Chapter 3

Applications of independent component analysis

Erkki Oja, Juha Karhunen, Aapo Hyvärinen, Petteri Pajunen, Ricardo Vigário, Harri Valpola, Jaakko Särelä, Ella Bingham, Mika Inki, Antti Honkela, Tapani Raiko, Karthikesh Raju, Alexander Ilin, Rãzvan Cristescu, Simona Mãlãroiu, Kimmo Kiviluoto, Mika Ilmoniemi, Mark Girolami, Ata Kabán, Maria Funaro

3.1 Decision trees using independent component analysis

Petteri Pajunen and Mark Girolami

Decision trees are typically used as computationally efficient representations in classification or regression. For example binary trees allows one to reach a leaf node in $\log_2 n$ decisions when the tree contains n nodes in total. If the decisions are easy to compute, this may result in a very light computational process.

Typical problems involving decision trees deal with multivariate data where the components are usually attribute values with a clear interpretation. Decisions can be simply implemented as threshold values on one of the attributes. An example is an attribute $height(k)$ which is the height of person k. A decision could be a threshold $height(k) > 175cm$, which would split the data in two classes based on height.

One of the key properties of Independent Component Analysis is the ability to find linear combinations of multivariate data that have certain information-theoretic properties. Often these linear combinations represent underlying sources, which cannot be directly observed. If the data contains such hidden sources, decision based on thresholding observed data components may not be very effective.

In [1], single-component ICA was used to implement binary decisions in a decision tree. The key idea is that the independent components have more structure than the observed components, and therefore can be expected to be better candidates for linear threshold decisions. Since only single ICA component needs to be computed at each tree node, computational complexity resulting from a large number of ICA components can be avoided.

References

[1] P. Pajunen and M. Girolami. Implementing Decisions in Binary Decision Trees using Independent Component Analysis. In *Proceedings of the 2nd Int. Workshop on Independent Component Analysis and Blind Source Separation*, pages 483–487, Espoo, Finland, June 2000.

3.2 ICA for text mining

Ella Bingham, Ata Kabán, Mark Girolami

Independent component analysis (ICA) was originally developed for signal processing applications. Recently it has been found out that ICA is a powerful tool for analyzing text document data as well, if the text documents are presented in a suitable numerical form. This opens up new possibilities for automatic analysis of large textual data bases: finding the topics of documents and grouping them accordingly.

First approaches of using ICA in the context of text data considered the data static. In our recent study, we concentrated on text data whose topic changes over time. Examples of dynamically evolving text are chat line discussions or newsgroup documents. The dynamical text stream can be seen as a time series, and methods of time series processing may be used to extract the underlying characteristics — here the topics — of the data.

As a preprocessing step, the text stream is split into short windows, and from each window a T-dimensional vector is formed, where T is the size of the vocabulary; T is typically several thousands of terms. The i-th element of the vector indicates (some function of) the frequency of the i-th vocabulary term in the window. The high dimensionality of the data is reduced by singular value decomposition, as is often done before applying ICA-type algorithms on the data.

Our method of finding the topics of dynamical text is based on the *complexity pursuit* algorithm presented by Hyvärinen [3]. This algorithm is a generalization of projection pursuit to time series, and it is closely related to ICA. The algorithm gives us the directions into which the multidimensional time series data should be projected; these projections are then our estimates of the topic time series.

We give here an example of analyzing chat line discussions[1]. Results on newsgroup data with comparisons to other algorithms are presented in [1] and [2]. Figure 3.1 shows how different topic time series are activated at different times. We can see that the topics clearly are autocorrelated in time. The time span of Figure 3.1 is almost 24 hours; some topics are more or less persistent during the whole period and some will come up again after a few hours.

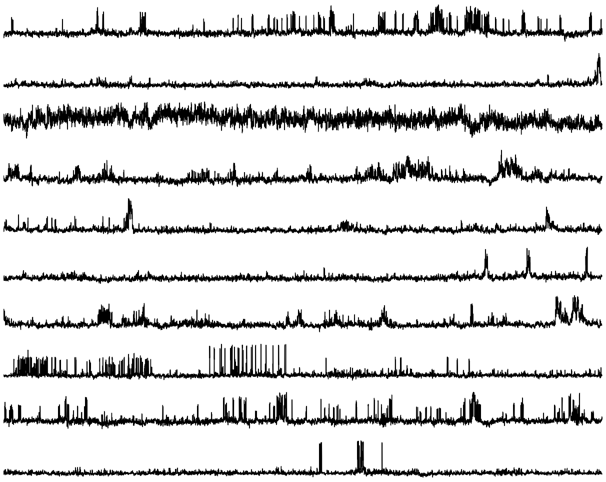


Figure 3.1: Activity of topics (vertical axis) in each chat window (horizontal axis). The uppermost time series corresponds to topic 1, the second to topic 2 etc.

