

# Semi and Nonparametric Models in Econometrics

## Part 2: Estimation of semiparametric models

Xavier d'Haultfoeuille  
CREST-INSEE

# Outline

Introduction

First strategy: getting rid of  $h_0$

Second strategy: two-step estimators

Third strategy: sieve estimation

# Outline

## Introduction

First strategy: getting rid of  $h_0$

Second strategy: two-step estimators

Third strategy: sieve estimation

# Introduction

- ▶ In this section, we consider semiparametric models where the distribution of the variables depend on a  $\beta_0 \in \mathbb{R}^k$  and an infinite dimensional parameter  $h_0 \in H$ .
- ▶ This is very common when one does not want to impose arbitrary distribution restrictions.
- ▶  $h_0$  is usually a *nuisance parameter*, i.e. not directly a parameter of interest. Otherwise, we rather talk of *semi-nonparametric* models.

# Introduction

- ▶ There is not a unique way to conduct inference in these models, but several general strategies.
- ▶ A general result is that despite  $h_0$ , we can often achieve root-n consistency of  $\widehat{\beta}$ .
- ▶ An important issue that is hardly addressed here is efficiency, i.e. find an estimator such that the asymptotic variance is minimal.

## Introduction

- ▶ Example 1: single index models. The dependence of  $Y$  in  $X$  only depends on  $X'\beta_0$ :  $f_{Y|X} = f_{Y|X'\beta_0}$ . In this case, equivalently, one can write

$$Y = g(X'\beta_0, \varepsilon), \text{ with } X \perp\!\!\!\perp \varepsilon.$$

where  $g$  and the distribution of  $\varepsilon$  are possibly unknown. Such models encompass several cases:

- ▶ binary models with independent errors: in this case  $g(u, v) = \mathbb{1}\{u + v \geq 0\}$ ,  $X \perp\!\!\!\perp \varepsilon$  but the distribution of  $\varepsilon$  is not specified.
- ▶ tobit models: as before but  $g(u, v) = \max(u + v, 0)$ .
- ▶ transformation models:  $\lambda(Y) = X'\beta_0 + \varepsilon$  for some strictly increasing  $\lambda$ , and  $X \perp\!\!\!\perp \varepsilon$  (example: Box-Cox models). In this case  $g(u, v) = \lambda^{-1}(u + v)$ .

## Introduction

- ▶ Example 2: semiparametric sample selection model. Suppose that we observe  $Y$  only when  $D = 1$ , with:

$$\begin{cases} Y &= X'\beta_0 + \varepsilon \\ D &= \mathbb{1}\{Z'\gamma_0 + \eta \geq 0\} \end{cases} \quad (1)$$

where  $(\varepsilon, \eta) \perp\!\!\!\perp X$  and, in general,  $\varepsilon$  and  $\eta$  are dependent.

- ▶ Example: Heckman (1974)'s model of female labor supply. Heckman supposes that  $(\varepsilon, \eta)$  are jointly normal. Is it possible to drop this assumption?
- ▶ Such a model is related to partially linear models of the kind

$$Y = X'\beta_0 + h_0(T) + \varepsilon, \text{ with } E(\varepsilon|X, T) = 0.$$

Indeed, in (1), we have

$$E(Y|D = 1, X) = X'\beta_0 + E(\varepsilon|X, \eta \geq -Z'\gamma_0) = X'\beta_0 + h_0(Z'\gamma_0).$$

## Introduction

- ▶ Example 3: treatment effects under exogeneity. Suppose that we are interested in the effect of a treatment  $T \in \{0, 1\}$  on an outcome. Let  $Y_i$  denote the outcome that would arise if  $T = i$  and  $Y = TY_1 + (1 - T)Y_0$  be the observed outcome.
- ▶ Suppose also that the treatment is exogenous, i.e.,

$$(Y_0, Y_1) \perp\!\!\!\perp T | X \quad (2)$$

- ▶ We are interested in the average effect of  $T$ ,  $\Delta_0 = E(Y_1 - Y_0)$ . Under Condition (2),

$$\begin{aligned} \Delta_0 &= E[E(Y_1|X) - E(Y_0|X)] \\ &= E[E(Y_1|T = 1, X) - E(Y_0|T = 0, X)] \\ &= E[E(Y|T = 1, X) - E(Y|T = 0, X)]. \end{aligned}$$

- ▶ Let  $h_{0j}(x) = E(Y|T = j, X = x)$ , then  $\Delta_0$  depends on  $(h_{00}, h_{01})$  since  $\Delta_0 = E(h_{01}(X) - h_{00}(X))$ .

