E C O N O M I C S B U L L E T I N

Investigating the structure of expansions and recessions in US business cycle: a modified recursive partitioning approach

Carmela Cappelli
University of Naples

Abstract

In this paper the problem of identifying the structure of expansions and recessions in the US economy is placed in the framework of recursive partitioning and discriminant analysis. The classification provided by the National Bureau ofEconomic Research (NBER) is considered. Using as covariates themain variables and indicators deemed useful to predict the business cycle, a modified recursive partitioning approach isproposed; at each step (tree node) the method identifies the linear combination of the covariates that discriminates the mostbetween being in and out of a recession; this new covariate is then used to split the data. The application to the case of the US business cycle and the comparison with classical logisticregression shows the merits of the proposed approach that represents a useful to tool to identify and to interpret thestructure of expansionsand recessions.

Special thanks go to G.M.Gallo for supplying the data and for his useful suggestions and encouragement.
Citation: Cappelli, Carmela, (2004) "Investigating the structure of expansions and recessions in US business cycle: a modified recursive partitioning approach." *Economics Bulletin*, Vol. 3, No. 48 pp. 1−9
Submitted: August 14, 2004. **Accepted:** December 28, 2004.
URL: http://www.economicsbulletin.com/2004/volume3/EB−04C40006A.pdf

1. Introduction

The issue of identifying expansions and recessions in the business cycle has attracted a lot of attention in the literature, with major efforts devoted to the identification of indicators capable of signaling expansions or contractions in the economy. In the case of the US, the research has developed some interesting aspects: on the one hand, the National Bureau of Economic research (NBER) closely monitors economic and financial variables and issues statements about the state of the economy. The Dating Committee officially declares dates in and out of a recession providing an undisputed classification which can be used as a reference. On the other hand, given the delay with which such dates are announced, there is a strong interest in developing methodologies capable of synthesizing the information available into early warnings about the inception and duration of a recession or of providing explanations for business cycle behavior. In this area, for many years now, Stock & Watson (1989,1993) have developed a methodology of synthesis of a number of indicators in order to derive a coincident index and leading index with some variants.
In this paper the extraction of the information from the financial and economic indicators of the US economy is placed within the framework of discriminant analysis and decision trees. The goal is to show that having available a response variable which can take on the values 1 or 0 according to whether the economy is officially declared in a recession or not (following the NBER dates) one can derive one or more linear combinations of the indicators and an appropriate number of thresholds capable of classifying a new vector of observations as a recession.
In particular, a modified recursive partitioning is adopted to investigate the structure of expansions and recessions. At each step (tree node) this method identifies the combination of variables and indicators that discriminate the most between being in and out of a recession. This combination represents the direction in the covariate space of maximum separation and the therefore the data are partitioned along it.
The resulting tree structure can be seen as a set of rules i.e., conditions in terms of covariate's values, leading to recession or expansion. Therefore, the tree defines *profiles* can be used to classify a new vector of observations as a recession or an expansion.
The paper is organized as follows: in section 2 the proposed approach is presented; section 3 the application to the case of US business cycle is described and commented by comparing the results to those from classical logistic regression; concluding remarks follow in section 4.

2. Recursive partitioning along directions

Recursive partitioning methods, or tree-based methodologies, have proven to be effective in discovering general classification rules from a set of examples. To summarize, the procedure consists of a recursive binary partition of a set of objects described in terms of explanatory variables (either numerical or and categorical) and a response variable, therefore the procedure starts with a training set $\{(x_{1i}, \ldots, x_{ji}, \ldots, x_{pi}, y_i)\}_{i=1}^{n}$, *i.e.*, a set of n examples characterized as a p-dimensional vector of features or attributes with each example belonging

to one of G known classes according to the value assumed by y_i. The algorithm examines each attribute and a cut point along that attribute, choosing the "best" one on the basis of a *goodness of split* measure. The corresponding value reflects how well that attribute and cut point discriminate classes into two mutually exclusive subsets. Because of the evident analogy with the graph theory, a subset of observations is called node and nodes that are not split are called terminal nodes or leaves.
Given an *impurity measure* (Breiman *et* al., 1984) that expresses how homogeneous a given node t is with respect to the response variable, the decrease in impurity generated by a candidate split s of node t into its left and right descendants (t_l and t_r respectively) is evaluated as follows

