Revista de Economía del Rosario

A Note on the Bandwidth Choice When the Null Hypothesis is Semiparametric

Jorge Barrientos-Marín*
University of Alicante, Spain

Recibido: abril 2005 - Aprobado: mayo 2005

Abstract. This work presents a tool for the additivity test. The additive model is widely used for parametric and semiparametric modeling of economic data. The additivity hypothesis is of interest because it is easy to interpret and produces reasonably fast convergence rates for non-parametric estimators. Another advantage of additive models is that they allow attacking the problem of the curse of dimensionality that arises in non- parametric estimation. Hypothesis testing is based in the well-known bootstrap residual process. In nonparametric testing literature, the dominant idea is that bandwidth utilized to produce bootstrap sample should be bigger that bandwidth for estimating model under null hypothesis. However, there is no hint so far about how to choose such bandwidth in practice. We will discuss a first step to find some rule of thumb to choose bandwidth in that context. Our suggestions are accompanied by simulation studies.

Keywords: additive models, bootstrap, bootstrap test, kernel smoothing, nonparametric regression.
JEL Classification: C13, C14, C52

Resumen. Este artículo presenta un contraste de aditividad. El modelo aditivo es usado para modelar estructuras paramétricas y semiparamétricas. La hipótesis de aditividad es interesante porque es fácil de interpretar y produce unas tasas de convergencia razonablemente rápidas de estimadores noparamétricos. Una ventaja adicional de las estructuras aditivas es que permite atacar directamente el problema de la maldición de la dimensionalidad que surge en estimación noparamétrica. El procedimiento que proponemos para el contraste de hipótesis esta basado en un proceso de remuestreo (bootstrap) de los residuales del modelo aditivo. La idea dominante en la selección de la banda usada para generar las muestras bootstrap, es que esta debe ser más grande que la banda utilizada para la estimación del modelo aditivo. No obstante, hasta el momento la literatura existente no suministra ayuda alguna. Nosotros discutimos, como un primer paso, un tipo de regla para elegir tal banda en este contexto.

Palabras Clave: modelos de aditividad, bootstrap, test de bootstrap, suavizamiento Kernel, regresión noparamétrica.
Clasificación JEL: C13, C14, C52

* I acknowledge the Spanish Ministry of Science and Technology for their funding of project BEC2001-0535. I am also grateful with my advisor Stefan Sperlich, from the Department of Economics, Universidad Carlos III, for his suggestions and ideas to improve the present work. This paper was presented at the Spring Meeting of Young Economists in Warsaw, Poland. 2004.

Profesor Asistente, Departamento de Fundamentos del Análisis Económico. Dirección de correspondencia: 03080, Campus de San Vicente del Raspeig. E-mail: jbarr@merlin.fae.ua.es.

1. Motivation

Smoothing techniques such as density estimation have an important role in the current development of theoretical econometrics. The usual practice when constructing regression models is to specify a parametric family. The most usual of these families is the linear model. But there is no reason to limit ourselves to this kind of model: since it belongs to a continuum of possible functional forms, there is a probability close to zero that we will choose correctly. A way to avoid the misspecification is to assume a non-functional form. Data can give us all the information we need to investigate functional forms, using, e.g. the kernel estimator for the regression function. This approach is known as a nonparametric estimation. A popular semiparametric model that has been investigated in recent years is the additive one. The estimation procedure for this kind of structure uses nonparametric techniques. The additive structure is present in many models of economic behavior, including the usual parametric estimation. Then given a data set, one could be interested in knowing what kind of structure follows the data.

Härdle and Marron (1991) propose a technique to construct confidence intervals by a bootstrap method. We take advantage from this procedure to construct the additivity test. Their approach consists in resampling the estimated residuals, $\hat{\varepsilon}_i = Y_i - \hat{m}_g(X_i)$, and then using these data to construct an estimator, whose distribution will approximate the distribution of the original estimator. Such a procedure allows for selection of two smoothing parameters, *g* and *h*, where *g* is the selected smoothing parameter for the bootstrap estimation and *h* is the bandwidth for the model under the null hypothesis. The band *g* must be oversmoothed. To test the hypothesis, Dette, Von Lieres and Sperlich (2003) used the bootstrap to construct statistics, and evaluated its performance. In fact, for our simulation studies, we are going to use the same tests statistic.

Deaton and Muellbauer (1980) provide many microeconomic examples in which a separable structure is convenient for analysis and important for interpretability. It has reasonably fast convergence rates for nonparametric estimators. Another advantage of the additive model is that it allows us to attack the problem of the *curse of dimensionality* that arises in nonparametric estimation. Let $m : \mathbb{R}^d \to \mathbb{R}$ be a smooth function and suppose that we want to estimate:

$$E[Y \mid X = x] = m(x) \tag{1}$$

where $m(x) = c + \sum_{\alpha \in \Lambda} m_\alpha(x_\alpha)$, c is a constant, m_α are *d* functions each one defined on $\mathbb{R}$ for all $\alpha \in \Lambda = \{1, \ldots, d\}$, and Λ is a set of indexes. Note that in this model we need to estimate functions of one dimension. Therefore, it is possible to speak of dimensionality reduction through additive modeling. Stone (1985) shows that the optimal rate for estimating the nonparametric function regression (1) is $n^{-\ell/(2\ell+d)}$ where ℓ is an index of smoothness of *m* and *n* is the sample size.

The performance of several statistical tests under the null and the alternative hypotheses are also studied. The alternative model is the well-known Nadaraya-Watson kernel regression function estimator. The null model is the aforementioned additive model. The method's procedure is to estimate a multidimensional functional of *m* first and then use the internal marginal integration to get the marginal effect. Under the additive structure this procedure yields m_α, $\alpha \in \Lambda$, plus a constant (see Linton and Nielsen, 1995). The asymptotic power of a test of H_0 is often investigated by deriving the asymptotic probability that the test rejects H_0 against an alternative model.

The objective of this work is to propose a rule of thumb to choose oversmoothed bandwidth in bootstrap estimation. To test such rule, we estimate the null model with an optimal bandwidth and after that we construct the statistics to test additivity, and to estimate the additive model we use a technique known as marginal integration estimation. Additional motivations for this work are: first, due to the advantages additive models offer to empirical researcher there is an increased interest in testing the additive structure. Second, at present there is not much theoretical work about testing and hardly empirical studies on internal marginal integrated estimator (IMIE). Sperlich, Tjostheim and Yang (2002) introduce a bootstrap based additivity test applying the marginal integration. In Dette, Von Lieres and Sperlich (2003) various statistical tests to check additive separability are introduced; they concentrate on the differences that result from the use of a different smoother in marginal integrations.

