Statistica Sinica **25** (2015), 1459-1476
doi:http://dx.doi.org/10.5705/ss.2013.030

A NEW INFORMATION CRITERION BASED ON LANGEVIN MIXTURE DISTRIBUTION FOR CLUSTERING CIRCULAR DATA WITH APPLICATION TO TIME COURSE GENOMIC DATA

Xing Qiu, Shuang Wu and Hulin Wu

University of Rochester

Abstract: Common pre-processing procedures for time course microarray analysis such as standardization and gene filtering based on the functional F-test, often result in directional data that lie on a sphere S^{d-1}. While there have been some efforts in designing spherical clustering algorithms, few researchers have developed methods for selecting the number of clusters for spherical cluster analysis. In this paper, we focus on circular data on S^1 and propose a novel information-based criterion ICCC (information criterion for circular clustering) to determine the number of clusters when clustering circular data. This new criterion, ICCC, is based on a finite mixture model of Langevin distributions and is derived from the asymptotic properties of the maximum likelihood of the Langevin mixture distribution. Through the study of both simulated data and a large set of time course microarray data, we demonstrate that the ICCC criterion provides better estimates of the number of clusters than such existing methods: AIC, BIC, the Gap criterion, and the Maitra-Ramler criterion.

Key words and phrases: Circular statistics, clustering, information criterion, Langevin distribution, mixture model, model selection.

1. Introduction

It is well known that the immune response to viral (such as influenza) infection involves the activation, co-expression, and interaction of many genes. The emergence of large-scale time course gene expression data on influenza infection (Huang et al. (2011); Pommerenke et al. (2012)) presents an opportunity for researchers to understand how mammalian immune systems control the influenza infection. Although time course data are well studied in statistics, the “large p, small n” nature of these data presents a unique challenge. As an example, the data used in Huang et al. (2011) are collected from a cohort of 17 healthy human volunteers who received intranasal inoculation of influenza H3N2/Wisconsin. A total of $m = 11,961$ gene expression profiles were measured on whole peripheral blood drawn from each subject at 15-16 time points after inoculation, covering 108-132 hours.

There are two notable features about these data: for each subject, the temporal patterns of gene expressions can be grouped into several well-separated clusters; the heterogeneity of these temporal patterns between subjects is substantial. This high-level of between-subject variation was reported but not studied in detail in Huang et al. (2011). In this study, we focus on examining the subject-specific features of immune response. Specifically, we identify significant genes and cluster them into co-expressed modules for each subject. We then analyze the between-subject differences at the gene- and module-levels. A good cluster analysis is a critical step in this study.

Clustering analysis has been applied to microarray gene expression data since the dawn of microarray technologies (Eisen et al. (1998); Tavazoie et al. (1999)). As is typical, we apply such standard microarray pre-processing procedures as background correction, normalization, summarization, and logarithmic transformation prior to the statistical analysis. Since our aim is to produce *functionally related* gene groups (Eisen et al. (1998); Dortet-Bernadet and Wicker (2008)), a gene-wise standardization procedure (z-transformation) is applied before clustering analysis to ensure that every gene has mean 0 and variance 1 in the time direction. Although the between-gene correlation is invariant under this standardization, it has a profound impact on the clustering analysis because it alters the *geometry* of the sample space. The standardized gene expressions now reside in S^{J-2}, a sphere of dimension $J-2$, where J is the number of time points. Furthermore, the significance test we used in this study, as well as such other tests as t-tests or the ANOVA F-test, selects significant genes with high signal-to-noise ratio. This implies that the top two functional principal components (FPCs, Ramsay and Silverman (2002)) scores of significant genes reside in a one-dimensional sub-manifold, $S^1 \subset S^{J-2}$, approximately (see Section S6 of Supplementary Materials for more details). In other words, the standardized temporal gene expressions in our study can be considered as circular data that are essentially the measurement of directions.

Most classical clustering algorithms, such as the K-means algorithm and a multitude of its variants (MacQueen (1967); Tavazoie et al. (1999)) are designed for classifying a set of observations on the Euclidean space, and are known to perform poorly for circular/spherical data (Strehl, Ghosh, and Mooney (2000)). In recent years there has been increased interest in designing clustering algorithms specifically for spherical data (Banerjee et al. (2006); Dortet-Bernadet and Wicker (2008); Maitra and Ramler (2010)). One of the most popular methods is the spherical K-means (SK-means) clustering algorithm (Dhillon and Modha (2001); Banerjee et al. (2006); Maitra and Ramler (2010)), which replaces the Euclidean distance used in the K-means algorithm by the within-cluster *cosine similarity* that is more relevant for spherical data. It has been shown that the

K-means method is equivalent to a model-based probabilistic clustering algorithm, namely the Gaussian mixture model with isotropic and equal covariance structure (Celeux and Govaert, 1993). Likewise, the SK-means method has a probabilistic interpretation based on a finite mixture of Langevin distributions (*a.k.a.* von Mises-Fisher distribution) on S^{d-1}, the $(d-1)$-dimensional sphere embedded in $\mathbb{R}^d$ (Banerjee et al. (2006); Maitra and Ramler (2010)).

Standard model selection methods such as AIC and BIC have been widely used to determine the number of clusters K (Bozdogan and Sclove (1984); Fraley and Raftery (1998)). However, in practice we often observe systematically overestimated numbers of clusters when using both AIC and BIC. In a recent study of clustering time course gene expression data, we found that the first two FPCs of the standardized gene expression trajectories followed a circular distribution (see Figures S6, S15, and Section S6 of the Supplementary Materials for more details). Applying the SK-means algorithm to cluster these two FPCs, we observed that BIC kept decreasing until it failed to converge, at which point K exceeded 300. The failure of using BIC to select K in clustering spherical data has also been documented in Dortet-Bernadet and Wicker (2008).

Other methods for estimating the number of clusters include the Gap criterion (GAP) proposed by Tibshirani, Walther, and Hastie (2001). Although the theoretical derivation of GAP is based on the K-means method, the principle is flexible enough to be applicable to any clustering algorithms. The Gap criterion compares the change in within-cluster dispersion to that expected under an appropriate reference null distribution. Since the null distribution can be very complex and/or involve nuisance parameters, the authors recommended using the resampling method, which poses a significant computational burden. Maitra and Ramler (2010) developed a criterion (MR) specifically for the SK-means clustering algorithm based on the largest relative change in the locally optimized objective functions. This method is computationally efficient, but it is *ad hoc* and lacks theoretical justification.

