Journal of Econometrics 209 (2019) 18–34

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

# Nearly weighted risk minimal unbiased estimation☆

Ulrich K. Müller [a,*], Yulong Wang [b]

[a] Economics Department, Princeton University, United States
[b] Economics Department, Syracuse University, United States

ARTICLE INFO

*Article history:*
Received 26 July 2017
Received in revised form 7 August 2018
Accepted 27 November 2018
Available online 18 December 2018

*JEL classification:*
C13, C22

*Keywords:*
Mean bias
Median bias
Autoregression
Quantile unbiased forecast

ABSTRACT

Consider a small-sample parametric estimation problem, such as the estimation of the coefficient in a Gaussian AR(1). We develop a numerical algorithm that determines an estimator that is nearly (mean or median) unbiased, and among all such estimators, comes close to minimizing a weighted average risk criterion. We also apply our generic approach to the median unbiased estimation of the degree of time variation in a Gaussian local-level model, and to a quantile unbiased point forecast for a Gaussian AR(1) process.


## 1. Introduction

Competing estimators are typically evaluated by their bias and risk properties, such as their mean bias and mean-squared error, or their median bias. Often estimators have no known small sample optimality. What is more, if the estimation problem does not reduce to a Gaussian shift experiment even asymptotically, then in many cases, no optimality claims can be made even in large samples. For example, the analysis of the bias in the AR(1) model has spawned a very large literature (see, among others, Hurvicz (1950), Marriott and Pope (1954), Kendall (1954), White (1961), Phillips (1977, 1978), Sawa (1978), Tanaka (1983, 1984), Shaman and Stine (1988) and Yu (2012) ), with numerous suggestions of alternative estimators with less bias (see, for instance, Quenouille (1956), Orcutt and Winokur (1969), Andrews (1993), Andrews and Chen (1994), Park and Fuller (1995), MacKinnon and Smith (1998), Cheang and Reinsel (2000), Roy and Fuller (2001), Crump (2008), Phillips and Han (2008), Han et al. (2011) and Phillips (2012)). With the exception of the conditional optimality of Crump's (2008) median unbiased estimator in a model with asymptotically negligible initial condition, none of these papers make an optimality claim.

In this paper we consider parametric small sample problems, and seek estimators that come close to minimizing a weighted average risk (WAR) criterion, under the constraint of having uniformly low bias. Our general framework allows for a wide range of loss functions and bias constraints, such as mean or median unbiasedness. The basic approach is to finitely discretize the bias constraint. For instance, one might impose zero or small bias only under $m$ distinct parameter values. Under this discretization, the derivation of a WAR minimizing unbiased estimator reduces to a Lagrangian problem with $2m$ non-negative Lagrange multipliers. The Lagrangian can be written as an integral over the data, since both the objective and

☆ The authors thank seminar participants at Harvard/MIT, Princeton, University of Kansas and University of Pennsylvania, an associate editor and two referees for helpful comments and suggestions. Müller gratefully acknowledges financial support by the NSF via grant SES-1226464.

* Corresponding author.
*E-mail address:* umueller@princeton.edu (U.K. Müller).

https://doi.org/10.1016/j.jeconom.2018.11.016
0304-4076/

the constraints can be written as expectations. Thus, for given multipliers, the best estimator simply minimizes the integrand for all realizations of the data, and so is usually straightforward to determine. Furthermore, it follows from standard duality theory that the value of the Lagrangian, evaluated at arbitrary non-negative multipliers, provides a lower bound on the WAR of any estimator that satisfies the constraints. This lower bound holds a fortiori in the non-discretized version of the problem, since the discretization amounts to a relaxation of the uniform unbiasedness constraint.

We use numerical techniques to obtain approximately optimal Lagrange multipliers, and use them for two conceptually distinct purposes. On the one hand, close to optimal Lagrange multipliers imply a large, and thus particularly informative lower bound on the WAR of any nearly unbiased estimator. On the other hand, close to optimal Lagrange multipliers yield an estimator that is nearly unbiased in the discrete problem. Thus, with a fine enough discretization and some smoothness of the problem, this estimator also has uniformly small bias. Combining these two usages then allows us to conclude that we have in fact identified a nearly WAR minimizing estimator among all estimators that have uniformly small bias.

These elements – discretization of the original problem, analytical lower bound on risk, numerical approximation – are analogous to Elliott et al.'s (2015) approach to the determination of nearly weighted average power maximizing tests in the presence of nuisance parameters. Our contribution here is the transfer of the same ideas to estimation problems. We do not consider the construction of corresponding confidence intervals, but focus exclusively on the point estimation problem.

There are two noteworthy special cases that have no counterpart in hypothesis testing. First, under squared loss and a mean bias constraint, the Lagrangian minimizing estimator is linear in the Lagrange multipliers. WAR then becomes a positive definite quadratic form in the multipliers, and the constraints are a linear function of the multipliers. The numerical determination of the multipliers thus reduces to a positive definite quadratic programming problem, which is readily solved even for large $m$ by well-known and widely implemented algorithms.[1]

Second, under a median unbiasedness constraint and in absence of any nuisance parameters, it is usually straightforward to numerically determine an exactly median unbiased estimator by inverting the median function of a suitable statistic, even in a non-standard problem. This approach was prominently applied in econometrics in Stock (1991), Andrews (1993) and Stock and Watson (1998), for instance. However, the median inversion of different statistics typically yields different median unbiased estimators. Our approach here can be used to determine the right statistic to invert under a given WAR criterion: Median inverting a nearly WAR minimizing nearly median unbiased estimator typically yields an exactly median unbiased nearly WAR minimizing estimator, as the median function of the nearly median unbiased estimator is close to the identity function.

In addition to the estimation of the AR(1) coefficient, we apply our general approach to two other small sample time series estimation problems. The first is the estimation of the degree of time variation in a Gaussian local-level model. After taking first differences, this becomes the problem of estimating the coefficient in a Gaussian MA(1). When the true MA(1) coefficient is close to being non-invertible, the maximum likelihood estimator suffers from the so-called pile-up problem, that is the coefficient is estimated to be exactly unity with positive probability. Stock and Watson (1998) derive an exactly median unbiased estimator by median inverting the Nyblom (1989) statistic. We derive an alternative, nearly WAR minimizing median unbiased estimator under absolute value loss and find it to have very substantially lower risk unless the degree of time variation is very small.

The second additional problem concerns forecasts from a stationary Gaussian AR(1) model. We use our framework to determine nearly quantile unbiased forecasts, that is, in repeated applications of the forecast rule for the $\alpha$ quantile, the future value realizes to be smaller than the forecast with probability $\alpha$, under all parameter values. See Phillips (1979), Stock (1996), Kemp (1999), Gospodinov (2002), Elliott (2006) and Müller and Watson (2016) for related studies.

After imposing appropriate invariance constraints, the parameter space in our examples is always one dimensional. Our general approach successfully determines an exactly or very nearly unbiased small sample estimator uniformly over this scalar parameter, with a WAR that is within 1% of the lower bound. We provide corresponding Matlab code in the replication files for sample sizes as small as $T = 5$ and as large as $T = 400$. For a given sample size, the computation of the Lagrange multipliers takes seconds (for the nearly mean unbiased estimator) or a few minutes (for the nearly median or quantile unbiased estimators) on a modern computer, suggesting that it might be possible to apply the technique also to, say, a two dimensional problem.

The risk profile of our new estimators is comparable or more attractive than previously suggested, often biased estimators. Doss and Sethuraman (1989) show that if no mean unbiased estimator exists, then the variance of a nearly mean unbiased estimators is necessarily large. Our results thus suggest that an exactly mean unbiased estimator of the coefficient in a Gaussian AR(1) exists. It would be interesting to corroborate this conjecture. At the same time, we view it as a strength of our generic computational approach that it does not require the ingenious derivation of a (nearly) unbiased estimator.[2]

An unbiasedness constraint may be motivated in a variety of ways. A first motivation may simply stem from the definition of unbiasedness. For example, the average of individual unbiased AR(1) estimators in a panel with many independent individuals but a common value of the autoregressive coefficients takes on values close to the true parameter with high probability by the law of large numbers. Also, regulatory or other institutional constraints might make it desirable that, in repeated quantile forecasts, the realized value takes on values smaller than the forecast $100\alpha\%$ of the time.

1 In this special case, and with a weighting function that puts all mass at one parameter value, the Lagrangian bound on WAR reduces to a Barankin (1946)-type bound on the MSE of an unbiased or biased estimator, as discussed by McAulay and Hofstetter (1971), Glave (1972) and Albuquerque (1973).

2 A recent example of an ingenious construction is Andrews and Armstrong (2017), who derive a mean unbiased estimator for the structural parameter under potentially weak instruments.

A second potential motivation relies on unbiasedness as a device to discipline estimators. As is well known, minimizing weighted risk without any side constraint leads to the Bayes estimator that, for each data draw, minimizes posterior expected loss, with a prior proportional to the weight function. The weight function then has an enormous influence on the resulting estimator; for example, a degenerate weight with all mass on one parameter value leads to an estimator that entirely ignores the data. In contrast, imposing unbiasedness limits the influence of the weight function on the resulting estimator. For instance, in the Gaussian shift experiment, imposing mean unbiasedness yields the MLE as the unique non-randomized estimator, so the MLE is weighted risk minimizing among all unbiased estimators, for any weight function. Correspondingly, in our applications, we find that the weight function plays a very limited role, with the risk of the nearly unbiased risk minimizing estimator with all weight on one parameter value only slightly below the risk of the nearly unbiased estimator under a more diffuse weighting function.

Finally, one may simply point to the long tradition of evaluating competing estimators by their bias. For instance, the large literature on the estimation of the AR(1) parameter, as partially reviewed above, focuses heavily on the mean or median bias. Under this "revealed preference" it makes sense to take a systematic approach to the derivation of estimators that perform nearly optimally under this criterion.

The remainder of the paper is organized as follows. Section 2 sets up the generic problem, derives the lower bound on WAR, and discusses the numerical implementation. Section 3 considers the two special cases of mean unbiased estimation under squared loss, and of median unbiased estimation without nuisance parameters. Section 4 extends the framework to invariant estimators. Throughout Sections 2–4, we use the small sample problem of estimating the coefficient in a Gaussian AR(1) as our running example. In Section 5, we consider the two additional problems of median unbiased estimation of the degree of time variation in a local-level model, and the quantile AR(1) forecast problem. Section 6 briefly investigates the performance of the new estimators in misspecified models with non-Gaussian innovations. Section 7 concludes.

## 2. Estimation, bias and risk

### 2.1. Set-up and notation

We observe the random element $X$ in the sample space $\mathcal{X}$. The density of $X$ relative to some $\sigma$-finite measure $\nu$ is $f_\theta$, where $\theta \in \Theta$ is the parameter space. We are interested in estimating $\eta = h(\theta) \in H$ with estimators $\delta : \mathcal{X} \mapsto H$. For scalar estimation problems, $H \subset \mathbb{R}$, but our set-up allows for more general estimands. Estimation errors lead to losses as measured by the function $\ell : H \times \Theta \mapsto [0, \infty)$ so that $\ell(\delta(x), \theta)$ is the loss incurred by the estimate $\delta(x)$ if the true parameter is $\theta$. The risk of the estimator $\delta$ is given by its expected loss, $r(\delta, \theta) = E_\theta[\ell(\delta(X), \theta)] = \int \ell(\delta(x), \theta) f_\theta(x) d\nu(x)$.

Beyond the estimator's risk, we are also concerned about its bias. For some function $c : H \times \Theta \mapsto H$, the bias of $\delta$ is defined as $b(\delta, \theta) = E_\theta[c(\delta(X), \theta)]$. For instance, for the mean bias, $c(\eta, \theta) = \eta - h(\theta)$, so that $b(\delta, \theta) = E_\theta[\delta(X)] - h(\theta)$, and for the median bias of a scalar parameter of interest $\eta$, $c(\eta, \theta) = \mathbf{1}[\eta > h(\theta)] - \frac{1}{2}$, so that $b(\delta, \theta) = P_\theta[\delta(X) > h(\theta)] - \frac{1}{2}$.

We are interested in deriving estimators $\delta$ that minimize risk subject to an unbiasedness constraint. In many problems of interest, a uniformly risk minimizing $\delta$ might not exist, even under the bias constraint. To make further progress, we thus measure the quality of estimators by their weighted average risk $R(\delta, F) = \int r(\delta, \theta) dF(\theta)$ for some given non-negative finite measure $F$ with support in $\Theta$.

In this notation, the weighted risk minimal unbiased estimator $\delta^*$ solves the program

$$\min_\delta R(\delta, F) \quad \text{s.t.} \tag{1}$$

$$b(\delta, \theta) = 0 \; \forall \theta \in \Theta. \tag{2}$$

More generally, one might also be interested in deriving estimators that are only approximately unbiased, that is solutions to (1) subject to

$$-\varepsilon \leq b(\delta, \theta) \leq \varepsilon \; \forall \theta \in \Theta \tag{3}$$

for some $\varepsilon \geq 0$. Allowing the bounds on $b(\delta, \theta)$ to depend on $\theta$ does not yield greater generality, as they can be subsumed in the definition of the function $c$. For instance, a restriction on the relative mean bias to be no more than 5% is achieved by setting $c(\eta, \theta) = (\eta - h(\theta))/h(\theta)$ and $\varepsilon = 0.05$.

An estimator is called *risk unbiased* if $E_{\theta_0}[\ell(\delta(X), \theta_0)] \leq E_{\theta_0}[\ell(\delta(X), \theta)]$ for any $\theta_0, \theta \in \Theta$. As discussed in Chapter 3 of Lehmann and Casella (1998), risk unbiasedness is a potentially attractive property as it ensures that under $\theta_0$, the estimate $\delta(x)$ is at least as close to the true value $\theta_0$ in expectation as measured by $\ell$ as it is to any other value of $\theta$. It is straightforward to show that under squared loss $\ell(\delta(x), \theta) = (\delta(x) - h(\theta))^2$ a risk unbiased estimator is necessarily mean unbiased, and under absolute value loss $\ell(\delta(x), \theta) = |\delta(x) - h(\theta)|$, it is necessarily median unbiased. From this perspective, squared loss and mean unbiased constraints, and absolute value loss and median unbiased constraints, form natural pairs, and we report results for these pairings in our examples. The following development, however, does not depend on any assumptions about the relationship between loss function and bias constraint, and other pairings of loss function and constraints might be more attractive in specific applications.

