INTERNATIONAL ECONOMIC REVIEW
Vol. 45, No. 1, February 2004

TESTING FOR CONVERGENCE CLUBS IN INCOME PER CAPITA: A PREDICTIVE DENSITY APPROACH*

BY FABIO CANOVA[1]

Universitat Pompeu Fabra and CEPR

The article proposes a technique, based on the predictive density of the data, conditional on the parameters of the model, to jointly tests for groups of unknown size in a panel and to estimate the parameters of each group. The procedure is applied to the problem of identifying convergence clubs in scaled income per capita data. The steady-state distribution of European regional data clusters around four poles of attraction with different economic features. The distribution of income per capita of OECD countries has two poles of attraction and each group clearly identifiable economic characteristics.

We share the uncommonness of being different.

J. P. Roche

1. INTRODUCTION

Recent theories of growth and development have suggested that the distribution of income per capita of countries and/or regions may display convergence clubs, i.e., a tendency for the steady-states distribution to cluster around a small number of poles of attraction (see e.g., Ben David, 1994; Quah, 1996a; Galor, 1996). This tendency may be induced by several factors: the existence of some threshold level in the endowment of strategic factors of production; nonconvexities or increasing returns; similarities in preferences and technologies; and government policies, which become more similar over time within certain groups (e.g., EU or East Asian countries). Although there is anecdotal evidence supporting the view that clustering is an important feature of world income, to the best of my knowledge, only Durlauf and Johnson (1995), Paap and Van Djik (1998), and Desdoigts (1998) have attempted to formally document whether this tendency exists in the data.

This article proposes a new technique to formally examine whether the distribution of income per capita displays convergence clubs. The approach is general, determines the number of groups and the location of the break points when the

* Manuscript received October 2000; revised December 2002.

[1] I would like to thank three anonymous referees, Bruce Hansen, Hashem Pesaran, Russell Cooper, Christopher Croux, Albert Marcet, Evi Pappa, and the participants at seminars at Universitat Pompeu Fabra, the University of Southampton, Universite de Paris I-MAD, IGIER, University of Jerusalem, HHWA Hamburg, the NBER-NSF Summer workshop on "Empirical Methods in Macroeconomics" and the CEPR workshop "New Approaches to the Study of Business Cycles" for comments and suggestions. A previous version of the article has circulated with the title "Testing for heterogeneity in the cross-sectional dimension of a panel: A predictive density approach." Please address correspondence to: Fabio Canova, University of Pompeu Fabra. E-mail: fabio.canova@uni-bocconi.it.

appropriate ordering of the units in the cross section is unknown, and, at the same time, allows one to estimate the parameters of each group in a unified manner. The approach is based on the predictive density (marginal likelihood) of the data and has appealing features for both Bayesian and classical analysts.

The technique can be viewed as a natural extension of the standard approach used to determine the number of heterogeneous groups in a cross section (see e.g., the Goldfeld and Quandt test) when the number of groups, the location of the breaks, and the ordering of units are unknown. However, instead of assuming that the regression coefficients are the same for all units belonging to one group, I allow for a further layer of heterogeneity within groups. This second layer of heterogeneity takes the form of a prior that restricts the coefficients of the units in a group to have the same distribution, but allows the distribution of the coefficients of units in different groups to differ. Such a restriction implies that the distribution of steady states may display multiple basins of attraction. Since a similar restriction on the coefficients of the *entire* cross section implies that the distribution of steady states only has one attractor, testing for convergence clubs is equivalent to checking which of these two assumptions is more appropriate.

Once the optimal ordering, the number of groups, and the location of the break points in the cross section have been established, I provide a simple way to estimate the parameters of each group and to conduct inference. The approach I employ lies within the Empirical Bayes (EB) tradition: I use predictive densities to estimate the parameters and posterior analysis to draw conclusions about functions of the coefficients of the model. Posterior inference is appealing because it gives us a compact way to summarize both subjective and objective uncertainty about economically interesting functions of the coefficients of the model (convergence rates, long run multipliers, steady-states distribution, etc.). EB methods are simpler than standard hierarchical Bayesian approaches, since they do not require numerical integration, and advantageous when there is interest in estimating aspects of prior distribution, which is precisely the case considered here.

The methodological contribution of this article is linked to a number of articles, both in the classical and the Bayesian tradition, testing for the existence of a unknown break point in time series (see e.g., Ploberger et al., 1989; Bai, 1997; Polasek and Ren, 1997) and to the EB tradition of constructing posterior estimates of the coefficients of a model by plugging in estimates of the parameters of the prior (see Morris, 1983; Berger, 1985; Efron, 1996). The idea behind the grouping approach is related to Forni and Reichlin (1997), who attempt to estimate a reduced number of common latent factors from large dynamic cross-sectional data, and to Hansen (1999, 2000), who examines inferential problems in threshold models for cross sections, time series, or static panels. However, while the latter author characterizes the asymptotic distribution of the threshold parameter and studies hypotheses testing, given an ordering of the cross section when there is only one threshold, our interest is in designing a technique to find multiple thresholds when the correct ordering is unknown. Contrary to Durlauf and Johnson (1995), we allow for heterogeneity within groups (so that the steady-state distribution of a club need not be degenerate) at the costs of imposing restrictions on the time series properties of the data. Our approach is more formal than Quah's (1996a, 1996b),

but it requires more stringent assumptions on the structure of the dynamic model than his. Finally, the grouping procedure we employ has roots in the Bayesian literature on mixture densities (see Titterington et al., 1985 or Paap and Van Djik, 1998) and shares similarities with classification/cluster analyses (see e.g., Mardia et al., 1980). Four features distinguish the proposed approach from existing ones: the use of serially correlated data, the possibility that groups have different covariance matrices, the lack of knowledge about number of break points, and the criteria used to assign units to groups (predictive ability vs. within group variance).

I employ European regional income per capita data from the NUTS2 data set of Eurostat and OECD national income per capita from the Summer and Heston data set to determine whether the income distribution shows any tendency toward club convergence. Recent theories of economic growth have suggested a number of indicators that may determine the club a unit will join: for example, the initial conditions of income per capita and of the average human capital, the dispersion of the distribution of income and education within units, and the geographical location of a unit may be crucial to determine the pole of attraction around which it will gravitate in the long run. Human capital and policy variables are not available at the regional level. Therefore a search for clubs in these data is conducted using initial conditions and geographical and threshold externalities measures as grouping devices. At the country level, indicators for access to technologies, government policies, human capital, geography, and threshold externalities are available and all of them are used to search for clubs.

I find that the ordering based on the ranking of scaled income per capita in the presample period is the one that maximizes the predictive power of the model for both data sets. With this ordering, there is a natural clustering of units in four groups of regional income per capita and two groups of national income per capita. No further break is detected when other variables are used to reorder units within these groups. In both cases clubs are characterized by different parameters controlling the speed of adjustment to the steady state and the mean level of per capita steady-state income relative to the average. More precisely, poor units converge faster to their steady state than rich ones and they tend to cluster around a pole of attraction that is substantially below the average (see also Quah, 1996b). The dispersion of steady states around each basin of attraction is significant, suggesting that clustering is more prevalent than convergence even within groups. I show that even though groups have different long-run mobility indices, there is substantial immobility in the ranking of units within groups, confirming the strong persistence in inequality found by Canova and Marcet (1995). As a consequence of the persistence of the initial income characteristics and of the immobility in ranking, the steady-state distribution of income per capita will become polarized. Since poor units are also those featuring low initial average human capital, distributions of income and education are more polarized and are geographically located in the "South" or in the periphery of the industrialized world, the results provide a bleak picture over the possibility of equalizing income per capita both in EU and in OECD countries in the near future.

The rest of the article is organized as follows. The next section describes the details of the testing approach. Section 3 provides a technique to estimate the

parameters and to conduct posterior inference on functions of the coefficients of the dynamic model. Section 4 provides the link between economic theory and the proposed econometric procedure, emphasizing measurable factors that may determine club convergence. Section 5 examines European regional and OECD national income per capita data. Section 6 concludes. Appendix A provides some Monte Carlo evidence on the properties of the tests.

2. THE GROUPING PROCEDURE

The starting point of the analysis is the *a priori* belief that there may be significant heterogeneities in the cross section of a panel and a natural clustering of units around certain poles of attraction, in the sense that the coefficients of the statistical model are more similar within each group than across groups. For example, if units i and j belong to a group, the vector of coefficients of the model for the two units may have the same mean and the same dispersion. However, if units i and j do not belong to the same group, the vector of coefficients of the two units may have different means and different dispersions.

For the sake of generality, I assume that the ordering of cross-sectional units, which naturally gives rise to clustering, is unknown. In practice, clustering in income per capita may be linked to geographical, economic, or sociopolitical factors and modern growth theory provides a restricted set of ordering, which is worth examining. Let N be the size of the cross section, T the size of the time series, and $m = 1, 2, \ldots N!$ the particular ordering of the units of the cross section. It is assumed that there may be $q = 1, 2, \ldots Q$ break points in the cross section, Q being given. Each of the resulting $q + 1$ groups is characterized by a statistical model of the form

$$Y_{it} = \alpha_i + \rho_i(\ell)Y_{it-1} + \theta_i(\ell)W_{t-1} + u_{it} \tag{1}$$

$$\beta_i = \beta^p + \epsilon_i^p \tag{2}$$

where $i = 1, \ldots, n^p(m)$; $p = 1, \ldots, q + 1$, $u_{it} \sim (0, \sigma_{u_i}^2)$, $\epsilon_i^p \sim (0, \Sigma_p)$, $\rho_i(\ell)$ and $\theta_i(\ell)$, are polynomials in the lag operator of order r and d, $\beta_i = [\alpha_i, \rho_{i1}, \ldots, \rho_{ir}, \theta_{i1}, \ldots, \theta_{id}]$ is the vector of coefficients of unit i, $n^p(m)$ is the number of units in group p, given the mth ordering of the cross section, $\sum_p n^p(m) = N$, each m. I assume that Y_{it} is a vector of dimension s for each unit i, while W_{t-1} is a vector of exogenous variables of dimension v affecting all units of the cross section with a period lag. In (2), the vector of coefficients for each i is random but the coefficients of the $n^p(m)$ units belonging to group p have the same mean and same covariance matrix. Furthermore, I assume that the underlying structure may differ across groups: The coefficients of units belonging to different groups may be drawn from distributions with different parameters. Equations (1)–(2) therefore capture the idea that there may be clustering of units within groups but that groups may drift apart, implying heterogeneous dynamics in the cross section. For the rest of the article I refer to β^p and Σ_p as the hyperparameters of the model.