We propose a new information criterion for selecting the number of clusters based on the likelihood of Langevin mixture distribution. We focus on the circular model (on S^1) to derive the new criterion, dubbed ICCC (information criterion for circular clustering), from the asymptotic properties of the maximum likelihood of the Langevin mixture distribution. ICCC measures the difference between the observed maximum log-likelihood and its expectation under the uniform distribution on S^1. It can be considered as an extension of the Gap criterion by using the log-likelihood of Langevin mixture distribution as the dispersion measure and the circular uniform distribution as the null distribution. But there are also significant differences between these two criteria. First, ICCC has an analytic formula and does not need to resort to a resampling method, so the computational cost is minimum compared to that of GAP. In addition, GAP is a stepwise procedure, so it may be trapped by a local maximum, leading to an underestimated number of clusters. Our new criterion is a global procedure and is superior to GAP in the performance of estimating the correct number of clusters, especially when the clusters are not well separated. This is further illustrated with simulation studies in Section 3 and an application to time course gene expression data from Huang et al. (2011) in Section 4. We summarize the conclusions and possible extensions of our method in Section 5. Proof of the main theorem and other auxiliary materials can be found in the Supplementary Materials.

2. Methods

2.1. The spherical K-means clustering algorithm

From this point on we focus on circular data on S^1. We denote by $\mathbf{X}_n = \{\mathbf{x}_i\}_{i=1}^n$, $\mathbf{x}_i \in S^1$ the set of circular observations. For $\mathbf{x}_i$ and $\mathbf{x}_{i'} \in S^1$, their similarity can be measured by $\cos\theta(\mathbf{x}_i, \mathbf{x}_{i'}) = \langle \mathbf{x}_i,\ \mathbf{x}_{i'}\rangle$, where $\theta(\mathbf{x}_i, \mathbf{x}_{i'})$ is the angle between $\mathbf{x}_i$ and $\mathbf{x}_{i'}$ and $\langle\cdot,\cdot\rangle$ is the inner product on $\mathbb{R}^2$. For a pre-specified number of clusters K, the SK-means algorithm finds a set of class indicators $\boldsymbol{\zeta} = \{\zeta_i\}_{i=1}^n$ to maximize the within-cluster cosine similarity:

$$\mathrm{CS}(\boldsymbol{\zeta}, \boldsymbol{\mu}_K, \mathbf{X}_n) = \sum_{i=1}^n \sum_{k=1}^K 1(\zeta_i = k)\langle \mathbf{x}_i,\ \mu_k\rangle, \tag{2.1}$$

where $\boldsymbol{\mu}_K = \{\mu_k\}_{k=1}^K$ and μ_k is the spherical center of the kth cluster, $\mu_k = \|\sum_{i=1}^n 1(\zeta_i = k)\mathbf{x}_i\|^{-1} \cdot \sum_{i=1}^n 1(\zeta_i = k)\mathbf{x}_i$.

The probabilistic interpretation of the SK-means method is provided as follows. Suppose $\mathbf{x}_i \in S^1$ is generated from a one-dimensional Langevin distribution with density function

$$f(\mathbf{x}|\mu_{\zeta_i}, \kappa) = c^{-1}(\kappa)e^{\kappa\langle \mathbf{x},\, \mu_{\zeta_i}\rangle} = \frac{e^{\kappa\langle \mathbf{x},\, \mu_{\zeta_i}\rangle}}{2\pi I_0(\kappa)}, \tag{2.2}$$

where μ_{ζ_i} is the mean direction, κ is the concentration parameter, and $I_0(\cdot)$ is the first kind of the modified Bessel function of order 0. When $\kappa = 0$, the Langevin distribution degenerates to the uniform distribution on S^1.

The joint log-likelihood of n independent observations can be written as

$$\log L(\boldsymbol{\mu}_K, \kappa \big| \mathbf{X}_n, K) = -n\left(\log(2\pi) + \log I_0(\kappa)\right) + \overbrace{\kappa \sum_{i=1}^n \sum_{k=1}^K 1(\zeta_i = k)\langle \mathbf{x}_i,\ \mu_k\rangle}^{=\kappa\cdot \mathrm{CS}(\boldsymbol{\mu}_K, \mathbf{X}_n)}. \tag{2.3}$$

From (2.3), it is apparent that maximizing $\log L$ is equivalent to maximizing (2.1). In fact, it has been shown by Banerjee et al. (2006) that the solution of SK-means problem is the maximum likelihood estimate (MLE) of a mixture model based on Langevin distributions with K different angular centers but the same dispersion parameter. This deep connection enables us to derive information-theoretic criteria to select K for the SK-means algorithm based on the asymptotics of the maximum likelihood.

2.2. A new information criterion for circular cluster analysis

In this section, we study the asymptotic properties of the log-likelihood function (2.3) under the uniform distribution on S^1 and use it to build a new model selection criterion.

Since the SK-means clustering has a probabilistic interpretation, ideally we should follow the derivation of the AIC formula to find an unbiased estimator of $D_{KL}(P^*, \hat{P}(K))$, the Kullback-Leibler divergence between P^*, the true model, and $\hat{P}(K)$, the model estimated by the maximum likelihood method with K clusters. However, this approach is quite challenging because the true distribution P^* is unknown and many asymptotic techniques Akaike employed in his seminal work (Akaike (1973)) do not work due to the differences between $\mathbb{R}^d$ and S^1.[1] We adopt an alternative approach to construct the new criterion ICCC based on the asymptotic property of the maximum likelihood $\hat{L}_n(K)$ under the null hypothesis.

Since S^1 is the compact Lie group of rotations of $\mathbb{R}^2$, a good cluster analysis should be equivariant under these rotations. From this point of view, the circular uniform distribution, which is the Haar measure of S^1, serves as a natural candidate for the null distribution on S^1. We denote this null hypothesis by H_0. This distribution is simple and free of nuisance parameter. As a comparison, the symmetry group of $\mathbb{R}^d$ is the Euclidean group E_d and it contains the translational group T_d as a subgroup; T_d is not compact so no probability measure is invariant under T_d, and there is no "natural" null distribution on $\mathbb{R}^d$. This explains why a resampling method is needed to compute the Gap statistic for the K-means clustering algorithm.