# Outline

Introduction

First strategy: getting rid of  $h_0$

Second strategy: two-step estimators

Third strategy: sieve estimation

## Well known examples

- ▶ Best examples: linear models or quantile models considered so far.
- ▶ Note that these models are truly semiparametric.
- ▶ In the linear model  $Y = X'\beta_0 + \varepsilon$  with  $E(\varepsilon|X) = 0$ , for instance, the nuisance parameter  $h_0$  is the distribution of  $\varepsilon$  conditional on  $X$ .

## Example 1: the maximum rank correlation estimator

- ▶ Consider Example 1 and suppose that  $g(u, v) = D \circ F(u, v)$ , where  $F(., .)$  is strictly increasing in both arguments and  $D$  is increasing. Then, under regularity conditions, one can show that

$$\begin{aligned}\beta_0 &= \arg \max_{\beta} \text{Corr}(F_Y(Y), F_{X'\beta}(X'\beta)) \\ &= \arg \max_{\beta} E(F_Y(Y)F_{X'\beta}(X'\beta)).\end{aligned}$$

- ▶ The maximum rank correlation estimator is then defined as

$$\hat{\beta} = \arg \max_{\beta} \sum_{i=1}^n R(Y_i)R(X'_i\beta),$$

where, for a r.v.  $U$ ,  $R(U_i)$  denotes the rank of  $U_i$  in  $(U_1, \dots, U_n)$ .

## Example 2: a simple difference method

- ▶ Consider the partially linear model  $Y = X'\beta_0 + h_0(T) + \varepsilon$ , where  $T$  is real.
- ▶ To estimate very easily  $\beta_0$  in this model, proceed as follows:
  - ▶ Sort the data according to  $T$ , so that  $(T_{(1)} \leq \dots \leq T_{(n)})$ ;
  - ▶ Compute the differences  $\Delta Y_i = Y_{(i)} - Y_{(i-1)}$  and  $\Delta X_i = X_{(i)} - X_{(i-1)}$ ;
  - ▶ Regress  $\Delta Y_i$  on  $\Delta X_i$ .
- ▶ One can show that if  $h_0$  is smooth, the corresponding estimator  $\hat{\beta}$  is root- $n$  consistent and normal. The idea behind is that  $h_0(T_{(i)}) - h_0(T_{(i-1)})$  is small in this case.

## Example 2: a simple difference method

Thus, we can estimate the semiparametric sample selection as follows:

- ▶ Estimate  $\gamma_0$  (up to scale) by the maximum rank correlation estimator. Let  $\hat{\gamma}$  be the corresponding estimator and let  $\hat{T}_i = Z_i' \hat{\gamma}$ ;
- ▶ Sort the data according to the  $\hat{T}_i$ . Let  $\Delta Y_i = Y_{(i)} - Y_{(i-1)}$  and  $\Delta X_i = X_{(i)} - X_{(i-1)}$ ;
- ▶ Regress  $\Delta Y_i$  on  $\Delta X_i$ .

## General results in this case

- ▶ In general, in such cases, asymptotic results of  $\hat{\beta}$  follow from general results on M- or Z-estimators.
- ▶ For the maximum rank correlation example, this is a bit more involved but one can show that  $\hat{\beta}$  is root-n consistent and asymptotically normal (see Sherman, 1993).
- ▶ The advantage of this approach is that it does not require estimating  $h_0$ , which may be difficult and usually depends on tuning parameters (such as the bandwidth in kernel estimation) which are difficult to choose in practice.
- ▶ A drawback is often an efficiency loss, as with the maximum rank correlation estimator or the difference estimator considered above.
- ▶ Besides, computation may not be that easy (Manski's maximum score, maximum rank correlation...).

