Econometrics **2013**, *1*, 157-179; doi:10.3390/econometrics1020157

OPEN ACCESS

econometrics

ISSN 2225-1146

www.mdpi.com/journal/econometrics

Article

Parametric and Nonparametric Frequentist Model Selection and Model Averaging

Aman Ullah * and Huansha Wang

Department of Economics, University of California, Riverside, CA 92521-0427, USA;
E-Mail: hwang022@ucr.edu

* Author to whom correspondence should be addressed; E-Mail: aman.ullah@ucr.edu;
Tel.: +1-951-827-1591; Fax: +1-951-827-5685.

Received: 27 June 2013; in revised form: 17 July 2013 / Accepted: 13 September 2013 / Published: 20 September 2013

Abstract: This paper presents recent developments in model selection and model averaging for parametric and nonparametric models. While there is extensive literature on model selection under parametric settings, we present recently developed results in the context of nonparametric models. In applications, estimation and inference are often conducted under the selected model without considering the uncertainty from the selection process. This often leads to inefficiency in results and misleading confidence intervals. Thus an alternative to model selection is model averaging where the estimated model is the weighted sum of all the submodels. This reduces model uncertainty. In recent years, there has been significant interest in model averaging and some important developments have taken place in this area. We present results for both the parametric and nonparametric cases. Some possible topics for future research are also indicated.

Keywords: nonparametric; model selection; model averaging

---

1. Introduction

Over the last several years many econometricians and statisticians have persistently devoted their efforts in finding various paths to the true model. The uncertainty in correctly specifying the regression model has resulted in a large amount of literature in two major directions: firstly, what variables are to be included and secondly, how they are related with the dependent variable in the model. Thus

"what" refers to determining the variables to be included in constructing the model and "how" refers to finding the correct functional form, e.g., parametric (specifications like linear, quadratic, *etc.*), or in general, nonparametric smoothing methods that do not require specifying a parametric functional form but instead let the data search for a suitable function that describes well the available data, see [1,2] among others.

To determine "what", model selection was first introduced, and it has a huge literature in statistics and econometrics. In fact, in recent years, model selection (variable selection) procedures have become more popular due to the emergence of econometric and statistical models with high dimension (large number) variables. As examples, in labor economics, wage equations can have a large number of regressors [3] and in financial econometrics, portfolio allocation may be among hundreds or thousands of stocks [4]. Such models raise additional challenges of econometric modeling and inference along with the selection of variables. Different tools have been developed based on various estimation criteria. The majority of such procedures involve variable selection by minimizing penalized loss functions based on the least squares and the log-likelihood, and their variants. The adjusted R^2 and residuals sum of squares are the usual variable selection procedures without any penalization. Among the penalized procedures we have Akaike information criterion (AIC) [5], Mallows C_p procedure [6], Bayesian information criterion (BIC) by [7], cross-validation method by [8], generalized cross-validation (GCV) by [9], and the focused information criterion (FIC) by [10]. We note that the traditional AIC and BIC are based on least squares (LS), maximum likelihood (ML), or Bayesian principles, and the penalization is based on the l_0-norm for the parameters entering in the model, with the result penalization is proportional to the number of nonzero parameters. Both AIC and BIC are variable selection procedures and do not provide estimators simultaneously. On the other hand the bridge estimator in [11,12] uses the l_q-norm ($q > 0$), and for $0 < q \leq 1$ provides a way to combine variable selection and parameter estimation simultaneously. Within this class the least absolute shrinkage and selection operator (LASSO; $q = 1$) has become the most popular. For $q = 2$ we get the ridge estimator [13]. For a detailed review of model selection in high dimensional modeling, see [14], and the books [15,16]. Similarly, in the context of empirical likelihood estimation and generalized methods of moments estimators, model selection criteria have been introduced by [17,18], among others.

Model selection is an important step for empirical policy evaluation and forecasting. However, it may produce unstable estimators because of bias in model selection. For example, a small data perturbation or an alternative selection procedure may give a different model. Reference [19] shows that AIC selection results in distorted inference, and [20] explores the negative impact on confidence regions. Reference [21] gives conditions under which post model selection estimators are adaptive, but see [22,23] for their comments that they cannot be uniformly estimated. For a selected model with unstable estimators, [24] provides bagging or bootstrap averaging procedure to reduce their variances for the i.i.d. data, and by [25] for the dependent time series data. But this averaging does not always work, e.g., for large samples and/or in entire parameter space.

Taking the above reasons into consideration, model averaging is introduced as an alternative to model selection. Unlike in model selection, where the model uncertainty is dealt with by econometricians selecting one model from a set of models, in model averaging, we resolve the uncertainty by averaging over the set of models. There is large recent literature on Bayesian model averaging (BMA) and more

recently, on frequentist model averaging (FMA). Among the BMA contributions, model uncertainty is considered by setting a prior probability to each candidate model, see [26–30]; for interesting applications in econometrics, see, e.g., [31–33]. Also, see [10] for comments on the BMA approach. The main focus here is on the FMA method, which is totally determined by data only and assumes no priors, and it has received much attention in recent years, see [34–41]. Reference [10] provides asymptotic theory. For applications, see [16,42,43]. The concept behind the FMA estimators is related to the ideas of combining procedures based on the same data, which have been considered before in several research areas. For instance, [44] introduces forecast combination and [45,46] suggest combining parametric and kernel estimators of density and regression respectively. Other works include bootstrap based averaging ("stacking") by [24,47,48], information theoretic method to combine density by [49,50], and the mixing of experts models by [51,52]. Similar kinds of combining have been used in computational learning theory by [53,54] and in information theory by [55].

Related to "how", or rather determining the unknown functional forms of econometric models, we use data based nonparametric procedures (e.g., kernel, smoothing spline, series approximation). See, for example,[1,2,56,57], for kernel smoothing procedures, [58] for the spline methods, and [59,60] for the series methods. These procedures help in dealing with the problems of bias and inconsistency in estimation and testing due to misspecifying functional forms. Because of this recent developments on nonparametric model selection and model averaging have taken place.

The current paper is hence focused on a review of parametric and nonparametric approaches to model selection and model averaging mainly from a frequentist point of view, and for independently and identically distributed (i.i.d.) observations. Earlier [14] provides a review of parametric model selections, [61] surveys the FMA estimation, and [62] provides variable selection in semiparametric regression models. To distinguish, our paper hence concentrates on the review of frequentist model selection and model averaging under both parametric and nonparametric settings.

The paper is organized as follows. We first introduce a review of parametric model selection and parametric model averaging in Section 2. Then, in Section 3 we present nonparametric model selection and model averaging procedures. A conclusion follows in Section 4.

2. Parametric Model Selection and Model Averaging

2.1. Model Selection

Let us consider y_i as a dependent variable and $x_i = (x_{i1,}..., x_{iq})'$ a $q \times 1$ vector of explanatory variables/covariates. Then the linear regression model can be written as

$$y_i = x_i'\beta + u_i = \sum_{j=1}^{q} x_{ij}\beta_j + u_i,\ i = 1, ..., n \tag{1}$$

or

$$y = X\beta + u \tag{2}$$

where y is $n \times 1$, X is $n \times q$, $\beta = (\beta_1, ..., \beta_q)'$, and u is $n \times 1$.

Among the well known procedures for model selection, often used routinely, we are looking at the goodness of fit R^2, adjusted R^2 (R_a^2), and residuals sum of squared (RSS) given by

$$R^2 = 1 - \frac{\sum \hat{u}_i^2}{\sum (y_i - \bar{y})^2}, \; R_a^2 = 1 - \frac{(n-1)\sum \hat{u}_i^2}{(n-q)\sum (y_i - \bar{y})^2}, \; RSS = \sum (\hat{u}_i)^2 \tag{3}$$

where $0 \leq R^2 \leq 1$. The model with the highest R^2 (or R_a^2) or smallest RSS is chosen. However R^2 increases or RSS decreases, monotonically as q increases. Further, between R^2 and R_a^2, $Bias(R_a^2) \leq Bias(R^2)$ but $V(R_a^2) \geq V(R^2)$. Thus R_a^2 may not always be statistically more efficient ($MSE(R_a^2) \leq MSE(R^2)$), see [63] for further detail. Thus R_a^2 and RSS are not preferred measures of goodness of fit or model selection. Recently [64] develops a model selection procedure based on the "mean squared prediction error" denoted by MSPE. Consider $(x_{i1}, ..., x_{iq}, z_i)$, $i = 1, ..., n$, as a new observed sample in which z_i is the "new observed value" and $\hat{y}_i$ is such that $MSPE = \sum E(z_i - \hat{y}_i)^2/n = \sigma_u^2(n+q+1)/n$. When a model has $q = 0$ (no explanatory variable), $MSPE = \sigma_y^2(n+1)/n$. Then, using the unbiased estimator of $MSPE_0 = FPE_0 = s_y^2(n+1)/n$, and of $MSPE = FPE$ as $s_{\hat{u}}^2(n+q+1)/n$, in [64] introduces

$$R_{FPE}^2 = 1 - \frac{FPE}{FPE_0} = \frac{(n-1)(n+q+1)R^2 - 2qn}{(n-q-1)(n+1)}$$

such that $R_{FPE}^2 \leq R_a^2 \leq R^2$ where FPE represents final prediction error. The statistical properties of the bias and MSE of R_{FPE}^2, compared to those of R_a^2 and R^2, are analyzed in [65]. Reference [64] has demonstrated that one of the exciting advantages of R_{FPE}^2 is that it can be used for choosing a model with the best prediction ability. Furthermore, R_{FPE}^2 not only overcomes inflation in R^2, it also avoids the problem of selecting an overfitted model with some irrelevant explanatory variables due to using R_a^2. In addition, they indicate that R_{FPE}^2 and AIC, discussed below, are asymptotically equivalent and in model selection R_{FPE}^2 is perfectly consistent with using AIC and is closest with BIC. Thus R_{FPE}^2 can be used simultaneously for goodness of fit as well as for model selection.

