Clustering Regression Functions in a Panel

Farshid Vahid
Department of Econometrics and Business Statistics
Monash University
Clayton, Victoria 3800
Australia.
Phone: (+61) (3) 9905-2412
Fax: (+61) (3) 9905-5474
E-mail: Farshid.Vahid@BusEco.monash.edu.au

Abstract

This paper suggests a hierarchical clustering algorithm with a global objective function, to partially pool regressions when the overall pooling restriction is rejected by the data. In addition to the lack of fit and lack of parsimony, the objective function also penalizes lack of conformity with theoretical priors and imprecision in the estimated parameters. This algorithm is used for clustering the gasoline demand functions of OECD countries. The results are compared with those of an alternative method based on a classification and regression tree (CART) procedure.

Keywords: Medium sized panels, cluster analysis, information criteria, minimum message length, classification and regression tree (CART).

1. Introduction

We introduce a procedure for classifying linear conditional expectation functions in "medium-sized" panels, i.e. panels with a moderately large time dimension, and not too large a cross sectional dimension. We are motivated by regional and sectoral panel studies (such as the 50 states of the United States, the 18 OECD countries, 20 two digit manufacturing industries, etc.) where, for each region or sector, the parameters of the conditional expectation function of one (dependent) variable given some other (independent) variables are of interest, and regions or sectors are independent enough so that there are no feedbacks from one region or sector to the other, and hence a VAR analysis on the entire panel would not be appropriate[1].

A central question in the study of panels with a relatively large time dimension is *when* and *how* the homogeneity assumption, which was particularly (or at least partially) made in earlier panel data studies because the time dimension was small, might be relaxed. This question has attracted some attention under the title of "To pool or not to pool" in the literature. The answer to *when*, seems, at least on the surface, easy: test the null hypothesis that coefficients are equal across the cross sectional units, and do not assume homogeneity if the null is rejected. There is a large body of applied work based on cross country or sectoral data which has tested and rejected the pooling restriction. For example, Burnside (1996) rejected the hypothesis that production function parameters across the US manufacturing sectors were equal, and Baltagi and Griffin (1997) rejected the hypothesis that gasoline demand elasticities across the OECD countries were equal. However in both cases, they noticed that the fully heterogeneous model, which assumed no cross equation restrictions, led to very imprecise parameter estimates, which in some cases had the wrong signs. In addition, Burnside observed that the general conclusion about the returns to scale of the US manufacturing sector would be quite different if one used the fully homogeneous, or the fully heterogeneous models. Baltagi and Griffin also noticed that the squared out of sample long-horizon forecast error of the pooled model was considerably smaller than that of the fully heterogeneous model.

[1] The question of dimension reduction in linear and non-linear VAR's have attracted considerable attention in the last ten years. See, inter alia, Engle and Granger (1987), Ahn and Reinsel (1988), Tiao and Tsay (1989), Vahid and Engle (1993, 1997) and Anderson and Vahid (1998).

This paper studies what can be done when the pooling hypothesis is rejected. We start with the observation that the alternative hypothesis to the pooling restriction is that at least one of the cross section units has different parameters from the rest. Therefore the rejection of the pooling hypothesis does not provide any evidence about the extent of heterogeneity across the cross sections, and in particular does not point to the alternative of a fully heterogeneous model. Of course, in the absence of any a-priori reason to group a subset or subsets of cross sectional units, if there is a lot of structure in the time series data, in the sense that parameters can be precisely estimated from the individual time-series models, then full heterogeneity will be a safe way to proceed[2]. However, our focus is the case where individual dynamic models do not lead to sharp inference about the parameters. The motivating application underlying this paper is that if an econometrician in country X is asked to predict the short and long run effects of a consumption tax on the demand for commodity Y, and if the time series data in that country does not lead to precise estimates of the price elasticity, should the econometrician stop looking at the historical evidence from other countries because the pooling restriction of equality of demand elasticities for *all* countries for which data are available is rejected? Our answer is no, and we suggest a systematic way in which cross sectional units can be partially pooled into homogeneous clusters[3].

There are many ways that N cross sectional units can be partitioned into non-empty groups. Ideally one would like to find the partition that minimizes some loss function, subject to the constraint that the null hypothesis that the parameters of the cross sectional unit within each sub-group are equal, is not rejected. However, since a global search over all possible partitions might be infeasible, we suggest a hierarchical clustering procedure. It starts from the fully heterogeneous model, and pools the two cross sectional units that are "closest" to each other conditional on their distance being less than a threshold value c, and continues until there are no two groups within c of each other. The measure of inter-cluster

[2] For example, Ramanathan et al (1997) estimate 24 fully heterogeneous models for conditional expectation of hourly electricity load given temperature, for the 24 hours of the day. In real time forecasting, their model has outperformed all competing linear and nonlinear pooled models.

[3] An alternative solution, suggested by Maddala et al (1997), is to assume that cofficients are random across the cross sectional units with a specific distribution. This will, in effect, lead to the shrinkage of the indiviual country estimates towards the pooled estimate.

distance that we use is the value of the χ^2 test statistic for the null that the parameters of the two clusters are equal. The threshold value that determines when two units can be pooled together is the 5% critical value of this null distribution.

It is conceivable that there are more than one non-nested partitions (i.e. partitions that one is not the same as the other with two or more of its sub-groups pooled together) of N cross sectional units for which the homogeneity of cross sectional units within each sub-group is not rejected. Hence it is important to have an appropriate loss function to be able to compare competing partitions and choose the "best" one. We combine features of measures of complexity introduced in coding and information theory in order to develop a loss functions that in addition to goodness of fit and parsimony, explicitly takes account of the (lack of) correspondence to theoretical priors and the (lack of) precision of parameter estimates, which are the two major concerns that have made the researchers wary of the fully heterogeneous models.

