Journal of Econometrics 167 (2012) 38–46

Contents lists available at SciVerse ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Jackknife model averaging

Bruce E. Hansen [a,*], Jeffrey S. Racine [b]

[a] Department of Economics, Social Science Building, University of Wisconsin, Madison, WI 53706-1396, USA

[b] Department of Economics, Kenneth Taylor Hall, McMaster University, Hamilton, ON, Canada L8S 4M4

ARTICLE INFO

Article history:
Received 28 January 2008
Received in revised form
3 June 2011
Accepted 10 June 2011
Available online 4 November 2011

ABSTRACT

We consider the problem of obtaining appropriate weights for averaging M approximate (misspecified) models for improved estimation of an unknown conditional mean in the face of non-nested model uncertainty in heteroskedastic error settings. We propose a "jackknife model averaging" (JMA) estimator which selects the weights by minimizing a cross-validation criterion. This criterion is quadratic in the weights, so computation is a simple application of quadratic programming. We show that our estimator is asymptotically optimal in the sense of achieving the lowest possible expected squared error. Monte Carlo simulations and an illustrative application show that JMA can achieve significant efficiency gains over existing model selection and averaging methods in the presence of heteroskedasticity.

1. Introduction

Confronting parametric model uncertainty has a rich heritage in the statistics and econometrics literature. The two most popular approaches toward dealing with model uncertainty are 'model selection' and 'model averaging', while both Bayesian and non-Bayesian approaches have been proposed.

Model selection remains perhaps the most popular approach for dealing with model uncertainty. This method has the user first adopt an estimation criterion such as the Akaike information criterion (AIC, Akaike, 1970) and then select from among a set of candidate models that which scores most highly based upon this criterion. Unfortunately, a variety of such criteria have been proposed in the literature, different criteria favor different models from within a given set of candidate models, and seasoned practitioners know that some criteria will favor more parsimonious models (e.g., the Schwarz–Bayes information criterion (BIC); Schwarz, 1978) while others will favor more heavily parameterized models (e.g., AIC).

Model averaging, on the other hand, deals with model uncertainty not by having the user select one model from among a set of candidate models according to a criterion such as AIC or BIC, but rather by averaging over the set of candidate models in a particular manner. Many readers will no doubt be familiar with the Bayesian model averaging literature, and we direct the interested reader to Hoeting et al. (1999) for a comprehensive review of this literature. There is also a rapidly-growing literature on frequentist methods for model averaging, including Buckland et al. (1997), Juditsky and Nemirovski (2000), Yang (2001, 2004), Hansen (2007), Goldenshluger (2009) and Wan et al. (2010). Most of these methods involve sample splitting, which can be inefficient, and all exclude heteroskedasticity, which limits their applicability.

In this paper we propose a frequentist model averaging method which we term "jackknife model averaging" (hereafter JMA) that selects the weights by minimizing a cross-validation criterion. In models that are linear in the parameters the cross-validation criterion is a simple quadratic function of the weights, so the solution is found through a standard application of numerical quadratic programming. (Delete-one) cross-validation for selection of regression models was introduced by Allen (1974), Stone (1974), Geisser (1974) and Wahba and Wold (1975), and its optimality demonstrated by Li (1987) for homoskedastic regression and Andrews (1991) for heteroskedastic regression. After our initial submission we discovered that the idea of using cross-validation to select model averaging weights has been proposed before by Wolpert (1992) and Breiman (1996), but these papers had no theoretical justification (only simulation evidence) for their proposed methods.

Applying the theory developed in Li (1987), Andrews (1991) and Hansen (2007), we establish the asymptotic optimality of the JMA estimator allowing for bounded heteroskedasticity of unknown form. Our results are the first (to our knowledge) theoretical results for model averaging allowing for heteroskedasticity. We show that the JMA estimator is asymptotically optimal in the sense of achieving the lowest possible expected squared error over the class of linear estimators constructed from a countable set of weights. The class of linear estimators includes but is not

* Correspondence to: Department of Economics, University of Wisconsin, 1180 Observatory Drive, 53706 Madison, WI, USA. Tel.: +1 608 263 3880; fax: +1 608 263 3876.

E-mail addresses: behansen@wisc.edu (B.E. Hansen), racinej@mcmaster.ca (J.S. Racine).

0304-4076/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2011.06.019

limited to linear least-squares, ridge regression, Nadaraya–Watson and local polynomial kernel regression with fixed bandwidths, nearest neighbor estimators, series estimators, estimators of additive interaction models, and spline estimators. Our results apply both to nested and nonnested regression models, extending the theory of Hansen (2007) whose proof was limited to the nested regression case. An important limitation is that our theoretical results are limited to random samples (and thus excludes time-series). While we expect that the methods are applicable to time-series regression with martingale difference errors, the theoretical extension would be quite challenging.

Our proof method follows Hansen (2007) by restricting the weight vectors to a discrete grid while allowing for an unbounded number of models. An alternative proof method used by Wan et al. (2010) allows for continuous weights but at the cost of greatly limiting the number of allowable models.

We demonstrate the potential efficiency gains through a simple Monte Carlo experiment. In a classic regression setting, we show that JMA achieves lower mean squared error than competitive methods. In the presence of homoskedastic errors, JMA and Mallows model averaging (MMA) are nearly equivalent, but when the errors are heteroskedastic, JMA has significantly lower MSE.

We also illustrate the method with an empirical application to cross-section earnings prediction. We find that JMA achieves out-of-sample prediction squared error which is either equivalent or lower than that achieved by all other methods considered.

The remainder of the paper is organized as follows. Section 2 presents averaging estimators, and Section 3 presents the jackknife weighting method. Section 4 provides an asymptotic optimality theory under high-level conditions for a broad class of linear estimators. Section 5 presents the optimality theory for linear regression models. Section 6 reports the Monte Carlo simulation experiment, and Section 7 an application to predicting earnings. Section 8 concludes, and the mathematical proofs are presented in the Appendix.

2. Averaging estimators

Consider a sample of independent observations (y_i, x_i) for $i = 1, \ldots, n$. Define the conditional mean $\mu_i = \mu(x_i) = \mathbb{E}(y_i \mid x_i)$ so that

$$y_i = \mu_i + e_i \qquad (1)$$
$$\mathbb{E}(e_i \mid x_i) = 0.$$

Define $\mathbf{y} = (y_1, \ldots, y_n)'$, $\mu = (\mu_1, \ldots, \mu_n)'$, and $\mathbf{e} = (e_1, \ldots, e_n)'$. Let $\sigma_i^2 = \mathbb{E}\left(e_i^2 \mid x_i\right)$ denote the conditional variance which is allowed to depend on x_i.

Suppose that we have a set of linear estimators $\left\{\hat{\mu}^1, \hat{\mu}^2, \ldots, \hat{\mu}^{M_n}\right\}$ for μ. By linear, we mean that the m'th estimator can be written as $\hat{\mu}^m = \mathbf{P}_m\mathbf{y}$ where $\mathbf{P}_m$ is not a function of $\mathbf{y}$. As mentioned in the introduction, this class of estimators includes linear least-squares, ridge regression, Nadaraya–Watson and local polynomial kernel regression with fixed bandwidths, nearest neighbor estimators, series estimators, estimators of additive interaction models, and spline estimators. This restriction to linear estimators will be used for our optimality theory, but is not essential to the definition of the JMA estimator.

Our primary focus will be on least-squares estimators, in which case $\mathbf{P}_m = \mathbf{X}^m\left(\mathbf{X}^{m\prime}\mathbf{X}^m\right)^{-1}\mathbf{X}^{m\prime}$ and the i'th row of $\mathbf{X}^m$ is x_i^m, a $k_m \times 1$ function of x_i. An estimator (or model) corresponds to a particular set of regressors x_i^m. In some applications the regressor matrices will be nested so that $\operatorname{span}(\mathbf{X}^m) \subset \operatorname{span}\left(\mathbf{X}^{m+1}\right)$, while in other applications the regressor sets will be non-nested.

The typical application we envision is when the potential regressor set x_i is large, and the regressors x_i^m are subsets of x_i. Another potential application is series estimation, for example a spline. In this case, x_i^m is a set of basis transformations of a (low-dimensional) regressor x_i.

In practice the number of estimators M_n can be quite large. To allow for this possibility, our optimality theory does not impose a bound on M_n and we view this as one of its important strengths.

