian distribution greatly reduce the variance of the parameter estimates relative to direct estimation, which compensates for the bias induced by misspecification. For the highest level of asymmetry in Table 6, $c = \frac{3}{2}$, the iterated forecast is significantly better than the direct forecasts for all multi-period forecasts. Only at $h = 1$ is the difference not significant.

The middle and lower panels of Table 6 present the corresponding results when the underlying structure is based on an AR(5) and an AR(3) model. One might expect these specifications to be more misspecified than the HAR model and that is indeed what the forecasting results suggests. The AR(5) forecasts are generally inferior to the HAR forecasts and the AR(3) forecasts are even worse. As our theoretical results predict, misspecification is more harmful to the likelihood-based forecasts than the direct forecasts based on innate estimation, and this is indeed what we observe in the middle and lower panels of Table 6. The iterated forecasts are inferior to the direct forecasts with very few exceptions. Only with a very asymmetric loss functions and a short horizon, such as $(c, h) = (\frac{3}{2}, 1)$, do we see cases where the sample performance was (slightly) better for the direct forecast.

The empirical results highlight another important point: At a given level of misspecification, one cannot necessarily conclude that the likelihood-based forecasts will be inferior to the direct forecast. The relative performance will depend on the actual objective, e.g. the loss function in a forecasting problem. So, the direct approach may yield superior forecasts for one objective, and inferior forecasts for another objective. This is precisely the situation we observe in the upper panel of Table 6.

We can relate these empirical observations to our theoretical results in Section 2. In this application, the objective is characterized by c and h, and these parameters will define the relative advantages that likelihood-based estimation will have under correct specification. This advantage is directly related to the smaller asymptotic variance of the likelihood-based estimator. The difference in asymptotic criterion risk, $\frac{1}{2}\text{tr}\{A^{-1}B\} - \frac{1}{2}\text{tr}\{A\tilde{A}^{-1}\}$, is the theoretical quantity that measures this advantage and this quantity depends on c and h. The corresponding quantity that embodies the negative impact of misspecification is given by $\frac{1}{2}\text{tr}\{Abb'\}$, and this quantity also depends on c and h. It is therefore not surprising that the first term dominates the second term for some configurations of (c, h), while the reverse is true for other configurations of (c, h). This empirical finding is consistent with these theoretical results and shows that a given level of misspecification can be detrimental for some objectives but relatively harmless for other objectives.

5 Conclusion

In this paper we have studied parameter estimation in the situation where the estimated model is intended to be applied to a different sample than the one used for estimation. The leading example is the problem of forecasting, where parameters are estimated from currently available observations, while the objective is to predict features of future observations. However, this structure is not specific to forecasting. Much empirical research is motivated by applying the estimated model to different samples than the one used for estimation.

In our framework the objective is defined by a criterion function, and the question is how this criterion function should be incorporated in the estimation method. The innate estimator is a very natural estimator because it employs the same criterion for estimation as the objective. If a different estimator is to be employed, we showed that the notion of coherency – between the estimation criterion and the actual objective – is essential. A coherent criterion can be crafted from a maximum likelihood estimator, and we showed that the likelihood-based estimator is asymptotically efficient in the present context. This result is analogous to, and deduced from, the classical Cramer-Rao bound. The superiority of the likelihood-based estimator relies on the likelihood function being correctly specified. When the likelihood function is misspecified, the likelihood-based estimator can be inferior to the innate estimator. However, the key problem is not that the likelihood-based estimator does not achieve the Cramer-Rao lower bound under misspecification, but a distortion of the mapping of likelihood parameters to criterion parameters. The latter is the Achilles heel of the likelihood-based estimator in this context.

