Measuring the Quality of Shifting and Scaling Patterns in Biclusters

B. Pontes[1] - R. Giráldez[2] - J. S. Aguilar-Ruiz[2]

[1]Department of Computer Science, University of Seville
bepontes@us.es
[2]School of Engineering, Pablo de Olavide University
{giraldez,aguilar}@upo.es

PRIB 2010

Expression Matrices

- Gene expression data are organized in matrices from microarrays, where rows represent experimental conditions and columns represent genes
- Each element in the matrix refers to the expression level of a particular gene under a specific condition
- Clustering has been applied to expression matrices
- However, relevant genes are not necessarily related to every condition, and the same holds for conditions
- Clustering should then be addressed in two dimensions simultaneously $\Rightarrow$ **biclustering**
- A bicluster is a sub-matrix from a microarray

Clustering and Biclustering

conditions

genes

9	1	10	1	5	7	3	8
0	2	3	9	5	3	2	5
3	2	6	1	8	6	1	2
3	8	3	4	1	5	3	1
4	1	3	2	2	5	2	2
5	7	2	4	2	6	7	2
11	3	2	7	3	8	4	3
4	2	12	5	7	0	4	6
4	2	6	4	2	7	8	2
3	7	3	7	5	2	4	6

Clustering and Biclustering

conditions

genes

9	1	10	1	5	7	3	8
0	2	3	9	5	3	2	5
3	2	6	1	8	6	1	2
3	8	3	4	1	5	3	1
4	1	3	2	2	5	2	2
5	7	2	4	2	6	7	2
11	3	2	7	3	8	4	3
4	2	12	5	7	0	4	6
4	2	6	4	2	7	8	2
3	7	3	7	5	2	4	6

Cluster containing genes $1, 3$ and 5

Clustering and Biclustering

conditions

genes

9	1	10	1	5	7	3	8
0	2	3	9	5	3	2	5
3	2	6	1	8	6	1	2
3	8	3	4	1	5	3	1
4	1	3	2	2	5	2	2
5	7	2	4	2	6	7	2
11	3	2	7	3	8	4	3
4	2	12	5	7	0	4	6
4	2	6	4	2	7	8	2
3	7	3	7	5	2	4	6

Bicluster contains rows $1, 3$ and 5 and columns $2, 4$ and 7.

Clustering and Biclustering

conditions

genes

9	1	10	1	5	7	3	8
0	2	3	9	5	3	2	5
3	2	6	1	8	6	1	2
3	8	3	4	1	5	3	1
4	1	3	2	2	5	2	2
5	7	2	4	2	6	7	2
11	3	2	7	3	8	4	3
4	2	12	5	7	0	4	6
4	2	6	4	2	7	8	2
3	7	3	7	5	2	4	6

Bicluster containing genes $1, 3$ and 5, for the experimental conditions $2, 4$ and 7.
Gen 5 overlapped.

Types of Biclusters

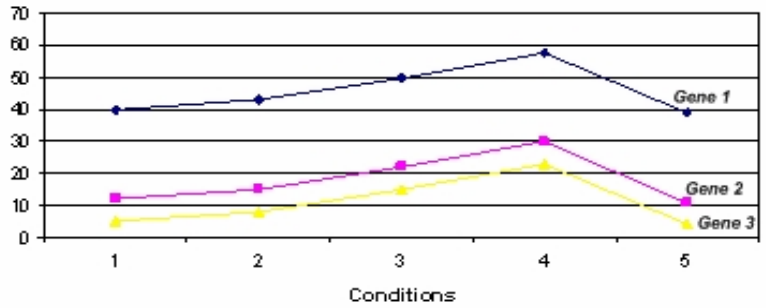
- **Constant values:** $b_{ij} = \pi$
- **Constant values on rows or columns**
  - Additive: $b_{ij} = \pi + \beta_i$, $b_{ij} = \pi + \beta_j$
  - Multiplicative: $b_{ij} = \pi \times \alpha_i$, $b_{ij} = \pi \times \alpha_j$
- **Coherent values on both rows and columns**
  - Additive: $b_{ij} = \pi + \beta_i + \beta_j$
  - Multiplicative: $b_{ij} = \pi \times \alpha_i \times \alpha_j$
- **Coherent evolutions:** Evidence that a subset of genes is up-regulated or down-regulated across a subset of conditions without taking into account their actual expression values

