A Novel Dimensionality Reduction Technique Based on Independent Component Analysis for Modeling Microarray Gene Expression Data

Han Liu
Department of Computer Science
University of Toronto
Toronto, ON M5S 3G4

Rafal Kustra
Department of Biostatistics
University of Toronto
Toronto, ON M5S 1A8

Ji Zhang
Department of Computer Science
University of Toronto
Toronto,ON M5S 3G4

Abstract

DNA microarray experiments generating thousands of gene expression measurements, are being used to gather information from tissue and cell samples regarding gene expression differences that will be useful in diagnosing disease. But one challenge of microarray studies is the fact that the number n of samples collected is relatively small compared to the number p of genes per sample which are usually in thousands. In statistical terms this very large number of predictors compared to a small number of samples or observations makes the classification problem difficult. This is known as the "curse of dimensionality problem". An efficient way to solve this problem is by using dimensionality reduction techniques. Principle Component Analysis(PCA) is a leading method for dimensionality reduction of gene expression data which is optimal in the sense of least square error. In this paper we propose a new dimensionality reduction technique for specific bioinformatics applications based on Independent component Analysis(ICA). Being able to exploit higher order statistics to identify a linear model result, this ICA based dimensionality reduction technique outperforms PCA from both statistical and biological significance aspects. We present experiments on NCI 60 dataset to show this result.

Keywords-*gene expression data, dimensionality reduction, independent component analysis, latent regulatory factors*

1. Introduction

In the specific area of computational biology, *ie.* tumor classification, analysis of high dimensional datasets is frequently encountered. For example, DNA microarray experiments generating thousands of gene expression measurements, are being used to gather information from tissue and cell samples regarding gene expression differences that will be useful in diagnosing disease[1][2].

This high dimension presents a great challenge for modeling and analysis of the data. Mathematically, when viewing the modeling problem in a regression framework. Some specific applications can be modelled as follows: the response variable (e.g. the prostrate cancer cell line) is expressed by predictor or explanatory variables (gene expression measurements) by a multiple linear regression model

$$y_i = \beta_0 + x_{i1}\beta_1 + ... + x_{ip}\beta_p + \varepsilon_i, i = 1, ..., n. \quad (1)$$

n is the number of observations (*ie.* cell lines), $x_i = (1, x_{i1}, ..., x_{ip})^\top$ are collected as rows in a matrix $\mathbf{X}$ containing the predictor variables, $y = (y_1, ..., y_n)^\top$ is the response variable, $\beta = (\beta_0, \beta_1, ..., \beta_p)^\top$ are the regression coefficients which are to be estimated, and $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^\top$ is the error term. The differences $y_i - \beta_0 - x_{i1}\beta_1 - ... - x_{ip}\beta_p$ express the deviation of the fit to the observed values and are called residuals. Traditionally, the regression coefficients are estimated by minimizing the sum of squared residuals $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - ... - \beta_q x_{i1})^2 = (y - \mathbf{X}\beta)^\top(y - \mathbf{X}\beta)$. This criterion is called *least squares* (LS) criterion[3], and the coefficient minimizing the criterion turns out to be

$$\widehat{\beta}_{LS} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top y. \quad (2)$$

Since the inverse of $\mathbf{X}^\top\mathbf{X}$ is needed in Equation (2), problems will occur if the rank of $\mathbf{X}$ is lower than $p + 1$. This happens if the predictor variables are highly correlated or if there are linear relationships among the variables. This situation is called *multicollinearity*[4], and often a generalized inverse is then taken for estimating the regression coefficients. The inverse of $\mathbf{X}^\top\mathbf{X}$ also appears when computing the standard errors and the correlation matrix of the regression coefficients estimator $\widehat{\beta}_{LS}$. In a near-singular case the standard errors can be inflated considerably and cause doubt on the interpretability of these coefficients. Also note that the rank of $\mathbf{X}$ is always lower than $p + 1$ if the number of observations is less than or equal to the number of variables ($n \leq p$). This is a frequent problem which occurs in many applications e.g. one feature of microarray studies is the fact that the number n of samples collected is relatively small compared to the number p of genes per sample which are usually in the thousands. In statistical terms this very

large number of predictors or variables (genes) compared to a small number of samples or observations (microarrays) makes most of classical "class prediction" methods unemployable, unless a preliminary variable selection step is performed[7].