In many examples, the risk of good estimators is far from constant, as the information in $X$ about $h(\theta)$ varies with $\theta$. This makes risk comparisons and the weighting function $F$ more difficult to interpret. Similarly, also the mean bias of an estimator is naturally gauged relative to its sampling variability. To address this issue, we introduce a normalization function $n(\theta)$ that roughly corresponds to the root mean squared error of a good estimator at $\theta$. The *normalized* risk $r_n(\delta,\theta)$ is then given by $r_n(\delta,\theta) = E_\theta[(\delta(X)-\theta)^2]/n(\theta)^2$ and $r_n(\delta,\theta) = E_\theta[|\delta(X)-h(\theta)|]/n(\theta)$ under squared and absolute value loss, respectively, and the normalized mean bias is $b_n(\delta,\theta) = (E_\theta[\delta(X)] - h(\theta))/n(\theta)$. The weighted average normalized risk with weighting function $F_n$, $\int r_n(\delta,\theta)dF_n(\theta)$, then reduces to $R(\delta,F)$ above with $dF(\theta) = dF_n(\theta)/n(\theta)^2$ and $dF(\theta) = dF_n(\theta)/n(\theta)$ in the squared loss and absolute value loss case, respectively, and $b_n(\delta,\theta) = b(\delta,\theta)$ with $c(\eta,\theta) = (\eta - h(\theta))/n(\theta)$. The median bias, of course, is readily interpretable without any normalization.

*Running example:* Consider a Gaussian autoregressive process of order 1, $Y_t = \theta Y_{t-1} + \varepsilon_t$, $t = 1,\ldots,T$, with $Y_0 = 0$ and $\varepsilon_t \sim i.i.d.\mathcal{N}(0,\sigma^2)$. We assume for now that the variance of the innovations $\varepsilon_t$ is known and equal to unity (we relax this in Section 4). With $X = (Y_1,\ldots,Y_T)$, the density is then given by

$$f_\theta(x) = \exp[-\tfrac{1}{2}\sum_{t=1}^{T}(y_t - \theta y_{t-1})^2]. \tag{4}$$

The aim is to estimate $\theta$ in a way that (nearly) minimizes weighted risk under a mean or median bias constraint. We set the parameter space equal to $\Theta = [-0.95, 1]$. The lower bound of $-0.95$ avoids complications that arise for roots very close to minus one, which have very little empirical relevance. We rule out explosive roots to ensure comparability to the stationary model below. While $\Theta$ is compact, estimators may take arbitrary values in $H = \mathbb{R}$.

The usual OLS estimator for $\theta$ (which is also equal to the MLE) is $\delta_{OLS}(x) = \sum_{t=1}^{T} y_t y_{t-1}/\sum_{t=1}^{T} y_{t-1}^2$. For $\theta$ not too close to one and $T$ large, $\delta_{OLS}(X) \overset{a}{\sim} \mathcal{N}(\theta,(1-\theta^2)/T)$. We thus use the normalization function $n(\theta) = \sqrt{(1-\theta^2)/T + 8\theta^2/T^2}$. The additional term $8\theta^2/T^2$ ensures that $n(\theta) > 0$ also for $\theta = 1$, and with this choice, the normalized mean squared error of good estimators turns out to be roughly equal to unity. ▲

### *2.2. A lower bound on weighted risk of unbiased estimators*

In general, it will be difficult to analytically solve (1) subject to (3). Both $\mathcal{X}$ and $H$ are typically uncountable, so we are faced with an optimization problem in a function space. Moreover, it is usually difficult to obtain analytical closed-form expressions for the integrals defining $r(\delta,\theta)$ and $b(\delta,\theta)$, so one must rely on approximation techniques, such as Monte Carlo simulation. For these reasons, it seems natural to resort to numerical techniques to obtain an approximate solution.

There are potentially many ways of approaching this numerical problem. For instance, one might posit some sieve-type space for $\delta$, and numerically determine the (now finite-dimensional) parameter that provides the relatively best approximate solution. But to make this operational, one must choose some dimension of the sieve space, and it is unclear how much better of a solution one might have been able to find in a different or more highly-dimensional sieve space.

It would therefore be useful to have a lower bound on the weighted risk $R(\delta,F)$ that holds for *all* estimators $\delta$ that satisfy (3). If an approximate solution $\hat{\delta}$ is found that also satisfies (3) and whose weighted risk $R(\hat{\delta},F)$ is close to the bound, then we know that we have found the nearly best solution overall.

To derive such a bound, we relax the constraint (3) by replacing it by a finite number of constraints: Let $G_i$, $i = 1,\ldots,m$ be probability distributions on $\Theta$, and define the weighted average bias $B(\delta,G)$ of the estimator $\delta$ as $B(\delta,G) = \int b(\delta,\theta)dG(\theta)$. Then any estimator that satisfies the uniform unbiasedness constraint (3) clearly also satisfies

$$-\varepsilon \le B(\delta,G_i) \le \varepsilon \text{ for all } i = 1,\ldots,m. \tag{5}$$

A special case of (5) has $G_i$ equal to a point mass at $\theta_i$, so that (5) amounts to imposing (3) at the finite number of parameter values $\theta_1,\ldots,\theta_m$. In some problems, it is computationally more attractive to rely on non-degenerate $G_i$. In that case, (5) only imposes that all $G_i$-weighted averages of the bias are close to zero, so negative and positive biases might cancel in each average.

Now consider the Lagrangian for the problem (1) subject to (5),

$$L(\delta,\lambda) = R(\delta,F) + \sum_{i=1}^{m}\lambda_i^u(B(\delta,G_i)-\varepsilon) + \sum_{i=1}^{m}\lambda_i^l(-B(\delta,G_i)-\varepsilon) \tag{6}$$

where $\lambda = (\lambda_1,\ldots,\lambda_m)$ and $\lambda_i = (\lambda_i^l,\lambda_i^u)$. By writing $R(\delta,F)$ and $B(\delta,G_i)$ in terms of their defining integrals and by assuming that we can change the order of integration, we obtain

$$\begin{aligned} L(\delta,\lambda) = \int \Bigg( \int f_\theta(x)\ell(\delta(x),\theta)dF(\theta) &+ \sum_{i=1}^{m}\lambda_i^u\Big(\int f_\theta(x)c(\delta(x),\theta)dG_i(\theta) - \varepsilon\Big) \\ &+ \sum_{i=1}^{m}\lambda_i^l\Big(-\int f_\theta(x)c(\delta(x),\theta)dG_i(\theta) - \varepsilon\Big)\Bigg) d\nu(x). \end{aligned} \tag{7}$$

Let $\delta_\lambda$ be the estimator such that for a given $\lambda$, $\delta_\lambda(x)$ minimizes the integrand on the right hand side of (7) for each $x$. Since minimizing the integrand at each point is sufficient to minimize the integral, $\delta_\lambda$ necessarily minimizes $L$ over $\delta$.

This argument requires that the change of the order of integration in (7) is justified. By Fubini's Theorem, this is always the case for $R(\delta, F)$ (since $\ell$ is non-negative), and also for $B(\delta, G_i)$ if $c$ is bounded (as is always the case for the median bias, and for the mean bias if $H$ is bounded). Under squared loss and mean bias constraints, and unbounded $H$, it suffices that there exists some estimator with uniformly bounded mean squared error (MSE). A sufficient condition is a compact parameter space $\Theta$.

Standard Duality Theory for optimization implies the following lemma.

**Lemma 1.** *Suppose $\tilde{\delta}$ satisfies (5), and for arbitrary $\lambda \geq 0$ (that is, each element in $\lambda$ is non-negative), $\delta_\lambda$ minimizes $L(\delta, \lambda)$ over $\delta$. Then $R(\tilde{\delta}, F) \geq L(\delta_\lambda, \lambda)$.*

**Proof.** For any $\delta$, $L(\delta, \lambda) \geq L(\delta_\lambda, \lambda)$ by definition of $\delta_\lambda$, so in particular, $L(\tilde{\delta}, \lambda) \geq L(\delta_\lambda, \lambda)$. Furthermore, $R(\tilde{\delta}, F) \geq L(\tilde{\delta}, \lambda)$ since $\lambda \geq 0$ and $\tilde{\delta}$ satisfies (5). Combining these inequalities yields the result. ■

Note that the bound on $R(\tilde{\delta}, F)$ in Lemma 1 does not require an exact solution to the discretized problem (1) subject to (5). Any $\lambda \geq 0$ implies a valid bound $L(\delta_\lambda, \lambda)$, although some choices for $\lambda$ yield better (i.e., larger) bounds than others. Since (5) is implied by (3), these bounds hold a fortiori for any estimator satisfying the uniform unbiasedness constraint (3).

If $\lambda^* \geq 0$ is such that $\delta_{\lambda^*}$ satisfies (5) and the complementarity slackness conditions $\lambda_i^{u*}(B(\delta, G_i) - \varepsilon) = 0$ and $\lambda_i^{l*}(-B(\delta, G_i) - \varepsilon) = 0$ hold, then (6) implies $L(\delta_{\lambda^*}, \lambda^*) = R(\delta_{\lambda^*}, F)$, so that by an application of Lemma 1, $L(\delta_{\lambda^*}, \lambda^*)$ is the best lower bound.

### 2.3. Numerical approach

Solving the program (1) subject to (5) thus yields the largest, and thus most informative bound on weighted risk $R(\delta, F)$. In addition, solutions $\delta_{\lambda^*}$ to this program also plausibly satisfy a slightly more relaxed version of original non-discretized constraint (3). The reasoning is as follows. Suppose the bias function $b(\delta_{\lambda^*}, \theta)$ of $\delta_{\lambda^*}$ is smooth in $\theta$. If the $G_i$ are equal to points masses on $\theta_i$ that form a fine discretization of $\Theta$, then low bias in the discretized problem $B(\delta_{\lambda^*}, G_i) = b(\delta_{\lambda^*}, \theta_i)$ for $i = 1, \ldots, m$ necessarily implies that $|b(\delta_{\lambda^*}, \theta)|$ is small uniformly in $\theta \in \Theta$. The same applies for $G_i$ that are non-degenerate but with supports equal to (potentially overlapping) small neighborhoods of $\theta_i$, as the degree of cancellation in the averages $B(\delta_{\lambda^*}, G_i) = \int b(\delta_{\lambda^*}, \theta) dG_i(\theta)$ cannot be large if $b(\delta_{\lambda^*}, \theta)$ is smooth. Of course, $b(\delta_{\lambda^*}, \theta)$ might not be smooth, but this can be checked numerically.

These considerations suggest the following strategy to obtain a *nearly weighted risk minimizing nearly unbiased estimator*, that is, for given $\varepsilon_B > 0$ and $\varepsilon_R > 0$, an estimator $\hat{\delta}$ that (i) satisfies (3) with $\varepsilon = \varepsilon_B$; (ii) has weighted risk $R(\hat{\delta}, F) \leq (1 + \varepsilon_R) R(\delta, F)$ for any estimator $\delta$ satisfying (3) with $\varepsilon = \varepsilon_B$.

1. Discretize $\Theta$ by point masses or other distributions $G_i$, $i = 1, \ldots, m$.
2. Obtain approximately optimal Lagrange multipliers $\hat{\lambda}^\dagger$ for the problem (1) subject to (5) for $\varepsilon = \varepsilon_B$, and associated value $\underline{R} = L(\delta_{\hat{\lambda}^\dagger}, \hat{\lambda}^\dagger)$.
3. Obtain an approximate solution $(\hat{e}^*, \hat{\delta}^*)$ to the problem

$$\min_{e_B \geq 0, \delta} e_B \text{ s.t. } R(\delta, F) \leq (1 + \varepsilon_R) \underline{R} \tag{8}$$

$$\text{and } -e_B \leq B(\delta, G_i) \leq e_B, i = 1, \ldots, m \tag{9}$$

   and check whether $\hat{\delta}^*$ satisfies the uniform unbiasedness constraint (3). If it does not, go back to Step 1 and use a larger $m$ or more concentrated distributions $G_i$. If it does, $\hat{\delta} = \hat{\delta}^*$ has the desired properties by an application of Lemma 1.

Importantly, neither $\delta_{\hat{\lambda}^\dagger}$ nor $\hat{\delta}^*$ have to be exact solutions to their respective programs to be able to conclude that $\hat{\delta}^*$ is indeed a nearly weighted risk minimizing nearly unbiased estimator as defined above, that is, it satisfies the uniform near unbiasedness property (3) (and not only the discretized unbiasedness property (5)), and its WAR is no more than a multiple $(1 + \varepsilon_R)$ larger than the WAR of any such estimator.

The solution $\hat{\delta}^*$ to the problem in Step 3 has the same form as $\delta_\lambda$, that is $\hat{\delta}^*$ minimizes a weighted average of the integrand in (7), but $\lambda$ now is the ratio of the Lagrange multipliers corresponding to the constraints (9) and the Lagrange multiplier corresponding to the constraint (8). These constraints are always feasible, since $\delta = \delta_{\hat{\lambda}^\dagger}$ satisfies (9) with $e_B = \varepsilon_B$ (at least if $\hat{\lambda}^\dagger$ is the exact solution), and $R(\delta_{\hat{\lambda}^\dagger}, F) = \underline{R} < (1 + \varepsilon_R) \underline{R}$. The additional slack provided by $\varepsilon_R$ is used to tighten the constraints on $B(\hat{\delta}^*, G_i)$ to potentially obtain a $\hat{\delta}^*$ satisfying (3). A finer discretization implies additional constraints in (5) and thus (weakly) increases the value of $\underline{R}$. At the same time, a finer discretization also adds additional constraints on the bias function of $\hat{\delta}^*$, making it more plausible that it satisfies the uniform constraint (3).

We suggest using simple fixed point iterations to obtain approximate solutions in Steps 2 and 3, similar to Elliott, Müller and Watson's (2015) approach to numerically approximate a least favorable distribution. See Appendix B for details. Once the Lagrange multipliers underlying $\hat{\delta}^*$ are determined, $\hat{\delta}^*(x)$ for given data $x$ is simply the minimizer of the integrand on the right hand side of (7).

