Bioinformatics for the management, analysis and interpretation of microarray data

Bologna, November 27 - 28, 2003

Cell classification using gene expression data: a strategy based on independent component analysis

D.G. Calò, G. Galimberti, M. Pillati, C. Viroli

Dipartimento di Scienze Statistiche, Università di Bologna
{calo,galimberti,pillati,viroli}@stat.unibo.it

NETTAB

KEY FEATURES OF GENE EXPRESSION DATA

"LARGE p , SMALL n" PROBLEM:

- *Very large number of variables (genes): from 5,000 to 10,000*
- *Small number of observations (cells): less than 100*

MANY VARIABLES ARE NOISY OR NOT RELEVANT TO CLASS PREDICTION

- *Pearson correlation coefficient*
- *Distance based metrics*
- *Class separation: $BSS(X_i)/WSS(X_i)$*

NETTAB

KEY FEATURES OF GENE EXPRESSION DATA

COMPLEX INTERACTIONS BETWEEN GENES; REDUNDANCY

Genes are points in a n-dimensional space

MOST OF THE ABNORMALITIES IN CELL BEHAVIOUR ARE DUE TO IRREGULAR GENE ACTIVITIES

Look for "outlying" genes

GENE EXPRESSION PROFILES ARE TYPICALLY NON-GAUSSIAN

Independent Component Analysis

NETTAB

INDEPENDENT COMPONENT ANALYSIS (ICA)

THE MODEL: $\boldsymbol{X} = \mathbf{A}\boldsymbol{S}$

$\boldsymbol{X}=(X_1, X_2, \ldots, X_n)$ observed variables $(E(\boldsymbol{X}) = \boldsymbol{0})$

$\boldsymbol{S}=(S_1, S_2, \ldots, S_k)$ latent variables $(E(\boldsymbol{S}) = \boldsymbol{0})$

$\mathbf{A}$= $n \times k$ mixing matrix, $k \leq n$

THE VARIABLES S_j ARE ASSUMED:

- *To be statistically independent:*

$$\text{pdf}\,(S_1, S_2, \ldots, S_k) = \prod \text{pdf}\,(S_j)$$

- *To have non-Gaussian distributions*

NETTAB

INDEPENDENT COMPONENT ANALYSIS (ICA)

$$\boldsymbol{Y = WX}$$

Restrictions on the extracted variables Y_j:

- $E(Y_h\, Y_i)=E(Y_h)E(Y_i) \qquad h \neq i \qquad h, i=1, \ldots, k$
  (in order to reduce the number of free parameters)
- $E(Y_j^2)=1$
  (conventional assumption)

For uncorrelated variables:

- $I\,(Y_1, Y_2, \ldots, Y_k) = J\,(Y_1, Y_2, \ldots, Y_k) - \Sigma J\,(Y_j)$
  (where *I* denotes mutual information and *J* denotes negentropy)

NETTAB

Therefore,

the less dependent are the most non-Gaussian ones

ICA leads to meaningful results whenever the probability distribution of *X* is far from Gaussian

ICA *IN GENE EXPRESSION DATA ANALYSIS:*

- G. Hori *et al.* (2002) - *ISMB2002*
- W. Liebermeister (2002) - *Bioinformatics*
- X. Liao *et al.* (2002) - *ICASSP 2002, IEEE International Conference*

NETTAB

THE PROPOSED SOLUTION

- *k* independent components are extracted from the training set
- the *p* genes are sorted in increasing order according to their absolute scores on each component
- these *k* marginal rankings are summarized by taking, for each gene, the highest value
- select the *m* ($m \ll p$) genes located in the last *m* positions of this joint ranking

NETTAB

AN APPLICATION TO A REAL DATA SET

THE LYMPHOMA DATA SET (Alizadeh et al., 2000):

C = 3 classes

p = 4026 genes

n = 62 cells (n_1=11, n_2=9, n_3=42)

