UNIVERSITY OF CALIFORNIA
RIVERSIDE

Essays on Semiparametric Ridge-Type Shrinkage Estimation, Model Averaging and Nonparametric Panel Data Model Estimation

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Economics

by

Huansha Wang

June 2014

Dissertation Committee:

Professor Aman Ullah, Chairperson
Professor Tae-Hwy Lee
Professor Gloria Gonzalez-Rivera

The Dissertation of Huansha Wang is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

The Author is very thankful to her great dissertation committee members, Professors Aman Ullah, Tae-Hwy Lee and Gloria Gonzalez-Rivera, for their guidance and help along the way. Especially, the author is grateful for the consistent care and excellent support that she gets from her advisor and mentor, Professor Ullah, in her journey of life as a Ph.D. student. The author is also very grateful for the education, help and care that she has received from the members of the department of Economics in UC Riverside. The faculty, staff and friends there have provided her a loving environment where she lives and learns. Especially, Professor Jang-Ting Guo has provided her lots of support whenever it is needed. Moreover, the author appreciates all the love and care that she has received from her family members and dear friends over the years.

To my family.

ABSTRACT OF THE DISSERTATION

Essays on Semiparametric Ridge-Type Shrinkage Estimation, Model Averaging and
Nonparametric Panel Data Model Estimation

by

Huansha Wang

Doctor of Philosophy, Graduate Program in Economics
University of California, Riverside, June 2014
Professor Aman Ullah, Chairperson

This dissertation is composed with 4 essays. They explore modelling uncertainty following two major directions. The former 2 contains topics on ordinary and general ridge-type shrinkage estimation developed from model averaging and kernel density estimation. The third one critically reviews recent literature in the areas of model averaging and model selection both parametrically and nonparametrically and proposes topics for future work. The last one focuses on nonparametric panel data estimation with random effects. In chapter 2, ordinary ridge-type shrinkage estimation is extensively studied, where a class of well-behaved ordinary ridge-type semiparametric estimators is proposed. Monte Carlo simulations, theoretical derivations, as well as empirical out-of-sample forecasts are all investigated to prove their usefulness in reducing mean squared errors, i.e. risks. Chapter 3 develops the works in Chapter 2 to the general ridge regressions. By connecting general ridge regression with kernel density estimation, an asymptotically optimal semiparametric ridge-type estimator is built. By connecting general ridge regression with model averaging, a class of model averaging ridge-type estimators are obtained. These estimators are observed to have different improvements upon the feasible general ridge estimators when model uncertainties, i.e., the error variances are different. To encourage better understanding on model averaging and model selection, Chapter 4 gives a comprehensive literature review and analysis on these topics from a frequentist's point of view. Parametric and nonparametric procedures in the recent developments are explored. Chapter 5 starts investigating panel data estimation by introducing nonparametrics in the picture. The proposed two-stage estimator shows good behaviors in Monte Carlo simulation. In addition, illustrative empirical examples in health economics and environmental economics are also introduced.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The study of data analysis is mainly composed of two major parts: selecting a model and making inferences from this selected. In the first stage, various models are considered to provide a closer insight to the real world despite all kinds of disturbances occur along the way. We could either impletement the nonparametric estimation and forecasting in the picture, where no specific functional format between the regressor and the regressand is assumed, thus leaving the model with a closer approximation to the real world. This leads to the popularity in the use of nonparametric models in recent years. Various topics have been explored by econometricians to bring more "efficiency", i.e. smaller mean squared errors (MSE) in modelling. Especially, with large datasets, such as the panel datasets, nonparaemetric estimation procedures are widely adopted.

On the other hand, for a series of models with specific/unspecific functional assumptions, a search of model can be carried out. Bearing this aim, model selection (choosing one model and define that as the "best") and model averaging (averaging across several models, i.e. submodels, to get an aggregated final model with weighs assigned to each candidate submodel) are brought up on the table to provide us a chance to look through the noises and see the signal. It is true that "when the noise level is very low relative to the signal, there is little difficulty identifying the best model and, accordingly model selection does a very good job"(Yang, 2001), yet as noises level becomes higher or when in fact it's hard to tell the noise from the signal, model averaging is considered a better choice.

Bearing this in mind, this dissertation covers several interesting essays on (i) semiparametric ordinary and general Ridge-type shirinkage estimation, along with their empirical applications in forecasting excess stock returns, etc.; (ii) a review of model averaging, model selection in pamametric and nonparametric frameworks, mainly from the view of a frequentist, with proposals of future developments and extensions in the line; (iii) nonparametric two-stage panel data model estimation with random effects where less MSE is achieved. On the one hand, these works contribute to the asymptotic theories and methodologies related to the topics and brings inspirations for future directions. On the other hand, the applications in this dissertaion covers areas such as financial economics, health economics, and enviromental economics, where the author proposes procedures to utilize the theories developed and brings interesting real world insights from new angles.

In the real world, what we face are lots of noises and the difficulty of telling them from the truth. Thus, with these uncertainties, most economists agree that model averaging could be more efficient and effective at explaining and forecasting in empirical works compared to model selection and other frameworks. Focusing on investigating the gains from model averaging, in Chapter 2, a class of easy-to-implement semiparametric ordinary ridge estimators of regression coefficients is introduced. Their properties are investigated and simulation results are provided to investigate their behaviors when the error variances are small and relatively large, respectively. The semiparametric estimators outperform the Hoerl, Kennard and Baldwin (1975) estimator in the sense that they give less risk (total mean squared error). In addition to this, a new method that minimizes the unbiased estimator of the MSE of the the regressand is also introduced, along with its theoretical derivations. Both Monte Carlo simulation and empirical application are also presented to demenstrate the usefulness of this method. Not surprisingly yet interestingly, this new criterion proposed works in a similar way compared with the Mallows criterion (Mallows, (1973)).

To extend the above work, Chapter 3 introduces general ridge regression with Mallows model averaging (GRRA) and Jackknife model averaging (GRRJ). Moreover, an asymptotic optimal semiparametric ridge estimation (AOSP) is also proposed. More specifically, we propose a new semi-parametric estimator of regression coefficients, which is in the form of a feasible generalized ridge estimator by Hoerl and Kennard (1970b) but with different biasing factors. We prove that the generalized ridge estimator is algebraically identical to the model average estimator. Further, the biasing factors that determine the properties of both the generalized ridge and semi-parametric estimators are directly linked to the weights used in model averaging. These are interesting results for the interpretations and applications of both semi-parametric and ridge estimators. Furthermore, we demonstrate that these estimators based on model averaging weights can have properties superior to the well-known feasible generalized ridge estimator in a large region of the parameter space. The GRRA/GRRJ and the AOSP estimators outperform each other given different model uncertainties that are controlled by the error variances. Monte Carlo simulations and two empirical examples are presented to demonstrate the usefulness of the new methods and which to choose given different model uncertainties.

In view of all the existing literature in the areas of model averaging and model selection, Chapter 4 presents a guidance for economists who are interested in the fundamental theories and most recent developments in this realm, mainly from a frequentist's point of view. While there is an extensive literature on the parametric case we provide the recently developed results in the context of nonparametric models. However, in applications, the 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. An alternative to model selection is the model averaging where the estimated model is the weighted sum of all the sub models. This reduces the model uncertainty. In recent years, there has been a significant interest in the model averaging and some important developments have taken place in this area. We present the results for both the parametric and nonparametric cases. Some possible future topics of research are also indicated.

Compared with model averaging, nonparametric estimation itself relaxes modelling assumptions from the setup of the model and thus brings more freedom to economists. Especially, the advancement of technology has brought more possible means to obtain and store data; thus, the use of big data, such as panel datasets, has become one of the top trends in econometrics. Chapter 5 introduces a new two step estimation procedure in the nonparametric regression function where the errors follow a general error variance-covariance matrix. By incorporating the information from the error variance-covariance matrix, efficient estimators of the nonparametric regression function (conditional mean) and its derivative are developed. The results for a special case of the nonparametric panel data regression with random effects are then presented. For this case Monte Carlo simulation results are performed to examine the finite sample performance of our proposed estimators. The results show that our proposed estimators outperform the local linear least squares estimator and many other existing estimators in the sense of smaller mean squared errors (risk). The asymptotic properties of the proposed estimator are established. Two real data applications on the relationship between health expenditure and education, as well as the environmental Kuznets curve, are also performed to illustrate the usefulness of this procedure.

Chapter 6 provides the concluding remarks for this thesis. And the appendix demonstrates the mathematical derivations in more detail.

Chapter 2

A Class of Semiparametric Ordinary Ridge Estimators of Regression Coefficients[*]

2.1 Introduction

In the ordinary least squares (OLS) estimation, if the prediction vectors from X, where $X = (x_1, x_2, ...)$ are not orthogonal, there is a high probability that the OLS estimators may be unsatisfactory. In particular, the estimated coefficients tend to be abnormally large in absolute value and sometimes have the wrong sign. To fix this problem, the ridge estimation was introduced by Hoerl (1962) and Hoerl and Kennard (1970). By

[*] The text of this chapter, in part or in full, is a reprint of the material as is appears in H, Wang. (2013) *Journal of Quantitative Economics* 11(12): 15-27.

adding an extra increment to the original $X'X$, the ridge estimation circumvents the non-orthogonality problem. Compared to the methods of principal components, computation of 2^p regressions, some subset of all regressions using fractional factorials, a branch and bound technique, ridge estimation "gives an insight into the structure of the factor space and the sensitivity of the results to the particular set of data at hand" (Hoerl and Kennard (1970)). In addition, most multiple regression models suffer multicollinearity to some degree. Thus, under these circumstances, introducing the ridge regression could gain, in the sense that the ridge estimators will outperform the OLS estimator and alleviate the multicollinearity problem.

As one can expect, the choice of the increment to $X'X$ is significant in the implementation of the ridge estimation. In the ordinary ridge estimation, we usually use kI, where the I being the identity matrix, to denote it. Many different choices of the biasing parameter k have been proposed in the literature, such as Hoerl, Kennard and Baldwin (1975), see Vinod and Ullah (1981).

In this chapter, a class of semiparametric ordinary ridge estimators is proposed. Starting from kernel density estimator of the regressors, these estimators bear more information than the OLS estimator and both simulation and empirical application results in the following content show the usefulness of these easy-to-implement estimators. The properties of these estimators are also investigated. The rest of this paper is arranged as follows. Section 2 introduces the class of ordinary semiparametric (OSP) estimators. Section 3 develops the approximate and exact unbiased mean squared errors (MSE) of the OSP estimators, and proposed the choices of window-width by minimizing them. Section

4 provides some simulation results comparing this class of OSP estimators with the OLS estimator and the estimator proposed by Hoerl, Kennard and Baldwin (1975). Section 5 gives the results for one empirical application. The last section concludes.

2.2 Semiparametric Estimator of Regression Coefficients

Consider a population multiple regression model

$$\begin{aligned} y &= x_1\beta_1 + \cdots + x_q\beta_q + u \qquad (2.1) \\ &= x'\beta + u \end{aligned}$$

where y is a scalar dependent variable, $x = [x_1, ..., x_q]'$ is a vector of q regressors, β is an unknown vector of regression coefficients, and u is a disturbance with $Eu = 0$ and $V(u) = \sigma^2$.

If we minimize $Eu^2 = E(y - x'\beta)^2$ with respect to β, we obtain

$$\beta = [Exx']^{-1}Exy \qquad (2.2)$$

where Exx' is a $q \times q$ moment matrix of q variables with the j-th diagonal element and j, j'-th off diagonal elements, respectively, given by

$$\begin{aligned} Ex_j^2 &= \int_{x_j} x_j^2 f(x_j)dx_j, \ j = 1, ..., q, \qquad (2.3) \\ Ex_jx_{j'} &= \int_{x_j}\int_{x_{j'}} x_jx_{j'}f(x_j, x_{j'})dx_jdx_{j'}, \ j \neq j' = 1, ..., q. \end{aligned}$$

Suppose we have the sample observations $\{y_i, x_{i1}, ..., x_{iq}\}$, $i = 1, ..., n$. Then the population averages in (2.3) can be estimated by their sample averages as

$$\hat{E}x_j^2 = \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2, \ \hat{E}x_jx_{j'} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ij'}. \qquad (2.4)$$

The result $\hat{E}x_jy = \sum_{i=1}^{n} x_{ij}y_i/n$ follows similarly.

Therefore we get[†]

$$\begin{aligned}\hat{\beta} &= (\hat{E}xx')^{-1}\hat{E}xy \qquad (2.5)\\ &= (X'X)^{-1}X'Y\end{aligned}$$

where X is an $n \times q$ matrix of observations on q variables, Y is an $n \times 1$ vector of n observations and $\hat{\beta}$ is the well known ordinary least squares (OLS) estimator.

Now we consider the estimation of Ex_j^2 and $Ex_jx_{j'}$ by using a smooth nonparametric kernel density estimation instead of empirical distribution function. In this case,

$$\begin{aligned}\tilde{E}x_j^2 &= \int_{x_j} x_j^2\tilde{f}(x_j)dx_j \qquad (2.6)\\ &= \frac{1}{nh}\sum_{i=1}^{n}\int_{x_j} x_j^2 k(\frac{x_{ij}-x_j}{h})dx_j\\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{\Psi_{ij}}(x_{ij}^2 + h^2\Psi_{ij}^2 - 2x_{ij}h\Psi_{ij})k(\Psi_{ij})d\Psi_{ij}\\ &= \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 + h^2\mu_2\end{aligned}$$

where $\tilde{f}(x_j) = \frac{1}{nh}\sum_{i=1}^{n} k(\frac{x_{ij}-x_j}{h})$ is a kernel density estimator, $\Psi_{ij} = \frac{x_{ij}-x_j}{h}$ is a transformed variable, $\mu_2 = \int v^2k(v)dv > 0$ is the second moment of kernel function, $k(\Psi_{ij})$ is a symmetric second order kernel, and h is window-width. For implementation, kernel is chosen as normal or Epanechnikov quadratic function, see Pagan and Ullah (1999).

[†] With the intercept, the expression for the $\hat{\beta}_{q-1}$ for the $(q-1)$ estimators could be defined as $(X'_{q-1}X_{q-1} + nh^2\mu_2 I - \bar{X}_{q-1}\bar{X}'_{q-1})^{-1}(X'_{q-1}Y - \bar{X}_{q-1}\bar{Y}')$ where $\bar{X}_{q-1} = (\bar{X}_2, ..., \bar{X}_q)$; and $\hat{\beta}_1 = \bar{Y} - \bar{X}_{q-1}\hat{\beta}_{q-1}$.

Similarly, it can easily be shown that

$$\begin{aligned}\tilde{E}(x_j x_{j'}) &= \int_{x_j}\int_{x_{j'}} x_j x_{j'} \tilde{f}(x_j, x_{j'}) dx_j dx_{j'} \qquad (2.7)\\ &= \frac{1}{nh^2}\sum_{i=1}^{n}\int_{x_j}\int_{x_{j'}} x_j x_{j'} k(\frac{x_{ij}-x_j}{h})k(\frac{x_{ij'}-x_{j'}}{h}) dx_j dx_{j'}\\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{\Psi_{ij}}\int_{\Psi_{ij'}} (x_{ij}-h\Psi_{ij})(x_{ij'}-h\Psi_{ij'})k(\Psi_{ij})k(\Psi_{ij'})d\Psi_{ij}d\Psi_{ij'}\\ &= \frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ij'},\end{aligned}$$

and

$$\tilde{E}(x_j y) = \frac{1}{n}\sum_{i=1}^{n} x_{ij} y_i \qquad (2.8)$$

where we have used product kernels without any loss of generality and $\Psi_{ij'} = \frac{x_{ij'}-x_{j'}}{h}$.Also, $\tilde{E}(x_j) = \frac{1}{n}\sum_{i=1}^{n} x_{ij} = \bar{x_j}$.

Thus, using (2.6) to (2.8) in (2.2), a new semiparametric estimator of β can be introduced as

$$\begin{aligned}\tilde{\beta}(h) &= (\tilde{E}xx')^{-1}\tilde{E}xy \qquad (2.9)\\ &= (X'X + nh^2\mu_2 I)^{-1}X'Y\\ &= (X'X + D)^{-1}X'Y\end{aligned}$$

where $D = nh^2\mu_2 I$ is a diagonal matrix. We refer this estimator as the ordinary semiparametric (OSP) estimator.

We note that both OLS and OSP estimators are obtained by first considering the population regression (2.1), in which the regression coefficient vector depends on the population moments of vector x and scalar variable y, and then estimating these moments by two different methods with the help of sample data. These are the estimators of the regression coefficients in the sample linear regression model

$$Y = X\beta + U \tag{2.10}$$

where the sample is drawn from the population linear regression model (2.1), and U is an $n \times 1$ vector of random errors with $EU = 0$ and $EUU' = \sigma^2 I_n$.

The class of ordinary ridge estimator due to Hoerl and Kennard (1970a) is defined as

$$\hat{\beta}(k) = (X'X + kI)^{-1}X'Y \tag{2.11}$$

where k is an unknown parameter. An operational ordinary ridge estimator, from Hoerl, Kennard and Baldwin (1975) is defined with $k = qs^2/\hat{\beta}'\hat{\beta}$ and $s^2 = \frac{(Y-X\hat{\beta})'(Y-X\hat{\beta})}{n-q}$. We refer this estimator as the HKB estimator in the following content.

2.3 Unbiased Estimation of MSE and Optimal Window-width Choice

In this section, the choices of window-width h are considered. These are based on the minimization of the approximate total MSE, risk, and the unbiased estimator of the exact MSE of $\tilde{\beta}(h) = E[(\tilde{\beta}(h) - \beta)'(\tilde{\beta}(h) - \beta)]$. Also we determine the choice of h based on the minimization of the total MSE of the predictor of y, which is $MSE(\tilde{\mu}(h)) = E[(\tilde{\beta}(h) - \beta)'X'X(\tilde{\beta}(h) - \beta)]$, where $\tilde{\mu}(h) = X\tilde{\beta}(h)$.

2.3.1 An Approximate Estimator of the MSE of $\tilde{\beta}(h)$

Theorem 1. *Under the conditions A1-A6 in the appendix, with $A \equiv n\mu_2(X'X)^{-1}$, approximate MSE (AMSE) of $\tilde{\beta}(h)$* is

$$AMSE(\tilde{\beta}(h)) = \sigma^2 tr(X'X)^{-1} - 2h^2\sigma^2\frac{trA^2}{n\mu_2} + h^4[2\beta' A^2\beta + 3\sigma^2\frac{trA^3}{n\mu_2}].$$

Proof: See the Appendix.

Remark 1: By minimizing the AMSE with respect to h^2, the first order condition gives the optimal choice of h^2 as

$$h^2 = \frac{\sigma^2 trA^2}{2n\mu_2\beta' A^2\beta + 3\sigma^2 trA^3}. \tag{2.12}$$

From the AMSE, we could observe that $AMSE - MSE(\hat{\beta}) = -2h^2\sigma^2\frac{trA^2}{n\mu_2} + h^4(2\beta' A^2\beta + 3\sigma^2 trA^3)$, thus, as long as $0 \leq h^2 \leq \frac{2\sigma^2 trA^2}{2n\mu_2\beta' A^2\beta + 3\sigma^2 trA^3}$, $\breve{\beta}(h)$ will outperform the OLS estimator in the sense that it generates smaller mean squared error. We refer this estimator as the AOSP estimator.

Also, we consider an approximation of h^2 in (2.12) as

$$h_1^2 = \frac{\sigma^2 trA^2}{2n\mu_2\beta' A^2\beta}, \tag{2.13}$$

and thus, this $\breve{\beta}(h_1)$ is referred as AOSP1.

2.3.2 An Exact Unbiased Estimator of the MSE of $\tilde{\beta}(h)$

Theorem 2. *Under the same conditions as in Theorem 1, the exact unbiased estimator of MSE of $\tilde{\beta}(h)$ is given by*

$$\widehat{MSE}(\tilde{\beta}(h) = s^2 tr(XD^2X') + (nh^2\mu_2)^2[\hat{\beta}' D^2\hat{\beta} - s^2 trD^2(X'X)^{-1}]$$

where $D = (X'X + n\mu_2 h^2 I)^{-1}$.

Proof: See the Appendix.

2.3.3 An Exact Unbiased Estimator of the MSE of $\tilde{\mu}(h)$

Theorem 3. *Under the same conditions as in Theorem 1, the exact unbiased estimator of MSE of* $\tilde{\mu}(h)$ *is given by*

$$\widehat{MSE}(\tilde{\mu}(h) = s^2 tr(XD'X')^2 + (nh^2\mu_2)^2[\hat{\beta}'D'X'XD\hat{\beta} - s^2 trD'X'XD(X'X)^{-1}]$$

where $D = (X'X + n\mu_2 h^2 I)^{-1}$.

Proof: See the Appendix.

Remark 2: The expression for the exact unbiased estimator of the MSE of $\tilde{\beta}(h)$ and $\tilde{\mu}(h)$ are nonlinear; thus in implementation, there's no closed form solution for the optimal window-width h; we will use the constraint optimization function built in *R version 2.13.1.* to approximate the optimal h. We refer these two estimators of $\tilde{\beta}(h)$ from the two $h's$ as the exact ordinary semiparametric (EOSP) estimator and an alternative exact ordinary semiparametric (EOSP1) estimators, respectively.

Another asymptotic optimal semiparametric under Mallows criterion is also considered in the simulation (see Hansen (2007) for reference on the Mallows criterion), where the optimal h^2 is obtained through minimize $(Y - X\tilde{\beta}(h))'(Y - X\tilde{\beta}(h)) + 2s^2 tr[(X'X + nh^2\mu_2 I)^{-1}X']$. We refer this estimator as the Mallows ordinary semiparametric (MOSP) estimator.

2.4 Simulation

Our Monte Carlo experiments are based on two DGP's.

DGP1: $y_i = \sum_{j=1}^{q} \theta_j x_{ij} + e_i$, x_{ij} are *iid* $N(0,1)$. The errors e_i are uncorrelated with $x's$ so it is set to be *iid* $N(0,1)$ and $N(0,25)$ respectively. This model is from Hansen (2007) with the values of θ_j considered here as $0.7071 j^{-3/2}$. Further, sample sizes taken are $n = 50$, $q = 11$ and $n = 150$, $q = 16$.

DGP2: In DGP 2, parameters and function are set as in DGP 1, but x_{2i} is set to be the sum of x_{3i} to x_{50i} plus an error which follows $N(0,1)$ so that this DGP incorporates near-perfect collinearity.

Under both DGP 1 and DGP 2, 1000 simulations are done. Further, normal kernel is selected, $K(\phi) = (2\pi)^{-1/2} \exp[-\frac{1}{2}\phi^2]$, and thus $\mu_2 = 1$.Performance of the estimators is evaluated in terms of total MSE (risk), $E(\hat{\beta}-\beta)'(\hat{\beta}-\beta)$, where $\hat{\beta}$ is an estimator. Simulation results are reported in Table 2.1.

From Table 2.1, we could observe that under different DGP settings, all classes of ordinary ridge estimators beat the OLS estimator. Compare column 4 with column 6 for example, as error variance increases, more gains are obtained through the usage of the semiparametric ordinary ridge estimators, especially MOSP and EOSP/EOSP1 estimators since much smaller risks are obtained. And another interesting, yet not surprising phenomenon we observe, is that the MOSP and EOSP1 generate very similar risks under different DGP's. This is due to the fact that both criteria are unbiased estimators of the MSE of $\tilde{\mu}(h)$.

2.5 An Empirical Application

2.5.1 Forecasting Excess Stock Returns

The data is the same as in Campbell and Thompson (2008). In the monthly data from January 1950 to December 2005 the total sample size is equal to 672. The dependent variable Y is the excess stock returns, which is defined as the difference between the monthly stock returns and the risk-free rate. We consider 12 regressors, default yield spread, treasury bill rate, new equity expansion, term spread, dividend price ratio, earnings price ratio, long term yield, book-to-market ratio, inflation, return on equity, lagged dependent variable, smoothed earnings price ratio. The independent variables are ordered in according to their correlation with the dependent variable.

