8

Parametric and Nonparametric Encompassing Procedures

Christophe Bontemps, Jean-Pierre Florens and Jean-François Richard [1]

Abstract

This chapter studies the asymptotic behaviour of encompassing statistics : Situations where a parametric model is tested against another parametric model, are now well known, but the situations testing two nonparametric models or "crossed" situations where a parametric model is tested against a nonparametric one, haven't been treated previously, and are presented here.

In each of the four cases presented, the encompassing statistic is based on the difference, (properly normalized) between an estimator of $\mathcal{M}_2$ parameters (eventually functional), and its pseudo-true value under $\mathcal{M}_1$.

The specification tests between nonparametrically estimated models that we introduce only have a meaning if the smoothing parameter is not arbitrarily chosen. In the encompassing methods we develop, the window-widths are calculated by an automatic empirical method (cross-validation). Then, the window-width determination becomes a whole part of the estimation procedure. This leads us to define the pseudo-true window-width, associated to the pseudo-true value.

8.1 Introduction

Encompassing corresponds to a familiar notion in most sciences : an essential criterion in validating a new model or theory lies in its capacity to account for ("encompass") findings or failures of earlier models. In econometrics, encompassing exercises - mostly informal - are frequently found in applications, whose authors reappraise conclusions of earlier work in the light of their most recent results.The notion of encompassing has re-

[1]We are grateful to Christian Gourieroux, Grayham Mizon, Eric Renault and Pascal Lavergne for helpful comments.

cently been the object of theoretical developments - see e.g. Mizon and Richard (1986), White (1990) or, for a survey Hendry and Richard (1989). Within a Bayesian framework encompassing can formally be reinterpreted as a concept of sufficiency among models ; see Florens, Hendry and Richard (1996) for further discussion.

The concept which is used as the starting point of the present chapter is parametric encompassing, as defined by Mizon and Richard and formalized further by White. Let $\mathcal{M}_1$ and $\mathcal{M}_2$ denote two models with respective data density $f(y|\beta)$ and $g(y|\gamma)$. Let $\hat{\gamma}$ denote a statistic of interest within the context of $\mathcal{M}_2$, typically an estimator of γ and let $\gamma(\beta)$ be its pseudo-true value on $\mathcal{M}_1$. For the object of our analysis, the latter is defined as the plim of $\hat{\gamma}$ on $\mathcal{M}_1$. One then computes an encompassing difference, which is typically of order $n^{\frac{-1}{2}}$, between $\hat{\gamma}$ and $\gamma(\hat{\beta})$, an estimator of its pseudo-true value within $\mathcal{M}_1$. We shall say that $\mathcal{M}_1$ (parametrically) encompasses $\mathcal{M}_2$ with respect to $\hat{\gamma}$ if that difference is not significant relative to its (asymptotic) sampling distribution on $\mathcal{M}_1$.

This general definition does not raise specific conceptual problems when applied to a pair of marginal models, each being characterized by its own assumptions relative to the distribution of a common random vector y. Numerous econometric models, however, are (conditional) regression models, in which the distribution of the conditioning variables is usually left unspecified. In such cases the derivation of the pseudo-true values and the evaluation of (asymptotic) sampling distributions for encompassing differences, inevitably requires the introduction of specific assumptions relative to the conditioning process. See e.g. the contributions by Gouriéroux, Monfort and Trognon (1983) or White (1980b), White (1982a) or Florens, Hendry and Richard (1996) which are directly relevant to our current analysis.

We mention that Mizon and Richard (1986) do apply the notion of encompassing to (static) linear regression models. Within that restricted class of models, plims are replaced by (finite) sample expectations, conditional on the observed values of the conditioning variables. Within the context of Gourieroux and Monfort (1994), Gouriéroux and Monfort (1995b), and Gouriéroux, Monfort and Trognon (1983) these expectations can be reinterpreted as finite sample pseudo-true values.

We shall examine below whether these conditional results remain valid when the regressors are treated as random variables.

The object of our chapter is to apply the notion of encompassing to general regression models which, in particular, need not be linear or Gaussian.

We shall, therefore, have to introduce explicit assumptions regarding the conditioning process[2].

[2]Conditioning variables are often selected on the basis of some exogeneity assumptions, such as weak exogeneity in the terminology of Engle, Hendry and Richard (1983). Note, however, that

For the ease of presentation we shall assume that the joint process which generates all relevant variables is i.i.d. but it will become obvious that our analysis extends to a broader class of stationary ergodic processes[3], in line with the recent contribution of White (1990).

We shall first briefly reconsider linear regression models within the context of an L_2–space (which does not require the introduction of more specific distributional assumptions such as normality). This enables us to reinterpret the results derived by Mizon and Richard (1986) in a broader context.

This being done, a key concern in the analysis of regression models is to know whether or not exclusion restrictions are robust with respect to the choice of functional forms for the regression functions. In particular, if the main goal in an encompassing analysis is to validate the conditional independence assumptions which are built within $\mathcal{M}_1$, we propose to adopt a non–parametric framework, in which no specific assumptions are made regarding the specific form of the regression functions. We shall, therefore, introduce nonparametric encompassing test procedures and discuss their implementation. We also discuss here of the hypotheses tested, and analyse the differences between "null" and "implicit null" hypo theses.

The chapter is organized as follows :

In the following section we define notations and assumptions. The estimation procedure and the pseudo-true values in linear and nonparametric cases are defined in section 8.3. Section 8.4 deals with the asymptotic behaviour of the encompassing statistic in each of the four cases we consider here. The nonparametric estimation procedure we use leads us to take into account the smoothing parameter choice as an important part of the estimation procedure, as we will see in section 8.5.

---

linear approximations and regression functions, which are the object of interest in the context of the current discussion, are defined independently of the exogeneity of the conditioning variables. Lack of exogeneity may result in inefficient estimation of these objects but will not bias encompassing test statistics which are aimed at testing the "validity" of the exclusion of z from M_1. Within the semi–parametric framework we adopt here, $\mathcal{P}$ is left unspecified beyond the linear or conditional independence assumptions associated with M_1, so that exogeneity assumptions are essentially irrelevant. Hence our systematic usage of "*conditioning*" instead of "*exogenous*" variables.

[3]Note, however, that the expressions we shall obtain for the variances or our encompassing differences assume that the relevant characteristics of the conditioning process, such as second order moments, are not subject to overidentifying restrictions, as it can be the case within dynamic models - see e.g. Govaerts, Hendry and Richard (1994). This issue is discussed later in the chapter. In short, accounting for overidentifying restrictions reduces sampling variances but complicates actual computations ; consistency of our encompassing differences is, however, preserved.

8.2 Notation and models

8.2.1 Notation

Observations consist of a sequence of vectors $(s_i)_{i=1,..,n} = (y_i, x_i, z_i)_{i=1,...,n}$ with $y_i \in \mathbb{R},\ x_i \in \mathbb{R}^p$ and $z_i \in \mathbb{R}^q$.

Let w_i denotes a basis of the subspace generated by (x_i, z_i), essentially x_i and z_i are the explanatory variables associated with $\mathcal{M}_1$ and $\mathcal{M}_2$ respectively.

Formally $(s_i)_{i=1,..,n}$ constitutes a square integrable process defined on a probability space $(\Omega, \mathcal{A}, \mathcal{P})$. The probability $\mathcal{P}$ is unknown and we shall restrict our attention to parameters or functions defined from it. The process $(s_i)_{i=1,..,n}$ is assumed to be centred.

We shall assume that the process $(s_i)_{i=1,..,n}$ is i.i.d. Hence, its distribution is fully characterized by a single observation which is itself described by its density [4] $\varphi(s_i)$ with respect to the Lebesgue measure in R^{p+q+1}. In order to simplify our study, we shall assume that the matrix $E(w_i w_i')$ is regular (rank $p+q$). The latter assumption can be relaxed, allowing for common components in x_i and z_i or, more generally, for lack of linear independence between components of x_i and z_i. In such cases, the density φ would be taken with respect to the Lebesgue measure restricted to the appropriate subspace of R^{p+q+1}.

It is important for a correct interpretation of our results to recognize that linear (least-squares) approximations can be used independently of the linearity of the regression functions.

Let $\mathbf{L}(y_i|x_i)$ denote L^2 - projection of y_i on the subspace generated by x_i :

$$\mathbf{L}(y_i|x_i) = \beta' x_i \quad with \quad \beta = [\mathbf{E}(x_i x_i')]^{-1}\mathbf{E}(x_i y_i)$$

The vector of parameters β is a function valued in R^P of the (unknown) density φ and, hence, of the probability $\mathcal{P}$. Other projections of interest are :

$$\mathbf{L}(y_i|z_i) = \gamma' z_i \quad with \quad \gamma = [\mathbf{E}(z_i z_i')]^{-1}\mathbf{E}(z_i y_i)$$

$$\mathbf{L}(y_i|x_i, z_i) = \alpha' w_i \quad with \quad \alpha = [\mathbf{E}(w_i w_i')]^{-1}\mathbf{E}(w_i y_i)$$

where w_i denotes a basis of the subspace generated by (x_i, z_i).

For practical purposes, we shall assume in the rest of chapter that $w_i = (x_i, z_i^*)$ where z_i^* is a subset of z_i. Expectations relative to $\mathcal{P}$, are generically represented by the letter $\mathbf{E}$. The conditional expectation of y_i given x_i is :

[4] For ease of notation φ will be generically used to represent the joint density of s_i, as well as its marginal and conditional densities, all ambiguties being revolved by the list of arguments.

$$\mathbf{E}(y_i|x_i) = \int y_i \varphi(y_i|x_i) dy_i$$

$\mathbf{E}(y_i|z_i)$ and $\mathbf{E}(y_i|x_i, z_i)$ are defined accordingly.

The short-hand notation $f(x_i), g(z_i)$ and $r(x_i, z_i)$ will be used to note $\mathbf{E}(y_i|x_i)$, $\mathbf{E}(y_i|z_i)$ and $\mathbf{E}(y_i|x_i, z_i)$ respectively. The process $(s_i)_{i=1,..,n}$ being integrable, f, g and r are themselves square integrable.

