Gene selection in classification problems via projections onto a latent space

Marilena Pillati and Cinzia Viroli

Statistics Department,
University of Bologna, Italy
pillati@stat.unibo.it, viroli@stat.unibo.it

Abstract. The analysis of gene expression data involves the observation of a very large number of variables (genes) on a few units (tissues). In such a context the recourse to conventional classification methods may be hard both for analytical and interpretative reasons. In this work a gene selection procedure for classification problems is addressed. The dimensionality reduction is based on the projections of genes along suitable directions obtained by Independent Factor Analysis (IFA). The performances of the proposed procedure are evaluated in the context of both supervised and unsupervised classification problems for different real data sets.

1 Introduction

The numerous statistical questions posed by the analysis of gene expression measurements, as currently determined by the microarray technology, include both the problem of distinguishing between cancer classes and the problem of identifying and discovering various subclasses of cancer. These are two distinct classification problems, supervised and unsupervised respectively, that can be addressed by discriminant analysis and clustering techniques.

The peculiarity of gene expression data is the very large number of variables (genes) with respect to the number of units (tissues or cells). A reduction in dimensionality is needed not only to allow the employment of standard statistical methods but also for a biological interpretation. As many genes result to be not relevant to the tumor classification, a natural choice may consist in performing a variable selection to avoid the inclusion of not relevant or noisy genes. Their presence may assert a negative influence on the overall performances of an estimated classification rule or hide some meaningful patterns or structures in the data. There is a vast literature on gene selection for cell classification; a comparative study of several discrimination methods based on filtered sets of genes can be found in Dudoit *et al.* (2002).

Following Caló *et al.* (2005), we propose an unsupervised multivariate strategy that allows to take into account gene interactions, and that may be employed in order to solve both discriminant and clustering problems. We move from a well known result within the biological community: only few genes have distinct levels of activity between conditions of interest (such as cancer and non cancer or different types of disorders). Most of the genes

demonstrate a "regular" expression profile and so they are not relevant to class prediction or to recovering subclasses. Therefore, we think that a reasonable criterion for dimension reduction in this context could consist in detecting and selecting the genes showing a behavior across the cells that most differs from that of all the other genes. We start from regarding genes as points in a n-dimensional space. Then, the genes are projected onto a lower dimensional space. In order to highlight the genes showing the greatest expression variability with respect to the tissues, the directions of the projections should exhibit non-gaussian gene expression profiles. We propose to use Independent Factor Analysis (IFA) to identify the latent space. Finally, a rank of the genes in the independent factor space is derived to detect subsets of relevant genes. Section 2 provides a brief introduction to independent factor analysis. Section 3 describes the proposed gene selection procedure. Finally, some applications to real data sets are illustrated.

2 Independent Factor Analysis (IFA)

The historical background of IFA can be found in the signal processing context, as a solution to the so called "blind source separation" problem (Attias, 1999). It identifies a situation in which a number of signals emitted by some physical sources are observed: the objective is to recover the unobserved sources from their signal mixtures. Despite its origin, IFA can be reinterpreted as a particular latent variable model with independent and non gaussian factors. In fact, the p observed variables x_j are modelled in terms of a smaller set of k unobserved independent latent variables y_i and an additive specific term u_j:

$$x_j = \sum_{i=1}^{k} \lambda_{ji} y_i + u_j, \tag{1}$$

where $j = 1, ..., p$, $i = 1, ..., k$. In compact form the IFA model is $\mathbf{x} = \boldsymbol{\Lambda}\mathbf{y} + \mathbf{u}$, where the random vector $\mathbf{u}$ represents the noise, assumed to be normally distributed, $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$ and the factor loading matrix $\boldsymbol{\Lambda} = \{\lambda_{ji}\}$ is also termed as *mixing matrix*. The density of the i^{th} factor is modelled by a mixture of n_i gaussians with mean μ_{i,q_i}, variance ν_{i,q_i} and mixing proportions w_{i,q_i} $(q_i = 1, ..., n_i)$. The parameter estimation problem may be quite promisingly solved by the EM algorithm. The identification of the latent structure dimensionality can be achieved according to the so called information criteria, which are based on penalized forms of the likelihood.

