Detecting Multiple Mean Breaks At Unknown Points With Atheoretical Regression Trees

[1]Cappelli, C., [2]R.N. Penny and [3]M. Reale

[1]University of Naples "*Federico II*", [2]Statistics New Zealand, [3]University of Canterbury
E-Mail: Richard.Penny@stats.govt.nz

Keywords: *Partitioning, official statistics, X-12-ARIMA, time series.*

EXTENDED ABSTRACT

In this paper we propose a computationally effective approach to detect multiple structural breaks in the mean occurring at unknown dates. We propose a non-parametric approach that exploits, in the framework of least squares regression trees, the contiguity property of the Fisher grouping method (1958) proposed for grouping a single real variable. The proposed approach is applied to study the possibility of using the series of anomalous observation C17 provided by the seasonal adjustment procedure implemented in X12-ARIMA.

1. INTRODUCTION

The detection of structural breaks is challenging and lot of effort has been devoted to this task both in the statistic and econometric literature (for a review see Hansen, 2001). In this paper we focus on the problem of detecting multiple breaks in the mean occurring at unknown dates. To this aim we propose a nonparametric approach based on regression trees. Given a continuous response variable *Y* and a set of *p* predictors $X\{1\},\mathrm{K}\ ,X\{p\}$, regression trees model the relationship between the response and the covariates employing a recursive partitioning approach that results into a partition of *Y* based upon the values of the predictor variables. Our procedure makes use of an artificial covariate (so that *p*=1) that is an arbitrary strictly ascending (or descending) sequence of numbers thus we call it Atheorethical Regression Trees (so forth denoted by ART) (Cappelli and Reale, 2005) because it is theory-free being the covariate not a predictor variable but rather a counter. In what follows we will show that the use of such covariate in least square regression trees (Breiman *et al*., 1984) resorts to a sequential use of the Fisher's method of exact optimization (1958) proposed for grouping *n* elements into *g* mutually exclusive and exhaustive subsets having maximum homogeneity i.e., minimizing the within-groups sum of squares. Fisher's algorithm is designed for situations in which the data points are ordered and groups consist of intervals of data. Two subclass of problems are considered: the *unrestricted* case when the observations can be ordered according to their numerical values, and the *restricted* one when an *a priori* ordering is given. Time series data belong to the second case as the ordering is provided by the time and observations are not

exchangeable. In this case seeking the minimum within sum of square partition corresponds to segment the series into homogeneous subperiods that contrast with each other i.e., it corresponds to detect breaks in the mean. A drawback of the Fisher's method is that it can deal with moderate-sized values of *n* and *g* while ART overcomes these limitations because it corresponds to a sequential application of the Fisher's algorithm to a problem of *g*=2 subperiods. Furthermore, whereas Fisher's method produces a single partition and it is advisable to create several partitions by varying *g*, ART produces a hierarchical structure. The final partition and the corresponding set of break dates can result either from automatic procedure such as pruning along with popular model selection criteria or from subjective choice of the applied scientist based on *a priori* knowledge.

2. ATHEORETHICAL REGRESSION TREES

In least square regression trees (LSRT) a node *t* is split into the left and right descendants t_1 and t_r to reduce the deviance of the response variable. Thus the algorithm selects the split *s* for which $SS(t)-\left[SS(t_1)+SS(t_r)\right]$ is maximum, where

$$SS(t)=\sum_{y_i\in t}(y_i-\bar{y}(t))\ i=1,\mathrm{K}\ ,n \qquad (1)$$

is the sum of squares for node *t*, and $SS(t_1)$ and $SS(t_r)$ are the sums of squares for the left and right descendants, respectively. The splitting criterion is equivalent to maximize the between-groups sum of squares *BSS*(*t*) that can be written

$$BSS=\frac{n(t_l)n(t_r)}{n(t)}(\bar{y}(t_l)-\bar{y}(t_r))^2 \qquad (2)$$

being

$$\bar{y}(t)=\frac{n(t_l)\bar{y}(t_l)+n(t_r)\bar{y}(t_r)}{n(t)}. \qquad (3)$$

Thus, in LSRT the splitting criterion searches for the child nodes consisting of subsets of *y* values whose means are as far as possible. Once the binary partition of a node is found, the splitting process is applied separately to each subgroup, and so on recursively until the subgroups either reach a minimum size or no improvement of the criterion can be achieved. The resultant tree usually is overly large so that a pruning method is applied to trim it back. Minimizing the within-group sum of squares is a natural criterion for partitioning a single real variable. This is the case in the Fisher's algorithm of exact optimization (1958) whose key aspect it's the concepts of *contiguous partitions*. Let *i*, *e* and *h* be three data points that have order $i\le e\le h$; a partition is said to be contiguous if it consists of groups that satisfy the following condition: if *i* and *h* are assigned to the same class then *e* must be also assigned to that class. For ordered data only contiguous partitions require to be considered to detect the optimal one minimizing the within-group sum of squares. In the restricted case of time series data the contiguity applies to time i.e., only subsequent intervals in terms of the ordering specified by time are admissible. The number of possible contiguous partitions of *n* (whatever) ordered objects into *g* groups makes it is unfeasible a global search but Fisher shows that the number of computations can be substantially

reduced by exploiting the additivity property of the sum of squares criterion by means a dynamic programming approach that allows to deal with the problem of finding the optimal partition into g groups making use of the results obtained while dealing with the problem of g-1 groups.

