Figure 1

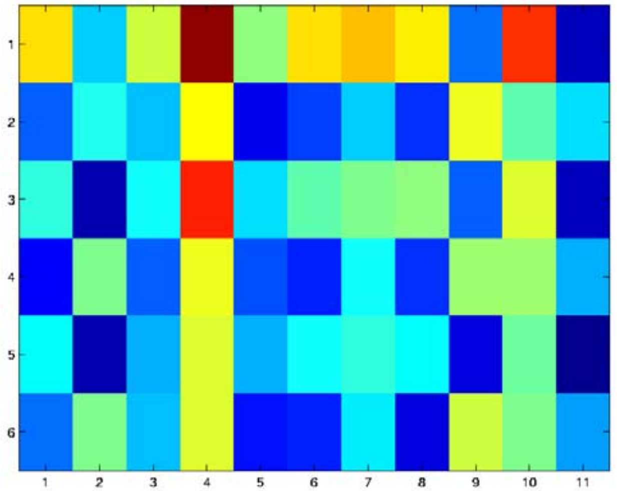
This figure gives an overview of important parts of the biclustering process. Part A shows the problem: shuffling a gene expression matrix to reveal a checkerboard pattern associating genes with conditions. Part B shows how this problem can be approached through solving an “eigenproblem.” If a gene expression matrix *A* has a checkerboard structure, applying it to a step-like condition classification vector *x* will result in a step-like gene classification vector *y*. Moreover, if one then applies A^T to y, one will regenerate a step-like condition classification vector with the same partitioning structure as *x*. This suggests one can determine if *A* has a checkerboard structure through solving an eigenvalue problem. In other words, if *A* has a (hidden) checkerboard structure there exist some piecewise constant partition vectors $x = v_*$ and $y = u_*$ such that $A^T A v_* = \lambda^2 v_*$ and $AA^T u_* = \lambda^2 u_*$ (bottom quadrant of part B). To reveal whether the data has checkerboard structure one can inspect if some of the pairs of monotonically sorted gene and tumor eigenvectors v_i and u_i have an approximate stepwise (piecewise) constant structure. The outer product $u_* v_*^T$ of the sorted partitioning eigenvectors gives a checkerboard structure. Part C shows how rescaling of matrix *A* can lead to improved co-partitioning of genes and conditions.

(On Next Page...)

(A) The Problem: Identifying Marker Genes Associated with Certain Conditions

Matrix of raw data

Genes



Conditions

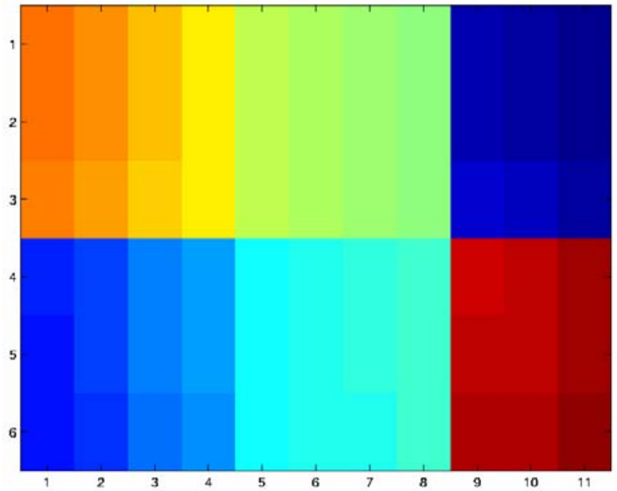
?

Shuffled Matrix

(containing checkerboard "biclusters" of conditions with marker genes)

Reordered Genes

(Sorted according to a classification vector)



Reordered Conditions

(Sorted according to a classification vector)

(B) Identifying checkerboard matrices by their action on classification vectors: Formulation as “eigenproblem”

Gene Classification Vector $\boldsymbol{y}$

Checkerboard Matrix $\boldsymbol{A}$

Genes

$$\begin{pmatrix} 8 & 8 & 8 & 8 & 7 & 7 & 7 & 7 & 3 & 3 & 3 \\ 8 & 8 & 8 & 8 & 7 & 7 & 7 & 7 & 3 & 3 & 3 \\ 8 & 8 & 8 & 8 & 7 & 7 & 7 & 7 & 3 & 3 & 3 \\ 6 & 6 & 6 & 6 & 4 & 4 & 4 & 4 & 5 & 5 & 5 \\ 6 & 6 & 6 & 6 & 4 & 4 & 4 & 4 & 5 & 5 & 5 \\ 6 & 6 & 6 & 6 & 4 & 4 & 4 & 4 & 5 & 5 & 5 \end{pmatrix} \begin{pmatrix} a \\ a \\ a \\ a \\ b \\ b \\ b \\ b \\ c \\ c \\ c \end{pmatrix} = \begin{pmatrix} D \\ D \\ D \\ E \\ E \\ E \end{pmatrix}$$