Behavioural patterns for gene expression data

Shifting: $b_{ij} = \pi_j + \beta_i$

Shifting Pattern

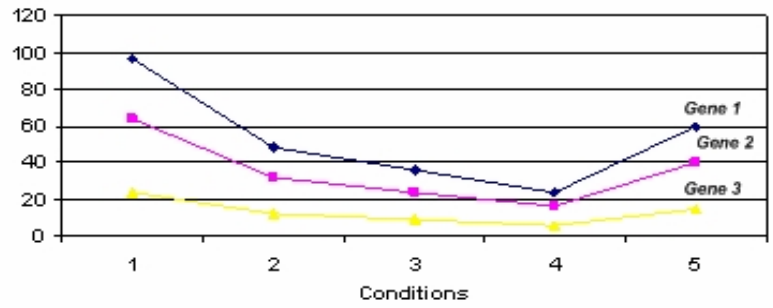
$$B = \begin{pmatrix} 40 & 12 & 5 \\ 43 & 15 & 8 \\ 50 & 22 & 15 \\ 58 & 30 & 23 \\ 39 & 11 & 4 \end{pmatrix} = \begin{pmatrix} 40+0 & 12+0 & 5+0 \\ 40+3 & 12+3 & 5+3 \\ 40+10 & 12+10 & 5+10 \\ 40+18 & 12+18 & 5+18 \\ 40-1 & 12-1 & 5-1 \end{pmatrix}$$



Scaling: $b_{ij} = \pi_j \times \alpha_i$

Scaling Pattern

$$B = \begin{pmatrix} 96 & 64 & 24 \\ 48 & 32 & 12 \\ 36 & 24 & 9 \\ 24 & 16 & 6 \\ 60 & 40 & 15 \end{pmatrix} = \begin{pmatrix} 12\times 8 & 8\times 8 & 3\times 8 \\ 12\times 4 & 8\times 4 & 3\times 4 \\ 12\times 3 & 8\times 3 & 3\times 3 \\ 12\times 2 & 8\times 2 & 3\times 2 \\ 12\times 5 & 8\times 5 & 3\times 5 \end{pmatrix}$$



Combined: $b_{ij} = \pi_j \times \alpha_i + \beta_i$

Cheng & Church's Mean Squared Residue

$$MSR(B) = \frac{1}{I \cdot J} \sum_{i=1}^{i=I} \sum_{j=1}^{j=J} (b_{ij} - b_{iJ} - b_{Ij} + b_{IJ})^2$$

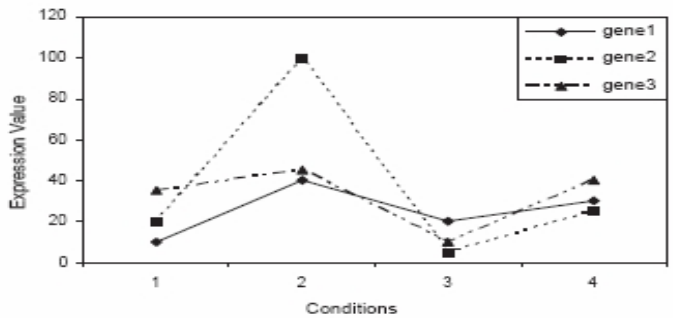
Cheng & Church's Mean Squared Residue

$$MSR(B) = \frac{1}{I \cdot J} \sum_{i=1}^{i=I} \sum_{j=1}^{j=J} (b_{ij} - b_{iJ} - b_{Ij} + b_{IJ})^2$$

