A New Information Criterion Based on Langevin Mixture Distribution for Clustering Circular Data with Application to Time Course Genomic Data

Xing Qiu[1], Shuang Wu[1], and Hulin Wu[1]

[1] *Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY 14642*

Supplementary Material

This document contains supplementary materials for manuscript “A New Information Criterion Based on Langevin Mixture Distribution for Clustering Circular Data with Application to Time Course Genomic Data”.

S1 Brief Summary

The following sections are presented in this document.

1. Brief Summary: A brief summary of the materials included in this document.

2. Additional Simulation Results. This section contains the following two subsections.

   (a) Illustrations of model selection procedures for main simulations. This subsection contains Figures S1 to S5 which illustrate the optimum number of clusters selected by various model selection methods for five simulated data (**SIM**.**K1**, **SIM**.**K5**, **SIM**.**K25**, **SIMBIO**.**A**, and **SIMBIO**.**B**.

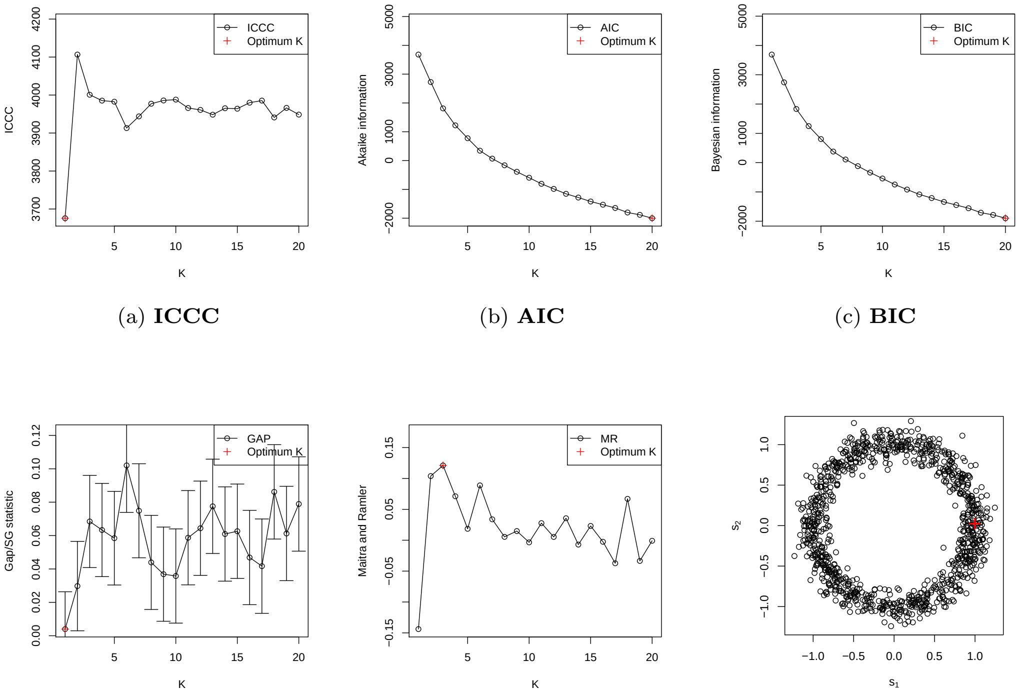
   (b) A comparison between the mixture Gaussian distribution and mixture Langevin distribution. We conducted a simulation study which illustrates the key differences between $\mathbb{R}^d$ and S^1.

3. Technical Details of Biological Data Analysis. We provide detailed information about the biological data used in this study. We also describes the hypothesis testing procedure and functional principal component analyses (fPCA) in two subsections (Identifying Significant Genes and Functional Principal Component Analysis). Table S1 summaries the number of identified significant genes for each subject; Table S2 lists the proportion of variance explained by the first two eigen-functions in fPCA. Figures S7 to S20 are included in this section to illustrate the optimum number of clusters selected by various model selection methods.

4. Functional Enrichment Analyses. We summarize the results of functional enrichment analysis in this section. Tables S3, S4, and S5 lists all the significant pathways identified in these analyses.

5. The Proof of the Main Theorem. This section contains the proofs of supporting lemmas and the main theorem (Theorem 2.1). The latter part is organized in subsection S5.1.

6. The Circular Shape of the First Two Functional Principal Components. We raised an important question the main manuscript: Why do the principal component scatter plots show circular pattern, and why is this pattern clearer for the symptomatic subjects than the asymptomatic ones? We provide an answer to this question in this section.



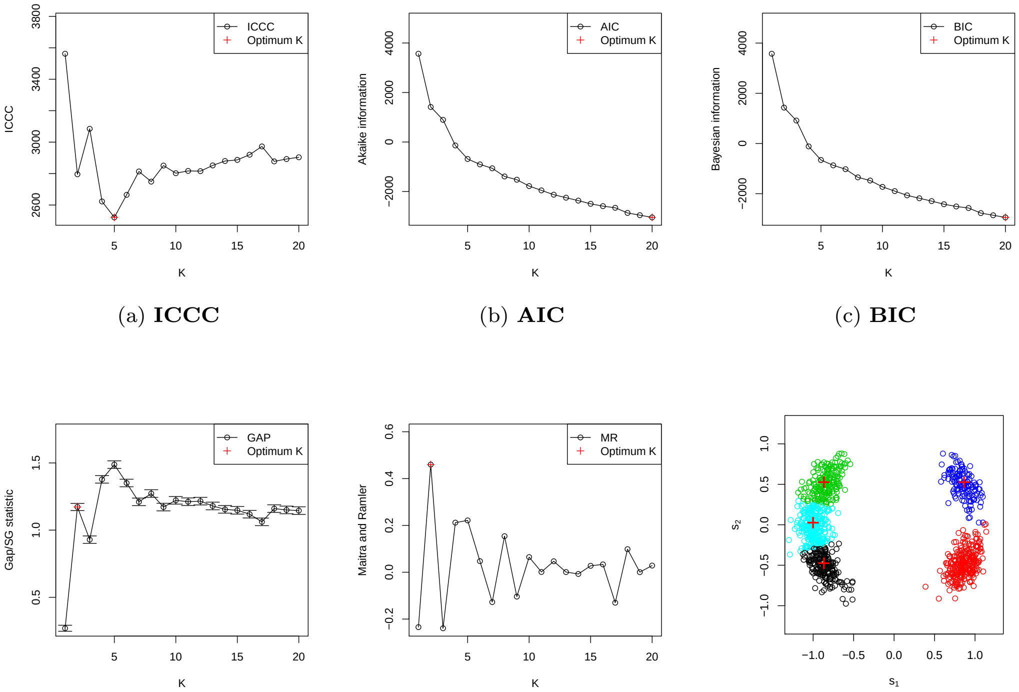
(a) **ICCC** (b) **AIC** (c) **BIC**

(d) **GAP** (e) **MR** (f) **Results**

Figure S1: Selecting the optimum K for **SIM.K1** by different model selection criteria. True number of clusters $K = 1$. The last panel (f) shows the results of running SK-means algorithm with number of clusters determined by ICCC ($K = 1$).

S2 Additional Simulation Results

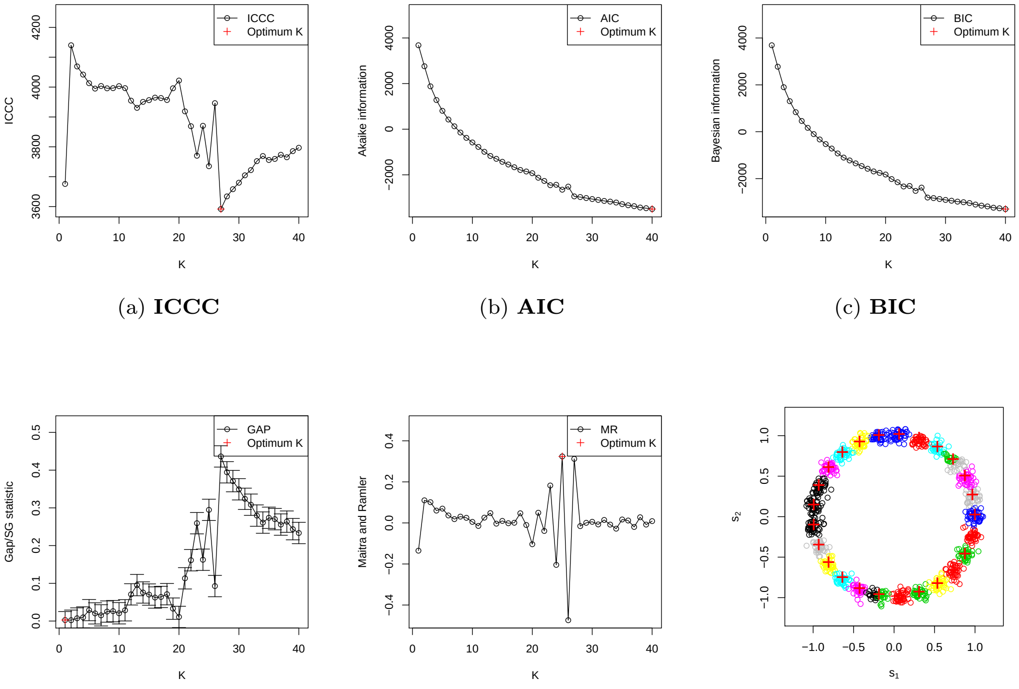
S2.1 Illustrations of model selection procedures for main simulations



(a) **ICCC** (b) **AIC** (c) **BIC**

(d) **GAP** (e) **MR** (f) **Results**

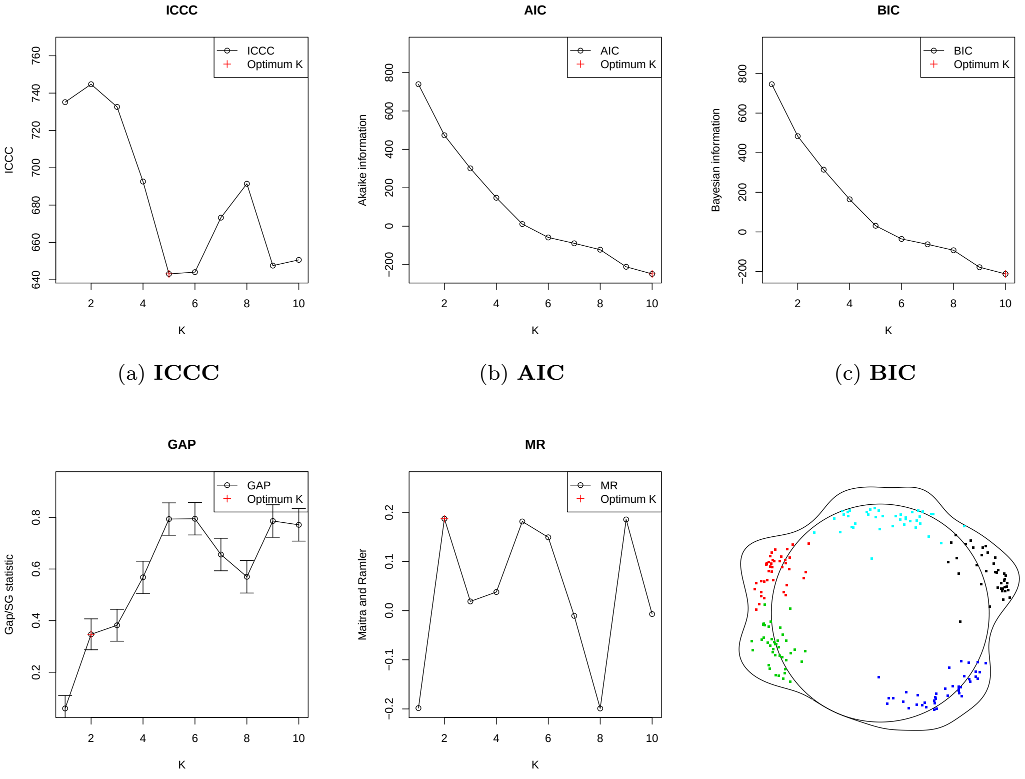
Figure S2: Selecting the optimum K for **SIM.K5** by different model selection criteria. True number of clusters: $K = 5$. The last panel (f) shows the results of running SK-means algorithm with number of clusters determined by ICCC ($K = 5$).



(a) **ICCC** (b) **AIC** (c) **BIC**

(d) **GAP** (e) **MR** (f) **Results**

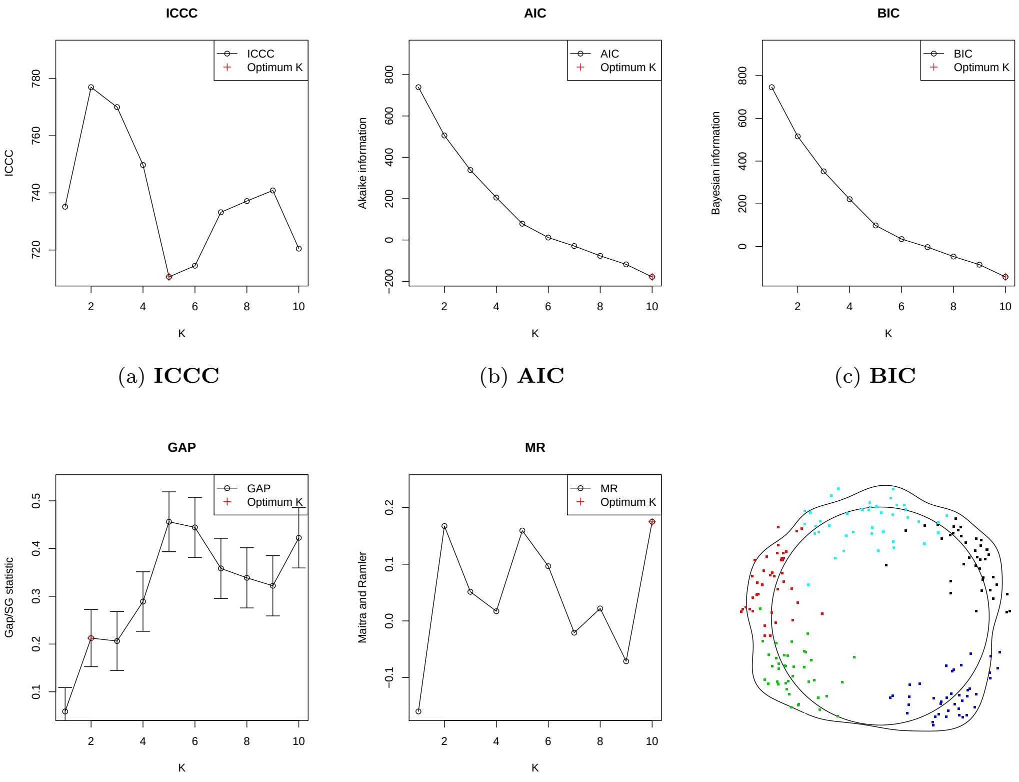
Figure S3: Selecting the optimum K for **SIM.K25** by different model selection criteria. True number of clusters: $K = 25$. The last panel (f) shows the results of running SK-means algorithm with number of clusters determined by ICCC ($K = 27$).



(a) **ICCC** (b) **AIC** (c) **BIC**

(d) **GAP** (e) **MR** (f) **Results**

Figure S4: Selecting the optimum K for **SIMBIO.A** by different model selection criteria. True number of clusters: $K = 5$. The last panel (f) shows the results of running SK-means algorithm with number of clusters determined by ICCC ($K = 5$).



(a) **ICCC** (b) **AIC** (c) **BIC**

(d) **GAP** (e) **MR** (f) **Results**

Figure S5: Selecting the optimum K for **SIMBIO.B** by different model selection criteria. True number of clusters: $K = 5$. The last panel (f) shows the results of running SK-means algorithm with number of clusters determined by ICCC ($K = 5$).

