Unbiased Estimation as a Public Good

David M. Kaplan*

September 17, 2019

Abstract

Bias and variance help measure how bad (or good) an estimator is. When considering a single estimate, minimizing variance plus squared bias (i.e., mean squared error) is optimal in a certain sense. Sometimes a smoothing parameter is explicitly chosen to produce such an optimal estimator. However, important parameters in economics are often estimated multiple times, in many studies over many years, collectively contributing to a public body of evidence. From this perspective, the bias of each single estimate is relatively more important, even if mean squared error minimization remains the goal. This suggests some tension between the single best estimate a paper can report and the estimate that contributes most to the public good. Simulations compare instrumental variables and linear regression, as well as different levels of smoothing for instrumental variables quantile regression.

JEL classification: C44, C52

Keywords: bias, mean squared error, meta-analysis, optimal estimation, science

1 Introduction

Mean squared error (MSE) is the most common way to measure optimality of an econometric estimator. MSE equals an estimator's variance plus the square of its bias. From a decision-theoretic view, given a quadratic loss function, minimizing MSE is equivalent to minimizing expected loss (i.e., risk). Although other loss functions and/or Bayesian frameworks could be studied, MSE is the focus here. Usually an asymptotic approximation of MSE is used; the ideas below remain the same.

Given the MSE criterion, increased bias can be good if it corresponds to a big enough decrease in variance, decreasing overall MSE. This strategy to reduce MSE is often suggested by theoretical statisticians and econometricians (including me in Kaplan and Sun, 2017). For example, this idea underlies the shrinkage, empirical Bayes, and averaging approaches (e.g., Cheng, Liao, and Shi, 2019; DiTraglia, 2016; Hansen, 2017; James and Stein, 1961; Stein,

*Email: kaplandm@missouri.edu. Mail: Department of Economics, University of Missouri, 118 Professional Bldg, 909 University Ave, Columbia, MO 65211-6040, United States.

1956). Additionally, this bias–variance tradeoff is unavoidable in nonparametric estimation, where the smoothing parameter (like a bandwidth) is usually chosen to minimize MSE. Even more basically, the MSE criterion can judge between two different methods. For example, for linear regression with endogeneity, ordinary least squares (OLS) has larger bias but smaller variance than instrumental variables (IV) regression in many cases.[1] Although OLS is not consistent asymptotically, it may be "better" (smaller MSE) in finite samples if the variance difference exceeds the squared bias difference.

However, MSE minimization only considers a single estimate in isolation, not the scientific process in which different researchers produce different estimates (of the same parameter) that may eventually be averaged together, either formally or informally.[2] In the latter case, bias is relatively more important, so the optimal estimator should have relatively less bias and more variance. Intuitively, having multiple studies from multiple datasets is like a larger sample size, which means lower variance. Details are given later.

A qualitatively similar point has been made in another setting. Most notably, Goldstein and Messer (1992) study optimal estimation of functionals (i.e., functions of functions) given an underlying nonparametric kernel estimator $\hat{f}(\cdot)$ of function $f(\cdot)$. The functional estimator might average the nonparametric estimates at the n observations, like $n^{-1}\sum_{i=1}^{n}\hat{f}(X_i)$. They find, "For some classes of functionals, $\hat{f}$ is [ideally] undersmoothed relative to what would be used to estimate f optimally" (p. 1306), shown formally in Theorems 4.1 and 4.2 (p. 1317). "Optimally" means "minimum MSE," so "undersmoothed" means less smoothing and thus less bias than the MSE-optimal amount of bias.

The same idea of optimal undersmoothing applies to divide-and-conquer approaches to big data. However, with samples large enough to warrant divide-and-conquer, estimation precision is usually not an issue, so the practical importance seems small. See Kaplan (2019).

The results here do not apply universally to all types of scientific learning processes. For example, they do not seem to apply (even qualitatively) to a pure stepwise Bayesian process. As another possibility, Frankel and Kasy (2018) write (*emphasis* added), "In a world without constraints, the first-best rule would be for all results – or *even better, all raw data* – to be published." I do not consider optimal estimation with many datasets. However, if the

[1]There are some technical caveats to this example: Kinal (1980) shows cases where the IV estimator does not even have a well-defined mean (and thus bias and MSE), and Hirano and Porter (2015) show finite-sample unbiased linear IV estimators do not exist with an unrestricted parameter space; however, Andrews and Armstrong (2017) restrict the parameter space to a known first-stage sign to propose an unbiased IV estimator. In other cases, like with a binary instrument and binary endogenous regressor, the finite-sample IV bias is well-defined (conditional on the estimator itself being well-defined) but not zero.

[2]In economics, this is less relevant for the treatment effect approach; e.g., Heckman and Vytlacil (2007, p. 4788) write, "Knowledge does not cumulate across treatment effect studies whereas it accumulates across studies estimating common behavioral or technological parameters."

meta-analysis were (for convenience) constrained to averaging estimates from the different datasets, then my results would still apply.

Sections 2 and 3 present and discuss theoretical results, and Section 4 shows simulated examples. Abbreviations include those for mean squared error (MSE), approximate mean squared error (AMSE), ordinary least squares (OLS), instrumental variables (IV), instrumental variables quantile regression (IVQR), meta-analysis (MA), probability density function (PDF), and data-generating process (DGP).

2 Choice of Unbiased or Biased Estimates

The main point of this section is that an unbiased estimator may be preferred when aggregating multiple studies' estimates even if the biased estimator has lower MSE (for a single estimate). Some simplified setups are presented to capture this phenomenon. Results are derived and discussed qualitatively.

2.1 Averaging many identical estimators

2.1.1 Formal results

Imagine J different studies estimating the same parameter θ, each using one of the following estimators. For $j = 1, \dots, J$,

$$\mathrm{Bias}(\hat{\theta}_j^u) = 0, \quad V_u \equiv \mathrm{Var}(\hat{\theta}_j^u), \quad B \equiv \mathrm{Bias}(\hat{\theta}_j^b), \quad V_b \equiv \mathrm{Var}(\hat{\theta}_j^b), \tag{1}$$

where $\mathrm{Bias}(\hat{\theta}) \equiv \mathrm{E}(\hat{\theta}) - \theta$, and superscripts u and b stand for "unbiased" and "biased." Also,

$$\mathrm{Cov}(\hat{\theta}_j^u, \hat{\theta}_k^u) = \mathrm{Cov}(\hat{\theta}_j^b, \hat{\theta}_k^b) = 0, \quad \text{for any } j \neq k. \tag{2}$$

For example, the J estimates may all come from independently sampled datasets.