The idea is to construct a limited set of *k* components $z_1, ..., z_k$ which are linear combinations of the original variables. So there are existing vectors b_j such that $z_j = \mathbf{X}b_j$ for $1 \leq j \leq k$. Let $\mathbf{Z} = (z_1, ..., z_k)$ be the $n \times k$ matrix having the components in its columns. For ease of notation, we ask these components to be centered, so $\mathbf{1}_n^\top \mathbf{Z}$. with 1_n a column vector with all n components equal to 1. Moreover, the well known method PCA(or Karhunen-Loeve expansion in pattern recognition[5]) also ask these components to be uncorrelated and to have unit variance:

$$\mathbf{Z}^\top \mathbf{Z} = \frac{1}{n-1}\mathbf{I}_k, \tag{3}$$

where I_k stands for an identity matrix of rank k. These components will then serve as now predictor variables in the regression model. Not that, due to (3) the multicollinearity problem has completely vanished when using a regression model with $z_1, ..., z_k$ as predictor variables. Moreover, when k is small relative to p, one has significantly reduced the number of predictor variables, leading to a more parsimonious regression model. PCA based dimensionality reduction method is up to second order statistics(covariance, correlation), However, higher order statistics contain significant complementary information. This is the case in particular when the distribution of data differs significantly from gaussian, which turns out to happen quite often in microarray expression data. Indeed, some particular genes may happen to be significantly over-expressed(under-expressed) in some conditions, which yields "heavy tail" distribution[6]. Therefore, we propose an ICA based dimensionality reduction method to try to exploit such higher order statistics for the analysis of expression data. Our approach models logarithms of expression profile of a specific as linear combination of "latent" regulatory factors which are statistically independent. Our method could not only provide useful information in term of discrimination or clustering of conditions, this ICA based method also provides a useful mathematical framework for processing and modeling genome-wide expression data, in which both the mathematical variables and operations could be assigned biological meaning and could be explained easily.

The next section of this paper will introduce mathematical framework of ICA based dimensionality reduction method for gene expression profile. In section 3, a detailed analysis and theoretical comparison with PCA based method will be given out. Experiment and result analysis are given out in section 4, conclusions and possible extensions are summarized in section 5.

2. Mathematical Framework– ICA

2.1 General Framework

The relative expression levels of p genes of a model organism, which may constitute almost the entire genome of this organism, in a single sample, are probed simultaneously by a single microarray. A series of N arrays, which are almost identical physically, probe the genome-wide expression levels in N different samples. Let the $p \times N$ matrix $\mathbf{X}$ denote the full expression profile,every $x_{ij} = \log_2(R_{ij}/G_{ij})$ represents the log ratio of red(experiment) and green(reference) intensities . p represents p-genes while N represents N arrays. Each element x_{ij} for all $1 \leq i \leq p$ and $1 \leq j \leq N$ denotes the relative expression level of the ith gene in the jth sample as measured by the jth array. The vector in the ith row of the matrix $\mathbf{X}$ lists the relative expression of the ith gene across the different samples which correspond to the different arrays; while the vector in the jth column of the matrix $\mathbf{X}$ lists the genome-wide relative expression measured by the jth array.

