

A model-based dimension reduction approach to classification of gene expression data

Luca Scrucca and Avner Bar-Hen

Abstract The monitoring of the expression profiles of thousands of genes have proved to be particularly promising for biological classification, particularly for cancer diagnosis. However, microarray data present major challenges due to the complex, multiclass nature and the overwhelming number of variables characterizing gene expression profiles. We introduce a dimension reduction method for determining the classes among gene expression data obtained from a finite mixture of Gaussian densities. Information on the dimension reduction subspace is obtained from the variation on group means and, depending on the estimated mixture model, on the variation on group covariances. We show that the combination of Slice Inverse Regression and model based clustering is a powerful technique to find group among gene expression data. This a joint work with Luca Scrucca (Univ. Perugia)

Introduction

By monitoring the activity of many genes from a sample in a single experiment microarrays offer a unique perspective for explaining the global genetic picture of a variant, whether a diseased individual or a sample subject to whatever stressing conditions. One of the main result of an experiment with microarrays is a list of genes that are differentially expressed between two conditions, or that allow samples to be classified according to their phenotypic features. An overview of classification methods for microarray data is given in [Dudoit and Fridlyand, 2003] while a comparison among several of them is provided in [Dudoit *et al.*, 2002].

Luca Scrucca
Università degli Studi di Perugia, 06100 Perugia, Italy.

Avner Bar-Hen
Université Paris Descartes, 45 rue des Saints-Pères, 75270 Paris, France.

Classification of microarray data is particularly problematic due to under-resolution, i.e. the presence of a large number of features (genes) from which to predict classes compared to the relatively small number of observations (samples). Moreover the expression levels of many genes are often highly correlated. Classical procedures for dimensionality reduction are principal components analysis and factor analysis, both of which reduce dimensionality by forming linear combinations of the features. The first method seeks a lower-dimensional representation that account for the most variance of the features, while the second looks for the most correlation among the features. Alternatively, inverse regression (SIR) is a dimension reduction method which exploits the information from the inverse mean function (see [Li, 1991]. SIR directions span at least part of the dimension reduction subspace [Cook, 1998, Chap. 6] and they provide an intuitive and useful basis for constructing summary plots [Scrucca, 2007].

Dimension reduction techniques have already been studied in the microarray context. A two-stage approach based on sliced SIR was proposed by [Chiaromonte and Martinelli, 2002], with the first stage aimed at eliminating the dimension redundancy through an eigenanalysis which identified a set of linear combinations of the original genes which retained the main variability structure. Then, the SIR method was applied on such reduced set of eigengenes [Li and Li, 2004]. A final gene selection stage was proposed based on a ranking of genes in accordance to the concordance with the final estimated directions. A related approach was also discussed by [Bura and Pfeiffer, 2003], in which they applied both SIR and SAVE, i.e. sliced average variance estimation, to microarray data.

Clustering algorithms based on probability models offer an alternative to heuristic based algorithms. In particular, the model based approach assumes that the data are generated by a finite mixture of underlying probability distributions such as multivariate normal distribution. The Gaussian mixture model has been shown to be a powerful tool for many applications [Banfield and Raftery, 1993, McLachlan and Basford, 1998, Yeung *et al.*, 2001, McLachlan *et al.*, 2002].

In this paper, we introduce a model based clustering on the low dimensional space obtained by SIR projection of the the high-dimensional gene expression data. SIR components allow to project the gene expression data in a lower dimension space where the direction are uncorrelated. This provided a robust framework to classify the data. On the other hand, the model based clustering part gives the flexibility to catch the class structure within the gene expression data.

Methods

Gene expression data on p genes for n mRNA samples can be summarized by a $n \times p$ matrix $X = [x_{ij}]$, where x_{ij} denotes the expression level of gene (variable) j in mRNA sample (observation) i . Note that in this representation the data matrix is the transpose of that adopted in the microarray literature. The expression levels might be

either absolute (e.g. oligonucleotide arrays) or relative with respect to the expression levels of a suitably defined common reference sample (e.g. cDNA microarrays).