# Outline

Introduction

First strategy: getting rid of  $h_0$

**Second strategy: two-step estimators**

Third strategy: sieve estimation

## Introduction

- ▶ In some cases we cannot get rid of the estimation of  $h_0$ .  $\beta_0$  is then defined as a function of the data and  $h_0$  (cf. treatment effects).
- ▶ Then a solution is to estimate  $h_0$  nonparametrically in a first step.
- ▶ We consider here cases where  $\beta_0$  satisfies the moment condition:

$$E [g(U, \beta_0, h_0)] = 0.$$

where  $U$  denotes the data for an observation (i.e.,  $(X, Y)$  in standard cases).

- ▶ we consider the estimator  $\hat{\beta}$  defined by

$$\frac{1}{n} \sum_{i=1}^n g(U_i, \hat{\beta}, \hat{h}) = 0,$$

where  $\hat{h}$  is a first step nonparametric estimator of  $h_0$ .

## Introduction

- ▶ Example 3 (continued):  $g(U, \Delta, h_0) = \Delta - h_{01}(X) - h_{00}(X)$ .  $h_{0j}(\cdot) = E(Y|T = j, X = \cdot)$  can be estimated, e.g., by the kernel estimator

$$\hat{h}_j(x) = \frac{\sum_{i/T_i=j} Y_i K\left(\frac{x-X_i}{h_n}\right)}{\sum_{i/T_i=j} K\left(\frac{x-X_i}{h_n}\right)}.$$

Thus  $\hat{h}_j(x)$  is a weighted mean of the  $Y$ s, with a weight which depends on the distance between  $X$  and  $x$ .

- ▶ In this example  $\hat{\Delta}$  takes a simple form

$$\hat{\Delta} = \frac{1}{n} \sum_{i=1}^n \hat{h}_1(X_i) - \hat{h}_0(X_i).$$

## Informal discussion

- ▶ An issue, at first glance: nonparametric estimators of  $h_0$  are not root-n consistent in general.
- ▶ Example: estimators of the regression function  $h_0(\cdot) = E(Y|X = \cdot)$  cannot converge at a rate faster than  $n^{2/5}$  if  $h_0$  is twice differentiable and  $X \in \mathbb{R}$ .
- ▶ So we may think that the estimator of  $\beta_0$  is not root-n consistent either.
- ▶ However, we can prove under reasonable restrictions that  $\beta_0$  is root-n consistent and asymptotically normal.

## Informal discussion

- ▶ Intuition behind: consider for instance the quantity  $\hat{\beta} = \frac{1}{n} \sum_{i=1}^n \hat{h}(X_i)$ , where  $\hat{h}$  is a kernel nonparametric estimate of  $h_0$  with bandwidth  $h_n$ . Then, letting  $\bar{h}(\cdot) = E(\hat{h}(\cdot))$ ,

$$\begin{aligned} \sqrt{n} \left( \hat{\beta} - E(h_0(X)) \right) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left( \hat{h} - \bar{h} \right) (X_i) + \frac{1}{\sqrt{n}} \sum_{i=1}^n \left( \bar{h} - h_0 \right) (X_i) \\ &\quad + \frac{1}{\sqrt{n}} \sum_{i=1}^n \left( h_0(X_i) - E(h_0(X)) \right). \end{aligned}$$

- ▶ The third term tends to a normal distribution. The second is a “bias term” of order  $\sqrt{nh_n^2}$ . Thus we can make it tends to zero quickly enough by choosing a small bandwidth parameter.

## Informal discussion

- ▶ The first term  $T_1$  is a “variance term”. Usually the variance of  $\hat{h}(u)$  at a *given point*  $u$  is proportional to  $1/nh_n$ . Thus, if  $X_i = u$  for all  $i$ , we would have

$$T_1 = \sqrt{n} \left( \hat{h}(u) - \bar{h}(u) \right) \propto \frac{1}{h_n} \rightarrow \infty.$$

- ▶ However here we are averaging over different  $X_i$ . If the different terms in the sum are not “too dependent”, the variance of this term will be also small.