The in-sample estimation periods $T1$ are set to be 144, 180, 216, and 336 respectively. We define the out-of-sample R^2 as

$$R^2 = 1 - \frac{\sum_{t=T1}^{T-1}(Y_{t+1} - \hat{Y}_{t+1})^2}{\sum_{t=T1}^{T-1}(Y_{t+1} - \bar{Y}_{t+1})^2}$$

where $\hat{Y}_{t+1}$, $\bar{Y}_{t+1}$ are the one-period-ahead prediction and historical average, respectively, using the sample of size $T1$.

2.5.2 Forecasting Results

The out-of-sample R^2 are reported in Table 2.2.

From Campbell and Thompson (2008), we know that when no restriction is put on the sign coefficients and return forecasts, the OLS estimator will generate forecasts that cannot beat the historical average. But considering the positivity restriction they tend to show that the restricted OLS estimators beat the historical average. However, with the ordinary ridge estimators and no constraints imposed on the signs, we could still generate forecasts beating the historical average for most of the cases. This may be due to the fact that the ordinary ridge estimators are restricted to be bounded, which makes them stable. The reason to use the ridge estimator is that, compared to Campbell and Thompson (2008), where only one independent variable is considered in each structure to forecast the equity premium, in this application, more than one explanatory variables are included, multicollinearity, if not perfect, exists and needs to be taken care of. Especially, the behaviors of EOSP1 and MOSP are also identical, which coincides with the simulation results and their forecasts outperform all other estimators.

2.6 Concluding Remarks

In this article, a class of semiparametric ordinary ridge estimators is proposed. Through the kernel density estimation, we are able to derive the estimator and obtain the estimator of regression coefficients in the ridge form. The properties of the estimators have also been investigated. Easy to implement, this class of estimators outperforms both the OLS estimator and the ordinary ridge estimator proposed by Hoerl, Kennard and Baldwin (1975) in both simulation and empirical applications. Also, one of the semiparametric estimator proposed, the EOSP1 estimator, generates almost the same result as the Mallows ordinary ridge estimator, due to the fact that both estimators are obtained through the minimization of unbiased estimators of MSE of the predictor of y, $\mu(h)$. This is an interesting result and we expect to see more applications of our estimator in the future research.

Table 2.1: Risk for Each Estimator

DGP	Estimators	$\sigma = 1$		$\sigma = 5$	
		$n = 50$	$n = 150$	$n = 50$	$n = 150$
1	OLS	0.0282	0.0074	0.6443	0.1923
	HKB	0.0191	0.0064	0.2000	0.0684
	AOSP	0.0237	0.0067	0.4982	0.1448
	AOSP1	0.0198	0.0064	0.2619	0.0786
	EOSP	0.0202	0.0065	0.1167	0.0439
	EOSP1	0.0193	0.0065	0.0912	0.0412
	MOSP	0.0193	0.0065	0.0912	0.0411
2	OLS	0.0279	0.0078	0.6289	0.1918
	HKB	0.0189	0.0069	0.1906	0.0660
	AOSP	0.0234	0.0072	0.4856	0.1455
	AOSP1	0.0193	0.0070	0.2406	0.0764
	EOSP	0.0195	0.0071	0.1069	0.0421
	EOSP1	0.0186	0.0071	0.0905	0.0393
	MOSP	0.0186	0.0071	0.0904	0.0393

Table 2.2: Out-of-Sample R^2

Estimator	$T1 = 144$	$T1 = 180$	$T1 = 200$	$T1 = 216$
OLS	-0.0625	-0.0123	-0.0340	-0.0479
HKB	0.0021	0.0532	0.0351	0.0190
AOSP1	-0.0475	0.0051	-0.0153	-0.0322
EOSP1	0.0485	0.0664	0.0071	-0.0325
MOSP	0.0485	0.0664	0.0071	-0.0325

Chapter 3

A Semiparametric Generalized Ridge Estimator and Link with Model Averaging[*]

3.1 Introduction

Ordinary least squares (OLS) is a widely used estimator for the coefficients of a linear regression model in econometrics and statistics (Schmidt (1976); Greene (2011)). It is shown here that the OLS estimator can also be obtained by estimating population moments (variances and covariances) of the economic variables involved in the regression by using empirical densities of their data sets. Further, we propose a new estimator of

[*] This chapter is a joint work with Dr. Aman Ullah, Dr. Alan K. Wan, Dr. Xinyu Zhang and Dr. Guohua Zou.

the regression coefficients by estimating population moments based on smooth kernel nonparametric density estimation. This proposed estimator, in contrast to the OLS estimator, is robust to multicollinearity, and we refer to this as the semi-parametric (SP) estimator of the regression coefficients. Although there are differences, this SP estimator turns out to be in the form of a generalized ridge regression (GRR) estimator developed by Hoerl and Kennard (1970b). Ridge regression (RR) (Hoerl and Kennard (1970a, b)) is a common shrinkage technique in linear regression when the covariates are highly collinear, and among the various ridge techniques, the GRR estimator is arguably the one that has attracted the most attention. The GRR estimator allows the biasing factor, which controls the amount of ridging, to be different for each coefficient; when the biasing factors are the same for all coefficients, the GRR estimator reduces to the ordinary RR estimator. However, since the biasing factors are unknown, the GRR estimator is not feasible. This is not the case for the SP estimator which is based on the information contained in the kernel density estimation of regressors, and hence the biasing factors are calculated using the data-based window-widths of regressors. Thus, the SP estimator, in contrast to the GRR estimator, is a feasible estimator. This SP estimator is compared with Hoerl and Kennard's (1970b) feasible GRR (FGRR) estimator based on the first step of a data-based iterative procedure for estimating the biasing factors. We note from Hemmerle and Carey (1983) that the FGRR estimator is more efficient than the estimator based on the closed form solution of Hoerl and Kennard's iterative method. For more details of GRR estimators, see Vinod and Ullah (1981) and Vinod, Ullah and Kadiyala (1981).

Yet another independently developed technique closely related to shrinkage estimation is model averaging, which is an alternative to model selection. While the process of model selection is an attempt to find a single best model for a given purpose, model averaging compromises across the competing models, and by so doing includes the uncertainty associated with the individual models in the estimation of parameter precision. Bayesian model averaging (BMA) has long been a popular statistical technique. In recent years, frequentist model averaging (FMA) has also been garnering interest. A major part of this literature is concerned with ways of weighting models. For BMA, models are usually weighted by their posterior model probabilities, whereas FMA weights can be based on scores of information criteria (e.g. Buckland, Burnham and Augustin (1997); Claeskens, Croux and van Kerckhoven (2006); Zhang and Liang (2011); Zhang, Wan and Zhou (2012)). Other FMA strategies that have been developed include adaptive regression by mixing by Yang (2001), Mallows model averaging (MMA) by Hansen (2007, 2008) (see also Wan, Zhang and Zou (2010)), optimal mean square error averaging by Liang, Zou, Wan and Zhang (2011), and Jackknife model averaging (JMA) by Hansen and Racine (2012) (see also Zhang, Wan and Zou (2013)). As well, Hjort and Claeskens (2003) introduced a local misspecification framework for studying the asymptotic properties of FMA estimators.

Given these two independent, but parallel, developments of research in ridge type shrinkage estimators and FMA estimators, the objective of this paper is to explore a link between them. An initial attempt in establishing this connection was made by Leamer and Chamberlain (1976), where a relationship between the ridge estimator and a model average estimator (which they called "search estimator") was noted. However, we emphasize that the ridge and model averaging estimators of Leamer and Chamberlain (1976) are respectively different from the ridge and model averaging estimators in our paper. More importantly, our results permit an exact connection between model averaging weights and ridge biasing factors, whereas their results do not. In addition, we propose a new SP ridge estimator and investigate its properties. The biasing factors of the SP estimator are also linked to the FMA weights. On the basis of these relationships, the selection of biasing factors in the GRR and SP estimators may be converted to the selection of weights in the FMA estimator. Our finding also implies that if the goal is to optimally mix the competing models based on a chosen criterion, e.g., Hansen's (2007) Mallows criterion, then there is always a GRR estimator that matches the performance of the resultant FMA estimator. We demonstrate via a Monte Carlo study that the GRR estimators with biasing factors derived from the weights used for Hansen's (2007) MMA and Hansen and Racine's (2012) JMA estimators perform well, in terms of risk, in a large region of parameter space.

This chapter is organized as follows. In Section 2, we present the SP and GRR estimators of the regression coefficients. In Section 3, we derive the exact algebraic relationship between the biasing factors of the SP and GRR estimators and the weights in the FMA estimator. Section 4 presents asymptotically optimal procedures for choosing window-widths. Section 5 reports the results of a Monte Carlo study comparing the risks of the SP and FGRR estimators with biasing factors based on weights of the MMA and JMA estimators. Section 6 provides two empirical applications of the SP and GRR estimators using the equity premium data in Campbell and Thompson (2008) and the wage data from Wooldridge (2003). Section 7 offers some concluding remarks.

3.2 Semiparametric Estimator of Regression Coefficients

Let us consider a population multiple regression model

$$\begin{aligned} y &= x_1\beta_1 + \cdots + x_q\beta_q + u \qquad (3.1) \\ &= x'\beta + u, \end{aligned}$$

where y is a scalar dependent variable, $x = (x_1, ..., x_q)'$ is a vector of q regressors, β is an unknown vector of regression coefficients, and u is a disturbance with $Eu = 0$ and $V(u) = \sigma^2$.

If we minimize $Eu^2 = E(y - x'\beta)^2$ with respect to β, we obtain

$$\beta = (Exx')^{-1}Exy, \qquad (3.2)$$

where Exx' is a $q \times q$ moment matrix of q variables with the j-th diagonal element and (j, j')-th off diagonal elements given by

$$\begin{aligned} Ex_j^2 &= \int_{x_j} x_j^2 f(x_j)dx_j, \ j = 1, ..., q, \qquad (3.3) \\ \text{and } \ Ex_jx_{j'} &= \int_{x_j}\int_{x_{j'}} x_jx_{j'}f(x_j, x_{j'})dx_jdx_{j'}, \ j \neq j' = 1, ..., q, \end{aligned}$$

respectively.

Suppose we have the sample observations $\{y_i, x_{i1}, ..., x_{iq}\}$, $i = 1, ..., n$. Then the population averages in (3.3) can be estimated by their sample averages

$$\hat{E}x_j^2 = \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2, \text{ and } \hat{E}x_jx_{j'} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ij'}. \qquad (3.4)$$

It is straightforward to note that

$$
\begin{aligned}
\hat{E}x_j^2 &= \int_{x_j} x_j^2 \hat{f}(x_j)dx_j = \int_{x_j} x_j^2 d\hat{F}(x_j) \qquad (3.5)\\
&= \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2
\end{aligned}
$$

by using the empirical distribution of $\hat{F}(\cdot)$. The results for $\hat{E}x_j x_{j'}$ in (3.4) and $\hat{E}x_j y = \sum_{i=1}^{n} x_{ij}y_i/n$ follow similarly.

Using (3.4) and (3.5) in (3.2), we obtain, for all j and j',

$$
\begin{aligned}
\hat{\beta} &= (\hat{E}xx')^{-1}\hat{E}xy \qquad (3.6)\\
&= (X'X)^{-1}X'Y,
\end{aligned}
$$

where X is an $n \times q$ matrix of observations on q variables, Y is an $n \times 1$ vector of n observations and $\hat{\beta}$ is the well-known ordinary least squares (OLS) estimator.

Now we consider the estimation of Ex_j^2 and $Ex_jx_{j'}$ by a smooth nonparametric kernel density instead of the empirical distribution function. This results in

$$
\begin{aligned}
\tilde{E}x_j^2 &= \int_{x_j} x_j^2 \tilde{f}(x_j)dx_j \qquad (3.7)\\
&= \frac{1}{nh_j}\sum_{i=1}^{n}\int_{x_j} x_j^2 k(\frac{x_{ij}-x_j}{h_j})dx_j\\
&= \frac{1}{n}\sum_{i=1}^{n}\int_{\Psi_{ij}} (x_{ij}-h_j\Psi_{ij})^2 k(\Psi_{ij})d\Psi_{ij}\\
&= \frac{1}{n}\sum_{i=1}^{n}\int_{\Psi_{ij}} (x_{ij}^2+h_j^2\Psi_{ij}^2-2x_{ij}h_j\Psi_{ij})k(\Psi_{ij})d\Psi_{ij}\\
&= \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 + h_j^2\mu_2,
\end{aligned}
$$

where $\tilde{f}(x_j) = \frac{1}{nh_j}\sum_{i=1}^{n} k(\frac{x_{ij}-x_j}{h_j})$ is a kernel density estimator, $\Psi_{ij} = \frac{x_{ij}-x_j}{h_j}$ is a transformed variable, $\mu_2 = \int v^2k(v)dv > 0$ is the second moment of kernel function, $k(\Psi_{ij})$ is a symmetric second order kernel, and h_j is window-width. For implementation, h_j can be selected by biased cross-validation based on the Normal or Epanechnikov kernel as in Scott and Terrell (1987). For more details, see Pagan and Ullah (1999).

Similarly, it can be shown easily that

$$\begin{aligned}
\tilde{E}(x_j x_{j'}) &= \int_{x_j}\int_{x_{j'}} x_j x_{j'} \tilde{f}(x_j, x_{j'}) dx_j dx_{j'} \qquad (3.8)\\
&= \frac{1}{nh_j h_{j'}} \sum_{i=1}^{n} \int_{x_j}\int_{x_{j'}} x_j x_{j'} k(\frac{x_{ij}-x_j}{h_j}, \frac{x_{ij'}-x_{j'}}{h_{j'}}) dx_j dx_{j'}\\
&= \frac{1}{nh_j h_{j'}} \sum_{i=1}^{n} \int_{x_j}\int_{x_{j'}} x_j x_{j'} k(\frac{x_{ij}-x_j}{h_j}) k(\frac{x_{ij'}-x_{j'}}{h_{j'}}) dx_j dx_{j'}\\
&= \frac{1}{n} \sum_{i=1}^{n} \int_{\Psi_{ij}}\int_{\Psi_{ij'}} (x_{ij} - h_j \Psi_{ij})(x_{ij'} - h_{j'}\Psi_{ij'}) k(\Psi_{ij}) k(\Psi_{ij'}) d\Psi_{ij} d\Psi_{ij'}\\
&= \frac{1}{n} \sum_{i=1}^{n} x_{ij} x_{ij'}
\end{aligned}$$

and

$$\tilde{E}(x_j y) = \frac{1}{n} \sum_{i=1}^{n} x_{ij} y_i, \qquad (3.9)$$

where the product kernels have been used without loss of generality and $\Psi_{ij'} = \frac{x_{ij'}-x_{j'}}{h_{j'}}$. Also, $\tilde{E}(x_j) = \frac{1}{n}\sum_{i=1}^{n} x_{ij} = \bar{x}_j$.

Thus, by using (3.7) to (3.9) in (3.2), we obtain the following new estimator of β:

$$\begin{aligned}
\tilde{\beta} &= (\tilde{E}xx')^{-1}\tilde{E}xy \qquad (3.10)\\
&= (X'X + D)^{-1} X'Y,
\end{aligned}$$

where $D = diag(d_1, ..., d_q)$ is a diagonal matrix with $d_j = nh_j^2 \mu_2$ as its j-th element ($j = 1, ..., q$). We refer to $\tilde{\beta}$ as the SP estimator.

The estimators in (3.7) and (3.8) are based on kernel density estimation assuming that the continuous regressors have support in the entire Euclidean space. In this paper, we assume that all regressors satisfy this property. However, when the regressors have a bounded support, it is well-known that the kernel density estimator is asymptotically biased and one should use bias adjusted kernels instead; see Li and Racine (2007) and Darolles, Fan, Florens and Renault (2011). When the variables are discrete, the estimator in (3.8) remains the same, but the estimator of Ex_j^2 can be written as $\sum_i x_i^2 p(x_i) = \sum_i \sum_j x_i^2 I(x_j = x_i)/n = \sum_i x_i^2/n$, where $I(x_j = x_i) = 1$ if $x_j = x_i$ and 0 otherwise. In this case, the estimator in (3.10) reduces to the OLS estimator. On the other hand, when the regressor matrix contains a mixture of discrete and continuous regressors, the estimator again has the form of (3.10), except that the matrix D is re-defined with its diagonal elements corresponding to the discrete variables set to zero. This can be explained by noting, for example, when x_1 is continuous and x_2 is discrete, that the estimator of

$$\begin{aligned} E(x_1x_2) &= E_{x_2}[x_2E(x_1|x_2)] \\ &= \sum_i \int_{x_1} K((x_{i1}-x_1)/h_1)dx_1 E[(x_2I(x_{i2}=x_2)/p(x_2)]/nh_1 \\ &= \sum_i \sum_j x_{i1}x_{j2}I(x_{i2}=x_{j2})/n \\ &= \sum_i x_{i1}x_{i2}/n. \end{aligned}$$

Note that both the OLS and SP estimators are based on the population regression (3.1), where the regression coefficient vector depends on the population moments of the vector x and the scalar variable y. These moments are then estimated using sample data by two different methods. This leads to estimators of the regression coefficients in the sample linear regression model

$$Y = X\beta + U, \tag{3.11}$$

where the sample is drawn from the population linear regression model (3.1), and U is an $n \times 1$ vector of random errors with $EU = 0$ and $EUU' = \sigma^2 I_n$. By standard eigenvalue decomposition, we can write $X'X = G\Lambda G'$, where G is an orthogonal matrix and $\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_q)$.

From Hoerl and Kennard (1970a, b), the GRR estimator of β is

$$\hat{\beta}(K) = (X'X + GKG')^{-1}X'Y, \tag{3.12}$$

where $K = diag(k_1, k_2, ..., k_q)$ is a diagonal matrix with $k_j \geq 0$, $j = 1, ..., q$. The $k_j's$ are the biasing factors controlling the amount of ridging in $\hat{\beta}(K)$. When $k_1 = k_2 = \cdots = k_q = k$, $\hat{\beta}(K)$ is commonly called the ordinary ridge regression estimator. We note that the SP estimator in (3.10) is in the form of the GRR estimator but these two estimators are not exactly the same . However, one may define an alternative SP-type estimator by equating the diagonal matrix D to the diagonal of the matrix GKG'. Thus, the elements of D can be determined from the biasing factors of the GRR estimator. Of course, if $K = kI$, then $D = K$ and the SP estimator is identical to the GRR estimator.

Define $Z = XG$ and $\alpha = G'\beta$. Then $Z'Z = \Lambda$ and model (3.11) may be reparameterized as

$$Y = Z\alpha + U. \tag{3.13}$$

Correspondingly, the GRR estimator of α is

$$\hat{\alpha}(K) = (Z'Z + K)^{-1}Z'Y = (\Lambda + K)^{-1}Z'Y = BZ'Y, \tag{3.14}$$

where $B = (\Lambda + K)^{-1}$ is a diagonal matrix. It is straightforward to show that

$$\hat{\alpha}(K) = G'\hat{\beta}(K). \tag{3.15}$$

Hence

$$E(\hat{\alpha}(K)-\alpha)'(\hat{\alpha}(K)-\alpha)=E(\hat{\beta}(K)-\beta)'(\hat{\beta}(K)-\beta). \tag{3.16}$$

That is, the trace of the MSE matrix (or equivalently, the risk under squared error loss) of the GRR estimator of α is the same as that of β, and the matrix K that minimizes the risk of $\hat{\alpha}(K)$ also minimizes that of $\hat{\beta}(K)$. It is well-known that the GRR estimator in (3.12) can be derived by minimizing $u'u$ with respect to β subject to the restriction that $\beta'GKG'\beta$ is bounded. Similarly, the SP estimator in (3.10), derived from using smooth kernel density estimators of moments, also results from minimizing $u'u$ with respect to β subject to a bounded restriction of $\beta'D\beta$. Note that both the GRR and SP estimators are robust to multicollinearity, a property not shared by the OLS estimator derived using empirical density estimation of moments. In Sections 4 and 5 we will show that the proposed SP and GRR estimators have superior performance to the OLS estimator in risk under squared error loss sense.

3.3 Connection between SP and Ridge Estimators and Model Averaging

To examine the connection between the SP and GRR estimators and model averaging, let us consider an averaging scheme across the sub-models

$$Y=Z_s\alpha_s+U,\ \ s=1,2,...,S, \tag{3.17}$$

where Z_s is a sub-matrix containing $q_s \leq q$ columns of Z, and α_s is the corresponding coefficient vector.

Least squares estimation of the models in (3.17) yields the OLS estimators

$$\hat{\alpha}_s = (Z_s' Z_s)^{-1} Z_s' Y. \tag{3.18}$$

Let us write $\alpha_s = A_s \alpha$, where $A_s = (I_{q_s} : 0_{q_s \times (q - q_s)})$ (or its column permutation) is a $q_s \times q$ selection matrix. Conformably, we write $Z_s = Z A_s'$.

The model averaging (MA) estimator of α,

$$\hat{\alpha}(w) = \sum_{s=1}^{S} w_s A_s' \hat{\alpha}_s, \tag{3.19}$$

where $w = (w_1, w_2, ..., w_S)'$ is the weight vector with $w_s \geq 0$ and $\sum_{s=1}^{S} w_s = 1$, is formed by a weighted combination of coefficient estimators across the S sub-models.

We can equivalently write $\hat{\alpha}(w)$ in (3.19) as

$$\begin{aligned} \hat{\alpha}(w) &= \sum_{s=1}^{S} w_s A_s' (A_s Z' Z A_s')^{-1} A_s Z' Y \qquad (3.20) \\ &= C Z' Y, \end{aligned}$$

where

$$\begin{aligned} C &= \sum_{s=1}^{S} [w_s A_s' (A_s Z' Z A_s')^{-1} A_s] \qquad (3.21) \\ &= \begin{pmatrix} w_1^* \lambda_1^{-1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_q^* \lambda_q^{-1} \end{pmatrix} \end{aligned}$$

and

$$w_j^* = \sum_{s=1}^{S} w_s I(j \in \Psi_s), \tag{3.22}$$

with $I(\cdot)$ being an indicator function that takes on 1 if $j \in \Psi_s$ and 0 otherwise, and Ψ_s being a set comprising the column indices of Z included in the s-th sub-model. For example, if the regressor matrix of the s-th sub-model comprises the first, second and fourth columns of Z, then $\Psi_s = \{1, 2, 4\}$. In view of the relationship between w_j^* and w_s, we can write (3.20) as

$$\hat{\alpha}(w^*) = CZ'Y = \hat{\alpha}(w) \tag{3.23}$$

where $w^* = (w_1^*, ..., w_q^*)'$.

Comparing equations (3.14) and (3.20), we notice an algebraic similarity between the GRR estimator $\hat{\alpha}(K) = BZ'Y$ and the MA estimator $\hat{\alpha}(w^*) = CZ'Y$. Clearly, $\hat{\alpha}(K) = \hat{\alpha}(w^*)$ if $B = C$, or more explicitly,

$$\begin{aligned} w_1^*\lambda_1^{-1} &= (\lambda_1 + k_1)^{-1} \qquad (3.24) \\ &\vdots \\ &\vdots \\ w_q^*\lambda_q^{-1} &= (\lambda_q + k_q)^{-1}. \end{aligned}$$

This is the essence of the algebraic equivalence between the GRR and MA estimators. Note that $\lambda's$ depend on the data, and $w^{*\prime}s$ can be determined by the MA weights $w's$ derived under a given criterion. Subsequently, the biasing factors $k's$ of the GRR estimator in (3.12) can be obtained from (3.24).

