Nearly Weighted Risk Minimal Unbiased Estimation*

Ulrich K. Müller and Yulong Wang
Economics Department
Princeton University
July 2017

Abstract

Consider a non-standard parametric estimation problem, such as the estimation of the AR(1) coefficient close to the unit root. 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 demonstrate the usefulness of our generic approach by also applying it to estimation in a predictive regression, estimation of the degree of time variation, and long-range quantile point forecasts for an AR(1) process with coefficient close to unity.

JEL classification: C13, C22
Keywords: Mean bias, median bias, local-to-unity, nearly non-invertible MA(1)

*The authors thank seminar participants at Harvard/MIT, Princeton, University of Kansas and University of Pennsylvania for helpful comments and suggestions. Müller gratefully acknowledges financial support by the NSF via grant SES-1226464.
1 Introduction

Much recent work in econometrics concerns inference in non-standard problems, that is, problems that don’t reduce to a Gaussian shift experiment, even asymptotically. Prominent examples include inference under weak identification (cf. Andrews, Moreira and Stock (2006, 2008)), in models involving local-to-unity time series (Elliott, Rothenberg, and Stock (1996), Jansson and Moreira (2006), Mikusheva (2007)), and in moment inequality models. The focus of this recent work is the derivation of hypothesis tests and confidence sets, and in many instances, optimal or nearly optimal procedures have been identified. In comparison, the systematic derivation of attractive point estimators has received less attention.

In this paper we seek to make progress by deriving 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 $m$ non-negative Lagrange multipliers. The Lagrangian can be written as an integral over the data, since both the objective and 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.
It is not possible to finely discretize a problem involving an unbounded parameter space. At the same time, many non-standard problems become approximately standard in the unbounded part of the parameter space. For instance, a large concentration parameter in a weak instrument problem, or a large local-to-unity parameter in a time series problem, induce nearly standard behavior of test statistics and estimators. We thus suggest “switching” to a suitable nearly unbiased estimator in this standard part of the parameter space in a way that ensures overall attractive properties.

These elements—discretization of the original problem, analytical lower bound on risk, numerical approximation, switching—are analogous to Elliott, Müller and Watson’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.

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 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, invariance considerations are more involved and potentially more powerful in

\(^1\)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).
an estimation problem compared to a hypothesis testing problem, since it sometimes makes sense to impose invariance (or equivariance) with respect to the parameter of interest. We suitably extend the theory of invariant estimators in Chapter 3 of Lehmann and Casella (1998) in Section 4 below.

We demonstrate our approach in four time series estimation problems. In all of them, we consider an asymptotic version of the problem.


Second, we consider a nearly non-invertible MA(1) model. The parameter estimator here suffers from the so-called pile-up problem, that is the MA(1) root 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 in the context of estimating the degree of parameter time variation. We derive an alternative, nearly WAR minimizing median unbiased estimator under absolute value loss and find it to have very substantially lower risk under moderate and large amounts of time variation.

Third, we derive nearly WAR minimizing mean and median unbiased estimators of the regression coefficient in the so-called predictive regression problem, that is with a weakly exogenous local-to-unity regressor. This contributes to previous analyses and estimator suggestions by Mankiw and Shapiro (1986), Stambaugh (1986, 1999), Cavanagh, Elliott, and Stock (1995), Jansson and Moreira (2006), Kothari and Shanken (1997), Amihud and Hurvich (2004), Eliasz (2005) and Chen and Deo (2009).

Finally, we use our framework to derive nearly quantile unbiased long-range forecasts from
a local-to-unity AR(1) process with unknown mean. The unbiased property here means that 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), Turner (2004) and Elliott (2006) for related studies of forecasts in a local-to-unity AR(1).

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

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 some of 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 AR(1) parameter, as partially reviewed above, focusses 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. In the AR(1) and predictive regression problems, despite numerous suggestions for alternative estimators, no mean unbiased estimator is known, so our approach comes close to solving these classic time series problems. For some applications, median unbiased estimators were previously derived, but it was not known
whether there exists alternative median unbiased estimators with much less risk. In the
problem of median unbiased estimation of the AR(1) coefficient with known mean, we find
that the suggestion of Andrews (1993) is essentially optimal. In contrast, in the problem of
estimating a nearly non-invertible MA(1) root our systematic approach yields a median un-
based estimator that outperforms a previously suggested and commonly employed estimator
by a substantial margin.

The remainder of the paper is organized as follows. Section 2 sets up the 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 problem of estimating the
AR(1) coefficient under local-to-unity asymptotics as our running example. In Section 5, we
provide details on the three additional examples mentioned above. Section 6 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 $\eta$, $c(\eta, \theta) = \mathbb{1}[\eta > h(\theta)] - \frac{1}{2}$, 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

$$\min_{\delta} R(\delta, F) \quad \text{s.t.} \quad b(\delta, \theta) = 0 \ \forall \theta \in \Theta. \quad (1)$$

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

$$-\varepsilon \leq b(\delta, \theta) \leq \varepsilon \ \forall \theta \in \Theta \quad (3)$$

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 an autoregressive process of order 1, \( y_t = \rho y_{t-1} + \varepsilon_t, t = 1, \ldots, T, \) where \( \varepsilon_t \sim iid \mathcal{N}(0, 1) \) and \( y_0 = 0 \). Under local-to-unity asymptotics, \( \rho = \rho_T = 1 - \theta/T \), this model converges to the limit experiment (in the sense of LeCam) of observing \( X \), where \( X \) is an Ornstein-Uhlenbeck process \( dX(s) = -\theta X(s)ds + dW(s) \) with \( X(0) = 0 \) and \( W \) a standard Wiener process. The random element \( X \) takes on values in the set of cadlag functions on the unit interval \( \mathcal{X} \), and the Radon-Nikodym derivative of \( X \) relative to Wiener measure \( \nu \) is

\[
f_\theta(x) = \exp[-\frac{1}{2}\theta(x(1)^2 - 1) - \frac{1}{2}\theta^2 \int_0^1 x(s)^2 ds].
\]

The aim is to estimate \( \theta \in \mathbb{R} \) in a way that (nearly) minimizes weighted risk under a mean or median bias constraint.

More generally, suppose that \( y_t = \rho y_{t-1} + u_t \) with \( y_0 = o_p(T^{1/2}), T^{-1/2} \sum_{t=1}^{[T]} u_t \Rightarrow \sigma W(\cdot) \), and \( \rho = 1 - \theta/T \). Then for any consistent estimator \( \hat{\sigma}^2 \) of the long run variance \( \sigma^2 \) of \( \{u_t\} \), \( X_T(\cdot) = T^{-1/2}\hat{\sigma}^{-1}y_{[T]} \Rightarrow X(\cdot) \). Thus, for any \( \nu \)-almost surely continuous estimator \( \hat{\sigma} : \mathcal{X} \rightarrow \mathbb{R} \) in the limiting problem, \( \hat{\sigma}(T^{-1/2}\hat{\sigma}^{-1}y_{[T]}) \) has the same asymptotic distribution as \( \delta(X) \) by the continuous mapping theorem. Therefore, if \( \delta(X) \) is (nearly) median unbiased for \( \theta \), then \( \delta(X_T) \) is asymptotically (nearly) median unbiased. For a (nearly) mean unbiased estimator \( \hat{\delta}(X) \), one could put constraints on \( \{u_t\} \) that ensure uniform integrability of \( \hat{\delta}(X_T) \), so that again \( E_\theta[\delta(X_T)] \rightarrow E_\theta[\delta(X)] \) for all \( \theta \). Alternatively, one might define “asymptotic mean bias” and “asymptotic risk” as the sequential limits

\[
\lim_{K \to \infty} \lim_{T \to \infty} E_\theta[1[|\delta(X_T)| < K]\hat{\delta}(X_T) - h(\theta))] \\
\lim_{K \to \infty} \lim_{T \to \infty} E_\theta[1[\ell(\delta(X_T), \theta) < K] \ell(\delta(X_T), \theta)]
\]

respectively. With these definitions, the bias and (weighted average) risk of \( \delta(X_T) \) converges to the bias and weighted average risk of \( \delta \) by construction, as long as \( \delta \) is \( \nu \)-almost surely continuous with finite mean and risk relative to (4).

The usual OLS estimator for \( \theta \) (which is also equal to the MLE) is \( \delta_{OLS}(x) = -\frac{1}{2}(x(1)^2 - 1)/\int_0^1 x(s)^2 ds \). As demonstrated by Phillips (1987) and Chan and Wei (1987), for large \( \theta \),
(2θ)^{-1/2} \delta_{OLS}(X) \sim N(0, 1) as \theta \to \infty$, yielding the approximation \delta_{OLS}(X) \sim N(\theta, 2\theta). We thus use the normalization function n(\theta) = \sqrt{2\theta + 10}$ (the offset of 10 ensures that $n(\theta) > 0$ even as $\theta = 0$).

We initially focus on estimating $\theta$ under squared loss and a mean bias constraint for $\theta \geq 0$; we return to the problem of median unbiased estimation in Section 3.2 below. Figure 1 depicts the normalized mean bias and the normalized mean squared error of $\delta_{OLS}$ as a function of $\theta$. As is well known, the bias of $\delta_{OLS}$ is quite substantial: even under fairly strong mean reversion of, say, $\theta = 60$, the bias is approximately on fifth of the root mean squared error of the estimator. ▲

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.