$$\Delta I(s,t) = i(t) - [i(t_l)p_l + i(t_r)p_r], \tag{1}$$

where p_l and p_r represent the weights, i.e. the proportion of examples in node t falling into its descendants. The most common impurity measure employed in classification trees is the Gini index of heterogeneity, defined at any node t as: $i(t) = 1 - \sum_{g=1}^{G} p_g^2(t)$, where $p_g(t)$ represents the proportion of group g cases falling into node t.
The split s^* that, according to the 1, produces the highest decrease in impurity is selected to partition node t.
This procedure is recursively applied to each descendant in an effort to classify correctly as many of the training cases as possible.
Once the tree is built, a response value or a class label is assigned to each terminal node; in particular, in the case of classification each node is assigned to the class which presents the highest proportion of observations.
We see that splits are binary questions of the form: "Is x_j in A?", where $j = 1, \ldots, p$, so that, if x_j is ordered, this set includes all questions: "Is $x_j \leq c_j$?", for appropriate cut points c_j ranging over the domain of x_j. Therefore the standard recursive partitioning is based on univariate spits but sometimes the structure of the data suggests the use of combinations of covariates because covariates are correlated and more of them affect the response at the same time.
In the literature there are suggestions about substituting functions of the form:

$$\sum_j a_j x_j \tag{2}$$

to the x_j's above, where the a_j's are coefficients to be determined. In particular, combination procedures employing normal theory-based discriminant analysis have been introduced, see for example Loh & Kim (2001).
In what follows, we will adopt the approach by Cappelli & Conversano (2002) that resorts to canonical variate analysis in order to identify the functions. This method, that in the broad class of discriminant analysis techniques, is useful for dimension reduction (McLachlan, 1992), searches for those linear combinations (so called canonical variates) of the original variables which provide the optimal configuration (in a reduced space) of the relationships among the covariates and the grouping variable.
Let $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_g, \ldots, \boldsymbol{\mu}_G$ be the (column) vectors of group means and

$\mathbf{B}_o = \frac{1}{G-1}\sum_{g=1}^{G}(\boldsymbol{\mu}_g - \boldsymbol{\mu})(\boldsymbol{\mu}_g - \boldsymbol{\mu})'$ be the between group variance matrix.

$\boldsymbol{\Sigma}$ be the (common) covariance matrix.

The method seeks for a p-dimensional column vector $\mathbf{w}_k$ ($k = 1, .., d$; $d = \min(G-1, p)$) that maximizes the ratio of the between-group sum of squares to the (common) within group sum of squares.
Thus, vector $\mathbf{w}_k$ is the eigenvector corresponding to the w-th largest (nonzero) eigen-value of the product matrix $\boldsymbol{\Sigma}^{-1}\mathbf{B_o}$.
Since in practice both $\boldsymbol{\Sigma}$ and $\mathbf{B_o}$ are unknown, they are estimated using the training data. Therefore, in the sample version of a canonical analysis, vectors $\mathbf{w}_k$ are computed considering the sample estimates of $\boldsymbol{\Sigma}$ and $\mathbf{B_o}$, denoted $\mathbf{S}$ and $\mathbf{B}$ respectively.
The projections of the original covariates onto the space spanned by the eigenvectors of $\mathbf{S^{-1}B}$ define the canonical variates:

