

Enhancement of Non-stationary Time-Series Clustering Analysis Using Projection Pursuit Regression Method

Tsair-chuan Lin, Associate Professor and Chair of Department of Statistics, National Taipei University, Taiwan

ABSTRACT

The theory and application of time-series clustering analysis is an effective explanatory technique in various research fields. To overcome the limitations and many assumptions in conventional model-based clustering, this study utilizes the projection pursuit regression method as explanatory tool for formulating, identifying and estimating nonlinear models to approach the complex regression surface and then applied the projection pursuit method and an agglomerative scheme to cluster time series based on a similarity measure. This clustering can be applied to nonlinear, non-stationary, non-Gaussian models, and models involving interactions in predictor variables. Simulation results and real data analysis for categorizing the collection of average personal income of 25 states in the US demonstrate that this scheme compares favorably with other methods for similar clustering tasks.

INTRODUCTION

The study of time-series clustering analysis has garnered considerable attention in data mining, and its theory has been applied in fields such as biology, medicine, economy, finance, machine learning, signal analysis, and gene recognition. Cluster analysis is a general data-classification approach that researchers use to simplify and group data. Generally, the aim of clustering is to identify similar series based on certain characteristics such that elements within a cluster are highly similar and very dissimilar to elements in other clusters. Time-series clustering (Xiong, 2004) can be divided into two major classes: distance-based and model-based methods. The distance-based schemes assume that each series can be represented as a point in a multi-dimensional space of fixed dimensionality, and then, based on a distance measurement, the dataset is grouped. For instance, the well-known K-mean technique assumes the data point with its each dimension is independent; Euclidean distance is then applied for clustering. Unfortunately, no natural distance function exists for time-series data. Additionally, problems, such as unequal length, time delay, premature cutoff, and correlation among each dimension in a time series, will be overly emphasized by the distance measure.

Generally, model-based time-series clustering adopts a probability model or statistical model to describe the mechanism for generating data. Under assumptions regarding a joint probability density function for series data and prior information, the probability model clustering maximizes the posterior probability of a cluster model as the foundation of clustering. For instance, the Markov method and hidden Markov method (HMM) are two common probability-model-based clustering methods. The study of statistical-model-based clustering of time-series models usually assume realizations as linear parameterized stationary autoregressive (AR) or autoregressive and moving-average (ARMA) models, and then clusters time courses based on a similarity measure of a feature function based on fitted models. The definition of feature functions between stochastic models, such as the weighted Euclidean distance between the autoregressive model parameters (WAR), autocorrelation functions (ACF), principle component analysis (PCA), discrete Fourier transform (DFT), discrete wavelet transforms (DWT) and series spectral transform, are an important subject. Since statistical-model-based clustering methods consider the stochastic properties of a time series, the interpretations of their clustering results are better and more reasonable than those of probability-model-based methods.

In recent years, nonparametric time-series analysis has become a powerful statistical tool for exploring the underlying structure in a dataset and gained attention due to the limitations of autoregressive and moving-average models in describing various natural phenomena such as asymmetrical limit cycles, time irreversibility, amplitude-

dependent frequency, and chaos. Let $\{Y_t\}_{t=1}^n$ be a time series, the most general nonparametric p-th order autoregressive model can be defined as

$$Y_t = f(Y_{t-1}, Y_{t-2}, \dots, Y_{t-p}) + \varepsilon_t, \quad (1)$$

where the random error ε_t is i.i.d. with a zero mean and variance σ^2 . In a linear regression setup, one assumes the response surface f can be expressed as a linear combination of predictor variables. If inadequate, model scope can be extended by adding terms, such product or high-order predictors, to the model. Parametric regression models have advantages such as ease in computations, interpretations and forecasting. However, guessing which terms should be included in the function when many predictors exist in the model, and which one is the most appropriate functional form is difficult when just looking at the data.