Classification methods aim at assigning a sample to one of K given classes based on information from a set of p predictor variables observed on n samples. Thus, a categorical response variable Y with K levels representing biological outcomes, such as tumors type, is also recorded along with gene expression levels. An overview of classification methods for microarray data is given in [Dudoit and Fridlyand, 2003] while a comparison among several of them is provided in [Dudoit *et al.*, 2002].

Dimension reduction methods

The general goal of a dimension reduction method is to find a $p \times d$ matrix β , with $d \leq p$, such that

$$Y \perp\!\!\!\perp X | \beta^\top X,$$

i.e. the classification response Y is independent of the predictors X given the linear reduction $\beta^\top X$. The subspace of dimension d spanned by the matrix β is called the sufficient dimension reduction (SDR) subspace for the regression of Y on X . It always exists, since we can trivially set $\beta = I$, but it is not unique. The central dimension-reduction (CDR) subspace is defined as the intersection over all the SDR subspaces; under mild conditions, it exists and it is the unique minimum SDR subspace [Cook, 1998, Chap. 6].

Several methods have been proposed to estimate the CDR subspace, such as sliced inverse regression (SIR) [Li, 1991] and sliced average variance estimation (SAVE) [Cook and Weisberg, 1991].

SIR, which is perhaps the most commonly used DR method, estimates the basis of the CDR subspace by the first d eigenvectors of the decomposition of $\Sigma_{X|Y} = \text{Var}(E(X|Y))$ with respect to $\Sigma = \text{Var}(X)$, i.e.

$$\begin{aligned} \Sigma_{X|Y} v_i &= l_i \Sigma v_i, \quad v_i^\top \Sigma v_j = 1 \text{ if } i = j, \text{ and } 0 \text{ otherwise,} \\ l_1 &\geq l_2 \geq \dots \geq l_d > 0. \end{aligned} \tag{1}$$

SIR requires the linearity condition on the marginal distribution of the predictors, i.e. requires that $E(b^\top X | \beta^\top X)$ is linear in $\beta^\top X$ for all $b \in \mathbb{R}^p$. This condition is satisfied if the predictors follow an elliptical symmetric distribution, such as the multivariate normal. Furthermore, this condition is not a severe restriction, since most low-dimensional projections are close to being normal [Hall and Li, 1993].

Usually, a sliced version of the response variable is needed, but in classification settings Y is taken directly as the class indicator with K levels. Sample estimates are then used to compute the unknown quantities in (1), and directions estimated as $\hat{\beta}_{\text{SIR}} \equiv [\hat{v}_1, \dots, \hat{v}_d]$. In this case SIR can find at most $\min(K-1, p)$ directions. Thus, for a binary response variable SIR estimates at most a single direction.

Regularized estimation of covariance matrix

Usually the number of genes is larger than the number of available samples. Hence, direct application of SIR is not possible because of the singularity of the predictors covariance matrix. This fact was not recognized in [Bura and Pfeiffer, 2003], while [Chiaromonte and Martinelli, 2002] addressed this problem by using singular value decomposition as an intermediate step to produce the so-called eigen-genes. Other approaches are based on a form of regularization for the covariance matrix. [Zhong *et al.*, 2005] introduced a ridge estimator with a ridge parameter selected via bootstrap. A similar shrinkage estimator has been also employed by [Scrucca, 2007], where the proposed estimator is obtained as a convex linear combination between the full covariance matrix and the diagonal matrix having the average total variation along the diagonal; the coefficient that determines the convex linear combination is selected by cross-validation. An L2-regularization was proposed by [Li and Yin, 2008] through a least-squares formulation of SIR, with the tuning parameter selected by an information criteria. Recently, [Bernard *et al.*, 2009] discussed a family of regularization procedures based on a Gaussian prior distribution, with some of the earlier proposals as special cases; however, the selection of prior parameters is still under investigation.