By viewing the expression pattern of each gene across different arrays as a random variable,we model the transcription level of all gene expressions in a cell as a mixture of latent regulatory factors which are statistically independent. Mathematically, suppose that a specific gene is governed by k independent latent factors. $S = (s_1, ..., s_k)^T$, Each of which can be viewed as a regulatory factor. ICA is then a generative model which can be viewed as a linear transformation of the expression data from the p-genes $\times$ N-array space to the reduced k-"regulatory factor" $\times$ N-array space, where $k \leq \min\{p, N\}$. By defining a model whereby the expression profile of each different gene x_i can be expressed as linear combinations of the k latent regulatory factors: $x_i = a_{i1}s_1 + a_{i2}s_2 + ... + a_{ik}s_k$. We can express this model consistently in the generative form of ICA.

$$X = AS, \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} = \begin{pmatrix} a_{11} & ... & a_{1k} \\ a_{21} & ... & a_{2k} \\ \vdots & \vdots & \vdots \\ a_{p1} & ... & a_{pk} \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_k \end{pmatrix} \tag{4}$$

Equation (4) corresponds to a generative model of interactions between latent regulatory factors. The linear transformation matrix A can be viewed as a loading matrix of the regulatory factors for each gene.

Since these latent regulatory factors are assumed to be statistically independent, each of the vector $s_1, ..., s_k$ can be viewed as an independent random source. Then, ICA can be applied to find a matrix W that provides the transformation $Y = (y_1, ..., y_k)^T = WX$ of the observed matrix $\mathbf{X}$ under which the transformed random variables $y_1, ..., y_k$ called the estimated independent components, are as inde-

pendent as possible[8]. Under certain mathematical conditions(will be discussed later), the estimated random variables $y_1, ..., y_k$ are close approximations of $s_1, ..., s_k$ up to permutation and scaling. Denote this in matrix notation(Equation 5)

$$Y_{k\times N} = W_{k\times p} X_{p\times N} \quad (5)$$

From equation (5), we see that the data are mapped from a $p \times N$ space to a reduced $k \times N$ space, when $k << p$, the dimension was reduced greatly. In the new space, the data are represented the matrix **Y**, the k rows of this matrix can be represented as k independent latent regulatory factors. By setting the number k, the dimension can be reduced from p to k, how to select such a k is important, we will give detail discussion about this issue.

2.2 Methodology

Under the general mathematical framework, given a $p \times N$ matrix which is a microarray of p genes under N arrays, the following procedures will be performed:

Step1-Data Preprocessing: The preprocessing of data is a standard but necessary procedure for microarray expression data modeling. The first step is to apply logarithmic corrections to the data, the main reason for this is that some effects under study are likely to have a multiplicative behavior, which becomes linear after being log transformed. Under our framework, we get every x_{ij} in the matrix as $x_{ij} = \log_2(R_{ij}/G_{ij})$, where R_{ij} and G_{ij} represent red(experiment) and green(reference) intensities respectively. Another important preprocessing step is treating the missing data, such as in NCI 60 cancer cell line dataset, the mean percentage of missing data points per array is 6.6%. There're different approaches proposed for imputation the missing data, the reader is refereed to the work of Troyanskaya[9], our method is described in the excrement part.

Step2-Gene Standardization: The gene expression data were standardized so that the observations(genes) have mean 0 and variance 1 across different arrays. Standardizing the data in this fashion achieves a location and scale normalization of the different gene expressions. This kind of scale adjustment is desirable in some cases to prevent the expression levels for one particular gene from dominating the average expression levels across different genes. In fact, this standardization process is just the so called "centering" and "whitening" processes, which are two very useful preprocessing steps for applying ICA estimation. By whitening(or"sphereing"), the unmixing matrix **W** should be an orthogonal one, thus reduce the parameters to be estimated greatly. Our method is based on *eigenvalue decomposition(EVD)* as shown in [8].

Step3-ICA Based Dimensionality Reduction: We denote **X** the corrected logarithms of expression profile after standardization, and start from a model of the form $X = AS$, where the S are independent sources and A is the mixing matrix. ICA algorithm will estimate out an unmixing matrix W such that $Y = WX$ and makes Y as approximate S as possible. The ICA algorithm we adopt is called **FastICA** which was developed by Hyvarinen and Oja[10]. For a linear transformation $Y_{k\times N} = W_{k\times p} X_{p\times N}$, which search the corresponding W by minimizing the mutual information as follows:

$$I(y_1, ..., y_k) = \sum_{i=1}^{k} H(y_i) - H(X) + \log|\det(W)| \quad (6)$$

where $H(y)$ represents the entropy for random variable y with density $f(y)$ and defined as $H(y) = -\int f(y)\log f(y)dy$. After this step, the data have already been mapped into a new feature space and when $k << p$, the dimension is reduced.

