# Data Transformation for Sum Squared Residue

Hyuk Cho

Computer Science, Sam Houston State University,
Huntsville, TX 77341-2090, USA
hyukcho@shsu.edu
http://www.cs.shsu.edu/~hxc005/

**Abstract.** The sum squared residue has been popularly used as a clustering and co-clustering quality measure, however little research on its detail properties has been performed. Recent research articulates that the residue is useful to discover shifting patterns but inappropriate to find scaling patterns. To remedy this weakness, we propose to take specific data transformations that can adjust latent scaling factors and eventually lead to lower the residue. First, we consider data matrix models with varied shifting and scaling factors. Then, we formally analyze the effect of several data transformations on the residue. Finally, we empirically validate the analysis with publicly-available human cancer gene expression datasets. Both the analytical and experimental results reveal column standard deviation normalization and column Z-score transformation are effective for the residue to handle scaling factors, through which we are able to achieve better tissue sample clustering accuracy.

**Keywords:** Data Transformation, Sum Squared Residue, Z-score Transformation, Scaling Pattern, Shifting Pattern.

## 1 Introduction

Hartigan's pioneering work, *direct clustering* [13], stimulated a vast amount of research on co-clustering. Co-clustering aims at identifying homogeneous local patterns, each of which consists of a subset of rows and a subset of of columns in a given two dimensional matrix. This idea has attracted genomic researchers, because it is compatible with our understanding of cellular processes, where a subset of genes are coregulated under a certain experimental conditions [5]. Madeira and Oliveira [15] surveyed biclustering algorithms and their applications to biological data analysis.

Cheng and Church [7] are considered to be the first to apply co-clustering, *biclustering*, to gene expression data. The greedy search heuristic generates biclusters, one at a time, which satisfy a certain homogeneity constraint, called *mean squared residue*. Since then, several similar approaches have been proposed to enhance the original work. For example, Cho *et al.* [8] developed two minimum sum squared co-clustering (MSSRCC) algorithms: one objective function is based on the partitioning model proposed by Hartigan [13] and the other one

M.J. Zaki et al. (Eds.): PAKDD 2010, Part I, LNAI 6118, pp. 48–55, 2010.

is based on the squared residue formulated by Cheng and Church [7]. The later is the residue of interest and defined in the next chapter.

Recently, Aguilar-Ruiz [1] shows that the mean squared residue depends on the scaling variance in the considered data matrix. This finding issues the weakness of the residue and the need of new approaches to discover scaling patterns. Motivated by the work, we propose a simple remedy to find scaling patterns, still using the same residue measure. We suggest to take specific data transformations so as to handle hidden scaling factors. In this paper, we apply several data transformations to data matrix models derived from varied scaling and shifting factors and analyze the effect of data transformations on the the second residue, RESIDUE(II) [8]. Furthermore, using MSSRCC, we empirically demonstrate the advantage of the data transformations with publicly available human cancer microarrays. Both analysis and experimental results reveal that column standard deviation normalization and column Z-score transformation are effective.

The rest of this paper is organized as follows: In Section 2 we introduce some definitions and facts used in this paper. We describe the considered data transformations in Section 3. Then, we formally analyze the effects of data transformations and summarize the analysis results in Section 4. We discuss the experimental results with human cancer gene expression datasets in Section 5. Finally, the paper is concluded with some remark.

## 2 Definition

We adapt the following definitions in Agular-Ruiz [1], Cheng and Church [7], and Cho *et al.* [8] to fit for our context.

**Data matrix.** *A data matrix is defined as a real-valued rectangular matrix* $\boldsymbol{A} \in \mathbb{R}^{m \times n}$*, whose* $(i, j)$*-th element is denoted by* $a_{ij}$*.*

For example, a microarray can be defined with two finite sets, the set of genes and the set of experimental conditions. Note that Aguilar-Ruiz [1] describes the microarray whose rows represent experimental condition and columns represent genes. However, in this paper, we will consider a microarray which consists of examples of genes in rows and attributes as experimental conditions in columns.

