UNIVERSITY OF LIVERPOOL
Management School

Working Paper in Economics

# 202106

Improving the Estimation and Predictions of Small Time Series Models

Gareth Liu-Evans

https://www.liverpool.ac.uk/management/people/economics/

Improving the estimation and predictions of small time series models

Gareth Liu-Evans
Management School, University of Liverpool, Chatham Street, Liverpool, L69 7ZH
Email: gareth.liu-evans@liverpool.ac.uk
Tel: 07791947279

Abstract

A new approach is developed for improving the point estimation and predictions of parametric time-series models. The method targets performance criteria such as estimation bias, root mean squared error, variance, or prediction error, and produces closed-form estimators focused towards these targets via a computational approximation method. This is done for an autoregression coefficient, for the mean reversion parameter in Vasicek and CIR diffusion models, for the Binomial thinning parameter in integer-valued autoregressive (INAR) models, and for predictions from a CIR model. The success of the prediction targeting approach is shown in Monte Carlo simulations and in out-of-sample forecasting of the US Federal Funds rate.

1. Introduction

A number of papers have addressed in different ways the difficulty in, or impossibility of, applying exact likelihood estimation to certain time-series models by providing approximate likelihood methods, see for example Aït-Sahalia (2002), likelihood-free methods based on simulation including Indirect Inference, see Gourieroux et al. (1993), Efficient Method of Moments, see Gallant and Tauchen (1996), and Approximate Bayesian Computation, see for example Martin et al. (2019). Many of the models considered are small but widely used and difficult to estimate. There is, moreover, a sizable literature on the correction of estimation bias for parameters of time series models, and a s ubstantial part of this has focused on methods involving asymptotic expansion and approximation of the true bias. A number of papers have addressed the estimation of continuous-time interest rate diffusion models recently, where the bias in estimation of the mean reversion parameter can be particularly severe, and a review can be found in Iglesias and Phillips (2019).

The aim in what follows is to demonstrate the effectiveness of a new approach to estimation and prediction improvement for parametric models, where simple closed-form correction terms similar to those obtained by asymptotic approximation, power series in $1/n$, are found computationally. The method relies on initial consistent estimates of the parameters being available. For the purpose of comparison with other methods the focus is mainly on the reduction of estimation bias, though it is illustrated how improvements in RMSE or variance can be targeted, and with a small modification the prediction error as well. The final section applies the approach to prediction improvement in a CIR model of the Federal Funds rate. When targeting a reduction in estimation bias, the new approach involves training a bias correction functional for a given model and estimator using Monte Carlo generated data and moment computation, with the overall aim being to obtain a closed-form correction to the initial estimator that can be applied subsequently to different initial estimates and a range of sample sizes. The approach is not limited to addressing estimation bias, and can be used to address more general risk objectives in relation to estimation performance.

The methodology is presented in Section 2 using least squares estimation of a first-order autoregressive model as an illustrative example. Sections 3 and 4 present further examples of the methodology: to estimation of the mean-reversion parameter in Vasicek and CIR diffusion models, where Maximum Likelihood (ML) methods can be severely biased, and to integer autoregressive (INAR) models, where the estimation is also biased. A further motivation for addressing estimation of diffusion models is that there are only a few cases where exact ML is possible, moreover we are able to compare the performance of the new approach with results in Tang and Chen (2009) for estimation by Bootstrap, Jackknife and Indirect Inference. A further motivation for addressing estimation of INAR models is that the exact ML estimation of INARMA models more generally is difficult, which motivated the Efficient Method of Moments (EMM) approach in Martin et al. (2014), and the new approach is applicable whenever initial consistent estimates of the model parameters are available. Section 5 applies the methodology to prediction of the Effective Federal Funds Rate for overnight lending in the United States, and Section 6 concludes.

2. Methodology

The AR(1) with constant is used here to illustrate the methodology as it is widely familiar, and has received substantial attention in the literature on correction of estimation bias. Kiviet and Phillips (2012, 2014) obtain theoretical results for asymptotic approximation of the estimation bias, of the variance and for analytically corrected estimation, while Chambers (2013) develops an improved jackknife methodology for autoregressions, see also Liu-Evans and Phillips (2012) who compare bootstrap, jackknife and analytical correction methods. Despite its relatively simple form, the AR(1) model continues to appear abundantly in empirical work, a recent example being Baltussen et al. (2019) on return predictability. The AR(1) also arises as a discrete-time counterpart to the Vasicek diffusion model for short term interest rates. Some further discussion relating to interpretation and generalisation of the methodology is in Section 2.3

2.1. Correcting OLS bias in estimation of an AR(1)

The following specification is considered:

$$y_t = \alpha + \lambda y_{t-1} + u_t, \tag{1}$$

$t = 1, \ldots, n$, where $u_t \overset{i.i.d.}{\sim} N(0, \sigma^2)$, $\sigma^2 < \infty$, and $|\lambda| < 1$. The bias in estimation of λ can be substantial, see in particular the % bias entries in Table 1 for $\hat{\lambda}$ at $n = 35$, which are in the range -12.7% to -27.7%. A sample size of 35 is small, but is consistent with other studies addressing the AR(1) estimation bias.[1] The estimation biases for diffusion models

[1] See for example Chambers (2013), where Monte Carlo results are presented for $n = 24, 48, 96$ and 192, and Kiviet and Phillips (2014) where $n = 20$ and 50.

in Section 3 are more severe at larger sample sizes, while the INAR models in Section 4 are designed for short series of count data.

Kendall (1954) and Marriott and Pope (1954) found that the bias in OLS estimation of λ in (1) could be asymptotically approximated as

$$b(\lambda) = -\frac{1+3\lambda}{n} + o(n^{-1}), \tag{2}$$

and this can be used to form a Corrected OLS (COLS) estimator

$$\hat{\lambda}_{COLS} := \hat{\lambda}_{OLS} + \frac{1}{n}(1 + 3\hat{\lambda}_{OLS}), \tag{3}$$

which is unbiased to order $O(n^{-1})$ in the sense that $E[\hat{\lambda}_{COLS} - \lambda]$ is $o(n^{-1})$. Similar bias-correction results have been obtained for other models and estimators, as noted in the Introduction, and the analytical approach has worked well in simulation experiments, see for example the early study by Orcutt and Winokur (1969).

Despite the success of the approach and its strong theoretical basis it might be possible to choose, according to some overall bias criterion, an even better correction function than the one in (3) implied by large-n asymptotic expansion. If attention is restricted to specific values of n in a small interval, for example, or just to a single value, this may seem quite plausible. There are, moreover, models and estimators where no asymptotic refinement to the bias is available. The investigator's primary interest may also not be in the bias, but in improving some other property of the estimator, such as the RMSE or variance, and analytical refinement towards one of these objectives may be challenging. Section 5 illustrates a case where a model used for prediction may be better served by an estimator focused specifically towards reduced prediction error rather than reduced estimation bias.

Continuing with the theme of bias reduction, a basic requirement is that the estimation bias be reduced from the original corresponding to ordinary least squares, and a comparison can also be made with the analytically corrected estimator in (3). Initially, our question is therefore whether it is possible to find a function g in (4) below, via a numerical optimisation, without knowing the bias approximation in (2), such that $\tilde{\lambda}$ is less biased in some overall sense than $\hat{\lambda}_{OLS}$:

$$\tilde{\lambda} := \hat{\lambda}_{OLS} + \frac{1}{n} g(\hat{\lambda}_{OLS}). \tag{4}$$

There are additional arguments besides $\hat{\lambda}_{OLS}$ that may be useful to have in g, and this issue is addressed in Section 2.3, but it is known that the bias in this case depends mainly on λ.[2]

The approach requires an overall performance measure to be decided for the new estimator $\tilde{\lambda}$ in (4), which should capture some aspect of estimation performance across different possible values of λ, α and σ^2, then the performance of $\tilde{\lambda}$ can be adjusted by choice of g. With a view toward reducing relative bias in estimation of λ, a loss of $L(\tilde{\lambda}, \lambda) = |\frac{\tilde{\lambda}-\lambda}{\lambda}|$ is defined for a given choice of λ, α and σ^2, and for a given sample size. Risk values $\mathbb{E}[L(\tilde{\lambda}, \lambda)]$ are then computed by Monte Carlo and collected at different points in the parameter space and at different sample sizes, all in a vector R, and the objective is to minimise a norm $||R||$ as a measure of overall performance. An ideal choice of g in (4) is then taken to be

$$g^\star := \underset{g \in \mathcal{G}}{\operatorname{argmin}} \, ||R|| \tag{5}$$

where $\mathcal{G}$ is a chosen class of approximating functions. The measure of overall performance can be viewed in terms of global risk, see for example Lehmann (1983), and this is outlined in Section 2.3. Beyond the main objective in (5), it may be preferable that the choice of g results in an estimator that performs no worse than the original in terms of bias or root mean square error. This relative performance constraint can be imposed at the parameterisations used for training g, and it is generally implemented in the examples that follow including those in the present section.[3]

Provided the chosen approximating functions can be parameterised, say by a vector w, then a numerical search can be used to minimise $||R||$. A minimisation of $||R||$ by choice of g in a space of polynomials, for example, could potentially yield $g^\star(\hat{\lambda}_{OLS}) = 1 + 3\hat{\lambda}$, which would make (4) the same as the COLS estimator in (3). Instead of polynomials, we mainly use univariate rational approximants in the Padé form, though a more general neural network approach is detailed in Section 2.3 and used in Section 3 for the Vasicek model. The idea of parameterising rational approximants in the Padé form computationally has been used in Chen et al. (2018), see in particular their RationalNet.

If the class of $[m_1/m_2]$ Pade approximants is used for g, then g as a mapping from $\hat{\lambda}$ is in the form

$$g(\hat{\lambda}) = \frac{\sum_{i=0}^{m_1} a_i \hat{\lambda}^i}{1 + \sum_{j=1}^{m_2} b_j \hat{\lambda}^j} \tag{6}$$

[2] Note that the parameters α and σ^2 do not enter (2), though they do enter the higher-order $O(n^{-2})$ bias approximation, see Bao (2007) and Kiviet and Phillips (2012).

[3] See Section 2.3 and the Appendix for details.

where a_i and b_j, $i = 0, \ldots, m_1$ and $j = 1, \ldots, m_2$, are the parameters in w to be selected by a numerical search. Analogous to a higher-order bias correction, see for example Bao (2007) and Kiviet and Phillips (2012), a $\frac{1}{n^2}$ term can be added to (4), then there are two mappings g_1 and g_2 to select as in (7). In this section we choose among estimators in the form

$$\tilde{\lambda} := \hat{\lambda}_{OLS} + \frac{1}{n} g_1(\hat{\lambda}_{OLS}) + \frac{1}{n^2} g_2(\hat{\lambda}_{OLS}) \tag{7}$$

where g_1 and g_2 are as in (6) with $m_1 = 4$ and $m_2 = 5$, so that there are 10 parameters to specify in each case. The following version is also considered, where the first two terms form the COLS estimator in (3), and the search is therefore for an improvement on the COLS estimator:

$$\tilde{\lambda}_{COLS} := \hat{\lambda}_{OLS} + \frac{1}{n}(1 + 3\hat{\lambda}_{OLS}) + \frac{1}{n} g_1(\hat{\lambda}_{OLS}) + \frac{1}{n^2} g_2(\hat{\lambda}_{OLS}). \tag{8}$$

The value for the overall performance $||R||$ at given choices of g_1 and g_2 will depend on the parameterisation and sample size choices used to obtain each element of R, and therefore these choices will shape the resulting estimator obtained by minimising $||R||$. The collection of parameter and sample size combinations used for each element of R is, in what follows, denoted by $\mathcal{T}$. These are training points for choosing g_1 and g_2, whose performance can later be assessed at other points in the parameter space and at other sample sizes. In the current section, g_1 and g_2 are trained on the three values of λ in $\{0.1, 0.5, 0.97\}$ with $\alpha = 0$ and $\sigma^2 = 1$, and on the two sample sizes in $\{20, 50\}$, then assessed at various other positive values of λ, at two choices of α, with σ^2 at 9 rather than 1, and at a sample size midway between the two sample sizes used for the training. Estimators in the form (7) and (8) are found for alternative objectives in Section 2.2, namely RMSE reduction and variance reduction, and for this reason the bias-reducing versions of $\tilde{\lambda}$ and $\tilde{\lambda}_{COLS}$ are denoted by $\tilde{\lambda}^{bias}$ and $\tilde{\lambda}^{bias}_{COLS}$.