## The method

- ▶ Let  $g_0(U, h) = g(U, \beta_0, h)$ . The main difficulty is to prove that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n g_0(U_i, \hat{h}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Omega). \quad (3)$$

- ▶ Then it suffices to use a first order Taylor expansion (I suppose here for simplicity that  $\dim g = \dim \beta_0$ ):

$$0 = \frac{1}{\sqrt{n}} \sum_{i=1}^n g(U_i, \hat{\beta}, \hat{h}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n g_0(U_i, \hat{h}) + \left[ \frac{1}{n} \sum_{i=1}^n \frac{\partial g}{\partial \beta}(U_i, \tilde{\beta}, \hat{h}) \right] \sqrt{n} (\hat{\beta} - \beta_0).$$

- ▶ The first term converges to a normal distribution by (3). With some additional efforts, we can prove that

$$\frac{1}{n} \sum_{i=1}^n \frac{\partial g}{\partial \beta}(U_i, \tilde{\beta}, \hat{h}) \xrightarrow{P} E \left[ \frac{\partial g}{\partial \beta}(U, \beta_0, h_0) \right].$$

- ▶ This finally yields:

$$\sqrt{n} (\hat{\beta} - \beta_0) \xrightarrow{\mathcal{L}} \mathcal{N} \left( 0, E \left[ \frac{\partial g}{\partial \beta}(U, \beta_0, h_0) \right]^{-1} \Omega E \left[ \frac{\partial g}{\partial \beta}(U, \beta_0, h_0) \right]^{-1} \right).$$

## The method

To prove (3), we proceed in four steps (cf. Newey and McFadden, 1994):

1. Linearization of  $g$ : prove that, for some norm  $\|\cdot\|$ , there exists a functional  $G(\cdot, \cdot)$  linear in its second argument, such that

$$\|g_0(U, h) - g_0(U, h_0) - G(U, h - h_0)\| \leq b(U) \|h - h_0\|^2.$$

2. Rate of convergence of  $\hat{h}$ : prove that  $n^{1/4} \|\hat{h} - h_0\| \xrightarrow{P} 0$ .

These two steps show that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n g_0(U_i, \hat{h}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n g_0(U_i, h_0) + \frac{1}{\sqrt{n}} \sum_{i=1}^n G(U_i, \hat{h} - h_0) + o_P(1).$$

The first term is easy to deal with, so it remains to handle the second one.

## The method

3. Stochastic equicontinuity: prove that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \left[ G(U_i, \hat{h} - h_0) - \int G(u, \hat{h} - h_0) dF_U(u) \right] \xrightarrow{P} 0.$$

4. Linearization of the integral: prove that there exists  $\delta(\cdot)$  such that  $E(\delta(U)) = 0$  and

$$\int G(u, \hat{h} - h_0) dF_U(u) = \frac{1}{n} \sum_{i=1}^n \delta(U_i) + o_P\left(\frac{1}{\sqrt{n}}\right).$$

These two steps prove that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n g_0(U_i, \hat{h}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n [g_0(U_i, h_0) + \delta(U_i)] + o_P(1).$$

## The method

Thus, by Slutski's lemma and the CLT,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n g_0(U_i, \hat{h}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, V(g_0(U, h_0) + \delta(U))).$$

The first term in the asymptotic variance comes from variations of  $g_0(U_i, h_0)$ , while the second accounts for the nonparametric estimation of  $h_0$ .

## The treatment effect example

- ▶ Instead of  $\Delta_0 = E(Y_1 - Y_0)$ , let us consider for simplicity  $\beta_0 = E(Y_1) = E[h_{01}(X)]$ .
- ▶ The kernel estimator  $\hat{h}_1$  of  $h_{01}$  is defined as a ratio  $\hat{N}(x)/\hat{D}(x)$ , with:

$$\hat{N}(x) = \frac{1}{nh_n} \sum_{i=1}^n Y_i T_i K\left(\frac{x - X_i}{h_n}\right)$$

$$\hat{D}(x) = \frac{1}{nh_n} \sum_{i=1}^n T_i K\left(\frac{x - X_i}{h_n}\right).$$

$\hat{N}(x)$  and  $\hat{D}(x)$  are estimates for  $N_0(x) = E(TY|X = x)f_X(x)$  and  $D_0(x) = E(T|X = x)f_X(x)$  respectively.

