NON-NESTED MODEL SELECTION VIA EMPIRICAL LIKELIHOOD

by

Cong Zhao

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Applied Mathematics

Charlotte

2017

Approved by:

____
Dr. Jiancheng Jiang

____
Dr. Yanqing Sun

____
Dr. Weihua Zhou

____
Dr. Moutaz Khouja

ABSTRACT

CONG ZHAO. Non-nested model selection via empirical likelihood. (Under the direction of DR. JIANCHENG JIANG)

In this dissertation we propose an empirical likelihood ratio (ELR) test to conduct non-nested model selection. It allows for heteroscedasticity and works for any two supervised statistical learning methods under mild conditions. We establish asymptotic properties for the ELR test used for model selection between two linear models, between a functional coefficient model and a non-parametric regression model, and between two general supervised statistical learning methods. Simulations demonstrate good finite sample performance of our model selection procedure. A real example illustrates the use of our methodology.

ACKNOWLEDGMENTS

Upon the completion of this thesis I would sincerely gratefully express my thanks to many people. I would like to give my deepest gratitude to my dissertation advisor, Dr. Jiancheng Jiang for his guidance, insights and encouragement throughout my dissertation research process. His attitude to work and to life deeply engraved in my heart and memory.

I would like to thank my committee members, Dr. Yanqing Sun and Dr. Weihua Zhou for their friendly guidance and suggestions. I would like to thank Dr. Moutaz Khouja for agreeing to serve on my doctoral dissertation committee as an outside department committee member.

I would also like to thank Graduate School and Mathematics Department for providing me financial support. I also need to offer special appreciation to Dr. Joel Avrin and Dr. Shaozhong Deng, who served as graduate coordinator during my time in this program.

I'd like to express my sincere gratitude to my parents, Kai Zhao and Zonghua Wang, for their love, patience, generosity and sacrifice.

Finally, I would dedicate my dissertation to my wife Yue Min and my daughter Vivian. You two are my life. Thank you and love you.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1: INTRODUCTION

In this dissertation, we develop an empirical likelihood approach to non-nested model selection via hypotheses testing. Most existing model selection criteria used penalized likelihood or least square approach, e.g. AIC, BIC, LASSO, etc. They are used widely in theory and make great success in practice, but generally cannot be applied to non-nested model selection. Consider, for example, selecting important genes in the non-Hodgkin's lymphoma data in Dave et al. (2004) using the famous Cox's model and the additive hazard model. Each model may give us a different group of important genes, but there is no general tool to judge which model is better. In this example, the two models are non-nested, one does not have likelihoods or model errors for comparison, and hence the AIC or BIC criteria cannot be used. Even if models are nested, one may have difficulty in making a decision on selecting the best model. For example, suppose there are two candidate models, which AIC values equal to 100 and 102, respectively. Then the model with an AIC value of 100 is preferred according to this criterion. However, one cannot conclude that it is definitely better because the AIC values are too close. In other words, one does not have a clear cutoff for the difference of AIC values to judge which model is significantly better. It is natural to develop a test that furnishes a critical values for model selection.

Some existing literatures use hypothesis testing in model selection. Cox(1961, 1962) introduced a likelihood ratio test when one of competing models is correctly specified. White (1982) used the Kullback-Leibler information criterion (KLIC), developed the asymptotic properties of quasi-maximum likelihood estimator (QMLE) when the model is specified or misspecified. An information matrix test for detecting model misspecification within a certain family of models was also introduced in the work. So if the null hypothesis of no misspecification is not rejected, one may have confidence that standard maximum likelihood techniques of estimation and inference are valid. Otherwise, one has an indication that the parameter estimator is inconsistent for the parameters of interest, so that the model specification must be carefully re-examined. However, the test requires knowledge of likelihood and only works for nested models.

Vuong (1989) also used KLIC to measure the closeness of a model to the truth, introduced a likelihood ratio test for the cases where the competing parametric models are non-nested, overlapping or nested and whether both, one, or neither contain the true law generating observations. He showed the asymptotic distribution of the likelihood ratio statistic is a weighted sum of chi-square distribution or a normal distribution depending on whether the distributions in the competing models closest to the truth are observationally identical. The procedure was motivated by the fact that KLIC measures the distance between a given distribution and the true distribution. So if the distance between a specified model and the true distribution is defined as the minimum of the KLIC over the distributions in the model, then the "best" model among a collection of competing models is defined to be the model that is closest to the true distribution. This approach in Vuong (1989) has the desirable property that it coincides with the usual classical testing approach when the models are nested. However, it requires availability of likelihood, so it works only for parametric models.

Some other literature introduced an extension to the likelihood ratio test. Fan, Zhang and Zhang (2001) proposed generalized likelihood ratio (GLR) tests and showed that the Wilks type of results hold for a variety of useful models, including univariate non-parametric model and varying-coefficient model and their extensions. The nonparametric maximum likelihood estimate (MLE) usually does not exist, or not optimal even it does exist. So the idea is to replace MLE by a nonparametric estimate GLR tests. Fan et al. (2001) showed that the GLR tests achieved the optimal rates of convergence and are adaptively optimal by using a simple choice of adaptive smoothing parameter. Inspired by this, Fan and Jiang (2005) developed GLR test for the additive model based on local polynomial fitting and a backfitting algorithm. A bias reducted version of GLR test was introduced, and a conditional bootstrap method for approximating the null distributions was conducted. A choice of optimal bandwidth was also seriously explored. Fan and Jiang (2005) along with Fan et al. (2001) showed the generality of the Wilks phenomenon and enriched the applicability of the GLR tests. However, the GLR tests work only for nested models and require the working models contains the truth.

In this dissertation, we use a more general approach to model selection. It is known that the prediction error (PE) criterion allows us to compare any two supervised statistical learning methods (parametric or nonparametric) under mild conditions. In practice, a statistical learning procedure with a smaller average prediction error (APE) is usually preferred. However, if the APEs are close among competing models? One does not know if the APEs are significant different. To address this problem and perform an accurate model selection, inspired by Owen (1988, 1989, 2001), Zhang and

Gijbels (2003), Fan and Zhang (2004), Xue and Zhu (2006) and Chen and Keilegom (2009), we introduce an empirical likelihood ratio (ELR) test. It is a nonparametric approach without requiring a specific parametric structure or likelihood. Furthermore it works for any two statistical learning procedures for the cases where the competing models are non-nested, overlapping or nested and whether both, one, or neither is misspecified. Because the process of ELR test needs no assumptions on the variance of the error term, this test allows for heteroscedasticity. We establish asymptotics for the proposed test, which provides us an easy-to-use model selection approach.

The rest of this dissertation is organized as follows. In Chapter 2 we consider the ELR model selection method between two parametric models. In Chapter 3 we extend out procedure to nonparametric models. In Chapter 4 we develop the ELR test for two general supervised statistical learning methods. In Chapter 5 we run simulations to evaluate finite sample performance of the proposed approaches. A real example is used to illustrate the use of our method in Chapter 6. And concluding remarks are presented in Chapter 7. Proofs are given in the Appendix.

CHAPTER 2: SELECTION OF PARAMETRIC MODELS

2.1 Empirical Likelihood Ratio

Parametric model selection is heavily studied, and many methodologies were introduced. Most of them are conducted on nested models, in which one model is included by another. Sometimes we are interested in not just two, but a family of parametric models' performances, and a lot of penalized likelihood or least square approaches are very efficient, e.g. AIC, BIC, LASSO, etc. However, none of them could be applied to non-nested model selection problems, in particular when the model is misspecified. Vuong (1989) used the Kuillback-Leibler Information Criterion (KLIC) and developed likelihood ratio tests for nested, non-nested or overlapping parametric models.

However, for nonparametric models, KLIC is not applicable, so we consider a more natural criterion, the prediction error (PE) criterion, to evaluate competing models. This criterion was used for parameters selection among stationary time series e.g. ARMA model in Rissanen (1986), Wei (1992) and Ing (2007), but it only works for a family of nested models and no test was used in the procedure. For any statistical learning method, we can use PE as a measurement of performance. Intuitively, one generally prefers a model with smaller average prediction error (APE). Now we need a technique to carry out the PE criterion model selection.

Empirical likelihood (EL) is a nonparametric technique for constructing confidence intervals and hypothesis test. It was introduced by Owen (1988), and the properties of EL in i.i.d. settings were described in Owen (1988, 1989, 1990), Hall (1990) and DiCiccio, Hall and Romano (1991), its properties on semi-parametric and nonparametric settings were studied in Zhang and Gijbels (2003), Fan and Zhang (2004), Xue and Zhu (2006) and Chen and Keilegom (2009). The book written by Owen (2001) made a great summary of the applications and possible extensions of EL. EL is an ideal platform to perform our PE criterion model selection. It is a nonparametric approach so we can apply it to any statistical learning method. To use EL, one must specify the estimating equations for the parameter of interest, but need not specify explicitly how to construct standard errors for them. The latter property saves us from the sensitivity of estimating some variability of a quantity like σ^2, and opens up a chance to study the cases of heteroscedasticity or asymmetric errors.

In this chapter, we only construct and study the asymptotic properties of a empirical likelihood ratio (ELR) test for a model selection between two linear models. Since linear model is the very basic and fundamental parametric model, and most of the parametric models are generalized from linear model. So the asymptotic properties of ELR on linear models can be easily extended to other parametric models.

Let's first consider a very simple linear models selection, to determine the response variable follows a linear model with predictor either X or Z. So we have the two models,

$$Y_i = Z_i^T \alpha + \varepsilon_i \tag{2.1}$$

and

$$Y_i = X_i^T \beta + \varepsilon_i \tag{2.2}$$

Choose one positive integers m, for example, $m = \lfloor 0.9n \rfloor$. We divide the data into 2 subseries according to the time order, with the first subserie having m observations and the second having $n - m$ observations. Train the models with the 1st subserie and compute the prediction errors (PE) for the second subserie.

So mathematically we calculate the average prediction error (APE) in the two models as

$$APE_1 = \frac{1}{n-m} \sum_{j=m+1}^{n} [Y_j - Z_j^T \hat{\alpha}]^2,$$

$$APE_2 = \frac{1}{n-m} \sum_{j=m+1}^{n} [Y_j - X_j^T \hat{\beta}]^2,$$

where $\hat{\alpha}$ and $\hat{\beta}$ are the corresponding least square estimators.

Let $\hat{\varepsilon}_{1_j} \equiv Y_j - Z_j^T \hat{\alpha}$, $\hat{\varepsilon}_{2_j} \equiv Y_j - X_j^T \hat{\beta}$, $\xi_j = \hat{\varepsilon}_{1_j}^2 - \hat{\varepsilon}_{2_j}^2$.

Following Owen (1988), we define the log empirical likelihood ratio as

$$R_n = -2 \log \sup \Big\{ \prod_{j=m+1}^{n} (n-m) p_j : \, p_j \geq 0, \sum_{j=m+1}^{n} p_j = 1, \sum_{j=m+1}^{n} p_j \xi_j = 0 \Big\}.$$

Using the Lagrange multiplier technique, we obtain that $p_j = \frac{1}{n-m} \frac{1}{1+\lambda \xi_j}$, where λ satisfies

$$\sum_{j=m+1}^{n} \frac{\xi_j}{1 + \lambda \xi_j} = 0. \tag{2.3}$$

Then the log empirical likelihood ratio becomes

$$R_n = 2 \sum_{j=m+1}^{n} \log(1 + \lambda \xi_j). \tag{2.4}$$

Denote the average prediction error for model (2.1) as APE_1 and the average prediction error for model (2.2) as APE_2, notice that $\bar{\xi} \equiv \frac{1}{n-m}\sum_{j=m+1}^{n}\xi_j = \text{APE}_1$ - APE_2, is the measurement of the performance difference between model (2.1) and model (2.2). The asymptotic properties of R_n is described in the next section.

2.2 Asymptotic results

For a random sample $\{Y_i, Z_i, X_i\}_{i=1}^{n}$, let $U_i = (Y_i, Z_i^T, X_i^T)^T$.