The problem of "model selection" is how to select an estimator from the set $\left\{\hat{\mu}^1, \hat{\mu}^2, \ldots, \hat{\mu}^{M_n}\right\}$. If a mean-square criterion is adopted, it is well-known that the optimal estimator is not necessarily the largest or most complete model. This should be quite apparent in context where the number of potential regressors exceeds the number of observations, but it is also true in any finite sample context. To minimize the mean-square estimation error, a balance must be attained between the bias due to omitted variables and the variance due to parameter estimation. Optimal model selection addresses this by designing a data-dependent rule to pick an individual estimator so that the selected estimator has low risk.

Further reductions in mean-squared error can be attained by averaging across estimators. Let $\mathbf{w} = \left(w^1, w^2, \ldots, w^{M_n}\right)'$ be a set of weights which sum to one. Given $\mathbf{w}$, an averaging estimator for μ takes the form

$$\hat{\mu}(\mathbf{w}) = \sum_{m=1}^{M_n} w^m\hat{\mu}^m = \hat{\mu}\mathbf{w} = \mathbf{P}(\mathbf{w})\,\mathbf{y} \qquad (2)$$

where $\hat{\mu} = \left(\hat{\mu}^1, \ldots, \hat{\mu}^{M_n}\right)$ is the $n \times M_n$ matrix of estimates and

$$\mathbf{P}(\mathbf{w}) = \sum_{m=1}^{M_n} w^m\mathbf{P}_m \qquad (3)$$

is a linear operator indexed by $\mathbf{w}$. Thus for fixed weights the averaging estimator $\hat{\mu}(\mathbf{w})$ is linear in $\mathbf{y}$ and is also linear in the weights $\mathbf{w}$.

By setting the weights $\mathbf{w}$ to be unit vectors ι_m (where $w^m = 1$ and $w^\ell = 0$ for $\ell \neq m$) the averaging estimator simplifies to a selection estimator $\hat{\mu}^m$. By allowing non-unit weights we generalize selection estimators and obtain smoother functions of the data. As shown by Hansen (2007), the estimator (2) can achieve lower mean-squared-error (MSE) than any individual estimator. The source of the improvement is similar to shrinkage, where the introduction of bias allows a reduction in estimation variance.

We require the weights to be non-negative and thus lie on the $\mathbb{R}^{M_n}$ unit simplex:

$$\mathcal{H}_n = \left\{\mathbf{w} \in \mathbb{R}^{M_n} : w^m \geq 0,\ \sum_{m=1}^{M_n} w^m = 1\right\}.$$

One might ask: Why restrict the weights to be non-negative? Can improved performance be attained by lifting this restriction and allowing for negative weights? The answer – at least in the context of linear regression – appears to be no. As shown by Cohen (1966) in a Gaussian setting and discussed by Li (1987) for regression, a necessary condition for the linear estimator $\mathbf{P}(\mathbf{w})\,\mathbf{y}$ to be admissible is that all eigenvalues of $\mathbf{P}(\mathbf{w})$ lie in the interval $[0, 1]$. For all common linear estimators (including regression), the eigenvalues of $\mathbf{P}_m$ lie in $[0, 1]$, so a sufficient condition for $\mathbf{P}(\mathbf{w})$ to have eigenvalues in $[0, 1]$ is that $\mathbf{w} \in \mathcal{H}_n$. Furthermore this restriction is necessary in the leading case of nested linear regression. To see this, rewrite

$$\mathbf{P}(\mathbf{w}) = \sum_{m=1}^{M_n} w_*^m\left(\mathbf{P}_m - \mathbf{P}_{m-1}\right)$$

where

$$w_*^m = \sum_{j=m}^{M_n} w^j.$$

When the models are nested and ordered, then $\mathbf{P}_m - \mathbf{P}_{m-1}$ are mutually orthogonal projection matrices. It follows that the eigenvalues of $\mathbf{P}(\mathbf{w})$ take only the values $\{0, w_*^1, w_*^2, \ldots, w_*^{M_n}\}$. Restricting these eigenvalues to $[0, 1]$ is equivalent to the restriction $0 \leq w^m \leq 1$ for all m, e.g. $\mathbf{w} \in \mathcal{H}_n$. It follows that for nested linear regression, the restriction of the weight vector to the unit simplex is a necessary condition for admissibility.

It is also instructive to investigate the form of the averaging estimator in the context of nonparametric series regression with orthogonal regressors.[1] Let x_{ij} be an orthogonal basis of functions of x_i and set $X_j = (x_{1j}, \ldots, x_{nj})'$ to be the $n \times 1$ vector of observations on the j'th regressor. Since the regressors are orthogonal then the OLS estimate of μ using the first m regressors is

$$\hat{\mu}^m = \sum_{j=1}^{m} X_j \hat{\beta}_j$$

where $\hat{\beta}_j = \left(X_j'X_j\right)^{-1} X_j'\mathbf{y}$ is least-squares on the j'th regressor. In this context a nested averaging estimator takes the simple form

$$\hat{\mu}(\mathbf{w}) = \left(w^1 + \cdots + w^{M_n}\right) X_1\hat{\beta}_1 + \left(w^2 + \cdots + w^{M_n}\right) X_2\hat{\beta}_2 + \cdots + w^{M_n} X_{M_n}\hat{\beta}_{M_n}.$$

Thus a nested averaging estimator necessarily assigns (weakly) declining weights to the components $X_j\hat{\beta}_j$. It follows that the ordering of the regressors X_j is of critical importance. Different orderings lead to different individual estimators, and thus to different averaging estimators.

While the order of the regressors is critical for averaging estimators, it is equally critical for traditional selection estimators. Averaging and selection methods are equally dependent on the individual estimators (models) and thus on the explicit ordering of the variables in nested modeling.

3. Jackknife weighting

In this paper we propose jackknife selection of $\mathbf{w}$ (also known as leave-one-out cross-validation). This requires the jackknife residuals for the averaging estimator, which we now derive.

The m'th jackknife estimator is $\tilde{\mu}^m = \left(\tilde{\mu}_1^m, \tilde{\mu}_2^m, \ldots, \tilde{\mu}_n^m\right)'$, where $\tilde{\mu}_i^m$ is the estimator $\hat{\mu}_i^m$ computed with the i'th observation deleted, and can be written as $\tilde{\mu}^m = \tilde{\mathbf{P}}_m\mathbf{y}$ where $\tilde{\mathbf{P}}_m$ has zeros on the diagonal. The jackknife residual vector for the m'th estimator is $\tilde{\mathbf{e}}^m = \mathbf{y} - \tilde{\mu}^m$.

In the least-squares example described in the previous section, $\tilde{\mu}_i^m = x_i^{m\prime}\left(\mathbf{X}_{(-i)}^{m\prime}\mathbf{X}_{(-i)}^m\right)^{-1}\mathbf{X}_{(-i)}^{m\prime}\mathbf{y}_{(-i)}$, where $\mathbf{X}_{(-i)}^m$ and $\mathbf{y}_{(-i)}$ denote the matrices $\mathbf{X}^m$ and $\mathbf{y}$ with the i'th row deleted. As shown in Eq. (1.4) of Li (1987), we have the simple relationship

$$\tilde{\mathbf{P}}_m = \mathbf{D}_m(\mathbf{P}_m - \mathbf{I}) + \mathbf{I} \tag{4}$$

where $\mathbf{D}_m$ is the $n \times n$ diagonal matrix with the i'th diagonal element equal to $\left(1 - h_{ii}^m\right)^{-1}$, and $h_{ii}^m = x_i^{m\prime}\left(\mathbf{X}^{m\prime}\mathbf{X}^m\right)^{-1}x_i^m$ is the i'th diagonal element of $\mathbf{P}_m$. (See also the generalization of (4) by Racine (1997).) It follows that the jackknife residual vector can be conveniently written as $\tilde{\mathbf{e}}^m = \mathbf{D}_m\hat{\mathbf{e}}^m$ where $\hat{\mathbf{e}}^m = \mathbf{y} - \mathbf{P}_m\mathbf{y}$ is the least squares residual vector. Using this representation, $\tilde{\mathbf{e}}^m$ can be computed with a simple linear operation and does not require n separate regressions.

The jackknife version of the averaging estimator is

$$\tilde{\mu}(\mathbf{w}) = \sum_{m=1}^{M_n} w^m\tilde{\mu}^m = \tilde{\mu}\mathbf{w} = \tilde{\mathbf{P}}(\mathbf{w})\mathbf{y}$$

where $\tilde{\mu} = \left(\tilde{\mu}^1, \ldots, \tilde{\mu}^{M_n}\right)$ and $\tilde{\mathbf{P}}(\mathbf{w}) = \sum_{m=1}^{M_n} w^m\tilde{\mathbf{P}}_m$. Note that the matrix $\tilde{\mathbf{P}}(\mathbf{w})$ has zeros on the diagonal.

