Optimal Smoothing in Divide-and-Conquer for Big Data

David M. Kaplan*

June 24, 2019

Abstract

With a large dataset ("big data"), it may be very slow or even infeasible to compute an estimator. The popular divide-and-conquer approach computes the estimator on smaller subsets of the data and then averages the subset estimates. This is simple and effective. However, for estimators involving a bandwidth (or other smoothing parameter), the optimal smoothing for an individual subset estimator differs from the optimum for the average. This note highlights that divide-and-conquer subset estimators should use less smoothing and suggests a simple adjustment in practice.

JEL classification: C13, C14

Keywords: bandwidth, bias, mean squared error, optimal smoothing

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. 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 (Kaplan and Sun, 2017, including me in). 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, 1956). Further, in nonparametric estimation, bias is unavoidable; proposed "optimal" smoothing parameters try to minimize MSE.

However, the optimal smoothing for a single estimate differs from the optimal smoothing in a divide-and-conquer procedure. Divide-and-conquer first divides the original large dataset into many subsets. Then, an estimate is computed from each data subset. Finally, the estimates are averaged (or otherwise combined). Intuitively, the averaging reduces variance

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

but not bias. However, the usual optimal smoothing parameters are unaware of this averaging step that reduces variance. Thus, they allow to much bias to reduce variance. That is, the optimal smoothing should be less (to lower bias) in the divide-and-conquer setting.

A qualitatively similar point has been made in the context of averaging many nonparametric estimates within the same dataset. It seems this point was first made by Goldstein and Messer (1992). They study optimal estimation of functionals (i.e., functions of functions) given an underlying nonparametric kernel estimator $\hat{f}(\cdot)$ of function $f(\cdot)$. The population function $f(\cdot)$ is either a regression (conditional mean) function or a probability density function. The functional estimator might average the nonparametric estimates at the n observations, like $n^{-1}\sum_{i=1}^{n}\hat{f}(X_i)$; e.g., the average derivative of a regression function. (More generally, they consider "smooth" functionals, which they contrast with "pointwise" or "atomic" functionals like $f(x)$, the function evaluated at a single point x; see page 1308.) They find, "For some classes of functionals, $\hat{f}$ is [ideally] undersmoothed relative to what would be used to estimate f optimally" (p. 1306). "Optimally" means "minimum MSE," so "undersmoothed" means less smoothing and thus less bias than the MSE-optimal amount of bias. Formally, their Theorems 4.1 and 4.2 (p. 1317) show that the MSE-optimal bandwidth is a smaller order of magnitude when this sort of averaging occurs.

Section 2 describes the idea, illustrated by simulation results in Section 3.

2 Idea

2.1 Individual estimator properties

Consider an estimator $\hat{\theta}_m$, where m is the sample size. Consider continuous smoothing parameter $h > 0$ that affects mean squared error (MSE). Specifically, MSE is approximated as a function of h as

$$\mathrm{MSE}(\hat{\theta}_m, h) \approx \mathrm{AMSE}(\hat{\theta}_m, h) = A_m + [B_m(h)]^2 + V_m(h), \tag{1}$$

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

$$\mathrm{Bias}(\hat{\theta}_m, h) \approx B_m(h), \quad \mathrm{Var}(\hat{\theta}_m, h) \approx A_m + V_m(h). \tag{2}$$

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

$$\text{as } h \to 0 : B_m(h) \to 0,\ V_m(h) \to \infty, \tag{3}$$

$$\text{as } h \to \infty : B_m(h) \to \infty,\ V_m(h) \to 0, \tag{4}$$

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

Usually the AMSE-optimal h can be found from a first-order condition. Most commonly, $B_m(h)$ and $V_m(h)$ are differentiable, and $[B_m(h)]^2 + V_m(h)$ is a convex function of h. Thus, the optimal h_m^* solves the first-order condition

$$0 = \left.\frac{d\,\text{AMSE}(h)}{dh}\right|_{h=h_m^*} = 2B_n(h_m^*)B_n'(h_m^*) + V_n'(h_m^*).$$

More specifically, the following is assumed.

