An Efficient Measure of Similarity between Gene Expression Profiles through Data Transformations

[1,2] Kyungpil Kim, [3] Keni Jiang, [3] Shibo Zhang, [4] Li Cai, [2] In-Beum Lee , [3] Lewis Feldman, [1] Haiyan Huang

[1] Department of Statistics, University of California, Berkeley, USA; [2] Department of Chemical Engineering, Pohang University of Science and Technology, Korea; [3] Department of Plant and Microbial Biology, University of California, Berkeley, USA; [4] Department of Biomedical Engineering, Rutgers University, USA

Correspondence should be addressed to: Haiyan Huang (hhuang@stat.berkeley.edu)

June 2006

Abstract

Background: Clustering methods have been widely applied to gene expression data in order to group genes sharing common or similar expression profiles into discrete functional groups. In such analyses, designing an appropriate (dis)similarity measure is critical. In this study, we aim to develop a new distance measure for gene expression profiles. The new measure is expected to be especially efficient when the shape of expression profile is vital in determining the gene relationship, yet the expression magnitude should also be accounted for to some extent. **Results:** The new measure, named *TransChisq*, was developed by separately modeling the shape and magnitude information and then using the estimated shape and magnitude parameters to define a distance measure in a new feature space. The feature space was constructed based on the specific clustering purpose of grouping genes with similar shape of expression curves, while the magnitude information should also be considered when determining the shape similarity. The new measure was employed into a *k*-means clustering procedure for performing clustering analyses. Results from applications to a simulation dataset, a developing mouse retina SAGE dataset, a small yeast sporulation cDNA dataset and a maize root affymetrix microarray dataset show the clear advantages of our method over others. **Conclusions:** The proposed method described in this paper shows great promise in capturing underlying biological relationship in gene expression profiles. This study also demonstrates that the construction of an appropriate feature space under certain clustering purpose is critical for a successful distance measure. We hope our method provides some new insights to further investigation in analyzing gene expression data. The clustering algorithms are available upon request.

Key words: clustering analysis; gene expression profiles; similarity measure; *PoissonC*

Background

With the explosion of various 'omic' data, a general question facing the biologists and statisticians is how to summarize and organize the observed data into meaningful structures. Clustering is one of the methods that have been widely explored for this purpose [1-3]. In particular, it is being generally applied to gene expression data to group genes sharing common or similar expression profiles into discrete functional clusters. Many clustering methods are available, including hierarchical clustering [3], *k*-means clustering [4-5], self-organizing maps [6], and various model based methods [7-9].

The focus of recent research in clustering analysis has been largely on the estimation of number of clusters in data with noise points [10-12] and the optimization of clustering algorithms [13-14]. In this present study, we focus on a different yet fundamental issue in clustering analysis: defining an appropriate measure of similarity for gene expression patterns.

The most common distance or similarity measure for analyzing gene expression data are the *Euclidean distance* and *Pearson correlation coefficient*, which are simple and easy to be implemented. However, in some situations, both measures could be unsuitable to explore the true gene relationship since *Pearson correlation* can be overly sensitive to the shape topology of an expression curve and *Euclidean distance* only cares about the magnitude of changes. For other model-based distance or similarity measures [15-17], their successes would highly depend on how well the assumed probability model fits the data and the clustering purpose.

In this study, we proposed a new distance measure, named *TransChisq*, to determine gene relationships concerning the shape of expression profiles. Moreover, the expression magnitude was also considered when measuring the shape similarity. The new method was designed based on a data transformation that emphasizes the shape of expression profiles and a distance measure *PoissonC* proposed for Serial Analysis of Gene Expression (SAGE) data in Cai et al. [18]. The new method should be applicable to other datasets besides SAGE data.

The detailed idea of the new method is to separately model the shape and magnitude of gene expression profiles, and use the estimated shape and magnitude parameters to define a chi-square based distance measure in a new feature space. The construction of an appropriate feature space under certain clustering purpose is the key for the success of the new distance measure, since an effective summary of data can greatly improve and simplify the extraction of relationship between genes. We explored several different transformation schemes to construct the feature space, including a space with features determined by the mutual differences of original expression components, a component space derived from a parametric covariance matrix, and the principal component space in PCA analysis [19]. Each of the measures defined in these spaces exhibits distinct characteristics. In order to evaluate these measures, we implemented them in a *k*-means clustering procedure and analyzed a simulation dataset and three experimental datasets. The experimental datasets include a

developing mouse retina SAGE dataset of 153 tags [18], a small yeast sporulation cDNA dataset [20] and a maize root affymetrix microarray dataset [21]. From the results, we found that the distance measure defined in the first feature space, named *TransChisq*, showed the best performance in producing more accurate clusters or clusters of more biological relevance. We called the measure defined in the third feature space (the principal component space) as *PCAChisq*.

We also implemented a set of widely used distance or similarity measures in the *k*-means clustering procedure for comparison. The measures we used include *Pearson correlation* (the corresponding algorithm named *PearsonC*), Euclidian distance (*Eucli*) and a chi-square based measure for Poisson distributed data (*PoissonC*). *TransChisq* was proved more powerful than other commonly used measures when the shape of expression profile is the key factor in determining the gene relationship, yet the expression magnitude should also be accounted for to some extent. The MATLAB source codes for all these algorithms are available upon request.

Results and discussion

In this section, we first used a maize expression dataset to illustrate the property of the new data transformations introduced in Method section. Next, for performance evaluation and comparison, we implemented *TransChisq*, *PCAChisq* (see Method section) and other commonly used distance or similarity measures into a *k*-means clustering procedure, and applied them to a simulation dataset, a yeast sporulation microarray dataset, and a mouse retinal SAGE dataset. The results demonstrate the success of our proposed method in practical applications.

Experimental maize gene expression data

The demonstration dataset consists of nine Affymetrix microarrays profiling the gene transcription activity in three maize root tissues with three biological replicates: the proximal meristem (PM), the quiescent center (QC), and the root cap (RC) [21]. We used the 2092 significantly differentially expressed genes, categorized into 6 classes of expression patterns by Jiang et al., to illustrate the properties of the newly proposed data transformation methods with a comparison to the traditional PCA.

We first applied the transformation employed in *TransChisq* to the above data. Figure 1(a)-(c) plot the expression profiles of the genes onto the new space, wherein each axis represents the gene expression difference in any two maize root tissues. The blue and red genes are from the two dominating classes (RC up or down regulated genes account for 94% of all genes) and the other four colors (orange, green, pink, light blue) correspond to the other four small classes (up- or down- regulated genes in QC or PM account for 6% of all genes). Three plots ascertain that the six classes can be recognized explicitly regardless of the relative size of each class in the new space.