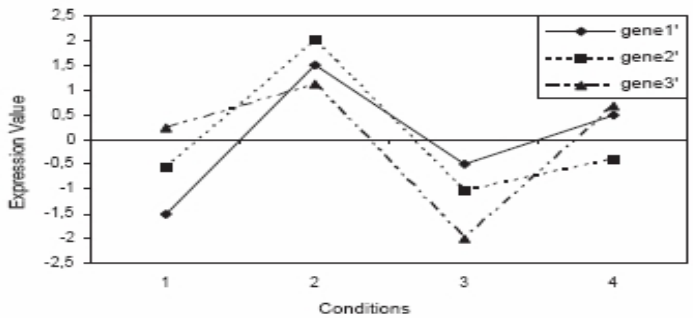
- The lower the MSR value the better the bicluster is
- Need of a γ threshold value for each dataset
- A shifting pattern has no effect on MSR. Biclusters with perfect shifting patterns have MSR equal to zero.
- A scaling pattern has significant effect on MSR.
- Using MSR for searching bicluster we might miss may high-quality biclusters
- The majority of the current biclustering algorithms use MSR for guiding the search

Virtual Error: Identifying shifting or scaling patterns

- Virtual Gene: Construct a new artificial gene consisting of the mean values of all the genes expression levels in the bicluster, thus capturing the general trend
- Standardization: Construct a new submatrix with the standardized elements of the original bicluster
  - All values are re-scaled to a similar range of values
  - The expression values for each gene are smoothed
  - The virtual gene is also standardized



(a) Bicluster B



(b) Standardized Bicluster B'

- Virtual Error: Quantify the numerical differences among the genes in the standardized bicluster and the standardized virtual gene

Transposed Virtual Error: Identifying all types of patterns

How to compute VE^t

1. Compute VE for the transposed bicluster
2. Perform the same process over the other dimension

Transposed Virtual Error: Identifying all types of patterns

How to compute VE^t

1. Compute VE for the transposed bicluster
2. Perform the same process over the other dimension

- Virtual Condition: Construct a new artificial condition
- Standardization: Construct a new submatrix, standardizing with regard to the conditions
  - The virtual condition is also standardized
- Transposed Virtual Error: Quantify the numerical differences among the conditions in the standardized bicluster and the standardized virtual condition

Analysis

We have analytically proved that a bicluster presenting either perfect shifting, scaling o combined pattern has VE^t equals to zero

Proof for the perfect combined pattern:

- Mean and deviation for each condition

$$\mu_{c_i} = \alpha_i \times \mu_\pi + \beta_i \quad ; \quad \sigma_{c_i} = \alpha_i \times \sigma_\pi$$

- Standardized values for b_{ij}

$$\hat{b}_{ij} = \frac{b_{ij} - \mu_{c_i}}{\sigma_{c_i}} = \frac{\pi_j \times \alpha_i + \beta_i - \alpha_i \times \mu_\pi + \beta_i}{\alpha_i \times \sigma_\pi} = \frac{\pi_j - \mu_\pi}{\sigma_\pi}$$

- Mean and deviation for the virtual condition

$$\mu_\rho = \mu_\pi \times \mu_\alpha + \mu_\beta \quad ; \quad \sigma_\rho = \mu_\alpha \times \sigma_\pi$$

- Standardized values for the virtual condition

$$\hat{\rho}_j = \frac{\rho_j - \mu_\rho}{\sigma_\rho} = \frac{\pi_j \times \mu_\alpha + \mu_\beta - \mu_\pi \times \mu_\alpha - \mu_\beta}{\mu_\alpha \times \sigma_\pi} = \frac{\pi_j - \mu_\pi}{\sigma_\pi}$$

VE^t for biclusters without perfect patterns

Equation for not perfect biclusters:

$$b_{ij} = \pi_j \times \alpha_i + \beta_i + \varepsilon_{ij}$$

VE^t for biclusters without perfect patterns

Equation for not perfect biclusters:

