A MATLAB TOOLBOX FOR DATA REDUCTION, VISUALIZATION, CLASSIFICATION AND KNOWLEDGE EXTRACTION OF COMPLEX BIOLOGICAL DATA

A. Mohammad-Djafari*, *G. Khodabandelou*† *and J. Lapuyade-Lahorgue* ‡

Laboratoire des signaux et systèmes (L2S)
UMR 8506 CNRS-SUPELEC-UNIV PARIS SUD
plateau de Moulon, 3 rue Joliot-Curie, 91192 GIF-SUR-YVETTE Cedex, France

ABSTRACT

In this paper, first we present A Matlab toolbox which gives the possibility to simulate the data for testing the algorithms such as: Principal Component Analysis (PCA), Factor Analysis(FA), Independent Component Analysis (ICA), Linear Discriminant Analysis (LDA) and many other classification methods which can be used in Data Reduction (DR), Data Visualization (DV), supervised and unsupervised classification of multivariate great dimensional biological data. Then, we describe some biological experiments related to studying the circadian cell cycles and cancer treatment where the biologists observe different kind of data such as the variations of temperature, activity, hormones, genes and proteins expressions. These data are often complex: multivariate, great dimensionality, heterogeneous, with missing data, and observed at different sampling rates. The classical methods of PCA, FA, ICA and LDA can not directly handle these data. In this paper, we show how this toolbox can help them to visualize, to analyse and to do classifications on these data and finally to extract some knowledge from them.

Keywords: Data visualization, Dimensionality reduction, Principal Component Analysis, Factor Analysis, Independent Component Analysis, Linear Discriminant Analysis, Bayesian inference, Sources separation, Inverse problems.

1. INTRODUCTION

In many biological experiments, we are always face to data sets which are heterogeneous, of great dimensionality with missing and outliers data. To understand these data, first we need to visualize them, but the great dimensionality of these data needs a Data Reduction (DR) step. Principal Component Analysis (PCA), Factor Analysis (FA), Independent Component Analysis (ICA) and Linear Discriminant Analysis (LDA) methods are the main classical methods for analyzing high dimensional data [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]. PCA, FA and ICA are mainly used for dimensionality reduction and LDA for supervised classification. Even if these methods are well defined, still there exist different algorithms for their practical usage: PCA and FA are the most stable ones because they use quadratic criteria and L2 norms (second order statistics in statistical interpretation and Gaussian hypothesis in probabilistic interpretation) and so they are very simple to implement. The characteristics of the results obtained by PCA and FA are well known. For example, we know that the factors are obtained up to a rotation indetermination. ICA is more complex because the criteria used to be optimized are often non quadratic (Kullback-Leibler divergence) and use higher order statistics (HOS) and non Gaussian probability laws. The corresponding algorithms are then more sophisticated. However the common properties of independent components are that they are obtained up to a permutation and scale factor indetermination. LDA can be considered as a particular supervised classification method where we know the number of classes.

In this paper, in a first step, we present, very shortly, but in a unifying way of forward and inverse problem, different multivariate data analysis tools. Then, we present a Matlab toolbox: to generate different factors with different properties; to generate different data sets with linear or non linear dependencies; to add different kind of errors; to apply different algorithms of PCA, FA, ICA, LDA, ... and to compare the obtained results. In a second step, we show some preliminary results for real data set obtained by biologists working on circadian and cell cycle influence on cancer. This work is done in collaboration within the European project EraSysBio.

2. A UNIFYING PRESENTATION OF MULTIVARIATE DATA ANALYSIS METHODS THROUGH FORWARD AND INVERSE MODELING

PCA, FA, ICA and LDA are classical methods of dimensionality reduction and data analysis. Due to the origin of these methods, there have been many different presentations and interpretations. Here, we present them in an unifying context of forward modeling and inversion. To do this, we start by defining the factors $\boldsymbol{f}(t) = [f_1(t), \cdots, f_N(t)]$ which is an N-dimensional vector of time series. Here, we choosed time series due to our final application. However, the time index can be anything else, for example, just the index of experiments of a position on a line, in a plane or in space.

In a first step, we assume that the observed data $\boldsymbol{g}(t) = [g_1(t), \cdots, g_M(t)]$ are obtained via a mixing (or loading) matrix $\boldsymbol{A}$ of dimensions $[M \times N]$ through the forward model

$$\boldsymbol{f}(t) \longrightarrow \boxed{\begin{array}{c}\text{Forward}\\ \text{model } \boldsymbol{A}\end{array}} \longrightarrow \overset{\downarrow \boldsymbol{\epsilon}}{\bigoplus} \longrightarrow \boldsymbol{g}(t) = \boldsymbol{A}\,\boldsymbol{f}(t) + \boldsymbol{\epsilon}(t),\; t = 1, \cdots, T \tag{1}$$

where $\boldsymbol{\epsilon}$ represents the errors of modeling and T is the total number of observed samples.

Using this forward model, the objective of many data analysis methods such as PCA, FA, ICA and LDA is to obtain the factor $\boldsymbol{f}$ and the loading matrix $\boldsymbol{A}$. Described as such, we see that this estimation problem is very ill-posed in the sense that we can find many

Senior Researcher at Laboratoire des signaux et systmes (L2S)

PhD candidate at Laboratoire des signaux et systmes (L2S)

Post-doc at Laboratoire des signaux et systmes (L2S)

This work is a part of ERASYSBIO-C5Sys European project ”Circadian and cell cycle clock systems in cancer”: `http://www.erasysbio.net/index.php?index=272`

combinations of factors and loading matrix which can satisfy this model. In the following, we use this model to explain the differences between PCA, FA, ICA and LDA.