**Co-cluster.** *Let* $I \subseteq \{1, 2, \ldots, m\}$ *and* $J \subseteq \{1, 2, \ldots, n\}$ *denote the set of indices of the rows in a row cluster and the set of indices of the columns in a column cluster. A submatrix of* $\boldsymbol{A}$ *induced by the index sets* $I$ *and* $J$ *is called a co-cluster and denoted as* $\boldsymbol{A}_{IJ} \in \mathbb{R}^{|I| \times |J|}$*, where* $|I|$ *and* $|J|$ *denote the cardinality of index set* $I$ *and* $J$*, respectively.* In reality, rows and columns in a co-cluster are not necessary to be consecutive. However, for brevity we consider the co-cluster, $\boldsymbol{A}_{IJ}$, whose entries consist of first $|I|$ rows and first $|J|$ columns in $\boldsymbol{A}$.

### 2.1 Sum Squared Residue

In order to evaluate the coherence of such a co-cluster, we define RESIDUE(II) of an element $a_{ij}$ in the co-cluster determined by index sets $I$ and $J$ as below.

**Residue.** *RESIDUE(II) is defined as $h_{ij} = a_{ij} - a_{iJ} - a_{Ij} + a_{IJ}$, where the mean of the entries in row $i$ whose column indices are in $J$ is computed by $a_{iJ} = \frac{1}{|J|}\sum_{j \in J} a_{ij}$, the mean of the entries in column $j$ whose row indices are in $I$ by $a_{Ij} = \frac{1}{|I|}\sum_{i \in I} a_{ij}$, and the mean of all the entries in the co-cluster whose row and column indices are in $I$ and $J$ by $a_{IJ} = \frac{1}{|I||J|}\sum_{i \in I, j \in J} a_{ij}$.*

**Sum squared residue (SSR).** *Let $\boldsymbol{H}_{IJ} \in \mathbb{R}^{|I| \times |J|}$ be the residue matrix whose entries are described by RESIDUE(II). Then, the sum squared residue of $\boldsymbol{H}_{IJ}$ is defined as $SSR = \|\boldsymbol{H}_{IJ}\|^2 = \sum_{i \in I, j \in J} h_{ij}^2$, where $\|\boldsymbol{X}\|$ denotes the Frobenius norm of matrix $\boldsymbol{X}$, i.e., $\|\boldsymbol{X}\|^2 = \sum_{i,j} x_{ij}^2$.*

### 2.2 Patterns

We assume $\boldsymbol{A}$ contains both scaling and shifting factors. We borrow the concepts of "local" and "global" scaling and shifting from Cheng and Church [7], Cho *et al.* [8], and Aguilar-Ruiz [1] and generalize the definition of data patterns in [1].

**Global/local scaling and global/local shifting patterns.** *A bicluster contains both scaling and shifting patterns when it expresses $a_{ij} = \pi_i \times \alpha_j + \beta_j$, where $\pi_i$ is the base value for row (e.g., gene) $i$, $\alpha_j$ is the scaling factor for column (e.g., experimental condition) $j$, and $\beta_j$ is the shifting factor for column (e.g., experimental condition) $j$. We classify the expression into the following four patterns: global scaling (gsc) and global shifting pattern (gsh) when $a_{ij} = \pi_i \times \alpha + \beta$; global scaling (gsc) and local shifting pattern (lsh) when $a_{ij} = \pi_i \times \alpha + \beta_j$; local scaling (lsc) and global shifting pattern (gsh) when $a_{ij} = \pi_i \times \alpha_j + \beta$; and local scaling (lsc) and local shifting pattern (lsh) when $a_{ij} = \pi_i \times \alpha_j + \beta_j$.*

## 3 Data Transformations

Raw data values have a limitation that raw values do not disclose how they vary from the central tendency of the distribution. Therefore, transformation of the raw data is considered one of the most important steps for various data mining processes since the variance of a variable will determine its importance in a given model [16]. In this study, we investigate the following data transformations.

