Feature Extraction from POS Transaction Data by Using Local Independent Components *

Katsuhiro Honda, and Hidetomo Ichihashi
Graduate School of Engineering, Osaka Prefecture University,
Gakuen-cho 1-1, Sakai, Osaka, 599-8531, Japan

Abstract

Independent component analysis (ICA), which has been developed mainly in signal processing, is a useful technique for Projection Pursuit as well. For some nonlinearly distributed data, the data set is partitioned into several groups using Fuzzy c-Varieties clustering method before applying the ICA algorithm, which constitute a fuzzy version of Fast ICA by Hyvärinen *et al.*. This paper discusses feature extraction from local independent components by applying the techniques to POS (point-of-sales) transaction data.

1 Introduction

Independent Component Analysis (ICA) is an unsupervised technique, which in many cases characterizes data in a natural way, and is a useful technique for Projection Pursuit as well [1]. In the general formulation of ICA, the purpose is to transform an observed vector linearly into the vector whose components are statistically as independent from each other as possible. The mutual dependence of the components is classically measured by their non-Gaussianity. Maximizing the non-Gaussianity gives us one of the independent components. Projection Pursuit is also a technique developed in statistics for finding "interesting" features of multivariate data. In Projection Pursuit, the goal is to find the one-dimensional projections of multivariate data which have "interesting" distributions for visualization purposes. Typically, the interestingness is measured by the non-Gaussianity. Therefore, the basis vectors of ICA should be especially useful in Projection Pursuit and in extracting characteristic features from natural data.

In spite of its usefulness, the linear ICA models are often too simple to describe real-world data and provide only a crude approximation for general nonlinear data distributions. Karhunen *et al.* proposed local ICA models [2] that were used in conjunction with some suitable clustering algorithms. In the local ICA models, the data are grouped into several clusters based on the similarities between the observed data, ahead of the preprocessing of linear ICA, by using some hard clustering algorithms such as k-Means algorithm. However, the observed data are assumed to be the linear combinations of source signals in linear ICA models. So the clustering methods that partition the data into some spherical clusters like by k-Means are not suitable for the extraction of local independent components, and the data set should be divided into linear clusters.

A technique that uses Fuzzy c-Varieties (FCV) clustering method [3] for extracting local independent components is proposed [4]. The FCV algorithm partitions an observed data set into linear fuzzy clusters based on the similarities of the mixing matrices. Because FCV can be regarded as a simultaneous approach to clustering and Principal Component Analysis (PCA), the FCV algorithm also performs the preprocessing of Fast ICA proposed by Hyvärinen *et al.* [5].

In this paper, we apply the ICA algorithms to a POS (point-of-sales) transaction data and compare the results derived by them.

2 ICA Formulation and Fast ICA Algorithm

Denote that $\boldsymbol{v}$ is the M dimensional observed data vector and $\boldsymbol{s}$ is the N dimensional source signal vector corresponding to the observed data with $N \leq M$. Under the constraints that the elements of source signals

*Knowledge-Based Intelligent Information Engineering Systems & Allied Technologies, ed. by N. Baba, L. C. Jain and R. J. Howlett, IOS Press, 27-31 (2001)

$(s_1, s_2, \cdots, s_N)$ are mutually statistically independent and have zero-means, the observed data are assumed to be the linear mixtures of s_i as follows:

$$\boldsymbol{v} = \boldsymbol{A}\boldsymbol{s}, \tag{1}$$

where the unknown $M \times N$ matrix $\boldsymbol{A}$ is called the mixing matrix. In ICA, we try to estimate the source signals s_i and the mixing matrix $\boldsymbol{A}$ using only the observed data $\boldsymbol{v}$.