3 Gene selection in the independent factor space

In order to perform gene selection by IFA, we start regarding genes as points in the n-dimensional space of the tissues (in doing this the role of units and

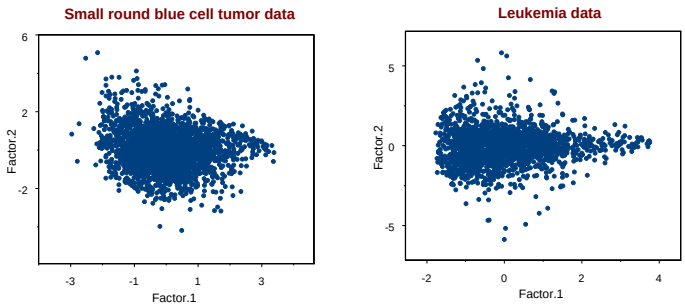

Fig. 1. Latent spaces detected by IFA (*left graph*: first two estimated factors out of four for the SRBCT data; *right graph*: latent space of the LK data).

that of variables are exchanged). When considering genes as units, the distribution of gene expression levels in the cell space must be taken into account. Empirical evidence shows that these distributions are typically leptokurtic, with heavy tails and a pronounced central peak. This implies that the observed variables (in this new perspective, the expression profiles across cells) have non-Gaussian distributions and hence the variance-covariance matrix does not suffice to describe the relations between them, but it is necessary to consider also higher-order moments. This is the reason why we look for a subspace in which the projections of the genes along the latent directions exhibit non-gaussian expression profiles. In this perspective Caló *et al.* (2005) have suggested to use independent component analysis (Comon, 1994), but have left the problem of selecting the correct number of independent components, *i.e.* the latent space dimension, still unsolved.

Following our proposal, the p n-dimensional gene expression data are projected by IFA on a k dimensional space, $k << p$, where k is chosen on the basis of some information criteria (AIC, BIC). The most "irregular" genes are then detected by looking at the tails of the distributions of gene projections along the IFA directions. In fact, highly induced or repressed genes should lie on the tails of the distributions of the y_i $(i = 1, ..., k)$.

Genes are ranked in the reduced space according to their maximum absolute scores across the factors, after rescaling the factor scores so that they have the same range. The genes located in the last m positions of this ranking (with $m << p$) should be used to class prediction or to discover subclasses. Ranking the genes according to the above criterion is equivalent to order the genes on the basis of their distance from the mean vector in the latent space in terms of the Minkowski metric with parameter $l \to \infty$. Therefore, as alternatives, we consider other distance measures, such as those obtained from the Minkowski metric for $l = 1$ (Manhattan distance), $l = 2$ (Euclidean distance) and $l = 3$ (cubic distance), in order to evaluate the sensitivity of the procedure to the choice of the metric.

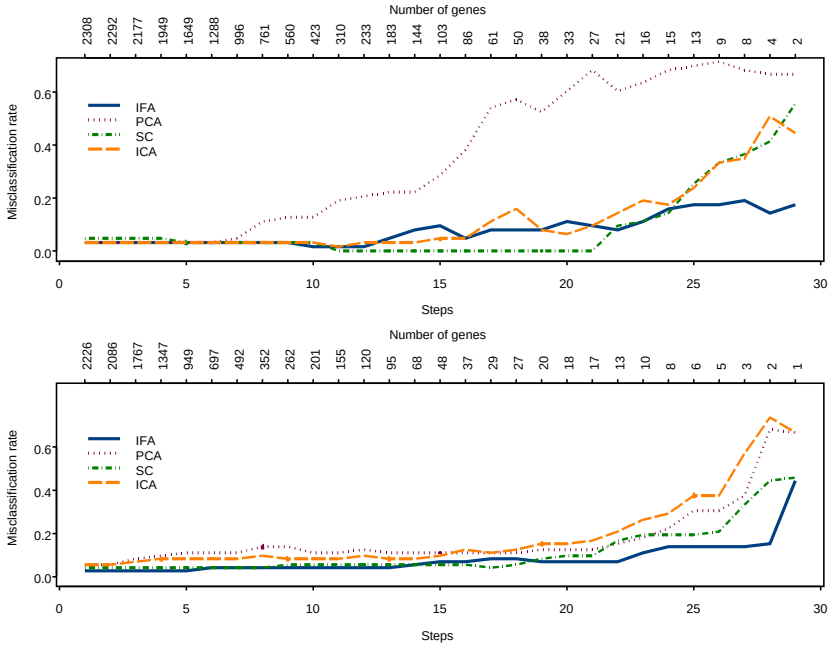

Fig. 2. Cross-validated misclassification rates for different subsets of genes (*first picture*: SRBCT data set; *second picture*: LK data set).

