SUMMER RESEARCH REPORT: OPTIMAL SELECTION OF SIEVE SIZE IN NONPARAMETRIC IV ESTIMATION

DAVID M. KAPLAN

Abstract. Nonparametric estimation allows researchers to estimate models where the functional form is unknown. The "method of sieves" provides theoretical support for consistency of estimation restricted to a finite space of functions (a sieve) that asymptotically grows dense in the full functional space. However, without a good technique for selecting the sieve size, a researcher is just assuming a parametric form and convincing others that it's okay by using tricks. In the endogenous case, critical for economic research, an optimal technique for selecting sieve size is an open question. Here, we present simulation results for a range of data-driven selection criteria. Initial results are discouraging, in that a simple residual sum of squares (RSS) selection criterion seems to outperform methods designed to account for endogeneity and prevent overfitting. There remains an opening for continued work on this project. Section 1 provides the context for our problem; Section 2 poses the problem; Section 3 is a brief review of the literature; Section 4 describes specific methods examined; Section 5 presents simulation results; and Section 6 discusses results and future work.

1. Basic Setting

Consider the nonparametric instrumental variable model:

$$Y = h(X) + u, \quad E(u|Z) = 0 \tag{1}$$

where X is the endogenous variable, Z is the instrumental variable, and $h(\cdot)$ is the unknown function of interest. Our observations are (X_i, Y_i, Z_i) for $i = 1, 2, \ldots, n$.

To estimate the unknown function $h(x)$, we employ the method of sieves as in Blundell et al. (2007). We assume the unknown function $h(\cdot)$ is the unique minimizer of the optimization problem:

$$\min_h E\left\{(E[Y - h(X)|Z])^2\right\}$$

for almost all z in the support of Z. That is, $h(\cdot)$ is the unique solution to the functional equation:

$$E(Y|Z) = E(h(X)|Z).$$

We approximate the unknown function and the conditional expectations using sieve bases in their respective functional spaces. Commonly used functional spaces are the Sobolev space and Hölder space. Let

$$P(Z_i) = [P_0(Z_i), P_1(Z_i), \ldots, P_{J_n}(Z_i)]^T$$
$$Q(X_i) = [Q_0(X_i), Q_1(X_i), \ldots, Q_{k_n}(X_i)]^T$$

be the two sieve bases evaluated at the i-th observation and P and Q be the corresponding matrices:

$$P_{n\times J_n} = \begin{pmatrix} [P(Z_1)]^T \\ [P(Z_2)]^T \\ \cdots \\ [P(Z_n)]^T \end{pmatrix}, Q_{n\times k_n} = \begin{pmatrix} [Q(X_1)]^T \\ [Q(X_2)]^T \\ \cdots \\ [Q(X_n)]^T \end{pmatrix};$$

Date: Sept. 14, 2009.

Under immensely helpful, patient guidance from Prof. Yixiao Sun, University of California-San Diego, who provided the initial project idea as well as an initial draft of sections 1, 2, 4.1, and 4.2; and funded generously by the economics department of the University of California-San Diego.

then $E(Y|Z)$ and $E(h(X)|Z)$ can be estimated by

$$\widehat{E}(Y|Z) = P\left(P'P\right)^{-1}P'Y$$

and

$$\widehat{E}[h(X)|Z] = \widehat{E}(Q|Z)\,\Pi = P\left(P'P\right)^{-1}P'Q\Pi$$

respectively.

Running the regression

$$P\left(P'P\right)^{-1}P'Y = P\left(P'P\right)^{-1}P'Q\Pi + error$$

we obtain the OLS estimator

$$\begin{aligned}\hat{\Pi} &= \left(Q'P\left(P'P\right)^{-1}P'P\left(P'P\right)^{-1}P'Q\right)^{-1}Q'P\left(P'P\right)^{-1}P'P\left(P'P\right)^{-1}P'Y \\ &= \left(Q'P\left(P'P\right)^{-1}P'Q\right)^{-1}Q'P\left(P'P\right)^{-1}P'Y\end{aligned}$$

and

$$\hat{h}(X) = Q(X)\hat{\Pi} = Q\left(Q'P\left(P'P\right)^{-1}P'Q\right)^{-1}Q'P\left(P'P\right)^{-1}P'Y$$

If we impose the restriction that

$$\Pi'C\Pi \leq D_n$$

for some D_n, then the restricted OLS estimator of Π is

$$\hat{\Pi} = \left(Q'P\left(P'P\right)^{-1}P'Q + \lambda_n C\right)^{-1}Q'P\left(P'P\right)^{-1}P'Y$$

and the corresponding estimator of $h(X)$ is now:

$$\hat{h}(X) = Q(X)\left(Q'P\left(P'P\right)^{-1}P'Q + \lambda_n C\right)^{-1}Q'P\left(P'P\right)^{-1}P'Y. \tag{2}$$

Here λ_n is the Lagrangian multiplier for the constraint $\Pi'C\Pi \leq D_n$. Selecting D_n is equivalent to selecting λ_n. We consider the following penalty matrices:

$$C = C_r = \begin{cases} \frac{1}{n}\sum_{i=1}^n Q(X_i)\left[Q(X_i)\right]^T & r = 0 \\ \frac{1}{n}\sum_{i=1}^n Q'(X_i)\left[Q'(X_i)\right]^T & r = 1 \\ \frac{1}{n}\sum_{i=1}^n Q''(X_i)\left[Q''(X_i)\right]^T & r = 2 \\ \frac{1}{n}\sum_{i=1}^n Q'(X_i)\left[Q'(X_i)\right]^T + \frac{1}{n}\sum_{i=1}^n Q''(X_i)\left[Q''(X_i)\right]^T & r = 3 \end{cases}$$

where $Q(X_i)$, $Q'(X_i)$ and $Q''(X_i)$ are the basis functions and their first- and second-order derivatives.

2. Sieve Size Determination (smoothing parameter choice)

The estimator $\hat{h}$ depends on smoothing parameters k_n, J_n, and λ_n.

Define the average mean squared error (AMSE) as

$$AMSE = \frac{1}{n}\sum_{i=1}^n E\left\{\left[h(X_i) - \hat{h}(X_i)\right]^2\right\} = E\left\{\left[h(X) - \hat{h}(X)\right]^2\right\},$$

where the $E\{\}$ in the first equation is with respect to X_i and the $E\{\}$ in the second equation is with respect to X, and we assume that the data $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ is a random sample from the distribution of (X, Y, Z). In this paper, we propose to fix λ_n at some small value and select the smoothing parameters k_n and J_n optimally to minimize the AMSE:

$$(k_n^*, J_n^*) = \arg\min_{k_n, J_n} AMSE\,(k_n, J_n)\,.$$

In principle, we can set λ_n to be zero and the asymptotic results will remain valid. However, in small samples, it is sometimes advantageous to set λ_n to a small value. We assume that λ_n is so small that its effect on the AMSE is asymptotically negligible.

Note that

$$\begin{aligned}\left[h(X_i) - \hat{h}(X_i)\right]^2 &= \left[Y_i - \hat{h}(X_i) - u_i\right]^2 \\ &= \left[Y_i - \hat{h}(X_i)\right]^2 + u_i^2 - 2\left[Y_i - \hat{h}(X_i)\right] u_i,\end{aligned}$$

so the average squared error (ASE)

$$\begin{aligned}ASE\left(k_n, J_n\right) &= \frac{1}{n}\sum_{i=1}^{n}\left[h(X_i) - \hat{h}(X_i)\right]^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \hat{h}(X_i)\right]^2 - \frac{2}{n}\sum_{i=1}^{n}\left[Y_i - \hat{h}(X_i)\right] u_i + \frac{1}{n}\sum_{i=1}^{n} u_i^2\end{aligned}$$

provides a consistent estimator for the $AMSE\left(k_n, J_n\right)$. If $ASE\left(k_n, J_n\right) = AMSE\left(k_n, J_n\right) + o_p\left(1\right)$ uniformly over the smoothing parameters (k_n, J_n), then the smoothing parameters that minimize ASE will converge to the optimal $(k_n^*, J_n^*, \lambda_n^*)$.

