Süleyman Demirel Üniversitesi,
Fen Bilimleri Enstitüsü Dergisi,
15-3(2011),186-191

Çoklu Yapısal Kırılmalar Yardımıyla Sürekli Değişkenin Kesikleştirilmesinde Deneysel Bir Çalışma

H. Hatice ÖZKOÇ*
Muğla Üniversitesi, Fen Fakültesi İstatistik Bölümü/ MUĞLA
Alınış Tarihi:29.06.2011, Kabul Tarihi:11.11.2011

Özet: Bir sınıflama algoritması ve ekonometrik modeler kesikli değerler alan değişkenler yardımıyla gerçekleştirilmektedir. Bu algoritmaların kullanılması durumunda sürekli değişkenlerin kesikli hale getirilmesi gerekmektedir. Bu çalışmada sürekli değişkenin kesikleştirilmesinde çoklu yapısal kırılmaların kullanımı gösterilmiştir.

Anahtar Kelimeler: Kesikleştirme, Çoklu Yapısal Kırılma, Bai-Perron, Kesikli değer

An Experimental Study on Discretizing Continuous Variables using Multiple Structural Changes

Abstract: Many classification algorithms and econometric models require that training examples contain only discrete values. In order to use these algorithms when some variables have continuous values, the numeric variables must be converted into discrete ones. This paper describes a new way of discretizing numeric values using multiple structural changes.

Keywords: Discretization, Multiple Structural Changes, Bai-Perron, Discrete value

1. Introduction

Learning from quantitative data is often less effective and less efficient than learning from qualitative data. Discretization addresses this issue by transforming quantitative data into qualitative data.

Discretization has been widely used in many classification algorithms such as artificial intelligence data mining, computer simulation, etc.

Most real-world applications of classification algorithm contain continuous attributes. When the feature space of data includes continuous attributes only or mixed type of attributes (continuous type along with discrete type), it makes the problem of classification vitally difficult. For example, classification methods based on instance-based measures are generally difficult to apply to such data because the similarity measures defined on discrete values are usually not compatible with similarity of continuous values. Alternative methodologies such as probabilistic modeling, when applied to continuous data, require an extremely large amount of data (Lee, 2005, p. 493).

The classification and thereby arrangement of continuous variables is a frequently used method in econometric studies, as well as computer sciences. Making a continuous variable discretization is important especially while working with ordinal models. The continuous variable, which is obtained as a result of discretization, becomes a categorical variable.

Use of categorical variables is preferred in many of econometric studies. Usually one way have a continuously observed dependent variable like individual buying behavior in terms of dollar sales, but in the end, one might be interested only in understanding which variables explain low-volume, medium-volume and high-volume buyers. One way then wants to construct an Ordinal Regression Model instead of a standard regression model (Franses and Cramer, 2010, p. 125). Categorization (making discrete) of continuous variables is required especially in income survey or during studying with ordinal models. In these studies, although percentiles are practiced, this type of categorization cannot provide the desired results and explanations.

The discretization that is encountered in the cross-section analysis is also encountered in time series in a similar way. The studies performed with time series aim to determine the effect of time on the change that occurs on the series. Thus the dates during which the series show a change according to time are tried to be found. These dates are expressed as dates, during which a break is seen on the series in the time series analyses. There are several methods in literature for the determination of the break dates. While some of them take only one break in the series into consideration (such as Perron and Zivot – Andrews unit root tests), others take multiple breaks into consideration.

Bai and Perron (1998, 2003a) have developed an alternative method to test the multiple structured fractions. Bai and Perron have also developed an effective algorithm acquiring global minimum values of error sum of squares. This algorithm bases on dynamical programming basic and requires Least Squares Method for each break point (Cevik and Erdogan, 2009, p. 31).

* hgorgulu@mu.edu.tr

The method developed by Bai and Perron (1998) to predict and to test the multiple structural change in time series will be used to convert continuous variables into discrete varieties in this study.

2. Material and Method

Discretization

Discretization is to divide the range of the continuous variable into intervals. Every interval is labeled a discrete value, and then the original data will be mapped to the discrete values. Discretization is a data-processing procedure that transforms quantitative data into qualitative data. This process is an essential task of the data preprocessing, not only because some learning methods do not handle continuous attributes, but also for other important reasons:

- The data transformed in a set of intervals are more cognitively relevant for a human interpretation,
- The computation process goes faster with a reduced level of data,
- The discretization can provide non-linear relations.