PCA and FA methods try to find uncorrelated factors $\widehat{\boldsymbol{f}}$. Because correlation describes a linear dependence, the main assumption is then that $\widehat{\boldsymbol{f}}$ has to be obtained through a linear combination of the data: $\widehat{\boldsymbol{f}}(t) = \boldsymbol{B}\,\boldsymbol{g}(t)$, where the matrix $\boldsymbol{B}$ is called separating (or demixing or deloading) matrix.

$$\boldsymbol{g}(t) \longrightarrow \boxed{\begin{array}{c}\text{Inference}\\ \text{PCA, FA, ICA}\\ \text{LDA, Bayes}\end{array}} \begin{array}{l}\longrightarrow \widehat{\boldsymbol{A}} \text{ or } \widehat{\boldsymbol{B}}\\ \\ \longrightarrow \widehat{\boldsymbol{f}}(t) = \widehat{\boldsymbol{B}}\,\boldsymbol{g}(t)\end{array} \qquad (2)$$

Here, we are not going to describe these algorithms which are described elsewhere in details [11, 12, 13, 14, 15], but we present a Matlab toolbox in which we implemented all these methods.

3. PRESENTATION OF THE MATLAB SIMULATION TOOLBOX

We have developed a menu driven simulation tool, which has, as the main menu, the following steps:
– Generation of different sources (factors) with different properties (Uniform, Gaussian, Mixture of Gaussian, ... ,
– Generation of different data sets with linear or nonlinear dependencies,
– Addition of different kind of errors,
– Application of different algorithms of PCA, FA, ICA, LDA, ... and
– Visualization and evaluation tools which give possibility to evaluate the performances of a given method or to compare the results obtained by two different methods.

As tools to measure the performances of these methods, we propose the following scheme:

$$\boldsymbol{f} \longrightarrow \boxed{\begin{array}{c}\text{Forward}\\ \text{model } \boldsymbol{A}\end{array}} \longrightarrow \overset{\downarrow\,\boldsymbol{\epsilon}}{\oplus} \longrightarrow \boldsymbol{g}$$

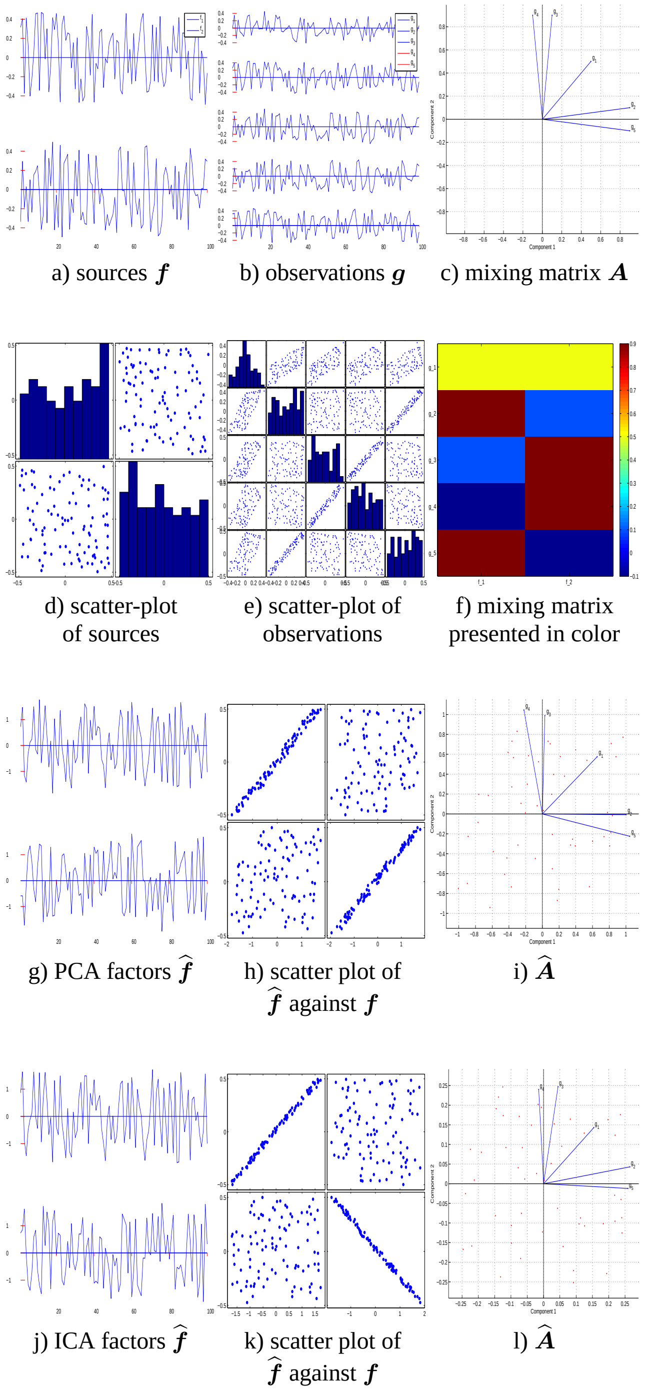
$$\boldsymbol{g} \longrightarrow \boxed{\begin{array}{c}\text{Inference}\\ \text{PCA, FA, ICA}\\ \text{LDA, Bayes}\end{array}} \begin{array}{l}\longrightarrow \widehat{\boldsymbol{A}}\\ \longrightarrow \widehat{\boldsymbol{f}}\end{array} \longrightarrow \boxed{\begin{array}{c}\text{Estimated}\\ \widehat{\boldsymbol{A}}\end{array}} \longrightarrow \widehat{\boldsymbol{g}}$$

and then compare $\widehat{\boldsymbol{g}}$ with $\boldsymbol{g}$, $\widehat{\boldsymbol{f}}$ with $\boldsymbol{f}$, $\widehat{\boldsymbol{A}}$ with $\boldsymbol{A}$, ...

As an example of using this simulation tool, we show here a complete set of figures detailing the different steps of simulation and inversion. Figure 1 shows an example of two sources $\boldsymbol{f}$ (generated via a mixture of two Gaussian model) and five data set $\boldsymbol{g}$ obtained via a mixing matrix $\boldsymbol{A}$ and addition of some noise $\boldsymbol{\epsilon}$ using the forward model $\boldsymbol{g} = \boldsymbol{A}\boldsymbol{f} + \boldsymbol{\epsilon}$ and then the results obtained by FA and ICA.