Step4-Interpretation of ICA results:As a result, the ICA method yields latent regulatory factors which are statistically independent. There are mainly two aspects we are of great interest: Given the model $X_{p\times N} = A_{p\times k} S_{k\times N}$, where X is expression profile and S is the independent sources. The mixing matrix A is of great interest to analysis. For a specific gene, one of the elements a_{ij}(where $1 \le i \le p$ and $1 \le j \le k$) represents the effect of the jth regulatory factor on the ith gene under N different conditions(arrays). If the generative model does hold, based on this information, we can predicate to which extent a specific latent regulatory factor regulates the expression level of a gene under different conditions or whether this factor is (positive or negative)"active" under the conditions. The other aspect is when fixing a specific regulatory factor, the distribution of the elements of matrix A could be a good indication for analyzing the behavior of specific genes in different regulatory factors. Given a threshold, the distribution of gene expression profile in a given regulatory factor generally features a small number of significantly over-expressed or under-expressed genes, which kind of "dominate" this regulator factor.

3 Discussions and Related Work

In this section, we will focus on some specific discussions and compare this ICA method with the PCA based method:

ICA vs. PCA: Using PCA(or SVD decomposition) in microarray analysis was first introduced in [11]. They decomposed a matrix X of p genes $\times$ N experiments into the product $X^T = UDV^T$ of a $N \times L$ orthogonal matrix U, a diagonal matrix D, and a $p \times L$ orthogonal matrix V, where L=rank(X). The columns of U are called eigengenes, and

the columns of V are called eigenarrays. Both eigengenes and eigenarrays are uncorrelated. In [11] they assume that each eigengene represents a transcriptional regulator and the corresponding eigenarray represents the expression pattern in samples where the regulator is overactive or underactive. By written their equation as $V^T X = UD$ and get out the fist k columns of V, we could get $(V^T)_{k\times p} X_{p\times N} = U_{k\times L} D_{L\times N}$. Thus, reduce the dimension.

Not like PCA, ICA models the $p \times N$ matrix X as a generative model $X = AS$(Equation 4), where S is a $k \times N$ matrix. whose rows are statistically independent regulatory factors. The main difference between ICA and PCA is that PCA only finds the k uncorrelated regulatory factors, while ICA could finds k independent regulatory factors. Uncorrelated is only partially independent. These two mathematical conditions are equivalent only for Gaussian random variables. But most microarray data are non-gaussian[1], Based on Central Limited Theorem, we can conclude that the distributions of regulatory factors are also non-gaussin. We hypothesized that different latent regulatory factors are highly statistically independent, and therefore should be best separated by ICA. Our experiment based on real-world dataset also illustrate this.

Comparison with Related Work: Some other researchers also applied ICA for microarray analysis. Liebermeister[6] and Chiappetta[12] first proposed using linear ICA for microarray analysis to extract expression modes, where each mode represents a linear influence of a hidden cellular variable. Su-In Lee[13]gave out a systematic analysis of the applicability of ICA as an analysis tool in diverse datasets. Given a $p \times N$ microarray expression profile matrix X, not like our method, instead using the model $X = AS$, they assume a generative model $(X^T)_{N\times p} = A_{N\times k} S_{k\times p}$, By this way, they view the expression $X = (x_1, ..., x_N)$ as a post-linear mixture of the underlying independent biological processes. Based on the assumption that the independent source vector $S = (s_1, ..., s_k)$ are k independent biological processes which are expressed by the whole p-gene wide expression profile, their method could be used for dimensionality reduction. Thus essentially not the same with ours.

4. Experiment and Analysis

To evaluate the performance of our method, the ICA based method has been applied to real world dataset,we now discuss results obtained with NCI 60 dataset.