As a simple illustration, suppose that $q = 2$ in model (3.11) and the data observations are such that $\lambda_1 = 1$ and $\lambda_2 = 1.5$. In this case, the model average is a combination of $S = 3$ candidate models including the full model. The two sub-models contain the first and second regressors respectively, while the full model contains both regressors. Now, suppose that the weights assigned to the three models are $\hat{w}_1 = 0.5$, $\hat{w}_2 = 0.2$ and $\hat{w}_3 = 0.3$

respectively. By (3.22), we have

$$\begin{aligned}
\hat{w}_1^* &= \sum_{s=1}^{3} \hat{w}_s I(1 \in \Psi_s) = \hat{w}_1 + \hat{w}_3 = 0.8 \quad \text{and} \\
\hat{w}_2^* &= \sum_{s=1}^{3} \hat{w}_s I(2 \in \Psi_s) = \hat{w}_2 + \hat{w}_3 = 0.5.
\end{aligned}$$

Then

$$\begin{aligned}
\hat{k}_1 &= \hat{w}_1^{*-1}\lambda_1 - \lambda_1 = 0.25 \quad \text{and} \\
\hat{k}_2 &= \hat{w}_2^{*-1}\lambda_2 - \lambda_2 = 1.5.
\end{aligned}$$

Equation (3.24) also shows that when $k_1 = k_2 = \cdots = k_q = 0$ such that the GRR estimator reduces to the OLS estimator, the MA estimator reduces to the OLS estimator in the full model. It should be mentioned that although (3.22) allows unique w_j^* to be determined from the given values of $w_j's$, the converse need not to be true. Thus, while one can obtain unique GRR biasing parameters from the MA weights using (3.24), the reverse derivation of unique MA weights from the GRR biasing parameters is not always feasible.

Note that the connection between model averaging and ridge estimators has been established on the basis of the orthogonal model. If we apply model averaging to the original regressors X directly, we cannot write the resulting model averaging estimator as a GRR estimator (see (3.12)), especially since $X'X + GKG'$ is not a diagonal matrix. It is only through orthogonalization that the GRR estimator (3.14) and model averaging estimator (3.20) have a common structure, i.e., a diagonal matrix multiplied by $Z'Y$. Due to the convenience it offers, orthogonalization is commonly used in the ridge literature (see Vinod and Ullah (1981)). It has also been used in recent model averaging studies (e.g., Magnus, Powell and Prufer (2010) and Magnus, Wan and Zhang (2011)).

It is also instructive to note that if model averaging is applied to the original regressors, no direct connection can be established for the SP estimator in (3.12) and the model averaging estimator since $X'X + D$ is not a diagonal matrix. Additionally, the estimator for the orthogonal model is $\tilde{\alpha} = (\Lambda + GDG')^{-1}Z'y$, for which no algebraic relationship with the model averaging estimator is apparent. However, if we write model (3.1) as $y = x'GG'\beta + u = z'\alpha + u$, with $z' = x'G$ and $\alpha = G'\beta$, then by using the technique of moments based on kernel density estimation with respect to (3.7) and (3.8), we can obtain $\hat{\alpha} = (Z'Z + D_z)^{-1}Z'y = (\Lambda + D_z)^{-1}Z'y$, where D_z is identical to D in (3.10) except that h_j, the window-width for the j-th variable x_j, is replaced by the window-width h_{jz} used for the density estimation of the j-th variable z_j. Thus, there is a direct linkage between the SP estimator applied to the transformed population model and the model averaging estimator. However, $\tilde{\beta}(D_z) = G'^{-1}\hat{\alpha} = (X'X + GD_zG')^{-1}X'y$, which is identical to the GRR estimator except for the replacement of D_z by K, is not the same as the SP estimator $(X'X + D)^{-1}X'y$ unless $X'X + D = X'X + GD_zG'$, i.e., they are identical only when $D = GD_zG'$. Our simulation results show that these two different looking estimators yield some very similar risk performance. Furthermore, as D and K are diagonal matrices, the optimal choice of K will uniquely determine the optimal choice of D_z; in other words, k_j uniquely determines h_j.

3.4 Asymptotically Optimal Selection of Window-Width in $\tilde{\beta}$

3.4.1 Unbiased Estimator of Exact Risk of SP Estimator and Prediction

From (3.10) and (3.11),

$$\tilde{\beta} - \beta = (X'X + D)^{-1}(X'u - D\beta), \tag{3.25}$$

which yields

$$(\tilde{\beta}-\beta)'(\tilde{\beta}-\beta) = \beta'D(X'X+D)^{-2}D\beta + u'X(X'X+D)^{-2}X'u - 2\beta'D(X'X+D)^{-2}D\beta. \tag{3.26}$$

Therefore, by taking expectations on both sides of (3.26), we can write

$$R(h) = R(\tilde{\beta}) = \beta'A_1\beta + \sigma^2 trA_2, \tag{3.27}$$

where $A_1 = D(X'X + D)^{-2}D$, $A_2 = (X'X + D)^{-2}X'X$ and $h = (h_1^2, ..., h_q^2)'$.

Now, note that an unbiased estimator of $\beta'A_1\beta$ is

$$\hat{\beta}A_1\hat{\beta} - \hat{\sigma}^2 tr(A_1(X'X)^{-1}), \tag{3.28}$$

where $\hat{\sigma}^2 = (Y - X\hat{\beta})'(Y - X\hat{\beta})/(n - q)$ is an unbiased estimator of σ^2. Thus, an unbiased estimator of $R(h)$ is

$$\hat{R}(h) = \hat{\beta}'A_1\hat{\beta} + \hat{\sigma}^2 tr(A_2 - A_1(X'X)^{-1}). \tag{3.29}$$

This expression can be used to find an optimal h. However we note that

$$tr(A_2 - A_1(X'X)^{-1}) = 2tr((X'X + D)^{-1}) - tr((X'X)^{-1}). \tag{3.30}$$

Therefore, it can be verified that,

$$\begin{aligned} \hat{R}(h) &= \hat{\beta}' A_1 \hat{\beta} + 2\hat{\sigma}^2 tr((X'X+D)^{-1}) \qquad (3.31) \\ &= (\tilde{\beta}-\hat{\beta})'(\tilde{\beta}-\hat{\beta}) + 2\hat{\sigma}^2 tr((X'X+D)^{-1}) \end{aligned}$$

is an unbiased estimator of $R(h)$ up to a constant $tr((X'X)^{-1})$ which does not depend on h. Thus the optimization of h based on (3.31) is the same as that obtained from (3.29).

Similarly it can be shown that an unbiased estimator of the predictive risk of $\tilde{\mu} = X\tilde{\beta}$, $E((\tilde{\beta}-\beta)'X'X(\tilde{\beta}-\beta)) = E((\tilde{\mu}-\mu)'(\tilde{\mu}-\mu)) = R_1(h)$, is

$$\tilde{R}_1(h) = \hat{\beta}' A_3 \hat{\beta} + \hat{\sigma}^2 tr(A_4 - A_3(X'X)^{-1}) \qquad (3.32)$$

where $A_3 = D(X'X+D)^{-1}X'X(X'X+D)^{-1}D$ and $A_4 = ((X'X+D)^{-1}X'X)^2$. Further, the minimization of $\tilde{R}_1(h)$ with respect to h is the same as the minimization of Mallows criterion

$$(\tilde{\mu}-Y)'(\tilde{\mu}-Y) + 2\hat{\sigma}^2 tr((X'X+D)^{-1}X'X),$$

which is an unbiased estimator of $R_1(h)$ given by $\tilde{R}_1(h)$ up to a constant.

In the following subsections we show that h obtained by minimizing $\hat{R}(h)$ or $\tilde{R}_1(h)$ is asymptotically optimal. Further, we refer $\tilde{\beta}(h)$ based on $\hat{R}(h)$ as AOSP, and based on $\tilde{R}_1(h)$ as AOSP$_1$.

3.4.1.1 Asymptotically Optimal h Using $\tilde{R}_1(h)$ (Mallows Criterion)

Let $P(h) = X(X'X+D)^{-1}X'$. Then from Section 3.4.1:

$$\tilde{\mu}(h) = X\tilde{\beta} = P(h)Y, \qquad (3.33)$$

the squared error loss function as $L(h) = (\tilde{\mu}(h)-\mu)'(\tilde{\mu}(h)-\mu)$ and the corresponding risk as $R_1(h) = E(L(h))$. We consider the choice of h by a minimization of the following Mallows criterion ($\tilde{R}_1(h)$ in (3.32)):

$$\tilde{R}_1(h) = (\tilde{\mu}(h) - Y)'(\tilde{\mu}(h) - Y) + 2\hat{\sigma}^2 tr(P(h)). \tag{3.34}$$

When minimizing $\tilde{R}_1(h)$, we restrict h to the set $H \in R^q$. Thus, the selected h is

$$\hat{h} = \text{argmin}_{h \in H} \tilde{R}_1(h). \tag{3.35}$$

Let $\xi = \inf_{h\in H} R_1(h)$.

$$\mu'\mu = O(n), X'U = O_p(n^{1/2}) \text{ and } n^{-1}X'X \to \Phi, \tag{3.36}$$

where Φ is a positive definite matrix, and

$$\xi \to \infty, \, \xi^{-2}\mu'\mu = o(1). \tag{3.37}$$

By using conditions (3.33)-(3.34), and the proof steps of Theorem 2.2 of Zhang, Wan and Zou (2013), we obtain the following asymptotic optimality property:

$$\frac{L(\hat{h})}{\inf_{h\in H} L(h)} \to^p 1. \tag{3.38}$$

Proof of (3.38). See the Appendix.

3.4.1.2 Asymptotic Optimal h Using $\tilde{R}(h)$

We restrict h in a set $H \in R^q$. So the selected h is

$$\tilde{h} = \arg\min_{h\in H} \hat{R}(h).$$

Let $\tilde{L}(h) = (\tilde{\beta}(h) - \beta)'(\tilde{\beta}(h) - \beta)$ be the squared loss function and $\tilde{\xi} = \inf_{h \in H} R(h)$, and $\bar{h} = \max(h)$. We assume the following conditions

$$\bar{h} \to 0, \;\; X'U = O_p(n^{1/2}) \text{ and } n^{-1}X'X \to \phi \text{ where } \phi \text{ is a positive definite matrix,} \quad (3.39)$$

$$n^{1/2}\tilde{\xi} \to \infty. \quad (3.40)$$

By using the conditions (3.39)-(3.40), we can obtain the following asymptotic optimality

$$\frac{\tilde{L}(\tilde{h})}{\inf_{h \in H} \tilde{L}(h)} \to^p 1. \quad (3.41)$$

Proof of (3.41). See the Appendix.

From (3.41), the proof of (3.42) in Zhang, Zou, Liang and Carroll (2014), and an additional condition that $(\tilde{L}(\tilde{h}) - \xi_n)\xi_n^{-1}$ is uniformly integrable, we further have

$$\frac{R(\tilde{h})}{\inf_{h \in H} R(h)} \to^p 1. \quad (3.42)$$

3.5 A Monte Carlo Study

The purpose of this section is to demonstrate via a Monte Carlo study the finite sample properties of GRR estimators with biasing factors obtained based on model weights of the Mallows MA (MMA) and Jackknife MA (JMA) estimators. As mentioned previously, these MA estimators were proposed by Hansen (2007) and Hansen and Racine (2012). We denote the corresponding GRR estimators as GRRM and GRRJ estimators respectively.

The weights of the MMA estimator are obtained by minimizing the quadratic form $(Y - Z\hat{\alpha}(w))'(Y - Z\hat{\alpha}(w)) + 2\hat{\sigma}^2 tr(ZCZ')$, where $\hat{\sigma}^2 = (Y - Z\hat{\alpha}_f)'(Y - Z\hat{\alpha}_f)/(n-q)$ and $\hat{\alpha}_f$ is the OLS estimator of α in the full model. On the other hand, the weights of the JMA estimator are determined by minimizing the leave-one-out least squares cross-validation function $CV_n(w) = (Y - \hat{g}(w))'(Y - \hat{g}(w))/n$, where $\hat{g}(w) = \sum_{s=1}^{S} w_s \hat{g}_s$, with $\hat{g}_s = (\hat{g}_{1s}, ..., \hat{g}_{ns})'$, $\hat{g}_{is} = x_i^{s\prime}(X_{-i}^{s\prime} X_{-i}^{s})^{-1} X_{-i}^{s\prime} Y_{-i}$, and X_{-i}^{s} and Y_{-i} being respectively the matrices X^s (the regressor matrix of the s-th submodel) and Y with the i-th element deleted. Following Hansen (2007), we assume that the candidate models in the model average are nested.

Our interest is focused on the risk performance under squared error loss of estimators in terms of the β space in the original model. For purposes of comparisons, we also evaluate the risks of the OLS estimator, the FGRR estimator $\hat{\alpha}_j$, where $\hat{\alpha}_j = \hat{\sigma}^2/\hat{\alpha}_{j,f}^2$ with $\hat{\alpha}_{j,f}$ being the j-th element of $\hat{\alpha}_f$, the asymptotically optimal GRR (AOGRR) estimator, with $k_j's$ obtained by directly minimizing the Mallows criterion $(Y - Z\hat{\alpha}(K))'(Y - Z\hat{\alpha}(K)) + 2\hat{\sigma}^2 tr(ZBZ')$ as a function of K, and the asymptotically optimal SP (AOSP$_1$) estimator, with window-widths obtained by minimizing the Mallows criterion (Section 3.4.1.1) $(Y - Z\hat{\alpha}(D))'(Y - Z\hat{\alpha}(D)) + 2\hat{\sigma}^2 tr(ZB_1Z')$ as a function of D, where $\hat{\alpha}(D) = (\Lambda + G'DG)^{-1}Z'Y = B_1 Z'Y$ is the SP estimator and $B_1 = (\Lambda + G'DG)^{-1}$. When implementing the GRRM, AOGRR, AOSP and AOSP$_1$ estimators, we made use of a constrained optimization routine available in R.version. 2.13.1. We used $k's$ from the FGRR method as the initial values for computing the AOGRR estimator. In Section 3.4.1 we have shown that the optimization under the Mallows criterion is equivalent to the optimization with respect to unbiased estimator of the predictive risk of $\tilde{\beta}(h)$. Therefore, in our simulation, we also consider the AOSP estimator in Section 3.4.1.2 based on the optimization of risk of $\tilde{\beta}(h)$.

Our Monte Carlo experiments are based on following data generating processes (DGP's):

DGP1: $y_i = \sum_{j=1}^{q} \theta_j x_{ij} + e_i$, $i = 1, \cdots n$, with x_{ij} being *iid* $N(0,1)$, e_i being *iid* $N(0,1)$ and $N(0,25)$ and are uncorrelated with $x's$. The same DGP was considered by Hansen (2007) in his Monte Carlo study. We let $\theta_j = 0.7071 j^{-3/2}$, and consider the following pairs of $(n,q) = (50,11)$ and $(150,16)$. To facilitate interpretation of the SP estimates, without loss of generality, we assume the DGP contains no intercept.

DGP2: The set-up is the same as DGP1, except x_{i2} is taken to be the sum of $x_{i3}, \cdots, x_{i50}$ plus an $N(0,1)$ distributed error term. The regressors are thus nearly perfectly correlated.

Our analysis is based on 100 replications. We adopt the Gaussian kernel with $K(\phi) = (2\pi)^{-1/2} \exp[-\frac{1}{2}\phi^2]$, resulting in $\mu_2 = 1$. Following Scott and Terrell (1987), we compute the window-widths of the SP estimator based on biased cross-validation, and the window-widths for the AOSP$_1$ and AOSP based on Mallows criterion$(\tilde{R}_1(h))$ and $\hat{R}(h)$ respectively, see Section 3.4. We have also considered such other window-widths as the naive and AIC cross-validation window-widths, but the biased cross-validation windownwidths generally delivered the best results in risk sense.

The simulation results reported in Table 3.1 show that although the SP, AOSP$_1$ and AOSP estimators behave well when the error variance is small, the GRRM and GRRJ are clearly the preferred estimators when the error variance is large, and often by a large margin. This finding is consistent with our intuition that the large variance associated with the true model makes it difficult to identify the best model, thus making model averaging, which shields against choosing a bad model, a more viable strategy. It is also apparent from Table 3.1 that FGRR and AOGRR estimators yield similar risk performance. This is perhaps attributable to the fact that the biasing factors chosen for the FGRR estimator are optimal in risk sense. See Vinod , Ullah, and Kadiyala (1981, p.363) and Hoerl and Kennard (1970b, p.63). Further, we observe that values of AOSP are smaller compared to AOSP$_1$. This may be due to h used in the AOSP$_1$ estimator is based on minimizing predictive risk ($\tilde{R}_1(h)$) instead of estimator's risk ($\tilde{R}(h)$). By comparing DGP 1 and DGP 2, we notice that when the error variance is large, all estimators in DGP 2 deliver larger risk deductions compared to DGP 1.

3.6 Empirical Applications

This section considers two empirical applications of the proposed methods. The first application uses the methods as forecasting devices for excess stock returns while the second considers wage forecasts.

3.6.1 Forecasting Excess Stock Returns

The data for this example are taken from Campbell and Thompson (2008). The same data set was also used by Jin, Su and Ullah (2012) and Lu and Su (2012) in their studies. This dataset contains $n = 672$ monthly observations between January 1950 and

December 2005 of Y, the monthly excess stock returns of S&P 500 Index, defined as the difference between the monthly stock returns and the risk-free rate. In addition, data observations over the same period are also provided for the following twelve regressors variables, ordered by the magnitude of their correlations with Y, as: default yield spread, treasury bill rate, new equity expansion, term spread, dividend price ratio, earnings price ratio, long term yield, book-to-market ratio, inflation, return on equity, the one-period lag of excess returns and smoothed earnings price ratio. We order these 12 regressors by the magnitude of their correlations with Y. Our model average thus contains the following 13 nested models: $\{1\}, \{1, x_1\}, \{1, x_1, x_2\} ..., \{1, x_1, x_2, ..., x_{12}\}$.

Our estimation is based on $n_1 = 144$, 180, 216, 336 and 456 observations and we use the remaining $n_2 = n - n_1$ observations for out-of-sample forecast evaluation purpose. We measure forecast accuracy based on the out-of-sample R^2 defined as follows:

$$R^2 = 1 - \frac{\sum_{t=n_1}^{n-1} (Y_{t+1} - \hat{Y}_{t+1})^2}{\sum_{t=n_1}^{n-1} (Y_{t+1} - \bar{Y}_{t+1})^2},$$

where $\hat{Y}_p$ is the prediction of Y_p based on a given forecast method and $\bar{Y}$ is the average of Y across the sample of the n_1 observations used for estimating the model. The out-of-sample R^2 is thus negative (positive) when the forecast method yields a larger (smaller) sum of squared forecast errors than does $\bar{Y}$. Table 3.2 reports the out-of-sample R^2 based on the six estimators considered in Section 5 and the selected n_1 values. The results show that except when $n_1 = 180$, the OLS forecasts are inferior to forecasts based on the historical average. This is consistent with the findings of Welch and Goyal (2008) for this data set, that the historical mean gives better forecasts when no restrictions are imposed In all but one case, the FGRR, AOGRR and AOSP$_1$ estimators are also inferior to the historical average in terms of prediction accuracy. On the other hand, the GRRJ and GRRM model averaging estimators result in positive out-of-sample R^2 in the large majority of cases, with GRRJ being the slightly better estimator of the two. The result based on AOSP estimator is not presented here because it does not perform as well as AOSP$_1$. This may be because our evaluation here is based on predictive risk instead of risk of $\tilde{\beta}(h)$.

3.6.2 Forecasting Wages

We use the data given in Wooldridge (2003) containing a cross sectional sample of 526 observations from the U.S. Current Population Survey from year 1976. The dependent variable of interest is the logarithm of average hourly earnings. We consider the following ten regressors, ordered according to their correlation with the dependent variable: professional occupation, education, tenure, female, service occupation, married, trade, SMSA, services, and clerk occupation based on their correlations with the dependent variable. We consider model averages based on 11 nested models in the same manner described in the last example, and $n_1 = 100$, 200, 300, 400..

Table 3.3 reports the out-of-sample R^2 for the six methods. It is apparent from the results that all six estimators yield more accurate forecasts than the historical average. However, the advantage of the GRRJ and GRRM estimators observed in the last example does not extend to the present case, where it is found that the FGRR, AOGRR and AOSP$_1$ estimators all result in more accurate forecasts than the two model averaging estimators. This can be explained by noting that $R^2 = .509$ for the wage data is much higher than that for the equity premium data, which is .097. Thus the standard deviation of errors in the wage data is much smaller compared to the standard deviation of error for the equity premium data, and our simulations suggest that for the small standard deviation case the GRRM and GRRJ estimators are outperformed by other estimators considered. This is also the reason why in the equity premium data, the GRRM and GRRJ estimators prevail since the standard deviation for this data set is much higher. Of the two model averaging estimators, the GRRM estimator is slightly preferred to the GRRJ estimator.

3.7 Conclusions

We have proposed a new SP estimator of regression coefficients which is in the form of the GRR estimator of Hoerl and Kennard (1970b). However, in contrast to the GRR, the biasing factors in our SP estimator are easily implemented by the window-width and the second moment of the kernel function used in the kernel density estimation. The selection of window-width that minimizes Mallows criterion(predictive risk) as well as estimator's risk are also proposed. We also show that the GRR estimator is in fact a model average estimator, and there is an algebraic relationship between the biasing factors of GRR and SP estimators and the model average weights. Naturally, the SP and GRR estimators that select the biasing factors based on this relationship have the same properties as the corresponding model average estimator. This is an interesting finding for the future application and interpretations of the SP and GRR estimators. Our Monte Carlo results demonstrate that some of the recently introduced weight choice strategies for model averaging can result in more accurate estimators than the well-known FGRR and OLS estimators over a wide range of parameter space.

Table 3.1: MSE for Each Estimator

DGP	Estimators	$\sigma = 1$		$\sigma = 5$	
		$n = 50$	$n = 150$	$n = 50$	$n = 150$
1	OLS	0.0251	0.0076	0.6775	0.1802
	FGRR	0.0201	0.0071	0.3374	0.0974
	GRRM	0.0377	0.0236	0.1065	0.0435
	GRRJ	0.0373	0.0236	0.0980	0.0432
	AOGRR	0.0217	0.0079	0.2591	0.0764
	SP	0.0167	0.0064	0.3514	0.1248
	AOSP_1	0.0135	0.0046	0.2220	0.0713
	AOSP	0.0126	0.0045	0.2170	0.0706
2	OLS	0.0253	0.0076	0.6846	0.1807
	FGRR	0.0198	0.0072	0.3388	0.0997
	GRRM	0.0341	0.0228	0.1066	0.0426
	GRRJ	0.0335	0.0228	0.0979	0.0424
	AOGRR	0.0212	0.0080	0.2592	0.0797
	SP	0.0166	0.0064	0.3488	0.1247
	AOSP_1	0.0140	0.0046	0.2251	0.0704
	AOSP	0.0136	0.0045	0.2157	0.0701

Table 3.2: Out-of-Sample R^2 for Stock Returns

Estimator	$n_1 = 144$	$n_1 = 180$	$n_1 = 216$	$n_1 = 336$	$n_1 = 456$
OLS	-0.0390	0.0062	-0.0434	-0.0425	-0.0208
FGRR	-0.0375	-0.0369	-0.0398	-0.0610	-0.0621
GRRM	0.0408	0.0895	0.0564	0.0103	-0.0003
GRRJ	0.0692	0.1079	0.0701	0.0180	0.0020
AOGRR	-0.0375	-0.0369	-0.0398	-0.0610	-0.0621
AOSP_1	-0.0302	0.0195	-0.0271	-0.0170	-0.0148

Table 3.3: Out-of-Sample R^2 for the Wage Data

Estimator	$n_1 = 100$	$n_1 = 200$	$n_1 = 300$	$n_1 = 400$
OLS	0.4516	0.4465	0.4656	0.4450
FGRR	0.4514	0.4440	0.4658	0.4410
AOGRR	0.4509	0.4418	0.4642	0.4390
GRRM	0.3964	0.3366	0.3390	0.3644
GRRJ	0.3877	0.3357	0.3375	0.3627
AOSP_1	0.4550	0.4477	0.4664	0.4470

Chapter 4

Parametric and Nonparametric Frequentist Model Selection and Model Averaging*

4.1 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

*Acknowledgment: The text of this chapter, in part or in full, is a reprint of the material as is appears in Ullah, A. and Wang, H. (2013) *Econometrics* 1(2): 157-179. The co-author Aman Ullah listed in that publication directed and supervised the research which forms the basis for this chapter.

