

# Nonparametric Multivariate Methods: an Application to the Empirics of Economic Growth

*Metodi non Parametrici Multivariati: un'Applicazione al Caso della Crescita*

Barbara Pacini, Guido Pellegrini

Dip. Scienze Statistiche,

Università di Bologna,

e-mail: [pacini@stat.unibo.it](mailto:pacini@stat.unibo.it) - : [pellegrini@stat.unibo.it](mailto:pellegrini@stat.unibo.it)

**Riassunto:** L'analisi non parametrica della forma e della dinamica della distribuzione di una variabile è stata recentemente utilizzata per identificare meccanismi di sviluppo e di polarizzazione (Quah, 1997a,b). La metodologia utilizzata per lo studio di quale variabile sia responsabile della presenza di tali polarizzazioni, nota come *conditioning*, è stata sviluppata principalmente per il caso univariato. Nel presente lavoro tale proposta viene rivista alla luce delle analogie con le tecniche di regressione non parametrica basate su criteri di approssimazione locale. In tale contesto trova applicazione anche lo studio dell'effetto congiunto di più variabili sulla forma della distribuzione. Un'esemplificazione viene fornita con riferimento allo studio della distribuzione del tasso di occupazione nei 784 sistemi locali del lavoro italiani.

**Keywords:** Nonparametric multivariate analysis, Conditioning, Growth empirics.

## 1. Introduction

Quah (1997a,b) recently proposed a new methodological approach to empirics of economic dynamics based on nonparametric analysis of the distribution of growth variables across countries and regions. Quah suggests two methodological tools: the stochastic kernel, a mathematical operator mapping distributions to distributions, and the conditioning scheme, that is analogous to explain, in a regression, a left-hand side variable by some right-hand side conditioning regressors. In models of distribution dynamics, the variable to be explained on the left-hand side is conceptually the whole distribution. The empirical finding in a univariate context is not completely satisfactory: actually, the shape of growth variable distributions is affected by several factors, and a multivariate approach is needed in order to assess the influence of each one. The aim of this paper is to propose a feasible multivariate approach to the conditioning scheme. The idea is to exploit the similarities between a suitable multivariate conditioning scheme and a local nonparametric estimator of a regression function. We present an empirical application on local labour market growth at a very disaggregate territorial level. The conditioning scheme is applied to the distribution of (non agricultural) employment rate in 784 Italian local labour markets (LLMs) in 1996. We show how proximity, population density and agricultural employment share affect the shape of the

employment rate distribution and explain its bimodality (the “twin peaks”, using Quah’s terminology).

## 2. The univariate conditioning scheme

The conditioning scheme, originally proposed by Quah (1997b), is the following: given a random variable  $X$  (the employment rate), observed on a set  $I$  (the Italian local labor markets, LLMs),  $X = \{X(i) : i \in I\}$ , we define a new variable, denoted by  $X^{**}$ , which is a weighted average of the original observations on  $X$  with weights  $w(i)$  obtained as suitable functions of other relevant variables:

$$X^{**} \equiv \sum_{i \in I} w(i) X(i). \quad (1)$$

The conditioning scheme,  $G$ , requires to introduce a set of weights  $W$  and to specify the relation between  $X$  and  $X^{**}$  as the following:

$$X^* = X \mid G \equiv \varphi\{X(i), X^{**}(i)\} \quad (2)$$

where  $X^*$  the conditioned variables and  $\varphi$  the operator that describes the functional relationship between the original data and the weighted data. The set of weights  $W$  can determine the relevance of each observation  $x(i)$  in  $I$  on  $x^{**}(j)$ . For example, we can assume the distance between labor markets as a relevant variable in determining the weighting structure: nearest neighbors (according to the specified measure of distance  $d_{ij}$ ) will have more influence in estimating  $x^{**}(j)$  than other observations:

$$x_j^{**} \equiv \sum_{i \in I} \left( \frac{n \cdot d_{ij}}{\sum_j \sum_i d_{ij} \cdot x_i} \right) \cdot x_i. \quad (3)$$

The weights represent the relative importance of the distance from each observation to  $x(j)$ . In our experiment, the weights are the simple inverse of the distance from the specific LLM. The distance is calculated by the centroids of the two LLM. Therefore, in our case, the conditioning procedure is based on the creation for each LLM of a “twin” LLM, where the “conditioned” employment rate is equal to a weighted sum of the simple employment rates of all LLMs, normalized to the average of the conditioned employment rate. For density and population, the weight is equal to 1 for the specific LLS, and 0 otherwise. We use the ratio between the original data and the weighted data to specify the operator  $\varphi$ :

$$x_j^* = \varphi(x_j, x_j^{**}) = \frac{x_j}{x_j^{**}}. \quad (4)$$

The weighted mean can influence the distribution shape in different ways depending on the choice of the conditioning variable and on the weighting structure. We consider now a different approach to the choice of weights: we can reformulate the problem referring to the literature on local nonparametric estimation of regression functions (see, for example, Wand and Jones, 1995). Let assume that the conditional mean of  $X$  given  $Z=z$  can be obtained as a local weighted mean of the response variables with weights depending on the conditioning variable:

$$\hat{m}(z) \equiv E(X | Z = z) \equiv X^{**} \equiv \sum_{i \in I} w(i, Z) X(i) \quad (5)$$