And for any prediction point U, define

$$\xi(U) = \hat{\varepsilon}_1^2(U) - \hat{\varepsilon}_2^2(U) = h(F_m, U),$$

where F_m is the empirical distribution of the training sample $\{U_1, \cdots, U_m\}$. Denote $\mu_\xi = E[\xi(U)], \sigma_\xi^2 = Var[\xi(U)]$.

Given the pair of competing models model (2.1) and model (2.2), it is natural to select the model which has a smaller APE. Notice that even though a model is selected, it may not be correctly specified. So given the above measure of distance, we consider the following hypotheses,

$$H_0 : \mu_\xi \to 0$$

meaning that model (2.1) and model (2.2) are equivalent, against

$$H_a : \mu_\xi \to A \neq 0$$

meaning that one model is sufficiently better than another one.

For model (2.1) and model (2.2) discussed in this chapter, we have the following asymptotic theorem for ELR test statistic R_n.

Theorem 2.1. *If there exists a small* $\delta > 0$ *such that* $E[\xi^{2+\delta}(U)] < \infty$,

(I)Under H_0,

$$R_n \to \chi_1^2.$$

(II)Under H_a,

$$R_n \to +\infty.$$

From Theorem 2.1, given a significant level α, we can conduct a model selection procedure based on the following decision rule.

If $R_n < \chi_1^2(\alpha)$, then we can not reject $H_0 : \mu_\xi = 0$, we say the two models are asymptotically equivalent.

If $R_n > \chi_1^2(\alpha)$, so one model is sufficiently better than another one.

Further more, if $R_n > \chi_1^2(\alpha)$ and $APE_1 - APE_2 < 0$, model (2.1) is better than model (2.2).

If $R_n > \chi_1^2(\alpha)$ and $APE_1 - APE_2 > 0$, model (2.2) is better than model (2.1).

To give an insight of Lemma A.1 and Theorem 2.1, let's consider that the true value of Y is generated from a mixture of model (2.1) and model (2.2). i.e.

$$Y_i = (1-\theta)\alpha Z_i + \theta\beta X_i + \varepsilon_i,$$

where $0 \leqslant \theta \leqslant 1$, ε_i is independent of Z_i and ε_i is independent of X_i.

With some simple algebra, we can show that $\mu_{XY} = (1-\theta)\alpha\mu_{XZ} + \theta\beta\Sigma_X$, $\mu_{ZY} = (1-\theta)\alpha\Sigma_Z + \theta\beta\mu_{XZ}$. So

$$\begin{aligned}\mu_\xi &= \Sigma_X^{-1}\mu_{XY}^2 - \Sigma_Z^{-1}\mu_{ZY}^2 + O(\frac{1}{m}) \\ &= -(1-\theta)^2\alpha^2\Sigma_Z(1-\Sigma_Z^{-1}\Sigma_X^{-1}\mu_{XZ}^2) + \theta^2\beta^2\Sigma_X(1-\Sigma_Z^{-1}\Sigma_X^{-1}\mu_{XZ}^2) + O(\frac{1}{m}),\end{aligned}$$

where $\mu_{XY} = E[X_i Y_i]$, $\mu_{ZY} = E[Z_i Y_i]$, $\Sigma_X = E[X_i^2]$, $\Sigma_Z = E[Z_i^2]$.

Furthermore, for two centered variables X and Z, let ρ be the correlation coefficient between X and Z, $\rho = \frac{Cov(X,Z)}{\sqrt{Var(X)Var(Z)}} = \frac{\mu_{XZ}}{\sqrt{\Sigma_X \Sigma_Z}}$. Apply the above result,

$$\mu_\xi = -(1-\theta)^2\alpha^2\Sigma_Z(1-\rho^2) + \theta^2\beta^2\Sigma_X(1-\rho^2) + O(\frac{1}{m}) \tag{2.5}$$

If θ decreases within (0, 1), intuitively, Y is more affected by model (2.1), so model (2.1) should make a better prediction thus model (2.1) is preferred. Mathematically, if θ decreases, from equation (2.5), μ_ξ decreases and eventually μ_ξ will be less than 0. So Applying Theorem 2.1, $R_n \to +\infty$, and $APE_1 - APE_2 \to \mu_\xi < 0$. we are confident to say model (2.1) is better than model (2.2). If θ increases within (0, 1), follow the same argument, intuitively and mathematically, model (2.2) is better than model (2.1).

We can also see from equation (2.3), if ρ is closer to 0, which means that X and Z are less correlated, then μ_ξ is more distinguished from 0. Using Theorem 2.1, model (2.1) and model (2.2) are more likely to be distinguished. But if ρ is closer to 1 or -1, which means that X and Z are more correlated, then μ_ξ gets closer to 0. Again using Theorem 2.1, we might not be able to distinguish model (2.1) and model (2.2).

Numerical studies of this example are in Chapter 5, Example 1 through Example 4.

CHAPTER 3: SELECTION OF NONPARAMETRIC MODELS

3.1 Empirical Likelihood Ratio

In this chapter we extend the ELR test to selection of nonparametric models. Especially we are interested in the selection between the functional coefficient model and the nonparametric regression model inspired by the folllowing situation.

Cointegration relationship widely exists in the financial area, for instance Consumption and Income, Interest Rate and Money Demand. The definition of cointegration is that if two or more time series are individually integrated (in the time series sense) but some linear combination of them has a lower order of integration, then the series are said to be cointegrated. So studying the cointegration relationship is critical in financial area for the reason that we can estimate the relationship of non-stationary financial assets, and once the cointegrating relationship is identified, it can be used in a form of error-correction. Two popular families of models used to estimate the cointegration are functional coefficient model and nonparametric regression model.

Suppose we have a random sample $\{Y_i, X_i\}_{i=1}^n$ and have found that there exists some in-sample significant evidence of "nonlinearity" between Y_i and X_i. We are interested in further investigating whether the documented "nonlinearitry" is the true nonlinearity under the stationarity condition or the documented "nonlinearity" is due to the functional coefficients in a linear regression model.

For this reason we conduct a model selection between the functional coefficient model,

$$Y_i = \beta(Z_i)X_i + \varepsilon_i, \tag{3.1}$$

and the non-parametric regression model,

$$Y_i = m(X_i) + \varepsilon_i, \tag{3.2}$$

where $\{X_i\}$ and $\{Z_i\}$ are stationary.

Following Jiang (2014), we introduce the following training and testing procedure. Choose two positive integers l and q such that $n > lq$, for example, $l = \lfloor 0.1n \rfloor$ and $q = 4$. Divide the data into $q+1$ subseries according to the time order, with the first subseries having $m \equiv n - ql$ observations and each of the remaining q sebseries having l observations. Compute the one-step prediction errors for each of the remaining q subseries using the estimated model, based on the historical data.

Mathematically we define the average prediction error (APE) in model (3.1) and (3.2), respectively, by

$$APE_3 = \frac{1}{ql}\sum_{k=1}^{q}\sum_{j=n-kl+1}^{n-kl+l}(Y_j - \hat{\beta}_k(Z_j)X_j)^2,$$

$$APE_4 = \frac{1}{ql}\sum_{k=1}^{q}\sum_{j=n-kl+1}^{n-kl+l}(Y_j - \hat{m}_k(X_j))^2,$$

where $\hat{\beta}_k(z) = [\frac{1}{n-kl}\sum_{i=1}^{n-kl} X_i^2 K_{h_1}(Z_i - z)]^{-1}[\frac{1}{n-kl}\sum_{i=1}^{n-kl} X_i Y_i K_{h_1}(Z_i - z)]$ is the local linear estimator,

$\hat{m_k}(x) = [\frac{1}{n-kl}\sum_{i=1}^{n-kl} J_{h_2}(X_i - x)]^{-1}[\frac{1}{n-kl}\sum_{i=1}^{n-kl} J_{h_2}(X_i - x)Y_i]$ is the Nadaraya-Watson kernel estimator.

Moreover, $K_{h_1}(\cdot) = \frac{1}{h_1}K(\frac{\cdot}{h_1})$ and $J_{h_2}(\cdot) = \frac{1}{h_2}J(\frac{\cdot}{h_2})$ are kernel functions in model (3.1) and model (3.2) respectively, h_1 and h_2 are the corresponding bandwidths.

Let $\hat{\varepsilon}_{3,k,j} \equiv Y_j - \hat{\beta}_k(Z_j)X_j$, $\hat{\varepsilon}_{4,k,j} \equiv Y_j - \hat{m}_k(X_j)$, $\xi_{k,j} = \hat{\varepsilon}^2_{3,k,j} - \hat{\varepsilon}^2_{4,k,j}$.

Following Owen (1988), we define the log empirical likelihood ratio as

$$R_n = -2\log\sup\{\prod_{k=1}^{q}\prod_{j=n-kl+1}^{n-kl+l}(qlp_{k,j}) :\ p_{k,j} \geqslant 0, \sum_k\sum_j p_{k,j} = 1, \sum_k\sum_j p_{k,j}\xi_{k,j} = 0\},$$

Using the Lagrange multiplier technique, we obtain that $p_{k,j} = \frac{1}{ql}\frac{1}{1+\lambda\xi_{k,j}}$,

where λ satisfies

$$\sum_{k=1}^{q}\sum_{j=n-kl+1}^{n-kl+l}\frac{\xi_{k,j}}{1+\lambda\xi_{k,j}} = 0. \tag{3.3}$$

Then the log empirical likelihood ratio becomes

$$R_n = 2\sum_{k=1}^{q}\sum_{j=n-kl+1}^{n-kl+l}\log(1+\lambda\xi_{k,j}). \tag{3.4}$$

Notice that $\bar{\xi} \equiv \frac{1}{ql}\sum_{k=1}^{q}\sum_{j=n-kl+1}^{n-kl+l}\xi_{k,j}$ = APE_3 - APE_4, is the measurement of the performance difference between model (3.1) and model (3.2). The asymptotic properties of R_n are described in the next section.

3.2 Asymptotic results

The asymptotic results in this section are based on i.i.d. data, but they can be easily extended to stationary time series data. Similar to Chapter 2, for a random sample $\{Y_i, Z_i, X_i\}_{i=1}^n$, let $U_i = (Y_i, Z_i, X_i)^T$. For any prediction point U, define

$$\xi(U) = \hat{\varepsilon}^2_3(U) - \hat{\varepsilon}^2_4(U) = h(F_m, U),$$

where F_m is the empirical distribution of the training sample $\{U_1, \cdots, U_m\}$.

Denote $\mu_\xi = E[\xi(U)]$ and $\sigma_\xi^2 = Var[\xi(U)]$.

Given the competing models model (3.1) and model (3.2), it is natural to select the model with a smaller APE. Notice that even though a model is selected, it may not be correctly specified. Therefore, given the above measure of distance, we consider the following hypotheses,

$$H_0 : \mu_\xi \to 0$$

meaning that model (3.1) and model (3.2) are equivalent, against

$$H_a : \mu_\xi \to A \neq 0$$

meaning that one model is sufficiently better than another one.

Similar to parametric model selection, we have the following asymptotic results of the ELR test under different hypotheses.

Theorem 3.1. *If there exists a small $\delta > 0$ such that $E[\xi^{2+\delta}(U)] < \infty$,*

(I)Under H_0,

$$R_n \to \chi_1^2.$$

(II)Under H_a,

$$R_n \to +\infty.$$

From Theorem 3.1, given a significant level α, we can conduct a model selection procedure based on the following decision rule.

If $R_n < \chi_1^2(\alpha)$, then we can not reject $H_0 : \mu_\xi = 0$, we say the two models are asymptotically equivalent.

If $R_n > \chi_1^2(\alpha)$, so one model is sufficiently better than another one.

Further more, if $R_n > \chi_1^2(\alpha)$ and $APE_3 - APE_4 < 0$, model (3.1) is better than model (3.2).

If $R_n > \chi_1^2(\alpha)$ and $APE_3 - APE_4 > 0$, model (3.2) is better than model (3.1).

To give an illustration of the model selection procedure mentioned above, we consider the following three examples.

