

Detecting contemporaneous mean co-breaking via ART

Carmela Cappelli and Francesca Di Iorio

Abstract In this paper we propose a procedure to detect the presence of co-breaking i.e. of a common structural break occurring at an unknown date in a vector of time series. Co-breaking occurs if a linear combination of the time series cancels the break. Our procedure employs a regression tree based approach namely ART, to detect the presence of breaks in the mean of each component of a multidimensional time series and Principal Component Analysis to estimate the linear combinations of the vector series. On each linear combination ART is performed again to detect the presence of a break. The combination that "hides" the co-breaking time is the one minimizing the employed splitting criterion. A simulation study has been carried out to evaluate the performance of the proposed approach whose results are presented and discussed.

Key words: Time series, Co-breaking, Regression Trees.

1 Introduction

In recent years the analysis of common features in economic and financial time series has emerged as a relevant topic of research (for an overview see Urga, 2007). Indeed economic or financial time series might have several distinctive features and the knowledge that a group of variables presents the same type of feature offers valuable information for modelling, policy making and forecasting.

The notion of common features was formalized by Engle and Kozicki (1993): *a feature possessed by a group of series is said to be common to those series if a linear combination of the series does not have the feature*. Developments of common

Carmela Cappelli and Francesca Di Iorio
Dipartimento di Scienze Statistiche, Università di Napoli Federico II e-mail:
carmela.cappelli@unina.it, fdiorio@unina.it. The authors gratefully acknowledge financial support from Dipartimento di Scienze Statistiche, Università di Napoli Federico II

features include *co-breaking* introduced by Hendry (1996) and Hendry and Mizon (1998). In particular *contemporaneous mean co-breaking* is defined as the cancellation of location shifts of individual series across linear combinations of these variables. In general, in the last two decades the literature of structural breaks has flourished and relevant contributions have been provided by Bai and Perron who have presented in various papers (1998, 2003, 2006) a comprehensive discussion of the issue providing estimation methods, testing procedures and confidence intervals for dealing with multiple structural breaks occurring at unknown dates in a regression framework. In this streamline, Cappelli and Reale (2005) proposed a fast method called ART (Atheoretical Regression Trees) that mimics Bai and Perron break dates estimation procedure by employing Least Squares Regression Trees to locate the breaks. Extensive simulation studies, comparison with Bai and Perron's and other current methods as well as applications to various real time series have provided evidence of the usefulness of the approach (see Cappelli *et al* 2008 and Rea *et al* 2009). In this paper we present a ART-based procedure to detect contemporaneous mean co-breaking. In particular, ART is employed for the preliminary identification of breaks in the mean of each one-dimensional component of a multidimensional time series. The corresponding break dates represent candidate co-breaking dates and thus they are employed to delimit an interval on which Principal Component Analysis (PCA) is recursively performed to generate the linear combinations of the series. On each linear combination ART is performed again to identify the best one that "hides" the co-breaking date. Although the use of PCA is not new in the field of common feature analysis, our approach represents an enhancement as it enables the date of the contemporaneous mean co-breaking to be *a priori* unknown i.e. to be estimated along the procedure. According to the current literature on the topic we focus on a single level shift and we present the results of a simulation study carried out considering contemporaneous mean co-breaking in various models.

2 Atheoretical Regression Trees for structural break analysis

In this section we discuss the issue of detecting multiple structural breaks and we introduce the ART method proposed by Cappelli and Reale (2005) to locate multiple breaks in the mean occurring at unknown dates; finally we show how the procedure can be employed for co-breaking analysis.

Let y_t be a series characterized by G regimes and $G - 1$ breaks, with $t = 1, \dots, T$. The objective is to estimate a set of break-dates $(T_1, \dots, T_g, \dots, T_{G-1})$ that define a partition of the series $P(G)$ into homogeneous subintervals with respect to some characteristic. In case of level shifts, the subseries are such that $\mu_g \neq \mu_{g+1}$. To identify these type of breaks a common estimation criterion is based on the least square principle that yields to

$$P^*(G) = \arg \min_{P(G)} \sum_{g=1}^G \sum_{t=T_{g-1}-1}^{T_g} (y_t - \mu_g)^2 \quad (1)$$

with $T_0 = 0$ and $T_G = T$. Cappelli and Reale (2005) showed that Regression Trees can be employed to identify multiple breaks in mean occurring at unknown dates. The procedure called Atheoretical Regression Trees (ART) works within the framework of Least Square Regression Trees (hereafter denoted LSRT) that are based on the same objective function as the 1. In LSRT a node h is split into its left and right descendants h_l and h_r to reduce the deviance of the response variable y_t fitting to each node the mean of the corresponding y 's values (for complete discussion the reader is referred to Breiman *et al*, 1984). Thus, the algorithm selects the split that maximizes the difference:

