Working Paper Series

Using Independent Component Analysis for Feature Extraction and Multivariate Data Projection

Andreas Weingessel
Martin Natter
Kurt Hornik

Working Paper No. 16
August 1998

Working Paper Series

August 1998

SFB
'Adaptive Information Systems and Modelling in Economics and Management Science'

Vienna University of Economics
and Business Administration
Augasse 2–6, 1090 Wien, Austria

in cooperation with
University of Vienna
Vienna University of Technology

http://www.wu-wien.ac.at/am

This piece of research was supported by the Austrian Science Foundation (FWF) under grant SFB#010 ('Adaptive Information Systems and Modelling in Economics and Management Science').

Using Independent Component Analysis for Feature Extraction and Multivariate Data Projection

Andreas Weingessel* Martin Natter† Kurt Hornik*

Abstract

Deriving low-dimensional perceptual spaces from data consisting of many variables is of crucial interest in strategic market planning. A frequently used method in this context is Principal Components Analysis, which finds uncorrelated directions in the data. This methodology which supports the identification of competitive structures can gainfully be utilized for product (re)positioning or optimal product (re)design. In our paper, we investigate the usefulness of a novel technique, Independent Component Analysis, to discover market structures. Independent Component Analysis is an extension of Principal Components Analysis in the sense that it looks for directions in the data that are not only uncorrelated but also independent. Comparing the two approaches on the basis of an empirical data set, we find that Independent Component Analysis leads to clearer and sharper structures than Principal Components Analysis. Furthermore, the results of Independent Component Analysis have a reasonable marketing interpretation.

1 Introduction

Marketing data often contain a large number of variables. In the first phase of analyzing multivariate data the marketing analyst typically applies methods for exploratory data analysis (Jain & Dubes, 1988), such as (stochastic) multidimensional scaling (DeSarbo & Rao, 1986; Jedidi & DeSarbo, 1991), factor analysis (Muthèn, 1978; Bartholomew, 1980; Hagerty, 1985), or correspondence analysis (Greenacre, 1984). Principal Component Analysis (PCA) is a well-known method for feature extraction, data compression, and multivariate data projection. Feature extraction can avoid the "curse of dimensionality" or improve the generalization ability of classifiers. Data projection methods enable us to visualize high dimensional data to better understand the underlying market structure, explore the intrinsic dimensionality, and analyze clustering tendency of multivariate data (Mao & Jain, 1995).

*Institut für Statistik und Wahrscheinlichkeitstheorie, Technische Universität Wien, Wiedner Hauptstraße 8-10/1071, A-1040 Wien, Austria; email:*firstname.lastname*@ci.tuwien.ac.at

†Abteilung für Industrielle Informationsverarbeitung, Wirtschaftsuniversität Wien, Pappenheimgasse 35/3/5, A-1200 Wien, Austria; email: Martin.Natter@wu-wien.ac.at

Independent Component Analysis (ICA) is a novel technique which identifies the directions in the input vector space where the signal components are independent random variables or at least as independent as possible, see for example Comon (1994); Karhunen (1996); Karhunen et al. (1997). Until now, ICA has mostly been used for the task of Blind Source Separation (BSS) which separates a mixture of signals into its independent sources, cf. for example Oja & Hyvärinen (1996); Vigário et al. (1996).

Alternatively, ICA could also be used for the task of projection pursuit as mentioned in Karhunen et al. (1997), but this has not been widely investigated thus far.

We use Independent Component Analysis for the identification of market structures. The data set analyzed consists of interviews about the usage of household cleaners.

In the next section we give details on the ICA algorithm, Section 3 describes a pilot application and Section 4 summarizes and interprets our findings.

2 Independent Component Analysis

2.1 Comparison between PCA and ICA

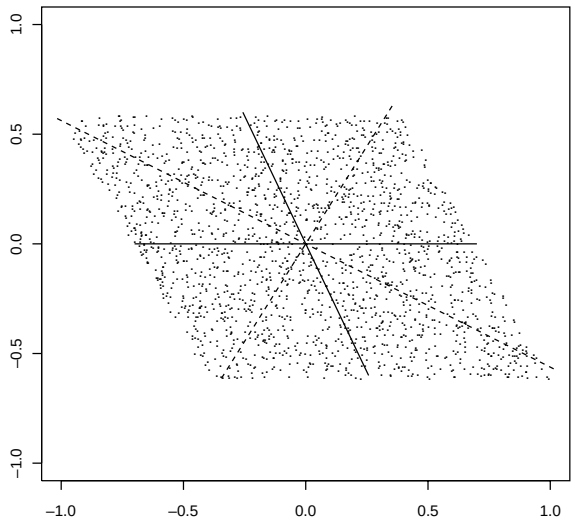

Figure 1: Comparison between PCA and ICA

Principal Component Analysis is a well-known technique to project high-dimensional data to a lower dimensional subspace by finding the most "important" directions, i.e., the directions where the variance is maximal. The