\footnote{For a given normalized weighting function $F_n$, different values for the offset amount to different effective weighting functions.}
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 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 (3) clearly also satisfies

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

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 $\theta_1, \ldots, \theta_m$.

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

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

where $\lambda = (\lambda_1, \ldots, \lambda_m)$ and $\lambda_i = (\lambda_i^u, \lambda_i^l)$. 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

$$L(\delta, \lambda) = \int \left( \int f_\theta(x)\ell(\delta(x), \theta)dF(\theta) + \sum_{i=1}^{m} \lambda_i^u(\int f_\theta(x)c(\delta(x), \theta)dG_i(\theta) - \varepsilon) \right) d\nu(x).$$

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 $\delta$. 

9
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 \( \mathbb{E} \), for all estimators whose normalized bias is smaller than 0.06 uniformly in \( \lambda \), it suffices that there exists some estimator with uniformly bounded mean squared error.

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).

**Running Example:** Suppose \( m = 1, F_n \) puts unit mass at \( \theta_0 = 0 \) and \( G_1 \) puts point mass at \( \theta_1 > \theta_0 \). With square loss \( \ell(\eta, \theta) = (\eta - \theta)^2 \), normalized mean bias constraint \( c(\eta, \theta) = (\eta - \theta)/n(\theta), \varepsilon = 0 \) and \( \lambda_i^u = 0, \delta_\lambda(x) \) minimizes
\[
\int f_\theta(x)\ell(\delta(x), \theta)dF(\theta) - \lambda_1^1 \int f_\theta(x)c(\delta(x), \theta)dG_i(\theta) = f_\theta(x)(\delta(x) - \theta_0)^2/n(\theta_0)^2 - \lambda_1^1 f_\theta(x)\delta(x) - \theta_1/n(\theta_1).
\]
This is solved by \( \delta_\lambda(x) = \theta_0 + \frac{1}{2}\lambda_1 n(\theta_0)^2 f_\theta(x)/(f_\theta(x)n(\theta_1)) \). For \( \theta_1 = 3 \) and \( \lambda_1^1 = 1.5 \), a numerical calculation yields \( R(\delta_\lambda, F) = r_n(\delta_\lambda, \theta_0) = 0.48 \) and \( B(\delta_\lambda, G_1) = b_n(\delta_\lambda, \theta_1) = 0.06 \), so that \( L(\delta_\lambda, \lambda) = 0.39 \). We can conclude from Lemma 1 that no estimator that is unbiased for all \( \theta \in \mathbb{R} \) can have risk \( R(\delta, F) = r_n(\delta, \theta_0) \) smaller than 0.39, and the same holds also for all estimators \( \delta \) that are unbiased only at \( \theta_1 \). Moreover, this lower risk bound also holds for all estimators whose normalized bias is smaller than 0.06 uniformly in \( \theta \in \mathbb{R} \), and more generally for all estimators whose normalized mean bias is smaller than 0.06 in absolute value only at \( \theta_1 \).

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 reason
is that if the bias $b(\delta_{\lambda^*}, \theta)$ does not vary greatly in $\theta$, and the distributions $G_i$ are sufficiently
dense in $\Theta$, then any estimator that satisfies (5) cannot violate (3) by very much.

This suggests 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 $\hat{R} = L(\delta_{\hat{\lambda}^\dagger}, \hat{\lambda}^\dagger)$.

3. Obtain an approximate solution $(\hat{e}^*, \hat{\delta}^*)$ to the problem

$$\min_{e \geq 0, \delta} e \text{ s.t. } R(\delta, F) \leq (1 + \varepsilon_R)\hat{R} \quad (8)$$
$$\text{and } -e \leq B(\delta, G_i) \leq e, \ i = 1, \ldots, m \quad (9)$$

and check whether $\hat{\delta}^*$ satisfies (3). If it doesn’t, go back to Step 1 and use a finer
discretization. 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 $e = \varepsilon_B$ (at least if $\hat{\lambda}^\dagger$ is the exact solution), and $R(\delta_{\hat{\lambda}^\dagger}, F) = \hat{R} < (1 + \varepsilon_R)\hat{R}$. 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 \( R \). At the same time, a finer discretization also adds additional constraints on the bias function of \( \delta^* \), 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 EMW’s approach to numerically approximate an approximately 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). We provide corresponding tables and matlab code for all examples in this paper under www.princeton.edu/~umueller.

An important class of non-standard estimation problems are limit experiments that arise from the localization of an underlying parameter \( \psi \in \Psi \) around problematic parameter values \( \Psi_N \subset \Psi \), such as the local-to-unity asymptotics for the AR(1) estimation problem in the running example. Other examples include weak instrument asymptotics (where the limit is taken for values of the concentration parameter local-to-zero), or local-to-binding moment conditions in moment inequality models. In these models, standard estimators, such as, say, (bias corrected) maximum likelihood estimators, are typically asymptotically unbiased and efficient for values of \( \psi \notin \Psi_N \). Moreover, this good behavior of standard estimators typically also arises, at least approximately, under local asymptotics with localization parameter \( \theta \) as \( \varsigma(\theta) \to \infty \) for some appropriate function \( \varsigma : \Theta \to \mathbb{R} \). For instance, as already mentioned in Section 2.1, Phillips (1987) and Chan and Wei (1987) show that the OLS estimator for the AR(1) coefficient \( \psi = \rho \) is approximately normal under local-to-unity asymptotics \( \rho = \rho_T = 1 - \theta / T \) as \( \varsigma(\theta) = \theta \to \infty \). This can be made precise by limit of experiment arguments, as discussed in Section 4.1 of EMW. Intuitively, the boundary of the localized parameter space \( \Theta \) typically continuously connects to the unproblematic part of the parameter space \( \Psi \setminus \Psi_N \) of the underlying parameter \( \psi \).

It then makes sense to focus the numerical strategy on the truly problematic part of the local parameter space, where the standard estimator \( \delta_S \) is substantially biased or inefficient. As in EMW, consider a partition of the parameter space \( \Theta \) into two regions \( \Theta = \Theta_N \cup \Theta_S \). For large enough \( \kappa_S \), \( \Theta_S = \{ \theta : \varsigma(\theta) > \kappa_S \} \) is the (typically unbounded) standard region where \( \delta_S \) is approximately unbiased and efficient, and \( \Theta_N = \Theta \setminus \Theta_S \) is the problematic non-standard region. Correspondingly, consider estimators \( \hat{\delta} \) that are of the following switching
form

\[ \hat{\delta}(x) = \chi(x)\delta_S(x) + (1 - \chi(x))\delta_N(x) \quad (10) \]

where \( \chi(x) = 1[\hat{\zeta}(x) > \kappa_x] \) for some estimator \( \hat{\zeta}(X) \) of \( \zeta(\theta) \). By choosing a large enough cut-off \( \kappa_x > \kappa_S \), we can ensure that \( \delta_S \) is only applied to realizations of \( X \) that stem from the relatively unproblematic part of the parameter space \( \Theta_S \) with high probability. The numerical problem then becomes the determination of the estimator \( \delta_N : \mathcal{X} \mapsto H \) that leads to overall appealing bias and efficiency properties of \( \hat{\delta} \) in (10) for any \( \theta \in \Theta \). As discussed in EMW, an additional advantage of this switching approach is that the weighting function \( F \) can then focus on \( \Theta_N \) and the “seam” part of \( \Theta_S \) where \( P_\theta(\chi(X)) \) is not close to zero or one. The good efficiency properties of \( \hat{\delta} \) on the remaining part of \( \Theta_S \) are simply inherited by the good properties of \( \delta_S \).

Incorporating this switching approach in the numerical strategy above is straightforward: For given \( \kappa_x \), the problem in Step 3 now has \( \delta \) constrained to be of the form (10), that is for all \( x \) that lead to \( \chi(x) = 1 \), \( \hat{\delta}^*(x) = \delta_S(x) \), and the numerical challenge is only to determine a suitable \( \delta_N = \hat{\delta}_N^* \). By not imposing the form (10) in Step 2, we still obtain a bound on weighted risk that is valid for any estimator \( \delta \) satisfying \(-\varepsilon_B \leq b(\delta, 0) \leq \varepsilon_B \) for all \( \Theta \) (that is, without imposing the functional form (10)). Of course, for a given switching rule \( \chi \), it might be that no nearly minimal weighted risk unbiased estimator of the form (10) exists, because \( \delta_S \) is not unbiased or efficient enough relative to the weighting \( F \), so one might try instead with larger values for \( \kappa_S \) and \( \kappa_x \).

**Running Example:** Taking the local-to-unity limit of Marriott and Pope’s (1954) mean bias expansion \( E[\hat{\rho}_{OLS} - \rho] = -2\rho/T + o(T^{-1}) \) for the AR(1) OLS estimator \( \hat{\rho}_{OLS} \) in a model without intercept suggests \( \delta_S(x) = \delta_{OLS}(x) - 2 \) as a good candidate for \( \delta_S \). And indeed, it can be inferred from Figure 1 that the normalized mean bias of \( \delta_S \) is smaller than 0.01 for \( \theta > \kappa_S = 20 \). Furthermore, setting \( \hat{\zeta}(x) = \delta_{OLS}(x) \) and \( \kappa_x = 40 \), we find that \( P_\theta(\hat{\zeta}(X) > \kappa_x) < 0.1\% \) for all \( \theta \leq \kappa_S \), so draws from the problematic part of the parameter space are only very rarely subject to the switching.