Two overall meta-analysis (MA) estimators are considered. Either all J unbiased estimates are averaged, or all J biased estimates are:

$$\bar{\hat{\theta}}_J^u \equiv \frac{1}{J} \sum_{j=1}^{J} \hat{\theta}_j^u, \quad \bar{\hat{\theta}}_J^b \equiv \frac{1}{J} \sum_{j=1}^{J} \hat{\theta}_j^b. \tag{3}$$

The estimator with lower MSE is desired.

The bias, variance, and MSE of the two MA estimators can be derived from the properties of the individual estimators. For the bias,

$$\begin{aligned}
\mathrm{Bias}(\bar{\hat{\theta}}^u_J) &= \mathrm{E}\left[\frac{1}{J}\sum_{j=1}^{J}\hat{\theta}^u_j\right] - \theta = \frac{1}{J}\sum_{j=1}^{J}\overbrace{[\mathrm{E}(\hat{\theta}^u_j) - \theta]}^{=0 \text{ by } (1)} = 0, \\
\mathrm{Bias}(\bar{\hat{\theta}}^b_J) &= \mathrm{E}\left[\frac{1}{J}\sum_{j=1}^{J}\hat{\theta}^b_j\right] - \theta = \frac{1}{J}\sum_{j=1}^{J}\overbrace{[\mathrm{E}(\hat{\theta}^b_j) - \theta]}^{=B \text{ by } (1)} = B.
\end{aligned} \tag{4}$$

For the variance and MSE, applying the same formulas for $\bar{\hat{\theta}}^b_J$ as shown for $\bar{\hat{\theta}}^u_J$,

$$\mathrm{Var}(\bar{\hat{\theta}}^u_J) = \frac{1}{J^2}\sum_{j=1}^{J}\sum_{k=1}^{J}\overbrace{\mathrm{Cov}(\hat{\theta}^u_j, \hat{\theta}^u_k)}^{=0 \text{ if } j\neq k} = \frac{1}{J^2}\sum_{j=1}^{J}\overbrace{\mathrm{Var}(\hat{\theta}^u_j)}^{=V_u} = V_u/J, \qquad \mathrm{Var}(\bar{\hat{\theta}}^b_J) = V_b/J, \tag{5}$$

$$\mathrm{MSE}(\bar{\hat{\theta}}^u_J) \equiv [\mathrm{Bias}(\bar{\hat{\theta}}^u_J)]^2 + \mathrm{Var}(\bar{\hat{\theta}}^u_J) = V_u/J, \qquad \mathrm{MSE}(\bar{\hat{\theta}}^b_J) = B^2 + V_b/J. \tag{6}$$

Proposition 1. *Assume* (1)–(3) *hold. (i) Given V_u, V_b, and J,* $\mathrm{MSE}(\bar{\hat{\theta}}^u_J) < \mathrm{MSE}(\bar{\hat{\theta}}^b_J)$ *if and only if $B^2 > (V_u - V_b)/J$. (ii) Given V_u, V_b, and B, with $B \neq 0$,* $\mathrm{MSE}(\bar{\hat{\theta}}^u_J) < \mathrm{MSE}(\bar{\hat{\theta}}^b_J)$ *if and only if $J > J_0 \equiv (V_u - V_b)/B^2$. With $B = 0$, then J is irrelevant:* $\mathrm{MSE}(\bar{\hat{\theta}}^u_J) < \mathrm{MSE}(\bar{\hat{\theta}}^b_J)$ *if and only if $V_u < V_b$.*

Proof of Proposition 1. For (i), using (6), the MSE inequality becomes $V_u/J < V_b/J + B^2$, which is equivalent to $B^2 > (V_u - V_b)/J$. For (ii), further multiply each side by J/B^2 if $B \neq 0$. If instead $B = 0$, then the MSE inequality is $V_u/J < V_b/J$, which is equivalent to $V_u < V_b$ since $J > 0$. □

2.1.2 Discussion and special cases

Proposition 1 looks at the MSE comparison from two perspectives. First, given J estimates being averaged and given the different variances, it shows how much bias there would have to be in order to prefer the unbiased estimator for MA. Second, given the variances and bias, Proposition 1 shows how many estimates would need to be averaged before the unbiased estimator becomes preferred for MA. If there are fewer than J_0 different estimates, then the biased estimator is preferred; if there are more than J_0, then the unbiased estimator is preferred.

Proposition 1 shows the effects of the variance difference and the bias on J_0. First, the larger the variance difference $V_u - V_b$, the larger is J_0. That is, holding bias fixed, if the unbiased estimator has relatively larger variance, then a larger number of estimates (J) is needed. Second, the larger the squared bias B^2, the smaller is J_0. That is, holding variances fixed, if the magnitude of bias increases, then fewer estimates (smaller J) are needed before the unbiased estimator is preferred for MA.

Consider some special cases of Proposition 1. If $V_u = V_b$ and $B \neq 0$, then $J_0 = (V_u - V_b)/B^2 = 0$, so the condition for preferring the unbiased estimator is $J > 0$, which is always true. This is well known: if the variance is the same, then unbiasedness yields lower MSE.

More often, $V_u > V_b$, so there is a bias–variance tradeoff. If $B^2 \to 0$, then $J_0 \to \infty$. That is, if the bias is technically non-zero but practically negligible, then the biased estimator is preferred for MA even with large J.

Proposition 1 characterizes when the biased estimator is better for $J = 1$ but not $J \geq 2$, i.e., $1 < J_0 < 2$. Using the formula for J_0 from Proposition 1,

$$J_0 = 1 \iff B^2 = V_u - V_b, \quad J_0 = 2 \iff B^2 = (V_u - V_b)/2, \tag{7}$$

$$1 < J_0 < 2 \iff \frac{V_u - V_b}{2} < B^2 < V_u - V_b. \tag{8}$$

Thus, whenever the squared bias is between the variance difference and half the variance difference, the single biased estimator has smaller MSE, but the unbiased estimator is preferred for MA even with only two estimates ($J = 2$).