Example (i) Consider, if

$$Y_i = sin(\pi Z_i) + cos(2\pi Z_i)X_i + \varepsilon_i,$$

where $Z_i \sim U[0,1]$, $X_i \sim U[0,2]$.

Since Y is generated from model (3.1), it should be preferred. Simple calculation leads to $\mu_\xi = -\frac{8}{\pi^2}$. Applying Theorem 3.1, we get that $R_n \to +\infty$, and $APE_3 - APE_4 \to \mu_\xi < 0$. Hence, model (3.1) is better than model (3.2).

Example (ii) Suppose

$$Y_i = exp(X_i)cos(X_i) + \varepsilon_i,$$

where $X_i \sim U[0,2]$.

Intuitively, Y is generated from model (3.2), so it should be preferred. By simple calculation, we get $\mu_\xi \approx 0.234$. Applying Theorem 3.1, we obtain that $R_n \to +\infty$, and $APE_3 - APE_4 \to \mu_\xi > 0$. Thus, model (3.2) is better than model (3.1).

Example (iii) If Y is generated from a model which is included in both model (3.1) and model (3.2), that is a linear model with constant coefficients. For example,

$$Y_i = X_i + \varepsilon_i.$$

Intuitively, Y can be regarded as being generated from either Model 3 or Model 4, so it's hard for us to distinguish the two models. By simple calculation, we get that $\mu_\xi = 0$. Applying Theorem 3.1, $R_n \to \chi_1^2$, it's very likely that model (3.1) and model (3.2) can not be distinguished.

A numeric study of these examples mentioned above is included in Chapter 5 Example 5.

CHAPTER 4: EMPIRICAL LIKELIHOOD RATIO TEST

4.1 Basic Framework

In this chapter we discuss the ELR test on model selection for two general statistical learning procedures. This framework allows for comparison of parametric or nonparametric, nested, non-nested or overlapping statistical models with response Y. It benefits from the fact that the prediction error (PE) criterion can be applied to any supervised models.

Suppose we have two supervised statistical learning models M_1, M_2.

Define $\hat{\varepsilon}_{k_j} = Y_j - \hat{Y}_{k_j}$,

where $\hat{Y}_{k_j}$ is the predicted value of Y_j under model M_k, $k = 1, 2$.

We use the same simple process as in Chapter 2 to calculate prediction errors and more importantly to illustrate the process. For a random sample $\{U_i\}_{i=1}^{n}$, where U is the vector of all predictive variables and response variable, choose one positive integers m, for example, $m = \lfloor 0.9n \rfloor$. Divide the data into 2 subseries according to the time order, with the first subserie having m observations and the second having $n - m$ observations. Train the models with the 1st part and compute the prediction errors (PE) for the second part.

We calculate APEs for model M_1 and M_2 as in section 2.

$$APE_1 = \frac{1}{n-m} \sum_{j=m+1}^{n} \hat{\varepsilon}_{1_j}^2,$$

$$APE_2 = \frac{1}{n-m}\sum_{j=m+1}^{n} \hat{\varepsilon}_{2_j}^2.$$

At a prediction point U, define $\xi(U) = \hat{\varepsilon}_1^2(U) - \hat{\varepsilon}_2^2(U)$, $\mu_\xi = E[\xi(U)]$, and $\sigma_\xi^2 = Var[\xi(U)]$. We use μ_ξ as the measurement of performance difference of the two learning procedures, since $\mu_\xi = E[APE_1 - APE_2]$.

Similar to Chapter 2 and Chapter 3, we consider the following hypotheses and definitions,

let $\tau_n = n^{\frac{1}{2}}\mu_\xi/\sigma_\xi$,

$$H_0 : \tau_n \to 0$$

meaning that model M_1 and model M_2 are equivalent, against

$$H_{a1} : \tau_n \to \tau \neq 0$$

or

$$H_{a2} : \tau_n \to \pm\infty$$

meaning that one model is sufficiently better than another one.

Following Owen (1988), we define the log empirical likelihood ratio as

$$R_n = -2\log\sup\Big\{\prod_{j=m+1}^{n}(n-m)p_j : p_j \geq 0, \sum_{j=m+1}^{n} p_j = 1, \sum_{j=m+1}^{n} p_j\xi_j = 0\Big\},$$

where $\xi_j = \xi(U_j)$. Using the Lagrange multiplier technique, we obtain that $p_j = \frac{1}{n-m}\frac{1}{1+\lambda\xi_j}$, where λ satisfies

$$\sum_{j=m+1}^{n} \frac{\xi_j}{1+\lambda\xi_j} = 0. \tag{4.1}$$

Then the log empirical likelihood ratio becomes

$$R_n = 2 \sum_{j=m+1}^{n} \log(1 + \lambda \xi_j). \tag{4.2}$$

4.2 Asymptotic results

For any prediction point U, define

$$\xi(U) = \hat{\varepsilon}_1^2(U) - \hat{\varepsilon}_2^2(U) = h(F_m, U)$$

where F_m is the empirical distribution of $\{U_1, \cdots, U_m\}$.

Theorem 4.1. *If the following conditions hold,*

(1) $\exists \delta > 0$ *such that* $E[\xi^{2+\delta}(U)] < \infty$,

(2) $Var\{E[h(F_m, U)|F_m]\} = o(\frac{1}{n})$,

(3) $E\{Var[h(F_m, U)|U]\} = o(1)$.

Then

$$R_n \overset{a}{\sim} \chi_1^2(\tau_n^2),$$

where τ_n^2 *is the noncentrality parameter.*

Notice that following Theorem 4.1, under $H_0 : \tau_n \to 0$, $R_n \to \chi_1^2$; under $H_{a1} : \tau_n \to \tau \neq 0$, R_n asymptotically follows a noncentral chi-square distribution $\chi_1^2(\tau_n^2)$; and under $H_{a2} : \tau_n \to \pm\infty$, $R_n \to \infty$.

From Theorem 4.1, given a significant level α, we can conduct a model selection procedure based on the following decision rule.

If $R_n < \chi_1^2(\alpha)$, then we can not reject $H_0 : \tau_n \to 0$, we say the two models are asymptotically equivalent.

If $R_n > \chi_1^2(\alpha)$, so one model is sufficiently better than another one.

Further more, if $R_n > \chi_1^2(\alpha)$ and $APE_1 - APE_2 < 0$, model M_1 is better than model M_2.

If $R_n > \chi_1^2(\alpha)$ and $APE_1 - APE_2 > 0$, model M_2 is better than model M_1.

To give an insight of Theorem 4.1 and ELR test, in the proof of Theorem 4.1 in Appendix, under H_0, the asymptotic leading term of R_n is $(APE_1 - APE_2)^2$ divided by $Var[APE_1 - APE_2]$. So if given the true variance of difference of APEs in the two models, ELR test-statistics is asymptotically equivalent to standardized difference of APEs. And notice that $Var[APE_1 - APE_2]$ is never known in reality, an estimate of $Var[APE_1 - APE_2]$ must be used to carry out a standardized difference of APEs test. But in ELR test, the structure of empirical likelihood needs no estimation of the variance, it saves us from this trouble.

The ELR test procedure for two general statistical learning methods is the same as what we described in Chapter 2 for two linear models and Chapter 3 for a functional coefficient model and a nonparametric regression model. The differences are the necessary of the three technical conditions.

To help people better understanding these conditions. Condition (1) implies the existence of more than fourth moment of prediction errors. Recall that the average of $\xi(U)$ is the difference of the average prediction error of the two competing models, when $E[\xi^2(U)] = \infty$, it means the two models are too much different. So we don't need to consider this model selection if Condition (1) is not satisfied. Condition (2) and (3) are technical conditions to prove Theorem 4.1. For a lot of statistical learning models, Condition (2) and (3) are satisfied. According to Lemma A.1 and Lemma A.2, linear model, functional coefficient model and nonparametric regression model satisfy Condition (2) and Condition (3). And it's not hard to check that many generalized linear models like logistic regression and many other kernel regression satisfy Condition (2) and (3). This gives a wide application of Theorem 4.1. There might be more statistical learning procedures satisfying the conditions in Theorem 4.1 and it needs future work to discover.

CHAPTER 5: SIMULATIONS

5.1 Two Linear Models

The asymptotic result from Chapter 2 through Chapter 4 are based on i.i.d. data, but our results can be extended to time series data. In this chapter, all the data we used were stationary time series. To study finite sample behaviours of our ELR test, from Example 1 to Example 4, we use the first 90% data in the time order as training set, and the last 10% as test set.

Example 1:

Consider a model selection between

$$Y_i = \alpha_0 + \alpha_1 Z_i + \varepsilon_i \tag{5.1}$$

and

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i. \tag{5.2}$$

We generate Y_i from model (5.1) with $\alpha_0 = 0$ and $\alpha_1 = 4$,

or model (5.2) with $\beta_0 = 0$ and $\beta_1 = 4$, where

Z_i and $\delta_i \sim AR(0.7), \varepsilon_i \sim N(0, 1)$,

and $X_i = bZ_i + c\delta_i$.

Different values of (b, c) were used to make $\{Z_i\}$ and $\{X_i\}$ have different correlation coefficients ρ but the same variance.

Using the notation in Chapter 2, we define false positive (FP) and false negatives (FN) as

FP={model (5.1) and model (5.2) are equivalent but either model (5.1) or model (5.2) is preferred},

FN_1={model (5.1) is better but not preferred},

FN_2={model (5.2) is better but not preferred}.

We conducted 500 simulations. The sample size n was set as 200, 500, and 1000, we take significant level $\alpha = 5\%$. In the time order, the first 90% of the data were used as training sample, and the last 10% were used as test sample. The simulation results are reported in Table 1.

As the correlation between Z and X increases, the "distance" between model (5.1) and model (5.2) decreases, and hence it gets harder for us to distinguish the two models. However, it's seen from Table 1 that the percentage that we made a wrong specification is still very low, i.e. the ELR testing was carried out very well.

Table 1: ELR for Two Linear Models

	Data from model 5.1 (FN_1)			Data from model 5.2 (FN_2)		
ρ^2	n=200	n=500	n=1000	n=200	n=500	n=1000
0	0	0	0	0	0	0
0.5	0	0	0	0	0	0
0.95	0.084	0.002	0	0.078	0.002	0

Example 2:

We use the same setting as Example 1 but allows for heteroscedasticity in the true model. That is Y_i is generated from $Y_i = 4Z_i + Z_i\varepsilon_i$, or $Y_i = 4X_i + X_i\varepsilon_i$. The simulation result is summarized in Table 2.

From Table 2, we can see that our ELR test works still well with heteroscedasticity, and as sample size increases, the false negative rates decreases.

Table 2: Two Linear Models with Heteroscedasticity

	Data from model 5.1 (FN_1)			Data from model 5.1 (FN_2)		
ρ^2	n=200	n=500	n=1000	n=200	n=500	n=1000
0	0.002	0	0	0.004	0	0
0.5	0.098	0.01	0	0.094	0.012	0
0.95	0.238	0.048	0.014	0.244	0.052	0.014

Example 3: Same setting as Example 1 but Y is generated from a mixture of model (5.1) and model (5.2):

$$Y_i = (1-\theta)Z_i + \theta X_i + \varepsilon_i,$$

where $0 \leqslant \theta \leqslant 1$.

In this set up, as we discussed in Chapter 2, if $\theta = 0.5$, following simple algebra, we can get $\mu_\xi = 0$, then model (5.1) and model (5.2) are equivalent. If $0 \leqslant \theta < 0.5$, it's not hard to see that $\mu_\xi < 0$, so model (5.1) is better than model (5.2). And if $0.5 < \theta \leqslant 1$, following the same argument, we have $\mu_\xi > 0$, so model (5.2) is better than model (5.1).

In this simulation, we used $\theta = 0$, 0.2, 0.5, 0.8 or 1. When $\theta = 0.5$, since the two models are equivalent, H_0 is true, so we only report FP, in Table 3, as sample size increases, FP gets closer to the significant level 0.05. When $\theta = 0$ or 0.2, model (5.1) is better, so we only consider FN_1. And when $\theta = 0.8$ or 1, model (5.2) is better, so we only consider FN_2. From Table 3, as $|\theta|$ approaches 1, the power (Power = 1 - FN) of ELR test gets higher and when sample size is large enough, the power = 1.