Conditions

Condition Classification Vect. $\boldsymbol{x}$ →

$$\boldsymbol{A}^T \qquad \boldsymbol{y}$$

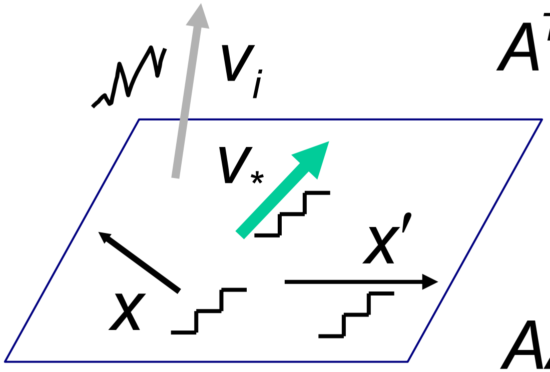
Conditions

$$\begin{pmatrix} 8 & 8 & 8 & 6 & 6 & 6 \\ 8 & 8 & 8 & 6 & 6 & 6 \\ 8 & 8 & 8 & 6 & 6 & 6 \\ 8 & 8 & 8 & 6 & 6 & 6 \\ 7 & 7 & 7 & 4 & 4 & 4 \\ 7 & 7 & 7 & 4 & 4 & 4 \\ 7 & 7 & 7 & 4 & 4 & 4 \\ 7 & 7 & 7 & 4 & 4 & 4 \\ 3 & 3 & 3 & 5 & 5 & 5 \\ 3 & 3 & 3 & 5 & 5 & 5 \\ 3 & 3 & 3 & 5 & 5 & 5 \end{pmatrix} \begin{pmatrix} D \\ D \\ D \\ E \\ E \\ E \end{pmatrix} = \begin{pmatrix} a' \\ a' \\ a' \\ a' \\ b' \\ b' \\ b' \\ b' \\ c' \\ c' \\ c' \end{pmatrix}$$

Genes

$\boldsymbol{x}'$

$$A^T A x = x'$$



$$A^T A v = \lambda^2 v$$

$$AA^T y = y'$$

$$AA^T u = \lambda^2 u$$

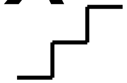
$$v_* \quad u_*^T$$

$$\begin{pmatrix} a \\ a \\ a \\ a \\ b \\ b \\ b \\ b \\ c \\ c \\ c \end{pmatrix} \begin{pmatrix} D & D & D & E & E & E \end{pmatrix} = \begin{pmatrix} aD & aD & aD & aE & aE & aE \\ aD & aD & aD & aE & aE & aE \\ aD & aD & aD & aE & aE & aE \\ aD & aD & aD & aE & aE & aE \\ bD & bD & bD & bE & bE & bE \\ bD & bD & bD & bE & bE & bE \\ bD & bD & bD & bE & bE & bE \\ bD & bD & bD & bE & bE & bE \\ cD & cD & cD & cE & cE & cE \\ cD & cD & cD & cE & cE & cE \\ cD & cD & cD & cE & cE & cE \end{pmatrix}$$

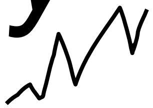
(C) A First Step of Matrix Normalization: Rescaling Rows to Same Mean

$$\begin{pmatrix} 8 & 8 & 8 & 8 & 7 & 7 & 7 & 7 & 3 & 3 & 3 \\ 16 & 16 & 16 & 16 & 14 & 14 & 14 & 14 & 6 & 6 & 6 \\ 8 & 8 & 8 & 8 & 7 & 7 & 7 & 7 & 3 & 3 & 3 \\ 12 & 12 & 12 & 12 & 8 & 8 & 8 & 8 & 10 & 10 & 10 \\ 6 & 6 & 6 & 6 & 4 & 4 & 4 & 4 & 5 & 5 & 5 \\ 6 & 6 & 6 & 6 & 4 & 4 & 4 & 4 & 5 & 5 & 5 \end{pmatrix} \begin{pmatrix} a \\ a \\ a \\ a \\ b \\ b \\ b \\ b \\ c \\ c \\ c \end{pmatrix} = \begin{pmatrix} D \\ 2D \\ D \\ 2E \\ E \\ E \end{pmatrix}$$

$$A_{raw} x = y$$



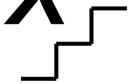
step-like



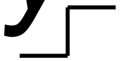
zigzag

$$\begin{pmatrix} .12 & .12 & .12 & .12 & .10 & .10 & .10 & .10 & .04 & .04 & .04 \\ .12 & .12 & .12 & .12 & .10 & .10 & .10 & .10 & .04 & .04 & .04 \\ .12 & .12 & .12 & .12 & .10 & .10 & .10 & .10 & .04 & .04 & .04 \\ .11 & .11 & .11 & .11 & .07 & .07 & .07 & .07 & .09 & .09 & .09 \\ .11 & .11 & .11 & .11 & .07 & .07 & .07 & .07 & .09 & .09 & .09 \\ .11 & .11 & .11 & .11 & .07 & .07 & .07 & .07 & .09 & .09 & .09 \end{pmatrix} \begin{pmatrix} a \\ a \\ a \\ a \\ b \\ b \\ b \\ b \\ c \\ c \\ c \end{pmatrix} = \begin{pmatrix} D \\ D \\ D \\ E \\ E \\ E \end{pmatrix}$$