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 Pagan and Ullah (1999), Li and Racine (2007), 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 (Belloni and Chernozhukov (2011)) and in financial econometrics, portfolio allocation may be among hundreds or thousands of stocks (Zhang, Fan and Yu (2011)). 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) (Akaike (1973)), Mallows C_p procedure (Mallows (1973)), Bayesian information criterion (BIC) by Schwarz (1978), cross-validation method by Stone (1974), generalized cross-validation (GCV) by (Craven and Wahba (1979)), and the focused information criterion (FIC) by Claeskens and Hjort

(2003). 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 Frank and Friedman (1993), Fu and Knight (2000) 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 (Hoerl and Kennard (1970)). For a detailed review of model selection in high dimensional modeling, see Fan and Lv (2010), and the books Bühlmann and Van de Geer (2011), Claeskens and Hjort (2008). Similarly, in the context of empirical likelihood estimation and generalized methods of moments estimators, model selection criteria have been introduced by Andrews and Lu (2001), Hall et al. (2007), 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 Pötscher (1991) shows that AIC selection results in distorted inference, and Kabaila (1995) explores the negative impact on confidence regions. Reference Bühlmann (1999) gives conditions under which post model selection estimators are adaptive, but see Leeb and P(ö)scher (2003, 2006) for their comments that they cannot be uniformly estimated. For a selected model with unstable estimators, Breiman (1996) provides bagging or bootstrap averaging procedure to reduce their variances for the i.i.d. data, and by Jin, Su and Ullah (2013) 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 Draper (1995), Hoeting et al. (1999), Clyde and George (2004), Geweke (2005, 2007); for interesting applications in econometrics, see, e.g., Brock, Durlarf and West (2003), Sala-i-Martin, Doppelhofer and Miller (2004) and Magnus, Powell and Pr¨(u)fer (2010). Also, see Claeskens and Hjort (2003) 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 Buckland, Burnham and Augustin (1997), Yang (2001), Burnham and Anderson (2002), Leung and Barron (2006), Yuan and Yang (2005), Hansen (2007), Hansen and Racine (2012) and Wan, Zhang and Zou (2010). Reference Claeskens and Hjort (2003) provides asymptotic theory. For applications, see Kapetanios, Labhard and Price (2006), Wan and Zhang (2009), Claeskens and Hjort (2008). 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, [?] introduces forecast combination and Olkin and Speigelman (1987), Fan and Ullah (1999) suggest combining parametric and kernel estimators of density and regression respectively. Other works include bootstrap based averaging ("stacking") by Wolpert (1992), Breiman (1996)and LeBlanc and Tibshirani (1996), information theoretic method to combine density by Yang (2000) and Catoni (1997), and the mixing of experts models by Jordan and Jacobs (1994) and Jiang and Tanner (2000). Similar kinds of combining have been used in computational learning theory by Vovk (1990, 1998) and in information theory by Merhav and Feder (1998).

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, Ullah (1988), Fan and Gijbels (1996), Pagan and Ullah (1999) and Li and Racine (2007), for kernel smoothing procedures, Eubank (1999) for the spline methods, and Geman and Hwang (1982) and Newey (1997) 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 Fan and Lv (2010) provides a review of parametric model selections, Wang, Zhang and Zou (2009) surveys the FMA estimation, and Su and Zhang (2013) 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 chapter 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.

4.2 Parametric Model Selection and Model Averaging

4.2.1 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

$$y_i = x_i'\beta + u_i = \sum_{j=1}^{q} x_{ij}\beta_j + u_i, \ i = 1, ..., n \tag{4.1}$$

or

$$y = X\beta + u \tag{4.2}$$

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

$$R^2 = 1 - \frac{\sum \hat{u}_i^2}{\sum (y_i - \bar{y})^2}, \ R_a^2 = 1 - \frac{(n-1)\sum \hat{u}_i^2}{(n-q)\sum (y_i - \bar{y})^2}, \ RSS = \sum (\hat{u}_i)^2 \tag{4.3}$$

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 Srivastava, Srivastava and Ullah (1995) for further detail. Thus R_a^2 and RSS are not preferred measures of goodness of fit or model selection. Recently Rousson and Gosoniu (2007) 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 $s_{\hat{u}}^2(n + q + 1)/n$, Rousson and Gosoniu (2007) 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 Wang (2013). Reference Rousson and Gosoniu (2007) 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.

4.2.1.1 AIC, TIC, and BIC

Now we turn to the methods of model selection, AIC in Akaike (1973), Takeuchi informaiton criterion (TIC) in Takeuchi (1976), and BIC in Schwarz (1978). 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 Maasoumi (1993) and Ullah (1996).

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 $2\hat{T}$:

$$AIC = -\frac{2L_n(\hat{\theta})}{n} + \frac{2q}{n}. \tag{4.4}$$

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 (1976), 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 non-normality. 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 Schwarz (1978) 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 Nishi (1984). 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 Hannan and Quinn (1979)). 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 Hurvich and Tsai (1989). 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 Kuha (2004), and for their properties, see Nishi (1984) and Stone (1977, 1979). 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.

4.2.1.2 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). Claeskens and Hjort (2003) provides MSE of $\hat{m_l}$. The basic idea of FIC is to develop a model selection criterion that chooses the model with the smallest 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 $\hat{\Psi}_l = \pi_l'(\pi_l \hat{L}^{-1}\pi_l')^{-1}\pi_l$, $\hat{L} = (Z'M_x Z)^{-1}$ 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 Claeskens and Hjort (2003),

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

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.

4.2.1.3 Mallows Model Selection

Let us write the regression model (4.2) 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 second 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.

4.2.1.4 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 $Eu^2_{n+1} = \sigma^2$ 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, $E\hat{u}^2_i = MSPE(q)$ for each i, and

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

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

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

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

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 Maddala (1988). This expression is useful for calculations. Also, see Stone (1979) for a link of $CV(q)$ with AIC.

4.2.1.5 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 Fan and Lv (2010) 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 (Tibshirani (1996)) 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 Hoerl and Kennard (1970) 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 Ullah et al. (2013).

Further, when $0 < p \leq 1$ we get the bridge estimator (Frank and Friedman (1993), Fu and Knight (2000)) 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 Su and Zhang (2013), Zou (2006), Zhang (2010) and Fan and Li (2001). Also, see the link of LASSO with the least angel regression selection (LARS) by Efron et al. (2004).

4.2.2 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 (4.1) or (4.2) 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.

4.2.2.1 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

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

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 Zhang, Wan and Zhou (2012).

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 Claeskens and Hjort (2003) and Zhang, Wan and Zhou (2012).

4.2.2.2 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

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

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 Li (1987), Hansen (2007) 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. Wan, Zhang and Zou (2010) 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 Hansen (2008) 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.

4.2.2.3 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 Hansen and Racine (2012) relaxes assumptions in Hansen (2007). 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 Hansen and Racine (2012) shows the asymptotic optimality, using Li (1987) and Andrews (1991), 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 Lu and Su (2012) does it for the quantile regression model. Reference Zhang, Wan and Zhou (2012) extends it for the dependent time series models or models with GARCH errors. Also, using MMA method in Hansen (2007), for models with endogeneity, Kuersteiner, and Okui (2010) 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.

4.3 Nonparametric (NP) Model Selection and Model Averaging

4.3.1 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 $P = P(h)$, 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 $i = 1, ..., n$, is an $n \times n$ matrix. 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

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

and its sample global estimator is given by

$$\begin{aligned} R^2 &= [1 - \frac{\sum \hat{u}_i^2}{\sum (y_i - \bar{y})^2}] = [1 - \frac{\hat{u}'\hat{u}}{y'M_2 y}] \\ &= 1 - \frac{y'M_1(h)y}{y'M_2 y} \\ &= \frac{y'M_1^*(h)y}{y'M_2 y} \end{aligned}$$

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 Yao and Ullah (2013) where they showed its application for the statistically significant variables selection in NP regression. R_2^2 is introduced in Su and Ullah (2013) and Huang and Chen (2008). 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.

4.3.1.1 AIC, BIC, and GCV

In the NP case the model selection (choosing q) using AIC is proposed by Hurvich, Simonoff and Tsai (1998). 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 (Racine and Li (2004)) 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 Hansen (2012). 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 Craven and Wahba (1979) 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 Craven and Wahba (1979) 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 4.2.1.3 and 4.3.1.2, respectively.

4.3.1.2 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 4.3.1. 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 4.2.1.3 and Mallows (1973) and Craven and Wahba (1979).

4.3.1.3 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, Hansen (2012) 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 4.2.1.4. 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 (Li and Racine (2007)), and $CV(q) = CV(h, q, d)$ continues to hold. For $d = 0$ we have a local constant LS (LCLS) estimator developed by

Nadaraya (1964) and Watson (1964). 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 Hall and Racine (2013). 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 Li (1987) 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, Härdle, Hall and Marron (1988) 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, Racine and Li (2004) 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 Li and Racine (2001) 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 Andrews (1991) shows that CV is optimal under heteroskedasticity. For GMM model selection which involves selecting moments conditions, see Andrews (1999). Also, see Chen, Hong and Shum (2007) for using minimization of empirical likelihood/KLIC and comments by Schennach (2007) claiming a fundamental flaw in the application of KLIC.

4.3.2 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 4.2.2.2) 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 4.2.2.3, 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 Li (1987). However, for $h = \hat{h}$ the validity of Li's result needs further investigation.

4.4 Concluding Remarks

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 Racine (2000) 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 Caner (2009), Caner and Fan (2011), Garcia (2011), Liao (2013) and Gautier and Tsybakov (2011). 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 Hansen (2012) for NP results of the estimators based on the sieve method.

Chapter 5

Efficient Two Stage Estimation for Nonparametric Panel with Random Effects: With Applications in Health and Environmental Models

5.1 Introduction

Nonparametric modeling has gained its popularity by relaxing the restrictions imposed on functional form compared to the parametric models. Recently, nonparametric modeling and estimation with panel data have attracted much attention among statisticians and econometricians. Among the works, nonparametric random effects panel model has raised a lot attention. Especially, a large number of literature has been developed where the information from the error variance-covariance is incorporated and hence more efficiency could be gained.

Recent works in nonparametric panel estimation with random effects include Ullah and Roy (1998), Lin and Carroll (2000), Ruckstuhl, Welsh and Carroll (2000), Wang (2003), Chen and Jin (2005), Chen, Fan and Jin (2008) and Su and Ullah (2007), among others. However, Henderson and Ullah (2005) showed that almost all these papers developed weighted local linear estimators in which weights involve the matrix of kernel function and error variance-covariance matrix in different ways. Further, the error variance-covariance matrix was estimated, sometimes differently, based on the first stage local linear least squares (LL) estimation. Thus, they were essentially two-step weights least squares estimators. Yet in an extensive simulation study Henderson and Ullah (2012) have found that some of these weighted estimators do not perform well and they may be beaten by the LL estimator which ignores the error variance-covariance matrix. More recently, Martins-Filho and Yao (2009, MY hereafter) proposed a two-step estimator by gaining information from the off-diagonal elements of the error variance-covariance matrix and demonstrated that it beats the LL estimator since the latter ignored information there. Then, Su, Ullah and Wang (2013, SUW hereafter) made a modification from the MY estimator by transforming the model in such a way that homoskedasticity was obtained. They showed more efficiency was gained by the SUW estimator both theoretically and empirically in Monte Carlo simulations. However, both MY and SUW estimators can also be written as a two-step weighted LL estimator with weights and transformed dependent variable different from those used in the weighted LL estimators referred above.

Meanwhile, Yao and Li (2013, YL hereafter) utilized the error information by doing a Cholesky decomposition of the variance-covariance matrix and incorporated a profiling least squares estimation method to introduce a new nonparametric panel estimator. In contrast to the two-step weighted LL estimators, they found the LL estimator of errors first, and then used them to estimate the nonparametric model and parameters of error variance-covariance matrix jointly in the second stage. By comparing against a class of nonparametric panel estimators, e.g. Wang (2003), Chen and Jin (2005), Lin and Carroll (2006), Chen, Fan and Jin (2008), YL showed that their estimator can outperform the other estimators when proper bandwidths were selected with finite sample performances.

Although a lot efforts have been done to incorporate the information from the error terms as mentioned above, most of the two-step weighted LL estimators transformed the model into a format where a matrix times the regression function of interest is put on the right hand side, which means that actually during the second stage, the weighted LL estimator of interest of the regression function at the original data points is obtained by the transformed data. In addition, transformation procedures in some were not straight forward. On the other hand, though YL utilized the information from the estimated error in the first stage, the profiling that motivated the model may also be too complicated to implement due to the following two reasons. First, it may involve large number of parameters to be estimated, hence providing lower degrees of freedom. Second Cholesky decomposition used introduces the problem of estimation with heteroskedasticity. In addition, it is not clear how YL procedure compares with SUW procedure both theoretically and in finite sample performances.

In this chapter, we provide a procedure incorporating the information in covariance matrix through a transformation which is both easy to implement in practice and intuitively straight forward to understand, and Monte Carlo simulation results have also shown its merits compared to both the SUW and the YL estimators. Moreover, the estimator is general in the sense that it can be applied for any regression model with non-scalar error variance-covariance matrix, although later we focus our discussion of its behavior in the one way error random effects panel model setting. Our Monte Carlo simulation shows efficiency gains compared to both the SUW and YL estimators. Asymptotic properties are also established. To illustrate the applicability of our efficient estimator, we have also applied the method in two real data settings. The first application investigates the relationship between health expenditure and education, delivers empirical results for 140 countries across 5 years, and broadens the scope of existing literature from a macroeconomic point of view. The second application estimates the famous environmental Kuznets curve and emphasizes the importance of our nonparametric estimation procedure in such modellings.

The remainder of this chapter is organized as follows. Section 2 first proposes our two-step estimation procedure with general error variance-covariance matrix and then gives its asymptotic properties. Section 3 applies the proposed estimation method to the random effects panel model with both non-profile and profile procedures considered. In addition, a modified new non-profile estimator developed from the YL estimator is introduced in the same section, where restrictions are put corresponding with the random effects panel model structure. Section 4 provides the Monte Carlo simulation results by comparing our estimator's finite sample behavior against that of others. In the Section 5 two empirical applications are introduced to illustrate the usefulness of our proposed estimator and interesting empirical conclusions, as well as policy inferences are presented. Finally, the last section concludes.

5.2 The Nonparametric Model and Two-Step Estimator with General Error Covariance

Let us start from the nonparametric regression model

$$y_i = m(x_i) + w_i,\ i = 1, ..., n \tag{5.1}$$

where y_i is the dependent variable, x_i is a $p \times 1$ vector of exogenous regressors and w_i is an error term such that $Ew_i = 0$ and $Ew_iw_j = \sigma_{ij}(\theta)$ in which θ determines σ_{ij}, $i, j = 1, ..., n$, and $m(\cdot)$ is an unknown function. Further, we assume that w_i is independent of x_i. But we permit the time series structure in either x_i or w_i. Moreover, we allow the non-identical distribution over the $i's$.

Denoting $m = (m(x_1), ..., m(x_n))'$ as an $n \times 1$ vector we can write (5.1) in a vector form as

$$y = m + w \tag{5.2}$$

where $y = (y_1, ..., y_n)'$ and $w = (w_1, ..., w_n)'$ is such that

$$Ew = 0 \text{ and } Eww' = \Sigma(\theta) \tag{5.3}$$

in which $\Sigma = \Sigma(\theta)$ is an $n \times n$ matrix. Now, if we write a local linear approximation with for x_i of all the datapoints in $(x_1, ..., x_n)'$, for $i = 1, ..., n$, we have $m(x_i) \simeq m(x) + (x_i - x)'\beta(x) = z_i(x)\delta(x)$, where $z_i(x) = [1 \ (x_i - x)']$ is $1 \times (p+1)$, $\beta(x) = \frac{\partial m(x)}{\partial x}$, $\delta(x) = (m(x) \ \beta'(x))'$. Then, for $i = 1, ..., n$, $y_i \simeq z_i(x)\delta(x) + w_i$ can be written as

$$y = Z(x)\delta(x) + w \tag{5.4}$$

where $Z(x)$ is $n \times (p+1)$. The well known LL estimator of $\delta(x)$ is obtained by minimizing $w'K(x)w = (y - Z(x)\delta(x))'K(x)(y - Z(x)\delta(x))$ with respect to $\delta(x)$, where $K(x) = K(x,h) = diag(K_h(x_1 - x), ..., K_h(x_n - x))$, $K_h(\cdot) = K(\cdot/h)/h^p$, $K(\cdot)$ is a kernel function, and h is the bandwidth parameter. This is given by

$$\hat{\delta}_{LL}(x) = (Z'(x)K(x)Z(x))^{-1}Z'(x)K(x)y. \tag{5.5}$$

From this the LL estimator of $m(x) = e'\delta(x)$ is

$$\hat{m}_{LL}(x) = e'\hat{\delta}_{LL}(x) \tag{5.6}$$

where $e = (1, 0, ..., 0)$ as a $(p+1) \times 1$ vector.

This LL estimator can still be improved upon since the information hidden in the error variance-covariance matrix Σ is not utilized For this reason, here we propose an efficient two step estimation procedure for $m(x)$ and $\delta(x)$ which uses the information in (5.3). This is as follows.

Starting from equation (5.2), we have

$$\begin{aligned} y &= m + w \qquad (5.7)\\ &= m - \Sigma^{-1/2}w + w + \Sigma^{-1/2}w \\ &= m + (I - \Sigma^{-1/2})w + u \\ &= m + Hw + u, \end{aligned}$$

where $u \equiv \Sigma^{-1/2}w$ is the transformed error term such that $Eu = 0$ and $Euu' = I$, and $H \equiv I - \Sigma^{-1/2}$. Define $y^* \equiv y - Hw$ we have

$$y^* = m + u. \qquad (5.8)$$

The two step more efficient (2S) estimators of $\delta(x)$ and $m(x)$ are given by

$$\hat{\delta}_{2S}(x) = (Z'(x)K(x)Z(x))^{-1}Z'(x)K(x)\hat{y}^* \qquad (5.9)$$

and

$$\hat{m}_{2S}(x) = e'\hat{\delta}_{2S}(x) \qquad (5.10)$$

where $\hat{y}^* = y - \hat{H}\hat{w}$ is based on the first stage estimator $\hat{m}_{LL}$ in which $\hat{w} = y - \hat{m}_{LL}$ and $\hat{H} = I - \hat{\Sigma}^{-1/2}$ with $\hat{\Sigma} = \Sigma(\hat{\theta})$ and θ is estimated by $\hat{\theta}$ at $\sqrt{N}$ rate.

Let $f_i(x)$ denote the marginal density of x_i and $\bar{f} = \bar{f}(x) = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} f_i(x)$. We have the following theorem depicting the asymptotic properties for the two-step estimator.

Theorem 1 *Under proper regularity conditions A1-A4 given in the appendix, we have*

$$\sqrt{nh^p}D_h(\hat{\delta}_{2S}(x) - \delta(x) - B_{2S}) \to^d N(0, \Omega_{2S})$$

where p equals the length of the multivariate x_i, $B_{2S} = \begin{pmatrix} \frac{\kappa_{21}}{2}h^2 \sum_{j=1}^{p} \frac{\partial^2 m(x)}{\partial x_j^2} \\ 0_{p\times 1} \end{pmatrix}$ and $\Omega_{2S} = \begin{pmatrix} (\kappa_{02})^p/\bar{f}(x) & 0_{1\times p} \\ 0_{p\times 1} & \frac{\kappa_{22}(\kappa_{02})^{p-1}}{\kappa_{21}^2 \bar{f}(x)} I_p \end{pmatrix}$, $f_i(x)$ is the marginal density of x_i, $D_h = diag(1, h, ..., h)$ is a $(p+1)\times(p+1)$ matrix, $\kappa_{ij} = \int t^i k(t)^j dt$ for $i, j = 0, 1, 2$.

Theorem 1 implies that for the two-step estimator of $m(x)$, $\hat{m}_{2S}(x)$

$$\sqrt{nh^p}(\hat{m}_{2S}(x) - m(x) - \frac{\kappa_{21}}{2}h^2 \sum_{j=1}^{p} \frac{\partial^2 m(x)}{\partial x_j^2}) \rightarrow^d N(0, (\kappa_{02})^p/\bar{f}(x))$$

and for the two-step estimator of the derivative $\beta(x)$, $\hat{\beta}_{2S}(x)$

$$\sqrt{nh^p}h(\hat{\beta}_{2S}(x) - \partial m(x)/\partial x) \rightarrow^d N(0, \frac{\kappa_{22}(\kappa_{02})^{p-1}}{\kappa_{21}^2 \bar{f}(x)} I_p).$$

The proof of **Theorem 1** is provided in the Appendix.

5.3 Nonparametric Panel with Random Effects

5.3.1 Nonparametric Two Step Estimator in Panel

To apply the above method in the panel model with random effects, let us start from the model

$$\begin{aligned} y_{it} &= m(x_{it}) + \alpha_i + \varepsilon_{it}, \; i = 1, ..., n, \; t = 1, ..., T \qquad (5.11) \\ &= m(x_{it}) + w_{it}, \end{aligned}$$

where y_{it} is the dependent variable, x_{it} is a $p \times 1$ vector of exogenous regressors, α_i captures the individual effect, which follows $N(0, \sigma_\alpha^2)$ and ε_{it} is the idiosyncratic error which follows $N(0, \sigma_\varepsilon^2)$. We assume for simplicity that ε_{it} is independent of α_i and x_{it}.

After vectorization, (5.11) can be rewritten as

$$\begin{aligned} y &= m + \alpha + \varepsilon \qquad (5.12) \\ &= m + w, \end{aligned}$$

where y and ε are $nT \times 1$ vectors, $\alpha = (\alpha_1 \iota_T', ..., \alpha_n \iota_T')'$, ι_T is a vector of ones with length T, $m = (m(x_{11}), ..., m(x_{1T}), ..., m(x_{n1}), ..., m(x_{nT}))'$, $w \equiv \alpha + \varepsilon = (w_1', ..., w_n')$ and $w_i = (w_{i1}, ..., w_{iT})'$, $i = 1, ..., n$. Thus, the variance-covariance matrix for the error w can be written as

$$\begin{aligned} Eww' &\equiv \Sigma = \sigma_\alpha^2 (I_n \otimes J_T) + \sigma_\varepsilon^2 (I_n \otimes I_T) \qquad (5.13) \\ &= \sigma_1^2 P + \sigma_\varepsilon^2 Q \end{aligned}$$

where I_l are identity matrices of dimension l, $l = n, T$ and nT respectively, J_T is a matrix of ones with dimension T, $P = I_n \otimes \bar{J}_T$, $Q = I_{nT} - P$ and $\bar{J}_T = J_T / T$, $\sigma_1^2 = T\sigma_\alpha^2 + \sigma_\varepsilon^2$. From the spectral decomposition as in Baltagi (2008) we note that

$$\Sigma^{-1/2} = \frac{1}{\sigma_1} P + \frac{1}{\sigma_\varepsilon} Q. \qquad (5.14)$$

Next, from (5.12) and (5.7) we can write

$$y^* = y - Hw = m + u \qquad (5.15)$$

where $u = \Sigma^{-1/2} w$ and $H = I - \Sigma^{-1/2}$, in which $\Sigma^{-1/2}$ is given in (5.13) and (5.14).

