INDENDENT COMPONENT ANALYSIS (ICA)

M.Sc. Tetiana Zinchenko, Supervisor: Dr. P. Miettinen
Saarland University (Germany)

What is ICA?

Before the actual explanation of all the aspects and algorithms of Indendent Component Analysis (ICA), it would be nice to outline what is all about.

Thus, what ICA actually is? The first intuition which comes to mind is that it is somehow connected with wide-used linear transformation method Principle Component Analysis. Of course, the analysis of components makes these two methods be ralated to the same sphere. However, ICA is comparatively recently developed technique. The main purpose of it is finding of linear representation of non-Gaussian data and in such a way which makes component statistically independent from each other. Or, at least, the independence of these components is maximal. In other words, ICA helps extracting latent factors which lie in a datasets of random variables, measurements or signals. Often these datasets are called **mixtures**.

Formally ICA of given matrix A of a size $n \times m$ can be defined as follows:

$$A = C\,F \qquad (1)$$

With:

C - $n \times m$ *and* F - $m \times m$

The rows of matrix F characterize m independent components of original matrix A; and matrix C with observed attributes of dataset [1].

ICA is very sound technique which is useful in separating distinct sources of several linearly mixed signals. On Fig.1 cartoon example depicts the global concept of ICA.

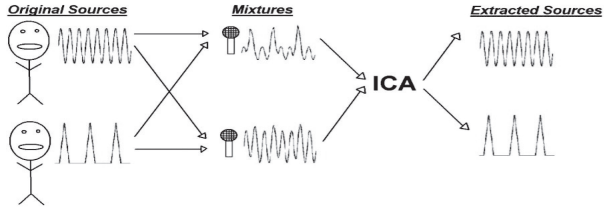

Fig. 1. ICA principle [14]

For example, in a record of electroencephalograms ICA is able to split artifacts which interfere the data, because they are usually independent. Another vivid example which describes usage of ICA is that the independence of the consequent hidden factors is useful, for example, in segregating mixtures of multiple sources. The simplest intuitive understanding of ICA implies that we require n different non-Gaussian sources to get positive outcomes.

Similarities and differences with studied methods

ICA has a lot in common with already discussed method during lectures – SVD. However, there are several meaningful differences between them.

Firstly, the factors which are picked up by SVD are uncorrelated. But, as it was already mentioned, in ICA components are meant to be statistically independent. Secondly, ICA does not provide the reduction of the size of dimensionality for dataset. More to say, if one chooses one scalar and multiply a row of F with it; and after this divide by it a corresponding column of C – decomposition still holds. Thus, the ambiguity can be observed. Finally, the permutation of rows in F matrix will not cause any troubles, if the columns of C are permutated reciprocally [1].

When talking about **assumptions** which are taken as a basis in ICA, firstly, it is worth mentioning the main one – independence of attributes.

Two variables x_1 and x_2 are considered to **be independent** iff their collective probability density function (PDF) is a product of two marginal pdfs [2].

So why do we actually focus on this **assumption**? The answer is quite simple. When one performs decomposition of matrix *A* into two matrices (*C* and *F*) in such a way that their product would give as close result as possible to original A, the assumption of independence is sufficient. In [1] the author provided a convincing proof. This concept delivers some ambiguity. On the one hand, the processes which are considered independent in reality should also perform mathematical independence when studying them. Thus, ICA will definitely be applicable and will perform prolific results.

On the other hand, it is intuitively understood that there some situations which at the first glance do not have any correlation amongst them but when statistically analyzing them, they can show some degree of dependency and connection of the components. That is why, statistical independence requires consideration for each particular data set.

One more essential assumption is **Non-Gaussianity**. This aspect should be underlined as well, since, with Gaussian variables initial independence of components cannot be reached. Thus, there should be at most one Gaussian distribution in independent attributes, because higher-order cumulants are zero for Gaussians and this violates independence principle. And this assumption can be considered as a weakness of ICA. Since, if it does not hold the method will not give meaningful result. More details about **Non-Gaussian** requirments can be looked up in [3] and [4].

Morover, one more assumption will be mentioned here. It is simplified that unknown mixing matrix has equal number of rows and columns and it is invertible (i.e. the number of the independent sources in matrix C is equal to the number of mixtures.). An ICA has also an assumption in it of zero mean data (as well as SVD). Make the observed signals zero mean and decorrelation of them remove the second-order dependencies between components and boosts independency.