It is now an empirical matter whether imposing the switching to \( \delta_S \) in this fashion is costly in terms of risk, and this can be evaluated with the algorithm above by setting the upper bound of the support of \( F \) to, say, 80, so that the weighted average risk criterion incorporates parameter values where switching to \( \delta_S \) occurs with high probability (\( P_\theta(\hat{\zeta}(X) > \kappa_x) > 99\% \) for \( \theta > 70 \)). ▲
3 Two Special Cases

3.1 Mean Unbiased Estimation under Quadratic Loss

Consider the special case of quadratic 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

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

(11)
a linear function of \( \tilde{\lambda}_i = \frac{1}{2}(\lambda^n_i - \lambda_i^\star) \). Plugging this back into the objective and constraints yields

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

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

where

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

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 \( \int f_\theta(x)/n(\theta) dG_i(\theta) \) cannot be written as a linear combination of \( \int f_\theta(x)/n(\theta) dG_j(\theta), j \neq i \) 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 \( \tilde{\lambda}^\star = \Omega^{-1}\omega \), where \( \omega = (\omega_1, \ldots, \omega_m)' \). Step 3 of the algorithm generally becomes a quadratically constrained quadratic program, but since \( e \) is scalar, it can easily be solved by conditioning on \( e \), with a line search as outer-loop. This remains true also if \( \delta \) is restricted to be of the switching form (10), where \( \delta_N \) is of the form (11), and \( R(\delta, F) \) and \( B(\delta, G_i) \) contain additional constants that arise from expectations over draws with \( \chi(x) = 1 \).

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 $[0, 80]$, $\varepsilon_B = \varepsilon_R = 0.01$, and restrict $\hat{\delta}^*$ to be of the switching form (10) with $\hat{\epsilon} = \delta_{OLS}$ and $\kappa_\chi = 40$ as described at the end of the last section. With $G_i$ equal to point masses at $\{0.0, 0.2^2, \ldots, 10^2\}$, the algorithm successfully identifies a nearly weighted risk minimizing estimator $\hat{\delta}^*$ that is uniformly nearly unbiased for $\theta \geq 0$. The remaining bias in $\hat{\delta}^*$ is very small: with the normalized mean squared error of $\hat{\delta}^*$ close to one, $\varepsilon_B = 0.01$ implies that about 10,000 Monte Carlo draws are necessary for the largest bias of $\hat{\delta}^*$ to be approximately of the same magnitude as the Monte Carlo standard error of its estimate.

Figure 2 plots the normalized bias and risk of the nearly weighted risk minimizing unbiased estimator $\hat{\delta}^*$. As a point of reference, we also report the normalized bias and risk of the OLS estimator $\delta_{OLS}$ (replicated from Figure 1), 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 $\theta \in [0, 80]$. As can be seen from Figure 2, however, this comes at the cost of a large mean bias.

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 estimator whose $G_i$ averaged normalized bias is no larger than $\varepsilon_B = 0.01$ 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 $\theta \in [0, 80]$ 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 (1959)). 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 monoticity of $m\hat{\delta}$, so if $P_\theta(\delta_{\lambda^1} = \theta) \leq 1/2$ 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 $\lambda^\dagger$ for the problem (1) subject to (5) for $\varepsilon = 0$, and associated value $\bar{R} = L(\delta_{\lambda^1}, \lambda^\dagger)$. Choose $G_i$ and $\lambda^\dagger$ to ensure that $P_\theta(\delta_{\lambda^1} = \theta) < 1/2$ for any boundary points of $\Theta$. If $m_{\delta_{\lambda^1}}$ 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 $\hat{\delta}_U(X) = m_{\hat{\delta}_U^{-1}}(\hat{\delta}_U(X))$.

If $R(\hat{\delta}_U^\dagger, F) > (1 + \varepsilon_R)R$, go back to Step 1 and use a finer discretization. Otherwise, $\hat{\delta}_U = \hat{\delta}_U^\dagger$ has the desired properties.

For the same reasons as discussed in Section 2.3, it can make sense to restrict the estimator of Step 2 of the switching form (10). For unbounded $\Theta$, it is not possible to numerically compute $m_{\hat{\delta}_U^{-1}}$ exactly, but if the problem becomes well approximated by a standard problem with standard unbiased estimator $\delta_S$ as $\varsigma(\theta) \to \infty$, then setting the inverse median function to be the identity for very large $\varsigma(X)$ leads to numerically negligible median biases.

**Running Example:** As discussed in Section 2.1, we combine the median unbiased constraint on $\theta \geq 0$ 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 $[0, 80]$, with risk now normalized as $r_n(\delta, \theta) = E_\theta[|\delta(X) - \theta|]/n(\theta)$, where $n(\theta) = \sqrt{2\theta + 10}$. Phillips (1974, 1975) provides Edgeworth expansions of $\hat{\rho}_{OLS}$, which implies that $\hat{\rho}_{OLS}/(1 - T^{-1})$ has a median bias of order $o(T^{-1})$ for $|\rho| < 1$. Taking the limit as $\rho \to 1$ suggests that the median bias of $\delta_S(x) = \delta_{OLS}(x) - 1$ should be small for moderate and large $\theta$, and a numerical analysis shows that this is indeed the case. We set $\varepsilon_R = 0.01$, and impose switching to $\delta_S$ when $\delta_{OLS}(x) > \kappa_x = 40$. 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 on the $m = 51$ knots $\{0, 0.2^2, \ldots, 10^2\}$ (with “not-a-knot” end conditions). We only compute $m_{\hat{\delta}_U^{-1}}$ for $\theta \leq 100$ and set it equal to the identity function otherwise; the median bias of $\hat{\delta}_U^\dagger$ for $\theta \geq 100$ is numerically indistinguishable from zero with a Monte Carlo standard error of the median bias estimate of approximately 0.2%.

Figure 3 reports the median bias and normalized risk of $\hat{\delta}_U^\dagger$, the OLS estimator $\delta_{OLS}$, the OLS-based median unbiased estimator $\delta_{U,OLS}(x) = m_{\hat{\delta}_{OLS}^{-1}}(\delta_{OLS}(x))$, 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. ▲
Figure 3: Median and Absolute Loss in Local-to-Unity AR(1) with Known Mean

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 $x \in \mathcal{X}$. Assume that the elements in $a$ are distinct in the sense that $g(x, a_1) = g(x, a_2)$ for some $x \in \mathcal{X}$ implies $a_1 = a_2$.

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 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.

4 Invariance
\( O : \mathcal{X} \mapsto A \) and \( \bar{O} : \Theta \mapsto A \) (as discussed on page 216-217 in Lehmann and Casella (1998)). Then by definition of a maximal invariant, \( M(M(x)) = M(x) \), \( \bar{M}(M(\theta)) = \bar{M}(\theta) \), and we have the decomposition

\[
\begin{align*}
x & = g(M(x), O(x)) \\
\theta & = \bar{g}(\bar{M}(\theta), \bar{O}(\theta)).
\end{align*}
\]

Equations (12) and (16) imply that any invariant estimator satisfies

\[
\delta(g(x,a)) = \bar{g}(\delta(x),a) \quad \text{for all } x \in \mathcal{X} \text{ and } a \in A.
\]

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 joint distribution of \( (M(X), \bar{g}(\theta, O(X)^-)) \) under \( \theta \) only depends on \( \theta \) through \( \bar{M}(\theta) \).

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

\[
\ell(\eta, \theta) = \ell(\hat{g}(\eta, a), \bar{g}(\theta, a)) \quad \text{for all } \eta \in H, \theta \in \Theta \text{ and } a \in A
\]

\[
c(\eta, \theta) = c(\hat{g}(\eta, a), \bar{g}(\theta, a)) \quad \text{for all } \eta \in H, \theta \in \Theta \text{ and } a \in A.
\]

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 \( \ell^i : H \mapsto H \), such as quadratic loss or absolute value loss, automatically satisfies (14). Similarly, constraints of the form \( c(\eta, \theta) = c^i(\eta, h(\theta)) \), such as those arising from mean or median unbiased constraints, satisfy (15).

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

\[
\delta(g(x,a)) = \hat{g}(\delta(x),a) \quad \text{for all } x \in \mathcal{X} \text{ and } a \in A.
\]

Equations (12) and (16) imply that any invariant estimator satisfies

\[
\delta(x) = \hat{g}(\delta(M(x)), O(x)) \quad \text{for all } x \in \mathcal{X}.
\]