Despite the saving, Fisher bounders the capacity of the algorithm in $n \leq 200$ and $g \leq 6$ and even with today's computers a complete enumeration and search it is possible only for $g=2$. The concept of contiguous partitions can be naturally exploited in the framework of least square regression trees. At this aim let k be an arbitrary ascending (or descending) sequence of completely ordered numbers, for sake of simplicity take $k=1,2,\ldots,i,\ldots,n$. The use of such sequence as covariate into least square regression trees resorts to create and check at any node t all the admissible binary partitions of the $y_i \in t$ whose number is $n(t)$-1 and thus it is treatable. Indeed, the contiguity property ensures that at any node t the best split for the given order lays in k and it will be identified by the splitting criterion. Note that in the original Fisher's method optimal partitions for different values of g need not to be hierarchically nested. In the ART method as splitting goes on, the previous partitions are fixed, but for many sets of data this is represents a reasonable approximation providing good partitions at a much less expensive computational cost. Indeed, the Fisher's method requires $O(n^2 g)$ steps, whereas ART, at any tree node requires $O(n(t))$ steps to identify the best split. Hartigan (1975) provides an excellent justification in favor of the binary division algorithm in the case of time series data: suppose that the time interval consists of g intervals within each of which the values are constant. Then there is a partition into g segments for which the within sum of squares is zero and it will be identified by the tree algorithm. ART generates a hierarchical structure and a nested sequence of partitions corresponding to candidates sets of break dates can be identified by means of pruning, that is the process of discarding terminal nodes whose contribute to reduction in deviance is negligible. In order to find the subtree whose terminal nodes provide the optimal partition corresponding to the actual number of break dates and distinct subperiods present in the data, we use classical model selection criteria (for discussion on the use of these criteria in tree methods see Su *et al.*, 2003). We also consider atheoretical regression trees in the in the context of *Smooth Transition Regression Trees* (Correa da Rosa *et al.*, 2005) with a splitting criterion based on a Lagrange multiplier stopping rule which is better suited for time series than cross-validation.

3. APPLICATION TO OFFICAL STATISTICS

National Statistical Offices (NSO) collect, collate and publish data for the use of researchers, policy analysts and the general public. The main concern of NSOs is to release data that reflect the social or economic concept that they are meant to represent, within the budget allocated for this work and, crucially, with little or no revisions after release. As an NSO is supplying time series for a range of users with varying needs and knowledge of statistical analysis much of its output will be descriptive, rather than analytical. The core descriptive output will be the time series collated from the data collected.

As the Chief Statistician of the Canadian NSO noted "Credibility plays a basic role in determining the value to users of the special commodity called statistical information. Indeed, few users can validate directly the data released by statistical offices." (Fellegi, 1996, p. 169). To enable users to use the data supplied by an NSO with confidence considerable work inside the NSO is done to check and report on the quality of the data produced. A key issue is to ensure that the series is a consistent throughout its length. By doing this the NSO can assure users that the series is a result of the data generating process, and not the way the NSO has collected, collated and published the data.

Much of the reporting on NSO outputs focuses on the movements in the time series, rather than the values. For any series that is seasonal often the largest part of the movement is caused by changes in the seasonal component. For this reason most statistical agencies provide the measured figure along with the seasonally adjusted value (where appropriate) and, increasingly, the trend estimates, and direct users to the latter series rather than the unadjusted figures.

To seasonally adjust series Statistics New Zealand estimates the unobserved seasonal (S), calendar (TD), trend (C) and irregular (I) components of the time series it releases. While state space modelling will provide estimates of these components many NSO, including Statistics New Zealand, use variants of the Census II Method of the U.S. Bureau of the Census (Shiskin et.al. 1967). Statistics New Zealand currently uses Census Method II Variant X-12, commonly called X-12 (Findley et.al. 1998). X-12 uses Henderson moving averages to decompose the original series into the set of unobserved components. As outliers can cause problems in time series analysis X-12 identifies outliers and records this information in a table, termed the C17 table. For complete details see Ladiray & Quenneville (2001).