2.1.1. AIC, TIC, and BIC

Now we turn to the methods of model selection, AIC in [5], Takeuchi informaiton criterion (TIC) in [66], and BIC in [7]. For this, we first note that if $f(y)$ is an unknown true density, and $g(y, \theta)$ is an assumed density then the Kullback-Leibler Information Criterion (KLIC) is given by

$$D(f, g) = KLIC(f, g) = E_f \log(\frac{f(y)}{g(y, \theta)}) = E_f \log f(y) - E_f \log g(y, \theta),$$

where E_f is the expectation with respect to $f(y)$. This is an expected "surprise" from knowing f is in fact the true density of y. We note that $D(f, g) \geq 0$ where equality holds if and only if $g = f$ almost everywhere. Further $E_f \log f(y)$ is called the entropy of distribution f; for more on entropy and information, see [67,68].

A concept related to entropy is the quasi maximum likelihood estimator (QMLE) $\hat{\theta}_{QML}$ which maximizes the quasi log-likelihood function

$$L(\theta) = L_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log g(y_i, \theta)$$

based on the random sample $\mathbf{Y} = (y_1, ..., y_n)$ from $f(y)$. Since $L_n(\theta) \to^p E_f[\log g(y_1, \theta)]$, it is expected that $\hat{\theta}_{QML}$ converges in probability to the maximizer θ^* of $E_f[\log g(y_1, \theta)]$ under suitable conditions. Since $E_f[\log f(y_1)]$ does not depend on θ, QMLE minimizes a random function which converges to

$$KLIC(f, g) = E_f \log f(y_1) - E_f \log g(y_1, \theta) = D(f, g)$$

Thus $\hat{\theta}_{QML} \to^p \theta^*$ where $\theta^* = arg \min_\theta D(f, g(\theta))$ is often referred to as the pseudo-true value of θ. It is well known that under some regularity conditions

$$\sqrt{n}(\hat{\theta}_{QML} - \theta^*) \to^d N(0, G(\theta^*)^{-1} I(\theta^*) G(\theta^*)^{-1})$$

where $G(\theta) = -E_g[\partial^2 \log g(y, \theta)/\partial\theta\partial\theta']$ and $I(\theta) = E_g[\partial \log g(y_1, \theta)\partial \log g(y_1, \theta)/\partial\theta\partial\theta']$. When $f(\cdot) = g(\cdot, \theta^*)$, $G(\theta^*) = I(\theta^*)$ and $\hat{\theta}_{QML}$ is the MLE and it is asymptotically efficient.

Now consider the fitted density $\hat{g}(y) = g(y, \hat{\theta}_{QML})$ and

$$\begin{aligned} KLIC(f, \hat{g}) &= E_f \log(\frac{f(y)}{\hat{g}(y)}) \\ &= c - E_y \log g(y, \hat{\theta}_{QML}) \end{aligned}$$

where $c = \int f(y) \log(f(y)) dy$ is free of the fitted model and $E_y(\cdot)$ denotes the expectation with respect to the true density of y, *i.e.*, $g(y)$ here. Then $E[KLIC(f, \hat{g})] = c - E_{\mathbf{Y}} E_y[\log g(y, \hat{\theta}_{QML})] = c - n^{-1} \sum E_{\mathbf{Y}} E_{y_i}[\log g(y_i, \hat{\theta}_{QML})]$ where $\mathbf{Y}$ and y are independent. The expected KLIC can be interpreted as the expected likelihood when $\mathbf{Y}$ is used for $\hat{\theta}_{QML}$, and an independent sample y (with one observation here) used for evaluation. In linear regression, the expected KLIC is the expected squared prediction error. Dropping c, and using second order Taylor expansion, it can be shown that

$$nT = E[KLIC(f, \hat{g})] = -E[L_n(\hat{\theta})] + tr[I(\theta^*)G(\theta^*)^{-1}].$$

Further, an asymptotically unbiased estimator of T can be written as

$$\hat{T} = -n^{-1}\{L_n(\hat{\theta}) - tr(\hat{I}\hat{G}^{-1})\}$$

where $L_n(\hat{\theta}) = \log g(\mathbf{Y}, \hat{\theta})$, $\hat{I}\hat{G}^{-1}$ is a consistent estimator of $I(\theta^*)G(\theta^*)^{-1}$ in which $\hat{I} = \frac{1}{n} \sum \frac{\partial \log g(y_i, \theta)}{\partial \theta} \frac{\partial \log g(y_i, \theta)}{\partial \theta'}$ and $\hat{G} = -\frac{1}{n} \sum \partial^2 \log g(y_i, \theta)/\partial\theta\partial\theta'$.

When the model is correctly specified, that is $g(y, \theta^*)$=$f(y)$, $G(\theta^*)$=$I(\theta^*)$ and $tr(I(\theta^*)G(\theta^*)^{-1}) = q$,

$$\hat{T} = -n^{-1} L_n(\hat{\theta}) + n^{-1} q$$

which is related with AIC given by $2\hat{T}$:

$$AIC = -\frac{2L_n(\hat{\theta})}{n} + \frac{2q}{n}. \tag{4}$$

Thus, we can think of AIC as an estimate of the expected 2KLIC based on the assumption that the model is correctly specified. Therefore, selecting a model based on the smallest AIC amounts to choosing the best-fitting model in the sense of having the smallest KLIC. A robust AIC by Takeuchi [66], known as the Takeuchi Information Criterion (TIC), is

$$TIC = -\frac{2L_n(\hat{\theta})}{n} + \frac{2tr(\hat{I}\hat{G}^{-1})}{n},$$

which, unlike AIC, does not require $g(y, \theta)$ to be correctly specified. In general, picking models with the smallest AIC/TIC is selecting fitted models whose densities are close to the true density.

We note that in a linear regression model, the minimization of the AIC reduces to the minimization of the following

$$AIC = \log \hat{\sigma}^2 + \frac{2q}{n}$$

where $\hat{\sigma}^2 = \frac{\hat{u}'\hat{u}}{n}$. It can be shown that $G(\theta^*) = I(\theta^*)$ if $u_i|x_i \sim N(0, \sigma^2)$. Thus AIC is more appropriate under normality, otherwise it is an approximation for the non-normal and heteroskedastic regression cases.

Further, in a linear regression case, the minimization of TIC can be shown as the minimization of

$$TIC = \log \hat{\sigma}^2 + \frac{2}{n\hat{\sigma}^2} \sum_{i=1}^{n} h_i \hat{u}_i^2 + \frac{\hat{k}_4}{n}$$

where $\hat{k}_4 = \frac{1}{n\hat{\sigma}^4} \sum_{i=1}^{n} (\hat{u}_i^2 - \hat{\sigma}^2)^2$ and $h_i = x_i'(X'X)^{-1}x_i$. When the errors are homoskedastic and normal,

$$TIC \simeq \log \hat{\sigma}^2 + \frac{2(q+1)}{n}$$

which is close to AIC. Although differences may arise under heteroskedasticity and nonnormality. However, as we change models, typically the results $\hat{u}_i^2$ and hence $\hat{k}_4$ may not change much. In this case, TIC and AIC may give similar model selection results.

We note that the BIC due to [7] is

$$BIC = \log \hat{\sigma}^2 + \frac{(\log n)q}{n}$$

in which the penalty term depends on the sample size and it is generally larger than the penalty term appearing in the AIC. BIC provides a large sample estimator of a transformation of the Bayesian posterior probability associated with the approximation model. In general, by choosing the fitted candidate model corresponding to the BIC criterion, one is selecting the candidate model with the highest posterior probability. A good property of BIC selection is that it provides consistent model selection, see for example [69]. That is, when the true model is of finite dimension, BIC will choose the model with probability tending to 1 as the sample size n increases.

