Independent Component Analysis Using Random Projection For Data Pre-Processing

Adeel Ansari[1], Afza Bt. Shafie[2] and Abas B Md Said[3]

[1] Computer Information Systems Department, Universiti Teknologi PETRONAS, Bandar Seri Iskandar, Tronoh, Perak, Malaysia

[2] Fundamentals and Applied Sciences Department, Universiti Teknologi PETRONAS, Bandar Seri Iskandar, Tronoh, Perak, Malaysia

[3] Computer Information Systems Department, Universiti Teknologi PETRONAS, Bandar Seri Iskandar, Tronoh, Perak, Malaysia

Abstract

There is an inherent difficulty of finding out latent structures within high dimensional data repository centers. It is assumed that this data is generated by these unknown latent variables and with the relationship and interaction between each of them. The task is to find these latent variables and the way they interact, given the observed data only. It is assumed that the latent variables do not depend on each other but act independently. A popular method for counteracting with the above stated problem scenario is independent component analysis (ICA). An ICA algorithm for analyzing complex valued signals is given; and an ICA-type algorithm is used for analyzing the topics in dynamically changing text data. Experimental results are given on all of the presented methods. Another, partially overlapping problem considered in this paper is dimensionality reduction. Empirical validation is given on a computationally simple method called random projection: it does not introduce severe distortions in the data. It is also proposed that random projection could be used as a preprocessing method prior to ICA, and experimental results are shown to support this claim.

Keywords: *Data Mining, Random Projection, Independent Component Analysis, ICA, RP.*

1. Introduction

The scope of the paper, considers the problem of seeking latent structures in high dimensional data sources. The term latent means hidden, unknown or unobserved; the term structure refers to some regularities in the data; high dimensional may be tens or tens of thousands of dimensions, depending on the situation; and data is any information that can be transformed into numerical values, most often represented as a matrix of multidimensional observations where each dimension corresponds to a variable whose value we can somehow measure. The aims in this thesis are to answer the question "What is there in the data?", to form a simple representation of a large data set that is difficult to analyze as such, and to present the data in a form that is understandable to a human observer[7].

Throughout the paper, it will be assumed that the observed data are generated by interactions between latent variables. The objective is to find out what these latent variables are and how they interact — this is the key to understanding what the data are about. The latent variables will be called components, sources or topics: the data are composed of these latent variables, or the latent variables are the sources of variability in the data, or in particular in text document data the latent variables are the topics of discussion. Depending on the point of view, the "structure" in the data we referred to in the beginning is either due to the values taken by the latent variables or due to the way the latent variables interact. Throughout this paper, we will assume that there are no inherent dependencies between the latent variables.

In addition to revealing the latent structure in high dimensional data, another aim of this paper is to present ways of reducing the dimensionality of the data. This aim overlaps partially with the first one: we wish to transform the data into a denser representation and only retain the most important aspects of the data. The approach/methodology is presented within the figure shown below [3,6,7]:

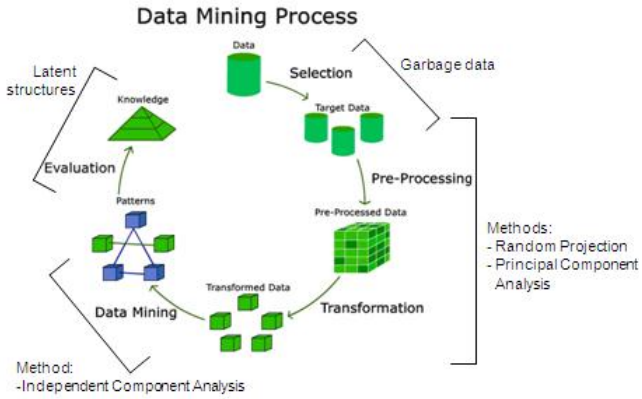

Figure 1: Data Mining Process Methodology

2. Related work

To rigorously define ICA (Jutten and Hérault, 1991; Comon, 1994), we can use a statistical "latent variables" model. Assume that we observe *n* linear mixtures $\mathbf{x_1}, \ldots, \mathbf{x_n}$ of n independent components

$$\mathbf{x_j} = \mathbf{a_{j1}s_1}+\mathbf{a_{j2}s_2}+\ldots+\mathbf{a_{jn}s_n} \qquad (1)$$