Note that for the purpose of smoothing parameter choice, we can ignore the last term in $ASE\left(k_n, J_n, \lambda_n\right)$ as it does not depend on k_n, J_n, or λ_n. This consideration leads to the definition of the following criterion functions:

$$\overline{L}\left(k_n, J_n\right) = E\{L\left(k_n, J_n\right)\} = E\left\{\left[Y - \hat{h}(X)\right]^2\right\} - 2E\left\{\left[Y - \hat{h}(X)\right] u\right\},$$

and

$$L\left(k_n, J_n\right) = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \hat{h}(X_i)\right]^2 - \frac{2}{n}\sum_{i=1}^{n}\left[Y_i - \hat{h}(X_i)\right] u_i.$$

Since the true function $h(X_i)$ is unknown (hence $u_i = Y_i - h(X_i)$ is unknown), we have to estimate $ASE\left(k_n, J_n, \lambda_n\right)$, or equivalently $L\left(k_n, J_n\right)$, in order to select the smoothing parameters.

3. Literature Review

Model selection is difficult because the mean squared prediction error (MSPE) of a model is different from the in-sample ASE, and so there is danger of overfitting; otherwise, simply minimizing the residual sum of squares (RSS) would be optimal. Many elegant criteria for statistical model selection exist; see, *e.g.*, Konishi and Kitagawa (2008). But endogeneity, a ubiquitous phenomenon in economic data, poses further difficulties that the Aikake information criterion (AIC), Mallows' C_p, and others do not account for.

Loader (1999) examines an analogous problem of selecting bandwidth in kernel density estimation. In particular, he compares "classical" methods, such as AIC and Mallows' C_p, with "plug-in" methods such as nonparametric bootstrap, concluding that classical methods are better. Efron (2004) does a similar comparison, denoting the same categories as "penalty methods" and "cross-validation" (which includes nonparametric bootstrap), and concluding that it is usually better to guess a parametric model and use a penalty method than use a nonparametric method. Ye (1998) provides a computational algorithm for computing "general degrees of freedom" (GDF) that can be plugged into the AIC to yield what he calls the extended AIC (EAIC), but he does not address endogeneity. Florens and Lestringant (2007) propose a method for the nonparametric IV model we examine here.

4. Methods

4.1. **Cross-validation (CV).** To perform CV:

(1) Select some "pilot" smoothing parameters k_0 and J_0 and construct the corresponding estimator $\tilde{h}(X_i)$ as in equation (2) above and the error term $\tilde{u}_i = Y_i - \tilde{h}(X_i)$. Two possible ways to select "pilot" parameters: (i) let k_0 be the one chosen by cross-validation ignoring endogeneity, let $J_0 = 2k_0$; (ii) choose k_0 and J_0 by cross-validation using the weak-norm criterion: $\min_h E\left\{(E[Y - h(X)|Z])^2\right\}$.

(2) For each smoothing parameter combination (k_n, J_n), construct $\hat{h}_{(-i)}(x)$ as in equation (2) as the estimator of $h(x)$ obtained after removing the i-th observation.

(3) Compute the cross-validation estimated risk:

$$\widehat{L}(k_n, J_n) = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \hat{h}_{(-i)}(X_i)\right]^2 - \frac{2}{n}\sum_{i=1}^{n}\left[Y_i - \hat{h}_{(-i)}(X_i)\right]\tilde{u}_i$$

or

$$ASE(k_n, J_n) = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \hat{h}_{(-i)}(X_i) - \tilde{u}_i\right]^2$$

(4) Select (k_n, J_n) to minimize $\widehat{L}(k_n, J_n)$ or $ASE(k_n, J_n)$:

$$(\hat{k}_n, \hat{J}_n) = \arg\min \widehat{L}(k_n, J_n)$$

or

$$(\hat{k}_n, \hat{J}_n) = \arg\min ASE(k_n, J_n)$$

Computing $\hat{h}_{(-i)}(X_i)$ for all $i = 1, 2, \ldots, n$ can be computationally intensive. We now establish the relationship between $\hat{h}_{(-i)}(X_i)$ and $\hat{h}(X_i)$. Using this relationship, the cross-validation risk $\widehat{L}(k_n, J_n)$ can be computed using a single nonparametric IV regression. Let

$$Y_j^i = \begin{cases} \hat{h}_{(-i)}(X_i), & \text{for } j = i \\ Y_j, & \text{for } j \neq i \end{cases}$$

and $Y^i = (Y_1^i, \ldots, Y_n^i)$. Y^i is the same as $Y = (Y_1, \ldots, Y_n)$ except that the i-th element is replaced by its leave-one-out predictive value $\hat{h}_{(-i)}(X_i)$.

Let $\hat{\Pi}_{-i}$ be the estimator of Π with the i-th observation left out. By defintion,

$$\begin{aligned}\hat{\Pi}_{-i} &= \arg\min_{\Pi}\left\|P_{-i}\left(P_{-i}'P_{-i}\right)^{-1}P_{-i}'\left(Y_{-i} - Q_{-i}\Pi\right)\right\| + \lambda_n\Pi'C\Pi \\ &= \arg\min_{\Pi}\left\|P\left(P'P\right)^{-1}P'\left(Y^i - Q\Pi\right)\right\| + \lambda_n\Pi'C\Pi.\end{aligned}$$

That is, $\hat{\Pi}_{-i}$ is the estimator of Π when the i-th row of the pseudo-regression

$$P\left(P'P\right)^{-1}P'Y = P\left(P'P\right)^{-1}P'Q\Pi + error$$

is left out.

Let

$$H =: (H_{ij})_{n\times n} = Q\left(Q'P\left(P'P\right)^{-1}P'Q + \lambda_n C\right)^{-1}Q'P\left(P'P\right)^{-1}P',$$

then $\hat{h}_{-i} = HY^i$ and

$$\begin{aligned}\hat{h}_{-i}(X_i) = Q_i\hat{\Pi}_{-i} = H_iY^i &= \sum_{j=1}^{n} H_{ij}Y_j^i = \sum_{j=1}^{n} H_{ij}Y_j + H_{ii}(Y_j^i - Y_i) \\ &= \hat{h}(X_i) + H_{ii}\hat{h}_{-i}(X_i) - H_{ii}Y_i\end{aligned}$$

So,

$$\hat{h}_{-i}(X_i) = \frac{\hat{h}(X_i) - H_{ii}Y_i}{1 - H_{ii}}$$

and

$$Y_i - \hat{h}_{-i}(X_i) = Y_i - \frac{\hat{h}(X_i) - H_{ii}Y_i}{1 - H_{ii}} = \frac{Y_i - \hat{h}(X_i)}{1 - H_{ii}}$$

Plugging this into $ASE(k_n, J_n)$ yields

$$\widehat{ASE}_{CV}(k_n, J_n) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_i - \hat{h}(X_i)}{1 - H_{ii}} - \tilde{u}_i\right]^2.$$

In the simulations, this "efficient" CV (as opposed to computing leave-one-out CV) is used with both the pilot $k0$ shown (labeled "EffCV0") as well as using $k_{\max}$ as the pilot ("EffCVm").

Following chapter 10 in Konishi and Kitagawa (2008), we also implemented a CV criterion that does not depend on pilot parameters:

$$\text{EffCVn}(k_n, J_n) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_i - \hat{h}(X_i)}{1 - H_{ii}}\right]^2.$$

4.2. **Bootstrap.** Since the true function $h(X_i)$ is unknown (hence $u_i = Y_i - h(X_i)$ is unknown), we have to estimate $L(k_n, J_n, \lambda_n)$ in order to select the smoothing parameters. There are many ways to estimate $L(k_n, J_n, \lambda_n)$ consistently. We propose to use the bootstrap method to estimate the the covariance term $2/n\sum_{i=1}^{n}\left[\hat{h}(X_i) - h(X_i)\right]u_i$. We outline the basic steps as follows:

(1) Select some pilot smoothing parameters k_0, J_0, and λ_0 and construct the corresponding estimator $\tilde{h}(X_i)$ as in equation (2) and the error term $\tilde{u}_i = Y_i - \tilde{h}(X_i)$

(2) For each smoothing parameter combination (k_n, J_n, λ_n), construct $\hat{h}(X_i)$ as in equation (2).