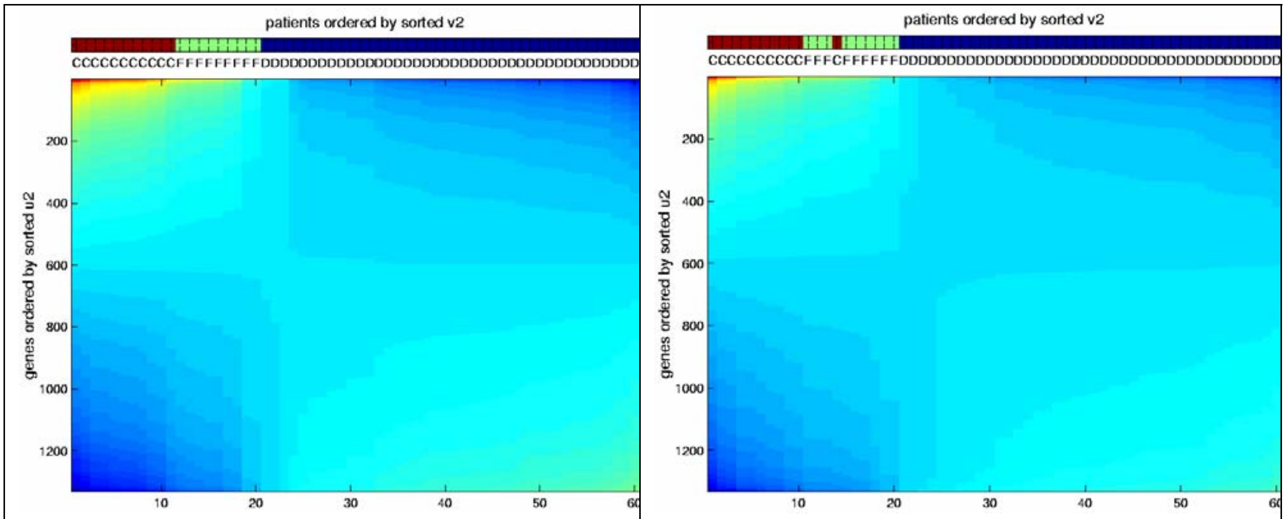
$$R^{-1} A_{raw} x = y$$



step-like

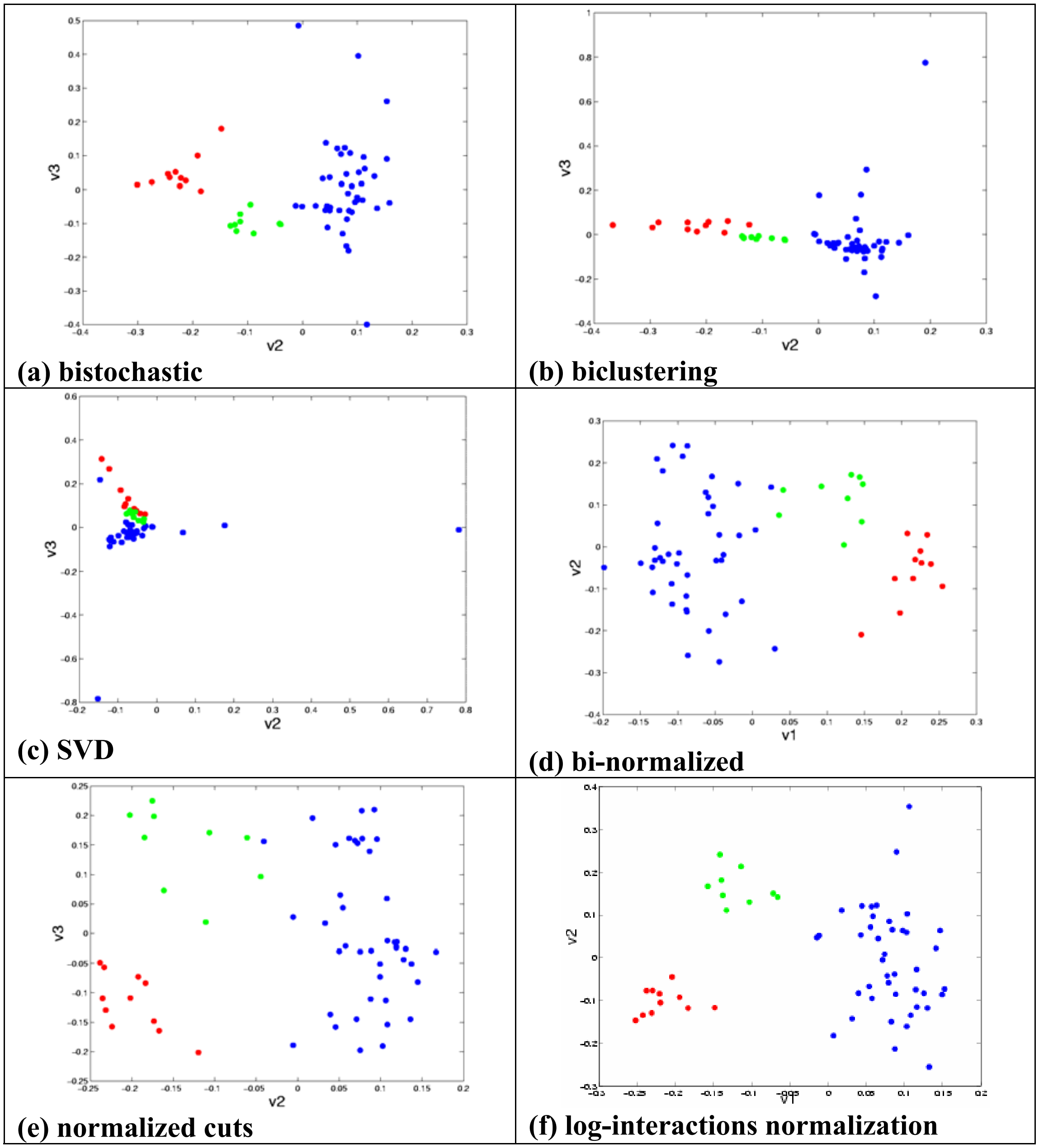


step-like



(a) bistochastic **(b) biclustering**

Figure 2 (a) The outer product of the sorted eigenvectors *u* and *v* of the 2^{nd} eigenvalue of the equal row- and column-sum bistochastic-like matrix *B* applied to dataset with three types of Lymphoma CLL(C), FL(F) and DLCL(D). Sorting of *v* orders the patients according to the different diseases. (b) as in (a) the 2^{nd} singular value contribution to the biclustering method ($C^{-1}A^{T}\ R^{-1}A$) of Lymphoma CLL(C), FL(F), DLCL(D) partitioned the patients according to their disease with one exception. We pre-selected all genes that had complete data along all experimental conditions (samples).