Lastly, discretization can harmonize the nature of the data if it is heterogeneous – eg., in text categorization, the attributes are a mix of numerical values and occurrence terms.

An expert realizes the best discretization because he can adapt the interval cuts to the context of the study and then he can make sense of the transformed attributes (Muhlenbach and Rakotomalala, 2005, p.397).

In addition, poorly discretized attributes prevent classification systems from finding important inductive rules. For example, if the ages between 15 and 25 mapped into the same interval, it is impossible to generate the rule about the legal age to start military service. Furthermore, poor discretization makes it difficult to distinguish the non-predictive case from poor discretization. In most cases, inaccurate classification caused by poor discretization is likely to be considered as an error originated from the classification method itself. In other words, if the numeric values are poorly discretized, no matter how good our classification systems are, we fail to find some important rules in databases.

Although discretization influences significantly the effectiveness of classification algorithms, not many studies have been done because it usually has been considered a peripheral issue. A simple method, called equal distance method, is to partition the range between the minimum and maximum values into N intervals of equal width. Another method, called equal frequency method, chooses the intervals so that each interval contains approximately the same number of training examples; thus, if N = 10, each interval would contain approximately 10% of the examples. However, with both of these discretizations, it would be very difficult or almost impossible to learn certain concepts (Lee, 2005, p. 494).

It is a process of transforming a continuous attribute values into a finite number of intervals and associating with each interval a discrete, categorical value. Discretization Algorithms can be mainly divided into following types: supervised or unsupervised (unsupervised algorithms do not consider the respective class labels); global or local (global algorithms group values of each feature into intervals by considering other features); static or dynamic (static algorithms discretize each feature in one iteration independent of other features, while dynamic algorithms search for all possible intervals for all features simultaneously). There are many specific algorithms of discretization.

3. The Model and Estimators

In the analysis of multiple structural change models, the most important contribution is the one of Bai and Perron (1998) who provides a comprehensive treatment of various issues. In particular, they consider tests and tabulate critical values using the derived asymptotic distributions. Bai and Perron (2000) and Jouini and Boutahar (2002) carry out some simulations to analyze the adequacy of these tests based on their asymptotic distributions. They find that the tests show size distortions especially for small values of the trimming when allowing for serial correlation and/or different distributions of the data and the errors across subsamples in the estimated regression models. Thus, the asymptotic distributions may be an unreliable guide to finite sample behavior and as a result the nominal levels of tests based on asymptotic critical values can be very different from the true levels. An alternative approximation is the bootstrap distribution that gives evidence on the adequacy of tests and often provides a tractable way to reduce or eliminate finite sample distortions of the sizes of statistical tests as suggested by Christiano (1992) and Diebold and Chen (1996) for the case of single structural change tests.

Consider the following structural change model with m breaks:

$$\begin{aligned} y_t &= x_t'\beta + z_t'\delta_1 + u_t, && t = 1,\ldots,T_1, \\ y_t &= x_t'\beta + z_t'\delta_2 + u_t, && t = T_1+1,\ldots,T_2, \\ &\vdots \\ y_t &= x_t'\beta + z_t'\delta_{m+1} + u_t, && t = T_m+1,\ldots,T. \end{aligned} \qquad (3.1)$$

In this model, y_t is the observed dependent variable at time t; $x_t(p\times 1)$ and $z_t(q\times 1)$ are vectors of covariates and β and $\delta_j(j=1,\ldots,m+1)$ are corresponding vectors of coefficients; u_t is the disturbance at time t. The indices $(T_1,\ldots,T_m)$, or the break points, are explicitly treated as unknown. The purpose is to estimate the unknown regression coefficients together with the break points when T observations on (y_t, x_t, z_t) are available. Note that this is a partial structural change model in the sense that the

parameter vector β is not subject to shifts and is effectively estimated using the entire sample. When $p=0$ all the coefficients are subject to change and we then obtain a pure structural change model.

Instant of break is suggested to be used in converting continues varieties into discrete variables. Here, each instant of break means transferring into a new category. So, category is acquired one more than instant of break number.

