Statistics and Its Interface Volume 1 (2008) 47–62

# Nonparametric Clustering of Functional Data

Haiyan Wang, James Neill and Forrest Miller

This paper presents a method for effectively detecting unknown patterns or clusters in high dimensional functional data. Examples of such data include gene expression levels measured over time from microarray experiments, functional magnetic resonance imaging (fMRI), mass spectrometry data from proteinomics, lipidomics etc. We define clusters through the unknown high dimensional multivariate distributions of all observations along each curve. Kullback-Leibler information and Mahalanobis generalized squared distance can fail to provide meaningful measure of distance between distributions in such high dimensional setting. We propose a new similarity measure and an agglomerative clustering algorithm, called PCLUST, to effectively differentiate among high dimensional populations. The algorithm produces invariant results under monotone transformations of data and does not require users to specify the number of clusters. Simulations show that PCLUST significantly outperforms 9 other popular algorithms in both clustering accuracy and robustness. An application in identifying biomarkers using time course gene expression data from Arabidopsis in response to environmental stresses is illustrated.

AMS 2000 subject classifications: Primary 62H30, 62G10, 62G35; secondary 62P10.
Keywords and phrases: Cluster analysis, Nonparametric inference, Hypothesis testing, Mixture model, High dimensional multivariate analysis, Time course gene expression microarray data, Lipid metabolism.

## 1. INTRODUCTION

This paper presents a method for effectively detecting patterns and clusters in high dimensional functional data. For example, in time course microarray experiments, thousands of gene expression data are taken over time. An important problem is to discover functionally related genes which could then be the target for new gene regulatory networks or functional pathways. Clustering such data can reveal groups of genes with similar expression patterns to identify such networks and pathways [13, 24, 1, 44, 10]. Similarly, data collected from mass spectrometers can contain thousands of different signal intensities plotted against the corresponding mass-to-charge ratios (m/z). Interest in such data includes, for example, the differentiation of genetically modified and conventional crops based on the entire metabolite composition [6]. This has been done through clustering techniques applied to such metabolome samples in order to determine similarities. In addition, significant effort has been seen in biomedical sciences with the objective of identifying or classifying disease status using mass spectra from biomedical samples (e.g. serum, blood plasma) (see for example, [16, 2]). In summary, the preceding data structures share some common characteristics. In particular, correlation generally exists among observations from each experimental unit along time points or different experimental conditions (e.g. microarray data), or the observations are spatially correlated (e.g. metabolome). It is of interest to identify unknown patterns or clusters where the number of clusters is typically unknown in practice.

The difficulties associated with clustering or classifying the types of data discussed above stem from both the high dimensionality (i.e. large number of variables) and the availability of only a small number of samples. For convenience, we will discuss correlation with reference to time. However, the methodology of this paper applies more generally in identifying patterns from correlated observations as discussed above. Available algorithms (e.g. k-means, hierarchical clustering, self organizing maps, fuzzy clustering) are used frequently in analyzing genomic data. However, these algorithms do not take into account the between time point correlation that is inherent in time course data and also require pre-specification of the number of appropriate clusters. Some recent algorithms are developed to accommodate the between time point correlation including smoothing based approaches (see [34, 21] and the references therein), Bayesian clustering [27], mixture model based clustering [15]. Effective use of smoothing based algorithms in real applications is hindered due to the extensive computational cost and the difficulty associated with the selection of the number of bases and knots [28]. Bayesian clustering [27] is based on an autoregression model which requires stationarity and the Markov property, and such properties are unlikely to hold for most time course microarray data. MCLUST [15] is promising in that it uses a general multivariate Gaussian mixture model to account for various possible correlation structures and automatically gives an estimate of the number of clusters by the Bayesian Information Criterion (BIC); however, such a model ignores the time order of observations and may fail to produce clustering when the number of time points is large. In addition, the performance of above algorithms can be very poor in time course microarray data even when the data are from

Gaussian distribution (see e.g., [35]). Further, the similarity measures used in above algorithms are not invariant to monotone transformations of the data. So the clustering applied to the original data or log transformed data often produces different results.

Here we introduce a dissimilarity measure that provides a robust and meaningful measure of distance for both categorical and numerical data. We also present a clustering approach, called PCLUST, that utilizes our dissimilarity measure to directly discover unknown patterns of functional data without a prior specification of either cluster number or pattern form. The combination of the robust dissimilarity measure together with PCLUST overcomes the aforementioned obstacles. In addition, PCLUST can use available grouping information in further clustering to facilitate clustering of very large data sets by splitting and combining the analysis. PCLUST provides the number of significantly different clusters and the experimental unit-to-cluster assignment. This makes PCLUST a very useful tool for clustering correlated data in practice.

## 2. SIMILARITY MEASURE AND THE MODEL

### 2.1 Notations and similarity measure

The objects to be clustered are realizations of random variables in $\mathbf{R}^b$ denoted by $\mathbf{X}_k = (X_{k1}, \ldots, X_{kb})'$, where $X_{kj}$ is the $j$-th measurement or feature of the $k$-th object. For example, for the time course microarray experiment $\mathbf{X}_k$ is a $b$-dimensional vector that records the measurements from the same gene over $b$ time points. Assume the vectors come from a mixture of $a$ unknown $b$-dimensional multivariate distributions.

To capture the dynamic nature of the experiments over time, a cluster is defined as the group of vectors that are generated by the same stochastic process (see e.g. [27]). That is, denoting the joint cumulative distribution function of $\mathbf{X}_k$ as $F_k(x_1, \ldots, x_b)$, we say $\mathbf{X}_{k_1}$ and $\mathbf{X}_{k_2}$ belong to the same cluster if and only if

$$F_{k_1}(x_1, \ldots, x_b) = F_{k_2}(x_1, \ldots, x_b), \forall x_1, \ldots, x_b.$$

To differentiate between distributions, various measures of distance have been used. For example, the Mahalanobis generalized squared distance ([33]) and the Kullback-Leibler (K-L) information are two commonly used such measures. However, for the high dimensional case, these measures can fail to provide a meaningful distance between distributions. For example, consider $m$-dimensional vectors $\mathbf{X}_i \sim MVN(\boldsymbol{\mu}_i, \mathbf{I}_m)$ for $i = 1, 2$ with $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_1 + m^{-\delta}\mathbf{1}_m$, where $\mathbf{I}_m$ and $\mathbf{1}_m$ are the $m-$dimensional identity matrix and a vector of ones. For $0 < \delta < 0.5$, the two distributions are nearly identical for large $m$ but the K-L distance and the Mahalanobis distance between these two distributions are $0.5m^{1-2\delta}$ and $m^{1-2\delta}$, respectively, which tend to infinity as $m \to \infty$. One can redefine these two distances by dividing by $m$. However, it is still necessary to estimate the high dimensional multivariate density function to use the K-L information and the inverse of the covariance matrix to use the Mahalanobis distance. It is generally not possible to obtain a reliable estimate of the high-dimensional multivariate density function or the large covariance matrix when the number of replications is small.

To avoid this problem, we suggest a new similarity measure. When two vectors $\mathbf{X}_{k_1}$ and $\mathbf{X}_{k_2}$ are from the same cluster (distribution), a valid test for the following hypotheses

$$H_0 : F_{k_1} = F_{k_2} \quad vs \quad H_a : F_{k_1} \neq F_{k_2} \tag{2.1}$$

would more often yield a large p-value as compared to when $\mathbf{X}_{k_1}$ and $\mathbf{X}_{k_2}$ are from different distributions. The power function assesses the ability of detecting deviations from the null hypothesis and therefore provides sample evidence to differentiate between clusters. One test for the hypotheses (2.1) is described in Section 3.1. We use the p-value produced from such test as the similarity measure between $\mathbf{X}_{k_1}$ and $\mathbf{X}_{k_2}$, and one minus the p-value as the dissimilarity measure (distance).

Similarly, when we have vectors $\mathbf{X}_1, \ldots, \mathbf{X}_{n_1}$ from a cluster characterized by a cumulative distribution function $F_1(x_1, \ldots, x_b)$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_{n_2}$ from another cluster characterized by a distribution $F_2(x_1, \ldots, x_b)$, we will define the similarity between the two clusters as the p-value for a valid test (see Section 3.2) for the following hypotheses:

$$H_0 : F_1 = F_2 \quad vs \quad H_a : F_1 \neq F_2.$$

If the observations of a vector from the same subject follow independent marginal distributions that remain unchanged over the varying experimental conditions, we call such a vector stochastically flat. For large datasets we have found it to be advantageous for efficiency purposes to first screen out stochastically flat vectors before clustering. The remaining vectors are then clustered based on the similarity measures from the outcomes of a sequence of hypothesis tests. Throughout the paper we use the terms vector and curve interchangeably, although we do not restrict the curve to be continuous.

### 2.2 The model

Before clustering, we have data of the form $X_{kj} \sim F_{kj}(x)$, $j = 1, \ldots, b, k = 1, \ldots, n$, where $X_{kj}$ is the $j$-th observation on the $k$th curve and $F_{kj}$ is an unknown distribution. We will use these notations in screening out flat curves.

Since the purpose of clustering is to cluster curves so that the between cluster variation at each time point is large and the within cluster variation is small, it is reasonable to allow different variances for curves in different groups and impose a constant variance condition for the curves to be clustered into the same group at the same time point. Specifically,

as different groups are formed during clustering, denote the data as

(2.2)
$$X_{ijk} \sim F_{ij}(x), \quad i = 1, \ldots, a,\ j = 1, \ldots, b,\ k = 1, \ldots, n_i,$$

where $X_{ijk}$ is the $j$-th observation on the $k$-th curve in the $i$-th group. The unknown distributions $F_{ij}$ are allowed to vary across groups and hence the variances of the observations for curves in different groups are allowed to vary. In this way, the curves at each corresponding time point may have large variation across groups but have low variation within clusters. If replications are available at each time point, it is possible to allow a curve specific variance at each time within clusters. Otherwise, as in this paper, we assume model (2.2). It is natural to assume further that the observations from different curves are independent and the dependence of the observations within each curve is due only to the fact that they are from the same subject (or gene). Note that we are working with the marginal model. Nothing is assumed for the joint distribution of observations. Both smooth and non-smooth curves can be analyzed in this model as $F_{ij}$ could be the distribution of a discrete random variable.

