

# Independent components in telecommunication call-detail records

P. Lakatos, B. Egri & Z. Cseh  
*DSS Consulting Ltd., Hungary*

## Abstract

Although Independent Component Analysis (ICA) has gained a recognized place among neural network models [1], data mining applications have been limited, especially for marketing purposes. In our paper we will present a unique ICA application, in which ICA outperforms other data mining and statistical methods.

We've applied ICA to the call-detail records of a fixed-line telecommunication company's customers. Telecommunication companies can observe the duration, time and tariff zone of the incoming/outgoing calls, which, in this context, represent the mixed input signals. These records are generated as a result of a (supposedly linear) mixture of independent signals representing the telecommunication needs of a family, like chatting with friends, business calls or Internet usage. The aim was to disclose these underlying independent components. These components can be interpreted according to the subscribers' independent telecommunications needs, potentially the most useful information for decision makers as it accurately describes overall customer behavior. ICA can show which components are most important for each customer.

The insights gained through these methods create new opportunities for identifying important market segments and facilitating customer demand-based product development and is an effective way of dealing with churn in the telecommunication realm.

In this paper we give a short introduction to the main features of the ICA method, focusing on its application in the telecommunications marketing area and briefly summarizing the results of our work on real data.

## 1 Introduction

Telecommunications companies strive to develop products that best suit customers' telecommunications needs. With ICA, these telecommunications needs can be determined solely from call records.

Customers' telecommunications behavior is influenced by many factors, some of which cannot be measured or are difficult to formalize, such as the customer's financial background or the weather. Nevertheless, the phone call data can be modeled as a linear combination of independent factors representing customers' reasons for placing calls. These factors, which we call needs, include business calls, keeping in touch with relatives, Internet usage, making plans for Saturday night or being available anytime. These needs may have different weights for each customer. Our task is to determine the independent factors that best describe the customer's telecommunications needs based on call records, which are noisy and contain only partial information about the customers.

Giving a thorough picture of the ICA algorithm wasn't our objective; readers can find very good and practical guidelines in [2]. Here we give a brief overview of the ICA method, including the basic techniques for obtaining the independent components and we give a short account of the particular method used. Finally, we present our results and some suggestions for further investigation and application.

## 2 Theoretical background

Feature extraction, the identification of low-dimensional sub-spaces containing most of the information in high dimensional data, is essential in data interpretation. The components obtained with the widely-used Principal Component Analysis method are uncorrelated (the second order statistics are zero). Principal components capture most of the variance in the data, so if, in addition, the resulting components are interpretable as well, the high-dimensional space can be successfully reduced to a lower one.

Independent Component Analysis finds statistically independent directions in the data (the higher order statistics are zero as well). This difference between the two methods also leads us to a novel interpretation of the ICA's output vectors, which in this case are not just high-variance extracted features. In our paper, we rely on the fact that the context-independence of the subscribers' call records can be interpreted as a telecommunication need or demand. In short, independent components themselves bear meaningful information- they are more than just a good projection of the original data.

Let us suppose now that the observable signals,  $\mathbf{c}$ , are a linear combination of statistically independent input vectors,  $\mathbf{s}$ , where  $\mathbf{M}$  stands for the mixing matrix. With these notations we have:

$$\mathbf{c} = \mathbf{M}\mathbf{s}$$

All we can observe are the  $\mathbf{c}$ -measurable input call records. We have no direct information about the mixing matrix, the original independent components, or the number of components. Given certain assumptions, the equation above could be transformed into

$$\mathbf{s} = \mathbf{M}^{-1}\mathbf{c}$$

The independent component analysis consists of the estimation of the  $\mathbf{M}^{-1}$  inverse matrix such that the  $\mathbf{s}$  signals should be as independent as they can be. In Section 2.3 we briefly summarize a few feasible methods for estimating the independent components.