## The treatment effect example

- ▶ Then let  $\hat{h} = (\hat{N}, \hat{D})$ ,  $h_0 = (N_0, D_0)$  and for all  $h = (N, D)$ ,  
 $g_0(U_i, h) = \beta_0 - N(X_i)/D(X_i)$ .
- ▶ Dealing with  $\hat{h}$  rather than  $\hat{h}_1$  is easier because  $\hat{N}$  and  $\hat{D}$  can be written as averages.
- ▶ We define (for instance)  $\|h\| = \max(\sup_x |N(x)|, \sup_x |D(x)|)$ .

## The treatment effect example

First step:

- ▶ we use a first order expansion of  $(N, D) \mapsto N/D$ :

$$g_0(u, h) = g_0(u, h_0) + \frac{1}{D_0(x)} [N(x) - N_0(x) - h_{01}(x)(D(x) - D_0(x))] + R(u)$$

The second term, which is linear in  $h = (N, D)$  corresponds to  $G(u, h - h_0)$ .

- ▶ Then one can show using standard analysis that the reminder term satisfies

$$\|R(u)\| \leq b(u) \|h - h_0\|^2.$$

## The treatment effect example

Second step:

- ▶ We use

$$\left\| \widehat{h} - h_0 \right\| \leq \left\| \widehat{h} - E(\widehat{h}) \right\| + \left\| E(\widehat{h}) - h_0 \right\|.$$

- ▶ When  $X$  is univariate,  $h_0$  is twice differentiable and the kernel satisfies  $\int uK(u)du = 0$ , we have  $\left\| \widehat{h} - E(\widehat{h}) \right\| = O_P(1/\sqrt{nh_n})$  and  $\left\| E(\widehat{h}) - h_0 \right\| \leq C_1 h_n^2 \ln(n)$ .
- ▶ Thus,  $n^{1/4} \left\| \widehat{h} - h_0 \right\| \xrightarrow{P} 0$  provided that  $n \ln^4(n) h_n^8 \rightarrow 0$  and  $nh_n^2 \rightarrow \infty$ .
- ▶ N.B.: one may have to impose strong additional restrictions to achieve a similar result when  $\dim(X)$  is large.

## The treatment effect example

- ▶ The third step is a bit tedious (relies on projection theorems of  $U$ -statistics).
- ▶ As for the fourth step, let  $K_{h_n}(u) = K(u/h_n)/h_n$  and  $Z_i = (T_i Y_i, T_i)$ , so that

$$\hat{h}(x) = \frac{1}{n} \sum_{i=1}^n Z_i K_{h_n}(x - X_i)$$

- ▶ We have  $G(u, \gamma) = v(x)\gamma(x)$  with  $v(x) = (1, -h_{01}(x))/D_0(x)$ . Thus,

$$\begin{aligned} \int G(u, \hat{h}) dF_U(u) &= \frac{1}{n} \sum_{i=1}^n \left[ \int v(x) K_{h_n}(x - X_i) dF_X(x) \right] Z_i \\ &= \frac{1}{n} \sum_{i=1}^n f_X(X_i) v(X_i) Z_i + B_i \end{aligned}$$

where  $B_i = \left[ \int K_{h_n}(x - X_i) v(x) f_X(x) dx - v(X_i) f_X(X_i) \right] Z_i$  is a bias term stemming from kernel smoothing.