Multivariate nonparametric smoothers only require a few assumptions; however, they frequently encounter the problem of “curse of dimensionality,” which is a neighborhood with a fixed number of points that become less local as the dimensions increase (Hastie and Tibshirani, 1990). Several nonparametric regression approaches have been developed in response to this dimensionality problem. Generalized additive models (GAMs),

$$Y_t = f_1(Y_{t-1}) + \dots + f_p(Y_{t-p}) + \varepsilon_t, \quad (2)$$

or projection pursuit regression (PPR) (Friedman and Switzer, 1981) overcomes this problem by using analogs of the Taylor expansion to approach a complex response surface. The mean function in GAMs is the summation of several univariate smooth and unknown functions $f_j(\cdot)$'s. Once the additive model is fitted to data, plots of smooth functions can be examined to assess the contributions of predictors in predicting a response. Although it can be applied to non-linear, non-Gaussian distributed or nonstationary cases, GAMs cannot deal with interactions between predictors. In such models, the projections are done onto individual predictors rather than onto a projection vector, which is the linear sum of the predictors, as in PPR. These projection vectors, instead of individual predictors, allow PPR to deal with interactions, which is the main property of PPR. Thus, this study applies the PPR model to overcome these problems, and estimate nonparametric models and clustering.

The remainder of this paper is organized as follows. The existing time-series clustering methods are introduced in Section 2. Section 3 describes the PPR clustering method and its basic concepts. This clustering method is simulated, and compared with other clustering analysis methods in Section 4. Average personal income in 1929–1999 in 25 US states is applied in Section 5.

LITERATURE REVIEW

Most time-series clustering studies applied the different models and assumptions to their proposed methods such as the HMM clustering method (Rabbiner et al., 1989), AR model-clustering method (Kalpakis et al., 2001; Xiong, 2004; Ramoni et al., 2002) and nonparametric model-clustering method (Luan et al., 2003; James et al., 2003; Ma et al., 2006).

HMM Clustering Method

The HMM is a double-layer random process that contains a hidden state layer and observable output layer. The HMM is a finite set of states, each of which is associated with a probability distribution. Transitions among states are governed by a set of probabilities called transition probabilities. An outcome or observation can be generated for a particular state according to the associated probability distribution. Although the hidden state layer cannot be observed directly, it can be estimated via aggregative estimation of the output sequence. The following three tasks typically occur when applying the HMM for clustering. (1) Given model parameters, compute the probability of a particular output sequence and the probabilities of the hidden state values given that output sequence. (2) Given model parameters, find the most likely sequence of hidden states that may have generated a given output sequence. (3) Given an output sequence or set of such sequences, find the most likely set of state transition and output probabilities. That is, identify HMM parameters given a dataset of sequences. The forward-backward procedure can be applied to reduce the number of calculations (Rabbiner et al., 1989); the Viterbi algorithm (Rabbiner et al., 1989) is applied to identify the most

probable hidden state under each time. The back-tracking method is then applied to identify the optimum route. The forward-backward calculation variable is employed in the Baum-Welch scheme is applied to solve the third problem.

Statistical Model Based Clustering

When the statistical-model-based clustering methods assume the series in the same cluster are from the same time-series model, such as common AR models, ARMA models or autoregressive integrated moving average (ARIMA) models. Ramoni et al. (2002) considered gene expression as an AR model, and applied the Bayesian method to perform clustering with an agglomerative algorithm. Xiong (2004) assumed data were mixtures of ARMA, and applied the Expectation Maximization (EM) algorithm (Dempster et al, 1976) to estimate model parameters and calculate maximum posterior probability for a clustering model. Kalpakis et al. (2001) adopted the ARIMA time-series model, and applied the linear predictive coding cepstrum (LPC) (Furui, 1989) coefficient to extract coefficients of the time-series model. Trend and seasonality components were first removed from data, since stationary data can be well fitted with an autoregressive model of a certain order; thus, the LPC can be acquired through estimated autoregressive coefficients. Sequentially, the cluster result can be obtained by applying the partitioning around method (PAM) to the LPC.

Nonparametric Model Clustering

Nonparametric regression allows one to estimate nonlinear fits between continuous variables with few assumptions about the functional space. This feature results in wide-ranging techniques that can be employed to numerous practical situations in diverse fields. In some current nonparametric clustering studies, the “time” factor is used as a predictive variable of a model; that is,

$$Y_t = f(t) + \varepsilon_t \quad (3)$$

is the generation mechanism producing the series in each cluster. Luan and Li (2003) applied this model to cluster time-series data, gene expression was analyzed by a nonparametric mixed-effects model, and a parameter model was obtained based on the B-splines (De Boor, 1978) transform. One crucial step in this clustering method is to apply the EM algorithm to obtain maximum likelihood estimates; optimum clusters can then be determined by the Bayesian information criteria (BIC) evaluation. Ma et al. (2006) also demonstrated that gene expression changes over time; thus, different gene sequences have different characterization functions, and each gene can have the same stochastic effect in the same clusters. Notably, nonparametric regression was applied to estimate the mean function of each series during clustering. James et al. (2003) proposed a clustering procedure that is applicable to various curve data but is especially useful when individuals are observed at a sparse set of time points. However, the independent variable of these nonparametric clustering methods is only time, t .