Example 4: We now consider a model selection between

$$Y_i = \alpha_0 + \alpha_1 Z_i + \varepsilon_i \tag{5.3}$$

and

$$Y_i = \beta_0 + \beta_1 Z_i + \beta_2 X_i + \varepsilon_i \tag{5.4}$$

Y is generated from model (5.4) with $\beta_0 = 0$, $\beta_1 = 1 - \theta$ and $\beta_2 = \theta$, where $\theta = 0$, 0.2, 0.4, 0.6, 0.8 or 1.

Let Z_i and $X_i \sim AR(0.7)$, Z_i and X_i are independent, $\varepsilon_i \sim N(0, 1)$.

Since model (5.3) is included in model (5.4), this is a model selection between two nested models. For this set up of dependent variable, we only need to consider FP and FN_2 defined in Example 1, because model (5.4) is at least as good as model (5.3) (with $\theta = 0$ meaning two models are equivalent).

Different θ values were used to show the power of this hypothesis testing. For $\theta > 0$, Power = 1 - FN_2. The simulation result is summarized in Table 3.

From Table 4, when $\theta = 0$, model (5.3) is equivalent to model (5.4), as sample size increases, the FP (Type I error) in the simulation gets very close to the significant level $\alpha = 0.05$. And when $\theta > 0$, model (5.4) is better than model (5.3) since the data were generated from model (5.4). As θ increases, the "difference" between model (5.3) and model (5.4) gets larger, so does the power (Power = 1 - FN_2) of the test, and the power goes to 1.

5.2 A Functional Coefficient Model vs A Non-parametric Model

Example 5:

Table 3: Dependent Variable is a Mixture of Two Models

θ	n=200	n=500	n=1000
	FN_1		
0	0.19	0.012	0
0.2	0.424	0.118	0.006
	FP		
0.5	0.106	0.064	0.054
	FN_2		
0.8	0.406	0.13	0.008
1	0.178	0.018	0

Table 4: ELR Type I Error and Power

	FP		
θ	n=200	n=500	n=1000
0	0.098	0.058	0.052
	FN_2		
θ	n=200	n=500	n=1000
0.2	0.852	0.792	0.7
0.4	0.664	0.484	0.232
0.6	0.45	0.208	0.032
0.8	0.354	0.034	0
1	0.2	0.016	0

Now we do a simulation of a model selection between time-varying coefficient model and non-parametric model,

$$Y_i = \beta_0(Z_i) + \beta_1(Z_i)X_i + \varepsilon_i \tag{5.5}$$

$$Y_i = m(X_i) + \varepsilon_i. \tag{5.6}$$

Y is generated from

$$Y_i = Z_i + \cos(10Z_i)X_i + \varepsilon_i,$$

in this setting model (5.5) is better than model (5.6),

$$Y_i = \exp(X_i)\cos(X_i) + \varepsilon_i,$$

in this setting model (5.6) is better than model (5.5), or

$$Y_i = X_i + \varepsilon_i,$$

in this setting the two models are equivalent.

Where $X_i, Z_i \sim AR(0.7)$, X_i and Z_i are independent, $\varepsilon_i \sim N(0, 1)$.

We divide the data into 5 subseries according to the time order, with the first subseries having 60% observations and each of the remaining 4 sebseries having 10% observations. Following the methods introduced in Chapter 3, compute the one-step prediction errors for each of the remaining 4 subseries using model (5.5) and model (5.6), based on the historical data. It means that according to the time order, we use the first 60% observations to predict and calculate the prediction errors on the data lying from 60% to 70% of the whole data set. And use the first 70% observations to predict and calculate the prediction errors on the data lying from 70% to 80% of the whole data set and so on. With the above process, we follow the construction in Chapter 3 to conduct the ELR test.

Follow the notation in Chapter 3, we define false positive (FP) and false negatives (FN) as

FP={model (5.5) and model (5.6) are equivalent but either model (5.5) or model (5.6) is preferred},

FN_1={model (5.5) is better but not preferred},

FN_2={model (5.6) is better but not preferred}.

500 simulations were conducted with different sample size n. Gaussian Kernel were applied to both two models, but bandwidths in the two models were optimized separately in the sense that the optimal bandwidth minimized the average prediction error in its own model. The simulation result is summarized in Table 5.

We could see from Table 5 that FN_1 and FN_2 get lower and eventually equal to 0 as sample sizes increase, at the mean time, the power of this test is very high and equal to 1 when $n = 1000$. The false positive rate gets very close to the significant lever 5% when the sample size is large enough.

Table 5: Time Varying Coefficient Model vs Non-parametric Model

	n=200	n=500	n=1000
FN_1	0.018	0.002	0
FN_2	0.066	0.014	0
FP	0.138	0.078	0.058

CHAPTER 6: A REAL EXAMPLE

We consider a real application of ELR test here. We have a subset of the Coronary Risk-Factor Study (CORIS) baseline survey, carried out in three rural areas of the Western Cape, South Africa (Rousseauw et al., 1983, Hastie, Tibshirani and Friedman 2009). The data can be downloaded from the website

http://statweb.stanford.edu/ tibs/ElemStatLearn/datasets/SAheart.data. The aim of the study was to establish the intensity of ischemic heart disease risk factors in that high-incidence region. The data represent white males between 15 and 64, there are 160 cases and a group of 302 controls. The response variable is the presence or absence of coronary heart disease (**chd**) at the time of the survey. The risk factors considered are systolic blood pressure (**sbp**), total lifetime tobacco usage in kilograms (**tobacco**), low density lipoprotein cholesterol (**ldl**), family history of heart disease (**famhist**), **obesity**, current **alcohol** consumption and **age** at onset.

Since the response variable **chd** is binary, it's natural to fit a logistic regression model by maximum likelihood, giving the results shown in Table 6. A insignificant p-value (greater than 5%) suggests a coefficient can be dropped from the model. Each of these correspond to a test of the null hypothesis that the coefficient is zero, while all the others are not.

We found some surprises in the coefficients in Table 6. Neither systolic blood pressure (**sbp**) nor **obesity** is significant. This confusion is a result of the correlation

Table 6: Logistic Linear Regression Fit to the South African Heart Disease Data

	Coefficient	Z value	P-value
(Intercept)	-4.130	-4.285	0
sbp	0.006	1.023	0.306
tobacco	0.080	3.034	0.002
ldl	0.185	3.219	0.001
famhist	0.939	4.177	0
obesity	-0.035	-1.187	0.235
alcohol	0.001	0.136	0.892
age	0.043	4.181	0

between the set of predictors, since on their own, both **sbp** and **obesity** are significant with positive sign, but with many other correlated variables, they are no longer needed.

Now we need to do some model selection, to find a subset of the variables that are sufficient for explaining their joint effect on the dependent variable **chd**. As suggested in Hastie, Tibshirani and Friedman (2009), one way is to drop the least significant coefficient, and refit the model, this is done repeatedly until no further terms can be dropped from the model. A better but more time consuming strategy is to refit each of the models with one variable removed at a time, and then perform an analysis of deviance to decide which variable to exclude. The residual deviance of a fitted model is minus twice its log-likelihood, and the deviance between two models is the difference of their individual residual deviances. The above two strategies gave the same final model with predictive variables **tabacco**, **ldl**, **famhist** and **age** (model 6.1) as shown in Table 7.

The second model we consider for this example is to explore the nonlinearities in the functions using natural splines. As suggested in Hastie, Tibshirani and Friedman

Table 7: Stepwise Logistic Regression Fit to the South African Heart Disease Data (model 6.1)

	Coefficient	Z value	P-value
(Intercept)	-4.204	-8.437	0
tobacco	0.081	3.163	0.002
ldl	0.168	3.093	0.002
famhist	0.924	4.141	0
age	0.044	4.521	0

(2009), we use four natural spline bases and three interior knots for each variable in the model except for variable **famhist**. Since **famhist** is a two-level factor, it is coded by a simple binary variable, and is associated with a single coefficient in the fit of the model.

We carried out a backward stepwise deletion process, dropping terms from this model while preserving the group structure of each term, rather than dropping one coefficient at a time. The AIC was used to drop terms, in the sense that all the terms remaining in the final model would cause AIC to increase if deleted from the model. The final model (model 6.2) is shown in Table 8. Notice that both **sbp** and **obesity** are included in model (6.2) while they are not in logistic linear model (6.1).

Table 8: Logistic regression model (6.2) after stepwise deletion of natural splines terms. The column AIC is the AIC when that term is deleted from the full model (labelled "none").

Terms	Df	AIC	P-value
none		502.09	
sbp	4	503.16	0.059
tobacco	4	506.48	0.015
ldl	4	508.39	0.006
famhist	1	521.44	0
obesity	4	502.24	0.086
age	4	517.86	0

Model (6.1) and model (6.2) are the "best" model in their own approach, one is through logistic linear regression, the other is from backward stepwise deletion of natural cubic splines. Their AICs are 495.44 and 502.09 respectively, so it's hard to distinguish these two models in AIC criterion. We also considered the Lasso approach and the group Lasso approach to the full model, and Lasso selects model (6.1) but group Lasso selects model (6.2), it makes us even harder to make a choice between these two models. To make a further comparison between Model 6.1 and Model 6.2, we consider the prediction error (PE) criterion and carry out our empirical likelihood ratio (ELR) test.

Following the discussion in Chapter 2, we denote PE_1 as the prediction error in model (6.1) and PE_2 as the prediction error in model (6.2). Define $\mu_\xi = E[PE_1 - PE_2]$. Our aim is to make a hypothesis testing among:

$$H_0 : \mu_\xi \to 0$$

meaning that model (6.1) and model (6.2) are equivalent, against

$$H_a : \mu_\xi \to A \neq 0$$

meaning that one model is sufficiently better than another one.

From the theorems in Chapter 2, given a significant level α, performing a ELR test, we can conduct a model selection procedure based on the following decision rule.

If $p - value > \alpha$, we say the two models work equivalently.

If $p - value < \alpha$ and $APE_1 - APE_2 < 0$, model (6.1) is better than model (6.2).

If $p - value < \alpha$ and $APE_1 - APE_2 > 0$, model (6.2) is better than model (6.1).

A 10 fold cross validation is used here to capture the prediction errors. The result of the ELR test is summarized in Table 9. The average prediction error (APE) is 0.2684 for model (6.1) and 0.3117 for model (6.2).

Table 9: ELR Test between model 6.1 and model 6.2

'-2LLR'	P-value	$APE_1 - APE_2$
6.1977	0.0128	-0.0433
	Model 6.1	Model 6.2
AIC	495.44	502.09
BIC	516.122	593.070
APE	0.2684	0.3117

The column '-2LLR' is the test-statistics in ELR test, which follows a chi-square distribution with d.f. = 1 under H_0. From Table 9 we can get, even though it's hard to distinguish model (6.1) and model (6.2) in AIC, BIC or LASSO method, the ELR test gives a sufficient conclusion that the logistic linear model (6.1) is better than the logistic natural cubic splines model (6.2) for the South African heart disease data in prediction error criterion.

Model (6.2) is slightly more generous than model (6.2) since both **sbp** and **obesity** are included. And it captures the nonlinearity of predictive variables. However, there are 1 + 1 + 4 * 5 = 22 splines are used in model (6.2), in other words, 22 parameters are included in model (6.2), compared with 5 parameters in model (6.1), Model (6.2) is overfitting. It has poor predictive performance, as it overreacts to minor fluctuations in a new training data.

CHAPTER 7: DISCUSSION

In this dissertation we have proposed an empirical likelihood ratio (ELR) test to conduct non-nested model selection. We have established the asymptotic properties of the ELR test-statistics for model selection between two linear models, between a functional coefficient model and a non-parametric regression model, and between two general statistical learning methods under mild conditions. It allows for heteroscedasticity in the error term. It is interesting to mention two future research topics related to this dissertation. First, there might be related technical conditions or more applicable methods to use the ELR test. Second, we can consider a ELR test with variance of prediction errors incorporated the estimation equation, so the ELR would have two constrains. This might improve the performance of the test on heteroscedasticity cases. We will explore these topics in the future.