Even with the additional requirement that at each stage the closest clusters should be fused only if the resulting partition has a smaller loss than before, our hierarchical clustering algorithm is still a local search algorithm. Similar to other local optimization procedures, this algorithm is not guaranteed to lead to the globally optimal partially pooled model, and may lead to different local optima depending on the path of the search. If N is so large that considering all possible hierarchical paths is not feasible, we use an exchange method on the final outcome of the hierarchical algorithm to make sure that the loss cannot be lowered by moving a unit from one cluster to another.

This paper is structured as follows. Section 2 is the main section of the paper, which gives a more formal and detailed description of the points raised in the introduction. In this section the hierarchical clustering algorithm, the loss function and the exchange method that are the three ingredients of our partial pooling procedure are discussed. Section 3 applies this algorithm to the data used in Baltagi and Griffin (1997) to partially pool gasoline demand schedules of the OECD countries. Section 4 reviews the related literature and one possible alternative method of clustering regressions. We have, rather unconventionally, left the discussion of the related literature to the section after the empirical example, so that the type of empirical questions that have motivated this paper, will be clearer for the reader.

The concluding Section 5 discusses some directions for future research on partial pooling of conditional models.

2. Partial pooling of dynamic regressions

2.1. The statement of the problem

Suppose we have time series observations of length T, on K variables across N cross sectional units. Let z_{it} denote the period t observation on the K-vector of random variables in cross sectional unit i, and partition z_{it} into the scalar random variable y_{it} (the "dependent variable") and the $K-1$ vector of random variables x_{it} (the "independent variables"). The objective is to provide conditional models for y_{it} given (x_{it}, Φ_{it-1}) for all $i = 1, ..., N$, where Φ_{it-1} is the information set of the cross sectional unit i at time $t-1$, i.e. $\Phi_{it-1} = z_{it-1}, z_{it-2}, ...$.

It is important to be clear right at the outset that this paper studies the question of whether some parameters of the conditional expectation functions across cross sectional units are equal, and it assumes that these functions are linear. Throughout the paper, we assume that z_{it} is strictly stationary and ergodic, and that the distribution of y_{it} given (x_{it}, Φ_{it-1}) is normal. We also assume that conditioning on a finite number of lags of z_{it}, completely specifies the dynamics. This rules out moving average errors, and implies autoregressive distributed lag models for every cross section $i = 1, ..., N$:

$$y_{it} = \alpha_i + x'_{it}\beta_i + \left(z'_{it-1}, ..., z'_{it-p}\right)\gamma_i + \varepsilon_{it}, \quad \varepsilon_{it} \sim iid\, N\left(0, \sigma_i^2\right), \forall t$$

The classic "homogeneity assumption" in panel data analysis is the $\theta_i \equiv (\beta_i, \gamma_i, \sigma_i^2)$ of all N cross sectional units are equal, after one allows for cross section specific fixed or random effects α_i. This assumption is absolutely necessary and non-testable for the analysis of panel data with very small T. With panels of moderately large T, this assumption can be tested by the usual tests of parameter restrictions (see Baltagi, 1995, Chapter 4). The null and the alternative of this test are:

$$\begin{aligned} H_0 &: \theta_1 = \theta_2 = ... = \theta_N \\ H_1 &: \text{at least one of the above does not hold} \end{aligned}$$

This paper studies cases where the null is rejected. We only consider the fixed effects case, i.e. the case where α_i, $i = 1, ..., N$ are fixed cross section specific constants. The reason for this is that when the time dimension is moderately large, the fixed effect assumption is not too costly. With this assumption, the analysis can be performed on the mean subtracted data (i.e. $z_{it} - \overline{z_{i.}}$), and the constant terms can be ignored. For simplicity, we assume that there is no contemporaneous correlation between ε_{it} and ε_{jt} for $i \neq j$. This, of course, does not mean that the dependent variables in different cross sectional units are not subject to common shocks (for example world-wide shocks when the cross sectional units are countries, or industry-wide shocks when they are sectors of an industry); it only means that such shocks influence the dependent variables solely through the channel of the independent variables.

The rejection of homogeneity does not mean that all θ_i are necessarily different from each other. Our objective is to partition the N cross sectional units into m sub-groups of homogeneous units, so that $\theta_i = \theta_j$ if i and j belong to the same group, and $\theta_i \neq \theta_j$ if i and j belong to different groups. The number of possible ways that N units could be partitioned into m non-overlapping blocks is called the Sterling number of the second kind[4] which can be calculated from the recursion:

$$\begin{aligned} S(N,m) &= S(N-1,m-1) + mS(N-1,m) \\ S(1,1) &= 1 \\ S(N,N) &= 1 \end{aligned}$$

Hence, the partitioning the 18 OECD countries into two non-overlapping blocks would require consideration of 131,071 possible configurations, and this number would go up to 5.63×10^{14} for the 50 states of the United States. If we want all possible ways that N objects can be partitioned into non-overlapping blocks, we need to sum the Sterling numbers over m, which will give us the so called Bell number of order N, which can be computed through the following recursion:

$$\begin{aligned} B(N) &= 1 + \sum_{m=1}^{N-1} \binom{N-1}{m} B(m) \\ B(1) &= 1 \end{aligned}$$

[4] See the Mathematics Forum web page at http://forum.swarthmore.edu.

This leads to 6.82×10^{11} possibilities for the 18 OECD countries, or 1.86×10^{47} possibilities for the 50 states of the United States! Clearly, it is not feasible to search over all possible partitions and choose all partitions which have homogeneous blocks. Our solution is to use a hierarchical agglomerative clustering algorithm with some additional requirements for linking the clusters, rather than just minimizing a distance measure. These requirements, which are explained below, can potentially reduce the number of possible clusters to a manageable number.