S2.2 A comparison between the mixture Gaussian distribution and mixture Langevin distribution

There are some important connections between the Langevin distribution and the Gaussian distribution (Mardia and Jupp, 2000). The mean direction μ_k and the inverse of the concentration parameter $1/\kappa$ of the one-dimensional Langevin distribution are the analogy of the mean vector and the variance of the bivariate Gaussian distribution, respectively. However, there are also significant differences between these two because the geometry of S^1 is fundamentally different from that of R^2. First, the circular space S^1 is curved and its curvature cannot be ignored for distributions that are not concentrated in a very small arc. In addition, summation and scaling have to be adapted so that these operators are well defined on S^1. Two fundamental limiting theorems in mathematical statistics, namely the law of large numbers and the central limit theorem on S^1 are formulated very differently than their counterparts on the Euclidean space.

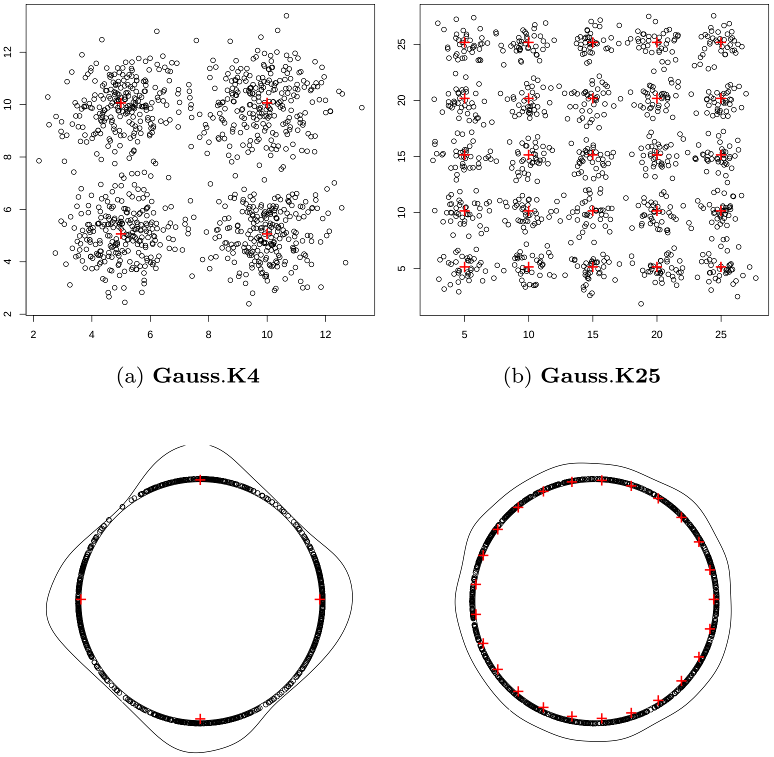
Secondly, the circular space S^1 is compact. This suggests that one cannot arbitrarily "squeeze in" many clusters on S^1 without making each cluster smaller (larger κ). The following simple example highlights this difference. Suppose that one has a Gaussian mixture model defined on $\mathbb{R}^2$ and a mixture Langevin model defined on S^1. Both consists of K clusters with the same shape parameter (σ^2 for Gaussian mixture and κ for Langevin mixture) but distinct location parameters. In the Gaussian mixture case, one can always spread the centers of these clusters on an equi-distance grid so the smallest distance between two cluster centers is always greater than a constant, for example 5σ. In this case all clusters are still distinguishable even for large K. However, the smallest distance between the centers of two clusters on S^1 must be less or equal to $\frac{2\pi}{K}$. So when K is large, the clusters are not distinguishable and the scatter plot would look like being generated from a uniform distribution on S^1, which is the least informative distribution for clustering.

We conduct a simulation study to demonstrate the different geometric properties of Langevin mixture distribution on S^1 and bivariate Gaussian mixtures. The simulated data are generated as follows.

Gauss.K4: $n = 1,000$ observations on $\mathbb{R}^2$ drawn from a Gaussian mixture of four clusters with equal size. The variance parameter is $\sigma^2 = 1$ for all clusters; the cluster centers lie on an equi-distance grid on $\mathbb{R}^2$ with unit length 5.

Gauss.K25: $n = 1,000$ observations on $\mathbb{R}^2$ drawn from a Gaussian mixture with 25 clusters with equal size. Like **Gauss.K4**, the variance parameter is $\sigma^2 = 1$ for all clusters and the cluster centers lie on an equi-distance grid on $\mathbb{R}^2$ with unit length 5.

Lang.K4 $n = 1,000$ observations on S^1 drawn from a Langevin mixture of four clusters with equal size. The concentration parameter is $\kappa = \dfrac{100}{\pi^2}$; the angular cluster centers lie on an equi-distance grid on S^1. The choice of κ is based on this consideration. When $K = 4$, the smallest between-cluster angular distance is $\pi/2$. To



(a) **Gauss**.**K4**

(b) **Gauss**.**K25**

(c) **Lang**.**K4**

(d) **Lang**.**K25**

Figure S6: A comparison of clustering Gaussian mixtures and Langevin mixtures. Red crosses mark cluster centers. Total number of observations for each cluster: $n = 1,000$. For better visual effects, empirical circular density plots are overlaid on top of scatter plots of Langevin mixtures.

match **Gauss**.**K4** the variance parameter should be $\sigma^2 = \left(\frac{1}{5}\frac{\pi}{2}\right)^2 = \dfrac{\pi^2}{100}$. Since $1/\kappa$ is the analogy of σ^2, we choose $\kappa = \dfrac{100}{\pi^2}$.

Lang.**K25**: Just like **Lang**.**K4**, except that it has 25 clusters with equal size centered on an equi-distance grid on S^1.

The results of this simulation is summarized in Figure S6. It is clear that when K becomes large, the (empirical) density function generated from a mixture of Langevin distributions is indistinguishable from the uniform distribution on S^1. This example highlights the compactness property of S^1.

S3 Technical Details of Biological Data Analysis

We applied the proposed ICCC model selection criterion and spherical K-means cluster algorithm to a large scale time course microarray data (Huang et al., 2011). In this study, a cohort of 17 healthy human volunteers received intranasal inoculation of influenza H3N2/Wisconsin and 9 of these subjects developed mild to severe symptoms. A total of $m = 11,961$ gene expression profiles were measured on whole peripheral blood drawn from all subjects at 16 time points. These time points cover a total of 132 hours of observation, including one measurement taken 24 hours before inoculation and $J = 15$ time points at an interval of roughly 8 hours post inoculation (hpi) through 108 hpi.

Hybridization and microarray data collection was performed using the Human Genome U133A 2.0 Array (Affymetrix, Santa Clara, CA). Data pre-processing was done using the robust multi-array (RMA) method (Bolstad et al., 2003). More detailed technical descriptions of these data can be found in Huang et al. (2011).

For convenience, we exclude the pre-inoculation measurement from our study. We also exclude subjects 8, 13, and 17 due to missing time points. Gene-wise standardization is applied before clustering analysis to ensure genes all have the same sample mean (zero) and variance (one).

S3.1 Identifying Significant Genes

We conduct functional F-test (Ramsay and Silverman, 2002; Storey et al., 2005) to identify genes whose gene expressions change significantly for this study. The null hypothesis of this testing problem is

$$H_{0,i}: \quad y_i(t) = c_{i0}, \quad 0 \leqslant t \leqslant 108,\ i = 1, 2, \ldots, m, \tag{S3.1}$$

where $y_i(t)$ represents the underlying true expression curve for the ith gene and the constant c_{i0} represents the "normal level" of this gene, should the exposure to influenza viruses has no effect to its expression.

In practice, we only observe a discrete and noisy representation of $y_i(t)$, *i.e.*, $w_{ij} = y_i(t_j) + \epsilon_{ij}$, for $j = 1, 2, \ldots, J$ time points. So $y_i(t)$ must be estimated. We can use $\hat{y}_i^0(t) \equiv \hat{c}_{i0} = \frac{1}{J}\sum_{j=1}^{J} w_{ij}$, the sample mean over $J = 15$ time points, as the estimate of c_{i0} under $H_{0,i}$. Since data has already been standardized, $\widehat{c_{i0}} \equiv 0$ for all genes.

Let SS_i^0 denote the residual sum of squares associated with the constant function approximation of the observed data. Due to the standardized nature of data, we have

$$SS_i^0 := \sum_{j=1}^{J} (w_{ij} - \hat{c}_{i0})^2 := (J-1)\hat{\sigma}^2(w_i) \equiv J - 1, \quad \text{for } i = 1, 2, \ldots, m. \tag{S3.2}$$

Under the alternative hypotheses, penalized B-splines are used to estimate $y_i(t)$ from the noisy microarray observations (Storey et al., 2005). This amounts to the following representation

$$H_{1,i}: \quad y_i(t) = c_i + \sum_{l=1}^{L} c_{il}\beta_l(t), \quad 0 \leqslant t \leqslant 108,\ i = 1, 2, \ldots, m, \tag{S3.3}$$

where $\beta_l(t)$, $l = 1, 2, \ldots, L$ are B-spline basis. Generalized cross-validation (GCV) principle is used for choosing the roughness penalty term. The estimated curves are denoted simply by $\hat{y}_i(t)$.

The following summary statistic is used for testing $H_{0,i}$:

$$F_i = \frac{SS_i^0 - SS_i^1}{SS_i^1} = \frac{J-1}{SS_i^1} - 1, \tag{S3.4}$$

where SS_i^0 and SS_i^1 are the residual sum of squares obtained from the null and alternative hypothesis, respectively. The null distribution is generated by $R = 5,000$ permutations of time labels. The unadjusted p-value for testing $H_{0,i}$ is calculated by

$$p_i = \sum_{r=1}^{R} \frac{\#\{r : F_i^r \geqslant F_i\}}{R}, \tag{S3.5}$$

where F_i^r is the summary statistic associated with the ith gene in the rth permutation.

According to Equation (S3.4), a gene is significant if and only if SS_i^1 is small. Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) is then applied to control the false discovery rate (FDR) at 0.05 level. Table S1 summarizes the number of significant genes for each individual.

	DEGs	Symptoms
subject1	110	Symptomatic
subject2	22	Asymptomatic
subject3	3	Asymptomatic
subject4	13	Asymptomatic
subject5	1345	Symptomatic
subject6	532	Symptomatic
subject7	163	Symptomatic
subject9	64	Asymptomatic
subject10	2504	Symptomatic
subject11	1	Asymptomatic
subject12	199	Symptomatic
subject14	99	Asymptomatic
subject15	200	Symptomatic
subject16	35	Asymptomatic

Table S1: Numbers of significant genes for each individual. Subjects with the most and least significant genes (10 and 11) are highlighted in the table.

One clear message of Table S1 is that the seven patients with visible symptom have much more significant genes on average compared with those without symptom. Some asymptomatic subjects have very few significant genes (such as subjects 3 and 11) which makes cluster analysis impossible. To facilitate cluster analysis, we use the most significant 200 genes if less than 200 genes are selected.

S3.2 Functional Principal Component Analysis

Once the significant genes are selected, functional principal component analysis (fPCA) (Ramsay and Silverman, 2002) is applied to $\hat{y}_i(t)$, the estimated temporal curves under the alternative hypothesis. Each $\hat{y}_i(t)$ is then represented by $\vec{s} := (s_1, s_2, \ldots, s_P)$, the first P principal component scores for the subsequent cluster analysis.

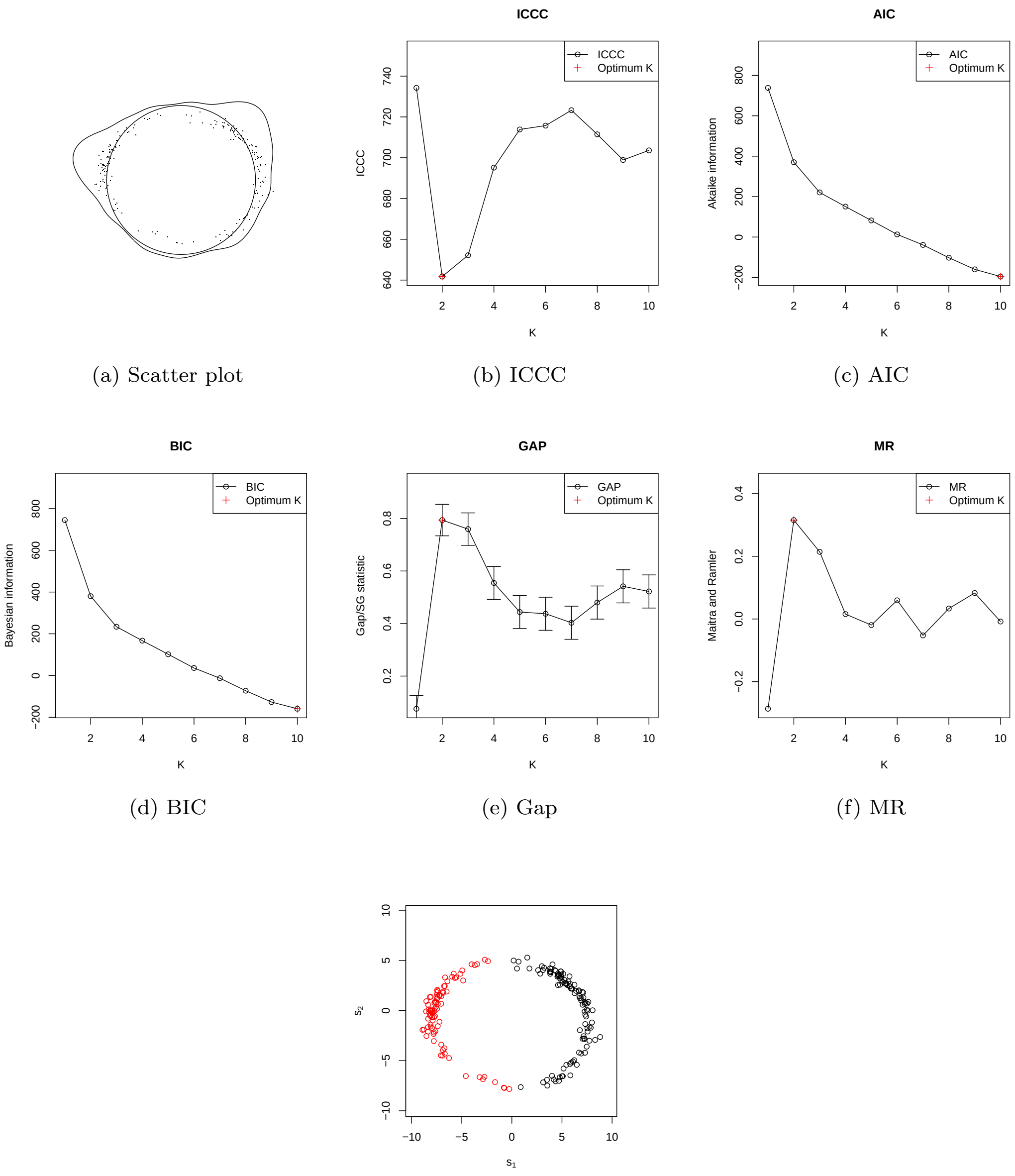
The fPCA serves the following purposes: 1. it transforms the functional objects into multivariate principal component scores so standard cluster analysis tools can be applied without modification; 2. it finds the most *parsimonious* basis representation of $\hat{y}_i(t)$, which by construction are vectors in the L-dimensional functional linear space.

Table S2 summarizes the proportion of total variance explained by the first two principal components. It is clear that these two PCs account for most of total variance for all subjects.