Throughout the paper, we assume the total number of clusters $a$ containing non-flat curves is small, and the number of time points $b$ is relatively large. The number of curves in each group, $n_i$, could be small or large because in some cases we might need to test whether several groups of curves should be combined as one group while each group contains only a small number of curves. Due to a totally unspecified covariance structure and possibly very small $n_i$, we assume each time series (curve) satisfies an $\alpha$-mixing condition. That is, assume for some sequence $\alpha_m \to 0$,

$$|P(A \cap B) - P(A)P(B)| \le \alpha_m,$$

holds for all $A \in \sigma(X_{i1k}, \ldots, X_{i\ell k})$, $B \in \sigma(X_{i,\ell+m,k}, X_{i,\ell+m+1,k}, \ldots)$, and all $i$, $k$, and $l \ge 1$ where $\sigma(\cdot)$ denotes the $\sigma$-field generated by the random variables. The $\alpha$-mixing condition basically requires the correlation between observations on the same curve to decay as the time lag $m$ increases (e.g. page 365 in [3]). We will make the assumption that the decay rate is $\alpha_m = O(m^{-5})$, which is weaker than the commonly used exponential decay rate as in the case of an autoregressive covariance structure. Many common time series models have been shown to satisfy this condition. In particular, both ARCH processes and additive AR processes with exogenous variables are $\alpha$-mixing under some mild conditions [22, 23]. Additionally, both stationary and nonstationary time series are allowed in this paper.

## 3. THE PCLUST ALGORITHM

### 3.1 Screening out flat curves

A curve is flat if the variation along the curve is due to only identically distributed independent random noise. This definition is also adopted in [27]. For this section, we confine our discussion to observations on a single curve.

The flat curves can be screened out by testing

(3.1)
$$H_0 : \{X_j, j = 1, \ldots, b\} \text{ iid} \sim A(x), \quad \text{for an unknown cdf } A(x).$$

Note that for a flat curve, the observations on the curve are not related to time in any aspect, including both first order property (mean) and higher order properties (variance, kurtosis, etc.). If the observations from a curve have constant mean but the variance varies over time, it should be classified as non-flat curve.

Notice that it is not reasonable to assume a common distribution for all nearly flat curves because we expect the subjects (e.g. genes) that produced the curves could behave differently. Therefore we need a goodness of fit test that is not restricted to any specific parametric distribution.

Our test statistic is developed as follows:

- Partition the observations on the same curve into $m$ groups where each group contains a small number of observations. We prefer to choose $m$ large so that the test is sensitive to departures from $H_0$ even for just a few observations. Let $m_i$ be the number of observations in the $i$-th group. In most cases, we can take $m_i = 2, 3$ or 4, and partition the observations such that the first $m_1$ observations (time-wise) are in group 1 and the next $m_2$ observations are in group 2 and so on.
- Denoting $F_i(x)$ to be the unknown distribution of the observations in the $i$-th group. The $F_i(x)$ for different curves can be different.
- Hypothesis (3.1) is equivalent to testing $F_1(x) = F_2(x) = \cdots = F_m(x)$ when each of the distribution has small number of independent observations. Now the flat curve screening problem becomes that of testing equality of $m$ distributions with $m$ large and the $m_i$ small.
- Rank all of the observations on the same curve and denote by $R_{ij}$ the overall (mid-) rank of the $j$-th observation in the $i$-th group. We will use the following results obtained in [42] to test $H_0$.

(3.2)
$$MST = \sum_{i=1}^{m} \frac{(\overline{R}_{i\cdot} - \widetilde{R}_{\cdot\cdot})^2}{m-1},$$
$$MSE = \sum_{i=1}^{m} \frac{S_i^2}{m m_i}, \quad F = \frac{MST}{MSE},$$

where $S_i^2 = (m_i - 1)^{-1} \sum_{j=1}^{m_i} (R_{ij} - \overline{R}_{i\cdot})^2$ and $\overline{R}_{i\cdot}$ is the average over the second index and $\widetilde{R}_{\cdot\cdot}$ is the average of $\overline{R}_{i\cdot}$ for all $i$.
Denote $H(x) = b^{-1} \sum_{i=1}^{m} m_i F_i(x)$, $\sigma_i^2 = \text{var}(H(X_{ij}))$, and

$$\nu^2 = \lim_{m\to\infty} \frac{1}{m}\sum_{i=1}^{m} \frac{1}{m_i}\sigma_i^2,$$
$$\tau = \lim_{m\to\infty} \frac{1}{m}\sum_{i=1}^{m} \frac{2\sigma_i^4}{m_i(m_i-1)}.$$

Assume $\nu^2 > 0$, then under $H_0$, $\sqrt{m}(F-1) \xrightarrow{d} N(0, \tau/\nu^4)$ as $m \to \infty$, if either the $m_i$ are uniformly bounded or go to infinity with $m$ such that $\min_{1\le i\le m} m_i / \max_{1\le i\le m} m_i = O(1)$. Note that $H(x)$ is the weighted average distribution and so $\sigma_i^2 < 1$ and $\nu^2$, $\tau$ always exist. They can be estimated consistently by the following statistics

$$\hat{\nu}^2 = m^{-1}\sum_{i=1}^{m} \frac{1}{m_i}\frac{S_i^2}{b^2}, \qquad \hat{\tau} = \sum_{i=1}^{m} \frac{2\widehat{\sigma}_i^4}{m_i(m_i-1)},$$

where $\hat{\sigma}_i^4$ is a Jackknife bias corrected unbiased estimator [11] using $R_{ij}/b$, $j = 1, \ldots, m_i$. Alternatively, in the case that $m_i \ge 4$, we can use an intuitively appealing estimator

$$\hat{\sigma}_i^4 = \frac{1}{4P_4^{m_i}} \sum_{j_1\ne j_2\ne j_3\ne j_4}^{m_i} \frac{(R_{ij_1} - R_{ij_2})^2 (R_{ij_3} - R_{ij_4})^2}{b^4},$$

where $P_4^{m_i}$ is the number of permutations of size 4 from $m_i$ objects.

The advantage of the above goodness of fit test is its nonparametric nature which is similar in spirit to the Kruskal-Wallis test. However, the Kruskal-Wallis test requires a large number of observations in each sample (group here) and the total number of groups is fixed. The Kruskal-Wallis test statistic is compared to a Chi-square distribution for critical values. Here the test is based on asymptotics for a large number of groups and compares the test statistic with the normal distribution for critical values.

Figure 1 gives a simulation study for type I error estimate when the true level is 0.05 using 1600 simulations. The goodness of fit test performs well for reasonably large $b$ under normal, lognormal, uniform, and Cauchy distributions.

It should be noted that only those curves that are not screened out will be clustered later and the screening is only to improve the efficiency for later clustering. If some of the flat curves fail to be screened out, they will eventually be clustered into one or multiple groups. However, if a non-flat curve is screened out, then it is permanently misclustered. For this reason, we would like to use a relatively conservative multiple comparison correction but effective enough to remove the majority of flat curves. The Bonferroni correction serves this purpose well and we reject $H_0$ when $p$-value $\le \alpha/N$, where $N$ is the total number of curves. We recommend to use a regular significance level such as $\alpha_A = 0.01$ or 0.05 when the number of time points is at least 100 when the underlying distribution of the data is not heavily skewed or heavy-tailed and at least 150 for heavily skewed or heavy-tailed. This is because the goodness of fit test gives good type I error estimate under such settings (see Figure 1). In cases that the number of time points is less than above threshold, we suggest to take smaller values for $\alpha_A$, such as $10^{-3}$, $10^{-5}$, etc, due to the fact that our goodness of fit test is liberal for small number of time points.

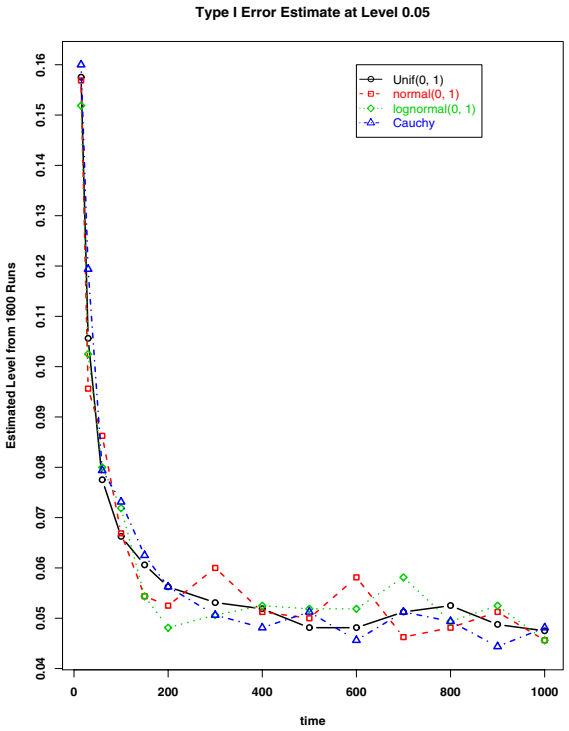


*Figure 1. Type I error estimate for the goodness of fit test given in Section 3.1 at level $\alpha = 0.05$. The test performs well for moderate to large number of time points (at least 150 for skewed distribution and at least 100 for other distributions).*

## 3.2 Comparing groups of curves

For $a$ groups of curves, we have $X_{ijk} \sim F_{ij}(x)$, $i = 1, \ldots, a$. To test if these groups are different, we test whether the distributions vary among groups at any time point. The hypothesis can be written as

$$H_0 : F_{ij}(x) = B_j(x), \text{ for all } i = 1, \ldots, a, j = 1, \ldots, b.$$

Let $R_{ijk}$ be the overall (mid-)rank of $X_{ijk}$ among all the observations and $\widetilde{R}_{.j.} = a^{-1}\sum_{i=1}^{a} \overline{R}_{ij.}$. Define

$$MSG = \frac{1}{(a-1)b}\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\overline{R}_{ij.} - \widetilde{R}_{.j.}\right)^2$$
$$MSE_G = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_i} \frac{(R_{ijk} - \overline{R}_{ij.})^2}{n_i(n_i-1)} \text{ and}$$
$$F_G = \frac{MSG}{MSE_G}.$$