(3) Perform nonparametric bootstrap, that is, randomly draw n observations with replacement from (X_i, Y_i, Z_i). Based on the bootstrapped sample $(X_i^*, Y_i^*, Z_i^*), i = 1, 2, \ldots, n$, construct the estimator $\hat{h}^*(X_i^*)$

(4) Let $(X_{ib}^*, Y_{ib}^*, Z_{ib}^*)$ be the b-th bootstrapped sample. For each bootstrap sample, compute

$$\frac{2}{n}\sum_{i=1}^{n}\left[\hat{h}^*(X_{ib}^*) - \tilde{h}(X_{ib}^*)\right]\tilde{u}_i(X_{ib}^*)$$

(5) Repeat steps (3) and (4) B times and estimate $L(k_n, J_n, \lambda_n)$ by

$$\hat{L}(k_n, J_n, \lambda_n) = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \hat{h}(X_i)\right]^2 + \frac{2}{nB}\sum_{b=1}^{B}\sum_{i=1}^{n}\left[\hat{h}^*(X_{ib}^*) - \tilde{h}(X_{ib}^*)\right]\tilde{u}_i(X_{ib}^*)$$

(6) Select k_n, J_n, and λ_n to minimize $\hat{L}(k_n, J_n, \lambda_n)$:

$$(\hat{k}_n, \hat{J}_n, \hat{\lambda}_n) = \arg\min \hat{L}(k_n, J_n, \lambda_n)$$

In the simulations, this bootstrap method is run for two pilot values, the given $k0$ and $k_{\max}$, respectively denoted "YixiaoBSk0" and "YixiaoBSkmax".

We also implemented a bootstrap version evaluating at the original sample points and using a direct bootstrap analog of the ASE, so that step (4) above computes

$$\frac{1}{n}\sum_{i=1}^{n}\left[\hat{h}^*(X_i)-\tilde{h}(X_i)\right]^2$$

and step (5) takes the average over all the bootstrap draws to get the estimated ASE criterion. In the simulation results below, this is (poorly) denoted "ASEhat0", and uses k_max as the pilot parameter.

4.3. **Weighted Bootstrap (WBS).** Mason and Newton (1990) first provided theory behind a bootstrap using generalized weights. The standard nonparametric bootstrap is WBS using multinomial weights. Since the multinomial weights sum to n and are integers, weighting the original data sample by these bootstrap weights is the same as resampling with replacement. Mason and Newton (1990) showed that independent, random weights can be used, which can make theoretical results easier, as noted by Ma and Kosorok (2005) in their paper on semiparametric M-estimation.

Here, we use independent weights drawn from a χ^2 distribution with mean 1 and variance 1. If W is a diagonal matrix of size $n \times n$ of these weights, then for each bootstrap draw of weights,

$$\widehat{E}(Y|Z)^* = P\left(P'WP\right)^{-1}P'WY$$
$$\widehat{E}[h(X)|Z]^* = \widehat{E}\left(Q|Z\right)\Pi = P\left(P'WP\right)^{-1}P'WQ\Pi$$

and so

$$\begin{aligned}\hat{\Pi}^* &= \left(Q'WP\left(P'WP\right)^{-1}P'WP\left(P'WP\right)^{-1}P'WQ+\lambda_n C\right)^{-1}Q'WP\left(P'WP\right)^{-1}P'WP\left(P'WP\right)^{-1}P'WY\\ &= \left(Q'WP\left(P'WP\right)^{-1}P'WQ+\lambda_n C\right)^{-1}Q'WP\left(P'WP\right)^{-1}P'WY\end{aligned}$$

The WBS criterion used in the simulation results below is

$$\text{WBS} = \frac{1}{nB}\sum_{b=1}^{B}\sum_{i=1}^{n}\left[\hat{h}(X_i)-\hat{h}^*(X_i)\right]^2$$

where $\hat{h}(X)$ is from equation (2) given (k_n, J_n), and $\hat{h}^*(X_i) = Q\hat{\Pi}^*$ given (k_n, J_n) and the weights W.

We also tried a WBS using a pilot $\tilde{h}(\cdot)$, similar to the standard bootstrap criterion above, but it did not yield encouraging results, so it was left out of the comprehensive simulation comparison below for purposes of computational time.

4.4. **Extended Aikake Information Criterion (EAIC).** The EAIC is taken directly from equation (12) in Ye (1998). It is based on the AIC, but uses Ye's generalized degrees of freedom (GDF) for the bias correction term. Thus,

$$\text{EAIC} = \text{RSS} - n\sigma^2 + 2D\sigma^2$$
$$\text{AEAIC} = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i-\hat{h}(X_i)\right]^2 - \sigma^2 + \frac{2}{n}D\sigma^2$$

where the AEAIC is just averaged (but yields identical smoothing parameter choice), and D is the GDF. For the simulation, we plugged in the true σ^2 from the DGP, so we denote this criterion "cheatAEAIC" in the results. To actually use it, we would either plug in an estimated $\hat{\sigma}^2$, or find an expression where σ^2 is implicit, as in the

derivation of generalized cross-validation (GCV) in Eggermont and LaRiccia (2009) in chapter 18, concluding with their equation (2.14).

We use $D = \mathrm{tr}H$, where H is the "hat matrix" $Q\left(Q'P\left(P'P\right)^{-1}P'Q + \lambda_n C\right)^{-1}Q'P\left(P'P\right)^{-1}P'$ that projects some observed Y onto its estimated values $\hat{Y}$ (thus H is also known as the "projection matrix"). The trace (which reduces to the number of regressors for OLS) can be used for closed-form estimators; Ye also provides a computational algorithm for more general cases.

4.5. **Generalized Cross-validation (GCV).** We used the GCV criterion as presented in Konishi and Kitagawa (2008), equation (10.13), citing Craven and Wahba (1979), though we prefer the derivation (of the identical criterion expression) presented in Eggermont and LaRiccia (2009) mentioned above.

$$\mathrm{GCV} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i - \hat{h}(X_i)\right\}^2}{\left\{1 - \frac{1}{n}\mathrm{tr}H(k_n, J_n)\right\}}$$

4.6. **Infeasible Rules.** We compare the performance of the proposed smoothing parameter selection rule with some infeasible rules. In the first infeasible rule, we simulate $AMSE\left(k_n, J_n, \lambda_n\right)$, find the theoretically optimal (on average) k_n^*, and use it throughout the rest of the simulation. In the second infeasible rule, we assume that the true function is known so that we can compute

$$ASE\left(k_n, J_n, \lambda_n\right) = \frac{1}{n}\sum_{i=1}^{n}\left[h(X_i) - \hat{h}(X_i)\right]^2$$

The optimal k_n^{**} that minimizes $ASE\left(k_n, J_n, \lambda_n\right)$ will depend on the realization of the DGP. Hence k_n^{**} is a random variable that changes from replication to replication. In our comparison, we also include the estimator that simply uses the pilot smoothing parameters (k_0, J_0, λ_0), denoted "k0" in the results.

Let $\left\{X_i^t\right\}$ be $(0.90)n$ target points to evaluate the performance of different estimators. In our simulation, $\left\{X_i^t, i = 1, 2, \ldots, n\right\}$ are iid with the same distribution as X, but with the lower and upper 5% (each) trimmed. Once generated, $\left\{X_i^t, i = 1, 2, \ldots, n\right\}$ is held fixed throughout simulation replications.

5. Simulation Results

The following are a couple specific results, followed by the comprehensive comparison among all the aforementioned methods. The data generating process (DGP) and model used in the first two subsections are described in the third subsection below.

5.1. **Weighted Bootstrap (WBS).** The WBS criterion seems to mirror the true ASE well for most k_n, but not for small values of k_n. It is possible that the implementation we used captures variance of the estimator quite well, but not bias, which would make sense since the bias term will dominate ASE for small k_n and variance will dominate for larger k_n. The fact that WBS can decrease with k_n, though, seems to indicate it is not just an estimator of variance. A comparison of the WBS criterion with the true ASE appears in Fig. 1. The dashed lines are confidence intervals around the true ASE. The "90" suffix implies calculation for only the middle 90% of points (omitting points close to the boundary of the sample). The plot is typical of the results we got comparing WBS to true ASE.