The multiple linear regression models given by (3.1) may be expressed in matrix form as

$$Y = X\beta + \bar{Z}\delta + U, \qquad (3.2)$$

where $Y=(y_1,\ldots,y_T)'$, $X=(x_1,\ldots,x_T)'$, $U=(u_1,\ldots,u_T)'$, $\delta=(\delta_1',\delta_2',\ldots,\delta_{m+1}')'$ and $\bar{Z}$ is the matrix which diagonally partitions Z at the m-partition $(T_1,\ldots,T_m)$, i.e. $\bar{Z}=diag(Z_1,\ldots,Z_{m+1})$ with $Z_i=(z_{T_{i-1}+1},\ldots,z_{T_i})'$. Bai and Perron (1998) impose some restrictions on the possible values of the break dates. Indeed, they define the following set for some arbitrary small positive number ε:

$$\Lambda_\varepsilon=\left\{(\lambda_1,\ldots,\lambda_m);\left|\lambda_{i+1}-\lambda_i\right|\geq\varepsilon,\lambda_1\geq\varepsilon,\lambda_m\leq 1-\varepsilon\right\}, \qquad (3.3)$$

to restrict each break date to be asymptotically distinct and bounded from the boundaries of the sample.

The estimation method considered is that based on the least-squares principle proposed in Bai and Perron (1998). This method is described as follows. For each m-partition $(T_1,\ldots,T_m)$, denoted $\{T_j\}$, the associated least-squares estimates of β and δ_j are obtained by minimizing the sum of squared residuals $\sum_{i=1}^{m+1}\sum_{t=T_{i-1}+1}^{T_i}(y_t-x_t'\beta-z_t'\delta_i)^2$. Let $\hat{\beta}(\{T_j\})$ and $\hat{\delta}(\{T_j\})$ denote the resulting estimates based on the given m-partition $(T_1,\ldots,T_m)$ denoted $\{T_j\}$. Substituting them in the objective function and denoting the resulting sum of squared residuals as $S_T(T_1,\ldots,T_m)$, the estimated break dates $(\hat{T}_1,\ldots,\hat{T}_m)$ are

$$(\hat{T}_1,\ldots,\hat{T}_m)=\arg\min_{(T_1,\ldots,T_m)} S_T(T_1,\ldots,T_m), \qquad (3.4)$$

where the minimization is taken over all partitions $(T_1,\ldots,T_m)$, such that $T_i-T_{i-1}\geq(\varepsilon T)$. The break point estimators are thus global minimizers of the objective function. Finally, the estimated regression parameters are the associated least-squares estimates at the estimated m-partition $\{\hat{T}_j\}$, *i.e.* $\hat{\beta}=\hat{\beta}(\{\hat{T}_j\})$ and $\hat{\delta}=\hat{\delta}(\{\hat{T}_j\})$ (Jouini and Boutahar, 2003, p. 4).

Bai and Perron (2003a) suggest three tests during determination of instant of break number. The hypothesis that does not include break for the extinction of some number of breaks is examined in first of these tests. In the second, the hypothesis that does not include break for unknown number of break is examined. And in last one, $l+1$ break for l is tested. With the help of these tests, the category number that will be used in converting continuous variables into categorical variable is determined so and tested.

A Test of No Break Versus Some Fixed Number of Breaks

Bai and Perron (1998) first consider the sup F type test of no structural break (m=0) versus the alternative hypothesis that there are m=k breaks. Let $(T_1,\ldots,T_k)$ be a partition such that $T_i=[T\lambda_i](i=1,\ldots,k)$.

$$F_T(\lambda_1,\ldots\lambda_k;q)=\left(\frac{T-(k+1)q-p}{kq}\right)\frac{\hat{\delta}'R'\left(R\left(\bar{Z}'M_X\bar{Z}\right)^{-1}R'\right)^{-1}R\hat{\delta}}{SSR_k} \qquad (3.5)$$

where R is the conventional matrix such that $(R\delta)'=(\delta_1'-\delta_2',\ldots,\delta_k'-\delta_{k+1}')$, $M_X=I-X(X'X)^{-1}X'$ and SSR_k is the sum of squared residuals under the alternative hypothesis, which depends on $(T_1,\ldots,T_k)$.