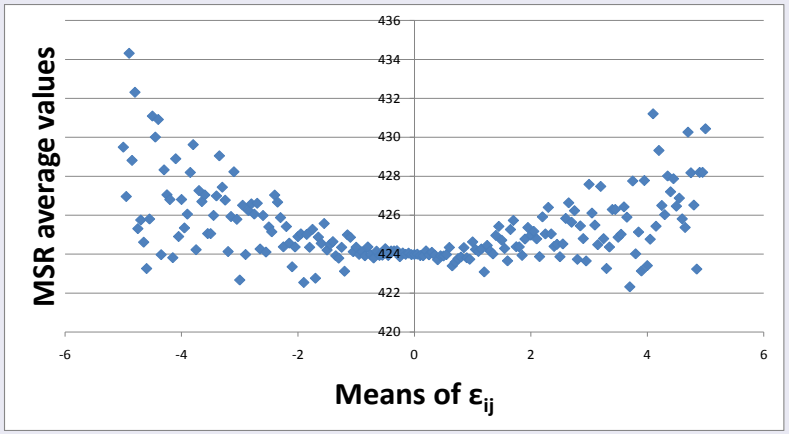
$$b_{ij} = \pi_j \times \alpha_i + \beta_i + \varepsilon_{ij}$$

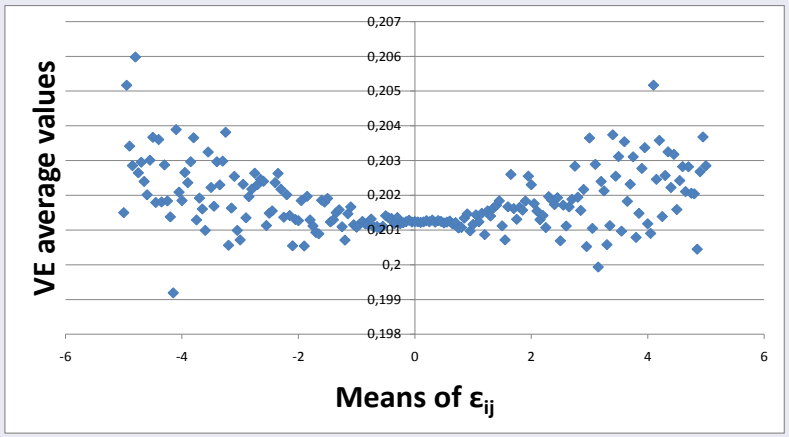
Experimental test to check the tendency of VE^t

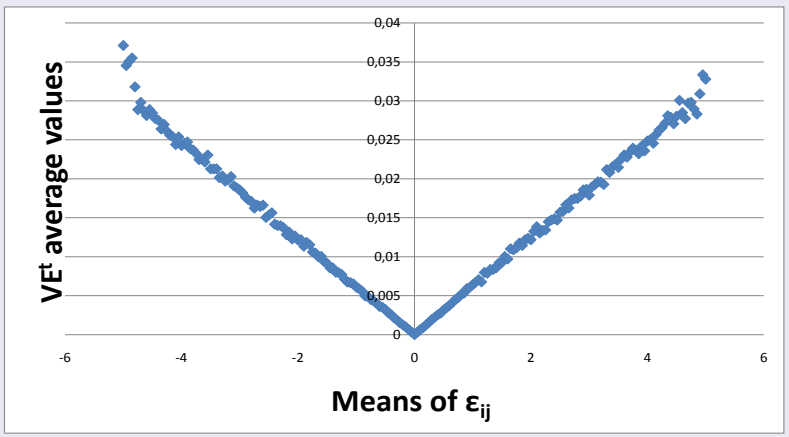
- Base bicluster with perfect combined pattern
- 100 synthetic biclusters adding random errors to the base one
- Repeat the process 200 times, with different amplitude of errors
- Addition of positive and negative errors
- Errors with uniform and normal distributions
- Report the mean of the MSR, VE and VE^t values of each group of 100 biclusters

Evaluation Measures' Behaviour(I)

- Mean of MSR, VE and VE^t values for positive and negative errors
- Uniform distribution