5.2. **Sensitivity to Pilot Parameters.** Most of the pilot-dependent methods look quite reliant on the choice of the pilot value of k_0, with the exception of the "ASEhat0" criterion (Fig. 2). Also note WBS performs poorly due to its problems at low values of k_n. Restricting the selected k_n to be bigger than 2 improves WBS significantly, but not the standard nonparametric bootstrap methods (Fig. 3).

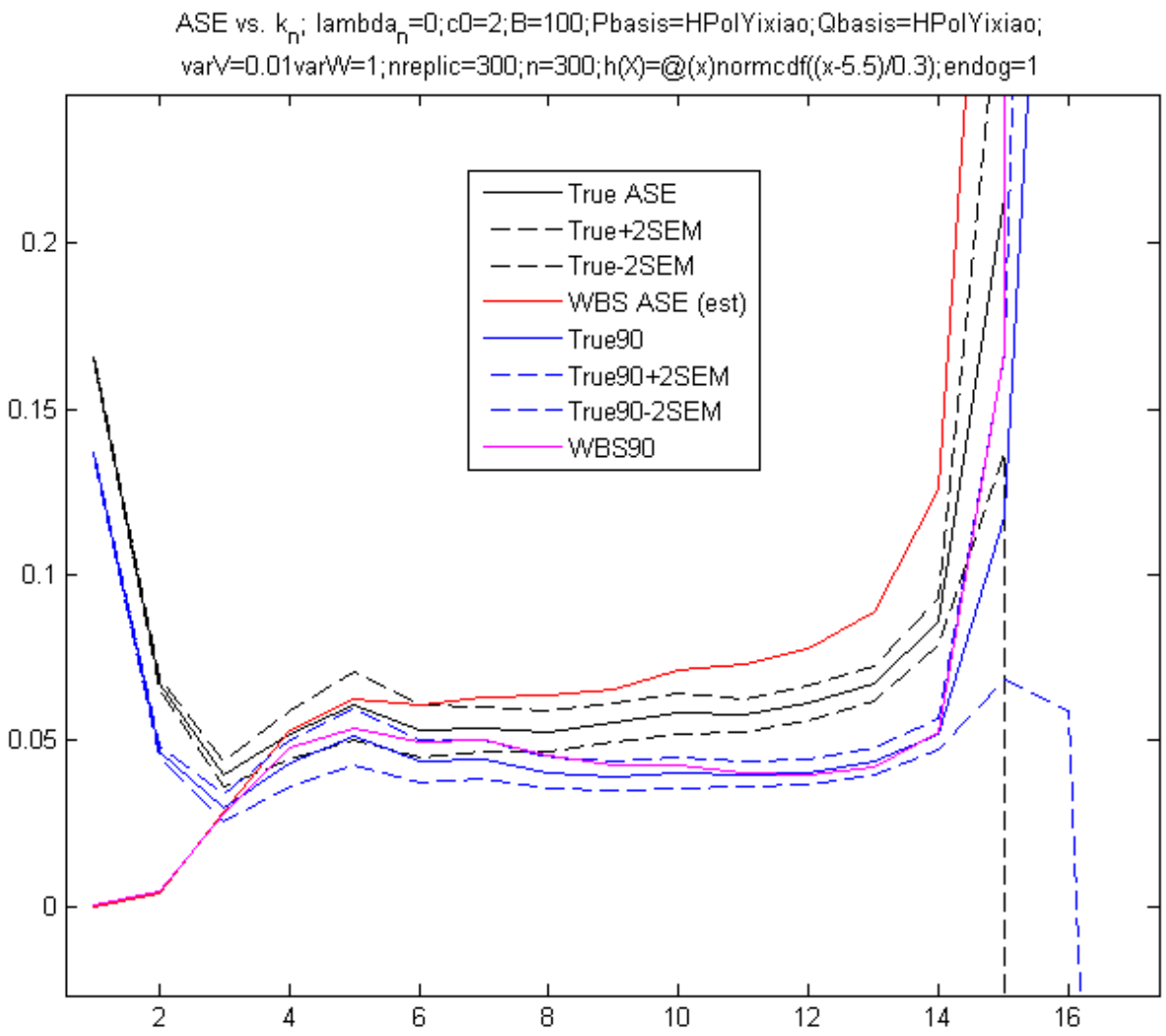

FIGURE 1. WBS criterion compared with true ASE

5.3. **Comparison Among All Methods.** We compared the performance of all criteria discussed above, using two different DGPs.

5.3.1. *BCK.* Blundell et al. (2007) have shown that the optimal orders of k_n and of J_n satisfy $J_n/k_n \to c_0 \in (1, \infty)$. In the simulation study, we set $c_0 = 2$ and set $J_n = c_0 k_n$ and $J_0 = c_0 k_0$. We also consider only a few value of λ_n, *i.e.* $\lambda_n = 0.001, 0.0001, 0$. Essentially, we search the optimal k_n, J_n, and λ_n over the following set

$$\{(k_n, J_n, \lambda_n) = (k, c_0 k, \lambda)\,, \text{ for } k = 1, 2, \ldots, k_{\max},\ c_0 = 2 \text{ and } \lambda = 0.001, 0.0001, 0\}$$

where $k_{\max}$ is the upper bound for k. In principle, we can do a full search, but the computational cost becomes increasingly high when we consider all possible combinations of (k_n, J_n, λ_n). In the simulation study here, we do not search over c_0 and λ_n. That is, we fix these two parameters and search for the optimal k_n.