	PC1	PC2	Both
subject1	75.99	23.02	99.00
subject2	60.81	36.93	97.74
subject3	85.90	12.57	98.47
subject4	78.49	19.79	98.27
subject5	90.04	9.21	99.24
subject6	76.95	22.20	99.15
subject7	88.21	11.04	99.25
subject9	84.69	14.19	98.88
subject10	89.07	9.97	99.04
subject11	67.50	26.66	94.15
subject12	77.41	21.94	99.34
subject14	85.96	13.37	99.33
subject15	92.44	5.58	98.02
subject16	83.52	15.79	99.30

Table S2: Percentage of total variance explained by the first two principal components.

For each subject we make a scatter plots of (s_{i1}, s_{i2}), the first two principal component scores of each gene.

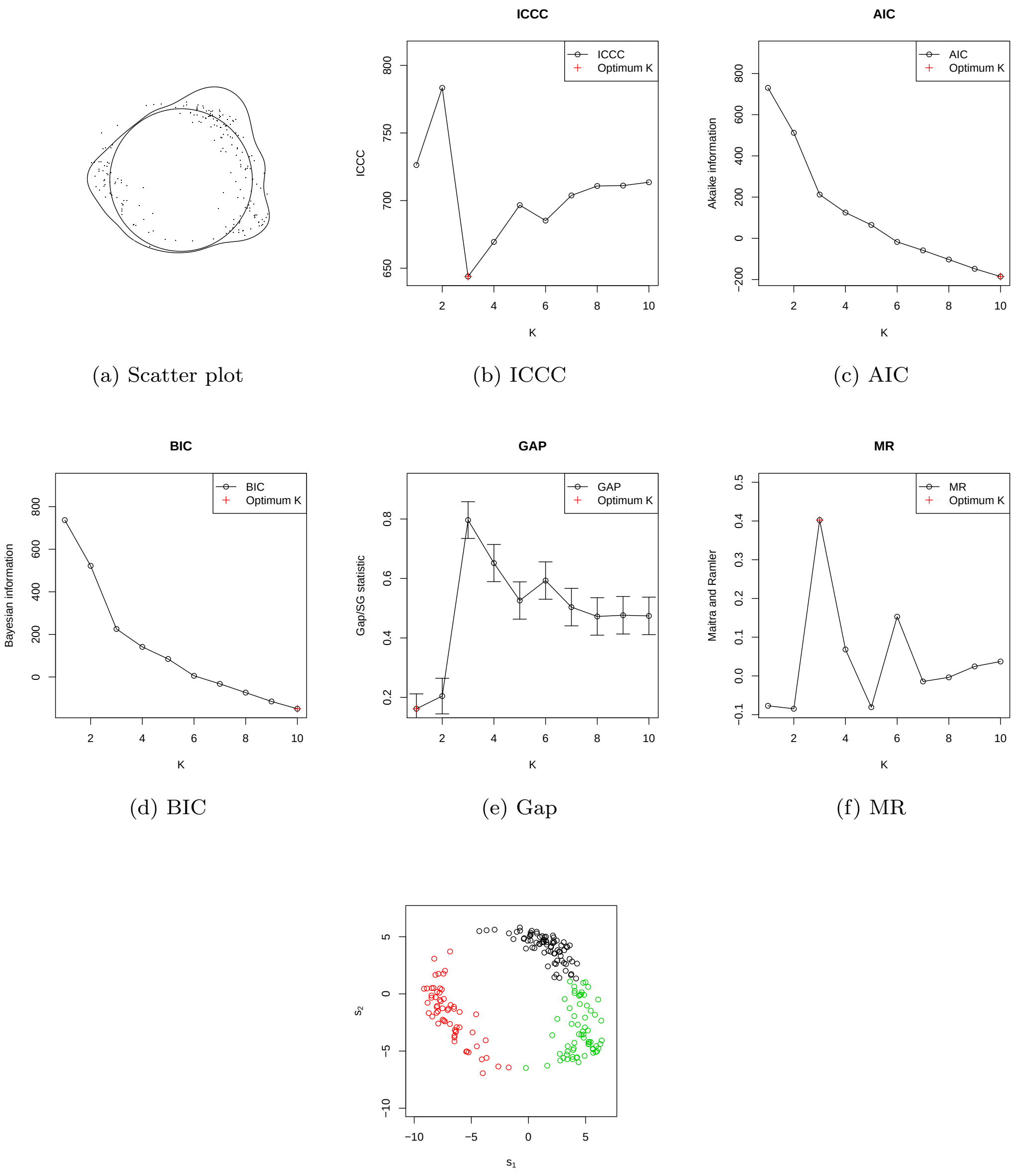


(a) Scatter plot (b) ICCC (c) AIC

(d) BIC (e) Gap (f) MR

(g) Results (ICCC)

Figure S7: Plots of model selection procedures for subject 1 (Symptomatic). Number of clusters estimated by ICCC: $K = 2$.

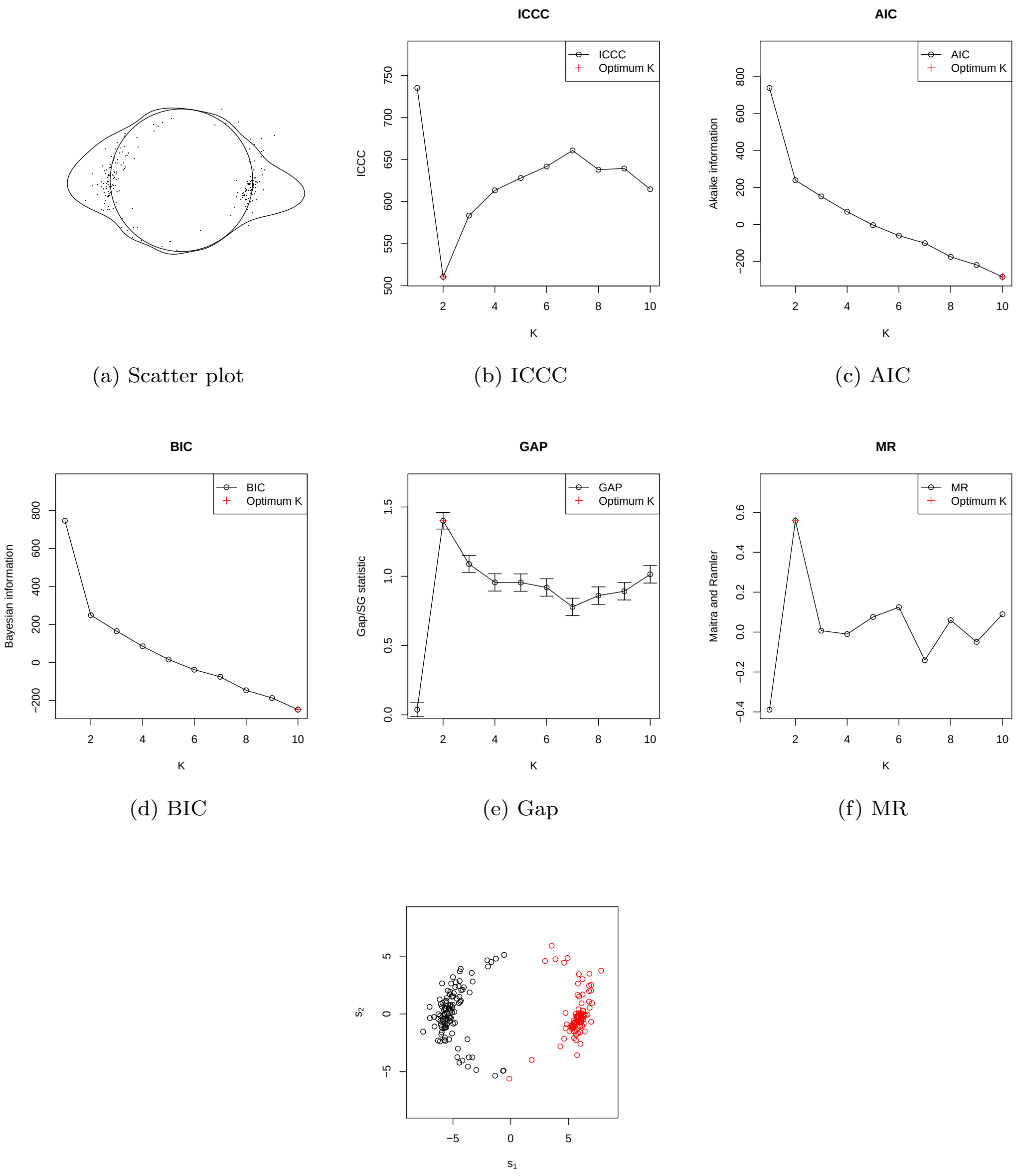


(a) Scatter plot (b) ICCC (c) AIC

(d) BIC (e) Gap (f) MR

(g) Results (ICCC)

Figure S8: Plots of model selection procedures for subject 2 (Asymptomatic). Number of clusters estimated by ICCC: $K = 3$.



(a) Scatter plot

(b) ICCC

(c) AIC

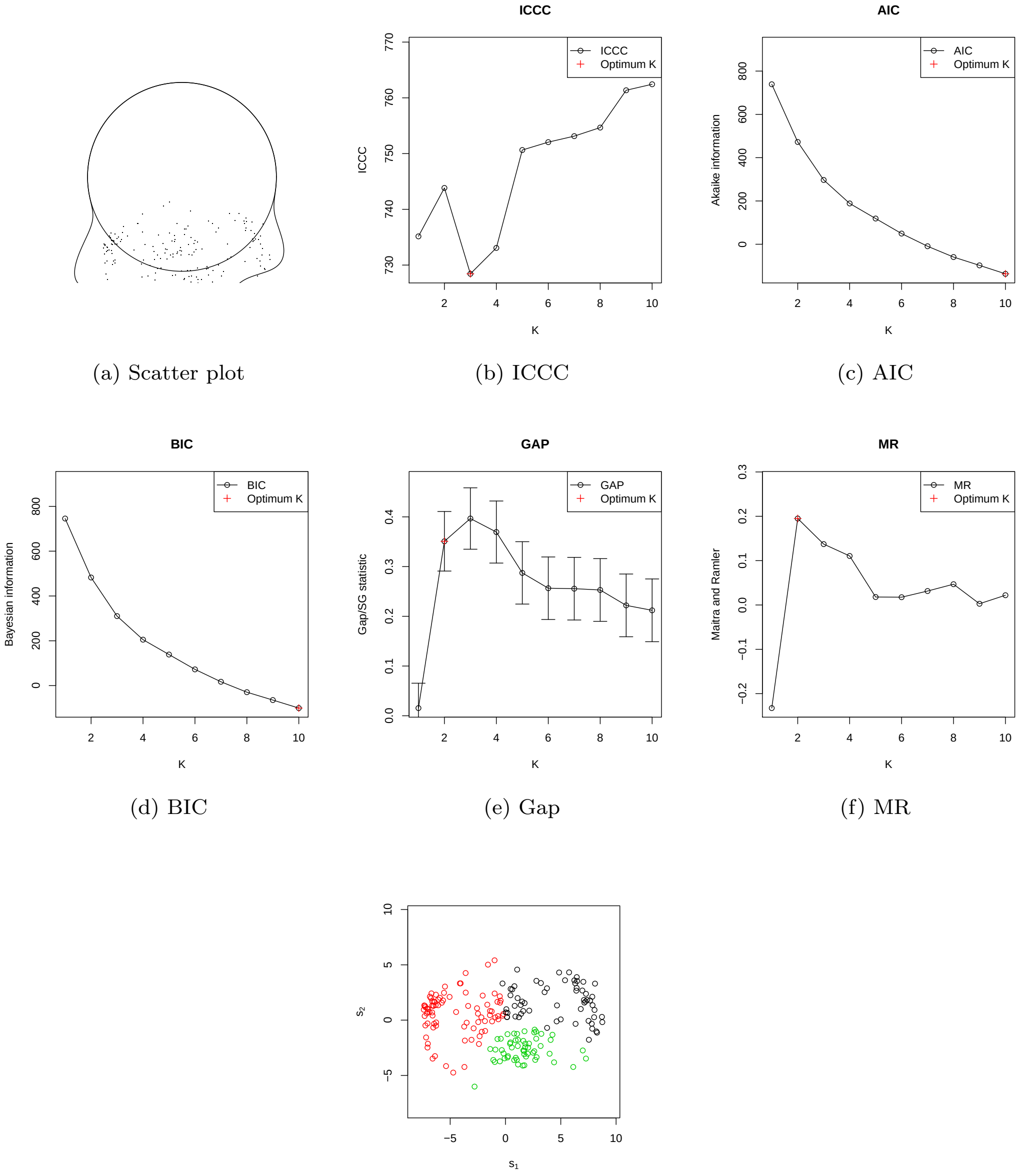
(d) BIC

(e) Gap

(f) MR

(g) Results (ICCC)

Figure S9: Plots of model selection procedures for subject 3 (Asymptomatic). Number of clusters estimated by ICCC: $K = 2$.

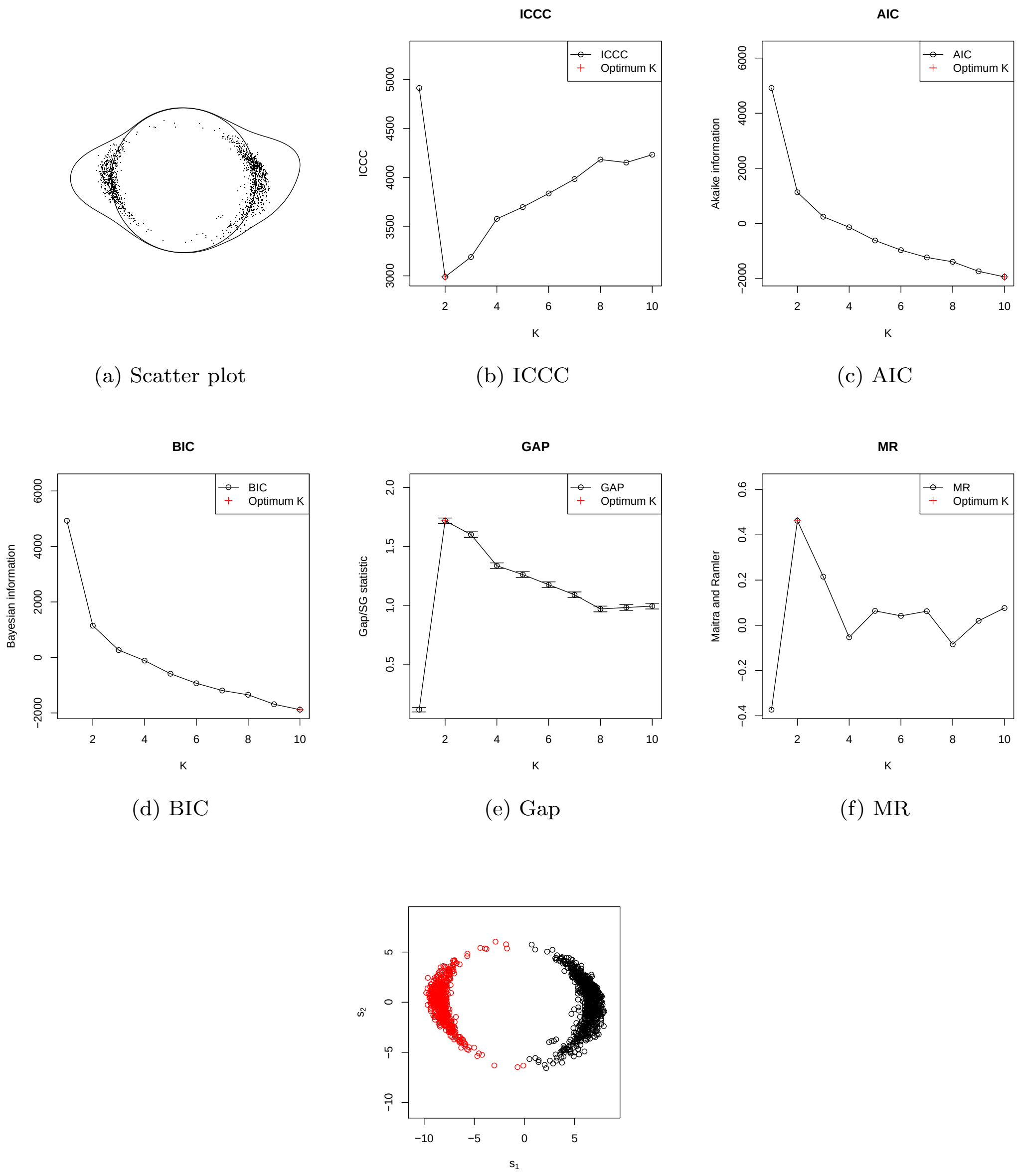


(a) Scatter plot (b) ICCC (c) AIC

(d) BIC (e) Gap (f) MR

(g) Results (ICCC)

Figure S10: Plots of model selection procedures for subject 4 (Asymptomatic). Number of clusters estimated by ICCC: $K = 3$.