$$SSR(h) - [SSR(h_l) + SSR(h_r)] \quad (2)$$

where $SSR(h) = \sum_{y_t \in h} (y_t - \hat{\mu}(h))^2$, is the residual sum of squares for node h , $\hat{\mu}(h)$ is the mean of the y_t values in node h and $SSR(h_l)$ and $SSR(h_r)$ are the corresponding quantity computed for the left and right descendants, respectively. Once the binary partition of a node is performed, the splitting process is applied to each subgroup separately, and so on recursively until either the subgroups reach a minimum size or no improvement of the criterion can be achieved. LSRT provides a practical tool for locating structural breaks in the mean of long time series data.

At this aim, let w be an arbitrary ascending (or descending) sequence of completely ordered numbers, for sake of simplicity take $w = 1, 2, \dots, T$. Tree regressing y_t on w yields to a partition of the series into homogeneous subperiods such that $\hat{\mu}_g \neq \hat{\mu}_{g+1}$. The split points in the tree identify the break-dates whereas the different regimes are given by the y_t values falling into the tree terminal nodes.

Note that, according to the common practice, breaks are searched into a proper interval, i.e. $\min_{obs} \leq w \leq T - \min_{obs}$, where $\min_{obs}$ is a minimum number of observations needed for modelling, estimating and testing purposes.

As in any tree procedure a criterion to select the relevant of set of splits, i.e. the set of final breaks and corresponding regimes, is needed. At this aim either retrospective pruning or a stopping rule based on a testing procedure can be employed (for details see Rea *et al*, forthcoming). ART mimics the well known break estimation method of Bai and Perron (1998, 2003) whose method is computationally expensive because it is a global minimizer of the sum of squares residuals. On the contrary, ART is not a global minimizer as Bai and Perron's method but it's much faster. Extensive simulation studies and comparison with Bai and Perron's method and other current methods can be found in Cappelli *et al* (2008) and Rea *et al* (forthcoming) where it is shown that ART provides comparable results and, in particular it performs well in the simulations for long series for which the compute times for the Bai and Perron's method may be prohibitive.

2.1 Co-breaking analysis via ART and PCA

The notion of *Common Structural Breaks* introduced by Hendry (1996) and Hendry and Mizon (1998) can be formalized as follows. Let $\mathbf{y}_t = (y_{1t}, \dots, y_{nt})'$ be a n -dimensional vector of time series with $t = 0, \dots, T$ and

$$E[\mathbf{y}_0] = \mu_0$$

the unconditional expectation around an initial parametrization.

A Location Shift is said to occur in $\mathbf{y}_t$ if, for some $t > 0$

$$E(\mathbf{y}_t - \mu_0) = \mu_t \in R^n$$

and $\mu_t \neq \mu_{t-1}$.

Contemporaneous Mean Co-breaking (CMC) occurs when

$$E[\Psi' y_t - \Psi' \mu_0] = \Psi' \mu_t = \mathbf{0}$$

where Ψ is a $(n \times r)$ matrix of rank (r) . Then the linear transforms $\Psi' \mathbf{y}_t$ are independent of the location shifts and thus they cancel the break.

It is worth noting that so far co-breaking literature has been developed assuming the knowledge of the location shift timing (for an overview see Hendry and Massmann, 2007). On the contrary we propose a procedure to deal with CMC occurring at an unknown date, i.e. to detect the presence and timing of co-breaking. The procedure employs Principal Component Analysis (PCA) to generate the linear transforms in the framework of ART and it can be summarized as follows:

1. Apply ART to each of the one-dimensional component of the time series vector $\mathbf{y}_t$ to detect a single break in the mean and let $\tau = (\tau_1 \dots \tau_j \dots \tau_n)'$ be the vector of the identified break points i.e. τ_j is the best split of the j -th series of $\mathbf{y}_t$ into two subintervals, where $j = 1, \dots, n$ and $\min_{obs} \leq \tau_j \leq T - \min_{obs}$.
2. Let τ_m and τ_M be the minimum and maximum among the τ_j 's values thus they represent the earliest and latest candidate co-breaking date respectively. Consider the interval $[t_a = (\tau_m - v); t_b = (\tau_M + v)]$ where $\pm v$ is added to expand the searching interval. For each possible partition obtained using $t_a \leq k \leq t_b$ as cut-point, estimate the PCA linear combination $\mathbf{z}_k = \mathbf{Y} \hat{\mathbf{e}}$, where $\hat{\mathbf{e}}$ is the $(n \times 1)$ eigenvector corresponding to the largest eigenvalue of the correlation matrix of the $\mathbf{y}_t$ series and $\mathbf{Y} = (y_1 \dots y_n)$ is the $(T \times n)$ matrix that collects the n observed time series.
3. On each linear combination $\mathbf{z}_k$ ART is performed and the best linear combination $\mathbf{z}_{k^*}$, i.e. the one "hiding" the co-breaking date, is the combination that shows the minimum value of the splitting criterion (2).
4. The estimated co-breaking date is the cut point k^* .