It can be seen from the left panel in Figure 1 that the new estimator $\tilde{\lambda}^{bias}$ is highly effective at bias reduction across all the sample sizes and λ values considered. The right panel plots the relative RMSE values for the new estimator compared with the initial estimator, and it can be seen that these are either around 1 or substantially lower than 1. There also does not appear to be any over-training at the three values of λ used in $\mathcal{T}$ or at the two specific sample sizes used in the training. By searching for single choices of g_1 and g_2 that work well at both $n = 20$ and $n = 50$ and at several different parameterisations, the numerical search has found a correction functional that works well for any n between the two values used in $\mathcal{T}$ and for a fine grid of positive values of λ between 0 and 1. These cases also use $\alpha = 10$ rather than the training value of 0, and $\sigma^2 = 9$ rather than the training value of 1.[4] Throughout the paper, a minimum of 20,000 replications are used for results in tables and figures.

<Figure 1 here>

Table 1 presents the bias and RMSE values for the initial estimator $\hat{\lambda}$, for the new reduced bias estimators $\tilde{\lambda}^{bias}$ and $\tilde{\lambda}^{bias}_{COLS}$, and for $\hat{\lambda}_{COLS}$. It can be seen that $\tilde{\lambda}^{bias}_{COLS}$, making use of the asymptotic approximation in addition to the methodology here, tends to do a little better than $\tilde{\lambda}^{bias}$, and that both seem marginally better than $\hat{\lambda}_{COLS}$ in terms of bias when $\lambda \geq 0.65$. There are only six training points in $\mathcal{T}$ in the current section, and better results could potentially be obtained by using more. This is tried in Section 3 for the mean reversion parameter in the Vasicek model.

[4] It may be unsurprising that these alternative values of α and σ^2 have a limited effect on the performance of the estimator, as they only enter asymptotic bias approximations for $\hat{\lambda}$ at order $O(n^{-2})$.

Table 1: Percentage bias and RMSE in estimation of λ with $\sigma^2 = 9$, $n = 35$

	α	λ	$\hat{\lambda}$	$\tilde{\lambda}^{bias}$	$\hat{\lambda}_{COLS}$	$\tilde{\lambda}^{bias}_{COLS}$	$\tilde{\lambda}^{RMSE}_{COLS}$
% Bias	0	0.15	-27.7	-2.30	-2.15	-2.04	7.05
		0.25	-19.3	-2.59	-1.77	-2.79	-2.89
		0.35	-16.5	-2.98	-1.20	-2.82	-4.71
		0.45	-14.8	-2.71	-1.26	-2.76	-5.10
		0.55	-13.7	-2.05	-1.26	-2.03	-4.93
		0.65	-13.0	-1.17	-1.21	-1.20	-4.56
		0.75	-12.8	-0.370	-1.50	-0.319	-4.57
		0.85	-12.7	0.622	-1.90	0.383	-5.14
		0.95	-13.5	1.23	-2.93	0.853	-6.87
	10	0.15	-27.7	-2.32	-1.98	-1.45	7.05
		0.25	-19.4	-2.64	-1.45	-2.72	-2.80
		0.35	-16.0	-2.50	-1.21	-2.92	-4.97
		0.45	-14.8	-2.69	-1.31	-2.67	-5.11
		0.55	-13.7	-1.99	-1.22	-1.89	-4.76
		0.65	-13.0	-1.23	-1.26	-1.19	-4.63
		0.75	-12.9	-0.521	-1.50	-0.493	-4.59
		0.85	-12.8	0.611	-1.92	0.442	-5.18
		0.95	-13.3	1.36	-2.92	0.843	-6.87
RMSE	0	0.15	0.170	0.170	0.179	0.170	0.143
		0.25	0.169	0.170	0.177	0.171	0.153
		0.35	0.170	0.171	0.174	0.171	0.160
		0.45	0.170	0.170	0.170	0.170	0.164
		0.55	0.169	0.169	0.164	0.169	0.163
		0.65	0.167	0.165	0.157	0.165	0.158
		0.75	0.166	0.161	0.149	0.160	0.150
		0.85	0.166	0.157	0.138	0.155	0.138
		0.95	0.173	0.154	0.131	0.151	0.130
	10	0.15	0.169	0.170	0.178	0.170	0.143
		0.25	0.170	0.170	0.176	0.171	0.154
		0.35	0.169	0.171	0.174	0.171	0.160
		0.45	0.170	0.170	0.170	0.170	0.164
		0.55	0.169	0.169	0.165	0.168	0.163
		0.65	0.167	0.165	0.157	0.165	0.158
		0.75	0.168	0.162	0.148	0.161	0.150
		0.85	0.167	0.157	0.139	0.155	0.138
		0.95	0.172	0.153	0.130	0.149	0.130

2.2. Reducing RMSE and Variance

It has been seen from Figure 1 and Table 1 that the reduced bias estimators tend to have better RMSE performance than the original estimator. It is possible, however, to target a reduction in RMSE directly, by changing the loss function L specified earlier to $L(\lambda, \tilde{\lambda}) = (\tilde{\lambda} - \lambda)^2$ and filling R with RMSE values $(\mathbb{E}[L(\lambda, \tilde{\lambda})])^{\frac{1}{2}}$, while keeping the rest of the setup unchanged. As the original reduced-bias estimator available from asymptotic expansion of the bias, $\hat{\lambda}_{COLS}$, already performs well in terms of bias correction, it seems interesting to ask whether some of this bias correction behaviour will remain after adding additional terms to improve the RMSE performance. The resulting estimator in the form (8) is denoted by $\tilde{\lambda}_{COLS}^{RMSE}$, and it can be seen in Table 1 and Figure 2 that the RMSE performance of this estimator is superior to the others while, in Table 1, the bias is still substantially reduced from the original OLS estimator.

<Figure 2 here>

It is possible to target a reduction in variance in the same way, still with the relative performance constraint controlling the bias performance at points in $\mathcal{T}$, and the resulting estimator is denoted by $\tilde{\lambda}_{COLS}^{Var}$. Variance results for all of the estimators are given in Table 2, and it can be seen that $\tilde{\lambda}_{COLS}^{Var}$ has substantially lower values. The left panel in Figure 3 depicts the variance of $\tilde{\lambda}_{COLS}^{Var}$ verses the OLS estimator, and it can be seen that the variance is almost halved for lower values of λ. The right panel presents a comparison of the absolute biases, and the reduced-variance estimator performs better in this respect as well for $\lambda \leq 0.6$, while being about the same (marginally worse) for $0.6 < \lambda < 1$.

<Figure 3 here>

Table 2: Variance in estimation of λ with $\sigma^2 = 9$, $n = 35$

	α	λ	$\hat{\lambda}$	$\tilde{\lambda}^{bias}$	$\hat{\lambda}^{bias}_{COLS}$	$\tilde{\lambda}^{bias}_{COLS}$	$\tilde{\lambda}^{RMSE}_{COLS}$	$\tilde{\lambda}^{Var}_{COLS}$
Variance × 10^2	0	0.15	2.71	2.88	3.21	2.88	2.04	1.59
		0.25	2.63	2.88	3.13	2.92	2.34	1.79
		0.35	2.57	2.92	3.02	2.91	2.54	1.97
		0.45	2.44	2.89	2.87	2.88	2.63	2.05
		0.55	2.29	2.83	2.69	2.83	2.58	2.04
		0.65	2.07	2.71	2.47	2.70	2.40	1.95
		0.75	1.85	2.60	2.21	2.57	2.12	1.78
		0.85	1.60	2.45	1.88	2.39	1.71	1.54
		0.95	1.39	2.37	1.63	2.27	1.28	1.33
	10	0.15	2.70	2.87	3.17	2.88	2.04	1.58
		0.25	2.64	2.90	3.11	2.92	2.36	1.79
		0.35	2.57	2.92	3.01	2.93	2.54	1.97
		0.45	2.44	2.88	2.90	2.88	2.62	2.05
		0.55	2.29	2.83	2.71	2.81	2.58	2.03
		0.65	2.06	2.70	2.46	2.72	2.41	1.96
		0.75	1.87	2.63	2.19	2.58	2.13	1.77
		0.85	1.60	2.46	1.90	2.40	1.70	1.56
		0.95	1.36	2.34	1.61	2.22	1.27	1.33

2.3. Further methodological notes

A neural network approach

The $AR(1)$ model is relatively simple, and the bias in estimation of λ mainly depends on one parameter, namely λ itself. This enables the use of univariate approximants for the bias reduction or other estimation improvement, but a more general approach is desirable. Given the problem of estimating a parameter θ whose estimation bias depends on parameters in a vector Θ, the general proposal is an estimator of θ in the following form

$$\tilde{\theta} = \hat{\theta} + G(\hat{\Theta}, r) \tag{9}$$

where

$$G(\hat{\Theta}, r) = \sum_{j=1}^{r} \frac{1}{n^j} g_j(\hat{\Theta}), \tag{10}$$

$\hat{\Theta}$ is an initial estimator of Θ, and r is a small number. The choice of G in (9) may depend on the interval of sample sizes considered, therefore the mappings G and g_j are implicitly indexed by n. In a typical situation where the initial estimator $\hat{\Theta}$ is $\sqrt{n}$-consistent, $\tilde{\theta}$ will have the same property under mild conditions on the sequences $\{g_{j,n}\}_n$. At large sample sizes it may even be reasonable to assume that zero mappings $g_{j,n} = 0$ are chosen for $j = 1, \ldots, r$.

Feedforward neural networks with one or more hidden layers can, for a sufficiently large number of hidden units, approximate any continuous function on a compact domain arbirarily closely and are therefore universal approximants, see for example Hornik (1991). Mappings of the following form with a single hidden layer for g_j are used in the Vasicek diffusion model application in Section 3.1:

$$g_j(\hat{\Theta}) = \sum_{i=1}^{m'} a_{ji} F(b_{ji} \cdot \hat{\Theta} + c_{ji}) \tag{11}$$

where F is the sigmoid activation function $F(v) = (1 + e^{-v})^{-1}$ and m' is, in the neural networks terminology, the number of hidden units. The parameters a_{ji}, b_{ji} and c_{ji}, for $j = 1, \ldots, r$ and $i = 1, \ldots, m'$, can be collected in a vector w in the same way as for Pade approximants earlier, with the numerical minimisation of $||R||$ again performed over w.

Interpretation in terms of point estimation theory

The methodology can be interpreted in terms of the theoretical framework in Lehmann (1983) relating to minimisation of global risk. We are interested in estimating θ, an element of $\Theta \in \mathcal{C} \subset \mathbb{R}^d$, and, in the notation of Lehmann, are seeking to choose among candidate estimators $\delta(X)$ that yield estimates $\delta(x)$ when given data x. In the same way as earlier, the cost associated with $\delta(x)$ for a given point Θ is denoted by $L(\Theta, \delta(x))$, and the average loss for a given Θ is measured by a risk function $R(\Theta, \delta) = E_\Theta[L\{\Theta, \delta(X)\}]$. If no restriction is put on the functional form of δ, as we have done by requiring it to be an initial estimator plus a power series in $1/n$, then the δ that minimises the average risk,

$$\int_{\mathcal{C}} R(\Theta, \delta) w(\Theta) d\Theta, \tag{12}$$

is by definition a Bayes estimator, provided the weight function w is specified as a prior for Θ, and it is otherwise a generalised Bayes estimator. Under a quadratic loss assumption $L(\Theta, \delta(x)) = (\delta(x) - \theta)^2$, for example, its estimate given observations x is known to be the posterior mean $\delta(x) = \int_{\mathcal{C}} \theta p(\Theta|x) d\Theta$, though this may be difficult or impossible to compute in practice.

