Nonparametric Sieve Regression: Least Squares, Averaging Least Squares, and Cross-Validation

Bruce E. Hansen*
University of Wisconsin†

www.ssc.wisc.edu/~bhansen

May 2012
Revised: August 2012

Abstract

This chapter concerns selection and combination of nonparametric sieve regression estimators. We review the concepts of series and sieve approximations, introduce least-squares estimates of sieve approximations, and measure the accuracy of the estimators by integrated mean-squared error (IMSE). We show that the critical issue in applications is selection of the order of the sieve, as the IMSE greatly varies across the choice. We introduce the cross-validation criterion as an estimator of mean-squared forecast error (MSFE) and IMSE. We extend the current optimality theory, by showing that cross-validation selection is asymptotically IMSE equivalent to the infeasible best sieve approximation.

We also introduce weighted averages of sieve regression estimators. Averaging estimators have lower IMSE than selection estimators. Following Hansen and Racine (2012) we introduce a cross-validation (or jackknife) criterion for the weight vector, and recommend selection of the weights by minimizing this criterion. The resulting jackknife model averaging (JMA) estimator is a feasible averaging sieve estimator. We show that the JMA estimator is optimal in the sense that it is asymptotically IMSE equivalent to the infeasible optimal weighted average sieve estimator. While computation of the JMA weights is a simple application of quadratic programming, we also introduce a simple algorithm which closely approximates the JMA solution without the need for quadratic programming.

*Research supported by the National Science Foundation. I thank Chu-An Liu for excellent research assistance, Guido Kuersteiner and Ryo Okui for pointing out an error in Hansen and Racine (2012), and the referee and editor for helpful comments and suggestions.

†Department of Economics, 1180 Observatory Drive, University of Wisconsin, Madison, WI 53706.

1 Introduction

One of the most popular nonparametric techniques in applied econometric analysis is sieve regression. A sieve is sequence of finite-dimensional models of increasing complexity. The most common examples of sieve regression are polynomials and splines. For a fixed order of complexity, the model can be estimated by classical (parametric) methods. An important difference with parametric regression is the the order of the sieve (the number of regressors) must be selected. This fundamentally changes both the distributional theory and applied practice.

In this chapter we consider selection and combination of nonparametric sieve regression estimators. We review the concepts of series and sieve approximations, introduce least-squares estimates of sieve approximations, and measure the accuracy of the estimators by integrated mean-squared error (IMSE). We show that a critical issue in applications is the order of the sieve, as the IMSE greatly varies across the choice.

We develop the relationship between IMSE and mean-squared forecast error (MSFE), and introduce the cross-validation criterion as an estimator of mean-squared forecast error (MSFE) and IMSE. A major theoretical contribution is that we show that selection based on cross-validation is asymptotically equivalent (with respect to IMSE) to estimation based on the infeasible best sieve approximation. This is an important extension of the theory of cross-validation, which currently has only established optimality with respect to conditional squared error.

We also introduce averaging estimators, which are weighted averages of sieve regression estimators. Averaging estimators have lower IMSE than selection estimators. The critical applied issue is the selection of the averaging weights. Following Hansen and Racine (2012) we introduce a cross-validation criterion for the weight vector, and recommend selection of the weights by minimizing this criterion. The resulting estimator – jackknife model averaging (JMA) – is a feasible averaging sieve estimator. We show that the JMA estimator is asymptotically optimal in the sense that it is asymptotically equivalent (with respect to IMSE) to the infeasible optimal weighted average sieve estimator. Computation of the JMA weights is a simple application of quadratic programming. We also introduce a simple algorithm which closely approximates the JMA solution without the need for quadratic programming.

Sieve approximation has a long history in numerical analysis, statistics, and econometrics. See Chui (1992) and de Boor (2001) for numerical properties of splines, Grenander (1981) for the development of the theory of sieves, Li and Racine (2007) for a useful introduction for econometricians, and Chen (2007) for a review of advanced econometric theory.

Nonparametric sieve regression has been studied by Andrews (1991a) and Newey (1995, 1997), including asymptotic bounds for the IMSE of the series estimators.

Selection by cross-validation was introduced by Stone (1974), Allen (1974), Wahba and Wold (1975), and Craven and Wahba (1979). The optimality of cross-validation selection was investigated by Li (1987) for homoskedastic regression and Andrews (1991b) for heteroskedastic regression. These authors established that the selected estimated is asymptotically equivalent to the infeasible best estimator, where "best" is defined in terms of conditional squared error.

Averaging estimators for regression models was introduced by Hansen (2007). A cross-validation (jacknife) method for selecting the averaging weights was introduced by Hansen and Racine (2012).

The organization of this chapter is as follows. Section 2 introduces nonparametric sieve regression, and Section 3 sieve approximations. Section 4 introduces the sieve regression model and least-squares estimation. Section 5 derives the IMSE of the sieve estimators. Section 6 is a numerical illustration of how the sieve order is of critical practical importance. Section 7 develops the connection between IMSE and MSFE. Section 8 introduces cross-validation for sieve selection. Section 9 presents the theory of optimal cross-validation selection. Section 10 is a brief discussion of how to pre-select the number of models, and Section 11 discusses alternative selection criteria. Section 12 is a continuation of the numerical example. Section 13 introduces averaging regression estimators, and Section 14 introduces the JMA averaging weights and estimator. Section 15 introduces methods for numerical computation of the JMA weights. Section 16 presents an optimality result for JMA weight selection. Section 17 is a further continuation of the numerical example. Section 18 concludes. Regularity conditions for the theorems are listed in Section 19, and the proofs of the theoretical results are presented in Section 20. Computer programs which create the numerical work is available on my webpage www.ssc.wisc.edu/~bhansen.

2 NonParametric Sieve Regression

Suppose that we observe a random sample (y_i, x_i), $i = 1, ..., n$, with y_i real-valued and $x_i \in \mathcal{X}$ possibly vector valued with $\mathcal{X}$ compact and density $f(x)$. We are interested in estimating the regression of y_i on x_i, that is, the conditional mean $g(x) = \mathbb{E}(y \mid x)$, which is identified almost surely if $\mathbb{E}|y| < \infty$. We can write the regression equation as

$$y_i = g(x_i) + e_i \tag{1}$$

$$\mathbb{E}(e_i \mid x_i) = 0. \tag{2}$$

The regression problem is nonparametric when $g(x)$ cannot be summarized by a finite set of parameters.

Notice that equations (1)-(2) do not impose any restrictions on the regression function $g(x)$ nor the regression error e_i (such as conditional homoskedasticity). This is because in a nonparametric context the goal is to be minimalistic regarding parametric assumptions. To develop distributional approximations for estimators it will be necessary to impose some smoothness and moment restrictions. But these restrictions are technical regularity conditions, not fundamental features of the nonparametric model.

A sieve expansion for $g(x)$ is a sequence of finite dimensional models $g_m(x)$, $m = 1, 2, ...$, with

increasing complexity. Particularly convenient are linear sieves, which take the form

$$\begin{aligned} g_m(x) &= \sum_{j=1}^{K_m} z_{jm}(x)\beta_{jm} \\ &= Z_m(x)'\beta_m \end{aligned}$$

where $z_{jm}(x)$ are (nonlinear) functions of x. The number of terms K_m indexes the complexity of the approximation, and plays an important role in the theory. Given a sieve expansion $Z_m(x)$ we define the $K_m \times 1$ regressor vector $z_{mi} = Z_m(x_i)$.

An important special case of a sieve is a series expansion, where the terms $z_{jm}(x)$ are not a function of the sieve order m. For example, a polynomial series expansion is obtained by setting $z_j(x) = x^{j-1}$. When the sieve is a series expansion then the models are nested in the sense that $m_2 > m_1$ implies that $g_{m_2}(x)$ contains $g_{m_1}(x)$ as a special case.

While polynomial series expansions are quite well known, better approximations can be typically achieved by a spline. A spline is a piecewise continuous polynomial, constrained to be smooth up to the order of the polynomial. There is more than one way to write out the basis of a regression spline. One convenient choice takes the form

$$g_m(x) = \sum_{j=0}^{p} x^j \beta_{jm} + \sum_{j=1}^{m} \beta_{p+j}(x - t_j)^p 1\left(x \geq t_j\right). \tag{3}$$

Here, p is the order of the polynomial. There are m constants $t_1, ..., t_m$ called knots which are the join points between the piecewise polynomials. Splines thus have $K_m = p + 1 + m$ coefficients, and a spline has similar flexibility to a $(p+m)$'th order polynomial. Splines require a rule to determine the location of the knots t_j. A common choice is to set the knots to evenly partition the support of x_i. An alternative is to set the knots to evenly partition the percentiles of the distribution of x_i (that is, if $m = 3$ then set t_1, t_2 and t_3 to equal the 25'th, 50'th, and 75'th percentile, respectively).

Typically, the order p of the spline is pre-selected based on desired smoothness (linear, quadratic and cubic are typical choices), and the number of knots m are then selected to determine the complexity of the approximation.

If the knots are set evenly, then the sequence of spline sieves with $m = 1, 2, 3, ...$, are non-nested in the sense that $m_2 > m_1$ does not implies that $g_{m_2}(x)$ contains $g_{m_1}(x)$. However, a sequence of splines can be nested if the knots are set sequentially, or if they are set to partition evenly but the number of knots doubled with each sequential sieve, that is, if we consider the sequence $m = 1, 2, 4, 8, ...$

In a given sample with n observations, we consider a set of sieves $g_m(x)$ for $m = 1, .., M_n$, where M_n can depend on sample size. For example, the set of sieve expansions could be the set of p'th order polynomials for $p = 1, ..., M$. Or alternatively, the sieve could be the set of p'th order splines with m knots, for $m = 0, 1, ..., M - 1$.

3 Sieve Approximation

We have been using the notation β_m to denote the coefficients of the m'th sieve approximation, but how are they defined? There is not a unique definition, but a convenient choice is the best linear predictor

$$\begin{aligned}\beta_m &= \underset{\beta}{\operatorname{argmin}}\, \mathbb{E}\left(y_i - z_{mi}'\beta\right)^2 \\ &= \left(\mathbb{E}\left(z_{mi}z_{mi}'\right)\right)^{-1}\mathbb{E}\left(z_{mi}y_i\right). \end{aligned} \tag{4}$$

Given β_m, define the approximation error

$$r_m(x) = g(x) - Z_m(x)'\beta_m,$$

set $r_{mi} = r_m(x_i)$ and define the expected squared approximation error

$$\phi_m^2 = \mathbb{E}r_{mi}^2 = \int r_m(x)^2 f(x)dx.$$

ϕ_m^2 measures the quality of $g_m(x)$ as an approximation to $g(x)$ in the sense that a smaller ϕ_m^2 means a better approximation. Notice that ϕ_m^2 is the variance of the projection error from the population regression of the true regression function $g(x_i)$ on the sieve regressors z_{mi}

$$\phi_m^2 = \int g(x)^2 f(x)dx - \int g(x)Z_m(x)'f(x)dx\left(\int Z_m(x)Z_m(x)'f(x)dx\right)^{-1}\int Z_m(x)g(x)f(x)dx.$$

It therefore follows that for nested series approximations, ϕ_m^2 is monotonically decreasing as K_m increases. That is, larger models mean smaller approximation error.

Furthermore, we can describe the rate at which ϕ_m^2 decreases to zero. As discussed on page 150 of Newey (1997), if $\dim(x) = q$ and $g(x)$ has s continuous derivatives, then for splines and power series there exists an approximation $\beta' z_m(x)$ such that $\left|g(x) - \beta' z_m(x)\right| = O\left(K_m^{-s/q}\right)$, uniformly in x. Thus

$$\phi_m^2 = \inf_\beta \mathbb{E}\left(g(x_i) - \beta' z_m(x_i)\right)^2 \leq \inf_\beta \sup_x \left|g(x) - \beta' z_m(x)\right|^2 \leq O\left(K_m^{-2s/q}\right).$$

This shows that the magnitude of the approximation error depends on the dimensionality and smoothness of $g(x)$. Smoother functions $g(x)$ can be approximated by a smaller number of series terms K_m, so the rate of convergence is increasing in the degree of smoothness.