An NSO is particularly interested in identifying any atypical changes in the behaviour of the series. These changes in the series can indicate changes in the data generating mechanism, but may arise through unforeseen effects in the data collection and collation process used by the NSO. Any method for identifying the breakpoints usually is applied to the original series, or some stationary series produced by differencing. Given the estimates of the unobserved components produced by X-12, as well as the information on outlier identification we have investigated the applicability of the method to these.

Is the C17 component a useful indicator for possible structural changes of the data generating process?

Our strategy to answer this question is to identify breakpoints in a time series using atheoretical regression trees and verifying if they were indicated as anomalous by the C17 component (values different from 1). In particular we focus our attention on the series of the irregulars, which is computed as residual from the other smoothed components. Monte Carlo simulations have been done to assess the capability of the C17 component to indicate structural changes and the promptness of ART to confirm that an anomalous value is actually a breakpoint. Finally we consider the real case of the *Quarterly Gross Domestic Expenditure* in New Zealand from June 1986 to March 2005. Figure 1 and 2 show a time plot of the original series and the irregular component respectively.

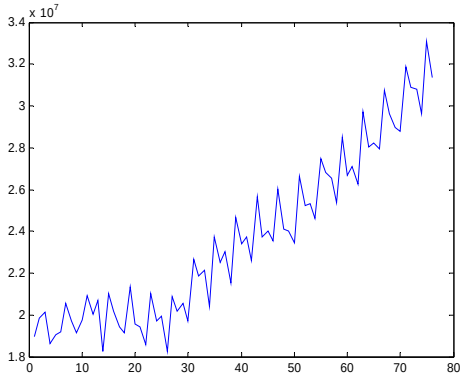

Figure 1: QGDE in NZ (1986-2005)

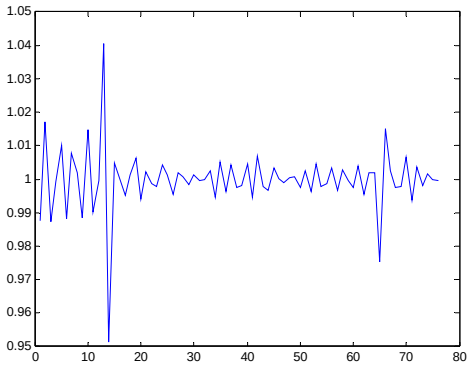

Figure 2: Irregular component of QGDE

The series C17 provided by X-12 contains 5 anomalous observations and ART identifies one of them, corresponding to March 1989 as an outlier.

4. CONCLUSION

Atheoretical regression trees are an effective way to identify structural breaks at unknown time. The use of smooth transition regression trees is better suited for time series. The application shows that the series of anomalous observations C17, provided by X12, provides useful information for the prompt identification of structural breaks.

5. ACKNOWLEDGMENTS

The authors wish to thank Marcelo Medeiros for kindly providing the MATLAB code for smooth transition regression trees.

6. REFERENCES

Breiman, L., J.H. Friedman, R.A. Olshen and C.J. Stone (1984), *Classification and Regression Trees*, Wadsworth & Brooks, 368 pp., Monterey (USA).

Cappelli, C. and M. Reale (2005), Detecting changes in mean with atheoretical regression trees, University of Canterbury Mathematics and Statistics Department Research Report 2005/02, March 2005, pp. 14.

Correa da Rosa, J., A. Veiga and M.C. Medeiros (2005), Tree-structured smooth transition regression models based on CART algorithm, mimeo, pp.35.

Fellegi, I.P. (1996), Characteristics of an Effective Statistical System, *International Statistical Review*, 64: 165–187.

Findley, D.F., B.C. Monsell, W.R. Bell, M.C. Otto and B.-C. Chen, (1998), New capabilities and methods of the X-12 ARIMA seasonal-adjustment program, *Journal of Business & Economic Statistics*, 16: 127–177.

Fisher, W.D. (1958), On grouping for maximum homogeneity, *Journal of the American Statistical Association*, 53: 789-798.

Hansen, B. (2001), The new econometrics of structural change: dating breaks in U.S. labor productivity, *Journal of Economic Perspectives*, 15: 117-128.

Hartigan, J.A. (1975), *Clustering Algorithms*, John Wiley & Sons, 366 pp., New York.

Ladiray, D. and B. Quenneville B. (2001), *Seasonal Adjustment with the X-11 Method*, Spinger-Verlag, 256 pp., New York.

Shiskin, J.; A.H. Young and J.C. Musgrave (1967), The X-11 Variant of the Census Method II Seasonal Adjustment Program, Technical Paper 15 (revised), Washington, D.C.: U.S. Bureau of the Census, pp.66.

Su, X.G., M. Wang and J.J. Fan (2004), Maximum likelihood regression trees, *Journal of Computational and Graphical Statistics*, 13: 586-598.