[1] `http://www.cnn.com/chat/channel/cnn_newsroom`

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
jackson	site	bush	religion	violenc
sharpton	web	ashcroft	god	report
child	net	vote	jesu	youth
stori	word	kennedi	bibl	children
drudg	parent	presid	religi	gun
rainbow	nanni	cnn	life	point
monei	internet	time	follow	home
mistress	block	gore	read	drug
coalition	kid	question	stori	famili
tonight	system	elect	univers	satcher
pregnant	access	god	exist	health
affair	child	senat	faith	risk
black	base	power	man	factor
chenei	chat	thing	book	surgeon
jessi	page	fact	earth	prevent
Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
flag	california	join	tax	free
move	power	discuss	cut	liber
citi	electr	est	exempt	opinion
ncaa	energi	tonight	monei	religion
offici	blackout	room	gop	form
atlanta	state	studio	hous	polit
count	deregul	cnn	congress	conserv
game	compani	conserv	pay	birth
night	crisi	american	interest	philosophi
georgia	price	nea	recess	establish
chang	plant	union	payer	narrow
lose	util	keen	secur	restrict
confeder	order	type	henri	independ
hehe	home	chat	hypocrit	orthodox
chenei	cost	newsroom	hyde	bound

Table 3.1: Keywords of chat line discussion topics related to the time series in Figure 3.1.

The most significant terms of the topics are listed in Table 3.1. It is seen that each keyword list indeed characterizes one distinct topic quite clearly. Topic 1 deals with Jesse Jackson and his illegitimate child, topic 2 is about parental control over children's web usage and topic 3 is a general discussion about G.W. Bush. Topic 4 is a religious discussion, topic 5 deals with problems of the youth such as violence and drug abuse, and topic 6 is about the controversial flag of the state of Georgia, US, due to which the NCAA basketball games risked cancellation in Atlanta. Topic 7 involves the energy shortage in California, topic 8 corresponds to comments given by the chat line moderator, topic 9 is about taxation and topic 10 is a short discussion dealing with the values of the politicians in the US.

To conclude, our method finds meaningful topics inherent in the data, and the experimental results suggest the applicability of the method to query-based retrieval from a temporally changing text stream.

References

[1] Bingham, E. Topic identification in dynamical text by extracting minimum complexity time components. In: *Proc. 3rd International Conference on Independent Component Analysis and Blind Signal Separation (ICA2001)*, December 9–13, 2001, San Diego, CA, USA, pp. 546–551.

[2] Bingham, E., A. Kabán, and M. Girolami. Topic identification in dynamical text by complexity pursuit. *Neural Processing Letters*. Submitted.

[3] Hyvärinen, A. Complexity pursuit: separating interesting components from time-series. *Neural Computation* **13**(4), 883–898.

3.3 ICA for analyzing financial time series

Simona Mãlãroiu, Kimmo Kiviluoto, Erkki Oja

It is a tempting alternative to try ICA on financial data. There are many situations in which parallel financial time series are available, such as currency exchange rates or daily returns of stocks, that may have some common underlying factors. ICA might reveal some driving mechanisms that otherwise remain hidden. If one could find the maximally independent mixtures of a set of original stocks, i.e. independent portfolios, this might help in minimizing the risk in the investment strategy.

Another promising application is the prediction of financial time series. The ICA transformation tends to produce component signals $s_j(t)$ that can be compressed with fewer bits than the original signals $x_i(t)$. They are thus more structured and regular. This gives motivation to try to predict the signals $x_i(t)$ by first going to the ICA space, doing the prediction there, and then transforming back to the original time series. The prediction can be done separately and with a different method for each independent component, depending on its time structure. Hence, some interaction from the user may be needed in the overall prediction procedure.

An algorithm for this was suggested in [1, 2, 3]. After subtracting the mean of each time series and prewhitening, the independent components $s_j(t)$ and the mixing matrix $\mathbf{A}$ are estimated using the FastICA algorithm. For each component $s_j(t)$, a suitable nonlinear filtering is applied to reduce the effects of noise. Each smoothed independent component is then predicted separately for a number of steps into the future, for instance using some method of AR modeling. The final predictions for the original time series $x_i(t)$ are obtained by inverting the ICA transformation.