The choice between the innate estimator and the likelihood-based estimator entails a trade-off between robustness and efficiency that arises in many statistical problems. While robustness of the innate estimator is appealing, we have shown that the likelihood-based estimator can be vastly better than the innate estimator even in simple problems. The flip-side is that the likelihood-based estimator can also be vastly inferior to the innate estimator if the underlying statistical model is misspecified. We showed that the types of misspecification that are most harmful, will vary from one criterion to another. For instance, we presented cases where the likelihood-based estimator is seriously impaired by a relatively modest level of misspecification, which would be difficult to detect statistically. Similarly, we documented cases with easy-to-detect levels of misspecification, where the likelihood-based estimator continued to dominate the innate estimator. Our analysis of local-to-correct specification added further insight about the problem. The likelihood-based estimator dominates the innate estimator at modest levels of misspecification. The threshold at which the innate estimator becomes superior to the likelihood-based estimator is context-specific and depends on many factors, including the nature of the

misspecification and the criterion function, Q. In our application based on the LinEx loss function we saw that the superiority of the likelihood-based estimator increases with the degree of asymmetry of the objective. For this reason, it takes a relatively high degree of 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. This shows that the selection problem is not merely equivalent to misspecification testing. The interaction between the criterion and the type of misspecification, must be taken into account.

Ultimately, we did not arrive at an unequivocal answer to the question: "How should parameter estimation be tailored to the objective?" Our theoretical results identified the likelihood-based estimator as the best choice if the likelihood function is correctly specified. But the likelihood-based estimator is a precarious choice unless measures are taken to minimize the harmful effects of misspecification. Our results highlight the benefits of using model diagnostics in the present context, especially diagnostics that can detect the forms of misspecification that distort the mapping from likelihood-parameters to criterion parameters. A direct comparison of the innate and likelihood-based estimates, $\hat{\theta}$ and $\theta(\tilde{\vartheta})$, is a simple way to guard against the types of misspecification that are most harmful to the objective. This could be implemented with a Hausman-type test or a score-based approach, where the derivative of $Q(\mathcal{X}, \theta)$ is evaluated at $\theta(\tilde{\vartheta})$. When harmful misspecification is found, it is not obvious if the best way to proceed is to adopt the innate estimator or to explore another (less misspecified) likelihood function. An extensive search over specifications can amplify problems related to overfitting that also play an important role in this context.

References

Akaike, H. (1978), 'On the likelihood of a time series model', *Journal of the Royal Statistical Society. Series D (The Statistician)* **27**, 217–235.

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

Archakov, I. (2016), 'Essays in applied econometrics of high frequency financial data', *PhD dissertation at the European University Instittute* .

Barndorff-Nielsen, O. E. (1977), 'Exponentially decreasing distributions for the logarithm of particle size', *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* **353**(1674), 401–419.

Barndorff-Nielsen, O. E. (1978), 'Hyperbolic distributions and distributions on hyperbolae', *Scandinavian Journal of Statistics* **5**, 151–157.

Barndorff-Nielsen, O. E., Blæsid, P., Jensen, J. L. and Sørensen, M. (1985), The fascination of sand, *in* A. Atkinson and S. Feinberg, eds, 'A Celebration of Statistics', Sprin, New York.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A. and Shephard, N. (2008), 'Designing realised kernels to measure the ex-post variation of equity prices in the presence of noise', *Econometrica* **76**, 1481–536.

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**, 746–785.

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

Christoffersen, P. and Jacobs, K. (2004), 'The importance of the loss function in option valuation', *Journal of Financial Economics* **72**, 291–318.

Clements, M. and Hendry, D. (1996), 'Multi-step estimation for forecasting', *Oxford Bulletin of Economics and Statistics* **58**, 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', *Journal of the Operational Research Society* pp. 199–207.

Granger, C. W. J. and Newbold, P. (1977), *Forecasting Economic Time Series*, Academic Press, Orlando.

Hansen, B. E. (2010*a*), 'Multi-step forecast model selection', *Working paper* .

Hansen, P. R. (2010*b*), 'A winner's 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, New York.

Hwang, S., Knight, J. and Satchell, S. (2001), 'Forecasting nonlinear functions of returns using LINEX loss functions', *Annals of economics and finance* **2**, 187–213.

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

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.

Schorfheide, F. (2005), 'VAR forecasting under misspecification', *Journal of Econometrics* **128**, 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**, 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**, 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.