Consider a numerical example of (8). Let $V_u = 2$, $V_b = 1$, $B^2 = 0.6$. MSEs can be computed with (6). If $J = 1$, then the unbiased MSE is $V_u + 0 = 2$, and the biased MSE is $V_b + B^2 = 1.6$, so the biased estimator is preferred. If $J = 2$, then the unbiased MSE is $V_u/2 = 1$, and the biased MSE is $B^2 + V_b/2 = 1.1$, so the unbiased estimator is preferred. The key is the squared bias $B^2 = 0.6$ is smaller than the variance difference $V_u - V_b = 2 - 1 = 1$, but it is larger than half the variance difference, $(V_u - V_b)/2 = 0.5$. When averaging $J = 2$ estimates, the variance is cut in half, but the bias is unchanged; this makes bias relatively more important, in this case important enough that the unbiased estimator is preferred when $J = 2$ (or more).

2.2 Weighted average with possibly biased estimator

2.2.1 Formal results

Now consider an MA estimator averaging a single unbiased or biased estimator with one other estimator. Imagine the other estimator uses existing data, so it is called the "existing estimator"; the choice is now whether to use the unbiased or biased estimator with a new dataset. The existing estimator could itself be an average of many estimates. If so, then it may make more sense for the MA estimator to put more than 1/2 weight on the existing estimator. For simplicity, there is a fixed weight w, $0 < w < 1$. The existing (subscript 0), new (subscript 1), and weighted average MA (subscript w) estimators are, respectively, $\hat{\theta}_0$, $\hat{\theta}_1^u$ (unbiased) or $\hat{\theta}_1^b$ (biased), and

$$\hat{\theta}_w^u \equiv w\hat{\theta}_0 + (1-w)\hat{\theta}_1^u \quad \text{or} \quad \hat{\theta}_w^b \equiv w\hat{\theta}_0 + (1-w)\hat{\theta}_1^b. \tag{9}$$

Although it is possible that $\text{Bias}(\hat{\theta}_w^u) \neq 0$, the name "unbiased MA estimator" refers to $\hat{\theta}_w^u$, while "biased MA estimator" refers to $\hat{\theta}_w^b$.

The properties of $\hat{\theta}_0$, $\hat{\theta}_1^b$, and $\hat{\theta}_1^u$ are defined as:

$$B_0 \equiv \text{Bias}(\hat{\theta}_0), \qquad B_1 \equiv \text{Bias}(\hat{\theta}_1^b), \qquad 0 = \text{Bias}(\hat{\theta}_1^u), \tag{10}$$

$$V_0 \equiv \text{Var}(\hat{\theta}_0), \qquad V_b \equiv \text{Var}(\hat{\theta}_1^b), \qquad V_u \equiv \text{Var}(\hat{\theta}_1^u), \tag{11}$$

$$\text{MSE}(\hat{\theta}_0) = V_0 + B_0^2, \qquad \text{MSE}(\hat{\theta}_1^b) = V_b + B_1^2, \qquad \text{MSE}(\hat{\theta}_1^u) = V_u. \tag{12}$$

It is also assumed that the existing estimator is independent of the new estimators (e.g., the new data was independently sampled), or more weakly that

$$\text{Cov}(\hat{\theta}_0, \hat{\theta}_1^u) = \text{Cov}(\hat{\theta}_0, \hat{\theta}_1^b) = 0. \tag{13}$$

The properties of $\hat{\theta}_w^u$ and $\hat{\theta}_w^b$ follow from (9)–(13). For the bias,

$$\text{Bias}(\hat{\theta}_w^b) \equiv \text{E}(\hat{\theta}_w^b) - \theta = \text{E}[w\hat{\theta}_0 + (1-w)\hat{\theta}_1^b] - w\theta - (1-w)\theta = wB_0 + (1-w)B_1, \tag{14}$$

and similarly

$$\text{Bias}(\hat{\theta}_w^u) = wB_0 + (1-w)(0) = wB_0. \tag{15}$$

For the variance, using (13),

$$\text{Var}(\hat{\theta}_w^u) = w^2V_0 + (1-w)^2V_u + \overbrace{2\,\text{Cov}(\hat{\theta}_0, \hat{\theta}_1^u)}^{=0}, \quad \text{Var}(\hat{\theta}_w^b) = w^2V_0 + (1-w)^2V_b. \tag{16}$$

Thus, the MSEs are

$$\text{MSE}(\hat{\theta}_w^u) = \overbrace{w^2B_0^2}^{\text{from (15), squared}} + \overbrace{w^2V_0 + (1-w)^2V_u}^{\text{from (16)}} = w^2\,\text{MSE}(\hat{\theta}_0) + (1-w)^2\,\text{MSE}(\hat{\theta}_1^u), \tag{17}$$

$$\begin{aligned} \text{MSE}(\hat{\theta}_w^b) &= \overbrace{[wB_0 + (1-w)B_1]^2]}^{=[\text{Bias}(\hat{\theta}_w^b)]^2 \text{ from (14)}} + \overbrace{w^2V_0 + (1-w)^2V_b}^{\text{from (16)}} \\ &= w^2B_0^2 + (1-w)^2B_1^2 + 2w(1-w)B_0B_1 + w^2V_0 + (1-w)V_b \\ &= w^2\,\text{MSE}(\hat{\theta}_0) + (1-w)^2\,\text{MSE}(\hat{\theta}_1^b) + 2w(1-w)B_0B_1. \end{aligned} \tag{18}$$

Continuing the idea that bias is relatively more important when averaging multiple estimators, there is an extra bias interaction term in (18) compared to (17). That is, when the unbiased $\hat{\theta}_1^u$ is used, regardless of the bias of $\hat{\theta}_0$, the weighted average MA estimator's MSE is simply a linear combination of the individual $\mathrm{MSE}(\hat{\theta}_0)$ and $\mathrm{MSE}(\hat{\theta}_1^u)$, weighted by w^2 and $(1-w)^2$, respectively. In contrast, when the biased $\hat{\theta}_1^b$ is used, the weighted average MA estimator's MSE is the same combination of individual MSEs *plus* $2w(1-w)B_0B_1$.

