Insper IBMEC São Paulo

Conditional Stochastic Kernel Estimation by Nonparametric Methods

Márcio Poletti Laurini

Pedro L. Valls Pereira

Insper Working Paper
WPE: 097/2007

Insper IBMEC São Paulo

CONDITIONAL STOCHASTIC KERNEL ESTIMATION BY NONPARAMETRIC METHODS

MÁRCIO POLETTI LAURINI
PEDRO L. VALLS PEREIRA

1. Introduction

In this article we made an application of nonparametric conditional density estimator proposed by Hyndman (1996) e Hyndman et al. (1996) to analyze the problem of conditional convergence studied in economic growth literature. The proposed methodology is a generalization of the methodology of Conditional Stochastic Kernel developed by Quah (1996,1998) to multiple and more general conditioning schemes and also of the method of Arbia et al. (2005) of analysis of unconditional convergence using nonparametric conditional density estimation.

We utilize this methodology to analyze conditional income convergence for Brazilian municipalities between 1970 and 1991, showing that the usually result of income club convergence is affected by the use of schooling as conditioning variable.

2. Methodology

The Distribution Dynamics methodology (Quah 1996) assumes that the density distribution ϕ_{t+1} for the variables under consideration, usually some measure of relative income, evolves according to a first order Markov process:

$$\phi_{t+1} = M \cdot \phi_t \tag{1}$$

In this formulation M is an operator mapping the transition between the income distribution existing in time t to the income distribution in time $t+1$.

The construction of the M operator can be done assuming that exists a finite number of states in ϕ_t distribution, using the model of Markov Transition Matrix (see Shorrocks (1978) for a definition of this method and some applications in mensuration mobility), or avoiding the state discretization and using a continuous state formulation for the M operator. This methodology is called Stochastic Kernel and can be defined (Quah (1996,1998)) as:

Definition: Let u and v be elements of B which are probabilities measures in $(\Re, \Re)$. A Stochastic Kernel relating u and v is a function $M_{(u,v)} : (\Re, \Re) \rightarrow [0, 1]$ such that:

- **(a):** for each $y \in \Re$, the restriction $M_{(u,v)}(y, \cdot)$ is a probability measure in $(\Re, \Re)$;
- **(b):** for each $A \in \Re$, the restriction $M_{(u,v)}(\cdot, A)$ is a measurable function in $\Re$;
- **(c):** for each$A \in \Re$, it is valid that $u(A) = \int M_{(u,v)}(y, A)\, dv(y)$.

Conditions (a) and (b) assure that stochastic kernel is a well defined mapping for $M_{(u,v)}$ and $(\Re, \Re)$ probability spaces. The principal concept is in (c) condition. Given as initial period t, for a given income y there is a fraction $dv(y)$ of economies with income close to y. In period $t+n$, part of economies contained in $dv(y)$ will move to a subset $A \subseteq R$. Normalizating this fraction of economies by the total number of economies, we have the stochastic kernel given by $M_{(u,v)}(y, A)$. The integral $\int M_{(u,v)}(y, A)\, dv(y)$ represents the total of economies which independently of initial income will be in the subset A of economies in $t+n$ period. The integral $M_{(u,v)}(y, A)$ represents the total of economies which migrates from y to A, and $dv(y)$ is the weighting associated to each $M(\cdot)$

First author - Ibmec São Paulo and Statistics Dept.-IMECC-UNICAMP. Second Author - Ibmec São Paulo. First author thanks Lizia Figueiredo and participants of UFMG Graduate Economics Research Seminar for useful comments. The second author acknowledge partial financial support from CNPq and Pronex-Temático CNPq-Fapesp project 2003/10105-2.

given by the marginal distribution of y. Quah (1998) formalizes the additional necessary conditions. The stochastic kernel can be understood as the continuous form of a transition matrix, where we have a continuum of rows and columns, and thus a continuum of states.