In general, a penalized function can only be consistent if its penalty term ($\log n$ in BIC) is a fast enough increasing function of n (see [70]). Thus AIC is not consistent as it always has some probability of selecting models that are too large. However, we note that in finite samples, adjusted versions of AIC can behave much better, see for example [71]. Further, since the penalty term of BIC is more stringent than the penalty term of AIC, BIC tends to form smaller models than AIC. However, BIC provides a large-sample estimator of the transformation of the Bayesian posterior probability associated with the approximating model, and AIC provides an asymptotically unbiased estimator of the expected Kullback discrepancy between the generating model and the fitted approximating model. In addition, AIC is asymptotically efficient in the sense that it asymptotically selects the fitted candidate model which minimizes the MSE of prediction, but BIC is not asymptotically efficient. This is because AIC can be advocated when the primary goal of the model is to induce meaningful factors influencing the outcome based on relative importance.

In summary, both AIC and BIC provide well-founded and self-contained approaches to model selection although with different motivations and penalty objectives. Both are typically good approximations of their own theoretical target quantities. Often, this also means that they will identify good models for observed data but both criteria can still fail in this respect. For a detailed simulation and empirical comparison of these two approaches, see [72], and for their properties see [69,73,74]. Both the AIC and the TIC are designed for the likelihood or quasi-likelihood context. They perform in a similar way. Their relationship is similar to the relationship between the conventional and the White covariance matrix estimators for the MSE/QMLE or LS. Unfortunately, despite the merit TIC has theoretically, it does not appear to be widely used perhaps because it needs a very large sample to get good estimates.

2.1.2. FIC

Let us start from the model

$$y_i = x_i'\beta + z_i'\gamma + u_i,\ i = 1, ..., n$$

or

$$y = X\beta + Z\gamma + u$$

where X is an $n \times p$ matrix of variables intended (focused) to be included all the time yet the variables in a $n \times q$ matrix Z may or may not be included. From the ML estimators $(\hat{\beta}_l, \hat{\gamma}_l)$, corresponding with the l-th model, the predictor for $m_l = x'\beta_l + z'\gamma_l$ can be written as $\hat{m}_l = x'\hat{\beta}_l + z'\hat{\gamma}_l$ at (x, z). In [10] provides MSE of $\hat{m}_l$. The basic idea of FIC is to develop a model selection criterion that chooses the model with the smalllest estimated MSE. Such an MSE-based FIC for the l-th submodel is

$$\widehat{FIC}_l = (\hat{\omega}'(I - \hat{\Psi}_l\hat{L}^{-1})\hat{\gamma})^2 + 2\hat{\omega}'\hat{\Psi}_l\hat{\omega}$$

where $\hat{\Psi}_l = \pi_l'(\pi_l\hat{L}^{-1}\pi_l')^{-1}\pi_l$, $\hat{L} = (Z'M_xZ)^{-1}$ where $M_x = I - X(X'X)^{-1}X'$, $\hat{\omega} = X(X'X)^{-1}x - z$, and π_l captures the projection mappings from the full model to the l-th submodel, such that $\omega_l = \pi_l\omega$.

In contrast, from [10],

$$AIC_l = -\hat{\gamma}'\hat{L}^{-1}\hat{\Psi}_l\hat{L}^{-1}\hat{\gamma} + 2\,|l|$$

where $|l|$ is the number of uncertain parameters in the l-th submodel, shows that when the estimand $m = \log f(y, \beta, \gamma)$ such that $f(y, \beta, \gamma)$ is the probability density function of the data, the MSE-based FIC is asymptotically equivalent to AIC.

2.1.3. Mallows Model Selection

Let us write the regression model (2) as

$$y = m + u$$

where $m = X\beta$. Then $\hat{m} = \hat{m}(q) = P(q)y$, where $P(q) = X(X'X)^{-1}X'$.

The objective is to choose q such that the average mean squared error (risk) $EL(q|X)$ is minimum, where

$$L(q) = \frac{1}{n}[m - \hat{m}(q)]'[m - \hat{m}(q)] = \frac{1}{n}(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta) = \frac{1}{n}u'P(q)u$$

such that

$$R(q) = E[L(q)|X] = \frac{1}{n}\sigma^2 tr(P(q)) = \frac{\sigma^2 q}{n}.$$

Mallows criterion for selecting q is to minimize

$$C(q) = \frac{\hat{u}'\hat{u}}{n} + \frac{2\sigma^2 q}{n}$$

where the seceond term on the right hand side is a penalty.

In fact, Mallows criterion is an unbiased estimator of the MSE of the predictive estimator $\hat{m}$ of m. This is because $E[L(q)|X] = E[(\hat{m}-m)'(\hat{m}-m)/n] = E[\frac{u'P(q)u}{n}] = \sigma^2 trP(q)/n$ and $E[C(q)|X] = \frac{\sigma^2(n-q)}{n} + \frac{2\sigma^2 q}{n} = \sigma^2 + \sigma^2 trP(q)/n$. But the minimization of $E[L(q)|X]$ with respect to q is the same as the minimization of $E[C(q)|X]$ since σ^2 does not depend on q.

Alternatively,

$$\begin{aligned}\frac{1}{n}(\hat{m}-m)'(\hat{m}-m) &= \frac{1}{n}(\hat{m}-y+y-m)'(\hat{m}-y+y-m)\\ &= \frac{1}{n}[\hat{u}'\hat{u}+u'u-2\hat{u}'u]\end{aligned}$$

and $E[\frac{1}{n}(\hat{m}-m)'(\hat{m}-m)] = \frac{1}{n}E[\hat{u}'\hat{u}+2\sigma^2 trP-\sigma^2]$. So, an unbiased estimator is $(\hat{u}'\hat{u}+2\sigma^2 q-\sigma^2)/n$ and its minimization is equivalent to the Mallows criterion.

2.1.4. Cross-Validation (CV)

CV is a commonly used procedure for model selection. According to this, the selection of q is made by minimizing

$$CV(q) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\hat{\beta}_{-i})^2$$

where $\hat{\beta}_{-i}$ is the LS estimator of β dropping the i-th observations y_i, x_i from the sample. It can be shown that $E[CV(q)] \simeq MSPE(q)$, where

$$MSPE(q) \simeq E(y_{n+1} - x_{n+1}'\hat{\beta})^2 = E\hat{u}_{n+1}^2$$

is the MSE of the forecast error $\hat{u}_{n+1} = y_{n+1} - \hat{y}_{n+1}$ with $\hat{y}_{n+1} = x_{n+1}\hat{\beta}$. Thus, CV is an almost unbiased estimator of $MSPE(q)$.

This can be shown by first writing the MSPE, based on an out of sample observation from the same distribution as the in sample observation, as

$$\begin{aligned}MSPE(q) &= E(y_{n+1} - x_{n+1}'\hat{\beta})^2 = E\hat{u}_{n+1}^2\\ &= Eu_{n+1}^2 + E[(\hat{\beta}-\beta)'x_{n+1}x_{n+1}'(\hat{\beta}-\beta)]\\ &= Eu_{n+1}^2 + MSE(q)\end{aligned}$$

where $MSE(q) = E[(\hat{m}(x_{n+1}) - m(x_{n+1}))'(\hat{m}(x_{n+1}) - m(x_{n+1}))] = E[(\hat{\beta}-\beta)'x_{n+1}x_{n+1}'(\hat{\beta}-\beta)]$. Since $Eu_{n+1}^2 = \sigma^2$ does not depend on q, its selection by $MSPE(q)$ and $MSE(q)$ are equivalent.

We observe that $\hat{u}_{n+1} = y_{n+1} - x_{n+1}\hat{\beta}$ is a prediction error based on first estimating $\hat{\beta}$ based on in sample n observations, and then calculating the error by using the out of sample observation $n+1$.

Therefore, $MSPE(q)$ is the expectation of a squared leave-one-out prediction error when the sample length is $n+1$. Using this idea we can also obtain a similar leave-one-out prediction error for each observation i. This is given by $\hat{u}_i = y_i - x_i'\hat{\beta}_{-i}$ based on n observations. Thus, $E\hat{u}_i^2 = MSPE(q)$ for each i, and

$$E[CV(q)] = E[\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2] = MSPE(q).$$

Further, since $E\hat{u}_{n+1}^2$ based on $n+1$ observations will be close to $E\hat{u}_i^2$ based on n observations, $CV(q)$ is an almost unbiased estimator of $MSPE(q)$.

The $CV(q)$ written above can be rewritten as

$$CV(q) = \frac{1}{n}\sum_{i=1}^{n}\frac{\tilde{u}_i^2}{1-h_{ii}}$$

where $\tilde{u}_i = y_i - x_i'\hat{\beta}$, h_{ii} is referred to as the leverage effect and it is the diagonal element of the projection matrix $X(X'X)^{-1}X'$, see [75]. This expression is useful for calculations. Also, see [74] for a link of $CV(q)$ with AIC.

2.1.5. Model Selection by Other Penalty Functions

The issue regarding the model selection has received more attention in recent years because of the challenging problem of estimating models with large numbers of regressors, which may increase with sample size, for example, earning models in labor economics with large number of regressors, financial portfolio models with large number of stocks, and VAR models with hundreds of macro variables.