It is useful to think about the right-hand side of (17) as inducing the invariance property: \textit{Any} function \( \delta_a : M(\mathcal{X}) \rightarrow H \) defines an invariant estimator \( \delta(x) = \hat{g}(\delta(M(x)), O(x)) \).
\( \hat{g}(\delta_a(M(x)), O(x)) \). Given that any invariant estimator satisfies (17), 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 after observing \( y_t = m_y + u_t \), and \( u_t = \rho u_{t-1} + \varepsilon_t \), where \( \varepsilon_t \sim iid \mathcal{N}(0,1) \) and \( u_0 = 0 \). Under local-to-unity asymptotics \( \rho = 1 - \eta/T \) for \( \eta \geq 0 \) and \( m_y = m_yT = \mu T^{1/2} \) for \( \mu \in \mathbb{R} \), the limiting problem involves the observation \( X \) on the space of continuous functions on the unit interval, where \( X(s) = \mu + J(s) \) and \( J(s) = \int_0^s e^{-(s-r)\eta}dW(r) \) is an Ornstein-Uhlenbeck process with \( J(0) = 0 \). This observation is indexed by \( \theta = (\eta, \mu) \), and the parameter of interest is \( \eta = h(\theta) \). The problem is invariant to transformations \( g(x, a) = x + a \) with \( a \in A = \mathbb{R} \), and corresponding group \( \hat{g}((\eta, \mu), a) = (\eta, \mu + a) \). One choice for maximal invariants is \( M(x) = g(x, -x(0)) \) and \( \tilde{M}((\eta, \mu)) = \hat{g}((\eta, \mu), -\mu) = (\eta, 0) \). The Lemma asserts that \( M(X) = X - X(0) \) and \( \hat{g}(\theta, O(X)^{-}) = (\eta, \mu - X(0)) \) have a joint distribution that only depends on \( \theta \) through \( \eta \). The induced group \( \hat{g} \) is given by \( \hat{g}(\eta, a) = \eta \) for all \( a \in A \). Under (14), the loss must not depend on the location parameter \( \mu \). Invariant estimators \( \delta \) are numerically invariant to translation shifts of \( X \), and can all be written in the form \( \delta(x) = \hat{g}(\delta_a(x - x(0)), x(0)) = \delta_a(x - x(0)) \) for some function \( \delta_a \) that maps paths \( y \) on the unit interval with \( y(0) = 0 \) to \( H \). ▲

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

\[
\begin{align*}
    r(\delta, \theta) &= E_\theta[\ell(\delta(X), \theta)] \\
    &= E_\theta[\ell(\delta(M(X)), \hat{g}(\theta, O(X)^{-}))] \quad \text{(by (12) and (14) with } a = O(X)^{-}) \\
    &= E_{\tilde{M}(\theta)}[\ell(\delta(M(X)), \hat{g}(\tilde{M}(\theta), O(X)^{-}))] \quad \text{(by Lemma 2)} \\
    &= E_{\tilde{M}(\theta)}[E_{\tilde{M}(\theta)}[\ell(\delta(M(X)), \hat{g}(\tilde{M}(\theta), O(X)^{-})) | M(X)]]
\end{align*}
\]

and similarly

\[
\begin{align*}
    b(\delta, \theta) &= E_\theta[c(\delta(X), \theta)] \\
    &= E_{\tilde{M}(\theta)}[E_{\tilde{M}(\theta)}[c(\delta(M(X)), \hat{g}(\tilde{M}(\theta), O(X)^{-})) | M(X)]]
\end{align*}
\]

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

\[
\begin{align*}
    \ell^*(x^*, \theta^*) &= E_{\theta^*}[\ell(\delta(X^*), \hat{g}(\theta^*, O(X)^{-})) | X^* = x^*] \quad (18) \\
    c^*(\delta(x^*), \theta^*) &= E_{\theta^*}[c(\delta(X^*), \hat{g}(\theta^*, O(X)^{-})) | X^* = x^*]. \quad (19)
\end{align*}
\]
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 (17), 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 \( \eta = h(\theta) \), and \( \tilde{g}(\theta, a) \) does not affect \( h(\theta), \ell(\delta(X^*), \tilde{g}(\theta^*, O(X^-)) = \ell(\delta(X^*), \theta^*) \) and \( c(\delta(X^*), \tilde{g}(\theta^*, O(X^-)) = c(\delta(X^*), \theta^*) \). Thus the starred problem amounts to estimating \( \eta \) from the observation \( X^* = M(X) = X - X(0) \) (whose distribution does not depend on \( \mu \)). Since \( X^* \) is equal to the Ornstein-Uhlenbeck process \( J \) with \( J(0) = 0 \), the starred problem is exactly what was previously solved in the running example. So in particular, the solutions \( \hat{\delta}^* \) and \( \hat{\delta}^* \), analyzed in Figures 2 and 3, now applied to \( X^* \), are also (nearly) weighted risk minimizing unbiased translation invariant estimators in the problem of observing \( X = \mu + J \).

A different problem arises, however, in the empirically more relevant case with the initial condition drawn from the stationary distribution. With \( y_t = m_y + u_t, u_t = \rho u_{t-1} + \varepsilon_t, \varepsilon_t \sim iid \mathcal{N}(0, 1) \) and \( u_0 \sim \mathcal{N}(0, (1 - \rho^2)^{-1}) \) independent of \( \{\varepsilon_t\} \) for \( \rho < 1 \) (and zero otherwise), the limiting problem under \( \rho = 1 - \eta/T \) and \( m_y = m_{y,T} = \mu T^{1/2} \) asymptotics is the observation of the process \( X(s) = \mu + J^S(s) \), where \( J^S(s) = Ze^{-\eta s/\sqrt{2\eta}} + J(s) \) and \( Z \sim \mathcal{N}(0, 1) \) is independent of \( J \) for \( \eta > 0 \) (and \( J^S = J \) otherwise). For \( \eta \geq 0 \), the density of \( X^* = X - X(0) \) is then given by (cf. Elliott (1999))

\[
 f_{\theta^*}(x^*) = \sqrt{\frac{2}{2 + \eta}} \exp \left[ -\frac{1}{2} \eta (x^*(1)^2 - 1) - \frac{1}{2} \eta^2 \int_0^1 x^*(s)^2 ds + \frac{1}{2} \eta (\eta \int_0^1 x^*(s) ds + x^*(1))^2 \right].
\]

(20)

Analogous to the discussion below (4), this asymptotic problem can be motivated more generally under suitable assumptions about the underlying processes \( y_t \).

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 (20). We use the normalization function \( n(\theta^*) = \sqrt{2\eta + 20} \), and make the same choices for \( F_n \) and \( G_i \) on \( \eta \) as in the problem with zero initial condition. Taking the local-to-unity limit of the approximately mean unbiased Orcutt and Winokur (1969) estimator of \( \rho \) in a regression that includes a constant suggests \( \delta_{OLS} - 4 \) to be nearly mean unbiased for large \( \eta \), where now \( \delta_{OLS}(x^*) = -\frac{1}{2} (\bar{x}^*(1)^2 - 1) / \int_0^1 \bar{x}^*(s)^2 ds \) with \( \bar{x}^*(s) = x^*(s) - \int_0^1 x^*(l) dl \). Similarly, Phillips’
(1975) results suggest $\delta_{OLS} - 3$ to be nearly median unbiased. We switch to these estimators if $\delta_{OLS}(x^*) > \kappa_X = 60$, set $\varepsilon_B = 0.01$ and $\varepsilon_R = 0.01$ for the nearly median unbiased estimator, and $\varepsilon_R = 0.02$ for the mean unbiased estimator.

Figures 4 and 5 show the normalized bias and risk of the resulting nearly weighted risk minimizing unbiased invariant estimators. We also plot the performance of the analogous set of comparisons as in Figures 2 and 3. 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 Figure 3, the WAR minimizing median unbiased estimator now has perceptibly lower risk than the OLS based median unbiased estimator for small $\theta$, but the gains are still fairly moderate.

Figure 6 compares the performance of the nearly mean unbiased estimator with some previously suggested alternative estimators: The local-to-unity limit $\delta_{OLS} - 4$ of the analytically bias corrected estimator by Orcutt and Winokur (1969), the weighted symmetric estimator analyzed by Pantula, Gonzalez-Farias, and Fuller (1994) and Park and Fuller (1995), and the MLE $\delta_{MLE}(x^*) = \arg\max_{\eta \geq 0} f_{\theta^*}(x^*)$ based on the maximal invariant likelihood (20) (called “restricted” MLE by Cheang and Reinsel (2000)). In contrast to $\hat{\delta}^*$, all of these previously suggested estimators have non-negligible mean biases for values of $\eta$ close to zero, with the smallest mean bias at $\eta = 0$ of nearly a third of the root mean squared error. ▲
Figure 5: Median and Absolute Loss in Local-to-Unity AR(1) with Unknown Mean

Figure 6: Mean Bias and MSE in Local-to-Unity AR(1) with Unknown Mean
5 Further Applications

5.1 Degree of Parameter Time Variation

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

\[ y_t = m + \phi \sum_{s=1}^{t} \varepsilon_s + u_t, \quad (\varepsilon_t, u_t) \sim iid \mathcal{N}(0, I_2) \]  

with only \( y_t \) observed. The parameter of interest is \( \phi \), the degree of time variation of the ‘local mean’ \( m + \phi \sum_{s=1}^{t} \varepsilon_s \). Under \( m = T^{-1/2} \mu \) and \( \phi = T^{-1} \eta \) asymptotics, we have the functional convergence of the partial sums

\[ X_T(s) = T^{-1/2} \sum_{t=1}^{[sT]} y_t \Rightarrow X(s) = \mu s + \eta \int_0^s W_\varepsilon(l)dl + W_u(s) \]

(22)

with \( W_\varepsilon \) and \( W_u \) two independent standard Wiener processes. The limit problem is the estimation of \( \eta \) from the observation \( X \), whose distribution depends on \( \theta = (\eta, \mu) \in [0, \infty) \times \mathbb{R} \). This problem is invariant to transformations \( g(x, a) = x + a \cdot s \) with \( a \in A = \mathbb{R} \), and corresponding group \( \tilde{g}((\eta, \mu), a) = (\eta, \mu + a) \). One choice for maximal invariants is \( M(x) = g(x, -x(1)) \) and \( \tilde{M}((\eta, \mu)) = \tilde{g}((\eta, \mu), -\mu) = (\eta, 0) \). The results of Section 4 imply that all invariant estimators can be written as functions of \( X^* = M(X) \), so that

\[ X^*(s) = \eta \int_0^s W_\varepsilon(l)dl + W_u(s) - s \left( \eta \int_0^1 W_\varepsilon(l)dl + W_u(1) \right) \]