4 Sieve Regression Model and Estimation

As we have described, for each sieve approximation there are a set of regressors z_{mi} and best linear projection coefficient β_m. The sieve regression model is then

$$y_i = z_{mi}'\beta_m + e_{mi} \tag{5}$$

where e_{mi} is a projection error and satisfies

$$\mathbb{E}\left(z_{mi}e_{mi}\right) = 0.$$

It is important to recognize that e_{mi} is defined by this construction, and it is therefore inappropriate to *assume* properties for e_{mi}. Rather they should be derived.

Recall that the approximation error is $r_{mi} = r_m(x_i) = g(x_i) - z_{mi}'\beta_m$. Since the true regression (1) is $y_i = g(x_i) + e_i$, it follows that the projection error is $e_{mi} = e_i + r_{mi}$, the sum of the true regression error e_i and the sieve approximation error r_{mi}.

The least-squares (LS) estimator of equation (5) is

$$\begin{aligned} \widehat{\beta}_m &= \left(\sum_{i=1}^n z_{mi}z_{mi}'\right)^{-1}\sum_{i=1}^n z_{mi}y_i \\ \widehat{g}_m(x) &= Z_m(x)'\widehat{\beta}_m. \end{aligned}$$

Least squares is an appropriate estimator as β_m is defined as the best linear predictor. The least-squares estimator is a natural moment estimator of the projection coefficient β_m.

5 Integrated Mean Squared Error

As a practical matter, the most critical choice in a series regression is the number of series terms. The choice matters greatly, and can have a huge impact on the empirical results.

Statements such as "the number of series terms should increase with the sample size" do not provide any useful guidance for practical selection. Applied nonparametric analysis needs practical, data-based rules. Fortunately, there are sound theoretical methods for data-dependent choices.

The foundation for a data-dependent choice is a (theoretical) criterion which measures the performance of an estimator. The second step is to constuct an estimator of this criterion. Armed with such an estimate, we can select the number of series terms or weights to minimize the empirical criterion.

Thus to start, we need a criterion to measure the performance of a nonparametric regression estimator. There are multiple possible criteria, but one particularly convenient choice is integrated

mean squared error (IMSE). For a sieve estimator $\widehat{g}_m(x)$ the IMSE equals

$$IMSE_n(m) = \int \mathbb{E}\left(\widehat{g}_m(x) - g(x)\right)^2 f(x)dx.$$

Using the fact that $\widehat{g}_m(x) - g(x) = z_m(x)'\left(\widehat{\beta}_m - \beta_m\right) - r_m(x)$ we can calculate that

$$\begin{aligned} & \int \left(\widehat{g}_m(x) - g(x)\right)^2 f(x)dx \\ = & \int r_m(x)^2 f(x)dx - 2\left(\widehat{\beta}_m - \beta_m\right)' \int x_m(x) r_m(x) f(x)dx \\ & + \left(\widehat{\beta}_m - \beta_m\right)' \int z_m(x) z_m(x)' f(x)dx \left(\widehat{\beta}_m - \beta_m\right). \end{aligned}$$

Note the the first term equals the expected squared approximation error ϕ_m^2. The second term is zero because $\int x_m(z) r_m(z) f(z)dz = \mathbb{E}\left(z_{mi} r_{mi}\right) = 0$. Defining $Q_m = \mathbb{E}\left(z_{mi} z_{mi}'\right)$, we can write

$$IMSE_n(m) = \phi_m^2 + \operatorname{tr}\left[Q_m \mathbb{E}\left(\left(\widehat{\beta}_m - \beta_m\right)\left(\widehat{\beta}_m - \beta_m\right)'\right)\right].$$

Asymptotically, $\mathbb{E}\left(\left(\widehat{\beta}_m - \beta_m\right)\left(\widehat{\beta}_m - \beta_m\right)'\right) \simeq \frac{1}{n} Q_m^{-1} \Omega_m Q_m^{-1}$ where $\Omega_m = \mathbb{E}\left(z_{mi} z_{mi}' \sigma_i^2\right)$ and $\sigma_i^2 = \mathbb{E}\left(e_i^2 \mid x_i\right)$. Making these substitutions, we expect that $IMSE_n(m)$ should be close to

$$IMSE_n^*(m) = \phi_m^2 + \frac{1}{n} \operatorname{tr}\left(Q_m^{-1} \Omega_m\right). \tag{6}$$

The second term in (6) is the integrated asymptotic variance. Under conditional homoskedasticity $\mathbb{E}\left(e_i^2 \mid x_i\right) = \sigma^2$ we have the simplification $\Omega_m = \mathbb{E}\left(z_{mi} z_{mi}'\right)\sigma^2 = Q_m \sigma^2$. Thus in this case $\frac{1}{n} \operatorname{tr}\left(Q_m^{-1} \Omega_m\right) = \sigma_m^2 K_m / n$, a simple function of the number of coefficients and sample size. That is, homoskedasticity implies the following simplification of (6)

$$IMSE_n^*(m) = \phi_m^2 + \sigma^2 \frac{K_m}{n}.$$

However, in the general case of conditional heteroskedasticity, (6) is the appropriate expression.

Hansen (2012) showed that $IMSE_n(m)$ and $IMSE_n^*(m)$ are uniformly close under quite general regularity conditions, listed in Section 19.

Theorem 1 *Under Assumption 1, uniformly across $m \leq M_n$,*

$$IMSE_n(m) = IMSE_n^*(m)(1 + o(1))$$

This shows that $IMSE_n^*(m)$ is a good approximation to $IMSE_n(m)$.

6 The Order of the Approximation Matters

The way that nonparametric methods are often presented, some users may have received the false impression that the user is free to select the order of the approximation m. So long as m increases with n, the method works, right? Unfortunately it is not so simple in practice. Instead, the actual choice of m in a given application can have large and substantive influence on the results.

To illustrate this point, we take a simple numerical example. We consider the following data generating process.

$$\begin{aligned} y_i &= g(x_i) + e_i \\ g(x) &= a \sin\left(2\pi x + \frac{\pi}{4}\right) \\ x_i &\sim U[0,1] \\ e_i &\sim N(0, \sigma_i^2) \\ \sigma_i^2 &= \sqrt{5} x_i^2 \end{aligned}$$

This is a simple normal regression with conditional heteroskedasticity. The parameter a is selected to control the population $R^2 = a^2/(2+a^2)$, and we vary $R^2 = 0.25$, 0.5, 0.75 and 0.9. We vary the sample size from $n = 50$ to 1000.

We consider estimation of $g(x)$ using quadratic splines, ranging the number of knots m from 1 to 5. For each R^2, n, and m, the integrated mean-squared error (IMSE) is calculated and displayed in Figure 1 as a function of sample size using a logarithmic scale. The four displays are for the four values of R^2, and each line corresponds to a different number of knots. Thus each line corresponds to a distinct sieve approximation m. To render the plots easy to read, the IMSE has been normalized by the IMSE of the infeasible optimal averaging estimator. Thus the reported IMSE's are multiples of the infeasible best.

One striking feature of Figure 1 is the strong variation with m. That is, for a given R^2 and n, the IMSE varies considerably across estimators. For example, take $n = 200$ and $R^2 = 0.25$. The relative IMSE ranges from about 1.4 (2 knots) to 2.1 (5 knots). Thus the choice really matters. Another striking feature is that the IMSE rankings strongly depend on unknowns. For example, again if $n = 200$ but we consider $R^2 = 0.9$ then the sieve with two knots performs quite poorly with IMSE=2.9, while the sieve with 5 knots has an relative IMSE of about 1.5.

A third striking feature is that the IMSE curves are U-shaped functions of the sample size n. When they reach bottom they tend to be the sieve with the lowest IMSE. Thus if we fix R^2 and vary n from small to large, we see how the best sieve is increasing. For example, take $R^2 = 0.25$. For $n = 50$, the lowest IMSE is obtained by the spline with one knot. The one-knot spline has the lowest IMSE until $n = 150$, at which point the two-knot spline has the lowest IMSE. The three-knot spline has lower IMSE for $n \geq 800$. Or, consider the case $R^2 = 0.75$. In this case, the two-knot spline has the lowest IMSE for $n < 100$, while the three-knot spline is best for $100 \leq n \leq 400$, with the four-knot spline for $n \leq 600$.

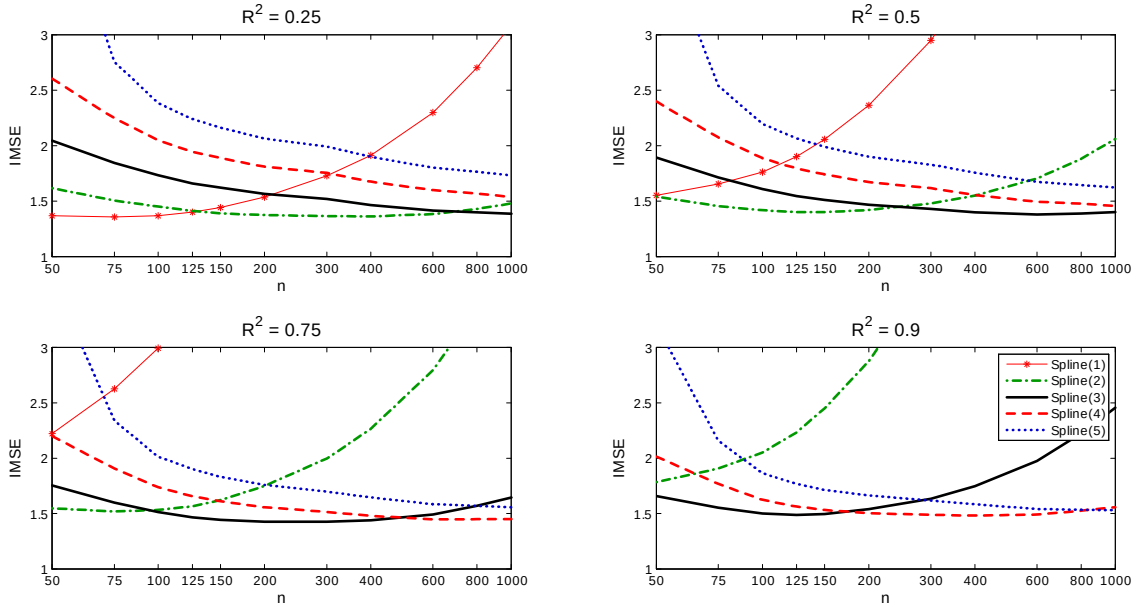

Figure 1: Integrated Mean-Squared Error of Spline Regression Estimators

The overall message is that the order of the series approximation matters, and it depends on features which we know (such as the sample size n) but also features that we do not know. Data-dependent methods for selection of m are essential, otherwise the selection between the estimators is arbitrary.

7 Mean Squared Forecast Error

A concept related to IMSE is the mean-squared-forecast-error (MSFE). This is the expected squared error from the prediction of an out-of-sample observation. Specifically, let (y_{n+1}, x_{n+1}) be an out-of-sample observation drawn from the same distribution as the in-sample observations. The forecast of y_{n+1} given x_{n+1} is $\widehat{g}_m(x_{n+1})$. The MSFE is the expected squared forecast error

$$MSFE_n(m) = \mathbb{E}\left(y_{n+1} - \widehat{g}_m(x_{n+1})\right)^2$$

which depends on the sample size n as well as the estimator $\widehat{g}_m$.

Making the substitution $y_{n+1} = g(x_{n+1}) + e_{n+1}$ and using the fact that e_{n+1} is independent of $g(x_{n+1}) - \widehat{g}_m(x_{n+1})$, we can calculate that the MSFE equals

$$MSFE_n(m) = \mathbb{E}\left(e_{n+1}^2\right) + \mathbb{E}\left(g(x_{n+1}) - \widehat{g}_m(x_{n+1})\right)^2.$$

The second term on the right is an expectation over the random vector x_{n+1} and the estimator $\widehat{g}_m(x)$, which are independent since the estimator is a function only of the in-sample observations.