2.2. The clustering algorithm

A clustering algorithm is based on a distance measure, and starts clustering the two objects which are "closest" to each other, and continues to do so until a desired number of clusters are formed, or another stopping criterion is reached. In our problem, the natural distance measure between clusters i and j, denoted by $d(C_i, C_j)$, would be the likelihood ratio test statistic (or the Chow (1960) test statistic) that the parameters of the two clusters are equal. The natural stopping rule would be to stop when there are no two blocks whose distance is less than the critical value of that test.

An additional plausible requirement, which we call sub-group consistency, ensures that the formation of clusters is independent of irrelevant alternatives:

Definition 2.1. *A block of cross-sectional units for which the homogeneity restriction is not rejected is* sub-group consistent *if the homogeneity restriction is not rejected for any arbitrary subset of the cross sectional units in that block.*

This requirement ensures that the pooling of the cross sectional units i and j in one block is independent of whether k is in that block. We think this is important, because if two researchers work on the same data, but one of them does not include the cross section unit k in the analysis, the one with smaller set of cross sectional units should not be surprised to see that the other one has pooled i and j together. This requirement leads us to consider hierarchical procedures which start with the fusion of two units, and only search for a possible third unit to be added to the group among those units that are close enough to both units in

that block. In that sense, our procedure is a complete linkage clustering algorithm (Everitt 1980).

This partial pooling procedure does not take into account the two main concerns that applied researchers have about heterogeneous models, namely that the parameters are imprecisely estimated and have the wrong signs. If partial pooling is based on the magnitude of the above test statistic only, the clustering algorithm may pool a group of cross sectional units that all happen to have imprecisely estimated parameters which have theoretically incorrect sign. Moreover, the between-variation of the right hand side variables of these cross sectional units may decrease standard errors of the estimated parameters of that block, and we may end up with a block with parameter estimates which seem to be precisely estimated, but have the wrong signs. In order to avoid this, we use an additional measure which leads the clustering algorithm to select, from all possible blocks that are close enough, one that leads to precisely estimated parameter estimates which accord with the researchers priors. In other words, we ask our clustering algorithm to move in a direction which reduces the "complexity" of the overall system. In the next sub-section we review how complexity is measured in stochastic systems, and we adopt a measure which explicitly accounts for priors about the parameters and their precision.

2.3. An ad hoc[5] measure of complexity

There is no agreement among different disciplines on what "complexity" means, if it can be measured, and how. The sense that we use the term here is closest to the statisticians' notion of "stochastic complexity" (see Rissanen 1987), which quantifies how complicated a conditional model is, and how unstructured the dependent variable is given the model[6]. One might initially propose the traditional information criteria such as Akaike or the Schwarz criteria as suitable candidates for measuring complexity, at least for linear models. However, such measures do not use information about what the parameters are supposed to measure.

[5] ad hoc: for the particular end or purpose at hand and without reference to wider application or employment (Webster dictionary).

[6] Notice that this is quite different from the notion of complexity in deterministic nonlinear dynamic systems. There, complexity reflects an intricate structure, which can predict the dependent variable perfectly. Here complexity reflects a confused structure, which cannot predict the dependent variable well.

For example, suppose two independent researchers are presenting the following two models (M1 and M2) as their estimated *demand* functions for the same commodity based on the same data,

$$\begin{array}{lccccc} \ln Q_t = & 2.02 & +0.50 & \ln RY_t & -0.01 & \ln RP_t + \hat{u}_t \\ & (.40) & (.25) & & (.10) & \end{array} \tag{M1}$$

$$\begin{array}{lccc} \ln Q_t = & 2.00 & +0.51 & \ln RY_t + \hat{u}_t \\ & (.40) & (.20) & \end{array} \tag{M2}$$

where Q, RY and RP denote quantity demanded, real income and relative prices respectively. The information criteria unanimously prefer M2 over M1, the price elasticity is clearly not significantly different from zero, and the omission of relative prices does not change the estimate of income elasticity significantly. However, M2 creates substantial confusion if presented to a panel of economists as a *demand* function. The presenter of M2 is likely to be asked more questions and is likely to need more time to describe and justify her model than the presenter of M1. A good measure of complexity should be able to reflect the idea that in *the context of demand functions*, M2 is more complex (or confusing) than M1 is. The analogy of complexity to the amount of description needed to present a model to a learned audience, motivates us to look at measures of complexity used in coding theory.

In coding theory, complexity is defined as the length of the shortest uniquely decipherable code that is needed to encode a data sequence. The two measures of complexity which have been developed on the basis of this principle are the minimum description length (MDL) proposed by Rissanen (1987) and the minimum message length (MML) developed by Wallace and Freeman (1987). For the application in this paper, the two measures are quite similar, and in what follows we provide a brief description of the MML measure.

For a discrete random variable Y which has n possible outcomes $(y_1, ..., y_n)$, the length of the most efficient code needed to code the value y_i is approximately[7] $-\log(\Pr(Y = y_i))$,

[7] The approximation relates to cases where $-\log(p_i)$ is not an integer. However, Shannon (1948) proves the existence of a uniquely decipherable code whose average length is at most one unit larger than $\sum_{i=1}^{n} -p_i \log(p_i)$, which is the average length of the most efficient code if fractional coding was possible (a code theoretic definition for the entropy of Y). Hence $-\log(p_i)$ is taken as the approximate code length, even if it is non-integer.