To test the method, we applied our algorithm on a set of 10 foreign exchange rate time series. The results were promising, as the ICA prediction performed better than direct prediction. Fig. 3.2 shows an example of prediction using our method. The upper figure represents one of the original time series (mixtures) and the lower one the forecast obtained using ICA prediction for a future interval of 50 time steps. The algorithm seemed to predict very well especially the turning points.

References

[1] Mãlãroiu, S., Kiviluoto, K. and Oja, E.: Time series prediction with Independent Component Analysis. *Proc. Int. Conf. on Advanced Investment Technology*, December 19 - 21, 1999, Queensland, Australia (1999).

[2] Mãlãroiu, S., Kiviluoto, K. and Oja, E.: ICA preprocessing for time series prediction. *Proc. 2nd Int. Workshop on Indep. Comp. Anal. and Blind Source Separation*, June 19 - 22, 2000, Helsinki, Finland, pp. 453 - 457 (2000)

[3] Oja, E., Kiviluoto, K. and Mãlãroiu, S.: Independent component analysis for financial time series. *Proc. IEEE 2000 Symp. on Adapt. Systems for Signal Proc., Comm. and Control AS-SPCC*, Oct. 1 - 4, 2000, Lake Louise, Canada, pp. 111 - 116 (2000).

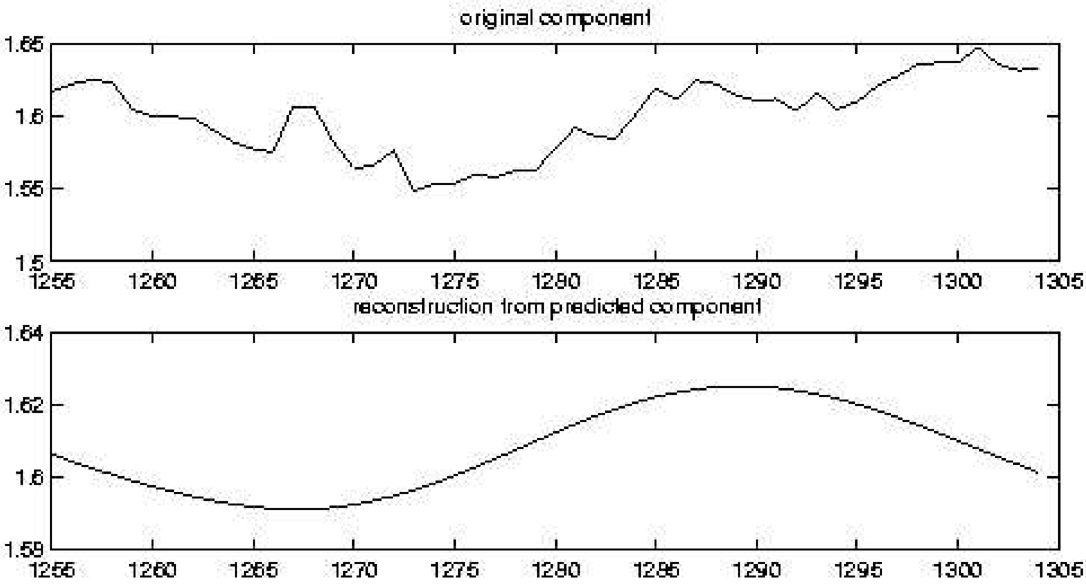


Figure 3.2: Prediction of real-world financial data: the upper figure represents the actual future outcome of one of the original mixtures and the lower one the forecast obtained using ICA prediction for an interval of 50 values.

3.4 ICA for astronomical data

Maria Funaro, Erkki Oja, Harri Valpola

A new approach in our group is to try ICA for removing artifacts from astronomical telescope images. This problem was introduced to us by Ms. Maria Funaro who finalized her Ph.D. Thesis in our lab in spring 2001.