## 2.1 Assumptions

We should outline now the assumptions we have to make. Statistically, independence means that no input vector provides probabilistic information about the value of any other. Although it isn't necessary to assume linear mixing, it's wholly reasonable in the context of telecommunications. No more than one independent component should have a gaussian distribution. This is the case here, as most real-life signals are non-gaussian. We can deal only with mixing matrices where the number of independent components doesn't exceed the number of measured input signals. We've applied a square mixing matrix here. Gradually extending the number of independent components introduces new components without altering the ones we've already obtained; therefore, we've put no restrictions on the number of independent components, except that it should not exceed the number of input signals. The interpretability of the independent components is the final measure of the model's utility.

## 2.2 Preprocessing

ICA requires some preprocessing. Centering and whitening are strongly suggested, but optional. We cropped the outliers, i.e. those customers who generate exceptionally large traffic in a specific time-tariff code section.

Due to centering, signals are transformed into zero-mean vectors. This is done only for our convenience and means are then factored back into to the independent components.

Whitening means that we transform the input vectors in order to obtain uncorrelated and unit-variance input signals. The resulting mixing matrix is orthogonal, so the number of parameters to be estimated is reduced by about half. A widely used method for whitening is the eigenvalue decomposition of the covariance matrix, which, in addition, allows us to reduce the dimensions and noise by neglecting those eigenvalues that are close to zero.

## 2.3 Methods

There are different methods [3] for estimating the independent components; they all have, however, an underlying common principle: nongaussianity. According to the Central Limit Theorem, under certain conditions the distribution of a sum

of independent random variables in the limit has a gaussian distribution. Therefore, our task is to find the de-mixing matrix that results in the  $\mathbf{s}$  output vector where the sum of the signals is the least gaussian. Kurtosis is an expedient measure of gaussianity, as it is zero for a gaussian variable and differs from zero for most of the nongaussian variables. We've used kurtosis-based independent component estimation as well; details can be found in the next section.

Entropy, or rather negentropy can also be used for measuring gaussianity, as gaussian variables have the largest entropy (i.e. are the least structured) of all continuous random variables with the same variance. Negentropy is simply the difference between the entropy of a gaussian variable with the same covariance matrix and the original one. Hence, with  $H$  as entropy, we have:

$$J(\mathbf{s}) = H(\mathbf{s}_{gauss}) - H(\mathbf{s})$$

In practice it is difficult to calculate with negentropy, so an approximation of it is used.

The mutual information-based approach from information theory is another method available for computing independent components, as it measures how much information one variable carries about another. Hence, the mutual information of independent variables equals zero. The  $I$  mutual information between  $s_i(n), i = 1..m$  ( $n$  stands for the customer ID) scalar variables, here representing the input signals, is

$$I(s_1, s_2, \dots, s_m) = \sum_{i=1}^m H(s_i) - H(\mathbf{s})$$

In our case,  $\mathbf{c} = \mathbf{M}\mathbf{s}$ , where the  $\mathbf{M}$  mixing matrix is assumed to be invertible, the mutual information could be rewritten as:

$$I(s_1, s_2, \dots, s_m) = \sum_{i=1}^m H(s_i) - H(\mathbf{s}) - \log|\det \mathbf{M}^{-1}|$$

And from this, taking into account the whitening preprocessing phase, derives the link between the mutual information and negentropy:

$$I(s_1, s_2, \dots, s_m) = const - \sum_{i=1}^m J(s_i)$$

Maximum likelihood estimation-based methods are also often used in independent component analysis. With our notation:

$$L = \sum_{n=1}^N \sum_{i=1}^m \log f_i((\mathbf{M}^{-1})_i^T \mathbf{c}(n)) + N \log|\det(\mathbf{M}^{-1})|$$

where  $f_i$  is the estimated density function of  $s_i(n)$ . It could also be shown that, with certain restrictions, the likelihood and mutual information are equivalent.

We've actually used a nonlinear PCA (gradient-based optimization) method. The optimization criterion is the approximation of the mutual information, which we try to minimize with gradients calculated for each element in the input signals.