**No transformation (NT).** No centering or scaling is taken. In other words, $a'_{ij} = a_{ij}$, $\forall i$ and $\forall j$, *i.e.*, the raw matrix is directly input to MSSRCC.

**Double centering (DC).** DC is defined as $a'_{ij} = a_{ij} - a_{i\cdot} - a_{\cdot j} + a_{\cdot\cdot}$, $\forall i$ and $\forall j$. Through DC, each entry of a data matrix $\boldsymbol{A}$ becomes $a'_{ij} = (\pi_i - \mu_\pi)(\alpha_j - \mu_\alpha)$. Note that we have $a'_{i\cdot} = a'_{\cdot j} = 0$ and consequently $a'_{\cdot\cdot} = 0$, since DC transforms the data matrix to have both row means and column means to be 0.

**Column/row mean centering (MC).** Column MC is defined as $a'_{ij} = a_{ij} - a_{\cdot j}$, $\forall i$ and $\forall j$. Through column MC, each entry becomes $a'_{ij} = \pi_i \alpha_j + \beta_j - \mu_{\cdot j}$. Therefore, row mean, column mean, and whole mean become $a'_{i\cdot} = \pi_i \mu_\alpha + \mu_\beta - a_{\cdot\cdot}$, $a'_{\cdot j} = \mu_\pi \alpha_j + \beta_j - a_{\cdot j}$, and $a'_{\cdot\cdot} = \mu_\pi \mu_\alpha + \mu_\beta - a_{\cdot\cdot}$, respectively. Similarly, row MC is defined similarly with $a_{i\cdot}$. Therefore, row mean, column mean, and whole mean become $a'_{i\cdot} = \pi_i \mu_\alpha + \mu_\beta - a_{i\cdot}$, $a'_{\cdot j} = \mu_\pi \alpha_j + \beta_j - a_{\cdot\cdot}$, and $a'_{\cdot\cdot} = \mu_\pi \mu_\alpha + \mu_\beta - a_{\cdot\cdot}$.

**Column/row standard deviation normalization (SDN).** Column SDN is defined as $a'_{ij} = \frac{a_{ij}}{\sigma_{\cdot j}}$, $\forall i$ and $\forall j$. Similarly, row SDN is defined with $\sigma^2_{i\cdot}$. Through column and row SDN each column and row has a unit variance, respectively.
**Column/row Z-score transformation (ZT).** Column ZT is defined as $a'_{ij} = \frac{a_{ij} - a_{\cdot j}}{\sigma_{\cdot j}}$, $\forall i$ and $\forall j$. Similarly, row ZT is defined with $a_{i\cdot}$ and $\sigma^2_{i\cdot}$. It is also called "autoscaling", where the measurements are scaled so that each column/row has a zero mean and a unit variance [14]. Through ZT, the relative variation in intensity is emphasized, since ZT is a linear transformation, which keeps the relative positions of observations and the shape of the original distribution.

## 4 Analysis

Now, we analyze the effect of the data transformations on the sum squared residue, RESIDUE(II). Because of space limitation, we focus on analysis on the three data transformations including NT, column SDN, and column ZT, which clearly demonstrate the effect of the specific data transformation.