PPR CLUSTERING METHOD

The primary concept underlying projection pursuit regression (PPR) is as follows. Let Y and $\underline{X} = (X_1, X_2, \dots, X_p)'$ be the response and explanatory vectors, respectively. Suppose one has observations y_i and corresponding predictors $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})'$, $i = 1, \dots, n$. Let $\alpha_1, \alpha_2, \dots, \alpha_{M_0}$ be p -dimensional unit “directional” vectors, and let $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$. The PPR function locates the $M = M_0$ directional vectors $\alpha_1, \alpha_2, \dots, \alpha_{M_0}$ and a good nonlinear transformation f_1, f_2, \dots, f_{M_0} , such that

$$y \approx \bar{y} + \sum_{m=1}^{M_0} \beta_m f_m(\alpha_m^T x)$$

provides a good model for data. Formally, y and x are assumed to satisfy the conditional expectation model,

$$E(Y | X_1, \dots, X_p) = \mu_y + \sum_{m=1}^{M_0} \beta_m f_m(\alpha_m^T \underline{X}) \quad (4)$$

where f_m has been standardized to have mean zero and unity variance:

$$E f_m(\alpha_m^T x) = 0, \quad E^2 f_m(\alpha_m^T x) = 1, \quad m = 1, \dots, M_0.$$

Model parameters $\beta_m, f_m, a_m, m = 1, \dots, M_0$ in Eq. (4) minimize mean squared error

$$E^2[Y - \mu_y - \sum_{m=1}^{M_0} \beta_m f_m(\alpha_m^T X)]^2. \quad (5)$$

For instance, suppose $E(Y|X_1, X_2) = X_1 X_2$. This is described by (4) with $\mu_y = 0, M_0 = 2, \beta_1 = \beta_2 = 1/4, a_1 = (1, 1)^T, a_2 = (1, -1)^T, f_1(x) = x^2, f_2(x) = -x^2$.

Thus, this PPR model is a simple interaction model. Due to the limitation of the space, researchers interested in the PPR estimation method can refer to Peter (1985), Friedman and Switzer (1981).

To apply the PPR model for time-series clustering, we assume the series in the same cluster are generated by the same model. Hence, if two sequences, S_i and S_j , are similar, the fitted models m_i and m_j can also reflect the similar structural relationship. Furthermore, if such a similar fitted model relationship exists, these fitted models should have similar predictive results for any sequence in two sequences. Thus, if predictive vector \hat{y}_j is acquired by predictive series S_j through fitted model m_j , a similar predictive vector, $\hat{\hat{y}}_j$, can be deduced through the m_i fitted model of this predictive vector; and *vice versa*. Therefore, clustering can be processed through recursively searching two similar series or cluster according their fitted models. Cross validation (CV) is a model-selection method that takes the predictive ability of a model as the basis for model selection. The basic purpose of CV is to divide a dataset into two parts—a training set and testing set. For each series, this work fits a PPR model using training data, and validates the model using testing data. Thus, the concept of CV is applied to determine the similarity between two fitted models during the clustering process. Let $\{S_1, \dots, S_n\}$ be the set of time-series data that is divided into k clusters $\{C_1, \dots, C_k\}$; the PPR clustering method is described as follows:

Initiate: Set the initial clusters $C_i = \{S_i\}, \forall 1 \leq i \leq n$.

For any two clusters, $\{C_i, C_j\}, \forall 1 \leq i < j \leq n$, and calculate the CV value to determine the similarity between any two series:

(a) Thus, the PPR method is individually applied to fit $\{C_i, C_j\}$ for obtaining the fitted model $\{m_i, m_j\}$ and fit vectors $\{\hat{y}_i, \hat{y}_j\}$.