Our main theorem states the asymptotic behavior of $\log \hat{L}_n(K)$ under H_0.

Theorem 1. *Under H_0, for a given large K, the observed maximum likelihood converges to a constant $G(K)$ with the approximation*

$$\frac{1}{n}\log \hat{L}_n(K) \xrightarrow{a.s.} G(K) \approx \log K - \frac{1}{2} - \frac{1}{2}\log\frac{2\pi^3}{3}. \tag{2.4}$$

[1] For those who are interested in this topic, we conducted a simulation study which illustrates the key differences between $\mathbb{R}^d$ and S^1 in Section 2.2 of Supplementary Materials.

Corollary 1. *For large K and n, we have*

$$\frac{1}{n}E\left(\log \hat{L}_n(K)\middle| H_0\right) \approx \log K + \text{Const}. \tag{2.5}$$

The proof of Theorem 1 can be found in Section S5 of the Supplementary Materials.

Based on (2.4) and (2.5), the penalty term must be $n \log K$ to "offset" the artificial gain of $\log \hat{L}_n(K)$ when K is large. Therefore, we propose a new model selection criterion ICCC (information criterion for circular clustering):

$$\text{ICCC}(K) = -2\log \hat{L}_n(K) + 2n\log K. \tag{2.6}$$

Equation (2.6) measures the difference between the maximum log-likelihood of the observed data and its expectation under the null hypothesis. ICCC differs from the Gap criterion proposed by Tibshirani, Walther, and Hastie (2001) in several ways. First, we use $\hat{L}_n(K)$ instead of the within-cluster cosine similarity as the similarity measure. Second, instead of using a resampling method to approximate $E\big(\log \hat{L}_n(K)|H_0\big)$, the expected within-cluster dispersion under the null hypothesis, we provide an analytical formula, that is computationally efficient. Third, GAP is a stepwise procedure and ICCC is a global optimization method. The optimal number of clusters by Gap is the smallest K such that $\text{Gap}(K) \geq \text{Gap}(K+1) - S_{K+1}$, where S_K is the standard deviation of $\log \hat{L}_n(K)$ under H_0. As pointed by Tibshirani, Walther, and Hastie (2001), the Gap statistic is good at identifying well-separated clusters, in which case one expects to observe a sharp increase in $\text{Gap}(K)$ when reaching the optimal K. However, if some clusters are not well separated, the Gap statistic may be trapped in local maximum and underestimate the number of clusters. Since ICCC requires minimum computations, we are able to find the global maximum of $\text{ICCC}(K)$ and therefore improve the likelihood of selecting the correct K.

3. Simulation Studies

Five sets of simulation studies (**SIM.K1**, **SIM.K5**, **SIM.K25**, **SIMBIO.A**, and **SIMBIO.B**), were conducted to compare the performance of ICCC and several other popular model selection methods in the context of circular clustering.

Each simulated data in **SIM.K1**, **SIM.K5**, and **SIM.K25** contains 1,000 observations, denoted by $\mathbf{x}_i := (x_{i1}, x_{i2})$, $i = 1, \ldots,$ 1,000, where the $\mathbf{x}_i$ were generated by adding bivariate Gaussian noise on the circular observations from Langevin distribution on S^1:

$$\mathbf{x}_i = R\mathbf{v}_i + \boldsymbol{\epsilon}_i, \qquad \mathbf{v}_i \sim M(\mu_{\zeta_i}, \kappa), \quad \boldsymbol{\epsilon}_i \sim MVN\left(\mathbf{0},\, \sigma_\epsilon^2 \mathbf{I}_{2\times 2}\right). \tag{3.1}$$

Here $M(\mu_{\zeta_i}, \kappa)$ refers to the Langevin distribution with angular center μ_{ζ_i} and concentration parameter κ, and R is the radius of the circle. Without loss of generality, we took $R = 1$ for all simulations. These simulated data were standardized to have mean 0 and variance 1 for each gene.

The parameters used for the three studies were as follows.

SIM.K1: $\mathbf{v}_i \sim M(0,0) = \mathrm{Unif}(S^1)$, $\sigma_\epsilon^2 = 0.1^2$, so all observations in this data set belonged to one uninformative cluster.

SIM.K5: 1,000 observations were divided into $K = 5$ clusters with angular centers $(\theta(\mu_1), \ldots, \theta(\mu_5)) = (30°, -30°, 150°, 180°, 200°)$ and cluster sizes $(n_1, \ldots, n_5) = (130, 250, 200, 220, 200)$, respectively. The concentration parameter for each cluster was $\kappa = (20/\pi)^2$ and the variance of Gaussian noise was $\sigma_\epsilon^2 = 0.1^2$.

SIM.K25: 1,000 observations were divided into $K = 25$ clusters with each $n_k = 40$, $k = 1, \ldots, 25$. The angular centers of these clusters formed an equi-distant grid on S^1, namely $\theta(\mu_k) = 2k\pi/25$, for $k = 1, \ldots, 25$. The concentration parameter for each cluster was $\kappa = (100/\pi)^2$ and the variance of Gaussian noise was $\sigma_\epsilon^2 = 0.05^2$.