A different method of variable selection and estimating such models is penalized least squares (PLS), see [14] for a review on this. In fact in this literature estimation of parameters and variables selections are done by using a criterion function involving loss function with a penalization function. Using l_p-penalized, the PLS estimator and variables selection problem are carried out as

$$\min_{\beta}[\sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \lambda(\sum_{j=1}^{q}|\beta_j|^p)^{1/p}]$$

where λ is a tuning or shrinkage parameter and the penalty is the restriction $(\sum_{j=1}^{q}|\beta_j|^p)^{1/p} \leq c$ (another tuning parameter). For $p=0$, the l_0-norm becomes $\sum_{j=1}^{q} I(\beta_j \neq 0)$ with $I(\cdot)$ as the usual indicator function which indicates the number of nonzero β_j for $j = 1, ..., q$. The AIC and BIC belong to this norm. For $p=1$, the l_p-norm becomes $\sum_{j=1}^{q}|\beta_j| \leq c$, which is used in the LASSO for simultaneous shrinkage estimation [76] and for variable selection. It can be shown analytically that the LASSO method estimates the zero coefficient as zero with positive probability as $n \to \infty$. Next, for $p=2$ the l_2-norm uses $\sum_{j=1}^{q}\beta_j^2 \leq c$ and provides ridge type [13] shrinkage estimation but not variable selection. However, if we consider the generalized ridge estimator under $\sum \hat{\lambda}_j\beta_j^2 \leq c$ then the coefficient estimates corresponding to $\hat{\lambda}_j \to \infty$ will tend to zero, see [77].

Further, when $0 < p \leq 1$ we get the bridge estimator [11,12] which provides a way to combine variable selection and parameter estimation together with $p=1$ as the LASSO. For adaptive LASSO and other forms of LASSO, see [62,78–80]. Also, see the link of LASSO with the least angel regression selection (LARS) by [81].

2.2. Model Averaging

Let us consider m be a parametric or nonparametric model, which can be a conditional mean or conditional variance. Let $\hat{m}_l$, $l = 1, ..., M$ be the set of estimators of m corresponding to the different sets of regressors considered in the problem of model selection. Consider w_l, $l = 1, ..., M$, to be the weights corresponding to $\hat{m}_l$, where $0 \leq w_l \leq 1$ and $\sum_{l=1}^{M} w_l = 1$. We can then define a model averaging estimator of m as

$$\hat{m}(w) = \sum_{l=1}^{M} w_l \hat{m}_l.$$

Below we present the choice of w_l in linear regression models. For the linear regression model consider the model in (1) or (2) where the dimension of β can tend to ∞, as $n \to \infty$. We take M models where l-th model contains q_l regressors, which is a subvector of x_i. The corresponding model could be written as

$$y = X_l \beta_l + u,$$

and the LS estimator of β_l is

$$\hat{\beta}_l = (X_l' X_l)^{-1} X_l' y.$$

This gives

$$\hat{m}_l = X_l \hat{\beta}_l = P_l y$$

where $P_l = X_l (X_l' X_l)^{-1} X_l'$. The model averaging estimator (MAE) of m is given as

$$\hat{m}(w) = \sum_{l=1}^{M} w_l \hat{m}_l = P(w) y$$

where $P(w) = \sum_{l=1}^{M} w_l P_l$. An alternative expression is

$$\hat{m}(w) = \sum_{l=1}^{M} w_l \hat{m}_l = \sum_{l=1}^{M} w_l X_l \hat{\beta}_l = X \hat{\beta}(w)$$

where we write $\tilde{\beta}_l = \begin{pmatrix} \hat{\beta}_l \\ 0 \end{pmatrix}$ such that $X_l \hat{\beta}_l = [X_l \; X_{-l}] \begin{pmatrix} \hat{\beta}_l \\ 0 \end{pmatrix} = X \begin{pmatrix} \hat{\beta}_l \\ 0 \end{pmatrix} = X \tilde{\beta}_l$ and $\hat{\beta}(w) = \sum_{l=1}^{M} w_l \tilde{\beta}_l = \begin{pmatrix} \sum_{l=1}^{M} w_l \hat{\beta}_l \\ 0 \end{pmatrix}$ is the MAE of β. Thus, for the linear model, the MAE of m corresponds to the MAE of β but this may not hold for the non-linear parameters model.

Now we consider the ways to determine weights.

2.2.1. Bayesian and FIC Weights

Under the Bayesian procedure we assume that there are M potential models and one of the models is the true model. Then, using the prior probabilities that each of the potential models is the true model, and considering the prior probability distributions of the parameters, the posterior probability distribution is obtained as the weighted average of the submodels where weights are the posterior probabilities that the given model is the true model given the data.

The two types of weights considered are then

$$w_l = \frac{\exp\{-\frac{1}{2}AIC_l\}}{\sum_{l=1}^{M}\exp\{-\frac{1}{2}AIC_l\}} \textit{ and } w_l = \frac{\exp\{-\frac{1}{2}BIC_l\}}{\sum_{l=1}^{M}\exp\{-\frac{1}{2}BIC_l\}}$$

where $AIC_l = -2\log L + 2q_l$ and $BIC_l = -2\log L + q_l \log n$. These are known as smoothed AIC (SAIC) and smoothed BIC (SBIC) weights. While the Bayesian model averaging estimator (BMAE) has a neat interpretation, it searches for the true model instead of selecting an estimator of a model with a low loss function. In simulations it has been found that SAIC and SBIC tend to outperform AIC and BIC estimators, see [82].

As for the FIC, consider the model averaging estimator as

$$\tilde{m} = \sum_{l=1}^{M} w_l \hat{m}_l$$

where

$$w_l = \exp(-\frac{1}{2}\frac{FIC_l}{\kappa\omega' L\omega})/\sum_{all\ l}\exp(\frac{1}{2}\frac{FIC_l}{\kappa\omega' L\omega})$$

and κ is an algorithmic parameter, bridging from uniform weighting (κ close to 0) to the hard-core FICC (κ is large). For this and further properties and applications of FIC, see [10] and [82].

2.2.2. Mallows Weight Selection Method

In the linear regression model, $\hat{m}(w) = P(w)y$ is a linear estimator with $w \in W_M$. So an optimal choice of w can be found following the Mallows criterion described above. The Mallows criterion for choosing weights w is

$$C(w) = \hat{u}(w)'\hat{u}(w) + 2\sigma^2 tr(P(w))$$

where $\hat{u}(w) = y - \hat{m}(w) = y - \sum_{l=1}^{M} w_l\hat{m}_l = \sum_{l=1}^{M} w_l(y - \hat{m}_l) = \sum_{l=1}^{M} w_l\hat{u}_l = \hat{U}w$ and

$$tr(P(w)) = \sum_{l=1}^{M} w_l trP_l = \sum_{l=1}^{M} w_l q_l = \mathbf{q}'w$$

in which $\mathbf{q} = (q_1, ..., q_M)'$, $w = (w_1, ..., w_M)'$, $\hat{u}_l$ is the residual vector from the l-th model and $\hat{U} = (\hat{u}_1, ..., \hat{u}_M)$ is an $n \times M$ matrix of residuals from all the models. Thus

$$C(w) = w'\hat{U}'\hat{U}w + 2\sigma^2\mathbf{q}'w$$

is quadratic in w. Thus

$$\hat{w} = arg \min_{w \in W_M} C(w),$$

which is obtained by using the quadratic programming procedure with inequality constraints using Gauss or MATLAB. Then Hansen's Mallows model averaging (MMA) estimator is

$$\hat{m}(\hat{w}) = \sum_{l=1}^{M} \hat{w}_l\hat{m}_l.$$

Following [83], [39] shows that

$$\frac{L(\hat{w})}{Inf_{w \in W_M^*} L(w)} \to 1$$

as $n \to \infty$, and $\hat{w}$ is asymptotically optimal in Li's sense, where $L(\hat{w}) = (m - \hat{m}(\hat{w}))'(m - \hat{m}(\hat{w}))$. However, Hansen's result requires weights belonging to a discrete set and the models to be nested. In [41] improves the result by relaxing discreteness and by not assuming that the models are nested. Their approach is based on deriving an unbiased estimator of the exact MSE of $\hat{m}(w)$.

Reference [84] also proposes a corresponding forecasting method, using Mallows model averaging (MMA). He proves that the criterion is an asymptotically unbiased estimator of both the in-sample and the out-of-sample one-step-ahead MSE.