The Stochastic Kernel estimation is realized using empirical measures for the elements of $\int M_{(u,v)}(y,A)\,dv(y)$ integral. The term $\int M_{(u,v)}(y,A)\,dv(y)$ can be estimated by the nonparametric estimation of joint density of relative incomes in periods t and $t+n$ using a bivariate kernel. This joint density is transformed in a Stochastic Kernel normalizating by the marginal distribution in period t, which is the empirical measure of $dv(y)$.

Arbia et al. (2005) points that a Stochastic Kernel also can be written as:

$$\phi_{t+\tau}(y) = \int_0^\infty f_\tau(y|x)\phi_t(x)dx \tag{2}$$

where y is the relative per capita income in period $t+\tau$ and x is the relative per capita income in period t. Now $f_\tau(y|x)$ is the conditional density which describes the probability of a region moves to a specific state of relative income, given the relative income in period t.

In order to construct an estimator for the Stochastic Kernel, methodologies for the estimation of conditional and marginal densities are necessary. A nonparametric estimator for the conditional density is proposed by Rosenblatt(1969), and further developed by Hyndman et al. (1996). This estimator is given by:

$$\widehat{f}_\tau(y|x) = \widehat{g}_\tau(x,y)/\widehat{h}_\tau(x) \tag{3}$$

where the estimator for the joint density $\widehat{g}_\tau(x,y)$ is given by:

$$\widehat{g}_\tau(x,y) = \frac{1}{nab}\sum_{i=1}^{n} K\left(\frac{\|x-X_i\|_x}{a}\right)\left(\frac{\|y-Y_i\|_y}{b}\right) \tag{4}$$

and the estimator for the marginal density $\widehat{h}_\tau(x)$ is:

$$\widehat{h}_\tau(x) = \frac{1}{na}\sum_{i=1}^{n} K\left(\frac{\|x-X_i\|_x}{a}\right) \tag{5}$$

In these estimations a and b are bandwidth parameters controlling the smoothness of fit, K is choosen kernel function and $\|x-X_i\|_x$ and $\|y-Y_i\|_y$ are usual Euclidian metrics.

The conditional density estimator can be rewritten as:

$$\widehat{f}_\tau(y|x) = \frac{1}{b}\sum_{i=1}^{n} w_i(x)K\left(\frac{\|y-Y_i\|_y}{b}\right) \tag{6}$$

where

$$w_i(x) = K\left(\frac{\|x-X_i\|_x}{a}\right)/\sum_{i=1}^{n} K\left(\frac{\|x-X_i\|_x}{a}\right) \tag{7}$$

This estimator is the Nadaraya-Watson kernel regression estimator. It shows that a conditional density can be obtained by the sum of n kernel functions in Y space weighted by the $w_i(x)$ in X space. The estimation of conditional mean by kernel regression method of Nadaraya-Watson is given by:

$$\widehat{m}_\tau(x) = \int y\widehat{f}_\tau(y|x)dy = \sum_{i=1}^{n} w_i(x)Y_i \tag{8}$$

Hyndmann et al (1996) proposed some changes in this method to correct the bias existing in this estimator, which is exacerbated when the conditional mean function has an exacerbate curvature or the points utilized in estimation are not regularly spaced. The estimator corrected by the bias proposed by Hyndmann et al (1996) is given by:

$$\widehat{f}_{\tau}^{*}(y|x) = \frac{1}{b}\sum_{i=1}^{n} w_i(x) K\left(\frac{\|y - Y_i^{*}(x)\|_y}{b}\right) \tag{9}$$

where $Y_i^{*}(x) = e_i + \widehat{r}(x) - \widehat{l}(x)$, $\widehat{r}(x)$ is the estimator of the conditional means $r(x) = E(Y|X = x)$, $e_i = y_i - \widehat{r}(x)$ e $\widehat{l}(x)$ is the mean of the conditional density estimates from $E(\cdot|X = x)$.

Note that the method of Hyndman (1996) and Hyndman et al. (1996) allow the introduction of more complex dependence structures in comparison to the methodology of Quah (1998), which allows for a unique conditioning factor and the dependence structure is imposed in the model and not estimated[1].