(b) Model m_i is applied to predict \hat{y}_j for obtaining $\hat{\hat{y}}_j$; at the same time, model m_j is applied to predict \hat{y}_i for obtaining $\hat{\hat{y}}_i$.

(c) Define $CV(i, j) = \frac{\sum_{i=1}^{n_i-p} (\hat{y}_i - \hat{\hat{y}}_i)^2 + \sum_{j=1}^{n_j-p} (\hat{y}_j - \hat{\hat{y}}_j)^2}{n_i + n_j - 2p}$, where n_i is the summation of all sequence lengths in cluster C_i .

Define the upper triangular matrix $CV = [CV(i, j)]$ of size $n \times n$.

Select two clusters, (C_{i^*}, C_{j^*}) , for meeting $(i^*, j^*) = \text{arg}(\min_{i < j} CV_{i,j})$ merged to the same clusters. Additionally, take CV_{i^*, j^*} as the CV^* of this iteration.

Cycle. Repeat steps 2–4 until the data merges as a cluster.

The CV^* value is calculated during clustering, as CV is a measurement of predictive ability of a fitted model. If the optimum number of clusters is obtained, variation between clusters increases and variation within clusters decreases; thus, the CV^* value increase markedly in the next recursion, and the optimum number of clusters can be determined based on these phenomena. To evaluate clustering quality or compare clustering results with other methods proposed for similar time series clustering tasks, this work uses the cluster similarity measure developed by Gavrilov (2000) to assess the performance of clustering methods. Given two clustering sets, $G = (G_1, \dots, G_c)$ and $A = (A_1, \dots, A_c)$, the cluster similarity measure is defined by

$$sim(G, A) = \frac{1}{C} \sum_{i=1}^C \max_{1 \leq j \leq C} Sim(G_i, A_j) \quad (6)$$

where $sim(G_i, A_j) = 2|G_i \cap A_j| / (|G_i| + |A_j|)$, G is the clustering for the “ground truth” and A is obtained by a cluster method under evaluation.

SIMULATION

The data generated under various conditions are considered in this section. Furthermore, the PPR clustering method and CV are applied to determine the number of clusters and obtain clustering results. The models are defined and labeled as follows:

$$\begin{aligned} M 1 : & (1 - \phi B)Y_t = \varepsilon_t, \\ M 2 : & (1 - \phi_1 B - \phi_2 B^2)Y_t = \varepsilon_t, \\ M 3 : & (1 - \phi_1 B - \phi_2 B^2 - \phi_3 B^3)Y_t = \varepsilon_t, \\ M 4 : & (1 - \phi_1 B - \phi_2 B^2)(1 - B)^d Y_t = \varepsilon_t, \\ M 5 : & (1 - \phi_1 B - \phi_2 B^2)Y_t = \beta t + \varepsilon_t, \\ M 6 : & Y_t = \beta_0 + \beta_1 Y_{t-1} Y_{t-2} + \varepsilon_t, \\ M 7 : & Y_t = \beta_0 + \beta_1 Y_{t-1} Y_{t-2} Y_{t-3} + \varepsilon_t, \end{aligned}$$

where M1–3 are the AR models; M4 is the non-stationary ARFIMA (2, d, 0) model; M5 contains the trend item; M6 and M7 interact in predictor variables. The number of clusters is set at 3; 5 time series are produced for each cluster in these models. The details of experiment are as follows:

Example 1. (Stationary model) linear AR (1) data.

The linear and stationary AR models M1 with coefficients ϕ of 3 clusters are 0.2, -0.2 and 0.8.

Example 2. (Mixture) linear AR (2) data

The linear AR models M2 with coefficient vectors (ϕ_1, ϕ_2) of 3 clusters are (0.2, -0.1), (-0.2, 0.7) and (0.8, 0.8).

The first and the second clusters are stationary and the third cluster is a non-stationary model.

Example 3. (Stationary model) linear AR (3) data

The linear and stationary AR model M3 with coefficient vectors (ϕ_1, ϕ_2, ϕ_3) of 3 clusters are (0.2, -0.1, 0.8), (-0.2, 0.7, 0.2) and (0.8, 0.8, -0.7), respectively.

Example 4. (Non-stationary model) ARFIMA (2,d,0)

For the non-stationary time-series model M4, generate ARFIMA (2, d, 0) time-series data for 3 clusters, where parameter vectors (ϕ_1, ϕ_2, d) are (.2, -.1, .7), (-.2, .7, .49) and (.8, .8, 1.5), respectively.