### 4.1 No Transformation (NT)

$(i, j)$-th entry of row $i \in I$ and column $j \in J$ of co-cluster $\boldsymbol{A}_{IJ}$ is described as $a_{ij} = \pi_i \alpha_j + \beta_j$. Then, the mean of the base values of $\boldsymbol{A}_{IJ}$ is computed by $\mu_{\pi_I} = \frac{1}{|I|} \sum_{i \in I} \pi_i$. and the mean of the scaling factors by $\mu_{\alpha_J} = \frac{1}{|J|} \sum_{j \in J} \alpha_j$, and the mean of the shifting factors by $\mu_{\beta_J} = \frac{1}{|J|} \sum_{j \in J} \beta_j$. Also, the mean of row $i$ is obtained by $a_{iJ} = \pi_i \mu_{\alpha_J} + \mu_{\beta_J}$, the mean of column $j$ by $a_{Ij} = \mu_\pi \alpha_j + \beta_j$, and the mean of all the elements by $a_{IJ} = \mu_\pi \mu_{\alpha_J} + \mu_{\beta_J}$. Using these values, we obtain RESIDUE(II), $h_{ij} = (\pi_i - \mu_{\pi_I})(\alpha_j - \mu_{\alpha_J})$. Consequently, the sum squared residue (SSR) can be computed as $SSR = \|\boldsymbol{H}_{IJ}\|^2 = \sum_{i \in I, j \in J} h^2_{ij} = \sum_{i \in I, j \in J} (\pi_i - \mu_{\pi_I})^2 (\alpha_j - \mu_{\alpha_J})^2 = |I||J| \sigma^2_{\pi_I} \sigma^2_{\alpha_J}$, where $\sigma^2_{\pi_I} = \frac{1}{|I|} \sum_{i \in I} (\pi_i - \mu_{\pi_I})^2$ and $\sigma^2_{\alpha_J} = \frac{1}{|J|} \sum_{j \in J} (\alpha_j - \mu_{\alpha_J})^2$.

In fact, SSR shown above is a revisit of Theorems in Aguilar-Ruiz [1]. It shows with no data transformation that SSR is dependent on both the variance of base values and the variance of scaling factors, but independent from shifting factors. Accordingly, any shifting operations such as DC and MC to the given data matrix should not contribute to RESIDUE(II). As also shown in [1], RESIDUE(II) itself has an ability to discover shifting patterns.

### 4.2 Column Standard Deviation Normalization (SDN)

Column SDN transforms $\boldsymbol{A}$ to have the constant global scaling factor, *i.e.*, 1, and the local shifting factors, *i.e.*, $\frac{\beta_j}{\alpha_j}$. To be more specific, $(i, j)$-th entry is transformed as $a_{ij} = \frac{1}{\sigma_{\cdot j}} (\pi_i \alpha_j + \beta_j) = \frac{1}{\sigma_\pi} \left( \pi_i + \frac{\beta_j}{\alpha_j} \right)$. Then, row mean, column mean, and whole mean of co-cluster $\boldsymbol{A}_{IJ}$ are computed by $a_{iJ} = \frac{1}{|J|}\sum_{j\in J}\frac{1}{\sigma_{\cdot j}}(\pi_i\alpha_j+\beta_j)$, $a_{Ij} = \frac{1}{|I|}\sum_{i\in I}\frac{1}{\sigma_{\cdot j}}(\pi_i\alpha_j+\beta_j)$, and $a_{IJ} = \frac{1}{|I||J|}\sum_{i\in I}\sum_{j\in J}\frac{1}{\sigma_{\cdot j}}(\pi_i\alpha_j+\beta_j)$, respectively. Therefore, using RESIDUE(II), we can capture the perfect co-cluster, *i.e.*, zero RESIDUE(II), for all the four expression patterns.

### 4.3 Column Z-Score Transformation (ZT)

Column ZT transforms $\boldsymbol{A}$ to have the constant global scaling factor, *i.e.*, 1, and the constant global shifting factor, *i.e.*, $-\mu_\pi$. To be more specific, $(i,j)$-th entry is transformed as $a_{ij} = \frac{1}{\sigma_{\cdot j}}(\pi_i\alpha_j+\beta_j-a_{\cdot j}) = \frac{1}{\sigma_\pi}(\pi_i-\mu_\pi)$. Then, row mean of co-cluster $\boldsymbol{A}_{IJ}$ is obtained by $a_{iJ} = \frac{1}{\sigma_\pi}(\mu_{\pi_i}-\mu_\pi) = a_{ij}$, and column mean and whole mean by $a_{Ij} = \frac{1}{\sigma_\pi}(\mu_{\pi_I}-\mu_\pi) = a_{IJ}$. Like column SDN, we obtain zero REDIDUE(II) for all the possible combinations of scaling and shifting patterns.