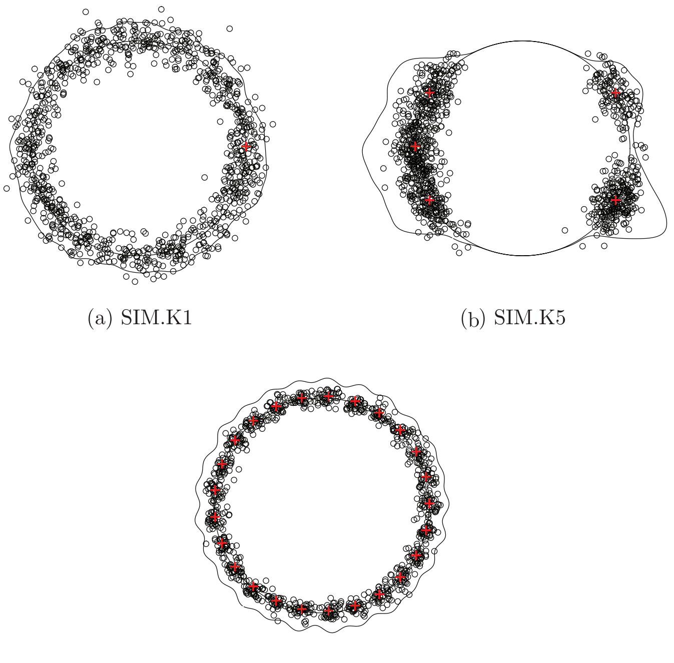
Scatter plots of the simulated data under these scenarios are shown in Figure 1. **SIM.K5** represents a typical clustering problem that commonly arises in applications. There are a few clearly visible clusters, such as the well-separated left and right clusters and the two clusters on the right, while some clusters are hard to be separated visually, such as the three clusters on the left. In **SIM.K25**, we chose a small σ_ϵ^2 so that the neighboring clusters were still distinguishable when K is large. This represents a "large K and small noise" structure.

The simulation studies **SIMBIO.A**, and **SIMBIO.B** were designed to match the time course microarray data (see Section 4 for more details). These data were generated by imposing random signals on five true cluster mean curves. The random signals were generated based on Subject 11 because the scatter plot of this subject (Figure 4(c)) resembles one uninformative cluster. More specifically, we first generated the expression levels w_{ij} for $i = 1, \ldots, 200$ genes and $j = 1, \ldots, 15$ time points according to the model

$$w_{ij} = M_{k(i)}(t_j) + z_i(t_j) + \varepsilon_{ij}. \tag{3.2}$$

Here $k(i)$ represents the cluster to which the ith gene belongs. We divided 200 genes into five clusters, each containing 40 genes. $M_{k(i)}(t_j)$ was the value of true mean curve of the kth cluster measured at time t_j, $z_i(t_j)$ was a random continuous temporal function, and the $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$ were *i.i.d.* random noise.

To best match the data, we randomly selected five genes from Subject 11 and used their smoothed temporal expression curves ($y_k(t)$, $k = 1, 2, \ldots, 5$) to



(a) SIM.K1

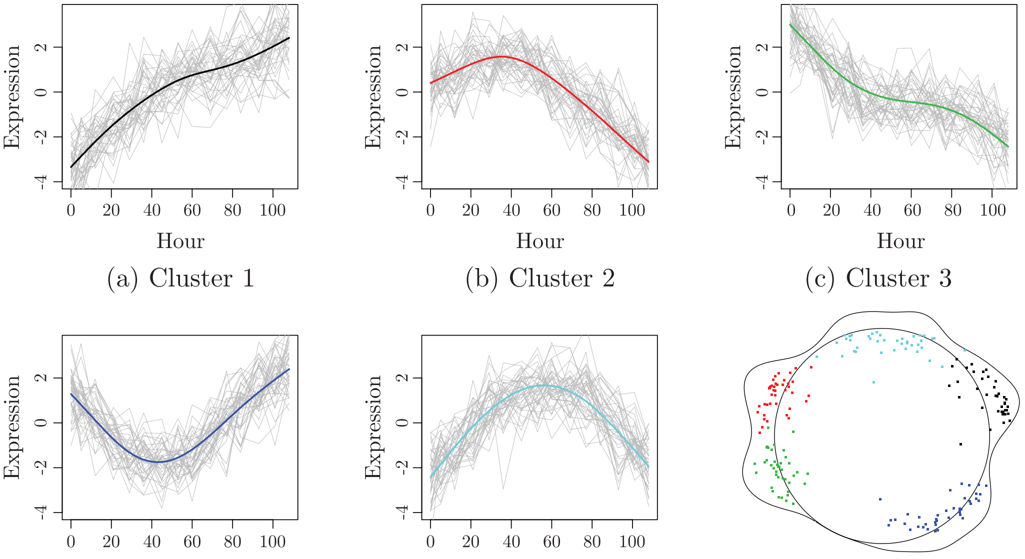
(b) SIM.K5

(c) SIM.K25

Figure 1. Circular density plots of simulated data. Red crosses represent the circular centers of the clusters. Empirical circular density functions are shown for better visual effects.

construct cluster mean curves. To ensure the signal-to-noise ratio was large enough for clustering analysis (none of these five randomly selected genes is significant in Subject 11), we took $M_k(t) = c{\cdot}y_k(t)$, where $c = 2.0$ for **SIMBIO.A** and $c = 1.6$ for **SIMBIO.B**. $z_i(t_j)$ was generated by randomly sampling (without replacement) from 200 smoothed temporal expression curves. We used the sample variance of the residuals from nonparametric smoothing of Subject 11 as the variance of random noise, $\sigma_\varepsilon^2 = 0.75$. These data were standardized to have mean 0 and variance 1 for each gene.

We used the estimated *percentage of signal*, $p_{\text{sig}} = \widehat{\sigma}_M^2/(\widehat{\sigma}_M^2 + \widehat{\sigma}_z^2 + \sigma_\varepsilon^2)$, to quantify the signal to noise ratio for these time course data. Here $\widehat{\sigma}_M^2$ and $\widehat{\sigma}_z^2$ are the mean sample variance of $M_{k(i)}(t_j)$ and $z_i(t_j)$, respectively. For **SIMBIO.A**, $p_{\text{sig}}^A = 0.721$. For **SIMBIO.B**, $p_{\text{sig}}^B = 0.624$. Next, we applied functional principal component analysis (fPCA, Ramsay and Silverman (2002)) to the data. Each gene was represented by the first two FPCs, $\vec{s}_i = (s_{i1}, s_{i2})$. Due to the nature of fPCA, the percentage of variance explained by the first two eigen-functions depends on the penalty (smoothing) parameter used in fPCA and is usually much higher than the corresponding multivariate PCA. In **SIMBIO.A**, the penalty