(a) Scatter plot

(b) ICCC

(c) AIC

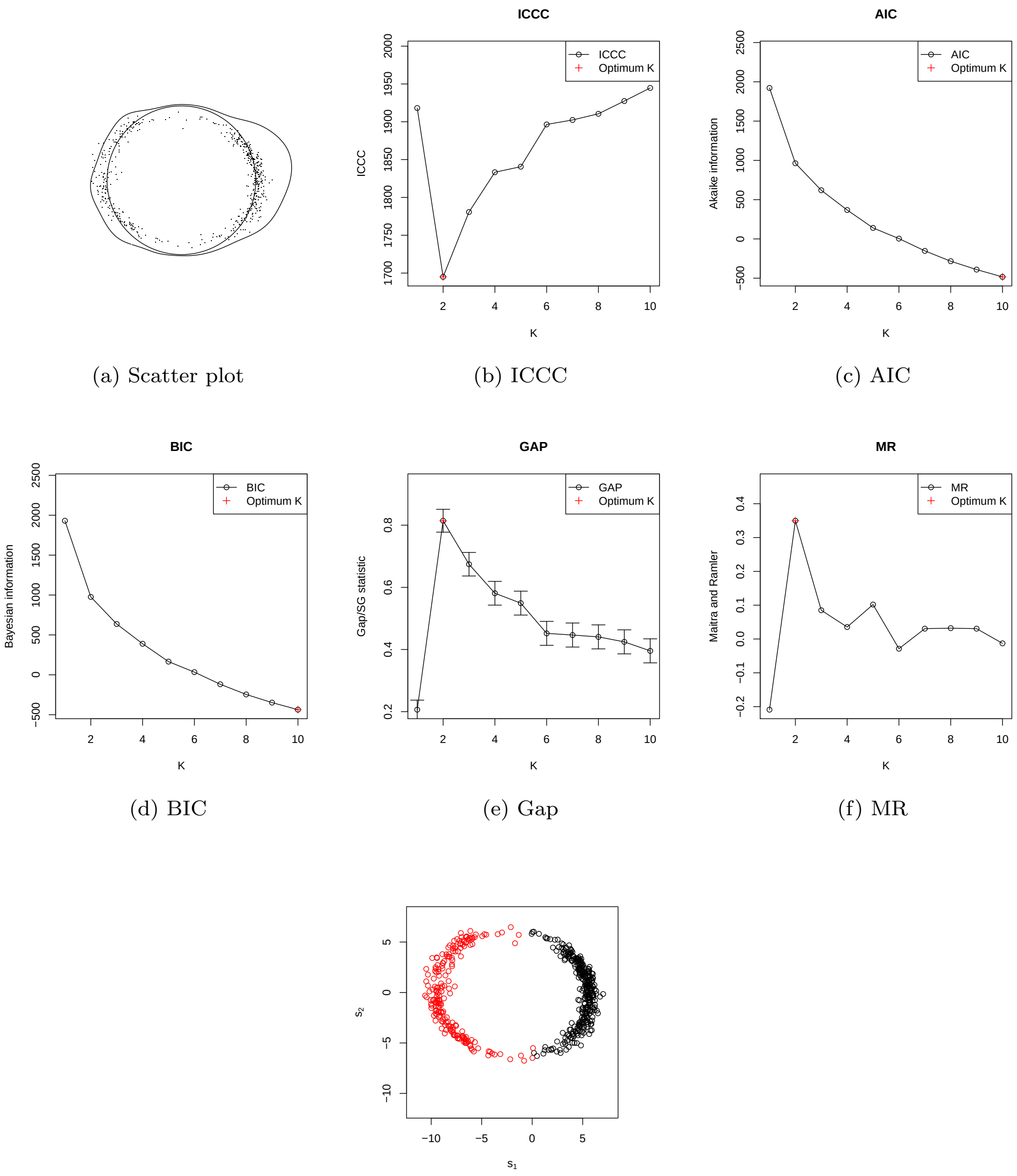
(d) BIC

(e) Gap

(f) MR

(g) Results (ICCC)

Figure S11: Plots of model selection procedures for subject 5 (Symptomatic). Number of clusters estimated by ICCC: $K = 2$.



(a) Scatter plot

(b) ICCC

(c) AIC

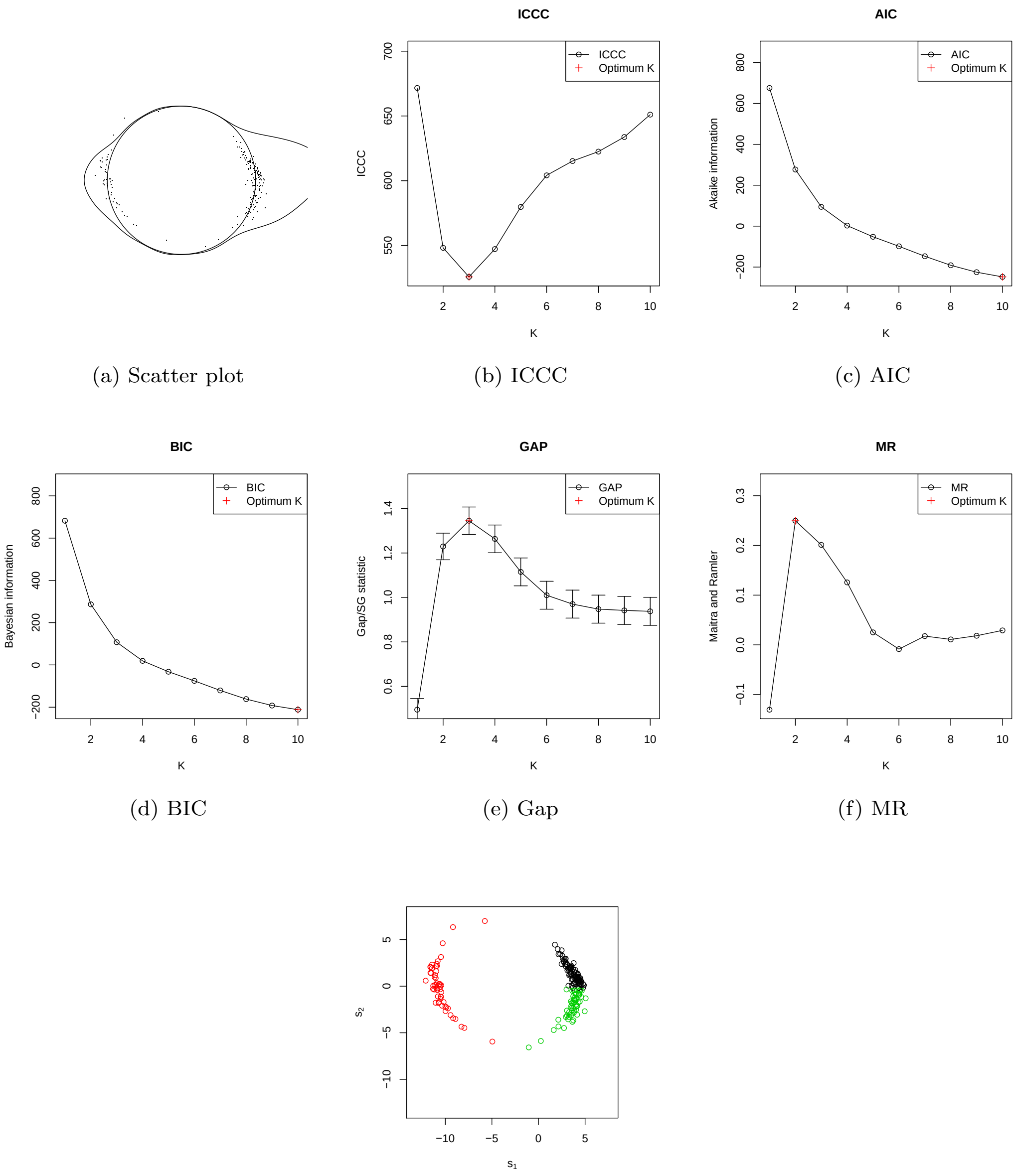
(d) BIC

(e) Gap

(f) MR

(g) Results (ICCC)

Figure S12: Plots of model selection procedures for subject 6 (Symptomatic). Number of clusters estimated by ICCC: $K = 2$.



(a) Scatter plot

(b) ICCC

(c) AIC

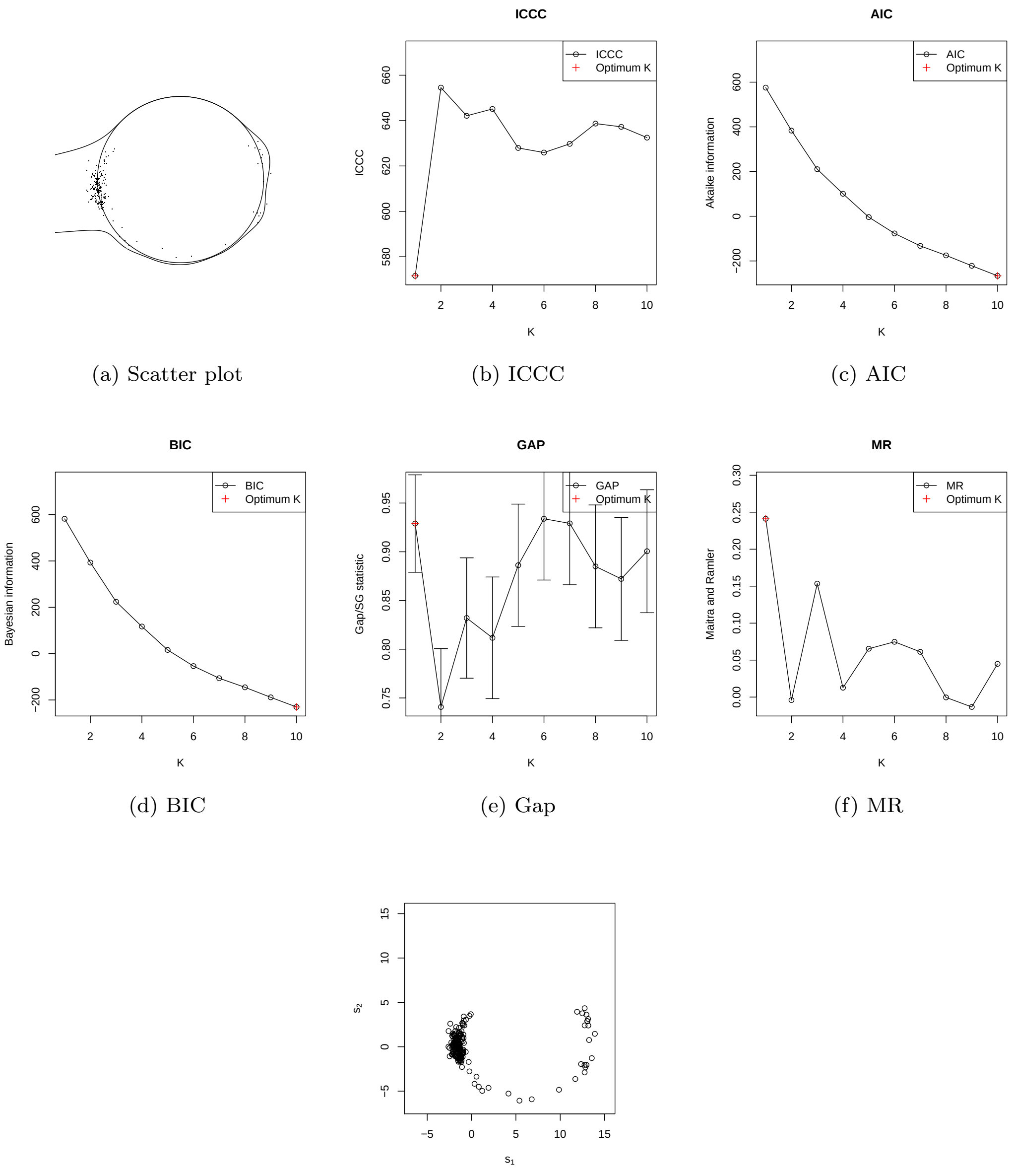
(d) BIC

(e) Gap

(f) MR

(g) Results (ICCC)

Figure S13: Plots of model selection procedures for subject 7 (Symptomatic). Number of clusters estimated by ICCC: $K = 3$.



(a) Scatter plot

(b) ICCC

(c) AIC

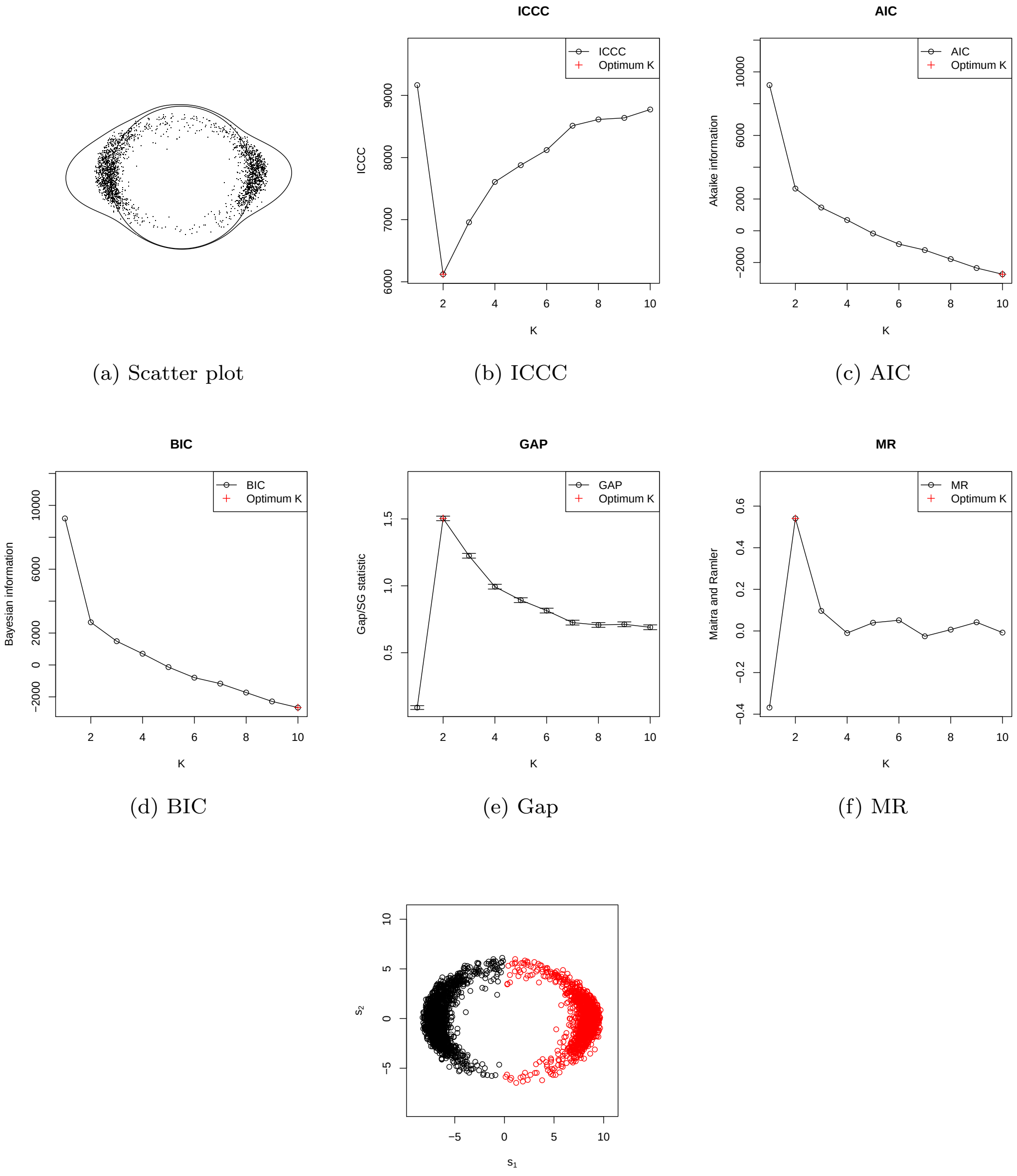
(d) BIC

(e) Gap

(f) MR

(g) Results (ICCC)

Figure S14: Plots of model selection procedures for subject 9 (Asymptomatic). Number of clusters estimated by ICCC: $K = 1$.