Fast ICA algorithm proposed by Hyvärinen *et al.* [5] is a useful algorithm that is very simple, does not depend on any user-defined parameters, and is fast to converge to the most accurate solution allowed by the data. Generally, a preprocessing of whitening and sphering by using PCA is applied. In the preprocessing, the observed data $\boldsymbol{v}$ are transformed into linear combinations $\boldsymbol{x}$,

$$\boldsymbol{x} = \boldsymbol{M}\boldsymbol{v} = \boldsymbol{M}\boldsymbol{A}\boldsymbol{s} = \boldsymbol{B}\boldsymbol{s}, \tag{2}$$

such that its elements $(x_1, x_2, \cdots, x_N)$ are mutually uncorrelated and all have unit variance, and $\boldsymbol{B} = \boldsymbol{M}\boldsymbol{A}$ is an orthogonal matrix. Thus we can reduce the problem of finding an arbitrary full-rank matrix $\boldsymbol{A}$ to the simpler problem of finding an orthogonal matrix $\boldsymbol{B}$, which gives $\boldsymbol{s} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{x}$.

To derive the matrix $\boldsymbol{B}$, Hyvärinen proposed the following objective function to be minimized or maximized.

$$J(\boldsymbol{w}_i) = E\{(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{x})^4\} - 3\|\boldsymbol{w}_i\|^4 + F(\|\boldsymbol{w}_i\|^2), \tag{3}$$

where $\boldsymbol{w}_i$ corresponds to one of the columns of the mixing matrix $\boldsymbol{B}$. The first two terms represent the fourth-order cumulant or kurtosis of the reconstructed signals which is a classical measure of non-Gaussianity. Maximizing the non-Gaussianity of the reconstructed signals gives us one of the independent components. Using the deviation of the objective function, we can obtain the fixed-point algorithm for ICA.

3 Extraction of Local Independent Components by Using Clustering

Even though linear ICA yields meaningful results in many cases, it can provide a crude approximation only for general nonlinear data distributions. Therefore, several techniques where local ICA models were applied to the data grouped by using some suitable clustering methods have proposed.

Karhunen *et al.* proposed local ICA models [2] that were used in conjunction with some hard clustering algorithms such as k-Means. k-Means is a clustering method that groups a data set into C spherical clusters and tries to minimize the criterion,

$$L = \sum_{c=1}^{C} \sum_{\boldsymbol{v}_k \in S_c} ||\boldsymbol{v}_k - \boldsymbol{m}_c||^2, \tag{4}$$

where $\boldsymbol{m}_c$ represents the mean vector of the cth cluster S_c. The minimum is achieved when the data vectors are divided into C clusters so that the overall mean-square error estimated from the data vectors is the smallest possible. Once we partition the data into C clusters, we apply linear ICA algorithms such as Fast ICA algorithm to each cluster.

Another technique for the extraction of local independent components uses the FCV algorithm [3] which are useful for the grouping of observed data taking the similarities of the mixing matrices into account [4]. The goal of FCV is to determine the memberships of the data in C fuzzy clusters $\boldsymbol{u}_c = (u_{c1}, u_{c2}, \cdots, u_{cJ})^{\mathrm{T}}, c = 1, \cdots, C$ and the M dimensional principal component vectors of each cluster $\boldsymbol{p}_{ci} = (p_{ci1}, p_{ci2}, \cdots, p_{ciM})^{\mathrm{T}}, c = 1, \cdots, C, i = 1, \cdots, N$. Where J represents the number of the observed data. The memberships $\boldsymbol{u}_c$ are constrained as follows:

$$\boldsymbol{u}_c = \{(u_{ck}) | \sum_{c=1}^{C} u_{ck} = 1, u_{ck} \in [0, 1], k = 1, \cdots, J\}. \tag{5}$$

