Using Trees to Investigate Structural Breaks

L'Identificazione di Break Strutturali mediante gli Alberi di Decisione

Carmela Cappelli [1]
Dipartimento di Scienze Statistiche
Università di Napoli Federico II

Marco Reale [2]
Mathematics and Statistics Department
University of Canterbury, New Zealand

Riassunto: Nel presente lavoro viene proposto un approccio per la individuazione di break strutturali in serie temporali quando si dispone di serie esplicative. Il metodo si articola in uno schema iterativo basato sull'uso alternato della analisi canonica e della regressione ad albero. In particolare, l'informazione contenuta nelle serie esplicative viene sintetizzata, a mezzo della analisi canonica, in combinazioni lineari. Queste ultime sono successivamente impiegate come covariate nella regressione ad albero in cui la variabile di risposta è la serie temporale di cui si desiderano identificare i break strutturali. La proposta sarà illustrata mediante uno studio di simulazione.

Keywords: Structural breaks, canonical variate analysis, regression trees.

1. Introduction

This paper continues the streamline of graphical modelling applications to time series, (see for example Reale and Tunnicliffe Wilson, 2001) focusing on the problem of structural break detection in univariate time series. We propose a novel approach based on the use of canonical variate analysis and regression trees which are a special class of directed acyclic graphs.

Let x_{jt} with $j = 1, \ldots, p$ be a set of covariates deemed useful to explain an observed series y_t. A structural break occurs when there is a change in the data generating process of y_t. This happens when the role played by the explanatory variables in different regimes changes or there is a change in the model parameter values. The econometric literature on structural breaks is vast; a review of the main contributions can be found in Hansen (2001).

In this paper we examine the problem of multiple changes at unknown times by making use of covariates. The need for explanatory variables although limitative in one way is a useful framework to test alternative hypotheses.

The effect of the covariates is summarized by linear combinations obtained by means of the canonical variate analysis. Then, the linear combinations are used as new covariates in a regression tree procedure. The underlying idea is that the structural break is contained in one of the linear combinations and the tree procedure will reveal the existing breaks and their dates. Indeed, the tree procedure chooses at each step the combination that splits the series so that the observations in the two subperiods are as distinct as possible.

[1] Indirizzo per corrispondenza: carmela.cappelli@unina.it

[2] Carmela Cappelli gratefully acknowledges research funds granted to the Dipartimento di Scienze Statistiche

2. Trees to Detect Structural Breaks

This section shows how the detection of structural breaks can be set within the framework of canonical variate analysis and regression trees.
Suppose that a time series is characterized by $g-1$ structural breaks. We can assume accordingly that the data have been drawn by g distinct multivariate populations. Then, a categorical variable with g levels can be created to indicate the population each observation belongs to. We do not know neither the number of structural breaks nor their date of occurrence and our aim is to determine both. We start by creating binary indicator variables considering possible splits of our series into two disjoint groups (subperiods). Notice that in this case we need to preserve the temporal structure since the observations are not mutually exchangeable. Hence we will consider only sequential splits. In other words we consider the $T-1$ dichotomous partitions of the series $y_t = \{y_1, \ldots, y_T\}$ of the form:

$$P_{1,m} = \{y_1, \ldots, y_m\} \;\; and \;\; P_{2,m} = \{y_{m+1}, \ldots y_T\},$$

.
where T is the number of observations. Then, we create an indicator variable $I_m, \;\; m = 1, \ldots, T-1$ such that:

$$\begin{aligned} I_m &= 0 \quad \text{if } y_t \in P_{1,m}, \\ I_m &= 1 \quad \text{otherwise.} \end{aligned}$$

The I_m's will be used as grouping variables in the canonical variate analysis that, in the broad class of discriminant analysis techniques, is a nonparametric approach useful for dimension reduction (McLachlan, 1992).
In practice, the relevant covariates in the different regimes are unknown, so meaningful combinations of all of them are considered.
According to the proposed approach the original covariates x_{jt}, $j = 1, \ldots, p$, are replaced by the canonical variates, i.e. the eigenvectors associated with the eigenvalues of the product matrix $\mathbf{S}^{-1}\mathbf{B}$, where $\mathbf{S}^{-1}$ is the (common) within group covariance matrix and $\mathbf{B}$ the between group variance matrix. Therefore, the canonical variates are linear combinations of the covariates that identify the directions of maximum separation among the groups in the space spanned by the covariates.
Indeed, when a structural break occurs, populations labelled by a grouping variable are as distinct as possible and the intuition is that the structural break is contained in one of the linear combinations.
The canonical variates are used as new covariates in a regression tree procedure where the response variable is the original series which breaks are under investigation.
In a tree algorithm, the data are successively split along coordinate axes of the covariates so that, at any node, the split which maximally distinguishes the response variable in the left and the right descendants is selected. Therefore, the data partitioning aims to reduce the heterogeneity in the y_t's. In least squares regression trees (Breiman *et al.* 1984), heterogeneity at a given node k is measured by the sum of squares $SS = \sum_{y_t \in k}(y_t - \bar{y}(k))^2$ where $\bar{y}(k)$ is the average of the y_t's falling into node k. Then, the decrease in heterogeneity induced by a candidate split s of node k into its left and right descendants (k_l and k_r respectively) is evaluated as follows