where the base of logarithm is the number of distinct symbols in the coding alphabet[8]. Conditional on some explanatory variables X, and within the class of parametric conditional models indexed by the parameter vector θ, the length of the most efficient code for encoding a sequence of observed values of Y given an estimate for θ, (say $\hat{\theta}$), will therefore be $-\sum_{i=1}^{n} \log f\left(y_i|X,\hat{\theta}\right)$. However, the parameters are unknown and are estimated from the sample, and they have to be coded as well. For simplicity assume that there is only one parameter. Assuming that the prior beliefs about θ is reflected by the density $h(\theta)$, the most efficient code for transmitting $\hat{\theta}$ to the precision δ (the smaller the δ the higher the precision), will be $-\log\left(h\left(\hat{\theta}\right)\delta\right)$. The precision level δ plays an important role, especially when the parameter space is not discrete, in which case, the coding of a parameter with infinite precision (i.e. $\delta = 0$) would require infinite code length. For a given precision δ, the average length of a code required for encoding the data given the model will be $\frac{1}{\delta}\int_{\hat{\theta}-\delta/2}^{\hat{\theta}+\delta/2} -\sum_{i=1}^{n} \log f\left(y_i|X,\theta\right) d\theta$. One can derive the optimal precision analytically (i.e. concentrate δ out of the message length formula), and use this to find the minimum message length required for encoding the parameters and the data, given the parameters. Following this logic, and after some simplifying approximations, for a multiple regression with k parameters, the MML criterion turns out to be[9]:

$$MML = -\sum_{i=1}^{n} \log f\left(y_i|X,\hat{\theta}\right) - \log\left(h\left(\hat{\theta}\right)\right) + \frac{1}{2}\log\left|\mathcal{I}_n\left(\hat{\theta}\right)\right| + \frac{k}{2} + \frac{k}{2}\log\kappa_k$$

where $\mathcal{I}_n\left(\hat{\theta}\right)$ is the empirical Fisher-information matrix[10] and κ_k is a constant, decreasing function of k.

How appropriate is the MML criterion for leading our clustering algorithm? The answer is that although it is an improvement over the usual information criteria, it is not ideal. Its advantage over AIC is the incorporation of prior beliefs. MML uses the priors of the sender and receiver to minimize the code length. This is another way of saying that when the estimated parameters are highly unusual given the theoretical priors of a learned audience[11],

[8] For binary codes, all logarithms will be in base 2 and the code length will be in "bits". Here, in order to compare this measure with information criteria, we choose natural logarithms.

[9] See Wallace and Freeman (1987) and Baxter and Dowe (1996) for details.

[10] Specifically, by empirical Fisher-information matrix we mean $\mathcal{I}_n\left(\hat{\theta}\right) = -\sum_{i=1}^{n} \frac{\partial^2 \log f\left(y_i|X,\hat{\theta}\right)}{\partial\theta\partial\theta'}$.

[11] By "theoretical priors" I mean restrictions implied by economic theory and generally expected by an

it takes a long discussion to justify the model. However, the MML criterion depends on the units of the regressors X, and it can be arbitrarily manipulated by changing the units of regressors. If the prior distribution is a proper prior, and the researcher is aware of the changes in the unit of measurement of X, then the prior will change accordingly and MML will stay the same. However, if we use MML to choose between competing models with qualitatively different regressors, it is easy to see that, given similar fits, the MML will choose the regressors with higher degree of collinearity. This is exactly the opposite to what we would like. This is not so much a flaw in the principle of MML, but rather a function of the assumptions made to derive the optimal precision at which the parameters are transmitted.

We remedy the above undesirable property of MML, by defining MML as the length of the shortest uniquely decodable code required to transmit the *standardized* coefficients $\psi = \Psi^{\frac{1}{2}}\theta$, $(\Psi = \text{diag}\left(E\left(\mathcal{I}_n^{-1}(\theta)/n\right)\right))$, and the model given the parameters. The prior used for transmitting the parameters is also the prior over the standardized coefficients. When the priors have finite variance MML will be invariant to such change in variables, but when the priors are improper (diffuse priors, or priors on the signs of the parameters only, as often is the case in economics), this change will make a significant difference. The modified MML will have the form:

$$\begin{aligned} MML^* &= -\sum_{i=1}^{n}\log f\left(y_i|X,\hat{\psi}\right) - \log\left(h\left(\hat{\psi}\right)\right) + \frac{1}{2}\log\left|\mathcal{I}_n\left(\hat{\psi}\right)\right| + \frac{k}{2} + \frac{k}{2}\log\kappa_k \\ &= -\sum_{i=1}^{n}\log f\left(y_i|X,\hat{\theta}\right) - \log\left(h\left(\hat{\psi}\right)\right) + \frac{1}{2}\sum_{j=1}^{k}\log\hat{\Psi}_{jj} + \frac{1}{2}\log\left|\mathcal{I}_n\left(\hat{\theta}\right)\right| + \frac{k}{2}\left(1+\log\kappa_k+\log n\right) \end{aligned}$$

In this form, MML has an interesting interpretation. The third and the fourth terms can be written as:

$$\begin{aligned} \frac{1}{2}\sum_{j=1}^{k}\log\hat{\Psi}_{jj} + \frac{1}{2}\log\left|\mathcal{I}_n\left(\hat{\theta}\right)\right| &= \frac{1}{2}\sum_{j=1}^{k}\log\hat{\Psi}_{jj} - \frac{1}{2}\log\left|\mathcal{I}_n^{-1}\left(\hat{\theta}\right)\right| \\ &= \sum_{j=1}^{k}H\left(\hat{\theta}_j\right) - H\left(\hat{\theta}_1,\hat{\theta}_2,...,\hat{\theta}_k\right) \end{aligned}$$

where H is the entropy, assuming that $\left(\hat{\theta}_1,\hat{\theta}_2,...,\hat{\theta}_k\right)$ are normally distributed. The difference between the sum of individual entropies and the joint entropy is a measure of dependence of

audience of economists. These would generally involve simple sign restrictions.

random variables. This difference is always non-negative (Shannon (1948)), and it is equal to zero if and only if $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k$ are independent. Van Emden (1975) used this difference as a measure of complexity of a covariance structure (Van Emden 1975, page 61). In the regression context, it can be seen as a penalty for the redundancy of the regressors (i.e. multicollinearity), and it is minimized when regressors are uncorrelated with each other. This property is highly desirable for our purpose, and MML^* is therefore the model selection criterion that we choose to direct our clustering algorithm.