(23)

which does not depend on any nuisance parameters. The distribution (23) arises more generally as the limit of the suitably scaled partial sum process of the scores, evaluated at the maximum likelihood estimator, of well-behaved time series models with a time varying parameter; see Stock and Watson (1998), Elliott and Müller (2006a) and Müller and Petalas (2010). Lemma 2 and Theorem 3 of Elliott and Müller (2006a) imply that the density of \( X^* \) with respect to the measure of a Brownian Bridge \( W(s) - sW(1) \) is given by

\[ f_\eta(x^*) = \sqrt{\frac{2\eta e^{-\eta}}{1 - e^{-2\eta}}} \exp \left[ -\eta \tilde{x}_\eta(1)^2 - \eta^2 \int_0^1 \tilde{x}_\eta(s)^2 ds - \frac{2\eta}{1 - e^{-2\eta}} \left( e^{-\eta} \tilde{x}_\eta(1) + \eta \int_0^1 e^{-\eta s} \tilde{x}_\eta(s)ds \right)^2 \\
+ \left( \tilde{x}_\eta(1) + \eta \int_0^1 \tilde{x}_\eta(s)ds \right)^2 \right] \]

where \( \tilde{x}_\eta(s) = x^*(s) - \eta \int_0^s e^{-\eta(s-l)}x^*(l)dl. \)
As discussed by Stock (1994), the local level model (21) is intimately linked to the MA(1) model 
\[ \Delta y_t = \phi e_t + \Delta u_t. \] 
Under small time variation asymptotics \( \phi = \eta / T, \Delta y_t \) is nearly noninvertible with local-to-unity MA(1) coefficient \( \zeta = 1 - \eta / T + o(T^{-1}) \). Both \( X_T'(s) = X_T(s) - sX_T(1) \) and the first differences \( \{\Delta y_t\}_{t=2}^T \) form small sample maximal invariants to translations of \( y_t \), so that they contain the same information. This suggests that \( X^* \) is also the limit problem (in the sense of LeCam) of a stationary Gaussian local-to-unity MA(1) model without a constant, and it follows from the development in Elliott and Müller (2006a) that this is indeed the case.

It has long been recognized that the maximum likelihood estimator of the MA(1) coefficient exhibits non-Gaussian limit behavior under non-invertibility; see Stock (1994) for a historical account and references. In particular, the MLE \( \hat{\zeta} \) suffers from the so-called pile-up problem \( P(\hat{\zeta} = 1 | \zeta = 1) > 0 \), and Sargan and Bhargava (1983) derive the limiting probability to be 0.657. The full limiting distribution of the MLE does not seem to have been previously derived. The above noted equivalence leads us to conjecture that under \( \zeta = 1 - \eta / T \) asymptotics,

\[ T(1 - \hat{\zeta}) \Rightarrow \delta_{MLE}(X^*) = \arg\max_{\eta} f_\eta(X^*). \] (24)

A numerical calculation based on (23) confirms that under \( \eta = 0 \), \( P(\delta_{MLE}(X^*) = 0) = 0.657 \), but we leave a formal proof of the convergence in (24) to future research.

In any event, with \( P(\delta_{MLE}(X^*) = 0) > 1/2 \) under \( \eta = 0 \), it is not possible to base a median unbiased estimator of \( \eta \in [0, \infty) \) on \( \delta_{MLE}(X^*) \), as the median function \( m_{\delta_{MLE}} \) is not one-to-one for small values of \( \eta \). Stock and Watson (1998) base their exactly median unbiased estimator on the Nyblom (1989) statistic \( \int_0^1 X^*(s)^2 ds \).

We employ the algorithm discussed in Section 3.2 to derive an alternative median unbiased estimator that comes close to minimizing weighted average risk under absolute value loss. Taking the local-to-unity limit of Tanaka’s (1984) expansions suggests \( \delta_S(x^*) = \delta_{MLE}(x^*) + 1 \) to be approximately median unbiased and normal with variance \( 2\eta \) for large \( \eta \). We thus consider risk normalized by \( n(\theta^*) = \sqrt{\eta + 6} \) (the offset 6 ensures that normalized risk is well defined also for \( \eta = 0 \)). We impose switching to \( \delta_S \) whenever \( \delta_{MLE}(x^*) > \kappa_x = 10 \), set \( F_n \) to be uniform on \([0, 30]\), choose the Lebesgue density of \( G_i \) to be proportional to the \( i \)th basis spline on the knots \( \{0, 0.2^2, \ldots, 8^2\} \), and set the inverse median function \( m_{\delta_{MLE}}^{-1} \) to the identity for arguments larger than 50 (the median bias of \( \delta_S \) for \( \eta > 50 \) is indistinguishable from zero with a Monte Carlo standard error of approximately 0.2%). With \( \varepsilon_R = 0.01 \), the
algorithm then successfully delivers a nearly weighted risk minimizing estimator $\hat{\delta}_U^+$ that is exactly median unbiased for $\eta \leq 50$, and very nearly median unbiased for $\eta > 50$.

Figure 7 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. Our estimator $\hat{\delta}_U^+$ is seen to have very substantially lower risk than the previously suggested median unbiased estimator by Stock and Watson (1998) for all but very small values of $\eta$. For instance, under $\eta = 15$ (a degree of time variation that is detected approximately 75% of the time by a 5% level Nyblom (1989) test of parameter constancy), the mean absolute loss of the Stock and Watson (1998) estimator is about twice as large.

### 5.2 Predictive Regression

A stylized small sample version of the predictive regression problem with a weakly exogenous and persistent regressor $z_t$ is given by

$$
y_t = m_y + b z_{t-1} + r \varepsilon_{zt} + \sqrt{1 - r^2} \varepsilon_{yt}$$

$$
z_t = m_z + u_t$$

$$
u_t = \rho u_{t-1} + \varepsilon_{zt} \quad u_0 \sim \mathcal{N}(0, (1 - \rho^2)^{-1})$$
with \((e_{zt}, e_{yt})' \sim iid \mathcal{N}(0, I_2)\) independent of \(u_0, -1 < r < 1\) known, and only \(\{y_t, z_{t-1}\}_{t=1}^T\) are observed. The parameter of interest is the regression coefficient \(b\). Other formulations of the predictive regression problem differ by not allowing an unknown mean for the regressor, or by different assumptions about the distribution of \(u_0\).

Under asymptotics with \(\rho = 1 - \gamma/T, b = \eta/T, m_y = \mu_y/\sqrt{T}\) and \(m_z = \mu_z \sqrt{T}\), we obtain

\[
\left(\frac{T^{-1/2} \sum_{t=1}^{[T]} y_t}{T^{-1/2} \sum_{t=1}^{[T]} z_t}\right) \Rightarrow X(\cdot) = \begin{pmatrix} X_1(\cdot) \\ X_2(\cdot) \end{pmatrix} = \begin{pmatrix} \mu_y s + \eta \int_0^s X_2(l)dl + rW_z(s) + \sqrt{1-r^2}W_y(s) \\ \mu_z + e^{-\gamma s} \sqrt{2\gamma} + \int_0^s e^{-\gamma(s-l)}dW_z(l) \end{pmatrix}
\]

with \(Z \sim \mathcal{N}(0,1)\) independent of the standard Wiener processes \(W_y\) and \(W_z\). The distribution of the limit process \(X\) depends on the unknown \(\theta = (\mu_y, \mu_z, \eta, \gamma) \in \mathbb{R}^3 \times (0, \infty)\), and \(\eta\) is the parameter of interest. The problem is seen to be invariant to the group of transformations indexed by \(a = (a_y, a_z, a_\eta) \in \mathbb{R}^3\)

\[
g(X, a) = \begin{pmatrix} X_1(\cdot) + a_y \cdot + a_\eta \int_0^s X_2(s)ds \\ X_2(\cdot) + a_z \end{pmatrix}
\]

\[
g(\theta, a) = \theta + (a_y, a_z, a_\eta, 0)
\]

and one choice for maximal invariants is

\[
\begin{align*}
M(X) &= X^* = \begin{pmatrix} X_1^*(\cdot) \\ X_2^*(\cdot) \end{pmatrix} = \begin{pmatrix} X_1(\cdot) - X_1(1) - \hat{\eta}_{OLS}(X) \int_0^1 X_2^*(s)ds \\ X_2(\cdot) - \int_0^1 X_2(s)ds \end{pmatrix} \\
\bar{M}(\theta) &= \theta^* = (0, 0, 0, \gamma)
\end{align*}
\]

where \(\hat{\eta}_{OLS}(X) = \int_0^1 X_2^*(s)dX_1(1)/\int_0^1 X_2^2(s)ds\). With \(\hat{g}(\eta, a) = \eta + a_\eta\) and \(O(X) = (X_1(1), \int_0^1 X_2(s)ds, \hat{\eta}_{OLS}(X), 0)\), (17) implies that all invariant estimators can be written in the form

\[
\hat{\eta}_{OLS}(x) + \delta_a(x^*),
\]

that is, they are equal to the OLS estimator \(\hat{\eta}_{OLS}(x)\), plus a “correction term” that is a function of the maximal invariant \(X^*\).

