

# Nonparametric Models for Convergence Analysis in Europe

*Modelli non Parametrici per l'Analisi della Convergenza in Europa*

Guido Pellegrini, Barbara Pacini

Dipartimento di Scienze Statistiche, Università di Bologna

pellegrini@stat.unibo.it, pacini@stat.unibo.it

**Riassunto:** Se si considera seriamente che la crescita è un concetto intrinsecamente multidimensionale e non riducibile ad un unico indicatore, al contrario di quanto avviene in letteratura, è necessario sviluppare nuovi strumenti per l'analisi empirica della convergenza. In questo lavoro viene proposta la definizione di convergenza  $\sigma$  generalizzata, da applicare in contesti multidimensionali, e vengono sviluppati strumenti non parametrici basati sull'approccio del condizionamento. Questi vengono applicati allo studio della distribuzione congiunta del prodotto pro capite e del tasso di occupazione in 107 regioni europee.

**Keywords:** Multivariate Adaptive Spline Regression, Conditioning, Convergence analysis.

## 1. Introduction

Growth and convergence are clearly multidimensional concepts. The inadequacy of the one-dimensional approach— based on a single indicator, mainly the GDP per capita – in assessing levels of economic development of regions and disparities is not only debated among academics, but it is also acknowledged by the European Union, that recommends the use of a “basket of structural indicators” in analysing social cohesion and in formulating regional policies. Unfortunately, the difficulties in managing a multiple distribution instead has lead the literature to develop the convergence analysis in a univariate framework. Therefore, if we take seriously the issue of multidimensionality in growth, we have to develop new concepts and new instruments for analysing a multidimensional convergence. The first problem is that the variable to be explained is a vector, described by a multivariate distribution. If the economy is characterized by polarizations, this distribution could be multimodal, and it cannot be approximated by a well-know multivariate distribution (as a multivariate Gaussian distribution). Therefore the parametric approach is unsuitable, and the analysis must be set in a non parametric framework. The second problem is related to the definition of convergence if the target variable is a vector. We propose in this paper a possible solution, by extending the concept of  $\sigma$ -convergence, based of the reduction of the target variable dispersion across regions, to generalized  $\sigma$ -convergence, where the reduction concerns the hyper-volume described by the multidimensional vector. The third question is related to the identifications of the factors affecting the shape of the distribution. Quah (1997) proposed a new nonparametric approach to empirics of economic growth based on the conditioning scheme that is analogous to explain, in a

regression, a left-hand side variable by some right-hand side conditioning regressors. In Pacini and Pellegrini (2000) we extend the univariate conditioning scheme proposed by Quah to a multivariate framework. In this paper we propose a new extension of the conditioning approach, developed for the multidimensional approach to convergence. An empirical applications of these concepts to the analysis of growth factors for 107 European regions, considering simultaneously per capita GDP and employment rate, shows their explanatory power .

## 2. Extended conditioning approach

The approach to the analysis of convergence presented in this paper is based on the conditioning scheme, originally proposed by Quah (1997). Consider a random variable  $X$ , observed on a set  $I$ , and define a new variable, denoted by  $X^{**}$ , which is a weighted average of the original observations on  $X$  with weights  $w(i)$  depending on one suitable explanatory variable:

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

The conditioning scheme,  $G$ , consists of introducing a set of weights  $W$  and specifying the relation between  $X$  and  $X^{**}$  as follows:  $X^* = X | G \equiv \varphi(X(i), X^{**}(i))$ , where  $X^*$  is the conditioned variables and  $\varphi$  the operator that describes the functional relationship between the original data and the weighted data. The ratio between the original data and the weighted data can be a suitable specification of  $\varphi$  , so that we have  $x_j^* = \varphi(x_j, x_j^{**}) = x_j / x_j^{**}$  .