The jackknife averaging residual is

$$\begin{aligned}\tilde{\mathbf{e}}(\mathbf{w}) &= \mathbf{y} - \tilde{\mu}(\mathbf{w}) \\ &= \sum_{m=1}^{M_n} w^m\tilde{\mathbf{e}}^m \\ &= \tilde{\mathbf{e}}\mathbf{w}\end{aligned}$$

where $\tilde{\mathbf{e}} = \left(\tilde{\mathbf{e}}^1, \ldots, \tilde{\mathbf{e}}^{M_n}\right)$.

The jackknife estimate of expected true error[2] is

$$\mathrm{CV}_n(\mathbf{w}) = \frac{1}{n}\tilde{\mathbf{e}}(\mathbf{w})'\tilde{\mathbf{e}}(\mathbf{w}) = \mathbf{w}'\mathbf{S}_n\mathbf{w} \tag{5}$$

where $\mathbf{S}_n = \frac{1}{n}\tilde{\mathbf{e}}'\tilde{\mathbf{e}}$ is $M_n \times M_n$. $\mathrm{CV}_n(\mathbf{w})$ is also known as the least-squares cross-validation criterion.

The jackknife (or cross-validation) choice of weight vector is the value which minimizes $\mathrm{CV}_n(\mathbf{w})$ over $\mathbf{w} \in \mathcal{H}_n^*$, where $\mathcal{H}_n^*$ is some subset of $\mathcal{H}_n$:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}\in\mathcal{H}_n^*}{\operatorname{argmin}}\ \mathrm{CV}_n(\mathbf{w}). \tag{6}$$

The jackknife model average (JMA) estimator of μ is $\hat{\mu}(\hat{\mathbf{w}}) = \hat{\mu}\hat{\mathbf{w}}$.

If we again consider unit weight vectors ι_m then $\mathrm{CV}_n(\iota_m)$ is the standard jackknife criterion for selection of regression models, and its minimizer $\hat{\iota}_m$ is the standard jackknife selected model. Thus the JMA estimator $\hat{\mu}(\hat{\mathbf{w}})$ is a generalization of jackknife model selection. It is a smoother function of the data, as the discrete selection of individual models is replaced by the smooth selection of weights across models.

The set $\mathcal{H}_n^*$ can be the entire unit simplex $\mathcal{H}_n$ or a constrained subset. For the theoretical treatment of the regression model in Section 5 we restrict $\mathcal{H}_n^*$ to consist of discrete weights w^m from the set $\{0, \frac{1}{N}, \frac{2}{N}, \ldots, 1\}$ for some positive integer N. For empirical practice, however, we set $\mathcal{H}_n^* = \mathcal{H}_n$, the unrestricted unit simplex.

Even though $\mathrm{CV}_n(\mathbf{w})$ is a quadratic function of $\mathbf{w}$, the solution to (6) is not available in closed form due to the inequality constraints on $\mathbf{w}$. This is a quadratic programming problem, for which numerical solutions have been thoroughly studied and algorithms are widely available. For example, in the R language (R Development Core Team, 2009) it is solved using the quadprog package, in GAUSS by the qprog command, and in MATLAB by the quadprog command. Even when M_n is very large the solution to (6) is nearly instantaneous using any of these packages.

Algebraically, (6) is a constrained least-squares problem. The vector $\hat{\mathbf{w}}$ is the $M_n \times 1$ coefficient vector obtained by the constrained regression of $\mathbf{y}$ on the $n \times M_n$ matrix $\tilde{\mu}$. We can write this regression problem as $\mathbf{y} = \tilde{\mu}\hat{\mathbf{w}} + \tilde{\mathbf{e}}$. The jackknife weight vector $\hat{\mathbf{w}}$ is the weights which find the linear combination of the different estimators which yields the lowest squared error.

4. Asymptotic optimality

Define the average squared error

$$L_n(\mathbf{w}) = \frac{1}{n}\left(\mu - \hat{\mu}(\mathbf{w})\right)'\left(\mu - \hat{\mu}(\mathbf{w})\right) \tag{7}$$

and the expected squared error (or risk)

$$R_n(\mathbf{w}) = \mathbb{E}\left(L_n(\mathbf{w}) \mid \mathbf{X}\right). \tag{8}$$

The average squared error (7) may be viewed as a measure of in-sample fit, while the risk (8) is equivalent to out-of-sample

[1] We thank the Associate Editor for this suggestion.

[2] For a detailed overview of expected apparent, true and excess error, we direct the reader to Efron (1982, Chapter 7).

prediction mean-squared-error. Both are useful measures of the accuracy of $\hat{\mu}(\mathbf{w})$ for μ.

Following Li (1987), Andrews (1991), and Hansen (2007), we seek conditions under which the jackknife selected weight vector $\hat{\mathbf{w}}$ is asymptotically optimal in the sense of making $L_n(\mathbf{w})$ and $R_n(\mathbf{w})$ as small as possible among all feasible weight vectors $\mathbf{w}$. Specifically, we wish to show that

$$\frac{L_n(\hat{\mathbf{w}})}{\inf_{\mathbf{w}\in\mathcal{H}_n^*} L_n(\mathbf{w})} \to_p 1 \tag{OPT.1}$$

and

$$\frac{R_n(\hat{\mathbf{w}})}{\inf_{\mathbf{w}\in\mathcal{H}_n^*} R_n(\mathbf{w})} \to_p 1. \tag{OPT.2}$$

(OPT.1) means that the average squared error of the jackknife estimator is asymptotically as small as the average squared error of the infeasible best possible averaging estimator. (OPT.2) is a similar statement about the expected squared error. These are conventional optimality criteria for selection of series terms and bandwidths for nonparametric estimation.

The optimality statements (OPT.1) and (OPT.2) are oracle properties—that the selected weight vector is asymptotically equivalent to the infeasible best weight vector. A limitation of these optimality statements is that they restrict attention to estimators which are weighted averages of the original estimators $\hat{\mu}^m$. Thus the optimality is conditional on the given set of estimators. The averaging estimator is not necessarily better than an estimator not included in the original set of estimators. For example, if we restrict attention to nested regression models, then the optimality of the averaging estimator will depend on the ordering of the regressors. Different orderings will lead to different estimator sets, and thus to different averaging estimators and optimality bounds.

In this section we establish optimality for the JMA estimator under a set of high-level conditions. Define the jackknife average squared error $\tilde{L}_n(\mathbf{w}) = \frac{1}{n}(\mu - \tilde{\mu}(\mathbf{w}))'(\mu - \tilde{\mu}(\mathbf{w}))$ and jackknife expected squared error $\tilde{R}_n(\mathbf{w}) = \mathbb{E}\left(\tilde{L}_n(\mathbf{w}) \mid \mathbf{X}\right)$. Let $\lambda(\mathbf{A})$ denote the largest absolute eigenvalue of a matrix $\mathbf{A}$. For some integer $N \geq 1$, assume the following conditions hold almost surely.

$$\inf_i \sigma_i^2 \geq \underline{\sigma}^2 > 0 \tag{A.1}$$

$$\sup_i \mathbb{E}\left(e_i^{4(N+1)} \mid x_i\right) < \infty \tag{A.2}$$

$$\overline{\lim_{n\to\infty}} \max_{1\leq m\leq M_n} \lambda(\mathbf{P}_m) < \infty \tag{A.3}$$

$$\overline{\lim_{n\to\infty}} \max_{1\leq m\leq M_n} \lambda(\tilde{\mathbf{P}}_m) < \infty \tag{A.4}$$

$$\sup_{\mathbf{w}\in\mathcal{H}_n^*} \left|\frac{\tilde{R}_n(\mathbf{w})}{R_n(\mathbf{w})} - 1\right| \to 0 \tag{A.5}$$

$$\sum_{\mathbf{w}\in\mathcal{H}_n^*} (nR_n(\mathbf{w}))^{-(N+1)} \to 0. \tag{A.6}$$

Theorem 1. *If (A.1)–(A.6) hold, then $\hat{\mathbf{w}}$ is asymptotically optimal in the sense that (OPT.1) and (OPT.2) hold.*

Condition (A.1) excludes degenerate heteroskedasticity. (A.2) is a strong conditional moment bound. Condition (A.3) is quite mild as typical estimators satisfy $\lambda(\mathbf{P}_m) \leq 1$. Condition (A.4) is an analog of (A.3). Condition (A.5) says that as n gets large, the difference between the risk of the regular and leave-one-out estimators gets small, uniformly over the class of averaging estimators. This is standard for the application of cross-validation.