It's worth noting that both the preliminary identification of the candidate co-breaking dates and the analysis of the linear combinations benefit from a fast approach such as ART to locate the break due to large memory and computing re-

quirements.

Also Note that in this context neither a pruning procedure or a stopping rule is needed because a single contemporaneous mean co-breaking is searched, thus only the first split is performed and evaluated.

3 Simulation study

In order to evaluate the ability of our procedure to detect a CMC occurring at an unknown date, we have carried out a simulation study considering two alternative DGP: a basic $AR(1)$ model and an $ARMA(2,1)$ model. In both cases we generated 1000 Monte Carlo replications of a 3-dimensional vector of time series $\mathbf{y}_t$ with $t = 1, \dots, 200$. Each series has a structural break in ϕ_0 that occurs at time t_0 . The $AR(1)$ model setting is as follows:

$$y_{j,t} = \begin{cases} \phi_{j,0} + \phi_{j,1}y_{j,t-1} + \varepsilon_{j,t} & 1 \leq t \leq t_0 \\ \phi_{j,0}^* + \phi_{j,1}y_{j,t-1} + \varepsilon_{j,t} & t_0 + 1 \leq t \leq 200 \end{cases} \quad j = 1, 2, 3. \quad (3)$$

where $\varepsilon_{j,t}$ are idiosyncratic gaussian error terms, $\phi_1 = \{0.5, 0.7, 0.4\}$, and, at the aim to obtain an heterogenous set up, $\phi_{j,0}$ and $\phi_{j,0}^*$ are generated from the Uniform variates: $\phi_{j,0} \sim Uniform(1; 1.3)$, $\phi_{j,0}^* \sim Uniform(3, 3.3)$ and their values are fixed along the Monte Carlo experiment.

In the $ARMA(2,1)$ setting

$$y_{j,t} = \begin{cases} \phi_{j,0} + \phi_{j,1}y_{j,t-1} + \phi_{j,2}y_{j,t-2} + \theta_{j,1}\varepsilon_{j,t-1} + \varepsilon_{j,t} & 1 \leq t \leq t_0 \\ \phi_{j,0}^* + \phi_{j,1}y_{j,t-1} + \phi_{j,2}y_{j,t-2} + \theta_{j,1}\varepsilon_{j,t-1} + \varepsilon_{j,t} & t_0 + 1 \leq t \leq 200 \end{cases} \quad j = 1, 2, 3. \quad (4)$$

$\varepsilon_{j,t}$ are idiosyncratic gaussian error terms and the coefficient values are $\phi_1 = \{1.5, 0.7, 0.2\}$, $\phi_2 = \{-0.7, -0.2, 0.3\}$ and, $\phi_{j,0}$ and $\phi_{j,0}^*$ are generated as in the previous case from the above seen Uniform variates.

Figure 1 depicts one of the 1000 Monte Carlo replications generated according to (3) and (4) respectively.

To give an insight into the problem, in Figures 2 and 3 we report for the two DGP the estimated PCA linear combination associated with the actual co-breaking time $t = 100$ (left panel) and the linear combination associated with $t = 80$ (right panel), that is 20 observations before the true co-breaking point.

As we can see in both cases the linear combination corresponding to the erroneous co-breaking date (time $t = 80$) does not cancel the break, on the contrary it shows a false appearance of 2 breaks whereas the linear combination corresponding to the actual co-breaking date shows no evidence of breaks.