In this paper we adopted a simpler approach, by estimating the covariance matrix as $\hat{\Sigma} = \text{diag}[s_i^2]_{i=1,\dots,p}$, where s_i^2 is the sample variance of the i -th gene. This is based on the fact that some of the most successful methods in microarray data analysis assume very weak dependence or even ignore correlation among genes [Dudoit *et al.*, 2002, Tibshirani *et al.*, 2002]. This also avoid the problem of selecting a tuning parameter.

Model-based SIR

The SIR estimator is based on the information coming from the variation variation on slice means. Thus, the distribution of the data within any class is summarized by the class means μ_k ($k = 1, \dots, K$). However, it may happen that a more complicated distribution of the data is needed. Finite mixture of Gaussian densities can be used to approximate the distribution of the predictors within any class, and then the kernel matrix can be computed using the components means [Scrucca, 2008].

Assume that within each class the data can be described by the finite mixture distribution

$$f(x|Y = k) = f_k(x) = \sum_{c=1}^{C_k} \pi_{kc} \phi(x; \mu_{kc}, \Sigma_{kc}),$$

where μ_{kc} and Σ_{kc} are the means and covariances for component c within class k , π_{kc} are the mixing weights such that $\pi_{kc} \geq 0$ and $\sum_c \pi_{kc} = 1$, and $\phi(\cdot)$ is the multivariate Gaussian density. Geometric features, such as shape, volume and orientation, of each component are determined by the covariance structure. [Banfield and Raftery, 1993]

proposed a representation of the covariance matrix for a finite mixture model, say in general Σ_g , in terms of its eigenvalue decomposition, $\Sigma_g = \lambda_g D_g A_g D_g^\top$, where D_g is the orthogonal matrix of eigenvectors which determines the orientation, A_g is a diagonal matrix of scales eigenvalues which determines the shape, and λ_g is a scalar determining the volume of the hyper-ellipsoid. Given that $p \gg n$, only a subset of all the possible models are allowed. In particular, only spherical and diagonal finite mixture models, obtained by imposing $D_g = I$, can be fitted. This, however, appears to be sufficiently flexible to accommodate data with different characteristics.

Estimation of parameters is obtained via the EM algorithm [McLachlan and Peel, 2000]. The number of components C_k , as well as the particular form of the covariance matrix Σ_{kc} , within the above mentioned restrictions on the possible models, may be chosen by BIC [Fraley and Raftery, 2002].

Consider the kernel matrix given by the covariance matrix of the between-component means

$$M = \sum_{k=1}^K \sum_{c=1}^{C_k} \omega_{kc} (\mu_{kc} - \mu)(\mu_{kc} - \mu)^\top,$$

where $\omega_{kc} = \pi_{kc} \Pr(Y = k)$ ($\omega_{kc} \geq 0$, $\sum_{k,c} \omega_{kc} = 1$), and the covariance matrix $\Sigma = n^{-1} \sum_{i=1}^n (x_i - \mu)(x_i - \mu)^\top$ with $\mu = \sum_{k=1}^K \sum_{c=1}^{C_k} \omega_{kc} \mu_{kc}$. An estimate of the CDR subspace is the solution of the following constrained optimization:

$$\operatorname{argmax}_{\beta} \beta^\top M \beta, \text{ subject to } \beta^\top \Sigma \beta = I,$$

where $\beta \in \mathbb{R}^{p \times d}$ is the spanning matrix, and I is the $(d \times d)$ identity matrix. This is solved through the generalized eigendecomposition

$$\begin{aligned} M v_i &= l_i \Sigma v_i, \quad v_i^\top \Sigma v_j = 1 \text{ if } i = j, \text{ and } 0 \text{ otherwise,} \\ l_1 &\geq l_2 \geq \dots \geq l_d > 0. \end{aligned} \tag{2}$$

The eigenvectors, corresponding to the first d largest eigenvalues provide a basis for the CDR subspace which shows the maximal variation between component means. There are at most $d = \min(p, C - 1)$ directions which span this subspace, where $C = \sum_{k=1}^K C_k$ is the total number of finite mixture components. Note that by imposing a single mixture component for each class (i.e., $C_k = 1$ for all $k = 1, \dots, K$) it reverts to the usual SIR estimator.