4.1 Dataset: NCI 60

In this study, cDNA were used to examine the variation in gene expression among the 60 cell lines from the National Cancer Institute's anticancer drug screen known as NCI 60 daataset[14].The 60 cell lines are derived from tumors with different sites of origin: 7 breast,6 central nervous system(CNS), 7colon, 6 leukemia, 8 melanoma, 9 nonsmall-cell lung carcinoma(NSCLC), 6 ovarian, 2 prostate, 8 renal, and 1 unknown(ADR-RES). Gene expression was studied using microarrays with 9,703 spotted cDNA sequences. In each hybridization, fluorescent cDNA targets were prepared from a cell line mRNA sample(fluorescent dye Cy5) and a reference mRNA sample obtained by pooling equal mixtures of mRNA from 12 of the cell lines(fluorescent dye Cy3). To investigate the reproducibility of the entire experimental procedure(e.g., cell culture, mRNA isolation, labeling, hybridization, scanning), a leukemia(K562) cell line and a breast cancer(MCF7) cell line were analyzed by three independent experiments. For our experiment, we make classification for eight classes(the two prostate cell line observations were excluded out because of their small class size). After screening out genes with missing data points, the data are collected into a $3,894 \times 57$ matrix $X = (x_{ij})$, where x_{ij} denotes the logarithmic of the Cy5/Cy3 fluorescence ration for gene i in mRNA sample j. Also, the standardization of the data have been performed as described above.

4.2 Experimental Result and Analysis

The Distributions of Gene Expression Profile Random variable: To test the distributions of the gene expression profile random variables, we randomly get some random variables from the rows of $p \times N$ matrix X to draw their QQ plot as shown in figure 1: From figure one, the three

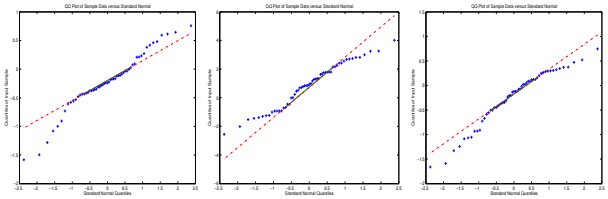

Figure 1: QQ plot of gene expression random variables

subfigures show that some distributions of gene are belong to the "heavy tail" family, "light tail" family and "skewed left" family. From which we could see that the distributions of the gene expression profile random variables are typically non-gaussian, thus,based on central limited theorem, we can get the conclusion that the distributions of the independent also be non-gaussian.

Analysis of the Unmixing Matrix W: We applied ICA to reduce the dimensionality of the matrix X from $3,894$ to 5 ,8 and 12 independent regulator factor components respectively.Given this $p \times N$ matrix X. When we assume there are 5 independent sources,by the ICA generative model $Y = WX$, we draw a picture of the distribution of

w_i ($1 \leq i \leq 5$) one of the rows in the unmixing matrix W as shown in figure 2.

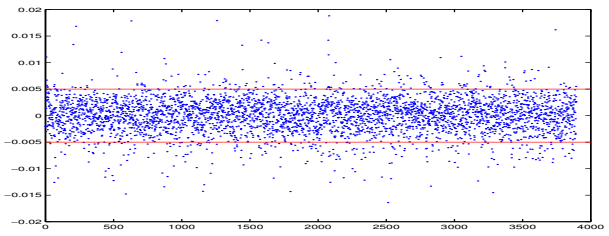

Figure 2: distributions of one row of unmixing matrix W

Because $y_i = w_i^T X$, every estimated regulatory factor y_i is a linear combination of the observed data X, where w_i represents the contribution of every genes to the regulatory factors. From figure 2, we can see that only a coherent group of genes "governing" an independent regulatory source. Between the two lines which includes 95% genes have contributes to the independent source less than 0.05. while only 5% genes have relative larger domination.