The sup F type test statistic is then defined as

$$\sup F_T(k;q)=\sup_{(\lambda_1,\ldots,\lambda_k)\in\Lambda_\varepsilon}F_T(\lambda_1,\ldots,\lambda_k;q)=F_T(\hat{\lambda}_1,\ldots,\hat{\lambda}_k;q).$$

where the break fraction estimates $\hat{\lambda}_1,\ldots,\hat{\lambda}_k$ minimize the global sum of squared residuals and are also obtained from the maximization of the following F statistic:

$$F_T(\lambda_1,\ldots\lambda_k;q)=\frac{1}{T}\left(\frac{T-(k+1)q-p}{kq}\right)\hat{\delta}'R'\left(R\hat{V}(\hat{\delta})R'\right)^{-1}R\hat{\delta}, \qquad (3.6)$$

where $\hat{V}(\hat{\delta})=\left(\bar{Z}'M_X\bar{Z}/T\right)^{-1}$ is the covariance matrix of $\hat{\delta}$ assuming spherical errors. Different versions of these tests can be obtained depending on the assumptions made with respect to the distribution of the regressors and the errors across segments (Bai and Perron, 2000 and 2003a).

A test of structural stability versus an unknown number of breaks

Bai and Perron (1998) also consider tests of no structural change against an unknown number of breaks given some upper bound M for m. The following new class of tests is called double maximum tests and is defined for some fixed weights $\{a_1,\ldots,a_M\}$ as

$$\begin{aligned}D\max F_T(M,q,a_1,\ldots,a_M)&=\max_{1\leq m\leq M}a_m\sup_{(\lambda_1,\ldots,\lambda_m)\in\Lambda_\varepsilon}F_T(\lambda_1,\ldots,\lambda_m;q)\\&=\max_{1\leq m\leq M}a_m F_T(\hat{\lambda}_1,\ldots,\hat{\lambda}_m;q)\end{aligned} \qquad (3.7)$$

The weights $\{a_1,\ldots,a_M\}$ reflect the imposition of some priors on the likelihood of various numbers of structural breaks. Firstly, they set all weights equal to unity, i.e. $a_m = 1$ and label this version of the test as $UD\max F_T(M,q)$. Then, they consider a set of weights such that the marginal $p-$values are equal across values of m. The weights are then defined as $a_1 = 1$ and for $m > 1$ as $a_m = c(q,\alpha,1)/c(q,\alpha,m)$ where α is the significance level of the test and $c(q,\alpha,m)$ is the asymptotic critical value of the test $\sup_{(\lambda_1,\ldots,\lambda_m)\in\Lambda_\varepsilon} F_T(\lambda_1,\ldots,\lambda_m;q)$. This version is denoted

$$WD\max F_T(M,q) = \max_{1\le m\le M} \frac{c(q,\alpha,1)}{c(q,\alpha,m)} \times \sup_{(\lambda_1,\ldots,\lambda_m)\in\Lambda_\varepsilon} F_T(\lambda_1,\ldots,\lambda_m;q). \quad (3.8)$$

Test of l ***versus*** $l+1$ ***Breaks***

The last test developed by Bai and Perron (1998) is a sequential test of l versus l+1 structural changes:

$$\sup F_T(l+1|l) = \left\{ S_T(\hat{T}_1,\ldots,\hat{T}_l) - \min_{1\le i\le l+1}\ \inf_{\tau\in\Lambda_{i,\eta}} S_T(\hat{T}_1,\ldots,\hat{T}_{i-1},\tau,\hat{T}_1,\ldots,\hat{T}_l) \right\} \Big/ \hat{\sigma}^2, \quad (3.9)$$

where

$\Lambda_{i,\eta} = \left\{ \tau; \hat{T}_{i-1} + \left(\hat{T}_i - \hat{T}_{i-1}\right)\eta \le \tau \le \hat{T}_i - \left(\hat{T}_i - \hat{T}_{i-1}\right)\eta \right\}$,