In the ICA model, we assume that each mixture $\mathbf{x_j}$ as well as each independent component $\mathbf{s_n}$ is a random variable, instead of a proper time signal. The observed values $\mathbf{x_j}$, e.g., the microphone signals in the cocktail party problem, are then a sample of this random variable. Without loss of generality, we can assume that both the mixture variables and the independent components have zero mean: If this is not true, then the observable variables $\mathbf{x_i}$ can always be centered by subtracting the sample mean, which makes the model zero-mean. It is convenient to use vector-matrix notation instead of the sums like in the previous equation. Using this vector-matrix notation, the above mixing model is written as

$$\mathbf{x} = \mathbf{As} \qquad (2)$$

(2) is the ICA generative model, where A is the mixing matrix, **s** is the source data and **x** is the mixed or observed data. After estimating the matrix **A**, we can compute its inverse, say **W**, and obtain the independent component simply by[1]:

$$\mathbf{s} = \mathbf{Wx} \qquad (3)$$

ICA Algorithms

There are number of ICA algorithms that are available to researchers such as:

- **FASTICA Algorithm.**

  The FastICA algorithm is a computationally highly efficient method for performing the estimation of ICA. It uses a fixed-point iteration scheme that has been found in independent experiments to be 10-100 times faster than conventional gradient descent methods for ICA. Another advantage of the FastICA algorithm is that it can be used to perform projection pursuit as well, thus providing a general-purpose data analysis method that can be used both in an exploratory fashion and for estimation of independent components (or sources) [11].

- **Complexity Pursuit.**

  Projection pursuit is a technique for exploratory data analysis with emphasis on visualization. It is based on finding low-dimensional projections of multivariate data that show highly non-gaussian distributions. Projection pursuit is technically very closely related to ICA[14].

Ambiguities of ICA

In the ICA model, it is easy to see that the following ambiguities will hold:

1. We cannot determine the variances (energies) of the independent components.

The reason is that, both **s** and **A** being unknown, any scalar multiplier in one of the sources could always be cancelled by dividing the corresponding column of **A** by the same scalar.

2. We cannot determine the order of the independent components.

The reason is that, again both **s** and **A** being unknown, we can freely change the order of the terms and call any of the independent components the first one [1].

3. Pre-processing Data for ICA

Estimating ICA in the original, high-dimensional space may lead to poor results. Therefore, it is often beneficial to reduce the dimensionalities of the observed data, using a high dimensionality reduction method, most preferably, in this paper, Random Projection method is considered and compared with other available methods as well for e.g. principal component analysis, singular value decomposition, discrete cosine transform and median filtering.

Random Projection

In many applications of data mining, the high dimensionality of the data restricts the choice of data processing methods. Such application areas include the analysis of market basket data, text documents, image data and so on; in these cases the dimensionality is large due to either a wealth of alternative products, a large vocabulary, or the use of large image windows, respectively[5]. A statistically optimal way of dimensionality reduction is to project the data onto a lower-dimensional orthogonal subspace that captures as much of the variation of the data as possible. The best (in mean-square sense) and most widely used way to do this is principal component analysis

(PCA); unfortunately it is quite expensive to compute for high-dimensional data sets.

Random projections have recently emerged as a powerful method for dimensionality reduction. Theoretical results indicate that the method preserves distances quite nicely; however, empirical results are sparse. Computing the PCA of a high-dimensional data set is computationally burdensome. In this thesis it is proposed that random projection (RP) is a suitable preprocessing method for ICA: using RP before PCA significantly reduces the computational load without introducing severe distortions in the data set. The performance of random projection was compared to several other methods of dimensionality reduction: principal component analysis, singular value decomposition, discrete cosine transform and median filtering[9].