We next applied the transformation in (7), suggested by the parametric covariance matrix,

to the same data (see Method). Figures 1(d)-(f) plot the expression profiles of the genes onto the (d) 1st and 2nd, (e) 2nd and 3rd, and (f) 1st and 3rd components in the new space. We see that the second and third components have correctly separated all six classes in Figure 1(e). The description of the six class separating regions, whose centers are the dotted lines in Figure 1(e), is provided in Table 1 (e.g., the genes around the line PC2 = $\sqrt{3} \cdot$ PC3 < 0 are expected to be PM up-regulated).

Both the above two transformations have a nice property: the information carried by each component is explicit and then the region in the new space corresponding to each class can be explicitly determined.

For comparison, we performed a PCA analysis to the same data. Figures 1(g)-(i) plot the expression profiles of the genes onto the (g) 1st and 2nd, (h) 2nd and 3rd, and (i) 1st and 3rd principal components. We see that the direct application of PCA is only able to separate the two dominating expression patterns and fails to recognize other patterns, even when exhausting all principal components. The failure of PCA could be attributed to the use of empirical sample covariance matrix for principal components determination. In this dataset, about 94% genes are RC up or down regulated genes, which causes the most variance in data. The principal components, determined from this sample covariance matrix, thus mainly capture the two dominating clusters and miss the meaningful class information for the other four small groups.

This example demonstrates the advantage of the proposed data transformations over PCA in keeping class information intact. These results shed a light on the successful applications of *TransChisq* in clustering analysis.

Simulation study

In the following, we call the modified *k*-means algorithms with the measures *TransChisq*, *PCAChisq*, *PoissonC*, *Pearson correlation coefficient* and *Euclidian distance* implemented as *TransChisq*, *PCAChisq*, *PoissonC*, *PearsonC* and *Eucli* respectively.

To evaluate the performance of these algorithms, we first applied them to a simulation dataset. The distributions used to generate the simulation dataset are described in Table 2. It consists of 46 vectors of dimension 5 with components independently generated from different Normal distributions. The mean (μ) and variance (σ^2) parameters of the Normal distributions are constrained by $\sigma^2 = 3\mu$. The 46 vectors belong to six groups (named A, B, C, D, E, and F) according to the Normal distributions they are generated from. The six groups are of size 3, 6, 6, 9, 7, and 15 respectively. Here, genes with similar expression profile shape are considered to be in the same group. Though the expression magnitude itself is not a factor for determining the gene clusters, its information is still useful and should be accounted for in comparing the expression profile shapes, i.e. the deviation penalty should be smaller for genes with larger expression magnitude.

The clustering results from different methods are shown in Figure 2. Only *TransChisq* has correctly grouped genes into six classes. *PCAChisq* (with all PCs used), *PoissonC*, and

PearsonC mix group A and group B together, and *Eucli* clusters genes mostly based on the magnitude of gene expression levels between data rather than the shape changes. To reduce the magnitude effects, we further applied *Eucli* to the rescaled data. The rescaling is performed in the way that the sum of the components within each vector is set the same. The clustering result of *Eucli* on rescaled data (Figure 2(f)) is improved over that on original data, though it is still not perfect.

We performed an additional 100 replications of the above simulation. *TransChisq, PCAChisq* and *PoissonC* correctly cluster 75, 37 and 43 of the 100 replicate simulation datasets respectively, while *PearsonC, Eucli* and *Eucli* on rescaled data never generate correct clusters. For *PCAChisq*, we have also tried different numbers and combinations of PCs to optimize the clustering results, which is however still not helpful to identify all the six classes.

This application evaluates the performance of our method on normally distributed data with Poisson-like properties: variance increases with mean. Success in this dataset would shed a light on more broad applications of our method.

Experimental mouse retinal SAGE data

For further validation, we applied *TransChisq*, *PCAChisq*, *PoissonC*, *PearsonC* and *Eucli* to a set of mouse retinal SAGE libraries. The raw mouse retinal data consists of 10 SAGE libraries (38818 unique tags with tag counts ≥ 2) from developing retina taken at 2 day intervals, ranging from embryonic to postnatal and adult [18, 22]. 1467 of the 38818 tags with counts ≥ 20 in at least one of the 10 libraries are selected. To effectively compare the clustering algorithms, a subset of 153 SAGE tags with known biological functions were further selected. These 153 tags fall into 5 clusters based on their biological function(s) (see Table 3(a)). 125 of these genes are developmental genes, which can be further grouped into four clusters by their expressions at different developmental stages. The other 28 genes are un-related to the mouse retina development. The average expression profiles for the five classes are shown in Figure 3.

TransChisq, *PCAChisq*, *PoissonC*, *PearsonC* and *Eucli* are applied to group these 153 SAGE tags into five clusters. Results show that *TransChisq* and *PCAChisq* outperform others (See Table 3(b)): 12, 12, 22, 26 and 38 of the 153 tags are wrongly clustered by *TransChisq*, *PCAChisq, PoissonC, PearsonC* and *Eucli* on rescaled data respectively. In general, *PCAChisq* would work well if the principal components can briefly capture the between-class variations. In this example, we found that the 5 different expression patterns can be well separated in the principal component space. The results from *Eucli* on original data are too messy to report the number of wrongly clustered tags. The performance of *TransChisq* can be further improved to give 8 wrongly clustered tags when we use the transformation associated with all the row-switching vectors of $\mathbf{e}_1,\ldots,\mathbf{e}_T$ (see Method section).

Microarray yeast sporulation gene expression data

To illustrate the effectiveness of our method at identifying genes with characterized patterns in a microarray analysis, we applied our method to a yeast sporulation dataset. Chu et al. [20] measured gene expression in the budding yeast *Saccharomyces cerevisiae* at seven time points during sporulation using spotted microarrays and identified seven distinct temporal patterns of induction. 39 representative genes for each of these seven patterns were used to define a model expression profile in that study. According to the property of each pattern, the seven patterns are named as Metabolic, Early I, Early II, Early-Mid, Middle, Mid-Late and Late. The average expression profiles for these seven patterns are presented in Figure 5. The profiles of Early I, Early II, Middle, Mid-Late and Late start induction at 0.5h, 2h, 5h, 7h and 9h, respectively, and sustain expression through the rest of the time course. Metabolic profile is also induced at 0.5h like Early I, but decayed afterwards. The genes in Early-Mid are induced not only at the 0.5h and 2h like Early genes, but also at 5h and 7h like the Middle and Mid-Late genes, which makes it difficult to separate this pattern from others. Due to the complex data structure, direct clustering analysis using *PearsonC* or *Eucli* turns out to be unsuccessful (results not shown).