REFERENCES

Blom, G. (1976). Some properties of incomplete U-statistics. *Biometrika*, 63(3):573–580.

Chen, S.X., Keilegom, I.V. (2009). A review on empirical likelihood methods for regression. *Test*, 18(3):415–447.

Dave, S. S. ET AL. (2004). Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. *N. Engl. J. Med.*, 351:2159–2169.

Fan, J., Zhang, C., Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *The Annals of Statistics*, 29(1):153–193.

Fan, J., Zhang, J. (2004). Sieve empirical likelihood ratio tests for nonparametric functions. *The Annals of Statistics*, 32(5):1858–1907.

Fan, J., Jiang, J. (2005). Nonparametric inference for additive models. *Journal of the American Statistical Association*, 100:890–907.

Hastie, Tibshirani, Friedman. (2009). *The Elements of Statistical Learning (2nd edition). Springer-Verlag.*

Hoeffding, W. (1948). A class of statistics with asymptotically normal distributions. *Annals of Statistics*, 19:293–325.

Jiang, J. (2014). Multivariate functional-coefficient regression models for nonlinear vector time series data. *Biometrika*, 101(3):689–702.

Owen, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249.

Owen, A.B. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics*, 18(1):90–120.

Owen, A.B. (1991). Empirical likelihood for linear models. *The Annals of Statistics*, 19(4):1725–1747.

Owen, A.B. (1995). Nonparametric likelihood confidence bands for a distribution function. *Journal of the American Statistical Association*, 90:516–521.

Owen, A.B. (2001). *Empirical Likelihood. CRC Press.*

Qin, J., L. J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, 22(1):300–325.

Qin, J., Wong. A. (1996). Empirical likelihood in a semi-parametric model. *Scandinavian Journal of Statistics*, 23(2):209–219.

Vuong, Q. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Journal of Econometrics*, 57(2):307–333.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25.

Xue, L., Zhu, L. (2006). Empirical likelihood for single-index models. *Journal of Multivariate Analysis*, 97:1295–1312.

Zhang, J., Gilbels, I. (2003). Sieve empirical likelihood and extensions of the generalized least squares. *Scandinavian Journal of Statistics*, 30(1):1–24.

APPENDIX A: SKETCH OF PROOFS

Lemma A.1

For the two linear models

$$Y_i = Z_i\alpha + \varepsilon_i \tag{2.1}$$

and

$$Y_i = X_i\beta + \varepsilon_i \tag{2.2}$$

To present the asymptotic distribution of ELR test statistic R_n, we introduce the following lemma.

Lemma A.1. *In the two models above, at any prediction point U, for dependent variable Y under an unknown distribution,*

$$\mu_\xi \equiv E[h(F_m, U)] = \Sigma_X^{-1}\mu_{XY}^2 - \Sigma_Z^{-1}\mu_{ZY}^2 + O(\frac{1}{m}),$$

$$\sigma_s^2 \equiv Var[E[h(F_m, U)|U]] = \Sigma_Z^{-4}\mu_{ZY}^4 Var(Z^2) + \Sigma_X^{-4}\mu_{XY}^4 Var(X^2) + 4\Sigma_Z^{-2}\mu_{ZY}^2 E(Z^2Y^2)$$

$$+ 4\Sigma_X^{-2}\mu_{XY}^2 E(X^2Y^2) - 2\Sigma_Z^{-2}\Sigma_X^{-2}\mu_{ZY}^2\mu_{XY}^2 Cov(Z^2, X^2) - 4\Sigma_Z^{-3}\mu_{ZY}^3 E(Z^3Y)$$

$$- 4\Sigma_X^{-3}\mu_{XY}^3 E(X^3Y) + 4\Sigma_Z^{-2}\Sigma_X^{-1}\mu_{ZY}^2\mu_{XY} E(Z^2XY) + 4\Sigma_Z^{-1}\Sigma_X^{-2}\mu_{ZY}\mu_{XY}^2 E(ZX^2Y)$$

$$- 8\Sigma_Z^{-1}\Sigma_X^{-1}\mu_{ZY}\mu_{XY} E(ZXY^2) + o(1),$$

$$\sigma_\xi^2 \equiv Var[h(F_m, U)] = \sigma_s^2 + O(\frac{1}{m}),$$

$$\sigma_t^2 \equiv Var[E[h(F_m, U)]|F_m]] = O(\frac{1}{m^2}),$$

where $\Sigma_Z \equiv E[Z_i^2]$, $\Sigma_X \equiv E[X_i^2]$, $\mu_{ZY} \equiv E[Z_iY_i]$, $\mu_{XY} \equiv E[X_iY_i]$.

Proof of Lemma A.1:

$$\xi_j = [Y_j - (\frac{1}{m}\sum_{i=1}^{m} Z_i^2)^{-1}\frac{1}{m}\sum_{i=1}^{m} Z_iY_iZ_j]^2 - [Y_j - (\frac{1}{m}\sum_{i=1}^{m} X_i^2)^{-1}\frac{1}{m}\sum_{i=1}^{m} X_iY_iX_j]^2$$
$$\overset{a}{\sim} \Sigma_Z^{-2}(\frac{1}{m}\sum Z_iY_i)^2Z_j^2 - \Sigma_X^{-2}(\frac{1}{m}\sum X_iY_i)^2X_j^2 - 2\Sigma_Z^{-1}\frac{1}{m}\sum Z_iY_iZ_jY_j$$
$$+ 2\Sigma_X^{-1}\frac{1}{m}\sum X_iY_iX_jY_j$$

So $\tilde{h}(F_m) = E[h(F_m, U_j)|F_m] = \Sigma_Z^{-1}(\frac{1}{m}\sum Z_iY_i)^2 - \Sigma_X^{-1}(\frac{1}{m}\sum X_iY_i)^2 - 2\Sigma_Z^{-1}\mu_{ZY}\frac{1}{m}\sum Z_iY_i +$ $2\Sigma_X^{-1}\mu_{XY}\frac{1}{m}\sum X_iY_i$.

Taking another expectation,

$$\mu_\xi = E\tilde{h}(F_m) = \Sigma_X^{-1}\mu_{XY}^2 - \Sigma_Z^{-1}\mu_{ZY}^2 + O(\frac{1}{m}).$$

In order to calculate $\sigma_t^2 = Var[\tilde{h}(F_m)]$, we need to calculate all the corresponding variances and covariances in the above expression of $\tilde{h}(F_m)$.

It is straight forward to show that

$$Var[(\frac{1}{m}\sum Z_iY_i)^2] = \frac{m(m-1)(m-2)(m-3)}{m^4}\mu_{ZY}^4 + \frac{C_2^4}{m}\mu_{ZY}^2E(Z^2Y^2)$$
$$- [\frac{m-1}{m}\mu_{ZY}^2 + \frac{1}{m}E(Z^2Y^2)]^2 + O(\frac{1}{m^2})$$
$$= \frac{m-6}{m}\mu_{ZY}^4 + \frac{6}{m}\mu_{ZY}^2E(Z^2Y^2) - \frac{m-2}{m}\mu_{ZY}^4 - \frac{2}{m}\mu_{ZY}^2E(Z^2Y^2)$$
$$+ O(\frac{1}{m^2})$$
$$= \frac{4}{m}\mu_{ZY}^2Var(ZY) + O(\frac{1}{m^2}).$$

Similarly, $Var[(\frac{1}{m}\sum X_iY_i)^2] = \frac{4}{m}\mu_{XY}^2Var(XY) + O(\frac{1}{m^2})$.

$$Cov[(\frac{1}{m}\sum Z_iY_i)^2,(\frac{1}{m}\sum X_iY_i)^2] = \frac{m-6}{m}\mu_{ZY}^2\mu_{XY}^2 + \frac{1}{m}\mu_{XY}^2E(Z^2Y^2)$$
$$+\frac{1}{m}\mu_{ZY}^2E(X^2Y^2) + \frac{4}{m}\mu_{ZY}\mu_{XY}E(ZXY^2) - \frac{m-2}{m}\mu_{ZY}^2\mu_{XY}^2 - \frac{1}{m}\mu_{XY}^2E(Z^2Y^2)$$
$$-\frac{1}{m}\mu_{ZY}^2E(X^2Y^2) + O(\frac{1}{m^2})$$
$$= \frac{4}{m}\mu_{ZY}\mu_{XY}E(ZXY^2) - \frac{4}{m}\mu_{ZY}^2\mu_{XY}^2 + O(\frac{1}{m^2}).$$

$$Cov[(\frac{1}{m}\sum Z_iY_i)^2,\frac{1}{m}\sum Z_iY_i] = \frac{m-3}{m}\mu_{ZY}^3 + \frac{3}{m}\mu_{ZY}E(Z^2Y^2) - \frac{m-1}{m}\mu_{ZY}^3$$
$$-\frac{1}{m}\mu_{ZY}E(Z^2Y^2)$$
$$= \frac{2}{m}\mu_{ZY}Var(ZY) + O(\frac{1}{m^2}),$$

follow the same argument, $Cov[(\frac{1}{m}\sum X_iY_i)^2,\frac{1}{m}\sum X_iY_i] = \frac{2}{m}\mu_{XY}Var(XY) + O(\frac{1}{m^2})$.

$$Cov[(\frac{1}{m}\sum Z_iY_i)^2,\frac{1}{m}\sum X_iY_i] = \frac{m-3}{m}\mu_{ZY}^3\mu_{XY} + \frac{1}{m}\mu_{XY}E(Z^2Y^2) + \frac{2}{m}\mu_{ZY}E(ZXY^2)$$
$$-\frac{m-1}{m}\mu_{ZY}^2\mu_{XY} - \frac{1}{m}\mu_{XY}E(Z^2Y^2) + O(\frac{1}{m^2})$$
$$= \frac{2}{m}\mu_{ZY}E(ZXY^2) - \frac{2}{m}\mu_{ZY}^2\mu_{XY} + O(\frac{1}{m^2}),$$

and $Cov[(\frac{1}{m}\sum X_iY_i)^2,\frac{1}{m}\sum Z_iY_i] = \frac{2}{m}\mu_{XY}E(ZXY^2) - \frac{2}{m}\mu_{ZY}\mu_{XY}^2 + O(\frac{1}{m^2})$.