Example 5. (Non-stationary model) with time trend data

For time-series model M5 with a trend component, the AR coefficient vectors (ϕ_1, ϕ_2, β) of the 3 clusters are (.2, .1, .1), (-.2, .7, .1) and (.8, .8, -.1), respectively.

Example 6. (Non-stationary model) with second-order interaction in predictor variables.

For time-series model M6 with an interaction item, coefficient vectors (β_0, β_1) of the 3 clusters are (0,1), (.1,1) and (-.24,.2), respectively.

Example 7. (Non-stationary model) with third-order interaction in predictor variables.

The order of interaction items in the generated model is 3, where coefficient vectors (β_0, β_1) of three clusters are (0, 1), (.1, 1) and (-.24, .2), respectively.

In simulations, the length of the time series is T and 100 iterations are applied to obtain average and standard deviation for clustering quality similarity (*Sim*). If the length of a generated time series is long for a given clustering model, its clustering similarity is high (Table 1). Additionally, if the length of a time series T is 400, its average clustering similarity is 0.9744. Furthermore AR(2) and AR(3) are generated in M2–5, and the data length is only 100 under a non-stationary scenario; thus, the similarity of the clustering simulation result is 0.97–1 (Table 2). Notably, if

the true model involves predictor interactions, even when data length is only 10, this clustering method can also determine that average clustering similarity is close to 1 (Table 3).

Table 1: Clustering similarity (Sim) of the models in Example 1.

Quantity	length T=100	length T=200	length T=400
Average (Sim)	0.6777	0.8161	0.9743
S.D.(Sim)	0.0707	0.1598	0.0830

Average (Sim) and S.D.(Sim) mean the average and standard deviation, respectively, of clustering quality similarity 100 iterations.

Table 2: Clustering similarity (Sim) of models in Examples 2–5.

Model	Quantity	length T=30	length T=50	length T=100
M2	Average(Sim)	0.7970	0.8947	0.9888
	S.D.(Sim)	0.1559	0.1348	0.497
M3	Average(Sim)	0.8187	0.9461	0.9841
	S.D.(Sim)	0.1608	0.1065	0.0616
M4	Average(Sim)	0.7099	0.8689	0.9767
	S.D.(Sim)	0.1678	0.1422	0.0700
M5	Average(Sim)	0.8235	0.9526	1.0000
	S.D.(Sim)	0.1544	0.0931	0

Table 3: Clustering similarity (Sim) of models M6 and M7

Model	Quantity	length T=30	length T=50	length T=100
M6	Average(Sim)	0.992	0.9858	1.0000
	S.D.(Sim)	0.0450	0.0609	0
M7	Average(Sim)	0.9582	0.9537	0.9888
	S.D.(Sim)	0.1145	0.1205	0.0634

To investigate the change in CV^* and determine the number of clusters, time series are generated based on the real number of clusters mixed with different models with 3–5 clusters. The following examples are used for explanation.

Example 8.

The real data are 5 series for 3 clusters of AR (2) models (total, 15 series), and the length of each time series is 100. Calculate the CV^* value of each recursion. Figure 1(a) shows the CV^* of each recursion. One can see from this plot that $CV^*(12) = 2.8206$ and $CV^*(13) = 6.0772$; thus, the CV^* value increases at the 12th iteration, and the optimum number of clusters is 3. This experimental result is same as the real number of clusters.

Example 9.

The data set contains 4 clusters, 3 AR(2) and 1 non-stationary ARFIMA time-series datum. There are five series with series lengths of $T=100$ for each of these 4 clusters. Figure 1(b) presents the change in the CV^* at each step of the agglomerative clustering procedure. The CV^* value in the first 16 steps remain roughly stable. In step 17, the CV^* value increases markedly; thus, the merging procedure should stop at step 16 and return the 4 clusters. This experimental result is same as the real number of clusters.

Example 10.

The data contains 5 clusters, 3 AR(2), 2 clusters of ARFIMA time series data, 3 series of each cluster (total, 15 time series) with a length $T=100$. Figure 1(c) lists the changes to the CV^* value at each step of the agglomerative search procedure. In the first 10 steps, the CV^* value increases slowly and linearly. In step 11, the CV^* value increases significantly; thus, that the merging procedure should stop at step 10 and return the 5 clusters. This experimental result is same as the real number of clusters.