**Theorem 3.1.** *Assume each curve $X_{ijk}, j = 1, 2, \ldots,$ is $\alpha$-mixing with $\alpha_m = O(m^{-5})$. Set $Y_{ijk} = H(X_{ijk})$, where $H(x) = N^{-1}\sum_{i=1}^{a}\sum_{j=1}^{b} n_i F_{ij}(x)$ with $N = \sum_{i=1}^{a} n_i b$, $\sigma_{ijj'} = cov(Y_{ijk}, Y_{ij'k})$, and define*

$$\zeta_1 = \frac{2}{a^2 b}\sum_{j=1}^{b}\sum_{j'=1}^{b}\sum_{i=1}^{a}\frac{\sigma_{ijj'}^2}{n_i(n_i-1)},$$

$$\zeta_2 = \frac{2}{a^2 b}\sum_{j=1}^{b}\sum_{j'=1}^{b}\sum_{i\neq i'}^{a}\frac{\sigma_{ijj'}\sigma_{i'jj'}}{n_i n_{i'}}.$$

*Then as $b \to \infty$ while $a$ remains fixed,*

*(1) if $n_i \geq 2$ are uniformly bounded, with*

$$\sigma^2 = \lim_{b\to\infty}\frac{E(MSE_G)}{N^2}, \qquad \tau_G^2 = \lim_{b\to\infty}\frac{\zeta_1 + \zeta_2/(a-1)^2}{N^4},$$

*under $H_0$,*

$$\sqrt{b}(F_G - 1) \xrightarrow{d} N\big(0, \tau_G^2/\sigma^4\big).$$

*(2) if $n_i \to \infty$ as $b \to \infty$, under the additional assumption $\max_i\{n_i\}/\min_i\{n_i\} = O(1)$, and let $n(a) = \min\{n_i, 1 \leq i \leq a\}$, and*

$$\sigma_*^2 = \lim_{b\to\infty}\frac{E(n(a)MSE_G)}{N^2},$$

$$\tau_{\gamma*}^2 = \lim_{b\to\infty} n^2(a)\frac{\zeta_1 + \zeta_2/(a-1)^2}{N^4},$$

*then under $H_0$,*

$$\sqrt{b}(F_G - 1) \xrightarrow{d} N\big(0, \tau_{\gamma*}^2/\sigma_*^4\big).$$

The above theorem is proved in [41]. The $\sigma_{ijj'}$ is the covariance between $H(X_{ijk})$ and $H(X_{ij'k})$ which account for the correlations among observations from the same curve. Through the $\alpha$-mixing condition, the time order of the correlations is taken into account without the need of specific form. To apply above theorem, we need a consistent estimate of the asymptotic variances. These can be achieved by replacing the $\sigma_{ijj'}$ term by an unbiased estimator. The sample covariance using $\widehat{H}(X_{ijk})$ and $\widehat{H}(X_{ij'k})$, $k = 1, \ldots, n_i$ can be used. Note that $N\widehat{H}(X_{ijk}) + 0.5$ is the overall rank of $X_{ijk}$ among all observations. So $\sigma_{ijj'}$ can be estimated by

$$\widehat{\sigma}_{ijj'} = (n_i - 1)^{-1}N^{-2}\sum_{k=1}^{n_i}(R_{ijk} - \overline{R}_{ij.})(R_{ij'k} - \overline{R}_{ij'.}).$$

## 3.3 Clustering

We now describe a clustering algorithm that uses the similarity measure defined in Section 2.1. The details of the algorithm are given below. Throughout all the tests involved, if there is no special statement, a Bonferroni correction is used for the significance level. The algorithm will stop when there are no groups to be combined.

The first step is to screen out the flat curves using the testing procedure given in Section 3.1. To further describe the algorithm, we introduce the following sets and operations.

- Let $S_0$ denote the set of all non-flat curves. These curves are ordered as they appear in the data set.
- $S$ is the updated set of unprocessed curves.
- $A$ is the updated ordered buffer set of curves.
- $\mathcal{C}$ is the updated ordered collection of clusters.
- $B$ is an empty set or contains a temporary cluster.

The following operations will be used. See the algorithm below.

- $O_1$: Remove the first curve from $S$ and place it into $A$ as the last curve.
- $O_2$: Compare the last curve in $A$ with each preceding curve until a compatible pair is found. If this occurs, remove the compatible pair from $A$ and place it into $B$. This comparison is done as follows. Calculate the difference between a pair of curves at each time point, and then apply the test in Section 3.1 on all such differences. If the test on the differences of the two curves yields a large $p$-value, we further apply a zero mean test to detect magnitude difference (i.e. large sample z-test of zero mean for i.i.d. samples; alternatively, the sign test could be used). The two curves will then be grouped together if the zero mean test is not significant. Theoretically, we can use the routine significance level as the threshold such as 0.05. However, each curve is assumed to be from a multivariate distribution and we need to judge whether two curves are from the same multivariate distribution with single observation. The parameters involved in each multivariate distribution are non-identifiable and so nonparametrically not solvable. Due to this reason, we use the additional parameter $\alpha_C$ to allow extra control on the initial pairing. The default threshold $\alpha_C$ for $p$-values in both tests is set to be 0.5 in the algorithm. It should be noted that this operation only provides initial pairing of the curves. It does not alone determine the final clustering. So we recommend to use a larger threshold to preclude the possibility of grouping unlike curves together. The larger, the more confidence we have on the alikeness of the paired curves. However, a threshold value too close to one may make the algorithm fail to find compatible pairs. In such case, the user can slightly reduce the value of $\alpha_C$ until the algorithm can proceed. We have tried $\alpha_C$ at values 0.99, 0.9, 0.8, 0.5, 0.05 as the threshold for the clustering in the real application in Section 5.1. They all produced identical clusters by the end of the algorithm. However, values smaller than 0.05 result in different clustering results. We suggest not to use too small values due to the reason stated above.

- $O_3$: If $\mathcal{C}$ is empty, remove the pair from $B$ and add it to $\mathcal{C}$ as a first cluster. Otherwise, use the test discussed in Section 3.2 to compare the cluster in $B$ with each successive cluster in $\mathcal{C}$. The first time compatibility (i.e., the test is nonsignificant) is achieved, remove the pair from $B$ and adjoin it to that cluster. If no compatibility is found, remove the pair from $B$ and add it to $\mathcal{C}$ as a new cluster. Note, at the end of $O_3$, $B$ is an empty set.
- $O_4$: Starting with the first curve in $A$, compare the curve with each cluster in $\mathcal{C}$ by comparing it with each member of that cluster. This is done using the same procedure as is given in $O_2$ but with the threshold being $\alpha_B = 0.001$ subject to Bonferroni correction. The first time compatibility is found with an entire cluster, remove the curve from $A$ and place it with that cluster. If the curve is not found to be compatible with any cluster of $\mathcal{C}$, adjoin it to $\mathcal{C}$ as a singleton cluster.
- $O_5$: Apply the compatibility test in Section 3.2 to pairs of clusters in $\mathcal{C}$ and combine compatible clusters to form a new cluster.

The tests in $O_3$, $O_4$, and $O_5$ have the same objective. In particular, the tests determine if any two clusters can be combined. The default threshold in the algorithm for these three operations is $\alpha_B = 0.001$ subject to Bonferroni correction (see also Section 3.2).

The following steps describe the algorithm.

1. Initialize $S = S_0$, $A = \emptyset$, $B = \emptyset$, $\mathcal{C} = \emptyset$.
2. Do operation $O_1$.
3. If $|A| = 1$ and $S \neq \emptyset$, go to 2.
4. If $|A| = 1$ and $S = \emptyset$, go to 11.
5. If $|A| \neq 1$, do operation $O_2$.
6. If $B = \emptyset$, go to 8.
7. If $B \neq \emptyset$, do operation $O_3$.
8. If $S \neq \emptyset$, go to 2.
9. If $S = \emptyset$ and $A \neq \emptyset$, go to 11.
10. If $S = \emptyset$ and $A = \emptyset$, go to 12.
11. Do operation $O_4$.
12. Do operation $O_5$.
13. Stop.

We remark that the clusters stored in $\mathcal{C}$ are meaningful groups at any stage of the algorithm. This follows since such groups have been judged to be significantly different, while the curves within each group have been determined to be similar by the statistical tests.

The algorithm is in part similar to agglomerative hierarchical clustering after stochastically flat curves have been removed. One difference is that PCLUST does not need the user to decide what is the appropriate number of clusters; another difference is that PCLUST does not need to search for the closest groups to merge, which is the major computational burden for agglomerative hierarchical clustering. In addition, each level of the hierarchy would represent a particular grouping of the data only if interest is confined to the details at that level. Instead, a user can avoid looking into the details of each level if the interest is in the final grouping formed when all curves are taken into account.

### 3.4 Properties of the clustering algorithm

We summarize the major properties of our clustering algorithm as below:

- The algorithm does not require a user to specify the number of clusters. Instead, this is determined automatically based on the user specified significance level for testing homogeneous distribution between groups of curves. This significance level is for controlling the family-wise type I error rate in case of Bonferronni correction. Alternatively, false discovery rate (FDR) may be used in screening for flat curves (see [12], [8] and the references therein ). This has dramatically simplified the difficulty associated with the user input as it is often very difficult to give a meaningful guess of the number of clusters. On the other hand, it has been a common practice to choose the level of significance in hypothesis testing.
- The algorithm always converges and the cost complexity of PCLUST is much less than traditional agglomerative clustering in that PCLUST merge any two groups that are judged to be generated by the same stochastic process at given confidence level while traditional hierarchical clustering searches for the closest two groups to merge. Such merge is meaningful as a result of the statistical equivalence of the two underlying multivariate distributions.
- Heteroscedasticity is allowed in the modelling so that each cluster can have its own variation at each time point. This general setup leads to higher power in detecting heterogeneous populations than using models assuming constant correlations (i.e. compound symmetry covariance structure) (see also [41]).
- Both the magnitude and shape of signal is taken into account in determination of clustering as is reflected in the test statistics.
- There is no smoothing involved and this leads to improved efficiency and clustering accuracy (see Section 4).
- In comparing groups with multiple curves, the hypotheses testing treats all curves in each group as a whole. This assures that group comparisons are done from a multivariate perspective rather than on a one-by-one basis.