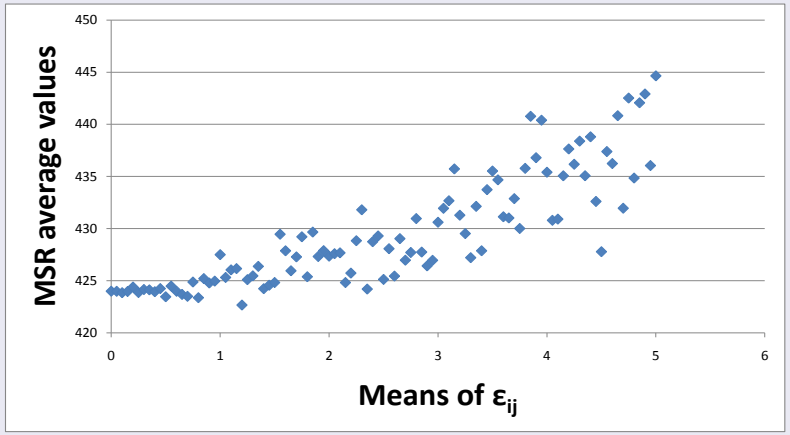


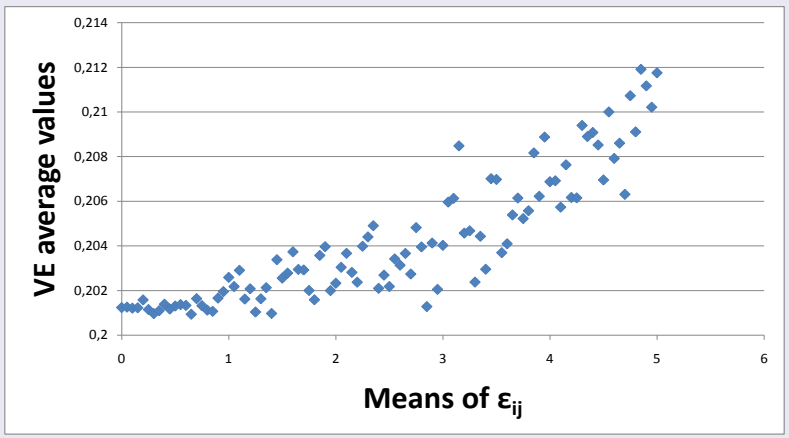


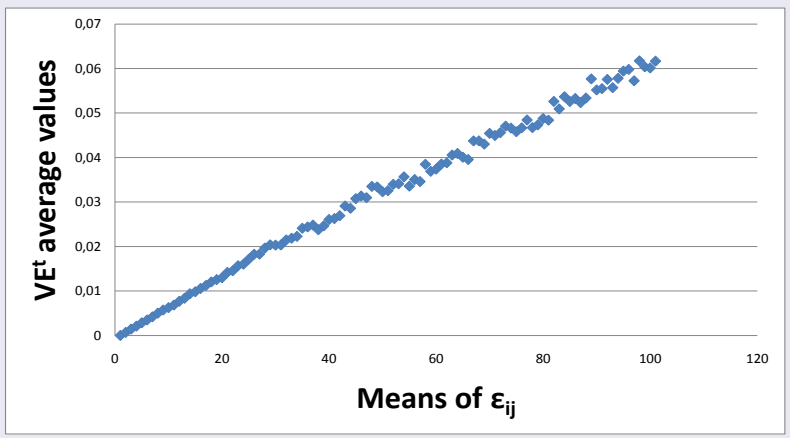


Evaluation Measures' Behaviour(II)

- Mean of MSR, VE and VE^t values for positive and negative errors simultaneously
- Uniform distribution