## The treatment effect example

- ▶ Similarly, one can show that

$$\int G(u, h_0) dF_U(u) = E(f_X(X)v(X)Z).$$

- ▶ As a result, letting  $\delta(U_i) = f_X(X_i)v(X_i)Z_i - E[f_X(X)v(X)Z]$ , we get:

$$\int G(u, \hat{h} - h_0) dF_U(u) = \frac{1}{n} \sum_{i=1}^n \delta(U_i) + B_i.$$

- ▶ It remains to show that  $(\sum_{i=1}^n B_i)/\sqrt{n} \xrightarrow{P} 0$ . For that purpose, note that by a change of variables,

$$\begin{aligned} \int K_{h_n}(x - X_i)v(x)f_X(x)dx &= \int K(u)v(X_i + uh_n)f_X(X_i + uh_n)du \\ &= v(X_i)f_X(X_i) + h_n(v \times f_X)'(X_i) \int uK(u)du \\ &\quad + \frac{h_n^2}{2}(v \times f_X)''(\tilde{X}_i) \int u^2K(u)du. \end{aligned}$$

## The treatment effect example

- ▶ The second term vanishes if  $\int uK(u)du = 0$ . In this case, letting

$$W_i = \left[ (v \times f_X)''(\tilde{X}_i) \int u^2 K(u)du \right] Z_i,$$

we get

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n B_i = (\sqrt{nh_n^2}) \frac{1}{n} \sum_{i=1}^n W_i.$$

- ▶ The second term tends to a constant, thus the result holds provided that  $nh_n^4 \rightarrow 0$ .
- ▶ To sum up, if  $nh_n^2 \rightarrow \infty$  and  $nh_n^4 \rightarrow 0$  (e.g.,  $h_n = n^{-1/3}$ ), we get:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n g_0(U_i, \hat{h}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, V(g_0(U, h_0) + \delta(U))). \quad (4)$$

## The treatment effect example

- ▶ Remark 1: as the intuition above indicated, compared to standard nonparametric estimation where the optimal bandwidth is  $h_n \propto n^{-1/5}$ , we have to “undersmooth” here by taking a smaller  $h_n$ .
- ▶ Remark 2: by definition of  $\hat{\beta}$ , Equation (4) directly translates into:

$$\sqrt{n} \left( \hat{\beta} - \beta_0 \right) \xrightarrow{\mathcal{L}} \mathcal{N} \left( 0, V(g_0(U, h_0) + \delta(U)) \right).$$

- ▶ Remark 3: we can rewrite the asymptotic variance  $V_{\text{as}}$  as

$$V_{\text{as}} = V \left( E(Y|T = 1, X) + \frac{T(Y - E(Y|T = 1, X))}{P(T = 1|X)} \right).$$

One can show that this is the efficiency bound (see, e.g., Hahn, 1998), i.e.  $\hat{\beta}$  is “asymptotically efficient”, as the maximum likelihood estimator in a parametric setting.

# Outline

Introduction

First strategy: getting rid of  $h_0$

Second strategy: two-step estimators

Third strategy: sieve estimation

## Introduction

- ▶ In Example 2, estimation would be easy if the distribution of  $(\varepsilon, \eta)$  were specified: we could resort to likelihood maximization for instance.
- ▶ Then an idea would be to choose a flexible parametric distribution for  $(\varepsilon, \eta)$ , and estimate the parameters by maximum likelihood.
- ▶ Sieve estimation is based on this idea: replace  $H$  by a flexible enough parametric space  $H_n$  and estimate the model as if it was parametric.
- ▶ Note that this idea is very general, and applies not only to MLE but also to GMM, M-estimation...
- ▶ An advantage is that we get estimates of both  $\beta_0$  and  $h_0$ .