We end this section by comparing MML^* with another model selection criterion developed on the basis of Van Emden's measure of covariance complexity. Bozdogan (1990) has introduced the following measure of informational complexity:

$$ICOMP(IFIM) = -\sum_{i=1}^{n} \log f\left(y_i|X, \hat{\theta}\right) + \frac{k}{2} \log\left(\text{trace}\left(\mathcal{I}^{-1}\left(\hat{\theta}\right)\right)/k\right) + \frac{1}{2} \log\left|\mathcal{I}\left(\hat{\theta}\right)\right|,$$

which can be expressed alternatively as:

$$ICOMP(IFIM) = -\sum_{i=1}^{n} \log f\left(y_i|X, \hat{\theta}\right) + \frac{k}{2} \log \frac{\bar{\lambda}_a}{\bar{\lambda}_g},$$

where $\bar{\lambda}_a$ and $\bar{\lambda}_g$ are the arithmetic average and the geometric average of the eigenvalues of $\mathcal{I}^{-1}\left(\hat{\theta}\right)$. In this criterion, $\frac{k}{2} \log\left(\text{trace}\left(\mathcal{I}^{-1}\left(\hat{\theta}\right)\right)/k\right)$ replaces $\frac{1}{2}\sum_{j=1}^{k} \log \hat{\Psi}_{jj}$. Van Emden (1975) introduces this substitution in order to make his measure of covariance complexity invariant to orthonormal transformations of the parameter vector. This does not create a concern in our context, because our procedure is designed for applications in which parameters have specific interpretations. Although their units may be arbitrary, arbitrary orthonormal transformations of them are of no interest. Note that $ICOMP$ is not invariant to changes in units of only one or a subset of parameters, whereas MML^* is. Table 1 compares the penalty assigned by AIC, Schwarz, Hannan-Quinn, MML, MML* and ICOMP in three situations: (i) when all regressors are independent; (ii) same as (i) but a quarter of the regressors are multiplied by 100; and (iii) when regressors are highly correlated ($X_j = 1.2X_1 + 0.5\varepsilon_j$, $j = 2, ..., k$, where $X_1, \varepsilon_2, ..., \varepsilon_k$ are *i.i.d.*). In each case, the penalty is shown for 8, 16, 24 parameters and 50, 150 and 250 observations.

Table 1: Penalty assigned by different model selection criteria

in a linear regression when regressors are independent and have equal variances

n	AIC	SC	HQ	$ICOMP$	MML	MML^*
Independent regressors with equal variances, $k = 8$						
50	8	15.65	10.91	1.46	0.51	10.53
150	8	20.04	12.89	1.15	5.24	14.65
250	8	22.09	13.67	1.14	7.27	16.68
Independent regressors with equal variances, $k = 16$						
50	16	31.30	21.82	3.17	−1.30	20.84
150	16	40.09	25.79	1.80	8.47	28.32
250	16	44.17	27.34	1.64	12.66	32.25
Independent regressors with equal variances, $k = 24$						
50	24	46.94	32.74	6.07	−4.55	32.08
150	24	60.13	38.68	2.91	10.95	42.30
250	24	66.26	41.01	2.33	17.30	47.87

Table 1 (continued): The case where regressors are independent but do not have equal variances

n	AIC	SC	HQ	$ICOMP$	MML	MML^*
Independent regressors with unequal variances, $k = 8$						
50	8	15.65	10.91	9.74	9.72	10.53
150	8	20.04	12.89	9.25	14.45	14.65
250	8	22.09	13.67	9.24	16.48	16.68
Independent regressors with unequal variances, $k = 16$						
50	16	31.30	21.82	19.37	17.12	20.84
150	16	40.09	25.79	18.02	26.89	28.32
250	16	44.17	27.34	17.84	31.08	32.25
Independent regressors with unequal variances, $k = 24$						
50	24	46.94	32.74	30.55	23.08	32.08
150	24	60.13	38.68	27.31	38.58	42.30
250	24	66.26	41.01	26.66	44.93	47.87

Table 1 (continued): The case of correlated regressors

n	AIC	SC	HQ	$ICOMP$	MML	MML^*
Correlated regressors, $k = 8$						
50	8	15.65	10.91	2.92	−4.35	11.76
150	8	20.04	12.89	2.70	0.38	15.86
250	8	22.09	13.67	2.75	2.41	17.93
Correlated regressors, $k = 16$						
50	16	31.30	21.82	4.81	−11.70	22.04
150	16	40.09	25.79	3.58	−1.93	29.63
250	16	44.17	27.34	3.38	2.26	33.55
Correlated regressors, $k = 24$						
50	24	46.94	32.74	7.80	−20.49	33.40
150	24	60.13	38.68	4.56	−4.99	43.56
250	24	66.26	41.01	4.09	1.36	49.16

This table shows the problem of MML with correlated regressors very clearly (MML actually favors overparameterized models in this case). One can also see the sensitivity of $ICOMP$ to change in the units of a subset of regressors. Table 1 also shows that the penalty assigned by MML^* is roughly between the Hannan-Quinn penalty, which is $k \ln \ln n$, and the Schwarz penalty, which is $\frac{k}{2} \ln n$. This suggests that MML^* is a consistent order selection criterion.

2.4. The exchange method and other modifications

Even though at each stage, from among all clusters that are within a critical distance of each other, our algorithm fuses two clusters which together form a subgroup consistent cluster and reduce the MML^* criterion the most, there is still no guarantee that it will lead to the globally optimal partition. Therefore, the algorithm should be amended with the usual enhancements to increase the chance of moving towards the global optimum. Depending on the size of the problem, these modifications might range from starting the clustering algorithm by fusing the second best options at the initial stage, to allowing for some possibility of non-optimal choice (like a mutation in a genetic algorithm) at each stage, and comparing the MML^* of the resulting final configurations.