A Appendix: Proof of Analytical Results in Section 2

Proof of Lemma 1. The consistency follows from the existing literature on M-estimators, see e.g. Amemiya (1985, Theorem 4.1.1). □

Proof of Lemma 2. The Gaussian limit distribution follows directly from Assumption 3, so we only have to determine the covariance matrix. To simplify notation, 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. It follows that the asymptotic variance of $n^{-1/2}\sum_{t=1}^{2n} s_t$ is $2B$, while the asymptotic variance of $n^{-1/2}\sum_{t=1}^{n} s_t$ and $n^{-1/2}\sum_{t=n+1}^{2n} s_t$ both equal B. Using the simple identity for the variance of a sum, it follows that the asymptotic covariance of $n^{-1/2}\sum_{t=1}^{n} s_t$ and $n^{-1/2}\sum_{t=n+1}^{2n} s_t$ must be 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 $n^{-1/2}\sum_{t=n+1}^{2n} s_t$. □

Proof of Lemma 3. Since $\tilde{\theta} \xrightarrow{p} \theta_0$, it follows by Assumptions 1 and 2 that $n^{-1}\sum_{t=1}^{n} q(\mathbf{x}_t, \tilde{\theta}) \xrightarrow{p} \mathbb{E}[q(\mathbf{x}_t, \theta_0)]$, which is strictly smaller than $\mathbb{E}[q(\mathbf{x}_t, \theta_*)]$, as a consequence of Assumption 2.*ii*. So that $\frac{1}{n}[Q[(\mathcal{Y}, \hat{\theta}) - Q(\mathcal{Y}, \tilde{\theta})] \xrightarrow{p} \mathbb{E}[q(\mathbf{x}_t, \theta_*)] - \mathbb{E}[q(\mathbf{x}_t, \theta_0)] > 0$ which shows that the divergence is at rate n, and the result follows. □

Proof of Theorem 1. To simplify notation, we write $Q_x(\theta)$ in place of $Q(\mathcal{X}, \theta)$, and similarly $H_x(\theta) = H(\mathcal{X}, \theta)$, $Q_y(\theta) = Q(\mathcal{Y}, \theta)$, $\tilde{Q}_x(\theta) = \tilde{Q}(\mathcal{X}, \theta)$, $S_x(\theta) = \sum_{t=1}^{n} s(\mathbf{x}_t, \theta)$, and $S_y(\theta) = \sum_{t=n+1}^{2n} s(\mathbf{x}_t, \theta)$, 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 $n^{-1/2}\{\tilde{S}_x(\theta), S_y(\theta)\} \xrightarrow{d} \{\tilde{B}^{1/2}Z_x, B^{1/2}Z_y\}$ where Z_x and Z_y are independent and both distributed as $\mathrm{N}(0, I)$. 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

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

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

Proof of Lemma 4. 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

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

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

Proof of Theorem 2. 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 $\mathrm{N}(0, I)$, with Z_y independent of $(Z_x, \tilde{Z}_x)$. This defines the random variable ξ. Two of the terms vanish after taking the expected value, which yields

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

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

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

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). □

Proof of Corollary 1. The proof is analogous to that of Theorem 2, albeit the comparison in question is here

$$-\mathrm{tr}\{\check{A}^{-1}A\check{A}^{-1}\check{B}\} + \mathrm{tr}\{\tilde{A}^{-1}A\tilde{A}^{-1}\tilde{B}\} = \mathrm{tr}\left\{A^{1/2}(\tilde{A}^{-1} - \check{A}^{-1}\check{B}\check{A}^{-1})A^{1/2}\right\} \leq 0,$$

where $\check{A}$ and $\check{B}$ are the "information" matrices associated with $\check{\theta}$. □

Proof of Theorem 3. With $P_n \to P_{\vartheta_0}$ we have $\theta(\vartheta_0) = \theta_*$. Thus with $\theta_0^{(n)} = \theta(\vartheta_0^{(n)})$ and a Taylor expansion of $Q(\mathcal{Y}, \tilde{\theta})$ about $Q(\mathcal{Y}, \theta_*)$ we find the limit distribution of $Q(\mathcal{Y}, \tilde{\theta}) - Q(\mathcal{Y}, \theta_*)$ to be given

from

$$S_y(\theta_*)'(\tilde{\theta} - \theta_0^{(n)} + \theta_0^{(n)} - \theta_*) + \tfrac{1}{2}(\tilde{\theta} - \theta_0^{(n)} + \theta_0^{(n)} - \theta_*)' H_y(\theta_*)(\tilde{\theta} - \theta_0^{(n)} + \theta_0^{(n)} - \theta_*).$$