$$\mathbf{cv}_k = \mathbf{X}\mathbf{w}_k. \tag{3}$$

where $\mathbf{X}$ is the $(n \times p)$ matrix of the observed covariate values.
The canonical variates are linear combinations of the original p covariates for which the groups are as much spread as possible.
The method enjoys two important properties: robustness to non normality and even to mildly not equal covariance matrices for the groups (Hastie *et al.*, 1995).
Indeed, the method does not depend on the assumption of normality requiring only the knowledge of the first and second order moments. In other words canonical variate analysis is optimal for ellipsoidal distributions completely described by these moments such as the Multivariate Normal one, but it does not relay on any distributional assumption (Hand, 1998). Thus, this approach preserves the nonparametric approach of tree based procedure, also it does not complicate the existing method because it uses the same the splitting procedure but, on the canonical variates.
Given that the canonical variates provide the directions in which the groups are best separated, it is convenient to seek the splits along them because they model the relation between the response and the covariates, whereas several univariate splits orthogonal to the axis might be required to approximate a single split on a linear combination. The splitting variables are then defined on the canonical variates rather than on the original covariates.
It is noteworthy that the coefficients are re-estimated at each step i.e., in the tree growing procedure the canonical variate analysis is recursively applied in order to generate new canonical variates which are derived considering the subgroups of observations falling in the current node. In this way, the coefficients of the combinations are updated at each run and in the nodes where only a subgroup of response classes is present, the canonical variates are built considering only the actual number of classes.
In the case of the US business cycle monthly data, it is $G = 2$ and each y_i is either equal to 1 when month i falls in one of the NBER recession periods or it is equal 0 otherwise.
The proposed approach appears useful because the canonical variates provide the best conjoint synthesis of the variables included in the original information set as business cycle predictors and it is reasonable to assume that not a single indicator but all of them are

responsible for a recession.
In the end, based on the identified combinations, the data space is partitioned into mutually disjoint subregions (terminal nodes) each labelled as recession or expansion. As it will be shown in the application, the tree generates paths of conditions defined in terms of binary questions on linear combinations that identify *profiles* of recessions and expansions. Therefore, once the tree has been generated, a new observation of unknown class (month which covariate's values are available), dropped down the tree reaches a terminal node, and it's labelled as an expansion or recession.
Note that in the current stage of the research the approach is static, in that only a single value for each indicator is considered at each point in time.

3. Data description and application

The time series used as covariates are chosen among the variables which signal business cycle activity, namely, the industrial production (IP), the Standard and Poor's 500 Stock Index (SP), the Total Employment (EMP), the Housing Starts (HOU), the Term Premium (TP), defined as the 10-year interest rate on Treasury bonds minus the equivalent rate on 1-year maturity, the Federal Funds Rate (FED), and some indicators which synthesize other variables related to the cycle, namely the Experimental Coincident Index (XCI), Experimental Recession Index (XRI), the Experimental Coincident Recession Index (XRIC), the Coincident Index (CI). The response variable is a binary variable based on the NBER classification of expansions and recessions. The series are monthly and have been transformed following Filardo (1994): IP and SP are expressed in growth rates; FED, TP, XCI, EMP, and CI are level differenced. The resulting available data run from Feb. 1961 to Jul. 2003. Out of these, we keep the observations up to Dec. 1999 in the training set while the remaining data are kept as a test set. This test set is particularly interesting because it starts at a date when the long expansion of the Clinton years was still present and it includes months characterized by uncertainty about the health of the US economy. The signs of slowdown and the burst of the stock exchange bubble in the year 2000 have generally been read just as symptoms of a recession. Its official inception was placed by the NBER (on Nov. 26, 2001) at Mar. 2001.
The tree structure is depicted in Figure 1. Nodes have been numbered according to the standard node numbering system, i.e., the descendants of any node node t are numbered $2t$ and $2t + 1$, respectively. Inside each node it is indicated the node number (in brackets) and the distribution of the response over the training set; test set misclassifications are also reported at the side of the node. Note that, being $G = 2$ only a singe canonical variate is estimated at each node and it is reported beneath the nodes, the superscript indicating the corresponding node of occurrence.
The tree with 5 terminal nodes is capable of explaining expansions and recessions with a very low error rate (7/467) especially considering that the wrong attributions are months at the beginning or the end of a recession (turning points), and therefore can be viewed as

consistent with the NBER dates. In fact, although the time dimension of the data is not explicitly considered, because the aim is to explain recession and expansions as a whole, the tree preserves this structure keeping subsequent periods together; it separates with a single split almost all of the periods of expansion that reach node 2. The remaining periods are recessions with some expansions and further splits are able to discriminate 51 recession months. Nodes 2 and 7 labelled as *strong expansions* and *strong recessions* are able to classify with no error most of the training and test set. As one may expect, however, the error rate on the test set is higher (2/43), the wrong classifications relating to July and August 2001. It is quite possible that there were early signs of recovery which were upset by the events of September 11 (as a matter of fact the official end of the recession was placed at Nov.2001).