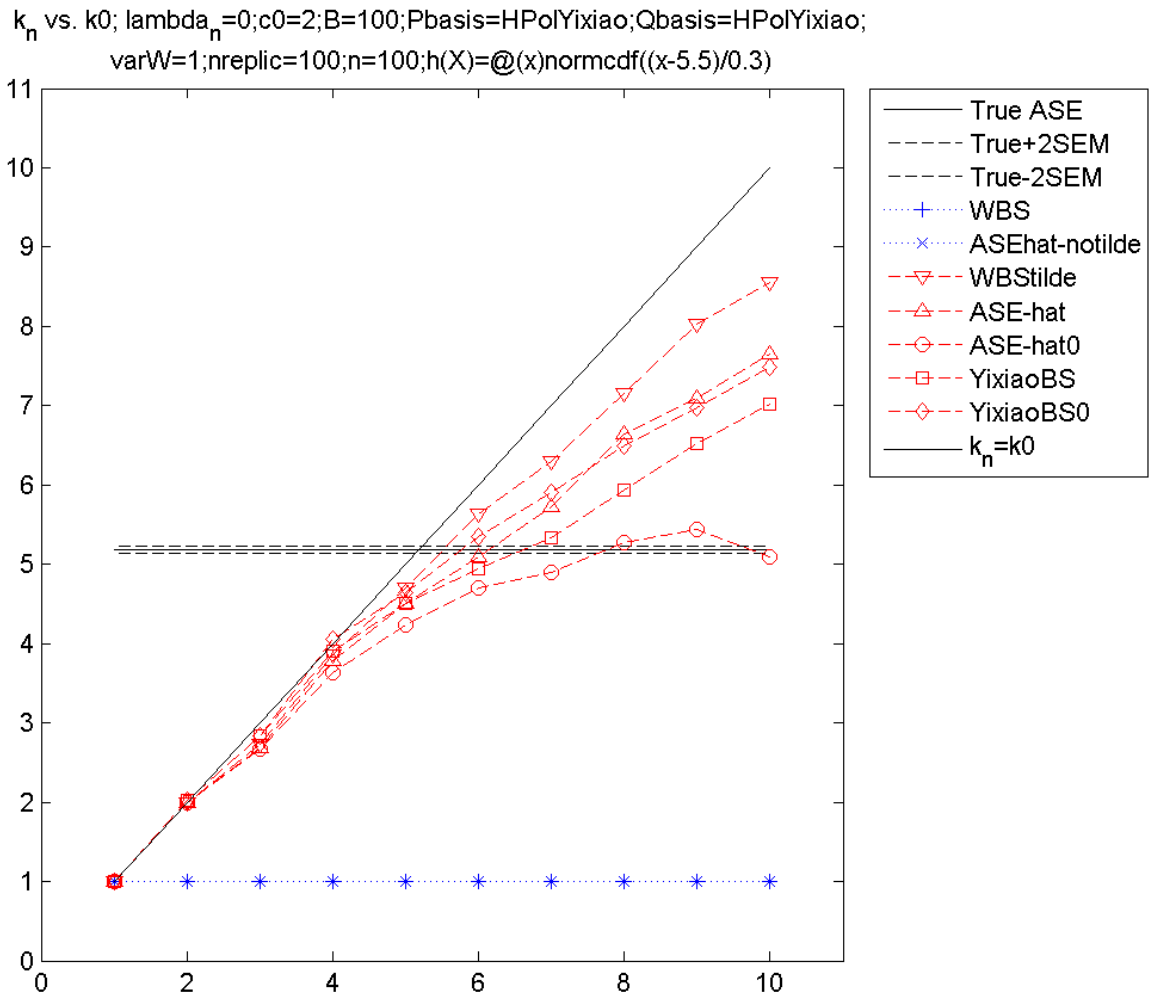

FIGURE 2. Selected k_n as a function of pilot k_0

In the simulation experiment, we employ the DGP in Blundell et al. (2007). We generate (X, Z) from the bivariate normal distribution with

$$\begin{aligned} mean &= (\mu_X, \mu_Z) = (5.3744, 5.7712) \\ sd &= (\sigma_X, \sigma_Z) = (0.4864, 0.5389) \end{aligned}$$

and correlation coefficient $\rho_{XZ} = 0.5111$. We take $h(\cdot)$ to be a Normal CDF,

$$h(x) = \Phi\left(\frac{x - 5.5}{0.3}\right).$$

For this specification, it is easy to show that

$$E\left(h(X)|Z\right) = \Phi\left(\frac{\mu_X + \rho_{XZ}\sigma_X(Z - \mu_Z)/\sigma_Z - 5.5}{\sqrt{0.3^2 + \left(1 - \rho_{XZ}^2\right)\sigma_X^2}}\right).$$

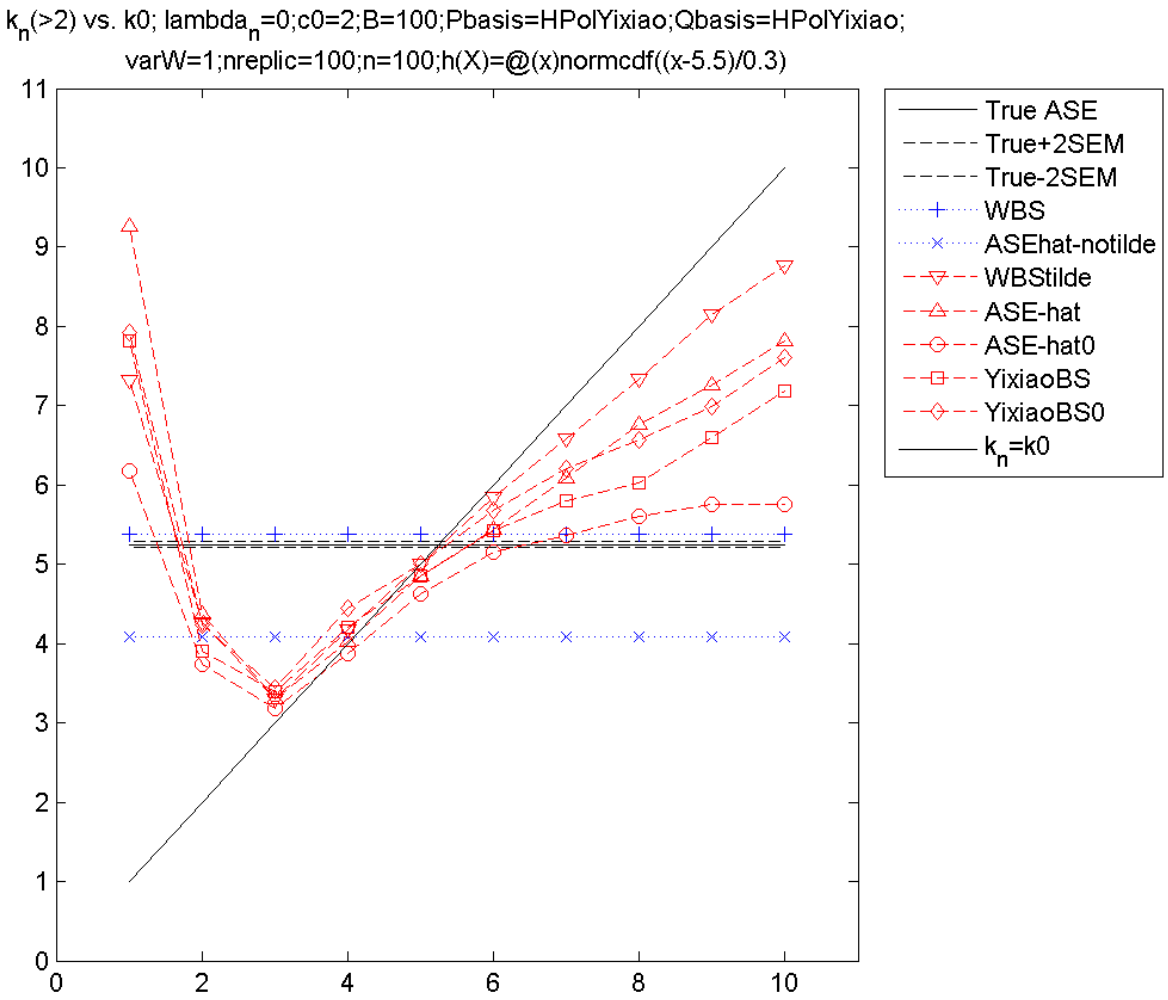

FIGURE 3. Selected k_n as a function of pilot k_0, restricting the selected $k_n > 2$

The dependent variable Y is generated according to

$$Y = h(X) + u$$

for

$$u = E(h(X)|Z) - h(X) + v$$

with $v \sim N(0, 0.01)$, which is independent of all other variables in the model. In Blundell et al. (2007), the sample size was $n = 628$, but in consideration of computational intensity, we have used $n = 300$. Unless otherwise noted, we ran 100 simulation replications.

TABLE 1. Results for various selection criteria using BCK setup and Hermite polynomial basis. $B = 100, \sigma_W^2 = 1, n = 300, nreplic = 100$

	$\mu(k_n)$	$\sigma(k_n)$	abias^2	avar	amse	$\mu(k_n)$	$\sigma(k_n)$	abias^2	avar	amse
	$(k_0 = 6, c_0 = 2, \lambda_n = 0.001, k_{\max} = 10)$					$(k_0 = 6, c_0 = 2, \lambda_n = 0.0001, k_{\max} = 15)$				
k*	3.00	0.00	0.0125	0.0188	0.0313	3.00	0.00	0.0111	0.0137	0.0249
k**	6.11	2.13	0.0003	0.0118	0.0121	5.69	2.38	0.0001	0.0099	0.0100
k0	6.00	0.00	0.0040	0.0366	0.0406	6.00	0.00	0.0019	0.0328	0.0346
EffCV0	5.95	0.26	0.0039	0.0366	0.0405	5.98	0.28	0.0019	0.0328	0.0346
EffCVm	8.32	1.92	0.0025	0.0324	0.0349	7.83	2.98	0.0019	0.0348	0.0368
EffCVn	5.86	1.91	0.0026	0.0135	0.0161	5.73	1.91	0.0019	0.0152	0.0171
ASEBS0	5.69	2.42	0.0016	0.0359	0.0375	5.38	2.58	0.0012	0.0315	0.0327
WBS	2.00	0.00	0.0417	0.0039	0.0456	2.00	0.00	0.0417	0.0055	0.0472
WBSgt3	5.50	1.98	0.0028	0.0233	0.0261	5.35	2.08	0.0027	0.0191	0.0219
EAIC	6.92	2.21	0.0031	0.0141	0.0171	8.95	3.66	0.0036	0.0166	0.0202
ARSS	6.96	2.23	0.0031	0.0141	0.0171	9.09	3.67	0.0036	0.0166	0.0202
YBS0	5.03	1.62	0.0021	0.0240	0.0262	5.00	1.61	0.0006	0.0216	0.0222
YBSm	7.06	2.28	0.0032	0.0143	0.0175	9.08	3.54	0.0036	0.0169	0.0206
GCV	6.72	2.27	0.0031	0.0138	0.0169	8.20	3.48	0.0034	0.0159	0.0193