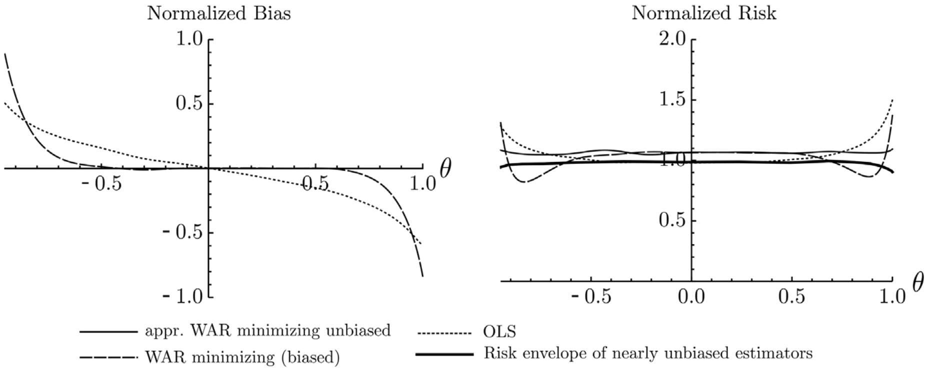


**Fig. 1.** Normalized mean bias and MSE in AR(1) with known mean and variance.

## 3. Two special cases

### 3.1. Mean unbiased estimation under squared loss

Consider the special case of squared loss $\ell(\eta, \theta) = (\eta - h(\theta))^2$ and normalized mean bias constraint $c(\eta, \theta) = (\eta - h(\theta))/n(\theta)$ (we focus on a scalar estimand for expositional ease, but the following discussion straightforwardly extends to vector valued $\eta$). Then the integrand in (7) becomes a quadratic function of $\delta(x)$, and the minimizing value $\delta_\lambda(x)$ is

$$\delta_\lambda(x) = \frac{\int f_\theta(x)h(\theta)dF(\theta) - \sum_{i=1}^m \tilde{\lambda}_i \int \frac{f_\theta(x)}{n(\theta)} dG_i(\theta)}{\int f_\theta(x)dF(\theta)}, \tag{10}$$

a linear function of $\tilde{\lambda}_i = \frac{1}{2}(\lambda_i^u - \lambda_i^l)$. Plugging this back into the objective and constraints yields

$$R(\delta_\lambda, F) = \tilde{\lambda}'\Omega\tilde{\lambda} + \omega_R, \Omega_{ij} = \int \frac{\int \frac{f_\theta(x)}{n(\theta)} dG_i(\theta) \int \frac{f_\theta(x)}{n(\theta)} dG_j(\theta)}{\int f_\theta(x)dF(\theta)} d\nu(x),$$

$$B(\delta_\lambda, G_i) = \omega_i - \Omega_i'\tilde{\lambda}, \omega_i = \int \left( \frac{\int \frac{f_\theta(x)}{n(\theta)} dG_i(\theta) \int f_\theta(x)h(\theta)dF(\theta)}{\int f_\theta(x)dF(\theta)} - \int \frac{f_\theta(x)}{n(\theta)} h(\theta)dG_i(\theta) \right) d\nu(x)$$

where

$$\omega_R = \int \left( \int h(\theta)^2 f_\theta(x)dF(\theta) - \frac{[\int f_\theta(x)h(\theta)dF(\theta)]^2}{\int f_\theta(x)dF(\theta)} \right) d\nu(x),$$

and $\Omega_i$ is the $i$th column of the $m \times m$ matrix $\Omega$. Note that $\Omega$ is symmetric and positive semi-definite. In fact, $\Omega$ is positive definite as long as $\int \frac{f_\theta(x)}{n(\theta)} dG_i(\theta)$ cannot be written as a linear combination of $\int \frac{f_\theta(x)}{n(\theta)} dG_j(\theta), j \neq i$ almost surely. Thus, minimizing $R(\delta, F)$ subject to (5) becomes a (semi)-definite quadratic programming problem, which is readily solved by well-known algorithms. In the special case of $\varepsilon = 0$, the solution is $\tilde{\lambda}^* = \Omega^{-1}\omega$, where $\omega = (\omega_1, \ldots, \omega_m)'$. Step 3 of the algorithm generally becomes a quadratically constrained quadratic program, but since $e_B$ is scalar, it can easily be solved by conditioning on $e_B$, with a line search as outer-loop.

Either way, it is computationally trivial to implement the strategy described in Section 2.3 under mean square risk and a mean bias constraint, even for a very fine discretization $m$.

*Running example:* We set $F_n$ uniform on $\Theta = [-0.95, 1]$ and $G_i$ equal to point masses at the 103 points $\{-0.95, 1.0\} \cup \{\tanh(-1.83 + 5.03i/100)\}_{i=0}^{100}$. This choice of grid is finer for values of $|\theta|$ close to one, where standard estimators exhibit larger normalized bias. We set $\varepsilon_B = 0.005$ and $\varepsilon_R = 0.01$, and focus in the main text on results for $T = 50$. The replication files contain tables for $\{\tilde{\lambda}_i\}_{i=1}^{103}$ that determine nearly unbiased WAR minimizing estimators with these choices for all $T \in \mathbb{T} = \{5,$ 6, 7, 8, 9, 10, 12, 14, 16, 18, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 120, 140, 160, 180, 200, 220, 240, 260, 280, 300, 320, 340, 360, 380, 400}. The remaining bias in $\hat{\delta}^*$ is very small: with the normalized mean squared error of $\hat{\delta}^*$ close to one, $\varepsilon_B = 0.005$ implies that about 40,000 Monte Carlo draws are necessary for the largest bias of $\hat{\delta}^*$ to be of the same magnitude as the Monte Carlo standard error of its estimate.

Fig. 1 plots the normalized bias and risk of $\hat{\delta}^*$ for $T = 50$. As a point of reference, we also report the normalized bias and risk of the OLS estimator $\delta_{OLS}$, and of the WAR minimizing estimator without any constraints. By construction, the unconstrained WAR minimizing estimator (which equals the posterior mean of $\theta$ under a prior proportional to $F$) has the smallest possible average normalized risk on $\theta \in \Theta$. As can be seen from Fig. 1, however, this comes at the cost of a substantive mean bias.

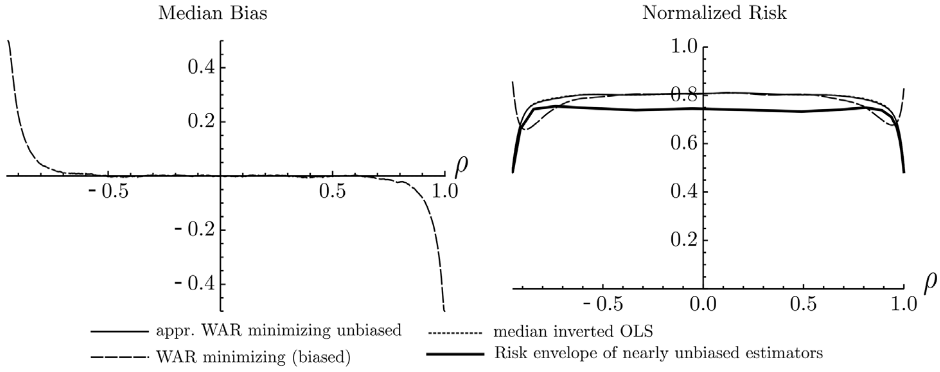


**Fig. 2.** Median bias and normalized MAD in AR(1) with known mean and variance.

The thick line plots a lower bound on the risk envelope for nearly unbiased estimators: For each $\theta$, we set the weighting function equal to a point mass at that $\theta$, and then report the lower bound on risk at $\theta$ among all estimators whose $G_i$ averaged normalized bias is no larger than $\varepsilon_B = 0.005$ in absolute value for all $i$ (that is, for each $\theta$, we perform Step 2 of the algorithm in Section 2.3, with $F$ equal to a point mass at $\theta$). The risk of $\hat{\delta}^*$ is seen to be approximately 10% larger than this lower bound on the envelope. This implies that our choice of a uniform normalized weighting function $F_n$ on $\theta \in \Theta$ has a fairly limited influence: no other weighting function can lead to a nearly unbiased estimator with substantially less risk, for any $\theta$. In fact, to the extent that 10% is considered small, one could call $\hat{\delta}^*$ approximately uniformly minimum variance unbiased. ▲

### 3.2. Median unbiased estimation without nuisance parameters

Suppose $h(\theta) = \theta$, that is there are no nuisance parameters, and $\Theta \subset \mathbb{R}$. Let $\delta_B$ be an estimator taking on values in $\mathbb{R}$, and let $m_{\delta_B}(\theta)$ be its median function, $P_\theta(\delta_B(X) \leq m_{\delta_B}(\theta)) = 1/2$ for all $\theta \in \Theta$. If $m_{\delta_B}(\theta)$ is one-to-one on $\Theta$ with inverse $m_{\delta_B}^{-1} : \mathbb{R} \mapsto \Theta$, then the estimator $\delta_U(X) = m_{\delta_B}^{-1}(\delta_B(X))$ is exactly median unbiased by construction (cf. Lehmann (1986), p. 23). In general, different estimators $\delta_B$ yield different median unbiased estimators $m_{\delta_B}^{-1}(\delta_B(X))$, which raises the question how $\delta_B$ should be chosen for $\delta_U$ to have low risk.

*Running example:* Stock (1991) and Andrews (1993) construct median unbiased estimators for the largest autoregressive root based on the OLS estimator. Their results allow for the possibility that there exists another median unbiased estimator with much smaller risk. ▲

A good candidate for $\delta_B$ is a nearly weighted risk minimizing *nearly* median unbiased estimator $\hat{\delta}$. A small median bias of $\hat{\delta}$ typically also yields monotonicity of $m_{\hat{\delta}}$, so if $P_\theta(\delta_{\hat{\lambda}^\dagger} = \theta) \leq 1/2$ at the boundary points of $\Theta$ (if there are any), then $m_{\hat{\delta}}$ has an inverse $m_{\hat{\delta}}^{-1}$, and $\hat{\delta}_U(x) = m_{\hat{\delta}}^{-1}(\hat{\delta}(x))$ is exactly median unbiased. Furthermore, if $\hat{\delta}$ is already nearly median unbiased, then $m_{\hat{\delta}}^{-1} : \mathbb{R} \mapsto \Theta$ is close to the identity transformation, so that $R(\hat{\delta}_U, F)$ is close to the nearly minimal risk $R(\hat{\delta}, F)$.

This suggests the following modified strategy to numerically identify an *exactly* median unbiased estimator $\hat{\delta}_U$ whose weighted risk $R(\hat{\delta}_U, F)$ is within $(1 + \varepsilon_R)$ of the risk of any other exactly median unbiased estimator.

1. Discretize $\Theta$ by the distributions $G_i$, $i = 1, \ldots, m$.
2. Obtain approximately optimal Lagrange multipliers $\hat{\lambda}^\dagger$ for the problem (1) subject to (5) for $\varepsilon = 0$, and associated value $\underline{R} = L(\delta_{\hat{\lambda}^\dagger}, \hat{\lambda}^\dagger)$. Choose $G_i$ and $\hat{\lambda}^\dagger$ to ensure that $P_\theta(\delta_{\hat{\lambda}^\dagger} = \theta) \leq 1/2$ for any boundary points of $\Theta$. If $m_{\delta_{\hat{\lambda}^\dagger}}$ still does not have an inverse, go back to Step 1 and use a finer discretization.
3. Compute $R(\hat{\delta}_U^\dagger, F)$ for the exactly median unbiased estimator $\hat{\delta}_U^\dagger(X) = m_{\delta_{\hat{\lambda}^\dagger}}^{-1}(\delta_{\hat{\lambda}^\dagger}(X))$. If $R(\hat{\delta}_U^\dagger, F) > (1 + \varepsilon_R)\underline{R}$, go back to Step 1 and use a finer discretization. Otherwise, $\hat{\delta}_U = \hat{\delta}_U^\dagger$ has the desired properties.

*Running example:* As discussed in Section 2.1, we combine the median unbiased constraint on $\theta \in \Theta = [-0.95, 1]$ with absolute value loss. We make largely similar choices as in the derivation of nearly mean unbiased estimators: the normalized weighting function $F_n$ is uniform on $\Theta$, with risk now normalized as $r_n(\delta, \theta) = E_\theta[|\delta(X) - \theta|]/n(\theta)$ with $n(\theta) = \sqrt{(1 - \theta^2)/T + 8\theta^2/T^2}$, as before. Under a median bias constraint and absolute value loss, using point masses for $G_i$ leads to a discontinuous integrand in the Lagrangian (7) for given $x$ . In order to ensure a smooth estimator $\delta_\lambda$, it makes sense to instead choose the distributions $G_i$ in a way that any mixture of the $G_i$'s has a continuous density. To this end we set the Lebesgue density of $G_i$, $i = 1, \ldots, m$ proportional to the $i$th third-order basis spline, with the knots on the 53 points $\{-0.95, 1.0\} \cup \{\tanh(-1.83 + 5.03i/50)\}_{i=0}^{50}$ (with "not-a-knot" end conditions).

We again focus on $T = 50$, with results for other $T \in \mathbb{T}$ relegated to the replication files. Fig. 2 reports the median bias and normalized mean absolute deviation (MAD) of $\hat{\delta}_U^\dagger$, the OLS-based median unbiased estimator $\delta_{U,OLS}(x) = m_{\delta_{OLS}}^{-1}(\delta_{OLS}(x))$, and the estimator that minimizes weighted risk relative to $F$ without any median unbiased constraints. The risk of $\hat{\delta}_U^\dagger$ is indistinguishable from the risk of $\delta_{U,OLS}$. Consequently, this analysis reveals $\delta_{U,OLS}$ to be (also) nearly weighted risk minimal unbiased in this problem. The thick line again plots the envelope for the normalized risk among all exactly median unbiased estimators, which is seen to be roughly 10% below the risk of $\hat{\delta}_U^\dagger$ and $\delta_{U,OLS}$. ▲

## 4. Invariance

In this section, we consider estimation problems that have some natural invariance (or equivariance) structure, so that it makes sense to impose the corresponding invariance also on estimators. We show that the problem of identifying a (nearly) minimal risk unbiased *invariant* estimator becomes equivalent to the problem of identifying an unrestricted (nearly) minimal risk unbiased estimator in a related problem with a sample and parameter space generated by maximal invariants. Imposing invariance then reduces the dimension of the effective parameter space, which facilitates numerical solutions.