Now to implement, we perform the LL estimation first to obtain the $\hat{w}$ as regression residuals. Also, we obtain $\hat{H} = I - \hat{\Sigma}^{-1/2}$ by estimating $\hat{\Sigma}$ from the estimators of $\hat{\sigma}_\alpha$ and $\hat{\sigma}_\varepsilon$. Following Henderson and Ullah (2005), among others, $\hat{\sigma}_\varepsilon^2 = \frac{1}{N(T-1)} \sum_i \sum_t (\hat{w}_{it} - \widehat{\bar{w}}_i)^2$ where

$\widehat{\hat{w}}_i = \sum_t \hat{w}_{it}/T$ and as before, $\hat{w}_{it}$ are the i,tth element from $\hat{w}$ .Further $\sigma_1^2 = T\sigma_\alpha^2 + \sigma_\varepsilon^2$, when $\hat{\sigma}_1^2 = T\sum_t \widehat{\hat{w}}_i^2/N$ is obtained, which gives $\hat{\sigma}_\alpha^2 = (\hat{\sigma}_1^2 - \hat{\sigma}_\varepsilon^2)/T$. This procedure gives $\check{y}^* = y - \hat{H}\hat{w}$ in operation, and our two-step estimator is

$$\hat{\delta}_{2S}(x) = (Z'(x)K(x)Z(x))^{-1}Z'(x)K(x)\check{y}^* \tag{5.16}$$

where the $nT \times (p+1)$ matrix $Z(x) = (1\ (x_{it} - x)')'$ for $i = 1, ..., n,\ t = 1, ..., T,\ K_x = diag\{K_h(x_{11} - x), ..., K_h(x_{nT} - x)\}$, and hence $\hat{m}_{2S}(x) = e'\hat{\delta}(x)$ and the derivatives $\hat{\beta}_{2S}(x)$ can also be obtained from $\hat{\delta}_{2S}(x)$.

This estimator is obtained from performing the LL in the first step and then use the residuals for the second step, thus, no profiling technique is adopted here. The following Theorem 2 gives the asymptotic properties of the estimator in the panel data setting.

Theorem 2 *Under proper regularity conditions A1-A4 given in the appendix, we have*

$$\sqrt{nTh^p}D_h(\hat{\delta}_{2S}(x) - \delta(x) - B_{2S}^{panel}) \rightarrow^d N(0, \Omega_{2S}^{panel})$$

where p equals the length of the multivariate x_{it},. $B_{2S}^{panel} = \begin{pmatrix} \frac{\kappa_{21}}{2}h^2\sum_{j=1}^p \frac{\partial^2 m(x)}{\partial x_j} \\ 0_{p\times 1} \end{pmatrix}$ and $\Omega_{2S}^{panel} = \begin{pmatrix} (\kappa_{02})^p/\frac{1}{T}\sum_{t=1}^T f_t(x) & 0_{1\times p} \\ 0_{p\times 1} & \frac{\kappa_{22}(\kappa_{02})^{p-1}}{\frac{\kappa_{21}^2}{T}\sum_{t=1}^T f_t(x)}I_p \end{pmatrix}$, $f_t(x)$ is the marginal density of x_{it}, $D_h = diag(1, h, ..., h)$ is a $(p+1)\times(p+1)$ matrix, $\kappa_{ij} = \int t^i k(t)^j dt$ for $i, j = 0, 1, 2$.

Theorem 2 implies that

$$\sqrt{nTh^p}(\hat{m}_{2S}(x) - m(x) - \frac{\kappa_{21}}{2}h^2\sum_{j=1}^p \frac{\partial^2 m(x)}{\partial x_j}) \rightarrow^d N(0, (\kappa_{02})^p/\frac{1}{T}\sum_{t=1}^T f_t(x))$$

and

$$\sqrt{nTh^p}h(\hat{\beta}_{2S}(x) - \partial m(x)/\partial x) \rightarrow^d N(0, \frac{\kappa_{22}(\kappa_{02})^{p-1}}{\frac{\kappa_{21}^2}{T}\sum_{t=1}^T f_t(x)}I_p).$$

The proof of **Theorem 2** could be found in the Appendix.

Alternatively, we can also use the profile least squares technique to implement this method and get the two-step estimator. Now, to simplify, let's rewrite $y = m + Hw + u$ as

$$y = m + \Lambda\theta + u \tag{5.17}$$

where $\Lambda = \begin{pmatrix} \hat{w}_{11} & \cdots & \hat{w}_{1t} & \cdots & \hat{w}_{n1} & \cdots & \hat{w}_{nT} \\ \sum_{t\neq 1}^{T} \hat{w}_{1t} & \cdots & \sum_{t\neq T}^{T} \hat{w}_{1t} & \cdots & \sum_{t\neq 1}^{T} \hat{w}_{nt} & \cdots & \sum_{t\neq T}^{T} \hat{w}_{nt} \end{pmatrix}'$, $\theta = \begin{pmatrix} c \\ d \end{pmatrix}$, $c = 1 - \frac{1}{\sigma_\varepsilon} + \frac{1}{T}(\frac{1}{\sigma_\varepsilon} - \frac{1}{\sigma_1})$, $d = \frac{1}{T}(\frac{1}{\sigma_\varepsilon} - \frac{1}{\sigma_1})$ and $\sigma_1^2 = T\sigma_\alpha^2 + \sigma_\varepsilon^2$. Note here that because of the setting of c and d, we need the estimate of d to be positive all the time.

As equation (5.17) shows, in $y^* = y - \hat{\Lambda}\theta = m + u$, with $Sy^* = m$, replace m, θ with $\hat{m}$, $\hat{\theta}$ respectively, we have

$$(I - S(x))y = (I - S(x))\Lambda\theta + u \tag{5.18}$$

where $S(x) = (1\ 0)(Z'(x)K(x)Z(x))^{-1}Z'(x)K(x)$.

So, we have

$$\hat{\theta} = \{\Lambda'(I - S(x))'(I - S(x))\Lambda\}^{-1}(I - S(x))'(I - S(x))y \tag{5.19}$$

and $\tilde{\theta} = \hat{\theta}$ if $\hat{d} \geq 0$, $\begin{pmatrix} \hat{c} \\ 0 \end{pmatrix}$ otherwise. Last, we use the local linear least squares method to do the regression to complete the estimation procedure. The asymptotic properties of this estimator can easily be extended from that of the two-step estimator and YL. And the simulations give comparable performances of these two estimators, the results are available from the authors on request.

5.3.2 The Non-Profile Estimator Developed from Cholesky Decomposition

In the YL estimation method in the panel setting, they used Cholesky decomposition of Σ to incorporate the information from the error terms. Specifically, Σ is decomposed by a lower triangle matrix Φ with ones on the main diagonal and $\phi_{i,t}$ as the i, tth off-diagonal element such that $\Phi\Sigma\Phi' = D \otimes I_n$ and $E(\Phi w_i w_i' \Phi') \equiv E e_i e_i'$, where $D = diag(d_1^2, ..., d_T^2)$ is a diagonal matrix and $e_i = \Phi w_i$. Through this transformation, their model can be written as

$$y_{del} = m_{del} + \hat{F}\phi + e_{del} \tag{5.20}$$

where $y_{del} = (y_{12}, ..., y_{nT})'$, $e_{del} = (e_{12}, ..., e_{nT})'$, $m_{del} = (m(x_{12}), ..., m(x_{nT}))'$ are obtained by deleting the first observation of each i from y, e and m, $\phi = (\phi_{21}, ..., \phi_{T,T-1})'$ with dimension $\frac{T(T-1)}{2} \times 1$ and $\hat{w}_{it}$, $i = 1, ..., n$, $t = 1, ..., T$, comes from the first stage LL estimation of y on x; and $\hat{F} = (\hat{F}_{12}, ..., \hat{F}_{1T}, ..., \hat{F}_{nT})'$ with $\hat{F}_{iT} = (0'_{(T-2)(T-1)/2}, \hat{w}_{i,1}, ..., \hat{w}_{i,T-1}, 0'_{(T-1)T/2-(T-1)T/2})$, 0_k being a vector of zeros with k length. It is interesting to note that the model considered by YL is essentially the model $y = m + (I - \Phi)w + \Phi w = m + H_c w + e$ where $H_c = I - \Phi$ based on Cholesky decomposition matrix.

Then, YL suggested a profile estimation procedure where $y^*_{del} = y_{del} - \hat{F}\phi = m(x_{del}) + e_{del}$. This procedure utilizes LL estimation and obtains the profile least squares estimator for ϕ from

$$(I - S_c(x_{del}))y_{del} = (I - S_c(x_{del}))\hat{F}\phi + e_{del} \tag{5.21}$$

where $S_c(x_{del}) = (1\ 0)(Z'(x_{del})W(x_{del})Z(x_{del}))^{-1}Z'(x_{del})W(x_{del})$, $W(x_{del}) = diag\{K_h(x_{12}-$

$x)/\hat{d}_2^2, ..., K_h(x_{1T} - x)/\hat{d}_T^2, ..., K_h(x_{nT} - x)/\hat{d}_T^2\}$ with $K_h(t) = h^{-1}K(t/h)$, and $K(\cdot)$ is the kernel function and h is the bandwidth, $\hat{d}_t$ is any consistent estimator of d_t. In operations, YL estimated $m(x)$ by adding the first observations back for each individual in the proper places, and obtained the estimator ϕ as

$$\hat{\phi} = \{\hat{F}'(I - S_c(x_{del}))\hat{G}^{-1}(I - S_c(x_{del}))\hat{F}\}^{-1}\hat{F}(I - S_c(x_{del}))'\hat{G}^{-1}(I - S_c(x_{del}))y \quad (5.22)$$

where $\hat{G} = diag(\hat{d}_2^2, ..., \hat{d}_T^2, ..., \hat{d}_2^2, ..., \hat{d}_T^2)$. Then from $\hat{y}^* = y - \hat{F}\hat{\phi} = m + e$, the profile YL (YL) estimator can be obtained. YL also provided the asymptotic properties of the their estimator.

However, when we apply the YL method, we notice that in the profiling process, a huge number of parameters in ϕ needs to be estimated, which reduces the degree of freedom in the model. If we do have some knowledge about the structure of the error variance-covariance matrix, e.g. one way error random effects as in this paper, then using this we can simplify the YL estimation. Now we propose the restricted non-profile YL (NYL) estimation method.

Since in the one way error component model with random effects, the structure of variance-covariance matrix of the error terms Σ in (5.13) could give the specification of ϕ as

$$\phi = (-\pi, \frac{\pi}{\pi - 1}, \frac{\pi}{\pi - 1}, \frac{\pi}{2\pi - 1}, ..., \frac{\pi}{(T-2)\pi - 1}) \quad (5.23)$$

where $\pi = -\frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2}$. For simplicity, let us rearrange $\hat{F}\phi$ as $\hat{F}_*\phi_*$ where $\hat{F}_*$, with dimension

$n(T-1) \times (T-1)$, is

$$\hat{F}_* = \begin{pmatrix} \hat{w}_{11} & 0 & \cdots & 0 \\ 0 & \sum_{i=1}^{2} \hat{w}_{1i} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \sum_{i=1}^{T-1} \hat{w}_{1i} \\ & \vdots & & \\ \hat{w}_{n1} & 0 & \cdots & 0 \\ 0 & \sum_{i=1}^{2} \hat{w}_{ni} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \sum_{i=1}^{T-1} \hat{w}_{ni} \end{pmatrix}, \qquad (5.24)$$

and

$$\begin{aligned} \phi_* &= (\phi_{21}, \phi_{31}, ..., \phi_{T1})' \qquad (5.25) \\ &= (-\pi, \frac{\pi}{\pi-1}, \frac{\pi}{2\pi-1}, ..., \frac{\pi}{(T-2)\pi-1})' \end{aligned}$$

without the repetitions as shown in the expression of ϕ. So the last step will be the local linear estimation corresponding with

$$y = m + \hat{F}_* \phi_* + e. \qquad (5.26)$$

As the Cholesky decomposition shows

$$d_t = diag(a, \frac{(a+b)(a-b)}{a}, \frac{(a+2b)(a-b)}{a+b}, ..., \frac{(a+(T-1)b)(a-b)}{a+(T-2)b}) \qquad (5.27)$$

where $a = \sigma_\alpha^2 + \sigma_\varepsilon^2$, $b = \sigma_\alpha^2$.

As shown in (5.25), to estimate ϕ_*, we just need to estimate two parameters, σ_α^2 and σ_ε^2. Also following Henderson and Ullah (2005) and others, $\hat{\sigma}_\alpha^2$ and $\hat{\sigma}_\varepsilon^2$ can be obtained as described in section 5.2. Thus, with $\hat{\sigma}_\alpha^2$ and $\hat{\sigma}_\varepsilon^2$ and $\hat{\pi} = -\frac{\hat{b}}{\hat{a}} = -\frac{\hat{\sigma}_\alpha^2}{\hat{\sigma}_\alpha^2 + \hat{\sigma}_\epsilon^2}$, $\hat{\phi}_* = (-\hat{\pi}, \frac{\hat{\pi}}{\hat{\pi}-1}, \frac{\hat{\pi}}{2\hat{\pi}-1}, ..., \frac{\hat{\pi}}{(T-2)\hat{\pi}-1})$ can be calculated directly. Then, inserting the first observations back, we have $\hat{y}^* = y - \hat{F}_*\hat{\phi}_* = m + w$. Then the last step follows that of section 5.2. The asymptotic properties follows that of the YL estimator, however, from the finite sample behavior in the next section we can still observe the improvement of the NYL upon the YL estimator. This is due to the reason that in the finite sample, the YL needs to do the profiling which reduces the degree of freedom, compared to the NYL estimation.

As mentioned in the introduction, to improve upon the LL estimator in (5.5), many works have been done by incorporating the information in the error variance-covariance matrix. Many of such estimators are written as $\hat{\delta}(x) = (Z'(x)W(x)Z(x))^{-1}Z'(x)W(x)y$ where $W(x)$ is the weighting matrix depending on the matrix of kernel function and error variance-covariance matrix. For example, Lin and Carroll (2000) used $\sqrt{K(x)}\Sigma^{-1}\sqrt{K(x)}$ or $\Sigma^{-1}\sqrt{K(x)}$ and Ullah and Roy (1998) considered $W(x) = \Sigma^{-1/2}K(x)\Sigma^{-1/2}$. Yet the finite sample simulations show that these estimators cannot always outperform the LL estimator, see Henderson and Ullah (2012). To cope with this, MY proposed a method where a two step estimator is developed by setting $W(x) = K(x)$ but the dependent variable is transformed to $H_{MY}P^{-1}y + (I - H_{MY}P^{-1})m$, where $PP' = \Sigma$ and H_{MY} is composed as a diagonal matrix where the information of the diagonal elements of the $\Sigma^{-1/2}$, $v_{11}, ..., v_{nn}$ are extracted and a transformed error with a new diagonal error variance-covariance matrix is obtained. MY showed that this estimator always dominates the LL estimator. To further improve upon the estimator by MY, SUW proposed a method where the diagonal error variance-covariance matrix is further transformed into an identity matrix thus more efficiency is gained. In their case, $Z(x) = H_{MY}^{-1}Z(x)$ and the dependent variable is transformed to $P^{-1}y + (H_{MY}^{-1} - P^{-1})m$. The efficiency gain over MY is also proven in SUW both in theory and by Monte Carlo simulation results. However, all these transformations are not quite straight forward in the sense that the integrity of m is affected during the transformation.

Another recent work that maintains the integrity of m is YL. As introduced in the next section, after the Cholesky decomposition, the information in the error term is incorporated. However, heteroskedasticity remains after their transformation, whereas in our proposed procedure, homoskedasticity is established after our transformation, which saves us from the trouble of estimating the standard deviation of diagonal elements of the transformed error term.

Furthermore, under the same assumption of the above theorem, the asymptotic variance of the univariate local linear estimator $\hat{m}_{LL}(x)$ with working independence correlation structure (Lin and Carroll (2000)) is $\frac{\kappa_{02}}{nhT}(\frac{1}{T}\sum_{t=1}^{T} f_t(x)\sigma_t^{-2})^{-1}$ where $Eww' = \sigma_t^2$. Thus, if $\sigma_t^2 \geq 1, t = 1, ..., T$, our proposed two-step estimator is asymptotically more efficient than the LL estimator. Moreover, the asymptotic variance of the YL estimator is $\frac{\kappa_{02}}{nhT}(\frac{1}{T}\sum_{t=1}^{T} f_t(x)d_t^{-2})^{-1}$, thus, if $d_t^2 \geq 1, t = 1, ..., T$, our estimator is also asymptotically more efficient than the YL estimator. The equality holds only when Σ is an identity matrix. In addition, the asymptotic variance of the univariate SUW estimator is $\frac{\kappa_{02}}{nhT}(\frac{1}{T}\sum_{t=1}^{T} f_t(x)v^2)^{-1}$, where $v = \frac{1}{\sigma_\varepsilon} - (1 - \frac{\sigma_\varepsilon}{\sigma_1})\frac{1}{T\sigma_\varepsilon}$, therefore, if $v^2 \leq 1$, our proposed two-step estimator could be asymptotically more efficient than the SUW estimator.

5.4 Simulation Results

Now we report the Monte Carlo simulation results to evaluate the finite sample performances of the estimators described in Section 3. In Yao and Li (2013), they compared the behavior of the YL estimator against a series of other estimators, e.g. Wang (2003), Chen and Jin (2005), Lin and Carroll (2006), Chen, Fan, Jin (2008) and showed that on average the YL estimator works in a comparable way, especially when the bandwidth is properly selected, the YL estimator outperforms that of the others. Hence for reporting purposes, we will focus on the comparison of the YL estimator and our proposed estimators.

Also, we will compare our results with the inclusion of another newly developed estimator by SUW. They improved upon the MY estimator by changing the error variance-covariance of the transformed model hence gained efficiency. Thus, in the following simulation results, we will report the comparisons between our proposed two-step estimator and the YL estimator, restricted NYL estimator, SUW estimator on their performances of both $m(x)$ and the derivative $\partial m(x)/\partial x$, i.e. $\beta(x)$. We measure the behaviors on the basis of "risk", i.e. the expected mean squared error (MSE). Also, we include the bias and standard deviation for each estimator for completeness.

Consider the following data generating process (DGP):

$$y_{it} = m(x_{it}) + \alpha_i + \varepsilon_{it},\ i = 1, ..., n,\ t = 1, ..., T, \tag{5.28}$$

where the univariate random variable x_i is generated independently from $N(0,1)$ and we specify two forms of $m(x)$. In specification 1, $m(x) = 1 - 0.9e^{-2x^2}$, and in specification 2, $m(x) = 0.5 + exp(-4x)/(1 + exp(-4x))$, as chosen by SUW. For the error terms, α_i and ε_{it}

are independently and identically distributed following (i.i.d.) $N(0, 1)$. For all estimators, Epanechnikov kernel is chosen. For the selection of the first step bandwidth h, we use the cross-validation under AIC criterion which is available in *R version 2.13.1.* Also as shown in SUW, in the second step the bandwidth h' used should be $h' = h^{\frac{4}{5}}$. To keep the consistency, we treat the rest of estimators in the same way. Based on the estimation results of $m(x)$ and the derivative $\partial m(x)/\partial x$ at all the data points, the mean squared error (MSE) are calculated and averaged across 240 repetitions. n is set to be 100 and T is fixed at 4. Table 1 reports the finite sample behaviors for the estimators we have: the naive LL estimator, the SUW estimator, the YL estimator, the NYL estimator, and our two-step estimator.

Several inferences could be made from Table 1. First, our two-step estimator works the best since it reduces "risk" to the largest extent compared to all other estimators for both $m(x)$ and $\partial m(x)/\partial x$. Especially compared with SUW estimator, the efficiency gain is obtained with the ease in implementation, without extracting elements in the H_{MY} matrix, which substantially facilitates the computation and reduces the time and resources required to perform the Monte Carlo and/or empirical estimations. Second, the efficiency gain of NYL over YL demonstrates that imposing restrictions in the YL method under the random effects panel model structure can improve upon the original estimator. This is due to the reduction of the number of parameters to be estimated. In the original YL method, $\phi_{11}, \phi_{21}, \phi_{22}$,$\phi_{31}$,..., ϕ_{nT} are all to be estimated, yet by investigating the structure of the variance-covariance matrix, we ease the problem and hence increase efficiency in the finite sample performances.

5.5 Empirical Illustrations

5.5.1 The Estimation Method Used in the Multivariate Scenario

It is straightforward to apply our estimation method in the multivariate scenario. Yet here in the empirical applications we will use the semiparametric framework to do the analysis. This choice helps in two ways. First, it avoids the "curse of dimensionality" that multivariate nonparametric regressions tend to encounter. Second, applying the semiparametric procedure significantly reduces the time that we need to complete the estimation.

Now let us start from the model

$$y_{it} = m(x_{it}) + r_{it}\zeta + w_{it},\ i = 1, ..., n,\ t = 1, ..., T \tag{5.29}$$

or after vectorization,

$$y = m(x) + r\zeta + w, \tag{5.30}$$

where ζ, a $q \times 1$ vector, is the coefficient vector of the variable r_{it}, which is of dimension $1 \times q$. q controls the number of independent variables that we have in the parametric part of the model. r_{it} is independent from w_{it}, r is of dimension $nT \times q$. We adopt the Robinson (1988) semiparametric estimation method to do the estimation of ζ in the first step. The estimation of ζ by Robinson (1988) is intuitive as follows. First apply the conditional expectation operator $E(\cdot|x_{it})$ and obtain

$$\begin{aligned} E(y_{it}|x_{it}) &= E(m(x_{it})|x_{it}) + E(r_{it}\zeta|x_{it}) + E(w_{it}|x_{it}) \\ &= m(x_{it}) + E(r_{it}|x_{it})\zeta \end{aligned} \tag{5.31}$$

with the law of iterated expectations. Define the conditional expectation g_y at point x as

$g_y(x) = E(y_{it}|x_{it} = x)$ and similarly $g_r(x) = E(r_{it}|x_{it} = x)$, then equation (5.31) can be rewritten as

$$g_y(x) = g_r(x)\zeta + m(x). \tag{5.32}$$

Subtract (5.32) from (5.30), we get

$$\begin{aligned} y - g_y(x) &= (r - g_r(x))\zeta + w \qquad (5.33) \\ \text{or } e_y &= e_r\zeta + w \end{aligned}$$

where $e_y = y - g_y(x)$ and $e_r = r - g_r(x)$ are self-evident. To get the estimator of ζ, $\hat{\zeta}$, we obtain $\hat{e}_y = y - \hat{g}_y(x)$ and $\hat{e}_r = r - \hat{g}_r(x)$ in the first stage and then run the LS of $\hat{e}_y$ on $\hat{e}_r$ in the next stage. $\hat{g}_y(x)$ and $\hat{g}_r(x)$ can be obtained through the LL estimation. Then we form the "new" dependent variable $y_{it} - r_{it}\hat{\zeta}$ in the following nonparametric estimations as what we have described in the previous two sections.

5.5.2 Relationship between Education and Health Expenditure

Starting from Grossman's (1972a, 1972b) model of health production, substantial evidence indicating the causal relationship between health expenditure and education has been found (Grossman and Kaestner 1997, Cutler and Lleras-Muney 2006, Yoo 2011). Increased education may lead to the reduction of health care expenditure and Grossman (1972a, 1972b) posited that increased educational attainment improves individual health through greater productivity efficiency: with given amount of increase in education, people receive higher education tend to produce "health" more efficiently. In other words, the marginal product of "health" is increasing. Also, Rosenzweig and Schultz (1983a, 1983b, RS hereafter) argued that individuals with higher education have more information to help them make the decisions on health maintenance. Empirical datasets of the US are also widely used to demonstrate their claims. For instance, the hypothesis that education reduces health expenditure are verified over the years, from one of the earliest literature, RS, which used 1967, 1968, 1969 National Natality Followback Surveys (USDHEW) to examine approximately 10,000 live births, to a latest one done by Yoo (2011), which was conducted with the Medical Expenditure Panel Survey (MEPS-HC).