Although in principle the interaction term could be good for MSE, realistically it is probably bad (positive). If $B_0 = 0$ (the existing estimator is unbiased), then the interaction disappears anyway. If $B_0B_1 < 0$, then the interaction is actually good (reducing MSE). This is intuitive: if the existing estimator has positive bias, and the new estimator has negative bias, then the biases partly cancel each other out. However, in practice, $B_0B_1 > 0$ seems more likely. For example, if the existing and new estimators both have omitted variable bias, that bias probably has the same sign in each case, so $B_0B_1 > 0$. Then $2w(1-w)B_0B_1 > 0$, so the effect is bad (higher MSE).

From (17) and (18), it can be seen whether the unbiased or biased MA estimator is preferred, given the values of bias and variance and the weight. Specifically, the unbiased MA estimator is preferred if and only if

$$0 < \mathrm{MSE}(\hat{\theta}_w^b) - \mathrm{MSE}(\hat{\theta}_w^u) = (1-w)^2[\mathrm{MSE}(\hat{\theta}_1^b) - \mathrm{MSE}(\hat{\theta}_1^u)] + 2w(1-w)B_0B_1.$$

Since $1 - w > 0$, this is equivalent to

$$0 < (1-w)[\mathrm{MSE}(\hat{\theta}_1^b) - \mathrm{MSE}(\hat{\theta}_1^u)] + 2wB_0B_1. \tag{19}$$

Proposition 2 states when this condition holds in different cases.

Proposition 2. *Assume* (9)–(13) *hold. (i) If* $B_0 = 0$ *or* $B_1 = 0$*, then* $\mathrm{MSE}(\hat{\theta}_w^u) < \mathrm{MSE}(\hat{\theta}_w^b)$ *if and only if* $\mathrm{MSE}(\hat{\theta}_1^u) < \mathrm{MSE}(\hat{\theta}_1^b)$*, for any* w*. (ii) If* $B_0 \neq 0$ *and* $B_1 > 0$*, then given values of* w*,* B_1*,* V_u*, and* V_b*,* $\mathrm{MSE}(\hat{\theta}_w^u) < \mathrm{MSE}(\hat{\theta}_w^b)$ *if and only if*

$$B_0 > \frac{(1-w)[\mathrm{MSE}(\hat{\theta}_1^u) - \mathrm{MSE}(\hat{\theta}_1^b)]}{2wB_1}. \tag{20}$$

If instead $B_1 < 0$*, the the condition is the same but with* $<$ *replacing* $>$*. (iii) In terms of* w*,* $\mathrm{MSE}(\hat{\theta}_w^u) < \mathrm{MSE}(\hat{\theta}_w^b)$ *if and only if*

$$w > \frac{\mathrm{MSE}(\hat{\theta}_1^u) - \mathrm{MSE}(\hat{\theta}_1^b)}{\mathrm{MSE}(\hat{\theta}_1^u) - \mathrm{MSE}(\hat{\theta}_1^b) + 2B_0B_1}, \tag{21}$$

with $<$ *replacing* $>$ *if the denominator is negative.*

Proof of Proposition 2. As shown, (19) follows from (9)–(13).

(i) If $B_0 = 0$, then the bias interaction term disappears, so (19) becomes $0 < (1 - w)[\text{MSE}(\hat{\theta}_1^b) - \text{MSE}(\hat{\theta}_1^u)]$. Dividing by $(1 - w)$ yields the result.

(ii) In (19), B_0 appears only in the bias interaction term, since the $\text{MSE}(\hat{\theta}_0)$ terms (that contain B_0) in (17) and (18) cancel each other out. Thus, from (19), terms can be moved to the other side, and then divided by $2wB_1$. The unbiased MA estimator is preferred iff

$$(1 - w)[\text{MSE}(\hat{\theta}_1^u) - \text{MSE}(\hat{\theta}_1^b)] < 2wB_0B_1. \tag{22}$$

If $B_1 > 0$, then this becomes

$$B_0 > \frac{(1 - w)[\text{MSE}(\hat{\theta}_1^u) - \text{MSE}(\hat{\theta}_1^b)]}{2wB_1}. \tag{23}$$

If $B_1 < 0$, the result is the same but with $<$ replacing $>$.

(iii) Rearranging (19),

$$\begin{aligned}
0 &< [\text{MSE}(\hat{\theta}_1^b) - \text{MSE}(\hat{\theta}_1^u)] + w[\text{MSE}(\hat{\theta}_1^u) - \text{MSE}(\hat{\theta}_1^b)] + 2wB_0B_1, \\
\text{MSE}(\hat{\theta}_1^u) - \text{MSE}(\hat{\theta}_1^b) &< w[\text{MSE}(\hat{\theta}_1^u) - \text{MSE}(\hat{\theta}_1^b) + 2B_0B_1], \\
w &> \frac{\text{MSE}(\hat{\theta}_1^u) - \text{MSE}(\hat{\theta}_1^b)}{\text{MSE}(\hat{\theta}_1^u) - \text{MSE}(\hat{\theta}_1^b) + 2B_0B_1}.
\end{aligned}$$

The $>$ changes to $<$ if the denominator is negative. □

2.2.2 Discussion

In Proposition 2, part (i) states bias conditions under which the choice of the unbiased or biased estimator is unaffected by the MA context. That is, the unbiased estimator $\hat{\theta}_1^u$ is preferred for use in the weighted average MA estimator if and only if the unbiased estimator ($\hat{\theta}_1^u$ itself) has smaller MSE than the biased estimator ($\hat{\theta}_1^b$ itself). Interestingly, $B_1 = 0$ is sufficient but not necessary; it could be that $B_1 \neq 0$ as long as $B_0 = 0$. When $B_0B_1 = 0$, the "myopic" choice based only on $\text{MSE}(\hat{\theta}_1^u) < \text{MSE}(\hat{\theta}_1^b)$ coincides with the optimal input to the weighted average MA estimator.

Part (ii) shows that the unbiased MA estimator is always preferred if B_0 is made large enough (and the same sign as B_1), all else equal. However, "large enough" may be extremely large in some cases. In the following, assume $\text{MSE}(\hat{\theta}_1^u) > \text{MSE}(\hat{\theta}_1^b)$, and let $B_1 > 0$ for simplicity.