Consider a group of transformations on the sample space $g : \mathcal{X} \times A \mapsto \mathcal{X}$, where $a \in A$ denotes a group action. We write $a_2 \circ a_1 \in A$ for the composite action $g(g(x, a_1), a_2) = g(x, a_2 \circ a_1)$ for all $a_1, a_2 \in A$, and we denote the inverse of action $a$ by $a^-$, that is $g(x, a^- \circ a) = x$ for all $a \in A$ and $x \in \mathcal{X}$.

Now suppose the problem is invariant in the sense that there exists a corresponding group $\bar{g} : \Theta \times A \mapsto \Theta$ on the parameter space, and the distribution of $g(X, a)$ under $X \sim P_\theta$ is $P_{\bar{g}(\theta,a)}$, for all $\theta$ and $a \in A$ (cf. Definition 2.1 of Chapter 3 in Lehmann and Casella (1998)). Let $M(x)$ and $\bar{M}(\theta)$ for $M : \mathcal{X} \mapsto \mathcal{X}$ and $\bar{M} : \Theta \mapsto \Theta$ be maximal invariants of these two groups. Assume that $M$ and $\bar{M}$ select a specific point on the orbit induced by $g$ and $\bar{g}$, that is $M(x) = g(x, O(x)^-)$ for all $x \in \mathcal{X}$ and $\bar{M}(\theta) = g(\theta, \bar{O}(\theta)^-)$ for all $\theta \in \Theta$ for some functions $O : \mathcal{X} \mapsto A$ and $\bar{O} : \Theta \mapsto A$ (as discussed on page 216–217 in Lehmann and Romano (2005)). Then by definition of a maximal invariant, $M(M(x)) = M(x)$, $\bar{M}(\bar{M}(\theta)) = \bar{M}(\theta)$, and we have the decomposition

$$x = g(M(x), O(x)) \tag{11}$$

$$\theta = \bar{g}(\bar{M}(\theta), \bar{O}(\theta)). \tag{12}$$

We further assume that group actions $a$ are distinct in the sense that $g(M(x), a_1) = g(M(x), a_2)$ for some $x \in \mathcal{X}$ implies $a_1 = a_2$.

By Theorem 6.3.2 of Lehmann and Romano (2005), the distribution of $M(X)$ only depends on $\bar{M}(\theta)$. The following lemma provides a slight generalization, which we require below. The proof is in Appendix A.

**Lemma 2.** *The distribution of $(M(X), \bar{g}(\theta, O(X)^-))$ under $\theta$ is the same as the distribution of $(M(X), \bar{g}(\bar{M}(\theta), O(X)^-))$ under $\bar{M}(\theta)$.*

Suppose further that the estimand $h(\theta)$ is compatible with the invariance structure in the sense that $h(\theta_1) = h(\theta_2)$ implies $h(\bar{g}(\theta_1, a)) = h(\bar{g}(\theta_2, a))$ for all $\theta_1, \theta_2 \in \Theta$ and $a \in A$. As discussed in Chapter 3 of Lehmann and Casella (1998), this induces a group $\hat{g} : H \times A \mapsto H$ satisfying $h(\bar{g}(\theta, a)) = \hat{g}(h(\theta), a)$ for all $\theta \in \Theta$ and $a \in A$, and it is natural for the loss function and constraints to correspondingly satisfy

$$\ell(\eta, \theta) = \ell(\hat{g}(\eta, a), \bar{g}(\theta, a)) \text{ for all } \eta \in H, \theta \in \Theta \text{ and } a \in A \tag{13}$$

$$c(\eta, \theta) = c(\hat{g}(\eta, a), \bar{g}(\theta, a)) \text{ for all } \eta \in H, \theta \in \Theta \text{ and } a \in A. \tag{14}$$

Any loss function $\ell(\eta, \theta)$ that depends on $\theta$ only through the parameter of interest, $\ell(\eta, \theta) = \ell^i(\eta, h(\theta))$ for some function $\ell^i : H \times \Theta \mapsto [0, \infty)$, such as quadratic loss or absolute value loss, automatically satisfies (13). Similarly, constraints of the form $c(\eta, \theta) = c^i(\eta, h(\theta))$, such as those arising from mean or median unbiased constraints, satisfy (14).

With these notions of invariance in place, it makes sense to impose that estimators conform to this structure and satisfy

$$\delta(g(x, a)) = \hat{g}(\delta(x), a) \text{ for all } x \in X \text{ and } a \in A. \tag{15}$$

Eqs. (11) and (15) imply that any invariant estimator satisfies

$$\delta(x) = \hat{g}(\delta(M(x)), O(x)) \text{ for all } x \in X. \tag{16}$$

It is useful to think about the right-hand side of (16) as inducing the invariance property: *Any* function $\delta_a : M(\mathcal{X}) \to H$ defines an invariant estimator $\delta(x)$ via $\delta(x) = \hat{g}(\delta_a(M(x)), O(x))$. Given that any invariant estimator satisfies (16), the set of all invariant estimators can therefore be generated by considering all (unconstrained) $\delta_a$, and setting $\delta(x) = \hat{g}(\delta_a(M(x)), O(x))$.

*Running example:* Consider estimation of the AR(1) coefficient in a stationary model with unknown mean and variance: $Y_t = \mu + u_t$, $u_t = \rho u_{t-1} + \varepsilon_t$, $\varepsilon_t \sim i.i.d. \mathcal{N}(0, \sigma^2)$ and $u_0 \sim \mathcal{N}(0, \sigma^2/(1-\rho^2))$. With $X = (Y_1, \ldots, Y_T)$, the distribution of $X$ is indexed by $\theta = (\rho, \mu, \sigma)$, and the parameter of interest is $\rho = h(\theta)$. The problem is invariant to transformations $g(x, a) = a_\sigma(x + a_\mu)$ with $a = (a_\mu, a_\sigma) \in A = \mathbb{R} \times (0, \infty)$, and corresponding group $\bar{g}((\rho, \mu, \sigma), a) = (\rho, a_\sigma(\mu + a_\mu), a_\sigma \sigma)$. One choice for maximal invariants are $M(x) = g(x, O(x)^-)$ where $O(x)^- = \left(-y_1, 1/s_y\right)$ and $s_y^2 = \sum_{t=1}^T (y_t - y_1)^2$, and $\bar{M}((\rho, \mu, \sigma)) = \bar{g}((\rho, \mu, \sigma), (-\mu, \sigma^{-1})) = (\rho, 0, 1)$. Lemma 2 asserts that $M(X) = (Y_1 - Y_1, \ldots, Y_T - Y_1)/s_y$ and $\bar{g}(\theta, O(X)^-) = (\rho, (\mu - Y_1)/s_y, \sigma/s_y)$ have a joint distribution that only depends on $\theta$ through $\rho$. The induced group $\hat{g}$ is given by $\hat{g}(\rho, a) = \rho$ for all $a \in A$. Under (13), the loss must not depend on the location and scale parameters $\mu$ and

$\sigma$. Invariant estimators $\delta$ are numerically invariant to scale and translation shifts of $X$, and can all be written in the form $\delta(x) = \hat{g}(\delta_a(M(x)), O(x)) = \delta_a((Y_1 - Y_1, \ldots, Y_T - Y_1)/s_y)$ for some function $\delta_a$. ▲

Now under these assumptions, we can write the risk of any invariant estimator as

$$\begin{aligned} r(\delta, \theta) &= E_\theta[\ell(\delta(X), \theta)] \\ &= E_\theta[\ell\left(\delta(M(X)), \bar{g}(\theta, O(X)^-)\right)] \quad \text{(by (11) and (13) with } a = O(X)^-) \\ &= E_{\bar{M}(\theta)}[\ell\left(\delta(M(X)), \bar{g}(\bar{M}(\theta), O(X)^-)\right)] \quad \text{(by Lemma 2)} \\ &= E_{\bar{M}(\theta)}[E_{\bar{M}(\theta)}[\ell\left(\delta(M(X)), \bar{g}(\bar{M}(\theta), O(X)^-)\right) |M(X)]] \end{aligned}$$

and similarly

$$\begin{aligned} b(\delta, \theta) &= E_\theta[c(\delta(X), \theta)] \\ &= E_{\bar{M}(\theta)}[E_{\bar{M}(\theta)}[c\left(\delta(M(X)), \bar{g}(\bar{M}(\theta), O(X)^-)\right) |M(X)]]. \end{aligned}$$

Now set $\theta^* = \bar{M}(\theta) \in \Theta^* = \bar{M}(\Theta)$, $h(\theta^*) = \eta^* \in H^* = h(\Theta^*)$, $x^* = M(x) \in \mathcal{X}^* = M(\mathcal{X})$ and

$$\ell^*(\delta(x^*), \theta^*) = E_{\theta^*}[\ell\left(\delta(X^*), \bar{g}(\theta^*, O(X)^-)\right) |X^* = x^*] \tag{17}$$

$$c^*(\delta(x^*), \theta^*) = E_{\theta^*}[c\left(\delta(X^*), \bar{g}(\theta^*, O(X)^-)\right) |X^* = x^*]. \tag{18}$$

Then the starred problem has exactly the same structure as the problem considered in Sections 2 and 3, and the same solution techniques can be applied to identify a nearly weighted average risk minimal estimator $\hat{\delta}^* : \mathcal{X}^* \mapsto H^*$ (with the weighting function a nonnegative measure on $\Theta^*$). This solution is then extended to the domain of the original sample space $\mathcal{X}$ via $\hat{\delta}(x) = \hat{g}(\hat{\delta}^*(M(x)), O(x))$ from (16), and the near optimality of $\hat{\delta}^*$ implies the corresponding near optimality of $\hat{\delta}$ in the class of all invariant estimators.

*Running example:* Since $\rho = h(\theta)$, and $\bar{g}(\theta, a)$ does not affect $h(\theta)$, $\ell(\delta(X^*), \bar{g}(\theta^*, O(X)^-)) = \ell(\delta(X^*), \theta^*)$ and $c(\delta(X^*), \bar{g}(\theta^*, O(X)^-)) = c(\delta(X^*), \theta^*)$. Thus the starred problem amounts to estimating $\rho$ from the observation $X^* = M(X) = (Y_1 - Y_1, \ldots, Y_T - Y_1)/s_y$ (whose distribution does not depend on $(\mu, \sigma)$). Let $\Sigma(\rho)$ be the $T \times T$ covariance matrix of a stationary AR(1) with coefficient $\rho$ and unit innovation variance, and let $e$ be a $T \times 1$ vector of ones. Then by King (1980) and Kariya (1980),

$$f_{\theta^*}(x^*) = C\frac{\left(x^{*\prime}[\Sigma(\rho)^{-1} - \Sigma(\rho)^{-1}e(e'\Sigma(\rho)^{-1}e)^{-1}e'\Sigma(\rho)^{-1}]x^*\right)^{-(T-1)/2}}{\sqrt{\det(e'\Sigma(\rho)^{-1}e)\det\Sigma(\rho)}} \tag{19}$$

$$= C\sqrt{\frac{1+\rho}{T(1-\rho)+2\rho}}\left(\hat{x}^{*\prime}\Sigma(\rho)^{-1}\hat{x}^*\right)^{-(T-1)/2} \tag{20}$$

where $C$ is a constant that does not depend on $\rho$ or $x^*$, and $\hat{x}^* = (\hat{x}_1^*, \ldots, \hat{x}_T^*)'$ are the GLS residuals of a regression of $x^*$ on $e$ with $\Sigma(\rho)^{-1}$ as the GLS weighting matrix (see the Appendix for an explicit expression). Note that $f_{\theta^*}(x^*)$ is well defined even at $\rho = 1$ (cf. Elliott (1998)).

The algorithms discussed in Section 3 can now be applied to determine nearly WAR minimizing invariant unbiased estimators in the problem of observing $X^*$ with density (19). We set $\Theta = [-0.95, 1]$, $n(\theta^*) = \sqrt{(1-\rho^2)/T + 8(\rho+0.4)^2/T^2}$, and make the same choices for $\varepsilon_B, \varepsilon_R, F_n$ and $G_i$ as in the problem with known mean and variance. We compute nearly WAR minimizing unbiased estimators for all $T \in \mathbb{T}$; see the replication files.

Figs. 3 and 4 show the normalized bias and risk of the resulting nearly weighted risk minimizing unbiased invariant estimators for $T = 50$. We also plot the performance of the analogous set of comparisons as in Figs. 1 and 2. The mean and median bias of the OLS estimator is now even larger, and the (nearly) unbiased estimators have substantially lower normalized risk. The nearly WAR minimizing mean unbiased estimator $\delta$ still has risk only about 10% above the (lower bound) on the envelope. In contrast to the case considered in Fig. 2, the WAR minimizing median unbiased estimator now has perceptibly lower risk than the OLS based median unbiased estimator for $|\rho|$ large, but the gains are still fairly moderate.