LK data	*2226*	*1347*	*262*	*68*	*37*	*20*	*17*	*10*	*6*	*3*
IFA	0.042	0.028	0.042	0.056	0.069	0.069	0.069	0.111	0.139	0.139
ICA	0.042	0.083	0.083	0.083	0.125	0.153	0.167	0.264	0.375	0.569
PCA	0.042	0.097	0.139	0.111	0.111	0.125	0.125	0.181	0.306	0.375
SC	0.042	0.042	0.056	0.056	0.056	0.083	0.097	0.194	0.194	0.333
SRBCT data	*2308*	*761*	*423*	*310*	*86*	*33*	*20*	*16*	*9*	*2*
IFA	0.032	0.032	0.016	0.016	0.048	0.111	0.079	0.111	0.175	0.175
ICA	0.032	0.032	0.032	0.016	0.048	0.063	0.143	0.190	0.333	0.444
PCA	0.032	0.111	0.127	0.190	0.381	0.603	0.603	0.635	0.714	0.667
SC	0.032	0.048	0.032	0.000	0.000	0.000	0.095	0.111	0.333	0.556

Table 1. *Supervised classification.* Cross-validated misclassification rates for different subsets of genes.

4 Some applications and concluding remarks

The proposed procedure has been applied on some publicly available data sets: the Leukemia (LK) data set of Golub *et al.*, (1999) and the Small Round Blue Cell Tumor (SRBCT) data set of Khan *et al.*, (2001). The first data set contains gene expression levels for p =2226 genes in 72 cells and consists

of 38 cases of B-cell acute lymphoblastic leukemia, 9 cases of T-cell acute lymphoblastic leukemia and 25 cases of acute myeloid leukemia. The small round blue cell data set contains gene expression levels for p =2038 genes in 63 cells and consists of 8 cases of Burkitt lymphoma, 23 cases of Ewing sarcoma, 12 cases of neuroblastoma and 20 cases of rhabdomyosarcoma.

Independent factor analysis has been performed on the two data sets and, according to the information criteria, 2-dimensional and 4-dimensional latent spaces for the LK and the SBRCT data sets respectively have been considered. The right graph of figure 1 displays the non gaussian projections onto the LK latent space. The left graph shows the plot of the genes projected onto the first two estimated factors for the SRBCT data set.

k-means	*2226*	*1347*	*262*	*68*	*37*	*20*	*17*	*10*	*6*	*3*
IFA	0.431	0.139	0.111	0.111	0.139	0.097	0.097	0.125	0.167	0.208
ICA	0.431	0.139	0.097	0.139	0.083	0.097	0.097	0.125	0.153	0.167
PCA	0.431	0.139	0.139	0.444	0.444	0.458	0.639	0.639	0.639	0.611
Ward method	*2226*	*1347*	*262*	*68*	*37*	*20*	*17*	*10*	*6*	*3*
IFA	0.167	0.180	0.111	0.083	0.153	0.069	0.069	0.153	0.167	0.250
ICA	0.167	0.125	0.139	0.069	0.083	0.056	0.139	0.153	0.180	0.153
PCA	0.167	0.194	0.056	0.389	0.403	0.444	0.639	0.639	0.639	0.611

Table 2. *Unsupervised classification.* LK data set: k-means and hierarchical Ward method misclassification rates for different values of selected genes (with 2 factors).

For each data set and for different distance measures, a gene ranking is produced. As the following empirical analysis shows, the proposed strategy seems to represent a useful and promising tool to detect subsets of relevant genes for cell classification based on microarray data. In fact the detected subsets of genes succeed in capturing the class structure in the data, both in a supervised and unsupervised perspective.

Supervised classification

A sequence of classification rules is performed for 30 different values of the number m of selected genes, ranging from p to 1.

The performances of the classification rules based on these subsets of genes are compared with those obtained by principal component analysis and independent component analysis, as an alternative in the projection phase. The nearest shrunken centroid (SC) method of Tibshirani *et al.*, (2002) is also evaluated on the same data sets.