The solution estimator to (12) under the quadratic loss assumption,

$$\delta(X) = \int_{\mathcal{C}} \theta p(\Theta|X) d\Theta, \tag{13}$$

is in the same form for each sample size, and therefore choosing $\delta(X)$ as in (13) for each sample size is the solution to minimisation of the following, where the risk values at different sample sizes in $\mathcal{N}$ are added together:

$$\int_{\mathcal{C}} \left(\sum_{n \in \mathcal{N}} R_n(\Theta, \delta) \right) w(\Theta) d\Theta. \tag{14}$$

If $\mathcal{C}$ in (14) is replaced by a training set of parameterisations $\tilde{\mathcal{C}} \subset \mathcal{C}$ and we set $w(\Theta) = 1$, this can be expressed as in (5) where R is the vector of risk values corresponding to points in $\mathcal{T} = \tilde{\mathcal{C}} \times \mathcal{N}$:

$$||R||_1 = \sum_{\Theta \in \tilde{\mathcal{C}}} \sum_{n \in \mathcal{N}} R_n(\Theta, \delta) \tag{15}$$

The method subsequently constrains $\delta(X)$ to be in the form $\delta(X) = \hat{\theta}(X) + G(\hat{\Theta}, r)(X)$ and searches numerically for the minimising choices of $g_1, g_2, \ldots, g_r$. While (13) is unlikely to be recovered for any given n, the result of minimising (15) with δ in its constrained form, where the loss and risk functions are defined as above, can be viewed for each n as a rough closed-form approximation of the posterior mean estimator in (13). It is not expected that this approach will yield accurate approximations of posterior moments, but the method avoids any significant computation having to be done for each given set of observations, and the resulting closed-form estimators can be assessed according to frequentist criteria. Moreover, the method makes it straightforward to compute estimators focused towards different choices of the global risk, e.g. via different choices of the loss function, without additional analytical derivation or posterior sampling. Constraints on the performance relative to a reference estimator can be imposed, along with constraints on the distribution of the resulting estimator.

The relative performance constraint

It was noted in Section 2.1 that a relative performance constraint can be placed on the choice of correction function when minimising $||R||$, and that we do this in most cases here. Similarly to the choice of loss function, the choice whether to include a relative performance constraint can be interpreted in terms of risk preferences. If particular solutions to (5) are avoided because the resulting estimator does not strictly outperform a reference set of performances by another estimator, this can be understood in terms of behavioural theory for decision making under risk where gains or losses are relative to a reference point, and where the decision maker is more sensitive to losses from a reference point than to gains, see for example Tversky and Kahneman (1992). For details about the implementation, see the Appendix.

3. Univariate interest rate diffusion models

We consider two special cases of the following univariate diffusion model:

$$dX(t) = \mu(X(t), t; \Theta)dt + \sigma(X(t), t; \Theta)dB(t) \tag{16}$$

where $B(t)$ is standard Brownian Motion, μ and σ are *drift* and *diffusion* functions, respectively, and Θ is an unknown parameter vector. In particular, we address the estimation of the Vasicek and the CIR models, which are relatively simple but widely used specifications for interest rates and other financial time series following Vasicek (1977) and Cox et al. (1985). Doing this allows a comparison with the Monte Carlo results in Tang and Chen (2009) for the parametric bootstrap estimator, Indirect Inference, and the $(m = 4)$ Quenouille jackknife.

The two models are as follows:

$$\begin{array}{ll} dX(t) = \kappa(\alpha - X(t))dt + \sigma dB(t) & \text{(Vasicek)} \\ dX(t) = \kappa(\alpha - X(t))dt + \sigma\sqrt{X(t)}dB(t) & \text{(CIR)} \end{array}$$

where $2\kappa\alpha/\sigma^2 > 1$ for the CIR model, see Cox et al. (1985). As in the preceding literature, the focus is on estimation of the mean reversion parameter κ, particularly the estimation performance at the small positive values of κ typical of short-term interest rate series. For the Vasicek model we only consider this situation, and only consider new estimators trained for a particular sample size - following some success with this, the CIR model is addressed more ambitiously in Section 3.2. In order to compare directly with the existing Monte Carlo results in Tang and Chen (2009), the Conditional ML estimator and Nowman pseudo-ML estimator, see Nowman (1997), are used as the initial estimators of κ when addressing the Vasicek and CIR models, respectively, and both are denoted by $\hat{\kappa}$ in the following. Details about the data generation can be found in Tang and Chen (2009).

3.1. Vasicek model

The Vasicek model has an exact Gaussian discretisation, and a Conditional ML estimation results in

$$\hat{\kappa} = -h^{-1}ln(\hat{\beta}_1)$$

with

$$\hat{\beta}_1 = \frac{n^{-1}\sum_{i=1}^{n} X_i X_{i-1} - n^{-2}\sum_{i=1}^{n} X_i \sum_{i=1}^{n} X_{i-1}}{n^{-1}\sum_{i=1}^{n} X_{i-1}^2 - n^{-2}(\sum_{i=1}^{n} X_{i-1})^2},$$

being an estimate of the autoregressive coefficient in the corresponding discrete model. Here n is the number of observations, h is the sampling interval, and nh the length of time over which the equally spaced observations are taken.[5] The new estimators take the form described in (9), (10) and (11) using the neural network approach with $r = 1$, $m' = 10$, but with $\hat{\Theta}$ replaced by $\hat{\kappa}$ by itself.

Since the Vasicek and CIR models are used mainly for modelling financial time-series with low mean reversion, it may be reasonable to search initially for an estimation improvement that works well just at relatively low values of κ. We do this in the present section with the Vasicek model, then consider the estimation of a wider range of κ values for the CIR model in Section 3.2. Three training schemes are considered for the Vasicek model, and the first two have this in mind. In all cases, we set $h = 1/12$. Case (1) uses just the two values of κ in $\{0.01, 0.3\}$, Case (2) uses a finer grid of low κ values in $\{0.001, 0.025, 0.05, 0.075, 0.1, 0.125, 0.15, 0.175, 0.2, 0.225, 0.25, 0.275, 0.3\}$ while Case (3) uses the three values of κ in $\{0.1, 0.5, 0.97\}$, which were the values used for λ in Section 2 (though we note the lower values now correspond to lower rather than higher speed of mean reversion, and stationarity requires $k > 0$, moreover κ can be far larger in principle than these values but this would be unusual within financial applications of the model). In all three cases, the training points in $\mathcal{T}$ use $\alpha = \sigma = 0.05$. Moreover, the estimator is trained to work at just a specific sample size $n = 500$. Section 3.2 on the CIR diffusion model will consider bias and RMSE reductions designed to work within an interval of sample sizes.

Table 3 presents results for the reduced-bias and reduced-RMSE estimators, $\tilde{\kappa}_n^{bias}$ and $\tilde{\kappa}_n^{RMSE}$ respectively, obtained using the Case (1) training setup. It can be seen that the initial bias values are much larger than for OLS estimation of the $AR(1)$ autoregressive

[5] A case with $n = 60$ and $h = 1/12$ would correspond to 5 years of monthly observations.

coefficient in Section 2, yet the new reduced-bias estimator has relatively low absolute bias values in the range 0-7% and substantially lower RMSE. The new reduced-RMSE estimator has even lower RMSE, less than half of the original in some cases. The results in Table 2 demonstrate that the new methodology can substantially improve the estimation performance of the mean reversion parameter in the Vasicek diffusion model, at a given sample size, with just a small number of training parameterisations.

Table 4 presents similar results using the training setups in Cases (1)-(3) for the two Vasicek models in Tang and Chen (2009) that have relatively low values of κ.[6] The values in Tang and Chen (2009) for the parametric bootstrap, the Quenouille jackknife and indirect inference, are included for comparison, the table also presents their values for the original estimator - ours were very similar. It can be seen that the new estimators compare well with the jackknife, the parametric bootstrap and indirect inference. The RMSE values of the two reduced-RMSE estimators are particularly good, being less than half the original and substantially lower than the bootstrap and indirect inference values. Curiously, despite the generally good reduction in bias among the new reduced-bias estimators, the best performance is obtained using the training setup in Case (3), where a wider interval for κ was considered, and where there were only 3 points in $\mathcal{T}$.

[6] ”Model 1” in Tang and Chen (2009) uses $\kappa = 0.858$.

Table 3: Percentage bias and RMSE in estimation of κ, Case (1) results, $n = 500$

$\kappa = 0.05$

		% Bias			RMSE		
α	σ	$\hat{\kappa}$	$\tilde{\kappa}_n^{bias}$	$\tilde{\kappa}_n^{RMSE}$	$\hat{\kappa}$	$\tilde{\kappa}_n^{bias}$	$\tilde{\kappa}_n^{RMSE}$
0.01	0.01	259	-2.01	112	0.177	0.131	0.0997
	0.05	261	-0.185	114	0.178	0.133	0.101
	0.1	260	-0.487	112	0.178	0.133	0.0999
0.05	0.01	260	-0.774	112	0.177	0.132	0.0999
	0.05	259	-2.18	113	0.177	0.132	0.100
	0.1	260	-0.547	112	0.178	0.132	0.0996
0.1	0.01	258	-3.38	113	0.176	0.131	0.0999
	0.05	258	-2.42	114	0.177	0.132	0.100
	0.1	257	-4.01	114	0.176	0.131	0.100

$\kappa = 0.15$

		% Bias			RMSE		
α	σ	$\hat{\kappa}$	$\tilde{\kappa}_n^{bias}$	$\tilde{\kappa}_n^{RMSE}$	$\hat{\kappa}$	$\tilde{\kappa}_n^{bias}$	$\tilde{\kappa}_n^{RMSE}$
0.01	0.01	77.8	-3.68	10.9	0.182	0.152	0.0831
	0.05	77.7	-3.78	11.1	0.181	0.151	0.0833
	0.1	78.3	-3.20	11.5	0.182	0.152	0.0836
0.05	0.01	78.3	-3.15	11.3	0.182	0.152	0.0829
	0.05	77.0	-4.54	11.3	0.180	0.150	0.0837
	0.1	77.6	-3.96	10.7	0.181	0.151	0.0835
0.1	0.01	78.0	-3.51	11.0	0.182	0.152	0.0828
	0.05	77.8	-3.74	11.1	0.181	0.152	0.0824
	0.1	78.2	-3.26	11.1	0.181	0.150	0.0827

$\kappa = 0.25$

		% Bias			RMSE		
α	σ	$\hat{\kappa}$	$\tilde{\kappa}_n^{bias}$	$\tilde{\kappa}_n^{RMSE}$	$\hat{\kappa}$	$\tilde{\kappa}_n^{bias}$	$\tilde{\kappa}_n^{RMSE}$
0.01	0.01	44.6	-0.786	-11.3	0.191	0.168	0.0851
	0.05	44.6	-0.792	-11.6	0.191	0.168	0.0847
	0.1	44.6	-0.863	-11.7	0.191	0.169	0.0841
0.05	0.01	44.7	-0.712	-11.5	0.191	0.169	0.0850
	0.05	44.1	-1.40	-11.5	0.189	0.167	0.0854
	0.1	44.2	-1.25	-11.7	0.191	0.170	0.0851
0.1	0.01	44.2	-1.24	-11.6	0.190	0.168	0.0838
	0.05	44.7	-0.735	-11.3	0.192	0.169	0.0854
	0.1	44.2	-1.23	-11.7	0.191	0.169	0.0847

Table 4 Comparison with *Tang and Chen (2009)*, Vasicek models, $n = 500$

	Tang and Chen (2009)				$\tilde{\kappa}_n^{bias}$			$\tilde{\kappa}_n^{RMSE}$	
					(Bias reducing)			(RMSE reducing)	
	$\hat{\kappa}$	J	B	I	(1)	(2)	(3)	(1)	(2)
Model 2									
% bias	*53.0*	*-5.23*	*0.861*	*-7.61*	-2.40	-2.91	-1.95	-5.68	-19.5
RMSE	*0.189*	*0.171*	*0.147*	*0.14*	0.162	0.147	0.154	0.080	0.065
Model 3									
% bias	*76.6*	*-7.7*	*2*	*-10.6*	-4.91	-1.81	-0.640	13.9	7.42
RMSE	*0.17*	*0.159*	*0.147*	*0.116*	0.148	0.164	0.142	0.084	0.049

Models 2 and 3 in Tang and Chen (2009) use $(\kappa, \alpha, \sigma^2) = (0.215, 0.0891, 0.0005)$ and $(0.140, 0.0891, 0.0003)$, respectively. Columns B, J and I are Monte Carlo results obtained by Tang and Chen (2009) for the parametric bootstrap method (ibid.) the Quenouille jackknife proposed for diffusion models in Phillips and Yu (2005), and the Indirect Inference methodology due to Gourieroux et al. (1993).

