

# A Clustering Method for Discovering Patterns Using Gene Regulatory Processes

Siyoung Park<sup>1</sup>

Parkc0@postech.ac.kr

Daewoo Choi<sup>2</sup>

dachoi@freechal.com

Chi-Hyuck Jun<sup>1</sup>

chjun@postech.ac.kr

<sup>1</sup> Department of Industrial Engineering, Pohang University of Science and Technology, San 31, Hyoja-dong Nam-gu Pohang, Kyungbuk, Korea 790-784

<sup>2</sup> Department of Informatics & Statistics, Hankuk University of Foreign Studies San 89, Wangsan-ri Mohyun Youngin, Kyounggi, Korea 449-791

**Keywords:** clustering, factor analysis, noise detection, patterns, gene expression data

## 1 Introduction

Clustering is a descriptive task that seeks to identify homogeneous groups of objects based on the values of their attributes (dimensions). The k-means and hierarchical as well as self-organizing maps have all been used for clustering expression profiles and a number of algorithms have been developed for expression data and applied to analyze it. These Clustering methods usually use metric distance for similarity measure. Correlation coefficient is also used but has a problem that it removes difference attributable to both the mean and the dispersion of the observations. Moreover, it may be unreasonable that every observation is assigned to one of clusters when the purpose is to find groups with similar pattern. Alter *et al.* [1] show that several significant eigengenes and the corresponding eigenarrays capture most of the expression information in field of genetics and some of the eigengenes represent independent regulatory programs or processes from its expression pattern across all arrays. Normalizing the data by filtering out the eigengenes (and the corresponding eigenarrays) that are inferred to represent noise or experimental artifacts enables meaningful comparison of the expression of different genes across different arrays in different expression. Such normalization may improve any further analysis of the expression data.

Q-mode factor analysis has been used to find groups like clustering analysis and could be a good method to find patterns. However, this approach to clustering is plagued with a number of problems [3]. Genes with similar expression profiles may have something in common in their regulatory mechanisms. In this study, Q-mode factor analysis is used to model the gene regulatory processes which control genes and gene products and we modify the Q-mode factor analysis for discovering useful patterns in gene expression data. As a result of the factor modeling of gene expression data, our method can improve the result of clustering by removing noises and produce characteristic values of expression data.

## 2 Proposed Method

The proposed method consists of two steps to find clusters or patterns in the data set. Summary of the proposed method are as follows.

[Step 1] Discriminate patterns and noises by factor model

- Q-mode factor analysis
- Noise detection by two-group clustering using communality
- Noise deletion

[Step 2] Find patterns using factor loadings and PC loadings

- PCA (principal component analysis) on covariance matrix input

- Hierarchical clustering factor loadings using Gap statistic
- Hierarchical clustering PC loadings using Gap statistic

Let  $Y$  be  $n \times p$  data matrix with these  $p$  variables and  $n$  observations. In the first step, we transpose  $Y$  and we make each column of the resulting matrix centered to have zero mean to eliminate the effect of mean. Let us denote the matrix by  $X$  and the  $i$ -th column vector or the  $i$ -th observation vector by  $x_i$  so that  $X = (x_1, x_2, \dots, x_n)$ . The independent regulatory processes described in the introduction can be also modeled by factor model expressed by matrix form as  $X^T = \Lambda F + E$ , where  $F = (f_1, f_2, \dots, f_m)^T$  is an  $m \times p$  matrix of unobserved variables called common factors,  $E$  is a matrix of error terms and  $\Lambda$  is  $n \times m$  matrix of unknown constants called factor loadings.

Factor loading matrix is the simple correlation between latent factors and indicator variables. Factor loading vector of  $x_i$ ,  $\lambda_i = [\lambda_{i1}, \lambda_{i2}, \dots, \lambda_{im}]$ , would be characteristic value of observations. The number of factors can be determined by scree test or parallel procedure and so on. In traditional Q-mode factor analysis, each factor is assumed to be a type of observations but here we defined the factors as independent regulatory processes for patterns. We will use these factor loadings for further analysis instead of the raw data. Since these values are obtained after eliminating the errors from the original data and factor loadings would represent more accurately the shape of some patterns. We define a noise as an observation that is less affected by common factors. Therefore, we could identify a noise by examining communality of each observation. Communality of the  $i$ -th observation is computed as  $c_i = \lambda_{i1}^2 + \lambda_{i2}^2 + \dots + \lambda_{im}^2$ . To discriminate noises from data, we perform two-group partitional clustering method using communalities of each observation. If observations are classified as noise group, we should eliminate these observations for improvement of further analysis.

In the second step, we employ hierarchical clustering method and Gap statistics using factor loading. However, The groups from hierarchical clustering method using the factor loadings cannot distinguish observations that have high correlation but different variance among them. Therefore, we should consider the covariance structure of data by PCA. After performing PCA for  $X$ , we obtain PC loadings for each observation like the factor loading. We perform hierarchical clustering method using both the factor loadings and the PC loadings. To determine the number of clusters, we use Gap statistics,  $\text{Gap}(k)$ , obtained by  $\text{Gap}(k) = 1/B \sum_b \log(W_{kb}^*) - \log(W_k)$  for  $k = 1, \dots, K$ , where  $k$  is the number of clusters,  $W_k$  is pooled within cluster sum of squares around the cluster means and  $W_{kb}^*$  is  $W_k$  of  $B$  reference datasets, where  $B$  is the number of reference datasets and  $K$  is the maximum number of clusters defined by user,  $b = 1, 2, \dots, B$ . Finally we perform hierarchical clustering method with the number of clusters obtained from Gap statistic. If we obtain different results from between factor loadings and PC loadings, we can conclude that there are a few patterns that have different variance structures.

### 3 Discussion

We generate a simulation data that there exist 8 underlying patterns and random noises to compare our method with sliding window gene shaving [2] which modified gene shaving clustering algorithm [4]. We applied our method and sliding window gene shaving to the simulation data. Our method would get better result than sliding window gene shaving because sliding window gene shaving missed a few patterns that have small variance compared to other patterns.

In analyzing gene expression data, it is important to discover the functional roles of different genes and cellular processes that they participate in. This work suggests that the application of Q-mode factor model is useful in analyzing gene expression data set. Removing noise could improve the result of clustering but have a risk to get rid of important genes to reveal regulatory processes. The number of factors plays a critical role in discriminating noise from data. So we should determine cautiously the number of factor and investigate the meaning of factors, regulatory processes. Finally this study highlights the underlying regulatory processes to discover the useful patterns of gene expression data.

## References

- [1] Alter, O., Brown, P.O., and Botstein, D., Singular value decomposition for genome-wide expression data processing and modeling, *Proc. Natl. Acad. Sci. USA*, 97(18):10101–10106, 2000.
- [2] Choi, D., Lee, H., and Jun, C., On combining clustering methods for microarray data analysis, *The Proceedings of International Statistical Institute*, 2001.
- [3] Dillon, W.R. and Goldstein, M., *Multivariate Analysis: Methods and Applications*, John Wiley & Sons, 1984.
- [4] Hastie, T., Tibshirani, R., Eisen, M.B., Alizadeh, A., Levy, R., Staudt, L., Chan, W.C., Botstein, D., and Brown, P., ‘Gene shaving’ as a method for identifying distinct sets of genes with similar expression patterns, *Genome Biology*, 1(2):RESEARCH0003, 2000.
- [5] Tibshirani, R., Walther, G., and Hastie, T., Estimating the number of clusters in a dataset via the Gap statistic, *Tech. Report, Dept of Statistics, Stanford Univ.*, 2000.