## 5 Experimental Results

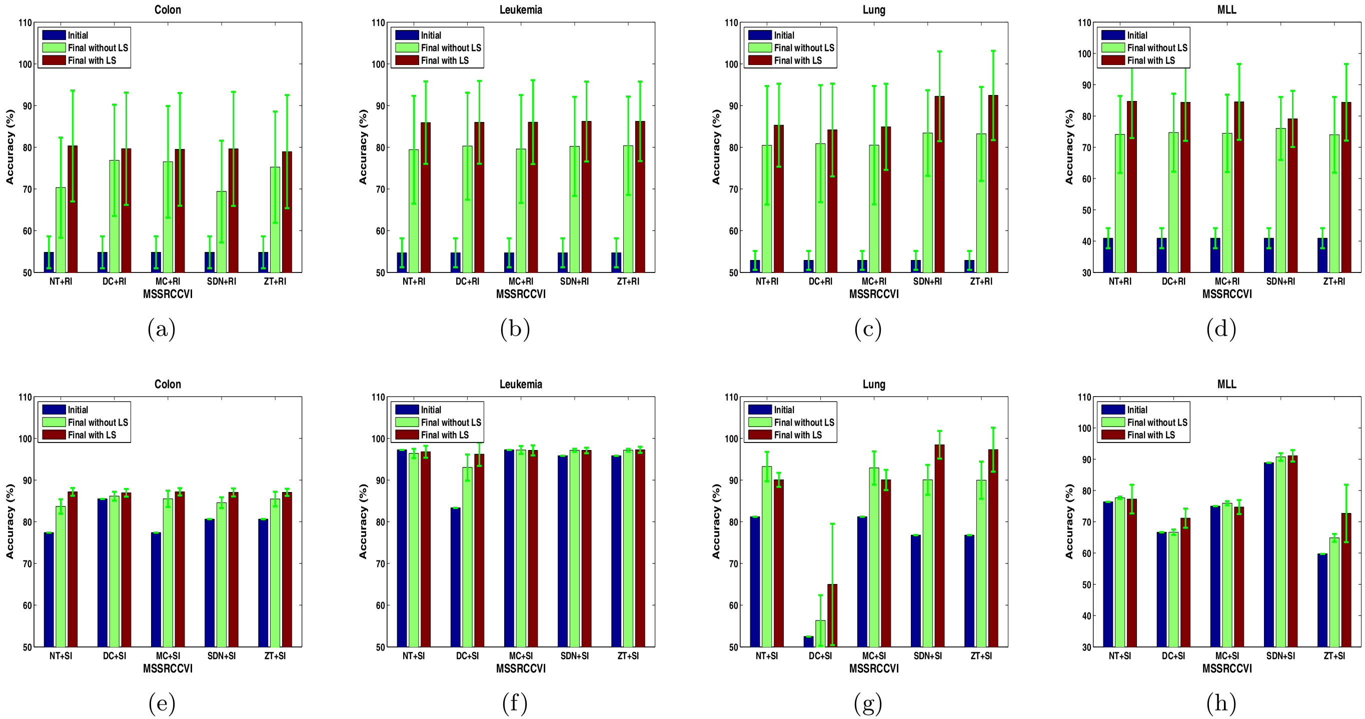
Now, we empirically show the effect of data transformations on the four publicly available human cancer microarray datasets including Colon cancer [2], Leukemia [12], Lung cancer [3], and MLL [3]. With MSSRCC [8][9], we generate $100 \times 2$ or $100 \times 3$ co-clusters with random and spectral initializations, setting $\tau = 10^{-3}\|\boldsymbol{A}\|^2$ and $\tau = 10^{-6}\|\boldsymbol{A}\|^2$ for batch and local search, respectively. Detailed algorithmic strategies and their contributions are discussed in [9].