Before analyzing the data, we first set the expression ratios below 0 to zero as in Figure 6(a). This simplifies the expression patterns, but keeps the key properties of each pattern intact. The clustering results are briefly summarized in Table 4. We see that *TransChisq* is superior to other methods: 3, 7, 8, 13, 14 and 17 of the 39 genes are wrongly clustered by *TransChisq*, *PoissonC, Eucli, PearsonC, PCAChisq* and *Eucli* on rescaled data respectively. It is interesting to see that *Eucli* on rescaled data is working worse than *Eucli* on original data, which is suggesting that the magnitude information should not be ignored to determine the seven classes. As we have discussed and shown in Figure 6(b)-(f), all methods fail to discriminate the genes in Early-Mid from the genes in Early I, Early II, Middle, Mid-Late and Late. Furthermore, *PCAChisq* and *PoissonC* mix up two different patterns from Metabolic and Early I due to their similar induction time at 0.5h (Figure 6(c) and (d)), and *PearsonC* even split Metabolic group further into two separate clusters (Figure 6(e)).

For *PCAChisq*, we have also tried different numbers and combinations of PCs to optimize the clustering results. The best result can be reached when first 5 PCs are used that results 3 of the 39 genes were incorrectly grouped. This optimal result is the same with that from *TransChisq*. However, it is not feasible to exhaust all possible combinations of PCs to search for the optimal clustering result in practice.

Conclusions

In this study, we proposed a new distance measure, named *TransChisq*, to determine gene relationships concerning the shape of expression profiles. Moreover, the expression magnitude was considered when measuring the shape similarity. The new method was designed based on a data transformation that emphasizes the shape of expression profiles and a distance measure *PoissonC* proposed for SAGE data [18]. Results from applications to a

variety of datasets show the clear advantages of *TransChisq* over other methods. This also demonstrates that the data transformation we utilized is effective in projecting the data into an informative space regarding the presentation of "pattern" for each class.

The proposed method described in this paper shows great promise but requires further study on possible data transformation schemes when the columns of original data matrices show complicated level of dependencies or when the clustering purpose is different. We hope our method provides some new insights to further investigation in gene expression experiments.

Methods

Our method was proposed based on the *PoissonC* in Cai et al. [18]. So before describing the new method, we first gave a brief review on *PoissonC*..

Review on the Poisson-based distance measure, PoissonC, for SAGE data

SAGE is one of the effective techniques for comprehensive gene expression profiling. The result of a SAGE experiment, called a SAGE library, is a list of counts of sequenced tags isolated from mRNAs that are randomly sampled from a cell or tissue. Ideally, each tag is uniquely mapped to a gene and its counts reflect the level of expression of the corresponding gene. SAGE data can be naturally modeled by Poisson distributions due to the data property; they are generated by "sampling," which results in counts. *PoissonC*, an effective distance measure for tag count profiles, was developed under this context [18]. The method was summarized below.

Let $Y_i(t)$ be the count of tag i in library t, and $\mathbf{Y}_i = (Y_i(1),\ldots, Y_i(T))$ be the vector of counts of tag i over a total of T libraries. $\mathbf{Y}_i$ is regarded as the count profile of tag i, which represents the expression profile of the corresponding gene. $Y_i(t)$ was assumed to be Poisson distributed with expected count γ_{it}, which can be further parameterized as $\gamma_{it} = \lambda_i(t)\theta_i$ to separately model the magnitude and shape of the count profiles. Here, θ_i represents the expression magnitude that is defined as the expected sum of counts of tag i over all libraries. $\lambda_i(t)$ reflects the expression profile shape, which can be considered as the contribution of tag i in library t to the sum θ_i expressed in percentage ($\Sigma_{t=1}^{T}\lambda_i(t)=1$). So $\lambda_i(t)\theta_i$ redistributes the tag counts according to the expression shape parameters ($\lambda_i(t)$'s), and the genes with similar $\lambda_i(t)$'s are considered to be in the same cluster. Under this model, the joint likelihood function for a cluster consisting of tags 1,2,…,m can be expressed as

$$L(\boldsymbol{\lambda},\boldsymbol{\theta}\,|\,\mathbf{Y}) \propto f(\mathbf{Y_1},\ldots,\mathbf{Y_m}\,|\,\boldsymbol{\lambda},\theta_1,\ldots,\theta_m) = \prod_{i=1}^{m}\prod_{t=1}^{T}\frac{\exp(-\lambda(t)\theta_i)(\lambda(t)\theta_i)^{Y_i(t)}}{Y_i(t)!}. \qquad (1)$$

The maximum likelihood estimates of $\boldsymbol{\lambda}$ and $\theta_1,\ldots,\theta_m$ are

$$\hat{\theta}_i = \sum_t Y_i(t), \text{ and } \hat{\lambda}(t) = \sum_{i=1}^{m} Y_i(t) \Big/ \sum_{i=1}^{m}\hat{\theta}_i = \sum_{i=1}^{m} Y_i(t) \Big/ \sum_{i=1}^{m}\sum_t Y_i(t). \qquad (2)$$

In order to evaluate how closely the observed samples are expressed to the estimated cluster model, a chi-square test statistic was adopted to measure the cluster dispersion:

$$S = \sum_i \sum_t \left(Y_i(t) - \hat{\lambda}(t)\hat{\theta}_i \right)^2 \Big/ \left(\hat{\lambda}(t)\hat{\theta}_i \right). \quad (3)$$

This method was called as *PoissonC*. The joint likelihood $f(\boldsymbol{Y_1},...,\boldsymbol{Y_m} \mid \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\theta}})$ can also be used to evaluate how well the observed samples $\boldsymbol{Y_1},\ldots,\boldsymbol{Y_m}$ fit the expected Poisson models:

$$L = -\sum_i \log f\left(\boldsymbol{Y}_i \middle| \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\theta}}_i \right). \quad (4)$$

This method was called as *PoissonL*. Chi-square test statistic makes the penalty for deviation from a large expected count smaller than that for a small expected count, a property consistent with the "mean = variance" nature of Poisson distributions. So *PoissonC* is expected to perform similarly to *PoissonL*. In general, the smaller the value of *S* or *L*, the more likely the tags have similar patterns and belong to the same cluster. We should also note that because of the separately estimated parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$, the statistics in (3) and (4) account for both the shape and magnitude information when measuring the cluster dispersion, though there is an emphasis on the shape.

