

# Significance Test of Clusters in Gene Expression Profile Data

Natsuko Kawase

natsu-ka@is.aist-nara.ac.jp

Shigeyuki Oba

shige-o@is.aist-nara.ac.jp

Shin Ishii

ishii@is.aist-nara.ac.jp

Graduate School of Information Science, Nara Institute of Science and Technology,  
8916-5 Takayama, Ikoma, Nara 630-0101, Japan

**Keywords:** statistical test, clustering, DNA microarrays

## 1 Introduction

Clustering is a popular technique used in gene expression analyses. However, conventionally well-used clustering methods often produce variety in results depending on, e.g., initial algorithmic condition. Due to this instability, analysis results may be questionable. For example [3], although B-cell lymphomas had been classified into two distinct classes using hierarchical clustering, another analysis by a different clustering method failed to find these classes [1, 2].

In spite of such instability, the preprocessing or visualization of data by a clustering technique is still important for gene expression analyses. For utilizing technique with certain reliability, it is required to develop not only a stable method but also a method to evaluate quantitatively the reliability of the clustering results.

In this report, we propose a quantitative evaluation method of clusters, which determines whether a data group, seemingly a cluster, constructs a statistically significant cluster. In particular, we propose a statistical test examining the existence possibility of a cluster within a parametric model of background noise. A seeming cluster is expressed by a window function represented by characteristic parameters. This statistical test can be extended to examining of feature spaces; we also propose a significance test of feature subspace extracted from PCA analysis.

## 2 Method and Results

### 2.1 Significance of the Window

First, we consider significance of the hypothesis that a certain cluster exists under the assumed background noise. Here, a seeming cluster is expressed by a window function that indicates whether each data point belongs to the assumed cluster or not. For a single window, we define a null hypothesis  $H_0$ : that window is not a cluster, and the number of data points in the window  $N_{in}$  out of all  $N$  data points can be explained by the background noise model. Let  $\beta$  be the probability that a randomly sampled data point from the background noise model to happens to enter the window. If the events are actually by chance, the number  $N_{in}$  obeys a binomial distribution  $B_i(N_{in}|N, \beta)$ . The P-value of the hypothesis  $H_0$  is calculated by integrating the beta distribution, which is an instance of well-known binomial tests. When the null hypothesis  $H_0$  is rejected by a certain level of significance, e.g.,  $\alpha = 0.01$ , the cluster expressed by the window is regarded as significant.

The significance test above can be applied to various window models and noise models. We here present an example, in which a conic window and an isotropic background noise are assumed. A conic window is defined by a direction vector  $m$  of the central axis and the angle  $\theta$  which represents the cone spread. A conic window method is equivalent to the assumption that each data point is mapped onto an unit supersphere so that the distance is measured along the supersphere; namely, this cluster is

correlation-based. Under the assumption of the isotropic background noise, we can analytically obtain the ratio  $\beta$  of a conic window with any angle  $\theta$  as a circular area on the supersphere. From that  $\beta$  value, the P-value for the null hypothesis is calculated.

## 2.2 Significance of the Feature Subspace

Extending the cluster significance test to a hierarchical manner, we can also test the significance of feature subspaces which are extracted by some feature extraction methods, e.g., PCA. Although we sometimes execute clustering analyses in a feature subspace, it should be examined beforehand that the data distribution in the subspace is sufficiently informative, e.g., involving cluster-like structures. If the subspace is non-informative, say uniform, further cluster analyses would be meaningless.

After the whole dataset is whitened so that the sample mean is zero and the sample covariance matrix is an identity, an isotropic background noise model is assumed. If a certain feature subspace is non-informative higher-order null hypothesis, the number  $N_s$  of significant windows out of  $N_r$  randomly sampled windows under the lower-order null hypothesis  $H_0$  and a certain level of significance  $\alpha$ , the expectation value of  $N_s$  becomes  $\alpha N_r$ . If  $N_s$  is much larger than  $\alpha N_r$ , the feature subspace is regarded as significantly anisotropic and then informative. The P-value of this hierarchical test is obtained similarly to the binomial test above.

## 2.3 Result

Fig. 1 illustrates the cluster test based on conic windows, for 1000 artificial samples that including two clusters in a two-dimensional space under the background noise. The curve around a circle denotes the number of data points entering the window centered at each direction. The broken-line circle denotes the expected data number in the window, explained by the background noise. The solid circle corresponds to the significance level  $\alpha = 0.01$ . When the curve goes over the solid circle, there is a significant cluster around that region. From this figure, we confirm that two clusters exist by a quantitative manner.

Fig. 2 shows the result of the significance test of the feature subspace. We applied PCA to a liver cancer dataset obtained by microarrays, that is 99 cases with 1812 genes. The upper panel in Fig. 2 shows eigenvalues in descending order. We tested the significance of two dimensional feature subspaces, (1st, 2nd), (2nd, 3rd)  $\dots$ , and the lower panel in Fig. 2 shows the number of significant windows found in the corresponding subspaces. The broken line denotes the significance level  $\alpha = 0.01$ . This figure shows that the first six principal subspaces extracted by PCA are significant and informative and the others do not have significant cluster-like structure.

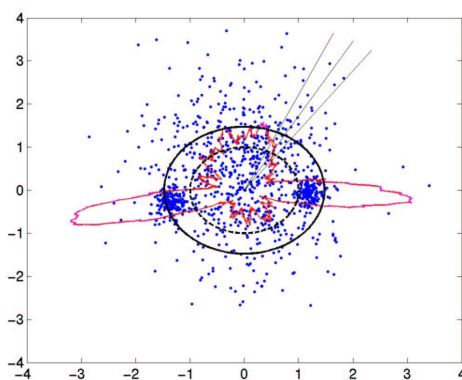


Figure 1:

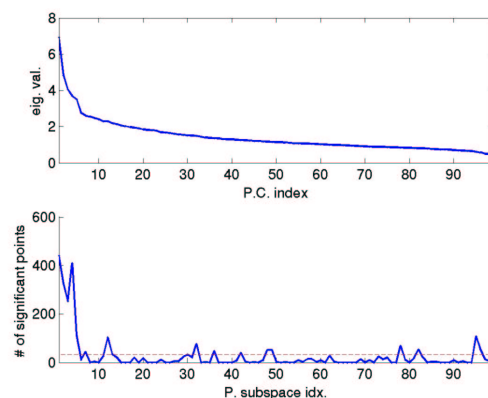


Figure 2:

## References

- [1] Alizadeh, A.A., *et al.*, Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature*, 403:503–511, 2000.
- [2] Shipp, M.A., *et al.*, Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning, *Nature Medicine*, 8:68–74, 2002.
- [3] Tilstone, C., News feature; Vital Statistics, *Nature*, 424:599–707, 2003.