(a) bistochastic

(b) biclustering

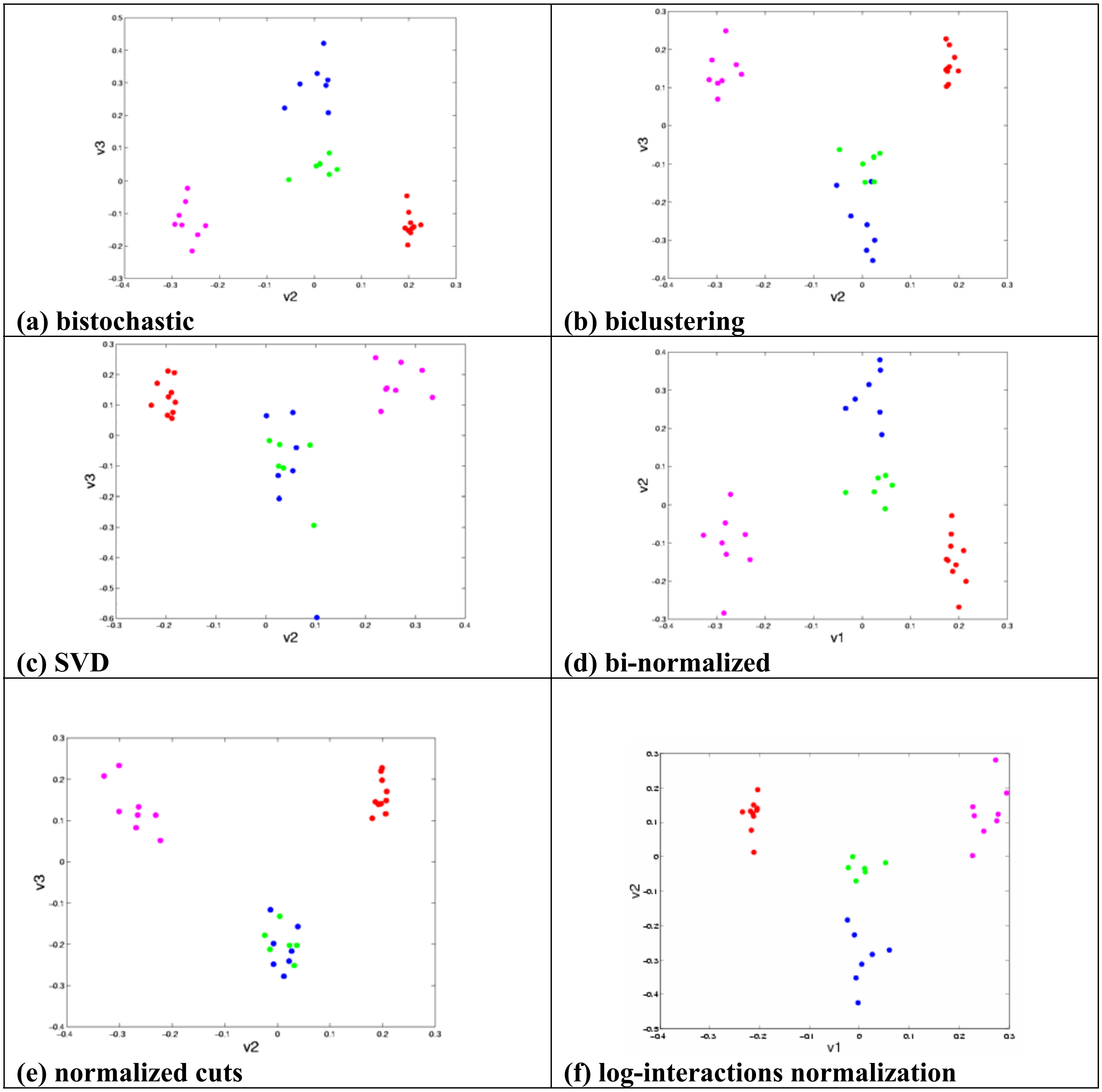
(c) SVD

(d) bi-normalized

(e) normalized cuts

(f) log-interactions normalization

Figure 3 Lymphoma: Scatter plot of experimental conditions of the two best class partitioning eigenvectors v_i, v_j. The subscripts (i,j) of these eigenvectors indicate their corresponding singular values . CLL samples are denoted by red dots, DLCL by blue dots, and FL by green dots. (a) Bistochastization: the 2nd and 3rd eigenvectors of BB^T (b) Biclustering: the 2nd and 3rd eigenvectors of $R^{-1}AC^{-1}A^T$ (c) SVD: the 2nd and 3rd eigenvectors of AA^T (d) normalization and SVD: the 1st and 2nd eigenvectors of $\overline{A}\overline{A}^T$ where $\overline{A}$ is obtained by first dividing each column of A by its mean and then standardizing each row of the column normalized matrix. (e) Normalized cut algorithm: 