By using Lagrange's method of indeterminate multiplier, the objective function with entropy regularization [6] is defined as follows:

$$\begin{aligned}
\min\ L &= \sum_{k=1}^{J}\sum_{c=1}^{C} u_{ck}\Big\{(\boldsymbol{v}_k - \boldsymbol{m}_c)^{\mathrm{T}}(\boldsymbol{v}_k - \boldsymbol{m}_c) - \sum_{i=1}^{N} \boldsymbol{p}_{ci}^{\mathrm{T}} R_{ck} \boldsymbol{p}_{ci}\Big\} \\
&\quad + \sum_{c=1}^{C}\sum_{i=1}^{N} \lambda_{ci}(\boldsymbol{p}_{ci}^{\mathrm{T}}\boldsymbol{p}_{ci} - 1) + \alpha \sum_{k=1}^{J}\sum_{c=1}^{C} u_{ck}\log u_{ck} + \sum_{k=1}^{J} \gamma_k \Big(\sum_{c=1}^{C} u_{ck} - 1\Big), \\
R_{ck} &= (\boldsymbol{v}_k - \boldsymbol{m}_c)(\boldsymbol{v}_k - \boldsymbol{m}_c)^{\mathrm{T}},
\end{aligned}$$

Table 1: Average Numbers of Customers for Each Day of Week

day of week	Mon.	Tues.	Wed.	Thur.	Fri.	Sat.	Sun.
supermarket A	613.4	693.7	439.4	827.6	592.0	748.6	720.8
supermarket B	595.9	671.0	478.3	652.0	562.1	750.3	759.9

where λ_{ci} and γ_k are Lagrangian multipliers and α is a positive constant. The entropy term is for fuzzification. The larger the value of α is, the fuzzier the membership assignments become.

Because the FCV algorithm is a simultaneous application of PCA and fuzzy clustering, we can also perform the preprocessing of the Fast ICA algorithm at the same time. After the preprocessing, a modified Fast ICA algorithm named Fuzzy Fast ICA is applied. Fuzzy Fast ICA extracts local independent components from each fuzzy cluster by using following objective function.

$$J_c(\boldsymbol{w}_c) = \frac{\sum_{j=1}^{J} u_{cj}(\boldsymbol{w}_c^{\mathrm{T}} \boldsymbol{x}_{cj})^4}{\sum_{j=1}^{J} u_{cj}} - 3\|\boldsymbol{w}_c\|^4 + F(\|\boldsymbol{w}_c\|^2),$$

where the first two terms represent fuzzy kurtosis defined by considering the memberships $\boldsymbol{u}_c$ derived by the FCV algorithm.

4 Feature Extraction from POS Transaction Data Set

We applied the ICA algorithms introduced in the previous sections to a POS data set to extract meaningful characteristic features and compare the results. The POS data set collected in 1997 at two supermarkets in Osaka is composed of 352 transactions with 9 values: 7 days of the week and the numbers of customers in each supermarket.

First, we applied Hyvärinen's Fast ICA algorithm and extracted two independent components. Figure 1 shows the projection onto the two-dimensional space spanned by the two independent components (IC1 and IC2). In the figure, the horizontal and the vertical axes were named based on the correlations between the independent components and the number of customers. For example, the first independent component (IC1) had positive correlation with the number of customers in both the two supermarkets. The data points are distributed on a line and we cannot extract useful knowledge. It is because the distribution of the data is too complicated to analyze with a single model.

Second, we analyzed the data both by local ICA using k-Means and Fuzzy Fast ICA using FCV. Figure 2 shows the projections of independent components obtained by local ICA and Figure 3 shows those by Fuzzy Fast ICA respectively. In the figures, "A-busy" ("B-busy") means supermarket A (B) had many customers while supermarket B (A) didn't have large correlation with the independent component, and vice versa. Figure 2-a and 3-a show the characteristic features that are common to both the two supermarkets. Figure 2-a (3-a) indicates that the number of customers is large on Monday (Sunday and Tuesday) but small on Wednesday and Friday (Wednesday and Friday). On the other hand, Figure 2-b and 3-b show the respective characteristics of each supermarket. In Figure 2-b, we can detect the following features. The number of customers is large at supermarket A but small at supermarket B on Thursday while it is large at supermarket B but small at supermarket A on Sunday. Figure 3-b shows that the number of customers is large at supermarket A but small at supermarket B on Thursday.