The sample version of MSIR is obtained by applying the eigendecomposition in (2) with the kernel matrix computed from the means estimated for each mixture component, and $\hat{\omega}_{kc} = n_{kc}/n$ is the observed fraction of observations belonging to the c -th mixture component within class k . Thus, the sample basis estimate is $\hat{\beta}_{\text{MSIR}} \equiv [\hat{v}_1, \dots, \hat{v}_d]$, and the projections onto such subspace are defined as $Z = X \hat{\beta}_{\text{MSIR}}$.

Preliminary filtering of genes

The large number of gene expression profiles usually collected in microarray experiments can be drastically reduced since many of them exhibit near constant expression levels across samples. In our analyses we performed a preliminary filtering step by selecting the genes which have the highest ratio of between- to within-groups sums of squares [Dudoit and Fridlyand, 2003, Scrucca, 2007].

Soft-thresholding

Soft-thresholding can be used to shrink coefficients toward zero [Tibshirani *et al.*, 2002]. This enables to identify subsets of genes relevant for the definition of a given set of MSIR directions.

Let $\hat{\beta}_{ij} \leftarrow \hat{\beta}_{ij}/s_i$ be the coefficient of the j -th direction for the i -th predictor standardized to unit standard deviation. The soft-thresholding function is defined as $s(x, \Delta) = \text{sign}(x)(|x| - \Delta)_+$, where $(u)_+ = u$ if $u > 0$ and zero otherwise. For a threshold value Δ_j ($j = 1, \dots, d$), we define the shrunk coefficients as $\tilde{\beta}_{ij} = s(\hat{\beta}_{ij}, \Delta_j)_+/s_i$, orthonormalized such that $\tilde{\beta}^\top \hat{\Sigma} \tilde{\beta} = I$, where $\tilde{\beta} = [\tilde{\beta}_{ij}] \in \mathbb{R}^{p \times d}$.

The shrunk directions $\tilde{\beta}$ can be computed for a grid of Δ_j values. Let $X\hat{\beta}$ and $X\tilde{\beta}$ be, respectively, the projection onto the subspace spanned by the estimated MSIR directions and the shrunk directions. The squared trace correlation [Li, 1991] provides a measure of the agreement between the two subspaces. For a single directions it is equivalent to the usual squared correlation coefficient. However, this measure cannot be used to select the amount of shrinkage since it is always decreasing as the threshold values is increased.

We suggest to select the amount of shrinkage applied on the basis of the maximum classes separation relative to their compactness on the projection subspace. This can be achieved by computing the Dunn index, which is given by the ratio of the smallest distance between observations not in the same class to the largest intra-class distance [Dunn, 1974]. For a given threshold value Δ_j we compute

$$D(\Delta_j) = \frac{\min_{i \in \mathcal{C}_k, l \notin \mathcal{C}_k} D_{il}}{\max_{i, l \in \mathcal{C}_k} D_{il}},$$

where $D_{il} = [(\tilde{\beta}^\top x_i - \tilde{\beta}^\top x_l)^\top (\tilde{\beta}^\top x_i - \tilde{\beta}^\top x_l)]^{1/2}$ and $\mathcal{C}_k = \{i : Y_i = k\}$ is the set of observations belonging to the k -th class ($k = 1, \dots, K$). Then, the amount of threshold Δ_{j^*} is chosen to maximize $D(\Delta_j)$. Careful study $D(\Delta_j)$ can also be useful to detect local maxima attained with smaller sets of genes.