We can write the expectation over x_{n+1} as an integral with respect to its marginal density $f(x)$, thus

$$\begin{aligned} MSFE_n(m) &= \mathbb{E}\left(e_{n+1}^2\right) + \int \mathbb{E}\left(\widehat{g}_m(x) - g(x)\right)^2 f(x)dx \\ &= \mathbb{E}\left(e_{n+1}^2\right) + IMSE_n(m). \end{aligned}$$

Thus $MSFE_n(m)$ equals $IMSE_n(m)$ plus $\mathbb{E}\left(e_{n+1}^2\right)$. Notice that $\mathbb{E}\left(e_{n+1}^2\right)$ does not depend on the estimator $\widehat{g}_m(x)$. Thus ranking estimators by MSFE and IMSE are equivalent.

8 Cross-Validation

Ideally, we want to select the estimator m which minimizes $IMSE_n(m)$ or equivalently $MSFE_n(m)$. However, the true MSFE is unknown. In this section we show how to estimate the MSFE.

Observe that

$$MSFE_n(m) = \mathbb{E}\left(\tilde{e}_{m,n+1}^2\right)$$

where $\tilde{e}_{m,n+1} = y_{n+1} - \widehat{g}_m(x_{n+1})$. This is a prediction error. Estimation is based on the sample $(y_i, x_i) : i = 1, ..., n$, and the error calculated on the out-of-sample observation $n + 1$. Thus $MSFE_n(m)$ is the expectation of a squared leave-one-out prediction error from a sample of length $n + 1$.

For each observation i, we can create a similar leave-one-out prediction error. For each i we can create a pseudo-prediction error by estimating the coefficients using the observations excluding i. That is, define the leave-one-out estimator

$$\widehat{\beta}_{m,-i} = \left(\sum_{j\neq i} z_{mj} z_{mj}'\right)^{-1} \sum_{j\neq i} z_{mj} y_j \tag{7}$$

and prediction error

$$\tilde{e}_{mi} = y_i - z_{mi}'\widehat{\beta}_{m,-i}. \tag{8}$$

The only difference between $\tilde{e}_{m,n+1}$ and $\tilde{e}_{mi}$ is that the former is based on the extended sample of length $n + 1$ while the latter are based on a sample of length n. Otherwise they have the same construction. It follows that for each i, $\mathbb{E}\tilde{e}_{mi}^2 = MSFE_{n-1}(m)$. Similarly, the sample average, known as the *cross-validation criterion*

$$CV_n(m) = \frac{1}{n}\sum_{i=1}^{n} \tilde{e}_{mi}^2$$

also has mean $MSFE_{n-1}(m)$. This is a natural moment estimator of $MSFE_{n-1}(m)$.

We have established the following result.

Theorem 2 $\mathbb{E}CV_n(m) = MSFE_{n-1}(m)$

As $MSFE_{n-1}(m)$ should be very close to $MSFE_n(m)$, we can view $CV_n(m)$ as a nearly unbiased estimator of $MSFE_n(m)$.

Computationally, the following algebraic relationship is convenient.

Proposition 1 *$\tilde{e}_{mi} = \hat{e}_{mi}(1 - h_{mi})^{-1}$, where $\hat{e}_{mi} = y_i - z_{mi}'\widehat{\beta}_m$ are the least-squares residuals and $h_{mi} = z_{mi}'\left(\sum_{i=1}^n z_{mi}z_{mi}'\right)^{-1} z_{mi}$ are known as the leverage values.*

While Proposition 1 is well known, we include a complete proof in Section 20 for completeness.

Proposition 1 directly implies the simple algebraic expression

$$CV_n(m) = \frac{1}{n}\sum_{i=1}^n \frac{\hat{e}_{mi}^2}{(1 - h_{mi})^2}. \tag{9}$$

This shows that for least squares estimation, cross-validation is a quite simple calculation, and does not require the explicit leave-one-out operations suggested by (7).

The estimator $\widehat{m}$ which is selected by cross-validation is the one with the smallest value of $CV(m)$. We can write this as

$$\widehat{m} = \underset{1\le m\le M_n}{\operatorname{argmin}}\ CV_n(m)$$

Computationally, we estimate each series regression $m = 1, ..., M_n$, compute the residuals $\hat{e}_{mi}$ for each, the CV criterion $CV_n(m)$ using (9) and then find $\widehat{m}$ as the value which yields the smallest value of $CV_n(m)$.

It is useful to plot $CV_n(m)$ against m to visually check if there are multiple local minima or flat regions. In these cases some statisticians have argued that it is reasonable to select the most parsimonious local minima or the most parsimonious estimator among near-equivalent values of the CV function. The reasons are diverse, but essentially the cross-validation function can be quite a noisy estimate of the IMSE, especially for high-dimensional models. The general recommendation is to augment automatic model-selection with visual checks and judgment.

To illustrate, Figure 2 plots the cross-validation function for one of the samples from Section 6. The cross-validation function is sharply decreasing until 2 knots, then flattens out, with the minimum $m = 2$ knots. In this particular example, the sample was drawn from the DGP of Section 6 with $n = 200$ and $R^2 = 0.5$. From Figure 1 we can see that the lowest IMSE is obtained by $m = 2$, so indeed the CV function is a constructive guide for selection.

9 Asymptotic Optimality of Cross-Validation Selection

Li (1987), Andrews (1991b) and Hansen and Racine (2012) have established conditions under which the CV selected estimator is asymptotically optimal, in the sense that the selected model is asymptotically equivalent to the infeasible optimum. The criterion they used to assess optimality

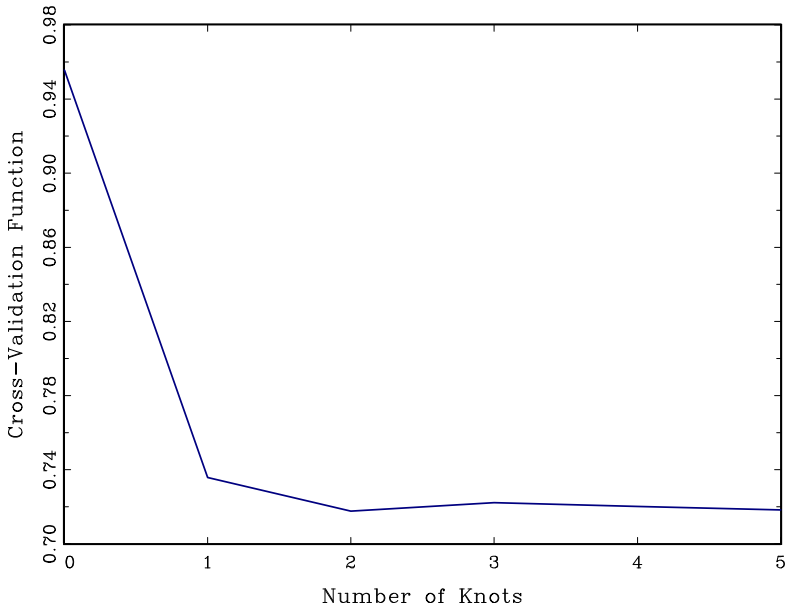

Figure 2: Typical Cross-Validaion Function, $n = 200$

is the conditional squared error fit

$$R_n(m) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left((\widehat{g}_m(x_i) - g(x_i))^2 \mid X \right) \tag{10}$$

where $X = \{x_1, ..., x_n\}$. This is similar to IMSE, but only assesses fit on the support points of the data. In contrast, the literature on sieve approximations focuses on IMSE. We now extend the asymptotic optimality theory and show that the CV selected estimator is asymptotically optimal with respect to IMSE.

Theorem 3 *Under Assumptions 1, 2, and 3, as* $n \to \infty$,

$$\left| \frac{IMSE_n(\widehat{m})}{\inf\limits_{1 \le m \le M_n} IMSE_n(m)} \right| \xrightarrow{p} 1$$

The assumptions and proof are presented in Sections 19 and 20, respectively.

Theorem 3 shows that in large samples, the IMSE of the CV-selected estimator $\widehat{g}_{\widehat{m}}(x)$ is equivalent to the IMSE of the infeasible best estimator in the class $\widehat{g}_m(x)$ for $1 \le m \le M_n$. This is an oracle property for cross-validation selection.

A critical assumption for Theorem 3 is that $\phi_m^2 > 0$ for all $m < \infty$ (Assumption 2 in Section 20). Equivalently, the approximation error is non-zero for all finite-dimensional models – that all models are approximations. If instead one of the finite-dimensional models is the true conditional mean (so that $\phi_m^2 = 0$ for some model m) then cross-validation asymptotically over-selects the model order with positive probability and is thus asymptotically sub-optimal. In this context consistent

model selection methods (such as BIC) are optimal. This classification was carefully articulated in the review paper by Shao (1997). Some researchers refers to cross-validation as a *conservative* selection procedure (it is optimal for the broad class of nonparametric models) and to BIC as a consistent selection procedure (it selects the correct model when it is truly finite dimensional).

10 Pre-Selection of the Number of Models

To implement cross-validation selection, a user first has to select the set of models $m = 1, ..., M_n$ over which to search. For example, if using a power series approximation, a user has to first determine the highest power, or if using a spline a user has to determine the order of the spline and the maximum number of knots. This choice affects the results, but unfortunately there is no theory about how to select these choices. What we know is that the assumptions restrict both the number of estimated parameters in each model K_m and the number of models M_n relative to sample size. Specifically, Assumption 1.5 specifies that for a power series $K_m^4/n = O(1)$ and for a spline sieve $K_m^3/n = O(1)$, uniformly for $m \leq M_n$. These conditions may be stronger than necessary, but they restrict the number of estimated parameters to be increasing at a rate much slower than sample size. Furthermore, Assumption 3.2 allows non-nested models, but controls the number of models. While these conditions do not give us precise rules for selecting the initial set of models, they do suggest that we should be reasonably parsimonious and not too aggressive in including highly parameterized models.

Unfortunately, these comments still do not give precise guidance on how to determine the number of models M_n. It may be a useful subject for future research to construct and justify data-dependent rules for determining M_n.

11 Alternative Selection Criteria

We have discussed the merits of cross-validation to select the sieve approximation, but many other selection methods have been proposed. In this section we briefly describe the motivation and properties of some of these alternative criteria.

The Mallows criterion (Mallows, 1973)

$$Mallows(m) = \frac{1}{n}\sum_{i=1}^{n} \hat{e}_{mi}^2 + 2\tilde{\sigma}^2 K_m$$

with $\tilde{\sigma}^2$ a preliminary estimate of $\mathbb{E}\left(e_i^2\right)$ is an alternative estimator of the IMSE under the additional assumption of conditional homoskedasticity $\mathbb{E}\left(e_i^2 \mid x_i\right) = \sigma^2$. Li (1987) provided conditions under which Mallows selection is asymptotically optimal, but Andrews (1991b) shows that its optimality fails under heteroskedasticity.

The Akaike information criterion (Akaike, 1973)

$$AIC(m) = n \log \left(\frac{1}{n} \sum_{i=1}^{n} \hat{e}_{mi}^2 \right) + 2K_m$$

is an estimate of the Kullback-Leibler divergence between the estimated Gaussian model and the true model density. AIC selection has similar asymptotic properties as Mallows selection, in that it is asymptotically optimal under conditional homoskedasticity but not under heteroskedasticity.

The corrected AIC (Hurvich and Tsai, 1989)

$$AIC_c(m) = AIC(m) + \frac{2K_m (K_m + 1)}{n - K_m - 1}$$

is a finite-sample unbiased estimate of the Kullback-Leibler divergence under the auxiliary assumption that the errors e_i are independent and Gaussian. Its asymptotic properties are the same as AIC, but has improved finite-sample performance, especially when the model dimension K_m is high relative to sample size.

The Bayesian information criterion (Schwarz, 1978)

$$BIC(m) = n \log \left(\frac{1}{n} \sum_{i=1}^{n} \hat{e}_{mi}^2 \right) + \log(n) K_m$$

is an approximation to the log posterior probability that model m is the true model, under the auxiliary assumption that the errors are independent and Gaussian, the true model is finite dimension, the models have equal prior probability, and priors for each model m are diffuse. BIC selection has the property of *consistent model selection*: When the true model is a finite dimensional series, BIC will select that model with probability approaching one as the sample size increases. However, when there is no finite dimensional true model, then BIC tends to select overly parsimonious models (based on IMSE).