The kurtosis or the fourth order cumulant of  $s$  :

$$K_4 = E(s^4) - 3 \cdot [E(s^2)]^2$$

We can sort the signals into three classes with respect to their kurtosis [4]:

$K_4 = 0$  : gaussian random variables

$K_4 > 0$  : supergaussian random variables (these have spiky probability distribution functions (pdf) with heavy tails, e.g. the Laplace distribution)

$K_4 < 0$  : subgaussian random variables (these have a flat pdf, e.g. uniform distribution).

The objective function of the gradient-based optimization algorithm is the approximation of the mutual information, the sum of the kurtosis of the mixed signals.

$$cum(s) = \sum_i K_4 = \sum_i \{E(s_i^4) - 3 \cdot [E(s_i^2)]^2\}$$

### 3 ICA in call records

#### 3.1 The cocktail party situation

If you go to a cocktail-party, you will hear a lot of sounds at the same time: People are talking, glasses are clinking, the music is playing and all this reaches your ears together, mixed up. Nevertheless, you are able to focus on any of these sources, i.e. to separate them. That is due to the human brain, which can separate the mixture of sounds into its original, independent sources. ICA, developed recently by researchers in artificial intelligence, also has this capability.

#### 3.2 Linking ICA and telecommunication needs

In telecommunications terms, the mixed sounds are the call-detailed records (CDR). For each customer we can measure how much they have communicated, broken down by call direction or network destination (i.e. location or modality of destination), peak/off-peak calling periods and day type. But CDRs are generated by the original, independent customer needs, so we are looking for these factors. These independent needs are mixed up, with different weightings for each customer and are hidden within the CDRs. The CDRs are the input to the algorithm, the goal of which is to isolate the original, independent customer needs which, in this context, are called independent components.

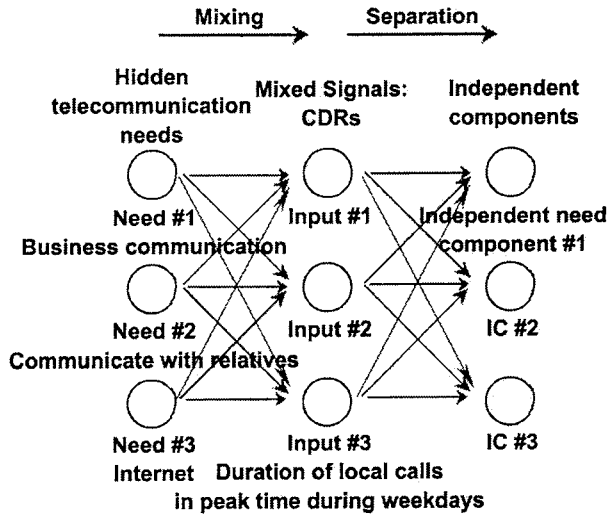


Figure 1: Independent components and call records.

All we can observe are the  $c_{i(n)}$  measurable input call records, where  $n = 1..N$  represents the subscribers; therefore our series of observations consists of the successive records of the subscribers rather than time series. The index  $i = 1..I$  represents the actual combination of time, day type and tariff zone. In our case, we had a sample of  $N = 50,000$  subscribers out of a total of 2.5 million. The combination of call record attributes consisted of  $I = 27$ , which means that we measured subscribers' calling behavior with 27 different microphones, where each microphone recorded the total sum of the duration of a specific type of call, like the sum of the calls of the  $n^{\text{th}}$  subscriber on weekdays in working hours to a mobile network destination. In addition, we've had to make another assumption: we've supposed that the same independent (need) components are valid for the whole subscriber base, while the intensity of the components might vary from subscriber to subscriber. This extra assumption allowed us to represent the need-mixture with the same matrix, and the method was invariant with regard to how we enumerated the subscribers in the  $c_{i(n)}$  input signal. It should be mentioned that no assumption about independent components is made.

## 4 Experimental evaluation

Independent Component Analysis was applied to the call-detail records of a fixed-line telecommunication company's customers. In what follows, we summarize the main results of this analysis, focusing on the kind of results that can be obtained using ICA. It is necessary to begin, however, by briefly introducing a few of the peculiarities of the Hungarian fixed-line telecommunications market.