## 4. COMPARISON WITH OTHER AVAILABLE METHODS

In this section, we compare PCLUST with nine available algorithms via simulations. These algorithms are KMEANS from R package stats, MCLUST(R package mclust), self-organizing map (R package som), clustering based on e-distance (R package energy), smoothing spline based clustering (SSCLUST from [21]), and several algorithms in R

package cluster (CLARA, FANNY, AGNES, and DIANA). Default values of the parameters are used for each algorithm except for those to be discussed in the latter part of this section.

## 4.1 Clustering error rate

Rand proposed the Rand index as the fraction of all pairs that are correctly put in the same cluster or correctly put in separate clusters to evaluate the similarity between two clustering partitions (1971). A problem with the Rand index is that the expected value of the Rand index of two random partitions does not take a constant value. [17] proposed the adjusted Rand index that takes the general form

$$\frac{\text{index} - \text{expected index}}{\text{maximum index} - \text{expected index}}.$$

The adjusted Rand index assumes the generalized hypergeometric distribution as the model of randomness. That is, the two partitions to be compared are picked at random such that the number of objects in the classes or clusters of each partition are fixed. The Rand index is typically much higher than the adjusted Rand index. The adjusted Rand index has expected value zero and maximum value of 1 which is achieved when the two partitions are identical up to renumbering of the subsets.

[25] evaluated many different indices for measuring agreement between two partitions in clustering analysis with different numbers of clusters, and they recommended the adjusted Rand index as the index of choice. We adopt the adjusted Rand index as our measure of agreement between clustering results and the truth in the simulation study to compare with other algorithms.

Five shapes of curves are generated, among which four groups of curves have non-flat shapes and the remaining one is flat. In each dataset, we generate 2,000 curves, 1,400 of which are stochastically flat curves and 150 are from each of the four other curve shapes. Among the ten algorithms, MCLUST and PCLUST can automatically estimate the number of clusters and we let them proceed without interference. KMEANS, DIANA, AGNES, ENERGYCLUST, FANNY, and CLARA requires the user to choose the number of clusters. We specified the number of clusters for these methods to be five, the true number of shapes, when we apply these algorithms on our generated data. As mentioned in the introduction, such prior information regarding the number of clusters is often not known so we can expect worse performance for these algorithms in practice as the guess in the number of clusters contributes additional uncertainty in applying these methods. SOM requires the user to give information about the two-dimensional grid of the clusters. We let the $x$-dimension of the map range among 3, 4 and 5, and let the $y$-dimension of the map range from 1, 2, 3, to 4. Each grid dimension specification results in a clustering and there are 12 such clusterings from the above grid ranges for each data set. We only collected the best adjusted Rand index (ARI) from the 12 clusterings. Since the true clustering is unknown in practice, a user cannot calculate ARI to guide the SOM clustering. Therefore the results reported for SOM are expected to be more optimistic than in practice.

SSCLUST will estimate the number of clusters after the user provides an initial number of clusters for the parameter "nclust". We assigned nclust = 4 and the number of chains to be 2. As commented by the authors of SSCLUST, it is faster but less accurate to use the number of chains to be 2. However, as SSCLUST is completely written in R, with the above less accurate specification, the computation requirement is still significantly higher than other algorithms we considered. In fact, SSCLUST could not provide a clustering result for 500 of our generated flat curves even after we increased the memory limit to 2,500 megabytes in R (with Intel Core 2 CPU 6400 @ 2.13 GHz and 2GB of RAM). Similarly, the algorithm stops with a memory error when the entire set of 2,000 curves or 800 curves are given to the algorithm. In Section 4.2 and 4.3, we apply SSCLUST only on a fraction of the generated data. In Section 4.4, we present a small simulation study for comparison of SSCLUST with PCLUST on 200 curves.

## 4.2 Clustering on Gaussian data

In this section, we compare the performance of the algorithms on normally distributed data. The observations of each stochastically flat curve are independent and identically distributed from a Gaussian distribution with mean $\nu$ and variance $\sigma_0^2$, where $\nu \sim Unif(-3,3)$ and $\sigma_0^2 \sim Unif(1.2,1.4)$. The non-flat curves are generated as a mean curve plus correlated random errors. The mean curves for each non-flat group are given below:

$$f_1(t) = 3\min\left\{\left(\frac{2-5t}{2}\right), \left[\left(\frac{5t-2}{3}\right)^2 + \sin\frac{5\pi t}{2}\right]\right\},$$
$$f_2(t) = -f_1(t), \quad f_3(t) = 3\cos(2\pi t), \quad f_4(t) = -f_3(t) \tag{4.1}$$

These four non-flat shapes were also used in [30]. However, only independent observations were considered by those authors. Here we consider the correlated case. Following [45] and [14], we sample the errors from a Gaussian process with zero mean and covariance function

$$cov(\epsilon_{ij}, \epsilon_{i'j'}) = \begin{cases} \sigma_j \sigma_{j'} e^{-|j-j'|/m}, & \text{if } i = i' \\ 0, & \text{if } i \neq i' \end{cases}, \tag{4.2}$$

where $m$ is the number of time points per curve and $\sigma_j \sim Unif(1.2,1.4)$. Under this model, the correlations among the observations from the same curve are high between successive observations and reduce as the time lag increases. This reflects the nature of biological processes. In practice, the number of observations per curve is often moderate. So we consider the case with 25 observations equally spaced over time on each curve.

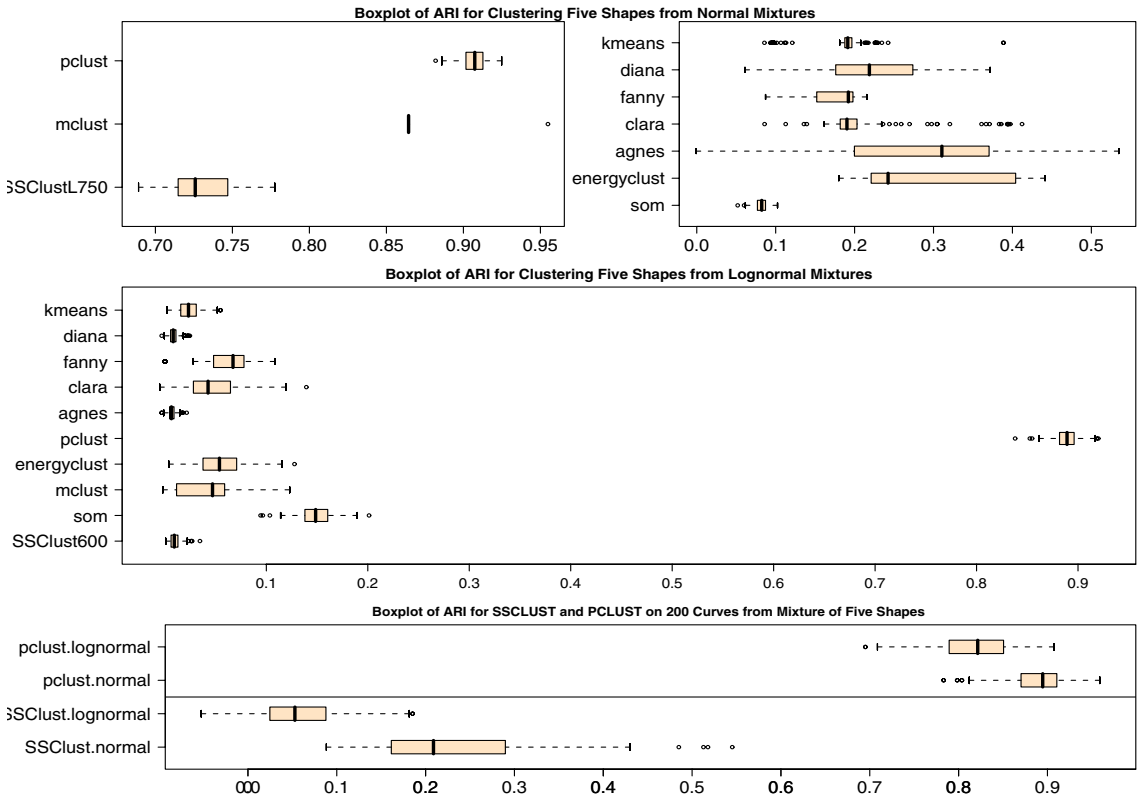


*Figure 2. Top two panels: Boxplot of ARI for clustering of 2000 curves from* **Normal** *mixtures of 5 shapes. Middle panel: Boxplot of ARI for clustering of 2000 curves from* **Lognormal** *mixtures of 5 shapes. Bottom panel: Boxplot of ARI for PCLUST and SSCLUST applied to 200 curves of normal and lognormal mixtures of five shapes. It is clear that PCLUST significantly outperforms the other algorithms in terms of both clustering accuracy and stability.*

*Table 1. Summary of ARI from 200 data sets for some algorithms. The performance data for the other algorithms are not reported here because their performance is not good (see Figure 2 for their graphical summary)*

| Algorithm | MixtureDist | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | stddev |
|---|---|---|---|---|---|---|---|---|
| SSClust(750) | normal | 0.689 | 0.715 | 0.726 | 0.731 | 0.747 | 0.777 | 0.019 |
| (600) | lognormal | 0.001 | 0.006 | 0.009 | 0.010 | 0.013 | 0.035 | 0.005 |
| | normal | 0.052 | 0.077 | 0.082 | 0.082 | 0.087 | 0.102 | 0.009 |
| som | lognormal | 0.094 | 0.138 | 0.149 | 0.149 | 0.160 | 0.201 | 0.018 |
| | normal | 0.864 | 0.864 | 0.864 | 0.865 | 0.864 | 0.955 | 0.006 |
| mclust | lognormal | −0.002 | 0.012 | 0.047 | 0.042 | 0.059 | 0.123 | 0.026 |
| | normal | 0.882 | 0.902 | 0.907 | 0.907 | 0.913 | 0.925 | 0.008 |
| pclust | lognormal | 0.838 | 0.882 | 0.889 | 0.889 | 0.896 | 0.920 | 0.012 |
| | normal | −0.001 | 0.201 | 0.311 | 0.299 | 0.370 | 0.535 | 0.108 |
| agnes | lognormal | −0.003 | 0.005 | 0.006 | 0.007 | 0.009 | 0.022 | 0.004 |

To compare the stability of the above algorithms, we repeat the data generation 200 times. The ARI of each clustering algorithm compared to the true membership is calculated for the 200 generated data sets and the boxplots of these ARI values are presented in the top panels of Figure 2. The numerical summaries are given in Table 1.