(a) Cluster 1

(b) Cluster 2

(c) Cluster 3

(d) Cluster 4

(e) Cluster 5

(f) FPC scores

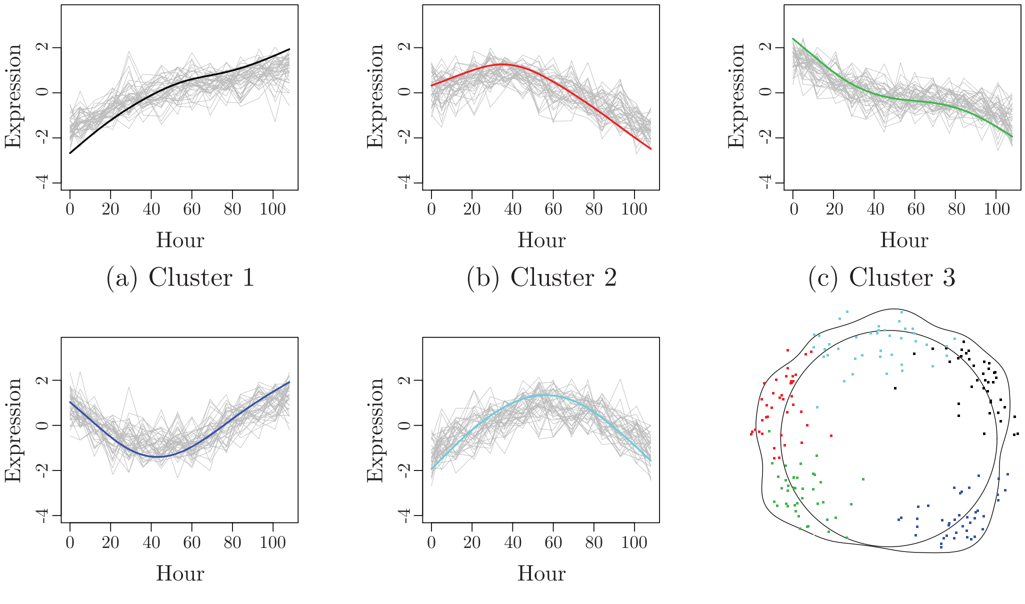
Figure 2. (a)−(e): Spaghetti plots of five clusters. Solid curves in the middle represent $M_k(t)$, true cluster mean curves. (f): FPC scores of these genes. These scores are coded in the same colors of cluster mean curves to which they belong. Empirical circular density functions are shown for better visual effects. Data used: **SIMBIO.A**.

parameter determined by the GCV principle was $\lambda = 10,000$, and the first two eigen-functions explained 99.3% of total variance. In **SIMBIO.B**, we manually set $\lambda = 10$, and the first two eigen-functions only explained 74.8 % of total variance. These data are illustrated by Figures 2 and 3, respectively. From these figures, it is clear that **SIMBIO.B** has weaker signals and is more difficult to cluster than **SIMBIO.A**.

We used the SK-means algorithm implemented in the `R` package `skmeans` (Hornik, Feinerer, and Kober (2012)) to cluster the simulated data. This package implements a genetic algorithm patterned after the genetic k-means algorithm described in Krishna and Murty (1999). ICCC and several existing model selection procedures, including AIC, BIC, GAP and MR, were used to determine the optimal number of clusters K. The upper limit of clusters was set to be 20 for **SIM.K1** and **SIM.K5** and 40 for **SIM.K25**. Each simulation was repeated for 100 times.

The MR criterion chooses K to maximize

$$\mathrm{MR}(K) = \frac{\mathrm{Obj}(K+1)}{\mathrm{Obj}(K)} - \frac{\mathrm{Obj}(K)}{\mathrm{Obj}(K-1)}, \tag{3.3}$$



(a) Cluster 1 (b) Cluster 2 (c) Cluster 3

(d) Cluster 4 (e) Cluster 5 (f) FPC scores

Figure 3. (a)−(e): Spaghetti plots of five clusters. Solid curves in the middle represent $M_k(t)$, true cluster mean curves. (f): FPC scores of these genes. These scores are coded in the same colors of cluster mean curves to which they belong. Empirical circular density functions are shown for better visual effects. Data used: **SIMBIO.B**.

where $\mathrm{Obj}(K) = n - \mathrm{CS}(\boldsymbol{\mu}_K, \mathbf{X}_n)$. As a special case, Obj(0) is set to be $2n$, the expected value of Obj(0) under the uniform distribution on S^1. The Gap statistic is

$$\mathrm{Gap}(K) = E\left(\log \mathrm{Obj}(K)|H_0\right) - \log \widehat{\mathrm{Obj}}(K), \tag{3.4}$$

where $\widehat{\mathrm{Obj}}(K) = n - \mathrm{CS}(\hat{\boldsymbol{\mu}}_K, \mathbf{X}_n)$. The optimal number of clusters is the smallest K such that

$$\mathrm{Gap}(K) \geq \mathrm{Gap}(K+1) - S_{K+1}, \tag{3.5}$$

where $S_K = \hat{\sigma}\left(\log \mathrm{Obj}(K)|H_0\right)$ and H_0 is the uniform distribution on S^1.

The results are reported in Table 1. Graphical illustrations of each model selection methods for one simulated dataset and the results of SK-means cluster analysis can be found in Figures S1 to S5 of the Supplementary Materials.

From Table 1, we can see that ICCC is better than other criteria under all simulation scenarios in terms of mean square error (MSE). The classical information criteria AIC and BIC failed; they selected the upper limits of clusters invariably, even for the uninformative case **SIM.K1**. We think that this is because these methods are not designed for circular cluster analysis.

Table 1. Mean and root mean square error (in parenthesis) of the estimated number of clusters. The first column shows the true number of clusters. Number of repetitions: 100.

	True K	Estimates				
		AIC	BIC	ICCC	GAP	MR
SIM.K1	1	20 (19)	20 (19)	1 (0)	1.18 (0.4)	2.77 (2.8)
SIM.K5	5	19.98 (15.0)	19.97 (15.0)	4.98 (0.1)	2 (3)	2.07 (3.0)
SIM.K25	25	39.92 (14.9)	39.90 (14.9)	26.81 (2.2)	1 (24)	26.39 (2.2)
SIMBIO.A	5	10 (5)	10 (5)	5.2 (0.45)	4.35 (1.45)	4.76 (1.29)
SIMBIO.B	5	10 (5)	10 (5)	5.09 (0.48)	2.26 (2.82)	4.79 (1.03)

GAP works well for **SIM.K1** but underestimates K for the other two cases, probably due to the stepwise nature of this procedure. Judging from Figures S2(d) and S3(d) in the Supplementary Materials, the Gap statistic is maximized at $K = 5$ and $K = 27$ for **SIM.K5** and **SIM.K25**, respectively. So a "global" GAP that selects the global maximum of the Gap statistic would work much better than its stepwise counterpart for these two scenarios. However, this global GAP criterion would overestimate the number of clusters for **SIM.K1**. In fact, the sample mean of K (averaged over 100 repetitions) selected by the global GAP for **SIM.K1** is 15.93, indicating that the global GAP is useless in this case.