The work is organized as follows. In section 2, we present the models to be estimated under both null and alternative hypotheses and the statistical tests to verify the hypothesis. In section 3 the procedure of estimation is described in detail. In the section 4, we provide some results based on simulations. In section 5 we present the conclusion and topics for further research. In the appendix 1 we show the results related to the simulations.

2. Models to Be Estimated

2.1. *The Internalized Nadaraya-Watson Estimator*

Let $\{(X_{i,}Y_i)\}_{i=1}^n \in \mathbb{R}^{d+1}$ be a finite sequence of random vectors and $m : \mathbb{R}^d \to \mathbb{R}$ an unknown Borel measurable function. Our goal is to estimate $m(x) = E(Y \mid X = x)$. Denoting $\varepsilon_i = Y_i - m(X_i)$ we get the regression model $Y_i = m(X_i) + \varepsilon_i$. Note that by construction $E[\varepsilon_i \mid X_i] = 0$. The regression function $m(.)$ takes the form: $m(x) = \int \frac{yf(x,\ y)}{f(x)} dx$, if $f(x) > 0$ and the marginal density of $f(x,\ y)$ becomes: $f(x) = \int f(x,\ y)dy$. The form of the (internalized) kernel regression estimator, developed by Nadaraya (1964) and Watson (1964), is:

$$\hat{m}_h(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{K_h(x - X_i)Y_i}{\hat{f}_h(X_i)} \text{ where } \hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n}K_h(x - X_i) \tag{2}$$

where $K_h(x-X_i)$ denotes $K\left(\frac{x-X_i}{h}\right)\frac{1}{h}$ and $K:\mathbb{R}^d\to\mathbb{R}$ is the *a priori* chosen *d*-dimensional (Lipschitz) continuous called kernel, whose compact support satisfies

$$\int|K(x)|\ dx<\infty,\ \int K(x)dx=1 \tag{3}$$

The kernel function used to estimate both null and alternative models is the Quartic kernel given by $K(\mathbf{x})=\frac{15}{16}(1-\mathbf{x}^2)^2$ for $|\mathbf{x}|<1$. In this work all procedures are stated in terms of this kernel. The results are very similar if we use other kernels. The smoothing parameter $h_n>0$ for all *n* is an *a priori* chosen sequence of numbers called *bandwidth* parameter that satisfies

$$\lim_{n\to\infty}h_n=0,\ \lim_{n\to\infty}nh_n^d=\infty \tag{4}$$

Provided that $\sup_x f(x)<\infty$, then the estimator $\hat{f}(x)$ is pointwise weakly consistent in every continuity point of $f(x)$. It results from the property of bounded convergence of $E\hat{f}_h(x)$, $E\hat{m}(x)\hat{f}_h(x)$, $nh_n^d\ \mathrm{var}(\hat{f}_h(x))$ and $nh_n^d\ \mathrm{var}(\hat{m}(x)\hat{f}_h(x))$. Then by (3) and (4) and provided that $\sup_x|m(x)|\ f_h(x)<\infty$ and $\sup\sigma_y^2(x)f_h(x)<\infty$ we have that $\hat{m}(x)\to m(x)$ in probability. If in addition $E(Y_j^4)<\infty$, $\sup\sigma_y^4(x)f_h(x)<\infty$ and $\sum_{n=1}^{\infty}n^{-2}h_n^{-3}<\infty$ then $\hat{m}(x)\to m(x)$ *a.s.*

The asymptotic distribution theory is established for the kernel regression approach, for details Bierens (1994). Suppose that $\sigma_u^2(x)f(x)$ are continuously and uniformly bounded, and $m^2(x)f(x)$ are continuously and uniformly bounded too; additionally, suppose that $\int xK(x)dx=0$, $\int xx^TK(x)dx=\Psi<\infty$ and $f_h(x)>0$. If $\lim_{n\to\infty}h_n^2\sqrt[2]{nh_n^d}=\mu$ with $0\le\mu<\infty$ then:

$$\sqrt[2]{nh_n^d}[\hat{m}(x)-m(x)]\to N\left\{\mu\frac{b(x)}{f_h(x)},\ \frac{\sigma_u^2(x)}{f_h(x)}\int K^2(z)dz\right\}$$

in distribution, where $b(x)=\frac{1}{2}tr\left\{\Psi H_{m(x)f(x)}\right\}-\frac{1}{2}f(x)tr\left\{H_{f(x)}\right\}$, *H* is the Hessian of the functions mf and f, respectively (Y was defined before). To selecth bandwidth, we define the usual mean integrated square error (MISE)

$$E\left\{\int[\hat{m}(x)\hat{f}_h(x)-m(x)f(x)]^2dx\right\} \tag{5}$$

which yields the optimal bandwidth of the form $\lim_{n\to\infty}h_n^2\sqrt[2]{nh_n^d}=\mu>0$. Thus, the band h_n, which gives the maximum rate of convergence in distribution is $h_n=h_0n^{-\frac{1}{d+4}}$, where h_0 is a constant. If we set $h_n=cn^{-\frac{1}{d+4}}$, then we have that: $n^{\frac{2}{d+4}}\left(\hat{m}(x)-m(x)\right)\to N\left\{c^2\frac{b(x)}{f(x)},\ c^{-d}\frac{\sigma_\varepsilon^2(x)}{f(x)}\int K^2(z)dz\right\}$. Note that the asymptotic rate of convergence in distribution, $n^{\frac{2}{d+4}}$, has a functional form and is similar to the optimal bandwidth h_n but they are very different quantities. Also note that $n^{\frac{2}{d+4}}$ is negatively related to the number of regressors. This feature is typical of nonparametric regression and is known as *curse of dimensionality.* Additive models are a good way to address this problem. More importantly, the optimal bandwidth for estimating is different from the bandwidth for generation the bootstrap samples.

2.2. *Additive Models and Marginal Integration*

Additive models can be presented as follows. Let $\left\{(X_{i,}Y_i)\right\}_{i=1}^{n} \in \mathbb{R}^{d+1}$ be a finite sequence of random vectors, $\Lambda = \{1,..., \ d\}$, $m : \mathbb{R}^d \to \mathbb{R}$, $m_\alpha : \mathbb{R} \to \mathbb{R} \ \ \forall \alpha \in \Lambda,$ are unknown Borel measurable functions. Then:

$$E[Y \mid X] = F\left\{\sum_{\alpha\in\Lambda} m_\alpha(X_\alpha)\right\} \quad \alpha \in \Lambda \tag{6}$$

is called Generalized Additive Models (GAM) which was considered by Winsberg and Winsberg and Ramsay (1980). In this work we are concerned with a less general structure. Let $E\left[Y^2\right] < \infty$ be and let *m* be the regression function, so that $m(x) = E(Y \mid X)$. If *m* is additive we can write it as follows:

$$E\left[Y \mid X\right] = m(x) = c + \sum_{\alpha\in\Lambda} m_\alpha(X_\alpha) \tag{7}$$