From Figure 2 and Table 1, we see that PCLUST has the best performance in terms of both clustering accuracy and stability. The average and median adjusted Rand indices are 0.907 for PCLUST. MCLUST achieved 0.864 for all but one data set. After multiple trial, we found that SSCLUST can run on 600 non-flat curves plus 150 flat curves when the data are generated from a Gaussian process. The results reported on the top left panel of Figure 2 for SSCLUST pertain to these 750 curves. SSCLUST obtained an average ARI of 0.731 for 750 curves. All of the other algorithms have ARI lower than 0.5 (with the exception of AGNES that have some ARIs between 0.5 and 0.535). The sample standard

deviation of the adjusted Rand indices reflects the stability of the clustering result. PCLUST, MCLUST, and SOM have comparable stability as the standard deviations from these algorithms are less than 0.01 for the 200 data sets. The remaining algorithms have at least twice the standard deviation as these three algorithms. AGNES and ENERGYCLUST are the least stable algorithms in their clustering performance.

### 4.3 Clustering on lognormal data

In microarray experiments, the data collected are often heavily skewed. Here we also examine the effect of skewness on clustering performance of the above algorithms. To do so, we double each observation generated in the simulation in Section 4.2 and transform it by the exponential function, i.e. exp(2 data) are our current observations, which follow lognormal distribution. We then apply each algorithm to such transformed data for clustering. Note that here we only make a monotone transformation and the curve shapes are preserved. This simple transform, however, poses extreme challenges to normal-based clustering methods. For example, before the transformation, SSCLUST can cluster 750 curves among which 600 are non-flat and 150 are stochastically flat curves. After the transformation, SSCLUST could not cluster these 750 curves. Instead, only 600 non-flat curves could be clustered by SSCLUST and the results reported in the middle panel of Figure 2 pertain to only these 600 curves. Other algorithms can still cluster the 2000 curves. But the accuracy significantly decreases except for PCLUST and SOM as compared to the results in Section 4.2 (see Figure 2 and Table 1). SOM shows marginal improvement but still exhibits low clustering accuracy. PCLUST is basically unaffected after the data are exponentially transformed. In fact, the adjusted Rand indices for all the other algorithms are very close to zero, the expected value of the adjusted Rand index. Thus we can say that the clustering assignments associated with all the algorithms except for PCLUST are only slightly better than making random guesses for these data sets.

### 4.4 Small simulation study to compare SSCLUST and PCLUST

Due to the severe computational cost demanded by SSCLUST, we give a small simulation study in this section to compare the performance of SSCLUST and PCLUST. Here we use the same five shapes and correlation structure as in Section 4.2. However, the number of curves for each shape is reduced to 1/10 of those generated in Section 4.2. That is, each data set contains 15 non-flat curves from each shape given in (4.2) and 140 stochastically flat curves. For accessing clustering stability, we repeat the data generation and clustering 200 times. Lognormally distributed data were obtained by exponentially transforming the normally distributed data with formula exp(2 normal data). The boxplots of the ARI as a measure of clustering consistency for each algorithm compared to the truth are summarized in the bottom panel of Figure 2. The mean (standard deviation) of the ARI for clustering the 200 datasets with normal mixtures is 0.229 (0.081) for SSCLUST and 0.890 (0.032) for PCLUST. For the lognormal mixture, SSCLUST failed to produce clustering result for 4 of the datasets. For the rest of the datasets, the mean (standard deviation) of the ARI for lognormal mixtures is 0.056 (0.046) for SSCLUST and 0.819 (0.042) for PCLUST. So PCLUST significantly outperforms SSCLUST in both normal and lognormal mixtures in terms of clustering accuracy and stability.

In summary, we compared the clustering performance of PCLUST with eight popular algorithms over 2,000 time series with five shapes and correlation ranging between 0.383 and 0.961 for observations along each time series. The results show that PCLUST has significantly much higher clustering accuracy and stability. In addition, PCLUST remains highly effective in differentiating shapes for highly skewed data as shown in the performance for the lognormal data. Further, PCLUST and SSCLUST are compared through a small simulation study. PCLUST significant outperforms SSCLUST in both clustering accuracy and stability.

## 5. CASE STUDY

In this section, we explore the application of PCLUST in two directions. In Section 5.1, we present the application of PCLUST in biomarker detection and in Section 5.2, we compare the performance of PLCUST with MCLUST in clustering of Arabidopsis gene expression data. The results in Section 5.1 is a small part of that obtained in Section 5.2.

Plants develop various biochemical and physiological mechanisms to respond and adapt to environmental stresses and thus acquire stress tolerance. For example, many plants increase in freezing tolerance in response to low non-freezing temperatures [36]. A fundamental goal of plant stress research is to identify genes that have roles in stress tolerance. Study for plant response to abiotic stress (cold/freezing, salt, drought, wounding, heat) has been an intensive research area for many years, but the molecular and cellular mechanisms are not well understood [36, 4]. Among these stresses, there have been many studies related to cold/freezing tolerance. Specifically, [40] summarized the currently identified genes that are cold regulated (mainly involved in two pathways, CBF and ZAT12), and indicated that there is strong evidence of additional cold-response genes and pathways yet to be discovered to advance our understanding of stress tolerance mechanisms.

### 5.1 Identifying biomarkers of Arabidopsis that respond to stress conditions

In this section, we will illustrate the use of the algorithm on contrast data, i.e. data collected from treatment versus control. This is typically useful for identifying biomarkers that respond to disease or treatment.

To further understand the mechanisms of response to stress, lipidomic analyses have been conducted in recent years as memberane lipids play both important structural and signalling roles in stress responses and regulate lipid composition and fatty acid saturation levels to optimize these functions. For example, it has been reported that lipid molecular species of phosphatidylcholine (PC), phosphatidylethanolamine (PE), phosphatidylglycerol (PG), phosphatidic acid (PA) and lysophospholipids level undergo changes when Arabidopsis thaliana are under cold or freezing stress [43]. It has also been found that digalactosyldiacylglycerol (DGDG) and monogalactosyldiacylglycerol (MGDG) molecular species levels change when Arabidopsis thaliana are subjected to wounding [5] or heat stress [7]. Experiments with mutants in specific gene have only partially explain these metabolic changes; the roles of many genes have yet to be elucidated. Here we will illustrate a use of analysis on gene expression data to aid in uncovering roles for membrane lipids in plant response to stresses and to identify in vivo functions of genes involved in lipid metabolism.

Possible metabolic reactions among above mentioned lipid species in the plant Arabidopsis thaliana are listed in Table 2 and the candidate gene classes that catalyze these reactions are given in the caption of Table 2. The average expression time series of these genes from wildtype Arabidopsis thaliana shoot tissue under control and stress (cold, salt, drought, wounding, and heat) were downloaded from the e-Northerns of Bio-Array Resource for Arabidopsis Functional Genomics database [37]. There are 34 time course measurements for each gene in both treatment and control (time points 1–6 are for cold stress, 7–12 for salt, 13–19 for drought, 20–26 for wounding and 27–34 for heat). The gene expression levels along each curve has big variations. For example, the mininum and maximum sample standard deviation along the curves is 0.617 and 917.2. The median and mean standard deviation along the curves are 19.48 and 116.1 respectively. So the data are heavily skewed.

The patterns under either treatment or control alone are not of much interest to the biologist. Instead, the change of pattern from control to treatment could give information of plant response to stresses. So we devote most of our efforts to interpretation of the results obtained from clustering the difference curve of the treatment and control. Genes showing differentially expressed pattern in comparing treatment with control obtained from PCLUST are reported in Table 3. For ease of discussion, we also included some genes that exhibit some pattern under the treatment or control.

Among the available clustering algorithms mentioned in the introduction, only EDGE proposed by [34] has the potential to detect differentially expressed genes by comparing treatment with control for time course data. When applied to this dataset, EDGE returned only error messages telling that the analysis returned one or more non-finite numbers. So the discussion in this section pertains to results from PCLUST only. The parameters used are $\alpha_A = 0.01$, $\alpha_B = 0.001$ and $\alpha_C = 0.9$. As commented in Section 3.3, $\alpha_C$ is only for initial pairing. For values at least 0.05 assigned to $\alpha_C$ (= 0.99, 0.9, 0.8, 0.5, or 0.05), the clustering results are identical. Different values of $\alpha_A$ and $\alpha_B$ could lead to different clustering results. These two values reflect the tolerance of an applied researcher on the family-wise type I error and need to be set with care. Here $\alpha_B = 0.001$ is the default value. A discussion about $\alpha_A$ is given in the next paragraph.

In Table 3, genes At3g25780, At4g35790, At4g00240, and At2g18730 are differentially expressed under both treatment and control as a function of time. However, only At4g35790 and At2g18730 have significantly different profiles of expression as a function of time in treatment and control. Therefore, simply studying gene expressions under the stress alone is not sufficient to discover which gene plays a crucial role in response to stress. Instead, it is important to compare treatment with a corresponding control to identify such genes. The gene expression profile for those genes differentially expressed in comparing treatment with control is given in Figure 3 and the differentially expressed genes under each stress are listed in Table 3. Genes At2g29980 and At2g42010 are reported to be differentially expressed only in the control for the current values of $\alpha_A$,$\alpha_B$, and $\alpha_C$. If $\alpha_A = 0.05$ is used as the threshold for screening of stochastically flat curves instead of 0.01, then both of them are also differentially expressed under the difference between treatment and control. In addition, there are three more genes that are differentially expressed under treatment, control or the difference between the treatment and control. The results discussed below are only for the $\alpha_A = 0.01$ case.

It is interesting to note that At3g11670 (DGD1) is found to be significantly different when comparing treatment and control but is not differentially expressed under control or treatment alone. This gene has been confirmed to be involved in up-regulation of relevent metabolites in [7] via comparing wildtype with DGD1 mutants. We summarize our findings below. We note that none of the differentially expressed genes reported below have been previously identified through gene expression data even though some have been validated by in vivo experiments.