As a second example, we show in Figure 2 two sources generated via a mixture of two Gaussian model. We then again used these sources to generate the data and applied different methods of PCA, FA, ICA (without using the class information) and LDA with using the class information.

As a third example, we show in Figure 3 two sources generated via a mixture of two uniforms model. We then again used these sources to generate the data and applied different methods of PCA, FA, ICA (without using the class information) and LDA with using the class information.



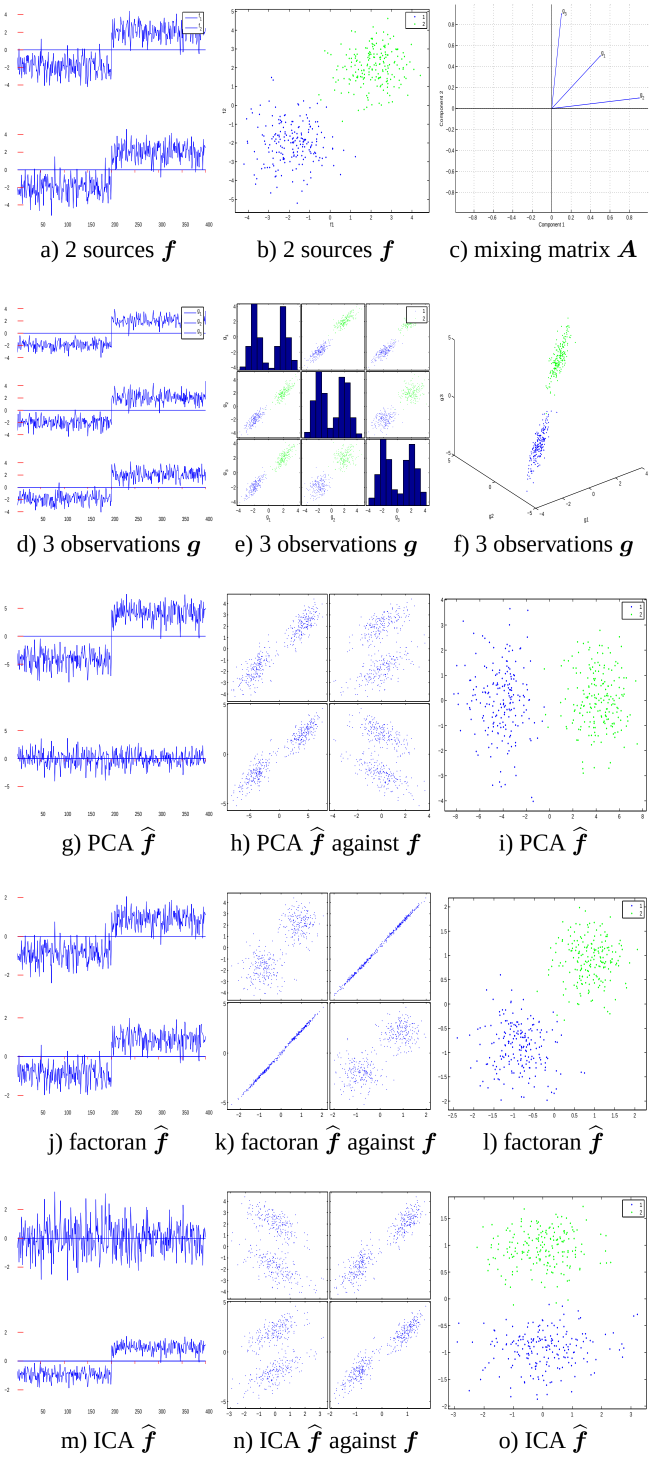
a) sources $\boldsymbol{f}$ b) observations $\boldsymbol{g}$ c) mixing matrix $\boldsymbol{A}$

d) scatter-plot of sources e) scatter-plot of observations f) mixing matrix presented in color

g) PCA factors $\widehat{\boldsymbol{f}}$ h) scatter plot of $\widehat{\boldsymbol{f}}$ against $\boldsymbol{f}$ i) $\widehat{\boldsymbol{A}}$

j) ICA factors $\widehat{\boldsymbol{f}}$ k) scatter plot of $\widehat{\boldsymbol{f}}$ against $\boldsymbol{f}$ l) $\widehat{\boldsymbol{A}}$

Fig. 1. Simulation of 2 sources $\boldsymbol{f}$ and 5 observations $\boldsymbol{g}$ with $T = 100$ samples: a) sources $\boldsymbol{f}$, b) observations $\boldsymbol{g}$, c) representation of the mixing matrix $\boldsymbol{A}$, d) scatter-plots of the sources, e) scatter-plots of the observations, f) color presentation of the mixing matrix, g) PCA factors $\widehat{\boldsymbol{f}}$, h) scatter-plot of $\widehat{\boldsymbol{f}}$ against $\boldsymbol{f}$, i) representation of the estimated mixing matrix $\widehat{\boldsymbol{A}}$, j) ICA factors $\widehat{\boldsymbol{f}}$, k) scatter-plot of $\widehat{\boldsymbol{f}}$ against $\boldsymbol{f}$, l) representation of the estimated mixing matrix $\widehat{\boldsymbol{A}}$.



a) 2 sources $\boldsymbol{f}$ b) 2 sources $\boldsymbol{f}$ c) mixing matrix $\boldsymbol{A}$