The above methods are all information criteria, similar in form to cross-validation. A different approach to selection is the class of penalized least squares estimators. Let z_i denote the $K_n \times 1$ vector of all potential regressors in all models, let $\beta = (\beta_1, ..., \beta_{K_n})$ denote its projection coefficient, and define the penalized least squares criteria

$$P_n(\beta, \lambda) = \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - z_i' \beta \right)^2 + \sum_{j=1}^{K_n} p_\lambda \left(\beta_j \right)$$

and the PLS estimator

$$\widehat{\beta}_\lambda = \underset{\beta}{\operatorname{argmin}} \, P_n(\beta, \lambda)$$

where $p_\lambda (u)$ is a non-negative symmetric penalty function and λ is a tuning parameter.

The choice of $p_\lambda (u)$ determines the estimator. In the recent literature a popular choice is

$p_\lambda(|u|) = \lambda|u|$ which yields the LASSO (least absolute shrinkage and selection operator) estimator, proposed by Tibshirani (1996). Different variants of LASSO have been proposed, including SCAD (smoothly clipped absolute deviation) (Fan and Li, 2001), and the adaptive LASSO (Zou, 2006).

PLS estimators are generally appropriate when the dimension of z_i is high (some estimators such as the LASSO are defined even when K_n exceeds n). The LASSO family are selection methods as the estimators $\widehat{\beta}_\lambda$ typically set most individual coefficients to zero. The non-zero coefficient estimates are the selected variables, and the zero coefficient estimates are the excluded variables. SCAD and the adaptive LASSO have optimality (oracle) properties when the true regression function is *sparse*, meaning that the true regression function is a finite dimensional series. When the true regression function is not sparse the properties of LASSO selection are unclear.

Among these methods, selection by cross-validation is uniquely the only method which is asymptotically optimal for general nonparametric regression functions and unknown conditional heteroskedasticity. Most of the other selection methods explicitly or implicity rely on conditional homoskedasticity, and some of the methods rely on sparsity (finite dimensionality), neither of which are generally appropriate for nonparametric estimation.

12 Numerical Simulation

We return to the simulation experiment introduced in Section 6. Recall that we reported the integrated mean-squared error of a set of least-squares estimates of a quadratic spline with given knots. Now we compare the IMSE of estimators which select the number of knots. We consider CV selection, and compare its performance with selection based on the Akaike information criterion (Akaike, 1973) and the Hurvich-Tsai (1989) corrected AIC.

For all methods, we estimate nonparametric quadratic splines with knots $m = 0, 1, ..., M_n$ with $M_n = 4n^{0.15}$. The selection criteria were calculated for each set of knots, and the model selected with the lowest value of the criteria.

We report the IMSE of the three methods in Figure 3 (along with the IMSE of the JMA method, to be discussed below). Again, the IMSE is normalized by the IMSE of the infeasible best averaging estimator, so all results are relative to this infeasible optimum.

One striking feature of this figure is that the three methods (CV, AIC and AIC_c) have similar performance for $n \geq 100$, though CV has slightly lower IMSE, especially for small n.

Another striking feature is that for $n \geq 100$, the IMSE of the selection methods is relatively unaffected by sample size n and the value of R^2. This is especially important when contrasted with Figure 1, where we found that the IMSE of individual sieve estimators depend greatly upon n and R^2. This is good news, it shows that the selection methods are adapting to the unknown features of the sampling distribution.

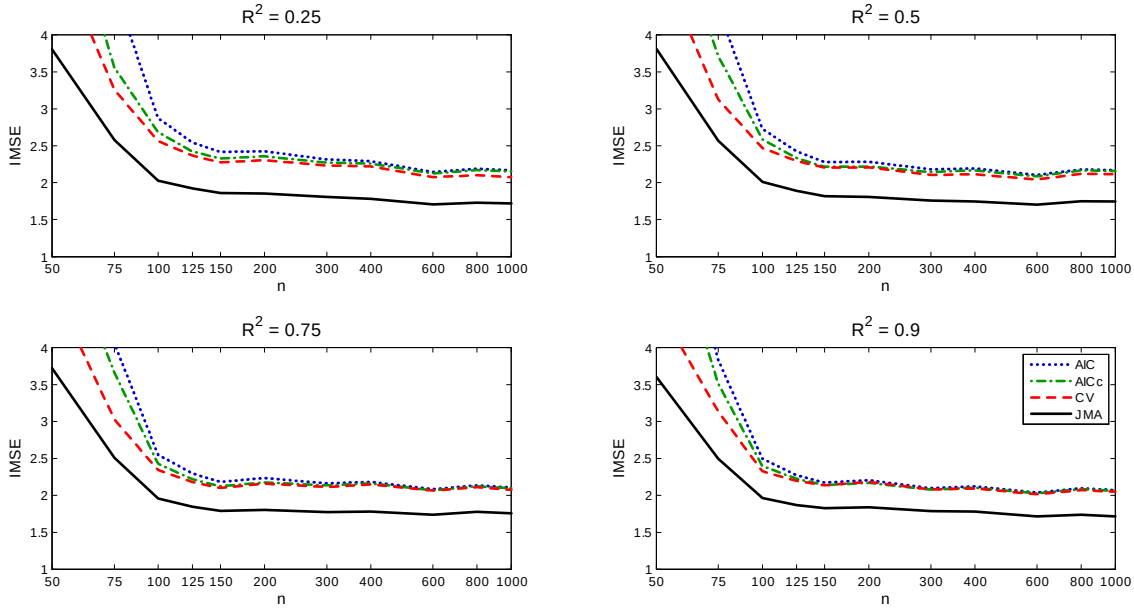

Figure 3: Integrated Mean-Squared Error, Selection and Averaging Estimators

13 Averaging Regression

Let $w = (w_1, w_2, ..., w_M)$ be a set of non-negative weights which sum to one, $\sum_{m=1}^{M} w_m = 1$. An averaging LS estimator is

$$\widehat{g}_w(x) = \sum_{m=1}^{M} w_m \widehat{g}_m(x). \tag{11}$$

The averaging estimator includes the m'th least-squares estimator as a special case by setting w to equal the unit vector with a weight of 1 in the m'th place.

For example, consider a set of spline estimators with $m = 0, 1, 2$, and 3 knots. The averaging estimator takes an average of these four estimators. In general, averaging is a smoother function of the data than selection, and smoothing generally reduces variance. The reduction in variance can result in estimators with lower IMSE.

We define the IMSE of the averaging estimator as

$$IMSE_n(w) = \int \mathbb{E}\left(\widehat{g}_w(x) - g(x)\right)^2 f(x)dx$$

which is a function of the weight vector.

It is recommended to constrain the weights w_m to be non-negative, that is, $w_m \geq 0$. In this case the weight vector w lies on $\mathcal{H}$, the unit simplex in $\mathbb{R}^{M_n}$. This restriction may not be necessary, but some bounds on the weights are required. Hansen and Racine (2012) suggested that in the case of nested models non-negativity is a necessary condition for admissibility, but they made a technical error. The actual condition is that $0 \leq \sum_{j=m}^{M} w_j \leq 1$ which is somewhat broader. (I

thank Guido Kuersteiner and Ryo Okui for pointing out this error to me.) It is unclear if this broader condition is compatible with the optimality theory, or what restrictions are permissible in the case of non-nested models.

Hansen (2012) provides an approximation to the IMSE of an averaging estimator.

Theorem 4 *Under Assumptions 1, 2, and 4, uniformly across* $w \in \mathcal{H}$,

$$IMSE_n(w) = IMSE_n^*(w)(1 + o(1))$$

where

$$\begin{aligned} IMSE_n^*(w) &= \sum_{m=1}^{M_n} w_m^2 \left(\phi_m^2 + \frac{1}{n}\operatorname{tr}\left(Q_m^{-1}\Omega_m\right)\right) \\ &\quad + 2\sum_{\ell=1}^{M_n}\sum_{m=1}^{\ell-1} w_\ell w_m \left(\phi_\ell^2 + \frac{1}{n}\operatorname{tr}\left(Q_m^{-1}\Omega_m\right)\right) \\ &= \sum_{m=1}^{M_n} w_m^* n\phi_m^2 + \sum_{m=1}^{M_n} w_m^{**}\operatorname{tr}\left(Q_m^{-1}\Omega_m\right) \end{aligned} \tag{12}$$

and

$$w_m^* = w_m^2 + 2w_m\sum_{\ell=1}^{m-1} w_\ell \tag{13}$$

$$w_m^{**} = w_m^2 + 2w_m\sum_{\ell=m+1}^{M_n} w_\ell. \tag{14}$$

14 JMA for Averaging Regression

The method of cross-validation for averaging regressions is much the same as for selection. First, note that the discussion about the equivalence of mean-square forecast error (MSFE) and IMSE from Section 7 is not specific to the estimation method. Thus it equally applies to averaging estimators. Namely the averaging forecast of y_{n+1} given x_{n+1} is $\widehat{g}_w(x_{n+1})$, with MSFE

$$\begin{aligned} MSFE_n(w) &= \mathbb{E}\left(y_{n+1} - \widehat{g}_w(x_{n+1})\right)^2 \\ &= \mathbb{E}\left(e_{n+1}^2\right) + IMSE_n(w) \end{aligned}$$

where the second equality follows by the same discussion as in Section 7.

Furthermore, the discussion in Section 8 about estimation of MSFE by cross-validation is also largely independent of the estimation method, and thus applies to averaging regression. There are some differences, however, in the algebraic implementation. The leave-one-out averaging prediction

errors are

$$\tilde{e}_{wi} = \sum_{m=1}^{M} w_m \tilde{e}_{mi}$$

where, as before, $\tilde{e}_{mi}$ is defined in (8) and Proposition 1. The cross-validation function for averaging regression is then

$$\begin{aligned} CV_n(w) &= \frac{1}{n}\sum_{i=1}^{n} \tilde{e}_{wi}^2 \\ &= \sum_{m=1}^{M}\sum_{\ell=1}^{M} w_m w_\ell \left(\frac{1}{n}\sum_{i=1}^{n} \tilde{e}_{mi}\tilde{e}_{\ell i}\right) \\ &= w'Sw \end{aligned}$$

where S is an $M \times M$ matrix with $m\ell$'th entry

$$\begin{aligned} S_{m\ell} &= \frac{1}{n}\sum_{i=1}^{n} \tilde{e}_{mi}\tilde{e}_{\ell i} \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{e}_{mi}\hat{e}_{\ell i}}{(1-h_{mi})(1-h_{\ell i})}. \end{aligned}$$

with $\hat{e}_{mi}$ the least-squares residuals for the $m'th$ estimator, and the second equality uses Proposition 1.

$CV_n(w)$ is also the jackknife estimator of the expected squared error, and thus Hansen and Racine (2012) call $CV_n(w)$ the jackknife model averaging (JMA) criterion.

15 Computation

The cross-validation or jackknife choice of weight vector w is the one which minimizes the cross-validation criterion $CV_n(w) = w'Sw$. Since the weights w_m are restricted to be non-negative and sum to one, the vector w lies on the M-dimension unit simplex $\mathcal{H}$, so we can write this problem as

$$\widehat{w} = \underset{w\in\mathcal{H}}{\operatorname{argmin}}\, w'Sw.$$

The weights $\widehat{w}$ are called the JMA weights, and when plugged into the estimator (11) yield the JMA nonparametric estimator

$$\widehat{g}_w(x) = \sum_{m=1}^{M} \widehat{w}_m \widehat{g}_m(x). \tag{15}$$

Since the criterion is quadratic in w and the weight space $\mathcal{H}$ is defined by a set of linear equality and inequality restrictions, this minimization problem is known as a quadratic programming problem. In matrix programming languages solution algorithms are available. For example, $\widehat{w}$ can be easily solved using the qprog command in GAUSS, the quadprog command in MATLAB, or the

quadprog command in R.

In other packages quadratic programming may not be available. However, it is often possible to call the calculation through an external call to a compatible language (for example, calling R from within STATA). This, however, may be rather cumbersome.