Example 11.

Repeat the simulation of Example 3; however, each cluster is increased to 5 series (total, 25 sequences of time series data) and length of each time series is 100. The CV^* value increases obviously between $CV^*(20) = 3.0224$

and $CV^*(21) = 4.1038$ (Fig. 1(d)); thus the merging procedure should stop at step 20 and return 5 clusters. Notably, the optimum number of clusters is 5, which is same as the real number of clusters.

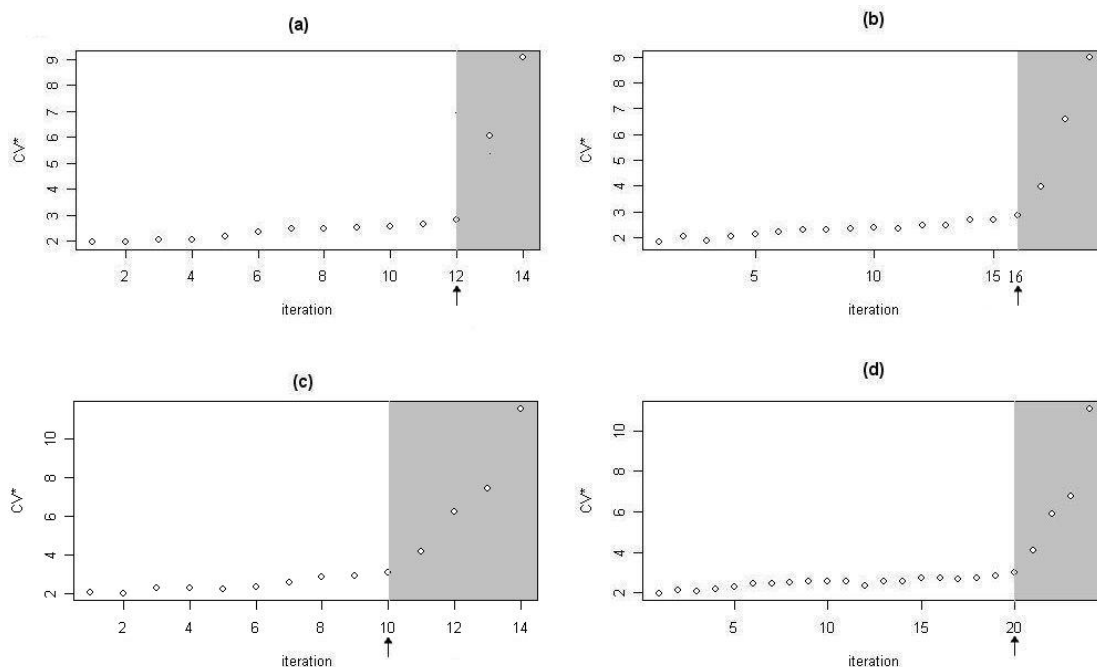


Figure 1: The CV^* of clustering in (a) Example 8, (b) Example 9, (c) Example 10, and (d) Example 11, with various numbers of clusters. The arrow indicates the real number of clusters.

To determine whether this technique improves existing clustering results, the proposed method is compared with WAR (Kalpakis, 2001), linear prediction cepstrum (LPC) (Kalpakis, 2001), and Bayesian AR clustering methods (BAR)(Ramoni, 2002; Medvedovic and Sivaganesan, 2002). For the same simulated data set, the WAR, LPC, BAR and PPR methods are applied for clustering and calculating similarity. Table 4 summarizes the performances of these methods based on 100 iterations during simulations. Experimental results demonstrate that the PPR clustering method compares favorably with other methods for similar time-series clustering tasks.

The data for demonstration analysis are personal annual average income for 25 states in the US in 1929–1999. Some economists divided high and low growth rates in personal income in 25 states into two groups. The first cluster includes 17 states (CT, DC, DE, FL, MA, ME, MD, NC, NJ, NY, PA, RI, VA, VT, WV, CA, IL) on the eastern seacoast; California and Illinois are areas with high growth rates in personal income. The second cluster includes 8 inland states (ID, TA, IN, KS, ND, NE, OK, and SD); these states are areas with low growth rates in personal income. In cluster analysis, this work applied the PPR method to group these 25 series into two clusters. In Table 5, p and M adopt the number of different candidate values used in the PPR clustering to assess the impact to performance (Table 5). The clustering similarity was 0.762–0.802; thus, analytical results did not change obviously.