One enhancement which is easy to implement, would be to add a so called exchange algorithm at the end, to check that the final configuration cannot be improved upon by moving one cross sectional unit from one partition to another, while preserving the sub-group consistency criterion. Suppose that the hierarchical algorithm partitions the n cross sectional units into $m\,(< n)$ groups and assigns cross sectional unit s to group j. An exchange method simply reevaluates MML^* using $\hat{\theta}_i (i = 1, ..., m; i \neq j)$ for the cross sectional unit s, and if the objective function is improved, s is moved from cluster j to cluster i. After all such exchanges have taken place, the parameters of the new partitions are re-estimated. Since re-estimation can only reduce the sum of squared errors of each cluster, the MML^* is guaranteed to improve at each round of exchange. The exchange method stops when no more MML^* improving exchanges are possible.

3. Empirical Example

Baltagi and Griffin (1983, 1997) consider the following model of gasoline demand for 18 OECD countries:

$$\ln\left(\frac{GAS}{CAR}\right)_{it} = \alpha_i + \beta_1^m \ln\left(\frac{GNP}{POP}\right)_{it} + \beta_2^m \ln\left(\frac{CAR}{POP}\right)_{it} + \beta_3^m \ln\left(\frac{P_{GAS}}{P_{GNP}}\right)_{it} + \beta_4^m \ln\left(\frac{GAS}{CAR}\right)_{it-1} + \beta_5^m \ln\left(\frac{GNP}{POP}\right)_{it-1} + \beta_6^m \ln\left(\frac{CAR}{POP}\right)_{it-1} + \varepsilon_{it}$$

where $\left(\frac{GAS}{CAR}\right)$ is the gasoline consumption per auto, $\left(\frac{GNP}{POP}\right)$ is the real income per-capita, $\left(\frac{CAR}{POP}\right)$ is cars per-capita and $\left(\frac{P_{GAS}}{P_{GNP}}\right)$ is the relative price of gasoline. The parameter α_i is the country specific fixed effect, and β_1^m to β_6^m are elasticities which are the same for all countries in group m.

Baltagi and Griffin consider only the two polar cases; the pooled case in which there is only one group which includes all countries, and the fully heterogeneous case in which each of the eighteen OECD countries is in a separate group. They estimate their model using data from 1960 to 1980, and use it to generate conditional forecasts for 1981 to 1990. They find that the pooled model leads to better forecasts than the individual country models, even though the pooling restriction is strongly rejected.

We ran their 1960 to 1980 data through the clustering algorithm explained in the previous section. The prior we used in the MML^* criterion was Baltagi and Griffin's (1997) prior that the long-run income and price elasticities should be positive and negative respectively. The algorithm produced the following final partition:

Group 1: US, Japan, Denmark, Ireland, Italy, Netherlands.

Group 2: Belgium, Spain, Sweden, Switzerland, Turkey.

Group 3: Austria, Germany, Norway, UK.

Group 4: Canada, France.

Group 5: Greece.

The resulting parameter estimates are included in Table 2. Had subgroup consistency not been a requirement, combining the first and the third groups would have been an MML^* enhancing fusion, with no possible move after that. The closeness of the first and third groups is also apparent from the elasticity estimates in Table 2.

Table 2: Parameter estimates for groups

(standard errors in parentheses)

	β_1	β_2	β_3	β_4	β_5	β_6
Group 1	0.289 (0.15)	−0.625 (0.13)	−0.119 (0.03)	0.789 (0.04)	−0.260 (0.15)	0.487 (0.13)
Group 2	0.192 (0.20)	−1.12 (0.08)	−0.286 (0.03)	0.461 (0.07)	−0.050 (0.20)	0.812 (0.08)
Group 3	0.248 (0.13)	−0.521 (0.19)	−0.186 (0.04)	0.634 (0.07)	−0.137 (0.13)	0.378 (0.16)
Group 4	0.456 (0.17)	−1.13 (0.14)	−0.276 (0.04)	0.728 (0.07)	−0.249 (0.17)	0.855 (0.13)
Group 5	0.571 (0.57)	−0.819 (0.25)	−0.402 (0.14)	−0.388 (0.20)	0.934 (0.71)	−0.155 (0.32)

Using these elasticity estimates, we produced conditional forecasts for gasoline demand for the years 1981 to 1990. The resulting average of the root mean squared forecast errors for all OECD countries is included in Table 3. This table also includes the corresponding measure for the forecasts based on the pooled (within) estimates of elasticities.

Table 3: Grouped vs. pooled[12] average of root mean squared forecast errors for all OECD countries

Year	Grouped	Pooled
1981	0.034	0.035
1982	0.045	0.046
1983	0.057	0.069
1984	0.063	0.075
1985	0.092	0.098
1986	0.098	0.107
1987	0.123	0.135
1988	0.140	0.155
1989	0.156	0.174
1990	0.177	0.189
10 year Average	0.0985	0.108

[12] The figures reported in the third column are different from those reported in Baltagi and Griffin (1997). They have made some adjustments to their forecasts that are not fully explained in their paper, and hence cannot be reproduced.

The table shows that for this particular model, there is only a very slight improvement of the out of sample performance of the grouped model over the pooled model. However, there is much more information in the grouping suggested by the data than in the two extreme cases of pooling all countries, or not pooling at all.

4. Related Literature and An Alternative Procedure

There is an extensive literature on classification and discrimination in statistics. However, most of this literature is concerned with the assignment of a new observation to one of k populations, given a training sample (see Anderson 1984, chapter 6). There is also a literature in time series analysis on the classification of ARMA processes (see Shumway 1982 for a survey). However, the problem of clustering time-series into an *unknown* number of groups, according to *estimated* parameters has not been studied. Our problem is more difficult; we do not know the number of possible classes, and we want to group cross sectional units according to the closeness of coefficients estimated from individual unit time-series data. Our solution is basically a hierarchical clustering algorithm (Everitt 1980). Previous applications of clustering procedures in economics include Hirschberg, et al (1991) who use them to develop a low dimensional multivariate quality of life index, and Haruvy (1997) who uses them to explore the levels of rationality of subjects in the analysis of experimental data.