For each estimator chosen by each selection criterion, we compute the average squared bias, average variance, and average squared error:

$$abias^2 = \frac{1}{n}\sum_{i=1}^{n}\left[h(X_i^t) - \frac{1}{n}\sum_{r=1}^{R}\hat{h}_r(X_i^t)\right]^2$$

$$avar = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{h}_r(X_i^t) - \frac{1}{n}\sum_{r=1}^{R}\hat{h}_r(X_i^t)\right]^2$$

$$ASE = abias^2 + avar$$

where $h(X_i^t)$ is the true function $h(X)$ evaluated at $X = X_i^t$, $\hat{h}_r(X_i^t)$ is the estimated function value at the r-th replication evaluated at X_i^t.

For consideration of space, we only show results from two basis functions instead of all five, but Table 1 is typical of the others.

5.3.2. *Additional DGP.* We tried one additional DGP with one function. The true function we used was $h(X) = 1 + X + X^2$. The iid data were generated by:

$$Z \sim N(0,1)$$

$$X = \rho_X Z + \epsilon_X\sqrt{1-\rho_X^2}, \quad \epsilon_X \sim N(0,1)$$

$$U = \sigma_U\left(\rho_U\epsilon_X + \epsilon_U\sqrt{1-\rho_U^2}\right), \quad \epsilon_U \sim N(0,1)$$

$$Y = h(X) + U$$

This DGP satisfies the requirements of the general nonparametric IV model described earlier in this paper, and allows for any $h(\cdot)$ to be used. Note that the data are standardized such that the mean of X is 0, variance 1. In the simulations, "varV" refers to the variance of the exogenous error; in terms of the above, varV is $\sigma_U^2(1-\rho_U^2)$. We used $\rho_X = 0.5$ and $\rho_U = 0.2, 0.8$.

TABLE 2. Results for various selection criteria using BCK setup and trigonometric polynomial basis. $B = 100, \sigma_W^2 = 1, n = 300, nreplic = 100$

	$\mu(k_n)$	$\sigma(k_n)$	abias^2	avar	amse	$\mu(k_n)$	$\sigma(k_n)$	abias^2	avar	amse
	$(k_0 = 6, c_0 = 2, \lambda_n = 0.001, k_{\max} = 10)$					$(k_0 = 6, c_0 = 2, \lambda_n = 0.0001, k_{\max} = 15)$				
k*	2.00	0.00	0.0030	0.0033	0.0062	1.00	0.00	0.0014	0.0051	0.0065
k**	3.49	1.86	0.0033	0.0021	0.0055	4.59	2.42	0.0022	0.0053	0.0074
k0	6.00	0.00	0.0068	0.0026	0.0094	6.00	0.00	0.0051	0.0074	0.0126
EffCV0	6.04	0.57	0.0069	0.0026	0.0094	6.10	0.77	0.0052	0.0074	0.0126
EffCVm	9.92	0.34	0.0116	0.0021	0.0137	14.01	1.87	0.0178	0.0045	0.0223
EffCVn	9.90	0.33	0.0117	0.0020	0.0138	14.23	1.56	0.0183	0.0044	0.0227
ASEBS0	6.68	1.20	0.0074	0.0028	0.0102	10.94	1.66	0.0126	0.0056	0.0182
WBS	8.45	2.29	0.0102	0.0027	0.0129	14.41	0.99	0.0185	0.0042	0.0227
WBSgt3	8.85	1.62	0.0108	0.0023	0.0131	14.41	0.99	0.0185	0.0042	0.0227
EAIC	9.97	0.17	0.0117	0.0020	0.0138	14.85	0.41	0.0192	0.0042	0.0234
ARSS	9.98	0.14	0.0117	0.0020	0.0138	14.85	0.41	0.0192	0.0042	0.0234
YBS0	2.90	1.49	0.0043	0.0033	0.0076	3.49	1.60	0.0047	0.0084	0.0131
YBSm	9.86	0.40	0.0117	0.0020	0.0137	14.82	0.54	0.0191	0.0042	0.0234
GCV	9.94	0.28	0.0117	0.0020	0.0138	14.81	0.46	0.0192	0.0042	0.0233

We ran this model using a few different basis functions. Using a polynomial basis (Fig. 4, Table 3), this is a good example of an "easy" problem where the true function is actually in the functional class used for the estimator, and the optimal k_n is clear (as long as the error is significant but not too large). With a Hermite polynomial basis (Fig. 5, Table 3), this looks like a more challenging problem than BCK, in that the true (simulated) AMSE is U-shaped with a smaller range of "good" k_n near the minimum of the U. With the trigonometric polynomial basis (Fig. 6, Table 4), interestingly, the true AMSE is almost completely invariant to k_n over the domain examined, and achieves a minimum close to the minimum AMSE for a Hermite polynomial basis (though bigger than a polynomial basis, of course), which seems to suggest that a trigonometric polynomial basis would be a good choice for a researcher with this problem, but not useful for our attempt to compare the performance of various smoothing parameter selection criteria. Also note that RSS does not perform as well with this basis.

6. Discussion and Next Steps

It is not encouraging that RSS seems to generally outperform other methods (Tables 1, 3, and 4), albeit with some exception (*e.g.*, Table 2), since it should not. This may be due to the small number of true $h(X)$ tried, but it is still surprising. The other top performers, EAIC and GCV, seem to mirror the RSS here (from examining the ASE plots), and EffCVn is quite related to GCV and RSS. Thus, the CV, bootstrap, and WBS criteria all performed the worst, generally. At this stage, there is also admittedly some chance of errors existing in the simulation programs.

There is still much room for work on this project. There are a few immediate steps toward reaching more robust simulation results and understanding the current results. It would be helpful to look at the ASE plots for individual replications, instead of just the averaged plots, to get a better sense of variance and accuracy. It would help to restrict the ASE criteria to the middle 90% or 95% of sample points, since points near the boundaries of the sample are known to be subject to large variance and thus may drive the overall ASE, when estimation near the boundary should not be paid any heed. This has been started with the WBS method. The number of bootstrap samples should be increased (say, to 1000) to see if we have been unfair to those methods. We should also try other functions, including some more difficult problems such as in Florens and Lestringant (2007), without weighting performance on hard-but-rare problems more than performance on easy-but-common problems. It would also be interesting to include selection among the basis functions in addition to smoothing parameter selection; note the