The weighted average in (1) can be alternatively stated as a problem of nonparametric conditional mean estimation of  $X$  given  $Z=z$  and can be obtained as a local weighted mean of the response variable with weights depending on the conditioning variable  $Z$ :

$$m(z) \equiv E(X | Z = z) \equiv X^{**} \equiv \sum_{i \in I} w(i, Z)X(i) \quad (2)$$

We now consider two (or even more) conditioning variables to be jointly used in determining a weighted mean vector  $\mathbf{X}^{**}$  of two (or even more) conditioned variables. Exploiting the suggestions in Pacini and Pellegrini (2002) the problem of determining  $\mathbf{X}^{**}$  can be reformulated as a nonparametric multivariate regression problem, where we now have a  $d$ -dimensional conditional mean vector depending on a vector  $\mathbf{Z}$  of  $m$  explanatory variables:

$$m(\mathbf{z}) \equiv \begin{pmatrix} \mu_{1|\mathbf{z}} \\ \vdots \\ \mu_{k|\mathbf{z}} \end{pmatrix} = E(\mathbf{X} | \mathbf{Z} = \mathbf{z}) \quad (3)$$

where the conditional expectation should be evaluated on the basis of the following  $d$ -dimensional conditional density function:

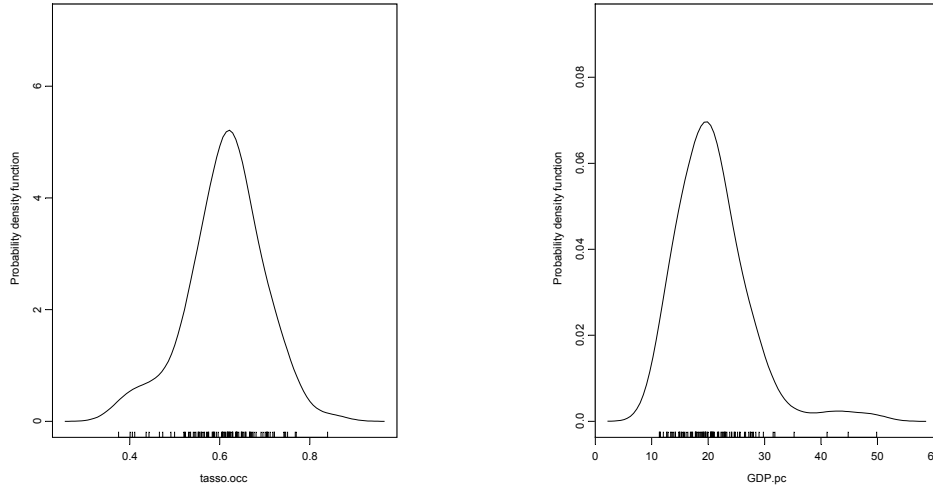
$$f(X_1, \dots, X_d | Z_1, \dots, Z_m) = \frac{f(X_1, \dots, X_d, Z_1, \dots, Z_m)}{f(Z_1, \dots, Z_m)} \quad (4)$$

Conditional expectations can be derived as weighted averages based on kernel functions of the Nadaraya-Watson type, which are straightforward to interpret and easy to implement. Other nonparametric techniques are, for example, based on locally weighted regression and regression splines. All these methods fit the regression function to the data points locally, so that at any point the fit depends only on some neighbor observations. Referring to the Multiple Adaptive Regression Splines (MARS) procedure, originally proposed by Friedman (1991), a multiple regression function can be locally approximated using linear splines and their tensor products, also allowing for multiple dependent variables.

#### 4. Empirical results and conclusions

The data set consists of 107 regional units, based on the classification proposed by Paci (1997), that considers regions at level NUTS 1, 2 and 3, in order to homogenize the geographical dimension of the grid single elements. The data refers on years 2000 (GDP and employment rate) and 1996 (other variables). As previously stated we locally approximate a multivariate regression function by the MARS algorithm, considering the GDP and the employment rate as dependent variables. Both the marginal density functions of unconditioned GDP and employment rate, smoothed by kernel density estimators, show modest signals of polarization and multimodality (Figure 1):