However, it turns out that $\widehat{w}$ can be found using a relatively simple set of linear regressions. First, let $\tilde{g}_i = (\tilde{g}_{1i}, ..., \tilde{g}_{Mi})'$ be the $M \times 1$ vector of leave-one-out predicted values for the i'th observation. Then note that $\tilde{e}_{wi} = y_i - \tilde{g}_i' w$, so the CV criterion can be written as

$$\begin{aligned} CV_n(w) &= \frac{1}{n}\sum_{i=1}^{n} \hat{e}_{wi}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} \left(y_i - \tilde{g}_i' w\right)^2. \end{aligned}$$

This is the sum-of-squared error function from a regression of y_i on the vector $\tilde{g}_i$, with coefficients w. Thus the problem of solving for $\hat{w}$ is algebraically equivalent to a constrained least-squares regression of y_i on $\tilde{g}_i$. We can write the least-squares regression as

$$y_i = \tilde{g}_i' w + \tilde{e}_{wi}$$

or in vector notation

$$y = \widetilde{G} w + \tilde{e}_w.$$

where $\widetilde{G}$ is an $n \times M$ matrix whose m'th columnn are the leave-one-out predicted values from the m'th series approximation.

The simple unconstrained least-squares estimator of w

$$\widetilde{w} = \left(\widetilde{G}'\widetilde{G}\right)^{-1} \widetilde{G}' y \tag{16}$$

will satisfy neither the summing up nor non-negativity constraints. To impose the constraint that the coefficients sum to one, letting $\mathbf{1}$ denote an $M \times 1$ vector of ones, then the least-squares estimator subject to the constraint $\mathbf{1}'w = 1$ is

$$\bar{w} = \widetilde{w} - \left(\widetilde{G}'\widetilde{G}\right)^{-1} \mathbf{1} \left(\mathbf{1}' \left(\widetilde{G}'\widetilde{G}\right)^{-1} \mathbf{1}\right)^{-1} \left(\mathbf{1}'\tilde{w} - 1\right). \tag{17}$$

Alternatively, subtract $\tilde{g}_{Mi}$ from y_i and $\tilde{g}_{1i}, ..., \tilde{g}_{M-1,i}$ and run the regression

$$y_i - \tilde{g}_{Mi} = \bar{w}_1 \left(\tilde{g}_{1i} - \tilde{g}_{Mi}\right) + \bar{w}_2 \left(\tilde{g}_{2i} - \tilde{g}_{Mi}\right) + \cdots + \bar{w}_{M-1} \left(\tilde{g}_{M-1,i} - \tilde{g}_{Mi}\right) + \tilde{e}_{wi} \tag{18}$$

and then set $\bar{w}_M = 1 - \sum_{m=1}^{M-1} \bar{w}_m$. (17) and (18) are algebraically equivalent methods to compute $\bar{w}$.

While the weights $\bar{w}$ will sum to one, they will typically violate the non-negativity constraints,

and thus is not a good estimator. However, a simple iterative algorithm will convert $\bar{w}$ to the desired $\widehat{w}$. Here are the steps.

1. If $\bar{w}_m \geq 0$ for all m, then $\widehat{w} = \bar{w}$ and stop.
2. If $\min_m \bar{w}_m < 0$, find the index $\bar{m}$ with the most negative weight $\bar{w}_{\bar{m}}$ (e.g. $\bar{m} = \operatorname{argmin} \bar{w}_m$)
3. Remove the estimator $\bar{m}$ from the set of M estimators. We are left with a set of $M-1$ estimators, with $\widetilde{G}$ an $n \times (M-1)$ matrix
4. Recompute $\widetilde{w}$ and $\bar{w}$ in (16) and (17) using this new $\widetilde{G}$.
5. Go back to Step 1 and iterate until all weights are non-negative.

This is a simple algorithm, and has at most M iteration steps, where M are the number of initial estimators, and is thus quite efficient. It is simple enough that it can be computed using simple least-squares methods and thus can be used in many packages.

16 Asymptotic Optimality of JMA Averaging

Hansen and Racine (2012) have established conditions under which the JMA weights are asymptotically optimal, in the sense that the selected averaging estimator is asymptotically equivalent to the infeasible optimal weights. They established optimality with respect to the conditional squared error fit (10). We now show that this can be extended to optimality with respect to IMSE.

As in Hansen (2007) and Hansen and Racine (2012), we only establish optimality with respect to a discrete set of weights. For some integer $N \geq 1$, let the weights w_j take values from the set $\{0, \frac{1}{N}, \frac{2}{N}, ..., 1\}$, and let $\mathcal{H}_n$ denote the subset of the unit simplex $\mathcal{H}$ restricted to these points. If N is large then this is not restrictive. This restriction is for technical reasons and does not affect how the method is implemented in practical applications.

Theorem 5 *Under Assumptions 1-4, as* $n \to \infty$,

$$\left| \frac{IMSE_n(\widehat{w})}{\inf\limits_{w \in \mathcal{H}_n} IMSE_n(w)} \right| \xrightarrow{p} 1$$

The assumptions and proof are presented in Sections 19 and 20, respectively.

Theorem 5 shows that in large samples, the IMSE of the JMA estimator $\widehat{g}_{\widehat{w}}(x)$ is equivalent to the IMSE of the infeasible best estimator in the class $\widehat{g}_w(x)$ for $w \in \mathcal{H}_n$. This is an oracle property for weight selection by cross-validation.

17 Numerical Simulation

We return to the simulation experiment introduced in Sections 6 and 12. Now we add the JMA estimator (15). The IMSE of the estimator is plotted in Figure 3 along with the other estimators. The IMSE of the JMA estimator is uniformly better than the other estimators, with the difference quite striking.

The plots display the IMSE relative to the IMSE of the infeasible optimal averaging estimator. The optimality theory (Theorem 5) suggests that the relative IMSE of the JMA estimator should approach one as the sample size n diverges. Examining the figures, we can see that the IMSE of the estimator is converging extremely slowly to this asymptotic limit. This suggests while the JMA is "asymptotically" optimal, there is considerable room for improvement in finite samples.

We illustrate implementation with the simulated sample ($n = 200$) from Section 12. We report the cross-validation function and JMA weights in Table 1. As we saw in Figure 2, the CV function is minimized at $m = 2$. However, the value of the CV function is quite flat for $m \geq 2$, and in particular its value at $m = 5$ is nearly identical to $m = 2$. This means that cross-validation ranks $m = 2$ and $m = 5$ quite similarly. The JMA weights account for this. Notice that JMA divides the weight between $m = 1$, $m = 2$ and $m = 5$, rather than putting all the weight on a single estimator. The estimators are plotted (along with the true conditional mean $g(x)$) in Figure 4. Both estimators are close to the true $g(x)$.

Table 1:
Cross-Validation Function and JMA weights

	$m=0$	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
$CV_n(m)$	0.955	0.735	0.717	0.722	0.720	0.718
$\widehat{w}_m$	0.02	0.17	0.35	0.00	0.00	0.46

18 Summary

Sieves are routinely used in applied econometrics to approximate unknown functions. Power series and splines are particularly popular and convenient choices. In all applications, the critical issue is selecting the order of the sieve. The choice greatly affects the results and the accuracy of the estimates. Rules of thumb are insufficient as the ideal choice depends on the unknown function to be estimated.

In regression estimation, a simple, straightforward and computationally easy method for selecting the sieve approximation is cross-validation. The method is also asymptotically optimal, in the sense that the CV-selected estimator is asymptotically equivalent to the infeasible best-fitting estimator, when we evaluate estimators based on IMSE (integrated mean-squared error).

Further improvements can be obtained by averaging. Averaging estimators reduce estimation variance, and thereby IMSE. Selection of the averaging weights is analogous to the problem of selection of the order of a sieve approximation, and a feasible method is again cross-validation.

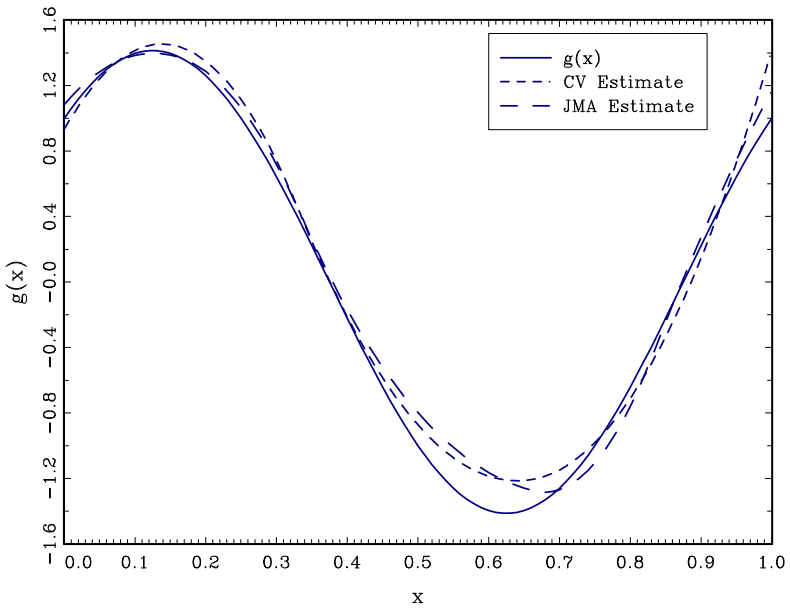

Figure 4: CV and JMA Feasible Series Estimators

Numerical computation of the averaging weights is simple using quadratic programming. Good approximations can be obtained by a simple iterative algorithm. The JMA weights selected by cross-validation are asymptotically optimal in the sense that the fitted averaging estimator is asymptotically equivalent (with respect to IMSE) to the infeasible best weighted average.

19 Regularity Conditions

In this section we list the regularity conditions for the theoretical results.

Assumption 1

1. *The support* $\mathcal{X}$ *of* x_i *is a Cartesian product of compact connected intervals on which the density* $f(x)$ *is bounded away from zero.*
2. $g(x)$ *is continuously differentiable on* $x \in \mathcal{X}$.
3. *For some* $\alpha > 0$, $\eta > 0$, *and* $\psi < \infty$, *for all* $\ell' Q_m \ell = 1$ *and* $0 \le u \le \eta$, $\sup_m \mathbb{P}\left(|\ell' z_{mi}| \le u\right) \le \psi u^{\alpha}$.
4. $0 < \underline{\sigma}^2 \le \sigma_i^2 \le \overline{\sigma}^2 < \infty$.
5. $\max_{m \le M_n} K_m^4 / n = O(1)$ *for a power series, or* $\max_{m \le M_n} K_m^3 / n = O(1)$ *for a spline sieve.*

Assumption 2 $\phi_m^2 > 0$ *for all* $m < \infty$.

The role of Assumption 1.1 is to ensure that the expected design matrix Q_m is uniformly invertible. Assumption 1.2 is used to ensure that $r_m(x)$ is uniformly bounded. Assumption 1.3 is unusual, but used to ensure that moments of the inverse sample design matrix $\left(n^{-1}\sum_{i=1}^n z_{mi}z_{mi}'\right)^{-1}$ exist. Assumption 1.4 bounds the extent of conditional heteroskedasticity, and Assumption 1.5 restricts the complexity of the fitted models.

Assumption 2 is quite important. It states that the approximation error is non-zero for all finite-dimensional models; thus all models are approximations. This is standard in the nonparametrics optimality literature. One implication is that $\xi_n = \inf_m nIMSE_n^*(m) \longrightarrow \infty$ as $n \to \infty$.

Let $q_{jn} = \#\{m : K_m = j\}$ be the number of models which have exactly j coefficients, and set $\overline{q}_n = \max_{j \le M_n} q_{jn}$. This is the largest number of models of any given dimension. For nested models, then $\overline{q}_n = 1$, but when the models are non-nested then $\overline{q}_n$ can exceed one.

Assumption 3 *For some $N \geq 1$*

1. $\sup_i \mathbb{E}\left(e_i^{4(N+1)} \mid x_i\right) < \infty$.
2. $\overline{q}_n = o(\xi_n^{1/N})$ *where* $\xi_n = \inf_m nIMSE_n^*(m)$.
3. $\max_{m \le M_n} \max_{i \le n} h_{mi} \longrightarrow 0$ *almost surely.*

Assumption 3.1 is a strengthing of Assumption 1.4. Assumption 3.2 allows for non-nested models, but bounds the number of models. Assumption 3.3 states that the design matrix cannot be too unbalanced. Under our conditions it is easy to show that $\max_{m \le M_n} \max_{i \le n} h_{mi} = o_p(1)$. The technical strengthening here is to almost sure convergence.