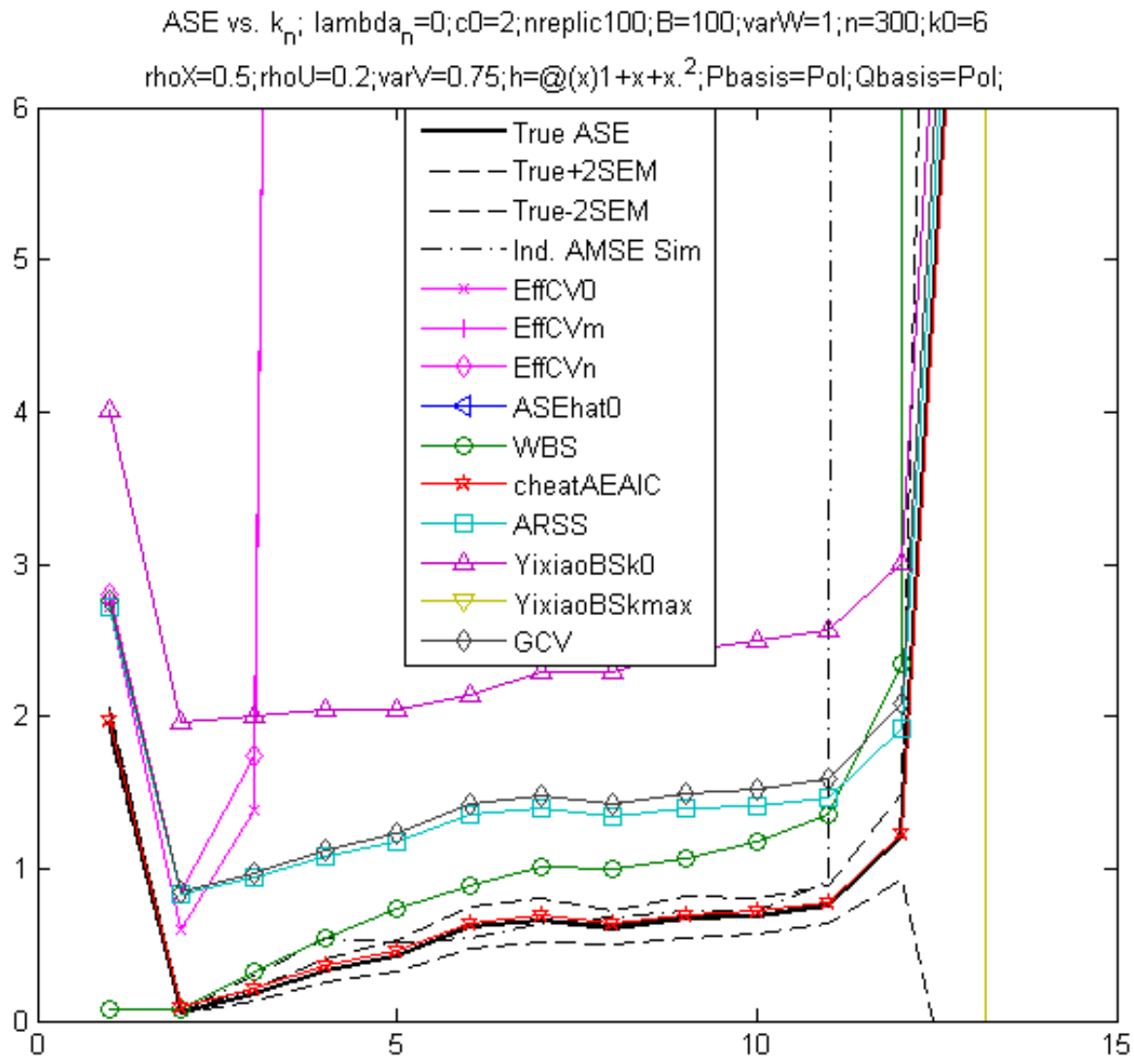

FIGURE 4. True ASE and various selection criteria, for second DGP and polynomial basis

significant difference in the problem for $h(X) = 1 + X + X^2$ when using a trigonometric polynomial basis versus a Hermite polynomial basis versus a polynomial basis.

Ultimately, the goal is a proof of the performance of some data-driven method for selecting the sieve size.

REFERENCES

Blundell, Richard, Xiaohong Chen, and Dennis Kristensen (2007), "Semi-nonparametric iv estimation of shape-invariant engel curves." *Econometrica*, 75, 1613–1669.

Efron, Bradley (2004), "The estimation of prediction error." *Journal of the American Statistical Association*, 99, 619–632.

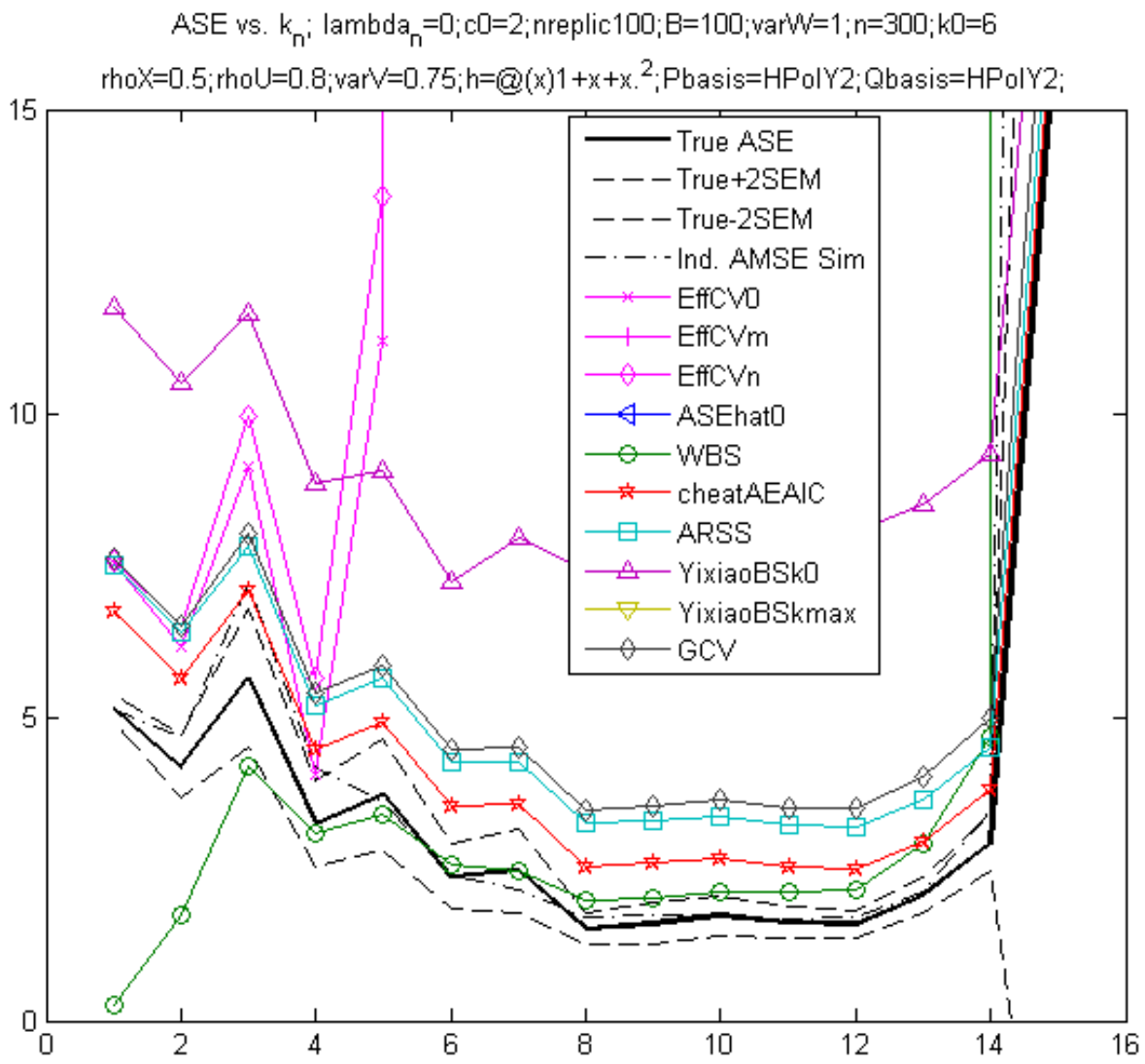

FIGURE 5. True ASE and various selection criteria, for second DGP and Hermite polynomial basis

Eggermont, P.P.B. and V.N. LaRiccia (2009), *Maximum Penalized Likelihood Estimation, Volume II: Regression*. Springer, New York.

Florens, Jean-Pierre and Renaud Lestringant (2007), "The practice of non parametric instrumental variables estimation." *Unpublished manuscript*.

Konishi, Sadanori and Genshiro Kitagawa (2008), *Information Criteria and Statistical Modeling*. Springer, New York.

Loader, Cliver R. (1999), "Bandwidth selection: Classical or plug-in?" *The Annals of Statistics*, 27, 415–438.

Ma, Shuangge and Michael R. Kosorok (2005), "Robust semiparametric m-estimation and the weighted bootstrap." *Journal of Multivariate Analysis*, 96, 190–217.