3.2. CIR model

The Nowman pseudo-ML method, which has been extended to Constant Elasticity of Variance (CEV) models in Iglesias and Phillips (2019), starts by making a discrete approximation to the diffusion function in the CIR model, setting $X(t) = X_{mh}$ for each h units of time while keeping $X(t)$ continuous in the drift term:

$$dX_t = \kappa(\alpha - X_t)dt + \sigma\sqrt{X_{mh}}dB(t)$$

for $t \in [mh, mh + h)$. The approximate process then has an exact discretisation in a convenient form for quasi maximum likelihood estimation, the result of which is a closed-form pseudo ML estimation of the CIR parameters:

$$\hat{\kappa} = -h^{-1}ln(\hat{\beta}_1)$$

where

$$\hat{\beta}_1 = \frac{n^{-2}\sum_{i=1}^{n} X_i \sum_{i=1}^{n} X_{i-1}^{-1} - n^{-1}\sum_{i=1}^{n} X_i X_{i-1}^{-1}}{n^{-2}\sum_{i=1}^{n} X_{i-1}\sum_{i=1}^{n} X_{i-1}^{-1} - 1}.$$

Following the success in Section 3.1 of finding improved estimators of κ for the relatively simple Vasicek model, designed to work at low values of κ and particular sample sizes, the objective for the CIR model is to find estimators that improve bias or RMSE performance across a wide range of κ values and sample sizes. A Pade approximant approach is used this time, with the new estimators taking the form in (9) and (10) but with g_j defined as in (6) for $j = 1, \dots, r$, moreover we set $m_1 = m_2 = 10$ and $r = 2$. Two training schemes are considered, based on the Case (3) parameterisation in the previous section. The first case, Case (1), is particularly similar and is designed for bias reduction over the same region of the parameter space at a specific sample size ($n = 500$ in Figure 4, and sample sizes 120, 300 and 500 in Table 5), but it uses the finer selection of κ training values in $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6,\ 0.7, 0.8, 0.9, 0.97\}$, while the second, Case (2), is an identical repeat of the previous section case but including the two sample sizes in $\{120, 500\}$ rather than just a sample size 500. Reduced-bias estimators are obained using the two different training setups, and the new estimator resulting from Case (2) is intended to be usable across sample sizes in the interval $[120, 500]$. A reduced-RMSE estimator is obtained using the Case (2) training setup aswell, where this time the relative performance constraint is not imposed, and this is denoted by $\tilde{\kappa}^{RMSE,\star}$ in Table 5.

Figure 4 plots the % bias of the Case (1) reduced-bias estimator and the original estimator across a range of different κ and σ values at $n = 500$, and plots the relative RMSE. It can be seen that the reduced-bias estimator has substantially lower bias, and the RMSE performance is also substantially better overall, there also does not appear to be any overtraining towards the particular parameterisations used in the training.

<Figure 4 here>

Table 5 compares Monte Carlo results for the new estimators with Tang and Chen (2009) again. The values of $(\kappa, \alpha, \sigma^2)$ in Models 1, 2 and 3 are, respectively, $(0.892, 0.09, 0.033)$, $(0.223, 0.09, 0.008)$ and $(0.148, 0.09, 0.005)$. While the remaining bias after bootstrap estimation in the three Tang and Chen (2009) models is as high as 39.6%, and the Indirect Inference bias is as high as 43.5%, the highest absolute biases for $\tilde{\kappa}^{bias}$ and $\tilde{\kappa}_n^{bias}$ are 5.99% and 3.10%, respectively. The new estimator $\tilde{\kappa}^{bias}$, which is trained to work at sample sizes in the interval $[120, 500]$, has lower bias than Indirect Inference for all three models, and lower bias than the bootstrap except in Model 2. The new estimator $\tilde{\kappa}_n^{bias}$, which is specialised for each sample size, has lower bias than the bootstrap and Indirect Inference in all three models. The RMSE values for the new reduced-bias estimators are marginally larger in most cases, but still typically lower than for the original estimator $\hat{\kappa}$. The RMSE of $\tilde{\kappa}^{RMSE}$ is half the original in two cases, where it also substantially lower than Bootstrap and Indirect Inference values, and is lower in all but one case than the original estimator, where the RMSE of the original is at its lowest. The setup could perhaps be adjusted, to make the improvement more uniform over different values of n.

Table 5: Comparison with *Tang and Chen (2009)*, CIR models

			Tang and Chen (2009)			$\tilde{\kappa}_n^{bias}$, $\tilde{\kappa}^{bias}$		$\tilde{\kappa}^{RMSE,\star}$
						(Bias reducing)		(RMSE reducing)
	n		$\hat{\kappa}$	B	I	(1)	(2)	(2)
Model 1	120	% bias	*52.0*	*0.178*	*2.68*	-0.104	-0.0783	-30.5
		RMSE	*0.780*	*0.651*	*0.603*	0.806	0.814	0.303
	300	% bias	*20.1*	*-0.447*	*-3.79*	0.0891	-0.120	-2.49
		RMSE	*0.380*	*0.326*	*0.328*	0.337	0.362	0.241
	500	% bias	*12.0*	*0.826*	*0.258*	0.183	-0.489	-0.188
		RMSE	*0.269*	*0.245*	*0.248*	0.255	0.259	0.204
Model 2	120	% bias	*228*	*13.6*	*43.5*	-1.22	-1.22	23.0
		RMSE	*0.719*	*0.502*	*0.495*	0.596	0.586	0.275
	300	% bias	*82.8*	*3.461*	*-14.92*	0.495	-5.99	16.2
		RMSE	*0.289*	*0.226*	*0.208*	0.242	0.235	0.235
	500	% bias	*48.6*	*1.325*	*-6.728*	-0.0329	-4.23	9.90
		RMSE	*0.183*	*0.15*	*0.14*	0.156	0.142	0.161
Model 3	120	% bias	*350*	*39.597*	*19.17*	3.10	0.067	28.4
		RMSE	*0.719*	*0.507*	*0.484*	0.577	0.595	0.314
	300	% bias	*129*	*4.459*	*17.67*	1.21	-3.45	18.4
		RMSE	*0.289*	*0.214*	*0.209*	0.237	0.202	0.240
	500	% bias	*74.5*	*1.83*	*-8.45*	-1.09	-2.14	11.4
		RMSE	*0.135*	*0.133*	*0.122*	0.148	0.116	0.160

4. Integer Autoregressive (INAR) models

The Integer Autoregressive (INAR) class of models was originally proposed in Al-Osh and Alzaid (1987) as a method of modeling dependent series of low counts, and there has been a growing interest in the area. Some recent contributions include Martin et al.

(2014), Sant‘Anna (2017) and Harris and McCabe (2018). As noted in the latter, INAR models have been used for applications in economics, medicine, environmental studies, and commerce. We address the estimation of the Binomial thinning parameter α in INAR(1) models with Poisson and Negative Binomial (NB) innovations. The INAR(1) model takes the following form:

$$X_t = \alpha \circ X_{t-1} + \varepsilon_t,$$

where

$$\alpha \circ X = \sum_{i=1}^{X} Y_i$$

counts the number of successes from X i.i.d. Bernoulli trials, each indendent of X. The sequence of Bernoulli random variables $\{Y_i\}_1^X$ has $P(Y_i = 1) = 1 - P(Y_i = 0) = \alpha$ and $\alpha\circ$ is known as the thinning operator, while ε_t is a Poisson(μ) or NB(μ, π) distributed innovations term where the latter allows "overdispersed" cases with a variance greater than μ. There is therefore dependence between the current count and the number ‘surviving’ from the previous period. INAR models are particularly natural when the count has the interpretation of being a stock variable, as noted for example in Harris and McCabe (2018), but they also have a more general applicability by providing a way to model dependence between current and past observations of an integer-valued variable. Gourieroux and Jasiak (2004), for example, use the approach to model insurance claim arrivals.

We address the estimation of α by starting with the Conditional Least Squares (CLS) estimator in Al-Osh and Alzaid (1987) and searching for improvements, first at a specific sample size $n = 30$ and then for any sample size in the interval $[20, 40]$. The new estimators use, as in Section 3.2, the Pade form of g_j defined in (6) with $r = 2$ and $m_1 = m_2 = 10$, this time with the CLS estimator $\hat{\alpha}$ as the initial estimator. Two training schemes are considered, and, in both, the same three-point training set of parameterisations is used for α as it was for κ in the diffusion models and λ in the AR(1). In both, each value of α is combined with the two different values of the disturbance parameter μ in $\{0.5, 4\}$, so that there are six training parameterisations in total, where the unconditional mean count for these, given by $E[X_t] = \mu/(1 - \alpha)$, ranges from 0.56 to 133. The Case (1) training setup uses just one sample size $n = 30$, and two estimators are obtained for this specific sample size: a reduced-bias estimator $\tilde{\alpha}_n^{bias}$, and a reduced-RMSE estimator $\tilde{\alpha}_n^{RMSE}$. The Case (2) training setup includes the two sample sizes in $\{20, 40\}$, making twelve training points in total, with the aim of finding a reduced-bias estimator $\tilde{\alpha}^{bias}$ that works well over all sample sizes in $[20, 40]$.

Table 6: Percentage bias and RMSE comparison, INAR(1), Poisson innovations, $n = 30$

	% Bias				RMSE			
α	$\hat{\alpha}$	$\tilde{\alpha}_n^{bias}$	$\tilde{\alpha}_n^{RMSE}$	$\tilde{\alpha}^{bias}$	$\hat{\alpha}$	$\tilde{\alpha}_n^{bias}$	$\tilde{\alpha}_n^{RMSE}$	$\tilde{\alpha}^{bias}$
0.2	-28.6	-3.34	4.88	-3.82	0.193	0.191	0.143	0.192
0.4	-19.5	-4.12	-8.33	-4.36	0.196	0.194	0.179	0.195
0.6	-16.7	-3.04	-6.16	-2.62	0.196	0.195	0.204	0.197
0.8	-15.9	-1.28	-2.29	0.0183	0.197	0.187	0.193	0.195

Table 6 compares the three estimators with the original for various α with μ set at 1 and $n = 30$. These are relatively low-count cases where the unconditional mean ranges from 1.25 in the case where $\alpha = 0.2$ to 5 where $\alpha = 0.8$. It can be seen that the reduced-bias estimators perform particularly well and the remaining biases are all less than 5%. There is a large reduction in RMSE at smaller values of α by using $\tilde{\alpha}_n^{RMSE}$. The Case (2) reduced-bias estimator $\tilde{\alpha}^{bias}$ was asked to work well at sample sizes 20 and 40 during the training, and seems to be working well at $n = 30$. Figure 5 confirms the performance of $\tilde{\alpha}^{bias}$ by plotting the percentage bias and RMSE results for $\tilde{\alpha}^{bias}$ and $\hat{\alpha}$ over a grid of α values and sample sizes. It can be seen that there is a large reduction in bias and, in most cases, a small improvement in RMSE performance. As was found in earlier sections, there does not appear to be any noticeable over-training at the particular parameterisations and sample sizes used in the training. The new estimator $\tilde{\alpha}^{bias}$ is a working closed-form estimator of α for the Poisson model at sample sizes in $[20, 40]$.