Assumption A1. In (1), $B_m(h) = h^q c_B$ and $V_m(h) = m^{-1}h^{-r}c_V$, where c_B and c_V may depend on the data generating process but not on m or h, and $rc_V > 0$.

Assumption A1 covers many settings. Sometimes AMSE is given for a scaled version of $\hat{\theta}_m$; since scaling by a constant doesn't change the optimum h, it can simply be rescaled to satisfy A1. For example, an r-dimensional nonparametric kernel density estimator with qth-order kernel satisfies A1, with $A_m = 0$ in (1). For the smoothed instrumental variables quantile regression of Kaplan and Sun (2017), after dividing by the sample size, equation (10) satisfies A1 with q the conditional smoothness of the error term's PDF (Assumption 3) and $r = -1$ with $c_V < 0$, so $rc_V > 0$.

Given A1, the AMSE-optimal bandwidth for $\hat{\theta}_m$ is in Lemma 1.

Lemma 1. *Given A1, the h that minimizes* $\text{AMSE}(\hat{\theta}_m, h)$ *is*

$$h_m^* = m^{-1/(2q+r)}\left(\frac{rc_V}{2qc_B^2}\right)^{1/(2q+r)}. \tag{5}$$

Proof of Lemma 1. Since AMSE is convex in h, the minimizer solves the first-order condition:

$$0 = \frac{d}{dh}\text{AMSE}(\hat{\theta}_m, h) = 2B_m(h)B_m'(h) + V_m'(h) = 2qc_B^2 h^{2q-1} - m^{-1}rc_V h^{-r-1},$$

$$\begin{aligned}
2qc_B^2 h^{2q-1} &= m^{-1}rc_V h^{-r-1}, \\
h^{2q+r} &= m^{-1}\frac{rc_V}{2qc_B^2}, \\
h_m^* &= m^{-1/(2q+r)}\left(\frac{rc_V}{2qc_B^2}\right)^{1/(2q+r)}.
\end{aligned}$$

□

2.2 Divide-and-conquer estimator

Now consider the divide-and-conquer estimator. An original sample of n iid observations is divided into s subsets with $m = n/s$ observations each. Subset estimates $\hat{\theta}_m^{(j)}$ are computed for subsets $j = 1, \ldots, s$. The final divide-and-conquer estimator is

$$\hat{\theta}_{m,s}^{\mathrm{DC}} = \frac{1}{s} \sum_{j=1}^{s} \hat{\theta}_m^{(j)}. \tag{6}$$

With iid sampling and randomly chosen subsets, the properties of $\hat{\theta}_{m,s}^{\mathrm{DC}}$ can be derived from those of $\hat{\theta}_m$. In this case, the mean and variance of $\hat{\theta}_m^{(j)}$ do not depend on j; they are the same as those of $\hat{\theta}_m$. Using the linearity of the expectation operator,

$$\mathrm{E}(\hat{\theta}_{m,s}^{\mathrm{DC}}) = \frac{1}{s} \sum_{j=1}^{s} \mathrm{E}(\hat{\theta}_m^{(j)}) = \mathrm{E}(\hat{\theta}_m), \tag{7}$$

so $\mathrm{Bias}(\hat{\theta}_{m,s}^{\mathrm{DC}}) = \mathrm{Bias}(\hat{\theta}_m)$. Also,

$$\mathrm{Var}(\hat{\theta}_{m,s}^{\mathrm{DC}}) = \mathrm{Var}\left(\frac{1}{s} \sum_{j=1}^{s} \hat{\theta}_m^{(j)} \right) = \frac{1}{s} \mathrm{Var}(\hat{\theta}_m). \tag{8}$$

The following formally states the appropriate bandwidth adjustment.