Fig. 5 compares the performance of the nearly mean unbiased estimator with some previously suggested alternative estimators: The analytically bias corrected estimator by Orcutt and Winokur (1969) $(T\delta_{OLS}(x) + 1)/(T - 3)$, the weighted symmetric estimator analyzed by Pantula et al. (1994) and Park and Fuller (1995), and the MLE $\delta_{MLE}(x^*) = \arg\max_{\rho\in[-0.95,1]} f_{\theta^*}(x^*)$ based on the maximal invariant likelihood (19) (called "restricted" MLE by Cheang and Reinsel (2000)). In contrast to $\hat{\delta}^*$, all of these previously suggested estimators have substantial biases for some values of $\rho$. ▲

## 5. Further applications

### *5.1. Degree of parameter time variation*

Consider the canonical local-level model in the sense of Harvey (1989),

$$y_t = \mu + \phi\sum_{s=1}^{t}\varepsilon_s + u_t, (\varepsilon_t, u_t) \sim i.i.d.\mathcal{N}(0, \sigma^2 I_2) \tag{21}$$

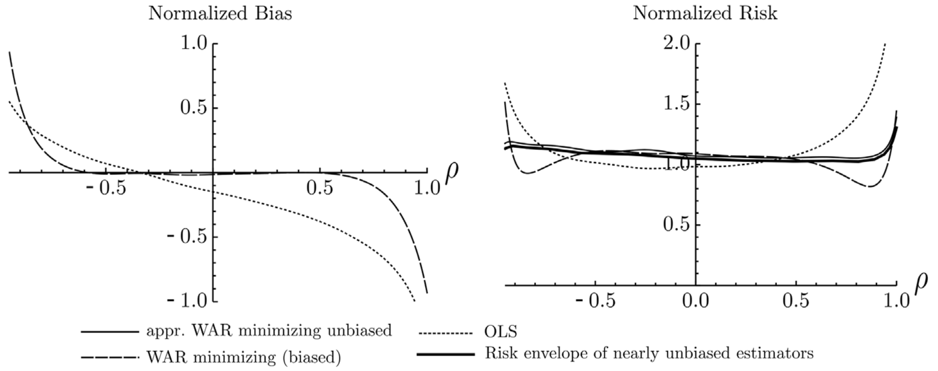


**Fig. 3.** Normalized mean bias and MSE in AR(1) with unknown mean and variance.

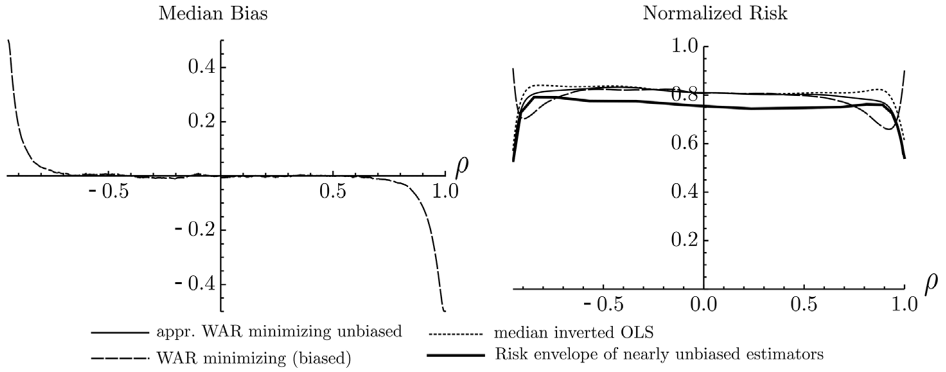


**Fig. 4.** Median bias and normalized MAD in AR(1) with unknown mean and variance.

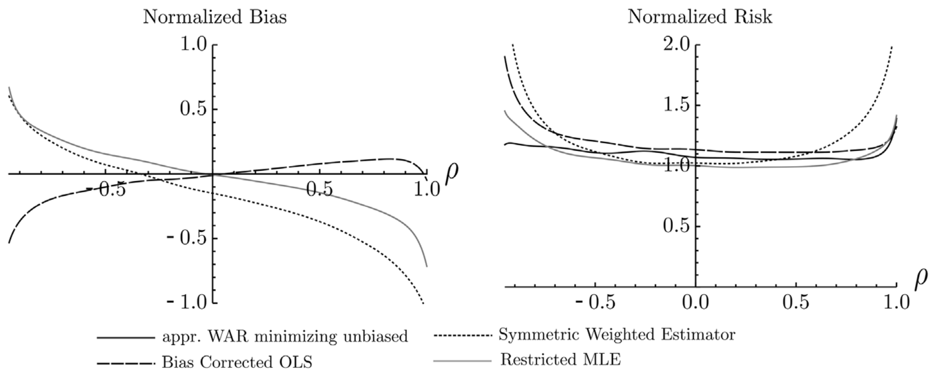


**Fig. 5.** Normalized mean bias and MSE in AR(1) with unknown mean and variance.

with only $y_t$ observed, $t = 1, \ldots, T$ and parameters $(\phi, \mu, \sigma^2)$. The parameter of interest is $\phi \geq 0$, the degree of time variation of the "local level" $\mu + \phi \sum_{s=1}^{t} \varepsilon_s$. As discussed by Stock (1994), (21) is intimately linked to the MA(1) model $\Delta y_t = \phi\varepsilon_t + \Delta u_t = \nu_t - \eta\nu_{t-1}$, where $\eta = \frac{1}{2}(2 + \phi^2 - \phi\sqrt{4 + \phi^2})$ and $\nu_t \sim i.i.d.\mathcal{N}(0, \sigma^2/\eta)$. Since the mapping from $\phi$ to $\eta$ is one-to-one, and $\{\Delta y_t/\sqrt{\sum_{s=2}^{T}(\Delta y_s)^2}\}_{t=2}^{T}$ forms a maximal invariant to the group of transformations $\{y_t\}_{t=1}^{T} \rightarrow \{a_\sigma(y_t + a_\mu)\}_{t=1}^{T}$ for $(a_\mu, a_\sigma) \in \mathbb{R} \times (0, \infty)$, the problem is recognized as equivalent to scale invariant inference about the MA(1) coefficient $0 \leq \eta \leq 1$ in a stationary zero-mean Gaussian MA(1) model.

It has long been recognized that the maximum likelihood estimator of the MA(1) coefficient exhibits non-Gaussian behavior even asymptotically under non-invertibility $\eta = 1$ (corresponding to $\phi = 0$ in (21)); see Stock (1994) for a historical account and references. In particular, the MLE $\hat{\eta}$ suffers from the so-called pile-up problem $P(\hat{\eta} = 1|\eta = 1) > 0$, and Sargan and Bhargava (1983) derive the limiting probability to be 0.657.

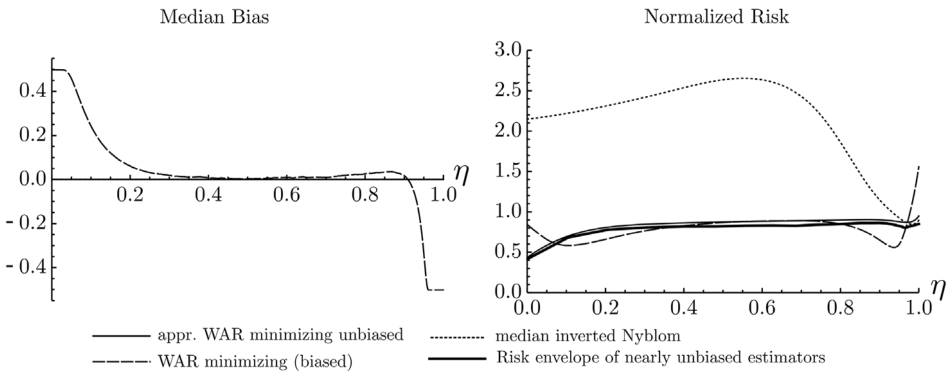


**Fig. 6.** Median bias and normalized MAD in local-level model.

With $P(\hat{\eta} = 1) > 1/2$ under $\eta = 1$, at least for large $T$, it is not possible to base a median unbiased estimator of $\eta$ on $\hat{\eta}$, as the median function $m_{\hat{\eta}}$ is not one-to-one for values of $\eta$ close to one. Stock and Watson (1998) derive an (asymptotically) exactly median unbiased estimator on the Nyblom (1989) statistic $\sum_{t=1}^{T}(\sum_{s=1}^{t}(y_s - \bar{y})^2)/\sum_{t=1}^{T}(y_t - \bar{y})^2$, which is also invariant to the transformations above.

We now determine an alternative median unbiased estimator for $\eta$ that comes close to minimizing weighted average risk under absolute value loss. With $X = (Y_1, \ldots, Y_T)$, we can choose the maximal invariants $X^* = M(X) = (Y_1 - Y_1, \ldots, Y_T - Y_1)/s_y$ and $\bar{M}((\eta, \mu, \sigma)) = (\eta, 0, 1)$, as in the example of the previous section. The density $f_{\theta^*}$ of $X^*$ is then of the same form as (19), with $\Sigma(\rho)$ replaced by the covariance matrix of $\{\phi \sum_{s=1}^{t} \varepsilon_s + u_t\}_{t=1}^{T}$ under $\phi = (1-\eta)/\sqrt{\eta}$ and $\sigma = 1$ (and by the covariance matrix of $\{\sum_{s=1}^{t} \varepsilon_s\}_{t=1}^{T}$ if $\eta = 0$). We provide an explicit expression in the Appendix.

Under standard large sample theory, one would expect $\hat{\eta} \overset{a}{\sim} \mathcal{N}(\eta, (1-\eta^2)/T)$. We set the parameter space equal to $[0, 1]$, $n(\theta^*) = \sqrt{(1-\eta^2)/T + 6\eta^2/T^2}$, $F_n$ uniform on $[0, 1]$, and choose the Lebesgue density of $G_i$ to be proportional to the $i$th basis spline on the 53 knots $\{0.0, 1.0\} \cup \{0.5 + 0.5\tanh(-3 + 6i/50)\}_{i=0}^{50}$ (with "not-a-knot" end conditions). With $\varepsilon_R = 0.01$, the algorithm successfully delivers an exactly median unbiased nearly weighted risk minimizing estimator $\hat{\delta}_U^{\dagger}$. Results for all $T \in \mathbb{T}$ are in the replication files.

Fig. 6 displays its median bias and normalized risk, along with Stock and Watson's (1998) median unbiased estimator and the weighted risk minimizing estimator without any bias constraints for $T = 50$. The new estimator $\hat{\delta}_U^{\dagger}$ is seen to have very substantially lower risk than the previously suggested median unbiased estimator by Stock and Watson (1998) for all but very large values of $\eta$. The risk envelope for exactly median unbiased estimators is never more than 10% below the risk of $\hat{\delta}_U^{\dagger}$. As in the previous examples, this again implies that the impact of our choice of $F$ is fairly limited, and that $\hat{\delta}_U^{\dagger}$ comes reasonably close to being uniformly median absolute deviation minimizing among all exactly median unbiased estimators.

### 5.2. Quantile forecasts from an AR(1)

The final application again involves a stationary Gaussian AR(1) process, but now we are interested in constructing quantile forecasts. The data generating process is as in Section 4, that is $Y_t = \mu + u_t$, $u_t = \rho u_{t-1} + \varepsilon_t$, $\varepsilon_t \sim i.i.d.\mathcal{N}(0, \sigma^2)$ and $u_0 \sim \mathcal{N}(0, \sigma^2/(1-\rho^2))$. The parameter is $(\rho, \mu, \sigma)$, and we observe $X = (Y_1, \ldots, Y_T)$. For some given $0 < \alpha < 1$ and horizon $\tau > 0$, we seek to estimate the conditional $\alpha$ quantile of the future value $Y_{T+\tau}$. In particular, it is potentially attractive to construct estimators $\delta$ that are quantile unbiased in the sense that

$$P_\theta(Y_{T+\tau} < \delta(X)) = \alpha \quad \text{for all } (\rho, \mu, \sigma) \in [-0.95, 1) \times \mathbb{R} \times (0, \infty). \tag{22}$$

This ensures that in repeated applications, $Y_{T+\tau}$ indeed realizes to be smaller than the estimator $\delta(X)$ of the $\alpha$ quantile with probability $\alpha$, irrespective of the true value of the parameters governing $X$.[3]

A standard calculation shows that $Y_{T+\tau}|X \sim \mathcal{N}(\mu_\tau, \sigma_\tau^2)$ with $\mu_\tau = \mu + (Y_T - \mu)\rho^\tau$ and $\sigma_\tau^2 = \sigma^2(1-\rho^{2\tau})/(1-\rho^2)$, so that the conditional quantile is equal to $\mu_\tau + \sigma_\tau z_\alpha$, where $P(\mathcal{N}(0, 1) < z_\alpha) = \alpha$. It is not possible to use this expression as an estimator, however, since $\mu_\tau$ and $\sigma_\tau$ depend on the unknown parameter $(\rho, \mu, \sigma)$. A simple plug-in estimator is given by $\delta_{PI}(x) = \hat{\mu}_\tau + \hat{\sigma}_\tau z_\alpha$ which replaces $(\rho, \mu, \sigma)$ by $(\hat{\rho}_{OLS}, \hat{\mu}, \hat{\sigma}_{OLS})$, where $\hat{\rho}_{OLS}$ and $\hat{\sigma}_{OLS}^2$ are the OLS estimators of $\rho$ and $\sigma^2$, and $\hat{\mu} = T^{-1}\sum_{t=1}^{T} y_t$.

In order to cast this problem in the framework of this paper, let $\xi = (Y_{T+\tau} - \mu)\sqrt{1-\rho^2}/\sigma$, so that $\xi \sim \mathcal{N}(0, 1)$. We treat $\xi \in \mathbb{R}$ as fixed and part of the parameter, $\theta = (\xi, \rho, \mu, \sigma)$. In order to recover the stochastic properties of $\xi$, we integrate

3 In contrast, Müller and Watson (2016) determine predictive sets that contain the future value with *at least* $1 - \alpha$ for all permissible data generating processes, so they do not penalize overcoverage beyond the weighted average length criterion.

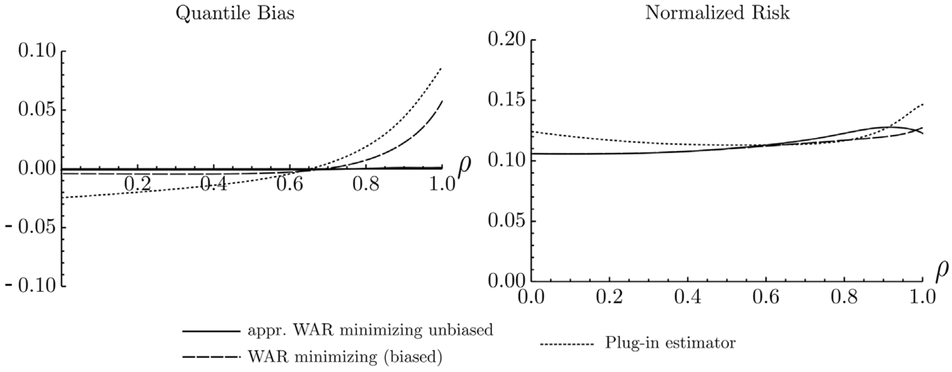


**Fig. 7.** Bias and risk in 10 step ahead 5% quantile forecast from AR(1).

over $\xi \sim \mathcal{N}(0, 1)$ in the weighting functions $F$ and $G_i$. In this way, weighted average bias constraints over $\theta$ amount to a bias constraint in the original model with $\xi$ stochastic and distributed $\mathcal{N}(0, 1)$.