The key assumption is (A.6). It requires the weight vector set $\mathcal{H}_n^*$ to be countably discrete, and thus must be a strict subset of the unit simplex $\mathcal{H}_n$. It also implicitly incorporates a trade-off between the number of models permitted, the set of potential weights, and the fit of the individual models.

The choice of N involves a trade-off between the conditional moment bound in (A.2) and the summation in (A.6). As N increases (A.6) becomes easier to satisfy, but at the cost of a stronger moment bound in (A.2).

Theorem 1 is an application of results in Li (1987) and Andrews (1991). It is similar to Theorem 5.1 of Li (1987) and Theorem 4.1* of Andrews (1991) which also give high-level conditions for the asymptotic optimality of cross-validation, but makes two different assumptions. One difference is that our condition (A.6) is somewhat stronger than theirs, as (A.6) specifies uniform convergence over the index set $\mathcal{H}_n^*$, while Li and Andrews only require such convergence for sequences $\mathbf{w}_n$ for which $R_n(\mathbf{w}_n) \to 0$ or $\tilde{R}_n(\mathbf{w}_n) \to 0$. This distinction does not appear to exclude relevant applications, however. The other difference is that Li and Andrews assume that $\inf_{\mathbf{w}\in\mathcal{H}_n^*} R_n(\mathbf{w}) \to 0$, while this assumption is not required in our Theorem 1. Their assumption means that as $n \to \infty$ the optimal expected squared error declines to zero. In practice this means that the index set $\mathcal{H}_n^*$ needs to expand with n in a way so that the estimate $\hat{\mu}(\mathbf{w})$ gets close to the unknown mean μ. Our Theorem 1 is consistent with but does not require this extra assumption.

A condition which appears to be necessary for (A.6) is

$$\xi_n = \inf_{\mathbf{w}\in\mathcal{H}_n} nR_n(\mathbf{w}) \to \infty \tag{A.7}$$

almost surely as $n \to \infty$. This is identical to the condition in Hansen (2007) for Mallows weight selection, is quite similar to the conditions of Li (1987) and Andrews (1991), and its central role is emphasized by Shao (1997). It requires that all finite dimensional models are approximations, and thus the trade-off between bias and variance is present for all sample sizes.

In the context of regression (1), it means that there is no finite dimensional model which is correctly specified. This corresponds to the case of an infinite-order regression $\mu_i = \sum_{j=1}^{\infty} \theta_j x_{ji}$ with an infinite number of non-zero coefficients.

We can use (A.7) to obtain a crude but useful primitive condition for (A.6). Let $C_n = \#\left(\mathcal{H}_n^*\right)$ denote the cardinality of $\mathcal{H}_n^*$, the number of distinct weight vectors. For example, if $\mathcal{H}_n^*$ is a grid on the unit simplex with N grid points on each axis then $C_n = O\left(N^{M_n}\right)$. It is not hard to see that (A.7) plus $C_n\xi_n^{-(N+1)} \to 0$ are sufficient for (A.6). A useful implication of this bound is that (A.6) can hold for any sequence ξ_n if we take N sufficiently large and if C_n diverges sufficiently slowly. Furthermore, this bound holds regardless of the estimators used. The downside is that the allowable rate for C_n is quite slow. For example, if $\xi_n = n^{\delta}$ for some $0 < \delta < 1$, then $C_n\xi_n^{-(N+1)} \to 0$ requires $C_n = o\left(n^{N\delta}\right)$. This is compatible with a grid on the unit simplex only if $M_n = O(\ln n)$, which is highly restrictive. In the next section, we develop alternative primitive conditions for the case of linear regression estimators which allow M_n to be unbounded.

5. Linear regression

In this section we focus attention on linear regression estimates. These are estimators which take the form $\hat{\mu}^m = \mathbf{P}_m\mathbf{y}$ where $\mathbf{P}_m = \mathbf{X}^m\left(\mathbf{X}^{m\prime}\mathbf{X}^m\right)^{-1}\mathbf{X}^{m\prime}$ and $\mathbf{X}^m$ is an $n \times k_m$ matrix of regressors. In this setting it is relatively straightforward to find primitive conditions for assumptions (A.3)–(A.6).

To apply Theorem 1 we need to construct a countably discrete subset $\mathcal{H}_n^*$ of $\mathcal{H}_n$. Following Hansen (2007), we set $\mathcal{H}_n^*$ to be an equal-spaced grid on $\mathcal{H}_n$. For N defined in (A.2) let the weights w^m

be restricted to the set $\{0, \frac{1}{N}, \frac{2}{N}, \ldots, 1\}$, and let $\mathcal{H}_n^*$ be the subset of $\mathcal{H}_n$ restricted to this set of weights. We view this limitation as a technical artifact of the proof technique. This restriction becomes less binding as N increases (as $\mathcal{H}_n^*$ is dense in $\mathcal{H}_n$). For empirical practice we ignore the discrete restriction and select $\hat{\mathbf{w}}$ by minimizing $CV_n(\mathbf{w})$ over the unrestricted unit simplex.

One of our important advances over Hansen (2007) is that we allow both nested and non-nested regressors. The primary difficulty with the non-nested regression case is that the number of potential models can be quite large, and this is difficult to reconcile with Condition (A.6). The solution we propose is to limit the number of models for each dimension, but we do not restrict the number of models or the dimension of the largest model. Specifically, let $q_{jn} = \#\{m : k_m = j\}$, the number of models which have exactly j parameters. For example, if there are R_n regressors and the models include all regressions with a single regressor in addition to the intercept, then $q_{2n} = R_n$. If the models include all pairs of regressors in addition to the intercept, then $q_{3n} = R_n(R_n - 1)/2$. In general, if all subsets of r regressors are included models, then

$$q_{r+1,n} = \binom{R_n}{r} = \frac{R_n!}{r!\,(R_n - r)!}.$$

Set

$$\bar{q}_n = \max_{j \le M_n} q_{jn}. \tag{9}$$

This is the largest number of models of any given dimension. Our optimality theory restricts the rate of growth of $\bar{q}_n$, specifically that

$$\bar{q}_n = o\left(\xi_n^{1/N}\right) \tag{A.8}$$

where ξ_n is defined in (A.7). In the special case of nested models, then $\bar{q}_n = 1$ so (A.8) is automatically satisfied, even if there is a infinity of potential regressors. In general, (A.8) limits the number of non-nested models.

While Condition (A.7) requires that ξ_n diverges to infinity, its rate of divergence will be quite slow. Thus the allowable rate of divergence for $\bar{q}_n$ given in (A.8) is also slow. Hence (A.8) effectively restricts $\bar{q}_n$ to be bounded, or diverging quite slowly with sample size. As discussed above, in non-nested regression the number of potential subset models increases rapidly with the number of variables, thus implementation of (A.8) either restricts the number of variables R_n quite severely, or restricts the number of permitted subset models. While these restrictions may seem strong, they are still a significant advance in allowing for an unbounded number of non-nested models.

Finally, let $h_{ii}^m = x_i^{m\prime}\left(\mathbf{X}^{m\prime}\mathbf{X}^m\right)^{-1} x_i^m$ denote the i'th diagonal element of $\mathbf{P}_m$. Assume that

$$\max_{1 \le m \le M_n} \max_{1 \le i \le n} h_{ii}^m \to 0, \tag{A.9}$$

almost surely, as $n \to \infty$. This requires that the self-weights h_{ii}^m be asymptotically negligible for all models considered. This is a reasonable restriction, as it excludes only extremely unbalanced designs (where a single observation remains relevant asymptotically). Instead of (A.9), Li (1987) and Andrews (1991) assume the uniform bound $h_{ii}^m \le \Lambda k_m/n$ for some $\Lambda < \infty$, which is similar to (A.9), but neither condition is strictly stronger than the other. Conditions of this form are typical for the application of cross-validation.

Theorem 2. *Suppose that Conditions (A.1), (A.2), (A.7), (A.8) and (A.9) hold. Then Conditions (A.3)–(A.6) hold, and hence* $\hat{\mathbf{w}}$ *is asymptotically optimal in the sense that (OPT.1) and (OPT.2) hold.*

Theorem 2 is our main result. It shows that under minimal primitive conditions, the JMA estimator is asymptotically optimal in the class of weighted averages of linear estimators. The assumptions imply a trade-off between the moments of the error e_i and the number of grid points. Condition (A.2) means that the number of permitted grid points N on each axis is smaller than one-fourth of the number of moments of the error. When e_i has all moments finite then N can be selected to be arbitrarily large.