The implications of imposing invariance are quite powerful in this example: the problem of determining the nearly weighted average risk minimizing unbiased estimator relative to (18) and (19) is only indexed by \(\theta^*\), that is by the mean reversion parameter \(\gamma\). Invariance
thus also effectively eliminates the parameter of interest $\eta$ from the problem. Solving for $\ell^*$ and $c^*$ in (18) and (19) requires the computation of the conditional expectation of the loss function and constraint evaluated at $\hat{g}(\hat{M}(\theta), O(X^-))$ given $X^*$ under $\theta^*$. With $\ell$ and $c$ only depending on $\theta$ through $\eta$, this requires derivation of the conditional distribution of $\hat{\eta}_{OLS}(X)$ given $X^*$. A calculation shows that under $\hat{\theta}^*$ (that is, with the true value of $\hat{\theta} = (\mu_y, \mu_z, \eta, \gamma)$ equal to $\theta^* = (0, 0, 0, \gamma)$)

$$
\hat{\eta}_{OLS}(X) | X^* \sim \mathcal{N}\left(r\int_0^1 X_2^*(s)dX_2^*(s) \int_0^1 X_2^*(s)^2ds + r\gamma, \frac{1 - r^2}{\int_0^1 X_2^*(s)^2ds}\right).
$$

Equation (26) implies that under $\theta^*$, the unconditional expectation of $\hat{\eta}_{OLS}(X)$ is equal to $-r$ times the bias of the OLS estimator $\hat{\gamma}_{OLS} = -\int_0^1 X_2^*(s)dX_2^*(s)/\int_0^1 X_2^*(s)^2ds$ of the mean reversion parameter $\gamma$. This tight relationship between the bias in the OLS coefficient estimator, and the OLS bias in estimating the autoregressive parameter, was already noted by Stambaugh (1986). What is more, let $\tilde{\gamma}$ be an alternative estimator of the mean reverting parameter based on $X_2^*$, and set $\delta_a$ in (25) equal to $r(\hat{\gamma}_{OLS} - \tilde{\gamma})$. Then the unconditional bias of the resulting estimator of $\eta$ is equal to $-r$ times the mean bias of $\tilde{\gamma}$, and its mean squared error is equal to $r^2$ times the mean squared error of $\tilde{\gamma}$ plus $E[(1 - r^2)/\int_0^1 X_2^*(s)^2ds]$. Thus, setting $\delta_a(x^*)$ equal to $r(\hat{\gamma}_{OLS} - \tilde{\gamma})^*$ for $\hat{\gamma}^*$ the nearly WAR minimizing mean unbiased estimator derived at the end of Section 4 yields a nearly WAR minimizing mean unbiased estimator of $\eta$ in the predictive regression problem. Correspondingly, the Orcutt and Winokur (1969)-based bias correction in the estimation of $\eta$ suggested by Stambaugh (1986) and the restricted maximum likelihood estimator suggested by Chen and Deo (2009) have mean biases that are equal to $-r$ times the mean biases reported in Figure 6. We omit a corresponding plot of bias and risk for brevity, but note that just as in Figure 4, a comparison with the (lower bound on the) risk envelope shows that our estimator comes quite close to being uniformly minimum variance among all (nearly) unbiased estimators.

The median bias of $\hat{\eta}_{OLS} + r(\hat{\gamma}_{OLS} - \tilde{\gamma})$, however, is not simply equal to $-r$ times the
median bias of \( \hat{\gamma} \). One also cannot simply invert the median function of an estimator of \( \eta \) due to the presence of the nuisance parameter \( \gamma \). We thus apply our general algorithm to obtain a nearly WAR minimizing nearly median unbiased estimator. We use \( n(\theta^*) = \sqrt{2\gamma + 20} \), \( \varepsilon_B = \varepsilon_R = 0.01 \), set \( F_n \) uniform on \([0, 80]\) and let \( G_i \) be point masses on \( \gamma \in \{0, 0.2, \ldots, 10^2\} \) (the Gaussian noise in (26) induces smooth solutions even for point masses, in contrast to direct median unbiased estimation of the mean reverting parameter). For \( \hat{\gamma}_{OLS} > k_X = 60 \) we switch to the estimator \( \delta_a = 3r \), which is nearly median unbiased for large \( \gamma \) and \( |r| \) close to 1.

Figure 8 reports the results for \( r = -0.95 \), along with the same analogous set of comparisons\(^3\) already reported in Figure 3.

5.3 Quantile Long-range Forecasts from a Local-To-Unity Autoregressive Process

The final example again involves a local-to-unity AR(1) process, but now we are interested in constructing long-run quantile forecasts. To be specific, assume \( y_t = m + u_t, u_t = \rho u_{t-1} + \varepsilon_t, \varepsilon_t \sim iid \mathcal{N}(0,1) \) and \( u_0 \sim \mathcal{N}(0, (1 - \rho^2)^{-1}) \) independent of \( \varepsilon_t \). We observe \( \{y_t\}_{t=1}^T \), and given some \( 0 < \alpha < 1 \) and horizon \( \tau \), seek to estimate the conditional \( \alpha \) quantile of \( y_{t+\tau} \). Under

\(^3\)We could not replicate the favorable results reported by Eliasz (2005) for his conditionally median unbiased estimator, derived for a closely related model with zero initial condition on \( z_0 \).
asymptotics with $m = T^{1/2} \mu$, $\rho = 1 - \gamma / T$ and $\tau = \lfloor rT \rceil$, the limiting problem involves the observation of the stationary Ornstein-Uhlenbeck process $X$ on the unit interval,

$$X(s) = \mu + \frac{Ze^{-\gamma s}}{\sqrt{2\gamma}} + \int_0^s \exp(-\gamma(s - l))dW(l)$$

with $Z \sim \mathcal{N}(0, 1)$ independent of $W$. The problem is indexed by the unknown parameters $(\mu, \gamma) \in \mathbb{R} \times (0, \infty)$, and the object of interest is the $\alpha$ quantile of $X(1 + r)$ conditional on $X$. It is potentially attractive to construct estimators $\hat{\beta}$ that are quantile unbiased in the sense that

$$P_{(\mu, \gamma)}(X(1 + r) < \hat{\beta}(X)) = \alpha \quad \text{for all } (\mu, \gamma) \in \mathbb{R} \times (0, \infty). \quad (27)$$

This ensures that in repeated applications, $X(1 + r)$ indeed realizes to be smaller than the estimator $\hat{\beta}(X)$ of the $\alpha$ quantile with probability $\alpha$, irrespective of the true value of the parameters governing $X$.

A straightforward calculation shows that $X(1 + r) | X \sim \mathcal{N}(\mu_{X(1+r)}X, \sigma^2_{X(1+r)}X)$ with

$$\mu_{X(1+r)}X = \mu + (X(1) - \mu)e^{-r\gamma} \quad \text{and} \quad \sigma^2_{X(1+r)}X = (1 - e^{-2r\gamma})/2\gamma,$$

so that the conditional quantile is equal to $\mu_{X(1+r)}X + \sigma_{X(1+r)}X z_\alpha$, where $P(Z < z_\alpha) = \alpha$. It is not possible to use this expression as an estimator, however, since $\mu_{X(1+r)}X$ and $\sigma_{X(1+r)}X$ depend on the unknown parameters $(\mu, \gamma)$. A simply plug-in estimator is given by $\hat{\beta}_{P1}(X) = \hat{\mu}_{X(1+r)}X + \hat{\sigma}_{X(1+r)}X z_\alpha$ which replaces $(\mu, \gamma)$ by $(\hat{\mu}, \hat{\gamma}_{OLS}) = (\int_0^1 X(s)ds, -\frac{1}{2}(\hat{X}(1)^2 - 1)/\int_0^1 \hat{X}(s)^2ds)$, $\hat{X}(s) = X(s) - \hat{\mu}$.

In order to cast this problem in the framework of this paper, let $\xi_1 = (X(1) - \mu)\sqrt{2\gamma}$ and $\xi_2 = (X(1 + r) - \mu_{X(1+r)}X)/\sigma_{X(1+r)}X$, so that $\xi = (\xi_1, \xi_2) \sim \mathcal{N}(0, I_2)$. We treat $\xi$ as fixed and part of the parameter, $\theta = (\xi, \mu, \gamma)$. In order to recover the stochastic properties of $\xi$, we integrate over $\xi \sim \mathcal{N}(0, I_2)$ 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 $\mathcal{N}(0, I_2)$.

Now in this notation, $X(1 + r)$ becomes non-stochastic and equals

$$h(\theta) = \mu + \frac{e^{-r\gamma}\xi_1 + \sqrt{1 - e^{-2r\gamma}}\xi_2}{\sqrt{2\gamma}}$$

and the density $f_\theta(x)$ depends on $\xi_1$ (but not $\xi_2$). Furthermore, the quantile unbiased constraint (27) is equivalent to weighted average bias over $\xi \sim \mathcal{N}(0, I_2)$ with $c(\eta, \theta)$ equal to $c(\eta, \theta) = 1[h(\theta) < \eta] - \alpha$. We use the usual quantile check function to measure the loss of estimation errors, $\ell(\eta, \theta) = |h(\theta) - \eta| \cdot |\alpha - 1[h(\theta) < \eta]|$. 