(a) Scatter plot

(b) ICCC

(c) AIC

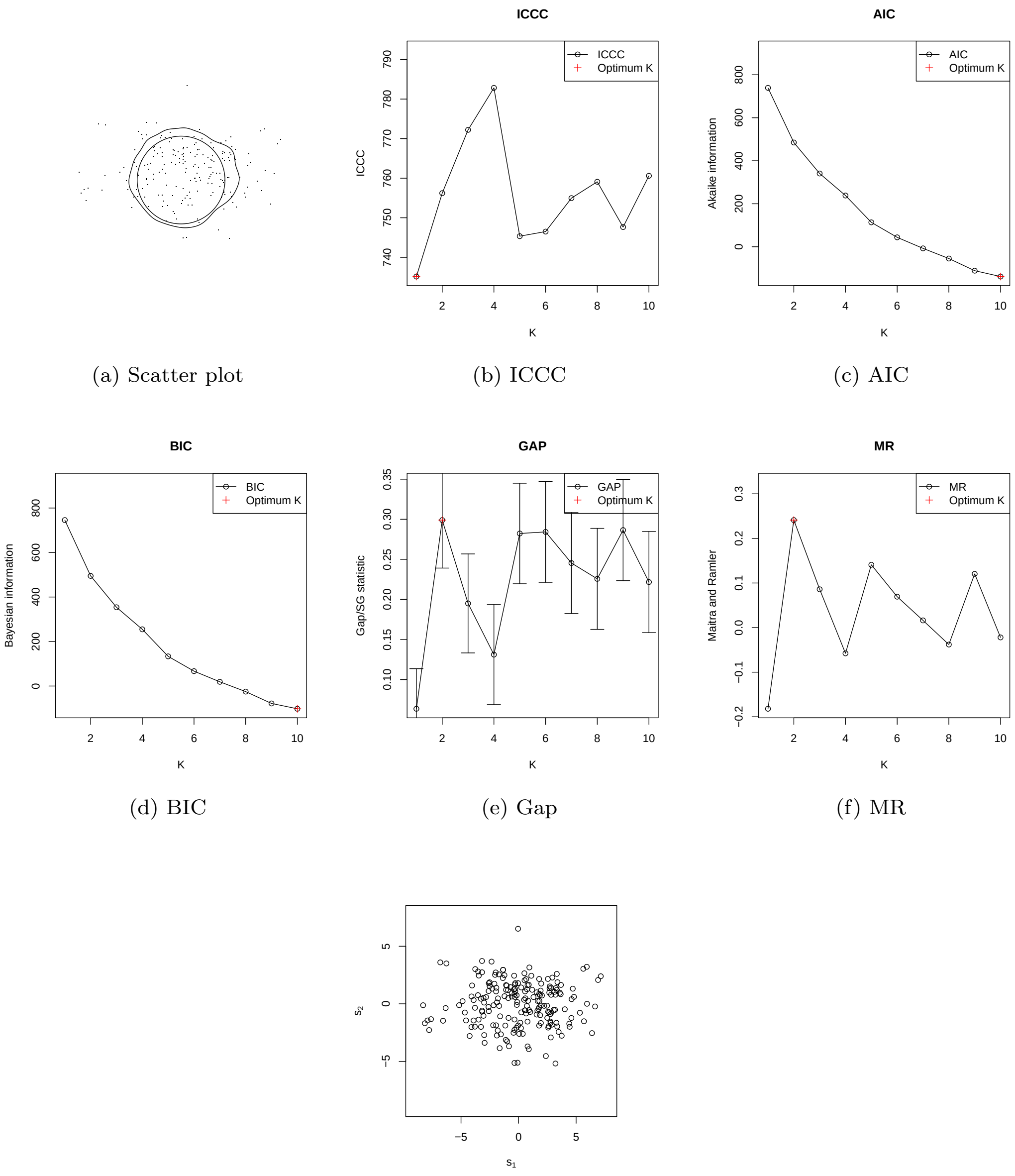
(d) BIC

(e) Gap

(f) MR

(g) Results (ICCC)

Figure S15: Plots of model selection procedures for subject 10 (Symptomatic). Number of clusters estimated by ICCC: $K = 2$.



(a) Scatter plot

(b) ICCC

(c) AIC

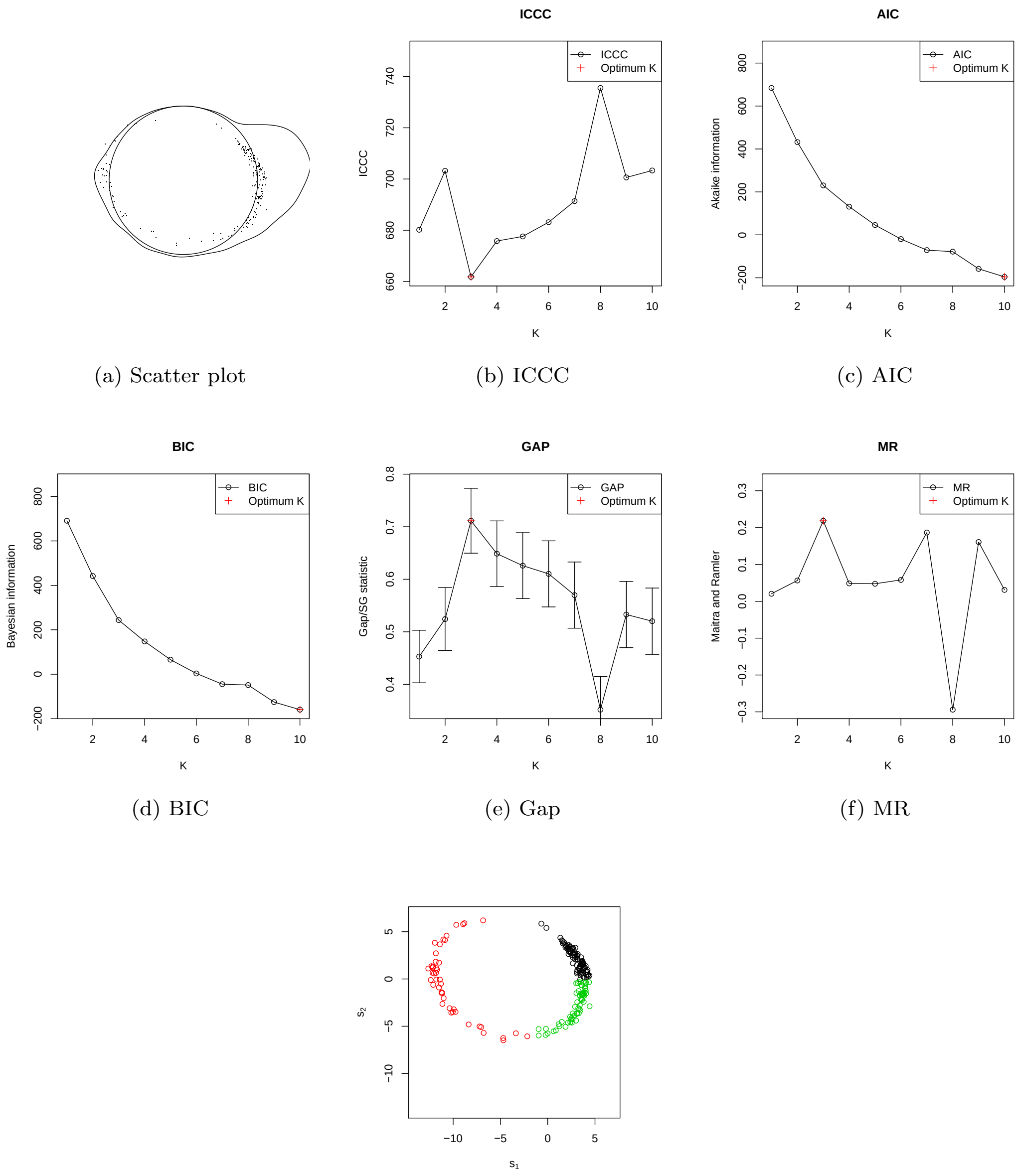
(d) BIC

(e) Gap

(f) MR

(g) Results (ICCC)

Figure S16: Plots of model selection procedures for subject 11 (Asymptomatic). Number of clusters estimated by ICCC: $K = 1$.



(a) Scatter plot

(b) ICCC

(c) AIC

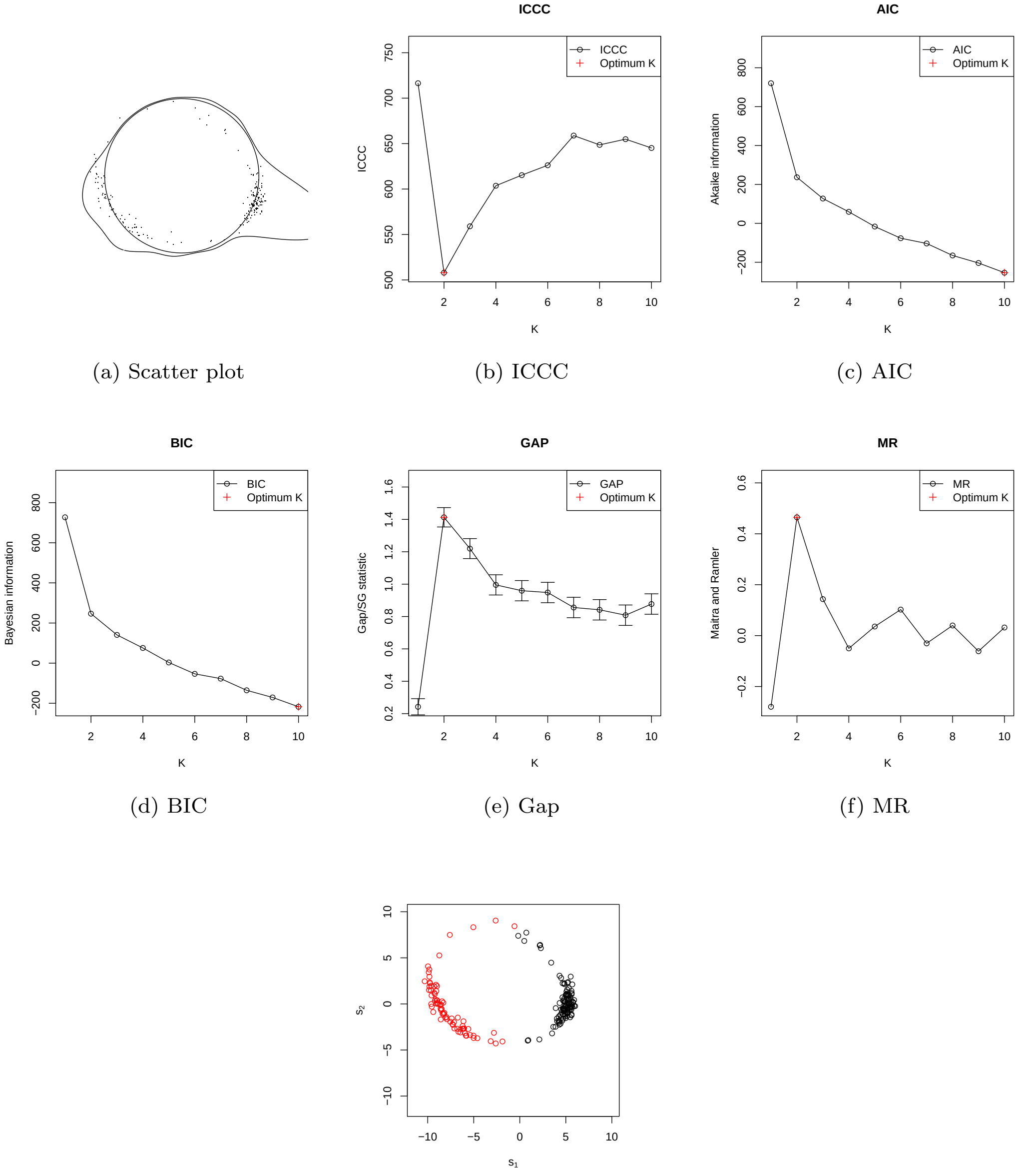
(d) BIC

(e) Gap

(f) MR

(g) Results (ICCC)

Figure S17: Plots of model selection procedures for subject 12 (Symptomatic). Number of clusters estimated by ICCC: $K = 3$.