Even if *m* is not genuinely additive; an additive approximation to *m* may be sufficiently accurate for a given application as well as being readily interpretable. If *m* is additive, of course, then $m^* = m$ and $m^*_\alpha = m_\alpha$ for $\alpha \in \Lambda$ with $E_{X_\alpha}\left\{m_\alpha(X_\alpha)\right\} = \int m_\alpha(x) f_\alpha(x) dx = 0,$ $\forall \alpha \in \Lambda,$ and $E\left\{Y\right\} = E\left\{m_{-\alpha}(X_{-\alpha})\right\} = c$ for identification. Stone (1985) explains that the optimal rate to estimate such regression curves *m* is the one-dimensional rate of convergence with $n^{-\ell/(2\ell+1)}$ and does not increase with dimension. The marginal integration estimator is defined noting that (see Hardle, Müller, Sperlich and Werwatz, 2004):

$$E_{X_{-\alpha}}\{m(x_\alpha, \ X_{-\alpha})\} = \int m(x_\alpha, \ x_{-\alpha}) f_{-\alpha}(x_{-\alpha}) dx_{-\alpha} = c + m_\alpha(x_\alpha) \tag{8}$$

Example 1 *Consider the following data generation process, let* $\mathbf{X} \in \mathbb{R}^3$ *we define:*

$$Y = 1 + X_1^2 + 2X_2 + 3\cos\left(X_3\pi\right) + \varepsilon \tag{9}$$

where $X_\alpha \sim U[-1, \ 1] \ \ \forall \alpha \in \Lambda = \left\{1, \ 2, \ 3\right\}$, *are independent,* $E[\varepsilon \mid X = \mathbf{x}] = 0$ *a.s and possibly heteroscedastic. We have:*

$$m(X) = E\left[Y \mid X = \mathbf{x}\right] = 1 + X_1^2 + 2X_2 + 3\cos\left(X_3\right)$$

Consequently, taking the marginal expectation, then:

$$E_{X_{-1}}\left\{m(X)\right\} = \int_{-1}^{1}\int_{-1}^{1}\left(1 + X_1^2 + 2u + 3\cos\left(v\pi\right)\right) \ \frac{1}{4} du dv = 1 + X_1^2$$

$$E_{X_{-2}}\left\{m(X)\right\} = \int_{-1}^{1}\int_{-1}^{1}(1 + u + 2X_2 + 3\cos\left(v\pi\right))\frac{1}{4} du dv = 1 + 2X_2$$

$$E_{X_{-3}}\left\{m(X)\right\} = \int_{-1}^{1}\int_{-1}^{1}(1 + u + 2v + 3\cos\left(X_3\pi\right))\frac{1}{4} du dv = 1 + 3\cos\left(X_3\pi\right)$$

we get the component functions $m_1(X_1) = 1 + X_1^2$, $m_2(X_2) = 2X_2 - 1$ *and* $m_3(X_3) = 3\cos\left(X_3\pi\right) - 1$ *and* $c = 2$ *. Then we can see that we always get the marginal effects of the explanatories variable X plus a constant* $c = 2$.

The corresponding internalized version of marginal integration estimator, and denoted as m^I, is given by:

$$\hat{m}_\alpha^I(x_\alpha) = \frac{1}{n}\sum_{i=1}^{n} K_h(X_{i\alpha} - x_\alpha)\hat{f}_{\alpha|-\alpha}^{-1}(X_\alpha \mid X_{-\alpha})Y_i \quad (10)$$

Where $\hat{f}_{\alpha|-\alpha}^{-1}$ is an estimate of the inverse of the conditional density $f_{\alpha|-\alpha}(X_\alpha \mid X_{-\alpha})$. Under null hypothesis of additivity $\hat{m}_\alpha$ and $\hat{m}_\alpha^I$ are consistent estimates of m_α with $\alpha \in \Lambda$. For the sake of simplicity, we assume that the constant $c_n = c + O_p\left(n^{-1/2}\right)$, for instance $\hat{c} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. In the internalized version $\hat{f}_h$ appears internally to the summation, in opposition to the external method where estimated density appears to be external to the summation (see Jones, Davies and Park, 1994).

2.3. *Testing Additivity*

In this section we are going to investigate several tests statistics in order to testing the additivity hypothesis under the null and alternative hypotheses of non-additivity, but we will focus on statistics based on residuals coming from an internal marginal integration (IMIE). Let $\Xi^a \cup \Xi^r$ be the potential random sample outcomes, where Ξ^r is the rejection region, and $\Xi^a \cap \Xi^r = \emptyset$, where $\emptyset$ is the empty set. Let $\mathfrak{D}$ be a family of additive models:

$$\mathfrak{D} = \left\{ m \in \mathbb{R} : m(x) = \sum_{\alpha \in \Lambda}\left[\int m(x_\alpha, x_{-\alpha}) f_{-\alpha}(x_{-\alpha}) dx_{-\alpha}\right]\right\} \quad [11] \quad (11)$$

Then the hypothesis is represented as follows:

$H_0 : m \in \mathfrak{D}$

$H_1 : m \notin \mathfrak{D}$

Note that if $x \in \Xi^r \Rightarrow$ we will reject Ho and if $x \notin \Xi^r \Rightarrow$ we will not reject Ho. The regression estimator based on the IMIE is defined by:

$$\hat{m}_0^I(x) = \sum_{\alpha \in \Lambda} \hat{m}_\alpha^I(x_\alpha) + (d-1)\ \hat{c} \quad (12)$$

where $\hat{m}_\alpha^I$ is given by the expression (10), and the residuals for this regression function are defined by:

$$\hat{e}_i = Y_i - \hat{m}_0^I(X_i) \quad (13)$$

The estimated alternative model is the multidimensional Nadaraya-Watson estimator, given by (2), with bandwidth *k*. For instance one from the three statistical tests considered here, which is based on the estimators under null alternative hypothesis, is given by:

$$T = \int (\hat{m}_k(x) - \hat{m}_0^I(x))^2 w(x) dx, \quad (14)$$

$$\hat{T}^* = \int (\hat{m}_k^*(x) - \hat{m}_0^{*g,\,h}(x))^2 w(x) dx, \quad (15)$$

where $w(x)$ denotes a weight function. This one serves to trim the boundaries or regions of sparse data. The statistic T is the unknown test value, because $m_h(x)$ is unknown; $\hat{T}$ is the original estimated test, and $\hat{T}^*$ is B bootstrap test; $\hat{m}_k$ is the general model estimator; $\hat{m}_0^h$ is the null model; $\hat{m}_k^*$ and $\hat{m}_0^{*\,g,\,h}$ are general and null models, respectively, based in bootstrap residual. The subscripts g, h tell us that $\hat{m}_0^{*\,g,\,h}$ was generated with g and estimated with the optimal bandwidth h. Of course, this selection was made using the rule that we are proposing as a criterion and will be justified in next section. According to that, we have that $\hat{g} > h_{opt}$ is satisfied. For instance, the sample version of statistical data T_n and $\hat{T}^*$ is given by the formula:

$$\hat{T}_{1n}^* = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{m}^{*\,I}(X_i) - \hat{m}_{0,\,h}^{*\,I}(X_i)\right]^2 \tag{16}$$