Model (1)–(2) is sufficiently general to include several models studied in the panel data literature as special cases. For example, a standard switching regression model is obtained by setting $\epsilon_i^p = 0$, $\forall i$. A fixed effect model is obtained by restricting $\rho_i = \rho$ and $\epsilon_i^p = 0$, $\forall i$ whereas a random effect model is obtained by setting $\rho_i = \rho$ and $\epsilon_i^p = [\epsilon_{i1}^p, 0, \ldots, 0]$. Also, for future reference, I take the alternative to (1)–(2) to be a model with homogeneous dynamics in the cross section. In this case $Q = 1$ and I replace Equation (2) with

$$\beta_i = \beta + \epsilon_i \qquad i = 1, \ldots N \tag{3}$$

where $\epsilon_i \sim N(0, \Sigma)$. In other words, in the alternative β and Σ are the same for all i, so that there is an exchangeable structure for all units of the cross section. The limiting case of this alternative is a pooled model, which can be obtained by setting $\epsilon_i = 0$, $\forall i$.

Within this general setup, I study two issues. First, I am interested in providing a framework for verifying the hypothesis that there are heterogeneities in the cross section in a situation where the number of groups, the location of the breaks, and the permutation, which naturally give rise to the clustering, are unknown. Once I have established the "submodel" of interest, i.e., the number of groups, the location of the breaks, and the ordering of the cross section, I will be concerned, at a second stage, with the problem of estimating the hyperparameters and $\sigma_{u_i}^2$ for each i, which are unknown to the investigator and needed to construct posterior estimates of important functions of the β_i.

Let Y be a $(N * T * s) \times 1$ vector of the LHS variables in (1) ordered to have the N cross sections for each $t = 1, \ldots T$, s times, X be a $(N * T * s) \times (N * k)$ matrix of the regressors, $k = s * r + v * d + 1$, β be a $(N * k) \times 1$ vector of coefficients, U be a $(N * T * s) \times 1$ vector of disturbances, β_0 be a $(q + 1) * k \times 1$ vector of means of β, A be a $(N * k) \times (q + 1) * k$ matrix, $A = \text{diag}\{A_p\}$, where A_p has the form $\iota \otimes I_k$ where I_k is a $k \times k$ identity matrix and ι is a $n^p(m) \times 1$ vector of ones. Given an ordering m, the number of groups q, and the location of the break point $h^p(m)$ we can rewrite (1)–(2) as

$$Y = X\beta + U \qquad U \sim (0, \Sigma_u) \tag{4}$$

$$\beta = A\beta_0 + E \qquad E \sim (0, \Sigma_E) \tag{5}$$

where the dimension of Σ_u is $(N * T * s) \times (N * T * s)$ and $\Sigma_E = \text{diag}\{\Sigma_p\}$ is a $(N * k) \times (N * k)$ matrix. Using (5) into (4) we arrive at

$$Y = \tilde{X}\beta_0 + W \qquad W \sim (0, \Sigma_W) \tag{6}$$

where $\tilde{X} = X * A$ and $W = XE + U$. In (5) I have expressed the dependent variable Y as a linear combination of the Xs and of the hyperparameters β_0 with errors that have a heteroskedastic structure. Since $q + 1 \ll N$, this operation has effectively reduced the dimensionality of the model.

To complete the specification in a Bayesian sense one must provide prior distributions for (β_0, Σ_E), for the covariance matrix Σ_u, and for the underlying

submodel characteristics $\mathcal{M}$, which are indexed by $(m, q, h^p(m))$. Structured in this way, our specification has a standard hierarchical structure and inference about any function of β can be conducted, in line with the Bayes theorem, by averaging over models and hyperparameters. Similarly, the selection of a submodel $\mathcal{M}_r$ requires averaging the joint density of the model over β and the hyperparameters. Roughly speaking, if we let ω denote the vector containing, stacked, the elements of the hyperpameters and of the covariance matrix, submodel section requires the construction of $L(Y \mid \mathcal{M}_r) = \int L(Y \mid \mathcal{M}_r, \beta) p(\beta \mid \omega) p(\omega)\, d\beta\, d\omega$ where $L(Y \mid \mathcal{M}_r, \beta)$ is the conditional likelihood of the data. Hence, a prior for the hyperparameters and for the submodels must be provided and numerical integration techniques are needed to compute these averages. Although these two steps are feasible in certain setups (see e.g., Canova and Ciccarelli, 1999), in general, the calculation of $L(Y \mid \mathcal{M}_r)$ is very demanding. The approach we present takes a short cut to these complications and proceeds using point estimates of those nuisance parameters that are integrated out in the fully hierarchical Bayesian procedure.

To intuitively understand what the approach involves, assume that β_0, Σ_E, Σ_u are known. Then, our approach to group units proceeds in three steps. First, given an ordering of the units of the cross section, I examine how many groups there are using the sequential testing approach described below. Second, given an ordering of units and the number of groups, I attempt to find the location of the break points by maximizing the predictive density (marginal likelihood) of the model with respect to the location of the breaks. Third, I iterate on the first two steps, altering the ordering of units in the cross section. The selected submodel is the one that maximizes the predictive density of the data over orderings, groups, and break points.

Formally speaking, let $L(Y \mid H_0)$ be the predictive density of the data under the assumption that the hyperparameters are the same in each group, i.e., $\beta_0 = \iota_1 \otimes \gamma_0$ where ι_1 is a $(q+1) \times 1$ vector of ones and γ_0 a $k \times 1$ vector, and $\Sigma_p = \Sigma, \forall p$. Furthermore, let I^q be the set of possible break points when there are q groups and J be the set of possible orderings of the cross section. Let $L(Y^p \mid H_q, h^p(m), m)$ be the predictive density for group p, under the assumption that there are q break points with location $h^p(m)$, using ordering m and let $L(Y \mid H_q, h^p(m), m) = \prod_{p=1}^{q+1} L(Y^p \mid H_q, h^p(m), m)$ be the total predictive density for the sample under the assumption that there are q break points. These predictive densities can be easily obtained from (5) once distributional assumptions for the error term are made. Define the following quantities:

- $L^+(Y \mid H_q, m) \equiv \sup_{i \in I^q} L(Y \mid H_q, i, m)$,
- $L^\dagger(Y \mid H_q) \equiv \sup_{j \in J} L^+(Y \mid H_q, j)$,
- $L^{Aq}(Y \mid H_q, m) \equiv \sum_{i \in I^q} \pi_i^p(m) L(Y \mid H_q, i, m)$,

where $\pi_i^p(m)$ is the prior probability that, for group p of ordering m, there is a break at location i. The first expression gives the maximized value of the predictive density with respect to the location of break points for each q and m; the second, the maximized value of the predictive density, for each q, once the location of the break point and the ordering of the data are chosen optimally. The last expression

gives the average predictive density under the assumption that there are q breaks: The average is calculated over all possible locations of the break points, using the prior probability that there is a break point in each location as weight. In general, ignorance about the location of the break points leads us to assume that $\pi_i^p(m)$ is uniform over each p, m.

To examine the hypothesis that the dynamics of the cross section are group-based we will use a posterior odds (PO) ratio.[2] I consider first the null that there are no break points against the alternative that there are at most Q breaks and then, if the alternative is more likely, sequentially verify a series of hypotheses where the null is that there are $q-1$ break points and the alternative that there are q break points, $q = 1, \ldots Q$. Given m, the statistics to verify the first hypothesis are

$$\text{PO}(m) = \frac{\pi_0 L(Y \mid H_0)}{\sum_{q=1}^{Q} \pi_q L^{Aq}(Y \mid H_q, m)\mathcal{P}_1(N)} \tag{7}$$

where π_0 is the prior probability that there are no breaks and π_q is the prior probability that there are q breaks and $\mathcal{P}_1(N)$ a penalty function that accounts for the fact that a model with Q breaks is more densely parametrized than a model with no breaks. H_0 is preferred to H_1 when $\text{PO}(m) \gg 1$. The statistics for the hypotheses that there are $q-1$ versus q breaks in the cross section are

$$\text{PO}(m, q-1) = \frac{\pi_{q-1} L^{A(q-1)}(Y \mid H_{q-1}, m)}{\pi_q L^{A(q)}(Y \mid H_q, m)\mathcal{P}_2(N)} \tag{8}$$

where $\mathcal{P}_2(N)$ is a penalty function. Similarly, $q-1$ breaks are preferred when $\text{PO}(m, q-1) \gg 1$. We can also test the null hypothesis that there are q break points at particular locations against the alternative that there is a further break point at a particular location i using a posterior odds ratio of the form

$$\text{PO}(m, q-1^*) = \frac{\pi_{q-1} L^{+}(Y \mid H_{q-1}, m)}{\pi_q \pi_i^p L^{+}(Y \mid H_q, m)\mathcal{P}_3(N)} \tag{9}$$

When $\pi_q = \pi_{q-1} = 0.5$; $\mathcal{P}_3(N) = 1$, (9) is the PIC criterion of Phillips and Ploberger (1994).

To put the testing problem in an alternative perspective, one can ask what is the prior probability on the null hypothesis one must entertain so that his/her beliefs will not be overturned by the data. For example, it may be of interest to know how much confidence one should have on the hypothesis that the sample is distributionally homogeneous so that an overall exchangeable prior is sufficient to characterize the data. This prior probability, which I call $\hat{\pi}$, can be found for any of the hypotheses considered by setting PO in (7)–(9) equal to 1 and solving for $\hat{\pi}_0, \hat{\pi}_{q-1}, \hat{\pi}_{q-1}$, respectively.

[2] As an alternative one could use a Wilks likelihood ratio (WL) criterion (see e.g., Efron, 1996) or the modified likelihood ratio (ML) of Hansen (2000).

The testing procedure I have described leaves the value of Q unspecified. Following Hartigan's (1975) rule of thumb, I set $Q \ll \sqrt{(N/2)}$.

To find the location of the break point, given that there are q breaks, I assign units to groups so as to provide the highest total predictive density, i.e., I compute $L^{+}(Y \mid H_q, m)$. Since there are m possible permutations of the cross section over which to search for clustering, I take the optimal permutation rule of units in the cross section to be the one that achieves $L^{\dagger}(Y \mid H_q)$.

Bai (1997) shows that proceeding sequentially in testing for breaks, i.e., test first for one break against no breaks; then conditional on the results of the first test, test for the existence of one break in each of the two subsamples and so on, produces consistent estimates of the number and the location of the breaks. However, when there are multiple groups and one tests for the presence of two groups only, the estimated break point is consistent for *any* of the existing break points and its location depends on which of the breaks is "stronger." If this is the case Bai suggests refining of the estimate of the break points. That is, if two breaks are identified at i_1 and i_2, it is convenient to reestimate i_1 over $[1, i_2]$ and i_2 over $[i_1, N]$. Each refined estimator of the location of the break has then the same properties as the estimator obtained in the case the sample has a single point.