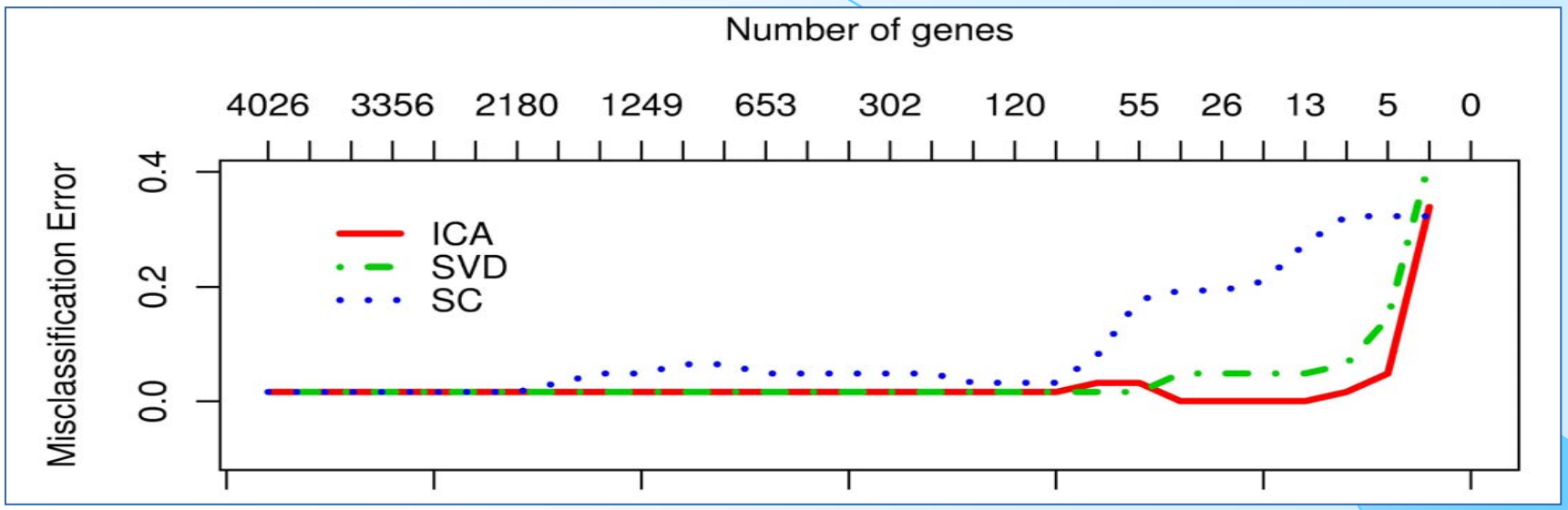
NEAREST SHRUNKEN CENTROIDS METHOD (Tibshirani et al., 2002)

Nearest centroid classification after shrinking each centroid to the overall one.

If a gene is shrunken to zero for all classes, then it is dropped from the allocation rule.

NETTAB

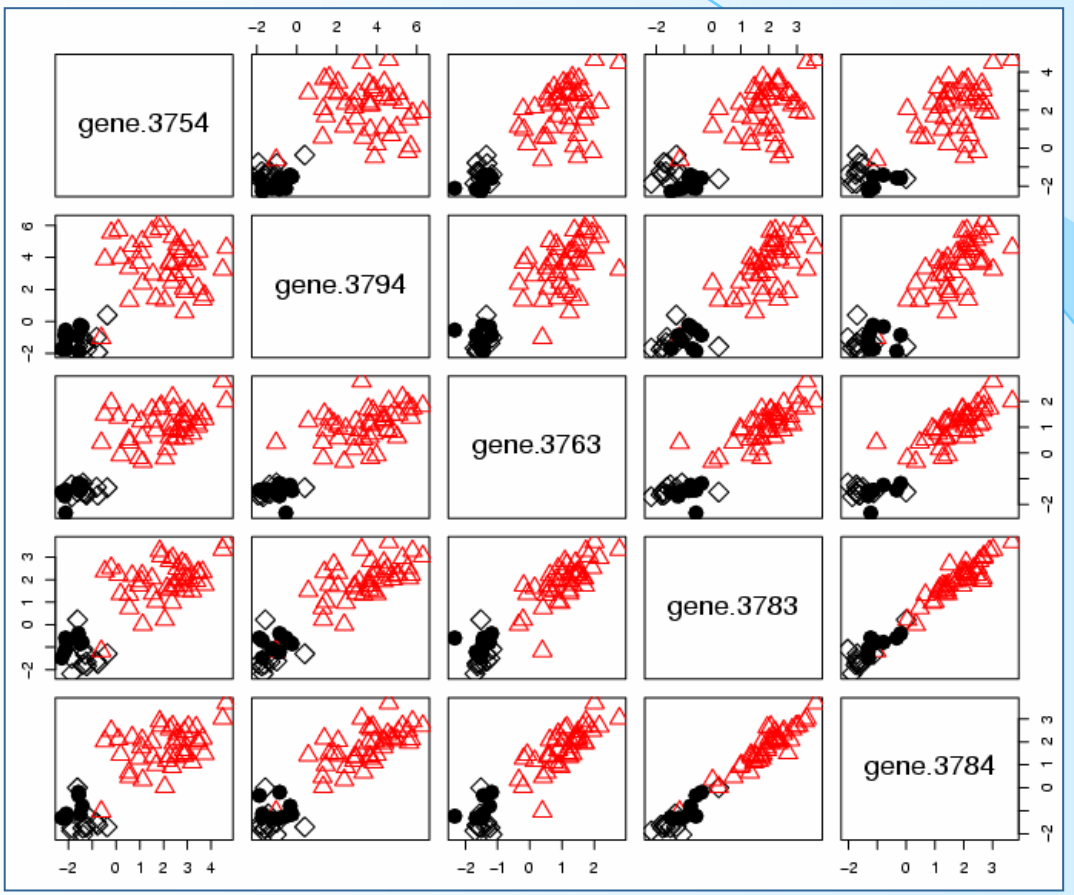
Lymphoma data set: cross-validated misclassification rates (as function of *m*)