An important limitation of Theorem 2 is the restriction to random samples. The assumptions do not allow time-series dependence, even though a natural application would be to time-series models with unknown lag order, such as (vector) autoregressions or GARCH models. Extending this theory to allow dependent data would be desirable but technically challenging.

6. Finite-sample performance

In this section we investigate the finite sample mean squared error of the JMA estimator via a modest Monte Carlo experiment.

We follow Hansen (2007) and consider an infinite-order regression model of the form $y_i = \sum_{j=1}^{\infty} \theta_j x_{ji} + \epsilon_i$, $i = 1, \ldots, n$. We set $x_{1i} = 1$ to be the intercept, while the remaining x_{ji} are independent and identically distributed $N(0, 1)$ random variates. The error ϵ_i is also $N(0, \sigma_i^2)$ and is independent of the x_{ji}. For the homoskedastic simulation we set $\sigma_i = 1$ for all i, while for the heteroskedastic simulation we set $\sigma_i = x_{2i}^2$ which has a mean value of 1. (Other specifications for the heteroskedasticity were considered and produced similar qualitative results.)

The parameters are determined by the rule $\theta_j = c\sqrt{2\alpha}j^{-\alpha-1/2}$. The difficult aspect here is determining the appropriate truncation J for $\sum_{j=1}^{J} \theta_j x_{ji}$ (α is unknown). The issue of 'appropriate truncation' occurs frequently in model selection and is also related to the problem of bandwidth selection in kernel regression (smaller values of J could correspond to larger bandwidths and vice versa). The sample size is varied between $n = 25, 50, 75$, and 100. The parameter α is varied from $1/2$, 1, and $3/2$.[3] Larger values of α imply that the coefficients θ_j decline more quickly with j. The number of models M is determined by the rule $M = 3n^{1/3}$ (so $M = 9$, 11, 13, and 14 for the four sample sizes considered herein). The coefficient c was selected to control the population $R^2 = 2\alpha\zeta(1 + 2\alpha)c^2/(1 + \zeta(1 + 2\alpha)c^2)$ where $\zeta(k)$ is the Zeta function. We fix α and set c so that R^2 varies on a grid between approximately 0.1 and 0.8. The simulations use nested regression models with variables $\{x_{ji}, j = 1, \ldots, M_n\}$. We consider five estimators: (1) AIC model selection (AIC), (2) BIC model selection (BIC), (3) leave-one-out cross-validated model selection (CV), (4) Jackknife model averaging (JMA), and (5) Mallows model averaging (Hansen, 2007, MMA). To evaluate the estimators, we compute the risk (expected squared error). We do this by computing averages across 10,000 simulation draws. For each parameterization, we normalize the risk by dividing by the risk of the infeasible optimal least squares estimator (the risk of the best-fitting model m when averaged over all replications).

6.1. Homoskedastic error processes

We first compare risk under homoskedastic errors. The risk calculations are summarized in Fig. 1. The four panels in each graph display results for a variety of sample sizes. In each panel, risk (expected squared error) is displayed on the y axis and $c^2/(1+c^2)$ is displayed on the x axis.

It can be seen from Fig. 1 that for the homoskedastic data generating process, the proposed method dominates it peers for all

[3] We report results for $1/2$ only for space considerations. All results are available on request from the authors.

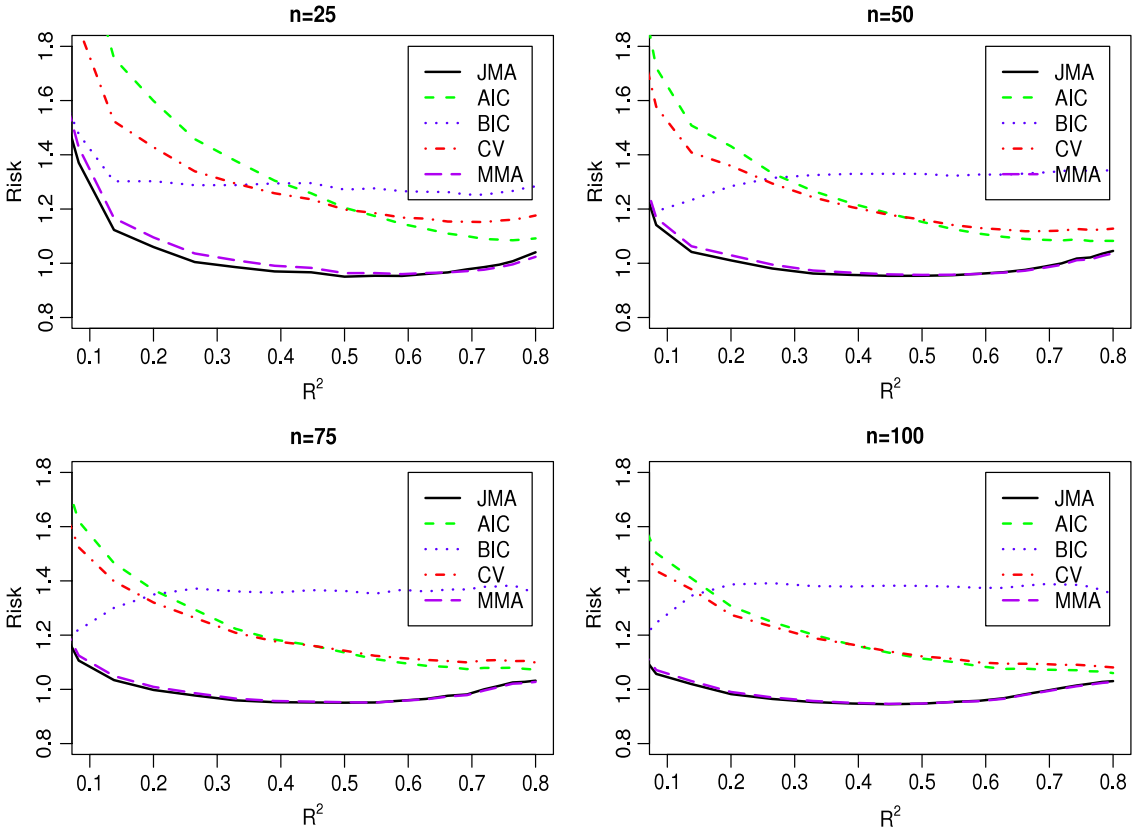

Fig. 1. Finite-sample performance, homoskedastic errors, $\alpha = 1/2$, $\sigma_i = 1$.

sample sizes and ranges of the population R^2 considered, though there is only a small gain to be had relative to the MMA method which performs very well overall and becomes indistinguishable as n increases beyond some modest level.

6.2. Heteroskedastic error processes

Next, we compare risk under heteroskedastic errors, and the risk calculations are summarized in Fig. 2.

It can be seen from Fig. 2 that for the heteroskedastic data generating process considered herein, the proposed method dominates it peers for all sample sizes and ranges of goodness of fit considered. Furthermore, the gain over that of the MMA approach is substantially larger than for the homoskedastic case summarized in Fig. 1, especially for small values of $c^2/(1+c^2)$.

7. Predicting earnings

We employ Wooldridge's (2003, pg. 226) popular 'wage1' dataset, a cross-section consisting of a random sample taken from the US Current Population Survey for the year 1976. There are 526 observations in total. The dependent variable is the log of average hourly earnings ('lwage'), while explanatory variables include dummy variables nonwhite, female, married, numdep, smsa, northcen, south, west, construc, ndurman, trcommpu, trade, services, profserv, profocc, clerocc, servocc, non-dummy variables educ, exper, tenure, and interaction variables nonwhite × educ, nonwhite × exper, nonwhite × tenure, female × educ, female × exper, female × tenure, married × educ, married × exper, and married × tenure.[4]

We presume there exists uncertainty about the appropriate model but need to predict earnings for a set of hold-out data. We consider two cases, (i) a set of thirty models ranging from the unconditional mean ($k = 1$) through a model that includes all variables listed above ($k = 30$), and (ii) a set of eleven models that use dummy variables female and married, and non-dummy variables educ, exper, and tenure that range from simple bivariate models that have as covariates each non-dummy regressor ($k = 2$) through a model that contains third-order polynomials in all non-dummy regressors and allows for interactions among all variables ($k = 64$).