However, two limitations remain. One is that they analyzed the relationship between education and health expenditure from the microeconomic point of view, rather than a broader, global prospective. Also, among many other works, most work drew their conclusion based on the US data only, which lacks generality. Second, the way in which the analysis was done could be improved by introducing nonparametrics, which gives more freedom than constructing a parametric model based on assuming the error terms follow a specific distribution.

Taking the above concerns into account, we emphasize the merit our model has in estimating the relationship between education and health expenditure. We will use the Greene (2004) panel dataset from the World Health Organization to perform our analysis, where n, number of countries is set to be 140 and $T = 5$ since we have years 1993 to 1997. The dependent variable is per capita public and private health care expenditure (HEXP) measured in 1997 dollars and the average years of education (EDUC) is our independent variable x_{it}. To partial out the effects related to other independent variables, we use r_{it} as the composite measure of health care delivery (COMP)*. More details could be found from Greene (2004). Thus the model becomes

$$HEXP_{it} = m(EDUC_{it}) + COMP_{it}\zeta + w_{it},\ i = 1, ..., n,\ t = 1, ..., T \tag{5.34}$$

where all variables are measured in log terms. This is due to two reasons. First, as pointed by Yoo (2011), health care expenditure is highly positively skewed, whose skewness could be reduced by taking the log terms. Second, by measuring in log format, the derivatives that we get will give us the elasticity of education on health expenditure directly.

The estimated mean elasticity for our two-step estimator of education is 0.7414, whose positivity states that over the 140 countries we have, actually education increases health expenditure, which is in contrast to what Grossman (1972a, 1972b), RS(1983a, 1983b), Yoo (2011) had as the conclusion for their hypothesis. Yet remember that when they analyzed the elasticity of education, they used the US data only; by looking at the US observations, the mean elasticity that we estimated with our model is -0.1810, which is consistent with their empirical findings. More specifically, Figure 1 captures the elasticity of all countries over the 5 years. Several inferences could be drawn from Figure 1 here. First, the elasticity curve has several intersections with the 0 line, which demonstrates that overall, the elasticity is not a constant term across countries, which supports the usage of nonparametric estimation. Second, in countries with low average level of education ranging

[*] Here, personal income, gini coefficient, population density, the country's demoncracy level, government efficiency, percentage of health care paid by the government and whether a country belongs to OECD, etc. are all heterogeneity indicators of cross country. Thus, they all enter into the term α_i, which is to be estimated in the second stage.

from no education to about 4 years of education, e.g. Bolivia, generally more education leads to the increase of health expenditure. This could be explained by the fact that more educated people usually have more access to better medicine and health facilities, and thus correspondingly they will spend more on their health issues. And for most countries, especially where people receive higher education (e.g. more than 7.3 years), the hypothesis that there's the inverse relationship between education and health expenditure no longer holds. People are more willing to pay for their health as they are better educated, their lives worth "more" than the others as they have better access to health care facilities or more "capacity" to receive better health care. This "capacity" could be understood as their purchasing power, or ability to communicate with the doctors and follow the doctors' instructions more effectively.

Figure 2 captures the elasticities for OECD countries. From this figure, we can observe that for these OECD countries, most people receive higher education than people from the rest of the world since people have more purchasing power to pay for education in developed countries. Also, as education increases, the elasticity of education on health care expenditure increases from negative to positive. This transition could also be understood by the logic provided before.

Now let's look at the elasticity of education in the US as shown in Figure 3. Here we can see that the negative values we have as elasticities support the hypotheses made by Grossman (1972a, 1972b), RS (1983a, 1983b) and Yao (2011). Within the US, it is indeed true that as education increases, health expenditure decreases. However, when we look at the elasticity of the European countries, this is not the case. The positivity of the elasticities for Greece in Figure 4 shows that education serves as one of the key factors increasing health expenditure. Also, over the 5-year period we can almost observe an increase in elasticity, which means that over the years the given the same increase in education, more health care expenditure needs to be made.

One of the possible explanation for the positivity among the OECD countries, especially the European countries, could be understood by their high social welfare system. Their system doesn't motivate them to use "information on health" more efficiently since they will get their medical care no matter what. Thus, people with more education are not constrained by this "budget" as the American people are. Yet for the European people, the government will take care of it with the near perfect social welfare plan, thus this correspondence within the US system no longer exists.

5.5.3 The Environmental Kuznets Curve (EKC)

As the human race is building the world and developing economy, environmental costs seem to be inevitable. What kind of relation does the cost, such as environmental damage and/or pollution, and the developing economy have, and how is this relation evolving raise the interests of economists. One of the ideas that the economists hold serving as an explanation would be the Kuznets curve hypothesis. It posits that there's an bell-shaped relation between environmental damage/pollution and Gross Domestic Product (GDP). At the initial stages of an economic development, GDP will be increased as more environmental damages are done, which happens because the marginal benefit of economic gains exceeds the social marginal cost of doing damage to the environment. Yet as marginal cost is increasing and marginal benefit is decreasing, a threshold exists. Later when a threshold is met, GDP will reach its peak for a given environmental condition and as more damages are done, GDP will be decreasing.

Many works have been done to study whether the Kuznets curve hypothesis could be verified. Mostly cited as the first literature, Malenbaum (1978) derived an inverted U-shaped relationship between the intensity of metal use and income. Yet the topic gained its popularity in the 1990s. Grossman and Krueger (1991) investigated the relationship between air quality and economic growth in a random effect panel model with 42 countries' data. Among the three air pollutants they used as indicator, sulfur dioxide (SO_2) and smoke revealed the bell shape that Kuznets curve proposed. For a more detailed literature review, see Dinda (2004). As early works were done in the parametric setting, later economists began the use of quadratic, cubic parametric models and nonparametric models to study the hypothesis. For instance, Azomahou, Laisney and Van (2006) used both parametric and nonparametric panel estimation to verify the existence of the Kuznets curve. While it is indeed true that using a parametric model, with CO_2 emissions as the environmental indicator, they found evidence of the Kuznets curve; however with nonparametric fixed effects analysis, the hypothesis was rejected. Also, parametric model was rejected against the nonparametric model specification. For more views on semiparametric and nonparametric studies on EKC, see Zapata, Paudel and Moss (2008).

Here in our application, we will use Particulate Matter 10 (PM10) as our environmental indicator as in Nigaru (2012), and economic development is measured by the Gross Domestic Product (GDP). In this panel data, $n = 160$ and $T = 4$. The dependent variable

y_{it} is PM10 with unit $\mu g/m^3$, while the independent variable x_{it} is real GDP per capita measured in constant 2000 US dollars. Both are measured in log terms. Three controls are selected as trade openness measured in percentage in 2005 constant price, coal consumption per capita measured in ton, and proportion of urban population as a percentage of total population, which forms the vector of r_{it}. Figure 5 reports the estimation result.

When we look at the estimation result that we get, it's very obvious that when we use PM10 as the indicator, we generate similar conclusion as in Azomahou, Laisney and Van (2006). In other words, it is possible that when GDP per capital is high yet not too high, we are going to observe a fact that even when the economy has already evolved into a more developed phase, environmental damage may still stimulate the economy to reach a higher GDP level. And in general, we have the negative relation between PM10 and GDP, which means environmental damage would harm the economy most of the time. Thus, policy makers should still focus on directing the economy to a balanced growth path where sustainable economic development should be encouraged.

5.6 Concluding Remarks

We have developed a new two stage local estimation procedure for regression functions and their derivatives in the nonparametric setting. This is straight forward and easy to implement in practice. Especially, application of this method using random effects panel data with one-way error components has also been demonstrated. Monte Carlo simulation results are also provided to show the efficiency gain associated with our estimator. Monte Carlo simulation results and empirical applications have also been provided to demonstrate the usefulness of the estimation procedure. Moreover, the method can be easily extended to other frameworks, such as varying-coefficients models, etc. The estimation procedure proposed in our paper can also be used for the cluster and seemingly unrelated regression models, or panel with serial correlations and/or heteroskedasticity. Further, it is straight forward to use for the random effects panel with conditional variance-covariance $\Sigma(x)$ estimated nonparametrically. Such extensions are subjects for future research.

Table 5.1: Simulation results for estimators of $m(x)$ and $\partial m(x)/\partial x$

Specification 1	Estimator		LL	SUW	YL	NYL	2S
	$m(x)$	MSE	**0.0766**	**0.0510**	**0.0539**	**0.0534**	**0.0503**
		Std	0.2766	0.2254	0.2320	0.2314	0.2239
	$\partial m(x)/\partial x$	MSE	**0.2496**	**0.1688**	**0.1738**	**0.1717**	**0.1653**
		Std	0.4996	0.4109	0.4169	0.4143	0.4065
Specification 2	$m(x)$	MSE	**0.0683**	**0.0463**	**0.0489**	**0.0485**	**0.0457**
		Std	0.2613	0.2152	0.2211	0.2202	0.2138
	$\partial m(x)/\partial x$	MSE	**0.1804**	**0.1200**	**0.1273**	**0.1248**	**0.1185**
		Std	0.4246	0.3463	0.3567	0.3531	0.3441

Figure 5.1: Elasticities of Education for 140 Countries

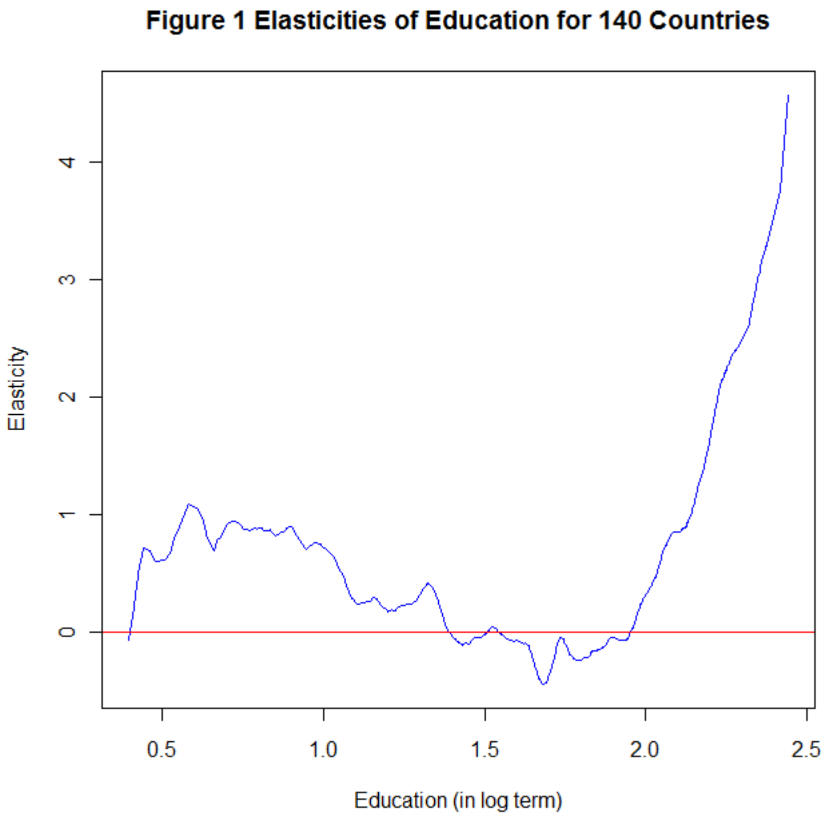


Figure 5.2: Elasticities of Education for OECD Countries

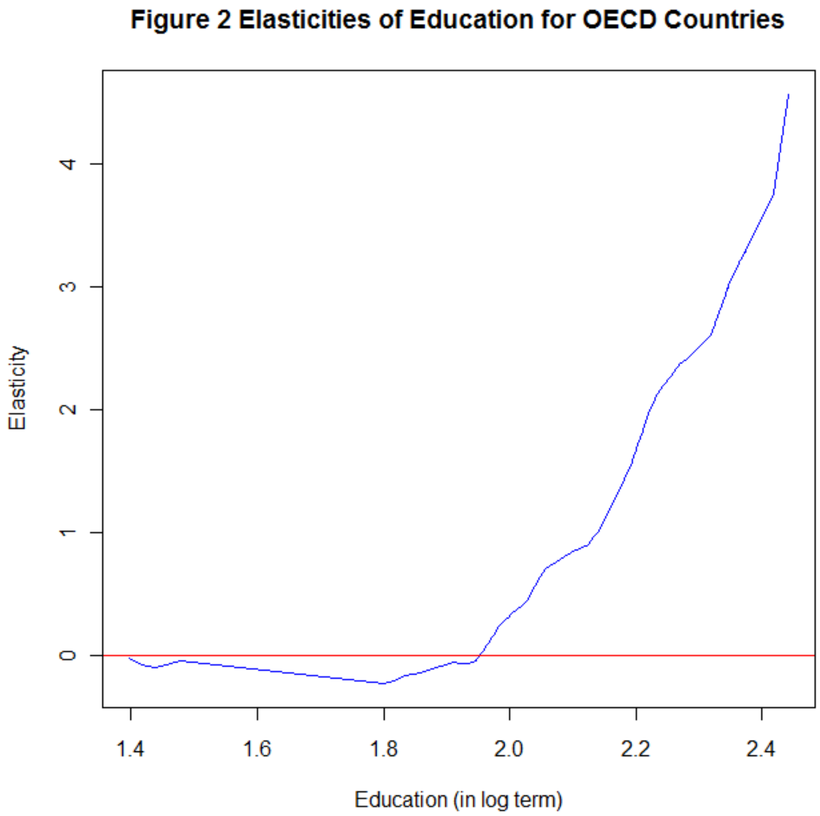


Figure 5.3: Elasticities of Education for the US

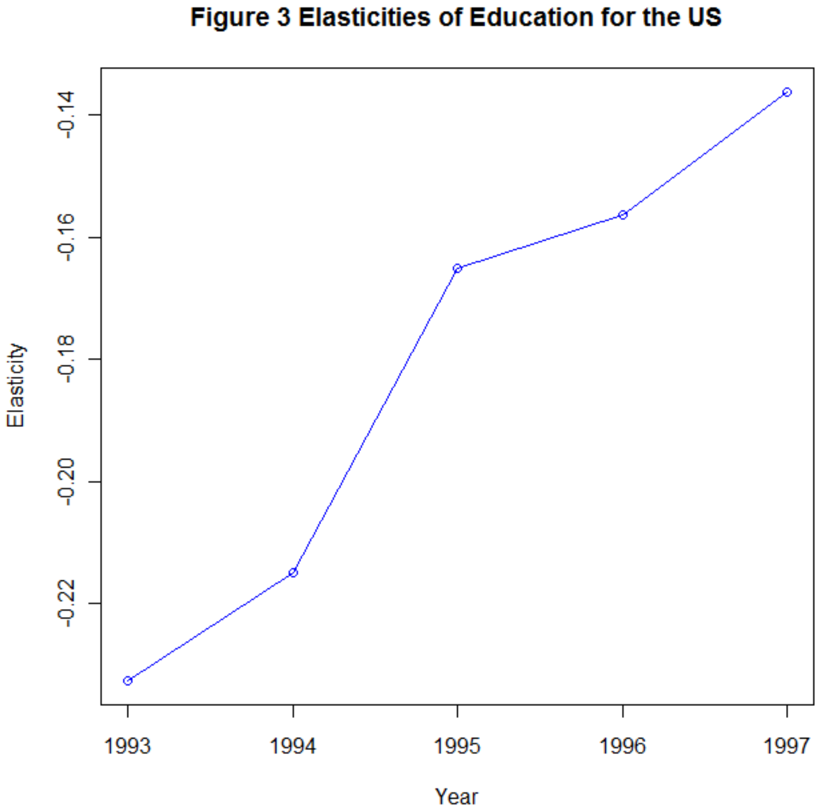


Figure 5.4: Elasticities of Education for the Greece

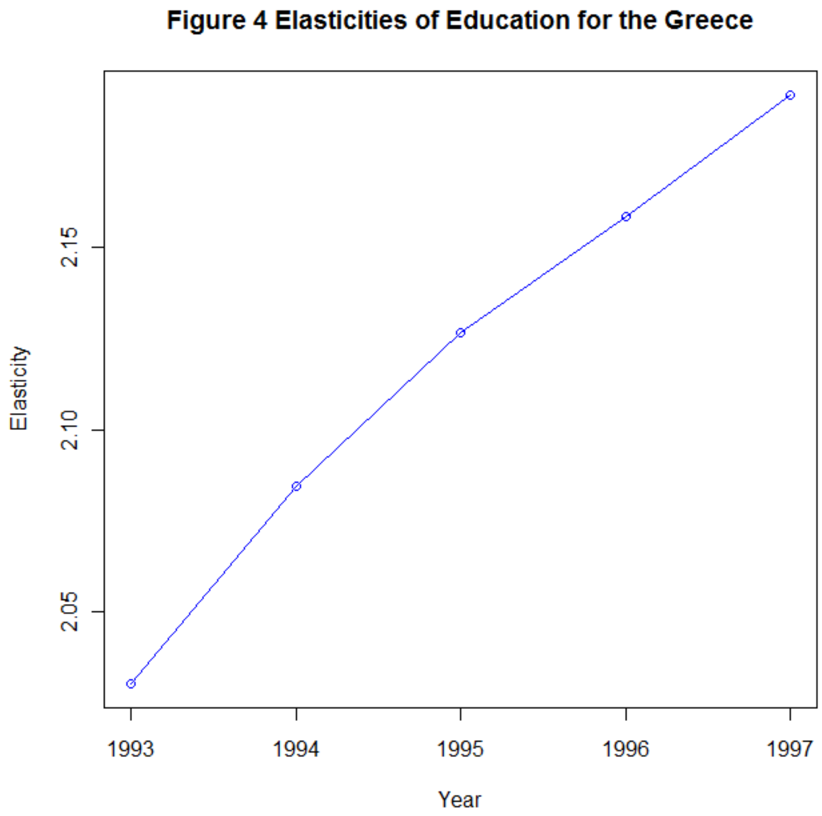
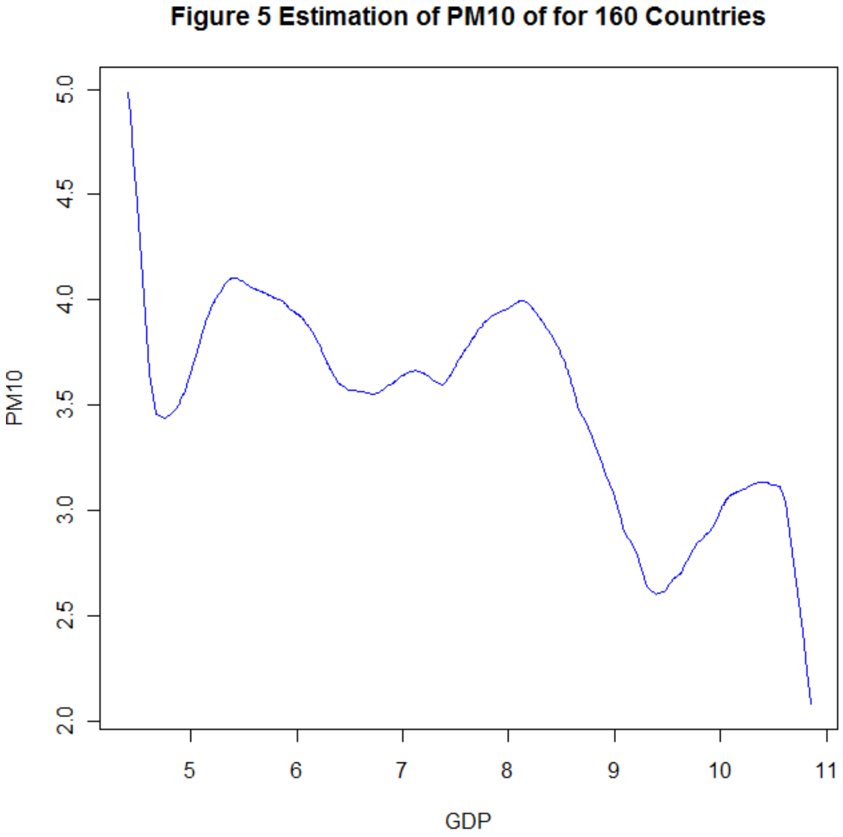


Figure 5.5: Elasticities of PM10 for 160 Countries



Chapter 6

Conclusion

This thesis focuses on the topics of ridge-type shrinkage estimation from both the point of view of both semiparametric and model averaging, as well as the nonparametric estimation with panel data random effects. By relaxing the assumptions in functional forms between the regressors and the regressand, and regressors to be considered, the topics bring more possiblities for economists to deal with model uncertainty, parametrically, semiparametrically, and nonparametrically. Chapter 2-3 discusses the Ridge-type shrinkage estimation in both ordinary and general ridge regression settings. Both simulation results and the empirical examples are adopted to prove the efficiency of the proposed estimators. Chapter 4 provides a comprehensive review on topics related to model selection, model averaging, parametrically and nonparametrically. In addition to exploring the recent developments in these areas, this chapter also proposes interesting extensions and future directions along the line. Chapter 5 brings economists' attention to panel data estimation where the random effects are considered. A two stage estimator is proposed, which could result in a smaller MSE when certain conditions are satisfied. To illustrate the usefulness of the estimation procedure, two empirical applications are also considered.

More specifically, chapter 2 proposes a class of ordinary ridge estimation procedures where we start from kernel estimation, and by building up the connection between ridge estimation and the kernel density estimator of the coefficients, MSE is reduced in the new estimator. More over, through minimizing th unbiased estimator of MSE of the predictor, another new estimation procedure provides us with nice regression results which coincide with the Mallows criterion (Mallows, (1973)). This interesting theoretical results also have proven the usefulness of the new criterion. Empirical application where the forecasting behavior of the proposed estimators are also reported, where large improvements of the out of sample R squared upon the feasible ordinary ridge estimation are observed.

Chapter 3 extends the work of Chapter 2 into the general ridge regressions' framework. Here an asymptotically optimal semiparametric ridge regression procedure is extensively studied. In addition to providing the asymptotic properties of this estimator, Monte Carlo simulations with different DGP are also done to show the numerical behavior. Also, a class of general ridge estimators incorporating Mallows model averaging and Jackknife model averaging are also proposed. The latter estimators performs relatively well when the model uncertainty passes a threshold. Thus, we complete the estimation recommendation by using the AOSP when model uncertainty is small and using the GRRM/GRRJ when the error variances are large. Two empirical applications, where we forecast wage and excess stock returns, are considered to demonstrate the real life behaviors of the proposed estimators.

Chapter 4 explores the literature in the realm of model averaging and model selection for both parametric and nonparametric framework by introducing the concepts and recent developments. Mainly focused on, yet not confined with, frequentists' works in model averaging and model selection, this review is comprehensive and inspiring. In addition to introducing major works on parametric model averaging and model selection, nonparametric model averaging estimation procedure is also proposed to extend the readers' interests.

In chapter 5 of the thesis, a two-stage panel random effects estimation procedure is studied where reduction of the MSE is achieved when certain conditions are satisfied. This work improves upon the recent proposed estimation method proposed by Li and Yao (2013). Theoretical proof, Monte Carlo simulations are provided to illustrate the good behavior of the proposed estimator. In addition, two empirical applications are considered to provide possible ways to adopt the estimation procedure. The first estimation investigates the relationship between one's educational level and his health expenditure; the second estimates the environmental Kuznets curve nonparametrically. Both applications provide interesting implications in the real life.