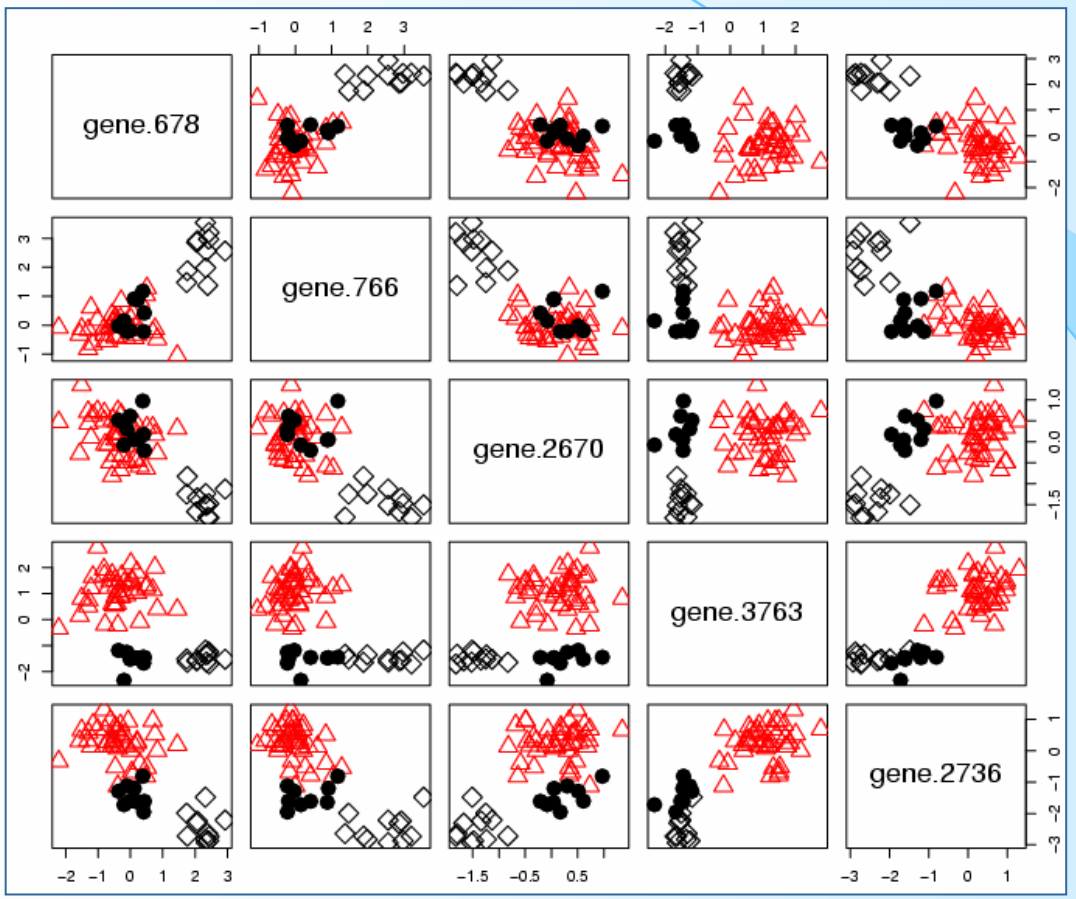
m	*222*	*164*	*120*	*93*	*72*	*55*	*34*	*26*	*19*	*13*	*10*	*5*
SC	0.048	0.032	0.032	**0.032**	0.081	0.177	0.194	0.194	0.210	0.274	0.323	0.323
SVD	0.016	0.016	0.016	0.016	0.016	**0.016**	0.048	0.048	0.048	0.048	0.065	0.145
ICA	0.016	0.016	0.016	0.016	0.032	0.032	0.000	0.000	0.000	**0.000**	0.016	0.048

NETTAB

Lymphoma data set: scatter plot matrix of the last 5 genes surviving the shrinkage procedure



Lymphoma data set: scatter plot matrix of the last 5 genes of the ranking obtained by ICA



OPEN ISSUES

- *Alternatives to the criterion for building the joint ranking*
- *How to choose the number k of the components*
- *How to choose the number m of retained genes*
- *Possible interactions between the proposed selection method and different allocation rules*

NETTAB