The expectation of the first term is zero, and the limit distribution of the second term is

$$\tfrac{1}{2}(\tilde{A}^{-1}\tilde{B}^{1/2}Z_x + b)'[-A](\tilde{A}^{-1}\tilde{B}^{1/2}Z_x + b),$$

so that the asymptotic criterion risk is

$$\tfrac{1}{2}\mathbb{E}(\tilde{A}^{-1}\tilde{B}^{1/2}Z_x + b)'A(\tilde{A}^{-1}\tilde{B}^{1/2}Z_x + b) = \tfrac{1}{2}\mathrm{tr}\{A\tilde{B}^{-1}\} + \tfrac{1}{2}\mathrm{tr}\{Abb'\} = \tfrac{1}{2}\mathrm{tr}\{A(\tilde{B}^{-1} + bb'\}$$

where we used that $\mathbb{E}Z_x = 0$, $\mathbb{E}Z_x Z_x' = I$ and the information matrix equality $\tilde{A} = \tilde{B}$, which also holds under local-to-correct specification. □

Proof of Theorem 4. 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. □

B Appendix: Proof of Auxiliary Results

B.1 Derivations Related to the Linex Case With Correct Specification

The expression for A follows by

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

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

For B we note that

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

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

B.2 Proof of (6) in Section 3.2

Lemma B.1. *Suppose that* $X \sim \mathrm{NIG}(\lambda,\delta,\alpha,\beta)$ *and let* $\theta_0 = \lambda + \frac{\delta\beta}{\gamma} + \frac{c}{2}\delta\frac{\alpha^2}{\gamma^3}$, *where* $\gamma = \sqrt{\alpha^2-\beta^2}$. *Then*

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

solves $\min_\theta \mathbb{E}[\exp\{c(X-\theta)\}-c(X-\theta)-1]$. *Moreover,* $\theta_*-\theta_0\to 0$ *if either (a)* $c\to 0$*; or (b)* $\alpha\to\infty$ *with* $\sigma^2=\delta/\alpha$ *constant and* $\beta=0$.

Proof. Using the moment generating function for the NIG-distribution the objective is 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 first result.

By the l'Hospital rule we find

$$\lim_{c\to 0}\frac{\sqrt{\alpha^2-\beta^2}-\sqrt{\alpha^2-(\beta+c)^2}}{c} = \frac{\lim_{c\to 0}2(\beta+c)\frac{1}{2}[\alpha^2-(\beta+c)^2]^{-1/2}}{1} = \beta/\gamma,$$

which proves that $\theta_* \to \lambda + \frac{\delta\beta}{\gamma}$ as $c \to 0$.

With $\beta = 0$ we have $\gamma = \alpha$, $\delta = \sigma^2\alpha = \mathrm{var}(X)$ and $\lambda = \mu = \mathbb{E}(X)$, so that

$$\theta_* = \mu + \frac{\sigma^2\alpha}{c}\left[\sqrt{\alpha^2} - \sqrt{\alpha^2 - c^2}\right] = \mu + \frac{\sigma^2}{c}\frac{1-\sqrt{1-xc^2}}{x},$$

where $x = \alpha^{-2}$. The limit as $\alpha \to \infty$, i.e. $x \to 0$, is by the l'Hospital rule

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

which completes the proof. □

Theorem B.1. *Consider the* $\mathrm{NIG}(\lambda, \delta, \alpha, \beta)$*, 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*

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

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 $\mathrm{NIG}(\lambda, \delta, \alpha, \beta)$ 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 vanishes 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

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

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

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

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

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

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 $\mathrm{N}(\mu, \sigma^2)$. □

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

Proof. The results follow from Theorem B.1, or directly by observing that the CF for $\mathrm{NIG}(\mu, \sigma^2\alpha, \alpha, 0)$ is

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

Now by l'Hospital's rule note 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