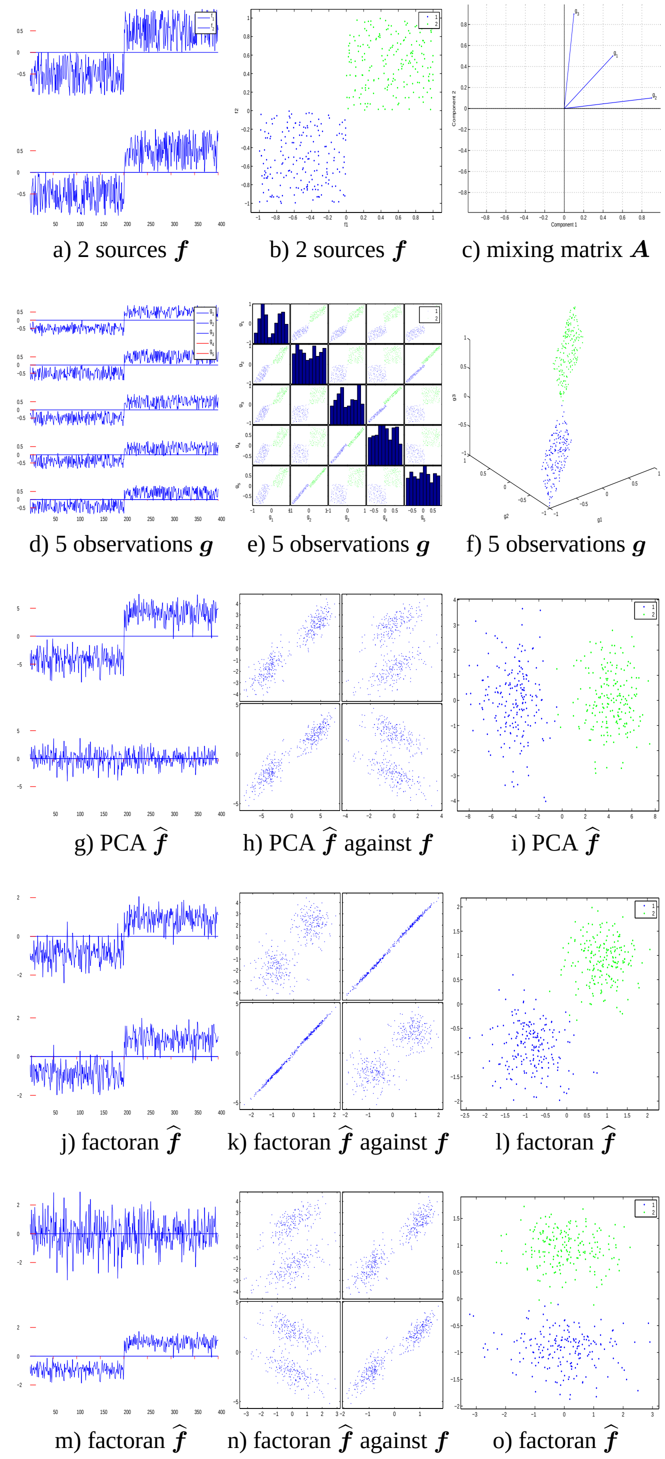
d) 3 observations $\boldsymbol{g}$ e) 3 observations $\boldsymbol{g}$ f) 3 observations $\boldsymbol{g}$

g) PCA $\widehat{\boldsymbol{f}}$ h) PCA $\widehat{\boldsymbol{f}}$ against $\boldsymbol{f}$ i) PCA $\widehat{\boldsymbol{f}}$

j) factoran $\widehat{\boldsymbol{f}}$ k) factoran $\widehat{\boldsymbol{f}}$ against $\boldsymbol{f}$ l) factoran $\widehat{\boldsymbol{f}}$

m) ICA $\widehat{\boldsymbol{f}}$ n) ICA $\widehat{\boldsymbol{f}}$ against $\boldsymbol{f}$ o) ICA $\widehat{\boldsymbol{f}}$

Fig. 2. Simulation of 2 sources (mixture of two Gaussian distributions) $\boldsymbol{f}$ and 3 observations $\boldsymbol{g}$ with $T = 400$ samples: a) sources $\boldsymbol{f}$, b) observations $\boldsymbol{g}$, c) representation of the mixing matrix $\boldsymbol{A}$, d) scatter-plots of the sources, e) scatter-plots of the observations, f) spatial structure of the 3 sources, g) PCA factors $\widehat{\boldsymbol{f}}$, h) scatter-plot of $\widehat{\boldsymbol{f}}$ against $\boldsymbol{f}$ and i) spatial structure of the PCA factors $\widehat{\boldsymbol{f}}$, j,k,l) the same with FA, m,n,o) the same with ICA.



a) 2 sources $\boldsymbol{f}$ b) 2 sources $\boldsymbol{f}$ c) mixing matrix $\boldsymbol{A}$

d) 5 observations $\boldsymbol{g}$ e) 5 observations $\boldsymbol{g}$ f) 5 observations $\boldsymbol{g}$

g) PCA $\widehat{\boldsymbol{f}}$ h) PCA $\widehat{\boldsymbol{f}}$ against $\boldsymbol{f}$ i) PCA $\widehat{\boldsymbol{f}}$

j) factoran $\widehat{\boldsymbol{f}}$ k) factoran $\widehat{\boldsymbol{f}}$ against $\boldsymbol{f}$ l) factoran $\widehat{\boldsymbol{f}}$