(a) Scatter plot

(b) ICCC

(c) AIC

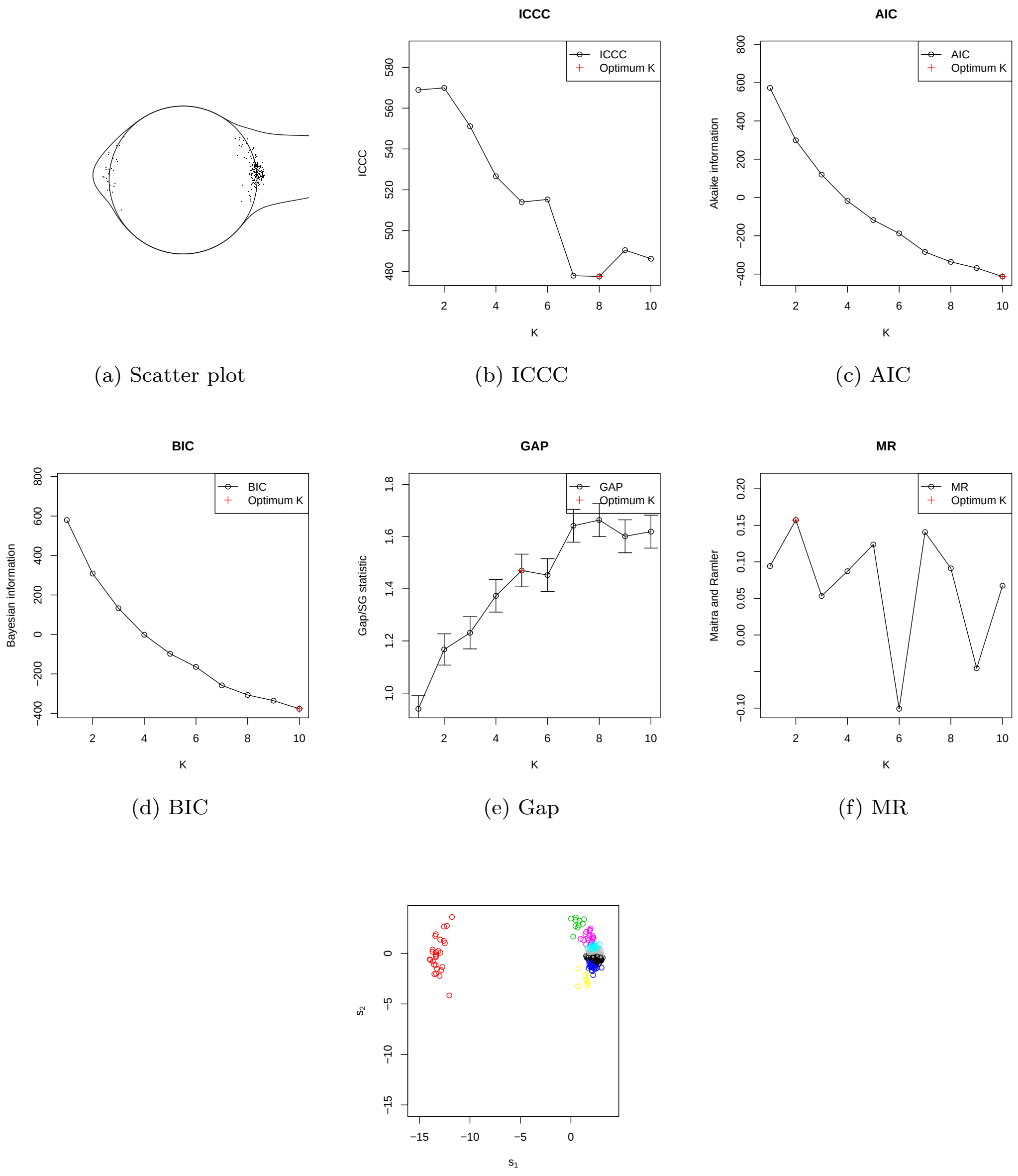
(d) BIC

(e) Gap

(f) MR

(g) Results (ICCC)

Figure S18: Plots of model selection procedures for subject 14 (Asymptomatic). Number of clusters estimated by ICCC: $K = 2$.



(a) Scatter plot

(b) ICCC

(c) AIC

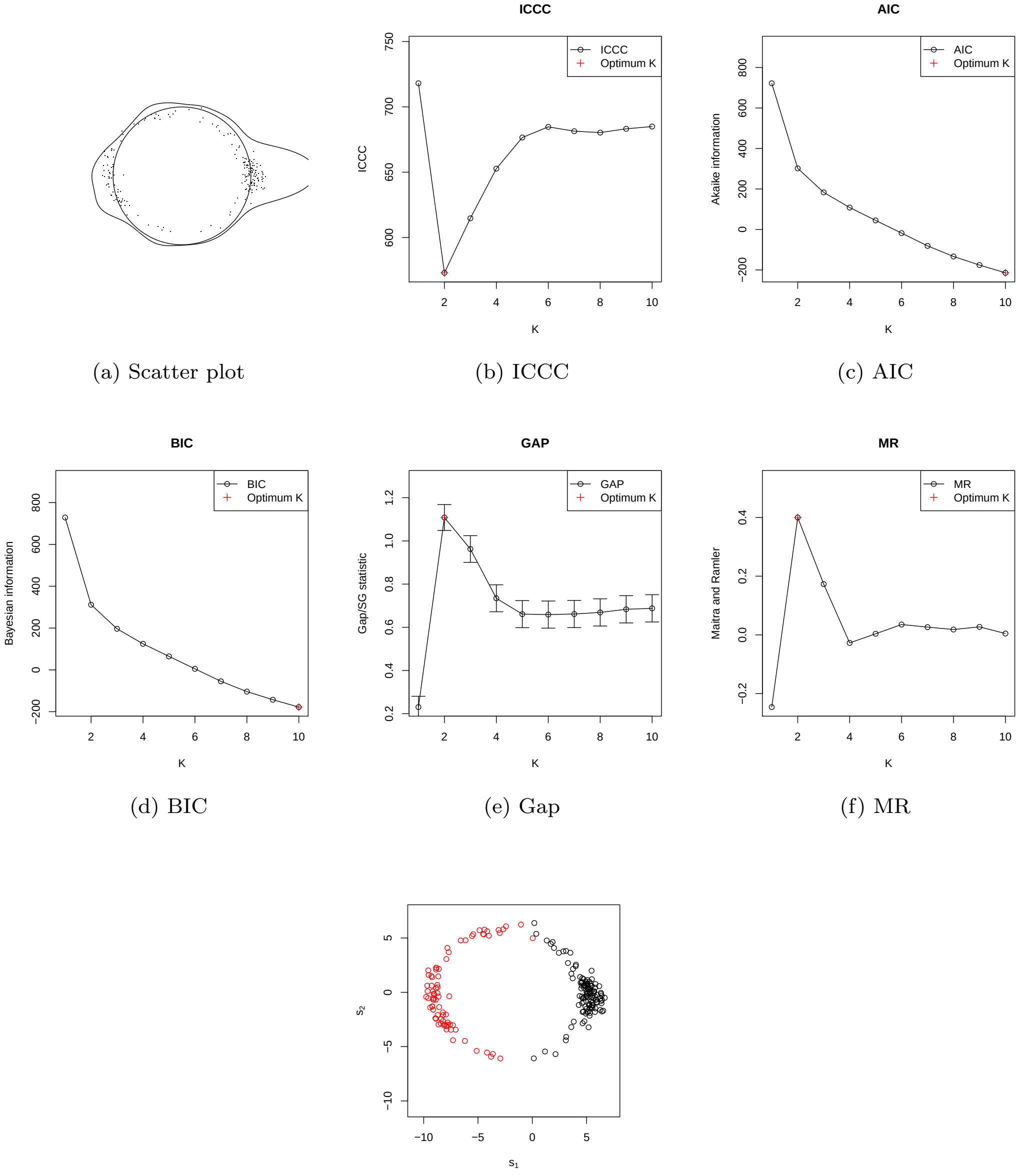
(d) BIC

(e) Gap

(f) MR

(g) Results (ICCC)

Figure S19: Plots of model selection procedures for subject 15 (Symptomatic). Number of clusters estimated by ICCC: $K = 8$.



(a) Scatter plot

(b) ICCC

(c) AIC

(d) BIC

(e) Gap

(f) MR

(g) Results (ICCC)

Figure S20: Plots of model selection procedures for subject 16 (Asymptomatic). Number of clusters estimated by ICCC: $K = 2$.

S4 Functional Enrichment Analyses

We conduct functional enrichment analyses on the classified gene clusters by DAVID Huang et al. (2009). Specifically, for each subject, we identified the enriched functions and pathway annotations of each cluster in Gene Ontology Ashburner (2000), KEGG Ogata et al. (1999) and REACTOME Joshi-Tope et al. (2005) curated pathway databases. The Bonferroni multiple testing procedure is selected to control the familywise error rate at level 0.05.

For Subject 10 (Symptomatic), its 2,504 DEGs are divided into two clusters based on ICCC, GAP, and MR. A total of 118 significant pathways are identified and these results are summarized in Table S3. For Subject 2 (Asymptomatic), ICCC divides its DEGs (200 most significant genes) into three clusters (same as MR) but GAP groups them into one cluster. For Subject 11 (Asymptomatic), ICCC integrates all DEGs (200 most significant genes) into one cluster but both GAP and MR divide them into two clusters. Based on ICCC, we identified six significant pathways from Subject 2 ($K = 3$) and four significant pathways from Subject 11 ($K = 1$). These results are summarized in Table S4. Based on the alternative model selection criteria, we identified three significant pathways from Subject 2 ($K = 1$, GAP) and four significant pathways from Subject 11 ($K = 2$, GAP/MR). These results are summarized in Table S5.

Term	Name	Adj.pval	Cluster
GO:0006955	immune response	0.00000	cluster2
GO:0016568	chromatin modification	0.00000	cluster1
GO:0006325	chromatin organization	0.00000	cluster1
GO:0051276	chromosome organization	0.00000	cluster1
REACT_578	apoptosis	0.00000	cluster2
GO:0006952	defense response	0.00000	cluster2
GO:0009615	response to virus	0.00000	cluster2
REACT_6850	cdc20	0.00000	cluster2
GO:0045321	leukocyte activation	0.00000	cluster1
REACT_13635	regulation of activated pak-2p34 by proteasome mediated degradation	0.00001	cluster2
REACT_11045	signaling by wnt	0.00001	cluster2
REACT_9035	apc/c	0.00001	cluster2
GO:0051443	positive regulation of ubiquitin-protein ligase activity	0.00001	cluster2
GO:0043123	positive regulation of i-kappab kinase/nf-kappab cascade	0.00001	cluster2
hsa04666	fc gamma r-mediated phagocytosis	0.00001	cluster1
GO:0051351	positive regulation of ligase activity	0.00002	cluster2
hsa04640	hematopoietic cell lineage	0.00002	cluster1
GO:0051437	positive regulation of ubiquitin-protein ligase activity during mitotic cell cycle	0.00003	cluster2
GO:0051439	regulation of ubiquitin-protein ligase activity during mitotic cell cycle	0.00006	cluster2

GO:0051438	regulation of ubiquitin-protein ligase activity	0.00006	cluster2
GO:0043122	regulation of i-kappab kinase/nf-kappab cascade	0.00006	cluster2
GO:0051436	negative regulation of ubiquitin-protein ligase activity during mitotic cell cycle	0.00007	cluster2
GO:0031145	anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process	0.00007	cluster2
REACT_383	dna replication	0.00007	cluster2
GO:0007242	intracellular signaling cascade	0.00010	cluster1
GO:0001775	cell activation	0.00010	cluster1
GO:0051352	negative regulation of ligase activity	0.00012	cluster2
GO:0051444	negative regulation of ubiquitin-protein ligase activity	0.00012	cluster2
hsa05010	alzheimer's disease	0.00012	cluster2
GO:0031397	negative regulation of protein ubiquitination	0.00013	cluster2
GO:0051340	regulation of ligase activity	0.00013	cluster2
GO:0030097	hemopoiesis	0.00013	cluster1
GO:0051056	regulation of small gtpase mediated signal transduction	0.00015	cluster1
GO:0031401	positive regulation of protein modification process	0.00018	cluster2
GO:0016569	covalent chromatin modification	0.00021	cluster1
GO:0031398	positive regulation of protein ubiquitination	0.00025	cluster2
GO:0044093	positive regulation of molecular function	0.00027	cluster2
GO:0016570	histone modification	0.00034	cluster1
GO:0007010	cytoskeleton organization	0.00035	cluster1
GO:0046649	lymphocyte activation	0.00037	cluster1
hsa04662	b cell receptor signaling pathway	0.00040	cluster1
GO:0010740	positive regulation of protein kinase cascade	0.00042	cluster2
GO:0048534	hemopoietic or lymphoid organ development	0.00042	cluster1
hsa03050	proteasome	0.00045	cluster2
GO:0051186	cofactor metabolic process	0.00051	cluster2
GO:0032270	positive regulation of cellular protein metabolic process	0.00058	cluster2
GO:0031400	negative regulation of protein modification process	0.00060	cluster2
GO:0051247	positive regulation of protein metabolic process	0.00062	cluster2
GO:0006954	inflammatory response	0.00089	cluster2
GO:0045087	innate immune response	0.00090	cluster2
GO:0043065	positive regulation of apoptosis	0.00099	cluster2
GO:0022900	electron transport chain	0.00101	cluster2
hsa04070	phosphatidylinositol signaling system	0.00122	cluster1
GO:0002684	positive regulation of immune system process	0.00125	cluster1
GO:0043068	positive regulation of programmed cell death	0.00125	cluster2
GO:0031396	regulation of protein ubiquitination	0.00140	cluster2
GO:0006917	induction of apoptosis	0.00143	cluster2

GO:0010942	positive regulation of cell death	0.00146	cluster2
GO:0012502	induction of programmed cell death	0.00156	cluster2
GO:0051248	negative regulation of protein metabolic process	0.00182	cluster2
GO:0002521	leukocyte differentiation	0.00182	cluster1
GO:0032269	negative regulation of cellular protein metabolic process	0.00236	cluster2
GO:0042110	t cell activation	0.00251	cluster1
GO:0002520	immune system development	0.00273	cluster1
hsa04660	t cell receptor signaling pathway	0.00275	cluster1
hsa04664	fc epsilon ri signaling pathway	0.00298	cluster1
GO:0016071	mrna metabolic process	0.00400	cluster1
GO:0042981	regulation of apoptosis	0.00436	cluster2
hsa04722	neurotrophin signaling pathway	0.00468	cluster1
GO:0030217	t cell differentiation	0.00469	cluster1
hsa04670	leukocyte transendothelial migration	0.00507	cluster1
GO:0030098	lymphocyte differentiation	0.00512	cluster1
hsa04130	snare interactions in vesicular transport	0.00518	cluster2
GO:0009611	response to wounding	0.00529	cluster2
REACT_6185	hiv infection	0.00539	cluster2
GO:0043085	positive regulation of catalytic activity	0.00567	cluster2
GO:0031399	regulation of protein modification process	0.00583	cluster2
GO:0048584	positive regulation of response to stimulus	0.00632	cluster2
GO:0043067	regulation of programmed cell death	0.00642	cluster2
GO:0008219	cell death	0.00710	cluster2
GO:0010941	regulation of cell death	0.00761	cluster2
GO:0032680	regulation of tumor necrosis factor production	0.00788	cluster2
GO:0010498	proteasomal protein catabolic process	0.00801	cluster2
GO:0043161	proteasomal ubiquitin-dependent protein catabolic process	0.00801	cluster2
GO:0050778	positive regulation of immune response	0.00845	cluster2
REACT_1538	cell cycle checkpoints	0.00861	cluster2
GO:0016265	death	0.00913	cluster2
hsa04142	lysosome	0.00999	cluster2
GO:0030036	actin cytoskeleton organization	0.01225	cluster1
REACT_6900	signaling in immune system	0.01262	cluster2
GO:0010033	response to organic substance	0.01485	cluster2
GO:0016044	membrane organization	0.01602	cluster2
GO:0001819	positive regulation of cytokine production	0.01619	cluster2
GO:0002263	cell activation during immune response	0.01645	cluster1
GO:0002366	leukocyte activation during immune response	0.01645	cluster1
hsa04012	erbb signaling pathway	0.01726	cluster1
hsa05220	chronic myeloid leukemia	0.01845	cluster1
hsa00190	oxidative phosphorylation	0.01964	cluster2
GO:0006468	protein amino acid phosphorylation	0.01973	cluster1
REACT_71	gene expression	0.02076	cluster1
REACT_125	processing of capped intron-containing pre-mrna	0.02122	cluster1

GO:0016192	vesicle-mediated transport	0.02166	cluster2
hsa05221	acute myeloid leukemia	0.02187	cluster1
GO:0001817	regulation of cytokine production	0.02338	cluster2
REACT_152	cell cycle, mitotic	0.02366	cluster2
hsa05215	prostate cancer	0.02439	cluster1
GO:0007015	actin filament organization	0.02479	cluster1
GO:0007265	ras protein signal transduction	0.02516	cluster1
GO:0007049	cell cycle	0.02761	cluster1
GO:0046578	regulation of ras protein signal transduction	0.02914	cluster1
hsa05223	non-small cell lung cancer	0.03243	cluster1
REACT_604	hemostasis	0.03701	cluster1
GO:0032675	regulation of interleukin-6 production	0.04054	cluster2
hsa05012	parkinson's disease	0.04090	cluster2
GO:0043086	negative regulation of catalytic activity	0.04256	cluster2
GO:0006357	regulation of transcription from rna polymerase ii promoter	0.04257	cluster1
GO:0032755	positive regulation of interleukin-6 production	0.04398	cluster2
GO:0010557	positive regulation of macromolecule biosynthetic process	0.04791	cluster1

Table S3: Gene set enrichment analysis results. Subject 10 (Symptomatic). Number of clusters: $K = 2$ (ICCC, GAP, MR).