Combining the above equations leads to

$$\begin{aligned}
\sigma_t^2 &= \frac{4}{m}\Sigma_Z^{-2}\mu_{ZY}^2 Var(ZY) + \frac{4}{m}\Sigma_X^{-2}\mu_{XY}^2 Var(XY) + \frac{4}{m}\Sigma_Z^{-2}\mu_{ZY}^2 Var(ZY) \\
&+ \frac{4}{m}\Sigma_X^{-2}\mu_{XY}^2 Var(XY) - \frac{8}{m}\Sigma_Z^{-1}\Sigma_X^{-1}\mu_{ZY}\mu_{XY}E(ZXY^2) + \frac{8}{m}\Sigma_Z^{-1}\Sigma_X^{-1}\mu_{ZY}^2\mu_{XY}^2 \\
&- \frac{8}{m}\Sigma_Z^{-2}\mu_{ZY}^2 Var(ZY) - \frac{8}{m}\Sigma_X^{-2}\mu_{XY}^2 Var(XY) + \frac{8}{m}\Sigma_Z^{-1}\Sigma_X^{-1}\mu_{ZY}\mu_{XY}E(ZXY^2) \\
&- \frac{8}{m}\Sigma_Z^{-1}\Sigma_X^{-1}\mu_{ZY}^2\mu_{XY}^2 + \frac{8}{m}\Sigma_Z^{-1}\Sigma_X^{-1}\mu_{ZY}\mu_{XY}E(ZXY^2) - \frac{8}{m}\Sigma_Z^{-1}\Sigma_X^{-1}\mu_{ZY}^2\mu_{XY}^2 \\
&- \frac{8}{m}\Sigma_Z^{-1}\Sigma_X^{-1}\mu_{ZY}\mu_{XY}E(ZXY^2) + \frac{8}{m}\Sigma_Z^{-1}\Sigma_X^{-1}\mu_{ZY}^2\mu_{XY}^2 + O(\frac{1}{m^2}) \\
&= O(\frac{1}{m^2})
\end{aligned}$$

In the following we calculate $\sigma_s^2 = Var[E[h(F_m, U_j)|U_j]]$. First of all,

$$\begin{aligned}
h^*(U_j) &= E[h(F_m, U_j)|U_j] \\
&= \Sigma_Z^{-2}\mu_{ZY}^2 Z_j^2 - \Sigma_X^{-2}\mu_{XY}^2 X_j^2 - 2\Sigma_Z^{-1}\mu_{ZY}^2 Z_jY_j + 2\Sigma_X^{-1}\mu_{XY}^2 X_jY_j + o(1)
\end{aligned}$$

Then

$$\sigma_s^2 = Var[h^*(U_j)] = \Sigma_Z^{-4}\mu_{ZY}^4 Var(Z^2) + \Sigma_X^{-4}\mu_{XY}^4 Var(X^2) + 4\Sigma_Z^{-2}\mu_{ZY}^2 Var(ZY)$$

$$+ 4\Sigma_X^{-2}\mu_{XY}^2 Var(XY) - 2\Sigma_Z^{-2}\Sigma_X^{-2}\mu_{ZY}^2\mu_{XY}^2 Cov(Z^2, X^2) - 4\Sigma_Z^{-3}\mu_{ZY}^3[E(Z^3Y) - \Sigma_Z\mu_{ZY}]$$

$$+ 4\Sigma_Z^{-2}\Sigma_X^{-1}\mu_{ZY}^2\mu_{XY}[E(Z^2XY) - \Sigma_Z\mu_{XY}] + 4\Sigma_Z^{-1}\Sigma_X^{-2}\mu_{ZY}\mu_{XY}^2[E(ZX^2Y) - \Sigma_X\mu_{ZY}]$$

$$- 4\Sigma_X^{-3}\mu_{XY}^3[E(X^3Y) - \Sigma_X\mu_{XY}] - 8\Sigma_Z^{-1}\Sigma_X^{-1}\mu_{ZY}\mu_{XY}[E(ZXY^2) - \mu_{ZX}\mu_{XY}] + o(1)$$

$$= \Sigma_Z^{-4}\mu_{ZY}^4 Var(Z^2) + \Sigma_X^{-4}\mu_{XY}^4 Var(X^2) + 4\Sigma_Z^{-2}\mu_{ZY}^2 E(Z^2Y^2)$$

$$+ 4\Sigma_X^{-2}\mu_{XY}^2 E(X^2Y^2) - 2\Sigma_Z^{-2}\Sigma_X^{-2}\mu_{ZY}^2\mu_{XY}^2 Cov(Z^2, X^2) - 4\Sigma_Z^{-3}\mu_{ZY}^3 E(Z^3Y)$$

$$- 4\Sigma_X^{-3}\mu_{XY}^3 E(X^3Y) + +4\Sigma_Z^{-2}\Sigma_X^{-1}\mu_{ZY}^2\mu_{XY} E(Z^2XY) + 4\Sigma_Z^{-1}\Sigma_X^{-2}\mu_{ZY}\mu_{XY}^2 E(ZX^2Y)$$

$$- 8\Sigma_Z^{-1}\Sigma_X^{-1}\mu_{ZY}\mu_{XY} E(ZXY^2) + o(1).$$

By a similar argument, we have

$$Var[h(F_m, U_j)|U_j] = \frac{4}{m}\Sigma_Z^{-4}\mu_{ZY}^2 Var(ZY)Z_j^4 + \frac{4}{m}\Sigma_X^{-4}\mu_{XY}^2 Var(XY)X_j^4$$

$$+ \frac{4}{m}\Sigma_Z^{-2}Var(ZY)Z_j^2Y_j^2 + \frac{4}{m}\Sigma_X^{-2}Var(XY)X_j^2Y_j^2 - \frac{8}{m}\Sigma_Z^{-2}\Sigma_X^{-2}E(ZXY^2)\mu_{ZY}\mu_{XY}Z_j^2X_j^2$$

$$+ \frac{8}{m}\Sigma_Z^{-2}\Sigma_X^{-2}\mu_{ZY}^2\mu_{XY}^2 Z_j^2X_j^2 - \frac{8}{m}\Sigma_Z^{-3}\mu_{ZY}Var(ZY)Z_j^3Y_j$$

$$+ \frac{8}{m}\Sigma_Z^{-2}\Sigma_X^{-1}E(ZXY^2)\mu_{ZY}Z_j^2X_jY_j - \frac{8}{m}\Sigma_Z^{-2}\Sigma_X^{-1}\mu_{ZY}^2\mu_{XY}Z_j^2X_jY_j$$

$$+ \frac{8}{m}\Sigma_Z^{-1}\Sigma_X^{-2}E(ZXY^2)\mu_{XY}Z_jX_j^2Y_j - \frac{8}{m}\Sigma_Z^{-1}\Sigma_X^{-2}\mu_{XY}^2\mu_{ZY}Z_jX_j^2Y_j$$

$$- \frac{8}{m}\Sigma_X^{-3}\mu_{XY}Var(XY)X_j^3Y_j - \frac{8}{m}\Sigma_Z^{-1}\Sigma_X^{-1}E(ZXY^2)Z_jX_jY_j^2$$

$$+ \frac{8}{m}\Sigma_Z^{-1}\Sigma_X^{-1}\mu_{ZY}\mu_{XY}Z_jX_jY_j^2 + O(\frac{1}{m^2})$$

Hence,

$$
\begin{aligned}
E[Var[h(F_m, U_j)|U_j]] &= \frac{4}{m}\Sigma_Z^{-4}\mu_{ZY}^2 Var(ZY)E(Z^4) + \frac{4}{m}\Sigma_X^{-4}\mu_{XY}^2 Var(XY)E(X^4) \\
&+ \frac{4}{m}\Sigma_Z^{-2}Var(ZY)E(Z^2Y^2) + \frac{4}{m}\Sigma_X^{-2}Var(XY)E(X^2Y^2) \\
&- \frac{8}{m}\Sigma_Z^{-2}\Sigma_X^{-2}E(ZXY^2)\mu_{ZY}\mu_{XY}E(Z^2X^2) + \frac{8}{m}\Sigma_Z^{-2}\Sigma_X^{-2}\mu_{ZY}^2\mu_{XY}^2E(Z^2X^2) \\
&- \frac{8}{m}\Sigma_Z^{-3}\mu_{ZY}Var(ZY)E(Z^3Y) + \frac{8}{m}\Sigma_Z^{-2}\Sigma_X^{-1}E(ZXY^2)\mu_{ZY}E(Z^2XY) \\
&- \frac{8}{m}\Sigma_Z^{-2}\Sigma_X^{-1}\mu_{ZY}^2\mu_{XY}E(Z^2XY) + \frac{8}{m}\Sigma_Z^{-1}\Sigma_X^{-2}E(ZXY^2)\mu_{XY}E(ZX^2Y) \\
&- \frac{8}{m}\Sigma_Z^{-1}\Sigma_X^{-2}\mu_{XY}^2\mu_{ZY}E(ZX^2Y) - \frac{8}{m}\Sigma_X^{-3}\mu_{XY}Var(XY)E(X^3Y) \\
&- \frac{8}{m}\Sigma_Z^{-1}\Sigma_X^{-1}E(ZXY^2)^2 + \frac{8}{m}\Sigma_Z^{-1}\Sigma_X^{-1}\mu_{ZY}\mu_{XY}E(ZXY^2) + O(\frac{1}{m^2}) \\
&= O(\frac{1}{m})
\end{aligned}
$$

Thus from law of total variance, $\sigma_\xi^2 = Var[E[h(F_m, U_j)|U_j]] + E[Var[h(F_m, U_j)|U_j]] = \sigma_s^2 + O(\frac{1}{m})$. □

Lemma A.2

For the functional coefficient model,

$$Y_i = \beta(Z_i)X_i + \varepsilon_i, \tag{3.1}$$

and the non-parametric regression model,

$$Y_i = m(X_i) + \varepsilon_i. \tag{3.2}$$

Lemma A.2. *In the two models above, at any prediction point U, for dependent* *variable Y under an unknown distribution,*

$$\mu_\xi \equiv E[h(F_m,U)] = \frac{m+1}{m}[\mu_Y^2 - \Sigma_X^{-1}\mu_{XY}^2] + \frac{1}{mh_1}\Sigma_X^{-1}v(K)E[X^2Y^2g(Z)^{-1}]$$

$$-\frac{1}{mh_2}v(J)E[Y^2f(X)^{-1}] + o(\frac{1}{m}),$$

$$\sigma_s^2 \equiv Var[E[h(F_m,U)|U]] = \Sigma_X^{-4}\mu_{XY}^4 Var(X^2) + 4\mu_Y^2 Var(Y) + 4\Sigma_X^{-2}\mu_{XY}^2 E(X^2Y^2)$$

$$-4\Sigma_X^{-3}\mu_{XY}^3 E(X^3Y) + 4\Sigma_X^{-2}\mu_{XY}^2\mu_Y E(X^2Y) + 4\Sigma_X^{-1}\mu_{XY}^2\mu_Y^2$$

$$-8\Sigma_X^{-1}\mu_{XY}\mu_Y E(XY^2) + o(1),$$

$$\sigma_\xi^2 \equiv Var[h(F_m,U)] = \sigma_s^2 + O(\frac{1}{m}),$$

$$\sigma_t^2 \equiv Var[E[h(F_m,U)]|F_m]] = o(\frac{1}{m}),$$

where $\Sigma_X(Z) \equiv E[X_i^2|Z_i = Z]$, $\mu_Y \equiv E[Y_i]$, $\mu_{XY} \equiv E[X_iY_i]$, $f(\cdot)$ *and* $g(\cdot)$ *are the true densities of* X *and* Z *respectively, and* $v(K) = \int K^2(u)du$, $v(J) = \int J^2(u)du$.

Proof of Lemma A.2:

By the definition in Chapter 3, we simply consider $\xi_j = h(F_m, U_j)$, and $h^*(U_j) = E[h(F_m, U_j)|U_j]$, where $m = n - ql$. So

$$\begin{aligned}
\xi_j &= (Y_j - \hat{\beta}(Z_j)X_j)^2 - (Y_j - \hat{m}(X_j))^2 \\
&= (Y_j - (\frac{1}{m}\sum_{i=1}^{m} X_i^2 K_{h_1}(Z_i - Z_j))^{-1}\frac{1}{m}\sum_{i=1}^{m} X_i Y_i K_{h_1}(Z_i - Z_j)X_j)^2 \\
&- (Y_j - (\frac{1}{m}\sum_{i=1}^{m} J_{h_2}(X_i - X_j))^{-1}\frac{1}{m}\sum_{i=1}^{m} J_{h_2}(X_i - X_j)Y_i)^2 \\
&\overset{a}{\sim} \Sigma_X(Z_j)^{-2}g(Z_j)^{-2}(\frac{1}{m}\sum X_i Y_i K_{h_1}(Z_i - Z_j))^2 X_j^2 \\
&- f(X_j)^{-2}(\frac{1}{m}\sum J_{h_2}(X_i - X_j)Y_i)^2 \\
&- 2\Sigma_X(Z_j)^{-1}g(Z_j)^{-1}\frac{1}{m}\sum X_i Y_i K_{h_1}(Z_i - Z_j)X_j Y_j \\
&+ 2f(X_j)^{-1}\frac{1}{m}\sum J_{h_2}(X_i - X_j)Y_i Y_j.
\end{aligned}$$