Eleven genes were cold regulated, and almost all (eight genes) were upregulated by cold stress (Table 3). These include three phospholipase D (PLD$\alpha$1, At3g15730; PLD$\zeta$2, At3g05630; PLD$\gamma$2, At4g11830), four diacylglycerol kinases (ATDGK1, At5g07920; ATDGK2, At5g63770; At2g20900; At2g18730), one allene oxide cyclase (AOC1, At3g25760), one fatty acid desaturase (FAD2, At3g12120), one patatin-like protein (PLP7), and one DGD synthase (DGD1). It is known that diacylglycerols (DAG) are converted to phosphatidic acid (PA) by diacylglycerol kinases [26]. PA can also be generated by direct hydrolysis of membrane phospholipids, such as phosphatidylcholine and

*Table 2. Possible reactions involved in response to stress [43, 5, 7]. Some PLA (Patatin-like Acyl-Hydrolase) candidates are At1g33270, At2g26560, At2g39220, At3g54950, At3g57140, At3g63200, At4g29800, At4g37050, At4g37060, At4g37070, At5g04040, and At5g43590. Some PLD candidates are At1g52570 (PLD alpha), At1g55180 (PLD alpha), At3g15730 (PLD alpha), At5g25370 (PLD alpha), At2g42010 (PLD beta), At4g00240 (PLD beta), At4g11830 (PLD gamma), At4g11840 (PLD gamma), At4g11850 (PLD gamma), At4g35790 (PLD delta), At3g05630 (PLD zeta), and At3g16785 (PLD zeta). DGK (Diacylglycerol Kinase) candidates are At2g18730, At2g20900, At4g28130, At4g30340, At5g07920, At5g57690, At5g63770. DGD synthase candidates are At3g11670 (DGD1), At4g00550 (DGD2). LOX/AOS (plastid) candidates are At1g17420 LOX3 At1g67560 (similar to LOX3), At1g72520 (similar to LOX3), At3g22400 (similar to LOX3), At3g45140 (LOX2), At5g42650 (AOS). Allene Oxide Cyclase candidates are At1g13280, At3g25760, At3g25770, At3g25780. FAD genes candidates are At3g15850 (FAD5), At4g30950 (FAD6), At3g11170 (FAD7), At5g05580 (FAD8), At3g12120 (FAD2), At2g29980 (FAD3)*

| Reactant | Product | Catalyzed by | Reactant | Product | Catalyzed by |
|---|---|---|---|---|---|
| MGDG34:3 | DGDG34:3 | DGD synthase | MGDG34:3 | PA34:3 | lipase, DGK |
| MGDG36:6 | DGDG36:6 | DGD synthase | PC34:3 | PA34:3 | PLD |
| DGDG36:6 | DGDG38:6 | LOX/AOS/AOC | PE34:3 | PA34:3 | PLD |
| PC34:2 | LysoPC16:0 | PLA | PG34:3 | PA34:3 | PLD |
| PC34:3 | LysoPC16:0 | PLA | PI34:3 | PA34:3 | PLD |
| PC36:2 | LysoPC18:1 | PLA | PS34:3 | PA34:3 | PLD |
| PC36:3 | LysoPC18:1 | PLA | PG34:4 | PA34:4 | PLD |
| PC36:4 | LysoPC18:1 | PLA | MGDG34:6 | PA34:6 | lipase, DGK |
| PC36:3 | LysoPC18:2 | PLA | PC36:2 | PA36:2 | PLD |
| PC36:4 | LysoPC18:2 | PLA | PC36:4 | PA36:4 | PLD |
| PC36:5 | LysoPC18:2 | PLA | PE36:4 | PA36:4 | PLD |
| PC36:4 | LysoPC18:3 | PLA | PC36:5 | PA36:5 | PLD |
| PC36:5 | LysoPC18:3 | PLA | PE36:5 | PA36:5 | PLD |
| PC36:6 | LysoPC18:3 | PLA | PI36:5 | PA36:5 | PLD |
| PE34:3 | LysoPE16:0 | PLA | PS36:5 | PA36:5 | PLD |
| PE36:4 | LysoPE18:2 | PLA | DGDG36:6 | PA36:6 | lipase, DGK |
| PE36:4 | LysoPE18:3 | PLA | MGDG36:6 | PA36:6 | lipase, DGK |
| PE36:6 | LysoPE18:3 | PLA | PC36:6 | PA36:6 | PLD |
| PG34:4 | LysoPG16:1 | PLA | PE36:6 | PA36:6 | PLD |
| PG34:4 | LysoPG18:3 | PLA | PI36:6 | PA36:6 | PLD |
| MGDG34:1 | MGDG34:2 | FAD5 | PS36:6 | PA36:6 | PLD |
| MGDG34:2 | MGDG34:3 | probably FAD6 | PC34:1 | PC34:2 | FAD2 |
| PC34:3 | MGDG34:3 | ? | PC34:2 | PC34:3 | FAD3 |
| MGDG34:3 | MGDG34:4 | probably FAD6 | PC36:4 | PC36:5 | probably FAD3 |
| MGDG34:4 | MGDG34:5 | probably FAD7 | PC36:5 | PC36:6 | FAD3 |
| MGDG34:5 | MGDG34:6 | FAD7 | PE34:1 | PE34:2 | FAD2 |
| PC36:4 | MGDG36:4 | ? | PE34:2 | PE34:3 | FAD3 |
| PC36:6 | MGDG36:6 | ? | PE36:4 | PE36:5 | probably FAD3 |
| MGDG36:6 | MGDG38:6 | LOX/AOS/AOC | PE36:5 | PE36:6 | FAD3 |
| PC34:2 | PA34:2 | PLD | PG34:1 | PG34:2 | FAD4 |
| PE34:2 | PA34:2 | PLD | PG34:2 | PG34:3 | probably FAD6 |
| PG34:2 | PA34:2 | PLD | PG34:3 | PG34:4 | FAD7 |
| PI34:2 | PA34:2 | PLD | PE34:2 | PS34:2 | PS synthase |
| PS34:2 | PA34:2 | PLD | PE34:3 | PS34:3 | PS synthase |
| DGDG34:3 | PA34:3 | lipase, DGK | PE42:2 | PS42:2 | PS synthase |
| | | | PE42:3 | PS42:3 | PS synthase |

phophatidylethanolamine by phospholipase D (PLD). There are accumulating lines of evidence that PA is involved in cold signaling, see for example, [43] and the references therein. Studies with Arabidopsis mutant or overexpression lines of PLDs supported their involvement in plant cold tolerance [43, 20]. PLD$\alpha$1, ATDGK1, ATDGK2 were also reported to be cold regulated in [19].

Six genes were found to be heat regulated. They are PLD$\gamma$2 (At4g11830), DGD1 (At3g11670), PLD$\alpha$1 (At3g15730), PLP7 (At3g54950), a putative diacylglycerol kinase (At2g18730), and FAD2(At3g12120). The role of *DGD*1 in response to heat stress has been confirmed by [7]. Metabolically, heat acclimation in wildtype Arabidopsis is associated with an increase in the relative amount of DGDG

*Table 3. The second column indicates differentially expressed genes under treatment, control, and difference. Genes differentially expressed in the treatment are marked as T# with the # giving the group the gene is clustered into. Similarly, genes differentially expressed in the control are marked as C#, and differentially expressed in comparing the treatment with control are marked as D#. The genes differentially expressed under the treatment are each clustered as a singleton cluster; the genes clustered together under the control exhibit cyclic behavior; and the genes clustered together in comparing treatment with control mostly show a dominant peak under the cold stress (see Figure 3). Column 3–7 indicate genes activated under cold, salt, drought, wounding, and heat stress, respectively. '+': positive sign indicates the gene expression levels stay up-regulated in corresponding stress compared to the control; '-': negative sign indicates the gene expression levels stay down regulated in corresponding stress compared to that in the control. '+3': means the gene expression levels are up-reguated for only the first three time points from the wounding stress*

| ID | Diff | Cold | Salt | D | W | Heat | Short Description |
|---|---|---|---|---|---|---|---|
| At2g20900 | D5 | − | | + | +3 | | diacylglycerol kinase |
| At1g72520 | T1 | | | | | | lipoxygenase, putative |
| At3g25780 | T2 C5 | | | | | | allene oxide cyclase3 |
| At5g25370 | C5 | | | | | | phospholipase D alpha3 |
| At1g13280 | T4 | | | | | | allene oxide cyclase4 |
| At3g25760 | C4 D1 | − | + | + | + | | allene oxide cyclase1 |
| At5g42650 | D4 | | | + | + | | AOS;hydro-lyase/ oxygen binding |
| At2g29980 | C5 | | | | | | fatty acid desaturase3; |
| At2g42010 | C3 | | | | | | phospholipase D beta1 |
| At3g12120 | D2 | + | + | | − | − | fatty acid desaturase2 |
| At4g35790 | T3 C5 D3 | | + | | | | phospholipase D delta |
| At4g00240 | T5 C2 | | | | | | phospholipase D beta2 |
| At4g11840 | C1 | | | | | | phospholipase D gamma3 |
| At3g11670 | D4 | + | | | | + | DGD1 |
| At3g05630 | C5 D4 | + | + | | | | phospholipase D zeta2 |
| At3g15730 | C5 D4 | + | + | | + | − | phospholipase D alpha1 |
| At5g07920 | C5 D4 | + | + | | +3 | | ATDGK1(diacylglycerol kinase1) |
| At3g54950 | D4 | + | + | + | +3 | − | patatin-like protein7 |
| At2g18730 | T6 C5 D4 | + | | | +3 | − | diacylglycerol kinase, putative |
| At2g26560 | C5 | | | | | | phospholipase A 2a; nutrient reservoir |
| At5g63770 | C5 D4 | + | | | +3 | | ATDGK2(diacylglycerol kinase2) |
| At4g11830 | C5 D4 | − | − | − | | + | phospholipase d gamma2 |

lipid species, a dramatic increase in the DGDG to MGDG ratio and a moderate increase in the saturation of fatty acids. DGD1 catalyzes the conversion of MGDG to DGDG in chloroplast. [7] report that mutants in the DGD1 gene have decreased basal thermotolerance and decreased ability to acquire thermotolerance to high temperature stress. Such mutants show reduced DGDG levels, the ratio of DGDG to MGDG, and blocked the accumulation of chlorophyll within cotyledons following a mild heat treatment.