Table 4: The similarity of models in Example 1–7 with various clustering methods.

Model	WAR	LPC	BAR	PPR
M1	0.94	0.89	1.0	0.94
M2	0.94	0.87	0.98	0.96
M3	0.95	0.91	0.97	0.99
M4	0.96	0.87	0.99	0.99
M5	0.80	0.73	0.89	0.98
M6	0.77	0.78	0.71	1.0
M7	0.93	0.85	0.93	0.98

The “WAR”, “LPC”, and “BAR” indicate the weighted Euclidean distance between the autoregressive model parameters, linear prediction cepstrum (Kalpakis, 2001), and Bayesian autoregressive clustering methods (BAR)(Ramoni, 2002; Medvedovic and Sivaganesan, 2002), respectively.

Table 5. Clustering quality of personal income for 25 US states with various orders in autoregressive model fits.

order	$M_0=1$	$M_0=2$	$M_0=3$	$M_0=4$	$M_0=5$
$p=2$	0.79	0.76	0.76	0.76	0.76
$p=3$	0.79	0.79	0.79	0.79	0.79
$p=4$	0.79	0.79	0.79	0.80	0.79

The order p and M_0 are the parameters used in equation (4).

REAL DATA ANALYSIS

The number of clusters is examined further. The arithmetic recursion is applied 24 times until all data are in one cluster, and the CV^* value of each time is calculated for each recursion. The CV^* value increases obviously between $CV^*(23)=374150.37$ and $CV^*(24)=1744352.57$ (Fig. 2); hence the optimum number of clusters is 2. Notably, this number of clusters agrees with the statement of 2 kinds of growth rates in average personal income by the economists.

Kalpakis (2001) first assumed all personal income data are from ARIMA models, and applied the different methods—LPC, DFT, DWT, PCV and MSE—to extract the coefficients, and then took the Euclidean distance between coefficients as the basis for clustering. Table 6 compares clustering results. The PPR clustering similarity is 0.81, which approaches the current optimum clustering result.

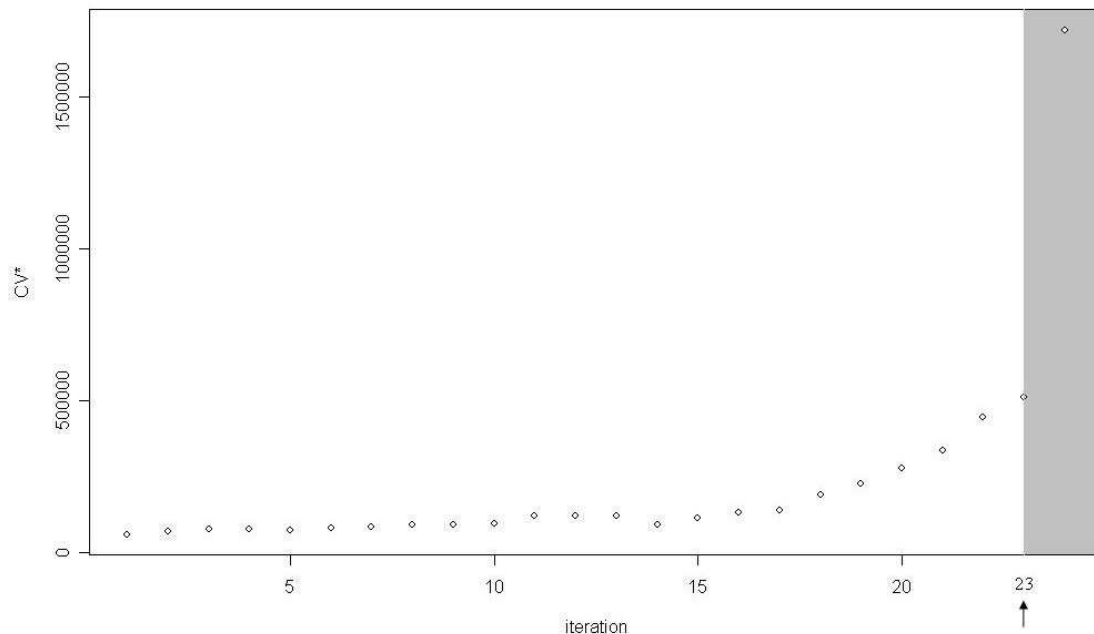


Figure 2: The CV^* value for personal income in 25 US states.