The measure of performance was the distortion in the similarity of randomly chosen data vectors that took place when the dimensionality of the data was reduced. The similarity of two data vectors was computed by using either their Euclidean distance or inner product. Also, the computational complexities of the dimensionality reduction methods were compared by measuring the number of floating point operations. The results indicate that random projection is a promising method for dimensionality reduction that does not introduce a great distortion in the data, while being computationally very simple[2].

Method

In random projection, the original *d*-dimensional data is projected to a *k*-dimensional (*k* << *d*) subspace through the origin, using a random $k \times d$ matrix $\mathbf{R}$ whose columns have unit lengths. Using matrix notation where $\mathbf{X_{d\times N}}$ is the original set of *N d*-dimensional observations,

$$\mathbf{X^{RP}_{k\times N}} = \mathbf{R_{k\times d}X_{d\times N}} \qquad (4)$$

is the projection of the data onto a lower *k*-dimensional Sub-space. The key idea of random mapping arises from the Johnson-Lindenstrauss lemma [15]: if points in a vector space are projected onto a randomly selected subspace of suitably high dimension, then the distances between the points are approximately preserved.

When comparing the performance of random projection to that of other methods of dimensionality reduction, it is instructive to see how the similarity of two vectors is distorted in the dimensionality reduction. We measure the similarity of data vectors either as their Euclidean distance or as their inner product.

After the random projection, this distance is approximated by the scaled Euclidean distance of these vectors in the reduced space:

$$\sqrt{(d/k)}\,\|\mathbf{Rx_1} - \mathbf{Rx_2}\| \qquad (5)$$

where **d** is the original and **k** the reduced dimensionality of the data set. The scaling term $\sqrt{(d/k)}$ takes into account the decrease in the dimensionality of the data: according to the Johnson-Lindenstrauss lemma, the expected norm of a projection of a unit vector onto a random subspace through the origin is $\sqrt{(k/d)}$ [2].

4. Simulation Results

The tool used is Matlab version version 7. A sample mixed signal is taken with a total number of iterations 224, as shown in figure 2.

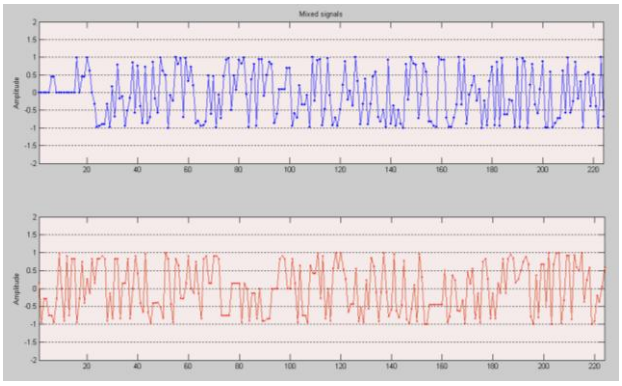

Figure 2: Mixed signals

The Eigen-value chart for both the two identified components are evaluated, as shown in figure 3.

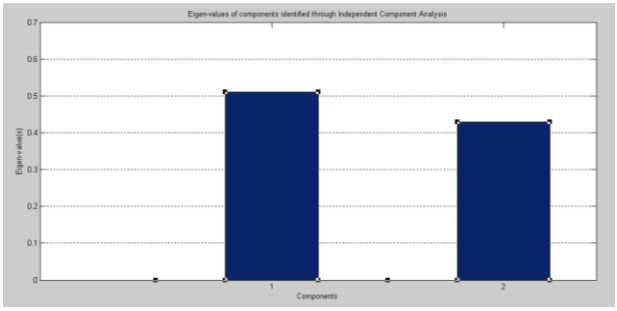

Figure 3: Eigen-value chart

The signals are now whitened or de-noised using Random Projection (RP) Method, as shown in figure 4.