In order to compare the results of our gene selection procedure with those obtained through the shrunken centroids, the nearest centroid method is used in class prediction, but any other postprocessing classifier could be applied. Given the small number of cells in the two data sets, the classification error rates have been estimated by balanced cross-validation. As the performances

k-means	*2308*	*761*	*423*	*310*	*86*	*33*	*20*	*16*	*9*	*2*
IFA	0.571	0.540	0.540	0.540	0.397	0.397	0.317	0.397	0.571	0.254
ICA	0.571	0..571	0.444	0.540	0.556	0.397	0.492	0.349	0.413	0.508
PCA	0.571	0.571	0.571	0.619	0.581	0.651	0.667	0.698	0.698	0.635
Ward method	*2308*	*761*	*423*	*310*	*86*	*33*	*20*	*16*	*9*	*2*
IFA	0.429	0.524	0.540	0.524	0.540	0.317	0.286	0.349	0.413	0.254
ICA	0.429	0.571	0.571	0.587	0.524	0.286	0.397	0.397	0.492	0.508
PCA	0.429	0.508	0.540	0.587	0.571	0.635	0.571	0.667	0.698	0.683

Table 3. *Unsupervised classification.* SRBCT data set: k-means and hierarchical Ward method misclassification rates for different values of selected genes (with 4 factors).

do not seem to be influenced by different choices of the metric when defining the gene ranking, we report only the results based on the Euclidean distance.

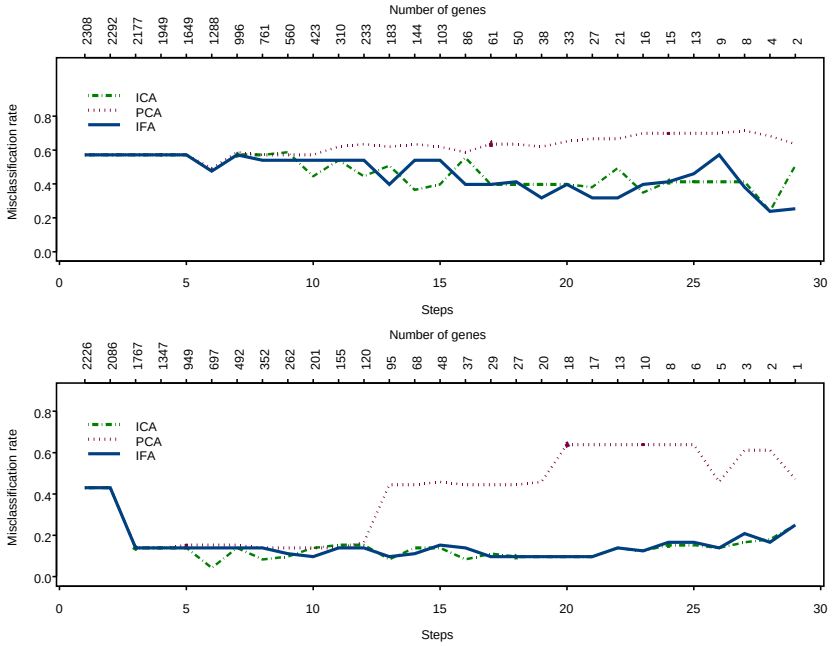

Fig. 3. K-means misclassification rates for different number of selected genes in SRBCT and LK data sets respectively.

The first graph of figure 2 clearly displays the superiority of the performances of classification rules based on subset of genes selected in non gaussian latent spaces. The worst performances of the PCA based solution

confirm the need to take into account also non-linear structures particularly for the SRBCT data set. The IFA-based procedure allows to obtain the better performances, and this is particularly evident for small values of m (see also table 1). The IFA-based procedure outperforms the others also in the LK data classification, where it allows to achieve small cross-validated errors with less than 20 genes.