Interpretation of the results

Usually, the Gaussian–shaped distributions are very comfortable to work with. Nevertheless, there are numerous domains which do not sustain such a distribution type. Thus, ICA is a useful tool for working with them. The main application sphere of ICA is **signal processing,** where distinction of different sound sources from noise is crucial. More to say, such areas as finances and biomedical sensing can benefit from ICA as well, because these spheres are quite similar. Firstly, it is relatively understood how many components should be in the data (i.e. we know what to expect from result and, thus, can assess the quality of it). Secondly, the noise is clear to be understood.

The **factor interpretation** of ICA results is clear enough. Matrix A contains rows which correspond to *n* signals recorded by each microphone within *m* time intervals. The rows if F matrix are the factors (separate voice lines within time span) mixed be elements of matrix C, which shows that the microphones recorded overlapped signals.

Two resulting matrices *F* and *C* can be interpret geometrically as well. So, the rows of matrix C represent coordinates in geometric space. Depiction of column from this matrix helps to visualize the structure of the data. Nevertheless, since the rows of F matrix are not axes (no ortogonality) metric cannot be applied blindfold in particular geometric space. Thus, two axes which can be erroneously considered as orthogonal in reality can be same way directed. This leads to misinterpreting of data and mistaken outcomes. Namely, if the luster is found, it is expected that it will be plot along an axes and every such a cluster would represent one independent component. Hence, ICA visualization will be looking better than SVD, only because of the fact that rows of F do not stand for axes and not orthogonal. [1]

Despite the fact that component interpretation of the results is not used with ICA, it is possible to determine the contribution of each independent to the initial data. Hence, as it was already mentioned, rows of matrix F represent one of the processes which is difused with original data

To discover the effect of ith process on entries of A matrix, it is sufficient to multiply ith column of matrix C and the corresponding row from matrix *F*.

Algorithm of ICA

Before applying ICA to any dataset it is useful to conduct preprocessing of the data. In general, there two essential steps covered in the literature. They are **centering** (preprocessing is to center A, i.e. subtract its mean vector, to make A a zero-mean variable) and **whitening** of the observed variables (transformation the observed vector *linearly* to get a **white** vector which, i.e. its components are uncorrelated and their variances equal unity). It can be implemented with a help of eigen-value decomposition (EVD). The author in [5] claimes that "…whitening solves half of the problem of ICA. Because whitening is a very simple and standard procedure, much simpler than any ICA algorithms, it is a good idea to reduce the complexity of the problem this way".

Moreover, before implementation of ICA it is very useful to perform reduction of dimensionality of the data. Only a few concealed components in the high-dimensional given data will give the best result. Thus, the data should be compressed before processing. Intact original high-dimensional data set might result in moderate results. This can be as a consequence of noise, which can be presented in several initial dimensions.

The main challenge when conducting dimension reduction is to lower the number of redundant dimentions without flatteneing the data structure, since the data, we are interested in, is usually projected to a lower dimensional space.

All existing **algorithms of ICA** can be separated in 4 groups (depending on techniques which they stand for):

1. maximization of **non-Gaussianity** of the components
2. minimization of mutual information
3. maximum likelihood estimation
4. tensorial methods

Three the most popular **algorithms of ICA** can be mentioned

- FastICA
- JADE (joint approximate diagonalization of eigenmatrices)
- Infomax

Conclusions

Independent component analysis is quite interesting method for discovering latent information from big data sets. It is a nice tool to be used in Data Mining. ICA methods are very successful and the most widely-used ones in blind source separation.

Despite some limitations which it has (can only separate **linearly mixed sources** and disability to work with Gaussian-distributed variable) with brainy use very often it gives impressive results.