Results

Leukemia data

[Golub *et al.*, 1999] analyzed data from high-density Affymetrix oligonucleotide arrays. There were 3571 genes and 38 samples: 27 in class ALL (acute lymphoblastic leukemia) and 11 in class AML (acute myeloid leukemia). The samples in class ALL could be further splitted into B-cell and T-cell types. After the preliminary filtering a subset of 892 genes was selected.

MSIR estimates a 3-components mixture for AML group and a 2-components mixture for the ALL group. There are four estimated directions, with corresponding eigenvalues equal to 273.3, 45.5, 40.8, and 24.3. Figure 1 shows a scatterplot of the samples projected along the first two MSIR directions, with marginal box-plots. The two tumor classes ALL and AML are clearly separated on the first direction. The second direction allows to split the ALL samples into B-cell (open circles) and T-cell (filled circles) types, with one unusual B-cell which is very close to the group of T-cell samples. This is interesting because MSIR estimation did not use the information on the different type of ALL tumors; nonetheless, it was able to show an underlying data structure.

We also performed SIR estimation; the only direction estimable by SIR presents a large agreement with the first one estimated by MSIR (the squared correlation coefficient is equal to 0.997).

Figure 2 shows the results from applying the soft-thresholding procedure to the first MSIR direction. The maximal separation measure by the Dunn index is achieved for a threshold value of $\Delta_j = 0.049$; at this value 8 genes can be used to reproduce the original estimated MSIR direction to a large extent (the squared correlation coefficient is equal to 0.9292).

Applying the soft-thresholding procedure to the first two MSIR directions lead to select the previous 8 genes, plus another gene (X03934) which is mainly responsible for the definition of the second direction. A plot analogous to Figure 1 is shown in Figure 3; here we plotted the samples projected along the first two MSIR directions after soft thresholding. By visual inspection we can see that the main data features are retained, but using only a small subset of genes.

Figure 4 shows a plot of expression levels for the selected genes. The first 8 genes (read from top to bottom) allow to separate the ALL from AML tumor types, and they contribute to the definition of the first MSIR direction. The last gene is the only needed to define the second MSIR direction: as we already mentioned, the B-cell and T-cell tumors for ALL are separated along this direction. This is clearly visible in the corresponding gene profile shown at the bottom of Figure 4, where samples for ALL T-cell are over-expressed, whereas those for ALL B-cell have expression levels around zero, except for one sample (number 17) which has a level comparable to the T-cell group. Note that this is the one outlying B-cell observation which appears closer to the group of T-cell in Figure 1 and Figure 3.