directions found by PCA are uncorrelated, that is, if X and Y are two random variables describing the position of a data point along two axes found by PCA, their covariance is 0. However, the covariance measures only the linear dependency between 2 random variables. Whereas for a multivariate normal distribution uncorrelatedness and independence are equal, because the normal distribution is completely defined by its mean and covariance structure, this is not true for other distributions. Especially, in marketing, typical data are often binary or categorical and the assumption of normality does not hold. That means that generally the uncorrelatedness of 2 random variables does not imply that they are independent. An example for the difference between uncorrelatedness and independence can be seen in Figure 1 (Karhunen et al., 1997) which shows points which are uniformly distributed inside a parallelogram. The directions found by PCA are depicted by dashed lines. One can see that these two lines are orthogonal, the positions (X, Y) of the points along the two lines are uncorrelated, that is $\mathrm{cov}(X, Y) = \mathbb{E}((X - \mathbb{E}\,X)(Y - \mathbb{E}\,Y)) = \mathbb{E}(XY) = 0$. However, X and Y are not independent, i.e., the distribution of Y given X is not equal the distribution of Y, if X is unknown. If, for example, X is known to be near the center, the range of possible values for Y is much larger than if X is at the end of the line.

Looking at the two solid lines which are parallel to the edges of the parallelogram, we see that the positions of the points along the two solid lines are independent. That is, the information about the position of a point along one axis yields no information about the probability distribution of the position along the other axis. The two directions given by the solid lines span the parallelogram. Therefore they are more suitable to "describe" this figure than the dashed lines found by PCA.

The goal of ICA is exactly to find such independent directions in the input space. This task is far more difficult than PCA, because the variance of the data along a direction (which is maximized by PCA) can be directly computed, whereas there is no simple measure for independence of directions in the input space. Therefore, suitable contrast functions have to be defined, see for example Comon (1994). These contrast functions are not easy to compute and therefore generally not applicable for a practical computation. Recently, neural networks algorithm for extracting independent components have been developed (Karhunen et al., 1997) which will be described in Section 2.3.

2.2 Neural Network Algorithms for PCA

Before we will describe the ICA algorithm it is necessary to review neural network algorithms for PCA. Within the last years a lot of such neural network algorithms for PCA have been developed. Whereas some of them are based on a linear autoassociative bottleneck neural network (see for example Baldi & Hornik, 1989), others are based on the so-called Oja rule (Oja, 1982).

The goal of PCA networks is to find a $p \times d$-matrix A which projects input vectors x which are samples of a d-dimensional distribution onto that p-dimensional subspace which is spanned by the first p principal components.

To achieve this the network tries to maximize the correlation between x and

its projection $v = Ax$. This is done by updating the weight matrix A by a Hebbian rule (see for example Haykin, 1994) with an additional decay term to prevent the output from growing out of bounds (Equation 1).

$$\Delta A \; \alpha \; vx' - vv'A \tag{1}$$

The $'$ operator denotes the transpose of a vector or a matrix. Several extensions to this basic rule have been proposed, the interested reader is referred to Diamantaras & Kung (1996). One of these extensions leads to neural networks which can be used to extract ICA as it will be described in the next section.

2.3 Neural Network Algorithms for ICA

Neural network algorithms for extracting the ICA are nonlinear extensions of Oja's subspace rule for PCA. The computation of the ICA consists of two parts as depicted in Figure 2. First the data x is prewhitened by computing the ordinary PCA and the first p principal components are extracted yielding the uncorrelated data v. From this data, the independent components y are extracted.