Referring to the traditional Nadaraya-Watson kernel regression estimator, we can define the following weights:

$$w(i, Z) = \frac{k_h(z - z_i)}{\sum_{i \in I} k_h(z - z_i)}, \text{ where } k_h(z - z_i) = \frac{1}{h} k\left(\frac{z - z_i}{h}\right). \quad (6)$$

The regression estimator for the  $j$ -th observation is given by:

$$\hat{m}(z_j) = \frac{\sum_{i \in I} k_h(z_j - z_i) \cdot x_i}{\sum_{i \in I} k_h(z_j - z_i)} \quad (7)$$

where  $h$  is the *bandwidth*, which controls the degree of smoothing, and  $k$  is the *kernel* function, which determines the shape of the weighting function. An alternative approach to the computation of a local weighed average is given by the local linear regression estimator (LOESS), which can be seen as the solution of a locally weighted least square problem. From (5) to (7) the similarity with the conditioning scheme is evident: we can use  $\hat{m}(z_j)$  as  $X^{**}$  in (1) and (4).

In doing so, most weight is given to the observation with value of  $z$  close to  $z_j$ . We divide each employment rate by a mean of observations which have substantially similar characteristics for  $z$ ; for example, if we consider as conditioning variable the population density in LLMs, each rate of employment will be divided by a mean of employment rates related to LLMs with a similar population density.

### 3. A multivariate approach to conditioning

The next step consists of considering two (or even more) conditioning variables to be jointly used in determining the weighted mean  $X^{**}$ . Suppose, for example, to be interested in showing how proximity and population density can affect the shape of the employment rate distribution. In this case  $X^{**}$  will be a weighted mean with weights  $w(i)$  in (1) obtained as a function of both variables, proximity and population density.

Following the Nadaraya – Watson kernel approach, we consider a vector  $\mathbf{Z}$  of  $d$  explanatory variables:

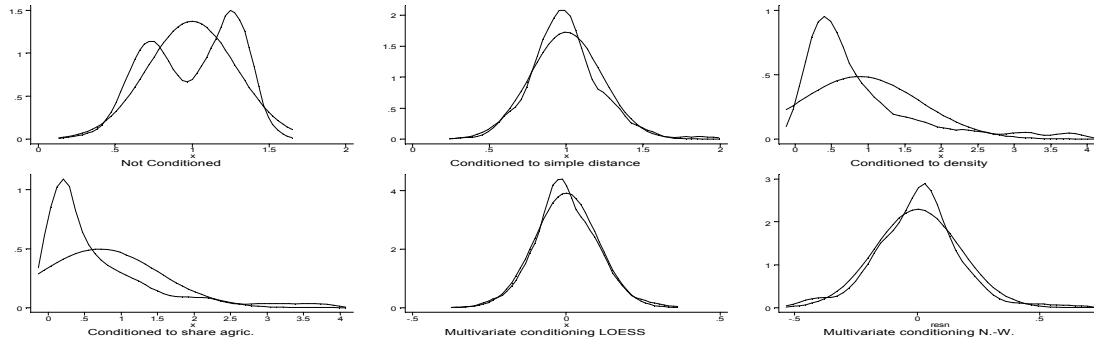
$$\hat{m}(\mathbf{z}) \equiv E(X | \mathbf{Z} = \mathbf{z}) \equiv \sum_{i \in I} w(i, \mathbf{Z}) X(i) \quad (8)$$

and define the weighting structure using a  $d$ -dimensional *bandwidth* matrix and a  $d$ -dimensional *kernel* function. In a similar way we can generalize the LOESS approach to include a vector  $\mathbf{Z}$  of explanatory variables.

## 4. Results and conclusions

In Figure 1 the density functions of unconditioned employment rate, employment rate conditioned to simple distance, population density and population in a univariate framework, and conditioned to the same variables in a multivariate framework (Nadaraya-Watson *kernel* estimator and the LOESS estimator), are plotted. The fitted values are taken as the  $X^{**}$  variable and, as before, we use the ratio between the original data and the weighted data to specify the operator  $\varphi$ . A normal density is also overlaid on the estimated density for comparison. The conditioning scheme changes radically the shape of the distribution, that shows a unique mode. The conditioning to distance appears to be strongly significant. As expected, the multivariate conditioned density function is closer to the normal density than in the univariate case.

**Figure 1:** *Unconditioned and conditioned kernel densities of the employment rate*



## References

- Quah D. T. (1997a), Empirics for Growth and Distribution: Polarization, Stratification and Convergence Clubs, *Journal of Economic Growth*, 2, March.  
 Quah D. T. (1997b), Regional cohesion from local isolated actions: II. Conditioning, *CEP Working Paper n. 379*, December.  
 Wand M.P. and M.C. Jones (1995), *Kernel Smoothing*, Chapman & Hall, London.