Assumption 4

1. $z_m(x)$ *is either a spline or power series, and is nested.*
2. $g(x)$ *has s continuous derivatives on $x \in \mathcal{X}$ with $s \geq q/2$ for a spline and $s \geq q$ for a power series.*

20 Technical Proofs

Proof of Proposition 1:

The key is the Sherman-Morrison formula (Sherman and Morrison, 1950) which states that for nonsingular $\boldsymbol{A}$ and vector $\boldsymbol{b}$

$$\left(\boldsymbol{A} - \boldsymbol{b}\boldsymbol{b}'\right)^{-1} = \boldsymbol{A}^{-1} + \left(1 - \boldsymbol{b}'\boldsymbol{A}^{-1}\boldsymbol{b}\right)^{-1}\boldsymbol{A}^{-1}\boldsymbol{b}\boldsymbol{b}'\boldsymbol{A}^{-1}.$$

This can be verified by premultiplying the expression by $\boldsymbol{A} - \boldsymbol{b}\boldsymbol{b}'$ and simplifying.

Let $\boldsymbol{Z}_m$ and $\boldsymbol{y}$ denote the matrices of stacked regressors and dependent variable so that the LS estimator is $\widehat{\beta}_m = (\boldsymbol{Z}'_m\boldsymbol{Z}_m)^{-1}\boldsymbol{Z}'_m\boldsymbol{y}$. An application of the Sherman-Morrison formula yields

$$\begin{aligned}\left(\sum_{j\neq i} z_{mj}z'_{mj}\right)^{-1} &= \left(\boldsymbol{Z}'_m\boldsymbol{Z}_m - z_{mi}z'_{mi}\right)^{-1} \\ &= \left(\boldsymbol{Z}'_m\boldsymbol{Z}_m\right)^{-1} + (1-h_{mi})^{-1}\left(\boldsymbol{Z}'_m\boldsymbol{Z}_m\right)^{-1} z_{mi}z'_{mi}\left(\boldsymbol{Z}'_m\boldsymbol{Z}_m\right)^{-1}.\end{aligned}$$

Thus

$$\begin{aligned}\tilde{e}_{mi} &= y_i - z'_{mi}\left(\boldsymbol{Z}'_m\boldsymbol{Z}_m - z_{mi}z'_{mi}\right)^{-1}\left(\boldsymbol{Z}'_m\boldsymbol{y} - z_{mi}y_i\right) \\ &= y_i - z'_{mi}\left(\boldsymbol{Z}'_m\boldsymbol{Z}_m\right)^{-1}\boldsymbol{Z}'_m\boldsymbol{y} + z'_{mi}\left(\boldsymbol{Z}'_m\boldsymbol{Z}_m\right)^{-1} z_{mi}y_i \\ &\quad - (1-h_{mi})^{-1} z'_{mi}\left(\boldsymbol{Z}'_m\boldsymbol{Z}_m\right)^{-1} z_{mi}z'_{mi}\left(\boldsymbol{Z}'_m\boldsymbol{Z}_m\right)^{-1}\boldsymbol{Z}'_m\boldsymbol{y} \\ &\quad + (1-h_{mi})^{-1} z'_{mi}\left(\boldsymbol{Z}'_m\boldsymbol{Z}_m\right)^{-1} z_{mi}z'_{mi}\left(\boldsymbol{Z}'_m\boldsymbol{Z}_m\right)^{-1} z_{mi}y_i \\ &= \hat{e}_{mi} + h_{mi}y_i - (1-h_{mi})^{-1} h_{mi}z'_{mi}\widehat{\beta}_m + (1-h_{mi})^{-1} h_{mi}^2 y_i \\ &= \hat{e}_{mi} + (1-h_{mi})^{-1} h_{mi}\hat{e}_{mi} \\ &= (1-h_{mi})^{-1}\hat{e}_{mi}\end{aligned}$$

the third equality making the substitutions $\widehat{\beta}_m = (\boldsymbol{Z}'_m\boldsymbol{Z}_m)^{-1}\boldsymbol{Z}'_m\boldsymbol{y}$ and $h_{mi} = z'_{mi}(\boldsymbol{Z}'_m\boldsymbol{Z}_m)^{-1} z_{mi}$, and the remainder collecting terms. ■

Define

$$\zeta_m = \sup_{x\in\mathcal{X}}\left(z_m(x)'Q_m^{-1}z_m(x)\right)^{1/2}, \tag{19}$$

the largest normalized Euclidean length of the regressor vector. Under Assumption 1, if z_{mi} is a power series then $\zeta_m^2 = O(k_m^2)$ (see Andrews (1991a)), and when z_{mi} is a regression spline then $\zeta_m^2 = O(k_m)$ (see Newey (1995)). For further discussion see Newey (1997) and Li and Racine (2006).

Without loss of generality, assume $Q_m = I_{K_m}$ throughout this section.

Proof of Theorem 3:

Assumptions (A.1), (A.2), (A.7), (A.9) and (A.10) of Hansen and Racine (2012) are satisfied under our Assumptions 1-3. Thus by their Theorem 2 with $N = 1$, CV selection is optimal with respect to the criterion $R_n(m)$, that is,

$$\left|\frac{R_n(\widehat{m})}{\inf_{1\leq m\leq M_n} R_n(m)}\right| \xrightarrow{p} 1$$

Furthermore, Theorem 1 shows that $IMSE_n^*(m)$ and $IMSE_n(m)$ are asymptotically equivalent. Thus for Theorem 3 it is thus sufficient to show that $R_n(m)$ and $IMSE_n^*(m)$ are asymptotically equivalent. To reduce the notation we will write $I_n(m) = IMSE_n^*(m) = \phi_m^2 + n^{-1}\operatorname{tr}(\Omega_m)$. Thus

what we need to show is

$$\sup_{1\leq m\leq M_n}\left|\frac{R_n(m)-I_n(m)}{I_n(m)}\right|\xrightarrow{p}0. \tag{20}$$

It is helpful to note the following inequalities:

$$\begin{aligned} n\phi_m^2 &\leq nI_n(m) && (21)\\ \operatorname{tr}(\Omega_m) &\leq nI_n(m) && (22)\\ 1 &\leq nI_n(m) && (23)\\ \frac{\zeta_m^2}{n} &\leq \frac{\zeta_m^2 K_m}{n}\leq\frac{\zeta_m^2K_m^2}{n}\leq\Psi<\infty. && (24)\end{aligned}$$

(21) and (22) follow from the formula $nI_n(m)=n\phi_m^2+\operatorname{tr}(\Omega_m)$. (23) holds for n sufficiently large since $\xi_n=\inf_m nI_n(m)\to\infty$. The first two inequalities in (24) holds since either $K_m\geq 1$ or $\zeta_m^2=0$, the third inequality holds for n sufficiently large under Assumption 1.5.

Set

$$\begin{aligned}\widehat{Q}_m &= \frac{1}{n}\sum_{i=1}^n z_{mi}z_{mi}'\\ \widehat{\gamma}_m &= \frac{1}{n}\sum_{i=1}^n z_{mi}r_{mi}\\ \widehat{\Omega}_m &= \frac{1}{n}\sum_{i=1}^n z_{mi}z_{mi}'\sigma_i^2.\end{aligned}$$

As shown in Andrews (1991a) and Hansen and Racine (2012)

$$nR_n(m)=\sum_{i=1}^n r_{mi}^2-n\widehat{\gamma}_m'\widehat{Q}_m^{-1}\widehat{\gamma}_m+\operatorname{tr}\left(\widehat{Q}_m^{-1}\widehat{\Omega}_m\right).$$

Then

$$\begin{aligned} n\left(R_n(m)-I_n(m)\right) &= \sum_{i=1}^n\left(r_{mi}^2-\phi_m^2\right)-n\widehat{\gamma}_m'\widehat{Q}_m^{-1}\widehat{\gamma}_m+\operatorname{tr}\left(\left(\widehat{Q}_m^{-1}-I_{K_m}\right)\Omega_m\right)\\ &\quad+\operatorname{tr}\left(\widehat{\Omega}_m-\Omega_m\right)+\operatorname{tr}\left(\left(\widehat{Q}_m^{-1}-I_{K_m}\right)\left(\widehat{\Omega}_m-\Omega_m\right)\right).\end{aligned}$$

and for any $J \geq 2$

$$\begin{aligned}
\left(\mathbb{E}\left|n\left(R_n(m) - I_n(m)\right)\right|^J\right)^{1/J} &\leq \left(\mathbb{E}\left|\sum_{i=1}^n \left(r_{mi}^2 - \phi_m^2\right)\right|^J\right)^{1/J} && (25)\\
&+ \left(\mathbb{E}\left|n\widehat{\gamma}_m' \widehat{Q}_m^{-1} \widehat{\gamma}_m\right|^J\right)^{1/J} && (26)\\
&+ \left(\mathbb{E}\left|\mathrm{tr}\left(\left(\widehat{Q}_m^{-1} - I_{K_m}\right)\Omega_m\right)\right|^J\right)^{1/J} && (27)\\
&+ \left(\mathbb{E}\left|\mathrm{tr}\left(\widehat{\Omega}_m - \Omega_m\right)\right|^J\right)^{1/J} && (28)\\
&+ \left(\mathbb{E}\left|\mathrm{tr}\left(\left(\widehat{Q}_m^{-1} - I_{K_m}\right)\left(\widehat{\Omega}_m - \Omega_m\right)\right)\right|^J\right)^{1/J} && (29)
\end{aligned}$$

We use some bounds developed in Hansen (2012) for the moment matrices which appear on the right-side of (25)-(29). A key bound is the matrix Rosenthal inequality (Theorem 1 of Hansen (2012)) which states that for any $J \geq 2$ there is a constant $A_J < \infty$ such that for any iid random matrix X_i

$$\begin{aligned}
\left(\mathbb{E}\left\|\sum_{i=1}^n \left(X_i - \mathbb{E}X_i\right)\right\|_2^J\right)^{1/J} &\leq \left[A_J\left(\left(n\mathbb{E}\left\|X_i\right\|_2^2\right)^{J/2} + \left(n\mathbb{E}\left\|X_i\right\|_2^J\right)\right)\right]^{1/J}\\
&\leq A_J^{1/J}\left(n\mathbb{E}\left\|X_i\right\|_2^2\right)^{1/2} + A_J^{1/J}\left(n\mathbb{E}\left\|X_i\right\|_2^J\right)^{1/J}. \qquad (30)
\end{aligned}$$

where the second inequality is the c_r inequality. Using this bound, Hansen (2012, Lemmas 2 and 3) established that for n sufficiently large

$$\begin{aligned}
\mathbb{E}\left\|\widehat{Q}_m^{-1}\right\|^J &\leq 2 && (31)\\
\left(\mathbb{E}\left\|\widehat{Q}_m^{-1} - I_{K_m}\right\|^J\right)^{1/J} &\leq A_{2J}^{1/J}\left(\frac{\zeta_m^2 K_m}{n}\right)^{1/2}. && (32)
\end{aligned}$$

We use (30)-(32) to bound the terms (25)-(29).