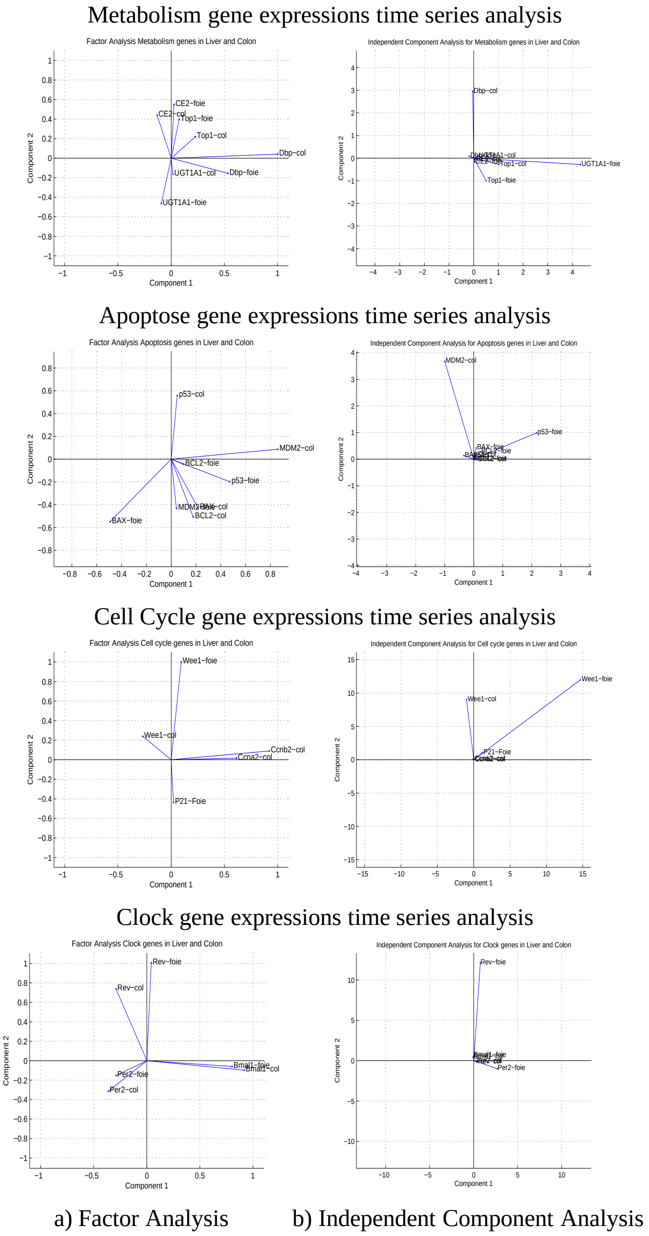
m) factoran $\widehat{\boldsymbol{f}}$ n) factoran $\widehat{\boldsymbol{f}}$ against $\boldsymbol{f}$ o) factoran $\widehat{\boldsymbol{f}}$

Fig. 3. Simulation of 2 sources (mixture of two uniform distributions) $\boldsymbol{f}$ and 3 observations $\boldsymbol{g}$ with $T = 400$ samples: a) sources $\boldsymbol{f}$, b) observations $\boldsymbol{g}$, c) representation of the mixing matrix $\boldsymbol{A}$, d) scatter-plots of the sources, e) scatter-plots of the observations, f) spatial structure of the 3 sources, g) PCA factors $\widehat{\boldsymbol{f}}$, h) scatter-plot of $\widehat{\boldsymbol{f}}$ against $\boldsymbol{f}$ and i) spatial structure of the PCA factors $\widehat{\boldsymbol{f}}$, j,k,l) the same with FA, m,n,o) the same with ICA.

4. APPLICATION ON REAL DATA

As we mentioned, we developed these tools for analyzing some biological data in relation with circadian cell cycle and evolution of cancer tumors in the context of the European project ERASYSBIO. A great number of experimentations have been done on mice. As an example, different quantities such as Temperature, Activity, different Hormones, different Genes expressions and different Proteins are measured during one or a few days and one of the problems addressed is finding the principal components or factors of some of these data.

In Figure 4, we show an example of such analysis on Gene expressions time series.


Metabolism gene expressions time series analysis

Apoptose gene expressions time series analysis

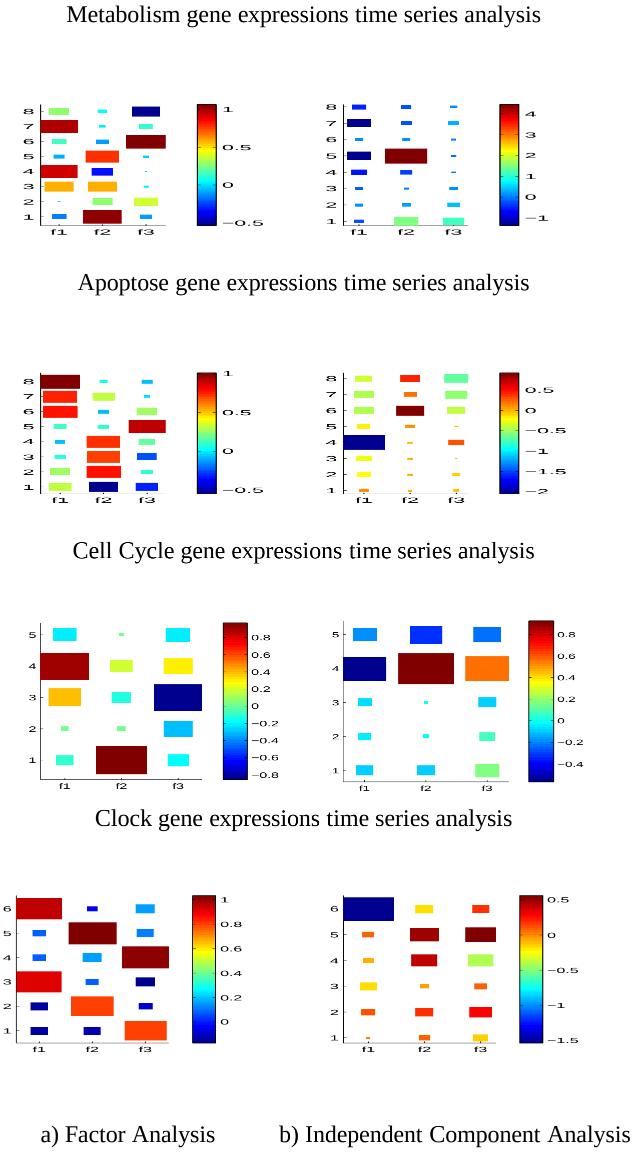
Cell Cycle gene expressions time series analysis

Clock gene expressions time series analysis

a) Factor Analysis b) Independent Component Analysis

Fig. 4. A comparison of FA and ICA on three sets of gene expression data. These results are obtained with two factors.

For now, we just applied these methods directly on the time series data without accounting for time structure which is very important. However, the results obtained seem to have some significant importance for biologists. Here, we assumed only two factors. As we can see it seems that there is a need to increase the number of factors.