<Figure 5 here>

Overdispersed models have received substantial attention in the INAR literature recently, as the Poisson assumption has sometimes been found unrealistic in applications, see for example Pavlopoulos and Karlis (2008). INAR models with Negative Binomial innovations allow for overdispersion, see for example Martin et al. (2014), in particular the variance of NB(μ,π) innovations is given by $Var(\epsilon_t) = \mu/(1-\pi)$. Figure 6 presents results for a Negative Binomial version of the estimator $\tilde{\alpha}^{bias}$. It is trained as earlier, but now also at two values of π in $\{0, 0.5\}$. The case $\pi = 0$ is the limiting Poisson model, while $\pi = 0.5$ allows the variance of the innovations to be twice the mean. The results are plotted for various sample sizes and α values at $\pi = 0.4$ with $\mu = 2.5$, and also for various π and α values at $n = 30$ with $\mu = 3.5$.

<Figure 6 here>

It can be seen from Figure 6 that the remaining biases at the various values of n, α and π are small, and that the RMSE is marginally lower in the majority of cases. In particular, the new estimator performs well across the different levels of overdispersion π. The largest bias at $n = 30$ across all of the (α, π) cases considered in Figure 6 is -5.10%. A comparison can be made with Table II in Martin et al. (2014), which presents mean and RMSE results for CLS, Efficient Method of Moments (EMM) and Maximum Likelihood (ML) under a Poisson assumption (corresponding to $\pi = 0$) at $n = 50$ for $\alpha = 0.3$ where $\mu = 3.5$, in particular the smallest bias was found to be via the Maximum Likelihood estimator, which had a mean value of 0.281, a bias of -6.33%. The case $\mu = 3.5$ is admittedly close to our training value of $\mu = 4$, but the bias results for $\mu = 2.5$ in the left-hand panel are comparable, and so are those for the specialised Poisson version of the estimator in Figure 5 where $\mu = 1$. As with the EMM estimation advanced in Martin et al. (2014), the computational approximation approach does not require specification of a likelihood function, which becomes difficult when the more general INARMA class of models is considered.

5. Application: forecasting the US Federal Funds rate

Given an initial estimator, it has been possible to find closed-form adjustments computationally that target specific aspects of estimation performance and which are reusable across an interval of sample sizes. The same general approach can also be used to target an improvement in point prediction, and we consider in particular predictions of the Federal Funds overnight lending rate using the CIR model considered in Section 3.2. A monthly-sampled series is used, as in Aït-Sahalia (1999), to avoid market microstructure effects, and this was obtained from the Federal Reserve Bank of St. Louis website for the period July 1st 1954 to January 1st 2020. The series is depicted in Figure 7.

<Figure 7 here>

Given estimates of the parameters in a CIR model for $r(t)$, forecasts s steps ahead can be obtained from the conditional mean,

$$E[r(t+s)|r(t)] = \alpha + \{r(t) - \alpha\}e^{-\kappa s}, \tag{17}$$

see e.g. Orlando et al. (2020). Somewhat surprisingly, the use of reduced-bias estimates of κ was found to result in relatively poor out-of-sample forecast performance, whether via the computational approximation method in Section 3.2 or via two other bias-correction methods that were tried. The poor forecast performance was accompanied by an increased prevalance and magnitude of negative estimates of κ, particularly in the case of the Quenouille jackknife bias correction, which can potentially be explained by overcorrection of the bias or by increased variance at small values of κ. The $\tilde{\kappa}^{bias}$ estimator was seen to over-correct the bias by relatively small amounts in Table 5, while Tables 2 and 3 of Tang and Chen (2009) show that the Quenouille jackknife over-corrects the bias more substantially. It can be seen from (17) that large negative estimates of κ may lead to poor predictions if the underlying d.g.p. is stationary, particularly in periods of high volatility where $r(t)$ deviates substantially from its mean α.

Table 7 illustrates the prediction performance resulting from bias-corrected estimation of κ over rolling windows of 300 and 500 monthly observations starting on July 1st, 1954.[7] Regardless of the method of bias correction, the performance in terms of Root Mean Square Prediction Error (RMSPE) is made worse overall. Besides the new approach in Section 3.2, the Quenouille Jackknife method suggested in Phillips and Yu (2005) was tried, along with a corrected estimator based on the asymptotic bias approximation for the Nowman estimator in Tang and Chen (2009)[8]. These are denoted in what follows by $\tilde{\kappa}^{bias}$, $\hat{\kappa}_{QJ}$, and $\hat{\kappa}_{TC}$, respectively. It can be seen that much of the addition to the RMSPE occured during the 1973-75 oil crisis, which indeed appears to be a period of high volatility in the Federal Funds rate - the adverse effect of over-correction of the bias on forecasting would be amplified during this period via the term $r(t) - \alpha$ in (17).

[7] As the Nowman estimator of α is unbiased to order $O(n^{-1})$, corrections are only made to estimation of κ.

[8] Tang and Chen (2009) found, see Theorem 3.2.3, that $E[\hat{\kappa}] = \kappa + 4T^{-1} + o(T^{-1})$ where $T = nh$ is the length of time over which the n observations are taken. We define $\hat{\kappa}_{TC} = \hat{\kappa} - 4T^{-1}$ where $T = 300/12 = 25$.

Table 7: CIR prediction performance by estimation method

	$n = 300$			$n = 500$		
	RMSPE	$\kappa^{\min}$	$\kappa^{\max}$	RMSPE	$\kappa^{\min}$	$\kappa^{\max}$
Nowman ($\hat{\kappa}$)	0.0065	-0.021	0.28	0.0027	-0.025	0.27
Bias target ($\tilde{\kappa}_{bias}$)	0.057	-0.10	0.027	0.045	-0.074	0.12
Bias corrected, TC ($\hat{\kappa}_{TC}$)	0.10	-0.18	0.12	0.078	-0.12	0.18
QJ ($\hat{\kappa}_{QJ}$)	0.69	-0.80	-0.038	0.27	-0.30	0.019
ML ($\hat{\kappa}_{ML}$)	0.0069	-0.00089	0.32	0.0033	-0.0056	0.27

Root Mean Squared Prediction Error for rolling window one-step forecasts, $n = 300$ and $n = 500$ monthly observations. The $\kappa^{\min}$ and $\kappa^{\max}$ columns record the smallest and largest estimates of κ yielded by each estimation method over the different time windows, respectively. $\hat{\kappa}^{(2)}_{ML}$ used Nelder-Mead with the grid search used in cases where convergence failed or where there was no movement from the starting value, and was used along with ML estimates $\hat{\alpha}_{ML}$ and $\hat{\sigma}^2_{ML}$.

Figure 8 presents Monte Carlo simulations of the RMSPE using the original and bias corrected estimators on data generated from CIR models with various values of the mean reversion parameter κ, and this further illustrates the issue. Reduced bias estimation of κ in a correctly specified CIR model, whether via the computational approximation approach in Section 3.2, the analytical approximation in Tang and Chen (2009) or the ($m = 4$) Quenouille jackknife, tends to reduce prediction performance at low levels of mean reversion, while use of the Quenouille jackknife bias correction also reduces prediction performance at higher levels of κ. The figure also shows the prediction performance using a new estimator, $\tilde{\kappa}^{pred}$, introduced in the next subsection, which adjusts the Nowman estimator specifically for the purpose of prediction performance - this prediction targeting estimator performs best throughout.

<Figure 8 here>

5.1. Prediction targeting

Some attempts were made at modifiying the correction term computed in Section 3.2 by placing a constraint on the (Monte Carlo estimated) probability of $\tilde{\kappa}$ being negative when generating the bias corrected estimator. The bias correction was then relatively conservative at lower values of κ, and this also affected the forecast performance adversely. To target the prediction performance directly, it is possible simply to modify the loss function L in the general procedure so that R in (5) is filled with the RMSPE values at different training parameterisations $\Theta = (\alpha, \kappa, \sigma^2)$, rather than with, as in Section 3.2, the relative bias or RMSE values.

Specifically, the vector R is comprised of the Monte Carlo computed values $(E[L(\tilde{\kappa}; \Theta)])^{\frac{1}{2}}$ at different training parameter points Θ, with the loss function now defined as

$$L(\tilde{\kappa}; \Theta) = (y^f - y)^2 \tag{18}$$

where y^f is the predicted value of y using a candidate estimator at the parameter point Θ. The estimator $\tilde{\kappa}^{pred}$ is then selected in the same way as in Section 3.2, by choosing the parameterisation for rational approximants g_1 and g_2 in Padé form that minimise $||R||$, possibly subject to constraints on performance relative to the original estimator or other estimators though this is not done here. To reflect the relatively small range of Nowman estimates of κ found at windows sizes $n = 300$ and $n = 500$ for the US Federal Funds rate, as seen by the minimum and maximum values in Table 7, the prediction targeting estimator was trained using the values $\kappa \in \{0.01, 0.1, 0.2\}$, while α and σ^2 were set as in Section 3.2 at 0.05.

The Monte Carlo performance of the prediction-targeting approach can be seen in Figure 8 alongside the prediction performance using other estimators of κ. As noted earlier, $\tilde{\kappa}^{pred}$ compares well in terms of RMSPE - it outperforms the other estimators at all values of κ considered, and even performs well at very low values of κ. Across the values of κ tried, the smallest percentage reduction in RMSPE using $\tilde{\kappa}^{pred}$ was found to be 22%, while the largest reduction in RMSPE was 36%.

Figure 9 illustrates the rolling-window out-of-sample forecast performance of the approach using the US Federal Funds rate series. The estimated root mean squared prediction errors based on $\tilde{\kappa}^{pred}$ are compared with those based on the Nowman estimator across a series of window sizes between 120 and 500. As in the Monte Carlo simulation, using the prediction targeting estimator $\tilde{\kappa}^{pred}$ results in superior out-of-sample forecasts from the CIR model. The performance is improved substantially at every window size.

<Figure 9 here>

5.2. Further investigation of $\tilde{\kappa}^{pred}$

The remaining analysis explores $\tilde{\kappa}^{pred}$ further, in order to understand its properties better. Figure 10 illustrates how the individual out-of-sample forecast errors for the Federal Funds rate relate to the initial Nowman estimates $\hat{\kappa}$, and provides the mapping of $\hat{\kappa}$

estimates to $\tilde{\kappa}^{pred}$ estimates and box plots for $\tilde{\kappa}^{pred}$. The plot of prediction error vs $ka\hat{p}pa$ appears to show a 'fanning' effect either side of estimates of κ slightly above zero, where there is also a relatively dense concentration of estimates - at $n = 150$ this happens at around $\kappa = 0.1$, while at $n = 450$ it is at around $\kappa = 0.02$. Meanwhile, it can be seen from the mapping of $\hat{\kappa}$ to $\tilde{\kappa}^{pred}$ that the estimates from the new methodology are far more concentrated near zero and, when they are negative, they are often much less negative. As noted, this seems important for making forecasts in volatile periods more reliable.

<Figure 10 here>

Figure 11 illustrates the performance of the prediction-targeting estimator in terms of bias, RMSE and variance, while Figure 12 plots the frequency distribution of the Nowman estimator, the reduced bias estimator and the prediction-targeting estimator at $\kappa = 0.02$ and $\kappa = 0.2$. The prediction targeting estimator is less biased, far less so at small values of κ, though it over-corrects the Nowman estimator on average, and is more biased than the reduced bias estimator $\tilde{\kappa}^{bias}$. It seems possible that the bias and distribution of $\tilde{\kappa}^{pred}$ itself could be improved by putting constraints on the minimal bias performance of $\tilde{\kappa}^{pred}$ at parameter points in the training set when computing the estimator. A wider training set than $\kappa \in \{0.01, 0.3\}$ could also be used, though the estimated values of κ are typically within this range in applications.