In the same experiment by [7], levels of other lipid species were also recorded in their Table 1 (page 1443). Beyond the obvious change in DGDG and MGDG as noted by these authors, it is clear that PA, PC and lyso lipids also showed significant increase for the mutants under heat acclimation. This point is difficult to detect for the wild type from Table 1 of [7] due to the large variation of these lipid species, even though the means of these lipid species under heat are higher than those under normal temperature. With our results on the gene expression data, we can offer an explanation of the increase in PA, PC and lyso lipids by looking at the metabolic reactions catalyzed by the significant genes found here. FAD2 catalyzes the reaction from PC34:1 to PC34:2, and PC34:2 is converted to PC34:3 (by FAD3), and then PC34:3 is converted to PA34:3 by PLD genes. PC, PE and PG may be converted to LysoPC, LysoPE, and LysoPG by the PLP enzymes. Diacylglycerol kinase is involved in conversion of MGDG and DGDG to PA species. Which specific genes in the PLD, PLP, and diacylglycerol kinase families catalyze these reactions under most specific stress conditions are not known. The genes we report here provide specific target/biomarkers for further assay by biologists.

For brevity, we report the significant genes found for the other stresses without detail (Table 3 and Figure 3).

Eight genes were salt regulated. They include FAD2 (At3g12120), four phospholipase D (PLD$\delta$, At4g35790; PLD$\alpha$1, At3g15730; PLD$\zeta$2, At3g05630; PLD$\gamma$, At4g11830), AOC1 (At3g25760); DGK1 (At5g07920); and PLP7 (At3g54950).

Five genes involved respond to drought stress. They are ALLENE OXIDE SYNTHASE (At5g42650), diacyl-

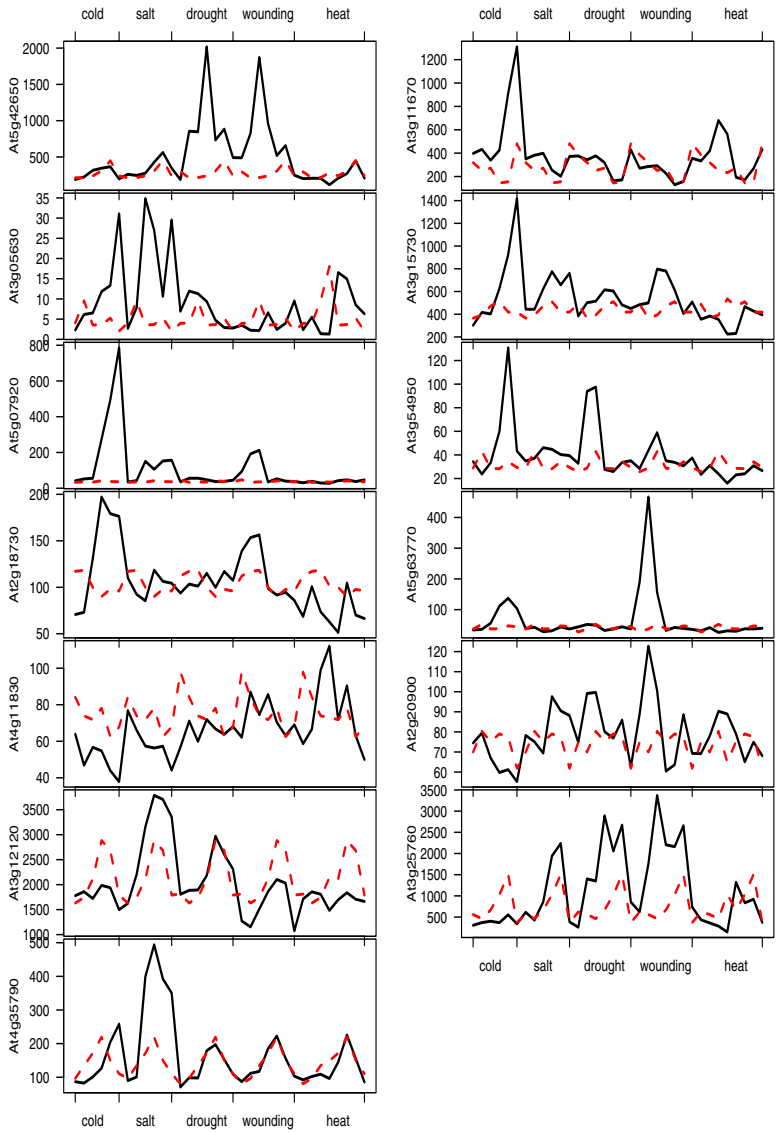


*Figure 3. This figure plots the profiles of differentially expressed genes in comparing treatment with control. - - -: red dashed line for control; — : black solid line for treatment. The ticks on the horizontal axis are marked at starting of the cold experiment and the end point of each stress. (Note: The beginning time of each stress is one unit after the end point of the previous stress. The total number of time points is 34 on each panel.)*

glycerol kinase (At2g20900), AOC1 (At3g25760), PLP7 (At3g54950), and PLD$\gamma$2 (At4g11830).

Nine genes were found to have induced change during wounding stress. These include Allene Oxide Synthase (At5g42650), AOC1 (At3g25760), PLD$\alpha$1 (At3g15730), PLP7 (At3g54950), ATDGK1 (At5g07920), ATDGK2 (At5g63770), diacylglycerol kinase (At2g20900), diacylglycerol kinase, putative (At2g18730), FAD2 (At3g12120).

We remark that the results indicate that a gene often has multiple roles over different stresses. For example, PLD$\alpha$1 (At3g15730) is upregulated in cold, salt, wounding but downregulated in the heat stress. Similarly, diacylglycerol kinase (At2g20900) showed differential expression in Arabidopsis under cold (downregulated), drought (upregulated) and wounding (upregulated) stresses. This gene has been previously studied in the analysis of leaves of Columbia-0 Arabidopsis thaliana transformed with an anti-sense mitochondrial alternative oxidase (AOX) in the acclimation to oxidative stress [38]. But its role under the acclimation to cold, drought, and wounding stresses has not been noted. It is difficult to put plants under multiple stress conditions in lab experiments even though such mutiple stress conditions are more natural in reality. The findings from the PCLUST algorithm applied to available gene expression data can offer insight for biologists to generate biological hypotheses for further lab experiments.

## 5.2 Clustering of gene expression data

By observing genes with similar or opposite expression patterns, one can infer that these genes are probably co-regulated. By analyzing time-dependent expression data, one can derive plausible causality relationships among genes [39, 9, 18], hence inferring initial information on how gene might be regulated. Appropriate clustering techniques can be used to extract and detect unknown patterns from a large amount of noisy data.

Raw time-course gene expression data for 13,236 genes in wild-type Arabidopsis thaliana shoot tissue grown under nine stress conditions (cold, osmotic, salt, drought, genotoxic, oxidative, UV-B, wounding, heat) and their corresponding control values were downloaded from the BioArray Resource for Arabidopsis Functional Genomics. The data contain gene expression levels at 58 time points under the stresses and 58 time points under corresponding control (the data in Section 5.1 is a subset of the current section). For each gene, the ratio of the expression level under a given stress to the control level was calculated for each time point. The PCLUST was applied to the time course ratio data. During the clustering, the Bonferroni correction was used to account for multiple comparison issue. Note that traditional parametric methods using Bonferroni correction applied to microarray data are often so conservative that alternative controls such as FDR approaches need to be used, but the current nonparametric method does not suffer from this drawback. When $\alpha_A = 10^{-7}, 10^{-5}, 0.001$, and 0.01, respectively, the number of detected stochastically non-flat genes is 1589, 2068, 2800, and 3338, respectively. The average number of false discoveries with above $\alpha_A$ values from 13,236 curves is $1.33 \times 10^{-6}$, $1.33 \times 10^{-3}$, 0.133, 13.236, and 132.36, respectively, provided that the flat curve screening test has correct type I error estimate. To avoid nonflat curves being screened out and considering the test is slightly liberal (type I error estimate at level 0.05 is between 0.07 and 0.08 for Uniform and normal distribution and between 0.10 and 0.12 for lognormal and Cauchy distribution, see Figure 1) when the number of time points is 58, we recommend to use $\alpha_A = 10^{-10}, 10^{-7}$ or $10^{-5}$. With $\alpha_A = 10^{-10}$, $\alpha_B = 0.05$, $\alpha_C = 0.95$, there are 1,041 genes found to be not stochastically flat (i.e. differentially expressed). These genes were clustered into 74 groups.

MCLUST failed to perform clustering for the whole set of 13,236 genes and reported memory allocation error. We then tried MCLUST on the 1,041 non-flat genes found by PCLUST. This time MCLUST finished running and produced three clusters. This is a very big contrast to the number of clusters obtained by PCLUST (74). Note that the dataset analyzed in Section 5.1 is a subset of the current section. It is clear from Figure 3 that there are more than three clusters in the data.

# 6. SUMMARY

In this article, we present an algorithm for clustering a large number of curves in the presence of possibly heteroscedastic temporal dependency (indeed, any underlying distributional changes along the time points are allowed). Unlike some other papers that have restricted their attention to autoregressive processes, we allow the dependency of temporal observations to have any form as long as the dependency decays suitably fast (at least at a polynomial rate) as the time lag between observations increases. In simulations, our algorithm achieved high clustering accuracy and stability. In addition, the simulation also confirms that taking the correlation into account in the modelling can improve clustering performance. We applied our algorithm to Arabidopsis data to study plant response to stress conditions in two scales. In the large scale study we illustrate the clustering of the whole set of genes and in the small scale study we give detailed discussion on our findings in connection to recent efforts in lipidomics.

Beyond the potential applications mentioned in the introduction, our algorithm fits into the framework of studies for fast functional magnetic resonance imaging (fMRI) data and ordered categorical functional data (such as DNA barcoding). FMRI contains data for analyzing event related physiological process at a time scale of seconds. Stochastically flat curves reflect noises in collecting fMRI data while detecting and grouping non-flat curves can reveal active regions of the brain in response to stimuli. DNA barcoding consists of a sequence of discrete genetic codes. It has been shown to be a useful tool for differentiating biological species (e.g. [29, 31]). The application of clustering and classification on such high dimensional categorical data offers the potential for rapid automated identification of known species and the discovery of new ones. The R code for the clustering algorithm is available upon request.