30
Figure 9: 5% Quantile Bias and Normalized Expected 5% Quantile Check Function Loss of Forecasts in Local-to-Unity AR(1) for 20% Horizon

With these choices, the problem is invariant to the group of translations \( g(x, a) = x + a \) and \( \tilde{g}(\theta, a) = (\xi, \mu + a, \gamma) \) for \( a \in \mathbb{R} \), with \( \tilde{g}(\eta, a) = \eta + a \). One set of maximal invariants is given by \( X^* = M(X) = X(s) - X(1) \) and \( \theta^* = \tilde{M}(\theta) = (\xi, 0, \gamma) \), and the density of \( X^* \) relative to the measure of \( W(\cdot) - W(1) \) is given by (cf. equation (25) in Elliott and Müller (2006b))

\[
\begin{align*}
    f_{\theta^*}(x^*) &= \exp\left[-\frac{1}{2}\gamma(x^*(0))^2 - 1\right] - \frac{1}{2}\gamma^2\int_0^1 x^*(s)^2 ds - \xi_1\sqrt{\frac{1}{2}\gamma} \left(\frac{1}{\sqrt{2}}\gamma \int_0^1 x^*(s) ds + x^*(0)\right) - \frac{1}{4}\xi_1^2 \gamma. 
\end{align*}
\] (28)

A calculation shows that with \( \gamma \) fixed and \( \xi \sim \mathcal{N}(0, I_2) \), the distribution of \( h(\theta^*) \) conditional on \( X^* = x^* \) is \( \mathcal{N}(\mu_{h|X^*}, \sigma_{h|X^*}^2) \) with \( \mu_{h|X^*} = (\gamma \int_0^1 x^*(s) ds + x^*(0))(1 - e^{-\gamma})/(2 + \gamma) \) and \( \sigma_{h|X^*}^2 = (1 - e^{-2\gamma} + \frac{2}{2+\gamma}(1 - e^{-\gamma})^2)/(2\gamma) \). The weighted averages of \( \ell^*(\theta^*, \eta) \) and \( c^*(\theta^*, \eta) \) in (18) and (19) with weighting function \( \xi \sim \mathcal{N}(0, I_2) \) then become

\[
\frac{\int f_{\theta^*}(x^*)c^*(\eta, \theta^*)\phi_{I_2}(\xi) d\xi}{\int f_{\theta^*}(x^*)\phi_{I_2}(\xi) d\xi} = \Phi\left(\frac{\eta - \mu_{h|X^*}}{\sigma_{h|X^*}}\right) - \alpha
\]

\[
\frac{\int f_{\theta^*}(x^*)\ell^*(\eta, \theta^*)\phi_{I_2}(\xi) d\xi}{\int f_{\theta^*}(x^*)\phi_{I_2}(\xi) d\xi} = \phi\left(\frac{\eta - \mu_{h|X^*}}{\sigma_{h|X^*}}\right)\sigma_{h|X^*} + (\eta - \mu_{h|X^*})\Phi\left(\frac{\eta - \mu_{h|X^*}}{\sigma_{h|X^*}}\right) - \alpha
\]

where \( \Phi, \phi \) and \( \phi_{I_2} \) are the cdf and pdfs of a scalar and bivariate standard normal, respectively.
With these expressions at hand, it becomes straightforward to implement the algorithm of Section 2.3. We normalize risk by 
\[ \varepsilon_B = \frac{1}{10}(2 + \gamma)^{-1/2}, \]
and in addition to the integration over \( \xi \sim N(0, I_2) \), choose \( \gamma \) uniform on \([0, 80]\) in \( F_n \), and let \( G_i \) equal to point masses on \( \gamma \in \{0, 0.2^2, \ldots, 10^2\} \). For \( \hat{\gamma}_{OLS} > \kappa_X = 60 \), we switch to an estimator of the unconditional quantile \( \hat{\delta}_{S}(x^*) = \hat{\mu} + \hat{\gamma}_a/\sqrt{2\hat{\gamma}} \), with \( \hat{\mu} = \int_0^1 x^*(s)ds \) and \( \hat{\gamma} = \hat{\gamma}_{OLS} - 3 \) and \( \hat{\gamma}_a = z_a \sqrt{(2\hat{\gamma} + 4)/(2\hat{\gamma} - z_a^2)} \) (the use of the approximately median unbiased estimator \( \hat{\gamma} \) and the replacement of \( z_a \) by \( \hat{\gamma}_a \) help correct for the estimation error in \( (\hat{\mu}, \hat{\gamma}) \); see the Appendix D.4 for details).

Figure 9 plots the results for \( r = 0.2 \) 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 (27) for small \( \gamma \): instead of the nominal 5%, \( X(1 + r) \) takes on a value smaller than the quantile estimator about 16% of the time.

6 Conclusion

The literature contains numerous suggestions for estimators in non-standard problems, and their evaluation typically considers bias and risk. For a recent example in the context of estimating a local-to-unity AR(1) coefficient, see, for instance, Phillips (2012).

We argue here that the problem of identifying estimators with low bias and risk may be fruitfully approached in a systematic manner. We suggest a numerical approximation technique to the resulting programming problem, 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.

The general approach might well be useful also for cross sectional non-standard problems, such as those arising under weak instruments asymptotics. In ongoing work, we consider an estimator of a very small quantile \( \alpha \) that has the property that in repeated applications, an independent draw from the same underlying distribution realizes to be smaller than the estimator with probability \( \alpha \). This is qualitatively similar to the forecasting problem considered in Section 5.3, but concerns a limiting problem of observing a finite-dimensional random vector with joint extreme value distribution.

\[ ^4\text{Andrews and Armstrong (2015) derive a mean unbiased estimator of the structural parameter under a known sign of the first stage in this problem.} \]
In all the examples we considered in this paper, it turned out that nearly unbiased estimators with relatively low risk exist. This may come as something of a surprise, and may well fail to be the case in other non-standard problems. For instance, as recently highlighted by Hirano and Porter (2012), mean unbiased estimation is not possible for the minimum of the means in a bivariate normal problem. Unreported results show that an application of our numerical approach to this problem yields very large lower bounds on weighted average risk for any estimator with moderately small mean bias. The conclusion then is that even insisting on small (but not necessarily zero) mean bias is not a fruitful constraint in that particular problem.

References


A Proofs

The proof of Lemma 2 uses the following Lemma.

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

Proof: Replacing $x$ by $g(x,a)$ in (12) yields

$$g(x,a) = g(M(g(x,a)), O(g(x,a)))$$

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

Alternatively, applying $a$ on both sides of (12) yields

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

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: Suppose $\theta_1$ and $\theta_2$ are such that $M(\theta_1) = M(\theta_2)$. Then there exists $a \in A$ such that $\theta_2 = \tilde{g}(\theta_1, a)$. Thus, for an arbitrary measurable subset $B \subset X \times \Theta$,

$$P_{\theta}(\{(M(X), \tilde{g}(\theta_2, O(X)^{-})) \in B\})$$

$$= P_{\tilde{g}(\theta_1, a)}((M(X), \tilde{g}(\theta_2, O(X)^{-})) \in B)$$

$$= P_{\theta_1}((M(g(X,a), \tilde{g}(\theta_2, O(g(X,a))^{-})) \in B)$$

(invariance of problem)

$$= P_{\theta_1}((M(X), \tilde{g}(\theta_2, O(X)^{-} \circ a^{-})) \in B)$$

(by $\theta_2 = \tilde{g}(\theta_1, a)$)

$$= P_{\theta_1}((M(X), \tilde{g}(\theta_1, O(X)^{-})) \in B). ■$$

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 $(\lambda^{1}, \lambda^{u})$ 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 $f_p(x) = (24 + m_p - 2)^{-1}(12f_{\theta_{p,1}}(x) + 12f_{\theta_{p,m}}(x) + \sum_{i=2}^{m_p-1} f_{\theta_{p,i}}(x))$ for some application-specific grid of $m_p$ parameter values $\{\theta_{p,i}\}_{i=1}^{m_p}$ (that is, under $f_p$, $X$ is generated by first drawing an index $J$ uniformly from $\{-10, -9, \ldots, m_p + 11\}$, then draw $X$ from $f_{\theta_{p,J^*}}$, where $J^* = \max(\min(J,1), m_p)$).

The overweighting of the boundary values $\theta_{p,1}$ and $\theta_{p,m}$ by a factor of 12 counteracts the lack of approximately constant 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

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

39
We further approximate $\int c(\eta, \theta)f_\theta(X_i)dG_i(\theta)$ for arbitrary $\eta$ by quadratic interpolation based on the closest three points in the grid $\mathcal{H}_i = \{\eta_{i,1}, \ldots, \eta_{i,m}\}$, $\eta_{i,i} < \eta_{i,j}$ for $i < j$ (we use the same grid for all $i = 1, \ldots, m$). The grid is chosen large enough so that $\delta_\lambda(X_i)$ is never artificially constrained for any value of $\lambda$ considered by the algorithm. Similarly, $R(\delta, F)$ is approximated by

$$R(\delta, F) \approx N^{-1} \sum_{i=1}^{N} \frac{\ell(\delta(X_i), \theta)f_\theta(X_i)dF(\theta)}{f_\theta(X_i)}$$

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

Furthermore, for given $\lambda$, the minimizer $\delta_\lambda(X_i)$ of the function $L_1: \mathbb{R} \rightarrow \mathbb{R}$

$$L_1(\eta) = \int \ell(\eta, \theta)f_\theta(X_i)dF(\theta) + \sum_{i=1}^{m} (\lambda_i^u - \lambda_i^l) \int c(\eta, \theta)f_\theta(X_i)dG_i(\theta)$$

is approximated by first obtaining the global minimum over $\eta \in \mathcal{H}_i$, followed by a quadratic approximation of $L_1(\eta)$ around the minimizing $\eta_{i,j}\ast \in \mathcal{H}_i$ based on the three values of $L_1(\eta)$ for $\eta \in \{\eta_{i,j-1}, \eta_{i,j-1}, \eta_{i,j+1}\} \subset \mathcal{H}_i$.