Term	Name	Adj.pval	Subject	Cluster
GO:0009615	response to virus	0.00000	subject2	cluster1
GO:0006955	immune response	0.00000	subject2	cluster1
GO:0007156	homophilic cell adhesion	0.00000	subject11	cluster1
GO:0016337	cell-cell adhesion	0.00002	subject11	cluster1
GO:0007155	cell adhesion	0.00146	subject11	cluster1
GO:0022610	biological adhesion	0.00150	subject11	cluster1
hsa04622	rig-i-like receptor signaling pathway	0.00175	subject2	cluster1
GO:0042742	defense response to bacterium	0.00416	subject2	cluster3
GO:0009617	response to bacterium	0.00893	subject2	cluster3
REACT_152	cell cycle, mitotic	0.03241	subject2	cluster3

Table S4: Gene set enrichment analysis results. Subjects 2 and 11 (Asymptomatic). Number of clusters: $K = 3$ (ICCC) for Subject 2, $K = 1$ (ICCC) for Subject 11.

One clear message from the above analyses is that the enriched pathways from the symptomatic subject (Subject 10) contain much richer information about immune response than that from the two asymptomatic subjects (Subjects 2, 11), 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 which is characterized by the mobilization of many pathways. The difference in these

Term	Name	Adj.pval	Subject	Cluster
GO:0006955	immune response	0.00000	subject2	cluster1
GO:0009615	response to virus	0.00000	subject2	cluster1
GO:0007156	homophilic cell adhesion	0.00000	subject11	cluster2
GO:0016337	cell-cell adhesion	0.00006	subject11	cluster2
GO:0007155	cell adhesion	0.00050	subject11	cluster2
GO:0022610	biological adhesion	0.00051	subject11	cluster2
GO:0006952	defense response	0.00216	subject2	cluster1

Table S5: Gene set enrichment analysis results. Subjects 2 and 11 (Asymptomatic). Number of clusters: $K = 1$ (GAP) for Subject 2, $K = 2$ (GAP, MR) for Subject 11.

functional enrichments implies the specificity of biological processes of asymptomatic and symptomatic subjects.

From the data-driven point of view, it suggests that stronger immune-response to influenza infection is accompanied by clearer circular cluster pattern. This observation is also confirmed in the circular cluster analysis of the other subjects used in this study. Among the 10 significant pathways identified from Subjects 2 and 11, three are also identified from Subject 10: a) GO:0009615 (response to virus); b) GO:0006955 (immune response); c) REACT 152 (Cell Cycle, Mitotic). All of them have good reasons to be expressed in both symptomatic and asymptomatic subjects. The more interesting results lie in the pathways that are activated in Subjects 2 and 11, but not in Subject 10. One of these pathways is hsa04622 (RIG-I-like receptor signaling pathway), which 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. Other differences can be related to either cell adhesion (GO:0007156; GO:0016337; GO:0007155; GO:0022610) or bacterium responses (GO:0042742; GO:0009617). Further studies are needed to understand the biological implications of these pathways.

Next, we use Tables S4 and S5 to compare ICCC with GAP and MR criteria. For Subject 2, six functional terms are enriched based on ICCC/MR ($K = 3$) and only three are enriched based on GAP ($K = 1$). Among them two pathways, GO:0009615 (response to virus) and GO:0006955 (immune response), are the same. 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 are enriched by the ICCC/MR approach only: GO:0009617 (response to bacterium), REACT_152 (cell cycle, mitotic), and hsa04622 (RIG-I-like receptor signaling pathway). As mentioned before, pathway hsa04622 may play an important role in differentiating the symptomatic and asymptomatic subjects.

For Subject 11, the enriched terms based on ICCC ($K = 1$) are exactly the same as those based on GAP/MR ($K = 2$). It shows that ICCC provides an identical but more parsimonious model than GAP and MR in this case.

S5 The Proof of the Main Theorem

In this sections, we denote the probability space of infinite sequences of circular data by $(\Omega, \mathcal{F}_\infty, P)$, where $\Omega = \prod_{i=1}^{\infty} S^1$, $\mathcal{F}_\infty$ is the infinite product σ-algebra, and P is the infinite product probability measure of the uniform distribution on S^1. A set of circular data can be considered as $\mathbf{X}_n(\omega) := \{\mathbf{x}_i(\omega)\}_{i=1}^n$, $\omega \in \Omega$. Without loss of generality, we assume that the circular centers are indexed in the anti-clockwise direction, i.e. $\theta(\mu_{k+1}) > \theta(\mu_k)$, for $k = 1, 2, \ldots, K-1$. Denote $\alpha_k = \theta(\mu_k, \mu_{k+1})$, the angle between μ_k and μ_{k+1}. For convenience, we allow k, the index of clusters, to "wrap around" after it is smaller than 1 or greater than K, i.e. $\hat{\mu}_0 \equiv \hat{\mu}_k$ and $\hat{\mu}_{K+1} \equiv \hat{\mu}_1$.

Lemma S5.1 (Midpoint rule)**.** *If the circular centers (denoted as $\boldsymbol{\mu}_K$) are given, the kth circular cluster defined by this rule maximizes the within-cluster cosine similarity*

$$C_k(\boldsymbol{\mu}_K) := \Big[\theta(\mu_k) - \frac{\theta(\mu_k) - \theta(\mu_{k-1})}{2},\ \theta(\mu_k) + \frac{\theta(\mu_{k+1}) - \theta(\mu_k)}{2}\Big). \tag{S5.6}$$

Note that for the special case $K = 2$, $\mu_{k-1} \equiv \mu_{k+1}$ for $k = 1, 2$.

Proof. By definition, $C_k(\boldsymbol{\mu}_K)$ covers the region from the midpoint of $\theta(\mu_{k-1})$ and $\theta(\mu_k)$ to the midpoint of $\theta(\mu_k)$ and $\theta(\mu_{k+1})$. So for every $\mathbf{x} \in C_k$, the nearest (in angular distance) circular cluster center is μ_k. Thus $\langle \mathbf{x},\ \mu_k\rangle \geqslant \langle \mathbf{x},\ \mu_{k'}\rangle$, for $k' \neq k$. In other words, changing the cluster membership of any observations will decrease the within-cluster cosine similarity. □

From Lemma S5.1, we can rewrite (2.1) as

$$\mathrm{CS}(\boldsymbol{\mu}_K, \mathbf{X}_n) = \sum_{i=1}^{n}\sum_{k=1}^{K} 1_{C_k(\boldsymbol{\mu}_K)}(\theta(\mathbf{x}_i))\langle \mathbf{x}_i,\ \mu_k\rangle, \tag{S5.7}$$

where $C_k = [(\theta(\mu_k) + \theta(\mu_{k-1}))/2,\ (\theta(\mu_k) + \theta(\mu_{k+1}))/2)$. This indicates that $\boldsymbol{\zeta}$ and $\boldsymbol{\mu}_K$ are essentially equivalent.

Let $\mathrm{CS}^*(K, \mathbf{X}_n) = \sup_{\mu_K \in \Omega} \mathrm{CS}(\boldsymbol{\mu}_K, \mathbf{X}_n)$ be the maximum cosine similarity. By Equation (2.3),

$$\frac{1}{n}\log \hat{L}_n(K) = -\log 2\pi - \log I_0(\hat{\kappa}) + \hat{\kappa}\frac{\mathrm{CS}^*(K, \mathbf{X}_n)}{n}. \tag{S5.8}$$

Therefore, in order to study the asymptotic property of $\log \hat{L}_n(K)$ we must first understand the asymptotic properties of $\mathrm{CS}^*(K, \mathbf{X}_n)$ and $\hat{\kappa}$.

Define a real valued function $\mathcal{C}(\boldsymbol{\mu}_K, \mathbf{x}) : \Omega \times S^1 \to R^1$:

$$\mathcal{C}(\boldsymbol{\mu}_K, \mathbf{x}) = \sum_{k=1}^{K}\langle \mu_k,\ \mathbf{x}\rangle \cdot 1_{C_k(\boldsymbol{\mu}_K)}(\theta(\mathbf{x})) = \max_k \cos\left(\theta(\mu_k, \mathbf{x})\right).$$

This function is the cosine similarity of $\mathbf{x}$ to its nearest cluster center and it serves as the "building block" of the within-cluster cosine similarity function (S5.7). It is easy to show that $\mathcal{C}(\boldsymbol{\mu}_K, \mathbf{x})$ is absolutely continuous for both $\boldsymbol{\mu}_K$ and $\mathbf{x}$.

Lemma S5.2. *Under H_0, for a given K,*

$$\frac{1}{n}\sum_{i=1}^{n}\mathcal{C}(\boldsymbol{\mu}_K,\mathbf{x}_i)\xrightarrow{a.s.}E\mathcal{C}(\boldsymbol{\mu}_K,\mathbf{x})=\frac{1}{2\pi}\sum_{k=1}^{K}\alpha_k\sin\frac{\alpha_k}{2}. \tag{S5.9}$$

Proof. The first part is just the strong law of large numbers applied to a bounded continuous random variable. Denote $\text{arc}_k=[\theta(\mu_k),\theta(\mu_{k+1}))$, $\text{arc}_k^-=\big[\theta(\mu_k),\theta(\mu_k)+\frac{\alpha_k}{2}\big)$ and $\text{arc}_k^+=\big[\theta(\mu_k)+\frac{\alpha_k}{2},\theta(\mu_{k+1})\big)$. The second part is computed as follows.

$$\begin{aligned}E\mathcal{C}(\boldsymbol{\mu}_K,\mathbf{x})&=\sum_{k=1}^{K}E\left(\langle\mu_k,\ \mathbf{x}\rangle\Big|\theta(\mathbf{x})\in\text{arc}_k\right)P(\theta(\mathbf{x})\in\text{arc}_k)\\&=\sum_{k=1}^{K}\frac{1}{\alpha_k}\left(\int_{\text{arc}_k^-}\cos(\theta-\theta(\mu_k))\mathrm{d}\theta+\int_{\text{arc}_k^+}\cos(\theta(\mu_{_{k+1}})-\theta)\mathrm{d}\theta\right)\frac{\alpha_k}{2\pi}\\&=\frac{1}{\pi}\sum_{k=1}^{K}\int_0^{\frac{\alpha_k}{2}}\cos\theta\mathrm{d}\theta=\frac{1}{\pi}\sum_{k=1}^{K}\sin\frac{\alpha_k}{2}.\end{aligned}$$

□

Lemma S5.3. *$E\mathcal{C}(\boldsymbol{\mu}_K,\mathbf{x})$ is maximized when $\boldsymbol{\mu}_K$ forms an equi-distant grid on S^1, namely $\alpha_{k-1}=\alpha_k=2\pi/K$ for all $k=1,2,\ldots,K$. Consequently,*

$$\max_{\boldsymbol{\mu}_K}E\mathcal{C}(\boldsymbol{\mu}_K,\mathbf{x})=\frac{K}{\pi}\sin\frac{\pi}{K}.$$

Proof. Since $E\mathcal{C}(\boldsymbol{\mu}_K,\mathbf{x})=\frac{1}{\pi}\sum_{k=1}^{K}\sin\frac{\alpha_k}{2}$ is continuous and the set $\mathcal{A}=\Big\{\boldsymbol{\alpha}:\alpha_k\in\mathbb{R}^+,\ \sum_{k=1}^{K}\alpha_k=2\pi\Big\}$ is compact, there exists an $\boldsymbol{\alpha}^*\in\mathcal{A}$ which maximizes this function. It suffices to show that none of the uneven grid can maximize $E\mathcal{C}$. To prove this, we show that for every uneven grid $\boldsymbol{\mu}_K$, there exists a grid $\boldsymbol{\mu}'_K$ such that $E\mathcal{C}(\boldsymbol{\mu}'_K,\mathbf{x})>E\mathcal{C}(\boldsymbol{\mu}_K,\mathbf{x})$.

Assume that $\boldsymbol{\mu}_K$ forms an uneven grid of S^1, there must exist a k^*, such that $\alpha_{k^*-1}\neq\alpha_{k^*}$. Without loss of generality, we may assume $\alpha_1\neq\alpha_2$. Now define μ'_2 to be the circular midpoint between μ_1 and μ_3 so that $\alpha'_1=\alpha'_2=\theta(\mu_1,\mu_3)/2=(\alpha_1+\alpha_2)/2$ and let $\boldsymbol{\mu}'_K=(\mu_1,\mu'_2,\mu_3,\ldots,\mu_k)$.

$$\begin{aligned}\pi\big(E\mathcal{C}(\boldsymbol{\mu}',\mathbf{x})-E\mathcal{C}(\boldsymbol{\mu},\mathbf{x})\big)&=2\sin\frac{\alpha_1+\alpha_2}{4}-\sin\frac{\alpha_1}{2}+\sin\frac{\alpha_2}{2}\\&=2\sin\frac{\alpha_1+\alpha_2}{4}\Big(1-\cos\frac{\alpha_1-\alpha_2}{4}\Big)\end{aligned}$$

Since $\alpha_1,\alpha_2\in(0,2\pi]$ and $\alpha_1\neq\alpha_2$, we have $\dfrac{\alpha_1-\alpha_2}{4}\in(0,\pi/2)$. So $\cos\dfrac{\alpha_1-\alpha_2}{4}<1$, which completes the proof. □

Note that Lemma S5.3 states that under $E\mathcal{C}$ is maximized when $\boldsymbol{\mu}_K$ forms an even grid. It does not say anything about the position of this grid, which is equivalent to the location of μ_1. In fact, $E\mathcal{C}$ is invariant under rotation. So under H_0, the location of μ_1 is highly unstable for different n.

Theorem S5.4. *Under H_0, for a given K,*

$$\frac{1}{n}\mathrm{CS}^*(K, \mathbf{X}_n) \xrightarrow{a.s.} \frac{K}{\pi}\sin\frac{\pi}{K}. \tag{S5.10}$$

Proof. Combining Lemmas S5.2 and S5.3, and together with the fact that $\mathcal{C}(\boldsymbol{\mu}_K, \mathbf{x})$ is absolutely continuous for both $\boldsymbol{\mu}_K$ and $\mathbf{x}$, we have

$$\frac{1}{n}\mathrm{CS}^*(K, \mathbf{X}_n) = \sup_{\boldsymbol{\mu}_K}\frac{1}{n}\sum_{i=1}^{n}\mathcal{C}(\boldsymbol{\mu}_K, \mathbf{x}_n) \xrightarrow{a.s.} \sup_{\boldsymbol{\mu}_K} E\mathcal{C}(\boldsymbol{\mu}_K, \mathbf{x}) = \frac{K}{\pi}\sin\frac{\pi}{K}.$$

□

Theorem S5.4 states that under H_0, as $n \to \infty$, the optimal clustering of $\mathbf{X}_n$ is given when $\boldsymbol{\mu}_K$ forms an equidistant grid on S^1. For given n, K, and a set of cluster center estimates $\hat{\boldsymbol{\mu}}_K = \{\hat{\mu}_k\}_{k=1}^K$ obtained by maximizing (S5.7), the dispersion parameter can be estimated by solving

$$\frac{I_1(\kappa)}{I_0(\kappa)} = R_e := \sqrt{\frac{n}{n-K}\Big(\frac{1}{K}\sum_{k=1}^{K}\|\bar{y}_k\|^2 - \frac{K}{n}\Big)}, \tag{S5.11}$$

where $I_0(x)$ and $I_1(x)$ are the first kind of the modified Bessel function of order 0 and 1, respectively and $\bar{y}_k = \big(\sum_{i=1}^n 1_{C_k(\hat{\boldsymbol{\mu}}_K)}(\theta(\mathbf{x}_i))\big)^{-1}\sum_{i=1}^n \mathbf{x}_i 1_{C_k(\hat{\boldsymbol{\mu}}_K)}(\theta(\mathbf{x}_i))$ (Mardia and Jupp (2000)).