We start with (25). Define $\overline{r} = \sup_m \sup_{x\in\mathcal{X}} |r_m(x)|$, which is bounded under Assumption 1.2. WLOG assume that $\overline{r} \geq 1$. Note that $|r_{mi}| \leq \overline{r}$. Applying (30) to (25), and then $\mathbb{E}r_{mi}^N \leq \overline{r}^{N-2}\mathbb{E}r_{mi}^2 \leq \overline{r}^{N-2}\phi_m^2$,

$$\begin{aligned}
\left(\mathbb{E}\left|\sum_{i=1}^n \left(r_{mi}^2 - \phi_m^2\right)\right|^J\right)^{1/J} &\leq A_J^{1/J}\left(n\mathbb{E}r_{mi}^4\right)^{1/2} + A_J^{1/J}\left(n\mathbb{E}r_{mi}^{2J}\right)^{1/J}\\
&\leq A_J^{1/J}\overline{r}\left(n\phi_m^2\right)^{1/2} + A_J^{1/J}\overline{r}^{2-2/J}\left(n\phi_m^2\right)^{1/J}. \qquad (33)
\end{aligned}$$

We next take (26). Note that (19) implies $\|z_{mi}\| \leq \zeta_m$. Then $\mathbb{E}\left\|z_{mi}r_{mi}\right\|^2 \leq \overline{r}^2\mathbb{E}\left\|z_{mi}\right\|^2 =$

$\overline{r}^2 \operatorname{tr}(Q_m) = \overline{r}^2 K_m$ and $\mathbb{E}\|z_{mi} r_{mi}\|^2 \leq \zeta_m^2 \phi_m^2$. Together,

$$\mathbb{E}\|z_{mi} r_{mi}\|^2 \leq \overline{r}\left(\frac{\zeta_m^2 K_m}{n}\right)^{1/2} \left(n\phi_m^2\right)^{1/2} \leq \overline{r}\Psi^{1/2}\left(n\phi_m^2\right)^{1/2} \tag{34}$$

where the second inequality uses (24). Similarly,

$$\frac{\mathbb{E}\|z_{mi} r_{mi}\|^{4J}}{n^{2J-1}} \leq \frac{\overline{r}^{4J-2}\zeta_m^{4J}\mathbb{E} r_{mi}^2}{n^{2J-1}} = \overline{r}^{4J-2}\left(\frac{\zeta_m^2}{n}\right)^{2J} n\phi_m^2 \leq \overline{r}^{4J-2}\Psi^{2J} n\phi_m^2. \tag{35}$$

Applying (30) to (26), and then (34) and (35) we find

$$\begin{aligned}
\left(\mathbb{E}\left\|n^{1/2}\widehat{\gamma}_m\right\|_2^{4J}\right)^{1/2J} &\leq A_{4J}^{1/2J}\mathbb{E}\|z_{mi} r_{mi}\|^2 + A_{4J}^{1/2J}\left(\frac{\mathbb{E}\|z_{mi} r_{mi}\|^{4J}}{n^{2J-1}}\right)^{1/2J} \\
&\leq A_{4J}^{1/2J}\overline{r}\Psi^{1/2}\left(n\phi_m^2\right)^{1/2} + A_{4J}^{1/2J}\overline{r}^{2-1/J}\Psi\left(n\phi_m^2\right)^{1/2J}.
\end{aligned} \tag{36}$$

Using the trace and Cauchy-Schwarz inequalities, (31), and (36)

$$\begin{aligned}
\left(\mathbb{E}\left|n\widehat{\gamma}_m'\widehat{Q}_m^{-1}\widehat{\gamma}_m\right|^J\right)^{1/J} &\leq \left(\mathbb{E}\left(\left\|\widehat{Q}_m^{-1}\right\|^J\left\|n^{1/2}\widehat{\gamma}_m\right\|_2^{2J}\right)\right)^{1/J} \\
&\leq \left(\mathbb{E}\left(\left\|\widehat{Q}_m^{-1}\right\|^{2J}\right)\mathbb{E}\left(\left\|n^{1/2}\widehat{\gamma}_m\right\|_2^{4J}\right)\right)^{1/2J} \\
&\leq (2A_{4J})^{1/2J}\overline{r}\Psi^{1/2}\left(n\phi_m^2\right)^{1/2} + (2A_{4J})^{1/2J}\overline{r}^{2-1/J}\Psi\left(n\phi_m^2\right)^{1/2J}.
\end{aligned} \tag{37}$$

Now we take (27). Using the trace inequality, (32), $\operatorname{tr}(\Omega_m) = \mathbb{E}\left|z_{mi}' z_{mi}\sigma_i^2\right| \leq \overline{\sigma}^2\mathbb{E}|z_{mi}' z_{mi}| = \overline{\sigma}^2 K_m$, and (24)

$$\begin{aligned}
\left(\mathbb{E}\left|\operatorname{tr}\left(\left(\widehat{Q}_m^{-1} - I_{K_m}\right)\Omega_m\right)\right|^J\right)^{1/J} &\leq \left(\mathbb{E}\left\|\widehat{Q}_m^{-1} - I_{K_m}\right\|^J\right)^{1/J}\operatorname{tr}(\Omega_m) \\
&\leq A_J^{1/J}\left(\frac{\zeta_m^2 K_m}{n}\right)^{1/2}\overline{\sigma}K_m^{1/2}\operatorname{tr}(\Omega_m)^{1/2} \\
&\leq \overline{\sigma}A_J^{1/J}\ \Psi^{1/2}\operatorname{tr}(\Omega_m)^{1/2}.
\end{aligned} \tag{38}$$

Next, take (28). Applying (30) to (28), using $\left|z_{mi}' z_{mi} \sigma_i^2\right| \leq \zeta_m^2 \overline{\sigma}^2$ and (24),

$$\begin{aligned}
&\left(\mathbb{E}\left|\operatorname{tr}\left(\widehat{\Omega}_m - \Omega_m\right)\right|^J\right)^{1/J} \\
\leq\ & A_J^{1/J}\left(\frac{\mathbb{E}\left|z_{mi}' z_{mi}\sigma_i^2\right|^2}{n}\right)^{1/2} + A_J^{1/J}\left(\frac{\mathbb{E}\left|z_{mi}' z_{mi}\sigma_i^2\right|^J}{n^{J-1}}\right)^{1/J} \\
\leq\ & \overline{\sigma} A_J^{1/J}\left(\frac{\zeta_m^2}{n}\right)^{1/2} \operatorname{tr}\left(\Omega_m\right)^{1/2} + \overline{\sigma}^{2(1-1/J)} A_J^{1/J}\left(\frac{\zeta_m^2}{n}\right)^{1-1/J} \operatorname{tr}\left(\Omega_m\right)^{1/J} \qquad (39) \\
\leq\ & \overline{\sigma} A_J^{1/J} \Psi^{1/2} \operatorname{tr}\left(\Omega_m\right)^{1/2} + \overline{\sigma}^{2(1-1/J)} A_J^{1/J} \Psi^{1-1/J} \operatorname{tr}\left(\Omega_m\right)^{1/J}. \qquad (40)
\end{aligned}$$

Finally, take (29). Using the trace inequality, Cauchy-Schwarz, (32), and (39),

$$\begin{aligned}
&\left(\mathbb{E}\left|\operatorname{tr}\left(\left(\widehat{Q}_m^{-1} - I_{K_m}\right)\left(\widehat{\Omega}_m - \Omega_m\right)\right)\right|^J\right)^{1/J} \\
\leq\ & \left(\mathbb{E}\left(\left\|\widehat{Q}_m^{-1} - I_{K_m}\right\|^J \left\|\widehat{\Omega}_m - \Omega_m\right\|^J\right)\right)^{1/J} K_m \\
\leq\ & \left(\mathbb{E}\left\|\widehat{Q}_m^{-1} - I_{K_m}\right\|^{2J}\right)^{1/2J} \left(\mathbb{E}\left\|\widehat{\Omega}_m - \Omega_m\right\|^{2J}\right)^{1/2J} K_m \\
\leq\ & A_{4J}^{1/2J}\left(\frac{\zeta_m^2 K_m}{n}\right)^{1/2}\left(\overline{\sigma} A_{2J}^{1/2J}\left(\frac{\zeta_m^2}{n}\right)^{1/2} \operatorname{tr}\left(\Omega_m\right)^{1/2} + \overline{\sigma}^{2(1-1/2J)} A_{2J}^{1/2J}\left(\frac{\zeta_m^2}{n}\right)^{1-1/2J} \operatorname{tr}\left(\Omega_m\right)^{1/2J}\right) K_m \\
\leq\ & \overline{\sigma} A_{4J}^{1/2J} A_J^{1/J} \Psi \operatorname{tr}\left(\Omega_m\right)^{1/2} + \overline{\sigma}^{2(1-1/2J)} A_{4J}^{1/2J} A_{2J}^{1/2J} \Psi^{3/2-1/2J} \operatorname{tr}\left(\Omega_m\right)^{1/2J}. \qquad (41)
\end{aligned}$$

Combining (33) and (37), and then applying (21) and (23) we find that

$$\begin{aligned}
&\left(\mathbb{E}\left|\sum_{i=1}^n \left(r_{mi}^2 - \phi_m^2\right)\right|^J\right)^{1/J} + \left(\mathbb{E}\left|n\widehat{\gamma}_m' \widehat{Q}_m^{-1} \widehat{\gamma}_m\right|^J\right)^{1/J} \\
\leq\ & C_1\left(n\phi_m^2\right)^{1/2} C_2\left(n\phi_m^2\right)^{1/J} + C_3\left(n\phi_m^2\right)^{1/2J} \qquad (42) \\
\leq\ & C_1\left(nI_n(m)\right)^{1/2} + C_2\left(nI_n(m)\right)^{1/J} + C_3\left(nI_n(m)\right)^{1/2J} \\
\leq\ & \left(C_1 + C_2 + C_3\right)\left(nI_n(m)\right)^{1/2} \qquad (43)
\end{aligned}$$

where $C_1 = A_J^{1/J}\overline{r} + \left(2A_{4J}\right)^{1/2J}\overline{r}\Psi^{1/2}$, $C_2 = A_J^{1/J}\overline{r}^{2-2/J}$, and $C_3 = \left(2A_{4J}\right)^{1/2J}\overline{r}^{2-1/J}\Psi$.

Similarly, combining (38), (40) and (41), and then applying (22) and (23),

$$\begin{aligned}
&\left(\mathbb{E}\left|\operatorname{tr}\left(\left(\widehat{Q}_m^{-1}-I_{K_m}\right)\Omega_m\right)\right|^J\right)^{1/J}+\left(\mathbb{E}\left|\operatorname{tr}\left(\widehat{\Omega}_m-\Omega_m\right)\right|^J\right)^{1/J}\\
&+\left(\mathbb{E}\left|\operatorname{tr}\left(\left(\widehat{Q}_m^{-1}-I_{K_m}\right)\left(\widehat{\Omega}_m-\Omega_m\right)\right)\right|^J\right)^{1/J}\\
\leq\ & C_4\operatorname{tr}(\Omega_m)^{1/2}C_5\operatorname{tr}(\Omega_m)^{1/J}+C_6\operatorname{tr}(\Omega_m)^{1/2J} \qquad (44)\\
\leq\ & C_4\left(nI_n(m)\right)^{1/2}+C_5\left(nI_n(m)\right)^{1/J}+C_6\left(nI_n(m)\right)^{1/2J}\\
\leq\ & \left(C_4+C_5+C_6\right)\left(nI_n(m)\right)^{1/2}. \qquad (45)
\end{aligned}$$

where $C_4=\overline{\sigma}A_J^{1/J}\left(2\Psi^{1/2}+A_{4J}^{1/2J}\Psi\right)$, $C_5=\overline{\sigma}^{2(1-1/J)}A_J^{1/J}\Psi^{1-1/J}$, and $C_6=\overline{\sigma}^{2(1-1/2J)}A_{4J}^{1/2J}A_{2J}^{1/2J}\Psi^{3/2-1/2J}$.

Setting $J=4$, (25)-(29), (43) and (45) imply that

$$\left(\mathbb{E}\left|n\left(R_n(m)-I_n(m)\right)\right|^4\right)^{1/4}\leq C\left(nI_n(m)\right)^{1/2}. \qquad (46)$$

where $C=C_1+C_2+C_3+C_4+C_5+C_6$.