An alternative to the clustering algorithm used in this paper could be a suitably modified variant of the classification and regression trees (CART) procedure (Breiman et al. 1984). While this procedure is designed for uncovering threshold nonlinearities in the conditional expectation function of cross sectional data, its fundamentals can be used to design a procedure for clustering cross sectional units in a panel. However, it relies on a pre-specified set of threshold variables to group the data. These variables can be a subset of exogenous variables in the model, or they can be given from outside. Therefore, CART is useful only for problems in which there are good reasons to believe that coefficients might be closely related to some specific attributes of the cross sectional units. For example, the coefficients of money demand are often assumed to be similar for "high inflation" countries, and similar for "low inflation" countries (see for example Vogel(1974) or Duck(1993)). Also, the regressions which study the

convergence or non-convergence of per-capita income in different countries are often grouped according to their initial level of income and development. Durlauf and Johnson (1995) use CART to cluster countries into homogenous groups and then they study convergence.

It is interesting to see what cluster arrangement will arise out of an alternative procedure such as CART, and to determine how it will perform in the above forecasting exercise. To use CART, we have to start by specifying a set of threshold variables. Since we have no reason to prefer one of the exogenous variables in the data set over any of the others, we let all three exogenous variables (income per-capita, cars per-capita and relative gasoline price) be possible threshold variables. Given that we have time-series observations on all of these attributes, we have to decide which observation on these attributes should be chosen as the guide for the CART procedure. For lack of any specific reason to do otherwise, we use the average over the estimation period (21 years).

The CART procedure[13] at each level, uses one of the threshold variables to split the observations of a cluster into two groups. All possible binary splits according to all specified threshold variables, and all possible threshold values are examined. The split with the smallest sum of squared errors (SSE) is chosen and the algorithm is repeated. With cross sectional data sets, these splits have to stop when there are just enough observations in each group to estimate the coefficients. However, in our situation, the coefficients for each unit can be estimated from the time-series dimension of that unit, and therefore each of the final nodes will contain only a single unit. Even though it is clear from the outset that the final split will lead to a fully heterogeneous model, it is the path to this final configuration, or the "tree", that matters, because in the next stage, this determines what can be "pruned". "Pruning", which is merging of the units that are the result of a former split, will allow a branch with a late split with a large effect on SSE to survive, while other branch with successive splits that have less impact on SSE to be pruned back. In the pruning stage, a penalty for the number of parameters is introduced, and the pruning stops when the increase in the SSE is no longer offset by the decrease in the penalty.

We have used the AIC criterion to guide us when to stop pruning. When there were only

[13] For a concise and clear explanation of CART, refer to the Technical Appendix in Durlauf and Johnson (1995).

two or three countries remaining in each group, the threshold variable could not be identified (i.e. sorting of countries based on any of the three attributes resulted in the same order), but the initial splits were all based on either income per-capita or relative price. The resulting final configuration was:

Group 1: Japan, France, Ireland, Italy, Netherlands, UK.

Group 2: Belgium, Norway.

Group 3: Austria, Denmark, Germany.

Group 4: US, Canada, Sweden, Switzerland.

Group 5: Spain, Turkey.

Group 6: Greece.

If the penalty for overparameterization is increased, groups 1 and 2 are merged first, then they are joined by 3, then by 4, and so on until we get back to the fully homogeneous model. This algorithm also chooses Greece as a clear outlier. However, since CART does not use information on the signs of actual parameters as a guide, two of the above groups, (groups 3 and 5), produce estimates with the "wrong" signs. The average mean squared error of the conditional forecasts produced from this configuration is 0.117, which is larger than the averages of the fully homogenous model and the partially pooled model reported in Table 3.

The purpose of the above comparison is not to show which algorithm is universally "better". After all, the aim of this paper is to argue that partial pooling should be taken seriously, and to suggest an algorithm which seems plausible for the specific applied problems which motivated this paper. The above comparison simply presents an alternative that might be useful under different circumstances. For example, if there was, in our data set, a measure which reflected the general attitude of the people of each country about energy conservation, then CART might have been a better algorithm than another which does not use this information. Different aspects of each algorithm can be mixed to produce better algorithms for other situations.

5. Summary and suggestions for further research

This paper studies the use of clustering methods for the partial pooling of cross sectional units in a medium sized panel. Here, clustering is based on the closeness of conditional expectation functions which, are assumed to remain constant over time for each cross sectional unit. This assumption will not always be plausible. For instance, the production structure of rapidly industrializing countries will be similar to that of the non-industrial countries during the earlier years of the sample, and then similar to the industrial countries later on. Future work should allow for one or more group switches over time, for every cross sectional unit.

It is well known that the final outcome of any hierarchical clustering algorithm is sensitive to its initial conditions. In this paper we suggest a global objective function, the modified message length criterion, which allows us to choose the best of all different partitions that the clustering algorithm may lead to. We choose this particular objective function because it can easily accommodate prior beliefs, and because it penalizes imprecision in the parameter estimates as well as lack of parsimony. Little work has been done on the properties of this criterion (see Baxter and Dowe, 1996). Future research should compare the performance of this criterion against other model selection criteria.

Since our clustering algorithm uses beliefs about the coefficients to influence the formation of groups, it is natural to think about using a fully Bayesian approach. However, given the enormous number of possible partitions, a fully Bayesian approach to this problem may become unwieldy. With uninformative prior on the possible partitions, and with prior only about the signs of coefficients, a fully Bayesian approach might not lead to a different result, given that there is at least one partition (i.e. the pooled regression) in which the estimated coefficients have the correct sign. Of course this is only a speculation, and the question deserves further research.