Another two tests statistics considered are:

$$\hat{T}_{2n} = \frac{1}{n}\sum_{i=1}^{n}\hat{e}_i\left[\hat{m}^{*\,I}(X_i) - \hat{m}_{0,\,h}^{*\,I}(X_i)\right] \tag{17}$$

$$T_{3n} = \frac{1}{n}\sum_{i=1}^{n}\left[(\hat{e}_i)^2 - (\hat{u}_i)^2\right] \tag{18}$$

where $\hat{e}_i = Y_i - \hat{m}_0^I(X_i)$ are the residuals under H_0. The estimated $\hat{T}_{2n}$ was introduced by Gozalo and Linton (2001). The statistic $\hat{T}_{n3}$ is essentially the difference of the estimators for the integrated variance function in the additive and general models –it was firstly proposed by Dette, von Lieres and Sperlich (2003) to test parametric structures of the regression function. In this estimate, the random variable $\hat{u}_i = Y_i - \hat{m}_k^I(X_i)$ denotes the corresponding residuals of the alternative model. Concerning the asymptotic distribution of (16)-(18) Dette, von Lieres and Sperlich (2003) have shown that under null hypothesis of additivity, *i.e.* $m \in \mathfrak{D}$, as $n \to \infty$, then $ng^{\frac{d}{2}}(T_{jn} - E\{T_{jn}\}) \to N(0,\ v_j^2)$ for $j = 1,\ 2,\ 3$. Moreover, if $m \in \mathfrak{D}$ they also show that $\sqrt[2]{n}(T_{jn} - M_j^2 - E\{T_{jn}\}) \to N(0,\ s_j^2)$, where v_j^2 and s_j^2 are the asymptotic variances and M_j^2 is a nonnegative measure of discrepancy.

3. The Resampling Problem

Let $\hat{m}_0^I = \hat{c} + \sum_{\alpha\in\Lambda}\hat{m}_\alpha(x_\alpha)$ be the estimation under null hypothesis. And denote by $\hat{m}_g^I$ the additive model estimator with bandwidth g. The approach is to be resampled from the estimated residual:

$$\hat{\varepsilon}_i = Y_i - \hat{m}_g^I(X_i),\ i = 1,\ldots,\ n \tag{19}$$

They are the differences between the observations and the regression function estimated under a null hypothesis, whose distributions will approximate the distribution of the original estimator. Here we use the idea of *wild bootstrap* (see Wu, 1986), *i.e* we used the so-called *golden cut construction*. This method is called *wild bootstrap* because in some sense, as suggested by Härdle and Marron (1991), the resampling distribution of ε_i^* can be thought of as attempting to reconstruct the distribution of each residual through the use of one single observation. After resampling, new observations are defined by

$$Y_i^* = \hat{m}_g^I(X_i) + \varepsilon_i^* \tag{20}$$

Let Ω be the sample space and $\varepsilon^* : \Omega \to \{a,\ b\}$ each bootstrap residual. Note that each ε^* is taken from the two points distribution, denoted for $G_{\{a,\ b\},\ i}$ such that:

$$\begin{aligned} P\{\varepsilon^*(\omega) = a\} &= \gamma \\ P\{\varepsilon^*(\omega) = b\} &= 1-\gamma \end{aligned} \tag{21}$$

for some $a, b \in \mathbb{R}$, and $\gamma \in [0,\ 1]$. Determine now a, b and γ subject to the restrictions:

$$E_{G_{\{a,\ b\},\ i}}\{\varepsilon_i^*\} = 0,\ \ E_{G_{\{a,\ b\},\ i}}\{(\varepsilon_i^*)^2\} = (\hat{\varepsilon}_i^2) \text{ and } E_{G_{\{a,\ b\},\ i}}\{(\varepsilon_i^*)^3\} = (\hat{\varepsilon}_i)^3 \tag{22}$$

For a two-point distribution $G_{\{a,\ b\},\ i}$, set $G_{\{a,\ b\},\ i} = \gamma\delta_a + (1-\gamma)\delta_b$ where δ_a and δ_b denote points measure at $\{a,\ b\}$. Some algebraic rule reveals that parameters a, b, and γ are given by: $a = \hat{\varepsilon}_i(1-\sqrt{5})/2$, $b = \hat{\varepsilon}_i(1+\sqrt{5})/2$ and $\gamma = (5+\sqrt{5})/10$, and this one satisfies the restrictions. Bandwidth for constructing the bootstrap sample, denoted by g, is different from the optimal bandwidth h, where

$$h = \min_{h \in H_n} E\left\{\int [\hat{m}(x)\hat{f}_h(x) - m(x)f(x)]^2 dx\right\} \tag{23}$$

where H_n is a set of possible bandwidths. The sample version of (23) is given by: $\frac{1}{n}\sum_{i=1}^n \{Y - \hat{m}_{h,-i}(X_i)\}^2$. This procedure to estimate h is known as cross-validation and gives us some idea about how to choose g, which has to be larger than the optimal bandwidth. Then, the kernel smoother m_h is applied to the bootstrapped data $\{(X_i, Y_i^*)\}_{i=1}^n \in \mathbb{R}^{d+1}$ using optimal bandwidth h. This estimation is denoted by $\hat{m}_h^*$. A number of replications of $\hat{m}_h^*$ can be used as a basis to construct the statistics, as the distribution of $\{\hat{m}_h(x) - m(x)\}$ is approximated by the distributions of $\{\hat{m}_h^*(x) - \hat{m}_g(x)\}$ which we can simulate. The symbol $Y \mid X$ to denote the conditional distribution of $\{Y_i\}_{i=1}^n \mid \{X_i\}_{i=1}^n$ and the symbol * , the bootstrap distribution $\{Y_i^*\}_{i=1}^n \mid \{(X_i,\ Y_i)\}_{i=1}^n$.