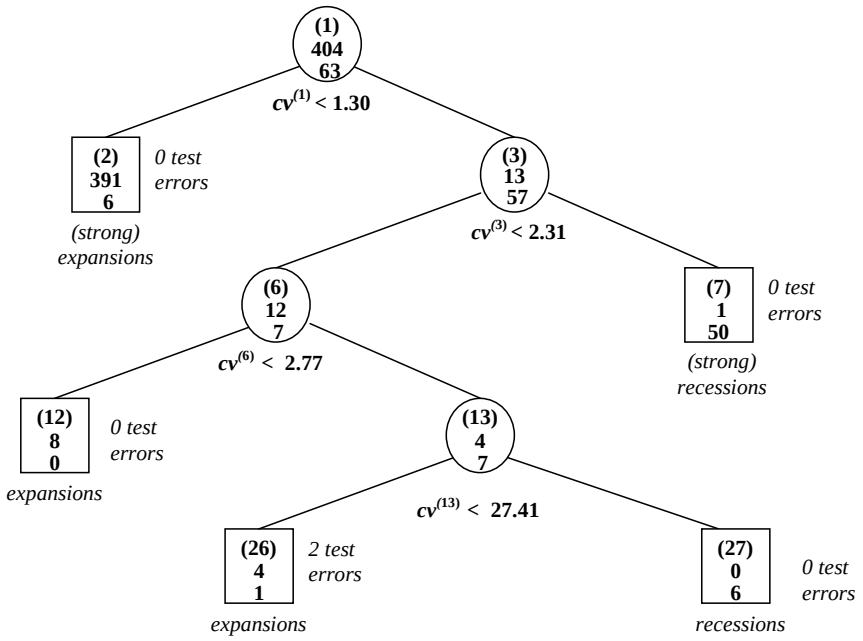

Figure 1. The canonical variate tree for the US business-cycle

In Table I we list the coefficients of the combinations used in the tree generation.
The combinations do not have an economic interpretation *per se* but some indication about the role played by the variables can be drawn, based on the corresponding coefficient. Note that, since the variables are measured on different scales, they have been standardized. Some profiles of recessions and expansions can be defined. The *strong expansions* are mostly discriminated on the basis of the values of the composite indexes indicators used by Stock and Watson, namely XRI.C, XRI and XCI with a less marked role played CI and of single covariates SP, EMP and FED. Once this split has taken place, composite indexes CI, XCI, XRI but also the single variables IP and EMP, explain the *strong recessions*. The same remark holds for the two further splits that account for the separation of very few recessions and expansions; these splits are indeed characterized by an increasing role played by single variables. In particular, note that the last split on node 13 is determined primarily by TP (with a very high coefficient), IP and EMP.

Table I. Coefficients of the variables and indicators in the canonical variates

	Coefficients			
Covariates	$cv^{(1)}$	$cv^{(3)}$	$cv^{(6)}$	$cv^{(13)}$
IP (industrial production)	0.034	0.717	1.175	13.340
SP(standard and poor's	-0.216	-0.057	-0.089	3.907
FED (federal funds rate)	-0.155	-0.069	-0.617	8.789
TP (term premium)	-0.036	-0.139	-0.409	50.478
XRI (exp. recession index)	0.651	0.717	1.269	-9.094
XCI (exp. coincident index)	0.480	0.948	3.50	-6.188
XRI.C (exp. coincident recession index)	1.433	0.522	1.571	0.068
EMP (total employment)	-0.180	-0.617	-1.091	14.508
CI (coicident index)	-0.284	-1.621	-4.154	6.755
HOU (housing starts)	0.014	0.283	0.544	-5.521

It's noteworthy that some of the individual variables are already included in the synthetic indices, therefore this result suggests the need for using different weights when building coincident indicators.
In order to provide a benchmark for the proposed approach, logistic regression has been fit to the data considering the same division between training and test set.
After a model that includes all covariates has been estimated by maximum likelihood, backward stepwise procedure has been used to select variables that have significant (at the %5 level) effect. This gave the final model shown in Table II.