However the use of square of deviation from expectation in (3) could somewhat torture the gene relationship concerning the shape of expression curves in some situations. For example, we considered an expression vector $\boldsymbol{Y}$ = (15, 30, 15), and its relationship with two clusters wherein the expected expression profiles of $\boldsymbol{Y}$ are $\boldsymbol{Y}_E^1$=(5, 50, 5) and $\boldsymbol{Y}_E^2$=(25, 10, 25) respectively. It is reasonable to expect that $\boldsymbol{Y}$ is closer to the first cluster because of the high expression observed on the middle component in both $\boldsymbol{Y}$ and $\boldsymbol{Y}_E^1$. The measures in (3) and (4), however, determine that $\boldsymbol{Y}$ has the same distance to $\boldsymbol{Y}_E^1$ and $\boldsymbol{Y}_E^2$. By (3), the distance between $\boldsymbol{Y}$ and $\boldsymbol{Y}_E^1$ is $(15-5)^2/5+(30-50)^2/50+(15-5)^2/5=48$; the distance between $\boldsymbol{Y}$ and $\boldsymbol{Y}_E^2$ is $(15-25)^2/25+(30-10)^2/10+(15-25)^2/25=48$. By (4), the distance between $\boldsymbol{Y}$ and $\boldsymbol{Y}_E^1$ is $-\log((\exp(-5)5^{15}/15!)(\exp(-50)50^{30}/30!)(\exp(-5)5^{15}/15!)) = 24.81295$ and between $\boldsymbol{Y}$ and $\boldsymbol{Y}_E^2$ is $-\log((\exp(-25)25^{15}/15!)(\exp(-10)10^{30}/30!)(\exp(-25)25^{15}/15!)) = 24.81295$. This result is clearly not desirable. The poor performance of (3) and (4) in this example can be attributed to the fact that they neglect the *direction of difference* when penalizing the deviation from the expected value and thus lose some shape information. In order to address this omission, we proposed to define a distance measure in a new feature space, wherein the expression profile shape can be more appropriately described and extracted. The construction of a proper feature space under certain clustering purpose is essential for defining an effective distance or similarity measure. In next section, we explored several different transformation schemes to construct new feature spaces. These spaces, with different characteristics, are expected to be useful in different situations.

Newly proposed distance measures: *TransChisq* ***and*** *PCAChisq*

Below we presented our new measures, which were proposed on the basis of the

probability model introduced in the previous section.

The distance measure based on a simple data transformation: TransChisq

Let us consider again the example presented in the previous section, where *PoissonC* and *PoissonL* consider $\boldsymbol{Y}$ = (15, 30, 15) to be equally distant from the two clusters with expected profile $\boldsymbol{Y}_E^1$=(5, 50, 5) and $\boldsymbol{Y}_E^2$=(25, 10, 25) respectively. The failure of *PoissonC* and *PoissonL* in this case is due to the neglects of the *direction of difference* when penalizing the deviation from the expected value. A simple yet natural feature space motivated by this example is then to consider the space consisting of the mutual differences of original vector components. That is, given a gene with expression profile/vector $\boldsymbol{Y}_i = (Y_i(1),\ldots,Y_i(T))$, the transformed vector $\boldsymbol{Z}_i$ is of dimension $T(T\text{-}1)/2$ with components in the form of $Y_i(t_1)\text{-}Y_i(t_2)$ for $t_1 = 1,\ldots,T\text{-}1$ and $t_2 = (t_1+1),\ldots,T$. These mutual differences can provide more profound interpretation regarding the shape changes of gene expression profiles and thus complement the weakness of *PoissonC* and *PoissonL*.

According to the Poisson model in the previous section, $E(Y_i(t_1)\text{-}Y_i(t_2)) = \left(\lambda_i(t_1) - \lambda_i(t_2)\right)\theta_i$ and $Var(Y_i(t_1)\text{-}Y_i(t_2)) = \left(\lambda_i(t_1) + \lambda_i(t_2)\right)\theta_i$. Then for a cluster consisting of tags 1, 2,…, m, we can define the following statistic to measure the cluster dispersion:

$$\begin{aligned} S_{trans} &= \sum_i \sum_{t_1,t_2} \Big(\big(Y_i(t_1) - Y_i(t_2)\big) - E\big(Y_i(t_1) - Y_i(t_2)\big) \Big)^2 \Big/ Var\big(Y_i(t_1) - Y_i(t_2)\big) \\ &= \sum_i \sum_{t_1,t_2} \Big(\big(Y_i(t_1) - Y_i(t_2)\big) - \big(\hat{\lambda}(t_1)\hat{\theta}_i - \hat{\lambda}(t_2)\hat{\theta}_i\big) \Big)^2 \Big/ \big(\hat{\lambda}(t_1)\hat{\theta}_i + \hat{\lambda}(t_2)\hat{\theta}_i\big), \end{aligned} \tag{5}$$

where $\hat{\lambda}(t)$ and $\hat{\theta}_i$ can be estimated by (2). We call this measure as *TransChisq*. In general, the smaller the value of S_{trans}, the more likely the tags have similar expression patterns and belong to the same cluster. For the previous example of determining the relationship between $\boldsymbol{Y}$ = (15, 30, 15) and two clusters with expected profiles $\boldsymbol{Y}_E^1$ = (5, 50, 5) and $\boldsymbol{Y}_E^2$ = (25, 10, 25), *TransChisq* considers that $\boldsymbol{Y}$ is closer to the first cluster $\boldsymbol{Y}_E^1$, which makes more sense intuitively.

The data transformation and distance measure based on a parametric covariance matrix

Here we consider the data transformation determined by a covariance matrix in the following parametric form:

$$\mathbf{R} = \mathrm{cov}(\mathbf{X}) = (\gamma_{ij})_{i,j=1,\ldots,T}, \text{ with } \gamma_{ij} = \alpha > 0 \text{ if } i = j \text{ and } \gamma_{ij} = \beta \text{ if } i \neq j, \tag{6}$$

where $\mathbf{X}$ is the data matrix with n observations on the rows and T variables on the columns, and the matrix $\mathbf{R}$ is the covariance matrix of the T variables. That the matrix $\mathbf{R}$ takes this form is equivalent to the conditions that the variables have identical variances, and that the covariance of every pair of variables is equal. These two conditions are biologically reasonable considering that normalized arrays have identical distributions and thus in

particular equal variances, and that all pairs of variables would exhibit equal covariance (or un-correlated when $\beta = 0$) if each component had been equally important (or independent) to determine a class.