The major stumbling block to the application of the procedure I have described is the dimensionality of maximization problem. When no information is available on the ordering of the units in the cross section and N is moderately large, the maximization problem may constitute a formidable task. However, this is not a binding constraint for many applications since economic theory guides the search for orderings and this considerably reduces the computational complexity of the problem. Note also that even in the case economic theory is silent and one engages in an unstructured search, the maximization problem requires a considerably smaller number of evaluation than $N!$, since many orderings are equivalent from the point of view of the predictive density. That is, once a particular grouping is found, searching for groups can be shrewdly conducted by reassigning units across groups around this local maxima[3].

3. PARAMETER ESTIMATION AND INFERENCE

Once the submodel characteristics have been determined, I will be interested in estimating the unknown matrix Σ_u and the hyperparameter vector. Let $\omega = [\beta_0', \text{vec}(\Sigma_E)', \text{vec}(\Sigma_u)']$ be the vector of the parameters of the model. The predictive

[3] To clarify the issue, suppose $N = 4$ so that we have a total of 24 possible orderings to examine. Suppose the initial ordering is 1234 and two groups are found: 1 and 234. Then all permutations of 234 with unit 1 coming ahead, i.e., 1243, 1342, etc., give the same predictive density (see Appendix A for a confirmation of this result in a Monte Carlo context). Similarly, permutations that leave unit 1 last need not to be examined, i.e., 2341, 2431, etc. This reduces the number of orderings to be examined to 13. By trying another ordering, say 4213, and finding, for example, two groups: 42 and 13, we can further eliminate all the orderings that consist of permutations of the elements of each group, i.e., 4132, 2341, etc.. It is easy to verify that once four carefully selected orderings have been tried and, say, two groups found in each trial, we have exhausted all possible combinations, as far as the predictive density is concerned. The example is rigged so that at each stage we find two groups. When this is not the case, the number of orderings to be examined is larger, but it does not exceed hN where h is the maximum number of breaks found with any of the permutation.

density $L^{\dagger}$ can now be used as a function of ω, for fixed Y, and maximized to construct the best possible fit of the model to the data. Maximizing $L^{\dagger}(\omega \mid Y, H_q)$ with respect to ω yields an EB estimator for ω (see Berger, 1985) which will be the starting point for obtaining posterior estimates of the coefficients of the dynamic model for each unit.

To a purist, the idea of estimating the parameters of the prior may sound, at least, strange. Information about the prior typically reflects subjective knowledge that an investigator has on the phenomenon under study. However, the prior may also reflect objective information, for example, information about the data itself. In this case, estimating features of the prior is a way to "tune it up" to the particular application. EB methods offer a simple way to obtain these estimates using predictive densities.

There are several ways to obtain estimates of ω under the assumption that the errors in (1)–(2) are normally distributed, or under the more general assumption that the errors come from a family of exponential distributions (see, e.g., Efron, 1996). For example, if the us are normally distributed, the elements of ω can be estimated as (see Maddala, 1991)

$$\hat{\beta}^p = \frac{1}{n^p(m)} \sum_{j=1}^{n^p(m)} \beta_{ols}^j \tag{10}$$

$$\hat{\Sigma}_p = \frac{1}{n^p(m)-1} \sum_{j=1}^{n^p(m)} \left(\beta_{ols}^j - \hat{\beta}^p\right)\left(\beta_{ols}^j - \hat{\beta}^p\right)' - \frac{1}{n^p(m)} \sum_{j=1}^{n^p(m)} (x_j x_j')^{-1} \hat{\sigma}_j^2 \tag{11}$$

$$\hat{\sigma}_j^2 = \frac{1}{T-k}\left(y_j' y_j - y_j' x_j \beta_{ols}^j\right) \tag{12}$$

where $p = 1, \ldots, q+1; i = 1, \ldots, N; x_j$ is the matrix of regressors, and y_j the vector of dependent variables for unit j, and β_{ols}^j is the OLS estimator of β^j obtained using only the information for unit j.

Given these estimates, one can construct Empirical Bayes (EB) posterior point estimates for the β vector by plugging in estimated values in standard formulas, i.e.,

$$\hat{\beta} = \left(X' \hat{\Sigma}_u^{-1} X + \hat{\Sigma}_E^{-1}\right)^{-1} \left(X' \hat{\Sigma}_u^{-1} Y + \hat{\Sigma}_E^{-1} A \hat{\beta}_0\right) \tag{13}$$

Alternatively, it is possible to show that, under normality of the us and the ϵs and after imposing a diffuse prior on the ω, it is possible to jointly estimate ω and the posterior mean of β as follows:

$$\hat{\beta}^p = \frac{1}{n^p(m)} \sum_{j=1}^{n^p(m)} \beta_j^* \tag{14}$$

$$\hat{\Sigma}_p = \frac{1}{n^p(m)-k-1} \left[R + \sum_{j=1}^{n^p(m)} \left(\beta_j^* - \hat{\beta}^p\right)\left(\beta_j^* - \hat{\beta}^p\right)' \right] \tag{15}$$

$$\hat{\sigma}_j^2 = \frac{1}{T+2}\left(y_j - x_j\beta_j^*\right)'\left(y_j - x_j\beta_j^*\right) \tag{16}$$

$$\beta_j^* = \left(\frac{1}{\hat{\sigma}_j^2}x_j'x_j + \hat{\Sigma}_p^{-1}\right)^{-1}\left(\frac{1}{\hat{\sigma}_j^2}x_j'x_j\beta_{ols}^j + \hat{\Sigma}_p^{-1}A_0\hat{\beta}^p\right) \tag{17}$$

where $p = 1, \ldots, q+1; i = 1, \ldots, N; j = 1, \ldots, n^p(m)$; and R is a diagonal matrix with small positive entries used here, as in ridgelike estimators, to insure that estimates of the dispersion matrix for each group are positive definite.

Note that, while the first approach only requires OLS estimates for each unit, so that posterior estimates can be computed in two steps, in (14)–(17) estimates of the prior parameters and of the posterior mean of β are obtained jointly using iterative methods.

The normal posterior generated with EB methods has a covariance matrix that underestimates the covariance matrix obtained from a fully hierarchical approach. This is because no allowance is made for the fact that the hyperparameters have been estimated and that the number of units in each group may be small. Morris (1983) and Carlin and Gelfand (1990) provide methods to correct for this problem. In many applications, among them the one presented here, researchers want to study functions of the posterior mean of β and are not necessarily interested in the spread of the posterior distribution of βs. In this case, the EB estimates given in (13) or (17) suffice.

I have run a Monte Carlo exercise to examine the ability of the procedure to detect breaks and of unbiasedly estimating the hyperparameters with simple DGPs. The results are presented in some detail in Appendix A. It turns out that, if the ordering is correctly specified, the predictive density approach I have suggested is able to correctly detect the number and the location of breaks when there are simple or multiple breaks in the data. However, the posterior odds ratio appears to be slightly biased when no heterogeneities are present and no penalty function is employed. This suggests that a conservative strategy to avoid the proliferation of groups is to choose the penalty function to make sure that the PO ratio is unbiased. When the ordering is unknown, maximization of the predictive density over permutations recovers the best ordering of units in the cross section, and once the ordering is found, the number and the location of the breaks is correctly identified. Estimates of the hyperparameters are biased when the size of the time series is small: Mean parameters are downward biased and variance parameters upward biased. When $T \geq 30$ most of these biases disappear.

4. LINKING THE ECONOMETRIC APPROACH TO GROWTH THEORY

Modern growth theory has suggested many mechanisms that may lead to convergence clubs. Galor (1996) provides a thoughtful and compact summary of the major implications of various theoretical models, stressing which economic indicators

may be helpful in detecting club convergence. To provide a link between growth theory and the approach described in the previous sections and, in particular, restrictions on the ordering of the units in the cross section and some guidelines to interpret the results, I next briefly summarize the causes of club convergence and the variables that may help to detect whether the theory has any bearing to the data.

Basic neoclassical growth models, with production functions exhibiting decreasing returns to scale to the capital–labor ratio, exogenous population growth, and fixed saving rate may generate convergence clubs in, at least, two circumstances: first, when saving rates out of wage and interest income differ, with the first being larger; and second, when the economy features heterogeneous agents. The first situation may be a consequence of heterogeneous factor endowments across individuals and of life-cycle considerations, whereas the second one, for example, is a standard feature of OG models. In both cases, multiplicity of stationary equilibria occurs and the distribution of initial income per capita determines the asymptotic club to which a particular unit will belong.

The incorporation of empirically important elements such as human capital or fertility in the basic neoclassical growth model, along with some type of market imperfections (externalities, imperfectly competitive markets, nonconvexities, and so on) produces additional channels that strengthen the possibility of club convergence. For example, social increasing returns with respect to human capital accumulation or capital market imperfections together with nonconvexities in the production of human capital will produce convergence clubs. In this case units that are similar in their characteristics and in their initial level of income may cluster around different steady-state equilibria because they have different endowments of human capital (see, e.g., Azariadis and Drazen, 1990). In some cases, it may be the within-unit distribution of human capital that determines the different steady state around which units will gravitate (see Galor and Zeira, 1993). The within-unit distribution of initial income may also be the reason why units converge to different clubs: Capital market imperfections together with some fixed cost in production may generate this outcome (see Quah, 1996b). A model with endogenous fertility, as in Barro and Becker (1989), can also generate the required outcome. In this case, the initial conditions with respect to the number of children and the level of human capital dictate the steady-state equilibria a unit will settle in. In other versions of such a model, it is the initial level of distribution of income that determines the distribution of the steady-state level of output per capita and fertility rates.

Quah (1996a) suggests that club convergence may be due to informational externalities that may occur at either the state or neighborhood level. That is, units that are either members of the same nation, share some borders, or belong to geographically homogenous areas may tend to cluster together because information flows more easily across units with these characteristics. Hence, the geographical location of a unit determines the convergence club it will join. This local externality hypothesis substantially differs from those that use increasing returns to scale in some factor of production and may generate converge clubs even under standard assumptions about preferences and technologies.

To summarize, the theoretical literature has suggested several mechanisms that may generate club convergence. At least four indicators (the initial level of income, the initial level of human capital, the initial distribution of income per capita, and human capital within the unit) could be used to proxy for the economic causes of these heterogeneities. In addition, geographical/location indicators can be used to try to measure the extent of neighborhood externalities. Finally, policy variables, such as the government expenditure-to-output ratio, could also be used to proxy for national effects.