MR worked very well for **SIM.K25** but it overestimated K for **SIM.K1** and underestimated K for **SIM.K5**. Strictly speaking, $\mathbf{MR}(K)$ is only well defined for $K \geq 2$. When $K = 1$, $\mathbf{MR}(1)$ depends on $\mathrm{Obj}(0) \equiv 2n$, which is prespecified and may not be reliable in practice. This explains why MR worked poorly for **SIM.K1**; this limitation is also well documented in Maitra and Ramler (2010). Moreover, MR is based on the ratio of the within-cluster similarity for consecutive K's. In **SIM.K5**, there are large changes in the within-cluster similarity when K increases from 1 to 2 and also from 2 to 3, but the changes are not prominent when K increases from 4 to 5 because the left three clusters are not well separated (see Figure S2(e) in Supplementary Materials). So MR tended to select a smaller number of clusters in **SIM.K5**.

In **SIMBIO.A** and **SIMBIO.B**, AIC and BIC failed again. ICCC and MR worked well in both cases; GAP worked well in **SIMBIO.A** but underestimated K for **SIMBIO.B**, in which the percentage of signal was smaller. In both cases, ICCC performs better than MR and GAP in terms of MSE, but its advantage was less than in the previous simulation studies because **SIMBIO.A** and **SIMBIO.B** were generated by resampling the time course microarray data, so the distribution of the first two FPCs (Figures 2(f) and 3(f)) follow the Langevin distribution only approximately. Since ICCC is derived from the Langevin distribution, the fact that it outperformed GAP and MR, neither of which depends on the parametric assumptions of the underlying distribution, shows that ICCC is robust to certain deviations from the Langevin model. Overall, it was the best method for estimating the number of clusters in all five simulation examples.

4. Analysis of Human Influenza Challenge Data

We illustrate the proposed ICCC model selection criterion with an application to the time course microarray data in Huang et al. (2011). In this study, a cohort of 17 healthy human volunteers received intranasal inoculation of influenza H3N2/Wisconsin and 9 of them developed mild to severe symptoms. A total of $m = 11,961$ gene expression profiles were measured on whole peripheral blood drawn from each subject at $J = 15$ time points after inoculation, covering about 108 hours. For the ith gene, we consider the expression measurements w_{ij} as noisy observations of the underlying true expression curve $y_i(t)$ as

$$w_{ij} = y_i(t_j) + \epsilon_{ij}, \qquad i = 1, \ldots, m, \quad j = 1, \ldots, J, \tag{4.1}$$

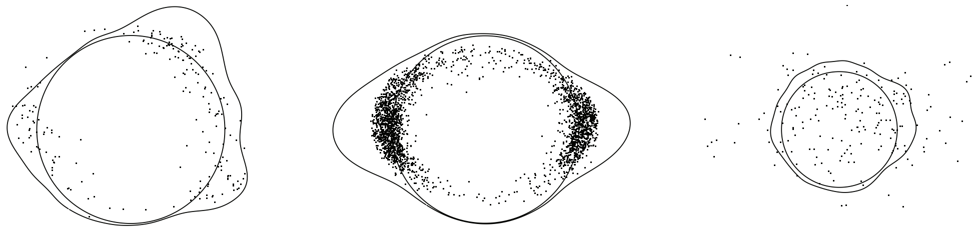
where t_j is the jth time point and the ϵ_{ij} are the noisy signals. The measurements are standardized to have mean 0 and variance 1 for each gene.

We first applied penalized B-splines to estimate $\hat{y}_i(t)$ from the noisy microarray observations and conducted functional F-test (Storey et al. (2005)) to identify differentially expressed genes. The multiple testing adjustment procedure proposed by Benjamini and Hochberg (1995) was then used to control the false discovery rate (FDR) at 0.05 level. When less than 200 genes were selected as significant, we included 200 top-ranked genes in the subsequent clustering analysis.

For the sake of clarity, we only present the results of three representative subjects: Subject 2 (asymptomatic, 22 significant genes, and 178 other top-ranked genes), Subject 10 (symptomatic, with 2,504 significant genes), and Subject 11 (asymptomatic, 1 significant gene, and 199 other top-ranked genes) in the main text. More technical details and the cluster analysis results of other subjects can be found in Section S3 of Supplementary Materials.

We applied fPCA to the selected genes for each subject. The first two functional principal components (FPCs) were chosen, explaining 97.74%, 99.04%, and 94.15% of the total variation for Subjects 2, 10, and 11, respectively. Each gene was represented by the first two FPCs, $\vec{s}_i = (s_{i1}, s_{i2})$.

The scatter plots of the first two FPCs for these subjects are displayed in Figure 4. The scatter plots of Subjects 2 and 10 exhibit circular patterns and this pattern is much more pronounced in Subject 10 (symptomatic) than Subject 2 (asymptomatic). The scatter plot of Subject 11 does not follow a circular pattern, but we can consider it as an example for the uniformly distributed data on S^1. We also observe that all symptomatic subjects and most asymptomatic subjects have circular shaped principal component scores, but generally the circular patterns



(a) Subject 2, Asymptomatic. (b) subject 10, Symptomatic. (c) subject 11, Asymptomatic.

Figure 4. Scatter plots of the first two principal component scores for Subjects 2, 10, and 11. Empirical circular densities are shown for better visual effects. Reasonable numbers of clusters (by visual examination): $K = 3$ for Subject 2, $K = 2$ for Subject 10, and $K = 1$ for Subject 11.

are much clearer for the symptomatic subjects than those of the asymptomatic ones (see Figures S7 to S20 in Supplementary Material). A natural question arises: Why do the principal component scatter plots show circular pattern, and why is this pattern clearer for the symptomatic subjects than the asymptomatic ones? The answer to this question is surprisingly nontrivial and we defer the related discussions to Section S6 of Supplementary Materials.