Using change of variable,

$$\begin{aligned}
\tilde{h}(F_m) = E[h(F_m, U_j)] &= \Sigma_X^{-1}\frac{1}{m^2}\sum_{i \neq k} X_i Y_i X_k Y_k \int K(u)K(v)dudv \\
&+ \Sigma_X^{-1}\frac{1}{m^2}\sum_i X_i^2 Y_i^2 \int g(Z_i + uh_1)^{-1}K^2(u)\frac{1}{h_1}du - \frac{1}{m^2}\sum_{i \neq k} Y_i Y_k \int J(u)J(v)dudv \\
&- \frac{1}{m^2}\sum_i Y_i^2 \int f(X_i + uh_1)^{-1}J^2(u)\frac{1}{h_2}du - \Sigma_X^{-1}\mu_{XY}\frac{2}{m}\sum_i X_i Y_i \int K(u)du \\
&+ \mu_Y \frac{2}{m}\sum_i Y_i \int J(u)du \\
&= \Sigma_X^{-1}\frac{1}{m^2}\sum_{i \neq k} X_i Y_i X_k Y_k + \Sigma_X^{-1}v(K)\frac{1}{m^2 h_1}\sum_i X_i^2 Y_i^2 g(Z_i)^{-1} - \frac{1}{m^2}\sum_{i \neq k} Y_i Y_k \\
&- v(J)\frac{1}{m^2 h_2}\sum_i Y_i^2 f(X_i)^{-1} - \Sigma_X^{-1}\mu_{XY}\frac{2}{m}\sum_i X_i Y_i + \mu_Y \frac{2}{m}\sum_i Y_i + o(\frac{1}{m}).
\end{aligned}$$

Taking another expectation,

$$\mu_\xi = E[h(F_m, U)] = \frac{m+1}{m}[\mu_Y^2 - \Sigma_X^{-1}\mu_{XY}^2] + \frac{1}{mh_1}\Sigma_X^{-1}v(K)E[X^2Y^2g(Z)^{-1}]$$
$$-\frac{1}{mh_2}v(J)E[Y^2f(X)^{-1}] + o(\frac{1}{m}).$$

To calculate $\sigma_t^2 = Var[\tilde{h}(F_m)]$, we firstly calculate all the corresponding variances and covariances in the expression of $\tilde{h}(F_m)$.

$$Var(\frac{1}{m^2}\sum_{i\neq k} X_iY_iX_kY_k) = \frac{m-6}{m}\mu_{XY}^4 + \frac{4}{m}\mu_{XY}^2E(X^2Y^2) - \frac{m-2}{m}\mu_{XY}^4$$
$$= \frac{4}{m}\mu_{XY}^2 Var(XY)$$
$$Var(\frac{1}{m^2}\sum_{i\neq k} Y_iY_k) = \frac{m-6}{m}\mu_Y^4 + \frac{4}{m}\mu_Y^2E(Y^2) - \frac{m-2}{m}\mu_Y^4 = \frac{4}{m}\mu_Y^2 Var(Y)$$
$$Var(\frac{1}{m}\sum_i X_iY_i) = \frac{1}{m}Var(XY)$$
$$Var(\frac{1}{m}\sum_i Y_i) = \frac{1}{m}Var(Y)$$

$Var(\frac{1}{m^2h_1}\sum_i X_i^2Y_i^2g(Z_i)^{-1}) = O(\frac{1}{m^2h_1}) = o(\frac{1}{m})$,

$Var(\frac{1}{m^2h_2}\sum_i Y_i^2f(X_i)^{-1}) = O(\frac{1}{m^2h_2}) = o(\frac{1}{m})$.

$$Cov(\frac{1}{m^2}\sum_{i\neq k} X_iY_iX_kY_k, \frac{1}{m^2}\sum_{i\neq k} Y_iY_k)$$
$$= \frac{m-6}{m}\mu_{XY}^2\mu_Y^2 + \frac{4}{m}\mu_{XY}\mu_Y E(XY^2) - \frac{m-2}{m}\mu_{XY}^2\mu_Y^2 = \frac{4}{m}\mu_{XY}\mu_Y Cov(XY, Y)$$
$$Cov(\frac{1}{m^2}\sum_{i\neq k} X_iY_iX_kY_k, \frac{1}{m}\sum_i X_iY_i)$$
$$= \frac{m-3}{m}\mu_{XY}^3 + \frac{2}{m}\mu_{XY}E(X^2Y^2) - \frac{m-1}{m}\mu_{XY}^3 = \frac{2}{m}\mu_{XY}Var(XY)$$
$$Cov(\frac{1}{m^2}\sum_{i\neq k} X_iY_iX_kY_k, \frac{1}{m}\sum_i Y_i)$$
$$= \frac{m-3}{m}\mu_{XY}^2\mu_Y + \frac{2}{m}\mu_{XY}E(XY^2) - \frac{m-1}{m}\mu_{XY}^2\mu_Y = \frac{2}{m}\mu_{XY}Cov(XY, Y)$$
$$Cov(\frac{1}{m^2}\sum_{i\neq k} Y_iY_k, \frac{1}{m}\sum_i X_iY_i)$$
$$= \frac{m-3}{m}\mu_{XY}\mu_Y^2 + \frac{2}{m}\mu_Y E(XY^2) - \frac{m-1}{m}\mu_{XY}\mu_Y^2 = \frac{2}{m}\mu_Y Cov(XY, Y)$$
$$Cov(\frac{1}{m^2}\sum_{i\neq k} Y_iY_k, \frac{1}{m}\sum_i Y_i)$$
$$= \frac{m-3}{m}\mu_Y^3 + \frac{2}{m}\mu_Y E(Y^2) - \frac{m-1}{m}\mu_Y^3 = \frac{2}{m}\mu_Y Var(Y)$$
$$Cov(\frac{1}{m^2}\sum_i X_iY_i, \frac{1}{m}\sum_i Y_i)$$
$$= \frac{m-1}{m}\mu_{XY}\mu_Y + \frac{1}{m}E(XY^2) - \mu_{XY}\mu_Y = \frac{1}{m}Cov(XY, Y).$$

Follow a similar calculation, we can get that other covariances in the expression of $\tilde{h}(F_m)$ are $o(\frac{1}{m})$.

Now we plug in all the variances and covariances above to calculate σ_t^2,

$$\sigma_t^2 = Var[\tilde{h}(f-m)] = \frac{4}{m}\Sigma_X^{-2}\mu_{XY}^2 Var(XY) + \frac{4}{m}\mu_Y^2 Var(Y) + \frac{4}{m}\Sigma_X^{-2}\mu_{XY}^2 Var(XY)$$
$$+ \frac{4}{m}\mu_Y^2 Var(Y) - \frac{8}{m}\Sigma_X^{-1}\mu_{XY}\mu_Y Cov(XY,Y) - \frac{8}{m}\Sigma_X^{-2}\mu_{XY}^2 Var(XY)$$
$$+ \frac{8}{m}\Sigma_X^{-1}\mu_{XY}\mu_Y Cov(XY,Y) + \frac{8}{m}\Sigma_X^{-1}\mu_{XY}\mu_Y Cov(XY,Y) - \frac{8}{m}\mu_Y^2 Var(Y)$$
$$- \frac{8}{m}\Sigma_X^{-1}\mu_{XY}\mu_Y Cov(XY,Y) + o(\frac{1}{m})$$
$$= o(\frac{1}{m})$$

In order to calculate $\sigma_s^2 = Var[h^*(U_j)]$, first of all, using change of variable,

$$h^*(U_j) = E[h(F_m, U_j)|U_j]$$
$$= \Sigma_X(Z_j)^{-2}g(Z_j)^{-2}X_j^2\mu_{XY}^2\frac{m-1}{m}\int K(u)K(v)g(Z_j+uh_1)g(Z_j+vh_1)dudv$$
$$+ \Sigma_X(Z_j)^{-2}g(Z_j)^{-2}X_j^2E(X^2Y^2)\frac{1}{mh_1}\int K^2(u)g(Z_j+uh_1)du$$
$$- \mu_Y^2\frac{m-1}{m}\int J(u)J(v)g(X_j+uh_1)g(X_j+vh_1)dudvf(X_j)^{-2}$$
$$- f(X_j)^{-2}E(Y^2)\frac{1}{mh_2}\int J^2(u)f(X_j+uh_2)du$$
$$- 2\Sigma_X(Z_j)^{-1}X_jY_jg(Z_j)^{-1}\mu_{XY}\int K(u)g(Z_j+uh_1)du$$
$$+ 2f(X_j)^{-1}Y_j\mu_Y\int J(u)f(X_j+uh_2)du$$
$$= \frac{m-1}{m}\Sigma_X(Z_j)^{-2}\mu_{XY}^2X_j^2 + \frac{1}{mh_1}\Sigma_X(Z_j)^{-2}E(X^2Y^2)v(K)g(Z_j)^{-1}X_j^2 - \frac{m-1}{m}\mu_Y^2$$
$$- \frac{1}{mh_2}E(Y^2)v(J)f(X_j)^{-1} - 2\Sigma_X(Z_j)^{-1}\mu_{XY}X_jY_j + 2\mu_YY_j + O(h_1^2+h_2^2)$$
$$= \Sigma_X(Z_j)^{-2}\mu_{XY}^2X_j^2 - \mu_Y^2 - 2\Sigma_X(Z_j)^{-1}\mu_{XY}X_jY_j + 2\mu_YY_j + o(1)$$

So

$$\sigma_s^2 = Var[h^*(U_j)] = \Sigma_X^{-4}\mu_{XY}^4 Var(X^2) + 4\Sigma_X^{-2}\mu_{XY}^2 Var(XY) + 4\mu_Y^2 Var(Y)$$

$$- 4\Sigma_X^{-3}\mu_{XY}^3 Cov(X^2, XY) + 4\Sigma_X^{-2}\mu_{XY}^2\mu_Y Cov(X^2, Y)$$

$$- 8\Sigma_X^{-1}\mu_{XY}\mu_Y Cov(XY, Y) + o(1)$$

$$= Var[h^*(U_j)] = \Sigma_X^{-4}\mu_{XY}^4 Var(X^2) + 4\mu_Y^2 Var(Y) + 4\Sigma_X^{-2}\mu_{XY}^2 E(X^2Y^2)$$

$$- 4\Sigma_X^{-3}\mu_{XY}^3 E(X^3Y) + 4\Sigma_X^{-2}\mu_{XY}^2\mu_Y E(X^2Y) + 4\Sigma_X^{-1}\mu_{XY}^2\mu_Y^2$$

$$- 8\Sigma_X^{-1}\mu_{XY}\mu_Y E(XY^2) + o(1).$$

Follow a similar calculation,

$$Var[h(F_m, U_j)|U_j] = \frac{4}{m}\Sigma_X^{-4}\mu_{XY}^4 X_j^4 + \frac{4}{m}\mu_Y^4 + \frac{4}{m}\Sigma_X^{-2}\mu_{XY}^2 X_j^2 Y_j^2 + \frac{4}{m}\mu_Y^2 Y_j^2$$
$$+ \frac{8}{m}\Sigma_X^{-2}\mu_{XY}^2\mu_Y^2 X_j^2 + \frac{8}{m}\Sigma_X^{-3}\mu_{XY}^3 X_j^3 Y_j - \frac{8}{m}\Sigma_X^{-2}\mu_{XY}^2\mu_Y X_j^2 Y_j$$
$$- \frac{8}{m}\Sigma_X^{-1}\mu_{XY}\mu_Y^2 X_j Y_j + \frac{8}{m}\mu_Y^3 Y_j + \frac{8}{m}\Sigma_X^{-1}\mu_{XY}\mu_Y X_j Y_j^2 + o(\frac{1}{m}).$$

Hence we can get

$$E[Var[h(F_m, U_j)|U_j]] = \frac{4}{m}\Sigma_X^{-4}\mu_{XY}^4 E(X^4) + \frac{4}{m}\mu_Y^4 + \frac{4}{m}\Sigma_X^{-2}\mu_{XY}^2 E(X^2Y^2) + \frac{4}{m}\mu_Y^2\Sigma_Y$$
$$+ \frac{8}{m}\Sigma_X^{-1}\mu_{XY}^2\mu_Y^2 + \frac{8}{m}\Sigma_X^{-3}\mu_{XY}^3 E(X^3Y) - \frac{8}{m}\Sigma_X^{-2}\mu_{XY}^2\mu_Y E(X^2Y)$$
$$- \frac{8}{m}\Sigma_X^{-1}\mu_{XY}^2\mu_Y^2 + \frac{8}{m}\mu_Y^4 + \frac{8}{m}\Sigma_X^{-1}\mu_{XY}\mu_Y E(XY^2) + o(\frac{1}{m})$$
$$= O(\frac{1}{m})$$

Thus by the law of total variance,

$$\sigma_\xi^2 = Var[E[h(F_m, U_j)|U_j]] + E[Var[h(F_m, U_j)|U_j]] = \sigma_s^2 + O(\frac{1}{m})$$

.