<Figure 11 here>

6. Conclusion

A simple computational approximation approach has been shown to work well for the reduction of estimation bias in small parametric time series models, inspired by existing correction methods based on asymptotic approximation. The methodology aims to find, via Monte Carlo and numerical search, a small-order adjustment to an initial estimator in a similar form to what might be found theoretically. The restriction on the form of the added terms seems to limit issues of overtraining at particular parameterisations or sample sizes. The approach has been found effective at removing estimation bias and reducing RMSE in small parametric time-series models that have received substantial attention in the bias-correction literature, and may be especially useful where no asymptotic expansion of the bias exists, or as a second layer of bias reduction after correcting to some asymptotic order via an existing asymptotic approximation. The new estimators share with corrected estimators based on asymptotic expansion the characteristic of being closed-form and fast to compute once found.

The approach has also been shown to work well when targeting a reduction in forecast error, in particular it has been possible to improve the one-step-ahead prediction from a CIR model both in Monte Carlo simulations and in out-of-sample forecasts of the Federal Funds rate over a wide range of window lengths. The presentation and examples here have focused on point estimation and point prediction, but it seems possible to extend the approach to interval estimation and prediction, mirroring the type of asymptotic corrections that can be obtained theoretically by methods such as Edgeworth expansion[9]. The adjusted coefficient estimates are also computationally simple and could potentially be bootstrapped. Regularised optimisation methods commonly used within the deep learning literature may offer a means to extend the methodology to substantially larger models, and this is being investigated in related work.

[9] See for example Rothenberg (1984) and Hausman and Palmer (2012).

References

Aït-Sahalia, Y., 1999. Transition densities for interest rate and other nonlinear diffusions. Journal of Finance LIV (4), 1361 – 1395.

Aït-Sahalia, Y., 2002. Maximum likelihood estimation of discretely sampled diffusions: a closed-form approximation approach. Econometrica 70 (1), 223 – 262.

Al-Osh, A., Alzaid, A., 1987. First-order integer-valued autoregressive (INAR(1)) process. Journal of Time Series Analysis 8(3), 261–275.

Baltussen, G., van Bekkum, S., Da, Z., 2019. Indexing and stock market serial dependence around the world. Journal of Financial Economics 132 (1), 26 – 48.

Bao, Y., 2007. The approximate moments of the least squares estimator for the stationary autoregressive model under a general error distribution. Econometric Theory 23(05), 1013–1021.

Chambers, M. J., 2013. Jackknife estimation of stationary autoregressive models. Journal of Econometrics 172 (1), 142 – 157.

Chen, Z., Chen, F., Lai, R., Zhang, X., Lu, C., 2018. Rational neural networks for approximating graph convolution operator on jump discontinuities. 2018 IEEE International Conference on Data Mining.

Cox, J., Igersoll, J., Ross, S., 1985. A theory of the term structure of interest rates. Econometrica 53, 385–407.

Gallant, A. R., Tauchen, G., 1996. Which moments to match? Econometric Theory 12 (4), 657681.

Gourieroux, C., Jasiak, J., 2004. Heterogeneous INAR(1) model with application to car insurance. Insurance: Mathematics and Economics 34 (2), 177 – 192.

Gourieroux, C., Monfort, A., Renault, E., 1993. Indirect inference. Journal of Applied Econometrics 8 (2), S85–S118.

Harris, D., McCabe, B., 2018. Semiparametric independence testing for time series of counts and the role of the support. Econometric Theory, 135.

Hausman, J., Palmer, C., 2012. Heteroskedasticity-robust inference in finite samples. Economics Letters 116 (2), 232 – 235.

Hornik, K., 1991. Approximation capabilities of multilayer feedforward networks. Neural Networks 4 (2), 251 – 257.

Iglesias, E. M., Phillips, G. D. A., 2019. Further results on pseudo-maximum likelihood estimation and testing in the constant elasticity of variance continuous time model. Journal of Time Series Analysis.

Kendall, M., 1954. Note on bias in the estimation of autocorrelation. Biometrika 61, 403–404.

Kiviet, J., Phillips, G., 2012. Higher-order asymptotic expansions of the least-squares estimation bias in first-order dynamic regression models. Computational Statistics and Data Analysis 56(11), 3706–3729.

Kiviet, J., Phillips, G., 2014. Improved variance estimation of maximum likelihood estimators in stable first-order dynamic regression models. Computational Statistics and Data Analysis 76, 424–448.

Lehmann, E., 1983. Theory of Point Estimation, 1st Edition. Wadsworth and Brooks/Cole, Belmont, California.

Liu-Evans, G., Phillips, G., 2012. Bootstrap, jackknife and COLS: bias correction and mean squared error in estimation of ARX models. Journal of Time Series Econometrics.

Marriott, F., Pope, J., 1954. Bias in the estimation of autocorrelations. Biometrika 61, 393–403.

Martin, G., McCabe, B., Frazier, D., Maneesoonthorn, W., Robert, C., 2019. Auxiliary likelihood-based approximate bayesian computation in state space models. Journal of Computational and Graphical Statistics.

Martin, V., Tremayne, A., Jung, R., 2014. Efficient method of moments estimators for integer time series models. Journal of Time Series Analysis 35, 491–516.

Nowman, K., 1997. Gaussian estimation of single-factor continuous-time models of the term structure of interest rates. Journal of Finance 52, 1695–1706.

Orcutt, G., Winokur, H., 1969. First order autoregression: inference, estimation, and prediction. Econometrica 37, 1–14.

Orlando, G., Mininni, R., Bufalo, M., 2020. Forecasting interest rates through Vasicek and CIR models: a partitioning approach. Journal of Forecasting, 1 – 11.

Pavlopoulos, H., Karlis, D., 2008. Inar(1) modeling of overdispersed count series with an environmental application. Environmetrics 19 (4), 369–393.

Phillips, P., Yu, J., 2005. Jackknifing bond option prices. Review of Financial Studies 18 (2), 707–742.

Rothenberg, T. J., 1984. Hypothesis testing in linear models when the error covariance matrix is nonscalar. Econometrica 52 (4), 827–842.

Sant'Anna, P. H. C., 2017. Testing for uncorrelated residuals in dynamic count models with an application to corporate bankruptcy. Journal of Business & Economic Statistics 35 (3), 349–358.

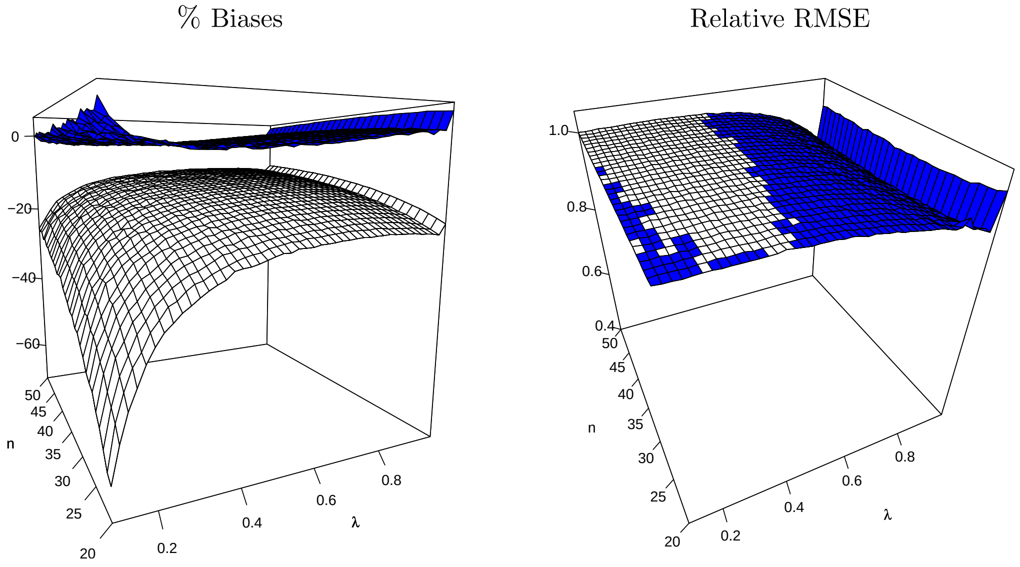
Tang, C., Chen, S., 2009. Parameter estimation and bias correction for diffusion processes. Journal of Econometrics 149 (1), 65 – 81.

Tversky, A., Kahneman, D., 1992. Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and Uncertainty 5 (4), 297 – 323.

Vasicek, O., 1977. An equilibrium characterization of the term structure. Journal of Financial Economics 5 (2), 177 – 188.

Figures

Figure 1: Bias targeting ($\tilde{\lambda}^{bias}$), percentage bias and relative RMSE comparison



Percentage bias for the OLS estimator (unshaded) vs $\tilde{\lambda}^{bias}$ (shaded). Relative RMSE values $RMSE(\tilde{\lambda}^{bias})/RMSE(\hat{\lambda})$ are shaded when less than 1.

Figure 2: RMSE targeting ($\tilde{\lambda}^{RMSE}_{COLS}$), $n = 35$.

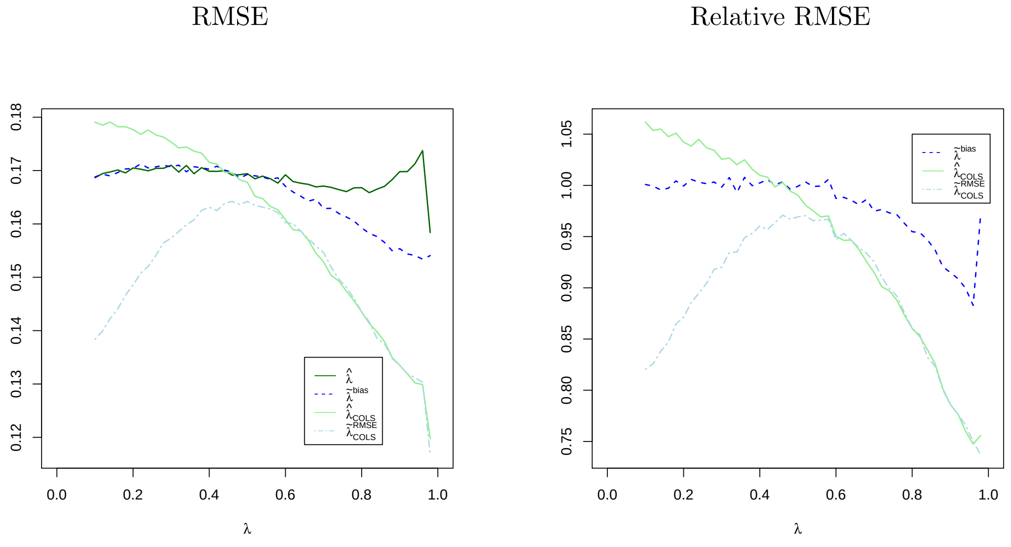
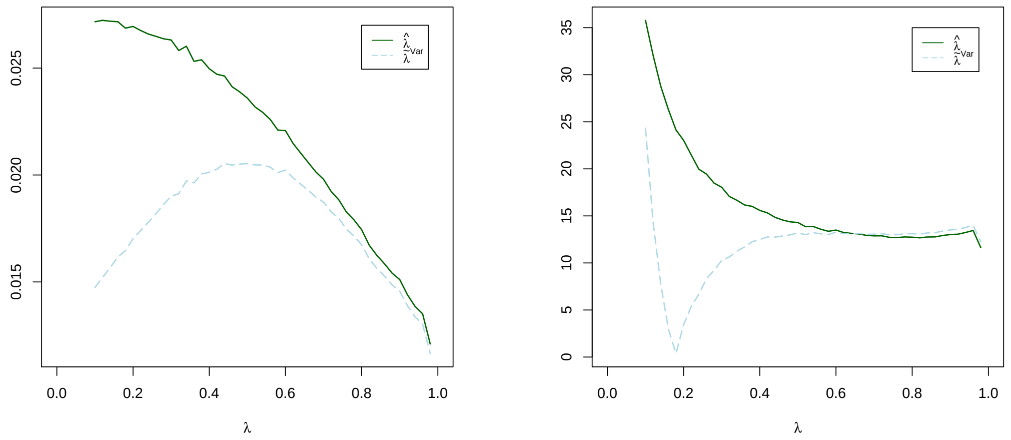
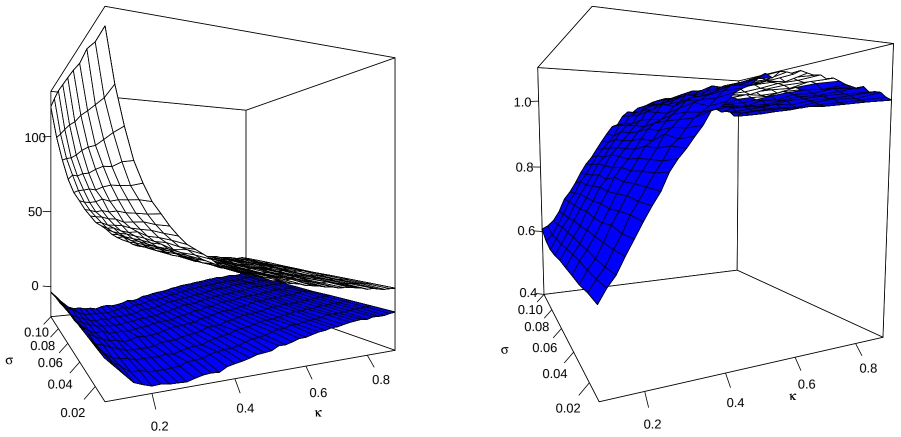