2nd and 3rd eigenvectors of the row-stochastic matrix P. P is obtained by first creating a distance matrix S using Euclidean distance between the standardized columns of A, transforming it to an affinity matrix with zero diagonal elements and off diagonal elements defined as $W_{ij} = \exp(-\alpha S_{ij}) / \max(S_{ij})$ and finally normalizing each row sum of the affinity matrix to one. (f) as in (c) but a with SVD analysis of the log interaction matrix K instead of A.



(a) bistochastic

(b) biclustering

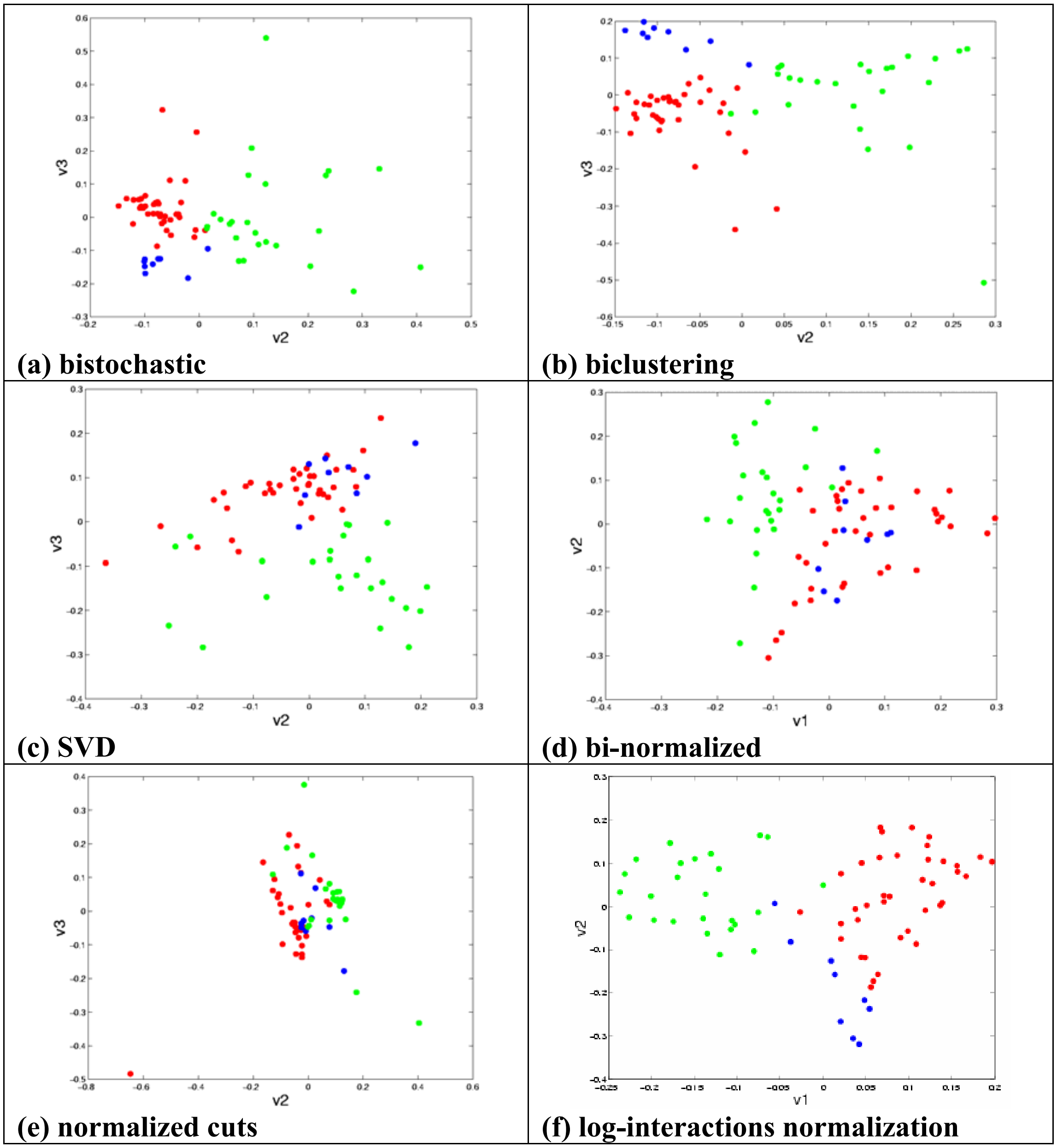
(c) SVD

(d) bi-normalized

(e) normalized cuts

(f) log-interactions normalization

Figure 4 Scatter plots as in Fig. 3 with another Lymphoma dataset generated using Affymetrix chips [9] instead of microarrays. DLCL samples are denoted by green dots, CLL by blue dots, FL by yellow dots and DLCL cell lines by magenta dots.