First, if w is arbitrarily close to zero, $(1 - w)/w$ is arbitrarily large, in which case B_0 is required to be arbitrarily large before the unbiased MA estimator is preferred; i.e., the biased MA estimator is practically always preferred. In the extreme, if $w = 0$, then the "average" is simply $\hat{\theta}_1^u$ or $\hat{\theta}_1^b$, so the biased MA estimator is preferred (since $\text{MSE}(\hat{\theta}_1^u) > \text{MSE}(\hat{\theta}_1^b)$ is assumed). Enough weight has to be put on the existing estimator to move away from this extreme case in order for the unbiased MA estimator to be preferred.

Second, if B_1 is near zero, then B_0 has to be very large to prefer the unbiased MA estimator. In the extreme with $B_1 = 0$, as in part (i), the biased MA estimator is always preferred (regardless of w). So B_1 has to be far enough from this extreme in order for the unbiased MA estimator to be preferred.

Third, if $\text{MSE}(\hat{\theta}_1^u) - \text{MSE}(\hat{\theta}_1^b)$ is very large, then B_0 must be very large to prefer the unbiased MA estimator. That is, if the individual unbiased estimator is much worse (in terms of MSE) than the individual biased estimator, then there must be substantial bias to outweigh this and prefer the unbiased MA estimator. Letting $w = 1/2$ for simplicity, the condition for preferring the unbiased MA estimator rearranges into $2B_0B_1 > \text{MSE}(\hat{\theta}_1^u) - \text{MSE}(\hat{\theta}_1^b)$, which more directly shows how the bias interaction term must outweigh the individual MSE difference. Phrased from the opposite perspective: even if the individual MSE difference is large, the unbiased estimator can still be better for MA if the bias is large.

Part (iii) shows that the unbiased MA estimator is preferred if the weight is mostly on the existing estimator. In practice, most likely $\text{MSE}(\hat{\theta}_1^u) - \text{MSE}(\hat{\theta}_1^b) > 0$ and $B_0B_1 > 0$. In that case, (21) is of the form $w > a/(a+b)$ with $a, b > 0$, implying $0 < a/(a+b) < 1$. (If $B_0B_1 = 0$, then the condition becomes $w > 1$, meaning the unbiased MA estimator is never preferred since it was assumed $w < 1$; this matches the result from part (i).) Thus, there is always some weight close enough to $w = 1$ such that the unbiased MA estimator is preferred. This is similar in spirit to the result from Proposition 1 that the unbiased MA estimator is preferred when J is large. In that case, the "existing estimator" that averages the first $J-1$ estimates has weight $w = (J-1)/J$, and the "new" Jth estimator has weight $1 - w = 1/J$. If J is large, then $w = (J-1)/J$ is very close to one, so the unbiased estimator is preferred; and this argument can be applied to each estimator number $j = 1, \ldots, J$ in turn. Further, from (21), the unbiased MA estimator is preferred for a larger range of w when the bias interaction term B_0B_1 is large compared to the individual MSE difference.

3 Choice of Smoothing Parameter

The main point of this section is that when averaging multiple estimates together, the optimal amount of smoothing is smaller (hence smaller bias) than when considering a single estimate. In a particular setting, the ratio of the MSE-optimal smoothing for a single estimate to the MSE-optimal smoothing for averaging J estimates is derived under certain assumptions.

3.1 Setting

Consider a smoothed estimator of θ. The reason for smoothing could be nonparametric estimation, e.g., of a regression function or density. Or, the reason could be to improve computation and/or efficiency, e.g., as in the smoothed maximum score estimator (Horowitz, 1992) or smoothed IV quantile regression (Kaplan and Sun, 2017). The amount of smoothing is controlled by the "smoothing parameter," also called a bandwidth in some settings.

Consider continuous smoothing parameter $h > 0$ that affects MSE. Specifically, given sample size n, for all $j = 1, \ldots, J$, the MSE is approximated as a function of h as

$$\mathrm{MSE}(\hat{\theta}_j, h) \approx \mathrm{AMSE}(\hat{\theta}_j, h) = A_n + V_n(h) + [B_n(h)]^2, \tag{24}$$

where A_n is part of the variance but does not depend on h (and often is zero), and V stands for "variance" and B for "bias." That is,

$$\mathrm{Bias}(\hat{\theta}_j, h) \approx B_n(h), \quad \mathrm{Var}(\hat{\theta}_j, h) \approx A_n + V_n(h). \tag{25}$$

The approximation $\approx$ may involve dropping smaller-order remainder terms and/or considering the asymptotic distribution of the estimator. Usually,

$$\text{as } h \to 0 : B_n(h) \to 0,\ V_n(h) \to \infty, \tag{26}$$

$$\text{as } h \to \infty : B_n(h) \to \infty,\ V_n(h) \to 0, \tag{27}$$

so the AMSE-minimizing h^* is finite and strictly positive.

To quantify the difference between the optimal smoothing for a single estimate versus the average of multiple estimates, the following assumptions are made.

Assumption A1. Estimators $\hat{\theta}_j$ for $j = 1, \ldots, J$ are mutually independent and each based on n observations, with the same approximate bias, variance, and MSE as in (24) and (25).

Assumption A2. In (24), $B_n(h) = h^q c_B$ and $V_n(h) = n^{-1}h^{-r}c_V$, where c_B and c_V may depend on the data generating process but not on n or h, and $rc_V > 0$.

Assumption A2 covers many settings. Sometimes AMSE is given for a scaled version of $\hat{\theta}_j$; since scaling by a constant doesn't change the optimum h, it can simply be rescaled to satisfy A2. For example, an r-dimensional nonparametric kernel density estimator with qth-order kernel satisfies A2, with $A_n = 0$ in (24). For the smoothed instrumental variables quantile regression of Kaplan and Sun (2017), after dividing by the sample size, equation (10) satisfies A2 with q the smoothness of the error term's conditional PDF (Assumption 3) and $r = -1$ with $c_V < 0$, so $rc_V > 0$.

Given Assumption A2, the AMSE-optimal smoothing parameter for a single $\hat{\theta}_j$ is in Lemma 3.