5. ARE THERE CONVERGENCE CLUBS?

This section attempts to shed light on two issues. First, I would like to examine whether income per capita data is consistent with the multiple steady-state version of modern growth theory.[4] Second, I would like to better understand the statistical properties of income per capita data. In particular, I am interested in examining the kind of heterogeneities the data displays: whether the average adjustment properties to the steady state and the average steady state are group dependent; and whether different groups display difference persistence of inequalities, in the sense that the relative ranking in the initial distribution is more important in determining the relative ranking in the steady-state distribution for some groups than for others. I study these issues using two different data sets: European regional income per capita from the Eurostat database and OECD national income per capita from the Summer and Heston database.

5.1. *European Regional Income Per Capita.* The data set used covers 144 European NUTS2 units[5] and refers to the period 1980–1992. Given that $N = 144$ I allow, at most, six groups (i.e., $Q = 5$). Income per capita is scaled by the European average to reduce both serial correlation and the effect of outliers, and logarithms are taken. AIC and BIC selection procedures indicate that an AR(1) with unit-specific parameters captures sufficiently well the dynamics of the scaled data and makes the residuals well behaved. Hence, $r = 1$ and no W_{t-1} variables are used.

For regional data there are few usable indicators to order units according to the suggestions of recent growth theories. For example, no measures of the average regional human capital (or its distribution) at the beginning of the sample are available, nor do I have regional measures of dispersions of income per capita. Furthermore, neither EU nor national expenditure (either for regional income support or for regional infrastructure and capital formation) are available on a consistent basis for all countries. Since the sample covers the 1980s and the regions

[4] What I examine here is a somewhat strong version of the convergence club hypothesis. A weaker version would predict the existence of convergence clubs in the distribution of *growth rates* of income per capita (see, e.g., Boldrin and Canova, 2001).

[5] Roughly speaking, the NUTS2 classification corresponds to regions. NUTS1 refers to larger territorial units (the "North," the "Centre," and the "South") whereas NUTS3 provides data at the provincial level.

belong to the EU, I conjecture that differences along these dimensions are unlikely to provide information for grouping units into convergence clubs.[6]

Given these limitations, I search for clubs ordering the cross section according to: (i) the magnitude of per capita income relative to the European average in 1979, with poor regions coming first; (ii) the magnitude of per capita income relative to the national average in 1979, with poor regions coming first; (iii) the magnitude of locally scaled income per capita in 1979 (Mediterranean regions and Ireland are scaled by their average and other regions by their average), with poor regions coming first; (iv) the magnitude of the average share of per capita income relative to the European average in the sample, with poor regions coming first; and (v) the magnitude of the average growth rate of per capita income in the sample, with regions growing slower coming first.

The first ordering attempts to capture the effect that initial conditions may have on the steady-state distribution of income per capita; the next two orderings try to verify whether geographical externalities (either at the country or at the south–north level) may be important to determine the basin of attraction of a unit; the last two classifications attempt to study the importance of threshold externalities, here proxied by the size of the share of income per capita in Europe or its growth rate. If geographical externalities are important, any tendency toward convergence clubs that may appear with (i) should be weakened or disappear with (ii) or (iii). Note also that, because of immobility in the initial ranking, the results obtained for orderings (i), (ii), and (iii) are insensitive to the choice of 1979 or earlier years as presample date.

Ordering units according to the initial distribution of income per capita maximizes the predictive density of the data. Given this ordering, I identify three breaks, corresponding to units 15, 23, and 120, and, consequently, four groups in the data. Within the first group there are 10 regions of Greece, 4 of Portugal, and 1 of Spain; in the second group there are 4 regions of Greece, 3 of Spain, and 1 of Italy; finally, the last group includes regions from 9 different countries but the majority are German (9) and Northern Italian (5). The exact composition of each group is given in Appendix B. The fourth and fifth orderings produce four and three groups, whose composition is very similar to these groups. Hence, the splitting produced is highly suggestive of the fact that European regions cluster into homogeneous groups along the poor–rich, south–north dimensions.

Figure 1 provides graphical evidence of the existence of groups. Using the initial conditions of income per capita as the ordering device, I plot the log of the predictive density as a function of the location of the break point, together with the log-predictive density obtained assuming no breaks (the dotted line). The first panel refers to the full sample, and the next two to the subsamples obtained

[6] As an informal check of this conjecture, I separately examined the case of regions in Italy and Spain, for which either educational data or government expenditure for infrastructures are available. I find that Italian regional differences in average human capital and in the distribution of human capital are small and typically unrelated to the time path of income per capita in the sample. Similarly, differences in goverment expnditure for infrastructures in Spain are unimportant as a grouping device.

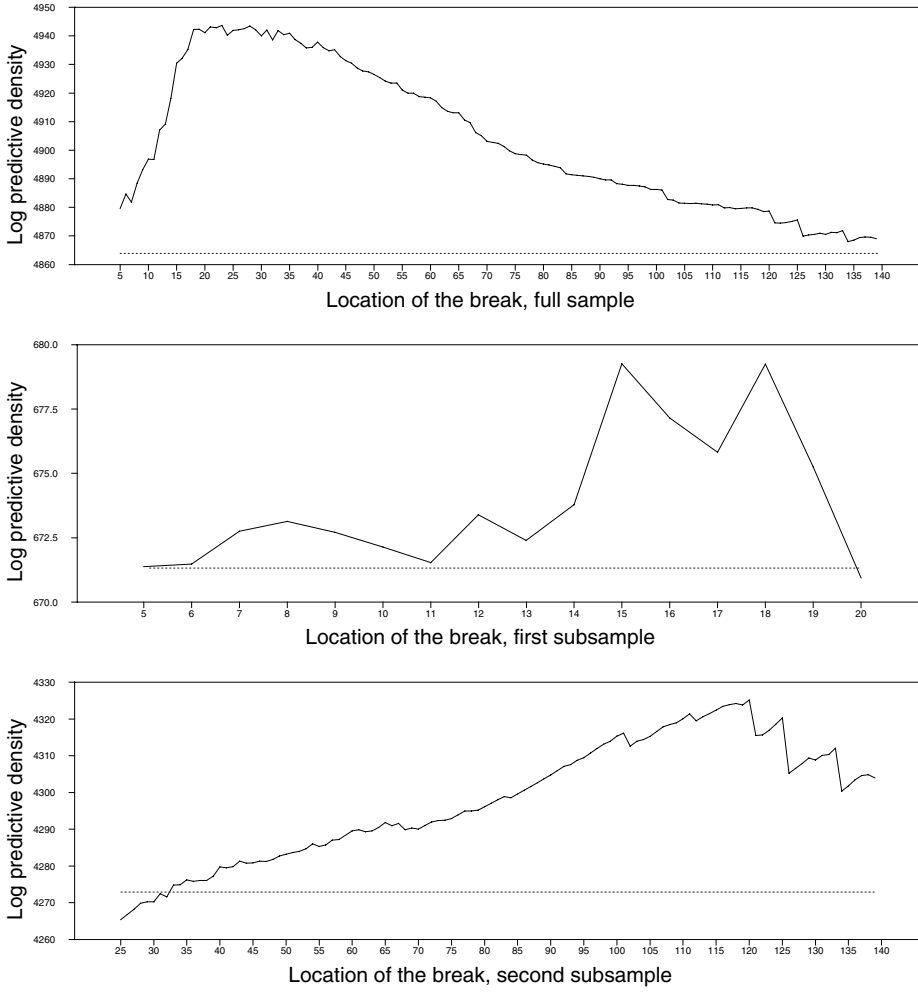

FIGURE 1

REGIONAL DATA, EUROPEAN SCALING

separating units according to the first optimal splitting. To interpret the graphs note that the horizontal entries give the location of the break and the vertical entries the value of the log predictive density. For example, entry 23 on the horizontal axis in the first panel indicates that assigning units 1–23 to the first group and units 24–144 in the second gives a value for the log of L^+ of 4943 (as compared to log $L^+ = 4863$ when no breaks are allowed). Similarly, the second panel indicates that splitting group 1 in two subgroups (1–15 and 16–23) could be beneficial (this produces a log $L^+ = 679$ as compared to log $L^+ = 670$ when no breaks are allowed). Finally, splitting group 2 in two subgroups (24–120, 121–144) gives log $L^+ = 4325$ (as compared to log $L^+ = 4272$ when no breaks are allowed).

Next, I examine whether these differences are significant using the posterior odds ratio. In each case, I use a penalty function of the form $\mathcal{P}(N) = \exp\{-0.5\ln(N)\}$ that resembles the one employed by the Schwartz approximation to the PO ratio, and assign equal prior probability to the null and the alternative. The results overwhelmingly suggest the presence of (at least) three breaks: These corresponding to units 23 and 120 have PO ratios in excess of 100, whereas the posterior probability for the break at unit 15 has a PO ratio of 91.38. In general, the fit of a model with three breaks is substantially better than the one without breaks: The log-predictive density is of an order of magnitude larger and a posterior odds ratio decisively favors the hypothesis of heterogeneities. Hence, we need very strong *a priori* expectation on the null for the data not to overturn our convictions (prior odds needs to be about 100 to 1). Also, these expectations do not substantially change as the number of break points we are testing for increases.

Economic differences among the groups also appear to be relevant. I present estimates of β^p for the whole sample and for each of the four selected groups in Table 1. It is clear that the four groups can be identified by both the value of the intercept and of the slope of the model. For example, the first group displays very low average persistence in relative income per capita (low ρ^p) and below average mean intercept (low and negative α^p). At the opposite end, the last group features higher average persistence and above average mean intercept (high ρ^p and positive α^p). Interestingly, the central group, which contains the largest number of units, has a mean value for the persistence parameter that is higher than that of the last group.

Dispersion measures are significant in three of the four groups, stressing the need to control for residual within-group heterogeneity, but vary substantially across groups. For example, differences in the persistence parameter are small in the second group (0.04) but large in the last one (0.64). In three of the four groups the dispersion is substantially smaller than the dispersion obtained by (weakly) pooling together all units with an exchangeable prior, suggesting a reduction of the residual heterogeneity once groups are identified. The last group, which includes

TABLE 1

ESTIMATES OF THE HYPERPARAMETERS AND OF STEADY STATES

	Hyperparameter Estimates					Posterior Estimates	
	α^p	σ_α^2	ρ^p	σ_ρ^2	$\sigma_{\alpha,\rho}$	Mean SS	Dispersion SS
Overall	−0.086	0.060	0.725	0.298	−0.071	−0.2712	0.6234
Group 1 (units 1–15)	−0.598	0.102	0.251	0.155	−0.031	−1.3171	0.3883
Group 2 (units 16–23)	−0.368	0.019	0.534	0.048	−0.042	−0.6369	0.0751
Group 3 (units 24–120)	−0.032	0.0004	0.686	0.193	−0.008	−0.1390	0.1468
Group 4 (units 121–144)	0.116	0.052	0.629	0.641	0.023	0.2922	0.2308

NOTES: The columns labeled "Hyperparameter estimates" report estimates of the hyperparameters obtained maximizing the predictive density of the data, viewed as function of the hyperparameters. The steady state for each region is computed as $\lim_{T\to\infty} \frac{\alpha_i * (1-\rho_i)^{T+1}}{1-\rho_i} + \alpha_i^T y_{i0}$ where α_i and ρ_i are posterior estimates. The columns named "Mean SS" and "Dispersion SS" report the mean and the standard deviation of steady states.

few outliers (Dutch oil producing regions), appears to be sufficiently heterogeneous to require a further subdivision. However, the procedure was unable to locate any further break in this group.