Finally, we compared the characteristic features extracted from local independent components with the average numbers of customers. Table 1 shows the average numbers of customers for each day of the week. The number of customers is large on Saturday and Sunday but small on Wednesday and Friday in both the supermarkets while it is large at supermarket A but small at supermarket B on Thursday. These features have apparently appeared from local independent components. Figure 3 shows the features more clearly. The preprocessing using the FCV algorithm is more effective than that using the k-Means algorithm for the feature extraction from the POS transaction data.

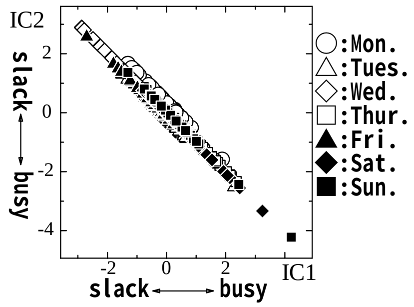
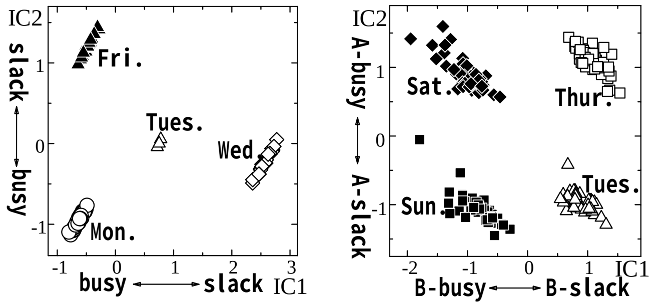
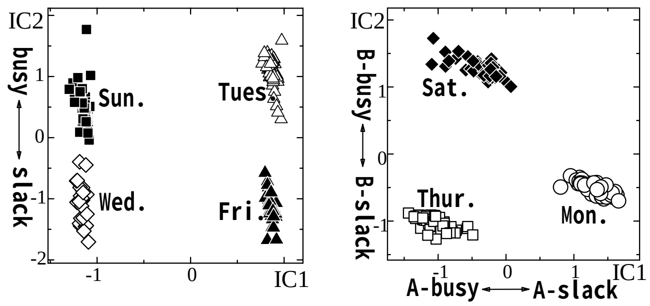


Figure 1: Projection of Independent Components derived by Fast ICA



a. Cluster 1 b. Cluster 2

Figure 2: Projections of Independent Components derived by Local ICA



a. Cluster 1 b. Cluster 2

Figure 3: Projections of Independent Components derived by Fuzzy Fast ICA

5 Conclusion

We have applied the ICA algorithms that extract local independent components by using clustering methods to a POS data set to find some meaningful characteristic features. It is a difficult task to search for interesting projections of multivariate real world data which have nonlinear data distribution, but we could successfully obtain useful knowledge from local models in each cluster. The results showed that the FCV algorithm is more suitable than the k-Means algorithm. The analysis of associations between the number of customers and other elements such as the meteorological elements is left for future work.

References

[1] J. Karhunen, E. Oja, L. Wang, R. Vigario and J. Joutsensalo, A Class of Neural Networks for Independent Component Analysis, *IEEE Trans. on Neural Networks* **8** (1997) 486-504.

[2] J. Karhunen and S. Malaroiu, Locally Linear Independent Component Analysis, *Proc. IJCNN'99*, 1999.

[3] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, NewYork, 1981.

[4] K. Honda, H. Ichihashi, M. Ohue and K. Kitaguchi, Extraction of Local Independent Components Using Fuzzy Clustering, *Proc. IIZUKA2000*, 2000, 837-842.

[5] A. Hyvärinen and E. Oja, A Fast Fixed-point Algorithm for Independent Component Analysis, *Advances in Neural Information Processing Systems* **9** (1997) 1483-1492.

[6] S. Miyamoto and M. Mukaidono, Fuzzy c-Means as a Regularization and Maximum Entropy Approach, *Proc. IFSA'97* **2**, 1997, 86-92.