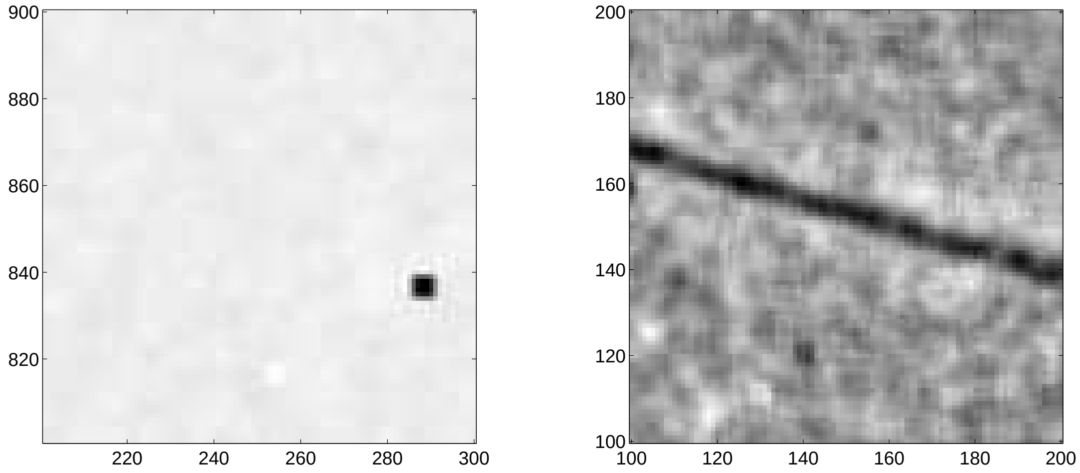
In modern astrophysics, one of the main research directions is understanding the dark matter in the universe. Of special interest are compact objects with substellar mass, such as black holes, dwarf stars, or planets. When such an object passes near the line of sight of a star, the luminosity of the star will increase – an effect called gravitational lensing, predicted by the general theory of relativity.

In studying other galaxies than our own, individual stars cannot be resolved, but a whole group of unresolved stars is registered in a single pixel element of a telescope ccd camera. In a new technique called pixel lensing (see [1]), the pixel luminosity variations over time are monitored, and using these time series the lensing events can yet be detected even in the case of unresolved stars.

A problem in the analysis of the images and luminosity variations is the presence of artefacts. One of the possible artefacts are the resolved or individual stars between the far-out galaxy and the camera, which emerge sharply from the luminosity background. Other artefacts are cosmic rays, atmospheric events, and noise in the CCD camera. Separating these artefacts from possible physical events is one of the necessary steps in the analysis of pixel lensing data.

The new idea proposed by us [2] is to use ICA for the artefact detection and removal. This is motivated by the fact that for astrophysical data, the independence of the artefacts is often theoretically guaranteed, and also the linear mixing model holds exactly. This is an almost ideal application for ICA.

In the astrophysical data, we have a number of digital images, recorded over consequent nights when the conditions are favourable, and carefully calibrated for geometrical and photometric alignments. ICA is then used for this image set to reveal independent components that might be artefacts. An example is given in Fig. 3.3.



a) b)

Figure 3.3: Two artefacts: (a) cosmic ray, (b) atmospheric effect.

References

[1] A. Gould, The theory of pixel lensing, *Astrophys. J.* 470, 1996, pp. 201 - 210.

[2] Funaro, M., Oja, E. and Valpola, H.: Artefact detection in astrophysical image data using Independent Component Analysis. *Proc. 3rd Int. Workshop on Indep. Comp. Anal. and Blind Source Separation*, Dec. 9 - 12, 2001, San Diego, USA, pp. 43 - 48 (2001).

3.5 ICA in CDMA communications

Rãzvan Cristescu, Karthikesh Raju, Juha Karhunen, Erkki Oja

In wireless communication systems, like mobile phones, an essential issue is division of the common transmission medium among several users. A primary goal is to enable each user of the system to communicate reliably despite the fact that the other users occupy the same resources, possibly simultaneously. As the number of users in the system grows, it becomes necessary to use the common resources as efficiently as possible. These two requirements have given rise to a number of multiple access schemes [2]. The traditional ones are FDMA and TDMA schemes, based on the use of either non-overlapping frequency or time slots assigned to each user, respectively.

During the last years, various systems based on CDMA (Code Division Multiple Access) techniques [1, 2] have become popular. In CDMA, there is no disjoint division in frequency or time slots, but the users are identified by their unique codes. CDMA systems offer several advantages over FDMA and TDMA techniques. Their capacity is larger, and it degrades gradually with increasing number of simultaneous users who can be asynchronous. On the other hand, CDMA systems require more advanced signal processing methods, and correct reception of CDMA signals is more difficult because of several disturbing phenomena [1, 2]. A serious problem is multipath propagation which is illustrated in Figure 3.4. In practice, one must typically resort to suboptimal techniques in processing CDMA signals for several reasons. Optimal methods may require prior information which is not available, they are usually computationally too demanding, and the assumptions made may not hold because of rapidly varying environment.