Regardless of the payment package, a customer has to pay for each type of call according to individual call duration and number of calls. Packages where unlimited local calling is provided for a fixed monthly price currently do not exist in Hungary. Different types of calls have different prices, and the call type always depends upon the call's destination, and more specifically, on the network destination code. The more expensive types are wireless (mobile), long-distance and international calls. In addition, mobile service providers have a separate network destination code, so fixed-to-mobile calls are always considered mobile calls and are never considered local. The Internet is also considered a specific call destination.

The duration, cost and total number of calls of each call type made by about 50 thousand private customers were aggregated for three months using the following aggregation keys:

- day types: working days (wd), holidays (n.wd)
- time intervals: peak (9am-3pm), morning/evening (n.w.time: 7am-9am and 3pm-6pm), reduced (5am-7am and 6pm-10pm), night: (10pm-5am)
- network destination: local, long-distance (L-dist), mobile (M), international, internet, data traffic, voice mail ...)

By combining the different values of the above aggregation keys, we distinguished 27 input variables associated with each customer. Data traffic, voice mail and certain types of call were not split by the two other aggregation keys due to the low numbers or short duration of these calls; in addition, holidays were divided into only two time intervals: reduced and night. These parameters represented the mixed input signals of our analysis. Data preparation and aggregation were merely heuristic, but after testing the different aggregations we concluded that the use of call-duration data provides the most reliable and interpretable results.

The maximum number of independent components that can be generated by ICA equals the number of input mixed signals. A sample of the independent components obtained is shown in the following table. We didn't show those  $c_i$  input vectors where the mixing matrix is zero for all of the outlined independent components. As the model is invariant under scalar multiplication, further analysis was needed to decide the right sign of the columns.

A component in this context is a calling behavior, built up as a linear combination of basic factors (the different call types, as input variables) describing the various aspects of telecommunication. Although only a theoretical pattern of telecom usage, each (statistically) independent component potentially represents the calling habits of real subscribers, and, as such, a *telecommunications need*.

High positive coefficient values (in italics) mean that in that component the specific call type is decisive, while high negative coefficient values (in underlined italics) suggest that it should *not* occur if it is a significant telecommunications need in the case of a given subscriber.

Table 1: A sample of the independent components obtained.

| Call types             | Component ID |            |            |            |            |
|------------------------|--------------|------------|------------|------------|------------|
|                        | 6            | 7          | 11         | 16         | 19         |
| L-dist., reduced, wd   | -3           | 1          | 0          | -16        | <u>-48</u> |
| L-dist., n.w. time, wd | -4           | -3         | 20         | 11         | <u>-25</u> |
| L-dist., nwd           | 1            | 2          | -5         | 3          | 100        |
| Local, peak, wd        | -2           | 100        | 0          | -2         | 1          |
| Local, reduced, wd     | -2           | 2          | -2         | 100        | 1          |
| Local, n.w. time, wd   | 1            | <u>-69</u> | 5          | <u>-21</u> | 1          |
| Local, nwd             | 0            | <u>-22</u> | -7         | <u>-70</u> | -4         |
| Internet, night        | 0            | 0          | 100        | 1          | 0          |
| M, peak, wd            | 9            | 72         | 7          | 31         | -1         |
| M, night, wd           | 80           | 3          | <u>-63</u> | -3         | 6          |
| M, reduced, wd         | 100          | -3         | 17         | -10        | 4          |

The component is not just a category into which some actual calling patterns fit and some do not – some fit (i.e. high positive values) and some fit negatively (i.e. high negative values), whereas coefficients around zero mean that the specific call type is irrelevant for that component.

Interpreting the independent components is not difficult if we can imagine what kind of call behavior (and what kind of caller) could lead to the high positive and negative values of the components. The components can be characterized in common-sense terms. For example, in the case of component ID 7, we get high positive values, i.e. strong presence of this telecommunications need for subscribers who communicate a lot mainly during peak hours of working days in local or mobile directions, but *do not* communicate in local directions in non-working time. Hence, it is most likely that the right interpretation of this component is *business communication*, even though these were private (non-business) customers. Previously, this behavior of a customer group was only presumed; applying ICA demonstrated that this telecommunication need is corroborated by the data.