Now in this notation, $Y_{T+\tau}|\theta$ becomes non-stochastic and equals

$$h(\theta) = \mu + \frac{\sigma}{\sqrt{1-\rho^2}}\xi.$$

Furthermore, the quantile bias in (22) is equivalent to weighted average bias over $\xi \sim \mathcal{N}(0, 1)$ with $c(\eta, \theta)$ equal to $\mathbf{1}[h(\theta) < \eta] - \alpha$. We use the usual quantile check function to measure the loss of estimation errors, $\ell(\eta, \theta) = |(h(\theta) - \eta)/\sigma| \cdot |\alpha - \mathbf{1}[h(\theta) < \eta]|$.

The problem is seen to be invariant to the group of translations $g(x, a) = a_\sigma(x + a_\mu)$ and $\bar{g}(\theta, a) = (\xi, \rho, a_\sigma(\mu + a_\mu), a_\sigma\sigma)$ for $a = (a_\mu, a_\sigma) \in A = \mathbb{R} \times (0, \infty)$. One set of maximal invariants is given by $X^* = M(X) = X_\Delta/s_y$ with $X_\Delta = (Y_1 - Y_1, \ldots, Y_T - Y_1)$ and $s_y^2 = \sum_{t=1}^T (Y_t - Y_1)^2$, and $\bar{M}((\xi, \rho, \mu, \sigma)) = \bar{g}((\xi, \rho, \mu, \sigma), (-\mu, \sigma^{-1})) = (\xi, \rho, 0, 1)$, as in the example discussed in Section 4. But in contrast to the case discussed there, the invariance here also leads to a corresponding change in the parameter of interest $h(\theta)$, $\hat{g}(\eta, a) = a_\sigma(\eta + a_\mu)$. Indeed, by (16), all invariant estimators can be written in the form

$$\delta(X) = Y_1 + \delta(X^*)s_y.$$

Note that $\int f_{\theta^*}(x^*)\phi(\xi)d\xi$ with $\phi$ the density of a standard normal is equal to the right-hand side of (19), since with $\xi \sim \mathcal{N}(0, 1)$, one recovers the distribution of $X^*$ without the conditioning on $\xi$. Furthermore, a calculation shows that $Y_{T+\tau} - Y_1|X_\Delta \sim \mathcal{N}(\omega' X_\Delta, \sigma_\Delta^2)$ with $\omega_i = (1-\rho)(1-\rho^\tau)/(T(1-\rho)+2\rho)$ for $i = 1, \ldots, T-1$, $\omega_T = (1+\rho^\tau(T(1-\rho)-1+2\rho))/(T(1-\rho)+2\rho)$, and $\sigma_\Delta^2 = ((1-\rho)T + 1 + 3\rho - 2(1+\rho)\rho^\tau - (1-\rho)(T-1)\rho^{2\tau})/((1-\rho^2)(T(1-\rho)+2\rho))$ (again without conditioning on $\xi$). Let $\hat{s}(\rho)^2 = \hat{x}^{*\prime}\Sigma(\rho)^{-1}\hat{x}^*$ be the weighted sum of squared GLS residuals as in (20). Furthermore, evaluating the weighted averages of $\ell^*(\delta(x^*), \theta^*)$ and $c^*(\delta(x^*), \theta^*)$ in (17) and (18) with weighting function $\xi \sim \mathcal{N}(0, 1)$ yields

$$\frac{\int f_{\theta^*}(x^*)c^*(\delta(x^*), \theta^*)\phi(\xi)d\xi}{\int f_{\theta^*}(x^*)\phi(\xi)d\xi} = E[\tilde{c}((\delta(x^*) - \omega' x^*)V/\hat{s}(\rho), \sigma_\Delta Z)] \tag{23}$$

$$\frac{\int f_{\theta^*}(x^*)\ell^*(\delta(x^*), \theta^*)\phi(\xi)d\xi}{\int f_{\theta^*}(x^*)\phi(\xi)d\xi} = E[\tilde{\ell}((\delta(x^*) - \omega' x^*)V/\hat{s}(\rho), \sigma_\Delta Z)] \tag{24}$$

where $\tilde{c}, \tilde{\ell} : \mathbb{R}^2 \mapsto \mathbb{R}$ with $\tilde{c}(\eta_1, \eta_2) = \mathbf{1}[\eta_2 < \eta_1] - \alpha$ and $\tilde{\ell}(\eta_1, \eta_2) = |\eta_1 - \eta_2| \cdot |\alpha - \mathbf{1}[\eta_2 < \eta_1]|$, $Z \sim \mathcal{N}(0, 1)$ and $V$ is an independent chi-distributed random variable with $T - 1$ degrees of freedom. We obtain an easily evaluated expression for these expectations in the Appendix.

The problem now is thus exactly of the form discussed in Section 2, with effective density equal to $\int f_{\theta^*}(x^*)\phi(\xi)d\xi$, bias and loss function equal to (23) and (24), and effective parameter space indexed only by $\rho \in [-0.95, 1.0]$. We hence implement the generic algorithm of Section 2.3. We normalize risk by $n(\theta^*) = \sqrt{(1-\rho^{2\tau})/(1-\rho^2)}$, set $\varepsilon_B = 0.002$ and $\varepsilon_R = 0.02$, and in addition to the integration over $\xi \sim \mathcal{N}(0, 1)$, choose $\rho$ uniform on $[-0.95, 1.0]$ in $F_n$, and $G_i$ equal to point masses on the same grid as had been used in Section 3.2. We generated nearly WAR minimizing estimators for $T \in \{10, 20, 50, 100, 200, 300, 400\}$, $\tau \in \{1, 2, 4, 10, 20\}$, $\tau \leq T/2$, and $\alpha \in \{0.01, 0.05, 0.1\}$.

Fig. 7 plots the results for $T = 50$, $\tau = 10$ and $\alpha = 0.05$, along with the plug-in estimator $\delta_{PI}$ defined above, and the WAR minimizing estimator without any quantile bias constraints. The plug-in estimator is seen to very severely violate the quantile unbiased constraint (22) for $\rho$ close to unity: instead of the nominal 5%, $Y_{T+\tau}$ takes on a value smaller than the quantile estimator about 13% of the time.

**Table 1**
Mean unbiased estimation of AR(1) coefficient with unknown mean and variance.

| | | Mean bias ×1000 | | | | | | MSE ×1000 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\rho$ | −0.3 | 0.5 | 0.8 | 0.90 | 0.95 | 0.99 | −0.3 | 0.5 | 0.8 | 0.90 | 0.95 | 0.99 |
| | Dist'n | $T=50$ | | | | | | | | | | | |
| New | $\chi_4^2-4$ | −0.9 | −0.1 | 2.2 | 1.8 | 1.4 | 2.3 | 19.5 | 17.7 | 11.8 | 9.6 | 8.6 | 8.7 |
| OW | $\chi_4^2-4$ | −1.6 | −0.5 | −2.7 | −7.5 | −13.0 | −20.6 | 19.8 | 17.4 | 11.8 | 9.7 | 8.6 | 8.3 |
| New | $\{-1,1\}$ | 1.7 | −2.0 | −1.5 | −1.3 | −1.3 | −1.3 | 20.3 | 19.0 | 12.6 | 10.1 | 9.1 | 8.8 |
| OW | $\{-1,1\}$ | 0.5 | −1.4 | −6.0 | −10.2 | −15.3 | −23.1 | 20.9 | 19.3 | 13.0 | 10.4 | 9.3 | 8.7 |
| | | $T=200$ | | | | | | | | | | | |
| New | $\chi_4^2-4$ | 0.7 | 0.4 | 0.6 | 0.0 | 0.2 | 0.4 | 4.8 | 3.8 | 2.0 | 1.3 | 0.8 | 0.5 |
| OW | $\chi_4^2-4$ | −0.2 | −0.2 | 0.0 | −0.8 | −1.5 | −4.3 | 4.7 | 3.9 | 2.0 | 1.3 | 0.9 | 0.6 |
| New | $\{-1,1\}$ | 1.0 | 0.5 | 0.1 | −0.2 | −0.1 | 0.0 | 4.8 | 4.0 | 2.1 | 1.3 | 0.9 | 0.6 |
| OW | $\{-1,1\}$ | 0.0 | −0.1 | −0.4 | −1.0 | −1.7 | −4.6 | 4.8 | 4.0 | 2.2 | 1.3 | 0.9 | 0.6 |

Notes: Entries are mean bias and mean square error in the estimation of the coefficient in a stationary AR(1) with i.i.d. innovations that are either mean-centered chi-squared with four degrees of freedom, or mean-zero discrete with two-point support on $\{-1, 1\}$. The two considered estimators are the "new" nearly mean unbiased estimator under Gaussian innovations derived here, and the analytically bias corrected estimator by Orcutt and Winokur (1969) "OW". Based on 100,000 simulations.

**Table 2**
Median unbiased estimation of AR(1) coefficient with unknown mean and variance.

| | | Median bias ×100 | | | | | | MAD ×100 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\rho$ | −0.3 | 0.5 | 0.8 | 0.90 | 0.95 | 0.99 | −0.3 | 0.5 | 0.8 | 0.90 | 0.95 | 0.99 |
| | Dist'n | $T=50$ | | | | | | | | | | | |
| New | $\chi_4^2-4$ | −1.7 | −0.9 | 0.1 | −0.1 | 0.2 | 0.4 | 11.0 | 10.4 | 8.3 | 7.2 | 6.2 | 4.7 |
| $\delta_{U,OLS}$ | $\chi_4^2-4$ | −1.8 | −0.7 | 0.2 | 0.1 | 0.1 | 0.5 | 11.0 | 10.4 | 8.6 | 7.6 | 6.6 | 5.1 |
| New | $\{-1,1\}$ | −0.1 | −0.1 | −0.5 | −0.5 | −0.3 | −0.4 | 11.3 | 10.8 | 8.5 | 7.4 | 6.4 | 4.9 |
| $\delta_{U,OLS}$ | $\{-1,1\}$ | −0.1 | 0.2 | −0.6 | −0.7 | −0.5 | −0.5 | 11.3 | 10.9 | 8.9 | 7.9 | 6.8 | 5.3 |
| | | $T=200$ | | | | | | | | | | | |
| New | $\chi_4^2-4$ | −2.3 | −0.3 | 0.9 | −0.2 | −0.1 | 0.3 | 5.4 | 4.9 | 3.5 | 2.8 | 2.2 | 1.5 |
| $\delta_{U,OLS}$ | $\chi_4^2-4$ | −2.3 | −0.3 | 0.8 | −0.3 | 0.0 | 0.4 | 5.4 | 4.9 | 3.5 | 2.8 | 2.3 | 1.6 |
| New | $\{-1,1\}$ | −1.5 | 0.2 | 1.0 | −0.2 | −0.4 | −0.1 | 5.4 | 5.0 | 3.6 | 2.8 | 2.2 | 1.5 |
| $\delta_{U,OLS}$ | $\{-1,1\}$ | −1.5 | 0.2 | 0.9 | −0.2 | −0.2 | 0.2 | 5.4 | 5.0 | 3.6 | 2.9 | 2.4 | 1.6 |

Notes: Entries are median bias and mean absolute deviation in the estimation of the coefficient in a stationary AR(1) with i.i.d. innovations that are either mean-centered chi-squared with four degrees of freedom, or mean-zero discrete with two-point support on $\{-1, 1\}$. The two considered estimators are exactly median unbiased under Gaussian innovations: the "new" nearly weighted risk minimizing estimator derived here, and the median inverted OLS estimator $\delta_{U,OLS}$. Based on 100,000 simulations.

## 6. Performance under non-Gaussian innovations

The estimators derived here have attractive unbiasedness and risk properties by construction in models with Gaussian innovations. In this section, we briefly explore the performance of these estimators in misspecified models due to non-Gaussian innovations.

We focus on two sample sizes $T = 50$ and $T = 200$, and two non-Gaussian distributions: the mean-centered chi-squared distribution with 4 degrees of freedom, and the mean-zero discrete distribution with support equal to $\{-1, 1\}$, taken from Andrews (1993). We report results for all previously considered examples, except that we omit the AR(1) example with known mean and variance for brevity. Tables 1–4 report the corresponding (non-normalized) biases and risks of the new estimators, as well as for a previously considered alternative estimator. Overall, the new estimators mostly perform quite comparably to the Gaussian case, where they are close to optimal by construction. A notable but unsurprising exception is the AR(1) forecast problem, where the non-Gaussian distribution of the future value induces severe quantile biases, especially for $\rho$ small.

## 7. Conclusion

We contend that the problem of identifying estimators with low bias and risk in parametric estimation problems may be fruitfully approached in a systematic manner. In particular, we suggest a numerical approximation technique to the resulting nonlinear program, and show that it delivers very nearly unbiased estimators with demonstrably close to minimal weighted average risk in a number of classic time series estimation problems.

All examples in the paper concern small sample problems, whose parametric structure is induced by an assumption of Gaussian innovations. It seems plausible that our numerical approach could also be usefully applied in the context of asymptotic limit problems, whose parametric structure either stems from LeCam type limit of experiment likelihood expansions,

**Table 3**
Median unbiased estimation of local-level parameter.

| | $\eta$ | Median bias $\times 100$ | | | | | | MAD $\times 100$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0.0 | 0.5 | 0.8 | 0.90 | 0.95 | 0.99 | 0.0 | 0.5 | 0.8 | 0.90 | 0.95 | 0.99 |
| | Dist'n | $T=50$ | | | | | | | | | | | |
| New | $\chi_4^2-4$ | 0.8 | 0.1 | −0.5 | −0.6 | 0.0 | 0.0 | 6.0 | 11.4 | 8.6 | 6.9 | 5.7 | 4.7 |
| SW | $\chi_4^2-4$ | −0.3 | 0.0 | −0.3 | −0.5 | −0.2 | 0.4 | 30.3 | 32.8 | 17.6 | 8.9 | 6.0 | 4.4 |
| New | $\{-1,1\}$ | 0.0 | −0.2 | 0.1 | 0.1 | 0.5 | 0.0 | 6.2 | 10.2 | 8.2 | 6.8 | 5.6 | 4.8 |
| SW | $\{-1,1\}$ | 0.2 | 0.1 | 0.3 | 0.3 | 0.5 | 0.4 | 30.8 | 32.8 | 17.1 | 8.6 | 5.9 | 4.5 |
| | | $T=200$ | | | | | | | | | | | |
| New | $\chi_4^2-4$ | 0.8 | 0.4 | 0.4 | 0.3 | −0.2 | −0.2 | 2.9 | 5.4 | 3.8 | 2.8 | 2.2 | 1.3 |
| SW | $\chi_4^2-4$ | 0.0 | 0.4 | 0.5 | −0.5 | −0.5 | −0.2 | 37.7 | 38.7 | 27.2 | 10.0 | 3.6 | 1.4 |
| New | $\{-1,1\}$ | −0.2 | 0.2 | 0.3 | 0.0 | −0.2 | −0.3 | 2.9 | 4.6 | 3.5 | 2.7 | 2.2 | 1.3 |
| SW | $\{-1,1\}$ | −0.1 | 1.1 | 0.6 | −0.2 | −0.3 | −0.2 | 37.6 | 38.6 | 27.0 | 9.7 | 3.6 | 1.4 |

Notes: Entries are median bias and mean absolute deviation in the estimation of $\eta$ in the local-level model (21) with $\varepsilon_t$ and $u_t$ independent and i.i.d. and either distributed mean-centered chi-squared with four degrees of freedom, or mean-zero discrete with two-point support on $\{-1, 1\}$. The two considered estimators are exactly median unbiased under Gaussian innovations: the "new" nearly weighted risk minimizing estimator derived here, and Stock and Watson's (1998) median inverted Nyblom statistic "SW". Based on 100,000 simulations.