Why bandwidth g, used in the construction of the bootstrap residual should be over-smoothed? Consider the mean of $\{\hat{m}_h(x) - m(x)\}$ under the $Y \mid X$ -distribution and $\{\hat{m}_h^*(x) - \hat{m}_g(x)\}$ under the * -distribution in the situation when the marginal density $f(x)$ is constant in neighborhood of x. Asymptotic analysis as in Rosenblatt (1969) shows that:

$$E^{Y|X}(\hat{m}_h(x) - m(x)) \approx h^2 \frac{\mu(K)}{2} m''(x) \tag{24}$$

$$E^*(\hat{m}_h^*(x) - \hat{m}_g(x)) \approx h^2 \frac{\mu(K)}{2} \hat{m}_g''(x) \tag{25}$$

(moreover we can see that $\frac{g}{h} \to 0$ for $n \to \infty$: this finding results from bootstrap consistency) where $\mu(K) = \left(\int u^2 K(u) du\right)$. Hence for these two distributions to have the same bias, we need $\{\hat{m}_g''(x) - m''(x)\} \to 0$. This requires choosing g going to zero at a slower rate than the *optimal bandwidth h* for estimating $m(x)$. (See Gasser and Müller, 1984 for details). Suppose that:

A1. $m(x)$, $f(x)$ and $\sigma^2(x) = E\{((Y - E(Y)) \mid X = x)\}^2$ are twice continuously differentiable.

A2. The kernel function K is symmetric and nonnegative, $\|K\| = \int K^2(u) \; du < \infty$ and $\mu(K) = \int u^2 K(u) \; du < \infty$.

A3. $\sup_x E\{\varepsilon^2 \mid X = x\} < \infty$.

A4. $f(x_i) > 0$ for $i \in \{i, \ldots, n\}$.

A5. $m(x)$, and $f(x)$ are four times continuously differentiable.

A6. K is twice continuously differentiable.

Under A1-A2, a reasonable choice for h will fall in the set

$$H_n = \left[h_0 n^{-1/(4+d)}, h_1 n^{-1/(4+d)}\right], 0 < h_0 < h_1 < \infty$$

For this choice of bandwidth, the kernel smoother $\hat{m}_h(x)$ is asymptotically optimal. For h_0 small and h_1 large, this assumption is not restrictive because it will be satisfied with probability to 1 if h is chosen by cross-validation. The rate of convergence of g must tend to zero at a slower rate than h. Hence it is assumed that g is chosen from the set $G_n = \left[n^{-1/(4+d)+\delta}, \; n^{-\delta}\right]$, $\delta > 0$. The following results characterize uniform convergence in h and g, in the spirit of Härdle and Marron (1991).

Let $(\Omega, \mathfrak{F}, P)$ be a fixed probability space, where $\mathfrak{F}$ is a non-empty collection of subsets of Ω, $x : (\Omega, \mathfrak{F}) \to (\mathbb{R}, \mathfrak{B})$, where $\mathfrak{B}$ is the Borel field generated by x. Then $\Phi_{\alpha,\,h}$ and $\Phi^*_{\alpha,\,h,\,g}$ should be random functions such that:

$$\sup_{h \in H_n} \sup_{g \in G_n} \left|\Phi_{\alpha,\,h}(z) - \Phi^*_{\alpha,\,h,\,g}(z)\right| \tag{26}$$

is a random variable defined for each n. Suppose that there is a null set N and $n \geq n_0(\omega, \; \varepsilon)$, $n \in \mathbb{N}$, both independent of h and g, where:

$$\Phi_{\alpha,\,h}(z) = P^{Y|X}\left\{\sqrt{nh}\left[\hat{m}_{\alpha,\,h}(x_\alpha) - m_\alpha(x_\alpha)\right] < z\right\} \tag{27}$$

$$\Phi^*_{\alpha,\,h,\,g}(z) = P^*\left\{\sqrt{nh}\left[\hat{m}^*_{\alpha,\,h}(x_\alpha) - \hat{m}_{\alpha,\,g}(x_\alpha)\right] < z\right\} \tag{28}$$

for a fixed z. Notice that $P^{Y/X}$ is the conditional distribution and P^* is the bootstrap distribution. On the other hand, it is important to keep in mind that the symbol Φ_α is to denote the conditional distribution of Y given $X_\alpha \in \mathbb{R}$ and $\Phi = \sum_\alpha \Phi_\alpha$ is to denote the conditional distribution of Y given $X \in \mathbb{R}^d$. In a similar way we get that $\Phi^* = \sum_\alpha \Phi^*_{\alpha,h,g}$

From Theorem 1 of Härdle and Marron (1991) we have that under A1-A4, $\forall z \in \mathbb{R}^n$ and $\forall \alpha \in \Lambda$

$$\sup_{h \in H_n} \sup_{g \in G_n} \left| \Phi_{\alpha,h}(z) - \Phi^*_{\alpha,h,g}(z) \right| \to 0 \tag{29}$$

This result tells us that under specified conditions $\Phi_{\alpha,h}(z)$ converge uniformly in h and g to $\Phi^*_{\alpha,h,g}(z)$ *almost sure*. The assumption on the speed of the bandwidth h ensures that each of the previous probabilities has a non-trivial limit. In fact, the result comes from showing that both $\sqrt{nh}\left[\hat{m}_{\alpha,h}(x_\alpha) - m_\alpha(x_\alpha)\right]$ and $\sqrt{nh}\left[\hat{m}^*_{\alpha,h}(x_\alpha) - \hat{m}_{\alpha,g}(x_\alpha)\right]$ have the same limiting normal distribution.[1] In other words, the result tells us that samples of wild bootstrap regression estimates $\hat{m}^*_{\alpha,h}$ centered around $\hat{m}_{\alpha,g}$ have nearly the same distribution as the regression function $\hat{m}_{\alpha,h}$ centered around m_α. It is important to say that the marginal density f is constant in a neighborhood of x and with a fixed z.

The main role of the pilot smooth is to provide a correct adjustment for the bias, in most of works the goal of bias estimation is used as a criterion. Recall that a bias in the estimation of $m_\alpha(x_\alpha)$ by $\hat{m}_{\alpha,h}(x_\alpha)$ is given by $b_{\alpha,h}(\hat{m}(x_\alpha)) = \{E^{Y|X}\hat{m}^*_{\alpha,h}(x_\alpha) - m_\alpha(x_\alpha)\}$. The bootstrap bias of the estimator constructed from the resampled data is:

$$\hat{b}_{\alpha,h,g}(\hat{m}(x_\alpha)) = \{E^*[\hat{m}^*_{\alpha,h}(x_\alpha)] - \hat{m}_{\alpha,g}(x_\alpha)\}$$

As usual, to find g it is necessary to minimize the mean square error. Again, from Härdle and Marron (1991) we have that under assumption A1-A6, along almost all sample sequences,

$$E\{(b_{\alpha,h}(x_\alpha) - \hat{b}_{\alpha,h,g}(x_\alpha))^2 \mid \{X_{\alpha i}\}_{i=1}^n\} \sim h^4\{An^{-1}g^{-5} + Bg^4\}$$

in the sense that the ratio tends in probability to 1, where A and B are constants with respect to n, h and g. A consequence of such a result is that differentiating the previously expected value with respect to g, the rate of convergence for $d = 1$ of g should be close to $n^{-1/9}$.