In addition to this estimator with reduces bias, Hyndman (1996) and Hyndman et al. (1996) proposed two new ways to visualize the conditional density. The first is by use of Stacked Plots in substitution to direct visualization of the conditional density. In this graphics the conditional density is showed for a grid of values of conditional variable, becoming easier to interpret the results of conditioning. The another way, introduced by Hyndman et al. (1996), is the use of High Density Region Plots. A region of high density is the smaller region in sample space containing a given probability. The advantage of this method is to make the visualization of multimodal densities more clear. The high density region presents multimodal densities as disjoints subsets in plane.

We made a application of this methodology to analyze conditional income convergence. Using the database about Brazilian municipalities (Atlas oh Human Development in Brazil) constructed by IPEA (Instituto de Pesquisas Econômicas Aplicadas) and João Pinheiro Foundation, we present an analysis of per capita income convergence for the years between 1970 and 1991.

3. Empirical Analysis

The data utilized in this study are a subset of Atlas of Human Development in Brazil, a comprehensive source of data on income, education and welfare compiled by United Nations Development Program, IPEA (Institute for Applied Economic Research), João Pinheiro Foundation and IBGE (Brazilian Institute of Geografy and Statistics). We use data on income municipality in 1971 and 1991, measured in terms of minimum wages proportion, and a measure of basic education given by the average number of years of schooling for individuals with less than 25 years in each municipality in 1970. Figure 1 shows the histogram of this variables.

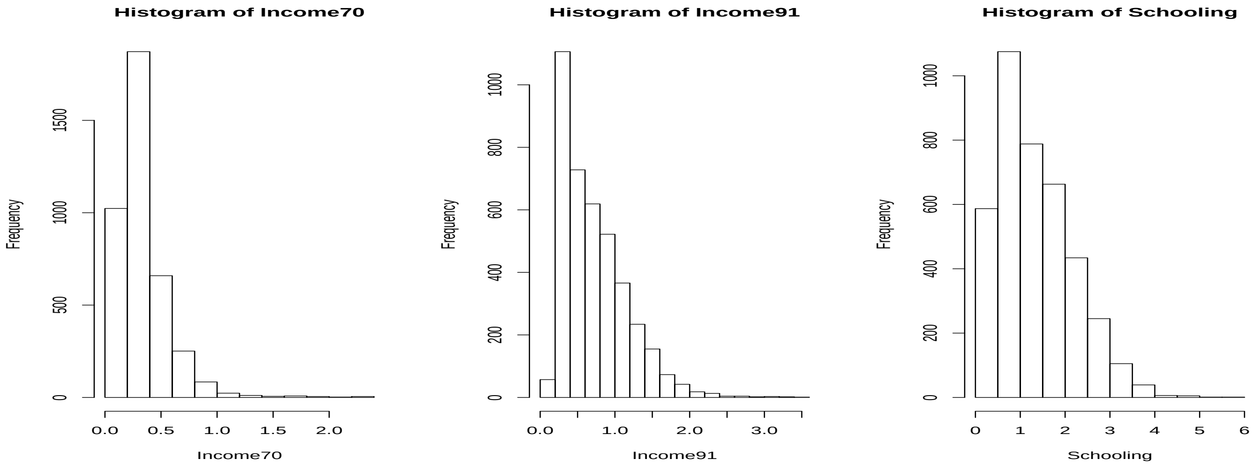
To illustrate the methodology, Figure 3 shows the non-parametric estimation of Stochastic Kernel for income, showing the evolution of average per capita family income of Brazilian Municipalities in 1991 conditioned in the existent income in 1970. The figure shows the Stacked Conditional Density Plot and the High Density Region Plot calculated using Hyndman (1996) e Hyndman et al. (1996) methods.

The Figure 3 shows that some stylized facts about convergence are captured in this figure: the estimated values shows that existence of a relative high mobility for the intermediary income categories. Note that conditional density estimated for intermediary income shows a high dispersion , showing the existence of high mobility to higher and lower categories of income. The figures shows that poor regions tend to remain poor and the rich regions tend to remain rich, given the higher probability for staying in the same region of relative income. These two facts correspond to the phenomena called "middle income disappearance" and the formation of Convergence Clubs, as interpretated as Quah (1996).