The following regularity condition is assumed :

Hypothesis $\mathcal{A}_o$: *There exists a continuous version of φ (and, hence of its marginal and conditional densities), as well as continuous versions of f, g and r.*

In the rest of our chapter, we shall implicitly restrict our attention to these (necessarily unique) continuous versions.

The exclusion of z_i^* from $\mathcal{M}_1$ can be assumed at different levels, which are discussed in the following section.

8.2.2 Hypotheses

The basic two hypotheses of concern are :

$$\mathbf{E}(y|x, z) = \mathbf{E}(y|x) \qquad (\mathcal{H}_1)$$

$$\mathbf{L}(y|x, z) = \mathbf{L}(y|x) \qquad (\mathcal{H}_2)$$

$\mathcal{H}_1$ corresponds to mean-conditional independence and $\mathcal{H}_2$ to conditional orthogonality (or linear independence). $\mathcal{H}_2$ does not imply that the regression functions themselves are linear. The latter requirement corresponds to a third hypothesis.

$$\mathbf{E}(y|x, z) = \mathbf{L}(y|x, z) \qquad (\mathcal{H}_3)$$

We shall eventually require an additional hypothesis relative to the square of y :

$$\mathbf{E}(y^2|x, z) = \mathbf{L}(y^2|x, z) \qquad (\mathcal{H}_4)$$

The pair $(\mathcal{H}_1, \mathcal{H}_4)$ implies the equality of the conditional variances $V(y|x, z)$ and $V(y|x)$ while the pair $(\mathcal{H}_2, \mathcal{H}_4)$ implies the equality of the variances of the "residuals" associated with the corresponding linear approximations

$$\mathbf{E}[(y - \mathbf{L}(y|x))^2|x, z] = \mathbf{E}[(y - \mathbf{L}(y|x))^2|x]$$

No additional assumptions will be introduced relative to $\mathcal{P}$.

Obviously, if s is jointly normally distributed, then $(\mathcal{H}_3, \mathcal{H}_4)$ holds and $\mathcal{H}_2$ is equivalent to $\mathcal{H}_1$.

Within a nonparametric framework, $\mathcal{M}_1$ will be characterized by the assumption $\mathcal{H}_1$. "Parametric" linearity can be either weak or strong, being characterized by $(\mathcal{H}_2, \mathcal{H}_4)$ or $(\mathcal{H}_2, \mathcal{H}_3, \mathcal{H}_4)$ respectively[5]

$\mathcal{M}_2$ denotes a rival model with z as sole regressors while $\mathcal{M}$ denotes a "nesting" model with w as regressors. From the perspective of the "owner" of $\mathcal{M}_1$, these models are of no special interest on their own, being either "misspecified" ($\mathcal{M}_2$) or "oversparametrized" ($\mathcal{M}$). They are essentially instrumental in the construction of encompassing test statistics aimed at validating $\mathcal{M}_1$.

Remark 1. *For the complete comprehension of our study, it is important to have in mind that "our" null hypotheses , $\mathcal{H}_1$ and $\mathcal{H}_2$ are different from some "implicit" null hypotheses defined bellow :*

$$\begin{aligned} (\mathcal{H}_1 bis) \qquad & \mathbf{E}\left[\mathbf{E}(y|x) \mid z\right] = \mathbf{E}(y|z) \\ (\mathcal{H}_2 bis) \qquad & \mathbf{L}\left[\mathbf{L}(y|x) \mid z\right] = \mathbf{L}(y|z) \end{aligned} \tag{8.1}$$

These are the "implicit encompassing hypotheses" we are testing in the following sections by estimating the pseudo-true value (left hand side) and comparing it with some estimation of the "parameter of interest[6]" in $\mathcal{M}_2$. Let us see the differences between theses hypotheses, we clearly have :

$$(\mathcal{H}_2) \Longleftrightarrow (\mathcal{H}_2 bis)$$

Proof :

We shall prove first that the hypothesis $\mathcal{H}_2 bis$ implies $\mathcal{H}_2$:

Under $\mathcal{H}_2 bis$ we have :

$$\mathbf{L}\left[\mathbf{L}(y|x) \mid z\right] - \mathbf{L}(y|z) = 0$$

or

$$\mathbf{L}\left[y - \mathbf{L}(y|x) \mid z\right] = 0$$

that means that $y - \mathbf{L}(y|x)$ is orthogonal to the subspace generated by z :

$$y - \mathbf{L}(y|x) \perp z$$

we also know, by definition, that $y - \mathbf{L}(y|x)$ is orthogonal to the subspace generated by x :

$$y - \mathbf{L}(y|x) \perp x$$

[5] In the rest of the chapter we shall regroup weak and strong linearity under the common heading of linear regression and be explicit regarding whether or not H_3 holds only when it matters to the argument.

[6] These parameters may be eventually functional.

These two orthogonal conditions lead to the result.

$$\left.\begin{array}{l} y - \mathbf{L}(y|x) \perp z \\ \\ y - \mathbf{L}(y|x) \perp x \end{array}\right\} \Longrightarrow y - \mathbf{L}(y|x) \perp x, z \quad (\mathcal{H}_2)$$

On the other hand, the hypothesis $\mathcal{H}_2$ says that :

$$\mathbf{L}(y|x, z) = \mathbf{L}(y|x)$$

If we consider the following expression (which is always true) :

$$\mathbf{L}\left[\mathbf{L}(y|x, z) \mid z\right] = \mathbf{L}\left[y \mid z\right]$$

under $\mathcal{H}_2$ this expression becomes :

$$\mathbf{L}\left[\mathbf{L}(y|x) \mid z\right] = \mathbf{L}\left[y \mid z\right] \quad (\mathcal{H}_2 bis)$$

□

On the contrary, the hypothesis $(\mathcal{H}_1 bis)$ is not equivalent to $(\mathcal{H}_1)$, in fact we only have an implication relation between these hypotheses, that is :

$$(\mathcal{H}_1) \Longrightarrow (\mathcal{H}_1 bis)$$

so that the implicit null hypothesis $(\mathcal{H}_1 bis)$ is weaker than $(\mathcal{H}_1)$.

Proof :

We have :

$$\mathbf{E}\left[\mathbf{E}(y|x, z) \mid z\right] = \mathbf{E}\left[y \mid z\right]$$

Under $\mathcal{H}_1$, the first term may be rewritten

$$\mathbf{E}\left[\mathbf{E}(y|x) \mid z\right] = \mathbf{E}\left[y \mid z\right]$$

so that it leads to $(\mathcal{H}_1 bis)$. □

Hence, it easy to find an example (see example 1 below) for which the testing hypothesis $(\mathcal{H}_1 bis)$ is true while $(\mathcal{H}_1)$ is not.

Example 1. *Let $x \sim \mathcal{N}(0, 1)$, $z \sim \mathcal{N}(0, 1)$ and suppose that $x \perp z$ and that the variable y follows also a normal distribution :*

$$y \sim \mathcal{N}\left(x \cdot z, I\right)$$

then $(\mathcal{H}_1 bis)$ is true because of the decomposition :

$$\begin{array}{rl} \mathbf{E}(y|x) & = \mathbf{E}\left[\mathbf{E}(y|x, z) \mid x\right] \\ & = \mathbf{E}\left[x \cdot z \mid x\right] \\ & = x \cdot \mathbf{E}\left[z \mid x\right] \\ & = 0 \end{array}$$

We also have (by symmetry) $\mathbf{E}(y|z) = 0$*, so that*

$$\mathbf{E}\left[\mathbf{E}(y|x) \mid z\right] = \mathbf{E}(y|z) = 0$$

while $(\mathcal{H}_1)$ *is not satisfied :*

$$\mathbf{E}(y|x,z) = x \cdot z \neq \mathbf{E}(y|x) = 0$$

□

It is important to notice that our test has no power against such models, in this case.

8.3 Estimation and pseudo-true value

In the context of our chapter, linear (approximations) and non linear regressions are the focus of interest. Linear regressions are estimated in the usual (parametric) way. Non linear regressions will be estimated non parametrically. In any event, our approach is fundamentally semi-parametric in the sense that the probability $\mathcal{P}$ on the reference space is left unspecified beyond assumptions $\mathcal{A}_0$ (regulatory), $\mathcal{H}_1$ or $\mathcal{H}_2$ (conditional or linear independence).

Consider first the linear version of model $\mathcal{M}_1$. A natural estimator of β, as defined in (2.1), is the OLS estimator :

$$\hat{\beta} = \left(\sum_{i=1}^{n} x_i x_i'\right)^{-1} \sum_{i=1}^{n} x_i y_i$$

Corresponding estimators of γ within $\mathcal{M}_2$ and α within the "nesting" model $\mathcal{M}$ are :

$$\begin{aligned} \hat{\gamma} &= \left(\sum_{i=1}^{n} z_i z_i'\right)^{1-} \sum_{i=1}^{n} z_i y_i \\ \hat{\alpha} &= \left(\sum_{i=1}^{n} w_i w_i'\right)^{-1} \sum_{i=1}^{n} w_i y_i \end{aligned}$$

Nonparametric estimation of regression functions is now widely used in econometrics (see e.g. Bierens, 1987 or, for more detailed analyses, Härdle, 1990, or Bosq and Lecoutre, 1987). A variety of techniques are now available but we shall focus our attention on kernel estimation (estimation by means of orthogonal, spline functions, or more generally δ-suits could be treated along similar lines).

A kernel estimator of $f(.) = E[Y|X = .]$ is given by :

$$\widehat{f}(x) = \frac{\frac{1}{n.h_n}\sum_{i=1}^{n} y_i K\left(\frac{x-x_i}{h_n}\right)}{\frac{1}{n.h_n}\sum_{i=1}^{n} y_i K\left(\frac{x-x_i}{h_n}\right)} \tag{8.2}$$

K denotes a Parzen-Rosenblatt kernel, i.e. an application from R^P in R which is integrable with respect to the Lebesgue measure, of integral equal to 1 and satisfies the limit condition

$$\lim_{\|x\|\to 0} \|x\|^P \cdot K(x) = 0$$

where $\|\cdot\|$denotes the Euclidian norm.

The positive real number h_n is the "window–width" of the estimator. It varies with the sample size n and tends towards zero as n tends toward infinity.