To summarize the features of the posterior distribution of the β, I report three economically interesting functions of the coefficients of the model: a scatter plot of speeds of adjustment to the steady state $(1 - \hat{\rho}_i)$ against the initial income per capita conditions for each of the four groups, the mean and the dispersion of estimated steady states for each group, and a long-run mobility index.

With the scatter plots I am interested in verifying whether units with below average initial conditions adjust faster or slower to their steady state than units with above average initial conditions. Recall that in the standard neoclassical growth model the speeds of adjustment do not depend on the initial conditions. The second statistic provides information on the core question of this article, i.e., whether the identified groups do cluster around different steady states. The mobility index, on the other hand, synthetically measures the likelihood of switching income classes in the long run (i.e., it measures the probability of "miracles and busts"). Such an index is of interest to policymakers concerned with, e.g., the evaluation of transfer programs to underdeveloped regions. Here I consider only two classes (above and below average income).[7] The mobility index is calculated as $M = 1 - p_{11} - p_{22}$ where p_{ii} is the estimated probability of staying in the class where a unit starts, $i = 1, 2$. Notice that $-1 \leq M \leq 1$, with > 0 indicating mobility and $M < 0$ supporting the idea that there is persistence of inequalities.

Figure 2 indicates that indeed there are striking differences in the relationship between speeds of adjustment and initial conditions in the four groups. Although for the first two groups the slope is strongly negative (estimates are -0.86 and -0.90, respectively) and the R^2 in relatively large, the slope for the third group is still negative (-1.03) but the dispersion around the line is very large, whereas the slope for the fourth group is positive (0.91) and the dispersion around the line is measurable. These differences are statistically significant. Notice also that there are a number of regions in the last two groups that have speeds of adjustment that are either negative or greater than 1, indicating possible nonstationary or oscillatory posterior dynamics.

Table 1 confirms that the identified groups do constitute different clubs. The means of the steady states are statistically different across groups (given equal prior probability, the posterior probability that they are equal is negligible for every pair except the first two) whereas the dispersion of steady states varies with the group. The economic significance of these differences is substantial: The mean steady state of the first group is around 45 percent of the average and the mean of the fourth group is about 15 percent above the average. Also, the steady-state distribution is far from collapsing for all but group 2, but there is a reduction of the steady-state dispersion once units are appropriately grouped.

The mobility index for the whole sample is equal to -0.24 (see Table 2), suggesting a very weak tendency to transit from the position where units start. Since

[7] Changing the threshold from the mean to the median do not change the qualitative features of the results.

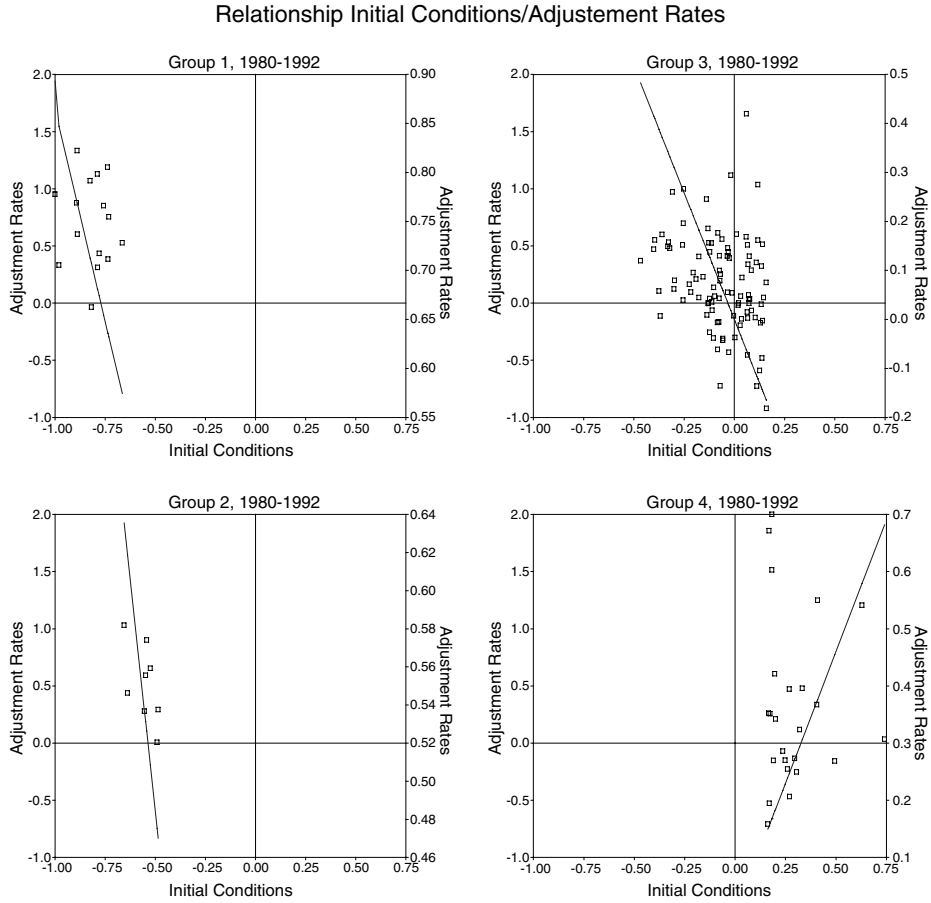

FIGURE 2

REGIONAL DATA, EUROPEAN SCALING.

such a tendency is much stronger for units, which starts above the mean, whereas poor regions tend to stay uniformly poor; busts are more probable than miracles. The four groups display different mobility characteristics. In the first group there is a strong tendency to stay in the low-income class and in the second group there is complete immobility. The third group mirrors, with minor differences, the tendencies of the whole sample, but 67 percent of the units starting above the average end up below it in the steady state. The fourth group also shows a tendency to slump: About 50 percent of those who started above average are expected to be below average in the steady states (curiously, most are French and German regions!).

Few general economic conclusions can be drawn from the analysis. Among the indicators suggested by theory, the distribution of income per capita at the beginning of the sample seems to be the one with the highest information content.

TABLE 2
MOBILITY INDICES

	Overall	Group 1	Group 2	Group 3	Group 4
M	−0.24	−0.41	−0.50	−0.18	0.00
P_{11}	0.83	0.91	1.00	0.85	0.00
P_{12}	0.17	0.09	0.00	0.15	0.00
P_{21}	0.59	0.00	0.00	0.67	0.50
P_{22}	0.41	0.00	0.00	0.33	0.50

NOTES: The M statistic is given by $M = 1 - P_{11} - P_{22}$. P_{11} is the probability that the unit starts below average and ends up below average in the steady state, P_{22} is the probability that the unit starts above and ends up above average in the steady state, P_{12} and P_{12} are the probabilities that the unit transits from a state to the other. In the case, the group is unbalanced, so that all units in the group are initially in one income class, the statistics M is computed as $M = 0.5 - P_{ii}$ where P_{ii} is the diagonal value different from zero.

This ordering emphasizes the North–South, rich–poor dimension, and produces club convergence with interesting characteristics. Income dynamics of initially poor regions are different from those of the initially rich and there is little tendency for the poor to move up in the income distribution ladder whereas the initially rich may fall back into mediocrity. Extrapolating this tendency far into the future implies that the distribution of income per capita will become more polarized with few very wealthy regions and the rest clustered in few groups below the average. Finally, the low mobility in the income distribution ladder of the majority of poor and very rich units confirms that inequalities in European regions will tend to persist over time (as noted by Canova and Marcet, 1995).

Quah (1996a) has argued that once geographical externalities are taken into account the tendency toward convergence clubs weakens. Does this occur in our sample? The answer is partially positive. In Figure 3 I plot the log of the predictive density as a function of the location of the break when regional income per capita is scaled by the national average and ordered according to the magnitude of the scaled initial conditions. There is evidence of only one significant break (producing two groups with units 1–93 and 94–141), but now estimates of the hyperparameters for the two groups are more similar. For example, the AR parameters have a mean of 0.597 in the first group and 0.713 in the second. Moreover, differences in estimated steady states are much smaller than those obtained scaling per capita income with the European average, and the dispersion around the two steady states is substantially reduced. Hence, geographical and/or informational externalities may be present: Once these effects are taken into account, the number of clubs is smaller and the economic differences among them significantly reduced.

Barro and Sala-i-Martin (1995) when analyzing EU regional data suggest that conditional convergence holds. Why are our results different? We can think of two reasons for why this is the case. First, the sample used and its composition differ: Barro and Sala-i-Martin used data for 73 regions (primarily German, French, and Italian) from the 1950s up to the beginning of 1980, whereas I am using data from 1980 to 1992 for 144 regions. It is a well-known fact (see, e.g., Boldrin and