Lemma 3. *Given A2, the h that minimizes* $\text{AMSE}(\hat{\theta}_j, h)$ *is*

$$h^* = n^{-1/(2q+r)} \left(\frac{rc_V}{2qc_B^2} \right)^{1/(2q+r)}. \tag{28}$$

Proof of Lemma 3. Since AMSE is convex in h, the minimizer solves the first-order condition:

$$\begin{aligned}
0 = \frac{d}{dh} \text{AMSE}(\hat{\theta}, h) = 2B_n(h)B_n'(h) + V_n'(h) = 2qc_B^2 h^{2q-1} - n^{-1} rc_V h^{-r-1}&, \\
2qc_B^2 h^{2q-1} = n^{-1} rc_V h^{-r-1}&, \\
h^{2q+r} = n^{-1} \frac{rc_V}{2qc_B^2}&, \\
h_n^* = n^{-1/(2q+r)} \left(\frac{rc_V}{2qc_B^2} \right)^{1/(2q+r)}&.
\end{aligned}$$

□

3.2 Results and discussion

Now consider the average of J different estimates. For simplicity, as in A1, these are assumed to all be from samples of the same size, n, and all mutually independent (e.g., from independently sampled datasets). Similar to before, the overall estimator is

$$\bar{\hat{\theta}}_J = \frac{1}{J} \sum_{j=1}^{J} \hat{\theta}_j. \tag{29}$$

Given A1 and A2, the properties of $\bar{\hat{\theta}}_J$ can be derived from those of $\hat{\theta}_j$. For the approximate bias,

$$\text{Bias}(\bar{\hat{\theta}}_J) = \text{E}\left[\frac{1}{J} \sum_{j=1}^{J} \hat{\theta}_j \right] - \theta = \frac{1}{J} \sum_{j=1}^{J} \overbrace{[\text{E}(\hat{\theta}_j) - \theta]}^{\approx B_n(h)} \approx B_n(h). \tag{30}$$

For the approximate variance,

$$\text{Var}(\bar{\hat{\theta}}_J) = \frac{1}{J^2} \sum_{j=1}^{J} \sum_{k=1}^{J} \overbrace{\text{Cov}(\hat{\theta}_j, \hat{\theta}_k)}^{=0 \text{ if } j \neq k} = \frac{1}{J^2} \sum_{j=1}^{J} \overbrace{\text{Var}(\hat{\theta}_j)}^{\approx A_n + V_n(h)} \approx [A_n + V_n(h)]/J. \tag{31}$$

The following formally states the smoothing adjustment to minimize AMSE.

Proposition 4. *Let Assumptions A1 and A2 hold. Consider the estimator in* (29). *If the AMSE-optimal smoothing parameter for $\hat{\theta}_j$ is h^* in Lemma 3, then the AMSE-optimal* *smoothing parameter for* $\bar{\hat{\theta}}_J$ *is* $J^{-1/(2q+r)}h^*$.

Proof of Proposition 4. Given (30) and (31), the bias is unchanged, but the variance is J times smaller, so the AMSE terms that depend on h are now

$$[B_n(h)]^2 + J^{-1}V_n(h) = c_B^2 h^{2q} + J^{-1}n^{-1}h^{-r}c_V. \tag{32}$$

This is identical to the AMSE terms for $\hat{\theta}_j$ except with n replaced by nJ. Thus, the AMSE-minimizing smoothing parameter replaces n with nJ in (28), yielding the result. $\square$

Proposition 4 says bias is relatively more important when averaging a group of estimates than for a single estimate. Essentially, averaging multiple estimates is like having a larger sample size, which decreases the variance without changing the bias. Thus, it becomes more important to reduce bias, by smoothing less.

For example, consider a scalar nonparametric density or regression estimator with a second-order kernel (like Gaussian or Epanechnikov). Then, $q = 2$ and $r = 1$, so $J^{-1/(2q+r)} = J^{-1/5}$, corresponding to the optimal $n^{-1/5}$ optimal bandwidth rate. If $J = 8$, then $J^{-1/5} = 0.66$. If eight studies produce eight estimates of θ that are eventually averaged, then it would be better (in terms of AMSE) to use a bandwidth 2/3 as big as the standard AMSE-optimal bandwidth. This smaller bandwidth helps more fully capture the benefit of averaging multiple estimates.

4 Simulations

The foregoing ideas are illustrated in the following simulations. The simulations concern regression with endogeneity. The estimators used are ordinary least squares (OLS), instrumental variables (IV), and smoothed IV quantile regression (IVQR). Most economists are familiar with the tradeoff between OLS and IV, related to Section 2.1: OLS is biased (due to endogeneity) but has lower variance. Actually, in finite samples, the IV estimator's bias may not even be defined (Kinal, 1980), in which case MSE is also undefined, but MSE is appropriate for the below example with binary instrument and binary endogenous regressor (if datasets with zero treated units are ignored). In addition to comparing OLS and IV, the choice of bandwidth (smoothing parameter) for smoothed IVQR relates to Section 3: generally, smoothing increases bias but reduces variance. The smoothed IVQR estimator was proposed and studied by Kaplan and Sun (2017), who provide code that computes a data-dependent plug-in bandwidth but also allows manual specification of the bandwidth. In the results tables, "$h = \infty$" refers to the IV estimate since the smoothed IVQR slope estimate approaches the usual IV slope estimate as $h \to \infty$ (Kaplan and Sun, 2017, §2.2, pp. 110–111).

The simulations generate data, estimates, and estimator properties as follows. Within each of $R = 1000$ replications, $J = 10$ datasets are sampled iid from a certain DGP (details below). Within each dataset, each estimator is computed (OLS, IV, and IVQR with various bandwidths). Additionally, for each estimator, a corresponding meta-analysis (MA) estimator is computed by averaging the J estimates. The simulated bias of an estimator is the difference between the true parameter and the average of all estimates in all replications (i.e., all RJ of them). (It is equivalent to average the individual estimators or the MA estimators, since the MA estimators themselves are averages.) The variance of the MA estimators is simply the variance of the R MA estimates from the R replications. The variance of the individual estimators is the variance of the RJ individual estimates. The MSE is variance plus squared bias in either case. Due to the possibility of IV not having finite variance in finite samples, nominally the MSE is trimmed (to make it finite) by allowing a maximum squared error $(\hat{\theta}-\theta)^2$ of 100, but the trimming does not actually happen in these simulations, i.e., the squared error is always below 100.