At an heuristic level, the denominator in (8.2) is an estimator of the marginal density $\varphi(x)$, and the numerator an estimator of the integral $\int y\varphi(x,y)dy$.

If, in particular, K is a probability density with zero mean, then the kernel estimator of $\varphi(x)$ is given by the density of the sum of two random variables, one which follows the empirical distribution of the x_i' and the other a distribution with density $h_n^{-p} \cdot K(\frac{1}{h_n}x)$. The convolution then follows a distribution whose density is given by the denominator in (8.2) and which, obviously at that heuristic level, tends towards the distribution of x as h_n tends towards zero.

The kernel estimators $\widehat{g}(z)$ and $\widehat{r}(x,z)$ of the regression functions $g(z)$ and $r(x,z)$ are defined accordingly. Note however that the bandwidth sequence associated to $\widehat{g}$, k_n, may be different from the one used in $\widehat{f}(x)$.

$$\widehat{g}(z) = \frac{\frac{1}{n.k_n}\sum_{i=1}^{n} y_i K\left(\frac{z-z_i}{k_n}\right)}{\frac{1}{n.k_n}\sum_{i=1}^{n} K\left(\frac{z-z_i}{k_n}\right)}$$

In order to alleviate notation, we shall use a common notation K for all kernels though they may obviously differ from each other, in particular for considerations of dimension.

We shall also assume that the window-width h_n and k_n satisfies the traditional convergence conditions :

$$\begin{array}{ll} \lim\limits_{n\to\infty} & nh_n^p = \infty \\ \lim\limits_{n\to\infty} & nk_n^q = \infty \end{array} \tag{8.3}$$

together with :

$$\lim_{n\to\infty} h_n = 0 \quad and \quad \lim_{n\to\infty} k_n = 0$$

In line with our earlier discussion of conditional independence assumptions, the kernel estimators $\hat{f}, \hat{g}$ and $\hat{r}$ are naturally associated with the models $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}$ respectively.

For the objectives of our encompassing analysis, we shall only derive their limits under the linear or mean-conditional independence assumptions associated with $\mathcal{M}_1$. These limits are given in the following two theorems.

The following hypotheses are maintained throughout the analysis : i.i.d., square integrability, continuous version $(\mathcal{A}_o)$ and limiting behaviour of h_n and k_n according to (8.3). We only list additional assumptions.

Theorem 1. *: Under $\mathcal{H}_1$,*

$$\begin{aligned}
\widehat{f}(x) &\longrightarrow f(x) && \forall x && i)\\
\widehat{g}(z) &\longrightarrow \mathbf{E}(f(x)|z) && \forall z && ii)\\
\widehat{r}(x,z) &\longrightarrow f(x) && \forall (x,z) && iii)\\
\hat{\gamma} &\longrightarrow [\mathbf{E}(zz')]^{-1}\mathbf{E}(zf') && && iv)
\end{aligned} \tag{8.4}$$

Proof : *The proof follows from theorem II.1 in Bosq and Lecoutre (1987, chapter 5).*

A kernel estimator of a conditional expectation tends in probability towards the latter in every point. Hence we have :

$$\begin{aligned}
\widehat{f}(x) &\longrightarrow \mathbf{E}(y|z) = f(x) && \forall x\\
\widehat{g}(z) &\longrightarrow \mathbf{E}(y|z) = g(z) && \forall z\\
\widehat{r}(x,z) &\longrightarrow \mathbf{E}(y|x,z) = r(x,z) && \forall (x,z)\\
\hat{\gamma} &\longrightarrow [\mathbf{E}(zz')]^{-1}\mathbf{E}(zf')
\end{aligned}$$

The results follow from the fact that under $\mathcal{H}_1$:

$$\begin{aligned}
r(x,z) &= f(x)\\
\mathbf{E}(y|z) &= \mathbf{E}(r(x,z)|z) = \mathbf{E}(f(x)|z)
\end{aligned}$$

□

The same asymptotic behaviour are observed under $\mathcal{H}_2$, they are given in the following theorem :

Theorem 2. *: Under $\mathcal{H}_2$,*

$$\hat{\beta} \longrightarrow \beta \qquad i)$$

$$\hat{\gamma} \longrightarrow [\mathbf{E}(zz')]^{-1} \cdot \mathbf{E}(zx') \cdot \beta \qquad ii)$$

$$\hat{\alpha} \longrightarrow (\beta' : 0) \qquad iii)$$

If, in addition,$\mathcal{H}_3$ holds, then

$$\hat{g}(z) \longrightarrow \beta' \cdot \mathbf{E}(x|z) \qquad \forall z$$

Proof : *Obvious.* □

Note that $\hat{\gamma}$ and $\hat{g}$ are estimators of (finite dimensional or functional) parameters associated with $\mathcal{M}_\mathbf{2}$, which "explains" y by z only. Their pseudo-true values are defined in accordance with the limit results in theorems 1 and 2.

Definition 1. *: The pseudo-true values associated with $\hat{\gamma}$ and $\hat{g}$ on $\mathcal{H}_\mathbf{1}$ are given by :*

$$\Gamma \quad f \longrightarrow [\mathbf{E}(zz']^{-1}\mathbf{E}(zf')$$

$$G \quad f \longrightarrow \mathbf{E}\left(f \mid z\right) = \int f(x_i)\varphi(x_i|z)dx_i$$

Definition 2. : *The pseudo-true value associated with $\hat{\gamma}$ on $\mathcal{H}_\mathbf{2}$ is given by*

$$\Gamma_L : \beta \longrightarrow [\mathbf{E}(zz')]^{-1} \cdot \mathbf{E}(zx') \cdot \beta$$

and that associated with $\hat{g}$ on $(\mathcal{H}_\mathbf{2}, \mathcal{H}_\mathbf{3})$ by

$$G_L : \beta \longrightarrow \beta' \cdot \mathbf{E}(x|z)$$

Both Γ_L and G_L take their value in $\mathbb{R}^p$. They can equally be interpreted as applications between linear functions since to the functional $x_i'\beta$ they associate the functions $z_i' \cdot \Gamma_L(\beta)$ and $z_i' \cdot G_L(\beta)$ respectively. The applications Γ, G, Γ_L and G_L are all linear and are projections between vector spaces of functions (linear or continuous, depending on each case).

The limits of the estimators $\hat{\beta}$ and $\hat{f}$ on $\mathcal{M}_\mathbf{1}$ do not depend on the underlying ("*true*") distribution $\mathcal{P}$ of the conditioning variables[7]. In contrast the $\mathcal{M}_\mathbf{1}$ limits of

[7] Except for the fact that the conditioning variables are assumed to be i.i.d. or, more generally, that their distribution is such that the law of large numbers applies.

estimators of "misspecified" models (such as $\mathcal{M}_2$ is relative to $\mathcal{M}_1$) critically do depend critically on that distribution, see Florens, Hendry and Richard (1996) or Florens and Larribeau (1991) for a more systematic comparison between "*true*" and "*approximate*" regression functions in relationship with the distribution $\mathcal{P}$ of the conditioning variables.

This dependence vanishes when the model to be encompassed $(\mathcal{M}_2)$ "nests" the encompassing model $(\mathcal{M}_1)$, as shown in parts (iii) of theorems 1 and 2 (Naturally relative to $\mathcal{M}_1$, $\mathcal{M}_2$ is not "misspecified" but simply "inefficient").

The situation where the rival models are "non-nested" $(\mathcal{M}_1/\mathcal{M}_2)$ deserves more attention. The pseudo-true values introduced in definitions 1 and 2 are theoretical in the sense that they depend on characteristics of the unknown probability $\mathcal{P}$. They need to be estimated for the purpose of constructing (encompassing) test statistics.

An important distinction then emerges between constrained and unconstrained estimation.

If the specification of the models $\mathcal{M}_1$ and $\mathcal{M}_2$ imposes no (cross) restriction on the distribution $\mathcal{P}$ of x and z, and in particular, on their moments, then population characteristics of $\mathcal{P}$ are simply replaced by their unconstrained sample estimates. We shall, in particular, use the following "unconstrained" estimators of the pseudo-true values introduced in definitions 1 and 2 :

Definition 3. *The estimators associated to the different pseudo-true values are :*

$$\widehat{\Gamma} \quad f \quad \longrightarrow (\sum_{i=1}^{n} z_i z_i')^{-1} \cdot \sum_{i=1}^{n} z_i f'(x_i) \qquad i)$$

$$\widehat{G} \quad f \quad \longrightarrow \frac{\dfrac{1}{n.k_n} \displaystyle\sum_{i=1}^{n} f(x_i) K\left(\frac{z - z_i}{k_n}\right)}{\dfrac{1}{n.k_n} \displaystyle\sum_{i=1}^{n} K\left(\frac{z - z_i}{k_n}\right)} \qquad ii)$$

$$\widehat{\Gamma}_L \quad \beta \quad \longrightarrow (\sum_{i=1}^{n} z_i z_i')^{-1} \cdot \sum_{i=1}^{n} z_i x_i' \cdot \beta \qquad iii)$$

$$\widehat{G}_L \quad \beta \quad \longrightarrow \frac{\dfrac{1}{n.k_n} \displaystyle\sum_{i=1}^{n} x_i' \cdot \beta K\left(\frac{z - z_i}{k_n}\right)}{\dfrac{1}{n.k_n} \displaystyle\sum_{i=1}^{n} K\left(\frac{z - z_i}{k_n}\right)} \qquad iv)$$

Applying the same asymptotic arguments than those used in the proofs of theorems 1 and 2 we can demonstrate that $\hat{\Gamma}, \hat{G}, \hat{\Gamma}_L$ and $\hat{G}_L$ are consistent estimators of Γ, G, Γ_L

and G_L respectively. The analysis in section 4 below is based on these unconstrained estimators.

There are, however, situations where the distribution $\mathcal{P}$ is naturally constrained. Though a detailed analysis of such cases goes beyond the objectives of the present chapter, we briefly discuss a pair of examples in order to illustrate the issues under consideration.

A first set of restrictions on the distribution of (x_i, z_i) may result from the presence of variables which are common to both models. Let, for example, $x_i = (x_i^*, \xi_i)$ and $z_i = (z_i^*, \xi_i)$.