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

□

Proof of (7). From $\xi = (1+\delta\gamma)^{-1/2}$ and $\chi = -\xi\frac{\beta}{\alpha}$, we note that $1/\xi^2 - 1 = (1+\delta\gamma) = \delta\gamma = \gamma^4/\alpha^2$ and $\xi^2 - \chi^2 = \xi^2(1-\beta^2/\alpha^2) = \xi^2\alpha^{-2}\gamma^2$ so that

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

Similarly,

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

which proves the identities in (7). □

B.3 Forecast error variance for autoregression

The optimal forecast under LinEx loss function entails an adjustment that is proportional to $\sigma_h^2 = \text{var}(Y_{t+h}|\mathcal{F}_t)$ under normality, see (4). In our empirical application, we seek the likelihood-based estimate of σ_h^2, that is based on autoregressive models (The HAR model corresponds to a restricted AR(22) model). Consider the autoregression $Y_t = \mu + \varphi_1 Y_{t-1} + \cdots + \varphi_p Y_{t-p} + \varepsilon_t$, $\varepsilon_t \sim iidN(0, \sigma_\varepsilon^2)$, expressed in the companion form $Y_t^* = \mu^* + \Phi Y_{t-1}^* + \varepsilon_t^*$, where

$$Y_t^* = \begin{pmatrix} Y_t \\ \vdots \\ Y_{t-p+1} \end{pmatrix}, \qquad \Phi^* = \begin{pmatrix} \varphi_1 & \varphi_2 & \cdots & \varphi_p \\ 1 & 0 & \cdots & 0 \\ & \ddots & \ddots & \vdots \\ 0 & & 1 & 0 \end{pmatrix}, \quad \varepsilon_t^* = \begin{pmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

and $\mu^* = (\mu', 0, \ldots, 0)'$. It follows that

$$Y_{t+h}^* = \Phi^h Y_t^* + (I + \Phi + \cdots + \Phi^{h-1})\mu^* + (\varepsilon_{t+h}^* + \Phi\varepsilon_{t+h-1}^* + \cdots \Phi^{h-1}\varepsilon_{t+h-(h-1)}^*),$$

and that $\mathbb{E}_t(Y_{t+h}) = a_{0,h} + a_{1,h}Y_t + \cdots + a_{p,h}Y_{t+1-p}$, where $a_{0,h}$ is given as μ times the upper-left element of $(I + \Phi + \cdots + \Phi^{h-1}) = \mu(1 + [\Phi]_{1,1} + \cdots + [\Phi^{h-1}]_{1,1})$, and the constants, $a_{i,h}$, $i = 1, \ldots, p$, are given from the first row of Φ^h. Thus, σ_h^2 is given by the variance of the first element of $\varepsilon_{t+h}^* + \Phi\varepsilon_{t+h-1}^* + \cdots \Phi^{h-1}\varepsilon_{t+h-(h-1)}^*$, which can be expressed as $\varepsilon_{t+h} + [\Phi]_{1,1}\varepsilon_{t+h-1} + \cdots + [\Phi^{h-1}]_{1,1}\varepsilon_{t+1}$. So, $\sigma_1^2 = \sigma_\varepsilon^2$, $\sigma_2^2 = (1 + \varphi_1^2)\sigma_\varepsilon^2$, $\sigma_3^2 = (1 + \varphi_1^2 + (\varphi_1^2 + \varphi_2)^2)\sigma_\varepsilon^2$, and the general expression is given by

$$\sigma_h^2 = (1 + [\Phi]_{1,1}^2 + \cdots + [\Phi^{h-1}]_{1,1}^2)\sigma_\varepsilon^2, \qquad h = 1, 2, \ldots. \tag{B.1}$$

The iterated forecast under linex loss is given by $\hat{\mathbb{E}}_t(Y_{t+h}) + \frac{c}{2}\hat{\sigma}_h^2$, which is computed with the expressions for $\mathbb{E}_t(Y_{t+h})$ and σ_h^2 using the maximum likelihood estimates of μ, $\varphi_1, \ldots, \varphi_p$, and σ_ε^2.