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

1. Generate i.i.d. draws $X_i$, $l = 1, \ldots, N$ with density $f_\theta$.
2. Compute and store $\int c(\eta, \theta)f_\theta(X_i)dG_i(\theta)$ and $\int \ell(\eta, \theta)f_\theta(X_i)dF(\theta)$, $\eta \in \mathcal{H}_i$, $i = 1, \ldots, m$, $l = 1, \ldots, N$.
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 $\delta_\lambda^{(k)}(X_i)$ 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 set $\lambda^\dagger = \lambda^{(K)}$ for $K = 300$. For Step 3 of the algorithm in Section 2.3, we first initialize and apply the above iterations 300 times for $e = \varepsilon_B$. We then continue to iterate the above algorithm for another 400 iterations, but
every 100 iterations increase or decrease $\epsilon$ via a simple bisection method based on whether or not $R(\delta_{(k)}, F) < (1 + \varepsilon_R)\mu$. After these 4 bisections for $\epsilon$, we continue to iterate another 300 times, for a total of $K = 1000$ iterations. The check of whether the resulting $\hat{\delta}^* = \delta_{(K)}$ 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 $\theta \in \Theta_g \subset \Theta$.

If the estimator is constrained to be of the switching form (10), the above algorithm is unchanged, except that $\delta_{(k)}(X_l)$ in Step 4.a is predetermined to equal $\delta_{(k)}(X_l) = \delta_S(X_l)$ for all draws $X_l$ with $\chi(X_l) = 1$.

We set the number of draws to $N = 250,000$ in all applications. Computations take no more than minutes on a modern PC, with a substantial fraction of time spent in Step 2.

C Details on Algorithms of Section 3

The matrix $\Omega$ and the vector $\omega$ are determined by the same importance sampling Monte Carlo approach as described in the previous section. Depending on the application, this takes seconds to at most a few minutes. Once $\Omega$ and $\omega$ are computed, the solution of the quadratic program is immediate for all practical purposes.

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

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

with $m_{\delta,\lambda}^*(\theta)$ 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 $m_{\delta,\lambda}^{-1}$ is obtained by linear interpolation between these points.

D Additional Application-Specific Details

D.1 AR(1) Coefficient

The process $X$ is approximated by a discrete time Gaussian AR(1) with $T = 2500$ observations. We use $\theta_{p,i} = (i-1)^2, i = 1, \ldots, 61$, $\Theta_g = \{(12j/500)^2\}_{j=0}^{500}$, and $H_i = \{(1/5(j-1)^2 + 4^{-j}(i+1)^2)_{j=0}^{49}\}_{i=0} \cup \{144\}$. 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 $G_i$. 

41
D.2 Degree of Parameter Time Variation

The process $X$ is approximated by partial sums of a discrete time MA(1) with $T = 800$ observations. We set $H_t = \{(i/5)^2 + 4j(i+1)^2\}_{j=0}^{40} \cup \{64\}$. After invariance, the effective parameter is $\eta$, and for $\eta$, we use the grids $\{(i/5)^2\}_{i=1}^{250}$ as proposal and $\{(j/50)^2\}_{j=0}^{500}$ to check the uniform bias property. 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 $G_i$.

D.3 Predictive Regression

The process $X_2^*$ is approximated by a discrete time Gaussian AR(1) with $T = 2500$ observations. We set $H_t = \{\int_0^1 X_{2,j}^*(s)^2ds^{-1/2}\}_{j=0}^{80}$ (corresponding to a range of the “correction term” in (25) of up to $\pm 8$ OLS standard deviations of $\hat{\gamma}_{OLS}$). After invariance and integration over $\xi \sim N(0, I_2)$, the effective parameter is $\gamma$, and for $\gamma$, we use the grids $\{(i/2)^2\}_{i=0}^{250}$ as proposal and $\{(12j/500)^2\}_{j=0}^{500}$ to check the uniform bias property.

To derive (26), note that under $\theta^*$, $X_1(s) = rW_z(s) + \sqrt{1-r^2}W_y(s)$ and $X_2^*$ is independent of $W_y$. Further $dX_2(s) = -\gamma X_2^*(s)ds + dW_z(s)$, so that $W_z(s) = X_2^*(s) + \gamma \int_0^s X_2^*(l)dl$. Thus,

\[
E[X_1(s)|X_2^*] = rX_2^*(s) + r\gamma \int_0^s X_2^*(l)dl
\]

\[
E[X_1(1)|X_2^*] = rX_2^*(1) + r\gamma \int_0^1 X_2^*(l)dl = rX_2^*(1)
\]

\[
E[\hat{\gamma}_{OLS}(X)|X_2^*] = \frac{r\int_0^1 X_2^*(l)dl}{\int_0^1 X_2^*(l)^2dl} + r\gamma
\]

so from $X_1^*(s) = X_1(s) - sX_1(1) - \hat{\gamma}_{OLS}(X) \int_0^s X_2^*(l)dl$,

\[
E[X_1^*(s)|X_2^*] = rX_2^*(s) - rsX_2^*(1) - r\frac{\int_0^1 X_2^*(l)dl}{\int_0^1 X_2^*(l)^2dl} \int_0^s X_2^*(l)dl.
\]

Furthermore, the conditional covariance between $X_1^*(s)$ and $\hat{\gamma}_{OLS}(X)$ given $X_2^*$ is

\[
E[(\hat{\gamma}_{OLS}(X) - E[\hat{\gamma}_{OLS}(X)|X_2^*])(X_1^*(s) - E[X_1^*(s)|X_2^*])|X_2^*]
\]

\[
= (1-r^2) \left[ \frac{\int_0^s X_2^*(l)dl}{(\int_0^1 X_2^*(l)^2dl)^2} \int_0^s X_2^*(l)dl - s \frac{\int_0^1 X_2^*(l)dl}{(\int_0^1 X_2^*(l)^2dl)^2} \int_0^1 X_2^*(l)dl \right] = 0
\]

so that by Gaussianity, $\hat{\gamma}_{OLS}(X)$ is independent of $X_1^*$ conditional on $X_2^*$. Thus, $E[\hat{\gamma}_{OLS}(X)|X^*] = E[\hat{\gamma}_{OLS}(X)|X_2^*]$, and (26) follows.

We now show that the conditional density of $X_1^*$ given $X_2^*$ does not depend on $\theta^*$. Clearly, $X_1^*$ is Gaussian conditional on $X_2^*$, so the conditional distribution of $X_1^*$ given $X_2^*$ is fully described by the conditional mean and conditional covariance kernel. For $0 \leq s_1 \leq s_2 \leq 1$, we obtain

\[
E[(X_1^*(s_1) - E[X_1^*(s_1)|X_2^*])(X_1^*(s_2) - E[X_1^*(s_2)|X_2^*])|X_2^*]
\]
Neither this expression, nor $E[X^*_1(s)|X^*_2]$ computed above, depend on $\theta^*$, which proves the claim.

D.4 Quantile Long-range Forecasts

The process $X^*$ is approximated by a discrete time Gaussian AR(1) with $T = 2500$ observations. We set $H_l = \{ (\hat{\mu}_{X(1+r)|X,l} + \hat{\sigma}_{X(1+r)|X,l}^2 \alpha + \frac{1}{12})\hat{\sigma}_{X(1+r)|X,l} \}_{i=0}^{80}$, where $\hat{\mu}_{X(1+r)|X,l}$ and $\hat{\sigma}_{X(1+r)|X,l}$ are estimators of $\mu_{X(1+r)|X}$ and $\sigma_{X(1+r)|X}$ that estimate $(\mu, \gamma)$ via $(\int_0^1 X_l(s)ds, \max(-\frac{1}{2}(X_l(1)^2 - 1)/\int_0^1 \tilde{X}_l(s)^2ds, 1))$, $\tilde{X}_l(s) = X_l(s) - \int_0^1 X_l(s)ds$. For $\gamma$, we use the grids $\{(i/12)^2\}_{i=0}^{500}$ as proposal and $\{(12j/500)^2\}_{j=0}^{500}$ to check the uniform bias property.

To further motivate the form of $\delta_S$, observe that for large $\gamma$, the unconditional distribution of a mean-zero Ornstein-Uhlenbeck process with mean reverting parameter $\gamma$ is $\mathcal{N}(0, (2\gamma)^{-1})$, suggesting $z_\alpha(2\gamma)^{-1/2}$ as the estimator of the $\alpha$ unconditional quantile. Recall from the discussion in Section 2.1 that for large $\gamma$, $\hat{\gamma}$ is approximately distributed $\tilde{\gamma} \overset{d}{\sim} \mathcal{N}(\gamma, 2\gamma)$. Further, $\hat{\mu} \overset{d}{\sim} \mathcal{N}(\mu, \gamma^{-2})$ independent of $\hat{\gamma}$. Thus, from a first order Taylor expansion,

$$P(\hat{\mu} + \tilde{z}_\alpha(2\gamma)^{-1/2} < \mu + (2\gamma)^{-1/2}Z) \approx P(\tilde{z}_\alpha(2\gamma)^{-1/2} < ((2\gamma)^{-1} + \gamma^{-2} + \tilde{z}_\alpha^2(2\gamma)^{-2})^{1/2}Z) = \alpha$$

which is solved by $\hat{z}_\alpha$ given in the main text.