Such restrictions are easily dealt with. Definitions 1 and 2 remain valid. In particular the pseudo-true value G in definition 1 which is rewritten as :

$$G(f)(z) = \int f(x_*, \xi)\varphi(x_*|\xi, z)dx$$

is still consistently estimated by $\hat{G}$ in definition 3. In the parametric case the components of Γ_L and $\hat{\Gamma}_L$ which correspond to the common variables ξ_i will be equal to their counterparts in β.

Other cases where the very nature of the models in presence determines, partially or totally, the distribution P may be more difficult to deal with at a general level. This frequently occurs with dynamic models and the following example is taken from Govaerts, Hendry and Richard (1994), to which we refer the reader for a general discussion of encompassing within the context of dynamic linear vector autoregressive processes. Let $\mathcal{M}_1$ be given by :

$$\mathcal{M}_1 : y_i = f(y_{i-1}) + u_i$$

Where $f(y_{i-1}) = E(y_i|y_0, \cdots, y_{i-1})$. Let also

$$\mathcal{M}_2 : y_i = g(y_{i-2}) + v_i$$

A kernel estimator of g is given by :

$$\widehat{g}(y) = \frac{\frac{1}{n.k_n}\sum_{i=1}^{n} y_i K\left(\frac{y - y_{i-2}}{k_n}\right)}{\frac{1}{n.k_n}\sum_{i=1}^{n} K\left(\frac{y - y_{i-2}}{k_n}\right)}$$

If M_1, is ergodic, then $\hat{g}$ converges towards

$$G(f) = \int f(y_{i-1})\varphi(y_i - 1|y_{i-2})dy_{i-1} = \mathbf{E}(y_i|y_{i-2})$$

and $\widehat{G}(f)$ is fully determined once $\mathcal{M}_1$ is estimated. The derivation of an asymptotic covariance matrix for such statistics as the encompassing difference $\sqrt{n}(\widehat{g} - \widehat{G}(\hat{f}))$, which we introduce below, implies accounting for such restrictions and will not be discussed further in the present chapter.

8.4 Asymptotic Analysis of Encompassing Statistics

We examine now the different encompassing test procedure between $\mathcal{M}_1$ and $\mathcal{M}_2$, allowing each one to be linear or non parametrically specified.

In each of the four cases [8], the encompassing test procedure we build is the same : We take the asymptotic distribution of the difference between the estimation of $\mathcal{M}_2$ and the estimation of its pseudo-true value, that is, the estimation of $\mathcal{M}_2$ in the belief that $\mathcal{M}_1$ is the real "*true*" model. This difference, properly normalized, converges to a zero mean normal law, from which we obtain, in the two situations where $\mathcal{M}_2$ is parametric a scalar statistic asymptotically χ^2.

We first recall the parametric (linear) case, already treated in Gouriéroux, Monfort and Trognon (1983) and Mizon and Richard (1986), then we shall examine the completely nonparametric case and then the crossed cases. The notations of the encompassing statistics and pseudo-true values are given in the following table :

$\mathcal{M}_1 \backslash \mathcal{M}_2$	Parametric	Nonparametric
Parametric	$\widehat{\delta}_{\beta,\gamma} = \widehat{\gamma} - \widehat{\Gamma}_L(\widehat{\beta})$	$\widehat{\delta}_{\beta,g} = \widehat{g}(z) - \widehat{G}_L(\hat{\beta})(z)$
Nonparametric	$\widehat{\delta}_{f,\gamma} = \widehat{\gamma} - \widehat{\Gamma}(\hat{f})$	$\widehat{\delta}_{f,g}(z) = \widehat{g}(z) - \widehat{G}(\hat{f})(z)$

(8.5)

In all this section we take an homoscedasticity hypothesis of the residuals :

$$Var[y|x,z] = \sigma^2 \qquad unknown \tag{8.6}$$

In the rest of the chapter, we shall analyse the asymptotic behaviour of encompassing statistics . $\mathcal{M}_1$ will always denote the "*encompassing*" model and $\mathcal{M}_2$ the one to be "*encompassed*".

[8] Parametric *versus* Parametric , Nonparametric *versus* Nonparametric, or "*crossed cases*" Parametric *versus* Nonparametric and Nonparametric *versus* Parametric.

8.4.1 The Completely Parametric Case : (P.P.)

Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be two linear models based on $x_i's$ and $z_i's$ respectively. The encompassing statistic is based on the difference between $\widehat{\gamma}$ an estimator of γ, and $\widehat{\Gamma}_L(\hat{\beta})$ an estimator of its pseudo-true value. That is :

$$\widehat{\delta}_{\beta,\gamma} = \widehat{\gamma} - \widehat{\Gamma}_L(\widehat{\beta})$$

We have the following asymptotic behaviour for our parametric encompassing statistic $\widehat{\delta}_{\beta,\gamma}$:

Theorem 3. :

Under $\mathcal{H}_2$,$\mathcal{H}_3$ and 8.6 we have the classical asymptotic behaviour for $\widehat{\delta}_{\beta,\gamma}$:

$$\sqrt{n}\,\widehat{\delta}_{\beta,\gamma} \xrightarrow{D} \mathcal{N}\left(0, \sigma^2 \cdot \Omega\right)$$

and

$$\frac{n}{\widetilde{\sigma}^2} \cdot \widehat{\delta}'_{\beta,\gamma} \cdot \widehat{\Omega^+} \cdot \widehat{\delta}_{\beta,\gamma} \xrightarrow{D} \chi^2_L$$

Where :

- $\Omega = Var(z)^{-1}E[Var(z|x)]Var(z)^{-1}$
- $\widehat{\Omega}$ *is an estimator of the matrix* Ω
- *"+" denote a generalized inverse*
- L *denote the rank of* Ω
- $\widetilde{\sigma}^2$ *is a convergent estimator*[9] *of* σ^2.

The proof is given in the mathematical appendix.

Let see now what happens when we let the regression functions be free of any parametric shape.

8.4.2 The "completely non parametric case" : (N.N.)

For the rest of the section, let $\mathcal{M}_1$ and $\mathcal{M}_2$ denote two nonparametric regression models based on x and z respectively.

In this case we need to introduce some additional assumptions on the nature of the underlying process, on the kernels structure, and on the smoothing parameters used.

Hypothesis 8.4.1. :

i) The densities and conditional means derived from (y_i, x_i, z_i) are d-times continuously differentiable and bounded.

[9] For example $\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{\beta}X_i\right)^2$

ii) The X_i's marginal density, φ, is compactly supported and is strictly positive on its support.

iii) The positive Parzen-Rosenblatt kernels introduced in section 8.3 must, moreover, satisfy :

$$\prod_{i=1}^{p} x_i^{j_i} K(x)dx = \begin{cases} 1 & if\ j_i = 0 \quad \forall i \\ 0 & if\ 0 < \sum_{i=1}^{n} j_i < d \end{cases}$$

and

$$\int |x_i|^{j_i} K(x)dx \ < \infty \ \forall x \ \in R^p$$

The window-width conditions (8.3) we fixed in the former section insure the consistency and asymptotic normality of Nadaraya-Watson estimators. However, we need some additional assumptions to "kill" the asymptotic bias, remaining in the limiting normal law (see Collomb, 1976 or Vieu, 1993 for a general discussion on that subject).

Hypothesis 8.4.2. *The asymptotic bias disappear if we impose an additional condition on the speed of convergence for our bandwidth ; of course this conditions involve the regularity order assumed for regression functions and densities, d.*

$$n \cdot {h_n}^{p+2d} \stackrel{n\to\infty}{\longrightarrow} 0$$

$$n \cdot {k_n}^{q+2d} \stackrel{n\to\infty}{\longrightarrow} 0$$

This assumption is usual in nonparametric estimation and is discussed in Vieu (1993) or Bierens (1987).

We keep these assumptions for the rest of the chapter.

Theorem 4. *Under these assumptions (8.4.1, 8.4.2) we have :*

$$\sqrt{n{h_n}^p}\,(\widehat{\varphi}(x) - \varphi(x)) \stackrel{D}{\longrightarrow} \mathcal{N}\left(0, \varphi(x) \cdot \int K^2\right)$$

$$\sqrt{n{h_n}^p}\left(\widehat{f}(x) - f(x)\right) \stackrel{D}{\longrightarrow} \mathcal{N}\left(0, \frac{\sigma^2 \cdot \int K^2}{\varphi(x)}\right)$$

Proof : The proof can be found explicitly in Bosq and Lecoutre (1987, pages 86 and 122). □

In this nonparametric case, the encompassing statistic is based on the later function :

$$\widehat{\delta}_{f,g}(z) = \widehat{g}(z) - \widehat{G}(\hat{f})(z)$$

We need some specific assumptions on the underlying process and on the relative speed of convergence of the two smoothing parameters used (one in each regression model), for the next asymptotic result concerning this nonparametric encompassing statistic.

Hypothesis 8.4.3.

If h_n is the window-width used in the Nadaraya-Watson estimator, $\hat{f}_h$, of f in $\mathcal{M}_1$, and k_n the one used for $\hat{g}_k$ in $\mathcal{M}_2$,

Then

$$\frac{{k_n}^q}{{h_n}^p} \overset{n\to\infty}{\longrightarrow} 0$$

In the univariate case $(p = q = 1)$ this last assumption means that the window-width used for g estimation, k_n must converge to zero faster than the one used for f, h_n.

Theorem 5. *Under $\mathcal{H}_1$, if the assumptions 8.6 (homoscedasticity), 8.4.1 (kernel regularity) are satisfied and if the bandwidth h_n and k_n satisfy 8.4.2 and 8.4.3 we have :*

$$\sqrt{n{k_n}^q}\,\widehat{\delta}_{f,g}(z) \xrightarrow{D} \mathcal{N}\left(0, \frac{\sigma^2 \cdot \int K^2}{\varphi(z)}\right)$$

Where :

- *$\varphi(z)$ is the marginal density of the Z_i in z,*
- *$\int K^2$ denotes the constant $\displaystyle\int_{-\infty}^{\infty} K^2(u)du$*

The proof is given in appendix.