2.2.3. Jackknife Model Averaging Method (CV)

Utilizing the leave-one-out cross validation (CV) procedure, which is also known as the Jackknife procedure, Jackknife model averaging (JMA) method of estimating $m(w)$ by [40] relaxes assumptions in [39]. The submodels are now allowed to be non-nested and also the error terms can be heteroskedastic. The sum-of-squared residuals in the JMA method is

$$CV(w) = \frac{1}{n}(y - \tilde{m}(w))'(y - \tilde{m}(w))$$

where $\tilde{m}(w)$ is the vector of the Jackknife estimator computed with the i-th element deleted. To be more specific, $\tilde{m}_l = X(X'_{l(-i)}X_{l(-i)})^{-1}X'_{l(-i)}y_{-i}$, where $X_{l(-i)}$ is equal to X_l with its i-th row deleted and y_{-i} is y with the i-th element deleted. Thus

$$\tilde{u}(w) = \sum_{l=1}^{M} w_l(y - \tilde{m}_l) = \sum_{l=1}^{M} w_l\tilde{u}_l = \tilde{U}w$$

where $\tilde{U} = (\tilde{u}_1, ..., \tilde{u}_M)$ is an $n \times M$ matrix, $\tilde{u}_l = (\tilde{u}_{1l,}..., \tilde{u}_{nl})'$ is an $n \times 1$ vector in which $\tilde{u}_{il}$ is computed with the i-th observation deleted. Then

$$CV(w) = \frac{1}{n}\tilde{u}(w)'\tilde{u}(w) = \frac{1}{n}w'\tilde{U}'\tilde{U}w$$

and JMA weights are obtained by minimizing $CV(w)$ with respect to $w = \tilde{w}_l$, and the JMA estimator is $\tilde{m}(w) = \sum_{l=1}^{M} w_l\tilde{m}_l$. Reference [40] shows the asymptotic optimality, using [83,85], in the sense of minimizing conditional risk which is equivalent to the out-of-sample prediction MSE.

There are many extensions of the JMA method to various other econometric models. Reference [86] does it for the quantile regression model. Reference [82] extends it for the dependent time series models or models with GARCH errors. Also, using MMA method in [39], for models with endogeneity, in [87] develops MMA based two-stage least squares (MATSLS), model averaging limited information maximum likelihood (MALIML), and model averaging Fuller (MAF) estimators.

However, it would be useful to have extensions of the MMA and JMA procedures to the models with GMM or IV estimator. In addition the sampling properties of the average estimators need to be developed for the purpose of statistical inference.

3. Nonparametric (NP) Model Selection and Model Averaging

3.1. NP Model Selection

Let us write the NP model as

$$y_i = m(x_i) + u_i$$

where x_i is i.i.d. with density f and the error u_i is independent of x_i.

We can write the local linear model as

$$\begin{aligned} y_i &= m(x) + (x_i - x)'\beta(x) + u_i \\ &= z_i(x)'\delta(x) + u_i \end{aligned}$$

or

$$y = Z(x)\delta(x) + u$$

where $z_i(x) = [1\ (x_i - x)']'$ so that $Z(x)$ is an $n \times (q+1)$ matrix and $\delta(x) = [m(x)\ \beta(x)]'$. Then the local linear LS estimator (LLLS) of $\delta(x)$ is

$$\hat{\delta}(x) = (Z'(x)K(x)Z(x))^{-1}Z'(x)K(x)y = P(x)y$$

where $P(x) = (Z'(x)K(x)Z(x))^{-1}Z'(x)K(x)$, $K(x) = diag(K((x_1 - x)/h), ..., K((x_n - x)/h))$ is a diagonal matrix in which the kernel $K((x_i - x)/h) = \prod_{j=1}^{q} K((x_{ij} - x_j)/h_j)$, and h_j is the window-width for the j-th variable. From this, pointwise $\hat{m}(x) = [1\ 0]\hat{\delta}(x)$, $\hat{\beta}(x) = [0\ 1]\hat{\delta}(x)$. Further, profiled $\hat{m} = (\hat{m}(x_1), ..., \hat{m}(x_n))'$ can be written as

$$\hat{m} = Py$$

where $P = P(h)$ is an $n \times n$ matrix generated by $[1\ 0]P(x_i) = [1\ 0](Z'(x_i)K(x_i)Z(x_i))^{-1}Z'(x_i)K(x_i)$, for $i = 1, ..., n$. If h is fixed then $\hat{m}$ is a linear estimator in y. But it will be a nonlinear estimator in y if $h = \hat{h}$ is either obtained by a plug-in estimator or by cross-validation.

With respect to the goodness of fit measures for the NP models we note that

$$V(y) = V(m(x)) + E[\sigma^2(x)]$$

So the global population goodness of fit is

$$\rho^2 = \frac{V(m(x))}{V(y)} = 1 - \frac{E[y - m(x)]^2}{V(y)},\ 0 \le \rho^2 \le 1$$

and its sample global estimator is given by

$$\begin{aligned} R^2 &= [1 - \frac{\sum \hat{u}_i^2}{\sum (y_i - \bar{y})^2}] = [1 - \frac{\hat{u}'\hat{u}}{y'M_2y}] \\ &= 1 - \frac{y'M_1(h)y}{y'M_2y} \\ &= \frac{y'M_1^*(h)y}{y'M_2y} \end{aligned}$$

where $\hat{u} = y - \hat{m} = y - P(h)y = M(h)y$ ($M(h) = I - P(h)$), $M_1(h) = M(h)'M(h)$, $M_1^*(h) = M_2 - M_1(h)$, and $M_2 = I - \frac{\iota\iota'}{n}$ with ι being an $n \times 1$ vector of unit elements. However, $0 \leq R^2 \leq 1$ may not be valid since $\sum(y_i - \bar{y})^2 \neq \sum(\hat{m}(x_i) - \bar{y})^2 + \sum \hat{u}_i^2$. Therefore, one can use the following modified $0 \leq R_1^2 \leq 1$ as

$$R_1^2 = R^2 I(a \leq 1)$$

where $a = \sum \hat{u}_i^2 / \sum(y_i - \bar{y})^2$ and $I(\cdot)$ is an indicator function.

Another way to define a proper global R^2 is to first consider a local $R^2(x)$. This is based on the fact that at the point x,

$$\sum(y_i - \bar{y})^2 K(\frac{x_i - x}{h}) = \sum(\hat{m}(x_i) - \bar{y})^2 K(\frac{x_i - x}{h}) + \sum \hat{u}_i^2 K(\frac{x_i - x}{h})$$

because $\sum u_i K(\frac{x_i - x}{h}) = 0$ and $\sum(x_i - x)u_i K(\frac{x_i - x}{h}) = 0$ due to local linear LS estimation. Thus a local $R^2(x)$ can be defined as

$$R^2(x) = \frac{\sum(\hat{m}(x_i) - \bar{y})^2 K(\frac{x_i - x}{h})}{\sum(y_i - \bar{y})^2 K(\frac{x_i - x}{h})} = \frac{SSR(x)}{SST(x)}$$

which satisfies $0 \leq R^2(x) \leq 1$. A global R_2^2 is then

$$R_2^2 = \frac{\int_x SSR(x)dx}{\int_x SST(x)dx}, \; 0 \leq R_2^2 \leq 1$$

The goodness of fit R_1^2 is considered in [88] where they showed its application for the statistically significant variables selection in NP regression. R_2^2 is introduced in [89,90]. For the variables selection it may be more appropriate to consider an adjusted R_1^2 as

$$R_{1a}^2 = R_a^2 I(b \leq 1)$$

where $R_a^2 = (1 - \frac{n-1}{trM_1(h)} \frac{y'M_1(h)y}{y'M_2y}) = 1 - b$. As a practical matter, the most critical choice in model selection in the nonparametric regression estimation above is the choice of the window-width h and the number of variables q. Further, if instead of considering the local linear estimator taken above and often used, we consider a local polynomial of degree d, then $Z(x)$ in $\hat{\delta}(x)$ would be a $n \times (qd + 1)$ matrix and we would need an additional selection for d. Thus the nonparametric goodness of fit measures described above should be considered as $R_1^2 = R_1^2(h, q, d)$ and $R_{1a}^2 = R_{1a}^2(h, q, d)$ and they can be used for choosing, say h, for fixed q and d, as the value which maximizes $R_{1a}^2(h, q, d)$. We note that $d = 0$ is the well known Nadaraya and Watson local constant estimator and for $d = 1$, it is the local linear estimator. Further, for given d and h, $R_1^2 = R_1^2(q)$ and $R_2^2 = R_2^2(q)$ can be used to choose q.

3.1.1. AIC, BIC, and GCV

In the NP case the model selection (choosing q) using AIC is proposed by [91]. This is based on the LCLS estimator,

$$AIC = \log \hat{\sigma}^2 + \frac{1 + trP(h)/n}{1 - (trP(h) + 2)/n}$$

where $\hat{\sigma}^2 = \hat{u}'\hat{u}/n = y'M_1(h)y/n$ in which $M_1(h) = M(h)'M(h)$ and $M(h) = I - P(h)$ where the (i, j)-th element of $P(h)$ is $P_{i,j}(h) = K_{ij} / \sum_{l=1}^n K_{il}$ and $K_{ij} = \prod_{s=1}^q h_s^{-1} K((x_{is} - x_{js})/h_s)$.