(a) bistochastic

(b) biclustering

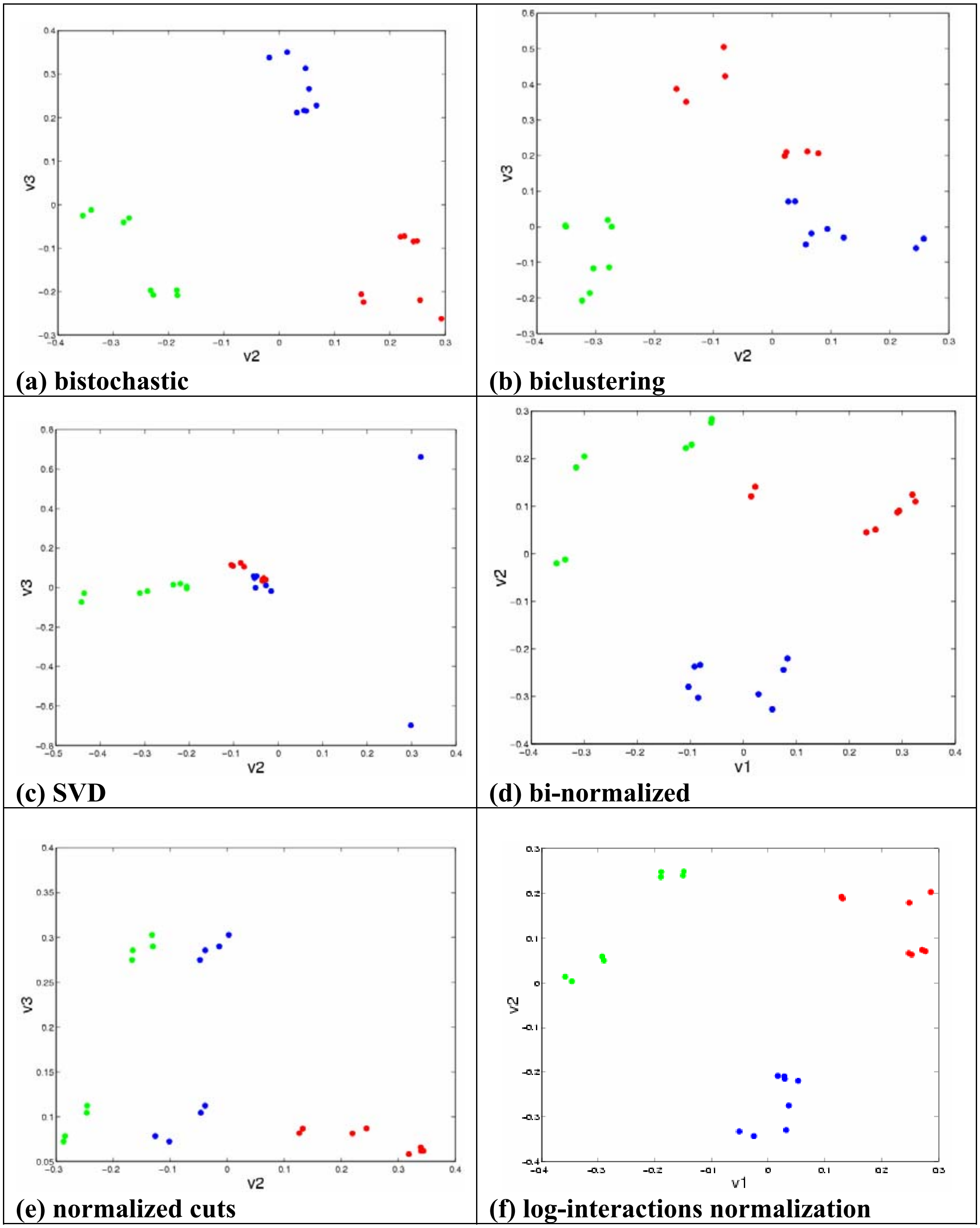
(c) SVD

(d) bi-normalized

(e) normalized cuts

(f) log-interactions normalization

Figure 5 Leukemia data is presented in the same format as in Fig. 3. B cell ALL samples are denoted by red dots, T cell ALL by blue dots, and AML by green dots. In this analysis we pre-selected all genes that had positive Affymetrix average difference expression levels.