Our procedure is motivated by empirical studies in which researchers have no a-priori reason to cluster the cross sectional units according to any specific variable. In some situations, there is a good reason to believe that one can group cross sectional units based on a single attribute. For example, countries are often grouped in high inflation and low inflation groups in money demand estimation, or grouped in low, medium and high income groups in

growth analysis. It is easy to incorporate this into our algorithm, by fusing, among all units whose distance is less than the critical value, those units which are closer to each other in some specified attribute. The partial pooling procedure, can also be used to check if such a-priori groupings are in fact suggested by the data. Alternatively, one can start from the grouping suggested by the algorithm in this paper, and investigate which particular attribute, or combination of attributes, can best determine which group a cross sectional unit belongs to.

In the empirical example of this paper, we clustered the OECD countries according to similarities in their short-run as well as their long-run elasticities. However, economic theory usually only provides good reasons to expect that long-run elasticities might be the same. Future research should investigate the partial pooling of the cross sectional units based on similarities in their long-run elasticities, while allowing the short-run elasticities to be different. Such an analysis would no longer require the lag specification for all cross sectional units to be the same, and this would allow considerable flexibility in the specification of the dynamics.

References

Ahn, S. K. and G. C. Reinsel (1988), "Nested Reduced Rank Autoregressive Models for Multiple Time Series", *Journal of the American Statistical Association*, 83, 849-856.

Anderson, H. M. and F. Vahid (1998), "Testing Multiple Equation Systems for Common Nonlinear Components", *Journal of Econometrics*, 84, 1, 1-36.

Anderson, T. W. (1984), *An Introduction to Multivariate Statistical Analysis*, 2nd Edition, Wiley.

Baltagi, B. H. (1995), *Econometric Analysis of Panel Data*, Wiley.

Baltagi, B. H. and J. M. Griffin (1983), "Gasoline Demand in the OECD: An Application of Pooling and Testing Procedures", *European Economic Review*, 22, 117-137.

Baltagi, B. H. and J. M. Griffin (1997), "Pooled Estimators vs. their Heterogeneous Counterparts in the Context of Dynamic Demand for Gasoline", *Journal of Econometrics*, 77, 303-327.

Baxter, R. A. and D. L. Dowe (1996), "Model Selection in Linear Regression Using the MML Criterion", Technical Report 96/276, Department of Computer Science, Monash University.

Bearse, P. M., H. Bozdogan and A. M. Schlottmann (1997), "Empirical Econometric Modelling of Food Consumption using a New Informational Complexity Approach", *Journal of Applied Econometrics*, 12, 563-592.

Bozdogan, H. (1990), "On the Information Based Measure of Covariance Complexity and its Application to the Evaluation of Multivariate Linear Models", *Communications in Statistics, Theory and Methods*, 19, 221-278.

Breiman, L., J. L. Friedman, R. A. Olshen and C. J. Stone (1984), *Classification and Regression Trees*, Wadsworth.

Burnside, C. (1996), "Production Function Regressions, Returns to Scale, and Externalities", *Journal of Monetary Economics*, 37, 177-201.

Chow, G. C. (1960), "Tests of Equality between Sets of Coefficients in Two Linear Regressions", *Econometrica*, 28, 591-605.

Duck, N. (1993), "Some International Evidence on the Quantity Theory of Money", *Journal of Money, Credit and Banking*, 25, 1-12.

Durlauf, S. N. and P. A. Johnson (1995), "Multiple Regimes and Cross-Country Growth Behaviour", *Journal of Applied Econometrics*, 10, 365-384.

Engle, R. F. and C. W. J. Granger (1987), "Cointegration and Error Correction: Representation, Estimation and Testing", *Econometrica*, 55, 251-276.

Everitt, B. (1980), *Cluster Analysis*, Second Edition, Wiley: New York.

Haruvy, E. (1997), "Testing Modes in the Population Distribution of Beliefs from Experimental Games", mimeo, Department of Economics, University of Texas at Austin.

Hirschberg, J. G., E. Maasoumi and D. J. Slottje (1991), "Cluster Analysis for Measuring Welfare and Quality of Life across Countries", *Journal of Econometrics*, 50, 131-150.

Maddala, G. S., R. P. Trost, H. Li and F. Joutz (1997), "Estimation of Short-Run and Long-Run Elasticities of Energy Demand From Panel Data Using Shrinkage Estimators", *Journal of Business and Economic Statistics*, 15, 90-100.

Ramanathan, R., R. F. Engle, C. W. J. Granger, F. Vahid and C. Brace (1997), "Short-run Forecasts of Electricity Loads and Peaks", *International Journal of Forecasting*, 13, 161-174.

Rissanen, J. (1987), "Stochastic Complexity", *Journal of the Royal Statistical Society, Series B*, 49, 223-239.

Shannon, C. E. (1948), "A Mathematical Theory of Communication", *Bell System Technical Journal*, 27, 379-423.

Shumway, R. H. (1982), "Discriminant Analysis for Time Series", in P. R. Krishnaiah and L. N. Kanal eds, *Handbook of Statistics*, Vol. 2, North Holland.

Tiao, G. C. and R. S. Tsay (1989), "Model Specification in Multivariate Time Series", *Journal of the Royal Statistical Society, Series B*, 51, 157-213.

Vahid, F. and R. F. Engle (1993), "Common Trends and Common Cycles", *Journal of Applied Econometrics*, 8, 341-360.

Vahid, F. and R. F. Engle (1997), "Codependent Cycles", *Journal of Econometrics*, 80, 199-221.

Van Emden, M. H. (1975), *An Analysis of Complexity*, Second Printing, Mathematical Center Tracts: Amsterdam.

Vogel, R. (1974), "The Dynamics of Inflation in Latin America", *American Economic Review*, 64, 102-114.

Wallace, C. S. and P. R. Freeman (1987), "Estimation and Inference by Compact Coding", *Journal of the Royal Statistical Society, Series B*, 49, 240-265.