DGP 1 has a single endogenous regressor with slope coefficient θ. The structural model is $Y = \beta_0 + \theta D + 5F^{-1}(U)$, where $\beta_0 = 0$, $\theta = 1$, $U \sim \text{Unif}(0, 1)$, and $F^{-1}(\cdot)$ is the inverse CDF (i.e., quantile function) of a χ^2_2 distribution. The endogenous regressor D is generated as $D = 0.05U + 0.5Z + 0.45V$, where the instrument $Z \sim \text{Unif}(0, 1)$ and also $V \sim \text{Unif}(0, 1)$, with Z, V, U mutually independent. Sampling is iid, with $n = 200$ observations per dataset. The IVQR estimator uses quantile index $\tau = 0.2$ (although here there is no heterogeneity: the slope is a constant $\theta = 1$ for all τ).

Table 1 shows the simulated properties of the various estimators under DGP 1. The column Bias^2 shows the squared bias, which is the same for the individual and MA estimators. The columns Var_{MA} and MSE_{MA} respectively show the variance and MSE for the meta-analysis estimator, while Var_n and MSE_n respectively show the variance and MSE for the estimator given a single dataset of n observations. In the first column, $\hat{h}$ refers to the plug-in bandwidth computed by the code from Kaplan and Sun (2017). The bandwidth $J^{-1/7}\hat{h}$ is the adjustment suggested by Proposition 4, given that Assumptions A1 and A2 hold and the optimal bandwidth rate is $n^{-1/7}$. Additionally, a grid of fixed bandwidths is used, as seen in the remaining rows. For example, in the row for $h = 10$, the bandwidth $h = 10$ is used for every single estimate, whereas $\hat{h}$ differs for each dataset. As noted by Kaplan and Sun (2017, p. 133), the optimal data-dependent bandwidth is always at least as good as the best fixed bandwidth (since "fixed" is a special case of data-dependent), so it is possible for the plug-in bandwidth to outperform the best fixed bandwidth in some cases (as seen below).

Table 1: Simulated properties of estimators, DGP 1.

Estimator	h	Bias^2	Var_{MA}	MSE_{MA}	Var_n	MSE_n
IV	n/a	0.013 028	2.400 217	**2.413 245**	24.930 940	24.943 968
OLS	n/a	10.416 644	1.293 459	11.710 103	13.214 261	**23.630 905**
IVQR	$\hat{h}$	0.000 378	0.405 289	0.405 667	4.429 437	**4.429 814**
IVQR	$J^{-1/7}\hat{h}$	0.000 200	0.403 862	**0.404 062**	4.450 957	4.451 157
IVQR	3	0.000 398	0.456 721	0.457 119	5.028 323	5.028 721
IVQR	4	0.000 410	0.423 006	0.423 416	4.663 993	4.664 403
IVQR	5	0.000 425	0.408 542	0.408 968	4.506 512	4.506 938
IVQR	5.50	0.000 463	0.406 989	**0.407 452**	4.487 934	4.488 397
IVQR	5.55	0.000 469	0.406 991	0.407 459	4.487 735	**4.488 203**
IVQR	5.60	0.000 474	0.407 018	0.407 492	4.487 803	4.488 277
IVQR	6	0.000 525	0.408 086	0.408 612	4.497 079	4.497 605
IVQR	10	0.001 353	0.458 564	0.459 917	4.991 569	4.992 922
IVQR	100	0.011 835	1.502 526	1.514 361	15.521 049	15.532 884
IVQR	1000	0.012 909	2.298 340	2.311 249	23.856 559	23.869 468
IVQR	∞	0.013 028	2.400 217	2.413 245	24.930 940	24.943 968

IV and OLS are the usual IV and OLS estimators.

Table 1 also illustrates Proposition 1. OLS is more biased than IV, but its variance is smaller. For the single dataset estimator, Var_n plays an important role in MSE_n, and in fact MSE_n is smaller for OLS. However, the variance is roughly J times smaller for the MA estimator, so the squared bias is relatively much more important for MSE_{MA}. Consequently, IV has much smaller MSE_{MA} than OLS.

Table 1 further illustrates some ideas from Section 3, as well as showing that the precise theoretical results are messier in practice. First, the table shows that adjusting the plug-in bandwidth by $J^{-1/7}$ makes MSE_n higher but MSE_{MA} lower, although partly due to luck here. As expected, for the single dataset estimator, the smaller (adjusted) bandwidth increases the variance Var_n but decreases the squared bias. Bias is relatively more important for MSE_{MA} than for MSE_n. More importantly here, though, the MA variance Var_{MA} happens to decrease with the adjusted bandwidth, which drives the reduction in MSE_{MA}. This is unexpected, but could be explained by higher-order terms and/or the fact that $\hat{h}$ is not the true MSE-optimal bandwidth, only a plug-in estimate.

Second, Table 1 shows qualitatively the same pattern with the grid of fixed bandwidths, albeit in very small magnitude. The bandwidth $h = 5.55$ produces the smallest MSE_n, whereas the smaller bandwidth $h = 5.50$ minimizes MSE_{MA}. However, the magnitude is very small, and the optimal bandwidth adjustment is not $J^{-1/7}$ here, presumably due to higher-order terms not captured in the AMSE.

DGP 2 is similar to DGP 1, but with a random slope coefficient. The unobserved component is again $U \sim \text{Unif}(0,1)$, and the endogenous regressor is $D = (U + Z)/2$, where $Z \sim \text{Unif}(0,1)$ is the instrument. The observed outcome is $Y = 3 + 3UD$, where $3U$ is the random slope coefficient. Interest is in the slope of the structural τ-quantile function, where $\tau = 0.7$; with $U = \tau$, the slope is $(3)(0.7) = 2.1$. The mean slope $\text{E}(3U) = 1.5$ is identified by the usual (non-quantile) IV approach; e.g., see Lewbel (2019, §5.2). Thus, unlike in DGP 1, the parameter of interest is different for IVQR estimation and for OLS/IV estimation. As a result, bias is relatively large for the smoothed IVQR estimator when h is large, since the smoothed IVQR slope estimate approaches the usual IV slope estimate as $h \to \infty$.

Table 2: Simulated properties of estimators, DGP 2.