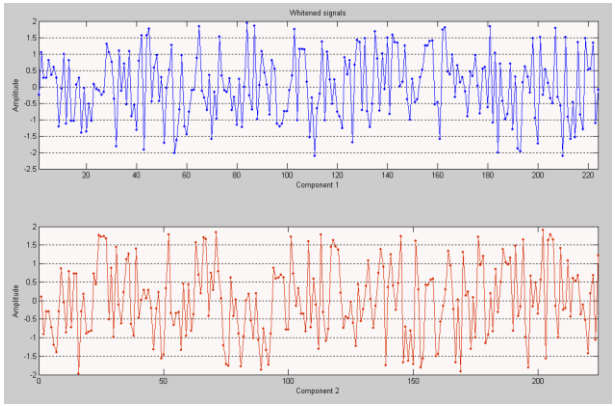

Figure 4: Whitened or de-noised using RP

The estimated source signals, as determined using FASTICA method, as shown in figure 5.

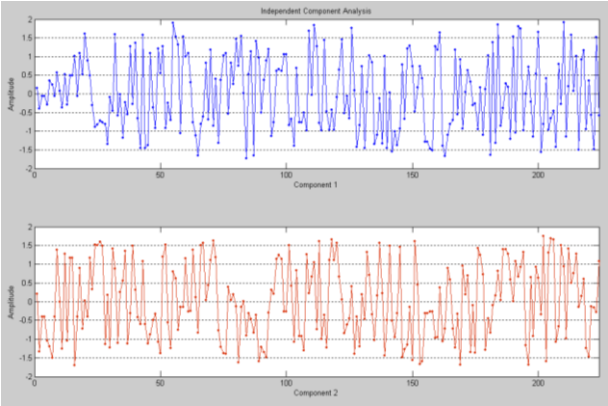

Figure 5: Estimated source signals using FASTICA

Convergence of ICA estimation of complex valued signals in comparisons with results attained from PCA (red line) and RP (blue line). Differences from the results of both pre-processing methods gave a very least dissimilarity with respect to data mining analysis, as observed from this figure, the first ten iterations varied with different amplitude levels, whilst the remaining coincided perfectly to each other. Therefore, both pre-processing methods have a 90% similarity level with a 10% dissimilarity level, as shown in figure 6.

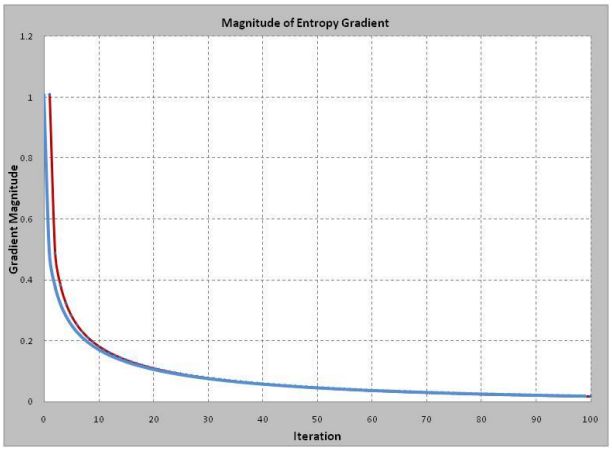

Figure 6: RP (blue curve) & PCA (red curve) sum of squared errors.

5. Conclusion

The problem addressed by this paper was to represent a large data set in a reduced and concise manner. A natural task in such a setting is to find out this internal or latent structure and thus obtain a simple representation of the data. Using the observed data only, to find these unknown latent variables and the way they interact. ICA has been applied more or less successfully to various different problems in a multitude of application areas.

The first extension was from real valued to complex valued signals. FASTICA algorithm was considered, which is simple but computationally efficient for separating complex valued linearly mixed signals.

The second extension was the use of ICA in textual data. It was shown that by using an ICA-type method developed for time-dependent signals, called complexity pursuit, the topics in such data can be found and visualized in a convenient way. Independent Component Analysis has been progressing with the advent of time and is being considered to focus on other problem settings. The extensions presented in this paper are by far not the only possible ones.