Remark 2. *In order to clarify the conditions on the bandwidth h_n and k_n, it may be interesting to rewrite them as :*

$$\begin{aligned} & h_n = n^{-\alpha} \\ and & \\ & k_n = n^{-b} \end{aligned}$$

where a and b are two positive real number characterizing the speed of convergence for the bandwidth.

The hypotheses 8.3 and 8.4.2 give the classical conditions on a and b :

$$\begin{aligned} & \frac{1}{p+2d} < a < \frac{1}{p} \\ and & \\ & \frac{1}{q+2d} < b < \frac{1}{q} \end{aligned}$$

while the hypothesis 8.4.3 on the relative speed of convergence of the two smoothing parameters gives :

$$pa < qb$$

In the univariate case $(p = q = 1)$*, and for an order of regularity* d *equal to two, we have :*

$$\frac{1}{5} < a < b < 1$$

□

8.4.3 The "crossed" cases (P.N.) and (N.P.)

Two "crossed" cases are now presented. The first, and simplest one, (case PN), is the case where $\mathcal{M}_1$ is chosen to be linear and $\mathcal{M}_2$ nonparametric.

8.4.3.1 The "Parametric versus Nonparametric" case.

In this situation, one may think that the encompassing will not often be realized but, if it is, the result is very powerful. In fact it reveals that a linear model on the X_i's space explains any results of a model based on Z_j's space, even if the later is very general.

Because of the functional structure of $\mathcal{M}_2$, the encompassing statistic is based on the difference between two nonparametric estimators :

$$\widehat{\delta}_{\beta,g} = \widehat{g}(z) - \widehat{G}_L(\hat{\beta})(z)$$

Theorem 6. *Under* $\mathcal{H}_2$*,* $\mathcal{H}_3$*, if the assumptions 8.6 (homoscedasticity), 8.4.1 (kernel regularity) are satisfied and if the bandwidth* k_n *satisfy 8.4.2 we have :*

$$\sqrt{nk_n{}^q}\,\widehat{\delta}_{\beta,g}(z) \xrightarrow{D} \mathcal{N}\left(0, \frac{\sigma^2 \cdot \int K^2}{\varphi(z)}\right)$$

Where :

- $\varphi(z)$ *is the marginal density the* Z_i*,*
- $\int K^2$ *denotes the following constant* $\displaystyle\int_{-\infty}^{\infty} K^2(u)du$

For the ease of the presentation the proof is reported in the mathematical appendix.

8.4.3.2 The "Nonparametric versus Parametric" case

We need an additional assumption relative to the speed of convergence of the unique bandwidth h_n :

Hypothesis 8.4.4.

$$\sqrt{n}.Max\left(\frac{\log(n)}{n.h_n^p}, h_n^{2p}\right) \overset{n\to\infty}{\longrightarrow} 0$$

This hypothesis correspond to the cases in which the bandwidth $h_n = n^{-\alpha}$ *satisfies :*

$$\frac{1}{4d} \leq a \leq \frac{1}{2p}$$

This condition impose a relation between d*, the order of regularity, and the dimension of* x*, that is :*

$$d \geq \frac{d}{2}$$

Then, in the univariate case, we have to assume that our functions are twice-continuously differentiable.

Like in the other cases the encompassing statistic is based on the difference between an estimator of γ and one of its pseudo-true value :

$$\widehat{\delta}_{f,\gamma} = \widehat{\gamma} - \widehat{\Gamma}(\hat{f})$$

Theorem 7. *Under* $\mathcal{H}_1$*, if the assumptions 8.6 (homoscedasticity) , 8.4.1 (kernel regularity) are satisfied and if the bandwidth* h_n *satisfy 8.4.2 and the additional condition 8.4.4, we get :*

$$\sqrt{n}\widehat{\delta}_{f,\gamma} \xrightarrow{D} \mathcal{N}\left(0, \sigma^2\Omega\right)$$

and

$$\frac{n}{\widetilde{\sigma}^2}\,\widehat{\delta}'_{f,\gamma}\widehat{\Omega}^{+}\widehat{\delta}_{f,\gamma} \xrightarrow{D} \chi^2_L$$

Where :

- $\Omega = Var(z)^{-1}E[Var(z|x)]Var(z)^{-1}$
- $\widehat{\Omega}$ *is an estimator*[10] *of the matrix* Ω
- $+$ *denotes a generalized inverse of a matrix.*
- L *is the rank of the matrix* Ω*.*
- $\widetilde{\sigma}^2$ *is a convergent estimator of* σ^2

The proof is given in the mathematical appendix

[10] $\widehat{\Omega}$ may be constructed on the following estimators :

$$\begin{aligned}
\widehat{Var}(z) &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\left(Z_i.Z_i' - \widehat{E}[Z|X_i].\widehat{E}[Z|X_i]'\right) \\
\widehat{E}[Z|X_i] &= \frac{\sum_{i=1}^{n} Z_i K\left(\frac{X_i - X_j}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - X_j}{h}\right)}
\end{aligned}$$

8.5 Window-widths and strategies

As we know, the window-width choice is a crucial component of nonparametric estimation (see Härdle, 1990, Härdle, Hall and Marron, 1992, or Marron, 1988). Being too small or too big, this parameter may induce some bias or variance error in the estimation (see Vieu, 1991, Stone, 1982). In the general context of estimated regression models comparison, this choice, or more exactly, **these** choice**s** may be arbitrary elements of the procedure, and therefore may be strategic. Several elements are to be underlined. First, the choice of the window-width is as important as the choice of the estimator itself, another point is that $\mathcal{M}_1$ cannot encompass $\mathcal{M}_2$ for any window-width choice of the latter. Finally, we have hidden the existence of a third window-width appearing in the pseudo-true value estimator $\widehat{G}(f)$.

This "pseudo-true window-width", m_n, had previously been fixed equal to k_n, but the basic encompassing difference should be written as :

$$\begin{aligned}\widehat{\delta}_{f,g}(z) &= \widehat{g_{\mathbf{k}}}(z) - \widehat{G_{\mathbf{m}}}(\widehat{f_{\mathbf{h}}})(z) \\ &= \frac{\sum_{i=1}^{n} Y_i \cdot K\left(\frac{Z_i - z}{\mathbf{k}_n}\right)}{\sum_{i=1}^{n} K\left(\frac{Z_i - z}{\mathbf{k}_n}\right)} - \frac{\sum_{i=1}^{n} \widehat{f_{\mathbf{h}}}(X_i) \cdot K\left(\frac{Z_i - z}{\mathbf{m}_n}\right)}{\sum_{i=1}^{n} K\left(\frac{Z_i - z}{\mathbf{m}_n}\right)}\end{aligned}$$

so that **three** bandwidth are involved :

- h_n associated to the estimator $\widehat{f_h}$
- k_n associated to the estimator $\widehat{g_k}$
- and m_n associated to the estimator $\widehat{G}_m(f)$ of the pseudo-true value $G(f)(z)$.

The importance of each of the window-width involved may be seen trough an empirical quadratic criterion Ξ :

$$\Xi = \sum_{l=1}^{L} \left(\widehat{\delta}_{f,g}(Z_l)\right)^2 \varpi\left(Z_l\right)$$

- Where $(Z_l)_{l=1}^{L}$ are L arbitrary values of Z, and ϖ is a weight function.

For this example, one can see in figures 8.1 and 8.2 that the criterion Ξ is increasing with h_n, so a simple strategy for $\mathcal{M}_1$ to have better chance to encompass $\mathcal{M}_2$ may be to choose an extremely small window-width.

Therefore, it seems more adequate to impose a common rule for the selection of the window-widths. In order to eliminate any strategic choice, this selection rule must be "objective", that is data-driven. The cross-validation may be a response to that objectivity preoccupation ; it is an automatic, data-driven, selection procedure (see Vieu, 1993,

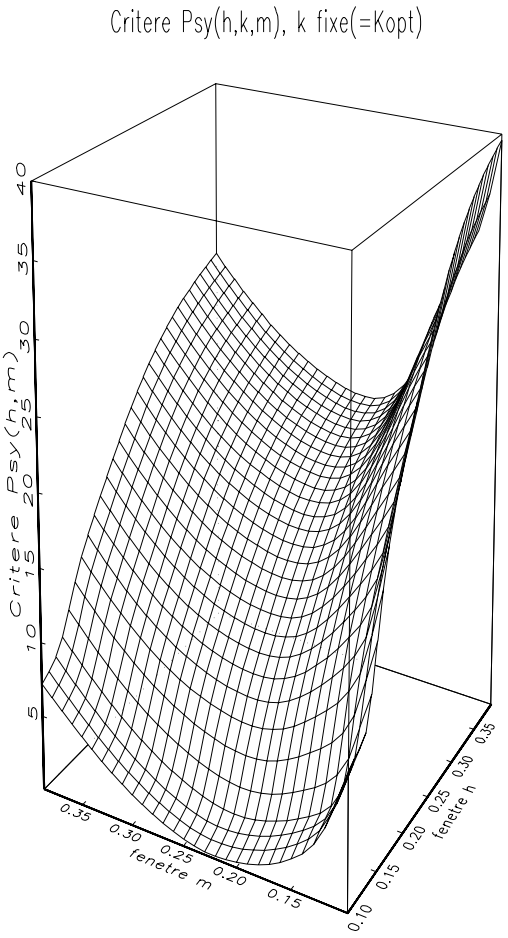

Figure 8.1 $\Xi(h,k,m)$,the bandwith k is fixed.