Comparison with PCA for multiclasses classification:For multiclass classification problem, we use two classifiers: logistic regression(represents parametric method) and k-nearest neighbor(represents nonparametric method). We use 2/3 of the original data as training cases and the other 1/3 as testing data, because there're altogether 8 classes and only about 4 training cases for one classes, the classification error is very high, the box plot of these two classifiers based on dimensionality reduction on ICA and PCA are given in figure 3 respectively:

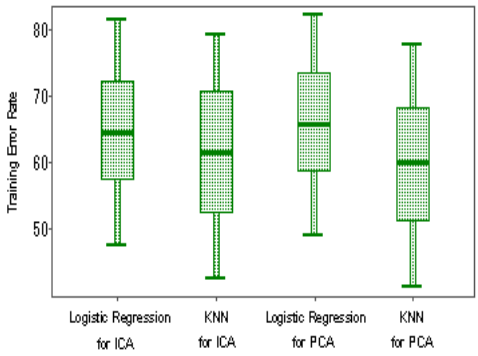

Figure 3: comparision with PCA and ICA for logistic regression and kNN classifier

From figure 3, we can see that the performance of ICA based method is better than PCA based method for logistic regression, but a little weaker for k-NN classifier. From the whole point, ICA is a promising method for dimensionality reduction.

5. Conclusion and Future Work

We have proposed an ICA based dimensionality reduction method in this paper, which could be viewed as an extension for PCA based method. Our method could be used to find latent "regulatory factors" which are statistically independent between each other, by utilizing our method on the real world dataset, we show that ICA based dimensionality reduction method is promising.

Because our current ICA model is a linear generative model which is based on the assumption that the interactions between different regulatory factors are linear, in fact, some of these processes could be unlinear. How to develop an interesting nonlinear ICA model maybe an interesting issue to give some further investigation, thus becomes our future work.

Acknowledgments

The authors would like to thank the referees for reviewing this paper, Han is supported by the graduate fellowship from Department of Computer Science of University of Toronto.

References

[1] Schena, M., Shalon,D. Davis, R.W. and Brown,P.O. *Quantitative monitoring of gene expression patterns with a complementary DNA microarray*, Science,vol 270, pp 467-470. 1995

[2] Lockhart,D., Dong,H.,etc.*Expression monitoring by hydrization to high density of oligonucleotide arrays.*, Nat.Biotechnol, 14, pp.1675-1680,1996

[3] Trevor Hastie, Robert Tibshirani, Jerome Friedman *The elements of statistical learning- Data ming, inference and prediction*, Springer 2001.

[4] Althauser, R. P.*Multicollinearity and non-additive regression models.*,In H. Blaock(ed) Causal models in the social sciences pp. 453-472. 1971

[5] Mallat,S.G. *A wavelet tour of signal processing*Academic, San Diego,2nd 1999

[6] Liebermeister,W. *linear modes of gene expression determined by independent component analysis*, Bioinformatics 18. pp.51-60,2002

[7] V.S.Cherkassky, I.F.Mulier *learning from data*,chapter 5. John Wiley & sons,1998

[8] Hyvarinen A. *a survey of indepenent component analysis*, Neural Computing Surveys (2), pp.94-128 1999

[9] Troyanskaya, O., Cantor,M. *missing value estimation methods for DNA microarrays*, Bioinformatics,17, pp.520-525,2001

[10] Hyvarinen, A. and Oja,E. *Indepdent component analysi: algorithm and applicatins*, Neural networks 13, pp.411-430 2000

[11] Alter O, Brown PO, Botstein D. *singular value decomposition for genome-wide expression data processing and modeling*, Pro Natl Acad Sci USA 97, pp10101-10106 2000

[12] Chippetta,P., Roubaud, M. and Torresani,B. *blind source seperation and the analysis of microarray data*, Proc of JOBIM'02 pp131-136, St Malo 2002

[13] Su-In Lee, Serafim Batzoglou *Application of independent compoenent analysis to microarrays*, Genome Biology Vol4, R76, 2003

[14] Ross,D.T., Scherf,U.,etc *Systematic variation in gene expression patterns in human cancer cell lines*, Nature Genetics, 24,pp227-234 2000