Assumption A2. Sampling is iid, and s subsets of size m are chosen randomly such that $\hat{\theta}_m^{(j)} \perp\!\!\!\perp \hat{\theta}_m^{(k)}$ for all $j \neq k$.

Proposition 2. *Let Assumptions A1 and A2 hold. Consider the divide-and-conquer estimator in* (6)*. If the AMSE-optimal bandwidth for $\hat{\theta}_m$ is h_m^* in Lemma 1, then the AMSE-optimal bandwidth for $\hat{\theta}_{m,s}^{DC}$ is $s^{-1/(2q+r)} h_m^*$.*

Proof of Proposition 2. Given (7) and (8), the bias is unchanged, but the variance is s times smaller, so the AMSE terms that depend on h are now

$$[B_m(h)]^2 + s^{-1} V_m(h) = c_B^2 h^{2q} + s^{-1} m^{-1} h^{-r} c_V. \tag{9}$$

This is identical to the AMSE terms for $\hat{\theta}_m$ except with m replaced by sm. Thus, the AMSE-minimizing bandwidth replaces m with sm in (5), yielding the result. □

Consider the earlier examples. For a second-order, one-dimensional kernel estimator, $q = 2$ and $r = 1$, with the well-known optimal bandwidth rate $h_m^* \propto m^{-1/(2q+r)} = m^{-1/5}$. The divide-and-conquer bandwidth should be $s^{-1/5} h_m^*$; e.g., if $s = 32$, then $h_{\mathrm{DC}}^* = h_m^*/2$. For smoothed instrumental variables quantile regression, the bandwidth from Kaplan and

Sun (2017) has $h_m^* \propto m^{-1/(2q+r)}$ with $r = -1$. They use $q = 4$ in their code, in which case $h_m^* \propto m^{-1/7}$ and $h_{\text{DC}}^* = s^{-1/7} h_m^*$; e.g., if again $s = 32$, then $h_{\text{DC}}^* = 0.61 h_m^*$.

3 Simulation

To illustrate the foregoing ideas, properties of a smoothed instrumental variables quantile regression (IVQR) estimator are simulated. The estimator was proposed and studied by Kaplan and Sun (2017). The idea is to replace the indicator function in the moment conditions with a smoothed version, to improve both computation and MSE. The smoothing adds bias but reduces variance. Their code automatically computes a data-dependent plug-in bandwidth but also allows manual specification of the bandwidth.

The DGP is as follows. Sampling is iid, with $m = 1000$ observations per subset and $s = 32$ subsets per full dataset; 1000 replications are run. There is one endogenous regressor, D, and five exogenous regressors, $X_j \sim \text{N}(0, 1)$ for $j = 1, \ldots, 5$, each X_j independent of all other variables. The unobserved term is $U \sim \text{Unif}(0, 1)$. The outcome is $Y = D\theta(U) + \alpha(U) + \mathbf{X}'\boldsymbol{\gamma}$, where $\boldsymbol{\gamma} = (1, 1, 1, 1, 1)'$, $\alpha(U)$ is a random intercept term equal to the U-quantile of a χ_3^2 distribution, and $\theta(U) = 25U$ is a random coefficient. The excluded instrument is $Z \sim \text{Unif}(0, 1)$. The endogenous regressor is $D = (U + Z)/2$. From the seminal IVQR identification results of Chernozhukov and Hansen (2005), for any $0 < \tau < 1$, the structural parameters $\theta(\tau)$, $\alpha(\tau)$, and $\boldsymbol{\gamma}$ are identified. Here, $\tau = 0.8$, and the focus is on $\theta \equiv \theta(0.8) = 20$.

The estimators and their properties are computed as follows. For each subset of m observations, $\hat{\theta}_m$ is computed. With 1000 replications and $s = 32$ subsets per replication, there are 32,000 such estimates total. The bias is the difference between the average of all those estimates and the true θ. The MSE is the sum of the squared bias and the variance of all those estimates. The divide-and-conquer estimator averages the s different $\hat{\theta}_m$ estimates from the s different subsets, so there are 1000 different divide-and-conquer esitmates. By construction, their bias is identical, but their variance and thus MSE are smaller. Such estimates are computed for a variety of different bandwidths.