We applied the SK-means algorithm to cluster the first two FPCs for these three subjects. The new criterion ICCC, together with AIC, BIC, GAP, and MR are used to select the number of clusters. The results of each model selection method are displayed in Figures S8, S15, and S16 of the Supplementary Materials, respectively. As in the simulation studies, AIC and BIC kept decreasing and failed to perform model selection for these subjects. We focus on the comparison between GAP, MR, and ICCC.

For Subject 2, both MR and ICCC selected $K = 3$ but GAP selected $K = 1$. If we use the global GAP criterion, it also selects $K = 3$. It is clear from Figure 4(a) that $K = 3$ is more reasonable than $K = 1$. For Subject 10, ICCC, GAP (both the stepwise and global versions), and MR all selected two clusters, which is in accordance with visual examination of Figure 4(b). This means that for very well defined clusters, all three criteria worked well. For Subject 11, ICCC selected $K = 1$ but GAP (both the stepwise and global versions) and MR selected $K = 2$. Judging from Figure 4(c), all points appear to lie on a large cluster, so we believe that $K = 1$ is a more reasonable choice.

Finally, we conducted functional enrichment analyses on the classified gene clusters by DAVID (Huang, Sherman, and Lempicki (2009)). Specifically, for each subject, we identified the enriched functions and pathway annotations of each cluster in Gene Ontology (Ashburner et al. (2000)), KEGG (Ogata et al. (1999)) and REACTOME (Joshi-Tope et al. (2005)) curated pathway databases.

The Bonferroni multiple testing procedure was applied to control the familywise error rate at level 0.05.

For Subject 10 (Symptomatic), a total of 118 significant pathways (see Table S3 the Supplementary Materials) were identified based on ICCC, GAP, and MR. For Subjects 2 and 11 (Asymptomatic), only ten significant pathways (Table S4 in the Supplementary Material) were identified based on ICCC and seven significant pathways (Table S5 in the Supplementary Material) were identified based on the alternative model selection criteria. The results from Subject 10 contained much richer information about immune response than that from the asymptomatic subjects, which is as predicted and in accordance with Huang et al. (2011). The biological implication is that the development of influenza symptom is driven by a complex biological procedure characterized by the mobilization of many pathways.

Overall, the clusters determined by ICCC provide clearer decomposition, and related genes are grouped in pathways with more specific immune responses functions. For example, for Subject 2, GO:0042742 (defense response to bacterium) was identified based on ICCC/MR, which is more specific than GO:0006952 (defense response) identified based on GAP. Three important pathways were enriched by the ICCC/MR approach only. Among them the most interesting one is hsa04622 (RIG-I-like receptor signaling pathway) because it was not enriched in the symptomatic subject (Subject 10). This pathway is responsible for making proteins (RIG-I, MDA5, and LGP2) that are vital for the synthesis of type I interferon and other inflammatory cytokines upon recognition of viral nucleic acids. The lack of the activation of this pathway in Subject 10 is conspicuous. For Subject 11, the enriched terms based on ICCC ($K = 1$) are exactly the same as those based on GAP/MR ($K = 2$). ICCC provided an identical but more parsimonious model than GAP and MR in this case. More discussions on these analyses can be found in Section S4 of Supplementary Materials.

5. Discussion

Cluster analysis is a powerful tool to reduce the complexity of large, high-dimensional data. Common pre-processing procedures employed in time course microarray analysis, such as standardization and gene filtering based on the functional F-test, often result in data that reside on a sphere. Such data are essentially directional data, meaning that the direction of the data vector is relevant, not its magnitude. Specialized cluster analysis method such as SK-means is most appropriate for such data.

A crucial element of a good cluster analysis is a good estimate of the number of clusters K. While there has been a considerable literature on this when clustering data on an Euclidean space, there has been very little work related to clustering spherical data. Classical model selection methods such as AIC and BIC do not work properly for spherical data because they tend to under-penalize the log-likelihood, as shown in our numerical examples. More specifically, AIC and BIC, as well as many other model selection methods such as AICc (Hurvich and Tsai (1989)) and MDL (minimum description length, Hansen and Yu (2001)) have the form

$$I = -2 \log \hat{L}_n(K) + \alpha(n)K, \tag{5.1}$$

where $\alpha(n)$ is either of order $O(1)$ (AIC, AICc, and MDL) or $\log n$ (BIC). The penalty term of ICCC is of order $n \log K$, which provides enough penalty for circular cluster analysis.

One culprit for the under-performing of AIC and BIC is the mixture model nature of SK-means. It has been known for a long time that the AIC/BIC formula do not hold in theory for finite mixture models (McLachlan and Peel (2000)) because regularity conditions do not hold for these models. Then too, $\mathbb{R}^d$ and S^1 have very different geometric properties which are reflected in the fundamentally different probabilistic models for the K-means and SK-means algorithms.

We have developed a novel model selection criterion to select K for clustering circular data, using the fact that the SK-means method is equivalent to a generative model consisting of a mixture of the Langevin distribution. Dubbed information criterion for circular clustering (ICCC), this criterion is derived from the asymptotic property of the maximum likelihood of the Langevin mixture distribution. The computation of ICCC is quite easy, which enables the selection of a globally optimal K. Through the study of simulated data and a time course microarray data set, we show that ICCC produces better estimates of K than such other existing methods as GAP and MR.

A natural next step is to extend ICCC to high-dimensional spheres (S^{d-1}, $d \geq 3$). This is not trivial. The derivation of ICCC depends on the "midpoint rule" in Lemma S5.1 and the convergence results in Lemma S5.3 which basically says that when $n \to \infty$, the best clustering of points generated from uniform distribution is an even partition of S^1.