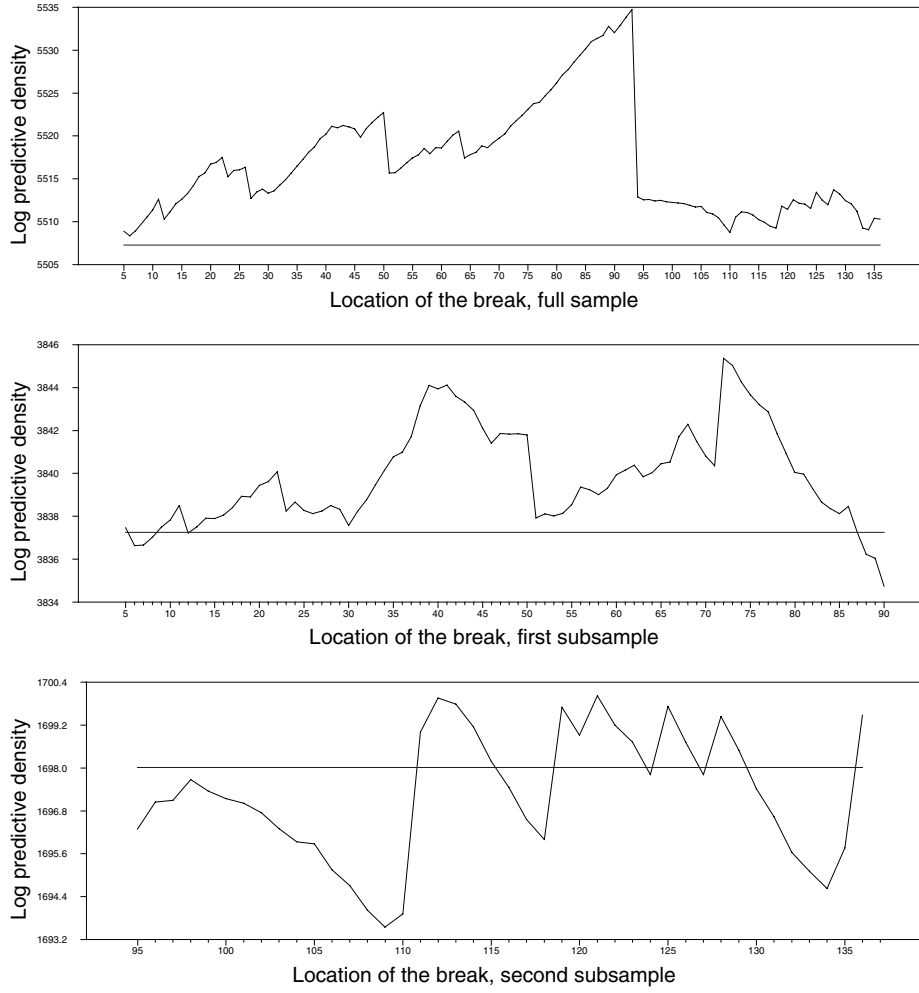

FIGURE 3

REGIONAL DATA, NATIONAL SCALING

Canova, 2001) that the convergence process in Europe stopped at the beginning of the 1980s. Furthermore, the majority of the 73 regions examined by these authors belong to the identified third convergence club. Therefore, it is perhaps not surprising that the two studies reach different conclusions. Second, and probably, more important, Barro and Sala-i-Martin do not use the information contained in the panel. Instead, they take averages of growth rates and run a cross-sectional regression on the initial conditions. Such an approach disregards those unit-specific heterogeneities that this and other articles found to be very important even in a group of regions with relatively similar institutional setups.

5.2. *OECD National Income Per Capita.* For this data set $N = 21$, time runs from 1951 to 1985 and at most three groups are allowed (i.e., $Q = 2$). Following Canova and Marcet (1995) a AR(1) with country-specific parameters is chosen for the log of income per capita scaled by the OECD average. Contrary to the case of regional data, useful information to order units is available at the country level. Hence, I search for clubs ordering units according to (i) the magnitude of the per capita GDP relative to the OECD average in 1950, with poor units coming first; (ii) the magnitude of the average human capital in 1950, measured as in Barro and Lee (1994), ordering units increasingly in their average endowment of human capital; (iii) the magnitude of the government expenditure share in 1950 or on average in the sample period; (iv) the dispersion of income distribution in 1950 (Gini index from the Luxemburg Income Study), with units displaying high dispersions coming first; (v) the dispersion of the distribution of human capital in 1950, (measured as the sum of the percentage of the population with primary and university education using Barro and Lee data), with units displaying high dispersions coming first; (vi) a center–periphery classification of the world economy (G-3 first and then the rest); (vii) a geographical criterion with European nations first and rest of the world afterward and Mediterranean countries preceding northern European countries in the order; and (viii) the average openness of the economies (measured as the ratio of import plus exports over GDP).[8]

When one break is allowed the maximized value of the log L^+ for the seven classifications is 2436, 2423, 2411, 2433, 2423, 2420, 2415, and 2430, respectively, suggesting that the predictive power of the model is maximized when units are ordered according to the initial conditions of income per capita. Therefore, consistent with Durlauf and Johnson (1995), the procedure prefers initial output over literacy rates as the most useful ordering device. Note, however, that differences in L^+ for three classifications are relatively small since the ordering of units is very similar in these cases. That is, countries that have low initial income conditions also have low average initial human capital, a distribution of income with high dispersion, and are geographically located in the "South" of the developed world and are less open to trade than average. The maximized value of log L^+ obtained using government expenditure shares as an ordering device is the lowest of all, and insignificantly different from the one with no breaks (2408), indicating that policy variables may have no role in shaping club convergence, at least for OECD countries. Attempts to refine membership in these two initial groups failed. For example, the predictive density obtained by reordering units within groups using literacy rates or government variables is indistinguishable from the one obtained in the baseline case.

Given this ordering, the posterior odds ratio establishes the presence of one break in the cross section, with a value of 0.979, given equal prior probabilities on the null and the alternative. In Figure 4, I plot log L^+ as a function of the location

[8] Substitution of the size of the population holding 50 percent of national wealth for Gini indices and the sum of the inverse of the percentage of the population with primary and the inverse of the percentage of population with secondary education for the percentage of the population holding primary and university degrees does not change the results. The ordering obtained with these new indices is practically identical to the ordering I use.

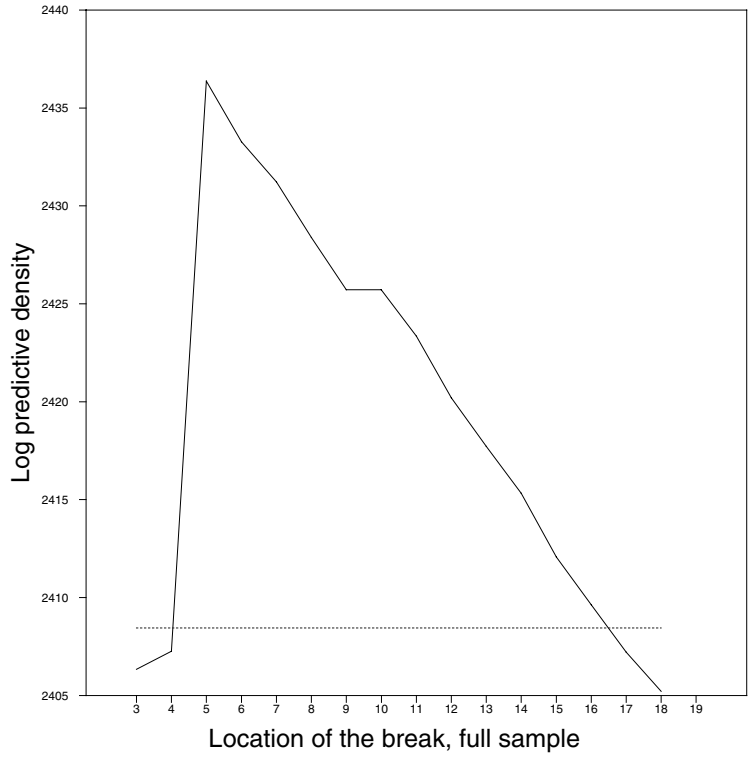

FIGURE 4

OECD NATIONAL DATA

of the break point for the ordering based on initial income per capita together with the predictive density obtained in the case of no breaks (dotted line). The first group contains the five poorest units (Turkey, Portugal, Greece, Spain, and Ireland) and the second group the rest.

Estimates of the hyperparameters for the two groups are $\beta^1 = [-0.162, 0.824]$ and $\beta^2 = [0.0004, 0.958]$, suggesting a much faster a priori average rate of convergence in the first group. The dispersion of estimates is small but nonnegligible (in particular, the dispersion of estimates of the AR parameter is 0.02 in the first group and 0.05 in the second group) indicating, once again, that clustering is more prevalent than convergence even after optimally splitting the sample.

The posterior characteristics of the two clubs differ. For example, the average posterior estimate of the steady state in the first group is -0.7647 and in the second group is 0.0498. This difference is statistically and economically large: It implies that there will be a permanent discrepancy in the average per capita income of units in the two groups of about 60 percent. The dispersion of estimated steady states around these poles of attraction is smaller than the one obtained when all units are (weakly) pooled together. However, differences of about 15–20 percent in steady-state income per capita in each group are still possible. Finally, the mobility

characteristics of the two groups are similar: Apart from a few exceptions, the ranking of units in the income distribution changes very little over time: The initially poor will still be the poorest in the steady state. What is interesting about this last observation is the fact that there is no evidence that the economic boom that took place in Ireland in the late 1990s and allowed the country to move up in the OECD income distribution ladder was forthcoming.

In sum, in agreement with what Quah (1996b) and Durlauf and Johnson (1995) have detected for a larger sample, I find that clustering along the poor–rich dimension is prevalent. Countries that were initially poor are also those having below average initial human capital, large income and educational inequalities, and are located in the "South" of the developed world. These characteristics are very persistent and produce a polarization in the steady-state distribution of income. The policy implications of results are striking: Unless some major changes occur, the initially poor will tend to cluster around a basin of attraction that is substantially below the OECD average and policy can do little to improve the situation of backward countries.

6. CONCLUSIONS

This article describes a procedure to examine the likelihood of convergence clubs in the distribution of income per capita. It proposed a unified approach to testing, estimation, and inference when the number of groups, the location of the breaks, and the ordering of units in the cross section are unknown. The methodology I outline has a number of applications, apart from the one considered in this article. For example, it could be used to examine the differential response of firms to monetary policy shocks or the international transmission of shocks across fixed and flexible exchange rate regimes. In general, the simplicity of the procedure, its easiness of implementation, and the good properties it demonstrates in a simple Monte Carlo exercise makes it a candidate to deal with the issue of grouping in a number of microeconomic and macroeconomic fields.

The procedure employs the predictive density of the data, conditional on the hyperparameters of the model. The use of predictive densities has a long tradition in Bayesian econometrics and provides a simple framework where interesting hypotheses can be verified. What is appealing about predictive densities is that, once hypotheses concerning the number of groups present in the data are examined, the location of the breaks, the best permutation in the data, and the hyperparameters of the model can be easily estimated. Once the hyperparameters are selected, inference can be conducted in an Empirical Bayes fashion and the properties of functions of the posterior estimates of the coefficients can be examined, plugging in hyperparameters estimates in the appropriate formulas.

I search for clubs using income per capita from European regions and OECD countries. I find that there are heterogeneities in European regional per capita income and a tendency of the steady-state distribution to cluster around four poles of attractions characterized by different dynamics, different posterior mean steady states, and different mobility features. Similarly, OECD national per capita income data present two convergence clubs. In both cases a rich–poor,

north–south dimension in the clubs emerge, supporting several versions of the theory we outlined in Section 4.