Härdle, 1990). The window-widths $\widehat{h}$ and $\widehat{k}$ resulting from this procedure are defined as :

$$\widehat{h} = Arg\min CV_X(h) \quad and \quad \widehat{k} = Arg\min CV_Z(k)$$

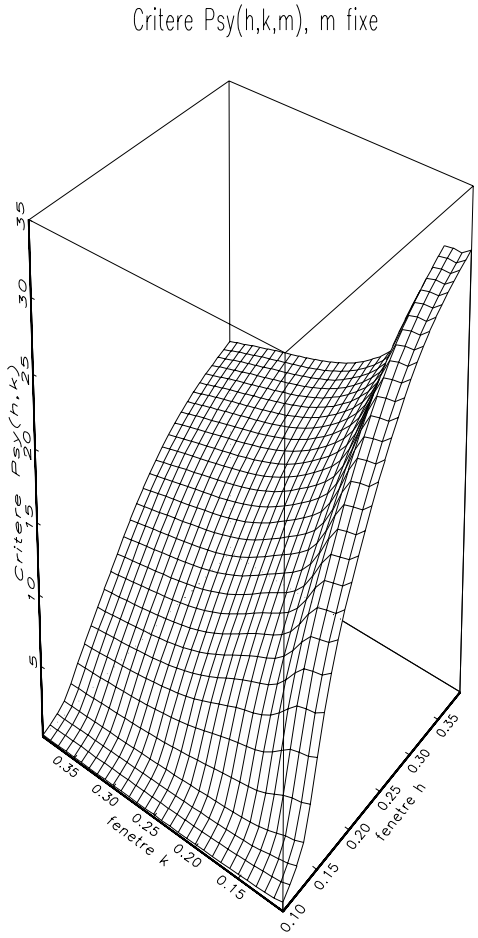

Figure 8.2 $\Xi(h,k,m)$,the bandwith m is fixed.

where

$$CV_X(h) = \sum_{i=1}^{n} \left(Y_i - \widehat{f_h^{-i}}(X_i)\right)^2 \quad and \quad CV_Z(h) = \sum_{i=1}^{n} \left(Y_i - \widehat{g_h^{-i}}(Z_i)\right)$$

and $\widehat{f_h^{-i}}$ and $\widehat{g_h^{-i}}$ are leave-one-out estimators.

A similar procedure is defined for the estimation of $\widehat{m}$.

$$\widehat{m} = Arg \min CV_f(m)$$

where

$$CV_f(m) = \sum_{i=1}^{n} \left(f(X_i) - \widehat{G_m^{-i}}(f)(Z_i) \right)^2$$

The use of these procedures may be useful in practical situation in order to resolve the window-width choice problem. However, one must modify the criterion $CV_f(m)$ using the estimate of f by a similar criterion with $\widehat{f_n}$ replacing f. This procedure should be used once the estimation of f (and of its window-width) has been done.

8.6 Conclusion

Using encompassing principle does not raise problem when applied to unconditional models, characterized by parameters, and is now well known. In line with the work of Hendry and Richard (1989), who applied encompassing to conditional models, we have extended this principle, allowing the parametric models to be free of any Gaussian distribution, and considering nonparametric models free of any functional shape.

The four statistics presented here, dealing with all the possible pairing of parametric and nonparametric models, have the same form. The encompassing statistic is the difference of two estimators in $\mathcal{M}_2$. Depending of the specified form of $\mathcal{M}_2$ (parametric or nonparametric), the statistic is either a parameter or a function difference. The formulation of these statistics can always be rewritten as regression estimators of estimated residuals in $\mathcal{M}_1$, upon regressors Z_i.

These residuals being either parametric or nonparametric, we obtain the four different situations : Linear regression on Z of parametric estimated residuals (case PP), nonparametric regression of nonparametric estimated residuals (case NN), nonparametric regression of linear residuals (case PN) and linear regression on Z of nonparametric estimated residuals (case NP). According to table 8.5, we may rewrite in table 8.7 the encompassing statistics in this optic.

$\mathcal{M}_1 \setminus \mathcal{M}_2$	Parametric	Nonparametric
Parametric	$\widehat{\delta}_{\beta,\gamma} = \frac{\sum_{i=1}^{n} Z_i\left(Y_i - X_i'\cdot\widehat{\beta}\right)}{\sum_{i=1}^{n} Z_i Z_i'}$	$\widehat{\delta}_{\beta,g} = \frac{\sum_{i=1}^{n}\left(Y_i - X_i'\,\widehat{\beta}\right)K\left(\frac{Z_i - z}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{Z_i - z}{h_n}\right)}$
Non-parametric	$\widehat{\delta}_{f,\gamma} = \frac{\sum_{i=1}^{n} Z_i\left(Y_i - \widehat{f_n}(X_i)\right)}{\sum_{i=1}^{n} Z_i Z_i'}$	$\widehat{\delta}_{f,g}(z) = \frac{\sum_{i=1}^{n}\left(Y_i - \widehat{f_n}(X_i)\right)K\left(\frac{Z_i - z}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{Z_i - z}{h_n}\right)}$

(8.7)

This approach shows important common features for encompassing test in parametric and nonparametric frameworks, and provides an unified analyse of encompassing in regression models. However, the analyse done in section 8.5 on the bandwidth choice, leading to the definition of a pseudo-true bandwidth, reveals how important is this parameter when using nonparametric estimator for model comparison. Moreover, we have done a distinction between the "null" and "implicit null" hypotheses tested in the encompassing procedures. This distinction reveals some important feature of encompassing tests and of their associated powers. A future study of the power of these tests under "well chosen" alternative would be of prime interest. Another feature, is the functional (and then local) form of nonparametric encompassing tests (case NN and PN). The recent development of a global encompassing test done by Bontemps and Florens (1995), may be an answer to the local behaviour of the nonparametric tests.

Wishes for future work may be to complete this table using semi-parametric models, or to follow a Bayesian approach to generalise this work (see Florens, Hendry and Richard, 1996), or following Govaerts, Hendry and Richard (1994), extensions to dynamic models may also constitute a reasonable wish for future research. It is therefore to be expected that a good number of papers will continue to appear in the near future.

8.7 Mathematical appendix

8.7.1 Proof of theorem 3 (Parametric vs Parametric)

The statistic $\widehat{\delta}_{\beta,\gamma}$ can be rewritten as :

$$\begin{aligned}
\widehat{\delta}_{\beta,\gamma} &= \left(\widehat{\gamma} - \widehat{\Gamma_L}(\widehat{\beta})\right) \\
&= \left(\frac{\sum_{i=1}^{n} Z_i Y_i}{\sum_{i=1}^{n} Z_i Z_i'} - \frac{\sum_{i=1}^{n} Z_i X_i' \cdot \widehat{\beta}}{\sum_{i=1}^{n} Z_i Z_i'} \right) \\
&= \left(\sum_{i=1}^{n} Z_i Z_i' \right)^{-1} \left(\sum_{i=1}^{n} Z_i \left(Y_i - X_i' \cdot \widehat{\beta} \right) \right)
\end{aligned}$$

Will shall decompose $\widehat{\beta}$ into $\widehat{\beta} = \beta + \left(\widehat{\beta} - \beta\right)$, that is :

$$\begin{aligned}
\widehat{\beta} &= \beta + \frac{1}{\sum_{i=1}^{n} X_i X_i'} \cdot \left(\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i X_i' \cdot \beta \right) \\
&= \beta + \frac{1}{\sum_{i=1}^{n} X_i X_i'} \cdot \left(\sum_{i=1}^{n} X_i \left(Y_i - X_i' \beta \right) \right)
\end{aligned}$$

from this latter expression we derive :

$$\begin{aligned}
\widehat{\delta}_{\beta,\gamma} &= \left(\sum_{i=1}^{n} Z_i Z_i' \right)^{-1} \left(\sum_{i=1}^{n} Z_i Y_i - \sum_{i=1}^{n} Z_i X_i' \cdot \beta - \frac{\sum_{i=1}^{n} Z_i X_i'}{\sum_{i=1}^{n} X_i X_i'} \cdot \left(\sum_{i=1}^{n} X_i \left(Y_i - X_i' \beta \right) \right) \right) \\
&= \left(\sum_{i=1}^{n} Z_i Z_i' \right)^{-1} \left(\sum_{i=1}^{n} Z_i \left(Y_i - X_i' \beta \right) - \frac{\sum_j Z_j X_j'}{\sum_j X_j X_j'} \sum_{i=1}^{n} X_i \left(Y_i - X_i' \beta \right) \right) \\
&= \left(\sum_{i=1}^{n} Z_i Z_i' \right)^{-1} \left(\sum_{i=1}^{n} \left(Z_i - \frac{\sum_j Z_j X_j'}{\sum_j X_j X_j'} \cdot X_i \right) \left(Y_i - X_i' \beta \right) \right)
\end{aligned}$$

Under $\mathcal{H}_2$ and $\mathcal{H}_3$, the general term of this sum satisfies :

$$E\left[\left(Z_i - \frac{\sum_j Z_j X_j'}{\sum_j X_j X_j'} \cdot X_i\right)(Y_i - X_i'\beta)\right] = 0$$

which has for a variance :

$$E\left[\left(Z_i - \frac{\sum_j Z_j X_j'}{\sum_j X_j X_j'} \cdot X_i\right)^2 (Y_i - X_i'\beta)^2\right]$$

or :

$$= E\left[E\left[\left(Z_i - \frac{\sum_j Z_j X_j'}{\sum_j X_j X_j'} \cdot X_i\right)^2 \cdot (Y_i - X_i'\beta)^2 \mid X_i, Z_i\right]\right]$$

using the homoscedasticity assumption (8.6), we get :

$$Var\left[\left(Z_i - \frac{\sum_j Z_j X_j'}{\sum_j X_j X_j'} \cdot X_i\right)(Y_i - X_i'\beta)\right] = \sigma^2 \cdot E\left[Var\left(Z \mid X\right)\right]$$

by application of the central limit theorem, the result holds . □

8.7.2 Proof of theorem 5 (Nonparametric vs Nonparametric)

As usual, in these demonstrations, we shall introduce the estimated residuals on $\mathcal{M}_1$, which are, here[11] : $\widehat{\varepsilon_i} = \left(Y_i - \widehat{f_n}(X_i)\right)$.

$$\begin{aligned}
\sqrt{n \cdot k_n^q} \cdot \widehat{\delta}_{f,g}(z) &= \sqrt{n \cdot k_n^q} \cdot \left(\widehat{g_n}(z) - \widehat{G}(\widehat{f})(z)\right) \\
&= \sqrt{n \cdot k_n^q} \cdot \left(\frac{\sum_{i=1}^n Y_i K\left(\frac{Z_i - z}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{Z_i - z}{h_n}\right)} - \frac{\sum_{i=1}^n \widehat{f_n}(X_i) \cdot K\left(\frac{Z_i - z}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{Z_i - z}{h_n}\right)}\right) \\
&= \sqrt{n \cdot k_n^q} \cdot \frac{\sum_{i=1}^n \left(Y_i - \widehat{f_n}(X_i)\right) K\left(\frac{Z_i - z}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{Z_i - z}{h_n}\right)}
\end{aligned}$$

[11] In the previous demonstration (theorem 3), the residuals appearing where those arising from the parametric regression on $\mathcal{M}_1$:

$$\widehat{\eta_i} = \left(Y_i - X_i' \cdot \widehat{\beta}\right)$$

.