This covariance matrix has two different eigenvalues $\eta_1 = \alpha + (T-1)\beta$ and $\eta_2 = \cdots = \eta_T = \alpha - \beta$. The corresponding orthonormal eigenvector to the eigenvalue η_1 is $\mathbf{e}_1 = (1/\sqrt{T}, \cdots, 1/\sqrt{T})^{\mathrm{T}}$. The orthonormal eigenspace of the eigenvalue $\eta_2 = \cdots = \eta_T$ is not unique. One set of column orthonormal eigenvectors, denoted by $\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_T$, was presented in the Appendix I. Given a gene expression profile $\boldsymbol{Y_i}$ = ($Y_i(1)$,…,$Y_i(T)$), we did the transformation of Y_i into the eigenspace of **R**, that is

$$\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iT}) = \mathbf{Y}_i \ (\mathbf{e}_1\ \mathbf{e}_2 \ldots \mathbf{e}_T). \tag{7}$$

This transformed space has nice properties that each component explicitly captures different aspects of the expression profile: the component related to $\mathbf{e}_1$ reflects the expression magnitude; the component associated with $\mathbf{e}_2$ represents the expression difference between $Y_i(1)$ and $Y_i(2)$; the component linked with $\mathbf{e}_3$ characterizes the expression of $Y_i(1)+Y_i(2)-2Y_i(3)$; etc.

We defined a distance measure based on the above transformation and the Poisson-based probability model. Under the Poisson model, $E(Z_{it}) = E(\boldsymbol{Y}_i)\mathbf{e}_t = (\lambda_i(1)\theta_i, \ \ldots, \lambda_i(T)\theta_i)\ \mathbf{e}_t$, $Var(Z_{it}) = (\lambda_i(1)\theta_i, \ \ldots, \lambda_i(T)\theta_i)\ \mathbf{e}_t^2$ and $Cov(Z_{it}, Z_{ik}) = 0$ when $t \neq k$. Then for a cluster consisting of tags 1, 2, …, m, we can measure the cluster dispersion by:

$$\begin{aligned} S_{trans\_N} &= \sum_i \sum_{t=1,..,T} \left(Z_{it} - E(Z_{it})\right)^2 \Big/ Var(Z_{it}) \\ &= \sum_i \sum_{t=2,..,T} \left(Z_{it} - \left(\hat{\lambda}(1)\hat{\theta}_i, ..., \hat{\lambda}(T)\hat{\theta}_i\right)\mathbf{e}_t\right)^2 \Big/ \left(\hat{\lambda}(1)\hat{\theta}_i, ..., \hat{\lambda}(T)\hat{\theta}_i\right)\mathbf{e}_t^2. \end{aligned} \tag{8}$$

We should note the connection between this measure and the *TransChisq* in (5). For the transformation in (7), as we have mentioned, the new component associated with $\mathbf{e}_2$ represents the expression difference between $Y_i(1)$ and $Y_i(2)$. However, the orthonormal eigenspace of a covariance matrix is not unique. An immediate alternative of the column vector $\mathbf{e}_2$ can be obtained by a row-switching transformation of $\mathbf{e}_2$, for which the associated component could represent the expression difference of any two original vector components, not necessary the $Y_i(1)$ and $Y_i(2)$. Under this consideration, the transformation associated with *TransChisq* is therefore equivalent to a transformation determined by all the possible row-switching vectors of $\mathbf{e}_2$.

This, however, also raises a shortcoming of S_{trans_N} in (8) and a not serious limitation of *TransChisq*. For S_{trans_N}, different eigenspaces could generate different values of S_{trans_N}. Though the problem could be overcome by using all the possible row-switching vectors of $\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_T$ (note that $\mathbf{e}_1$ is invariant under row-switching transformations) in principle, it is not computationally feasible since the computation cost increases exponentially with the dimension of the transformed data space. The problem of *TransChisq* lies in the use of only the row-switching vectors of $\mathbf{e}_2$, due to which it has the potential of losing the information carried by $\mathbf{e}_3$ or other eigenvectors. However, the application results showed that *TransChisq*

algorithm is efficient in terms of both performance and running time. This implies that the potential information loss in *TransChisq* is minor and could be ignored in most cases practically.

The distance measure based on the PCA data transformation: PCAChisq

We have explored new data transformation schemes useful for extracting desired information. For comparison, we also performed the Principal Components Analysis (PCA) [19]. PCA is a statistical technique for determining key features of a high dimensional dataset. In more detail, PCA approach uses the first few principal components (PCs), which are determined by the eigenspace of the sample covariance matrix, in data analysis. The first few PCs capture most of the variation in the original data set while the last few PCs are usually believed to capture only the residual noise in the data. Recently, PCA has been explored as a method for clustering gene expression data [23-29] and proved useful in simplifying the analysis of a high dimensional dataset in many situations. However, we should note that a blind application of PCA in clustering analysis could be dangerous, since PCA chooses principal component axes based on the empirical covariance matrix of overall data rather than class information, and thus it does not necessarily give good clustering results [30].

By substituting a set of PCs for the orthonormal eigenvectors $(\mathbf{e}_1\, \mathbf{e}_2 \ldots \mathbf{e}_T)$ in (8), we defined a new distance measure named *PCAChisq*. In some theoretical [31] and empirical [24] studies, it has been observed that the first few PCs determined by the empirical sample covariance matrix in PCA are not always helpful to extract biologically meaningful signals from data. Thus, we considered all PCs in this study unless *PCAChisq* shows better performance with the first few PCs. Results section gives some examples showing the positive and negative effects of applying PCA transformation.

Clustering analysis of microarray data

We would also like to explore the potential application of the proposed measures in clustering analysis of microarray data when Normal distribution is assumed and the Poisson-like property that variance increases with mean holds. Given a microarray dataset of expressions of n genes in T experiments, adopting the parameter notations in the Poisson model, we assumed that the expression of gene i in experiment t, $X_i(t)$, is normally distributed with mean $\mu_i(t) = \lambda_t(t)\theta_i$ and variance $\sigma_i^2(t) = k\lambda_t(t)\theta_i$, where k is a constant that can be estimated from data. The derivation of the maximum likelihood estimates (MLEs) of $\lambda_t(t)$ and θ_i under the normal model is rather involved. So we borrowed the estimators in (2); it can be shown that $\widehat{\theta}_i$ in (2) is unbiased and $\widehat{\lambda}_t$ in (2) is consistent under the above restricted normal model (See Appendix II). With $\widehat{\theta}_i$ and $\widehat{\lambda}_t$ available under the normal model, *TransChisq*, *PCAChisq* and *PoissonC* can then be employed in a clustering analysis of microarray data.

For both oligonucleotide and cDNA microarray data, the strong dependence of the

variance on the mean, in particular, variance increasing with mean, has been widely observed and studied [32-33]. Therefore, it is reasonable to expect that our restricted normal model is more or less valid and applicable to many microarray datasets. One example on the yeast sporulation dataset has been shown to demonstrate the power of *TransChisq* in analyzing microarray data (see Results and discussion section). But we should also note that *TransChisq* should not be used when the assumption on the relationship between the variance and the mean is seriously violated.