**Table 4**
Quantile forecasts for AR(1) process.

| | $\rho$ | Quantile bias $\times 100$ | | | | | | Quantile risk $\times 100$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | −0.3 | 0.5 | 0.8 | 0.90 | 0.95 | 0.99 | −0.3 | 0.5 | 0.8 | 0.90 | 0.95 | 0.99 |
| | Dist'n | $T=50$ | | | | | | | | | | | |
| New | $\chi_4^2-4$ | −3.7 | −3.8 | −2.0 | −1.4 | −1.2 | −0.7 | 26 | 29 | 50 | 70 | 84 | 98 |
| PI | $\chi_4^2-4$ | −4.5 | −4.0 | −0.3 | 3.3 | 5.6 | 8.5 | 38 | 31 | 45 | 65 | 85 | 114 |
| New | $\{-1,1\}$ | −4.9 | −1.7 | −0.1 | 0.1 | 0.0 | −0.1 | 9 | 11 | 20 | 27 | 32 | 36 |
| PI | $\{-1,1\}$ | −5.0 | −2.7 | 2.2 | 4.8 | 6.7 | 8.5 | 14 | 11 | 19 | 27 | 34 | 43 |
| | | $T=200$ | | | | | | | | | | | |
| New | $\chi_4^2-4$ | −3.8 | −4.1 | −2.2 | −1.7 | −1.5 | −1.4 | 25 | 28 | 43 | 58 | 70 | 83 |
| PI | $\chi_4^2-4$ | −4.6 | −4.5 | −2.2 | −0.8 | 0.3 | 1.5 | 38 | 31 | 43 | 57 | 70 | 85 |
| New | $\{-1,1\}$ | −5.0 | −2.4 | 0.0 | 0.1 | 0.1 | 0.2 | 9 | 10 | 17 | 23 | 27 | 32 |
| PI | $\{-1,1\}$ | −5.0 | −3.9 | 0.0 | 1.1 | 1.8 | 2.7 | 14 | 11 | 17 | 23 | 28 | 34 |

Notes: Forecasts $\delta(X)$ are about the 5% quantile of $Y_{T+10}$ (10 steps ahead). Entries are the bias $P(Y_{T+10} < \delta(X)) - 0.05$ and the risk $E[|Y_{T+10} - \delta(X)| \cdot |\mathbf{1}[Y_{T+10} < \delta(X)] - 0.05|]$ in an AR(1) with i.i.d. innovations that are either mean-centered chi-squared with four degrees of freedom, or mean-zero discrete with two-point support on $\{-1, 1\}$. The two considered estimators are the "new" nearly WAR minimizing unbiased estimator under Gaussian innovations, and the plug-in estimator PI described in Section 5.2. Based on 100,000 simulations.

or from large sample distributional approximations. For instance, the statistical problem of estimating the AR(1) coefficient from a Gaussian data set converges in the LeCam sense under local-to-unity asymptotics to the parametric problem of estimating the mean reversion parameter from the observation of an Ornstein–Uhlenbeck process. Similarly, Stock and Watson (1998) apply the functional central limit theorem to transform the small sample problem of estimating the degree of time variation into a limit problem that involves the observation of a Gaussian process on the unit interval.

Such asymptotics naturally lead to an unbounded local parameter space. This does not impede our general approach to yield lower bounds on risk of any (nearly) unbiased estimator in the parametric limit problem, which in turn are lower bounds on the asymptotic risk of estimators in the underlying small sample problem. This can be useful to assess the performance of a given candidate estimator. But the unbounded parameter space does preclude a purely computational approach to the determination of a feasible estimator. To make further progress, it would presumably be necessary to rely in part on an estimator that is known to perform well over most of the parameter space, and to focus the computational approach on the bounded problematic region, similar to the switching approach suggested by Elliott et al. (2015) in the context of hypothesis testing problems. We leave such extensions to future work.

## Appendix A. Proofs

The proof of Lemma 2 uses the following lemma.

**Lemma 3.** *For all $a \in A$ and $x \in \mathcal{X}$, $O(g(x,a))^- = O(x)^- \circ a^-$.*

**Proof.** Replacing $x$ by $g(x,a)$ in (11) yields

$$\begin{aligned} g(x,a) &= g(M(g(x,a)), O(g(x,a))) \\ &= g(M(x), O(g(x,a))). \end{aligned}$$

Alternatively, applying $a$ on both sides of (11) yields

$$g(x,a) = g(g(M(x), O(x)), a) = g(M(x), a \circ O(x)).$$

Thus $g(M(x), a \circ O(x)) = g(M(x), O(g(x,a)))$. By the assumption that group actions $a$ are distinct, this implies that $O(g(x,a)) = a \circ O(x)$. The result now follows from $(a_2 \circ a_1)^{-} = a_1^{-} \circ a_2^{-}$. ■

**Proof of Lemma 2.** For an arbitrary measurable subset $\mathcal{B} \subset \mathcal{X} \times \Theta$,

$$\begin{aligned} & P_\theta((M(X), \bar{g}(\theta, O(X)^{-})) \in \mathcal{B}) \\ =\; & P_{\bar{g}(\bar{M}(\theta),\bar{O}(\theta))}((M(X), \bar{g}(\theta, O(X)^{-})) \in \mathcal{B}) \text{ (by } \theta = \bar{g}(\bar{M}(\theta), \bar{O}(\theta))) \\ =\; & P_{\bar{M}(\theta)}((M(g(X, \bar{O}(\theta))), \bar{g}(\theta, O(g(X, \bar{O}(\theta)))^{-})) \in \mathcal{B}) \text{ (invariance of problem)} \\ =\; & P_{\bar{M}(\theta)}((M(X), \bar{g}(\theta, O(X)^{-} \circ \bar{O}(\theta)^{-})) \in \mathcal{B}) \text{ (invariance of } M \text{ and Lemma 3)} \\ =\; & P_{\bar{M}(\theta)}((M(X), \bar{g}(\bar{g}(\bar{M}(\theta), \bar{O}(\theta)), O(X)^{-} \circ \bar{O}(\theta)^{-})) \in \mathcal{B}) \text{ (by } \theta = \bar{g}(\bar{M}(\theta), \bar{O}(\theta))) \\ =\; & P_{\bar{M}(\theta)}((M(X), \bar{g}(\bar{M}(\theta), O(X)^{-})) \in \mathcal{B}). \quad ■ \end{aligned}$$

## Appendix B. Details on algorithm of Section 2.3

The basic idea of the algorithm is to start with some guess for the Lagrange multipliers $\lambda$, compute the biases $B(\delta_\lambda, G_i)$, and adjust $(\lambda_i^l, \lambda_i^u)$ iteratively as a function of $B(\delta_\lambda, G_i)$, $i = 1, \ldots, m$.

To facilitate the repeated computation of $B(\delta_\lambda, G_i)$, $i = 1, \ldots, m$, it is useful to employ an importance sampling estimator. We use the proposal density $f_p$, where $f_p$ is the mixture density $f_p(x) = (24 + m - 2)^{-1}(12 f_{\theta_{p,1}}(x) + 12 f_{\theta_{p,m}}(x) + \sum_{i=2}^{m-1} f_{\theta_{p,i}}(x))$, where $\theta_{p,i}$ are equal to the location of the $m$ knots or point masses of $G_i$, respectively. In other words, under $f_p$, $X$ is generated by first drawing an index $J$ uniformly from $\{-10, -9, \ldots, m+11\}$, then draw $X$ from $f_{\theta_{p,J^*}}$, where $J^* = \max(\min(J, m), 1)$. The overweighting of the boundary values $\theta_{p,1}$ and $\theta_{p,m}$ by a factor of 12 counteracts the lack of additional importance sampling points to one side, leading to approximately constant importance sampling Monte Carlo standard errors in problems where $\theta$ is one-dimensional, as is the case in all our applications once invariance is imposed.

Let $X_l$, $l = 1, \ldots, N$ be i.i.d. draws from $f_p$. For a given estimator $\delta$, we approximate $B(\delta, G_i)$ by

$$B(\delta, G_i) = E_{f_p}\left[\int c(\delta(X), \theta)\frac{f_\theta(X)}{f_p(X)} dG_i(\theta)\right] \approx N^{-1}\sum_{l=1}^{N}\int c(\delta(X_l), \theta)\frac{f_\theta(X_l)}{f_p(X_l)} dG_i(\theta).$$

We further approximate $\int c(\eta, \theta) f_\theta(X_l) dG_i(\theta)$ for arbitrary $\eta$ by quadratic interpolation based on the closest three points in the grid $\mathcal{H}_l = \{\eta_{l,1}, \ldots, \eta_{l,M}\}$, $\eta_{l,i} < \eta_{l,j}$ for $i < j$ (we use the same grid for all $i = 1, \ldots, m$). The grid is chosen large enough so that $\delta_\lambda(X_l)$ is never artificially constrained for any value of $\lambda$ considered by the algorithm. Similarly, $R(\delta, F)$ is approximated by

$$R(\delta, F) \approx N^{-1}\sum_{l=1}^{N}\frac{\int \ell(\delta(X_l), \theta) f_\theta(X_l) dF(\theta)}{f_p(X_l)}$$

and $\int \ell(\eta, \theta) f_\theta(X_l) dF(\theta)$ for arbitrary $\eta$ is approximated with the analogous quadratic interpolation scheme.

Furthermore, for given $\lambda$, the minimizer $\delta_\lambda(X_l)$ of the function $L_l : \mathbb{R} \mapsto \mathbb{R}$

$$L_l(\eta) = \int \ell(\eta, \theta) f_\theta(X_l) dF(\theta) + \sum_{i=1}^{m}(\lambda_i^u - \lambda_i^l)\int c(\eta, \theta) f_\theta(X_l) dG_i(\theta)$$

is approximated by first obtaining the global minimum over $\eta \in \mathcal{H}_l$, followed by a quadratic approximation of $L_l(\eta)$ around the minimizing $\eta_{l,j^*} \in \mathcal{H}_l$ based on the three values of $L_l(\eta)$ for $\eta \in \{\eta_{l,j^*-1}, \eta_{l,j^*}, \eta_{l,j^*+1}\} \subset \mathcal{H}_l$.

For given $\varepsilon$, the approximate solution to (1) subject to (5) is now determined as follows:

1. Generate i.i.d. draws $X_l$, $l = 1, \ldots, N$ with density $f_p$.
2. Compute and store $\int c(\eta, \theta) f_\theta(X_l) dG_i(\theta)$ and $\int \ell(\eta, \theta) f_\theta(X_l) dF(\theta)$, $\eta \in \mathcal{H}_l$, $i = 1, \ldots, m$, $l = 1, \ldots, N$.
3. Initialize $\lambda^{(0)}$ as $\lambda_i^{u,(0)} = \lambda_i^{l,(0)} = 0.0001$ and $\omega_i^{u,(0)} = \omega_i^{l,(0)} = 0.05$, $i = 1, \ldots, m$.
4. For $k = 0, \ldots, K - 1$
   (a) Compute $\delta_{\lambda^{(k)}}(X_l)$ as described above, $l = 1, \ldots, N$.
   (b) Compute $B(\delta_{\lambda^{(k)}}, G_i)$ as described above, $i = 1, \ldots, m$.
   (c) Compute $\lambda^{(k+1)}$ from $\lambda^{(k)}$ via $\lambda_i^{u,(k+1)} = \lambda_i^{u,(k)} \exp(\omega_i^{u,(k)}(B(\delta_{\lambda^{(k)}}, G_i) - \varepsilon))$, $\lambda_i^{l,(k+1)} = \lambda_i^{l,(k)} \exp(\omega_i^{l,(k)}(-B(\delta_{\lambda^{(k)}}, G_i) - \varepsilon))$, $i = 1, \ldots, m$.

(d) Compute $\omega_i^{u,(k+1)} = \max(0.01, 0.5\omega_i^{u,(k)})$ if $(B(\delta_{\lambda^{(k+1)}}, G_i) - \varepsilon)(B(\delta_{\lambda^{(k)}}, G_i) - \varepsilon) < 0$, and $\omega_i^{u,(k+1)} = \min(100, 1.03\omega_i^{u,(k)})$ otherwise, and similarly $\omega_i^{l,(k+1)} = \max(0.01, 0.5\omega_i^{l,(k+1)})$ if $(B(\delta_{\lambda^{(k+1)}}, G_i) + \varepsilon)(B(\delta_{\lambda^{(k)}}, G_i) + \varepsilon) < 0$, and $\omega_i^{l,(k+1)} = \min(100, 1.03\omega_i^{l,(k)})$ otherwise.

The idea of Step 4.d is to slowly increase the speed of the change in the Lagrange multipliers as long as the sign of the violation remains the same in consecutive iterations, and to decrease it otherwise.