References

1. http://www.mpi-inf.mpg.de/departments/d5/teaching/ss13/dmm/papers/skillicorn_ch7.pdf
2. *https://aaltodoc.aalto.fi/bitstream/handle/123456789/4610/isbn9789512298310.pdf?sequence=1*
3. http://cis.legacy.ics.tkk.fi/aapo/papers/IJCNN99_tutorialweb/node9.html]
4. http://www.informatica.si/PDF/35-1/10_Naik%20-%20An%20Overview%20of%20Independent%20Component%20Analysis%20and%20Its%20Applications.pdf
5. http://cis.legacy.ics.tkk.fi/aapo/papers/IJCNN99_tutorialweb/node26.html
6. http://en.wikipedia.org/wiki/Central_limit_theorem
7. http://en.wikipedia.org/wiki/Central_limit_theorem
8. http://en.wikipedia.org/wiki/Independent_component_analysis
9. http://sccn.ucsd.edu/~arno/indexica.html
10. http://www.cs.helsinki.fi/u/ahyvarin/whatisica.shtml
11. http://rsta.royalsocietypublishing.org/content/371/1984/20110534.full
12. http://www.informatica.si/PDF/35-1/10_Naik%20-%20An%20Overview%20of%20Independent%20Component%20Analysis%20and%20Its%20Applications.pdf

13. https://aaltodoc.aalto.fi/bitstream/handle/123456789/4610/isbn9789512298310.pdf?sequence=1
14. http://www.tqmp.org/Content/vol06-1/p031/p031.pdf

МОДЕЛЬ ПОТОКУ ДАНИХ АНАЛІЗУ ВІДНОСНОЇ КОНЦЕНТРАЦІЇ

Х.В. Ліп'яніна, аспірант
Тернопільський національний економічний університет

Рух до інформаційного суспільства супроводжується широким розповсюдженням розподілених інформаційних систем, які стають основним джерелом отримання знань та інформації. Розподілена інформаційна система – це набір незалежних вузлів (комп'ютерів), сполучених апаратно і взаємодіючих у рамках певної концепції, який для кінцевих користувачів виглядає як єдина централізована система. Потік даних – це засіб передачі скінченної або в граничному випадку, нескінченної кількості впорядкованих даних між компонентами обчислювальної моделі.

Концентрація виробництва дає можливість порівняти роль великих господарюючих суб'єктів (продавців) у виробництві конкретних товарів або наданні послуг. Вимірювання концентрації виробництва здійснюється в галузі за наслідками статистичних наглядів за підприємствами за такими показниками, як обсяг виробленої продукції у вартісному виразі, чисельність зайнятих, фонд зарплати, прибуток. Зазначені економічні характеристики по кожному з підприємств галузі зіставляються з галузевими показниками, після чого розраховується частка кожного підприємства в сукупному галузевому показнику.

Вимірювання ринкової концентрації припускає дослідження сфери товарного обігу — оптового або роздрібного ринку. Збіг показників концентрації виробництва і концентрації ринку буде тим більшим, чим вищі транспортні витрати, що обмежують переміщення товару в просторі; чим більш однорідний товар; чим більше поєднуються сфера виробництва і сфера обігу. Коли йдеться про концентрацію виробництва в масштабі національної економіки, то використовується термін «сукупна концентрація».

Коефіцієнт відносної концентрації розраховується як відношення часток найбільших підприємств ринку в загальній сумі підприємств до часткам продукції цих підприємств у загальному обсязі продукції, що випускається:

$$K = \frac{b}{a}$$

де K - коефіцієнт відносної концентрації;

b - частка найбільших підприємств ринку в загальній сумі підприємств, у відсотках;

a - частка продукції цих підприємств у загальному обсязі продукції, у відсотках.

Даний показник вимірюється в абсолютних значеннях. Чим ближче коефіцієнт до нуля, тим більш висока ступінь концентрації спостерігається на ринку. Легко переконатися в тому, що в разі ринку досконалої конкуренції, коли всі підприємства мають однакові і рівні частки, індекс дорівнює одиниці. Даний індекс володіє істотними перевагами, що вигідно відрізняють його від попереднього індексу, так як враховуються не тільки ринкові частки найбільших підприємств, але і число підприємств, що працюють на ринку. У той же час до цих пір невирішеною залишається проблема визначення числа найбільших підприємств, що включаються в цей індекс. Це може бути і три підприємства, і десять підприємств, і одне підприємство. Очевидно, що в кожному конкретному випадку потрібно самостійне визначення цього значення, що ускладнює практичне використання коефіцієнта відносної концентрації. До того ж, дуже складно дати тлумачення конкретним значенням коефіцієнта,