Decomposing $\widehat{f_n}$, at the observation point X_i into $\widehat{f_n}(X_i) = f(X_i) + \widehat{f_n}(X_i) - f(X_i)$, we get :

$$\begin{aligned}\sqrt{n \cdot k_n^q} \cdot \widehat{\delta}_{f,g}(z) &= \sqrt{n \cdot k_n^q} \cdot \frac{\sum_{i=1}^{n} (Y_i - f(X_i)) K\left(\frac{Z_i - z}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{Z_i - z}{h_n}\right)} \\ &+ \sqrt{n \cdot k_n^q} \cdot \frac{\sum_{i=1}^{n} \left(\widehat{f_n}(X_i) - f(X_i)\right) K\left(\frac{Z_i - z}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{Z_i - z}{h_n}\right)} \\ &= A + B\end{aligned}$$

A is the kernel estimator of the residuals $U_i = Y_i - f(X_i)$ upon Z_i. The asymptotic behaviour of this term under $\mathcal{H}_1$ is given in theorem 4 :

$$A \xrightarrow{D} \mathcal{N}\left(0, \frac{\sigma^2 \cdot \int K^2}{\varphi(z)}\right)$$

In order to conclude, we only need to show that B vanishes to zero. The hypothesis (8.4.1) implies that for $d = 2$, the kernel function K is positive , which allows the following inequality :

$$\begin{aligned}B = \sqrt{n \cdot k_n^q} \cdot \frac{\sum_{i=1}^{n} \left(\widehat{f_n}(X_i) - f(X_i)\right) K\left(\frac{Z_i - z}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{Z_i - z}{h_n}\right)} &\leq \sqrt{n \cdot k_n^q} \cdot \frac{\sum_{i=1}^{n} \left(\sup_{X_i} \left|\widehat{f_n}(X_i) - f(X_i)\right|\right) K\left(\frac{Z_i - z}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{Z_i - z}{h_n}\right)} \\ &\leq \sqrt{n \cdot k_n^q} \cdot \left(\sup_{X_i} \left|\widehat{f_n}(X_i) - f(X_i)\right|\right)\end{aligned}$$

Using a result given by Györfi, Härdle, Sarda and Vieu (1989) (see e.g. Bierens, 1987) we know that $\sup_{X_i} \left|\widehat{f_n}(X_i) - f(X_i)\right|$ converge to 0. The speed of convergence is, as usual in nonparametrics, the key for "killing bias term". Here, we have :

$$\sup_{X_i} \left|\widehat{f_n}(X_i) - f(X_i)\right| = O_p\left[\max\left(\frac{\sqrt{\log(n)}}{\sqrt{n \cdot h_n^p}} \ , \ h_n^d\right)\right]$$

Term B is then of order $\sqrt{n \cdot k_n^q} \cdot O_p\left[\max\left(\frac{\sqrt{\log(n)}}{\sqrt{n \cdot h_n^p}} \ , \ h_n^d\right)\right]$, that is :

$$\sup_{X_i} \left|\widehat{f_n}(X_i) - f(X_i)\right| = O_p\left[\max\left(\frac{\sqrt{\log(n)} \cdot \sqrt{n \cdot k_n^q}}{\sqrt{n \cdot h_n^p}} \ , \ \sqrt{n \cdot k_n^q} \cdot \ h_n^d\right)\right]$$

Under hypothesis (8.4.3), $B \xrightarrow{D} 0$ because :

$$\frac{\log(n) \cdot k_n^q}{h_n^p} \overset{n\to\infty}{\longrightarrow} 0$$

and

$$\sqrt{n \cdot k_n^q} \cdot \ h_n^d \leq \sqrt{n \cdot h_n^p} \cdot \ h_n^d$$

the bandwidth h_n being such that hypothesis (8.4.3) holds, and so we get :

$$\sqrt{n \cdot h_n^p} \cdot \ h_n^d \overset{n\to\infty}{\longrightarrow} 0$$

Using Slutsky's theorem (Serfling, 1980, p.19), we finally have :

$$A + B \xrightarrow{D} \mathcal{N}\left(0, \frac{\sigma^2 \cdot \int K^2}{\varphi(z)}\right)$$

□

8.7.3 Proof of theorem 6 (Parametric vs Nonparametric)

Here again, we have to rewrite our statistic $\widehat{\delta}_{f,g}(z)$ such that the parametric residual $\left(Y_i - X_i^{'}\widehat{\beta}\right)$ appears explicitly :

$$\begin{aligned}
\sqrt{n \cdot k_n^q} \cdot \widehat{\delta}_{\beta,g}(z) &= \sqrt{n \cdot k_n^q} \cdot \left(\widehat{g_n}(z) - \widehat{G_L}(\widehat{\beta})(z)\right) \\
&= \sqrt{n \cdot k_n^q} \cdot \left(\frac{\sum_{i=1}^{n} Y_i K\left(\frac{Z_i - z}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{Z_i - z}{h_n}\right)} - \frac{\sum_{i=1}^{n} X_i^{'}\widehat{\beta} \cdot K\left(\frac{Z_i - z}{k_n}\right)}{\sum_{i=1}^{n} K\left(\frac{Z_i - z}{k_n}\right)} \right) \\
&= \sqrt{n \cdot k_n^q} \cdot \frac{\sum_{i=1}^{n} \left(Y_i - X_i^{'}\widehat{\beta}\right) K\left(\frac{Z_i - z}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{Z_i - z}{h_n}\right)}
\end{aligned}$$

$\widehat{\beta}$ being obviously equal to $\widehat{\beta} = \beta + \left(\widehat{\beta} - \beta\right)$, we get :

$$\begin{aligned}
\sqrt{n \cdot k_n^q} \cdot \widehat{\delta}_{f,g}(z) &= \sqrt{n \cdot k_n^q} \cdot \frac{\sum_{i=1}^{n} \left(Y_i - X_i^{'} \beta\right) K\left(\frac{Z_i - z}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{Z_i - z}{h_n}\right)} \\
&\quad + \sqrt{n \cdot k_n^q} \cdot \frac{\sum_{i=1}^{n} \left(X_i^{'} \widehat{\beta} - X_i^{'} \beta\right) K\left(\frac{Z_i - z}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{Z_i - z}{h_n}\right)} \\
&= C + D
\end{aligned}$$

We shall use the technic elaborated in the previous proof (theorem 5) dealing with nonparametric statistic. Basically, C is a kernel estimator of the parametric residuals $U_i = Y_i - \beta X_i$ upon Z_i. The asymptotic behaviour of this term under $\mathcal{H}_2$ and $\mathcal{H}_3$ is given in theorem 4 :

$$C \xrightarrow{D} \mathcal{N}\left(0, \frac{\sigma^2 \cdot \int K^2}{\varphi(z)}\right)$$

while $D \xrightarrow{P} 0$, because :

$$\begin{aligned}
D &= \sqrt{n \cdot k_n^q} \cdot \frac{\sum_{i=1}^{n} \left(X_i' \widehat{\beta} - X_i' \beta\right) \cdot K\left(\frac{Z_i - z}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{Z_i - z}{h_n}\right)} \\
&= \left(\widehat{\beta} - \beta\right) \cdot \sqrt{n \cdot k_n^q}\, \widehat{E}\left[X' \mid Z = z\right] \\
&\xrightarrow{P} 0
\end{aligned}$$

Adding the term and using Slutsky's theorem, the result holds. □

8.7.4 Proof of theorem 7 (Nonparametric vs Parametric)

As usual, the nonparametric residuals $\left(Y_i - \widehat{f_n}(X_i)\right)$ shall appear from the decomposition of the statistic $\widehat{\delta}_{f,\gamma}$:

$$\begin{aligned}
\sqrt{n} \cdot \widehat{\delta}_{f,\gamma} &= \sqrt{n} \cdot \left(\widehat{\gamma} - \widehat{\Gamma}(\widehat{f_n})\right) \\
&= \sqrt{n} \cdot \left(\frac{\sum_{i=1}^{n} Z_i Y_i}{\sum_{i=1}^{n} Z_i Z_i'} - \frac{\sum_{i=1}^{n} Z_i \cdot \widehat{f_n}(X_i)}{\sum_{i=1}^{n} Z_i Z_i'} \right) \\
&= \sqrt{n} \cdot \frac{\sum_{i=1}^{n} Z_i \left(Y_i - \widehat{f_n}(X_i)\right)}{\sum_{i=1}^{n} Z_i Z_i'}
\end{aligned}$$

These residuals may be rewritten as :

$$\left(Y_i - \widehat{f_n}(X_i)\right) = (Y_i - f(X_i)) - \left(\widehat{f_n}(X_i) - f(X_i)\right)$$

and so :

$$\begin{aligned}
\sqrt{n} \cdot \widehat{\delta}_{f,\gamma} &= \frac{\sqrt{n}}{n} \cdot \frac{\sum_{i=1}^{n} Z_i \left(Y_i - f(X_i)\right)}{\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i'} \\
&\quad - \frac{\sqrt{n}}{n} \cdot \frac{\sum_{i=1}^{n} Z_i \left(\widehat{f_n}(X_i) - f(X_i)\right)}{\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i'} \\
&= E - F
\end{aligned}$$

These two terms E and F have different asymptotic behaviour under $\mathcal{H}_1$, they shall be treated separately, starting with F.

In short, a part of F will be added to E and give a new term G, giving the asymptotic normality, while the remaining part of will disappears asymptotically, that is :

- $F = F_1 + F_2$ and $F_2 \xrightarrow{\mathcal{P}} 0$, while
- $G = E + F_1 \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma^2 \cdot V\left(\widehat{\delta}\right)\right)$

First step :

The denominator in F is well known (it is essentially $Var(Z)$) and will be omitted for the rest of the demonstration, we call $F^+$the numerator.