Component ID 6 corresponds to the following telecommunications need: cost-sensitive communication in the mobile direction on weekday mornings and evenings. In this case, the contrast in the values of the last two parameters and the others is very sharp. This component looks to be a well-defined customer need.

Data traffic, voice mail, and a couple of other specific types of call were zero with more than 95% of the customers. Appropriately, the ICA algorithm discovered these extremities and treated them as independent components (not in the table).

The third component in the table, component ID 11, describes the “midnight internet surfers” need. It is interesting to note that wireless calls at night have a very low weighting in this customer need.

The fourth column, component ID 16, looks very strange at first glance. The two parameters that occur in this customer need are very dissimilar, and it is difficult to explain what kind of need it describes. Finding the correct explanation will require greater knowledge of customer demographics where this telecommunications need arises.

The need corresponding to the last component, component ID 19, probably describes long-distance weekend communication with friends and relatives. In this component the need for long-distance calls in any other day or time interval is very low. As can be seen above, ICA illuminates many different kinds of customer needs.

In general, we have distinguished four different groups of independent components:

The very rarely used call types (like data traffic), which are themselves independent components, fall into the first group.

The second group contains those “single rule” components that can be linked to only one call type, i.e. the column vector has only one element that differs greatly from zero. In these cases only one parameter is enough to determine an independent customer need. Surfing the Internet by night (Component ID11) is a good example of this category.

The third group is the so-called singularly composed components’ group. This means that the components belonging to this group are composed of more diverse call types, and none of the other components have these types of calls with anywhere near the same weightings (Component ID19). Thus, these components cannot be merged with any of the others. For example, our analysis revealed the component in which only long distance calls had great importance. None of the other components contained these parameters with roughly the same scalars.

## 5 Concluding remarks

Although research in Independent Component Analysis is intensive, data mining applications on large databases and with the aim of knowledge discovery are limited. ICA could bring successful results hardly replicable with other data mining methods. Applying ICA to telecommunications call records for marketing purposes results in components that are interpretable as independent telecommunications *needs*. Discovering the hidden telecommunication demand patterns in the case of a fixed line company is more challenging, because here the customers are primarily families rather than individuals, and the observable call records are already a mixture of the family members’ different needs. ICA, as projection pursuit and preprocessing for segmentation based on call-behavior data, proved to be very useful, leading to lower-dimensional and “more interesting” spaces and ultimately to more interpretable, meaningful segments.

Telecommunications needs revealed by ICA and additional information about the subscribers can, taken together, lead to a detailed understanding of customer demand.

## Acknowledgments

We are very grateful to Zsolt Györi at MATÁV Inc. for providing us with call records, and for comments on interpreting the independent components. We would like to thank our colleagues at DSS Consulting for their support and advice, especially Rita Ludvig and László Mocsár. We are also very grateful to Miklós Pintér at the Budapest University of Economic Sciences for bringing our attention to the capabilities of ICA and to Péter Aszalós at TextLab Ltd. for his technical support.

## References

- [1] Haykin, S., *Neural Networks, A Comprehensive Foundation*, Prentice Hall: New Jersey, 1999.
- [2] Lee, T., Girolami, M., Bell, A.J. & Sejnowski, T.J., A Unifying Information-theoretic Framework for Independent Component Analysis, *Computers & Mathematics with Applications*, **31**, pp. 1-21, 2000.
- [3] Hyvärinen, A. & Oja, E., Independent Component Analysis: Algorithms and Applications. *Neural Networks*, **13**, pp. 411-430, 2000.
- [4] Karhunen, J., Oja, E., Wang, L., Vigário, R. & Joutsensalo, J., A Class of Neural Networks for Independent Component Analysis, *Proc. of the 8<sup>th</sup> IEEE Transactions on Neural Networks*, pp. 487-504, 1997.
- [5] Pajunen, P. & Karhunen, J., Least-Squares Methods for Blind Source Separation Based on Nonlinear PCA, *International Journal of Neural Systems*, **8**, pp. 601-612, 1997.