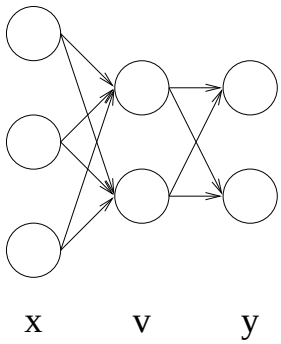

Figure 2: A Network to Compute ICA

Let $y = Wv$. Then, the weight matrix W which performs ICA is updated via the following algorithm.

$$\Delta W \; \alpha \; f(y) * (v' - f(y)'W) \tag{2}$$

where $f(y)$ denotes a suitable nonlinear function. By comparing the update equations for PCA (1) and ICA (2) we see that the only difference between these two algorithm is the application of the function f.

In order to achieve independent component analysis a suitable function f has to be chosen. ICA assumes that the observed values $v = (v_1, v_2, \ldots)'$ are a linear mixture of independent sources $y = (y_1, y_2, \ldots)'$ and tries to invert this mixture by finding a proper separating matrix W. In order to find a proper function f one has to know whether these sources are sub-Gaussian or super-Gaussian. Sub-Gaussian means that the distribution has a negative kurtosis, i.e., it has a flatter tail than the normal distribution. Super-Gaussian distributions have a

positive kurtosis, i.e., the distribution has heavy tails. Gaussian sources (i.e., sources from a normal distribution) pose a problem to ICA, because any linear combination of Gaussian distributed variables is again Gaussian distributed, so it is not possible to separate a mixture of Gaussian variables. It can be shown (Karhunen et al., 1997) that the maximization of

$$\sum_i |\operatorname{cum}(y_i^4)| := \sum_i |\mathbb{E}\, y^4 - 3(\mathbb{E}(y_i^2))^2|$$

yields ICA if the sign of the kurtosis is the same for all sources. For prewhitened output this is equivalent to minimize (for negative kurtosis) or maximize (for positive kurtosis) the sum of the fourth moments.

3 Pilot Application

3.1 Data Used

Our data set consists of usages of household cleaners in different usage situations. There are 7 different brands (*A–F*) and 5 different usage situations (*1–5*) giving a total of 35 combinations. Combinations with a small frequency have been removed, leaving 20 combinations, namely *A1, A2, A3, A5, B1, B2, B3, B5, C1, C3, C5, D1, D2, E3, E4, F4, G1, G2, G3, G4*. These 20 combinations correspond to 20 binary (0-1) variables. If *A1* equals 1, then brand *A* is used in situation *1*, if *A1* equals 0, this brand is not used in this situation. The respondents constitute a representative random sample of 831 housewives.

3.2 Experimental Results

Our experiments were performed as follows. First principal component analysis was performed. Then the ICA was extracted from the values of the principal components. The nonlinearity $f(y) = y - \tanh(y)$, see Karhunen et al. (1997), proved to yield suitable results for our application.

The main result obtained is the ability of the ICA algorithm to find variables belonging together and to cluster them accordingly. That is, the ICA algorithm projects all these variables onto one dimension and the values in this dimension form clusters according to the different combinations of these variables.

For example, ICA finds that the variables corresponding to one brand, as for example *A1, A2, A3*, and *A5*, belong together. That is, these 4 variables are projected onto one dimension. Small (large) values in this dimension indicate that the corresponding person never uses brand *A* (uses brand *A* in all situations) or vice versa. Values in between indicate that *A* is used only in 1, 2, or 3 of the possible 4 situations. Mostly, there are sharp boundaries between the number of situations a brand is used. The most important brands according to ICA are *G* and *B*, these are also the ones most frequently used. But if the ICA is computed with enough dimensions all brands except *F* are projected to their individual dimension.

The results of ICA are depicted in Figure 4 and compared with PCA in Figure 3. We see that 4 of the 6 dimensions of PCA can be interpreted as