Estimator	Bandwidth (h)	Bias^2	Var_{MA}	MSE_{MA}	Var_n	MSE_n
IV	n/a	0.0002	0.0052	**0.0055**	0.056	**0.056**
OLS	n/a	2.2457	0.0013	2.2470	0.013	2.259
IVQR	$\hat{h}$	0.0272	0.0082	0.0354	0.089	0.117
IVQR	$J^{-1/7}\hat{h}$	0.0111	0.0093	**0.0204**	0.101	**0.113**
IVQR	0.01	0.0072	0.0136	0.0208	0.145	0.152
IVQR	0.1	0.0071	0.0132	0.0202	0.142	0.149
IVQR	0.3	0.0070	0.0123	**0.0193**	0.133	0.140
IVQR	0.4	0.0082	0.0114	0.0195	0.123	0.131
IVQR	0.5	0.0110	0.0102	0.0212	0.111	0.122
IVQR	0.6	0.0158	0.0092	0.0250	0.100	0.116
IVQR	0.7	0.0230	0.0084	0.0313	0.092	**0.115**
IVQR	0.8	0.0325	0.0078	0.0403	0.086	0.118
IVQR	1	0.0585	0.0073	0.0658	0.079	0.137
IVQR	10	0.3258	0.0052	0.3310	0.055	0.381
IVQR	∞	0.3780	0.0052	0.3832	0.056	0.434

Table 2 illustrates Section 3 better than Table 1, although the finite-sample approximation error is still apparent. (The IV/OLS comparison is less interesting since the OLS bias dominates even MSE_n here.) First, the $J^{-1/7}$ adjustment to the plug-in bandwidth has the expected effect: smaller bias, but larger variance. Comparing the estimators with adjusted and unadjusted plug-in bandwidth, the MSE_n is very similar (the increased variance nearly cancels out the decreased squared bias from the adjustment), but the MSE_{MA} is much smaller for the adjusted bandwidth. Second, among the fixed bandwidths, $h = 0.7$ minimizes MSE_n, whereas the smaller $h = 0.3$ minimizes MSE_{MA}. Further, the MSE_{MA} for $h = 0.7$ is around 50% larger than the MSE_{MA} for $h = 0.3$; not only is the optimal bandwidth significantly different, but the MSE itself is significantly different. As noted before, according to Proposition 4, the approximately optimal adjustment should be to multiply by $J^{-1/7} = 0.72$, which suggests $h = (0.7)J^{-1/7} = 0.5$ should minimize the (approximate) MSE_{MA}. Again, this adjustment is not fully optimal, since $h = 0.3$ outperforms $h = 0.5$ for MSE_{MA}. However, the MSE_{MA} with $h = 0.5$ is much closer to that with $h = 0.3$ than that with $h = 0.7$, so the adjustment is still reasonable and helpful.

5 Conclusion

Different perspectives lead to different estimators being optimal, even with the same criterion of mean squared error. Specifically, when averaging multiple estimates, each individual estimator should have less bias than when considered in isolation. Thus, when contributing to a broader scientific process, it may help to report a less-biased estimate alongside the MSE-optimal estimate.

Future research could examine the effects of different studies having different sample sizes and estimators. For example, if one study has a particularly large or small sample size, does reducing bias matter more or less (or neither)? The uncertainty about the eventual total number of different studies could also be incorporated. More precise quantification could also be done for definitions of optimality other than mean squared error, e.g., with different loss functions or with posterior expected loss.

References

Andrews, I., Armstrong, T. B., 2017. Unbiased instrumental variables estimation under known first-stage sign. Quantitative Economics 8 (2), 479–503.
URL `https://doi.org/10.3982/QE700`

Cheng, X., Liao, Z., Shi, R., 2019. On uniform asymptotic risk of averaging GMM estimators. Quantitative Economics 10 (3), 931–979.
URL `https://doi.org/10.3982/QE711`

DiTraglia, F. J., 2016. Using invalid instruments on purpose: Focused moment selection and averaging for GMM. Journal of Econometrics 195 (2), 187–208.
URL `https://doi.org/10.1016/j.jeconom.2016.07.006`

Frankel, A., Kasy, M., 2018. Which findings should be published?, working paper available at `https://scholar.harvard.edu/kasy/publications/which-findings-should-be-published`.

Goldstein, L., Messer, K., 1992. Optimal plug-in estimators for nonparametric functional estimation. Annals of Statistics 20 (3), 1306–1328.
URL `https://projecteuclid.org/euclid.aos/1176348770`

Hansen, B. E., 2017. A Stein-like 2SLS estimator. Econometric Reviews 36 (6–9), 840–852. URL https://doi.org/10.1080/07474938.2017.1307579

Heckman, J. J., Vytlacil, E. J., 2007. Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation. In: Heckman, J. J., Leamer, E. E. (Eds.), Handbook of Econometrics. Vol. 6B. Elsevier, Ch. 70, pp. 4779–4874. URL https://doi.org/10.1016/S1573-4412(07)06070-9

Hirano, K., Porter, J. R., 2015. Location properties of point estimators in linear instrumental variables and related models. Econometric Reviews 34 (6-10), 720–733. URL https://doi.org/10.1080/07474938.2014.956573

Horowitz, J. L., 1992. A smoothed maximum score estimator for the binary response model. Econometrica 60 (3), 505–531. URL https://www.jstor.org/stable/2951582

James, W., Stein, C., 1961. Estimation with quadratic loss. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability. Vol. 1. University of California Press, Berkeley, CA, pp. 361–379. URL https://projecteuclid.org/euclid.bsmsp/1200512173

Kaplan, D. M., 2019. Optimal smoothing in divide-and-conquer for big data, working paper available at https://faculty.missouri.edu/~kaplandm.

Kaplan, D. M., Sun, Y., 2017. Smoothed estimating equations for instrumental variables quantile regression. Econometric Theory 33 (1), 105–157. URL https://doi.org/10.1017/S0266466615000407

Kinal, T. W., 1980. The existence of moments of k-class estimators. Econometrica 48 (1), 241–249. URL https://www.jstor.org/stable/1912027

Lewbel, A., 2019. The identification zoo: Meanings of identification in econometrics. Journal of Economic Literature XXX (XXX), XXX–XXX. URL https://www.aeaweb.org/articles?id=10.1257/jel.20181361

Stein, C., 1956. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability. Vol. 1. University of California Press, Berkeley, CA, pp. 197–206. URL https://projecteuclid.org/euclid.bsmsp/1200501656