Next, we shuffle the sample into a training set of size n_1 and an evaluation set of size $n_2 = n - n_1$, and select models via model selection based on AIC, BIC, and leave-one-out CV, then consider model average-based models using JMA and MMA. We also consider two nonparametric kernel estimators, namely, the local linear (NP_{ll}) and local constant (NP_{lc}) estimators with data-driven bandwidths modified to handle categorical and continuous data; see Li and Racine (2004) and Racine and Li (2004) for details. Note that the nonparametric model for case (i) contains more covariates that for case (ii), most of which are categorical. Finally, we evaluate the models on the independent hold-out data computing their average square prediction error (ASPE). We repeat this procedure 1000 times then report the median ASPE over the 1000 splits. We vary n_1 and consider $n_1 = 100, 200, 300, 400, 500$. Tables 1 and 2 report the median ASPE for this experiment. Entries greater than one indicate inferior performance relative to the JMA method.

Table 1 reveals that the proposed method delivers models that are no worse than existing model selection-based models or model-average based models across the range of sample sizes considered, often delivering an improved model. The MMA method works very well for case (i) and yields models comparable to those selected by the JMA method. However, this is not the situation

[4] See http://fmwww.bc.edu/ec-p/data/wooldridge/WAGE1.des for a full description of the data.

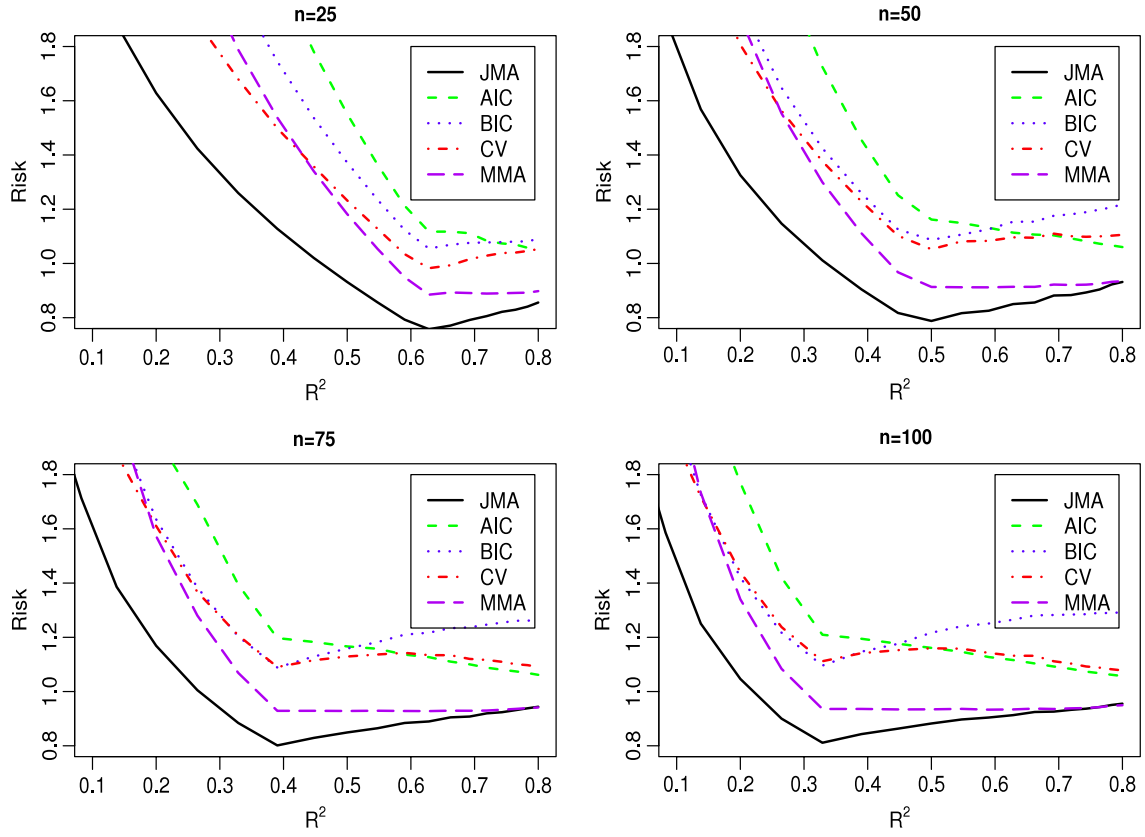

Fig. 2. Finite-sample performance, heteroskedastic errors, $\alpha = 1/2$, $\sigma_i = x_{i2}^2$.

Table 1
Case (i), relative out-of-sample predictive efficiency. Entries greater than one indicate inferior performance relative to the JMA method.

n	AIC	BIC	CV	MMA	NP_{ll}	NP_{lc}
100	1.10	1.34	1.07	1.01	2.76	1.77
200	1.04	1.04	1.02	1.00	4.08	3.01
300	1.03	1.01	1.02	1.00	3.06	2.89
400	1.01	1.01	1.03	1.00	3.34	2.69
500	1.00	1.01	1.01	1.00	2.78	2.56

Table 2
Case (ii), relative out-of-sample predictive efficiency. Entries greater than one indicate inferior performance relative to the JMA method.

n	AIC	BIC	CV	MMA	NP_{ll}	NP_{lc}
100	1.20	1.00	0.98	4.46	1.01	1.06
200	1.02	1.01	1.00	1.11	1.04	1.09
300	1.02	1.00	1.00	1.04	1.02	1.08
400	1.02	1.00	1.00	1.02	1.02	1.07
500	1.03	1.00	1.01	0.99	1.00	1.03

for case (ii), as Table 2 reveals. It would appear that the MMA method is somewhat sensitive to the estimate of σ^2 needed for its computation, and relying on a "large" approximating model Hansen (2007, pg. 1181) may not be sufficient to deliver optimal results. This issue deserves further study.

This application suggests that the proposed method could be of interest to those who predict using model-selection criterion. These results, particularly in light of the simulation evidence, suggest that the method could be recommended for general use though it would of course be prudent to base this recommendation on more extensive simulation evidence than that outlined in Section 6.

8. Conclusion

We propose a frequentist method for model averaging that does not preclude heteroskedastic settings, unlike many of its peers. The method is computationally tractable, is asymptotically optimal, and its finite-sample performance is better than that exhibited by its peers. An application to predicting wages is undertaken that demonstrates performance not worse than other methods of model selection considered by way of comparison, and performance that is often better for the range of sample sizes investigated.

There are many questions about the JMA estimator which are open for future research. What is the behavior of the optimal weights (the minimizers of (A.7)) as $n \to \infty$? What is the behavior of the selected weights? What is the asymptotic distribution of the parameter estimates? How can we construct confidence intervals for the parameters or the conditional mean? Can the theory be extended to allow for dependent data? These questions remain to be answered by future research.

Acknowledgments

We thank the Co-Editor, Associate Editor, and two referees for helpful comments. Hansen would like to gratefully acknowledge support from the National Science Foundation. Racine would like to gratefully acknowledge support from Natural Sciences and Engineering Research Council of Canada (NSERC:www.nserc.ca), the Social Sciences and Humanities Research Council of Canada (SSHRC:www.sshrc.ca), and the Shared Hierarchical Academic Research Computing Network (SHARCNET:www.sharcnet.ca).

Appendix. Mathematical proofs

Proof of Theorem 1. The Mallows criterion for a linear estimator $\mathbf{Pe}$ is the average squared residuals plus a penalty proportional to $\mathbb{E}\left(\mathbf{e}'\mathbf{Pe} \mid \mathbf{X}\right)$. Since $CV_n(\mathbf{w})$ defined in (5) is the average squared residual from the jackknife estimator $\tilde{\mu}(\mathbf{w}) = \tilde{\mathbf{P}}(\mathbf{w})\,\mathbf{y}$, and

$$\mathbb{E}\left(\mathbf{e}'\tilde{\mathbf{P}}(\mathbf{w})\,\mathbf{e} \mid \mathbf{X}\right) = \operatorname{tr}\left(\tilde{\mathbf{P}}(\mathbf{w})\,\boldsymbol{\Omega}\right) = 0 \tag{10}$$

since the diagonal elements of $\tilde{\mathbf{P}}(\mathbf{w})$ are zero and $\boldsymbol{\Omega} = \operatorname{diag}\left\{\sigma_1^2, \ldots, \sigma_n^2\right\}$ then $CV_n(\mathbf{w})$ is a Mallows criterion without the need for a penalty. As shown by Theorem 2.1 of Li (1987) as extended by Andrews (1991) to the heteroskedastic case, $\hat{\mathbf{w}}$, as the minimizer of $CV_n(\mathbf{w})$, is asymptotically optimal for the jackknife risk $\tilde{R}_n(\mathbf{w})$, that is

$$\frac{\tilde{R}_n\left(\hat{\mathbf{w}}\right)}{\inf_{\mathbf{w}\in\mathcal{H}_n^*}\tilde{R}_n(\mathbf{w})} \to_p 1, \tag{11}$$

under (A.1), (A.2),

$$\lim_{n\to\infty}\sup_{\mathbf{w}\in\mathcal{H}_n^*}\lambda(\tilde{\mathbf{P}}(\mathbf{w})) < \infty \tag{12}$$

and

$$\sum_{\mathbf{w}\in\mathcal{H}_n^*}\left(n\tilde{R}_n(\mathbf{w})\right)^{-(N+1)} \to 0 \tag{13}$$

almost surely. It is thus sufficient to verify (12) and (13) in order to establish (11).