To illustrate the effect of conditioning, we estimate the income density in 1991 conditioned in Schooling in 1970 (Figure 2). The Figure shows that regions in the extremes of relative income have higher probability to be generated by the respective extremes of schooling. Relative richer regions in 1991 have high probability to be generated by high schooling and poor regions by low schooling. Intermediary income regions can be generated by high and low schooling, given the

[1]See Quah(1998) for details in construction of conditioning scheme.

FIGURE 1. Histogram - Income 1970, 1991 and Schooling in 1970.



high dispersion of probabilities of relative income, and in this way we can obtain the convergence club formation, but conditioned by the education variable.

To show the effect of multiple conditioning variables, we join the analysis of Figures 2 and 3, formulating a conditional stochastic kernel of income in 1991 conditioned in income in 1970 and schooling in 1991, represented by the Figure 4. To graphically represent a multiple conditioning scheme, we fix the first variable and shows the stochastic kernel for the other variable. Figure 4 shows conditional stochastic kernel fixing the schooling average in 1,2, 3 and 4 years. This Figure shows that controlling for schooling and initial income, municipalities with low schooling and initial income has high probability to stay in the region of relative income below 1; municipalities with high schooling has probability to stay in regions with a higher income.

Note that conditioning on multiple variables, the multimodality presents in figures 2 and 3 disappears. This effect shows that the presence of multimodality in unconditional analyses can be interpreted as the result of omission relevant conditional factors. Controlling for the effect schooling effect, we have that the conditional distribution of income in function of the initial income is essentially unimodal.

4. Conclusions

We show that the estimation of conditional Stochastic Kernel using in this methodology surpasses some existing limitations in the original estimator considered by Quah (1998). This method allows the estimation using multiple conditioning factors and non-parametric estimators for the conditioning process. It also allows a more comprehensive analysis of the economic process of conditional convergence then non-parametric methods currently used and also generalize the methods considered for Quah (1998) and Arbia et al. (2005).

5. References

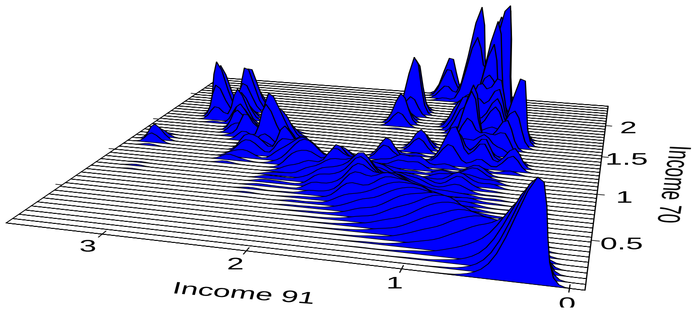
Arbia, G., Basile, R. e Piras, G. (2005). Analyzing Intra-Distribution Dynamics: A Reappraisal. REAL Discussion Papers 05-T-11, University of Illinois.

Heston,A, Summers, R. e Bettina, A. (2002) Penn World Table Version 6.1, Center for International Comparisons at the University of Pennsylvania (CICUP), October 2002.