Authors' contributions

KK participated in the design of the study and performed the analysis and drafted the manuscript. SZ, KJ and LJF provided the Maize root microarray data, which helped in motivating this research. SZ, KJ and LJF were responsible for the biological explanations on the results related to maize data. LC provided the developing mouse retina SAGE data and was responsible to the biological explanations on the clustering results related to SAGE data. IBL helped in formulating PCA related studies. HH conceived of this study, and participated in its design and coordination, and conceptualized the framework of the method and wrote the paper. All authors read and approved the final manuscript.

Acknowledgements

The work of K. Kim was supported by Pohang University of Science and Technology (POSTECH), Korea and NIH R01GM075312. The work of H. Huang was supported by NIH R01GM075312.

References

1. Brazma A, Vilo J: **Gene expression data analysis.** *FEBS Lett* 2000, **480:**17-24.
2. Quackenbush J: **Computational analysis of microarray data.** *Nat Rev Genet* 2001, **2:**418-427.
3. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95:**14863-14868.
4. Hartigan JA: *Clustering algorithms*. New York: John Wiley & Sons, Inc; 1975.
5. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM: **Systematic determination of genetic network architecture.** *Nat Genet* 1999, **22:**281-285.
6. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR: **Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation.** *Proc Natl Acad Sci USA* 1999, **96:**2907-2912.
7. McLachlan GJ, Basford KE: *Mixture models: inference and applications to clustering*. New York: Dekker; 1988.
8. Banfield JD, Raftery AE: **Model-based Gaussian and non-Gaussian clustering.** *Biometrics* 1993, **49:** 803-821.
9. Fraley C, Raftery AE: **Model-based clustering, discriminant analysis and density estimation.** *Journal of the American Statistical Association* 2002, **97:** 611-631.
10. Tibshirani R, Walther G, Hastie T: **Estimating the number of clusters in a data set via the gap statistic.** *J R Statist Soc B* 2001, **63:**411-423.
11. Feher M, Schmidt JM: **Fuzzy clustering as a means of selecting representative conformers and molecular alignments.** *J Chem Inf Comput Sci* 2003, **43:**810-818.
12. Okada Y, Sahara T, Mitsubayashi H, Ohgiya S, Nagashima T: **Knowledge-assisted recognition of cluster boundaries in gene expression data.** *Artif Intell Med* 2005, **35:**171-183.
13. Baccelli F, Kofman D, Rougier JL: **Self organizing hierarchical multicast trees and their optimization.** *Proceedings of IEEE Inforcom'99* 1999, **3:**1081-1089.
14. Jia L, Bagirov AM, Ouveysi I, Rubinov AM: **Optimization based clustering algorithms in multicast group hierarchies.** In *Proceedings of the Australian Telecommunications, Networks and Applications Conference (ATNAC)*: 2003; Melbourne Australia (published on CD, ISNB 0-646-42229-4).
15. Bussermaker HJ, Li H, Siggia ED: **Regulatory element detection using correlation with expression.** *Nat Genet* 2001, **27:**167-174.
16. Friedman N, Linial M, Nachman I, Pe'er D: **Using Bayesian networks to analyze expression data.** *J Comput Biol* 2000, **7:**601-620.
17. Lazzeroni L, Owen A: **Plaid models for gene expression data.** *Statistica Sinica* 2002, **12:**61-86.

18. Cai L, Huang H, Blackshaw S, Liu JS, Cepko C, Wong WH: **Cluster analysis of SAGE data using a Poisson approach.** *Genome Biology* 2004, **5:**R51.
19. Jolliffe IT: *Principal Component Analysis*. New York: Springer-Verlag; 1986.
20. Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO, Herskowitz I: **The transcriptional program of sporulation in budding yeast.** *Science* 1998, **282:**699-705.
21. Jiang K, Zhang S, Lee S, Tsai G, Kim K, Huang H, Chilcott C, Zhu T, Feldman LJ: **Transcription profile analysis identify genes and pathways central to root cap functions in maize.** *Plant Molecular Biology* 2006, **60:**343-363.
22. Blackshaw S, Harpavat S, Trimarchi J, Cai L, Huang H, Kuo WP, Weber G, Lee K, Fraioli RE, Cho S-H, Yung R, Asch E, Ohno-Machado L, Wong WH, Cepko CL: **Genomic analysis of mouse retinal development.** *PLoS Biology* 2004, **2:**e247.
23. Raychaudhuri S, Stuart JM, Altman RB: **Principal components analysis to summarize microarray experiments: application to sporulation time series.** *Pac Symp Biocomput* 2000, **5:**452-463.
24. Yeung KY, Ruzzo WL: **Principal component analysis for clustering gene expression data.** *Bioinformatics* 2001, **17:**763-774.
25. Alter O, Brown PO, Botstein D: **Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms.** *Proc Natl Acad Sci USA* 2003, **100:**3351-3356.
26. Alter O, Brown PO, Bostein D: **Singular value decomposition for genome-wide expression data processing and modeling.** *Proc Natl Acad Sci USA* 2000, **97:** 10101-10106.
27. Holter NS, Mitra M, Maritan A, Cieplak M, Banavar JR, Fedoroff NV: **Fundamental patterns underlying gene expression profiles: simplicity from complexity.** *Proc Natl Acad Sci USA* 2000, **97:**8409-8414.
28. Bicciato S, Luchini A, Di Bello C: **PCA disjoint models for multiclass cancer analysis using gene expression data.** *Bioinformatics* 2003, **19:**571-578.
29. Misra J, Schmitt W, Hwang D, Hsiao L-L, Gullans S, Stephanopoulos G, Stephanopoulos G: **Interactive exploration of microarray gene expression patterns in a reduced dimensional space.** *Genome Res* 2002, **12:**1112-1120.
30. Komura D, Nakamura H, Tsutsumi S, Aburatani H, Ihara S: **Multidimensional support vector machines for visualization of gene expression data.** *Bioinformatics* 2005, **21:**439-444.
31. Chang W-C: **On using principal components before separating a mixture of two multivariate normal distributions.** *Appl Statist* 1983, **32:**267-275.
32. Durbin BP, Hardin JS, Hawkins DM, Rocke DM: **A variance-stabilizing transformation for gene-expression microarray data.** *Bioinformatics* 2002, **18:** S105-S110.
33. Rocke DM: **Heterogeneity of variance in gene expression microarray data.**

University of California at Davis, Department of Applied Science and Division of Biostatistics; 2003. [http://www.cipic.ucdavis.edu/~dmrocke/papers/empbayes2.pdf]

Figures

Figure 1 - Plots of 2092 maize genes on to the three different feature spaces.