## Example 2 (continued)

- ▶ In the sample selection model, we maximize the joint likelihood:

$$\begin{aligned}
 (\hat{\beta}, \hat{\gamma}, \hat{h}) &= \arg \max_{\beta, \gamma, h \in H_n} \frac{1}{n} \sum_{i=1}^n D_i \ln \left( \int_{-Z_i' \gamma}^{\infty} h(Y_i - X_i' \beta, v) dv \right) \\
 &\quad + (1 - D_i) \ln \left[ \int_{-\infty}^{\infty} \left( \int_{-Z_i' \gamma}^{\infty} h(u, v) dv \right) du \right].
 \end{aligned}$$

- ▶ In the partially linear model  $Y = X' \beta_0 + h_0(T) + \varepsilon$ , the difficulty stems from function  $h_0$ .
- ▶ It is difficult to use the previous two-step estimation strategy here since  $h_0$  depends on  $\beta_0$ :  $h_0(T) = E(Y - X' \beta_0 | T)$ .
- ▶ Instead, consider a linear space  $H_n = \text{Vect}(q_1, \dots, q_{k_n})$ . Then compute the OLS estimator of  $Y$  on  $(X, q_1(T), \dots, q_{k_n}(T))$ .

## Introduction

- ▶ The difference with standard parametric estimation is that  $H_n$  increases with  $n$ :  $H_n \subseteq H_{n+1} \subseteq H$
- ▶ Besides,  $\cup_n H_n$  should be dense in  $H$ . This ensures that asymptotically, we can be as close as possible to any  $h \in H$ , avoiding thus any misspecification.
- ▶ Because  $H_n$  changes with  $n$ , standard parametric tools cannot be used to study the asymptotics of  $\hat{\beta}$ .
- ▶ As previously, the parametric part  $\hat{\beta}$  is usually root- $n$  consistent. It is also often possible to develop efficient sieve estimator of  $\beta_0$ .
- ▶ Another advantage is that we can easily impose restrictions on  $h$  such as positivity, monotonicity, convexity... It suffices indeed to ensure that  $H_n$  satisfies these constraints.

## Example of sieve bases

- ▶ Several choices are possible for the approximating space  $H_n$ .
- ▶ For functions on  $[0, 1]$  (or any bounded interval), one may use polynomials, trigonometric polynomials or *polynomial splines* (polynomials of order  $r + 1$  on  $m_n$  subintervals,  $r$  times differentiable on  $[0, 1]$ ).
- ▶ For functions on  $\mathbb{R}$ , one may use for instance the linear span of  $(x \mapsto \varphi^{(k)}(x))_{k \leq k_n}$ , where  $\varphi$  is the density of a  $\mathcal{N}(0, 1)$ .
- ▶ For functions on  $\mathbb{R}^D$ , one may consider the tensor product of real functions, i.e.  $\prod_{d=1}^D g_d(x_d)$ , where  $g_d$  belongs to the basis considered for real functions. For instance, for polynomials, we consider all functions of the form

$$(x_1, \dots, x_D) \mapsto \prod_{d=1}^D x_d^{j_d}, \quad \text{with } 1 \leq j_d \leq k_n.$$

## Consistency

- ▶ Let  $\theta = (\beta, h)$ ,  $\Theta = \mathbb{R}^k \times H$  and  $\Theta_n = \mathbb{R}^k \times H_n$ . Let also  $Q(\cdot)$  denote a real function such that  $Q(\theta_0) > Q(\theta)$  for all  $\theta \neq \theta_0$ , and let  $Q_n(\cdot)$  denote an estimator of  $Q(\cdot)$ .
- ▶ Consider the following *sieve extremum estimator*:

$$\hat{\theta} = \arg \max_{\theta \in \Theta_n} Q_n(\theta).$$

- ▶ This case includes  $M$ -estimation (and thus MLE), GMM...
- ▶ Let  $d$  be a metric on  $\Theta$ , then  $d(\hat{\theta}_n, \theta_0) \xrightarrow{P} 0$  if
  - ▶ (sieve spaces)  $\Theta_n \subseteq \Theta_{n+1} \subseteq \Theta$  for all  $n$ , and there exists a sequence  $\pi_n \theta_0 \in \Theta_n$  such that  $d(\theta_0, \pi_n \theta_0) \rightarrow 0$ .
  - ▶ (continuity)  $Q(\cdot)$  is uppersemicontinuous on  $\theta$  with respect to  $d$ .
  - ▶ (compactness)  $\Theta_n$  is compact under  $d$ .
  - ▶ (uniform convergence over sieves)  $\sup_{\theta \in \Theta_n} |\hat{Q}_n(\theta) - Q(\theta)| \xrightarrow{P} 0$ .