This makes precise the previous remark that indicated that g should be oversmoothed. Moreover, in general terms, the optimal rate for estimating the functions m and m_α and their derivatives for the additive models depends on the dimension d, the smoothness index of m, denoted by ℓ, the order of its derivatives, v, and the order of local polynomials used in the estimation, p. Sperlich and Severance Lossin (1999), showed that for $d = 1$, $\ell = 2$, $v = 0$ and, $p = 1$, the corresponding bandwidth is close to $n^{-\frac{1}{9}}$.

Under these assumptions, a reasonable choice of h will be optimal. This result gives us some indication on why g should be selected bigger than h. However, even theoreti-

[1] The reason for the uniform convergence in the previous result is important because this ensures that results still hold when h or g are replaced by random-driven data bandwidths.

cally band selection for the bootstrap is made as literature suggests; the question, not less important, is how, in the practice, we can do it. Since it is true that rate $n^{-1/9}$ is the optimal one, the selection of g involves a constant that multiplies the rate $n^{-1/9}$. Our goal is to provide a method to select a bandwidth, which is reasonable and can be easily applied. When we say reasonable, we mean that the bandwidth chosen gives us a good approximation to the original distribution of the bootstrap distribution functions.

Let h be the argument that minimizes the averages of cross-validation. For $d = 1$ we define the oversmoothing bandwidth $\hat{g} = hn^{-\frac{1}{9}}$. The following result tells us that if for some X_α the expression (26) holds, then for entire vector $X = (X_1, ..., X_d)^T$ the uniform convergence is satisfied. Thus we have:

Proposition 2 *Given A1-A6, if* (29) *holds, then* $\forall z \in \mathbb{R}^n$ *along almost all sample sequences we have*

$$\sup_{h \in H_n} \sup_{g \in G_n} \left| \Phi_h(z) - \Phi^*_{h,\, g}(z) \right| \to 0 \tag{30}$$

where

$$\Phi_h(z) = P^{Y|X} \left\{ \sqrt{nh^d} \left[\hat{m}_h(x) - m(x) \right] < z \right\}$$

$$\Phi^*_{h,\, g}(z) = P^* \left\{ \sqrt{nh^d} \left[\hat{m}^*_h(x) - \hat{m}_g(x) \right] < z \right\}$$

Proof. See appendix 2.

4. Simulation Results

The bootstrap procedure allows us to obtain critical values using $\hat{T}_{jn}$ for $j = 1,2,3$. This works as follows: i) generate independent $\{\varepsilon_i^*\}_{i=1}^n$, where distribution of ε_i^* is given by a two-point distribution; ii) construct the bootstrap data $\left\{(X_i, Y_i^*)\right\}_{i=1}^n$ and with this data compute the bootstrap statistics $\hat{T}^*_{jn}$; iii) repeat the process B times to obtain $\{T^*_{jn,\, b}\}_{b \in \beta}$ with $\beta = \{1, ..., B\}$, and use these B values to construct the empirical bootstrap distribution; iv) we use this empirical distribution to compute the empirical p-values. We reject H_0 if $\hat{T}^*_{jn} > \hat{T}_{jn}$. The bootstrap p-value[2] is computed as:

$$p_B = \frac{\left| \{ \mathbf{x} : \hat{T}^*_{jn,\, b} > \hat{T}_{jn,\, b},\ \ b \in \beta \} \right|}{B} \tag{31}$$

where $|.|$ denotes the cardinality of the set $\{\hat{T}^*_{jn,\, b} : \hat{T}^*_{jn,\, b} > \hat{T}_{jn}\}$. In our simulations we take $d = 3$ and consider performance differences between tests $\hat{T}_{1n}$ and $\hat{T}_{3n}$. Since IMIE allows oversmoothing in the nuisance directions, we set $b = qh$, with $q \in \mathbb{R}$. We choose

[2] Theoretically we define the p-value as: $p = \min_{\alpha} \left\{ \sup_{m \in \mathfrak{F}} P\left(x \in \Xi^r(\alpha);\ m_0 \in \mathfrak{F}\right) = \alpha \right\}$

q as in Dette, von Lieres and Sperlich (2003), and it means that $q = 6$ for $d = 3$. The cross-validation yields for the Nadaraya-Watson an optimal bandwidth $k = 0.9$, for IMIE an optimal bandwidth $h = 0.7$ and for nuisance directions $b = 6h$. Bandwidth selection for bootstrap estimation is carried out as the suggested criterion 1.

We report percentaje of the rejections for the 1%, 5%, 10% and 15% levels for all tests, without trimming (tr0) at the approximate 95% quantile (tr5) and 90% quantile (tr10) under null hypothesis of additivity. We report simulations for $n = 100$, $B = 500$ and results refer to 1000 simulations runs with a randomly drawn design for each run. Consider the following model with an interaction term lX_2X_3, where $l \in \{0,\ 2\}$ is an index that tells us if the model is null ($l = 0$) or alternative ($l = 2$), i.e.

$$\begin{aligned} &Y = X_1 + X_2^2 + 2\ \sin(\pi X_3) + lX_2X_3 + \varepsilon, \quad \text{with } \varepsilon \sim N(0,\ 1) \\ &X \sim N(0,\ \Sigma_\gamma),\ \ \Sigma_\gamma = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{21} & 1 & \rho_{23} \\ \rho_{31} & \rho_{32} & 1 \end{pmatrix},\ \ \gamma \in \{1,\ 2\} \end{aligned} \tag{32}$$

where $\gamma = 1$ means that there is no correlation between covariates, and $\gamma = 2$ means that covariates are correlated. Firstly, we carry out a simulation with $\gamma = 1$ thus $\rho_{12} = \rho_{21} = \rho_{13} = \rho_{31} = \rho_{23} = \rho_{32} = 0$. Secondly, we consider the correlated design, where $\gamma = 2,\ \rho_{12} = \rho_{21} = 0.2,\ \rho_{13} = \rho_{31} = 0.4,$ and $\rho_{23} = \rho_{32} = 0.6.$

In order to assess the validity of the statistical tests introduced above we set an analysis based on the p-value concept. p-value is defined as the smallest possible level of significance at which the null hypothesis will be rejected for the computed statistical test. Therefore, if we have a level of significance equal to α, we want the p-value to be close to this value under H_0 and to be equal to one under H_1.