Metabolism gene expressions time series analysis

Apoptose gene expressions time series analysis

Cell Cycle gene expressions time series analysis

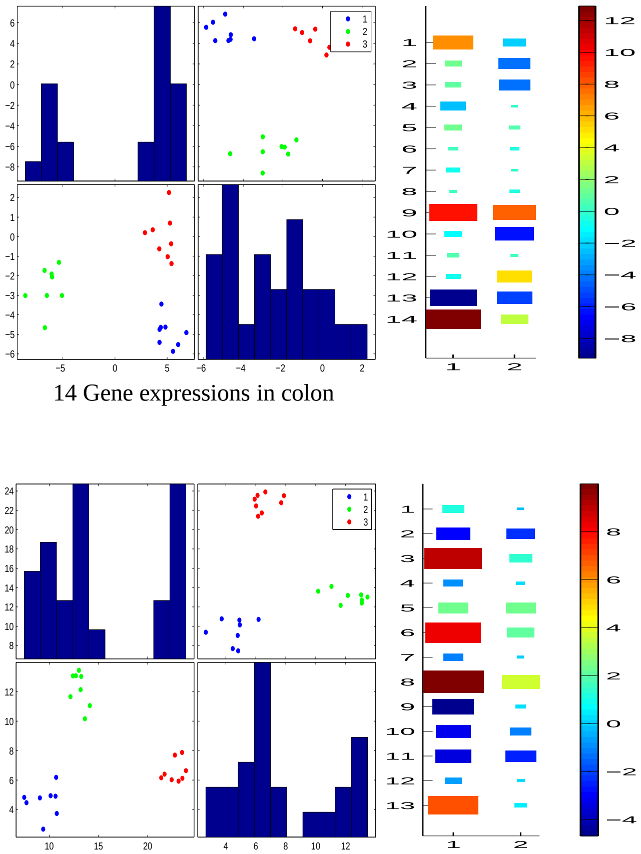
Clock gene expressions time series analysis

a) Factor Analysis b) Independent Component Analysis

Fig. 5. A comparison of FA and ICA on three sets of gene expression data. These results are obtained with three factors. Here, we used a different presentation of the loading matrix which is more appropriate for the cases where the number of factors are greater than two. This presentation is called Hinton where the values of the matrix are coded by color and by size of the patches.

In Figure 5, we show the same results with three factors. However, when the number of factors is greater than two, it is no more easy to represent them as bi-plot graphs of Figure 4. Here, we use a different presentation of the loading matrix which is more appropriate for the cases where the number of factors are greater than two. This presentation is called Hinton [16, 17] where the values of the matrix are coded by color and by size of the patches.

In Figure 6, we show two results of Linear Discriminant Analysis on 14 genes expressions in Colon and 13 genes expressions in liver. As we can see, here two factors are enough to discriminate the three classes of mice. On this figure, at left, we see this discrimination and at right the weights of these genes in these two factors.



14 Gene expressions in colon

13 Gene expressions in liver

Fig. 6. Discriminant Analysis on real mice data: 13 genes expressions in liver have been used. Two factors have been enough to discriminate the three classes of mice (left). The weights of these genes in these two factors are shown on the right.

The main difficulties in these data are: great dimensionality (more than fifty), non-homogeneity (Temperature, Activity, Hormones, Genes, Proteins), presence of outliers data, missing data and lack of synchronization (for example, temperature is measured every 15 minutes but Genes expressions every 3 hours). We need to adapt these methods to account for all these difficulties. We are working on these difficulties and will report soon in details on them.

5. CONCLUSIONS

In this paper, first we introduced a unifying presentation of many classical data analysis methods such as PCA, FA and ICA based on forward modeling and inversion. This unifying presentation facilitates the comprehension of these different methods. We then presented a simulation Matlab toolbox which has the possibilities of generating sources and observations, doing FA, PCA and ICA and evaluating the performances of the proposed methods. Finally, we used these tools for analyzing some biological data which seems giving important information, or at least confirm their intuition on the role of different quantities. We are still exploring these tools for the real application of biological data where we have to adapt more particularly these tools for the situations where:
- we have fewer number of data compared to the number of variables;
- the estimated covariance matrix of the data is not positive definite;
- the data are inhomogeneous;
- the data have different sampling rate;
- there are some non-observed values (missing data);
- there are outliers in the observed data (for example, measured temperature greater than 44 or less then 35, etc.).

6. PERSPECTIVES

When analyzing these biological data, the main questions we need to answer can be summarized as follows:

Variable section: One of the main questions asked very often is: If we had to redo other experiences, which ones of these quantities are the most importances to observe again. This is a very difficult question. The answer depends on the type of information we need to extract. Very often the quantities we have observed are linked (correlated or dependent). So, any selection of subset of variables causes, in some sense, loss of information. So, this question, very often, cannot be answered directly. We need modeling, the link between variables directly or in a transformed space, dimension reduction, clustering and classification, etc. Here are a few references concerning this subject [18, 19, 20, 21, 22]

Dimension reduction and Factor analysis: The second question is: Can we express the information content of all these data in a fewer set of factors or components? The main classical tools here are PCA, FA and ICA. One of the difficulties in these tools is the determination of the number of factors which is still an open problem [23, 9, 7]. When the number of factors is fixed, then these tools can be used easily. However, one of the drawbacks of these tools is the interpretation of the factors or components. Modeling the problem as an inverse problem of sources separation and using the the Bayesian approach are the promising tools to push farther these limitations [24, 25, 23, 26, 27, 28, 9, 29].

Discriminant Analysis:
Very often the observed data comes from different classes of subjects (male/female, healthy/Tumor,...) and we know the classes. In these cases, another question which arises is: Which of these variables or factors are the most discriminant between classes? Here are a few references concerning this subject [18, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39].

Clustering and classification:
Some times, in opposite of the previous case, we have only the data and we are asked to group or cluster them. This is also called *totally unsupervised classification*. In some other cases, we may know the number of classes and even the characteristics of each one of classes. The question is then to classify a given new observation. This is called *totally supervised classification*. When the number of classes is known, but the characteristics of each classe has to be *learned* from a *training set* of observations, then the problem is called *semi-supervised classification*. The estimation of number of classes is

related to *model selection*. Here are a few references concerning this subject [40, 33, 41, 8]

Graph of links and dependencies between variables:
One of the main steps of Knowledge extraction in studying biological data is producing a graph of dependencies between variables. To obtain such a graph we need to decide if two variables are dependent or not. We need then measures of dependencies to discover these dependencies [42, 43, 44, 45, 46, 47, 48]. One of the classical and most used is the Pearson's correlation ρ. When $|\rho|$ is near to one, we say that the two variables are dependent. However, when $|\rho|$ is near to zero or even zero, this does not mean that the two variables are independent. Indeed, $|\rho|$ measures only the linear dependence between those two variables. There are many other measures of dependencies that we can use which are more appropriate. For example, we use the Spearman's ρ_s and the Kendall τ jointly with Pearson's correlation ρ.