(a) bistochastic

(b) biclustering

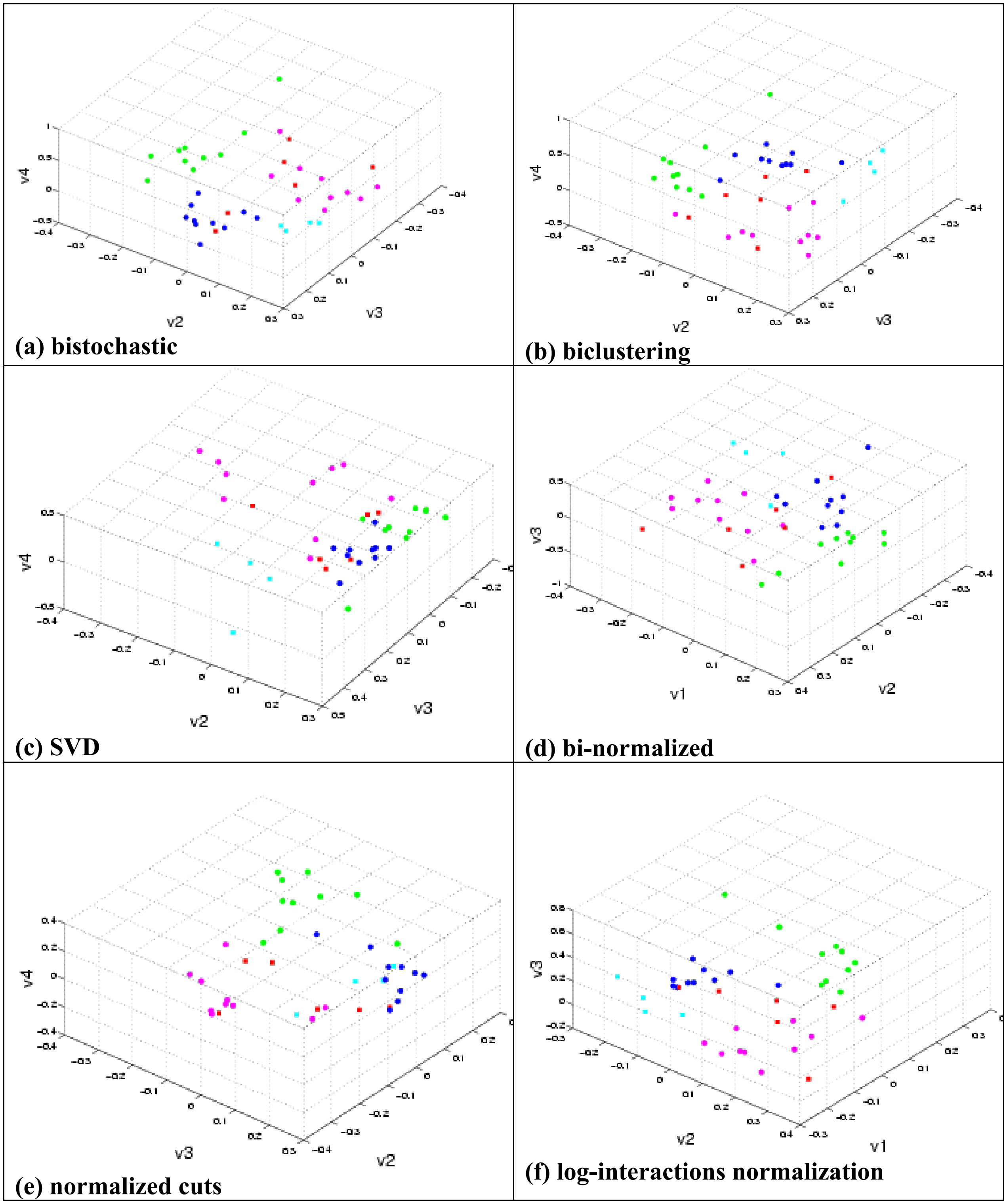
(c) SVD

(d) bi-normalized

(e) normalized cuts

(f) log-interactions normalization

Figure 6 Breast cell lines transfected with the CSF1R oncogene: Scatter plots as in Fig. 3 for mRNA ratios of benign breast cells and wild type cells transfected with the CSF1R oncogene causing them to invade and metastasize (A,a), ratios of cells transfected with a mutated oncogene causing an invasive phenotype and cells transfected with the wild type oncogene (C,c) and ratios of cells transfected with a mutated oncogene causing a metastatic phenotype and cells transfected with the wild type oncogene (D,d).



(a) bistochastic

(b) biclustering

(c) SVD

(d) bi-normalized

(e) normalized cuts

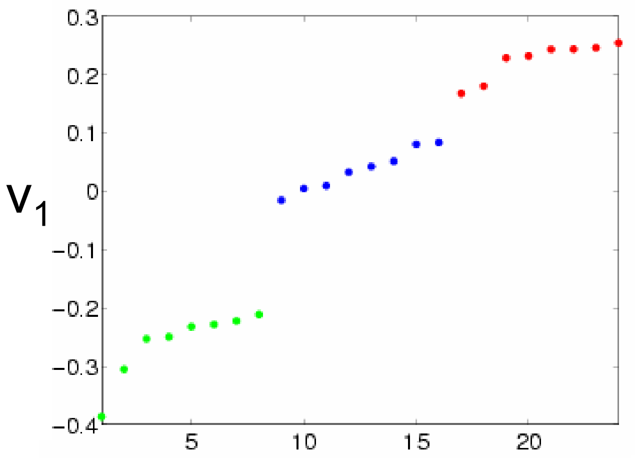
(f) log-interactions normalization

Figure 7 central nervous system embryonal tumor data generated using Affymetrix chips[10] of medulloblastoma (blue), malignant glioma (pink), normal cerebella (cyan), rhabdoid (green) and primitive neuro-ectodermal (red) tumors. Scatter plots of experimental conditions projected onto the three best class partitioning eigenvectors using the same format as in Fig. 3.