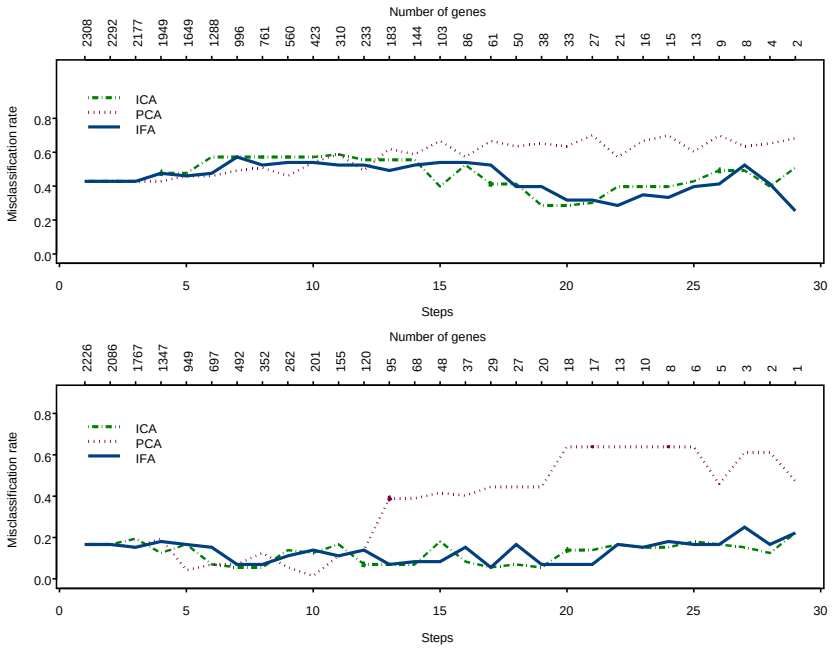

Fig. 4. *Unsupervised classification.* Hierarchical Ward method misclassification rates for different number of selected genes in SRBCT and LK data sets respectively.

SRBCT data set									
m=2308		*Actual*			m=20		*Actual*		
	BL	EWS	NB	RMS		BL	EWS	NB	RMS
Predicted					*Predicted*				
BL	4	3	0	0	BL	0	0	0	0
EWS	0	0	0	0	EWS	8	22	0	0
NB	4	6	9	6	NB	0	0	9	8
RMS	0	14	3	14	RMS	0	1	3	12

Table 4. *Unsupervised classification.* SRBTC data set: confusion matrices for k-means classification with 2308 and 20 selected genes in the IFA latent space.

LK data set							
m=2226		*Actual*		m=20		*Actual*	
	ALL B	ALL T	AML		ALL B	ALL T	AML
Predicted				*Predicted*			
ALL B	16	5	0	ALL B	32	0	0
ALL T	0	0	0	ALL T	6	9	4
AML	22	4	25	AML	0	0	21

Table 5. *Unsupervised classification.* LK data set: confusion matrices for k-means classification with 2226 and 20 selected genes in the IFA latent space.

Unsupervised classification

In order to check if the selected subsets of genes are able to recover the clustering structure of the data, we applied two different cluster analysis techniques to the same data sets. As shown in tables 2 and 3, the use of the whole gene set does not allow to accurately detect the tumor classes by clustering methods. Only after a selection of relevant genes, the performances improve and this fact confirms the usefulness of variable selection in unsupervised classification. The IFA-based gene selection allows to halve the classification error by reducing the number of genes from some thousands to just less than 20. The confusion matrices for all the genes and for a subset of 20 ones confirm the effectiveness of the proposed gene selection procedure.

References

ATTIAS, H. (1999): Independent Factor Analysis. *Neural Computation*, 11, 803–851.

CALÒ, D.G., GALIMBERTI, G., PILLATI, M. and VIROLI, C. (2005): Variable selection in classification problems: a strategy based on independent component analysis. In: M. Vichi, P. Monari, S. Mignani and A. Montanari (Eds.): *New Developments in Classification and Data Analysis*. Studies in Classification, Data Analysis, and Knowledge Organization, Springer-Verlag, Berlin, 21–30.

COMON, P. (1994): Independent component analysis, a new concept? *Signal Processing*, 36, 287–314.

DUDOIT, S., FRIDLYAND, J. and SPEED, T.P. (2002): Comparison of Discrimination Methods for the Classification of Tumors using Gene Expression Data. *Journal of the American Statistical Association*, 457, 77-87.

GOLUB, T.R., SLONIM, D.K., TAMAYO, P. *et al.* (1999): Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286, 531–537.

KHAN, J., WEI, J., RINGNER, M. *et al.* (2001): Classification and Diagnostic Prediction of Cancers Using Gene Expression Profiling and Artificial Neural Networks. *Nature Medicine*, 7, 673–679.

TIBSHIRANI, R., HASTIE, T., NARASIMHAN, B. and CHU, G. (2002): Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression, *Proceedings of the National Accademy of Sciences*, 99, 6567-6572.