C Appendix: Details Concerning the Simulations Designs

C.1 LinEx Loss

Proposition C.1 (Invariance of LinEx Simulation Study)**.** *The simulation study based on $X_i \sim iid\text{N}(\mu, \sigma^2)$ and LinEx loss L_d, is equivalent to that based on $X_i \sim iid\text{N}(0, 1)$ and LinEx loss L_c.*

Our simulation 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 d, is equivalent to a design based on $Z_i = (X_i - \mu)/\sigma$ with asymmetry parameter $c = \sigma d$. To establish this result we first show that the estimator, $\check{\theta}_d$, deduced from LinEx loss, L_d, and the sample $\mathcal{X} = (X_1, \ldots, X_n)$, is linearly related to the estimator $\check{\theta}_c$, deduced from LinEx loss, L_c, and the sample $\mathcal{Z} = (Z_1, \ldots, Z_n)$, by

$$\check{\theta}_d(\mathcal{X}) = \mu + \sigma\check{\theta}_c(\mathcal{Z}). \tag{C.1}$$

For the innate estimator this follows by

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

and similarly for the likelihood-based estimator we observed that

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

Hence

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

which proves that,

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

Since the scale, σ^2, is common for all estimators that satisfy (C.1), their relative performance is unaffected.

Details about the simulation study

We draw $2n$ independent observations $\mathrm{N}(0,1)$. The first n observations (in-sample) are used to compute the estimators $\hat{\theta}$ and $\tilde{\theta}$. The remaining n observation are used to compute the out-of-sample losses, including the losses resulting from the ideal parameter value θ_*. This is done in 500,000 independent replications, and the properties of $\hat{\theta}$ and $\tilde{\theta}$ are evaluated by averaging over the simulations. We have used $n = 100$ and $n = 1,000$ in the finite sample analysis. A range, $c \in \{0; 0.25; 0.5; 1; 1.5; 2; 2.5\}$, of values for the asymmetry parameter is used.

C.2 LinEx under NIG distribution

The simulations with the normal inverse Gaussian (NIG) distribution represent a case with local misspecification. We use random variables drawn from NIG$(\lambda, \delta, \alpha, \beta)$, that are normalized to have mean zero and unit variance. This is achieved by setting $\delta = \gamma^3/\alpha^2$ and $\lambda = -\delta\beta/\gamma$ where $\gamma = \sqrt{\alpha^2 - \beta^2}$, and this standardized NIG distribution may be parameterized by:

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

An asymmetric distribution is obtained with $\chi = -\xi^{3/2}$, leaving one free parameter, where the Gaussian case arises as $\xi \to 0$. 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. The advantage of this setup is that the NIG optimal predictor is always computable as long as $0 \leq \xi < 1$. The values of the parameters $\lambda, \delta, \alpha, \beta$ are then deduced from the relations above and the standardization constraints.

To set the misspecification level in the simulations design, we first deduce the rate of convergence of ξ to zero. By setting the LinEx asymmetry coefficient to 1, we solve for ξ so that $\sqrt{n}(\theta_0 - \theta_*) - b = 0$. From (4) and (6), and the normalizations $\mu = 0$ and $\sigma = 1$, we observe that $\theta_0 - \theta_* = \frac{1}{2} - \lambda - \delta\left[\gamma - \sqrt{\alpha^2 - (\beta+1)^2}\right]$. Next using that $\chi = -\xi^{3/2}$ and defining $a_\xi = \sqrt{\frac{1+\xi}{1-\xi}}$ we find that

$$\begin{aligned}
\alpha &= \xi\tfrac{\sqrt{1-\xi^2}}{\xi^2-\chi^2} = \xi\tfrac{\sqrt{1-\xi^2}}{\xi^2-\xi^3} = \tfrac{1}{\xi}\sqrt{\tfrac{1-\xi^2}{(1-\xi)^2}} = \tfrac{a_\xi}{\xi}, \\
\beta &= \chi\tfrac{\sqrt{1-\xi^2}}{\xi^2-\chi^2} = -\tfrac{1}{\sqrt{\xi}}\tfrac{\sqrt{1-\xi^2}}{1-\xi} = -\tfrac{a_\xi}{\sqrt{\xi}}, \\
\gamma &= \sqrt{\tfrac{1-\xi^2}{\xi^2-\chi^2}} = \tfrac{1}{\xi}\sqrt{\tfrac{1-\xi^2}{1-\xi}} = \sqrt{\tfrac{1+\xi}{\xi^2}} = \tfrac{\sqrt{1-\xi}}{\xi}\tfrac{\sqrt{1-\xi^2}}{1-\xi} = \tfrac{a_\xi\sqrt{1-\xi}}{\xi},
\end{aligned}$$