Graph of oriented dependencies between variables and causality: One of the last steps of Knowledge extraction in studying biological data is studying the oriented graph or causality [49, 50, 51]

7. REFERENCES

[1] J. Karhunen and J. Joutsensalo, "Representation and separation of signals using nonlinear PCA type learning," *Neural Networks*, vol. 7, no. 1, pp. 113–127, 1994.

[2] Pierre Comon, "Independent Component Analysis, a new concept ?," *Signal processing, Special issue on Higher-Order Statistics, Elsevier*, vol. 36 (3), pp. 287–314, Apr. 1994.

[3] D. J. C. MacKay, "Maximum likelihood and covariant algorithms for independent component analysis," Tech. Rep., University of Cambridge, Cavindish Laboratory, Cambridge, UK, 1996.

[4] K. Knuth, "Bayesian source separation and localization," in *SPIE'98 Proceedings: Bayesian Inference for Inverse Problems, San Diego, CA*, A. Mohammad-Djafari, Ed., July 1998, pp. 147–158.

[5] S. J. Roberts, "Independent component analysis: Source assessment, and separation, a Bayesian approach," *IEE Proceedings - Vision, Image, and Signal Processing*, vol. 145, no. 3, 1998.

[6] A. Hyvarinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, pp. 411–430, 2000.

[7] Ma Yi, P. Niyogi, G. Sapiro, and R. Vidal, "Dimensionality reduction via subspace and submanifold learning [from the guest editors]," *Signal Processing Magazine, IEEE*, vol. 28, no. 2, pp. 14 –126, march 2011.

[8] R. Vidal, "Subspace clustering," *Signal Processing Magazine, IEEE*, vol. 28, no. 2, pp. 52 –68, march 2011.

[9] K.M. Carter, R. Raich, W.G. Finn, and A.O. Hero, "Information-geometric dimensionality reduction," *Signal Processing Magazine, IEEE*, vol. 28, no. 2, pp. 89 –99, march 2011.

[10] L. Carin, R.G. Baraniuk, V. Cevher, D. Dunson, M.I. Jordan, G. Sapiro, and M.B. Wakin, "Learning low-dimensional signal models," *Signal Processing Magazine, IEEE*, vol. 28, no. 2, pp. 39 –51, march 2011.

[11] Y. Koren and L. Carmel, "Robust linear dimensionality reduction," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 10, no. 4, pp. 459 –470, 2004.

[12] A. Sharma and K.K. Paliwal, "Rotational linear discriminant analysis technique for dimensionality reduction," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 20, no. 10, pp. 1336 –1347, 2008.

[13] Jing Peng, Peng Zhang, and N. Riedel, "Discriminant learning analysis," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 38, no. 6, pp. 1614 –1625, 2008.

[14] P. Chaudhuri, A.K. Ghosh, and H. Oja, "Classification based on hybridization of parametric and nonparametric classifiers," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 7, pp. 1153 –1164, 2009.

[15] Taiping Zhang, Bin Fang, Yuan Yan Tang, Zhaowei Shang, and Bin Xu, "Generalized discriminant analysis: A matrix exponential approach," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 40, no. 1, pp. 186 –197, 2010.

[16] G.E. Hinton and R.R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504 – 507, July 2006.

[17] G.E. Hinton and R.R. Salakhutdinov, "Learning multiple layers of representation," *Trends in Cognitive Sciences*, vol. 11, pp. 428–434, 2007.

[18] J. Brezmes, P. Cabre, S. Rojo, E. Llobet, X. Vilanova, and X. Correig, "Discrimination between different samples of olive oil using variable selection techniques and modified fuzzy artmap neural networks," *Sensors Journal, IEEE*, vol. 5, no. 3, pp. 463 – 470, june 2005.

[19] C. Fevotte and S.J. Godsill, "Sparse linear regression in unions of bases via bayesian variable selection," *Signal Processing Letters, IEEE*, vol. 13, no. 7, pp. 441 –444, july 2006.

[20] T. Trappenberg, J. Ouyang, and A. Back, "Input variable selection: mutual information and linear mixing measures," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 18, no. 1, pp. 37 – 46, jan. 2006.

[21] J.-J. Fuchs and S. Maria, "A new approach to variable selection using the tls approach," *Signal Processing, IEEE Transactions on*, vol. 55, no. 1, pp. 10 –19, jan. 2007.

[22] Lu Chuan, A. Devos, J.A.K. Suykens, C. Arus, and S. Van Huffel, "Bagging linear sparse bayesian learning models for variable selection in cancer diagnosis," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 11, no. 3, pp. 338 –347, may 2007.

[23] Farahmand A., M., Szepesvári C., and Audibert J.-Y., "Manifold-Adaptive Dimension Estimation," Proceedings of the 24th International Conference on Machine Learning, 2007.

[24] A. Mohammad-Djafari, "Séparation de sources," in *Approche bayésienne en séparation de sources*, A. Mohammad-Djafari, Ed., Paris, 2006, Traité IC2, Série traitement du signal et de l'image, Hermès, (P. Common et Ch. Jutten ed.).

[25] D.P. Wipf and B.D. Rao, "An empirical bayesian strategy for solving the simultaneous sparse approximation problem," *Signal Processing, IEEE Transactions on*, vol. 55, no. 7, pp. 3704 –3716, july 2007.

[26] Erik G. Larsson and Yngve Selen, “Linear regression with a sparse parameter vector,” *Signal Processing, IEEE Transactions on*, vol. 55, no. 2, pp. 451 –460, feb. 2007.