$S_T(\hat{T}_1,\ldots,\hat{T}_{i-1},\tau,\hat{T}_1,\ldots,\hat{T}_l)$ is the sum of squared residuals resulting from the least-squares estimation from each $m-$partition $(T_1,\ldots,T_m)$,, and $\hat{\sigma}^2$ is a consistent estimator of σ^2 under the null hypothesis. The test amounts to the application of $(l+1)$ tests of the stability null hypothesis against the alternative hypothesis of a single break. It is applied to each segment $\left[\hat{T}_{i-1}+1, \hat{T}_i\right]$ for $i = 1,\ldots,l+1$, and with $\hat{T}_0 = 0$ and $\hat{T}_{l+1} = T$. We reject the null hypothesis and we conclude in favor of a model with $(l+1)$ structural breaks if the sum of squared residuals obtained from the estimated model with l changes is sufficiently larger than the overall minimal value of the sum of squared residuals (over all segments where an additional change is included) and the break point thus selected is the one associated with this overall minimum.

The asymptotic distributions of these three tests are derived in Bai and Perron (1998) and asymptotic critical values are tabulated in Bai and Perron (1998, 2003b) for ε = 0.05 (M = 9), 0.10 (M = 8), 0.15 (M = 5), 0.20 (M = 3), and 0.25 (M = 2). Note that these asymptotic distributions are derived without taking the trending regressors into account.

Regarding the selection of the model dimension, Bai and Perron (2003) suggested sequential model selection criteria that is based on the Bayesyen Information Criteria (BIC), LWZ criteria that is the modified version of the Schwarz criteria and finally the sequential supFT(l+1 | l) test that is developed by Bai-Perron.

4. Empirical Results

In this part of the study, an application regarding the use of Bai-Perron (1998) multiple structural break test as an alternative approach in the discretization of the continuous variable and the results obtained were discussed.

Incomes of 4946 workers participating in household budget survey done by TUIK in 2006 are used in this study. Income variables are used in this study in ascending sort. Descriptive statistics of continuous variables are given in Table 1.

Table 1. *Description of dataset*

	N	Minimum	Maximum	Mean	Std. Deviation
INCOME	4946	10.00	162000.00	5663.8993	5774.07720

Bai-Perron multiple structured breaks test results for income continuous variables are given in Table 2. According to the results acquired by BP test, the importance of $SupF_T(k)$ $(k = 1,\ldots,8)$, UDmax, WDmax and $SupF(l+1|l)$ tests are found at 1% level for continuous income variables. This result refers to the multiple breaks in the set.

As a consequence, it was determined that there was a break in the series, in other words, the series could be divided into at least two parts. Considering this consequence, supFT(l+1 | l) test was applied for the determination of the number of sections (parts-categories) in the variable. Although the BIC and LWZ criteria obtained a greater number of breaks, the sequential criteria found 3 break points in the income variable.

According to sequential supFT(l+1 | l) test, break points are found as 1519, 3551 and 4552 in the study in which three breaks are determined. Individuals corresponding the observations in the dataset sorted accordingly it; refer to the transition to new category. In other words, income variable is divided into 4 categories with these three break points: the lowest, low, high and the highest income levels.

Table 2. *Bai-Perron Multiple Structural Breaks*

Specification							
z_t = {1}	q=1	p=0	h=12	m=8	$\varepsilon = 0.10$		
sup F_T (1)	sup F_T (2)	sup F_T (3)	sup F_T (4)	sup F_T (5)	sup F_T (6)	sup F_T (7)	sup F_T (8)
106.9564*	92.0653*	74.1537*	66.9388*	79.3867*	41.7448*	33.0252*	46.7870*
UD max	*WD* max						
106.9564*	159.2634*						
sup F_T (2|1)	sup F_T (3|2)	sup F_T (4|3)	sup F_T (5|4)	sup F_T (6|5)	sup F_T (7|6)	sup F_T (8|7)	
58.3304	16.1389	4.7249	2.8169	12.1859	0.0011	0.0000	
Number of breaks selected by information criteria							
Sequential	**3**						
BIC	**7**						
LWZ	**5**						
The dates of the breaks ($\hat{T}$)							
1519	3551	4452					

Note: * significant at the level of 0.01.

Figure 1 shows the continuous variables and break points acquired at the end of Bai-Perron multiple structural breaks and categories together.