so that $\delta = \gamma^3/\alpha^2 = a_\xi\frac{\xi^2(1-\xi)^{3/2}}{\xi^3} = a_\xi\frac{(1-\xi)^{3/2}}{\xi} = \sqrt{1+\xi}\frac{(1-\xi)}{\xi} = \sqrt{\frac{1-\xi^2}{\xi}}\sqrt{\frac{1-\xi}{\xi}}$ and $\lambda = -\delta\beta/\gamma = a_\xi\frac{1-\xi}{\sqrt{\xi}} = \sqrt{\frac{1-\xi^2}{\xi}}$. Consequently,

$$\begin{aligned}
\theta_0 - \theta_* &= \tfrac{1}{2} - a_\xi\tfrac{1-\xi}{\sqrt{\xi}} - a_\xi\tfrac{(1-\xi)^{3/2}}{\xi}\left[\tfrac{a_\xi\sqrt{1-\xi}}{\xi} - \sqrt{a_\xi^2/\xi^2 - (1 - a_\xi/\sqrt{\xi})^2}\right] \\
&= \tfrac{1}{2} - \sqrt{\tfrac{1-\xi^2}{\xi}} - \sqrt{\tfrac{1-\xi^2}{\xi}}\sqrt{\tfrac{1-\xi}{\xi}}\left[\sqrt{\tfrac{1+\xi}{\xi^2}} - \sqrt{\tfrac{1+\xi}{1-\xi}/\xi^2 - (1 - \sqrt{\tfrac{1+\xi}{1-\xi}}/\sqrt{\xi})^2}\right] \\
&= \tfrac{1}{2}\xi^{3/2} - \tfrac{1}{8}\xi^2 + O(\xi^3),
\end{aligned}$$

equating with $n^{-1/2}b$ implies $\xi \simeq (2b)^{\frac{2}{3}}n^{-\frac{1}{3}}$. So in our simulations design we set $\xi = d \times n^{-1/3}$ where d defines the degree of local-misspecification, and we let d vary from 0 (correctly specified model) to 3.

Then we proceed with the following steps:

1. Given ξ we compute the parameters $(\lambda, \delta, \alpha, \beta)$ using the expressions provided in Section 3.2.
2. Draw a sample of size $2n$ from the NIG$(\lambda, \delta, \alpha, \beta)$ distribution.
3. The estimators $\hat{\theta}$ and $\tilde{\theta}$ are computed with (3) and (5) using the first n observations, while θ_* is computed with (6).
4. The criterion is then evaluated out-of-sample using the last n observations, using $\hat{\theta}$, $\tilde{\theta}$, and θ_*.
5. Repeat steps 2-4 for the desired number of repetitions (we use $100,000$ and $500,000$ in some cases).
6. Evaluate the out-of-sample properties by averaging over repetitions.

For the design resembling the asymptotic case, we use the sample size $n = 1,000,000$, and we have verified that $n = 100,000$ produced very similar results. The levels of asymmetry used are $c \in \{0; 0.25; 0.5; 1; 2\}$. A finite sample analysis with $n = 200$ is also performed.

In this where $\xi \propto n^{-1/3}$, we have $\chi = -\xi^{3/2} \propto -n^{-1/2}$,

$$\alpha = \frac{\sqrt{1-\xi^2}}{\xi^2 - \chi^2} \propto n^{-1/3} \frac{\sqrt{1-n^{-2/3}}}{n^{-2/3} - n^{-1}} \propto n^{1/3},$$

and $\beta = \alpha\chi/\xi \propto -n^{\frac{1}{3}-\frac{1}{2}+\frac{1}{3}} = -n^{1/6}$.