[27] E. Diederichs, A. Juditsky, V. Spokoiny, and C. Schutte, “Sparse non-gaussian component analysis,” *Information Theory, IEEE Transactions on*, vol. 56, no. 6, pp. 3033 –3047, june 2010.

[28] Mohammad-Djafari Ali and Knuth K.H., “Bayesian approaches,” in *Handbook of Blind Source Separation*, Pierre Comon and Christian Jutten, Eds., Elsevier Ltd, 2010, Academic Press.

[29] J. Lapuyade and A. Mohammad-Djafari, “Nearest neighbors and correlation dimension for dimensionality estimation. application to factor analysis of real biological time series data.,” in *19th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, Michel Verleysen, Ed. 2011, ESANN 2011 Proceedings.

[30] Yijuan Lu, Qi Tian, M. Sanchez, J. Neary, Feng Liu, and Yufeng Wang, “Learning microarray gene expression data by hybrid discriminant analysis,” *Multimedia, IEEE*, vol. 14, no. 4, pp. 22 –31, oct.-dec. 2007.

[31] Jian Yang, A.F. Frangi, Jing-Yu Yang, David Zhang, and Zhong Jin, “Kpca plus lda: a complete kernel fisher discriminant framework for feature extraction and recognition,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 2, pp. 230 –244, feb. 2005.

[32] Taiping Zhang, Bin Fang, Yuan Yan Tang, Zhaowei Shang, and Bin Xu, “Generalized discriminant analysis: A matrix exponential approach,” *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 40, no. 1, pp. 186 –197, feb. 2010.

[33] Pi-Fuei Hsieh, Deng-Shiang Wang, and Chia-Wei Hsu, “A linear feature extraction for multiclass classification problems based on class mean and covariance discriminant information,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 2, pp. 223 –235, feb. 2006.

[34] Xuelian Yu, Xuegang Wang, and Benyong Liu, “A direct kernel uncorrelated discriminant analysis algorithm,” *Signal Processing Letters, IEEE*, vol. 14, no. 10, pp. 742 –745, oct. 2007.

[35] Juwei Lu, K.N. Plataniotis, and A.N. Venetsanopoulos, “Face recognition using kernel direct discriminant analysis algorithms,” *Neural Networks, IEEE Transactions on*, vol. 14, no. 1, pp. 117 – 126, jan 2003.

[36] T. Kurita, K. Watanabe, and N. Otsu, “Logistic discriminant analysis,” in *Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on*, oct. 2009, pp. 2167 –2172.

[37] Jian Yang and Chengjun Liu, “Horizontal and vertical 2dpca-based discriminant analysis for face verification on a large-scale database,” *Information Forensics and Security, IEEE Transactions on*, vol. 2, no. 4, pp. 781 –792, dec. 2007.

[38] Chein-I Chang and Baohong Ji, “Fisher’s linear spectral mixture analysis,” *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 44, no. 8, pp. 2292 –2304, aug. 2006.

[39] Changyou Chen, Junping Zhang, and R. Fleischer, “Distance approximating dimension reduction of riemannian manifolds,” *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 40, no. 1, pp. 208 –217, feb. 2010.

[40] Nadler B., Lafon S., Coifman R.R., and Kevrekidis I.G., “Diffusion maps, spectral clustering and eigenfunctions of Fokker-Planck operators,” *Advances in Neural Information Processing Systems*, vol. 18, pp. 955–962, 2005.

[41] Tian Lan and D. Erdogmus, “Local linear ica for mutual information estimation in feature selection,” in *Machine Learning for Signal Processing, 2005 IEEE Workshop on*, sept. 2005, pp. 3 –8.

[42] G. Qu, S. Hariri, and M. Yousif, “A new dependency and correlation analysis for features,” *Knowledge and Data Engineering, IEEE Transactions on*, vol. 17, no. 9, pp. 1199 – 1207, sept. 2005.

[43] F. Chin and H.C. Leung, “Dna motif representation with nucleotide dependency,” *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, vol. 5, no. 1, pp. 110 – 119, jan.-march 2008.

[44] Deng Cai, Xiaofei He, and Jiawei Han, “Srda: An efficient algorithm for large-scale discriminant analysis,” *Knowledge and Data Engineering, IEEE Transactions on*, vol. 20, no. 1, pp. 1 –12, jan. 2008.

[45] P.C.H. Ma and K.C.C. Chan, “Inferring gene regulatory networks from expression data by discovering fuzzy dependency relationships,” *Fuzzy Systems, IEEE Transactions on*, vol. 16, no. 2, pp. 455 –465, april 2008.

[46] S. Zafeiriou and I. Pitas, “Discriminant graph structures for facial expression recognition,” *Multimedia, IEEE Transactions on*, vol. 10, no. 8, pp. 1528 –1540, dec. 2008.

[47] L. Yu, A. Mishra, and S. Ramaswamy, “Component co-evolution and component dependency: speculations and verifications,” *Software, IET*, vol. 4, no. 4, pp. 252 –267, august 2010.

[48] Fan Wenfei, F. Geerts, Jianzhong Li, and Ming Xiong, “Discovering conditional functional dependencies,” *Knowledge and Data Engineering, IEEE Transactions on*, vol. 23, no. 5, pp. 683 –698, may 2011.

[49] D.A. Bell, “From data properties to evidence,” *Knowledge and Data Engineering, IEEE Transactions on*, vol. 5, no. 6, pp. 965 –969, dec 1993.

[50] M.L. Raymer, T.E. Doom, L.A. Kuhn, and W.F. Punch, “Knowledge discovery in medical and biological datasets using a hybrid bayes classifier/evolutionary algorithm,” *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 33, no. 5, pp. 802 – 813, oct. 2003.

[51] Ghim-Eng Yap, Ah-Hwee Tan, and Hwee-Hwa Pang, “Discovering and exploiting causal dependencies for robust mobile context-aware recommenders,” *Knowledge and Data Engineering, IEEE Transactions on*, vol. 19, no. 7, pp. 977 –992, july 2007.