Tables 1 and 2 show the results of uncorrelated design (i.e. $\gamma = 1$) under H_0, with optimal bandwidth h and band g for the bootstrap. We obtain that the IMIE procedure yields too conservative tests, but holds the level for all tests when no trim is applied and $g = h_{opt}$. However the nominal level is not accurate enough. When analyzing the results for a correlated design (i.e, $\gamma = 2$), we mimic the same results, but tests are still more conservative than in the previous setting. See Tables 3 and 4. It is remarkable that T_1 performs better than T_2 and T_3 for all significance levels in both situations. Results under H_1 are given in tables 5 and 6 when $\gamma = 1$. Tables 7 and 8 present the results for $\gamma = 2.$ Under the alternative, we get reasonable power if there is no trim and $g = h_{opt}$. As before, we can see that all tests work worse for correlated design. With both H_0 and H_1, we obtain better results for all tests under consideration when we do not trim and we rather choose $g = h_{opt}$

Note that with γ =1, l equals 0 and 2 for any level of trimming and all significance levels when the bandwidth for the bootstrap is oversmoothing, results are more conservative than when we use the optimal bandwidth. See tables 1, 2, 5 and 6. However, we can get a reasonable power level. We hope results under H_1 to be closer to 1 with the oversmoothing band, but this is not the case. Results show that is more difficult to reject H_0 when bandwidth g is bigger than the optimal one. With $\gamma = 2, l = 0$ and $g = h_{opt}$, and trim at 5% and 10% it is remarkable that T_1 has a good performance. For $\gamma = 2, l = 2$ and for any level of trimming and all significance levels, results are bad and the power of test are too low. Although the statistical procedure works better in the case of variables uncorrelated and without trim, results obtained indicate that the statistical procedure we have proposed works reasonably well.

5. Conclusion

In this work we obtain a bandwidth for testing a semiparametric model against a nonparametric alternative. The bandwidth posed by the alternative model still deserves some discussions. In this paper, we are not interested in local alternatives but in a fixed one. Then, bandwidth selection for the alternative, k, does not affect the result about estimation under the null hypothesis. Therefore, k can be any of them. In particular, we take the band chosen as an argument that minimizes the cross-validation average. Moreover, since for this paper purposes it is not necessary to make a consistent estimation of any parameter, either with H_0 being true or false, the problem refers only to the selection of bootstrap bandwidth g.

References

Bierens, H. J. (1994). Topics in Advanced Econometrics. Cambridge University Press.

Deaton, A., Muellbauer, J. (1980). Economics and Consumer Behavior. Cambridge University Press, Cambridge.

Dette, H., Von Lieres, C., Sperlich, S. (2003). A Comparison of Different Nonparametric Methods for Inference on Additive Models. Working Paper.

Gasser, T., Müller, H. G. (1984). "Estimating Regression Functions and Their Derivatives by the Kernel Method". *Scandinavian Journal of Statistics 11*, 171-185.

Gozalo, P. L., Linton, O. B. (2001) "Testing Additivity in Generalized Nonparametric regression models with Estimated Parameters". *Journal of Econometrics 104*, 1-48.

Härdle, W., Marron, J. S. (1991). "Bootstrap Simultaneous Error Bars for Nonparametric Regression". *Annals of Statistics 19*, 778-796.

Härdle, W., Müller, M., Sperlich, S., Werwatz, A. (2004). An Introduction to Non-and Semi-parametric Models. Springer Series in Statistics. Springer-Verlag.

Jones, M. C., Davies, S. J., Park, B. U. (1994). "Versions of Kernel-Type Regression Estimators". *Journal of the American Statistical Association 89*, 825-832.

Linton, O. B., Nielsen, J. P. (1995). "A Kernel Method of Estimating Structured Nonparametric regression Based on Marginal Integration". *Biometrika 82,* 93-101.

Nadaraya, E. A. (1964). On Estimating Regression. *Theory of Probability and it's Applications 10*, 186-190.

Winsberg, S., Ramsay, J.O. (1980). "Monotonic Transformation to Additivity Using Splines". *Biometrika 67*, 669-674.

Rosenblatt, M. (1969). Conditional Probability Density and Regression Estimators in Krishnaiah (Ed.). *Multivariate Analysis II*. 25-31. Academic Press: New York.

Sperlich, S., Tjostheim, D., Yang, L. (2002). "Nonparametric estimation and Testing of Interaction in Additive Models". *Econometric Theory 18*, 197-251.

Severance-Lossin, E., Sperlich, S. (1999). Estimation of Derivatives for Additive Separable Models, Statistics 33: 241-265.

Stone, C. J. (1985). "Additive Regression And Other Nonparametric Regression Models". *Annals of Statistics 13*, 689-705.

Watson, G. S. (1964). Smooth Regression Analysis. *Sankhyâ Ser*. A, 26: 359-372.

Wu, C. (1986). "Jackknife, Bootstrap and Other Resampling Methods in Regression analysis" (with discussion), *Annals of Statistics 14*, 1261-1350.

Appendix 1.

Table 1. Percentage of rejection. $\gamma = 1.\ l = 0$ **Table 2.** Percentage of rejection. $\gamma = 1.\ l = 0$

$g = h_{opt}$	α	T_1	T_2	T_3	$g > h_{opt}$	α	T_1	T_2	T_3
tr0	15%	0.17	0.16	0.15	tr0	15%	0.06	0.01	0.006
	10%	0.09	0.075	0.06		10%	0.02	0.006	0.002
	5%	0.02	0.016	0.01		5%	0.004	0.002	0.001
	1%	0.001	0.0	0.0		1%	0.0	0.0	0.0
tr5	15%	0.08	0.05	0.05	tr5	15%	0.03	0.001	0.0
	10%	0.03	0.023	0.018		10%	0.01	0.0	0.0
	5%	0.008	0.005	0.004		5%	0.001	0.0	0.0
	1%	0.001	0.0	0.0		1%	0.0	0.0	0.0
tr10%	15%	0.079	0.053	0.04	tr10	15%	0.04	0.0	0.0
	10%	0.03	0.016	0.01		10%	0.02	0.0	0.0
	5%	0.014	0.007	0.006		5%	0.002	0.0	0.0
	1%	0.0	0.0	0.0		1%	0.0	0.0	0.0

Table 3. Percentage of rejection. $\gamma = 2.\ l = 0$ **Table 4.** Percentage of rejection. $\gamma = 2.\ l = 0$