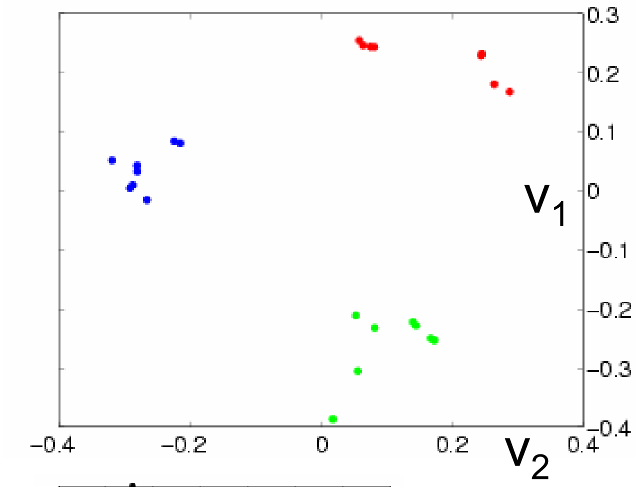
Figure 8 Optimal array partitioning obtained by the 1st singular vectors of the log-interaction matrix. The data consists of eight measurements of mRNA ratios for three pair of cell types: (A,a) benign breast cells and the wild-type cells transfected with the CSF1R oncogene causing them to invade and metastatize; (C,c) cells transfected with a mutated oncogene causing an invasive phenotype and cells transfected with the wild type oncogene; and (D,d) cells transfected with a mutated oncogene causing a metastatic phenotype and cells transfected with the wild type oncogene. In this case we pre-selected differentially expressed genes such that for at least one pair of samples the genes had a three fold ratio. The sorted eigen-gene v_1 and eigen-array u_1 have gaps indicating partitioning of patients and genes respectively. As a result, the outer product matrix sort(u_1) sort(v_1)^T has a "soft" block structure. The block structure is hardly seen when the raw data is sorted but not normalized. However it is more noticeable when the data is both sorted and normalized. Also, shown is the conditions projected onto the first two partitioning eigenvectors u_1 and u_2. Obviously, using the extra dimension gives a clearer separation.

(On next page...)

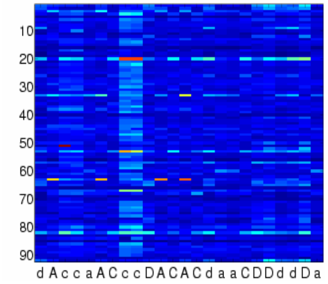
samples projected onto u_1



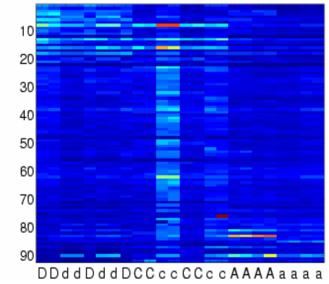
samples projected onto $u_{1,2}$

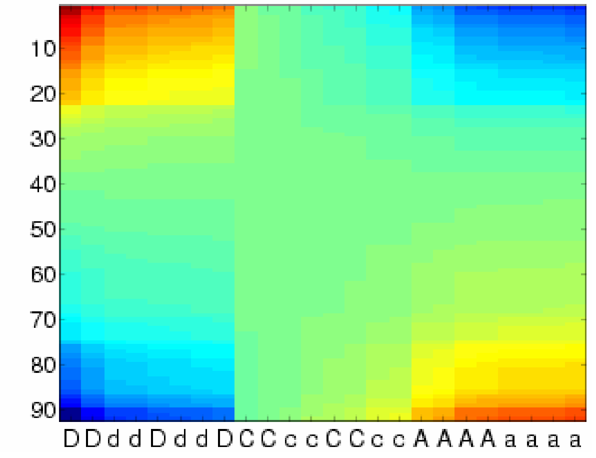


Raw data

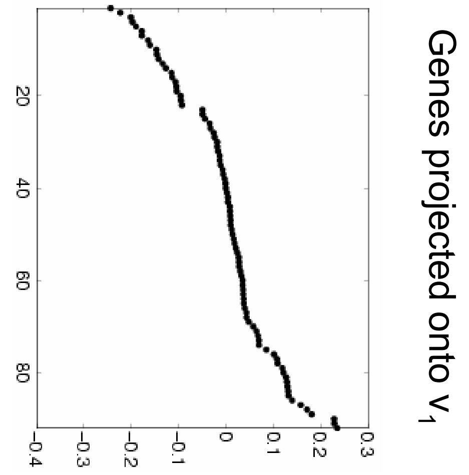


Sorted raw data



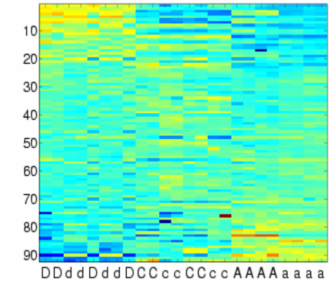


$u_1 v^T_1$



u_1

Sorted & normalized



10
20
30
40
50
60
70
80
90
D D d d D d d D C C c c C C c c A A A A a a a a