First, the inequality $\lambda(\mathbf{A}+\mathbf{B}) \le \lambda(\mathbf{A}) + \lambda(\mathbf{B})$ plus (A.4) shows that

$$\lambda(\tilde{\mathbf{P}}(\mathbf{w})) \le \sum_{m=1}^{M_n} w^m \lambda(\tilde{\mathbf{P}}_m) < \infty \tag{14}$$

uniformly in $\mathbf{w} \in \mathcal{H}_n^*$, almost surely as $n \to \infty$ establishing (12). Second, using (A.5) and (A.6)

$$\begin{aligned}\sum_{\mathbf{w}\in\mathcal{H}_n^*}\left(n\tilde{R}_n(\mathbf{w})\right)^{-(N+1)} &= \sum_{\mathbf{w}\in\mathcal{H}_n^*}(nR_n(\mathbf{w}))^{-(N+1)}\left(\frac{R_n(\mathbf{w})}{\tilde{R}_n(\mathbf{w})}\right)^{N+1}\\ &\le \left(\sum_{\mathbf{w}\in\mathcal{H}_n^*}(nR_n(\mathbf{w}))^{-(N+1)}\right)\\ &\quad\times\sup_{\mathbf{w}\in\mathcal{H}_n^*}\left(\frac{R_n(\mathbf{w})}{\tilde{R}_n(\mathbf{w})}\right)^{N+1}\\ &\to 0\end{aligned} \tag{15}$$

almost surely, establishing (13). Thus (11) holds. Combined with (A.5) this implies (OPT.2). Eq. (7.1*) of Andrews (1991) states that

$$\sup_{\mathbf{w}\in\mathcal{H}_n^*}\left|\frac{L_n(\mathbf{w})}{R_n(\mathbf{w})} - 1\right| \to 0$$

which holds under our assumptions. Combined with (OPT.2) this implies (OPT.1), completing the proof. □

Proof of Theorem 2. We will show that Conditions (A.3) through (A.6) hold, so an application of Theorem 1 yields (OPT.1) and (OPT.2).

First, since the $\mathbf{P}_m$ are idempotent, Condition (A.3) holds trivially.

Second, let $\bar{\lambda}_m$ and $\underline{\lambda}_m$ denote the largest and smallest diagonal elements of $\mathbf{D}_m$. Condition (A.9) implies that

$$\bar{\lambda}_m = 1 + o(1) = \underline{\lambda}_m \tag{16}$$

uniformly in $m \le M_n$. Using (4), this implies

$$\lambda(\tilde{\mathbf{P}}_m) \le 1 + \lambda(\mathbf{D}_m(\mathbf{P}_m - \mathbf{I})) \le 1 + o(1)$$

uniformly in m, so (A.4) holds.

We next show (A.5). We can calculate that

$$\begin{aligned}nR_n(\mathbf{w}) &= \mu'\left(\mathbf{I} - \mathbf{P}(\mathbf{w})'\right)\left(\mathbf{I} - \mathbf{P}(\mathbf{w})\right)\mu + \operatorname{tr}\left(\mathbf{P}(\mathbf{w})'\mathbf{P}(\mathbf{w})\,\boldsymbol{\Omega}\right)\\ &= \sum_{m=1}^{M_n}\sum_{\ell=1}^{M_n} w^m w^\ell\left[\mu'\left(\mathbf{I} - \mathbf{P}_m'\right)\left(\mathbf{I} - \mathbf{P}_\ell\right)\mu + \operatorname{tr}\left(\mathbf{P}_m'\mathbf{P}_\ell\boldsymbol{\Omega}\right)\right].\end{aligned} \tag{17}$$

The first equality is given in Andrews (1991) and the second uses (3). Similarly,

$$n\tilde{R}_n(\mathbf{w}) = \sum_{m=1}^{M_n}\sum_{\ell=1}^{M_n} w^m w^\ell\left[\mu'(\mathbf{I} - \tilde{\mathbf{P}}_m')(\mathbf{I} - \tilde{\mathbf{P}}_\ell)\mu + \operatorname{tr}(\tilde{\mathbf{P}}_m'\tilde{\mathbf{P}}_\ell\boldsymbol{\Omega})\right]. \tag{18}$$

It is sufficient to establish that

$$\operatorname{tr}(\tilde{\mathbf{P}}_m'\tilde{\mathbf{P}}_\ell\boldsymbol{\Omega}) = \operatorname{tr}\left(\mathbf{P}_m'\mathbf{P}_\ell\boldsymbol{\Omega}\right)(1 + o(1)) \tag{19}$$

and

$$\mu'(\mathbf{I} - \tilde{\mathbf{P}}_m')(\mathbf{I} - \tilde{\mathbf{P}}_\ell)\mu = \mu'\left(\mathbf{I} - \mathbf{P}_m'\right)\left(\mathbf{I} - \mathbf{P}_\ell\right)\mu\,(1 + o(1)) \tag{20}$$

where the $o(1)$ terms are uniform in $m \le M_n$ and $\ell \le M_n$. First take (19). Using (4) and the fact that the diagonal elements of $\tilde{\mathbf{P}}_m$ are zero, (16), again (4), and finally (16)

$$\begin{aligned}\operatorname{tr}(\tilde{\mathbf{P}}_m'\tilde{\mathbf{P}}_\ell\boldsymbol{\Omega}) &= \operatorname{tr}(\tilde{\mathbf{P}}_m'\mathbf{D}_\ell\mathbf{P}_\ell\boldsymbol{\Omega}) - \operatorname{tr}(\tilde{\mathbf{P}}_m'\mathbf{D}_\ell\boldsymbol{\Omega}) + \operatorname{tr}(\tilde{\mathbf{P}}_m'\boldsymbol{\Omega})\\ &= \operatorname{tr}(\tilde{\mathbf{P}}_m'\mathbf{D}_\ell\mathbf{P}_\ell\boldsymbol{\Omega})\\ &= \operatorname{tr}(\tilde{\mathbf{P}}_m'\mathbf{P}_\ell\boldsymbol{\Omega})(1 + o(1))\\ &= \left(\operatorname{tr}\left(\left(\mathbf{P}_m' - \mathbf{I}\right)\mathbf{D}_m\mathbf{P}_\ell\boldsymbol{\Omega}\right) + \operatorname{tr}\left(\mathbf{P}_\ell\boldsymbol{\Omega}\right)\right)(1 + o(1))\\ &= \left(\operatorname{tr}\left(\left(\mathbf{P}_m' - \mathbf{I}\right)\mathbf{P}_\ell\boldsymbol{\Omega}\right) + \operatorname{tr}\left(\mathbf{P}_\ell\boldsymbol{\Omega}\right)\right)(1 + o(1))\\ &= \operatorname{tr}\left(\mathbf{P}_m'\mathbf{P}_\ell\boldsymbol{\Omega}\right)(1 + o(1))\end{aligned}$$

establishing (19). Next, (4) implies $\mathbf{I} - \tilde{\mathbf{P}}_m = \mathbf{D}_m(\mathbf{I} - \mathbf{P}_m)$, which combined with (16) directly implies (20). We have established (19) and (20) which are sufficient for (A.5).

Finally, we show (A.6). Without loss of generality, arrange the models (estimators) so that they are weakly ordered by the number of parameters k_m, e.g. $k_1 \le k_2 \le \cdots \le k_{M_n}$. As in the proof of Theorem 1 of Hansen (2007), for integers $1 \le j_1 \le j_2 \le \cdots \le j_N$ let $\mathbf{w}_{j_1,j_2,\ldots,j_N}$ be the weight vector which sets $w^{j_l} = 1/N$ for $l = 1, \ldots, N$, and the remainder zero.