Figure 3: Variance targeting with bias constraint ($\tilde{\lambda}^{Var}_{COLS}$), $n = 35$


Variance

Absolute % bias

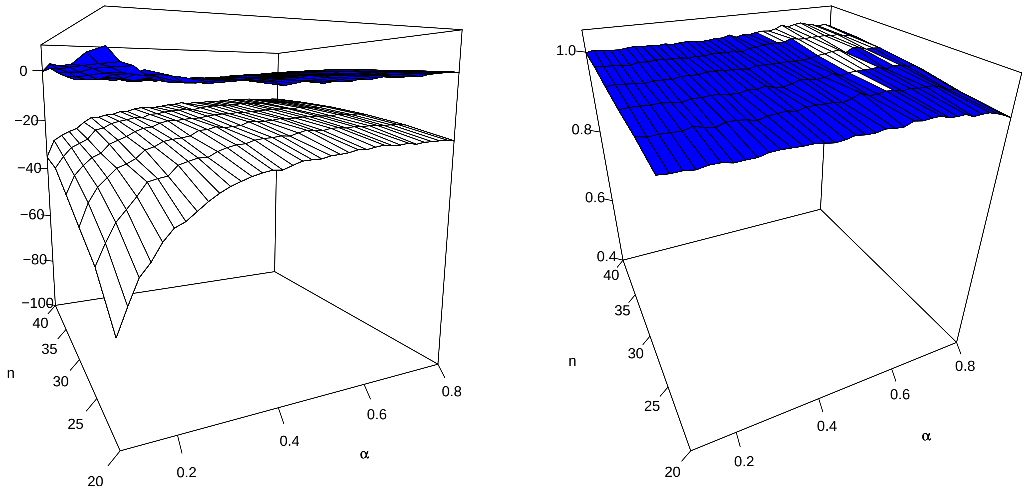
Figure 4: Bias targeting ($\tilde{\kappa}^{bias}$), percentage bias and relative RMSE comparison, $n = 500$


% Biases

Relative RMSE

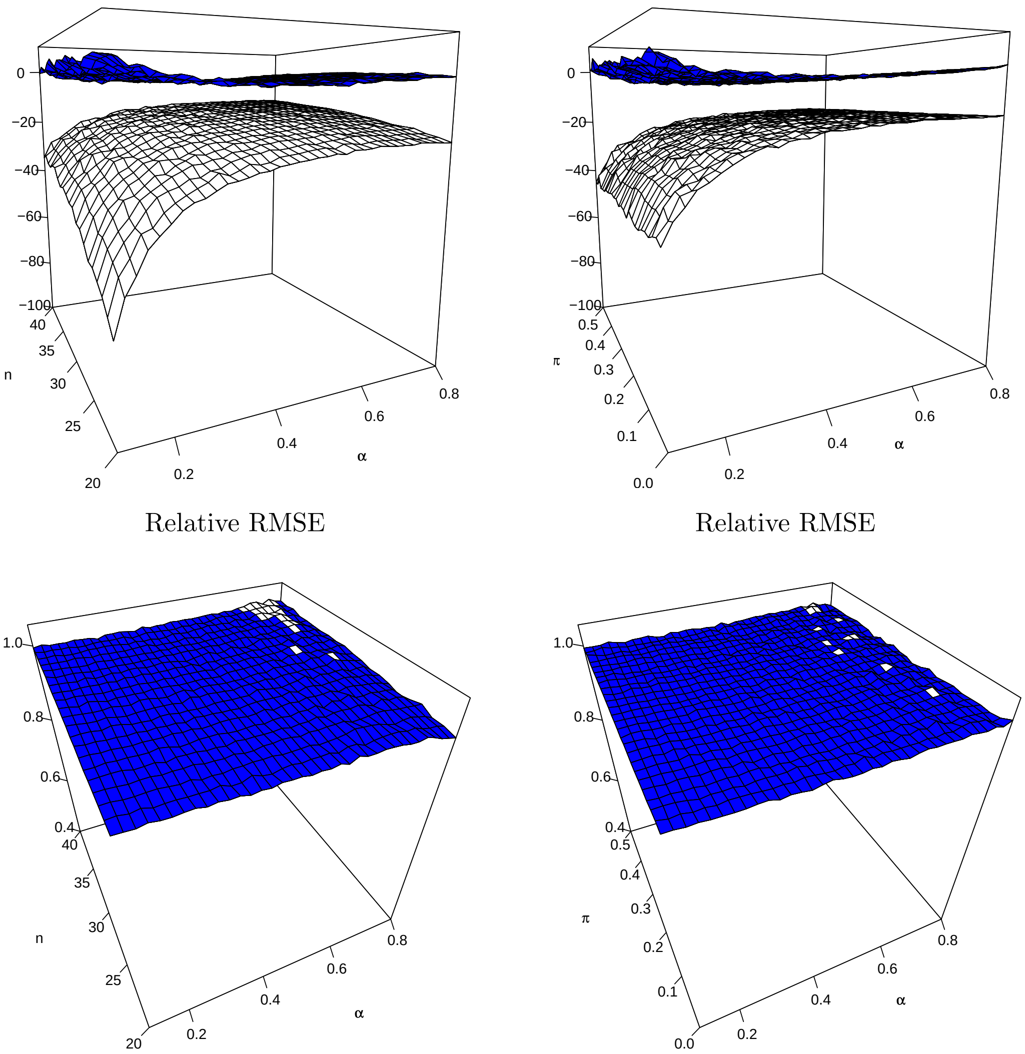
Percentage bias for $\hat{\kappa}$ (shaded) vs $\tilde{\kappa}^{bias}$ (unshaded). Relative RMSE values $RMSE(\tilde{\kappa}^{bias})/RMSE(\hat{\kappa})$ are shaded when greater than 1.

Figure 5: Bias targeting ($\tilde{\alpha}^{bias}$), percentage bias and relative RMSE comparison. Poisson innovations, $\mu = 1$



Percentage bias for the $\hat{\alpha}$ (unshaded) vs $\tilde{\alpha}^{bias}$ (shaded). Relative RMSE values $RMSE(\tilde{\alpha}^{bias})/RMSE(\hat{\alpha})$ are shaded when lower than 1.

Figure 6: Bias targeting ($\tilde{\alpha}^{bias}$), percentage bias and relative RMSE comparison. Negative Binomial innovations.



($\pi = 0.4$, $\mu = 2.5$)

% bias

($n = 30$, $\mu = 3.5$)

% bias

Relative RMSE

Relative RMSE

Percentage bias for the $\hat{\alpha}$ (unshaded) vs $\tilde{\alpha}^{bias}$ (shaded). Relative RMSE values $RMSE(\tilde{\alpha}^{bias})/RMSE(\hat{\alpha})$ are shaded when lower than 1.

Figure 7: US Federal Funds rate, July 1st 1954 to January 1st 2020

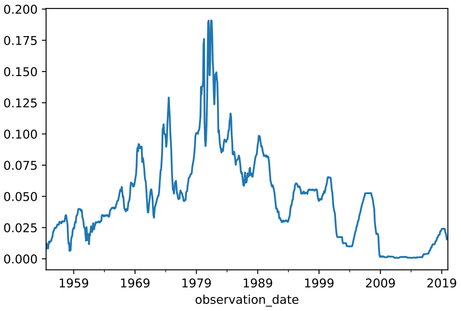


Figure 8: Simulated root mean square prediction error by κ value and estimation method, $n = 500$, $\alpha = \sigma = 0.05$.

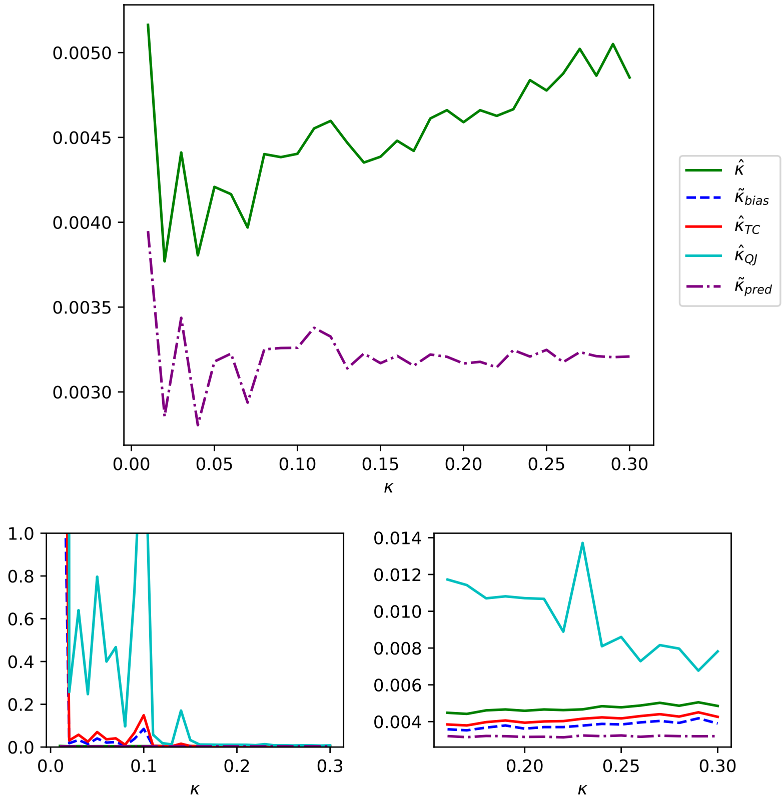


Figure 9: Out-of-sample prediction performance by window size.