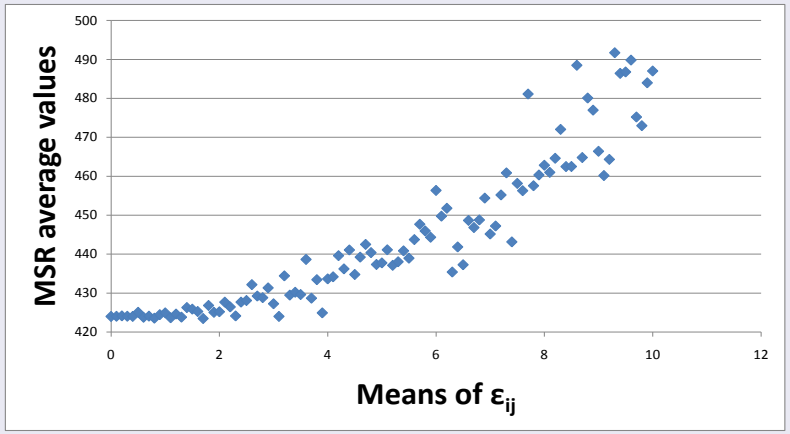


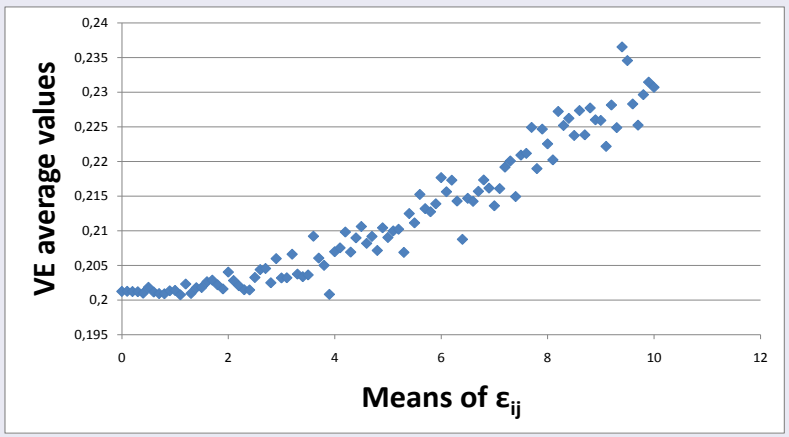


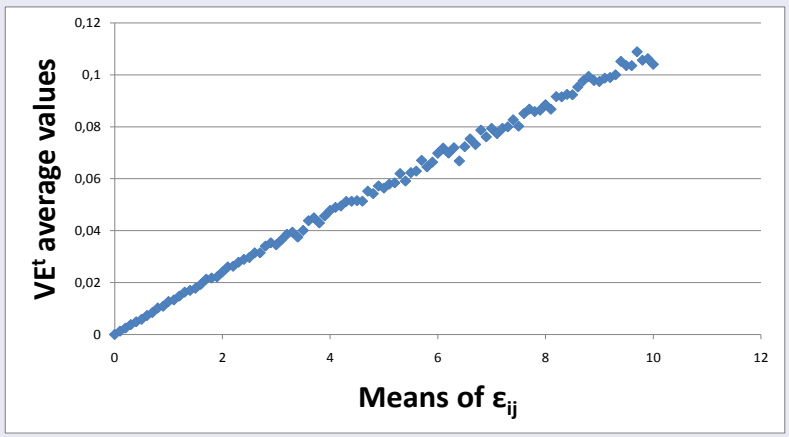


Evaluation Measures' Behaviour(III)

- Mean of MSR, VE and VE^t values for positive and negative errors simultaneously
- Normal distribution







Conclusions

1. Biologically, the interest resides in variable behaviours, represented by shifting and scaling patterns
2. Scaling behaviour is more probable in nature (e.g. Regulatory pathways)
3. MSR is highly affected by scaling behaviours
4. VE is able to recognize both shifting and scaling behaviours, but not simultaneously
5. VE^t can detect combined patterns
6. VE^t shows a linear tendency with regard to errors

Future Work

1. Include VE^t in several heuristics for the biclustering problem
2. Perform experimental tests with synthetic data
3. Make comparisons with other evaluation measures and heuristics
4. Experiments with real gene expression data
5. Biological validation of the results

Measuring the Quality of Shifting and Scaling Patterns in Biclusters

Beatriz Pontes
bepontes@us.es

THANK YOU!