It is important to stress that the article has demonstrated that the *scaled* distribution of regional and national income per capita shows a tendency to cluster around few poles of attraction when ordered according to the initial conditions of income per capita and that, even within the endogenously selected groups, level convergence is a rare phenomenon. These results do not imply that the *unscaled* level of per capita income shows these features, nor do they shed any light on the existence of a steady-state distribution of per capita income in levels or in growth rates. More important, they do not suggest that one type of economic theory (endogenous growth) is to be preferred to another one (exogenous type) or vice versa, since both theories can generate outcomes that are consistent with the findings of the article.

APPENDIX

A. In this appendix I present the results of a Monte Carlo exercise designed to examine the properties of the testing procedure to uncover breaks and estimation approach for the hyperparameters with data displaying properties similar to those considered in Section 5. For this reason I generate times series for $N = 144$ units, each of length $T = 13$, and assume that the data generating process is

$$y_{it} = \alpha_i + \rho_i y_{t-1} + e_{it} \tag{A.1}$$

$$\beta_i = \beta^1 + u_i^1 \, u_i^1 \sim N(0, \Sigma_1) \quad \text{if} \quad i \leq 50 \tag{A.2}$$

$$\beta_i = \beta^2 + u_i^2 \, u_i^2 \sim N(0, \Sigma_2) \quad \text{if} \quad 51 \leq i \leq 144 \tag{A.3}$$

where $\beta_i = [\alpha_i, \rho_i]$, $\beta_1 = [0.3, 0.8]$, $\beta_2 = [-0.3, 0.4]$, $\Sigma_1 = \text{diag}(0.052, 0.255)$, $\Sigma_2 = \text{diag}(0.102, 0.155)$, $\text{var}(e_{it}) = 0.1$ if $i \leq 51$ and $\text{var}\,(e_{it}) = 0.15$ otherwise. The initial conditions satisfy: $y_{i,0} \sim U[-0.10, 0.10]$.

On the panel of simulated data I estimate both AR(1) and AR(2) models for each $i = 1, \ldots, 144$ and apply the testing procedure to examine whether there is a break in the cross section using data in the order I have generated. The posterior odds ratio, giving equal chance to the possibility that there is one break and there are no breaks, is 0.4×10^{-8} for the two models. The hypothesis that there is a further group produced a posterior odds ratio of the order of 0.54×10^{7} for the two models, confirming the presence of one break only. Figure A.1 plots the log of L^+ as a function of the location of the break $\bar{\imath}$ for the AR(1) model (the first panel) where $h_1^1 = 10$, $h_2^1 = 135$. The peak is achieved at $\bar{\imath} = 50$, implying the presence of two groups comprising units 1–50 and 51–144, and there are no other peaks within the range I explore. Repeating the experiment 100 times, I find that in 100 percent of the cases the posterior odds ratio reveals the presence of two groups in the cross section and the predictive density is maximized in 85 percent of the time at $\bar{\imath} = 50$ (96 percent of the times for $\bar{\imath} \in [49, 51]$). The average posterior

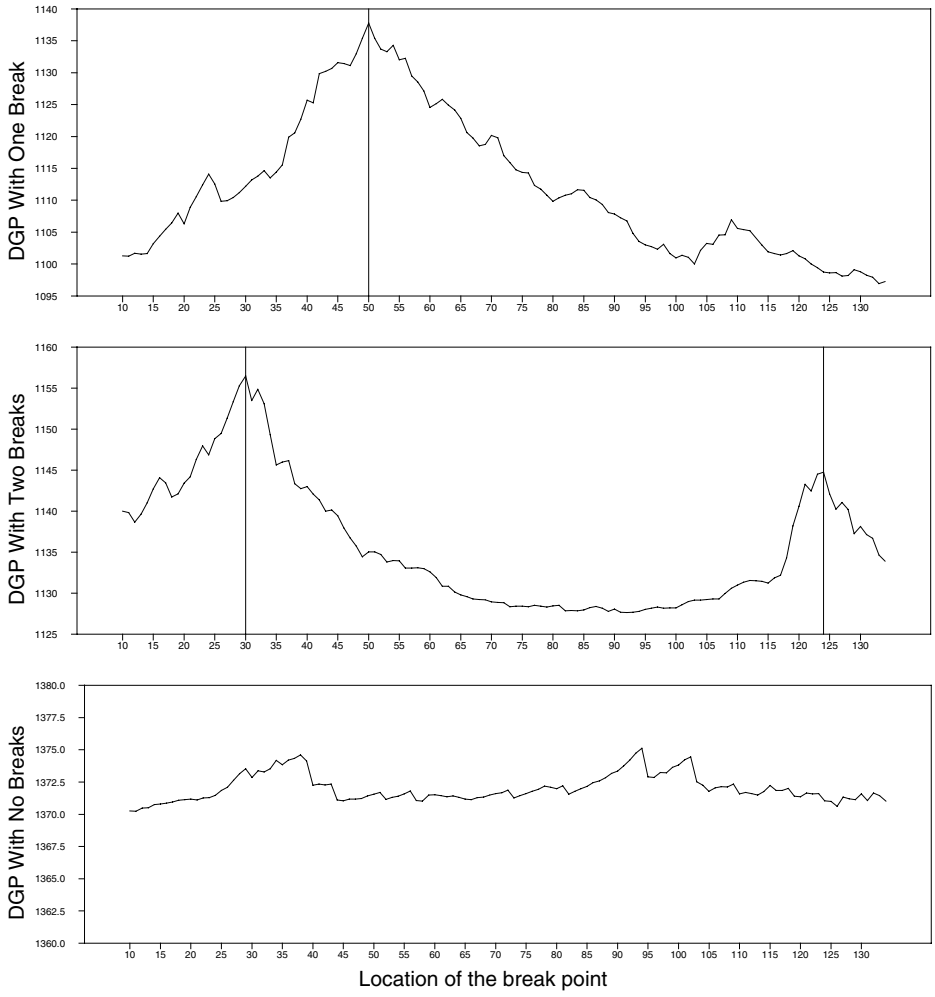

FIGURE A.1

SIMULATED DATA, PREDICTIVE DENSITIES

odds ratio for the hypothesis that there are three groups in the generated data is 0.8×10^8.

Next, I conducted three experiments: First, I randomized the order of the units within the two groups before estimation is undertaken. This did not change any of the results, confirming that, absent any information on the appropriate ordering of the data, the number of actual permutations to be tried is substantially less than N!. Second, I reshuffled the entire cross section, taking the first 20 units of the time series and putting them last. In this case the ordered data displays three groups with breaks at $i = 30$ and $i = 124$. Estimating an AR(1) model on the data, the posterior odds ratio finds two breaks, and the predictive density is maximized

at $\bar{\imath}_1 = 30$ (see plot of the log of L^+ as function of $\bar{\imath}$ in panel 2 of Figure A.1). The pattern displayed by the predictive density in this case is very well known from the break point literature (see Bai, 1997 or Hansen, 1999) and conveys information confirming the presence of three groups in the sample. In fact, conditional on having a break at $\bar{\imath}_1 = 30$, the posterior odds ratio for a second break is 0.4×10^{-6} and the location of the break is $\bar{\imath}_2 = 124$. Repeating this experiment 100 times I find that the neighborhood $\bar{\imath}_1 \in [29, 31]$ is identified as the first break point 80 percent of the times and that the neighborhood $\bar{\imath}_2 \in [123, 125]$ is identified as the first break point in 12 percent of the cases (average PO ratio for the hypothesis of one break is 0.3×10^{-7}). Conditional on having a break at $\bar{\imath}_1 = 30$, the latter neighborhood is identified as the second break in 84 percent of the cases (average PO ratio for the hypothesis of two breaks is 0.6×10^{-8}).

The design of this second experiment also allows examination of the power of the test when the cross-sectional data are not properly ordered. That is, suppose that the DGP is such that there are only two groups in the data, but an econometrician has available unordered data. Would the procedure be able to recognize the optimal permutation of the units in the cross section, select the correct ordering with only two groups, and find the location of the break point? To provide an idea of the properties of the approach in this case I assume that there is a break at unit 56 and reshuffle blocks of 28 units, so that I allow 5! combinations (120 trials) over which to search for the optimal ordering. Figure A.2 plots the log of L^+ and the selected location of break point as a function of the permutation $m = 1, \ldots, 120$. There is a plateau in the log of L^+, corresponding to the 12 permutations that correctly put the first two blocks first and the next three last and for the remaining cases the log of L^+ declines slowly. Notice also that for all permutations the log of L^+ is substantially higher than the likelihood under the null (the dotted line in the graph). Also, the procedure correctly identifies the location of the break in those 12 cases when the log of L^+ is maximized.

Finally, I study the properties of the testing procedure when the cross section is homogeneous (the parameters for the two groups are those for β_1, Σ_1, and e_1). The posterior odds ratio for the hypothesis of 0 versus 1 breaks gives a value of 2.92 and the predictive density as a function of $\bar{\imath}$ produces a plateau with little difference between the minimum and the maximum values (see the third panel of Figure A.1). Replicating the experiment 100 times I find that the average posterior odds ratio giving equal prior probabilities to the null and the alternative of two groups, is 2.33. The distribution of the break point is practically uniform in the interval [10, 135], confirming the results obtained with one experiment only.