The procedure described in section 2.1 has been applied setting for both DGP $min_{obs} = 80$ and $v = \pm 3$. The results, averaged over the 1000 MC replications, are summarized in Table 1 where, for the two DGP we report the Monte Carlo mean of the earliest and latest co-breaking date $\bar{\tau}_m$ and $\bar{\tau}_M$ respectively (the associated standard errors are given in brackets) and the percentages of correct identifications %ci

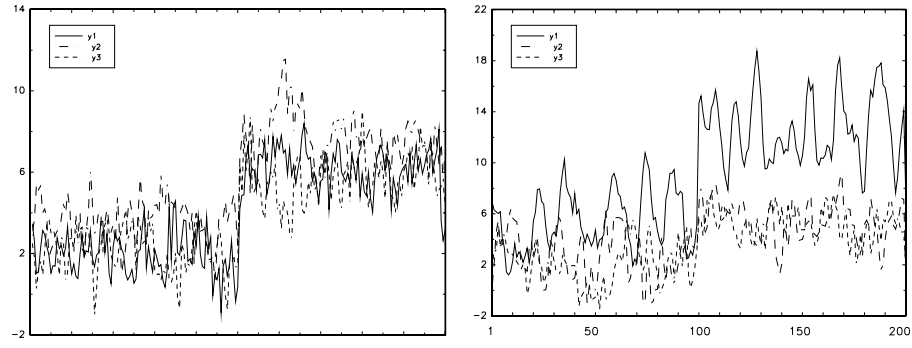


Fig. 1 Illustrative example of the simulated 3-dimensional vector of time series. Left panel $AR(1)$ model, right panel $ARMA(2,1)$ model.

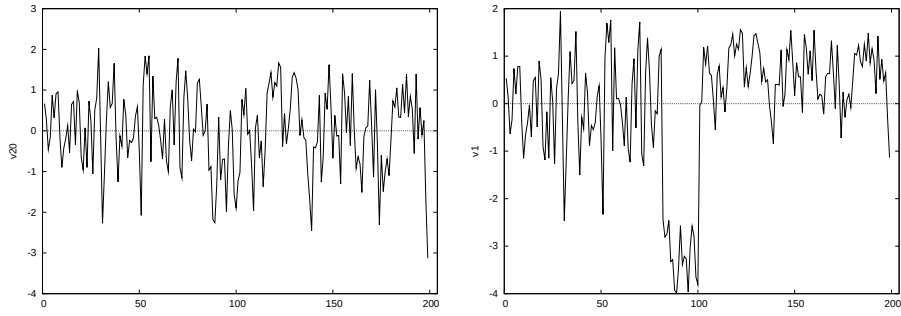


Fig. 2 $AR(1)$ model: linear combination corresponding to the actual co-breaking time ($t=100$) and to $t=80$.

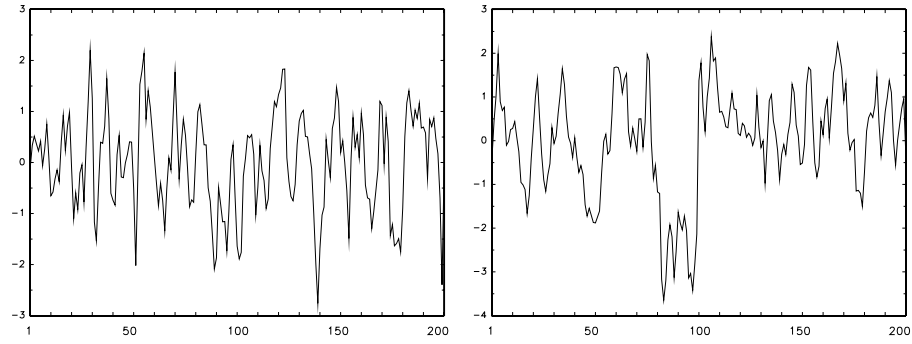


Fig. 3 $ARMA(2,1)$ model: linear combination corresponding to the actual co-breaking time ($t=100$) and to $t=80$.

i.e. the number of correctly identified CMC dates within various intervals around the actual date.

Table 1 Results for 1000 MC replications; co-breaking at time $0.5 \times T$ ($t = 100$)

	$\bar{\tau}_m$	$\bar{\tau}_M$	%ci ± 6	%ci ± 4	%ci ± 2	%ci ± 1
AR(1)	100.2 (1.68)	101.8 (1.64)	99.7	94.9	60.1	42.7
ARMA(2,1)	98.7 (4.68)	103.3 (4.51)	95.7	84	49.7	33.5

We see that the values of $\bar{\tau}_m$ and $\bar{\tau}_M$ i.e. of the mean earliest and latest break date are very close to each other (especially for the AR(1) model) and to the actual break date $t = 100$, confirming that ART is very effective in identifying the break date on individual series. As to its application to the linear combinations together with the selection criterion described in section 2.1 it provides percentages of correct identifications over or close to the 50% in the ± 2 interval around the actual break date reaching the 99.7% and 95.7% in the interval ± 6 . As expected the percentages of correct identifications tend to be higher for the AR(1) model but, for both DGP the results show that the proposed approach represents a promising fast tool to detect a relatively short interval to focus on for further analysis and investigations. In order to evaluate how sensitive is the method is to the position of the break we have carried out a further simulation considering a break at time $t = 150$. The results are reported in Table 2.