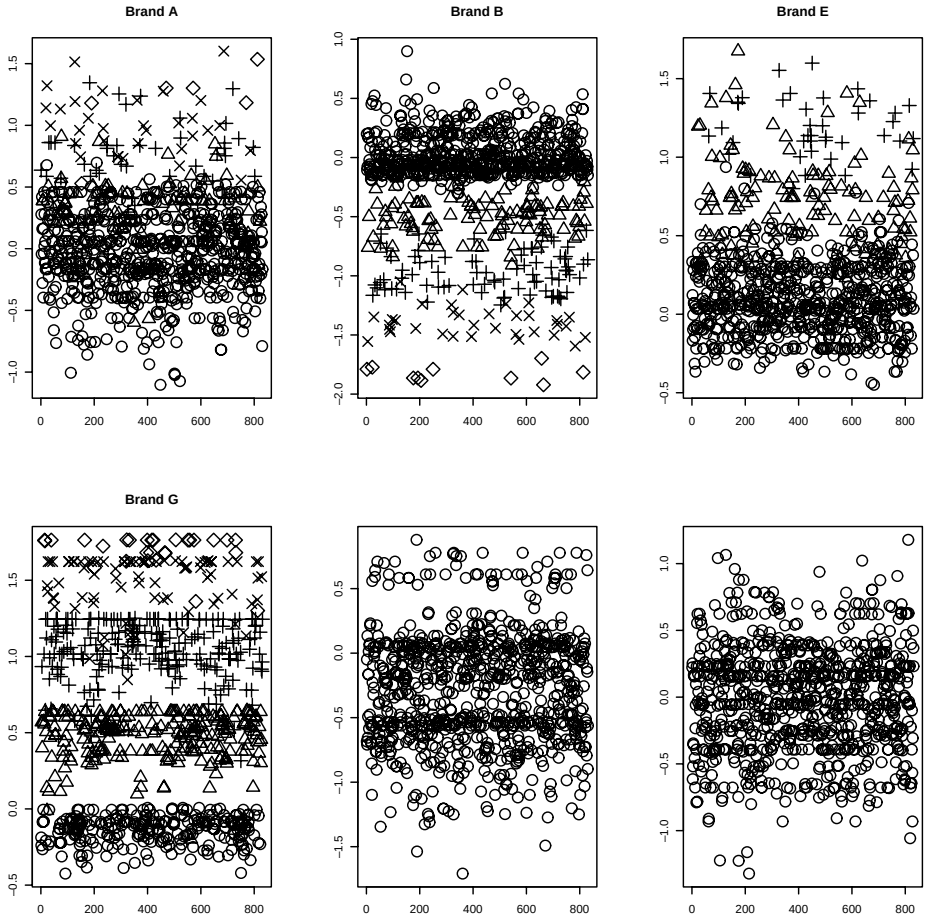

Figure 3: Results of PCA

belonging to one brand. Only two of them (brands B and G) show a strict separation, two others (A and E) show only a slight separation. 2 dimensions can not be interpreted at all. ICA, however, is able to assign every dimension to one brand, yielding a much sharper separation than PCA.

Although there are not any restrictions about the usage of the brands (all combinations of zeros and ones are allowed in the variables) ICA finds that the brands are independent factors on the usage of household cleaners. That is, ICA can be used to separate the variables describing the particular brands.