Appendix A

Mathematical Derivations

A.1 Derivations in Chapter 2

Let $f = f(x)$ denote the continuous density function of a random variable X at point x, and $x_1, x_2, ..., x_n$ be the observations from f. As in section 2, kernel density estimator $\tilde{f}(x_j) = \frac{1}{nh}\sum_{i=1}^{n} k(\frac{x_{ij}-x_j}{h}), where\ k(\cdot)$ is the kernel function. In the population $Y = X\beta + U$, Y is a scalar dependent variable, $X = [X_1, ..., X_q]'$ is a vector of q regressors, β is an unknown vector of regression coefficients, U is an $n \times 1$ vector of random errors. We make the following assumptions following Pagan and Ullah (1999):

A1. The observations $x_1, x_2, ..., x_n$ are independent and identically distributed (i.i.d.).

A2. The kernel $k(\cdot)$ is a symmetric function around zero satisfying

(i) $\int k(v)dv = 1$,

(ii) $\int v^2 k(v)dv = \mu_2 \neq 0$,

(iii) $\int k^2(v)dv < \infty$.

A3. The second order derivatives of f are continuous and bounded in some neighborhood of x.

A4. $h = h_n \to 0$ as $n \to \infty$.

A5. $nh_n \to \infty$ as $n \to \infty$.

A6. $EU = 0$ and $EUU' = \sigma^2 I_n$.

Proof of Theorem 1.

Proof. Under the above assumptions, since

$$\begin{aligned}\tilde{\beta}(h) &= (X'X + nh^2\mu_2 I)^{-1}X'Y \\ &= (X'X + nh^2\mu_2 I)^{-1}X'(X\beta + U) \\ &= (X'X + nh^2\mu_2 I)^{-1}((X'X + nh^2\mu_2 I - nh^2\mu^2 I)\beta + U),\end{aligned}$$

we have

$$\begin{aligned}\tilde{\beta}(h) - \beta &= (X'X + nh^2\mu_2 I)^{-1}(X'U - nh^2\mu_2\beta) \\ &= [I + (X'X)^{-1}nh^2\mu_2 I]^{-1}[(X'X)^{-1}X'U - (X'X)^{-1}nh^2\mu_2\beta].\end{aligned}$$

Let $A = n\mu_2(X'X)^{-1}$, then $A = A'$, hence

$$\tilde{\beta}(h) - \beta = (I + h^2A)^{-1}[(X'X)^{-1}X'U - h^2A\beta].$$

Since the window-width is small, or $h^2 \to 0$, we expand at 1, get a geometric series at the right hand side as

$$(I + h^2A)^{-1} = I - h^2A + h^4A^2 + O(h^6).$$

Thus

$$\begin{aligned}\tilde{\beta}(h)-\beta &\simeq [I-h^2A+h^4A^2][(X'X)^{-1}X'U-h^2A\beta] \\ &= (X'X)^{-1}X'U-h^2A\beta-h^2A(X'X)^{-1}X'U+h^4A^2\beta+h^4A^2(X'X)^{-1}X'U,\end{aligned}$$

and

$$Bias = E(\tilde{\beta}(h)-\beta) = h^4A^2\beta-h^2A\beta,$$

$$\begin{aligned}V(\tilde{\beta}(h)-\beta) &= E[(\tilde{\beta}(h)-\beta)(\tilde{\beta}(h)-\beta)'] \\ &= \sigma^2(X'X)^{-1}-h^2\sigma^2(X'X)^{-1}A'+h^4\sigma^2(X'X)^{-1}A'^2-h^2\sigma^2A(X'X)^{-1} \\ &\quad +h^4A\beta\beta'A'+h^4\sigma^2A(X'X)^{-1}A'+h^4\sigma^2A^2(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}-2h^2\sigma^2\frac{A^2}{n\mu_2}+h^4[A\beta\beta'A'+3\sigma^2\frac{A^3}{n\mu_2}].\end{aligned}$$

And hence the AMSE of $\tilde{\beta}(h)$ is

$$\begin{aligned}AMSE &= V(\tilde{\beta}(h)-\beta)+(Bias)^2 \\ &= V(\tilde{\beta}(h)-\beta)+(h^4A^2\beta-h^2A\beta)(h^4A^2\beta-h^2A\beta)' \\ &\simeq V(\tilde{\beta}(h)-\beta)+h^4A\beta\beta'A' \\ &= \sigma^2(X'X)^{-1}-2h^2\sigma^2\frac{A^2}{n\mu_2}+h^4[2A\beta\beta'A'+3\sigma^2\frac{A^3}{n\mu_2}].\end{aligned}$$

To minimize the risk, we need

$$\begin{aligned}E(\tilde{\beta}(h)-\beta)'(\tilde{\beta}(h)-\beta) &= tr(AMSE) \qquad \text{(A.1)} \\ &= \sigma^2tr(X'X)^{-1}-2h^2\sigma^2\frac{trA^2}{n\mu_2}+h^4[2\beta'A^2\beta+3\sigma^2\frac{trA^3}{n\mu_2}].\end{aligned}$$

The first order condition of equation (A1) gives $\frac{\partial trAMSE}{\partial h}=0$, and thus

$$h^2=\frac{\sigma^2 trA^2}{2n\mu_2\beta' A^2\beta+3\sigma^2 trA^3}.$$

■

Proof of Theorem 2.

Proof.

Under condition A1-A6 above, let $D=(X'X+nh^2\mu_2 I)^{-1}$, so $D=D'$ and

$$\tilde{\beta}(h)-\beta=(X'X+nh^2\mu_2 I)^{-1}X'Y-\beta.$$

Thus MSE of $\tilde{\beta}(h)$ is

$$\begin{aligned}
MSE(\tilde{\beta}(h)) &= E[(\tilde{\beta}(h)-\beta)'(\tilde{\beta}(h)-\beta)] \\
&= E[(U'X-nh^2\mu_2\beta')D'D(X'U-nh^2\mu_2\beta)] \\
&= E[U'XD^2X'U]+(nh^2\mu_2)^2\beta' D^2\beta \\
&= \sigma^2 tr(XD^2X')+(nh^2\mu_2)^2\beta' D^2\beta.
\end{aligned}$$

Since $\hat{E}(\hat{\beta}-\beta)'(\hat{\beta}-\beta)=s^2tr(D^2X'X)+h^2\hat{\beta}D^2\hat{\beta}$, where $\hat{\beta}\sim N(\beta,\sigma^2(X'X)^{-1})$,we have

$$\begin{aligned}
E(\hat{\beta}'D^2\hat{\beta}) &= E[(\hat{\beta}-\beta)'D^2(\hat{\beta}-\beta)]+\beta' D^2\beta \\
&= E[(\hat{\beta}-\beta)'(X'X)^{\frac{1}{2}}(X'X)^{-\frac{1}{2}}D'D(X'X)^{-\frac{1}{2}}(X'X)^{\frac{1}{2}}(\hat{\beta}-\beta)]+\beta' D^2\beta \\
&= E[Z'CZ]+\beta' D^2\beta \\
&= \sigma^2 trC+\beta' D^2\beta
\end{aligned}$$

where $Z\equiv(X'X)^{\frac{1}{2}}(\hat{\beta}-\beta)$, $C\equiv(X'X)^{-\frac{1}{2}}D'D(X'X)^{-\frac{1}{2}}$.

Thus the unbiased estimator of $\beta' D^2 \beta$ is $\hat{\beta}' D^2 \hat{\beta} - \sigma^2 trC$, and the unbiased estimator of the MSE of $\tilde{\beta}(h)$ is $s^2 tr(XD^2X') + (nh^2\mu_2)^2[\hat{\beta}'D^2\hat{\beta} - s^2 trD^2(X'X)^{-1}]$. ■

Proof of Theorem 3.

Proof. Under the same conditions with Theorem 2,

$$\begin{aligned}
MSE(\tilde{\mu}(h)) &= E[(\tilde{\mu}(h) - \mu)'(\tilde{\mu}(h) - \mu)] \\
&= E[(\tilde{\beta}(h) - \beta)'X'X(\tilde{\beta}(h) - \beta)] \\
&= E[(U'X - nh^2\mu_2\beta')D'X'XD(X'U - nh^2\mu_2\beta)] \\
&= E[U'XD'X'XDX'U + (nh^2\mu_2)^2\beta'D'X'XD\beta] \\
&= \sigma^2 tr[XD'X'XDX'] + (nh^2\mu_2)^2\beta'D'X'XD\beta.
\end{aligned}$$

Following the same logic in the proof of Theorem 2, the unbiased estimator of $(nh^2\mu_2)^2\beta'D'X'XD\beta$ is $(nh^2\mu_2)^2[\hat{\beta}'D'X'XD\hat{\beta} - s^2 tr(D'X'XD(X'X)^{-1})]$.

Thus, the unbiased estimator of $MSE(\tilde{\mu}(h))$ is $s^2 tr(XD'X')^2 + (nh^2\mu_2)^2[\hat{\beta}'D'X'XD\hat{\beta} - s^2 trD'X'XD(X'X)^{-1}]$. ■

A.2 Derivations in Chapter 3

Proof of (3.38)

Proof. Observe that

$$\begin{aligned}
\tilde{R}_1(h) &= (\hat{\mu}(h) - Y)'(\hat{\mu}(h) - Y) + 2\hat{\sigma}^2 tr(P(h)) \qquad (A.2) \\
&= L(h) + U'U - 2\mu'P(h)U + 2\mu'U + 2\hat{\sigma}^2 tr(P(h))
\end{aligned}$$

and

$$\begin{aligned} R_1(h) &= (P(h)\mu - \mu)'(P(h)\mu - \mu) + \sigma^2 tr(P^2(h)) \\ &= L(h) - U'P^2(h)U - 2(P(h)\mu - \mu)'P(h)U + \sigma^2 tr(P^2(h)). \end{aligned}$$

Hence to prove (3.38), it suffices to show that

$$\sup_{h\in H} \frac{|\mu' P(h)U|}{R(h)} = o_p(1), \tag{A.3}$$

$$\sup_{h\in H} \frac{\left|\hat{\sigma}^2 tr(P(h))\right|}{R(h)} = o_p(1), \tag{A.4}$$

$$\sup_{h\in H} \frac{\left|U'P^2(h)U\right|}{R(h)} = o_p(1), \tag{A.5}$$

$$\sup_{h\in H} \frac{|(P(h)\mu - \mu)'P(h)U|}{R(h)} = o_p(1) \tag{A.6}$$

and

$$\sup_{h\in H} \frac{\left|tr(P^2(h))\right|}{R(h)} = o_p(1). \tag{A.7}$$

Let $\lambda(A)$ be the largest eigenvalue of the matrix A. From condition (3.37) and the following formulae:

$$\begin{aligned} \sup_{h\in H} \lambda(P(h)) &\le \lambda(X(X'X)^{-1}X') = 1, \\ \sup_{h\in H} tr(P(h)) &\le tr(X(X'X)^{-1}X') = q, \\ \sup_{h\in H} U'P(h)U &\le U'X(X'X)^{-1}X'U, \\ (\mu'P(h)U)^2 &\le \mu'\mu U'P^2(h)U \le \mu'\mu\lambda(P(h))U'P(h)U \end{aligned}$$

and

$$((P(h)\mu-\mu)'P(h)U)^2 \leq (P(h)\mu-\mu)'(P(h)\mu-\mu)U'P^2(h)U \leq R(h)U'P(h)U,$$

we need only to show that

$$U'X(X'X)^{-1}X'U = O_p(1), \tag{A.8}$$

and

$$\hat{\sigma}^2 = O_p(1). \tag{A.9}$$

Equations (A.8) and (A.9) are implied by condition (3.36). The proof of (3.38) thus follows. If in addition $\{L(\hat{h})-\xi\}\xi^{-1}$ is uniformly integrable, then

$$\frac{R_1(\hat{h})}{\inf_{h\in H} R_1(h)} \rightarrow^p 1$$

follows from Zhang, Zou, Liang and Carroll (2014). ■

Proof of (3.41)

Proof. Observe that, with $\tilde{\beta}(h) = B(h)\hat{\beta}$ and $B(h) = (X'X+D)^{-1}X'X$,

$$\begin{aligned}
\hat{R}(h) &= (B(h)\hat{\beta}-\hat{\beta})'(B(h)\hat{\beta}-\hat{\beta}) + 2\hat{\sigma}^2 tr(B(h)(X'X)^{-1}) \\
&= \tilde{L}(h) + (\hat{\beta}-\beta)'(\hat{\beta}-\beta) - 2\hat{\beta}'B(h)(\hat{\beta}-\beta) + 2\beta'(\hat{\beta}-\beta) + 2\hat{\sigma}^2 tr(B(h)(X'X)^{-1}) \\
&\equiv \tilde{L}(h) + \Xi_1(h) \qquad\qquad (A.10)
\end{aligned}$$

and writing $R(h) = MSE(h)$, we have

$$\begin{aligned} R(h) &= (B(h)\beta - \beta)'(B(h)\beta - \beta) + \sigma^2 tr(B'(h)B(h)(X'X)^{-1}) \\ &= \tilde{L}(h) + (\hat{\beta} - \beta)'B'(h)B(h)(\hat{\beta} - \beta) \\ &\quad -2\hat{\beta}'B'(h)B(h)(\hat{\beta} - \beta) + 2\beta B(h)(\hat{\beta} - \beta) + \hat{\sigma}^2 tr(B'(h)B(h)(X'X)^{-1}) \\ &\equiv \tilde{L}(h) + \Xi_2(h). \end{aligned} \tag{A.11}$$

From the condition (3.39), we have $\hat{\beta} - \beta = O_p(n^{-1/2})$, which, together with the conditions (3.39)-(3.40) leads to

$$\sup_{h \in H} \frac{|\Xi_1(h)|}{R(h)} = o_p(1), \tag{A.12}$$

$$\sup_{h \in H} \frac{|\Xi_2(h)|}{R(h)} = o_p(1). \tag{A.13}$$

Hence, we can obtain (3.41). ■

A.3 Derivations in Chapter 5

Condition A1: The kernel $K(\cdot)$ is product kernel such that $K(x) = \prod_{i=1}^{p} k(x_i)$ where $k(\cdot)$ is a univariate symmetric kernel with compact support $\aleph$ such that (i) $\int k(x_i)dx_i = 1$; (ii) $\int x_i k(x_i)dx_i = 0$; (iii) $\int x_i^2 k(x_i)dx_i = \sigma_k^2$; (iv) for all $x_i, x_i' \in \aleph$ we have $|k(x_i) - k(x_i')| \leq C_k |u - v|$, $C_k \in [0, \infty)$.

Condition A2: (i) $f_i(x, \theta_0)$ is the marginal density of x_i evaluated at x, with $f_i(x, \theta_0) < c$ for all i; (ii) $\bar{f}(x) = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} f_i(x, \theta_0)$, and $0 < \bar{f}(x) < \infty$; (iii)

$f_i(x, \theta_0)$ is differentiable, and $\left|f_i^{(1)}(x, \theta_0)\right| < c$; (iv) $|f(x_i, \theta_0) - f(x_i', \theta_0)| \leq c\,|x - x'|$ for all x, x', and θ_0 denotes the true parameters.

Condition A3: $m^{\alpha}(x) < c$ for all x and $\alpha = 1, 2$, $m^{\alpha}(x)$ is the α-th order derivative of $m(x)$ evaluated at x.

Condition A4: As $n \to \infty$, $h \to 0$, $nh^{q+2} \to \infty$ and $nh^{q+6} \to 0$.

Proof of Theorem 1

Proof. Provided any consistent estimator of H, $\hat{H} = H + o_p(1)$ in the first stage, following MY, we can show that $\hat{\delta}_{2S}(x)$ is asymptotically equivalent to the infeasible estimator $\hat{\delta}_{2S}(x) = (Z'(x)K(x)Z(x))^{-1}Z'(x)K(x)y^*$. Thus we have

$$\begin{aligned}\hat{\delta}_{2S}(x) &= (Z'(x)K(x)Z(x))^{-1}Z'(x)K(x)y^* \\ &= (Z'(x)K(x)Z(x))^{-1}Z'(x)K(x)(m+u)\end{aligned}$$

where $Eu = 0$ and $Euu' = I$.

Do a Taylor expansion of m around point x and this gives us

$$\hat{\delta}_{2S}(x) = \delta(x) + (Z'(x)K(x)Z(x))^{-1}Z'(x)K(x)(B(x) + u) + o_p(h^2)$$

where $B(x)$ is a $n \times 1$ column vector whose ith element is given by $b_{x,i} = \frac{1}{2}(x_i - x)'\frac{\partial^2 m(x)}{\partial x^2}(x_i - x)$ and $\frac{\partial^2 m(x)}{\partial x^2}$ is the $p \times p$ Hessian matrix of $m(x)$.

It follows that

$$\begin{aligned}
& \sqrt{nTh^p} D_h(\hat{\delta}(x) - \delta(x)) && \text{(A.14)} \\
= \; & \sqrt{nTh^p} D_h(Z'(x)K(x)Z(x))^{-1} Z'(x)K(x)B(x) + \\
& \sqrt{nTh^p} D_h(Z'(x)K(x)Z(x))^{-1} Z'(x)K(x)u + o_p(1) \\
= \; & B_{2S}^{panel} + V_{2S}^{panel} + o_p(1)
\end{aligned}$$

where

$$B_{2S} \equiv \sqrt{nh^p} D_h(Z'(x)K(x)Z(x))^{-1} Z'(x)K(x)B(x)$$

and

$$V_{2S} \equiv \sqrt{nh^p} D_h(Z'(x)K(x)Z(x))^{-1} Z'(x)K(x)u.$$

To calculate the asymptotic bias, we follow SUW, set $S_n = n^{-1} D_h^{-1} Z'(x)K(x)Z(x)D_h^{-1}$ and assume that we use the product kernel function, then

$$S_n = n^{-1} \sum_{i=1}^{n} \begin{pmatrix} 1 & \frac{(x_i-x)'}{h} \\ \frac{(x_i-x)}{h} & \frac{(x_i-x)(x_i-x)'}{h^2} \end{pmatrix} K(x_i - x) \to^p \begin{pmatrix} \bar{f} & 0 \\ 0 & \kappa_{21} \bar{f} I_p \end{pmatrix}. \qquad \text{(A.15)}$$

Similarly,

$$\frac{1}{n} D_h^{-1} Z'(x)K(x)B(x) = \frac{1}{n} \begin{pmatrix} \sum_{i=1}^{n} K(x) b_{x,i} \\ \sum_{i=1}^{n} \frac{x_i - x}{h} K(x) b_{x,i} \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\kappa_{21} h^2 \sum_{j=1}^{p} \frac{\partial^2 m(x)}{\partial x_j^2} \\ 0_{p\times 1} \end{pmatrix} + o_p(h^2).$$

Next, $B_{2S} = \sqrt{nh^p} S_n^{-1} \frac{1}{n} D_h^{-1} Z'(x)K(x)B(x) = \begin{pmatrix} \sqrt{nh^p} \frac{\kappa_{21} h^2}{2} \sum_{j=1}^{p} \frac{\partial^2 m(x)}{\partial x_j^2} \\ 0_{p\times 1} \end{pmatrix} + o_p(h^2)$ can be obtained.

By equations (A.14) and (A.15), we get

$$V_{2S} = \sqrt{nh^p} S_n^{-1} \frac{1}{n} D_h^{-1} Z'(x)K(x)u = \frac{1 + o_p(1)}{\bar{f}} \sqrt{n^{-1}h^p} \sum_{i=1}^{n} K(x) \begin{pmatrix} u_i \\ (x_i - x)u_i \end{pmatrix}.$$

Then, conditioning on x, the asymptotic normality can be established by using the central limit theorem since $u's$ are independent and identically distributed with mean 0 and variance 1. $\blacksquare$

Proof of Theorem 2

Proof. From Li and Ullah (1998), it is straight forward to show that obtained from the nonparametric LL estimation, the residuals $\hat{w}$ give the consistent estimators of $\hat{\sigma}_1^2 = \sigma_1^2 + o_p(1)$, $\hat{\sigma}_\alpha^2 = \sigma_\alpha^2 + o_p(1)$ and $\hat{\sigma}_\varepsilon^2 = \sigma_\varepsilon^2 + o_p(1)$. These give $\hat{\Sigma} = \Sigma + o_p(1)$ and thus the the consistency of $\hat{H} = I - \hat{\Sigma}^{-1/2} = H + o_p(1)$ is obtained.

Next we derive the asymptotic bias and variance of $\hat{m}(x)$. From equation (5.16), following MY, we can show that $\hat{\delta}_{2S}(x)$ is asymptotically equivalent to the infeasible estimator $\hat{\delta}_{2S}(x) = (Z'(x)K(x)Z(x))^{-1}Z'(x)K(x)y^*$. Thus we have

$$\begin{aligned}\hat{\delta}_{2S}(x) &= (Z'(x)K(x)Z(x))^{-1}Z'(x)K(x)y^* \\ &= (Z'(x)K(x)Z(x))^{-1}Z'(x)K(x)(m+u)\end{aligned}$$

where $Eu = 0$ and $Euu' = I$.

Do a Taylor expansion of m around point x and this gives us

$$\hat{\delta}_{2S}(x) = \delta(x) + (Z'(x)K(x)Z(x))^{-1}Z'(x)K(x)(B(x) + u) + o_p(h^2)$$

where $B(x)$ is a $nT \times 1$ column vector whose i,tth element is given by $b_{x,it} = \frac{1}{2}(x_{it} - x)'m''(x)(x_{it} - x)$ and $m''(x)$ is the $p \times p$ Hessian matrix of $m(x)$.

It follows that

$$\begin{aligned}
&\sqrt{nTh^p}D_h(\hat{\delta}(x)-\delta(x)) && \text{(A.16)}\\
=\;&\sqrt{nTh^p}D_h(Z'(x)K(x)Z(x))^{-1}Z'(x)K(x)B(x)+\\
&\sqrt{nTh^p}D_h(Z'(x)K(x)Z(x))^{-1}Z'(x)K(x)u+o_p(1)\\
=\;&B_{2S}^{panel}+V_{2S}^{panel}+o_p(1)
\end{aligned}$$

where

$$B_{2S}^{panel}\equiv\sqrt{nTh^p}D_h(Z'(x)K(x)Z(x))^{-1}Z'(x)K(x)B(x)$$

and

$$V_{2S}^{panel}\equiv\sqrt{nTh^p}D_h(Z'(x)K(x)Z(x))^{-1}Z'(x)K(x)u.$$

To calculate the asymptotic bias, set $S_{nT}=(nT)^{-1}D_h^{-1}Z'(x)K(x)Z(x)D_h^{-1}$, and we have

$$\begin{aligned}
S_{nT} \;&=\; (nT)^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T}\begin{pmatrix}1 & \frac{(x_{it}-x)'}{h}\\ \frac{(x_{it}-x)}{h} & \frac{(x_{it}-x)(x_{it}-x)'}{h^2}\end{pmatrix}K(x_{it}-x) && \text{(A.17)}\\
&\to_p \begin{pmatrix}\frac{1}{T}\sum_{t=1}^{T}f_t(x) & 0\\ 0 & \frac{1}{T}\sum_{t=1}^{T}f_t(x)\kappa_{21}I_p\end{pmatrix}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\frac{1}{nT}D_h^{-1}Z'(x)K(x)B(x) \;&=\; \frac{1}{nT}\begin{pmatrix}\sum_{i=1}^{n}\sum_{t=1}^{T}K(x)b_{x,it}\\ \sum_{i=1}^{n}\sum_{t=1}^{T}\frac{x_{it}-x}{h}K(x)b_{x,it}\end{pmatrix}\\
&=\; \begin{pmatrix}\frac{1}{2T}\sum_{t=1}^{T}f_t(x)\kappa_{21}h^2m''(x)\\ 0_{p\times1}\end{pmatrix}+o_p(h^2).
\end{aligned}$$

Next, $B_{2S}^{panel}=\sqrt{nTh^p}S_{nT}^{-1}\frac{1}{nT}D_h^{-1}Z'(x)K(x)B(x)=\begin{pmatrix}\sqrt{nTh^p}\frac{\kappa_{21}h^2}{2}m''(x)\\ 0_{p\times1}\end{pmatrix}+o_p(h^2)$ can be obtained.