In the same way, we note that $AIC = AIC(h, q, d)$ and it can be used to select, for example, h given q and d ([92]) or q given h and d. In the latter case $AIC = AIC(q)$. The result for the $BIC = BIC(q)$ procedure in the NP model is not yet known. However, if one considers NP sieve regression of the type $m(x) = \sum_{j=1}^{q} z_j(x)\beta_j$ where $z_j(x)$ are nonlinear function of x and q, then BIC is similar to the BIC given in [96]. This includes, for example, special cases of a series expansion in which $z_j(x) = x^j$, and a spline regression in which $m(x) = \sum_{j=1}^{p} x^j\beta_j + \sum_{j=1}^{r} \beta_{p+j}(x - t_j)I(x \geq t_j)$ with $q = p + r$, t_j as j-th knot, and $I(x \geq t_j) = 1$ if $x \geq t_j$ and 0 otherwise.

In [9] an estimate of the minimizer of $EL(q)$, called the GCV, is proposed which does not require the knowledge of σ^2. This can be written as the minimization of

$$V(q) = \frac{n^{-1}\sum_{i=1}^{n}(y_i - \hat{m}(x_i))^2}{(1 - n^{-1}trP)^2}$$

with respect to q. It has been shown by [9] that $E[V(q)|x] - \sigma^2 \simeq E[L(q)|x]$ for large n, and the minimizer $\hat{q}$ of $EV(q)$ is asymptotically optimal in the sense that $EL(\hat{q})/\min_q EL(q) = 1$ as $n \to \infty$. That is, the MSE of $\hat{q}$ tends to be minimum as $n \to \infty$. We note that $L(q)$ in parametric and nonparametric cases are given in Sections *2.1.3* and *3.1.2*, respectively.

3.1.2. Mallows Model Selection

Let us write the regression model

$$y_i = m(x_i) + u_i$$

where $E[u_i|x_i] = 0$ and $E(u_i^2|x_i) = \sigma^2$. Then, for $m = (m(x_1), ..., m(x_n))'$, $y = (y_1, ..., y_n)'$ and $u = (u_1, ..., u_n)'$

$$y = m + u.$$

Let us consider the LLLS estimator of m, which is linear in y, as

$$\hat{m} = \hat{m}(q) = P(q)y$$

where $P = P(h) = P(q)$ as defined in section *3.1*. When $\hat{h} \to h$ for large n, $\hat{m}$ can become asymptotically linear.

Our objective is to choose q such that the average mean squared error (risk) $E[L(q)|x]$ is minimum where

$$L(q) = \frac{1}{n}(m - \hat{m}(q))'(m - \hat{m}(q)).$$

We note that for $\hat{u} = y - \hat{m}(q)$

$$\begin{aligned} L(q) &= \frac{1}{n}(m - \hat{m}(q)y)'(m - \hat{m}(q)y) \\ &= \frac{1}{n}[\hat{u}'\hat{u} + u'u - 2\hat{u}'u] \end{aligned}$$

and

$$R(q) = E(L(q)|x) = \frac{1}{n}E[\hat{u}'\hat{u} + 2\sigma^2 trP(q) - \sigma^2]$$

Further Mallows criterion for selecting q (number of variables in x_i) is by minimizing

$$C(q) = \frac{1}{n}(y - \hat{m}(q))'(y - \hat{m}(q)) + \frac{2\sigma^2}{n}trP(q)$$

where the second term on the right-hand side is the penalty. Essentially, the minimization of $C(q)$ is the same as the minimization of the unbiased estimator of $E[L(q)|x] = R$ since σ^2 does not depend on q, see Section *2.1.3* and [6,9].

3.1.3. Cross Validation (CV)

The CV method is one of the most widely used window-width selectors for NP kernel smoothing. We note that the cross-validation estimator of the integrated squared error weighted by the density $f(x)$,

$$ISE(q) = \int_x (\hat{m}(x) - m(x))^2 f(x)dx$$

is given by

$$CV(q) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{m}_{-i}(x_i))^2$$

where $\hat{m}_{-i}(x_i)$ is $\hat{m}(x_i)$ after deleting the i-th observations y_i, x_i from the sample. In fact,

$$CV(q) = \frac{1}{n}\sum_{i=1}^{n}(m(x_i) - \hat{m}_{-i}(x_i))^2 + \frac{2}{n}\sum_{i=1}^{n}(m(x_i) - \hat{m}_{-i}(x_i))u_i + \frac{1}{n}\sum_{i=1}^{n}u_i^2$$

where the first term on the right-hand side is a good approximation to $ISE(h)$, because the second term is generally negligibly small, and the third term converges to a constant $\sigma^2 = E[\sigma^2(x)]$ free from h. Therefore $CV(q) = ISE(q) + \sigma^2$ asymptotically.

Also, in the case where $m(x)$ is a sieve regression, [96] shows that CV is an unbiased estimator of the MSE of prediction error (MSEPE) of m, $MSEPE = E[y_{n+1} - \hat{m}_(x_{n+1})]^2$, see section 2.1.4. In addition, the minimization of MSEPE is equivalent to the minimization of MSE and integrated MSE (IMSE) of estimated m for conditional and unconditional x, respectively.

If, instead of the local linear of $m(x_i)$ we consider the local polynomial of order d, then $\hat{m}(x_i)$ is the LPLS estimator [2], and $CV(q) = CV(h, q, d)$ continues to hold. For $d = 0$ we have a local constant LS (LCLS) estimator developed by [98,99]. For $d = 1$ we have the LLLS estimator as considered above. In practice, the values of h and d can be determined by minimizing $CV(h, q, d)$ with respect to h and d for given q, which is developed by [100]. For a vector x_i, if the choice of $h_j = \hat{h}_j$ for any j tends to be infinity (very large) then the corresponding variable is an irrelevant variable. This can be observed from a simple example. Suppose the $\hat{m}(x)$ for two variables $x_{i1,}$ x_{i2}, considering the LCLS estimator is $\hat{m}(x_1, x_2) = \hat{m}(x) = \sum y_i K(\frac{x_{i1}-x_1}{h_1})K(\frac{x_{i2}-x_2}{h_2})/\sum K(\frac{x_{i1}-x_1}{h_1})K(\frac{x_{i2}-x_2}{h_2})$. Thus if $h_2 \to \infty$, then $K(\frac{x_{i2}-x_2}{h_2}) = K(0)$ is constant and $\hat{m}(x) = \hat{m}(x_1, x_2) = \sum y_i K(\frac{x_{i1}-x_1}{h_1})/\sum K(\frac{x_{i1}-x_1}{h_1})$. Thus a large estimated value of the window-width leads to the exclusion of variables, and hence variables selection.

In a seminal paper [83] shows that Mallows, GCV and CV procedures are asymptotically equivalent and all of them lead to optimal smoothing in the sense that

$$\frac{\int(\hat{m}(x, \hat{q}) - m(x))^2 dF(x)}{\inf_q \int(\hat{m}(x, q) - m(x))^2 dF(x)} \to^p 1$$

where $\hat{m}(x) = \hat{m}(x, \hat{q})$, given h and d, is an estimator of $m(x)$ with $\hat{q}$ obtained using one of the above procedures.

Also, [101] demonstrates that for the local constant estimator ($d = 0$ and given q), $CV = CV(h, q, 0)$ smoothing selectors of h are asymptotically equivalent to GCV selectors. In an important paper, in [92] shows the asymptotic normality of $\hat{m}(x) = \hat{m}(x, \hat{h})$, where $\hat{h}$ is obtained by the CV method and x_i is a vector of mixed continuous and discrete variables. Their extensive simulation results reveal (no theoretical proof) that AIC window-width selection criterion is asymptotically equivalent to the CV method, but for small samples AIC tends to perform better than the CV method. Further, with repect to the comparison of NP and parametric models, their results explain the observations of [102] which finds that NP estimators with smoothing parameters h chosen by CV can yield better prediction relative to commonly used parametric methods for the datasets of several countries. Reference [85] shows that CV is optimal under heteroskedasticity. For GMM model selection which involves selecting moments conditions, see [93]. Also, see [94] for using minimization of empirical likelihood/KLIC and comments by [95] claiming a fundamental flaw in the application of KLIC.

3.2. NP Model Averaging

Let us consider $\hat{m}_l$, $l = 1, ..., M$, to be the set of estimators of m corresponding to the different sets of regressors considered in the model selection. Then

$$\hat{m}(w) = \sum_{l=1}^{M} w_l \hat{m}_l = P(w)y$$

where $\hat{m}_l = P_l y$, $P(w) = \sum_{l=1}^{M} w_l P_l$ and P_l is the P matrix, as defined before, based here on the variables in the l-th model. Then the choice of w can be determined by applying Mallows criterion (see Section *2.2.2*) as

$$C(w) = w'\hat{U}'\hat{U}w + 2\sigma^2 \mathbf{q}^{*\prime} w$$

where $\mathbf{q}^* = (trP(q_1), ..., trP(q_M))$, and $\hat{U} = (\hat{u}_1, ..., \hat{u}_M)'$ is a matrix of NP residuals of all the models. Thus we get $\hat{m}(\hat{w}) = \sum_{l=1}^{M} \hat{w}_l \hat{m}_l$.