(Top line) Genes are plotted on to the new space wherein each axis represents the gene expression difference in any two maize root tissues (*TransChisq* method); (Middle and Bottom lines) Genes are plotted on to the (d)(g) 1st and 2nd, (e)(h) 2nd and 3rd, and (f)(i) 1st and 3rd components by parametric covariance matrix (Middle line), and *PCAChisq* (Bottom line). Blue/red dots: RC up-/down-regulated genes, cyanide/pink dots: PM up-/down-regulated genes, green/orange dots: QC up-/down-regulated genes, respectively.

Figure 2 - Graphs of clustering results for the simulation data.

Horizontal axis represents the index of the 46 genes, which belong to six groups (named A, B, C, D, E and F) that are marked at the top of the figure; vertical axis represents the index of the cluster that each gene has been assigned by each algorithm.

Figure 3 - Average expression profiles for the 153 SAGE tags.

153 tags falling into 5 classes based on their biological functions are chosen from a developing mouse retinal SAGE dataset, and the average expression profiles for each class are shown. 125 of these genes are developmental genes grouped into four clusters (Early I, Early II, Late I and Late II) by their expressions at different developmental stages and the other 28 genes are un-related to the mouse retina development.

Figure 4 - Clustering results for the SAGE data.

Horizontal axis represents the index of the 153 tags, which belong to five groups (named Early I, Early II, Late I, Late II and Non.) that are marked at the top of the figure; vertical axis represents the index of the cluster that each gene has been assigned by each algorithm.

Figure 5 - Average expression profiles for the 39 representative genes in the yeast sporulation data.

39 representative genes are chosen from each of the seven expression patterns of the yeast sporulation data, and average expression profiles for each set are shown.

Figure 6 – Clustering results for the yeast sporulation data.

(a) Original expression profiles of the 39 representative genes from 7 functional groups in the yeast data. (b)-(f) Expression profiles of 7 groups after applying different clustering algorithm. The resulting plots by *Eucli* on rescaled data were too messy to present. The x-axis represents different time points of 0h, 0.5h, 2h, 5h, 7h, 9h, 11.5h; the y-axis represents the normalized log-ratio expression levels.

Tables

Table 1 - Six expression patterns of maize gene expression dataset and their separating regions described by PC2 and PC3.

Table 2 - Five dimensional simulation dataset with Normal distributions ($\sigma^2 = 3\mu$).

Table 3(a) - Functional categorization of the 153 mouse retinal tags (125 developmental genes; 28 non-developmental genes).

Table 3(b) - Comparison of algorithms on the 153 SAGE tags.

Table 4 - Comparison of algorithms on the 39 yeast sporulation genes.

Additional files

Appendix I - One set of orthonormal eigenvectors

This file contains the derived one set of orthonormal eigenvectors referred in the Method section.

Appendix II – Proof of the properties of Poisson parameters under normal model

This file shows the proof that the $\widehat{\theta}_i$ in (2) is an unbiased estimator of θ_i and $\hat{\lambda}(t)$ in (2) is a consistent estimator of $\lambda(t)$ under the suggested normal model.

Table 1. Six expression patterns of maize gene expression dataset and their separating regions described by PC2 and PC3.

Class index	Expression patterns	Center of separating regions described by PC2 and PC3
1	PM > (QC ≈ RC)	PC2 = $\sqrt{3} \cdot$ PC3 < 0
2	PM < (QC ≈ RC)	PC2 = $\sqrt{3} \cdot$ PC3 > 0
3	QC > (PM ≈ RC)	PC2 = $-\sqrt{3} \cdot$ PC3 > 0
4	QC < (PM ≈ RC)	PC2 = $-\sqrt{3} \cdot$ PC3 < 0
5	RC > (PM ≈ QC)	PC2 = 0; PC3 > 0
6	RC < (PM ≈ QC)	PC2 = 0; PC3 < 0

Table 2. Five dimensional simulation dataset with Normal distributions $\sigma^2 = 3\mu$.

Group ID		Mean parameters of the Normal distributions (μ)				
Group A	a1 ~ a3	1	1	1	15	150
Group B	b1 ~ b6	15	1	1	1	150
Group C	c1 ~ c4	10	30	30	60	10
	c5 ~ c6	100	300	300	600	100
Group D	d1 ~ d7	200	70	70	10	10
	d8 ~ d9	2000	700	700	100	100
Group E	e1 ~ e5	210	120	10	10	10
	e6 ~ e7	2100	1200	100	100	100
Group F	f1 ~ f3	5	50	5	5	5
	f4 ~ f6	5	75	5	5	5
	f7 ~ f9	5	100	5	5	5
	f10 ~ f11	50	500	50	50	50
	f12 ~ f13	50	750	50	50	50
	f14 ~ f15	50	1000	50	50	50

Table 3 (a). Functional categorization of the 153 mouse retinal tags (125 developmental genes; 28 non-developmental genes).

	Function Groups					Total
	Early I	Early II	Late I	Late II	Non-dev.	
Number of tags	32	34	32	27	28	153

Table 3 (b). Comparison of algorithms on the 153 SAGE tags

Algorithm	Number of tags in incorrect clusters	Percentage of tags in incorrect clusters
TransChisq	12	7.8
PCAChisq	12	7.8
PoissonC	22	14.4
PearsonC	26	17.0
Eucli on rescaled data	38	24.8
Eucli	NA	NA

Clusters generated by *Eucli* are too messy.

Table 4. Comparison of algorithms on the 39 yeast sporulation genes.

Algorithm	Number of genes in incorrect clusters	Percentage of genes in incorrect clusters
TransChisq	3	7.7
PCAChisq	14	35.9
PoissonC	7	18.0
PearsonC	13	33.3
Eucli	8	20.5
Eucli on rescaled data	17	43.6

Appendix I.

One set of T column eigenvector of the covariance matrix in (6) is given by

$$\begin{bmatrix}\mathbf{e}_1 & \cdots & \mathbf{e}_T\end{bmatrix}=\left[\begin{array}{c|ccc}1 & -1 & \cdots & -1\\ \hline 1 & & & \\ \vdots & & \mathbf{I}_{(T-1)\times(T-1)} & \\ 1 & & & \end{array}\right], \tag{S1}$$

where $\mathbf{I}_{(T-1)\times(T-1)}$ is an (T-1)-dimensional identity matrix. Orthonormal eigenvectors can be further obtained by applying the Gramm-Schmidt procedure which orthogonalize each eigenvector ($\mathbf{e}_i$) with respect to all the other eigenvectors to give $\mathbf{e}_{i\perp}$:

$$\mathbf{e}_{i\perp}=\frac{\mathbf{e}_i-\sum_{j=1}^{i-1}\left(\mathbf{e}_i\circ\mathbf{e}_{j\perp}\right)\mathbf{e}_{j\perp}}{\left|\mathbf{e}_i-\sum_{j=1}^{i-1}\left(\mathbf{e}_i\circ\mathbf{e}_{j\perp}\right)\mathbf{e}_{j\perp}\right|} \tag{S2}$$