$$\Delta SS(s,k) = SS(k) - [SS(k_l) + SS(k_r)]. \tag{1}$$

The algorithm searches over all permissible splits and chooses the best split to be the one that maximizes equation (1).
Since any best split divides the series into two subperiods as distinct as possible, it is expected that the split points reveal the breakdates.
Let t^* be the time of occurrence of the detected structural break, the procedure is recursively applied to the sibling nodes, containing $1, \ldots, t^*$ and $t^* + 1, \ldots, T$ observations respectively, until no further splitting is possible.
Note that, given the equivalence of maximum likelihood estimation with least square estimation in normal linear models, a maximum likelihood splitting criterion chooses the same split as the criterion in equation (1) does. Therefore, if normality (or asymptotic normality) can be assumed, to avoid the well known problem of *overfitting*, the Chow test (Chow, 1960) can be adopted as stopping rule. For a priori known breakdate, the Chow test is useful to verify the hypothesis of equality of two sets of parameters. Indeed, in a tree regression, a split (candidate breakdate) of a given node k divides the sample data, belonging to the node, into two subperiods. In the testing procedure we use the tree residuals. If the null hypothesis of constancy of the parameters is not rejected, the procedure declares the corresponding node as terminal.

3. Simulation

Preliminary results from a simulation study to assess the performance of the proposed procedure are now shown. We consider 100 simulations from the following model including two structural breaks:

$$\begin{aligned}
\text{Regime1}: &\quad y_t = 0.6x_{1t} + 0.2x_{2t} + 0.2x_{3t} + \epsilon_t \quad t = 1, \ldots, 50 \\
\text{Regime2}: &\quad y_t = 0.2x_{1t} + 0.6x_{2t} + 0.2x_{3t} + \epsilon_t \quad t = 51, \ldots, 100 \\
\text{Regime3}: &\quad y_t = 0.2x_{1t} + 0.2x_{2t} + 0.6x_{3t} + \epsilon_t \quad t = 101, \ldots, 150
\end{aligned}$$

where $\epsilon_t \sim NID(0, 100)$ and

$$\begin{aligned}
x_1 &= \nu_1 \\
x_2 &= \text{trend} + \nu_2 \\
x_3 &= \text{trend}^2 + \nu_3,
\end{aligned}$$

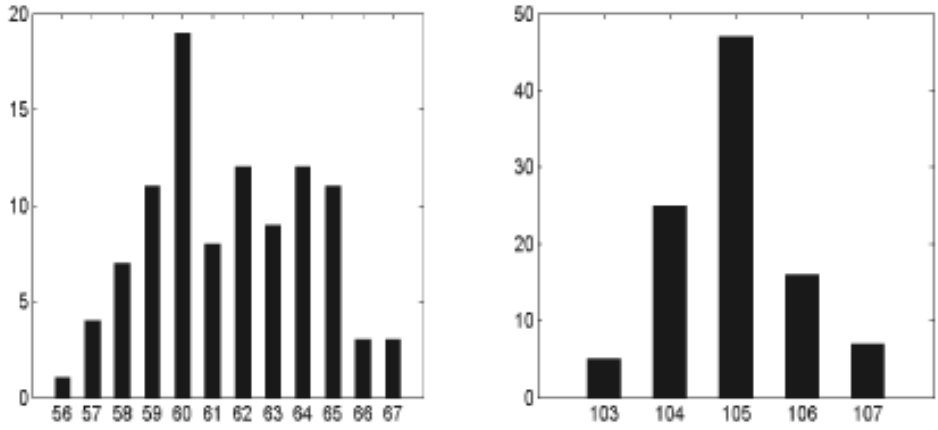
where the ν_i's, $i = 1, \ldots, 3$, are drawn from $NID(0, 100)$.
The difference in the regimes is given by the intercept and the trend: there is no trend in the first regime, a linear trend in the second and a quadratic trend in the third; the constants also differ from one regime to another.
Table 1 gives some summary statistics from the simulation results. For each break the actual date and the lower and upper detected times of occurrence, and the modal date over the 100 simulated time series are presented.

Table 1. *Actual breakdate, lower and upper time of occurrence and mode*

	Actual breakdate	Lower time	Upper time	Mode
Break 1	51	56	67	60
Break 2	101	103	106	105

Over all the 100 runs the procedure has detected always two structural breaks. It is noteworthy that the tree identifies in first place the second structural change as it is more apparent and also, the tree is much quicker in recognizing it: over the 100 runs at most at time 107 this break is identified. Also, notice that the actual breakdate is not within the interval between the lower and upper time of occurrence detected in the simulation study. This is not surprising because a break can be recognized only after its occurrence.
Further insights into the simulation results arise from figure 1 that depicts the histograms of the frequencies for the first and the second candidate breaks as detected by the tree.

Figure 1: *Histograms of the detected times of occurrence of break one and two.*



The histograms emphasize that the second structural change is more evident and therefore it is easier to detect. Indeed, for this break the detection times are less disperse and hence more precise, with a unimodal and symmetric distribution.
These preliminary results are encouraging and suggest extending the simulation study to understand the sampling properties in order to provide a diagnostic for breakdates.

References

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

Chow, G. (1960) Tests of equality between sets of coefficients in two linear regressions, *Econometrica*, 28, 591-605.

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

McLachlan G. (1992). *Discriminant Analysis and Statistical Pattern Recognition*, J. Wiley & Sons, New York.

Reale M. and Tunnicliffe Wilson G. (2001) Identification of vector AR models with recursive structural errors using conditional independence graphs, *Statistical Methods and Applications*,10, 49-65.