Estimates of the hyperparameters of the model are, in general, biased. In particular, the average values across 100 experiments, in the baseline case are $\beta^1 = [0.3614, 0.6931]$, $\beta^2 = [-0.3660, 0.2682]$, indicating that estimates of ρ are downward biased and, as a consequence, estimates of α are upward biased. This appears to be due to the small time series size of each cross section: If I increase the sample size to $T = 36$ (the size of the time series with OECD country data), most of these biases disappear. The variances of all the estimated coefficients are also upward biased by 25–50 percent. Again, the bias drops to 10–15 percent when $T = 36$. When there are three groups in the cross section, results are similar even

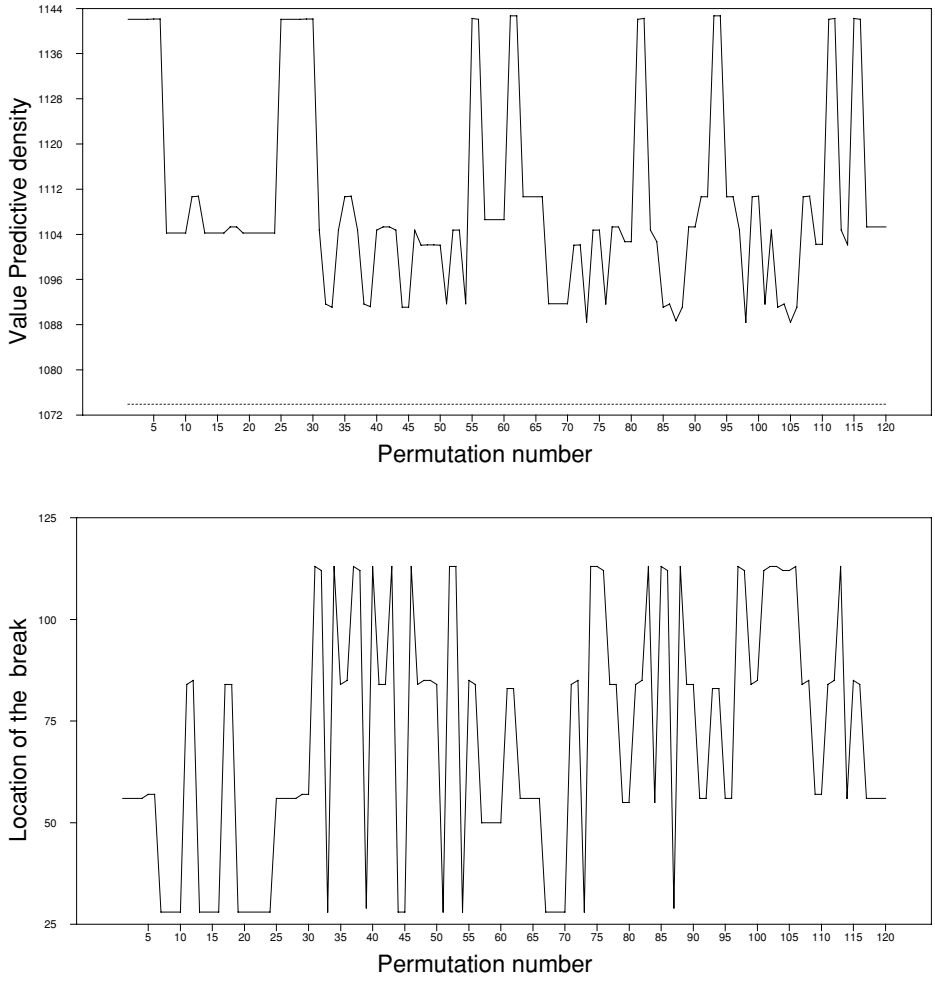

FIGURE A.2

SIMULATED DATA, DGP WITH ONE BREAK

though average estimates of the hyperparameters of the third group are more biased, probably because of the small number of units in this group. Finally, when the cross section is homogeneous, average estimates (across replications) are still biased but by a smaller amount (average $\beta = [0.3816, 0.7372]$) whereas variances of the estimated coefficients are similar to those obtained in the baseline case.

Overall, the results indicate that the testing procedure has a reasonable size and power properties against the particular alternative I consider. It also appears to be able to identify multiple groups and the location of the breaks with sufficient precision, even when the data are not correctly ordered. However, since the posterior odds ratio appears to be slightly biased when there are no heterogeneities and no

penalty is used, a conservative strategy would be to choose the penalty function so that the posterior odds ratio is, on average, 1. This implies a value for $\mathcal{P}(N)$ of about 0.31, which is less harsh than the one used in the article ($\exp\{-0.5\ln(N)\} = 0.08$).

When the time series size of each cross section is small, estimates of the autoregressive parameters are downward biased and averaging over the cross section does not help since estimates of all the units are downward biased (see also Pesaran and Smith, 1995). When the size of each cross section is greater than 30, estimates of the hyperparameters obtained by maximizing the predictive density of the data are sufficiently precise whereas estimates of the dispersion of the prior distribution are still significantly biased.

B. This appendix lists the NUTS2 regions belonging to the four groups we have found:

Group 1: Greece (10) Anatoliki Makedonia, Kentriki Makedonia, Dytiki Makedonia, Thessalia, Ipireos, Ionia Nisia, Dityki Ellada, Voreio Aigaio, Notio Aigaio, Kriti; Portugal (4) Norte, Centro, Alentejo, Algarve; Spain (1) Extremadura.

Group 2: Greece (3) Sterea Ellada, Peloponnisos, Attiki; Spain (4): Galicia, Castilla-La Mancia, Andalucia, Ceuta y Melilla; Italy (1): Calabria.

Group 3: Belgium (7): Hainault, Liege, Limburg, Luxemburg, Namur, Oost-Vlaanderen, West Vlaaanderen; Denmark (1); Germany (23) Freiburg, Tubingen, Niederbayern, Oberpfalz, Oberfranken, Unterfranken, Schwaben, Giessen, Kassel, Braunschweig, Hannover, Lunenburg, Weser-ems, Koln, Munster, Detmold, Arnsberg, Koblenz, Trier, Rheinhessen-pfalz, Saarland, Schleswig-Holstein; Spain (13) Asturias, Cantabria, Pais Vasco, Navarra, Rioja, Aragon, Mardid, Castillia-leon, Catalunia, Comunidad Valenciana, Baleares, Murcia, Canarias; France (21) Champagne-Ardenne, Picardie, Haute-Normadie, Centre-Normandie, Basse Normandie, Bourgogne, Nord-Pas de Calais, Lorraine, Alsace, France-Comte, Pays de la Loire, Bretagne, Poitou-Charantes, Aquitaine, Midi-Pyrenees, Limousin, Rhone-Alps, Auvergne, Languedoc-Roussilon, Provence-Alpes-Cote d'azur, Corse; Ireland (1); Italy (13) Veneto, Friuli-Venezia Giulia, Toscana, Umbria, Marche, Lazio, Campania, Abruzzi, Molise, Puglia, Baslicata, Sicilia, Sardegna; Netherlands (9) Friesland, Drenthe, Overijessel, Gerderland, Flevoland, Noord-Holland, Zuid-Holland, Zeeland, Noord-Brabant; Portugal (1) Lisboa e Vale do Tejo; UK (9) North, Yorkshire and Humberside, East Midlands, East Anglia, South West, West Midlands, North West, Wales, Scotland, Northern Ireland.

Group 4: Netherlands (3): Groningen, Utrecht, Limburg; Luxemburg (1); England (1): South East; Belgium (2): Antwerp, Brabant; Germany (9) Stuttgard, Karlsruhe, Oberbayern, Mittelfranken, Berlin, Bremen, Hamburg, Darmstadt, Dusseldorf; Italy (6) Piemonte, Valle d'Aosta,

Liguria, Lombardia, Trentino-Alto Adige, Emilia-Romagna; France (1): Ile de France.

REFERENCES

AZARIADIS, C., AND A. DRAZEN, "Threshold Externalities in Economic Development," *Quarterly Journal of Economics* 105 (1990), 501–26.

BAI, J., "Estimation of Multiple Breaks One at a Time," *Econometric Theory* 13 (1997), 315–52.

BARRO, R., AND G. BECKER, "Fertility Choice in a Model of Economic Growth," *Econometrica* 57 (1989), 481–501.

——, AND J.-W. LEE, Data set for a panel of 138 Countries, manuscript (1994).

——, AND X. SALA-I-MARTIN, *Economic Growth* (New York: McGraw Hill, 1995).

BEN DAVID, D., "Convergence Clubs and Diverging Economies," CEPR working paper 922, 1994.

BERGER, J., *Statistical Decision Theory and Bayesian Analysis* (New York: Springer Verlag), 1985.

BOLDRIN, M., AND F. CANOVA, "Inequalities and Convergence in Europe's Regions: Reconsidering European Regional Policies," *Economic Policy* 32 (2001), 205–45.

CANOVA, F., AND M. CICCARELLI, "Forecasting and Turning Point Predictions in a Bayesian Panel VAR model," CEPR working paper 2961 (1999).

——, AND A. MARCET, "The Poor Stay Poor: Non-Convergence Across Countries and Regions," CEPR working paper 1215 (1995).

CARLIN, B., AND A. GELFAND, "Approaches for Empirical Bayes Confidence Intervals," *Journal of the American Statistical Association* 85 (1990), 105–14.

DESDOIGTS, A., "Pattern of Economic Development and the Formation of Clubs," Universite' d'Evry-Val D'Essonne EPEE, working paper, 1998.

DURLAUF, S., AND P. JOHNSON, "Multiple Regimes and Cross Country Growth Behavior," *Journal of Applied Econometrics* 10 (1995), 365–84.

EFRON, B., "Empirical Bayes Methods for Combining Likelihoods," *Journal of the American Statistical Association* 91 (1996), 538–65 (with discussion).

FORNI, M., AND L. REICHLIN, "Let's Get Real: Dynamic Factor Analytical Approach to Disaggregated Business Cycles," *Review of Economics Studies* 65 (1997), 453–74.

GALOR, O., "Convergence? Inference from Theoretical Models," *Economic Journal* 106 (1996), 1056–69.

——, AND J. ZEIRA, "Income Distribution and Macroeconomics," *Review of Economic Studies* 60 (1993), 35–52.

HANSEN, B. E., "Threshold Effects in Non-Dynamic Panels: Estimation, Testing and Inference," *Journal of Econometrics* 93 (1999), 345–68.

——, "Sample Splitting and Threshold Estimation," *Econometrica* 68 (2000), 575–607.

HARTIGAN, J., *Clustering Algorithms* (New York: Wiley, 1975).

MADDALA, G. S., "To Pool or Not to Pool: That Is the Question," *Journal of Quantitative Economics* 7(2) (1991), 255–64.

MARDIA, K., J. KENT, AND J. BIBBY, *Multivariate Analysis* (New York: Academic Press, 1980).

MORRIS, C. N., "Parametric Empirical Bayes Inference: Theory and Applications," *Journal of American Statistical Association* 78 (1983), 47–59.

PAAP, R., AND H. VAN DJIK, "Distribution and Mobility of Wealth of Nations," *European Economic Review* 42 (1998), 1269–93.

PESARAN, H., AND R. SMITH, "Estimating Long Run Relationships for Dynamic Heterogeneous Panels," *Journal of Econometrics* 68 (1995), 79–113.

PHILLIPS, P. C. B., AND W. PLOBERGER, "Posterior Odds Testing for Unit Root with Data-Based Model Selection," *Econometric Theory* 10 (1994), 774–808.

POLASEK, W., AND L. REN, "Structural Breaks and AR Modelling with Marginal Likelihoods," University of Basel, manuscript, 1997.

PLOBERGER, W., W. KRAMER, AND K. KONTRUS, "A New Test for the Structural Stability in the Linear Regression Model," *Journal of Econometrics* 40 (1989), 307–18.
QUAH, D., "Convergence Empirics across Economies with Some Capital Mobility," LSE Center for Economic Performance 257 (1996a).
——, "Regional Convergence Clusters across Europe," CEPR working paper 1286 (1996b).
TITTERINGTON, J., R. MAKOV, AND J. SMITH, *Statistical Analysis of Finite Mixture Distributions* (New York: Wiley, 1985).