The statistic R_e used in Equation (S5.11) has the following asymptotic property.

Lemma S5.5. *Under H_0, for a given K,*

$$R_e = \sqrt{\frac{n}{n-K}\left(\frac{1}{K}\sum_{k=1}^{K}\|\bar{y}_k\|^2 - \frac{K}{n}\right)} \xrightarrow{a.s.} \frac{K}{\pi}\sin\frac{\pi}{K}.$$

Proof. For a given set of estimates $\hat{\boldsymbol{\mu}}_K = \{\hat{\mu}_k\}_{k=1}^K$, denote $\hat{\alpha}_k = \theta(\hat{\mu}_k, \hat{\mu}_{k+1})$, the angle between $\hat{\mu}_k$ and $\hat{\mu}_{k+1}$. By the strong law of large numbers and the continuous mapping theorem, we have

$$\begin{aligned}
\bar{y}_k|\hat{\boldsymbol{\mu}}_K &\xrightarrow{a.s.} E\left(\mathbf{x}|\mathbf{x}\in C_k(\hat{\boldsymbol{\mu}}_K)\right) = \frac{2}{\hat{\alpha}_{k-1}+\hat{\alpha}_k}\begin{pmatrix}\int_{C_k(\hat{\boldsymbol{\mu}}_K)}\cos\theta \mathrm{d}\theta\\ \int_{C_k(\hat{\boldsymbol{\mu}}_K)}\sin\theta \mathrm{d}\theta\end{pmatrix}\\
&= \frac{2}{\hat{\alpha}_{k-1}+\hat{\alpha}_k}\begin{pmatrix}\sin\left(\theta(\hat{\mu}_k)+\frac{\hat{\alpha}_k}{2}\right) - \sin\left(\theta(\hat{\mu}_k)-\frac{\hat{\alpha}_{k-1}}{2}\right)\\ -\cos\left(\theta(\hat{\mu}_k)+\frac{\hat{\alpha}_k}{2}\right) + \cos\left(\theta(\hat{\mu}_k)-\frac{\hat{\alpha}_{k-1}}{2}\right)\end{pmatrix},\\
\|\bar{y}_k|\hat{\boldsymbol{\mu}}_K\|^2 &\xrightarrow{a.s.} \|E\left(\mathbf{x}|\mathbf{x}\in C_k(\hat{\boldsymbol{\mu}}_K)\right)\|^2 = \left(\frac{2}{\hat{\alpha}_{k-1}+\hat{\alpha}_k}\right)^2\left(2-2\cos\left(\frac{\hat{\alpha}_{k-1}+\hat{\alpha}_k}{2}\right)\right)\\
&= \left(\frac{4}{\hat{\alpha}_{k-1}+\hat{\alpha}_k}\sin\frac{\hat{\alpha}_{k-1}+\hat{\alpha}_k}{4}\right)^2.
\end{aligned} \tag{S5.12}$$

By Lemma S5.3, $\hat{\alpha}_k \to \frac{2\pi}{K}$. So,

$$\|\bar{y}_k\|^2 \xrightarrow{a.s.} \left(\frac{K}{\pi}\sin\frac{\pi}{K}\right)^2.$$

Since $K \ll n$, R_e^2 and $\frac{1}{K}\sum_{k=1}^{K}\|\bar{y}_k\|^2$ share the same asymptotic property. □

When R_e is small (it can only happen when $K = 1$ and the observations roughly follow the uniform distribution), we obtain $\hat{\kappa}$ by solving (S5.11) with the function `uniroot()` in `R`. For most practical cases ($K \geq 2$ or $K = 1$ with non-uniform observations), R_e is close to 1. In this case, directly solving (S5.11) can be slow and numerically unstable. So we adopt the approximation formula proposed by Watamori (1995):

$$\hat{\kappa} \approx \frac{1}{R_e(1-R_e)(3-R_e)}. \tag{S5.13}$$

Equation (S5.13) provides a closed form representation of $\hat{\kappa}$ which facilitates the derivation of its asymptotic property.

Theorem S5.6. *Under H_0, for large K,*

$$\hat{\kappa} \xrightarrow{a.s.} \left(\frac{K}{\pi}\sin\frac{\pi}{K}\left(1-\frac{K}{\pi}\sin\frac{\pi}{K}\right)\left(3-\frac{K}{\pi}\sin\frac{\pi}{K}\right)\right)^{-1} \approx \frac{3K^2}{\pi^2}. \tag{S5.14}$$

Proof. When K is large, $\frac{\pi}{K} \approx 0$. Using the Taylor expansion, we have $\frac{K}{\pi}\sin\frac{\pi}{K} \approx 1 - \frac{\pi^2}{6K^2}$. By Lemma S5.5, we have

$$\hat{\kappa} = \frac{1}{R_e(1-R_e)(3-R_e)} \xrightarrow{a.s.} \left(\frac{K}{\pi}\sin\frac{\pi}{K}\left(1-\frac{K}{\pi}\sin\frac{\pi}{K}\right)\left(3-\frac{K}{\pi}\sin\frac{\pi}{K}\right)\right)^{-1} \approx \frac{3K^2}{\pi^2}.$$

□

Corollary S5.7. *Under H_0, for a given K,*

$$\log I_0(\hat{\kappa}) \xrightarrow{a.s.} \frac{3K^2}{\pi^2} - \log K - \frac{1}{2}\log\frac{6}{\pi}.$$

Proof. This corollary is a direct consequence of the Taylor expansion of $I_0(\kappa)$ provided in Abramowitz and Stegun (1964):

$$I_0(\kappa) \approx \frac{e^{\kappa}}{\sqrt{2\pi\kappa}}\left(1+\frac{1}{8\kappa}+O(\kappa^{-2})\right), \quad \log I_0(\kappa) \approx \kappa - \frac{\log 2\pi\kappa}{2} + O(\kappa^{-1}). \tag{S5.15}$$

□

S5.1 The Proof of Theorem 2.1

Proof. According to Equation (S5.8), Theorem 2.1 can be derived from Theorem S5.4, Theorem S5.6, and Corollary S5.7 immediately if $\log \hat{L}_n(K)$ is a bounded random variable.

Since both $\frac{1}{n}\mathrm{CS}^*$ and $\hat{\kappa}$ are bounded, we only need to show that $\log I_0(\hat{\kappa})$ is bounded according to equation (S5.8). It is well known that $I_0(x)$ is an increasing function and $I_0(0) = 1$ (Abramowitz and Stegun, 1964). So $0 < \log I_0(\hat{\kappa}) \leq \log I_0(\max(\hat{\kappa}))$, which completes the proof. □

S6 The Circular Shape of the First Two Functional Principal Components

For simplicity of presentation, we assume $(t_1, t_2, \ldots, t_J)$ form an even-grid on $[0, T]$ with step $\Delta t = \frac{T}{J}$. As in Equation (S3.1), the null hypothesis of testing differentially expressed genes can be defined as

$$H_{0,i}: \quad y_i(t) = c_{i0}, \quad t \in [0, T],\ i = 1, 2, \ldots, m, \tag{S6.16}$$

where the constant c_{i0} represents the "normal level" of the ith gene not affected by the exposure to influenza viruses. Again, the following F-statistic is used for testing $H_{0,i}$:

$$F_i = \frac{SS_i^0 - SS_i^1}{SS_i^1}, \tag{S6.17}$$

where $SS_i^0 = \sum_{j=1}^J (w_{ij} - \hat{c}_{i0})^2$ and $SS_i^0 = \sum_{j=1}^J (w_{ij} - \hat{y}_i(t_j))^2$, and $\hat{c}_{i0}$ and $\hat{y}_i(t)$ are the estimated gene expression profiles under the null and alternative hypothesis, respectively.

A reasonable way to estimate c_{i0} would be $\hat{c}_{i0} = \bar{w}_i$, the sample mean of $\mathbf{w}_i = (w_{i1}, \ldots, w_{iJ})$. Since the expression measurements are standardized, we have $SS_i^0 = \sum_{j=1}^J w_{ij}^2 = (J-1)\hat{\sigma}^2(\mathbf{w}_i) = J - 1$. Let $\mathrm{LS}_{\mathbf{1}} \subset \mathbb{R}^J$ be the 1-dimensional linear subspace spanned by vector $\mathbf{1} := (1, 1, \ldots, 1)$, $\mathrm{LS}_{\mathbf{1}}^{\perp}$ be the $(J-1)$-dimensional linear subspace orthogonal to $\mathrm{LS}_{\mathbf{1}}$, S^{J-1} be the standard sphere embedded in $\mathbb{R}^J$ and S^{J-2} be the standard sphere in $\mathrm{LS}_{\mathbf{1}}^{\perp}$. Simple algebra shows that

$$\frac{\mathbf{w}_i}{\sqrt{J-1}} \in S^{J-1} \cap \mathrm{LS}_{\mathbf{1}}^{\perp} \cong S^{J-2}. \tag{S6.18}$$

Define step functions $h_i(t) = \sum_{j=1}^{J-1} w_{ij} 1_{[t_j, t_{j+1})}(t)$ and $\hat{h}_i(t) = \sum_{j=1}^{J-1} \hat{y}_i(t_j) 1_{[t_j, t_{j+1})}(t)$, $t \in [0, T]$. Clearly,

$$\|h_i(t) - \hat{h}_i(t)\|_2^2 = \Delta t \sum_{j=1}^{J-1} (w_{ij} - \hat{y}_i(t_j))^2 \leqslant \Delta t \cdot SS_i^1. \tag{S6.19}$$

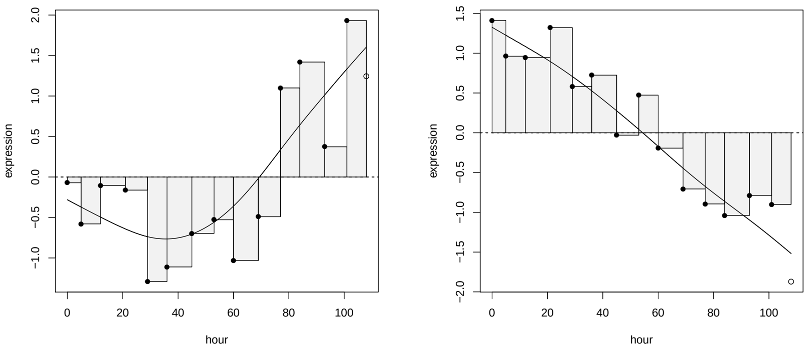
According to (S6.17), we know that SS_i^1 is small for significant genes. So for significant genes, when the number of time point J is large, we have

$$\begin{aligned} \|h_i(t) - \hat{y}_i(t)\|_2^2 &\leqslant \|h_i(t) - \hat{h}_i(t)\|_2^2 + \|\hat{h}_i(t) - \hat{y}_i(t)\|_2^2 \\ &\leqslant \Delta t \cdot SS_i^1 + \int_0^T \left(\hat{h}_i(t) - \hat{y}_i(t)\right)^2 \mathrm{d}t \approx 0, \end{aligned} \tag{S6.20}$$

indicating that $\|\hat{y}_i(t)\|_2^2 \approx \|h_i(t)\|_2$, $t \in [0, T]$.

By definition, the squared L^2-norm of $h_i(t)$ is

$$\|h_i(t)\|_2^2 := \int_{t_1}^{t_J} \left(\sum_{j=1}^{J-1} w_{ij} 1_{[t_j, t_{j+1})}(t) \right)^2 \mathrm{d}t = \Delta t \sum_{j=1}^{J-1} w_{ij}^2 = \frac{t_J - t_1}{J} \left(\|\mathbf{w}_i\|^2 - w_{iJ}^2 \right).$$



(a) Probe: `10062_at`.

(b) Probe: `1005_at`.

Figure S21: An illustration of step-function approximation of the fitted curve. Note that by construction, the last time point (w_{iJ}, marked as a circle in the figure) is not used in the step function. The Euclidean norm of both pre-processed expressions is $\|\mathbf{w}_i\| = \sqrt{14}$. The L^2-norm of the step function ($\|h_i(t)\|_2$) and the fitted curves ($\|\hat{y}_i(t)\|_2$), as well as the Euclidean length of the two principal components ($\sqrt{s_{i1}^2 + s_{i2}^2}$) are listed as follows:
a) Probe `10062_at`: $\|h_i(t)\|_2 = 9.83$, $\|\hat{y}_i(t)\|_2 = 8.32$, $\sqrt{s_{i1}^2 + s_{i2}^2} = 8.29$.
b) Probe `1005_at`: $\|h_i(t)\|_2 = 9.01$, $\|\hat{y}_i(t)\|_2 = 8.68$, $\sqrt{s_{i1}^2 + s_{i2}^2} = 8.67$.

If J is large and $w_{i,j}^2$ are not concentrated in the last time point, i.e. $w_{iJ}^2 \ll \|\mathbf{w}_i\|^2$, we can approximate w_{iJ}^2 by $\frac{\|\mathbf{w}_i\|^2}{J}$. Applying this approximation to the real data in Section 4, we have

$$\|h_i(t)\|_2^2 \approx \frac{t_J - t_1}{J}\left(\|\mathbf{w}_i\|^2 - \frac{\|\mathbf{w}_i\|^2}{J}\right) = \frac{(J-1)(t_J - t_1)}{J^2}\|\mathbf{w}_i\|^2 = \frac{14^2 \cdot 108}{15^2}.$$

Therefore, in our example,

$$\|\hat{y}_i(t)\|_2 \approx \|h_i(t)\|_2 \approx \sqrt{\frac{14^2 \cdot 108}{15^2}} \approx 9.70, \tag{S6.21}$$

which is a constant for all significant genes.

The first two functional principal components of subjects 2 and 10 explained 97.74% and 99.04% of the total variation, respectively. By (S6.21), we have $\sqrt{s_{i1}^2 + s_{i2}^2} \approx \|\hat{y}_i(t)\|_2 \approx 9.70$ for all significant genes. This explains the circular shape of the scatter plots of FPCs in Figure 5.4. These calculations are illustrated by Figure S21. The data used in this figure are the standardized expression levels of two typical significant gene (probe IDs: `10062_at` and `1005_at`) sampled from Subject 10. We see that $\sqrt{s_{i1}^2 + s_{i2}^2}$ are roughly the same for both genes.

The reason for the phenomenon that the FPCs of subject 10 show a more pronounced circular pattern than those of subjects 2 and 11 is largely related to the difference in the noise levels of these subjects. We observe that the gene expressions of subject 10 (symptomatic) have smaller noises than those of subjects 2 and 11 (asymptomatic), which implies that on average, SS_i^1 are relatively larger for subjects 2 and 11. So the approximation (S6.20) is not as good for these two subjects. This also explains why we detect much fewer significant genes for subjects 2 and 11 than subject 10.

Bibliography

Abramowitz, M. and Stegun, I. (1964). *Handbook of mathematical functions with formulas, graphs, and mathematical tables.* Dover publications.

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., Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature genetics.* **25**, 25-29.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B.* **57**, 289–300.

Bolstad, B., Irizarry, R., Astrand, M., and Speed, T. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics.* **19**, 185–193.

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 Genet.* **7**, e1002234.

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.

Mardia, K. and Jupp, P. (2000). *Directional statistics.* John Wiley & Sons Inc.

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

Ramsay, J. and Silverman, B. (2002). *Applied functional data analysis: methods and case studies.* Springer Verlag.

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.

Watamori, Y. (1995). Statistical inference of Langevin distribution for directional data. *Hiroshima Mathematical Journal.* **26**, 25-74.