By equations (A.16) and (A.17), we get

$$\begin{aligned} V_{2S}^{panel} &= \sqrt{nTh^p} S_{nT}^{-1} \frac{1}{nT} D_h^{-1} Z'(x) K(x) u \\ &= \frac{1 + o_p(1)}{\frac{1}{T}\sum_{t=1}^{T} f_t(x)} \sqrt{(nT)^{-1} h^p} \sum_{i=1}^{n} \sum_{t=1}^{T} K_x \begin{pmatrix} u_{it} \\ (x_{it} - x) u_{it} \end{pmatrix}. \end{aligned}$$

Therefore, conditioning on x, the asymptotic normality can be established by using the central limit theorem since given t, the $u_l' s$ are independent and identically distributed with mean 0 and variance 1. ■

Bibliography

1. Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. Editor, B.N., Petrov, F., Csaki. *International Symposium on Information Theory*: 267-281.

2. 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.

3. Andrews, D.W.K. (1999) Consistent moment selection procedures for generalized method of moments estimation. *Econometrica* 67: 543-564.

4. Andrews, D.W.K. and Lu, B. (2001) Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models. *Journal of Econometrics* 101: 123-164.

5. Azomahou, T., Laisney, F. and Van, P.N. (2006) Economic development and CO_2 emissions: A nonparametric panel approach. *Journal of Public Economics* 90: 1347-1363.

6. Baltagi, B.H. (2008) *Econometric Analysis of Panel Data.* Wiley: West Sussex, UK.

7. Bates, J.M. and Granger, C.W. (1969) The combination of forecasts. *Operations Research Quarterly* 20: 451-468.

8. Belloni, A. and Chernozhukov, V. (2011) L1-penalized quantile regression in high-dimensional sparse models. *Annals of Statistics* 39: 82-130.

9. Breiman, L. (1996) Heuristics of instability and stabilization in model selection. *Annals of Statistics* 24: 2350-2383.

10. Brock, W.A., Durlauf, S.N. and West, K.D. (2003) Policy evaluation uncertain economic environment. *Brookings Papers on Economic Activity* 2003: 235-301.

11. Buckland, S.T., Burnham, K.P. and Augustin, N.H. (1997) Model selection: An integral part of inference. *Biometrics* 53: 603-618.

12. Bühlmann, P. and Van de Geer, S. (2011) *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer: New York, US.

13. Bühlmann, P. (1999) Efficient and adaptive post-model-selection estimators. *Journal of Statistical Planning and Inference* 79: 1-9.

14. Burnham, K.P. and Anderson, D.R. (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretical Approach*. Springer-Verlag: New York, US.

15. Catoni, O. (1997) The mixture approach to universal model selection. *Technical report, Ecole Normale Superieure.*

16. Campbell, J.Y. and Thompson, S.B. (2008) Predicting excess stock returns out of sample: can anything beat the historical average? *Review of Financial Studies* 21: 1509-1531.

17. Caner, M. (2009) A Lasso type GMM estimator. *Econometric Theory* 25: 270-290.

18. Caner, M. and Fan, M. A near minimax risk bound: Adaptive Lasso with heteroskedastic data in instrumental variable selection. Working Paper. North Carolina State University.

19. Chen, K., Fan, J. and Jin, Z. (2008) Design-adaptive minimax local linear regression for longitudinal/clustered data. *Statistica Sinica* 18: 515-534.

20. Chen, X., Hong, H. and Shum, M. (2007) Nonparametric likelihood ratio model selection tTests between parametric likelihood and moment condition models. *Journal of Econometrics* 141: 109-140.

21. Chen, K. and Jin, Z. (2005) Local polynomial regression analysis for clustered data. *Biometrika* 92: 59-74.

22. Claeskens, G., Croux, C. and van Kerckhoven, J. (2006) Variable selection for logistic regression using a prediction-focused information criterion. *Biometrics* 62: 972-979.

23. Claeskens, G. and Hjort, N.L. (2003) The focused information criterion. *Journal of the American Statistical Association* 98: 900-945.

24. Claeskens, G. and Hjort, N.L. (2008) *Model Selection and Model Averaging.* Cambridge University Press: Cambridge, UK.

25. Clyde, M. and George, E.I. (2004) Model uncertainty. *Statistical Science* 19: 81-94.

26. Craven, P. and Wahba, G. (1979) Smoothing noisy data with spline functions. *Numerische Mathematik* 31: 377-403.

27. Cutler, D. M. and Lleras-Muney, A. (2006) Education and health: evaluating theories and evidence. Working Paper. National Bureau of Economic Research No.12352.

28. Darolles, S., Fan, Y., Florens, J. and Renault, E. (2011) Nonparametric instrumental regression. *Econometrica* 79: 1541-1565.

29. Dinda, S. (2004) Environmental Kuznets curve hypothesis: A survey. *Ecological Economics* 49: 431-455.

30. Draper, D. (1995) Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society* 57: 45-97.

31. Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression. *Annals of Statistics* 32: 407-499.

32. Eubank, R.L. (1999) *Nonparametric Regression and Spline Smoothing.* CRC Press: New York, US.

33. Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96: 1348-1360.

34. Fan, J. and Gijbels, I. (1996) *Nonparametric Estimation of Econometric Functionals.* Champman and Hall: London, UK.

35. Fan, J. and Lv, J. (2010) A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* 20: 101-148.

36. Fan, Y. and Ullah, A. (1999) Asymptotic normality of a combined regression estimator. *Journal of Multivariate Analysis* 71: 191-240.

37. Frank, I.E. and Friedman, J.H. (1993) A statistical view of some chemomtrics regression tools. *Technometrics* 35: 109-135.

38. Fu, W. and Knight, K. (2000) Asymptotics for Lasso-type estimators. *Annals of Statistics* 28: 1356-1378.

39. Garcia, P.E. Instrumental variable estimation and selection with many weak and irrelevant instruments. Working Paper. University of Wisconsin, Madison.

40. Gautier, E. and Tsybakov, A. high-dimensional instrumental variables regression and confidence sets. Working Paper. Centre de Recherche en Economie et Statistique.

41. Geweke, J.F. (2005) *Contemporary Bayesian Econometrics and Statistics.* John Wiley and Sons Inc.: New Jersey, US.

42. Geweke, J.F. (2007) Bayesian model comparison and validation. *American Economic Review Papers and Proceedings* 97: 60-64.

43. Geman, S. and Hwang, C. (1982) Diffusions for global optimization. *SIAM Journal on Control and Optimization* 24: 1031-1043.

44. Greene, W.H. (2011) *Econometric Analysis.* Prentice Hall: New Jersey, US.

45. Greene, W.H. (2004) Distinguishing between heterogeneity and inefficiency: stochastic frontier analysis of the World Health Organization's panel data on national health care systems. *Health Economics* 13: 959-980.

46. Grossman, M. (1972a) On the concept of health capital and the demand for health. *Journal of Political Economy* 80: 223-255.

47. Grossman, M. (1972b) *The Demand for Health: A Theoretical and Empirical Investigation.* Columbia University Press for the NBER: New York, US.

48. Grossman, G. and Krueger, A. (1991) Environmental impacts of a North American free trade agreement. Working Paper. National Bureau of Economic Research No.3914.

49. Grossman, M. and Robert, K. (1997) Effects of education on health. In J.R. Behman and N. Stacey, eds., *The Social Benefits of Education* University of Michigan Press, Ann Arbor, MI: 69-123.

50. Hall, A.R., Inoue, A., Jana, K. and Shin, C. (2007) Information in generalized method of moments estimation and entropy-based moment selection. *Journal of Econometrics* 138: 488-512.

51. Hall, P.G. and Racine, J.S. (2013) Infinite order cross-validated local polynomial regression. Working Paper. Department of Economic, McMaster University.

52. Hannan, E.J. and Quinn, B.G. (1979) The determination of the order of an autoregression. *Journal of the Royal Statistical Society* 41: 190-195.

53. Hansen, B.E. (2007) Notes and comments least squares model averaging. *Econometrica* 75: 1175-1189.

54. Hansen, B.E. (2008). Least squares forecast averaging. *Journal of Econometrics* 146: 342-350.

55. Hansen, B.E. (2014) Nonparametric sieve regression: Least squares, least squares averaging, and Cross-validation. *Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics.*

56. Hansen, B.E. and Racine, J. (2012) Jackknife model averaging. *Journal of Econometrics* 167: 38-46.

57. Härdle, W., Hall, P. and Marron, J.S. (1998) How far are automatically chosen regression smoothing parameters from their optimum? *Journal of American Statistical Association* 83: 86-99.

58. Hemmerle, W.J. and Carey M.B. (1983) Some properties of generalized ridge estimators. *Communications in Statistics: Computation and Simulation* 12: 239-253.

59. Henderson, D.J. and Ullah, A. (2005) A nonparametric random effects estimator. *Economics Letters* 88: 403-407.

60. Henderson, D.J. and Ullah, A. (2012) Nonparametric estimation in a one-way error component model: a Monte Carlo analysis. *Statistical Paradigms: Recent Advances and Reconciliations* (eds: A. Sengupta, T. Samanta, A. Basu), World Scientific Review.

61. Hoerl, A.E. (1962) Application of ridge analysis to regression problems. *Chemical Engineering Progress* 58: 54-59.

62. Hoerl, A.E. and Kennard, R.W. (1970a) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 12: 55-67.

63. Hoerl, A.E. and Kennard, R.W. (1970b) Ridge regression: application to nonorthogonal problems. *Technometrics* 12: 69-82.

Hoerl, A.E., Kennard. R.W. and Baldwin, K.F. (1975) Ridge regression: some simulations. *Communications in Statistics* 4: 105-123.

64. Hjort, N.L. and Claeskens, G. (2003) Frequentist model average estimators. *Journal of the American Statistical Association* 98: 879-899.

65. Hoeting, J.A., Madigan, D., Raftery, A.E. and Volinsky, C.T. (1999) Bayesian model averaging: A tutorial (with discussion). *Statistical Science* 14: 382-417.

66. Huang, L.H. and Chen, J. (2008) Analysis of variance, coefficient of determination and F-test for local polynomial regression. *Annals of Statistics* 36: 2085-2109.

67. Hurvich, C.M. and Tsai, C.L. (1989) Regression and time series model selection in small samples. *Biometrika* 76: 297-307.

68. Hurvich, C., Simonoff, J. and Tsai, C. (1998) Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society* 60: 271-293.

69. Jiang, X. and Tanner, M.A. (2000) On the asymptotic normality of hierarchical mixtures-of-experts for generalized linear models. *IEEE Transactions on Information Theory* 46: 1005-1013.

70. Jin, S., Su, L. and Ullah, A. (2014) Robustify financial time series forecasting with bagging. *Econometric Reviews* 33: 575-615.

71. Jordan, M.I. and Jacobs, R.A. (1994) Hiearchical mixtures of experts and the EM algorithm. *Neural Computation* 6: 181-214.

72. Kabaila, P. (1995) The effect of model selection on confidence regions and prediction regions. *Econometric Theory* 11: 537-549.

73. Kapetanios, G., Labhard, V. and Price, S. (2006) Forecasting using predictive likelihood model averaging. *Economics Letters* 91: 373-379.

74. Kuersteiner, G. and Okui, R. (2010) Constructing optimal instruments by first-stage prediction averaging. *Econometrica* 78: 697-718.

75. Kuha, J. (2004) AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods and Research* 33: 188-229.

76. Leamer, E.E. and Chamberlain, G. (1976) A Bayesian interpretation of pretesting. *Journal of Royal Statistical Society* 13: 85-94.

77. LeBlanc, M. and Tibshirani, R. (1996) Combining estimates in regression and classification. *Journal of American Statistical Association* 91: 1641-1650.

78. Leeb, H. and Pötscher, B.M. (2003) The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations. *Econometric Theory* 19: 100-142.

79. Leeb, H. and Pötscher, B.M. (2006) Can one estimate the conditional distribution of post-model-selection estimators? *Annals of Statistics* 34: 2554-2591.

80. Leung, G. and Barron, A.R. (2006) Information theory and mixing least-squares regressions. *IEEE Transaction on Information Theory* 52: 3396-3410.

81. Li, K.C. (1986) Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. *The Annals of Statistics* 14: 1101-1112.

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

83. Li, Q. and Racine, J.S. (2007) *Nonparametric Econometrics: Theory and Practice.* Princeton University Press: New Jersey, US.

84. Li, Q. and Racine, J. (2001) Empirical applications of smoothing categorical variables. Working Paper. Department of Economic, McMaster University.

85. Li, Q. and Ullah, A. (1998) Estimating partially linear panel data models with one-way error components. *Econometric Reviews* 17: 145-166.

86. Liang, H., Zou, G., Wan, A.T.K. and Zhang, X. (2011) Optimal weight choice for frequentist model average estimators. *Journal of the American Statistical Association* 106: 1053-1066.

87. Liao, Z. (2013) Adaptive GMM shrinkage estimation with consistent moment selection. *Econometric Theory* FirstView: 1-48.

88. Lin, X. and Carroll, R.J. (2000) Nonparametric function estimation for clustered data when the predictor is measured without/with error. *Journal of the American Statistical Association* 95: 520-534.

89. Lin, X. and Carroll, R.J. (2006) Semiparametric estimation in general repeated measures problems. *Journal of Royal Statistical Society* 68: 69-88.

90. Lu, X. and Su, L. (2013) Jackknife model averaging for quantile regressions. Working Paper. School of Economics, Singapore Management University.

91. Maasoumi, E. (1993) A compendium to information theory in economics and econometrics. *Econometric Reviews* 12: 137-181.

92. Maddala, G.S. (1988) *Introduction to Econometrics.* Macmillan: New York and London.

93. Magnus, J.R., Powell, O. and Prufer, P. (2010) A comparison of two model averaging techniques with an application to growth empirics. *Journal of Econometrics* 154: 139-153.

94. Magnus, J.R., Wan, A.T.K. and Zhang, X. (2011) Weighted average least squares estimation with nonspherical disturbances and an application to the Hong Kong housing market. *Computational Statistics and Data Analysis* 55: 1331-1341.

95. Mallows, C.L. (1973) Some comments on C_p. *Technometrics* 15: 661-675.

96. Malenbaum, W. (1978) World demand for raw materials in 1985 & 2000. McGraw-Hill: New York, US.

97. Martins-Filho, C. and Yao F. (2009) Nonparametric regression estimation with general parametric error covariance. *Journal of Multivariate Analysis* 100: 309-333.

98. Merhav, N. and Feder, M. (1998) Universal prediction. *IEEE Transaction on Information Theory* 44: 2124-2147.

99. Nadaraya, E.A. (1964) Some new estimates for distribution functions. *Theory of Probability and Its Applications* 9: 497-500.

100. Newey, W.K. (1997) Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics* 79: 147-168.

101. Nigatu, G. (2012) Economic growth and PM10 pollution: A nonparametric environmental Kuznets curve. Working Paper, Department of Economics, University of California, Riverside.

102. Nishi, R. (1984) Asymptotic properties of criteria for selection of variables in multiple regression. *Annals of Statistics* 12: 758-765.

103. Olkin, I. and Speigelman, C.H. (1987) A semiparametric approach to density estimation. *Journal of the American Statistical Association* 82: 858-865.

104. Pagan, A. and Ullah, A. (1999) *Nonparametric Econometrics.* Cambridge University Press: Cambridge, UK.

105. Pötscher, B.M. (1991) Effects of model selection on inference. *Econometric Theory* 7: 163-185.

106. Racine, J. (2000) Consistent cross-validatory model-selection for dependent data: Hv-block cross-validation. *Journal of Econometrics* 99: 39-61.

107. Racine, J. and Li, Q. (2004) Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics* 119: 99-130.

108. Robinson, P.M. (1988) Root-N-consistent semiparametric regression. *Econometrica* 56: 931-954.

109. Rosenzweig, M.R. and Schultz, T.P. (1983a) Consumer demand and household production: the relationship between fertility and child mortality. *American Economic Review* 73: 38-42.

110. Rosenzweig, M.R. and Schultz, T.P. (1983b) Estimating household production function: heterogeneity the demand for health Inputs, and their effects on birth weight. *Journal of Political Economy* 91: 723-46.

111. Rousson, V. and Gosoniu, N.F. (2007) An R-square coefficient based on final prediction error. *Statistical Methodology* 4: 331-340.

112. Ruckstuhl, A.F., Welsh, A.H. and Carroll, R.J. (2000) Nonparametric function estimation of the relationship between two repeatedly measured variables. *Statistica Sinica* 10: 51-71.

113. Sala-i-Martin, X., Doppelhofer, G. and Miller, R.I. (2004) Determinants of long-term growth: a Bayesian averaging of classical estimates (BACE) approach. *American Economic Review* 94: 813-835.

114. Schennach, S.M. (2007) Instrumental variable estimation of nonlinear errors-in-variables models. *Econometrica* 75: 201-239.

115. Schmidt, P. (1976) *Econometrics.* CRC Press: New York, US.

116. Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics* 6: 461-464.

117. Scott, D.W. (1992) *Multivariate Density Estimation: Theory, Practice and Visualization.* Wiley: New York, US.

118. Scott, D.W. and Terrell C.R. (1987) Biased and unbiased cross-validation in density estimation. *Journal of American Statistical Association* 82: 1131-1146.

119. Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society* 36: 111-147.

120. Stone, M. (1977) An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society* 39: 44-47.

121. Stone, M. (1979) Comments on model selection criteria of Akaike and Schwartz. *Journal of the Royal Statistical Society* 41: 276-278.

122. Srivastava, A.K., Srivastava, V.K. and Ullah, A. (1995) The coefficient of determination and its adjusted version in linear regression models. *Econometric Reviews* 14: 229-240.

123. Su, L. and Ullah, A. (2006) More efficient estimation in nonparametric regression with nonparametric autocorrelated errors. *Economic Theory* 22: 98-126.

124. Su, L. and Ullah, A. (2007) More efficient estimation of nonparametric panel data models with random effects. *Economics Letters* 96: 375-380.

125. Su, L.; Ullah, A. A Nonparametric Goodness-of-fit-based Test for Conditional Heteroskedasticity. *Econometric Theory* **2013**, *29*, 187-212.

126. Su, L, Ullah, A. and Wang, Y. (2013) Nonparametric regression estimation with general parametric error covariance: A more efficient two-step estimator. *Empirical Economics* 45: 1009-1024.

127. Su, L. and Zhang, Y. (2013) Variable selection in nonparametric and semiparametric regression models. *Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*, (eds: A. Ullah, J. Racine, and L. Su.), Oxford University Press: Oxford, UK.

128. Takeuchi, K. (1976) Distribution of information statistics and criteria for adequacy of models. *Mathematical Science* 153: 12-18. In Japanese.

129. Tibshirani, R. (1996) Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical* 58: 267-288.

130. Ullah, A. (1988) Nonparametric estimation of econometric functionals. *The Canadian Journal of Economics* 21: 625-658.

131. Ullah, A. (1996) Entropy, divergence and distance measures with econometric applications. *Journal of Statistical Planning and Inference* 49: 137-162.

132. Ullah, A. and Roy, N. (1998) Nonparametric and semiparametric econometrics of panel data. *Handbook of Applied Economics Statistics* (eds: A. Ullah and D.E.A. Giles) 1: 579-604, Marcel Dekker: New York, US.

133. Ullah, A., Wan, A.T.K., Wang, H., Zhang, X. and Zou, G. (2013) A semiparametric generalized ridge estimator and link with model averaging. Working Paper. Department of Economics, University of California, Riverside.

134. Vinod, H.D. and Ullah, A. (1981) *Recent Advances in Regression Methods.* Marcel Dekker: New York, US.

135. Vinod, H.D., Ullah, A. and Kadiyala, K. (1981a) A family of improved shrinkage factors for the ordinary ridge estimator. *The Economic Studies Quarterly* 32: 164-175.

136. Vinod, H.D., Ullah, A. and Kadiyala, K. (1981b) Evaluation of the mean squared error of certain generalized ridge estimators using confluent hypergeometric functions. *Sankhya* Series B 43: 360-383.

137. Vovk, V.G. (1990) Aggregateing strategies. *Proceedings of the 3rd Annual Workshop on Computational Learning Theory* 56: 371-383.

138. Vovk, V.G. (1998) A game of prediction with expert advice. *Journal of Computer and System Sciences* 56: 153-173.

139. Wan, A.T.K., Zhang, X. and Zou, G. (2010) Least squares model averaging by mallows criterion. *Journal of Econometrics* 156: 277-283.

140. Wan, A.T.K. and Zhang, X. (2009) On the use of model averaging in tourism research. *Annals of Tourism Research* 36: 525-532.

141. Wang, H., Zhang, X. and Zou, G. Frequentist model averaging estimation: A review. *Journal of Systems Science and Complexity* 22: 732-748.

142. Wang, N. (2003) Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika* 90: 43-52.

143. Wang, Y. (2013) 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.

144. Watson, G.S. (1964) Smooth regression analysis. *Sankhya* Series A 26: 359-372.

145. Welch, I. and Goyal, A. (2008) A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies* 21: 1455-1508.

146. Wolpert, D.H. (1992) Stacked generalization. *Neural Networks* 5: 241-259.

147. Wooldridge, J.M. (2003) *Introductory Econometrics: A Modern Approach.* Thompson South-Western: Kentucky, US.

148. Xiao, Z., Linton, O.B., Carroll, R.J. and Mammen, E. (2003) More efficient local polynomial estimation in nonparametric regression with autocorrelated errors. *Journal of the American Statistical Association* 98: 980-992.

149. Yang, Y. (2000) Mixing strategies for density estimation. *Annals of Statistics* 28: 75-87.

150. Yang, Y. (2001) Adaptive regression by mixing. *Journal of the American Statistical Association* 96: 574-586.

151. Yoo, M. (2011) Does increased education lower health care spending? Findings for self-managed health conditions. Working Paper. Department of Economics, Rutgers University.

152. Yao, W. and Li, R. (2013) New local estimation procedure for a non-parametric regression function for longitudinal data. *Journal of Royal Statistical Society* 75: 123-138.

153. Yao, F. and Ullah, A. (2013) A nonparametric R^2 test for the presence of relevant variables. *Journal of Statistical Planning and Inference* 143: 1527-1547.

154. Yuan, Z. and Yang, Y. (2005) Combining linear regression models: when and how? *Journal of the American Statistical Association* 100: 1202-1204.

155. Zapata, H., Paudel, K. and Moss, C. (2008) Functional form of the environmental Kuznets curve. *Advances in Econometrics* 25: 471-493.

156. Zhang, C. (2010) Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* 38: 894-942.

157. Zhang, C., Fan, J. and Yu, T. (2011) Multiple testing via FDRL for large-scale imaging data. *Annals of Statistics* 39: 613-642.

158. Zhang, X. and Liang, H. (2011) Focused information criterion and model averaging for generalized additive partial linear models. *Annals of Statistics* 39: 174-200.

159. Zhang, X., Wan, A.T.K. and Zhou, S.Z. (2012) Focused information criteria, model selection, and model averaging in a Tobit model with a nonzero threshold. *Journal of Business and Economic Statistics* 30: 132-143.

160. Zhang, X., Wan, A.T.K. and Zou, G. (2013) Model averaging by Jackknife criterion in models with dependent data. *Journal of Econometrics* 174: 82-94.

161. Zhang, X., Zou, G., Liang, H. and Carroll, R.J. (2014) Oracle model averaging estimation for sparse high-dimensional data. Chinese Academy of Sciences, under review.

162. Zou, H. (2006) The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* 101: 1418-1429.