Similarly, as in section *2.2.3*, if we calculate $\tilde{m}_l$ by deleting one element of each variable, then w can be determined by minimizing

$$CV(w) = \frac{1}{n} w'\tilde{U}'\tilde{U}w$$

in which the NP residuals matrix $\tilde{U} = (\tilde{u}_1, ..., \tilde{u}_M)'$ with $\tilde{u}_l = (\tilde{u}_{1l}, ..., \tilde{u}_{nl})'$, and $\tilde{u}_{il}$ is computed with the i-th observation deleted.

For the fixed window-width the optimality result of $\hat{w}$ can be shown to follow from [83]. However, for $h = \hat{h}$ the validity of Li's result needs further investigation.

4. Conclusions

Nonparametric and parametric models are studied in econometrics and practice. In all applications, the important issue is to reduce model uncertainty by using model selection or model averaging. This paper selectively reviews frequentist results on model selection and model averaging in the regression context.

It is clear that most of the results presented are under the i.i.d. assumption. It is useful to relax this assumption to allow dependence or heterogeneity in the data, see [103] for model selection in dependent time series models using various CV procedures. A systematic study of the properties of estimators based on FMA is warranted. Further, results need to be developed for more complicated nonparametric models, e.g., panel data models and models where variables are endogenous, although for the parametric case see [104–108]. Also, the properties of NP model averaging estimators, when the window-width in kernel regression is estimated are to be developed; although readers can see [96] for NP results of the estimators based on the sieve method.

Acknowledgements

The authors are thankful to L. Su, A.Wan, X. Zhang, and G. Zou for some discussions and references on the subject matter of this paper. They are also grateful to the guest editor, Tomohiro Ando, and anonymous referees for their constructive suggestions and comments. First author is also thankful to the Academic Senate, UCR for its financial support.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Pagan, A.; Ullah, A. *Nonparametric Econometrics*; Cambridge University Press: Cambridge, UK, 1999.
2. Li, Q.; Racine, J.S. *Nonparametric Econometrics: Theory and Practice*; Princeton University Press: Princeton, NJ, USA, 2007.
3. Belloni, A.; Chernozhukov, V. L1-penalized quantile regression in high-dimensional sparse models. *Ann. Stat.* **2011**, *39*, 82–130.
4. Zhang, C.; Fan, J.; Yu, T. Multiple testing via FDRL for large-scale imaging data. *Ann. Stat.* **2011**, *39*, 613–642.
5. Akaike, H. Information Theory and An Extension of the Maximum Likelihood Principle. In *International Symposium on Information Theory*; Petrov, B.N., Csaki, F., Eds.; Springer-Verlag: New York, USA, 1973; pp. 267–281.
6. Mallows, C.L. Some comments on C_p. *Technometrics* **1973**, *15*, 661–675.
7. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464.
8. Stone, M. Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc.* **1974**, *36*, 111–147.
9. Craven, P.; Wahba, G. Smoothing noisy data with spline functions. *Numer. Math.* **1979**, *31*, 377–403.
10. Claeskens, G.; Hjort, N.L. The focused information criterion. *J. Am. Stat. Assoc.* **2003**, *98*, 900–945.

11. Frank, I.E.; Friedman, J.H. A statistical view of some chemomtrics regression tools. *Technometrics* **1993**, *35*, 109–135.
12. Fu, W.; Knight, K. Asymptotics for lasso-type estimators. *Ann. Stat.* **2000**, *28*, 1356–1378.
13. Hoerl, A.E.; Kennard, R.W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **1970**, *12*, 55–67.
14. Fan, J.; Lv, J. A selective overview of variable selection in high dimensional feature space. *Stat. Sin.* **2010**, *20*, 101–148.
15. Bühlmann, P.; Van de Geer, S. *Statistics for High-Dimensional Data: Methods, Theory and Applications*; Springer: New York, NY, USA, 2011.
16. Claeskens, G.; Hjort, N.L. *Model Selection and Model Averaging*; Cambridge University Press: Cambridge, UK, 2008.
17. Andrews, D.; Lu, B. Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models. *J. Econom.* **2001**, *101*, 123–164.
18. Hall, A.R.; Inoue, A.; Jana, K.; Shin, C. Information in generalized method of moments estimation and entropy-based moment selection. *J. Econom.* **2007**, *138*, 488–512.
19. Pötscher, B.M. Effects of model selection on inference. *Econom. Theory* **1991**, *7*, 163–185.
20. Kabaila, P. The Effect of Model Selection on Confidence Regions and Prediction Regions. *Econom. Theory* **1995**, *11*, 537–549.
21. Bühlmann, P. Efficient and adaptive post-model-selection estimators. *J. Stat. Plan. Inference* **1999**, *79*, 1–9.
22. Leeb, H.; Pötscher, B.M. The finite-sample distribution of post-model-selection estimators and uniform *versus* nonuniform approximations. *Econom. Theory* **2003**, *19*, 100–142.
23. Leeb, H.; Pötscher, B.M. Can one estimate the conditional distribution of post-model-selection estimators? *Ann. Stat.* **2006**, *34*, 2554–2591.
24. Breiman, L. Heuristics of instability and stabilization in model selection. *Ann. Stat.* **1996**, *24*, 2350–2383.
25. Jin, S.; Su, L.; Ullah, A. Robustify financial time series forecasting. *Econom. Rev.* **2013**; in press.
26. Geweke, J.F. *Contemporary Bayesian Econometrics and Statistics*; John Wiley and Sons Inc.: Hoboken, NJ, USA, 2005.
27. Geweke, J.F. Bayesian model comparison and validation. *Am. Econ. Rev. Pap. Proc.* **2007**, *97*, 60–64.
28. Draper, D. Assessment and propagation of model uncertainty. *J. R. Stat. Soc.* **1995**, *57*, 45–97.
29. Hoeting, J.A.; Madigan, D.; Raftery, A.E.; Volinsky, C.T. Bayesian model averaging: A tutorial (with discussion). *Stat. Sci.* **1999**, *14*, 382–417.
30. Clyde, M.; George, E.I. Model uncertainty. *Stat. Sci.* **2004**, *19*, 81–94.
31. Brock, W.A.; Durlauf, S.N.; West, K.D. Policy evaluation uncertain economic environment. *Brook. Pap. Econ. Act.* **2003**, *2003*, 235–301.
32. Sala-i-Martin, X.; Doppelhofer, G.; Miller, R.I. Determinants of long-term growth: A Bayesian Averaging of Classical Estimates (BACE) approach. *Am. Econ. Rev.* **2004**, *94*, 813–835.
33. Magnus, J.R.; Powell, O.; Prüfer, P. A comparison of two model averaging techniques with an application to growth empirics. *J. Econom.* **2010**, *154*, 139–153.

34. Buckland, S.T.; Burnham, K.P.; Augustin, N.H. Model selection: An integral part of inference. *Biometrics* **1997**, *53*, 603–618.
35. Yang, Y. Adaptive regression by mixing. *J. Am. Stat. Assoc.* **2001**, *96*, 574–586.
36. Burnham, K.P., Anderson, D.R. *Model Selection and Multimodel Inference: A Practical Information-Theoretical Approach*; Springer-Verlag: New York, NY, USA, 2002.
37. Leung, G.; Barron, A.R. Information theory and mixing least-squares regressions. *IEEE Trans. Inf. Theory* **2006**, *52*, 3396–3410.
38. Yuan, Z.; Yang, Y. Combining linear regression models: When and how? *J. Bus. Econ. Stat.* **2005**, *100*, 1202–1204.
39. Hansen, B.E. Notes and comments least squares model averaging. *Econometrica* **2007**, *75*, 1175–1189.
40. Hansen, B.E.; Racine, J. Jackknife model averaging. *J. Econom.* **2012**, *167*, 38–46.
41. Wan, A.T.K.; Zhang, X.; Zou, G. Least squares model averaging by mallows criterion. *J. Econom.* **2010**, *156*, 277–283.
42. Kapetanios, G.; Labhard, V.; Price, S. Forecasting using predictive likelihood model averaging. *Econ. Lett.* **2006**, *91*, 373–379.
43. Wan, A.T.K.; Zhang, X. On the use of model averaging in tourism research. *Ann. Tour. Res.* **2009**, *36*, 525–532.
44. Bates, J.M.; Granger, C.W. The combination of forecasts. *Oper. Res. Q.* **1969**, *20*, 451–468.
45. Olkin, I.; Speigelman, C.H. A semiparametric approach to density estimation. *J. Am. Stat. Assoc.* **1987**, *82*, 858–865.
46. Fan, Y.; Ullah, A. Asymptotic normality of a combined regression estimator. *J. Multivar. Anal.* **1999**, *71*, 191–240.
47. Wolpert, D.H. Stacked generalization. *Neural Netw.* **1992**, *5*, 241–259.
48. LeBlanc, M.; Tibshirani, R. Combining estimates in regression and classification. *J. Am. Stat. Assoc.* **1996**, *91*, 1641–1650.
49. Yang, Y. Mixing strategies for density estimation. *Ann. Stat.* **2000**, *28*, 75–87.
50. Catoni, O. *The Mixture Approach to Universal Model Selection*; Technical Report; Ecole Normale Superieure: Paris, France, **1997**.
51. Jordan, M.I.; Jacobs, R.A. Hiearchical mixtures of experts and the EM algorithm. *Neural Comput.* **1994**, *6*, 181–214.
52. Jiang, X.; Tanner, M.A. On the asymptotic normality of hierarchical mixtures-of-experts for generalized linear models. *IEEE Trans. Inf. Theory* **2000**, *46*, 1005–1013.
53. Vovk, V.G. Aggregateing Strategies. In Proceedings of the 3rd Annual Workshop on Computational Learning Theory, Rochester, NY, USA, 06–08 August 1990; Volume 56, pp. 371–383.
54. Vovk, V.G. A game of prediction with expert advice. *J. Comput. Syst. Sci.* **1998**, *56*, 153–173.
55. Merhav, N.; Feder, M. Universal prediction. *IEEE Trans. Inf. Theory* **1998**, *44*, 2124–2147.
56. Ullah, A. Nonparametric estimation of econometric functionals. *Can. J. Econ.* **1988**, *21*, 625–658.