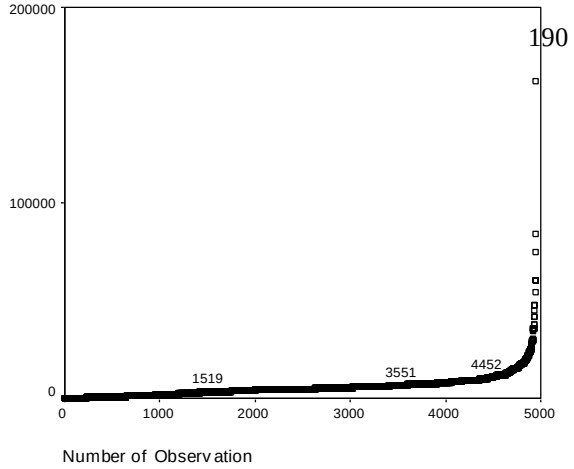

Figure.1 Discretization result

In Figure 1, the break points that were obtained by Bai-Perron multiple structural break test are marked on the series. In respect of its structure, the variable of income is a series where extreme values could be observed. These extreme values, which express low and high levels of income, could be observed from the break points. What needs to be paid attention here is that the series that is studied is not a time series, thus the break times in the variable which is arranged in an increasing way actually correspond to the observation number.

5. Conclusions

In this paper, we proposed a new way of discretizing numeric attributes, considering multiple structured break. As for econometric studies, use of categorical variables is preferred rather than continuous variables in many of them.

Categorical variables are especially effective while studying with ordered models. However, categorization of continuous variables is required by the data's own structure. In this study, Bai-Perron multiple structural break test is used for categorization of continuous variables. Although Bai-Perron is a break test developed for time series, it was not used for the discretization of any continuous variables in this study (without considering the fact that it is a time series or a cross-section data). It is more advantageous compared to other methods of discretization, since it can test the category number obtained in method-categorization statistically.

Bai-Perron multiple structured break tests are easy to apply because all it requires for users to do is to provide the maximum number of intervals. This method showed better performance than other traditional (for example equal width discretization) methods in most cases. Another benefit of this method is that it provides a concise summarization of

numeric attributes, an aid to increasing human understanding of the relationship between numeric features and the class attributes.

References

Bai, J., Perron, P., 1998. Estimating and Testing Linear Models with Multiple Structural Changes. Econometrica, 66, 47-78.

Bai, J., Perron, P., 2000. Multiple Structural ChangeModels: A Simulation Analysis, unpublished manuscript. Department of Economics, Boston University.

Bai, J., Perron, P., 2003a. Computation and Analysis of Multiple Structural Change Models. Journal of Applied Econometrics, 18, 1-22.

Bai, J., Perron, P., 2003b. Critical Values for Multiple Structural Change Tests. Econometrics Journal, 6, 72-78.

Cevik, E.I., Erdogan, S., 2009. Efficıency Of Bankıng Sector Stock Market: Structural Break And Long Memory. Doğuş Üniversitesi Dergisi, 10 (1), 26-40.

Christiano, L. J., 1992. Searching for a Break in GNP. Journal of Business and Economic Statistics, 10, 237-249.

Diebold, F. X., Chen, C., 1996. Testing Structural Stability with Endogenous Break Point: A Size Comparison of Analytic and Bootstrap Procedures. unpublished manuscript, Department of Economics, University of Pennsylvania.

Franses, P.H., Cramer, J.S. 2010. On the number of categories in an ordered regression model. Statistica Neerlandica, 64(1), 125-128.

Jouini, J., Boutahar, M., 2002. L'étude des modèles avec changements structurels. Document de Travail n◦ 02C01, GREQAM, Université de la Méditerranée, Marseille.

Jouini, J., Boutahar, M., 2003. Bootstrap Tests in Multiple Structural Change Models. E.D.G.E. (European Doctorate Group in Economics), University of Aix Marseille, p:28.

Lee, C.H. 2005. Discretizing Continuous Attributes Using Information Theory. Computer and Information Sciences – ISCIS, Lecture Notes in Computer Science, 3733, ,493-502.

Muhlenbach, F., Rakotomalala, R., 2005. Discretization for Continuous Attributes. In J Wang Editor. Encyclopedia of Data Warehousing and Mining, Idea Group Reference, 397-402.