## Consistency

- ▶ As usually, the most difficult condition to check is uniform convergence. If we restrict to sieve M-estimators, i.e.  $Q(\theta) = E(g(U, \theta))$  and

$$Q_n(\theta) = \frac{1}{n} \sum_{i=1}^n g(U_i, \theta),$$

then sufficient conditions for this to hold are the following:

- ▶  $E(\sup_{\theta \in \Theta_n} |g(U_i, \theta)|) < \infty$ .
- ▶ There exist  $s > 0$  and  $q(\cdot)$  such that  $E(|q(U)|) < \infty$  and

$$\sup_{\substack{(\theta, \theta') \in \Theta_n^2 \\ d(\theta, \theta') \leq \delta}} |g(U_i, \theta) - g(U_i, \theta')| \leq \delta^s q(U_i). \quad (5)$$

- ▶  $\ln N(\delta^{1/s}, \Theta_n, d) = o(n)$ .

Here, for a class of functions  $\mathcal{F}$ ,  $N(\delta, \mathcal{F}, d)$  is the *entropy without bracketing*, i.e. the minimum number of balls of size  $\delta$  (for the metric  $d$ ) needed to cover  $\mathcal{F}$ .

## Rate of convergence

- ▶ Consider a sieve M-estimator defined by

$$\hat{\theta} = \arg \max_{\theta \in \Theta_n} \frac{1}{n} \sum_{i=1}^n g(U_i, \theta)$$

- ▶ As often, there is a trade-off between the variance of  $\hat{\theta}$ , which is small when  $\Theta_n$  is small, and a bias term, which is small when  $\Theta_n$  is large.
- ▶ Suppose that, for a suitable norm  $\|\cdot\|$ :
  - ▶ There exists  $C_1$  and  $K$  such that for all  $\varepsilon < K$ ,

$$\sup_{\theta \in \Theta_n: \|\theta - \theta_0\| < \varepsilon} V(g(U, \theta_0) - g(U, \theta)) < C_1 \varepsilon^2.$$

- ▶ Condition (5) holds, with  $E(q(U)^2) < \infty$  and  $d(u, v) = \|u - v\|$ .

## Rate of convergence

- ▶ Besides, define

$\mathcal{F}_n = \{u \mapsto g(u, \theta) - g(u, \theta_0), \|\theta - \theta_0\| \leq \delta, \theta \in \Theta_n\}$  and for a given constant  $C$ , let

$$\delta_n = \inf \left\{ \delta \in (0, 1) : \frac{1}{\sqrt{n}\delta^2} \int_{\delta^2}^{\delta} \ln N_{[]} (w, \mathcal{F}_n, L_2) dw \leq C \right\},$$

where  $N_{[]} (w, \mathcal{F}_n, L_2)$  denotes the bracketing number already defined in part 1 of the course.

- ▶ Then:

$$\left\| \hat{\theta} - \theta_0 \right\| = O_P(\max(\delta_n, \|\theta_0 - \pi_n \theta_0\|)).$$

- ▶ As the size of  $\Theta_n$  increases,  $\delta_n$  increases while  $\|\theta_0 - \pi_n \theta_0\|$  decreases, whence the aforementioned trade-off.

## Asymptotic normality of $\widehat{\beta}$

- ▶ It is possible to prove root-n consistency and asymptotic normality of  $\widehat{\beta}$  using the previous result on the rates of convergence.
- ▶ The idea is quite similar to the four steps used to prove asymptotic normality of the two step estimator.
- ▶ The condition  $n^{1/4} \left\| \widehat{h} - h \right\| \xrightarrow{P} 0$  is replaced by

$$n^{1/2} \left\| \pi_n v^* - v^* \right\| \times \left\| \widehat{\theta} - \theta_0 \right\| \xrightarrow{P} 0,$$

for a particular  $v^*$  (see Chen, 2007, for more details).

- ▶ One can also prove that the estimator of  $\widehat{\beta}$  is asymptotically efficient: see Chen (2007) for details.