APPENDIX B: SKETCH OF PROOFS

Theorem 4.1

Under the following technical conditions:

(1) $\exists\delta > 0$, $C_n > 0$, $D > 0$, such that $E[\xi^{2+\delta}(U)] < D < \infty$, $\sigma_\xi^2 > C_n > 0$

(2) $Var\{E[h(F_m, U)|F_m]\}/\sigma_\xi^2 = o(\frac{1}{n})$

(3) $E\{Var[h(F_m, U)|U]\}/\sigma_\xi^2 = o(1)$

With the following notations:

$\xi(U) \equiv h(F_m, U)$, $\mu_\xi \equiv E[h(F_m, U)]$, $\sigma_\xi^2 \equiv Var[h(F_m, U)]$,

$h^*(U) \equiv E[h(F_m, U)|U]$, $\tilde{h}(F_m) \equiv E[h(F_m, U)|F_m]$,

$\sigma_s^2 \equiv Var[h^*(U)]$, $\sigma_t^2 \equiv Var[\tilde{h}(F_m)]$.

First of all, we derive the distribution of $\bar{\xi} = \frac{1}{n-m}\sum_{j=m+1}^{n} h(F_m, U_j)$.

Since $F_m, U_{m+1}, \cdots, U_n$ are independent, following Hajek Projection Principle, we define the Hajek projection of $\bar{\xi} = \frac{1}{n-m}\sum_{j=m+1}^{n} h(F_m, U_j)$ as

$$\begin{aligned}
\bar{\xi}^* &= E[\bar{\xi}|F_m] + \sum_{j=m+1}^{n} E[\bar{\xi}|U_j] - (n-m)E(\bar{\xi}) \\
&= \tilde{h}(F_m) + \sum_{j=m+1}^{n} E[\frac{1}{n-m}\sum_{k=m+1}^{n} h(F_m, U_k)|U_j] - (n-m)\mu_\xi
\end{aligned}$$

Notice that

$$E[h(F_m, U_k)|U_j] = \begin{cases} h^*(U_j) & \text{if } k = j \\ \\ \mu_\xi & \text{if } k \neq j \end{cases}$$

So

$$\bar{\xi}^* = \tilde{h}(F_m) + \frac{1}{n-m}\sum_{k=m+1}^{n} h^*(U_j) + \frac{(n-m)(n-m-1)}{n-m}\mu_\xi - (n-m)\mu_\xi$$
$$= \frac{1}{n-m}\sum_{k=m+1}^{n} h^*(U_j) + \tilde{h}(F_m) - \mu_\xi$$

The first term is the average of $n-m$ i.i.d. random variables $h^*(U_j)$, which has a mean of $E[h^*(U)] = E[E[h(F_m, U)|U]] = \mu_\xi$, and a variance of $Var[h^*(U_j)] = \sigma_s^2$. The remainder $\tilde{h}(F_m)$ is $o(\frac{\sigma_s^2}{\sqrt{n}})$ by Condition (2) and (3) and Markov's inequality. Thus by the central limit theorem,

$$\bar{\xi}^* \overset{a}{\sim} N(\mu_\xi, (n-m)^{-1}\sigma_s^2) \tag{B1}$$

Calculation of $Var(\bar{\xi})$ is a bit more involved, but not too bad. For $j \neq k$,

$$Cov(\xi_j, \xi_k) = Cov[h(F_m, U_j), h(F_m, U_k)]$$
$$= E[(h(F_m, U_j) - \mu_\xi)(h(F_m, U_k) - \mu_\xi)]$$
$$= E[h(F_m, U_j)h(F_m, U_k)] - \mu_\xi^2$$
$$= E[E[h(F_m, U_j)h(F_m, U_k)|F_m]] - \mu_\xi^2$$

Because U_j, U_k are i.i.d given F_m, taking expectation of the conditional expectation over F_m,the two terms in the conditional expectation are independent. So

$$Cov(\xi_j, \xi_k) = E[E[h(F_m, U_j)|F_m] \cdot E[h(F_m, U_k)|F_m]] - \mu_\xi^2$$
$$= E[\tilde{h}(F_m)\tilde{h}(F_m)] - \mu_\xi^2$$
$$= \sigma_t^2$$

Thus

$$
\begin{aligned}
Var(\bar{\xi}) &= \frac{1}{(n-m)^2}\sum_{j=m+1}^{n}\sum_{j=m+1}^{n} Cov(\xi_j,\xi_k) \\
&= \frac{1}{(n-m)^2}[\sum_{j} Var(\xi_j) + \sum_{j\neq k} Cov(\xi_j,\xi_k)] \\
&= \frac{1}{n-m}\sigma_{\xi}^2 + (1+\frac{1}{n-m})\sigma_t^2.
\end{aligned}
$$

And $\sigma_t^2 = o(\frac{\sigma_{\xi}^2}{\sqrt{n}})$ because of Condition (2).

Thus

$$
Var(\bar{\xi}) = \frac{1}{n-m}\sigma_{\xi}^2 + o(\frac{\sigma_{\xi}^2}{n^{\frac{3}{2}}}). \tag{B2}
$$

From Condition (3), we can get $\sigma_{\xi}^2/\sigma_s^2 \to 1$, and combine (B1), (B2), we have

$$
Var[\bar{\xi}]/Var[\bar{\xi}^*] \to 1.
$$

So using Hajek projection asymptotic theorem,

$$
\frac{\bar{\xi} - E[\bar{\xi}]}{\sqrt{Var[\bar{\xi}]}} - \frac{\bar{\xi}^* - E[\bar{\xi}^*]}{\sqrt{Var[\bar{\xi}^*]}} \xrightarrow{P} 0.
$$

Therefore using (B2) again, we have the asymptotic distribution of $\bar{\xi}$,

$$
\sqrt{n-m}\frac{\bar{\xi} - \mu_{\xi}}{\sigma_{\xi}} \xrightarrow{d} N(0,1) \tag{B3}
$$

So using (B3) and Markov's inequality, we have $\bar{\xi}/\sigma_{\xi} = O_p(n^{-\frac{1}{2}})$.

Under Condition (1) that $E[\xi^{2+\delta}(U)] < \infty$ for a small $\delta > 0$,

we can get $Var[h^2(F_m,U)|F_m] < \infty$, and thus $\hat{V} = O_p(1)$.

From (4.1)

$$0 = \frac{1}{n-m}\sum_j \frac{\xi_j}{1+\lambda\xi_j} = \frac{1}{n-m}\sum_j \frac{\xi_j(1+\lambda\xi_j)-\lambda\xi_j^2}{1+\lambda\xi_j}$$
$$= \bar{\xi} - \lambda\frac{1}{n-m}\sum_j \frac{\xi_j^2}{1+\lambda\xi_j} \leqslant \bar{\xi} - \frac{|\lambda|}{1+|\lambda|\xi^*}\hat{V}$$
$$So|\lambda|(\hat{V} - \xi^*\bar{\xi}) \leqslant \bar{\xi},$$

where $\xi^* = \max_j |\xi_j| = o_p(n^{\frac{1}{2}})$.

This is because $P((n-m)^{-\frac{1}{2}}\xi^* > \varepsilon) \leqslant \frac{E[(n-m)^{-\frac{2+\delta}{2}}\xi^{*2+\delta}]}{\varepsilon^{2+\delta}}$

$\leqslant (n-m)^{-\frac{\delta}{2}}\varepsilon^{-(2+\delta)}\frac{1}{n-m}\sum_j E(\xi_j^{2+\delta}) \to 0$ for a small $\delta > 0$ and δ satisfies $E(\xi_j^{2+\delta}) < \infty$.

And since $\bar{\xi}/\sigma_\xi = O_p(n^{-\frac{1}{2}})$, $\hat{V} = O_p(1)$,

thus $\lambda/\sigma_\xi = O_p(n^{-\frac{1}{2}})$.

Again from (4.1),

$$0 = \frac{1}{n-m}\sum_j \xi_j - \lambda\frac{1}{n-m}\sum_j \frac{\xi_j^2(1+\lambda\xi_j)}{1+\lambda\xi_j} + \frac{1}{n-m}\sum_j \frac{(\lambda\xi_j)^2\xi_j}{1+\lambda\xi_j}$$
$$= \bar{\xi} - \lambda\hat{V} + \frac{1}{n-m}\sum_j \frac{(\lambda\xi_j)^2\xi_j}{1+\lambda\xi_j}$$

Since $\xi^* = o_p(n^{\frac{1}{2}})$, $\lambda/\sigma_\xi = O_p(n^{-\frac{1}{2}})$, we have $\max|\lambda\xi_j| = o_p(\sigma_\xi)$. So the third term on the right side of the equation is $o_p(n^{-\frac{1}{2}}\sigma_\xi)$.

Thus $\lambda = \hat{V}^{-1}\bar{\xi} + \delta$, where $\delta/\sigma_\xi = o_p(n^{-\frac{1}{2}})$.

Plug this into (4.2), the log empirical likelihood ratio

$$
\begin{aligned}
R_n &= 2\sum_{j=m+1}^{n}\log(1+\lambda\xi_j) = 2\sum_{j=m+1}^{n}[\lambda\xi_j - \frac{1}{2}\lambda^2\xi_j^2 + O_p(\lambda^3\xi_j^3)] \\
&= 2(n-m)\hat{V}^{-1}\bar{\xi}^2 + 2(n-m)\delta\bar{\xi} - (n-m)\hat{V}^{-1}\bar{\xi}^2 - (n-m)\delta^2\hat{V} - 2(n-m)\delta\bar{\xi} \\
&+ 2\sum_j O_p(\lambda^3\xi_j^3) \\
&= (n-m)\hat{V}^{-1}\bar{\xi}^2 - (n-m)\delta^2\hat{V} + 2\sum_j O_p(\lambda^3\xi_j^3)
\end{aligned}
$$

Follow the notation in Chapter 4, let $\tau_n = n^{\frac{1}{2}}\mu_\xi/\sigma_\xi$.

The leading term is asymptotically equivalent to $\chi_1^2(\tau_n^2)$ because of (B3) and $\hat{V}/\sigma_\xi^2 \to 1$, the second term is $o_p(1)$ because $\delta/\sigma_\xi = o_p(n^{-\frac{1}{2}})$, and for some finite $C > 0$, the third term $\leqslant C\sum_j(\lambda^3\xi_j^3) \leqslant qmC|\lambda|^3\xi^*\hat{V} = o_p(1)$. Therefore

$$R_n \overset{a}{\sim} \chi_1^2(\tau_n^2).$$

Theorem 2.1 and Theorem 3.1

With Lemma A.1 and Lemma A.2, the technical conditions in Theorem 4.1 can be easily verified. Also from Lemma A.1 and A.2, we can get that for model selection between two linear models or between a functional coefficient model and a nonparametric regression model, $\sigma_\xi = O(1)$, and either $\mu_\xi \to 0$ or $\mu_\xi \to A \neq 0$. Hence, for $\tau_n = n^{\frac{1}{2}}\mu_\xi/\sigma_\xi$, either $\tau_n = 0$ or $\tau_n \to \pm\infty$. So we only need to consider $H_0 : \{\mu_\xi \to 0\} \Rightarrow \{\tau_n \to 0\}$ and $H_a : \{\mu_\xi \to A \neq 0\} \Rightarrow \{\tau_n \to \pm\infty\}$ in Chapter 2 and 3. Thus Theorem 2.1 and Theorem 3.1 are simply proved by Theorem 4.1.