Table II. Results from stepwise logistic regression

Selected covariate	**Coefficient estimate**	**Standard error**	Z **score**
constant	-4.256	0.571	-7.444
SP(standard and poor's	-0.773	0.260	-2.968
FED (federal funds rate)	-2.968	0.239	-2.164
XRI (exp. recession index)	1.385	0.247	5.593
XCI (exp. coincident index)	2.091	0.799	2.616
XRI.C (exp. coincident recession index)	1.751	0.363	4.822
EMP (total employment)	-1.037	0.410	-2.527
CI (coicident index)	-2.171	0.762	-2.850

The selected model does not include covariates IP, TP and HOU and, indeed, by comparing the model with the results reported in Table I, we see that the covariates included correspond to those having the highest coefficients in the first canonical variate cv_1 used to split the root node of the tree. In other words, the logistic model recovers the first split of the canonical variate tree, i.e., the split that separates most of recessions and expansions.
This conclusion is confirmed by the analysis of the confusion matrix associated to the model

and shown in Table III. The number of misclassifications and their distribution between the two response classes are close to that associated with node 2 and 3 of the canonical variate tree (see the distribution of recessions and expansions into nodes 2 and 3 in Figure 1).
Thus, the classification produced by logistic regression may be considered effective for the "simple" task, i.e., separating the strong recession form the strong expansions, but, it is not able to classify 11 recessions.

Table III. Confusion matrix from the logistic regression

	Predicted	
Observed	*Expansion*	*Recession*
Expansion	396	8
Recession	11	52

On the contrary the canonical variate tree due to further splits archives a higher accuracy correctly classifying 62 out of 63 recessions. Moreover, the logistic model misclassifications are not all turning points i.e., it misclassifies months within period of recession of expansions. The error rate on the test set (3/43) is slightly higher that the one associated to the canonical variate tree, the misclassifications not corresponding to the same months misclassified by the canonical variate tree.
The comparison with logistic regression that is a standard data analysis and inference tool, highlights the usefulness of the proposed approach: 1) it deals with correlated covariates by defining meaningful combinations of the original covariates; 2) makes use of the joint information present in the data; 3) being based on a recursive partitioning approach that successively divides up groups of observations it is likely to isolate and treat subgroups showing particular features.

4. Conclusions

The application to the case of the US business cycle and the comparison with classical logistic regression, has shown that the proposed approach provides an accurate classification of recessions and expansions explaining their structure in a way that is consistent with the undisputed classification provided by the NBER.
Since the process of declaring whether the economy is in or out of a recession may take several months, it is highly important to have available methods for early accurate detection of slowdowns.
In this respect the method, identifying profiles of recessions (and expansions) based on the values of the covariates jointly considered, represents a useful tool for a preliminary classification of new vectors of observations (months) as recessions or expansions.

References

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

Cappelli, C. and Conversano, C. (2002) Canonical Variates for Recursive Partitioning in Data Mining, in: Haerdle W. and Rnz B. (eds), *COMPSTAT 2002 Proceedings in Computational Statistics*, Physica-Verlag, Heidelberg, 213-218.

Filardo, A. (1994) Business-Cycle Phases and their Transitional Dynamics, *Journal of Business and Economic Statistics*,12, 299-308.

Hastie, T., A. Buja and R. Tibshirani (1995) Penalised discriminat analysis, *The Annals of Statistics*, 23 73-102.

Hand,H.J. (1998). *Construction and assessment of classification rules*, Wiley, New York.

Loh, S. and Kim, B. (1995) Classification Trees with Unbiased Multiway Splits, *Journal of the American Statistical a Association*, 44, 389-397.

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

Stock, J. O. and Watson, M.W. (1989) New Indexes of Coincident and Leading Economic Indicators, in: *NBER Macromeconomics Annual*, 351-394.

Stock, J. O. and Watson, M. W. (1993) A Procedure for Predicting Recessions with Leading Indicators: Economic Issues and Recent Performances, *Business Cycle, Indicators and Forecasting*, 25, 95-153.