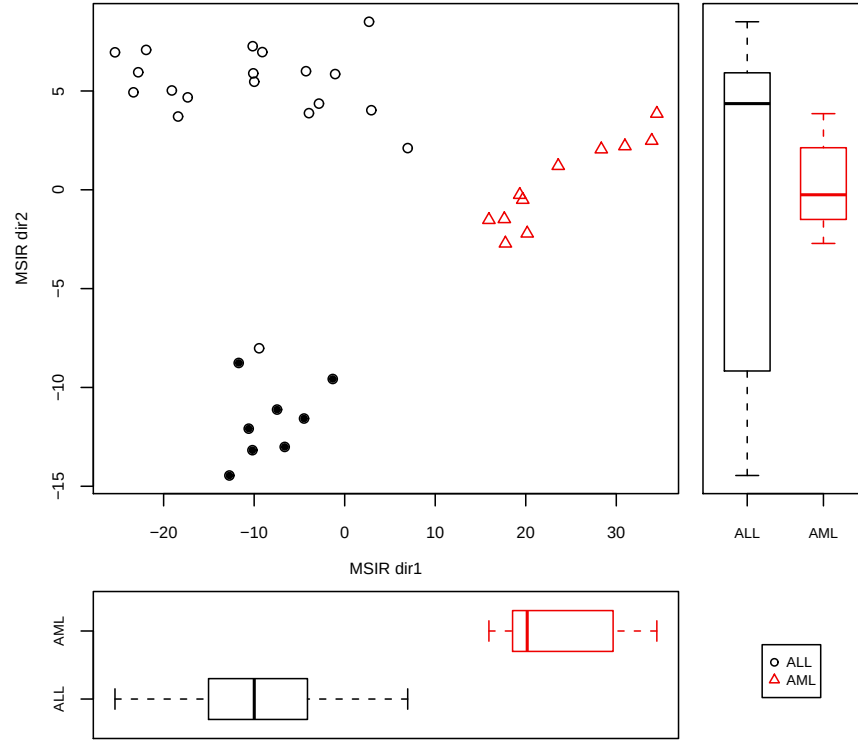


Fig. 1 Scatterplot with marginal box-plots for the Golub data projected along the first two MSIR directions, with points marked by tumor classes (the ALL tumors are further identified by open circles for B-cell type and filled circles for T-cell type).

Discussion

There has been emphasis on mixture model based approach to clustering as it provides a sound mathematical based method. However, in using this approach, the number of observations to be clustered needs to be sufficiently large in number relative to their dimension in order to prevent singular estimates of the component-covariance matrices occurring during the estimation process. Unfortunately, this is not the case with the problem of clustering tissues on the basis of gene expression levels, as the latter are typically much larger than the number of tissues to be clustered.

In this paper we addressed the problem of classification from model-based clustering on dimension reduced space. Generally these two steps (classification and model based clustering) are done separately leading to sub-optimal procedures. Robustness of dimension reduced space and flexibility of mixture model allow for significant improvement in class identification.

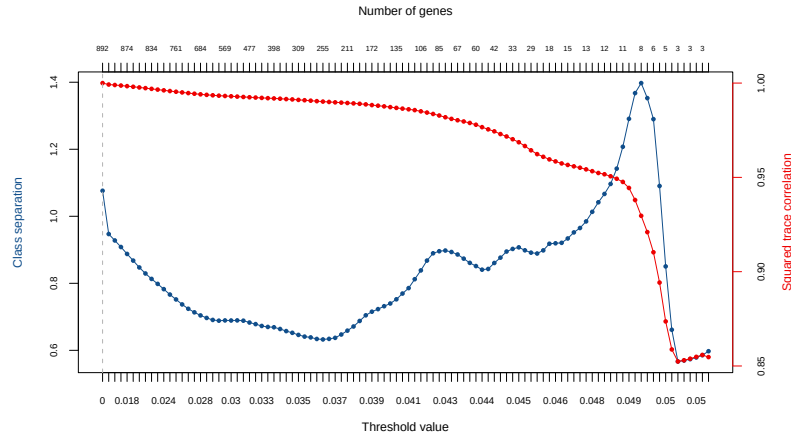


Fig. 2 Trace plot of class separation index and square trace correlation as the threshold value is increased.

The proposed approach is demonstrated on a well known data sets in the microarray literature, the leukaemia of [Golub *et al.*, 1999]. The aim was not to provide a detailed analysis of these sets, but rather to highlight the potential role and usefulness of our approach. In particular, we demonstrated how mixture model on reduced dimension space can identify various classes among tissues on the basis of gene expression levels.

The improvement is larger when within-cluster covariances are different and redundant or noise variables are present.

Acknowledgement

The authors thank the Working Group on Model-Based Clustering Summer Session (Paris, 2009) where this work was initiated.

References

- [Banfield and Raftery, 1993] Banfield, J. and Raftery, A. E. (1993) Model-based gaussian and non-gaussian clustering. *Biometrics*, **49**, 803–821.
- [Bernard *et al.*, 2009] Bernard-Michel, C., Gardes, L., and Girard, S. (2009) Gaussian regularized sliced inverse regression. *Statistics and Computing*, **19**, 85–98.
- [Bura and Pfeiffer, 2003] Bura, E. and Pfeiffer, R.M. (2003) Graphical methods for class prediction using dimension reduction techniques on DNA microarray data. *Bioinformatics*, **19**, 1252–1258.