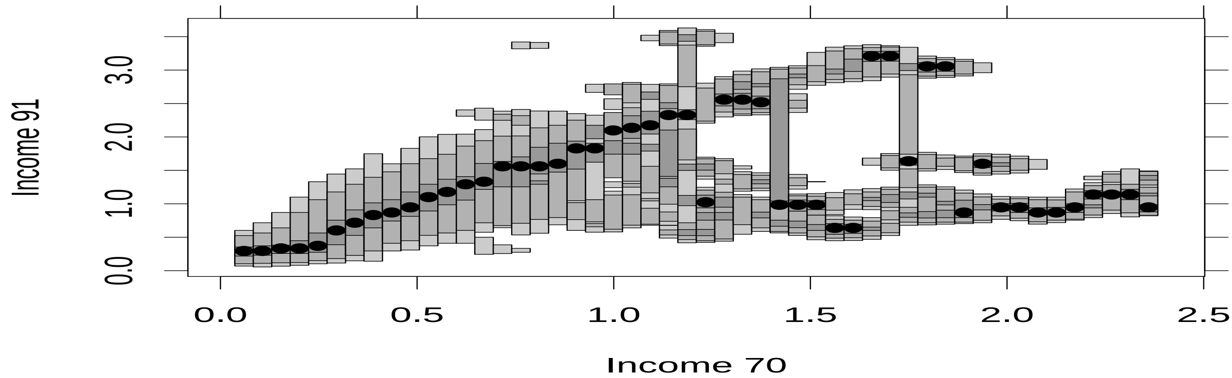
Hyndman, R.J. (1996). Computing and Graphing Highest Density Regions. The American Statistician 50:2, pp. 120-126.

Hyndman, R.J., Bashtannyk, D.M e Grunwald, G.K. (1996). Estimating and Visualizating Conditional Densities. Journal of Computacional and Graphical Statistics 5, pp. 315-336.

Figure 2. Stochastic Kernel - Income



(a) Stochastic Kernel



(b) Stochastic Kernel

Quah, D. (1996). Twin Peaks: Growth and Convergence in Models of Distribution Dynamics. The Economic Journal 106, pp 1045-1055.

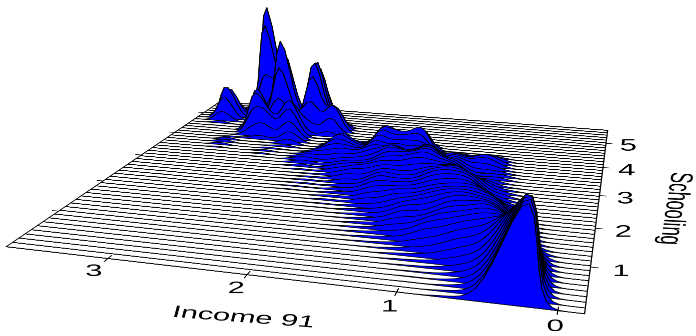
Quah, Danny. (1996), Regional Convergence Clusters across Europe, European Economic Review, 40, pp 951-958.

Quah. D. (1998). TsRf Reference Manual.

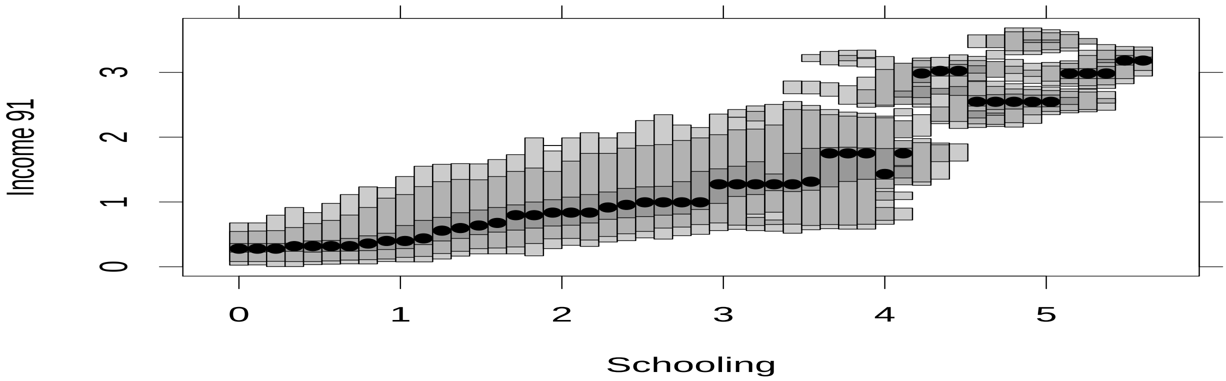
Shorrocks. A.F. (1978). The Measurement of Mobility. Econometrica 46:5, pp 1013-102

Current address: Ibmec São Paulo, Rua Quatá 300,Vila Olimpia,04546-042,São Paulo, SP, Brazil.