Applying Boole's inequality, Markov's inequality, and (46)

$$\begin{aligned}
\mathbb{P}\left(\sup_{1\leq m\leq M_n}\left|\frac{R_n(m)-I_n(m)}{I_n(m)}\right|>\eta\right) &= \mathbb{P}\left(\bigcup_{m=1}^{M_n}\left\{\left|\frac{R_n(m)-I_n(m)}{I_n(m)}\right|>\eta\right\}\right)\\
&\leq \sum_{m=1}^{M_n}\mathbb{P}\left(\left\{\left|\frac{n\left(R_n(m)-I_n(m)\right)}{nI_n(m)}\right|>\eta\right\}\right)\\
&\leq \eta^{-4}\sum_{m=1}^{M_n}\frac{\mathbb{E}\left|n\left(R_n(m)-I_n(m)\right)\right|^4}{\left(nI_n(m)\right)^4}\\
&\leq C^4\eta^{-4}\sum_{m=1}^{M_n}\frac{1}{\left(nI_n(m)\right)^2}.
\end{aligned}$$

Recall the definitions of $\overline{q}_n$ and ξ_n. Pick a sequence $m_n\to\infty$ such that $m_n\xi_n^{-2}\to 0$ yet $\overline{q}_n^2=o(m_n)$ which is possible since $\xi_n\to\infty$ and $\overline{q}_n^2=o(\xi_n^2)$ under Assumption 3.2. Then since $nI_n(m)\geq\xi_n$ and $nI_n(m)\geq\operatorname{tr}(\Omega_m)\geq\underline{\sigma}^2K_m\geq\underline{\sigma}^2m/\overline{q}_n$, the sum on the right-hand side is bounded by

$$m_n\xi_n^{-2}+\sum_{m=m_n+1}^{\infty}\frac{\overline{q}_n^2}{\underline{\sigma}^4m^2}\leq m_n\xi_n^{-2}+\frac{\overline{q}_n^2}{\underline{\sigma}^4m_n}\longrightarrow 0$$

as $n\to\infty$. This establishes (20) as desired, completing the proof. ■

Proof of Theorem 5

As in the proof of Theorem 2, it is sufficient to show that

$$\sup_{w\in\mathcal{H}_n}\left|\frac{R_n(w)-I_n(w)}{I_n(w)}\right|\xrightarrow{p}0. \qquad (47)$$

where we have written $I_n(w) = IMSE_n^*(w)$. WLOG, assume $Q_m = I_{K_m}$. For w_m^* and w_m^{**} defined in (13) and (14), observe that $\sum_{m=1}^{M_n} w_m^* = \sum_{m=1}^{M_n} w_m^{**} = 1$. Since w_m^* are non-negative and sum to one, they define a probability distribution. Thus by Liapunov's inequality, for any $s \geq 1$ and any constants $a_m \geq 0$

$$\sum_{m=1}^{M_n} w_m^* a_m^{1/s} \leq \left(\sum_{m=1}^{M_n} w_m^* a_m \right)^{1/s} \tag{48}$$

and similarly

$$\sum_{m=1}^{M_n} w_m^{**} a_m^{1/s} \leq \left(\sum_{m=1}^{M_n} w_m^{**} a_m \right)^{1/s} \tag{49}$$

As shown in Andrews (1991a) and Hansen and Racine (2012)

$$nR_n(w) = \sum_{m=1}^{M_n} w_m^* n\phi_m^2 - \sum_{m=1}^{M_n} w_m^* n\widehat{\gamma}_m' \widehat{Q}_m^{-1} \widehat{\gamma}_m + \sum_{m=1}^{M_n} w_m^{**} \operatorname{tr}\left(\widehat{Q}_m^{-1} \widehat{\Omega}_m \right).$$

Then applying Minkowski's inequality, (42), (44), and then (48) and (49)

$$\begin{aligned}
&\left(\mathbb{E} \left| n \left(R_n(m) - I_n(m) \right) \right|^J \right)^{1/J} \\
&\leq \sum_{m=1}^{M_n} w_m^* \left[\left(\mathbb{E} \left| \sum_{i=1}^{n} \left(r_{mi}^2 - \phi_m^2 \right) \right|^J \right)^{1/J} + \left(\mathbb{E} \left| n\widehat{\gamma}_m' \widehat{Q}_m^{-1} \widehat{\gamma}_m \right|^J \right)^{1/J} \right] \\
&+ \sum_{m=1}^{M_n} w_m^{**} \left[\left(\mathbb{E} \left| \operatorname{tr}\left(\left(\widehat{Q}_m^{-1} - I_{K_m} \right) \Omega_m \right) \right|^J \right)^{1/J} + \left(\mathbb{E} \left| \operatorname{tr}\left(\widehat{\Omega}_m - \Omega_m \right) \right|^J \right)^{1/J} \right. \\
&\left. + \left(\mathbb{E} \left| \operatorname{tr}\left(\left(\widehat{Q}_m^{-1} - I_{K_m} \right) \left(\widehat{\Omega}_m - \Omega_m \right) \right) \right|^J \right)^{1/J} \right] \\
&\leq C_1 \sum_{m=1}^{M_n} w_m^* \left(n\phi_m^2 \right)^{1/2} + C_2 \sum_{m=1}^{M_n} w_m^* \left(n\phi_m^2 \right)^{1/J} + C_3 \sum_{m=1}^{M_n} w_m^* \left(n\phi_m^2 \right)^{1/2J} \\
&+ C_4 \sum_{m=1}^{M_n} w_m^{**} \operatorname{tr}\left(\Omega_m \right)^{1/2} + C_5 \sum_{m=1}^{M_n} w_m^{**} \operatorname{tr}\left(\Omega_m \right)^{1/J} + C_6 \sum_{m=1}^{M_n} w_m^{**} \operatorname{tr}\left(\Omega_m \right)^{1/2J} \\
&\leq C_1 \left(\sum_{m=1}^{M_n} w_m^* n\phi_m^2 \right)^{1/2} + C_2 \left(\sum_{m=1}^{M_n} w_m^* n\phi_m^2 \right)^{1/J} + C_3 \left(\sum_{m=1}^{M_n} w_m^* n\phi_m^2 \right)^{1/2J} \\
&+ C_4 \left(\sum_{m=1}^{M_n} w_m^{**} \operatorname{tr}\left(\Omega_m \right) \right)^{1/2} + C_5 \left(\sum_{m=1}^{M_n} w_m^{**} \operatorname{tr}\left(\Omega_m \right) \right)^{1/J} + C_6 \left(\sum_{m=1}^{M_n} w_m^{**} \operatorname{tr}\left(\Omega_m \right) \right)^{1/2J} \\
&\leq C_1 \left(nI_n(w) \right)^{1/2} + C_2 \left(nI_n(w) \right)^{1/J} + C_3 \left(nI_n(w) \right)^{1/2J} \\
&+ C_4 \left(nI_n(w) \right)^{1/2} + C_5 \left(nI_n(w) \right)^{1/J} + C_6 \left(nI_n(w) \right)^{1/2J} \\
&\leq C \left(nI_n(m) \right)^{1/2}
\end{aligned}$$

where the final two inequalities use

$$\begin{aligned}
\sum_{m=1}^{M_n} w_m^* n\phi_m^2 &\leq nI_n(w) \\
\sum_{m=1}^{M_n} w_m^{**} \operatorname{tr}(\Omega_m) &\leq nI_n(w) \\
1 &\leq nI_n(w)
\end{aligned}$$

where the first two follow from the formula (12) for $nI_n(w)$, and the third holds for n sufficiently large since $\inf_w nI_n(w) \to \infty$.

Setting $J = 2(N+1)$, we have shown that

$$\mathbb{E}\left|n\left(R_n(w) - I_n(w)\right)\right|^{2(N+1)} \leq C^{1+N}\left(nI_n(w)\right)^{N+1}.$$

Then

$$\begin{aligned}
\mathbb{P}\left(\sup_{w\in\mathcal{H}_n}\left|\frac{R_n(w)-I_n(w)}{I_n(w)}\right| > \eta\right) &= \mathbb{P}\left(\bigcup_{w\in\mathcal{H}_n}\left\{\left|\frac{R_n(w)-I_n(w)}{I_n(w)}\right| > \eta\right\}\right) \\
&\leq \sum_{w\in\mathcal{H}_n}\mathbb{P}\left(\left\{\left|\frac{n\left(R_n(w)-I_n(w)\right)}{nI_n(w)}\right| > \eta\right\}\right) \\
&\leq \eta^{-2(N+1)}\sum_{w\in\mathcal{H}_n}\frac{\mathbb{E}\left|n\left(R_n(w)-I_n(w)\right)\right|^{2(N+1)}}{\left(nI_n(w)\right)^{2(N+1)}} \\
&\leq C^{1+N}\eta^{-2(N+1)}\sum_{w\in\mathcal{H}_n}\frac{1}{\left(nI_n(w)\right)^{N+1}}.
\end{aligned}$$

As shown in Hansen and Racine (2012), equations (23)-(25)-(28), the right-hand-side is $o(1)$. ■

By Markov's inequality, we have established (47), as desired.

References

[1] Akaike, H. 1973. "Information theory and an extension of the maximum likelihood principle." In B. Petroc and F. Csake, eds., *Second International Symposium on Information Theory.*

[2] Allen, David M. 1974. "The relationship betwen variable selection and data augmentation and a method for prediction," *Technometrics*, 16, pp. 125-127.

[3] Andrews, Donald W.K. 1991a. "Asymptotic normality of series estimators for nonparametric and semiparametric models," *Econometrica,* 59, pp. 307-345.

[4] Andrews, Donald W. K. 1991b. "Asymptotic optimality of generalized C_L, cross-validation, and generalized cross-validation in regression with heteroskedastic errors," *Journal of Econometrics*, 47, pp. 359-377.

[5] Chen, Xiaohong. 2007. "Large Sample Sieve Estimation of Semi-nonparametric Models" *Handbook of Econometrics*, Vol. 6B, Chapter 76, eds. James J. Heckman and Edward E. Leamer, North-Holland.

[6] Chui, Charles K. 1992. *An Introduction to Wavelets*. Academic Press.

[7] Craven P. and Grace Wahba. 1979. "Smoothing noisy data with spline functions," *Numerische Mathematik*, 31, pp. 377-403

[8] de Boor, Carl. 2001. *A Practical Guide to Splines.* Springer.

[9] Fan, Jianging and Runze Li. 2001. "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, 96, pp. 1348-1360.

[10] Grenander, U. 1981. *Abstract Inference*. New York: Wiley.

[11] Hansen, Bruce E. 2007 "Least squares model averaging," *Econometrica*, 75, pp. 1175-1189.

[12] Hansen, Bruce E. 2012. "A matrix Rosenthal-type inequality with an application to the integrated mean squared error of series regression," Working Paper. University of Wisconsin.

[13] Hansen, Bruce E. and Jeffrey S. Racine. 2012. "Jackknife model averaging," *Journal of Econometrics*, (2012), 167, pp. 38–46.

[14] Hurvich, Clifford M. and Chih-Ling Tsai. 1989. "Regression and time series model selection in small samples", *Biometrika*, 76, pp. 297-307.

[15] Li, Ker-Chau. 1987. "Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete Index Set," *Annals of Statistics*, 15, pp. 958-975.

[16] Li, Qi, and Jeffrey S. Racine. 2006. *Nonparametric Econometrics: Theory and Practice.* Princeton University Press.

[17] Mallows, C. L. 1973. "Some comments on C_p," *Technometrics*, 15, pp. 661-675.

[18] Newey, Whitney K. 1995. "Convergence rates for series estimators," in Maddalla, G.S., Phillips, P.C.B., Srinavasan, T.N. (eds.) *Statistical Methods of Economics and Quantitative Economics: Essays in Honor of C.R. Rao*. Backwell, Cambridge, pp. 254-275.

[19] Newey, Whitney K. 1997. "Convergence rates and asymptotic normality for series estimators," *Journal of Econometrics*, 79, pp. 147-168.

[20] Schwarz, G. 1978. "Estimating the dimension of a model," *Annals of Statistics*, 6, pp. 461-464.

[21] Shao, Jun. 1997. "An asymptotic theory for linear model selection," *Statistica Sinica*, 7, pp. 221-264.

[22] Sherman, Jack and Winifred J. Morrison. 1950. "Adjustment of an Inverse Matrix Corresponding to a Change in One Element of a Given Matrix," *Annals of Mathematical Statistics*, 21, pp. 124–127.

[23] Stone, M. 1974. "Cross-validatory choice and assessment of statistical predictions" (with discussion), *Journal of the Royal Statistical Society, Series B*, 36, pp. 111-147.

[24] Tibshirani, R. J. 1996. "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, pp. 267-288.

[25] Wahba, Grace and S. Wold. 1975. "A completely automatic French curve: Fitting spline functions by cross-validation," *Communications in Statistics*, 4, pp. 1-17.

[26] Zou, Hui. 2006. "The adaptive LASSO and its oracle properties," *Journal of the American Statistical Association*, 101, pp. 1418-1429.