# ACKNOWLEDGEMENTS

The authors are grateful to Professor Ruth Welti from the Division of Biology at Kansas State University for helpful comments and encouragement. Her patience and discussions have improved the first author's knowledge on lipid metabolism. The authors are also grateful to the referee, the Associate Editor, and the Editor for their sharp understanding of the issues and providing very useful comments.

*Received 14 November 2007*


# REFERENCES


[1] Arbeitman, M., Furlong, E., Imam, F., Johnson, E., Null, B., Baker, B., Krasnow, M., Scott, M., Davis, R. and White, K. (2002). Gene expression during the life cycle of drosophila melanogaster. *Science* **297** 2270–2275.

[2] Bao-Ling, A., Yinsheng, Q., John, W., Michael, D., Mary, A., Lisa, H., John, O., Paul, F., Yutaka, Y., Ziding, F. and George, L. (2002). Serum protein fingerprinting coupled with

a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res.* **62** 3609–3614.
[3] Billingsley, P. (1995). Probability and Measure, Third Edition. Wiley. MR1324786
[4] Browse, J. and Xin, Z. (2001). *Curr. Opin. Plant Biol.* **4** 241–246.
[5] Buseman, C., Tamura, P., Sparks, A., Baughman, E., Maatta, S., Zhao, J. and Roth, M., Esch, S., Shah, J., Williams, T. and Welti, R. (2006). Wounding stimulates the accumulation of glycerolipids containing oxophytodienoic acid and dinor-oxophytodienoic acid in arabidopsis leaves. *Plant Physiol.* **142** 28–39.
[6] Catchpole, G. S., Beckmann, M., Enot, D. P., Mondhe, M., Zywicki, B., Taylor, J., Hardy, N., Smith, A., King, R. D., Kell, D. B., Fiehn, O. and Draper, J. (2005). Hierarchical metabolomics demonstrates substantial compositional similarity between genetically modified and conventional potato crops. *Proc. Nat. Acad. Sci.* USA **102**(40) 14458–14462.
[7] Chen, J., Burke, J. J., Xin, Z., Xu, C. and Velten J. (2006). Characterization of the arabidopsis thermosensitive mutant atts02 reveals an important role for galactolipids in thermotolerance. *Plant, Cell and Environment* **29**(7) 1437–1448.
[8] Clarke, S. and Hall, P. (2007). Robustness of multiple testing procedures against dependence. Submited for publication.
[9] Covert, M. W., Schilling, C. H. and Palsson, B. (2001). Regulation of gene expression in flux balance models of metabolism. *J. Theor. Biol.* **213** 73–88.
[10] Dequeant, M.-L., Glynn, E., Gaudenz, K., Wahl, M., Chen, J., Mushegian, A., Pourqui, O. (2006). A complex oscillating network of signaling genes underlies the mouse segmentation clock. *Science* **314**(1595) 1595–1598.
[11] Efron, B. (1982). *The Jackknife, the bootstrap, and other resampling plans*, Volume 38 of CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM. MR0659849
[12] Efron, B.(2007). Correlation and large-scale simultaneous significance testing. *J. Amer. Statist. Assoc.* **102** 93–103. MR2293302
[13] Eisen, M., Spellman, P., Brown, P. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression. *Proc. Nat. Acad. Sci.* USA **95**(25) 14863–14868.
[14] Fan, J. and Zhang, J. (2000). Two-step estimation of functional linear models with applications to longitudinal data. *J. of Royal Statistical Society B* **62** 303–322. MR1749541
[15] Fraley, C. and Raftery, A. (2002). Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc.* **97** 611–631. MR1951635
[16] Guoan, C., Tarek, G., Chiang-Ching, H., Dafydd, G., Kerby, A., Jeremy, M., Sharon, L., David, E., Thomas, J. and Mark, D. I. et al. (2002). Proteomic analysis of lung adenocarcinoma: identification of a highly expressed set of proteins in tumors. *Clin. Cancer Res.* **8** 2298–2305.
[17] Hubert, L. and Arabie, P. (1985). Comparing partitions. *J. of Classification* **2** 193–218.
[18] Jamshidi, N., Edwards, J. S., Fahland, T., Church, G. M. and Palsson, B. O. (2001). Dynamic simulation of the human red blood cell metabolic network. *Bioinformatics* **17** 286–287.
[19] Lee, B., Henderson, D. A. and Zhu, J. K. (2005). The arabidopsis cold-responsive transcriptome and its regulation by ice1. *Plant Cell.* **17**(11) 3155–3175, doi: 10.1105/tpc.105.035568. PMCID: 1276035.
[20] Li, W., Li, M., Zhang, W., Welti, R. and Wang, X. (2004). The plasma membrane-bound phospholipase D enhances freezing tolerance in arabidopsis thaliana. *Nat. Biotechnol.* **22** 427–433.
[21] Ma, P., Castillo-Davis, C. I., Zhong, W. and Liu, J. S. (2006). A data-driven clustering method for time course gene expression data. *Nucleic Acids Research* **34**(4) 1261–1269.
[22] Masry, E. and Tjostheim, D. (1995). Nonparametric estimation and identification of nonlinear arch time series: Strong convergence and asymptotic normality. *Econometric Theory* **11** 258–289. MR1341250
[23] Masry, E. and Tjostheim, D. (1997). Additive nonlinear ARX time series and projection estimates. *Econometric Theory* **13** 214–252. MR1449941
[24] Menges, M., Hennig, L., Gruissem, W. and Murray, J. (2002). Cell cycle-regulated gene expression in arabidopsis. *J. Biol. Chem.* **277** 41987–42002.
[25] Milligan, G. W. and Cooper, M. C. (1986). A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research* **21** 441–458.
[26] Munnik, T. (2001). Phosphatidic acid: An emerging plant lipid second messenger. *Trends Plant Sci.* **6** 227–233.
[27] Ramoni, M., Sebastiani, P. and Kohane, P. (2002). Cluster analysis of gene expression dynamics. *Proc. Nat. Acad. Sci.* USA **99** 9121–9126. MR1909705
[28] Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric regression.* Cambridge University Press, Cambridge, UK. MR1998720
[29] Seifert, K., Samson, R., deWaard, J., Houbraken, J., Levesque, C., Moncalvo, J., Louis-Seize, G. and Hebert, P. (2007). Prospects for fungus identification using co1 DNA barcodes, with penicillium as a test case. *Proc. Nat. Acad. Sci.* USA **104** 3901–6.
[30] Serban, N. and Wasserman, L. (2005). Cats: clustering after transformation and smoothing. *J. Amer. Statist. Assoc.* **100** 990–999. MR2201025
[31] Smith, M., Wood, D., Janzen, D., Hallwachs, W. and Hebert, P. (2007). DNA barcodes affirm that 16 species of apparently generalist tropical parasitoid flies (diptera, tachinidae) are not all generalists. *Proc. Nat. Acad. Sci.* USA **104** 4967–4972.
[32] Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D. and Futcher, B. (1998). Comprehensive identification of cell cycle regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell* **9** 3273–3297.
[33] Srivastava, M. (1967). Comparing distances between multivariate populations – the problem of minimum distance. *Ann. Math. Statist.* **38** 550–556. MR0208757
[34] Storey, J., Xiao, W., Leek, J., Tompkins, R. and Davis, R. (2005). Significance analysis of time course microarray experiments. *Proc. Natl Acad. Sci.* USA **102** 12837–12842.
[35] Stuetzle, W. (2003). Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. *J. of Classification* **20** 25–47. MR1983120
[36] Thomashow, M. F. (1999). Annu. Rev. Plant Physiol. *Plant Mol. Biol.* **50** 571–599.
[37] Toufighi, K., Brady, S. M., Austin, R., Ly, E. and Provart, N. J. (2005). The botany array resource: e-northerns, expression angling, and promoter analyses. *The Plant Journal* **43**(43) 153–163, doi: 10.1111/j.1365-313X.2005.02437.x.
[38] Umbach, A., Fiorani, F. and Siedow, J. (2005). Characterization of transformed arabidopsis with altered alternative oxidase levels and analysis of effects on reactive oxygen species in tissue. *Plant. Physiol.* **139**(4) 1806–20. PMID: 16299171.
[39] Valdivia, R. H. (1999). Regulatory network analysis. *Trends Microbiol* **7** 398–9.
[40] Van Buskirk, H. and Thomashow, M. F. (2006). Arabidopsis transcription factors regulating cold acclimation. *Physiologia Plantarum* **126** 72–80.
[41] Wang, H. (2004). Testing in multi-factor heteroscedastic anova and repeated measures designs with large number of levels. Ph.D. Dissertation, The Pennsylvania State University.
[42] Wang, H. and Akritas, M. G. (2004). Rank tests for anova with large number of factor levels. *J. of Nonparametric Statistics* **16**(3–4) 563–589. MR2073042
[43] Welti, R., Li, W., Li, M., Sang, Y., Biesiada, H., Zhou, H., Rajashekar, C., Williams, T. and Wang, X. (2002). Profiling membrane lipids in plant stress responses: Role of phospholipase dalpha in freezing-induced lipid changes in arabidopsis. *J. Biol. Chem.* **277** 31994–32002.

[44] Whitfield, M. L., Sherlock, G., Saldanha, A., Murray, J., Ball, C., Alexander, K., Matese, J., Perou, C., Hurt, M., Brown, P. and Botstein, D. (2002). Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Molecular Biology of the Cell* **13**(1595) 1977–2000.
[45] Wu, C. and Chiang, C. (1998). Kernel smoothing on varying coefficient models with longitudinal dependent variable. Unpublished.

Haiyan Wang
Department of Statistics
101 Dickens Hall
Kansas State University
Manhattan, KS 66506
E-mail address: hwang@ksu.edu

James Neill
Department of Statistics
101 Dickens Hall
Kansas State University
Manhattan, KS 66506
E-mail address: jwneill@ksu.edu

Forrest Miller
Department of Mathematics
208 Cardwell Hall
Kansas State University
Manhattan, KS 66506
E-mail address: frm@math.ksu.edu