A structure worth studying is one in which some latent variables have inhibitory effects: for example, if some topic is active, then some other topic must stay inactive. Or, a topic favors the appearance of a term and inhibits the appearance of another. The field of bioinformatics has several problems where latent variable methods could prove useful, too.

Acknowledgments

The authors would like to express their gratitude and appreciation to those who have contributed and facilitated towards the success of this paper and more particularly to Universiti Teknologi PETRONAS for the facilities provided to carry out the simulation work.

References

[1] Ella Bingham, Heikki MannilaYu-Lu LIU, *" Random projection in dimensionality reduction: Applications to image and text data"*, 2009 Sixth International Conference on Fuzzy Systems and Knowledge.

[2] S. Kaski,"*Dimensionality reduction by random mapping*". In Proc. Int. Joint Conf. on Neural Networks, volume 1, pages 413–418, 1998.

[3] E. J. Keogh, M. J. Pazzani,"*A simple dimensionality reduction technique for fast similarity search in large time series databases"*, In 4th Pacific-Asia Conf. on Knowledge Discovery and Data Mining, 2000.

[4] Xiaoli Zhang Fern and Xiaoli Zhang Fern, *"Random Projection for High Dimensional Data Clustering: A Cluster Ensemble Approach"*. In Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.

[5] R. I. Arriaga and S. Vempala,"*An algorithmic theory of learning: robust concepts and random projection*", In Proc. 40th Annual Symp. on Foundations of Computer Science, pages 616–623. IEEE Computer Society Press, 1999.

[6] S. Dasgupta,"*Experiments with random projection*", In Proc. Uncertainty in Artificial Intelligence, 2000.

[7] P. Frankl, H. Maehara*,"The Johnson-Lindenstrauss lemma and the sphericity of some graphs*", Journal of Combinatorial Theory, Ser. B, 44:355–362, 1988.

[8] L. Sirovich, R. Everson,"*Management and analysis of large scientific datasets",* Int. Journal of Supercomputer Applications, 6(1):50–68, spring 1992.

[9] Thomas Kolenda, Lars Kai Hansen, and Jan Larsen,"*Signal detection using ICA: application to chat room topic spotting",* In Proceedings of the Third International Conference on Independent Component Analysis and Signal Separation (ICA2001), pages 540–545, 2001.

[10] Charles Lee Isbell, Paul Viola,"*Restructuring sparse high dimensional data for effective retrieval*", In Advances in Neural Information Processing Systems 11, pages 480–486, 1998.

[11] Aapo Hyvärinen, Erkki Oja,"*A fast fixed-point algorithm for independent component analysis*", Neural Computation, 9:1483–1492, 1997.

[12] Aapo Hyvärinen,"*Fast and robust fixed-point algorithms for independent component analysis*", IEEE Transactions on Neural Networks, 10(3):626–634, May 1999.

[13] David Hand, Heikki Mannila, and Padhraic Smyth,"*Principles of Data Mining*", The MIT Press, 2001.

[14] M. Jones and Robin Sibson,*"What is projection pursuit?",*Journal of the Royal Statistical Society, ser. A, 150:1–36, 1987.

[15] Jean-François Cardoso,"*High-order contrasts for independent component analysis*", Neural Computation, 11(1):157–192, 1999.

[16] Aapo Hyvärinen.*"Complexity pursuit: separating interesting components from time series. Neural Computation",* 13(4):883–898, 2001.

Adeel Ansari is a researcher at the Universiti Teknologi PETRONAS University in Malaysia. His area of expertise are in the field of signal processing and data mining.

Afza Bt. Shafie is an Associate Professor at the Universiti Teknologi PETRONAS University in Malaysia. Her research area and expertise are in the area of Fundamental mathematics and Applied Sciences. She has a number of publications in the field of CSEM - Seabed Logging Application.

Abas B Md Said is an Associate Professor at the Universiti Technologi PETRONAS University in Malaysia. His research area and expertise are in the field of computer graphics, Networking and signal processing algorithms. He has a number of publications in the field of Computer Information Sciences.