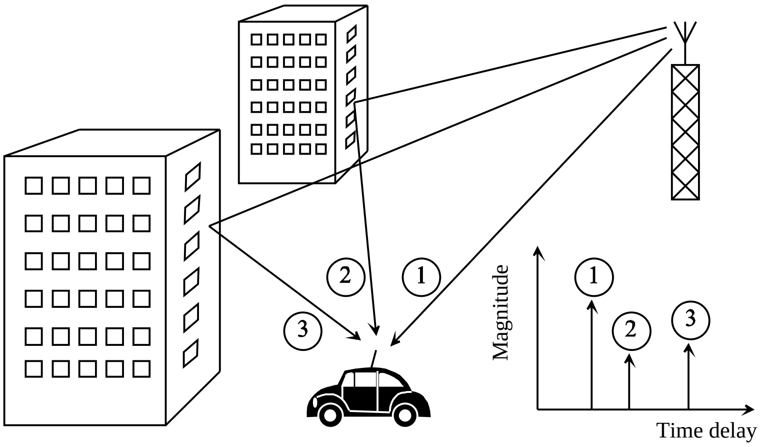


Figure 3.4: An example of multipath propagation in urban environment.

It turns out that direct sequence CDMA data can be cast in the form of a linear ICA/BSS data model [3]. However, the situation is not completely blind, because there is some prior information available. In particular, the transmitted symbols have a finite number of possible values, and the spreading code of the desired user is known. On the other hand, multipath propagation, possibly fading channels, and time delays make separation of the desired user's symbols a very challenging estimation problem which is more complicated than the standard linear ICA problem.

In the first stage of this project, we applied ICA and BSS methods to various problems in multiuser detection [1, 2], trying to take into account the available prior information whenever possible. In particular, we have considered estimation of the desired user's time delays in [5], estimation of fading channels in [4], and detection of the desired user's symbol sequence in [6]. The results are very promising, showing that ICA based methods can yield considerably better performances than more conventional methods based on second-order statistics. The work done during this stage is reviewed with the necessary background in Chapter 23 of the book [3].

In the second stage of the project, we have applied independent component analysis to blind suppression of interference caused by bit-pulsed jamming in a direct sequence CDMA communication system. This jamming problem is important in practical CDMA communication systems. We have taken into account both data modulation and temporally uncorrelated jamming, improving and extending earlier preliminary work on the same problem. Computer simulations show that the proposed method performs better than the well-known RAKE method, which is the standard choice for suppressing jammer signals. The results are reported in more detail in forthcoming conference papers [7, 8].

References

[1] S. Verdu, *Multiuser Detection.* Cambridge Univ. Press, 1998.

[2] J. Proakis, *Digital Communications.* McGraw-Hill, 3rd edition, 1995.

[3] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis.* Wiley, 2001, 481+xxii pages.

[4] R. Cristescu, J. Joutsensalo, J. Karhunen, and E. Oja, A complexity minimization approach for estimating fading Gaussian channel in CDMA communications. In *Proc. of the 2nd Int. Workshop on Independent Component Analysis and Blind Source Separation (ICA2000)*, Espoo, Finland, June 19-22, 2000, pages 527-532.

[5] R. Cristescu, T. Ristaniemi, J. Joutsensalo, and J. Karhunen, CDMA delay estimation using a fast ICA algorithm. In *Proc. of the IEEE Int. Symp. on Personal, Indoor, and Mobile Communications (PIMRC'00)*, London, United Kingdom, September 17-19, 2000.

[6] R. Cristescu, T. Ristaniemi, J. Joutsensalo, and J. Karhunen, Blind separation of convolved mixtures for CDMA systems. In *Proc. of the X European Signal Processing Conference (EUSIPCO 2000)*, Tampere, Finland, September 5-8, 2000, pages 619-622.

[7] T. Ristaniemi, K. Raju, and J. Karhunen, Jammer mitigation in DS-CDMA array systems using independent component analysis. To appear in *Proc. of the 2002 IEEE Int. Conf. on Communications (ICC2002)*, New York City, NY, USA, April 28–May 2, 2002.

[8] K. Raju, T. Ristaniemi, J. Karhunen, and E. Oja, Suppression of bit-pulsed jammer signals in DS-CDMA array systems using independent component analysis. To appear in *Proc. of the 2002 IEEE Int. Symp. on Circuits and Systems (ISCAS2002)*, Phoenix, Arizona, USA, May 26-29, 2002.