Pick a sequence ψ_n which satisfies

$$\psi_n = o\left(\xi_n^{1+1/N}\right) \tag{21}$$

yet

$$\bar{q}_n^{1+N} = o(\psi_n) \tag{22}$$

which is feasible since

$$\bar{q}_n^{1+N} = o\left(\xi_n^{1+1/N}\right)$$

under (A.8).

We then have

$$\sum_{\mathbf{w}\in\mathcal{H}_n^*} (nR_n(\mathbf{w}))^{-(N+1)}$$

$$= \sum_{j_N=1}^{M_n} \sum_{j_{N-1}=1}^{j_N} \cdots \sum_{j_1=1}^{j_2} \left(nR_n(\mathbf{w}_{j_1,j_2,\ldots,j_N})\right)^{-(N+1)} \le B_{1n} + B_{2n} \quad (23)$$

where

$$B_{1n} = \sum_{j_N=1}^{\psi_n} \sum_{j_{N-1}=1}^{j_N} \cdots \sum_{j_1=1}^{j_2} \left(nR_n(\mathbf{w}_{j_1,j_2,\ldots,j_N})\right)^{-(N+1)} \quad (24)$$

and

$$B_{2n} = \sum_{j_N=\psi_n+1}^{\infty} \sum_{j_{N-1}=1}^{j_N} \cdots \sum_{j_1=1}^{j_2} \left(nR_n(\mathbf{w}_{j_1,j_2,\ldots,j_N})\right)^{-(N+1)}. \quad (25)$$

Take (24). As the sum has less than ψ_n^N elements, we apply $nR_n(\mathbf{w}) \ge \xi_n$ from (A.7) and find that

$$B_{1n} \le \psi_n^N \xi_n^{-(N+1)} \le o(1) \quad (26)$$

under (21).

To bound (25) we use the simple inequality

$$\begin{aligned} nR_n\left(\mathbf{w}_{j_1,j_2,\ldots,j_N}\right) &\ge \operatorname{tr}\left(\mathbf{P}\left(\mathbf{w}_{j_1,j_2,\ldots,j_N}\right)\mathbf{P}\left(\mathbf{w}_{j_1,j_2,\ldots,j_N}\right)\boldsymbol{\Omega}\right) \\ &\ge \underline{\sigma}^2 \operatorname{tr}\left(\mathbf{P}\left(\mathbf{w}_{j_1,j_2,\ldots,j_N}\right)\mathbf{P}\left(\mathbf{w}_{j_1,j_2,\ldots,j_N}\right)\right) \\ &= \frac{\underline{\sigma}^2}{N^2}\left(\sum_{l=1}^{N} \operatorname{tr}\left(\mathbf{P}_{j_l}\right) + \sum_{l=1}^{N}\sum_{m\ne l} \operatorname{tr}\left(\mathbf{P}_{j_l}\mathbf{P}_{j_m}\right)\right) \\ &\ge \frac{\underline{\sigma}^2}{N^2}\left(\sum_{l=1}^{N} \operatorname{tr}\left(\mathbf{P}_{j_l}\right)\right) \\ &= \frac{\underline{\sigma}^2}{N^2}\left(\sum_{l=1}^{N} k_{j_l}\right) \\ &\ge \frac{\underline{\sigma}^2}{N^2} k_{j_N} \\ &\ge \frac{\underline{\sigma}^2}{N^2}\frac{j_N}{\bar{q}_n} \end{aligned} \quad (27)$$

where the first inequality is from (17), the second inequality uses $\operatorname{tr}(\mathbf{AB}) \ge \operatorname{tr}(\mathbf{A})\lambda_{\min}(\mathbf{B})$ and $\lambda_{\min}(\boldsymbol{\Omega}) \ge \underline{\sigma}^2$, and the following equality uses the definition of $\mathbf{P}(\mathbf{w})$ and the weights $\mathbf{w}_{j_1,j_2,\ldots,j_N}$. The third inequality is

$$\operatorname{tr}\left(\mathbf{P}_{j_l}\mathbf{P}_{j_m}\right) \ge \operatorname{tr}\left(\mathbf{P}_{j_l}\right)\lambda_{\min}\left(\mathbf{P}_{j_m}\right) = 0$$

since the matrices $\mathbf{P}_{j_m}$ are idempotent. The final inequality (27) follows from definition (9) and the ordering of the models by the number of parameters k_j.

Using inequality (27) we obtain

$$\begin{aligned} B_{2n} &\le \sum_{j_N=\psi_n+1}^{\infty} \sum_{j_{N-1}=1}^{j_N} \cdots \sum_{j_1=1}^{j_2} \left(\frac{\underline{\sigma}^2}{N^2}\frac{j}{\bar{q}_n}\right)^{-(N+1)} \\ &\le \frac{N^{2(N+1)}}{\underline{\sigma}^{2(N+1)}} \bar{q}_n^{N+1} \sum_{j=\psi_n+1}^{\infty} j^{-2} \\ &\le \frac{N^{2(N+1)}}{\underline{\sigma}^{2(N+1)}} \bar{q}_n^{1+N} \psi_n^{-1} \\ &\le o(1) \end{aligned} \quad (28)$$

the last inequality by (22). Eqs. (23), (26), and (28) establish (A.6).

Conditions (A.3)–(A.6) have been verified as desired. □

References

Akaike, H., 1970. Statistical predictor identification. Annals of the Institute of Statistical Mathematics 22, 203–217.

Allen, D.M., 1974. The relationship between variable selection and data augmentation and a method for prediction. Technometrics 16, 125–127.

Andrews, D.W.K., 1991. Asymptotic optimality of generalized C_L, cross-validation, and generalized cross-validation in regression with heteroskedastic errors. Journal of Econometrics 47, 359–377.

Breiman, L., 1996. Stacked regressions. Machine Learning 24, 49–64.

Buckland, S.T., Burnhamn, K.P., Augustin, N.H., 1997. Model selection: an integral part of inference. Biometrics 53, 603–618.

Cohen, A., 1966. All admissible linear estimates of the mean vector. Annals of Mathematical Statistics 37, 458–463.

Efron, B., 1982. The Jackknife, the Bootstrap, and Other Resampling Plans. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 19103.

Geisser, S., 1974. The predictive sample reuse method with applications. Journal of the American Statistical Association 70, 320–328.

Goldenshluger, A., 2009. A universal procedure for aggregating estimators. Annals of Statistics 37, 542–568.

Hansen, B.E., 2007. Least squares model averaging. Econometrica 75, 1175–1189.

Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: a tutorial. Statistical Science 14, 382–417.

Juditsky, A., Nemirovski, A., 2000. Functional aggregation for nonparametric estimation. Annals of Statistics 28, 681–712.

Li, K.-C., 1987. Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. The Annals of Statistics 15, 958–975.

Li, Q., Racine, J.S., 2004. Cross-validated local linear nonparametric regression. Statistica Sinica 14 (2), 485–512.

R Development Core Team, 2009. R: a language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN: 3-900051-07-0. URL: http://www.R-project.org.

Racine, J., 1997. Feasible cross-validatory model selection for general stationary processes. Journal of Applied Econometrics 12, 169–179.

Racine, J.S., Li, Q., 2004. Nonparametric estimation of regression functions with both categorical and continuous data. Journal of Econometrics 119 (1), 99–130.

Schwarz, G., 1978. Estimating the dimension of a model. The Annals of Statistics 6, 461–464.

Shao, J., 1997. An asymptotic theory for linear model selection. Statistica Sinica 7, 221–264.

Stone, C., 1974. Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society. Series B 36, 111–147.

Wahba, G., Wold, S., 1975. A completely automatic french curve: fitting spline functions by cross-validation. Communications in Statistics 4, 1–17.

Wan, A.T.K., Zhang, X., Zou, G., 2010. Least squares model averaging by mallows criterion. Journal of Econometrics 156 (2), 277–283.

Wolpert, D., 1992. Stacked generalization. Neural Networks 5, 241–259.

Wooldridge, J.M., 2003. Introductory Econometrics. Thompson South-Western.

Yang, Y., 2001. Adaptive regression by mixing. Journal of the American Statistical Association 96, 574–588.

Yang, Y., 2004. Aggregating regression procedures to improve performance. Bernoulli 10, 25–47.