**Data preprocessing.** Since utilizing sophisticated feature selection algorithms is not a main focus, we just apply the simple preprocessing steps usually adopted in microarray experiments as in [6][10][11] to detect differential expression. Details are summarized in Table 1. Further, the gene expression values in Colon dataset were transformed by taking the base-10 logarithm.

**Table 1.** Description of microarray datasets used in our experiments

| | Colon | Leukemia | Lung | MLL |
|---|---|---|---|---|
| # original genes | 2000 | 7129 | 12533 | 12582 |
| # samples | 62 | 72 | 181 | 72 |
| # sample classes | 2 | 2 | 2 | 3 |
| | Normal(20) | ALL(47) | ADCA(150) | ALL(24) |
| Sample class names | Tumor(42) | AML(25) | MPM(31) | AML(25) |
| | | | | MLL(23) |
| $\lvert max/min \rvert$ | 15 | 5 | 5 | 5 |
| $\lvert max - min \rvert$ | 500 | 500 | 600 | 5500 |
| # remaining genes | 1096 | 3571 | 2401 | 2474 |

Abbreviations: ALL – Acute Lymphoblastic Leukemia; AML – Acute Myeloid Leukemia; ADCA – Adenocarcinoma; MPM – Malignant Pleural Mesothelioma; and MLL – Mixed-Lineage Leukemia. The number after each sample class name denotes the number of samples.



(a) (b) (c) (d)

(e) (f) (g) (h)

**Fig. 1.** Average tissue sample clustering accuracy using MSSRCC with RESIDUE(II). The accuracy values are averaged over 1 to 100 gene clusters. (a)-(d) are averaged over 10 random runs and (e)-(h) are obtained with deterministic spectral initialization. Abbreviations: RI – Random Initialization; SI – Spectral Initialization; NT – No Transformation; DC – Double Centering; MC – (column) Mean Centering; SDN – (column) Standard Deviation Normalization; ZT – (column) Z-score Transformation; and LS – Local Search.

**Tissue sample clustering evaluation measure.** To evaluate the performance of sample clusterings, we quantify tissue sample clustering performance using the following clustering accuracy measure: $accuracy(\%) = \frac{1}{T}\left(\sum_{i=1}^{l} t_i\right) \times 100$, where $T$ denotes the total number of samples, $l$ the number of sample clusters, and $t_i$ the numbers of the samples correctly clustered into a sample class $i$.

**Performance comparison.** Figure 1 illustrates the average tissue sample accuracy using MSSRCC with RESIDUE(II). As reported in [9], spectral initialization and local search strategy play a significant role in improving MSSRCC performance. However, in this paper, we are more interested in how data transformations affect the tissue sample accuracy performance.

NT, DC, and MC with random initialization ((a)-(d)) and NT and MC with spectral initialization ((e)-(h)) result in nearly similar accuracy. Note that RESIDUE(II) is not affected by shifting factors, but still affected by the scaling factors as first articulated in [1] and also revisited in the analysis. To be more specific, the residue with NT on data matrices with local scaling factors is $(\pi_i - \mu_{\pi_I})(\alpha_j - \mu_{\alpha_J})$, on which interestingly the residue with DC or MC is also dependent. In our experiment, DC with random initialization generates compatible accuracy with that of other data transformations ((a)-(d)), however it is relatively less effective with spectral initialization ((f)-(h)). For all the considered datasets, MC presents compatible performance that of NT, but not better than that of either SDN or ZT.

As analyzed in the previous section, both column SDN and column ZT help MSSRCC with RESIDUE(II) capture perfect co-clusters, thus MSSRCC with column SDN or column ZT is supposed to generate similar accuracy and also better accuracy than those with NT, DC, or MC. Accordingly, they lead to the best accuracy values for most cases ((a)-(h)).