In the context of Step 2 of the algorithm in Section 2.3, we iterate as described above until the relative improvement of $\underline{R}$ over the last 25 iterations is less than 0.1%. For Step 3 of the algorithm in Section 2.3, we first initialize and apply the above iterations 200 times for $e_B = \varepsilon_B$. We then continue to iterate the above algorithm for another 400 iterations, but every 200 iterations increase or decrease $e_B$ via a simple bisection method based on whether or not $R(\delta_{\lambda^{(k)}}, F) < (1+\varepsilon_R)\underline{R}$. The check of whether the resulting $\hat{\delta}^* = \delta_{\lambda^{(K)}}$ satisfies the uniform bias constraint (3) is performed by directly computing its bias $b(\hat{\delta}^*, \theta)$ via the importance sampling approximation

$$b(\delta, \theta) \approx N^{-1} \sum_{l=1}^{N} c(\delta(X_l), \theta) \frac{f_\theta(X_l)}{f_p(X_l)}$$

over a fine but discrete grid $\theta \in \Theta_g \subset \Theta$.

The computation of the median function $m_{\delta_{\hat{\lambda}^\dagger}}$ in Step 2 of the algorithm in Section 3.2 is based on the importance sampling approximation

$$N^{-1} \sum_{l=1}^{N} \mathbf{1}[\delta_{\hat{\lambda}^\dagger}(X_l) < m_{\delta_{\hat{\lambda}^\dagger}}(\theta)] \frac{f_\theta(X_l)}{f_p(X_l)} \approx 1/2$$

with $m_{\delta_{\hat{\lambda}^\dagger}}(\theta)$ determined by a simple bisection algorithm, and is performed on the same grid that is employed to check the uniform bias property. The inverse function $m_{\delta_{\hat{\lambda}^\dagger}}^{-1}$ is obtained by linear interpolation between these points.

We set the number of draws to $N = 250{,}000$ in all applications. Computations for a given problem take no more than minutes on a modern PC, and seconds for the mean unbiased problems.

### B.1. Additional application-specific details

*AR(1) coefficient*

We set $\Theta_g = \{-0.95 + 1.95j/500\}_{j=0}^{500}$ and, for the median unbiased estimator, $\mathcal{H}_l = \mathcal{H}$ where $\mathcal{H}$ subdivides the knot locations into four equally long subintervals (so that $M = 4(m-1)+1$). Integration over $G_i$ and $F$ is performed using a Gaussian quadrature rule with 7 points separately on each of the intervals determined by the sequence of knots that underlie $G_i$.

A straightforward calculation yields $\hat{x}^{*\prime}\Sigma(\rho)^{-1}\hat{x}^* = \sum_{t=2}^{T}(x_t^* - \rho x_{t-1}^*)^2 + (1-\rho^2)(x_1^*)^2 - [(1-\rho)(x_1^* + x_T^* + (1-\rho)\sum_{t=2}^{T-1} x_t^*)^2]/(T(1-\rho)+2\rho)$.

*Degree of parameter time variation*

We set $\Theta_g = \{j/500\}_{j=0}^{500}$ and use the same construction for $\mathcal{H}_l$ as for the AR(1) coefficient.

Applying Lemma 4 in Elliott and Müller (2006) yields

$$f_{\theta^*}(x^*) = C \frac{\left(\sum_{t=1}^{T}(\tilde{x}_t^*)^2\right)^{-(T-1)/2}}{\sqrt{(1-\eta^{2T})/(T(1-\eta^2))}}$$

where $\tilde{x}_t^*$ are the residuals of a regression of $\{u_t^*\}_{t=1}^T$ on $\{1, \eta, \eta^2, \ldots, \eta^{T-1}\}$ with $u_t^* = \eta u_{t-1}^* + x_t^* - x_{t-1}^*$ and $u_1^* = x_1^*$.

*AR(1) quantile forecast*

We set $\Theta_g = \{-0.95 + 1.95j/500\}_{j=0}^{500}$, and $\mathcal{H}_l$ is an equally-spaced grid of 100 points such that the endpoints cover the quantile forecasts of level $0.1\alpha$ and $5\alpha$ conditional for all parameter values $\theta$ in the grid for $G_i$ whose likelihood is at least $10^{-5}$ of the average.

To obtain (24), note that with $\bar{g}(\theta^*, O(X)^-) = (\xi, \rho, -Y_1/s_y, 1/s_y)$,

$$E_{\theta^*}[\ell\left(\delta(X^*), \bar{g}(\theta^*, O(X)^-)\right) | X^* = x^*] = E_{\theta^*}[\tilde{\ell}\left(\delta(X^*)s_y, h(\theta^*) - Y_1\right) | X^* = x^*].$$

Now conditioning on $(s_y, X^*)$ and integrating out $\xi$, we obtain $h(\theta^*) - Y_1 \sim \mathcal{N}(s_y\omega' X^*, \sigma_\Delta^2)$ since $X_\Delta = s_y X^*$. Furthermore, the joint density of $(X_2^*, X_3^*, \ldots, X_T^*, s_y) \in S_{T-1} \times [0, \infty)$, where $S_{T-1}$ is the surface of the $T-1$ dimensional sphere, is proportional to $\exp[-\frac{1}{2}s_y^2\hat{s}(\rho)^2]s_y^{T-2}$ under $\theta^*$. Viewed as a function of $s_y$, this is recognized as the kernel of a chi-distributed random variable with $T-1$ degrees of freedom, scaled by $1/\hat{s}(\rho)$. Eq. (24) thus follows from the law of iterated expectations, and similarly for (23).

Let $f_{\chi,k}(s) = C_k \exp[-\frac{1}{2}s^2]s^{k-1}$ with $1/C_k = 2^{k/2-1}\Gamma(k/2)$ be the p.d.f. of a $\chi$ distributed random variable, and $F_{t,k}$ and $f_{t,k}$ be the c.d.f. and p.d.f. of a student-t variate with $k$ degrees of freedom. Then (23) is equal to

$$F_{t,T-1}\left(\sqrt{T-1}\frac{\delta(x^*)-\omega' x^*}{\hat{s}(\rho)\sigma_\Delta}\right)-\alpha.$$

To obtain a more explicit expression for (24), first condition on $V$ and integrate out $Z$ to obtain

$$\sigma_\Delta \phi(Vr) + \sigma_\Delta rV(\Phi(Vr) - \alpha)$$

where $\Phi$ and $\phi$ are the c.d.f. and p.d.f. of a standard normal variate, and $r = (\delta(x^*) - \omega' x^*)/(\hat{s}(\rho)\sigma_\Delta)$. Note that $\int_0^\infty \phi(rs)sf_{\chi,k}(s)ds = f_{t,k}(r\sqrt{k})\sqrt{k}$, so $\int_0^\infty \phi(sr)f_{\chi,T-1}(s)ds = \sqrt{T-2}\frac{C_{T-1}}{C_{T-2}}f_{t,T-2}(r\sqrt{T-2})$. Similarly, $\int_0^\infty \Phi(sr)sf_{\chi,T-1}(s)ds = \frac{C_{T-1}}{C_T}F_{t,T}(r\sqrt{T})$. Finally, $E[V] = \tilde{C}_{T-1} = \sqrt{2}\Gamma(T/2)/\Gamma((T-1)/2) = C_{T-1}/C_T$. Thus, (24) is equal to

$$\tilde{C}_{T-1}\sigma_\Delta[f_{t,T-2}(\sqrt{T-2}r)/\sqrt{T-2} + rF_{t,T}(\sqrt{T}r) - r\alpha].$$

## References


Albuquerque, J., 1973. The barankin bound: a geometric interpretation. IEEE Trans. Inform. Theory 19, 559–561.
Andrews, D.W.K., 1993. Exactly median-unbiased estimation of first order autoregressive/unit root models. Econometrica 61, 139–165.
Andrews, I., Armstrong, T.B., 2017. Unbiased instrumental variables estimation under known first-stage sign. Quant. Econ. 8, 479–503.
Andrews, D.W.K., Chen, H., 1994. Approximately median-unbiased estimation of autoregressive models. J. Bus. Econom. Statist. 12, 187–204.
Barankin, E.W., 1946. Locally best unbiased estimates. Ann. Math. Stat. 20, 477–501.
Cheang, W.-K., Reinsel, G.C., 2000. Bias reduction of autoregressive estimates in time series regression model through restricted maximum likelihood. J. Amer. Statist. Assoc. 95, 1173–1184.
Crump, R.K., 2008. Optimal conditonal inference in nearly-integrated autoregressive processes, Working paper. Federal Reserve Bank of New York.
Doss, H., Sethuraman, J., 1989. The price of bias reduction when there is no unbiased estimate. Ann. Statist. 17, 440–442.
Elliott, G., 1998. The robustness of cointegration methods when regressors almost have unit roots. Econometrica 66, 149–158.
Elliott, G., 2006. Forecasting with trending data. In: Elliott, G., Granger, C.W.J., Timmerman, A. (Eds.), Handbook of Economic Forecasting, Volume 1. North-Holland.
Elliott, G., Müller, U.K., 2006. Efficient tests for general persistent time variation in regression coefficients. Rev. Econom. Stud. 73, 907–940.
Elliott, G., Müller, U.K., Watson, M.W., 2015. Nearly optimal tests when a nuisance parameter is present under the null hypothesis. Econometrica 83, 771–811.
Glave, F.E., 1972. A new look at the barankin lower bound. IEEE Trans. Inform. Theory 18, 349–356.
Gospodinov, N., 2002. Median unbiased forecasts for highly persistent autoregressive processes. J. Econometrics 111, 85–101.
Han, C., Phillips, P., Sul, D., 2011. Uniform asymptotic normality in stationary and unit root autoregression. Econ. Theory 27, 1117–1151.
Harvey, A.C., 1989. Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press.
Hurvicz, L., 1950. Least squares bias in time series. In: Koopmans, T. (Ed.), Statistical Inference in Dynamic Economic Models. Wiley, New York, pp. 365–383.
Kariya, T., 1980. Locally robust test for serial correlation in least squares regression. Ann. Statist. 8, 1065–1070.
Kemp, G.C.R., 1999. The behavior of forecast errors from a nearly integrated AR(1) model as both sample size and forecast horizon become large. Econ. Theory 15 (2), pp. 238–256.
Kendall, M.G., 1954. Note on the bias in the estimation of autocorrelation. Biometrika 41, 403–404.
King, M.L., 1980. Robust tests for spherical symmetry and their application to least squares regression. Ann. Statist. 8, 1265–1271.
Lehmann, E.L., 1986. Testing Statistical Hypotheses, second ed. Wiley, New York.
Lehmann, E.L., Casella, G., 1998. Theory of Point Estimation, second ed. Springer, New York.
Lehmann, E.L., Romano, J.P., 2005. Testing Statistical Hypotheses. Springer, New York.
MacKinnon, J.G., Smith, A.A., 1998. Approximate bias correction in econometrics. J. Econometrics 85, 205–230.
Marriott, F.H.C., Pope, J.A., 1954. Bias in the estimation of autocorrelations. Biometrika 41, 393–402.
McAulay, R., Hofstetter, E., 1971. Barankin bounds on parameter estimation. IEEE Trans. Inform. Theory 17, 669–676.
Müller, U.K., Watson, M.W., 2016. Measuring uncertainty about long-run predictions. Rev. Econom. Stud. 83.
Nyblom, J., 1989. Testing for the constancy of parameters over time. J. Amer. Statist. Assoc. 84, 223–230.
Orcutt, G.H., Winokur, H.S., 1969. First order autoregression: inference, estimation, and prediction. Econometrica 37, 1–14.
Pantula, S.G., Gonzalez-Farias, G., Fuller, W.A., 1994. A comparison of unit-root test criteria. J. Bus. Econom. Statist. 12, 449–459.
Park, H.J., Fuller, W.A., 1995. Alternative estimators and unit root tests for the autoregressive process. J. Time Series Anal. 16, 415–429.
Phillips, P., 1977. Approximations to some finite sample distributions associated with a first order stochastic difference equation. Econometrica 45, 463–486.
Phillips, P., 1978. Edgeworth and saddlepoint approximations in the first-order noncircular autoregression. Biometrika 65, 91–108.
Phillips, P., 1979. The sampling distribution of forecasts from a first order autoregression. J. Econometrics 9, 241–261.
Phillips, P., 2012. Folklore theorems, implicit maps, and indirect inference. Econometrica 80, 425–454.
Phillips, P.C.B., Han, C., 2008. Gaussian inference in AR(1) time series with or without a unit root. Econ. Theory 24, 631–650.
Quenouille, M.H., 1956. Notes on bias in estimation. Biometrika 43, 353–360.
Roy, A., Fuller, W.A., 2001. Estimation for autoregressive time series with a root near 1. J. Bus. Econom. Statist. 19, 482–493.
Sargan, J.D., Bhargava, A., 1983. Maximum likelihood estimation of regression models with first order moving average errors when the root lies on the unit circle. Econometrica 51, 799–820.
Sawa, T., 1978. The exact moments of the least squares estimator for the autoregressive model. J. Econometrics 8, 159–172.
Shaman, P., Stine, R.A., 1988. The bias of autoregressive coefficient estimators. J. Amer. Statist. Assoc. 83, 842–848.
Stock, J.H., 1991. Confidence intervals for the largest autoregressive root in U.S. macroeconomic time series. J. Bus. Econ. Stat. 28, 435–459.
Stock, J.H., 1994. Unit roots, structural breaks and trends. In: Engle, R.F., McFadden, D. (Eds.), Handbook of Econometrics, Vol. 4. North Holland, New York, pp. 2740–2841.
Stock, J.H., 1996. VAR, error correction and pretest forecasts at long horizons. Oxford Bull. Econ. Stat. 58, 685–701.
Stock, J.H., Watson, M.W., 1998. Median unbiased estimation of coefficient variance in a time-varying parameter model. J. Amer. Statist. Assoc. 93, 349–358.
Tanaka, K., 1983. Asymptotic expansions associated with the AR(1) model with unknown mean. Econometrica 51, 1221–1231.
Tanaka, K., 1984. An asymptotic expansion associated with the maximum likelihood estimators in ARMA models. J. R. Statist. Soc. B 46, 58–67.
White, J.S., 1961. Asymptotic expansions for the mean and variance of the serial correlation coefficient. Biometrika 48, 85–94.
Yu, J., 2012. Bias in the estimation of the mean reversion parameter in continuous time models. J. Econometrics 169, 114–122.