57. Fan, J.; Gijbels, I. *Nonparametric Estimation of Econometric Functionals*; Champman and Hall: London, UK, 1996.
58. Eubank, R.L. *Nonparametric Regression and Spline Smoothing*; CRC Press: New York, NY, USA, 1999.
59. Geman, S.; Hwang, C. Diffusions for global optimization. *SIAM J. Control Optim.* **1982**, *24*, 1031–1043.
60. Newey, W.K. Convergence rates and asymptotic normality for series estimators. *J. Econom.* **1997**, *79*, 147–168.
61. Wang, H.; Zhang, X.; Zou, G. Frequentist model averaging estimation: A review. *J. Syst. Sci. Complex.* **2009**, *22*, 732–748.
62. Su, L.; Zhang, Y. Variable Selection in Nonparametric and Semiparametric Regression Models. In *Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*; Ullah, A., Racine, J., Su, L., Eds.; Oxford University Press: Oxford, UK, 2013; in press.
63. Srivastava, A.K.; Srivastava, V.K.; Ullah, A. The coefficient of determination and its adjusted version in linear regression models. *Econom. Rev.* **1995**, *14*, 229–240.
64. Rousson, V.; Gosoniu, N.F. An R-square coefficient based on final prediction error. *Stat. Methodol.* **2007**, *4*, 331–340.
65. Wang, Y. *On Efficiency Properties of An R-square Coefficient Based on Final Prediction Error*; Working Paper; School of International Trade and Economics, University of International Business and Economics: Beijing, China, 2013.
66. Takeuchi, K. Distribution of information statistics and criteria for adequacy of models. *Math. Sci.* **1976**, *153*, 12–18. In Japanese.
67. Maasoumi, E. A compendium to information theory in economics and econometrics. *Econom. Rev.* **1993**, *12*, 137–181.
68. Ullah, A. Entropy, divergence and distance measures with econometric applications. *J. Stat. Plan. Inference* **1996**, *49*, 137–162.
69. Nishi, R. Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Stat.* **1984**, *12*, 758–765.
70. Hannan, E.J.; Quinn, B.G. The determination of the order of an autoregression. *J. R. Stat. Soc.* **1979**, *41*, 190–195.
71. Hurvich, C.M.; Tsai, C.L. Regression and time series model selection in small samples. *Biometrika* **1989**, *76*, 297–307.
72. Kuha, J. AIC and BIC: Comparisons of assumptions and performance. *Sociol. Methods Res.* **2004**, *33*, 188–229.
73. Stone, M. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. R. Stat. Soc.* **1977**, *39*, 44–47.
74. Stone, M. 1979. Comments on model selection criteria of Akaike and Schwartz. *J. R. Stat. Soc.* **1979**, *41*, 276–278.
75. Maddala, G.S. *Introduction to Econometrics*; Macmillan: New York, NY, USA, 1988.
76. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat.* **1996**, *58*, 267–288.

77. Ullah, A.; Wan, A.T.K.; Wang, H.; Zhang, X.; Zou, G. *A Semiparametric Generalized Ridge Estimator and Link with Model Averaging*; Working Paper; Department of Economics, University of California: Riverside, CA, USA, 2013.
78. Zou, H. The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* **2006**, *101*, 1418–1429.
79. Zhang, C. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* **2010**, *38*, 894–942.
80. Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360.
81. Efron, B.; Hastie, T.; Johnstone, I.; Tibshirani, R. Least angle regression. *Ann. Stat.* **2004**, *32*, 407–499.
82. Zhang, X.; Wan, A.T.K.; Zhou, S.Z. Focused information criteria, model selection, and model averaging in a tobit model with a nonzero threshold. *J. Bus. Econ. Stat.* **2012**, *30*, 132–143.
83. Li, K.C. Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. *Ann. Stat.* **1987**, *15*, 958–975.
84. Hansen, B. Least-squares forecast averaging. *J. Econom.* **2008**, *146*, 342–350.
85. Andrews, D.W.K. Asymptotic optimality of generalized C_L, cross-validation, and generalized cross-validation in regression with heteroskedastic errors. *J. Econom.* **1991**, *47*, 359–377.
86. Lu, X.; Su, L. *Jackknife Model Averaging for Quantile Regressions*; Working Paper; School of Economics, Singapore Management University: Singapore, 2012.
87. Kuersteiner, G.; Okui, R. Constructing optimal instruments by first-stage prediction averaging. *Econometrica* **2010**, *78*, 697–718.
88. Yao, F.; Ullah, A. A nonparametric R^2 test for the presence of relevant variables. *J. Stat. Plan. Inference*, **2013**, *143*, 1527-1547.
89. Su, L.; Ullah, A. A nonparametric goodness-of-fit-based test for conditional heteroskedasticity. *Econom. Theory* **2013**, *29*, 187–212.
90. Huang, L.H.; Chen, J. Analysis of variance, coefficient of determination and f-test for local polynomial regression. *Ann. Stat.* **2008**, *36*, 2085–2109.
91. Hurvich, C.; Simonoff, J.; Tsai, C. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J. R. Stat. Soc.* **1998**, *60*, 271–293.
92. Racine, J.; Li, Q. Nonparametric estimation of regression functions with both categorical and continuous data. *J. Econom.* **2004**, *119*, 99–130.
93. Andrews, D.W.K. Consistent moment selection procedures for generalized method of moments estimation. *Econometrica* **1999**, *67*, 543–564.
94. Chen, X.; Hong, H.; Shum, M. Nonparametric likelihood ratio model selection tests between parametric likelihood and moment condition models. *J. Econom.* **2007**, *141*, 109–140.
95. Schennach, S.M. Instrumental variable estimation of nonlinear errors-in-variables models. *Econometrica* **2007**, *75*, 201–239.
96. Hansen, B. *Nonparametric Sieve Regression: Least Squares Averaging Least Squares, and Cross-validation*; Working Paper; University of Wisconsin: Madison, WI, USA, 2012.
97. Liang, H.; Zou, G.; Wan, A.T.K.; Zhang, X. Optimal weight choice for frequentist model average estimators. *J. Am. Stat. Assoc.* **2011**, *106*, 1053–1066.

98. Nadaraya, E.A. Some new estimates for distribution functions. *Theory Probab. Its Appl.* **1964**, *9*, 497–500.
99. Watson, G.S. Smooth regression analysis. *Sankhya Ser. A* **1964**, *26*, 359–372.
100. Hall, P.G.; Racine, J.S. *Infinite Order Cross-validated Local Polynomial Regression*; Working Paper; Department of Economic, McMaster University: Ontario, Canada, 2013.
101. Härdle, W.; Hall, P.; Marron, J.S. How far are automatically chosen regression smoothing parameters from their optimum? *J. Am. Stat. Assoc.* **1988**, *83*, 86–99.
102. Li, Q.; Racine, J. *Empirical Applications of Smoothing Categorical Variables*; Working Paper; Department of Economic, McMaster University: Ontario, Canada, 2001.
103. Racine, J. Consistent cross-validatory model-selection for dependent data: Hv-block cross-validation. *J. Econom.* **2000**, *99*, 39–61.
104. Caner, M. A lasso type GMM estimator. *Econom. Theory* **2009**, *25*, 270–290.
105. Caner, M.; Fan, M. *A Near Minimax Risk Bound: Adaptive Lasso with Heteroskedastic Data in Instrumental Variable Selection*; Working Paper; North Carolina State University: Raleigh, USA, 2011.
106. Garcia, P.E. *Instrumental Variable Estimation and Selection with Many Weak and Irrelevant Instruments*; Working Paper; University of Wisconsin: Madison, WI, USA, 2011.
107. Liao, Z. Adaptive GMM shrinkage estimation with consistent moment selection. *Econom. Theory* **2013**, *FirstView*, 1–48.
108. Gautier, E.; Tsybakov, A. *High-Dimensional Instrumental Variables Regression and Confidence Sets*; Working Paper; Centre de Recherche en Economie et Statistique: Malakoff Cedex, France, 2011.