**Figure 1:** *Unconditioned kernel densities of GDP and employment rate*

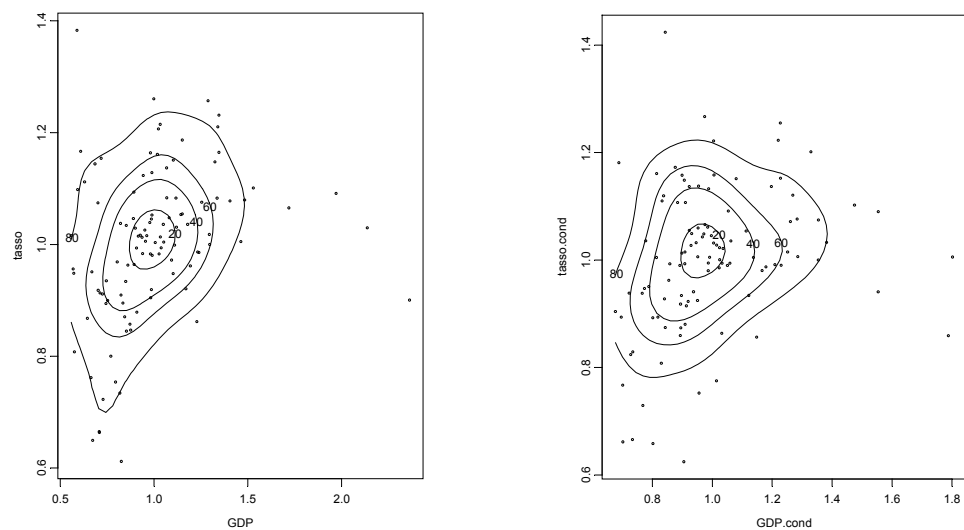


Polarization is more evident for per capita GDP: the Figure 1 shows the presence of a cluster of regions with a level of per capita income significantly lower than the average. In Figure 2 we plot the contour of the unconditioned bivariate density function estimation for gross domestic and employment rate and the same distribution conditioned to the following variables: structure (measured by the agricultural employment rate), technology (measured by the number of patents) and social capital (measured by the motorways/area ratio). For sake of brevity we do not report here

results based on different sets of variables. The conditioned bivariate density function is obtained by kernel multivariate density estimation applied on the two-dimension vector  $\mathbf{X}^*$ . The fitted values derived by MARS implementation are taken as the  $\mathbf{X}^{**}$  vector and the ratio between the original data and the weighted data is used to specify the operator  $\varphi$ .

The application of the multivariate conditioning scheme leads to a substantial reduction of the per capita GDP and employment rate dispersion across regions, showing that the endogenous variables have a role in explaining regional disparities: the generalized variance of the bivariate distribution is equal to 0.002 in the unconditioned distribution, to 0.0008 in the conditioned one. We also observe a decrease in the variance of the marginal distribution of both dependent variables, even if with a different intensity (from 0.095 to 0.047 for per capita GDP, from 0.0184 to 0.0180 for employment rate). Note that the reduction is stronger for the per capita GDP distribution, where the presence of polarization was more evident. Moreover, the bivariate conditioned distribution has clearly a unique mode and it is more concentrated. Therefore the multivariate conditioning appears to be strongly significant: some polarization in growth in Europe can be explained by differences in technology, social capital and structural development.

**Figure 2:** *Unconditioned and conditioned kernel bivariate densities – contour plot*



## References

- Friedman J. (1991), Multivariate Adaptive Regression Splines, *Annals of Statistics*, 19.  
Paci, R. (1997), More similar and less equal: economic growth in the European Regions, *Weltwirtschaftliches Archiv*, 133, 609-634.  
Pacini, B., Pellegrini, G. (2002), Metodi non parametrici multivariati: un'applicazione al caso della crescita, *Statistica*, 3.  
Quah D. T. (1997), Empirics for Growth and Distribution: Polarization, Stratification and Convergence Clubs, *Journal of Economic Growth*, 2, March.