## 6 Conclusion and Remark

Aguilar-Ruiz [1] issues the need of a new metric to discover both scaling and shifting patterns, showing that the sum squared residue can discover any shifted patterns but may not capture some scaled patterns. To answer this need, we propose a simple remedy that helps the residue resolve its dependency on scaling variances. We suggest to take specific data transformations through which the hidden scaling factors are implicitly removed. We analyze the effect of various data transformation on RESIDUE(II) [8] for data matrices with global/local scaling and global/shifting factors.

Both analysis and experimental results reveal that column standard deviation normalization and column Z-score transformation are effective for RESIDUE(II). To be more specific, through MSSRCC with RESIDUE(II) and the two data transformations, we are able to discover coherent patterns with both scaling and shifting factors. The transformed matrix contains the constant global scaling factor 1 and local shifting factors and gives the perfect residue score, *i.e.*, zero RESIDUE(II).

Note that RESIDUE(II) is a special case (scheme 6) of the six Euclidean co-clustering schemes in Bregman co-clustering algorithms [4]. Our formal analysis can be applicable to any clustering/co-clustering algorithm that has a closed-form of objective function, thus our potential future direction is to apply the proposed analysis to the remaining co-clustering models in Bregman co-clustering algorithms and formally characterize each of Bregman co-clustering algorithms.

## References

1. Aguilar-Ruiz, J.S.: Shifting and scaling patterns from gene expression data. Bioinformatics 21(20), 3840–3845 (2005)
2. Alon, U., et al.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. PNAS 96(12), 6745–6750 (1999)
3. Armstrong, S.A., et al.: MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nature Genetics 30, 41–47 (2002)
4. Banerjee, A., Dhillon, I.S., Ghosh, J., Merugu, S., Modha, D.: A Generalized maximum entropy approach to Bregman co-clustering and matrix approximation. Journal of Machine Learning Research 8, 1919–1986 (2007)
5. Ben-Dor, A., Chor, B., Karp, R., Yakhini, Z.: Discovering Local Structure in Gene Expression Data: The Order-Preserving Submatrix Problem. Journal of Computational Biology 10(3-4), 373–384 (2003)
6. Bø, T.H., Jonassen, I.: New feature subset selection procedures for classification of expression profiles. Genome Biology 3(4) (2002)
7. Cheng, Y., Church, G.M.: Biclustering of expression data. In: Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB), vol. 8, pp. 93–103 (2000)
8. Cho, H., Dhillon, I.S., Guan, Y., Sra, S.: Minimum sum squared residue based co-clustering of gene expression data. In: Proceedings of the Fourth SIAM International Conference on Data Mining (SDM), pp. 114–125 (2004)
9. Cho, H., Dhillon, I.S.: Co-clustering of human cancer microarrays using minimum sum-squared residue co-clustering (MSSRCC) algorithm. IEEE/ACM Transactions on Computational Biology and Bioinformatics (IEEE/ACM TCBB) 5(3), 114–125 (2008)
10. Dettling, M., Bühlmann, P.: Supervised clustering of genes. Genome Biology 3(12) (2002)
11. Dudoit, S., Fridlyand, J.: A Prediction-based Resampling Method for Estimating the Number of Clusters in a Dataset. Genome Biology 3(7) (2002)
12. Golub, T.R., et al.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science 286, 531–537 (2002)
13. Hartigan, J.A.: Direct clustering of a data matrix. Journal of the American Statistical Association 67(337), 123–129 (1972)
14. Kowalski, B.R., Bender, C.F.: Pattern recognition: A powerful approach to interpreting chemical data. Journal of the American Chemical Society 94(16), 5632–5639 (1972)
15. Madeira, S.C., Oliveira, A.L.: Biclustering algorithms for biological data analysis: a survey. IEEE Transactions on Computational Biology and Bioinformatics (IEEE ACM/TCBB) 1(1), 24–45 (2004)
16. Sánchez, F.C., Lewi, P.J., Massart, D.L.: Effect of different preprocessing methods for principal component analysis applied to the composition of mixtures: detection of impurities in HPLC-DAD. Chemometrics and Intelligent Laboratory Systems 25(2), 157–177 (1994)