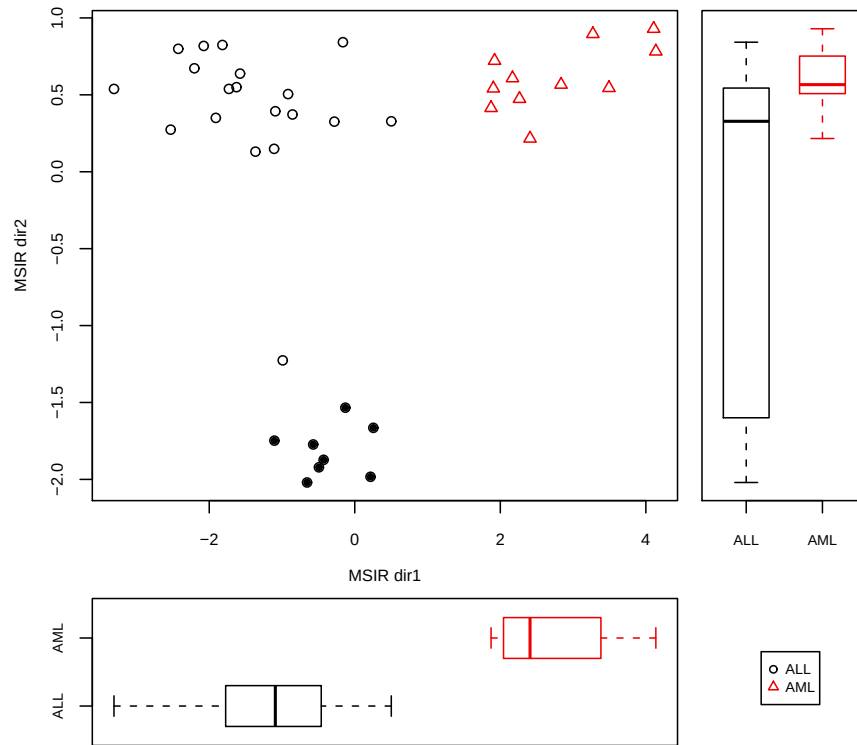


Fig. 3 Scatterplot with marginal box-plots for the Golub data projected along the first two MSIR directions after applying the soft-thresholding procedure.

- [Chiaromonte and Martinelli, 2002] Chiaromonte, F. and Martinelli, J. (2002). Dimension reduction strategies for analyzing global gene expression data with a response. *Mathematical Biosciences*, **176**, 123–144.
- [Cook, 1998] Cook, R. D. (1998) *Regression Graphics: Ideas for Studying Regressions Through Graphics*. Wiley, New York.
- [Cook and Weisberg, 1991] Cook, R. D. and Weisberg, S. (1991) Discussion of Li (1991). *Journal of the American Statistical Association*, **86**, 328–332.
- [Dudoit et al., 2002] 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*, **97**, 77–87.
- [Dudoit and Fridlyand, 2003] Dudoit, S. and Fridlyand, J. (2003) Classification in microarray experiments. In *Statistical analysis of gene expression microarray data* (ed. Speed, T. P), Chapman & Hall/CRC, Boca Raton, FL, p. 93–158.
- [Dunn, 1974] Dunn, J. (1974) Well separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, **4**, 95–104.
- [Fraley and Raftery, 2002] Fraley, C. and Raftery, A. E. (2002) Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, **97**, 611–631.