$g = h_{opt}$	α	T_1	T_2	T_3	$g > h_{opt}$	α	T_1	T_2	T_3
tr0	15%	0.06	0.05	0.04	tr0	15%	0.03	0.01	0.001
	10%	0.02	0.02	0.02		10%	0.01	0.02	0.0
	5%	0.008	0.006	0.003		5%	0.04	0.0	0.0
	1%	0.001	0.0	0.0		1%	0.0	0.0	0.0
tr5	15%	0.03	0.02	0.01	tr5	15%	0.01	0.001	0.0
	10%	0.01	0.007	0.006		10%	0.06	0.0	0.0
	5%	0.04	0.003	0.0		5%	0.003	0.0	0.0
	1%	0.0	0.0	0.0		1%	0.0	0.0	0.0
tr10%	15%	0.03	0.023	0.016	tr10	15%	0.019	0.01	0.001
	10%	0.01	0.009	0.007		10%	0.01	0.01	0.001
	5%	0.03	0.001	0.001		5%	0.002	0.01	0.0
	1%	0.001	0.0	0.0		1%	0.0	0.0	0.0

Table 5. Percentage of rejection. $\gamma = 1.\ l = 2$ **Table 6.** Percentage of rejection. $\gamma = 1.\ l = 2$

$g = h_{opt}$	α	T_1	T_2	T_3	$g > h_{opt}$	α	T_1	T_2	T_3
tr0	15%	0.98	0.98	0.98	tr0	15%	0.95	0.87	0.68
	10%	0.95	0.96	0.94		10%	0.85	0.68	0.45
	5%	0.80	0.76	0.66		5%	0.58	0.35	0.13
	1%	0.30	0.18	0.07		1%	0.12	0.03	0.002
tr5	15%	0.94	0.93	0.88	tr5	15%	0.86	0.70	0.40
	10%	0.88	0.83	0.74		10%	0.75	0.50	0.23
	5%	0.71	0.59	0.40		5%	0.53	0.27	0.06
	1%	0.29	0.14	0.04		1%	0.16	0.03	0.001
tr10%	15%	0.89	0.85	0.75	tr10	15%	0.83	0.62	0.32
	10%	0.82	0.74	0.57		10%	0.72	0.47	0.18
	5%	0.66	0.52	0.29		5%	0.57	0.25	0.05
	1%	0.28	0.13	0.03		1%	0.20	0.03	0.003

Table 7. Percentage of rejection. $\gamma = 2$. $l = 2$

$g = h_{opt}$	α	T_1	T_2	T_3
tr0	15%	0.39	0.39	0.37
	10%	0.26	0.26	0.22
	5%	0.12	0.10	0.08
	1%	0.01	0.009	0.002
tr5	15%	0.40	0.38	0.32
	10%	0.30	0.26	0.19
	5%	0.17	0.12	0.07
	1%	0.03	0.01	0.004
tr10%	15%	0.40	0.36	0.30
	10%	0.30	0.25	0.17
	5%	0.18	0.12	0.07
	1%	0.03	0.01	0.002

Table 8. Percentage of rejection. $\gamma = 2$. $l = 2$

$g > h_{opt}$	α	T_1	T_2	T_3
tr0	15%	0.33	0.23	0.16
	10%	0.20	0.14	0.07
	5%	0.09	0.04	0.01
	1%	0.005	0.001	0.0
tr5	15%	0.26	0.15	0.07
	10%	0.18	0.09	0.02
	5%	0.09	0.02	0.006
	1%	0.01	0.001	0.0
tr10	15%	0.29	0.17	0.08
	10%	0.21	0.10	0.03
	5%	0.11	0.04	0.01
	1%	0.01	0.003	0.0

Appendix 2.

Proof. Theorem 1 of Härdle and Marron (1991) implies that along almost all samples, $\forall z \in \mathbb{R}^n$, and for some $\alpha \in \Lambda$, $\forall t \in \mathbb{R}^n$:

$$\left|\Phi_{\alpha,\, h}(z) - \phi_{B,\, V}\left(Z\right)\right| \to 0$$

uniformly over $h \in H_n$, where $Z = \frac{z - t^T B}{\sqrt[2]{t^T V t}}$, and $\phi_{B,\, V}$ denotes the univariate standard normal *c.d.f* with mean *B* and variance *V* and the same for the * distribution. Thus, we have that:

$$\sup_{h \in H_n} \sup_{g \in G_n} \left|\Phi_{\alpha,\, h}(z) - \phi_{B,\, V}(Z)\right| \to 0,$$

$$\sup_{h \in H_n} \sup_{g \in G_n} \left|\Phi^*_{\alpha,\, h,\, g}(z) - \phi_{B,\, V}(Z)\right| \to 0$$

It means that $Y \mid X$ and * distribution have the same asymptotic distribution. Then for $\forall z \in \mathbb{R}^n$, and $\alpha \in \Lambda = \{1, \ldots,\ d\}$ with $d < \infty$, we can define

$$\left\{\sqrt{nh^d}\,[\hat{m}_h(x) - m(x)] < z\right\} = \cup_{\alpha \in \Lambda} \left\{\sqrt{nh^d}\,[\hat{m}_{\alpha,h}(x_\alpha) - m_\alpha(x_\alpha)] < z\right\}$$

and

$$\left\{\sqrt{nh^d}\,[\hat{m}^*_h(x) - m(x)] < z\right\} = \cup_{\alpha \in \Lambda} \left\{\sqrt{nh^d}\left[\hat{m}^*_{\alpha,h}(x_\alpha) - \hat{m}_{\alpha,g}(x_\alpha)\right] < z\right\}$$

Therefore, $\Phi_h(z) \le \sum_{\alpha \in \Lambda} \Phi_{\alpha,\, h}(z)$ and $\Phi^*_{h,\, g}(z) \le \sum_{\alpha \in \Lambda} \Phi^*_{\alpha,\, h,\, g}(z)$. Notice that:

$$\left|\Phi_h(z) - \Phi^*_{h,\, g}(z)\right| \le \left|\Phi_{\alpha,\, h}(z) - \phi_{B,\, V}(Z)\right| + \left|\phi_{B,\, V}(Z) - \Phi^*_{\alpha,\, h,\, g}(z)\right|$$

consequently we have:

$$\begin{aligned}\sup_{h \in H_n} \sup_{g \in G_n} \left|\Phi_h(z) - \Phi^*_{h,\, g}(z)\right| &\le \sum_{\alpha \in \Lambda} \sup_{h \in H_n} \sup_{g \in G_n} \left|\Phi_{\alpha,\, h}(z) - \Phi^*_{\alpha,\, h,\, g}(z)\right| \\ &\le \sum_{\alpha \in \Lambda} \sup_{h \in H_n} \sup_{g \in G_n} \left|\Phi_{\alpha,\, h}(z) - \phi_{B,\, V}(Z)\right| \\ &\quad + \sum_{\alpha \in \Lambda} \sup_{h \in H_n} \sup_{g \in G_n} \left|\phi_{B,\, V}(Z) - \Phi^*_{\alpha,\, h,\, g}(z)\right|\end{aligned}$$

then it follows that:

$$\sup_{h \in H_n} \sup_{g \in G_n} \left|\Phi_h(z) - \Phi^*_{h,\, g}(z)\right| \to 0$$