Table 6: Similarity of personal incomes in 25 US states clustered by various methods.

Method	PPR	LPC	DFT	DWT	PCV	MSE
Sim	0.81	0.84	0.68	0.60	0.68	78

CONCLUSION

Under any condition, the length of the time series is long in simulation analysis and, thus, the clustering effect is optimal. If real data are AR (1), the length of time series should be 400, and clustering similarity will approach 1; however, under other conditions, only time length is 100, and then clustering similarity will approach 1, even when the real model has interactive predictors.

Compared with conventional methods, if time-series data are the simplest linear data AR (1) because the model is excessively simple, the nonparametric model is applied for clustering simulation, and the effectiveness of the PPR

clustering method is not good relative to other methods. If time-series data are AR (2) and AR (3), no obvious difference exists between the PPR method and other methods. If there are trends and interactive items in the data, the PPR clustering method is better than other methods.

In real data analysis, such as that for personal income data for 25 US states in 1929–1999, the optimum cluster size 2, which was obtained by the PPR method proposed in this study for determining the number of clusters, is same as that put forward by the economist. The PPR clustering similarity is 0.81, which approaches the current optimum clustering result.

ACKNOWLEDGMENT

The authors would like to thank Ted Knoy for his editorial assistance.

REFERENCES

- De Boor, C., (1978) "A Practical Guides to Splines." Springer.
- Dempster A.P., Laird N., and Rubin D.B, (1976) "Maximum likelihood estimation from incomplete data using the EM algorithm (with discussion)." J. Roy. Statist. Soc. Series B, 39, pp. 1–38.
- Friedman J. H., and Stuetzle W., (1981) "Projection Pursuit Regression," Journal of the American Statistical Association, Vol. 76, pp. 817-823.
- Furui S., (1989) "Digital Speech Processing, Synthesis, and Recognition." Marcel Dekker, Inc., New York.
- Gavrilov M. D., Anguelov P. I., and Motwani R., (2000) "Mining the stock market: Which measure is best?" In Proc. of the KDD, pp. 487-496.
- Hastie T. J. and Tibshirani R. J., (1990) "Generalized Additive Models." Chapman & Hall, London.
- Huber Peter J., (1985) "Projection Pursuit." The annals of statistics, Vol. 13, No. 2., pp. 435-475.
- Kalpakis K, Gada D. and Puttagunta V., (2001) "Distance measure for effective clustering of ARIMA time series." In proc. of the IEEE ICDM, pp. 273-280.
- Luan Y., and Li H., (2003) "Clustering of time-course gene expression data using a mixed-effects model with B-splines." Bioinformatics. Vol 19, No. 4, pp. 474-482.
- Luan Y., and Li H., (2004) "Model-based methods for identifying periodically regulated genes based on the time course microarray gene expression data." Bioinformatics, 20, pp. 332–339.
- Ma P., Castillo-Davis C. I., Zhong W., and Liu J.S., (2006) "A data-driven clustering method for time course gene expression data. Nucleic Acids Res. Vol.34, No. 4, pp. 1261–1269.
- Medvedovic M., and Sivaganesan S., (2002) "Bayesian infinite mixture model based clustering of gene expression profiles." Bioinformatics, Vol.18, No.9, pp 1194–1206.
- Rabiner, L. R. (1989) "A tutorial on hidden Markov models and selected applications in speech recognition." Proc. of the IEEE, Vol. 77, No.2, pp. 257-286.
- Rabiner L. R., Lee C. H., Juang B. H, and Wilpon J. G., (1989) "HMM clustering for connected word recognition". IEEE. Int. conf. Acoust., Speech, Signal Processing, 1989, pp. 405-408.
- Ramoni M. F., Sebastiani P., and Kohane I. S., (2002) "Cluster analysis of gene expression dynamics." Proc Nat Acad Sci USA , Vol. 99, No. 14, pp. 9121-9126.
- Xiong Y., and Yeung D. Y., (2004) "Time series clustering with ARMA mixtures." Pattern Recognition, Vol. 37, No. 8, pp. 1675-1689.