Finally, we can obtain one set of orthonormal eigenvectors of (S1) as follows:

$$\begin{bmatrix}\mathbf{e}_{1\perp} & \mathbf{e}_{2\perp} & \mathbf{e}_{3\perp} & \cdots & \mathbf{e}_{T\perp}\end{bmatrix}=\begin{bmatrix}\frac{1}{\sqrt{T}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \cdots & \frac{1}{\sqrt{T(T-1)}}\\ \frac{1}{\sqrt{T}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \cdots & \frac{1}{\sqrt{T(T-1)}}\\ \frac{1}{\sqrt{T}} & 0 & -\frac{2}{\sqrt{6}} & \cdots & \frac{1}{\sqrt{T(T-1)}}\\ \frac{1}{\sqrt{T}} & 0 & 0 & \cdots & \frac{1}{\sqrt{T(T-1)}}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \frac{1}{\sqrt{T}} & 0 & 0 & \cdots & -\frac{T-1}{\sqrt{T(T-1)}}\end{bmatrix} \tag{S3}$$

For ease of notation, we used $\mathbf{e}_i$ instead of $\mathbf{e}_{i\perp}$ in the main text to denote the orthonormal eigenvectors. By randomly permuting the rows (performing row-switching transformations), we can obtain alternative orthonormal eigenspaces of the covariance matrix in (6).

Appendix II.

Let the expression of gene i at experiment t, $X_i(t)$, follow a Normal distribution with mean $\lambda(t)\theta_i$ and variance $k\lambda(t)\theta_i$, where k is a constant. We want to show that the $\hat{\theta}_i$ in (2) is an unbiased estimator of θ_i and $\hat{\lambda}(t)$ in (2) is a consistent estimator of $\lambda(t)$ under this normal model. By (2), $\hat{\theta}_i$ and $\hat{\lambda}(t)$ can be computed by

$$\hat{\theta}_i = \sum_{t=1}^{T} X_i(t), \quad \hat{\lambda}(t) = \frac{\sum_{i=1}^{n} X_i(t)}{\sum_{i=1}^{n}\sum_{t=1}^{T} X_i(t)}. \tag{S4}$$

Statement 1. $\hat{\theta}_i$ is unbiased.

Proof: $E\left(\hat{\theta}_i\right) = \sum_{t=1}^{T} E\left(X_i(t)\right) = \sum_{t=1}^{T}\left(\lambda(t)\theta_i\right) = \theta_i \sum_{t=1}^{T}\lambda(t) = \theta_i$. So $\hat{\theta}_i$ is an unbiased estimator of θ_i. ∎

Statement 2. $\hat{\lambda}(t)$ is a consistent estimator of $\lambda(t)$.

Proof: For $\hat{\lambda}(t)$ to be a consistent estimator of $\lambda(t)$, it is sufficient to show that $\hat{\lambda}(t) - \lambda(t)$ converges to 0 in probability. By (S4), we have

$$\hat{\lambda}(t) - \lambda(t) = \frac{\sum_{i=1}^{n} X_i(t) - \lambda(t)\sum_{i=1}^{n}\sum_{j=1}^{T} X_i(j)}{\sum_{i=1}^{n}\sum_{j=1}^{T} X_i(j)} = \frac{\frac{1}{n}\left[\sum_{i=1}^{n} X_i(t) - \lambda(t)\sum_{i=1}^{n}\sum_{j=1}^{T} X_i(j)\right]}{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{T} X_i(j)}. \tag{S5}$$

We first consider the numerator (M_n) of (S5).

$$\begin{aligned} E(M_n) &= \frac{1}{n}\sum_{i=1}^{n}\left[\lambda(t)\theta_i - \lambda(t)\left(\sum_{j=1}^{T}\lambda(j)\theta_i\right)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\lambda(t)\theta_i - \lambda(t)\theta_i\right] \\ &= 0 \end{aligned},$$

and

$$
\begin{aligned}
Var(M_n) &= \frac{1}{n^2}\sum_{i=1}^{n}\left[\left(1-\lambda(t)\right)^2 k\lambda(t)\theta_i + \lambda(t)^2\left(\sum_{j\neq t}^{T} k\lambda(j)\theta_i\right)\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\left[\left(1-\lambda(t)\right)^2 k\lambda(t)\theta_i + k\lambda(t)^2\left(1-\lambda(t)\right)\theta_i\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\left[k\left(1-\lambda(t)\right)\lambda(t)\theta_i\right] \\
&= \frac{k\left(1-\lambda(t)\right)\lambda(t)\theta_i}{n}
\end{aligned}
.
$$

So the numerator M_n converges to 0 in probability as n goes to infinity. Now we consider the denominator (D_n) in (S5). It is reasonable to assume that θ_i's are uniformly bounded. That is that there exists a positive real value A and B, such that $A \le |\theta_i| \le B$ for any i. Then we have

$$
0 < A \le E(D_n) = \frac{\sum_{i=1}^{n}\theta_i}{n} \le B, \quad Var(D_n) = \frac{\sum_{i=1}^{n} k\theta_i}{n^2} \le \frac{kB}{n} \quad \text{and} \quad \lim_{n\to\infty} Var(D_n) = 0.
$$

Consequently, for any $\varepsilon > 0$, we have

$$
\begin{aligned}
\lim_{n\to\infty} P\left(\left|\frac{M_n}{D_n}\right| > \varepsilon\right) &= \lim_{n\to\infty}\left(P\left(\left|\frac{M_n}{D_n}\right| > \varepsilon, |D_n - E(D_n)| > \frac{A}{2}\right) + P\left(\left|\frac{M_n}{D_n}\right| > \varepsilon, |D_n - E(D_n)| \le \frac{A}{2}\right)\right) \\
&\le \lim_{n\to\infty}\left(P\left(|D_n - E(D_n)| > \frac{A}{2}\right) + P\left(\left|\frac{M_n}{A/2}\right| > \varepsilon, |D_n - E(D_n)| \le \frac{A}{2}\right)\right) \\
&\le \lim_{n\to\infty}\left(\frac{Var(D_n)}{(A/2)^2} + P\left(|M_n| > \frac{A\varepsilon}{2}\right)\right) \\
&= \lim_{n\to\infty}\frac{Var(D_n)}{(A/2)^2} + \lim_{n\to\infty} P\left(|M_n| > \frac{A\varepsilon}{2}\right) \\
&= 0
\end{aligned}
,
$$

So $\hat{\lambda}(t) - \lambda(t) = \frac{M_n}{D_n}$ converges to 0 in probability as n goes to infinity, and then $\hat{\lambda}(t)$ is a consistent estimator of $\lambda(t)$. ∎