FIGURE 3. Conditional Stochastic Kernel - Income Conditioned on Schooling



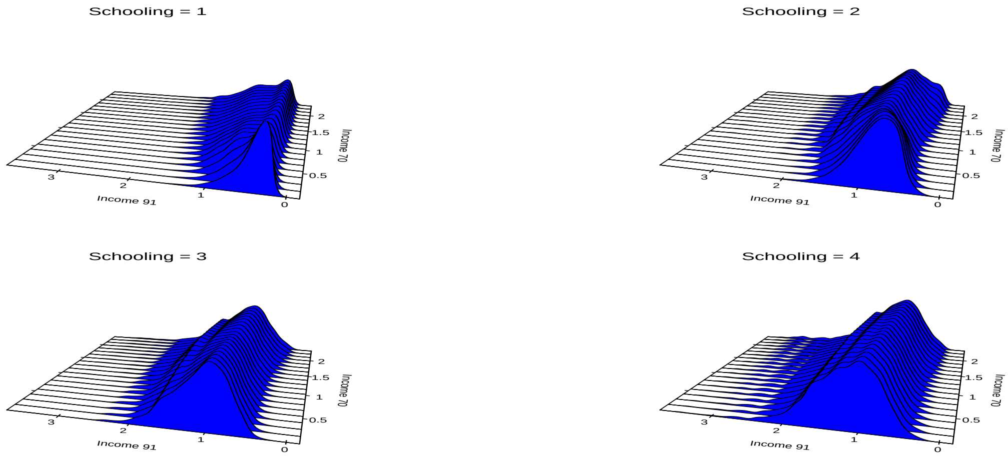
(a)



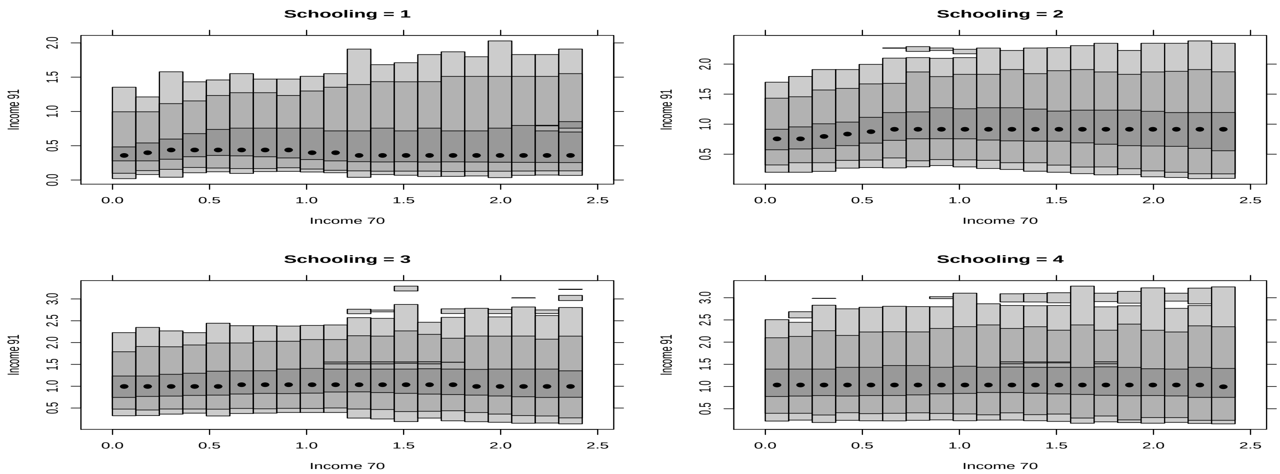
(b)

more easy

FIGURE 4. Conditional Stochastic Kernel - Income in 1991 Conditioned on Income and Schooling in 1970.



(a)



(b)