Table 2 Results for the 1000 MC replications, co-breaking at time $0.75 \times T$ ($t=150$)

	$\bar{\tau}_m$	$\bar{\tau}_M$	%ci ± 6	%ci ± 4	%ci ± 2	%ci ± 1
AR(1)	150.1 (1.82)	151.7 (1.75)	99.6	96.3	69.2	46.6
ARMA(2,1)	148.2.7 (6.68)	153.1 (4.00)	96.7	87.8	55.5	37.5

As in the previous case the values of $\bar{\tau}_m$ and $\bar{\tau}_M$ are close to each other and, surprisingly, the percentage of correct identifications are slightly higher. The simulations have been carried out with GAUSS routines, and, on a Intel Core Duo 1.8 GHz, the CPU time was 2min and 48sec for the break at time $t = 100$ and 3 min 12 sec $t = 150$.

Finally note that in principle, the approach could be applied separately to the sub-series identified by the co-breaking date to detect further mean co-breaking and we are currently investigating this issue. Preliminary simulations carried out considering two CMC show that the performance of the procedure depend on the size of the breaks: if one break is noticeably larger then the other, in a very high percentages of cases it is correctly detected by our procedure and thus one of the subseries contains the other break and the procedure tends to detect this other break too although with lower precision. Actually this is a common problem in structural break analysis because, in general, small breaks are difficult to detect; in the present case, further simulations are needed to assess the extent and burden of the difference between the two co-breaking sizes.

4 Conclusions

In this paper we have proposed a procedure to detect a contemporaneous mean co-breaking occurring at an unknown date in multivariate time series. Co-breaking represents a development of common features whose presence implies that a linear combination of the series cancels the feature.

Our approach employs a fast approach, namely ART, for the identification of breaks and Principal Component Analysis to generate the linear combinations of the series that are then analyzed again by ART to identify the one that contains the co-breaking date.

The results of the simulation studies have shown that the proposed procedure provides high percentages of correct identification of the contemporaneous mean co-breaking date and thus it helps to identify intervals that are likely to contain the co-breaking date.

However, further, extensive simulations are needed to assess the performance of the proposed approach when co-breaking size is moderate or small, when co-breaking occurs in the tails and in case of multiple breaks.

5 References

1. Bai J., Perron P. (1998), Estimating and testing linear models with multiple structural changes, *Econometrica*, 66, 47-78.
2. Bai J., Perron P. (2003), Computation and analysis of multiple structural change models, *Journal of Applied Econometrics*, 18, 1-22.
3. Bai J., Perron P. (2006), *Multiple structural change models: a simulation analysis*, in: Phillips P.C.B., Corbae D., Durlauf S.N., Hanse B.(eds.) *Econometric theory and practice: frontiers of analysis and applied research*, Cambridge University Press, 212-240.
4. Breiman L. , Friedman J.H., Olshen R.A., Stone C.J. (1984), *Classification and regression trees*, Wadsworth & Brooks, Monterey.
5. Cappelli C., Penny R.N., Rea W.S., Reale M. (2008), Detecting multiple mean breaks at unknown points with regression trees, *Mathematics and Computers in Simulation*, 78, 351-356.
6. Cappelli C., Reale M. (2005), Dating multiple structural breaks occurring at unknown dates via Atheoretical Regression Trees, in: Provati C. (ed.) *S.Co. 2005: Modelli Complessi e Metodi Computazionali Intensivi per la Stima e la Previsione*, Cleup, Padova, 479-484.
7. Engle R.F., Kozicki F. (1993), Testing for common features, *Journal of Business & Economic Statistics*, 11, 369-395.
8. Hendry D.F. (1996), A theory of Co-breaking, mimeo Nuffield College, University of Oxford.
9. Hendry D.F., Massmann M. (2007), Co-breaking: recent advances and a synopsis of the literature, *Journal of Business & Economic Statistics*, 25, 33-51.
10. Hendry D.F., Mizon G.E. (1998), Exogeneity, causality and co-breaking in economic policy analysis of a small econometric model of money in the UK, *Empir. Economics*, 23, 267-294.
11. Rea, W.S, Reale M., Cappelli C., Brown J.A. (2009), Identification of changes in mean with regression trees: an application to market research, *Econometric Reviews*, forthcoming.
12. Urga G. (2007), Common Features in economics and finance: an overview of recent developments, *Journal of Business & Economic Statistics*, 25, 2-11.