We can write the difference $\left(\widehat{f_n}(X_i) - f(X_i)\right)$ as :

$$\widehat{f_n}(X_i) - f(X_i) = \frac{\widehat{\Phi}(X_i)}{\widehat{\varphi}(X_i)} - \frac{\Phi(X_i)}{\varphi(X_i)}$$

Where $\Phi(\cdot) = \int \varphi(\cdot, y)dy$ and $\widehat{\Phi}(\cdot)$ is its nonparametric estimator

$$\widehat{\Phi}(\cdot) = \frac{1}{nh^p}\sum_{i=1}^{n} Y_i K\left(\frac{X_i - \cdot}{h}\right)$$

We will use the development of the function

$$p(x, y) \longrightarrow \frac{x}{y}$$

to treat this term ; this development is :

$$\begin{aligned} p(x,y) - p(x_1,y_1) &= (x - x_1)/y \\ &\quad -x/y^2 \cdot (y - y_1) \\ &\quad -(y - y_1)^2 \cdot x_\alpha / y_\alpha^3 \\ &\quad -\tfrac{1}{2}(x - x_1)(y - y_1)/y_\alpha \end{aligned}$$

where $x_\alpha \in [x, x_1]$ and $y_\alpha \in [y, y_1]$

$$\begin{aligned} \widehat{f_n}(X_i) - f(X_i) &= \tfrac{1}{\varphi(X_i)} \cdot \left(\widehat{\Phi}(X_i) - \Phi(X_i)\right) \\ &\quad -\tfrac{\Phi(X_i)}{\varphi(X_i)^2}\left(\widehat{\varphi}(X_i) - \varphi(X_i)\right) \\ &\quad +\left(\widehat{\varphi}(X_i) - \varphi(X_i)\right)^2 R_n^1(X_i) \\ &\quad +\left(\widehat{\varphi}(X_i) - \varphi(X_i)\right)\left(\widehat{\Phi}(X_i) - \Phi(X_i)\right) R_n^2(X_i) \end{aligned}$$

In this expression the remaining terms $R_n^1(X_i)$ and $R_n^2(X_i)$ are :

$$R_n^1(X_i) = -\frac{\alpha_n \widehat{\Phi}(X_i) + (1 - \alpha_n)\Phi(X_i)}{\left(\alpha_n \widehat{\varphi}(X_i) + (1 - \alpha_n)\varphi(X_i)\right)^3}$$

and

$$R_n^2(X_i) = \tfrac{-1}{2}\frac{1}{\left(\alpha_n \widehat{\varphi}(X_i) + (1 - \alpha_n)\varphi(X_i)\right)^2}$$

Or, if grouping the terms :

$$\widehat{f_n}(X_i) - f(X_i) = \frac{\widehat{\Phi}(X_i) - f(x_i)\cdot\widehat{\varphi}(X_i)}{\varphi(X_i)} + R_n(X_i)$$

with $\alpha_n \in [0,1]$, and R_n defined as :

$$\begin{aligned} R_n(X_i) &= \left(\widehat{\varphi}(X_i) - \varphi(X_i)\right)^2 R_n^1(X_i) \\ &\quad + \left(\widehat{\varphi}(X_i) - \varphi(X_i)\right)\left(\widehat{\Phi}(X_i) - \Phi(X_i)\right) R_n^2(X_i) \end{aligned}$$

then F^+ is the sum of two terms :

$$F^+ = \frac{\sqrt{n}}{n}\left(\sum_{i=1}^{n} Z_i \left(\frac{\widehat{\Phi}(X_i) - f(x_i)\cdot\widehat{\varphi}(X_i)}{\varphi(X_i)}\right)\right) + \frac{\sqrt{n}}{n}\sum_{i=1}^{n} Z_i \cdot R_n(X_i) \quad (8.8)$$

$$F^+ = F_1^+ + F_2^+$$

We first show that $F_1^+ = \frac{\sqrt{n}}{n}\sum_{i=1}^{n} R_n(X_i)\cdot Z_i \xrightarrow{\mathcal{P}} 0$, using the following inequality :

$$\left|\frac{\sqrt{n}}{n}\sum_{i=1}^{n} Z_i \cdot \left(\widehat{\varphi}(X_i) - \varphi(X_i)\right)^2 R_n^1(X_i)\right|$$

$$\leq \sqrt{n}\left(Sup_{X_i}\ |\widehat{\varphi}(X_i) - \varphi(X_i)|\right)^2 \cdot \frac{1}{n}\sum_{i=1}^{n} |Z_i|\left|R_n^1(X_i)\right|$$

In a compact neighbourhood of f , the functional term $\left|R_n^1(X_i)\right|$ is bounded by a function $\rho^1(X_i)$ independent of n. And so, $\frac{1}{n}\sum_{i=1}^{n} |Z_i|\cdot\rho^1(X_i)$ is of order $O_p(1)$.

Moreover, $Sup_{X_i}\ |\widehat{\varphi}(X_i) - \varphi(X_i)|$is of order $Max\left(\frac{\sqrt{\log(n)}}{\sqrt{n\cdot h^p}}\ , h^d\right)$.

Under (8.4.4),that speed of convergence is such that :

$$\sqrt{n}\left(Sup_{X_i}\ |\widehat{\varphi}(X_i) - \varphi(X_i)|\right)^2 \longrightarrow 0$$

The same technic is applied to R_n^2, and the announced result is proved.

Second step :

The asymptotic normality arises from the sum of the two terms E^+ and F_1^+ remaining in (8.8). These terms create G^+ :

$$G^+ = \frac{\sqrt{n}}{n}\cdot\sum_{i=1}^{n}\ Z_i\left(Y_i - f(X_i)\right) - \frac{\sqrt{n}}{n}\left(\sum_{i=1}^{n} Z_i\left(\frac{\widehat{\Phi}(X_i) - f(x_i)\cdot\widehat{\varphi}(X_i)}{\varphi(X_i)}\right)\right)$$

Using the nonparametric expression of $\widehat{\Phi}(X_i)$, we have :

$$\begin{aligned} G^+ &= \frac{\sqrt{n}}{n} \cdot \sum_{i=1}^{n} Z_i \left(Y_i - f(X_i)\right) \\ &- \frac{\sqrt{n}}{n} \cdot \sum_{i=1}^{n} \frac{1}{n} \sum\nolimits_{j=1}^{n} \frac{Z_i}{\varphi(X_i)} \left(Y_j - f(X_i)\right) \cdot \frac{1}{h^p} K\left(\frac{X_i - X_j}{h}\right) \end{aligned}$$

Eliminating the term $i = j$ does not affect the limit of the latter term and is usual in cross-validation :

$$\begin{aligned} G^+ &= \tfrac{\sqrt{n}}{n} \cdot \sum_{i=1}^{n} Z_i \left(Y_i - f(X_i)\right) \\ &- \tfrac{\sqrt{n}}{n \cdot (n-1)} \cdot \sum_{i,j \neq i}^{n} \tfrac{Z_i}{\varphi(X_i)} \left(Y_j - f(X_i)\right) \cdot \tfrac{1}{h^p} K\left(\tfrac{X_i - X_j}{h}\right) \end{aligned} \tag{8.9}$$

We are now able to treat this second term by U-statistics (see Serfling, 1980).
Let us define :

$$U_n = \frac{1}{n \cdot (n-1)} \sum_{i,j \neq i}^{n} \lambda_2 \left(S_i, S_j\right) \tag{8.10}$$

where $S_i = (X_i, Y_i, Z_i)$ and $\lambda_2 \left(S_i, S_j\right)$ the symmetric function defined by :

$$\begin{aligned} \lambda_2 \left(S_i, S_j\right) &= \tfrac{1}{2} \frac{Z_i}{\varphi(X_i)} \left(Y_j - f(X_i)\right) \cdot \frac{1}{h^p} K\left(\tfrac{X_i - X_j}{h}\right) \\ &+ \tfrac{1}{2} \frac{Z_j}{\varphi(X_j)} \left(Y_i - f(X_j)\right) \cdot \frac{1}{h^p} K\left(\tfrac{X_i - X_j}{h}\right) \end{aligned}$$

The U-statistic U_n is asymptotically equal to its projection $\widehat{U_n}$ upon the observation :

$$\widehat{U_n} = E\left[\lambda_2 \left(S_i, S_j\right)\right] + \frac{1}{n} \sum_{i=1}^{n} \lambda_i \left(S_i\right) \tag{8.11}$$

where $\lambda_i \left(s\right)$ is :

$$\lambda_i \left(s\right) = E\left[\lambda_2 \left(S_i, S_j\right) \mid S_i = s\right]$$

or,

$$\lambda_i \left(s\right) = E\left[\lambda_2 \left(S_i, S_j\right) \mid S_i = s\right] = E\left[E\left[\lambda_2 \left(S_i, S_j\right) \mid X_i = x\right] \mid Y_i = y, Z_i = z\right]$$

we obtain finally :

$$\lambda_i \left(S_i\right) = m(X_i) \left(Y_i - f(X_i)\right)$$

with $m(x) = E[Z \mid X = x]$.

This lead us to the conclusion because :

$$E[\lambda_n(S_i, S_j)] \longrightarrow 0$$

and then by definition of (8.10) and (8.11), we have :

$$\frac{\sqrt{n}}{n} \cdot \sum_{i,j \neq i}^{n} \lambda_n(\omega_i, \omega_j) \sim \frac{\sqrt{n}}{n} \cdot \sum_{i=1}^{n} m(X_i)(Y_i - f(X_i))$$

Finally, if we recombine the terms involved in (8.9), the statistic is :

$$\sqrt{n} \cdot \left(\widehat{\gamma} - \widehat{\Gamma}(\widehat{f_n})\right) = Var(z)^{-1} \cdot \left[\frac{\sqrt{n}}{n} \sum_{i=1}^{n} (Z_i - m(X_i))(Y_i - f(X_i))\right] + O_p(1)$$

Under $\mathcal{H}_1$, the general term of this i.i.d. sum is centred.

Using the same arguments than in the PP case, and under the homoscedasticity hypothesis (8.6), the variance of this term is $\sigma^2 . E[Var(Z \mid X)]$. □