Table 1 shows the simulated properties of the individual $\hat{\theta}_m$ estimator and the overall divide-and-conquer (DC) estimator, as indicated by the subscripts in the row headers. In the bandwidth column, $\hat{h}$ indicates the data-dependent plug-in bandwidth automatically computed by the code of Kaplan and Sun (2017). Since the optimal bandwidth rate is $n^{-1/7}$ and Assumptions A1 and A2 hold, the adjusted bandwidth is $s^{-1/7}\hat{h}$. 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

Table 1: Simulated properties of smoothed IVQR estimator.

Bandwidth (h)	$\text{Bias}^2_{\text{DC}}$	Var_{DC}	MSE_{DC}	Var_m	MSE_m
$\hat{h}$	0.065	0.145	0.209	4.693	4.758
$s^{-1/7}\hat{h}$	0.007	0.180	0.187	5.875	5.882
5	0.010	0.183	0.192	5.966	5.976
7	0.016	0.176	0.192	5.760	5.775
9	0.014	0.134	0.148	4.352	4.366
10	0.003	0.113	0.117	3.668	3.671
11	0.106	0.098	0.204	3.169	3.275
12	0.436	0.088	0.524	2.841	3.277
15	3.312	0.076	3.388	2.436	5.747

data subset.

Table 1 shows that adjusting the plug-in bandwidth by $s^{-1/7}$ makes MSE_m higher but MSE_{DC} lower. As expected, the smaller (adjusted) bandwidth increases variance but decreases bias. Bias is relatively more important for MSE_{DC} than for MSE_m. Thus, when adjusting the bandwidth to be smaller, MSE_m increases due to the increased variance, whereas MSE_{DC} decreases due to the decreased bias. That is, the original $\hat{h}$ is better for a single estimate with m observations (which is what it was designed for), whereas the adjusted $s^{-1/7}\hat{h}$ is better for the DC estimator, which is the goal of the adjustment.

Table 1 shows qualitatively the same pattern with the grid of fixed bandwidths. For MSE_m, the best bandwidth seems to be somewhere between $h = 11$ and $h = 12$. Despite having 10 times higher squared bias, the variance is enough lower to minimize MSE. In contrast, any of the bandwidths $5 \leq h \leq 10$ yield smaller MSE_{DC} than $h = 11$ or $h = 12$, since the bias is relatively more important. (The squared bias is not monotonically increasing in h since the bias switches from positive to negative around $h = 10$.) However, the suggested adjustment $s^{-1/7} \approx 0.61$ seems to overcompensate here. If MSE_m is minimized around $h^*_{\text{MSE}} \approx 11.5$, then $s^{-1/7}h^*_{\text{MSE}} = 7$, but in the table MSE_{DC} is smaller at $h = 9$ or $h = 10$. Presumably, this is due to the smaller-order terms ignored by Kaplan and Sun (2017). Although the adjustment is not optimal for this DGP, it still improves MSE_{DC}, which is the goal.

4 Conclusion

As shown, individual estimates within a divide-and-conquer procedure should have less smoothing (and thus less bias, but more variance) than usual. Presuming the criterion of mean squared error, an adjustment has been proposed. Work in progress (Kaplan, 2021) considers the same idea in the context of building scientific knowledge over time. Future work could consider alternative criteria of optimality, like with alternative (non-quadratic) loss functions.

References

Cheng, X., Liao, Z., Shi, R., 2019. On uniform asymptotic risk of averaging GMM estimators. Quantitative Economics XXX (XXX), XXX–XXX.

Chernozhukov, V., Hansen, C., 2005. An IV model of quantile treatment effects. Econometrica 73 (1), 245–261.
URL https://www.jstor.org/stable/3598944

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

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

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., 2021. Unbiased estimation as a public good, 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

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