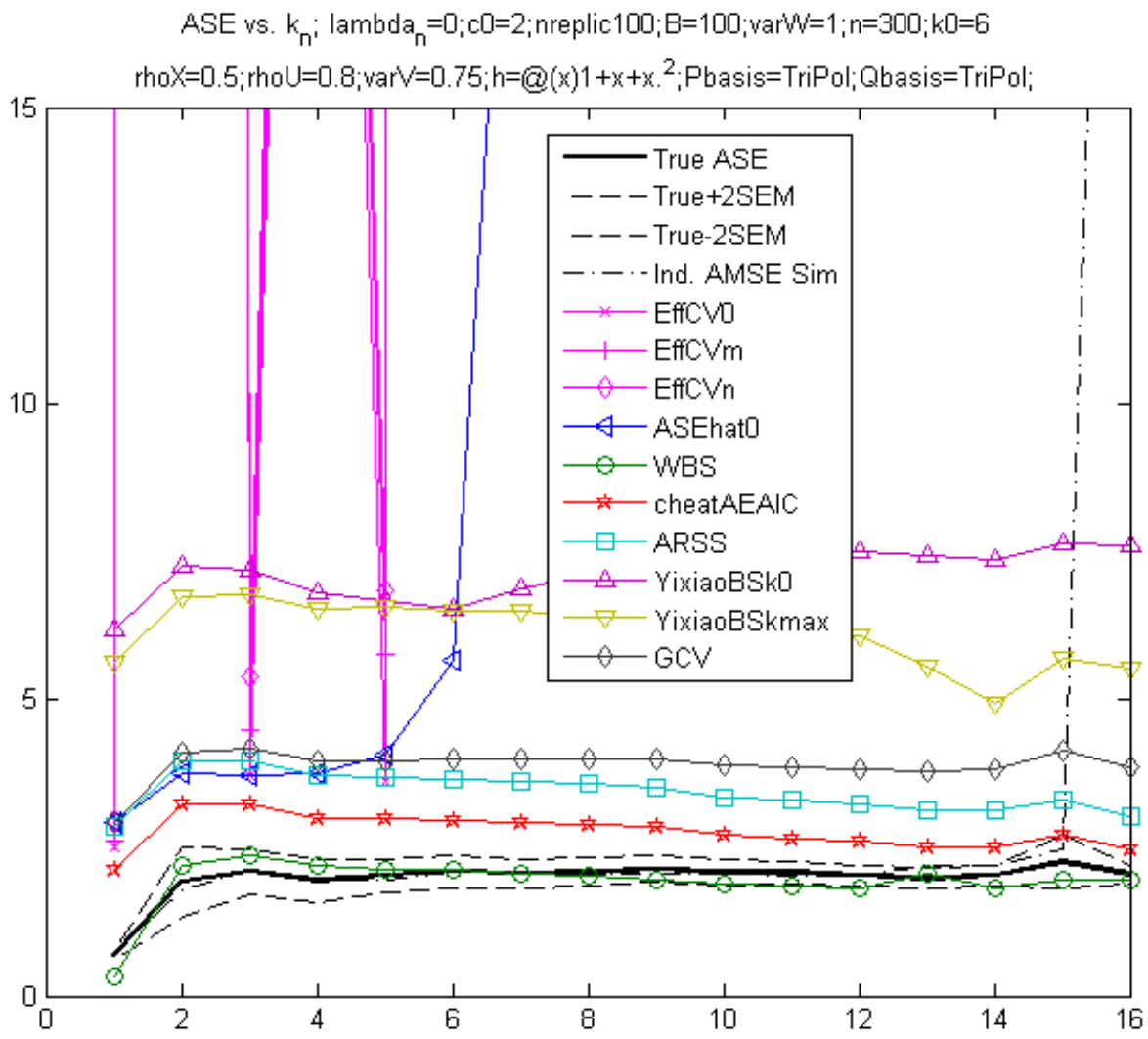

FIGURE 6. True ASE and various selection criteria, for second DGP and trigonometric polynomial basis

Mason, David M. and Michael A. Newton (1990), "A rank statistics approach to the consistency of a general bootstrap." *University of Washington Department of Statistics Technical Report*.

Ye, Jianming (1998), "On measuring and correcting the effects of data mining and model selection." *Journal of the American Statistical Association*, 93, 120–131.

E-mail address: dkaplan@ucsd.edu

TABLE 3. Results for various selection criteria using different basis functions, $h(X) = 1 + X + X^2$, second DGP. $\lambda_n = 0, c_0 = 2, B = 100, \sigma_W^2 = 1, n = 300, nreplic = 100, k_0 = 6, \rho_X = 0.5, \rho_U = 0.8, \sigma_V^2 = 0.75$

	$\mu(k_n)$	$\sigma(k_n)$	abias^2	avar	amse	$\mu(k_n)$	$\sigma(k_n)$	abias^2	avar	amse
	Polynomial basis					Hermite polynomial basis				
k*	2.00	0.00	0.0004	0.0244	0.0248	11.00	0.00	0.0114	0.9285	0.9399
k**	2.48	1.53	0.0004	0.0229	0.0233	8.88	2.58	0.0119	0.4519	0.4638
k0	6.00	0.00	0.0089	0.2512	0.2601	6.00	0.00	0.0853	1.5665	1.6518
EffCV0	3.57	1.78	0.0031	0.1379	0.1410	5.96	0.63	0.0695	1.5897	1.6592
EffCVm	3.34	2.22	0.0053	0.1103	0.1156	7.73	2.76	0.0277	1.0024	1.0301
EffCVn	2.06	0.28	0.0004	0.0245	0.0249	7.14	2.09	0.0027	0.6884	0.6910
ASEBS0	2.44	1.22	0.0087	0.1241	0.1328	6.79	2.31	0.0198	1.2359	1.2557
WBS	1.67	0.49	0.0605	0.1673	0.2278	1.05	0.22	1.7536	0.4909	2.2444
WBSgt3	4.89	1.76	0.0030	0.1288	0.1318	7.32	2.30	0.0107	0.7608	0.7715
EAIC	2.35	1.28	0.0003	0.0232	0.0235	9.57	2.57	0.0096	0.4673	0.4769
ARSS	2.79	1.82	0.0005	0.0248	0.0253	9.58	2.58	0.0095	0.4682	0.4777
YBS0	5.54	3.55	0.0057	0.2337	0.2394	6.52	2.74	0.0271	1.0855	1.1126
YBSm	9.01	4.51	0.0078	0.5087	0.5165	11.10	4.08	0.0298	1.2302	1.2600
GCV	2.34	1.27	0.0003	0.0232	0.0235	9.20	2.46	0.0112	0.4725	0.4838

TABLE 4. More results for various selection criteria using different basis functions, $h(X) = 1 + X + X^2$, second DGP. $\lambda_n = 0, c_0 = 2, B = 100, \sigma_W^2 = 1, n = 300, nreplic = 100, k_0 = 6, \rho_X = 0.5, \rho_U = 0.8, \sigma_V^2 = 0.75$

	$\mu(k_n)$	$\sigma(k_n)$	abias^2	avar	amse
	Trigonometric polynomial basis				
k*	1.00	0.00	0.1355	0.1839	0.3194
k**	1.96	1.97	0.0666	0.2206	0.2871
k0	6.00	0.00	0.0485	1.4674	1.5160
EffCV0	5.58	1.28	0.0445	1.3807	1.4251
EffCVm	5.63	4.86	0.0473	0.7069	0.7542
EffCVn	1.94	1.73	0.0641	0.2670	0.3311
ASEBS0	2.13	2.52	0.0746	0.4029	0.4775
WBS	1.00	0.00	0.1355	0.1839	0.3194
WBSgt3	10.58	4.37	0.0944	0.9395	1.0339
EAIC	6.13	4.96	0.0709	0.4697	0.5405
ARSS	7.22	5.13	0.0753	0.5299	0.6052
YBS0	4.04	2.65	0.0427	0.6048	0.6475
YBSm	7.51	5.45	0.0863	0.6864	0.7727
GCV	4.31	4.03	0.0592	0.3989	0.4581