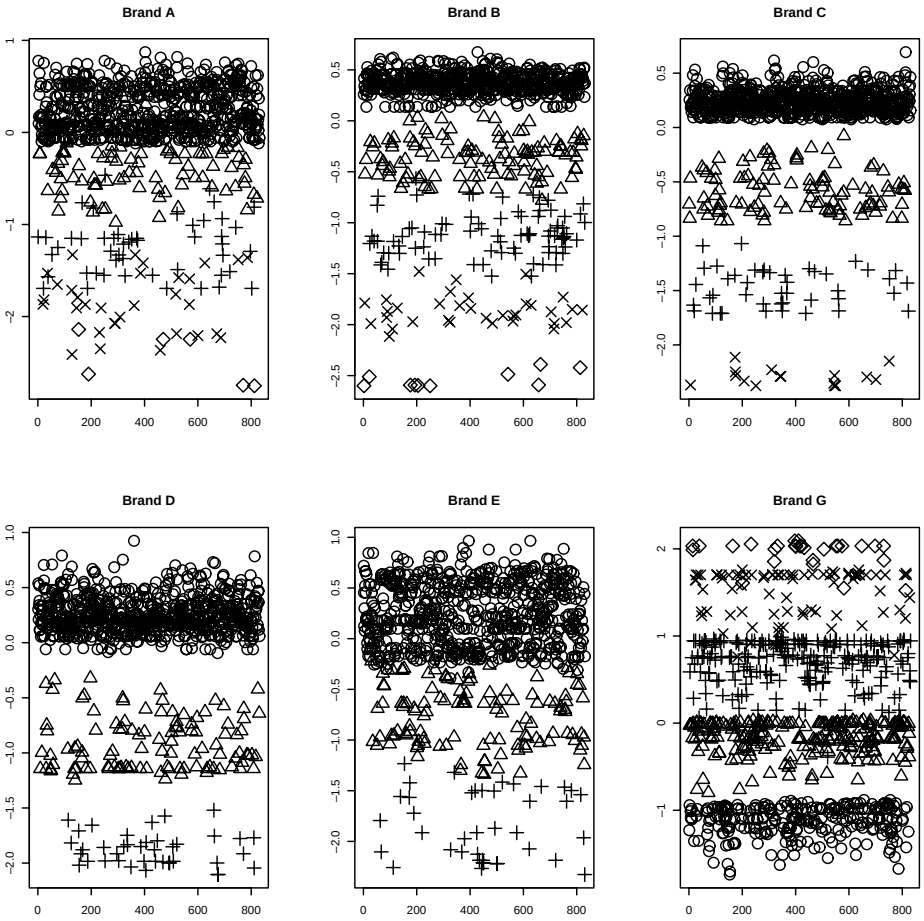

Figure 4: Results of ICA

4 Conclusion

We propose to use ICA as an alternative tool for the identification of market structures. A pilot application with binary household-cleaner usage data indicates that ICA can be a valuable tool for extracting a lower dimensional representation which offers new insight into market structure where PCA failed in

discovering interesting data views. This advantage over PCA may be explained by the fact that ICA looks for independent dimensions taking non-linearities into account.

One of the problems of the ICA approach is that an explicit linear mixture model $x = My$ is assumed, where only x is observed and the independent sources y are unknown. ICA tries to separate this mixture model again. However, one needs to make some assumptions on the sources y, especially the sign of the kurtosis should be equal for all sources and be known.

As in real data the true dimensionality and structure is unknown, it would be interesting to see future work that investigates this novel methodology in a Monte Carlo setting with different numbers of respondents, brands and dimensions.

References

Baldi, P. & Hornik, K. (1989). Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, **2**, 53–58.

Bartholomew, D. J. (1980). Factor analysis of categorial data. *Journal of the Royal Statistical Society, Series B*, **42**, 293–321.

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing*, **36**, 287–314.

DeSarbo, W. & Rao, V. R. (1986). A constrained unfolding methodology for product positioning. *Marketing Science*, **5**(1), 1–19.

Diamantaras, K. I. & Kung, S. Y. (1996). *Principal Component Neural Networks: Theory and Applications*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons, Inc.

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. Academic Press, London.

Hagerty, M. R. (1985). Improving the predictive power of conjoint analysis: The use of factor analysis and cluster analysis. *Journal of Marketing Research*, **22**, 168–184.

Haykin, S. (1994). *Neural Networks. A Comprehensive Foundation*. New York: Macmillan College Publishing.

Jain, A. K. & Dubes, R. C. (1988). *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice Hall.

Jedidi, K. & DeSarbo, W. (1991). A stochastic multidimensional scaling procedure for the spacial representation of three-mode, three-way pick any/j data. *Psychometrika*, **56**(3), 471–494.

Karhunen, J. (1996). Neural approaches to independent component analysis and source separation. In *4th European Symposium on Artificial Neural Networks*, pp. 249–266, Bruges, Belgium.

Karhunen, J., Oja, E., Wang, L., Vigário, R., & Joutsensalo, J. (1997). A class of neural networks for independent component analysis. *IEEE Transactions on Neural Networks*, **8**(3), 486–504.

Mao, J. & Jain, A. K. (1995). Artificial neural networks for feature extraction and multivariate data projection. *IEEE Transactions on Neural Networks*, **6**(2), 296–317.

Muthèn, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, **43**, 551–560.

Oja, E. (1982). A simplified neuron model as principal component analyzer. *Journal of Mathematical Biology*, **15**, 267–273.

Oja, E. & Hyvärinen, A. (1996). Blind signal separation by neural networks. In Amari, S.-i., Xu, L., Chan, L.-W., King, I., & Leung, K.-S. (eds.), *Progress in Neural Information Processing: Proceedings of ICONIP'96*, vol. 1, pp. 7–14, Hong Kong.

Vigário, R., Hyvärinen, A., & Oja, E. (1996). ICA fixed-point algorithm in extraction of artifacts from EEG. In *NORSIG 96*, pp. 383–386, Espoo, Finland. 1996 IEEE Nordic Signal Processing Symposium.