The partition of S^1 determined by a set of pre-determined centers and the midpoint rule has a clear analogy on S^{d-1}, namely the *Dirichlet cells* (or Voronoi cells) of S^{d-1}. Although these cells have been studied in the community of computational geometry for a long time (Okabe et al. (1992)), their large sample properties under the uniform distribution of the centers are currently unknown. The main difficulty is that there is no clear analogy of equi-distance partition on higher dimensional spheres. Taking S^2 as an example, it is not possible to divide it into K polygons with exactly the same shape. Euler's formula dictates that the vertices and edges of a partition of a sphere must satisfy certain constraints (Saff and Kuijlaars (1997)). For S^2, empirical evidences suggest that for large K, most Dirichlet cells are hexagons and a handful are pentagons (Saff and Kuijlaars (1997)). To the best of our knowledge, there is no general theory for the asymptotic behavior of Dirichlet cells for S^{d-1}.

A weak analogy of an equi-distance partition of S^1 on S^2 is a partition such that the smallest distance between the centers of parts is maximized. This is known as *Tammes's problem* and is also a member of the *packing problems* which are among the most active and challenging research areas in mathematics, see Conway, Sloane, and Bannai (1999).

We believe that our work represents a starting point for developing the right tools for mixture-model based cluster analysis on a manifold. Asymptotic properties of the likelihood function derived from manifold-valued models can be very different from those derived from multivariate Gaussian distributions on $\mathbb{R}^n$. We expect our work on S^1 can serve as a foundation for a unifying theory which is applicable for higher-dimensional spheres, or even more general compact manifolds.

Acknowledgement

This research is partially supported by NIH HHSN272201000055C, 5 R01 AI087135, and 2 R01 HL062826-09A2 from the National Institutes of Health; and the University of Rochester CTSI pilot award (UL1RR024160) from the National Center For Research Resources and the National Center for Advancing Translational Sciences of the National Institutes of Health. We would like to thank Dr. Zhi-Ping Liu for his valuable comments and suggestions.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory.* Springer Verlag **1**, 267-281.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**, 25-29.

Banerjee, A., Dhillon, I., Ghosh, J. and Sra, S. (2006). Clustering on the unit hypersphere using von Mises-Fisher distributions. *J. Mach. Learn. Res.* **6**, 1345.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57**, 289-300.

Bozdogan, H. and Sclove, S. (1984). Multi-sample cluster analysis using Akaike's information criterion. *Ann. Inst. Statist. Math.* **36**, 163-180.

Celeux, G. and Govaert, G. (1993). Comparison of the mixture and the classification maximum likelihood in cluster analysis. *J. Statist. Comput. Simulation* **47**, 127-146.

Conway, J., Sloane, N., and Bannai, E. (1999). *Sphere Packings, Lattices, and Groups.* **290** Springer Verlag.

Dhillon, I. and Modha, D. (2001). Concept decompositions for large sparse text data using clustering. *Machine learning* **42**, 143-175.

Dortet-Bernadet, J. and Wicker, N. (2008). Model-based clustering on the unit sphere with an illustration using gene expression profiles. *Biostatistics* **9**, 66.

Eisen, M., Spellman, P., Brown, P. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Nat. Acad. Sci.* **95**, 14863-14868.

Fraley, C. and Raftery, A. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer J.* **41**, 578-588.

Hansen, M. H. and Yu, B. (2001). Model selection and the principle of minimum description length. *J. Amer. Statist. Assoc.* **96**, 746-774.

Hornik, K., Feinerer, I. and Kober, M. (2012). *skmeans: Spherical k-Means Clustering.* R package version 0.2-1.

Huang, D. W., Sherman, B. T. and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4**, 44-57.

Huang, Y., Zaas, A. K., Rao, A., Dobigeon, N., Woolf, P. J., Veldman, T., Oien, N. C., McClain, M. T., Varkey, J. B., Nicholson, B., Carin, L., Kingsmore, S., Woods, C. W., Ginsburg, G. S. and Hero, III, A. O. (2011). Temporal dynamics of host molecular responses differentiate symptomatic and asymptomatic influenza a infection. *PLoS Genetics* **7**, e1002234.

Hurvich, C. M. and Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 297-307.

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath G. R., Wu, G. R., Matthews, L., lewis, S., Birney, E. and Stein, L. (2005). Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research* **33**, D428-D432.

Krishna, K. and Murty, M. N. (1999). Genetic K-means algorithm. *Systems, Man, and Cybernetics B: Cybernetics, IEEE Transactions* **29**, 433-439.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proc. 5th Berkeley Symposium On Mathematical Statistics and Probability* **1**, 281-297. Berkeley, University of California Press.

Maitra, R. and Ramler, I. (2010). A k-mean-directions algorithm for fast clustering of data on the sphere. *J. Comput. Graph. Statist.* **19**, 377-396.

Mardia, K. and Jupp, P. (2000). *Directional Statistics.* Wiley.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models.* Wiley-Interscience.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **27**, 29-34.

Okabe, A., Boots, B., Sugihara, K. and Chiu, S. (1992). *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams.* Wiley, Chichester.

Pommerenke, C., Wilk, E., Srivastava, B., Schulze, A., Novoselova, N., Geffers, R. and Schughart, K. (2012). Global transcriptome analysis in influenza-infected mouse lungs reveals the kinetics of innate and adaptive host immune responses. *PLoS One* **7**, e41169.

Ramsay, J. and Silverman, B. (2002). *Applied Functional Data Analysis: Methods and Case Studies.* Springer Verlag.

Saff, E. and Kuijlaars, A. (1997). Distributing many points on a sphere. *The Math. Intelligencer.* **19**, 5-11.

Storey, J. D., Xiao, W., Leek, J. T., Tompkins, R. G. and Davis, R. W. (2005). Significance analysis of time course microarray experiments. *Proc. Natl. Acad. Sci. USA* **102**, 12837-12842.

Strehl, A., Ghosh, J. and Mooney, R. (2000). Impact of similarity measures on web-page clustering. In *Workshop on Artificial Intelligence for Web Search* (AAAI 2000), 58-64.

Tavazoie, S., Hughes, J., Campbell, M., Cho, R. and Church, G. (1999). Systematic determination of genetic network architecture. *Nature Genetics* **22**, 281-285.

Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *J. Roy. Statist. Soc. Ser. B* **63**, 411-423.

Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY 14642, USA.

E-mail: xing_qiu@urmc.rochester.edu

Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY 14642, USA.

E-mail: shuang_wu@urmc.rochester.edu

Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY 14642, USA.

E-mail: hulin_wu@urmc.rochester.edu

(Received February 2013; accepted September 2014)