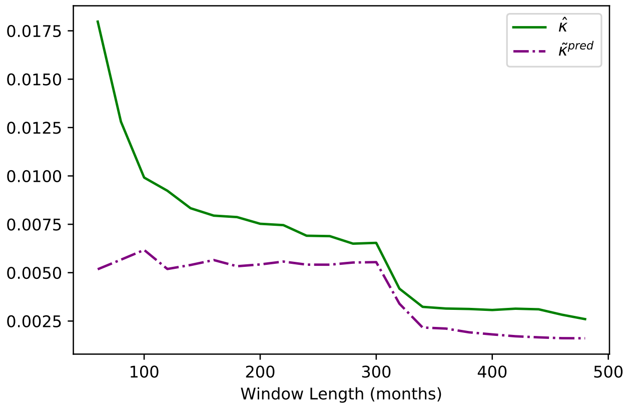
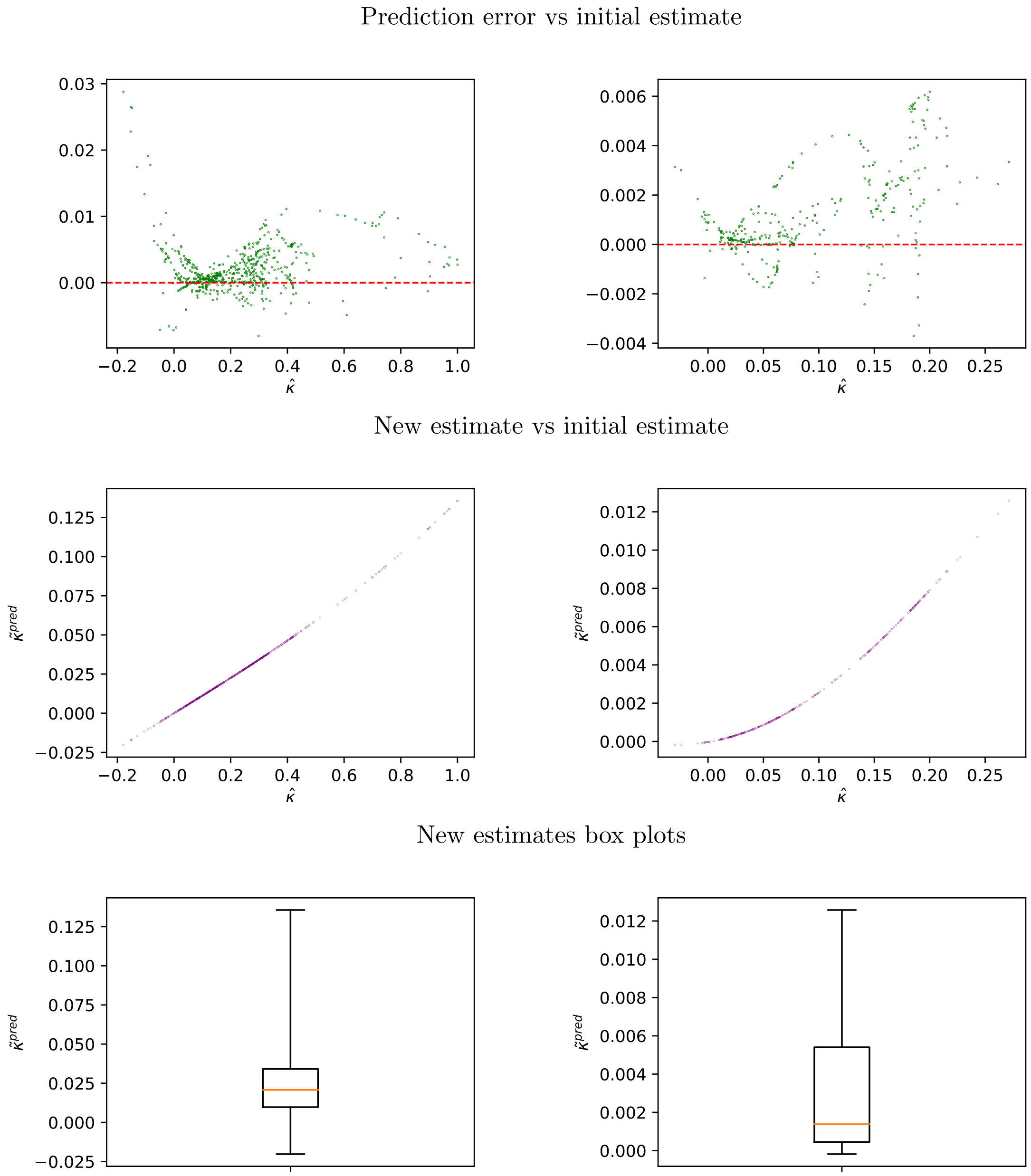


Figure 10: Out-of-sample prediction error vs κ estimates, the $\hat{\kappa}$ to $\tilde{\kappa}^{pred}$ mapping, and $\tilde{\kappa}^{pred}$ box plots. Window sizes $n = 150$ (left) and $n = 450$ (right).

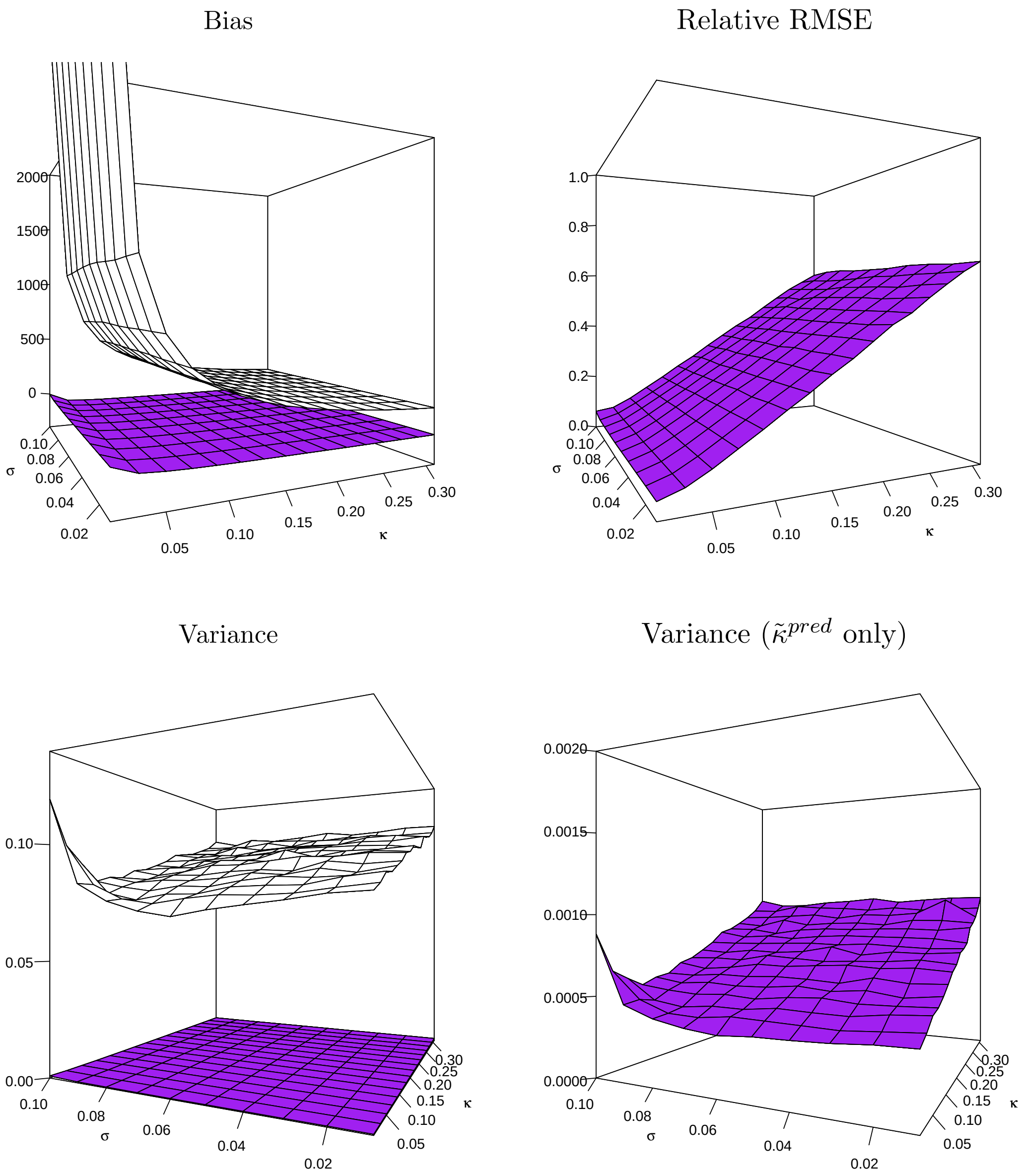


Prediction error vs initial estimate

New estimate vs initial estimate

New estimates box plots

Figure 11: Further Monte Carlo investigation of $\tilde{\kappa}^{pred}$ - Bias and RMSE, $n = 200$.


Bias

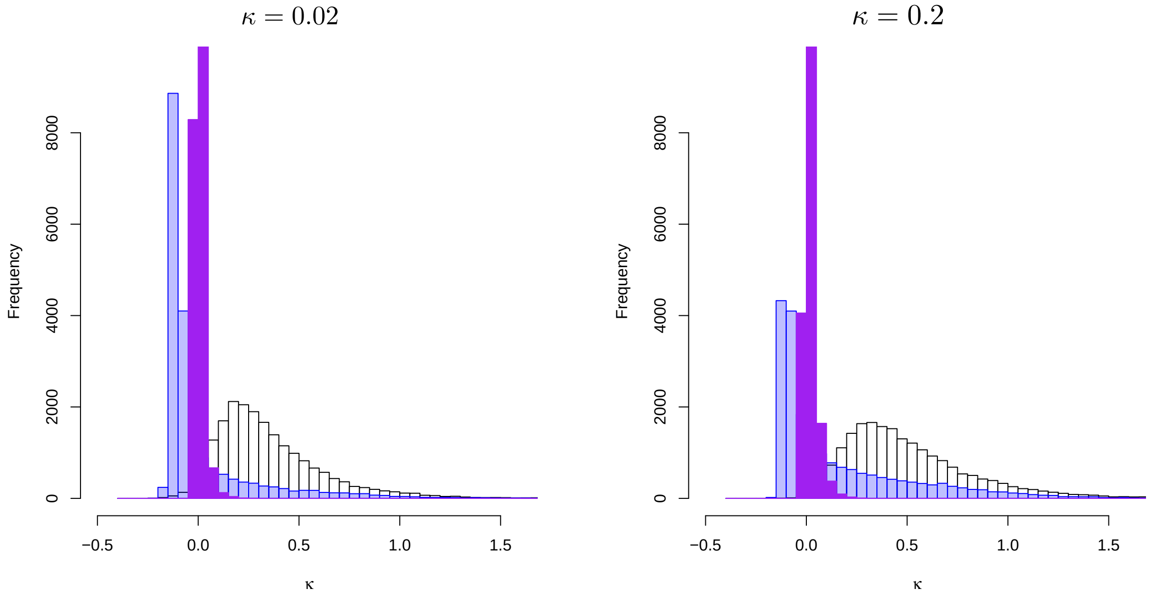
Relative RMSE

Variance

Variance ($\tilde{\kappa}^{pred}$ only)

Percentage bias for the $\hat{\kappa}$ (unshaded) vs $\tilde{\kappa}^{pred}$ (shaded). Relative RMSE values $RMSE(\tilde{\kappa}^{pred})/RMSE(\hat{\kappa})$ are shaded when lower than 1.

Figure 12: Further Monte Carlo investigation of $\tilde{\kappa}^{pred}$ - frequency distribution, $n = 200$.



Frequency distribution of $\hat{\kappa}$ (unshaded), $\tilde{\kappa}^{bias}$ (light blue) and $\tilde{\kappa}^{pred}$ (purple) estimates using 20,000 replications.

Appendix

Implementation of the relative performance constraint

The implementation in the examples of Sections 2-4 requires that the bias and RMSE values of the new estimator be no greater than those of the original at each point in the training set $\mathcal{T}$. Let ab_o and $RMSE_o$ denote the vectors of absolute bias and RMSE values for the original estimator corresponding to points in $\mathcal{T}$, while ab and $RMSE$ are similar vectors for the new estimator. Let d_{ab} and d_{RMSE} then denote the maximal (signed) elements of the vectors $ab - ab_o$ and $RMSE - RMSE_o$, respectively. It is required that $d_{ab} \leq 0$ and $d_{RMSE} \leq 0$, and to achieve this the parameter vector w, which defines G in (9) once $\mathcal{G}$ is chosen, is selected to minimise the value of the penalised objective function:

$$\mathcal{L} = ||E|| + \lambda\{max(d_{ab}, 0)^2 + max(d_{RMSE}, 0)^2\}$$

where $\lambda > 0$ is large. This simple penalty function method was sufficient for the applications that were considered, using the `subplex` global optimisation algorithm.

The data and code for the application and methods used in Section 5 are available at the author's GitHub page. There is no conflict of interest relating to the author and this paper.