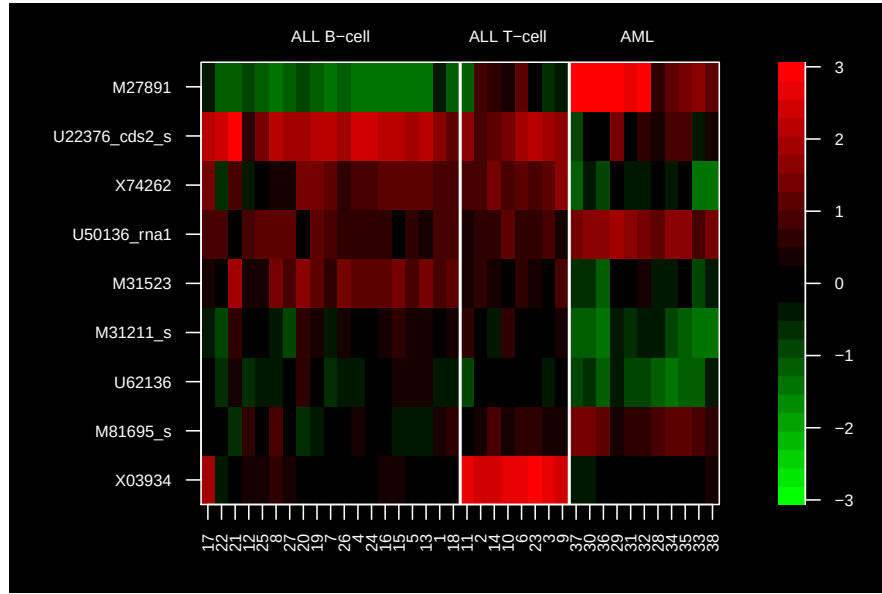


Fig. 4 Gene expression profiles for the selected genes. The display is a heat map, ranging from bright green (negative, under expressed) to bright red (positive, over expressed).

- [Golub *et al.*, 1999] Golub T.R., Slonim D.K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J.P., Coller H., Loh M.L., Downing J. R., Caligiuri M.A., Bloomfield C.D., Lander E.S. (1999) Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, **286**, 531-537.
- [Hall and Li, 1993] Hall, P. and Li, K.-C. (1993) On almost linearity of low dimensional projections from high dimensional data. *The Annals of Statistics*, **21**, 867-889.
- [Li, 1991] Li, K. C. (1991) Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, **86**, 316-342.
- [Li and Li, 2004] Li, L. Li, H. (2004) Dimension reduction methods for microarrays with application to censored survival data. *Bioinformatics*, **20**, 3406-3412.
- [Li and Yin, 2008] Li, L. and Yin, X. (2008) Sliced inverse regression with regularizations. *Biometrics*, **64**, 124-131.
- [McLachlan and Basford, 1998] McLachlan, G.J. and Basford, K.E. (1988) *Mixture Models: Inference and Applications to Clustering*. Dekker, New-York.
- [McLachlan *et al.*, 2002] McLachlan, G.J., Bean, R.W. and Peel, D. (2002) A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, **18**(3), 413-422.
- [McLachlan and Peel, 2000] McLachlan, G.J. and Peel, D. (2000) *Finite Mixture Models*. Wiley, New York.
- [Scrucca, 2006] Scrucca, L. (2006) Regularized sliced inverse regression with applications in classification. In Zani, S., Cerioli, A. and Vichi, M. editors, *Data Analysis, Classification and the Forward Search*, Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Berlin, 59-66.
- [Scrucca, 2007] Scrucca, L. (2007) Class prediction and gene selection for DNA microarrays using regularized sliced inverse regression. *Computational Statistics & Data Analysis*, **52**, 438-451.

- [Scrucca, 2008] Scrucca, L. (2008) A mixture model approach to dimension reduction in regression. *SIS2008, Proceedings of XLIV Scientific Meeting of the Italian Statistical Society*, pp. 233–240. Arcavacata di Rende, 25–27 June, 2008.
- [Tibshirani *et al.*, 2002] 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 Academy of Sciences*, **99**, 6567–6572.
- [Yeung *et al.*, 2001] Yeung, K.Y., Fraley, C., Murua, A., Raftery, A.E. and Ruzzo, W.L. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **17**, 977–987.
- [Zhong *et al.*, 2005] Zhong, W., Zeng, P., Ma, P., Liu, J. S., and Zhu, Y. (2005) RSIR: regularized sliced inverse regression for motif discovery. *Bioinformatics*, **21**, 4169–4175.