CATS:
Clustering After Transformation and Smoothing

Nicoleta Serban and Larry Wasserman[1]
Department of Statistics
Carnegie Mellon University

CATS – Clustering After Transformation and Smoothing – is a technique for nonparametrically estimating and clustering a large number of curves. Our motivating example is a genetic microarray experiment but the method is very general. The method includes: transformation and smoothing multiple curves, multiple nonparametric testing for trends, clustering curves with similar shape, and nonparametrically inferring the misclustering rate.

Key words and phrases: Multiple testing, False discovery rate, clustering,misclustering, smoothing, genetic microarrays.

1 Introduction

CATS – Clustering After Transformation and Smoothing – is a technique for nonparametrically estimating and clustering a large number of curves (or profiles). The basic idea is to first remove the curves which are nearly flat, smooth the remaining curves, and then cluster the smoothed curves. A novel feature of our method is that we estimate the error due to the fact that we are clustering the estimated rather than the true curves. We obtain an asymptotic confidence bound for the clustering estimation error based on estimated confidence balls of the non-constant curves. The method we use for confidence ball estimation was introduced by Beran and Dümbgen (1998).

CATS is quite general but, for clarity, we will discuss the method in the context of microarray experiments. This problem is challenging because of the small number of design points but large number of expression profiles, and the small signal-to-noise ratio. Indeed, our motivating example is a genetic microarray experiment conducted at the University of Pittsburgh.

[1]Research supported by NIH Grant R01-CA54852-07, NIH grant number MH57881, NSF Grant DMS-98-03433 and NSF Grant DMS-0104016. The authors are grateful to David Peters and Rob O'Doherty for allowing them to use the data from the fat cell experiment. The authors also thank Dan Handley, Clark Glymour, Peter Spirtes, Richard Scheines, Greg Cooper and the other members of the CMU-University of Pittsburgh Gene group for their invaluable input.

This experiment produced time series of gene expression levels for 5355 genes over 15 time points.

There is now a substantial literature on genetic microarrays on various topics such as clustering (Eisen et al, 1998; Hastie et al, 2000; Bar-Joseph, Gerber, Gifford and Jaakkola, 2002; Wakefield, Zhou, Self, 2002) and multiple testing (Dudoit et al, 2000; Efron, Storey and Tibshirani, 2001; Newton et al, 2001).

For related work on curve clustering in the context of microarray data see Bar-Joseph, Gerber, Gifford and Jaakkola(2002) and Wakefield, Zhou, Self(2002).

2 The Model

We consider data of the form,

$$Y_{ij} = f_i(t_{ij}) + \sigma_i \epsilon_{ij}, \quad i = 1, \ldots, N, \quad j = 1, \ldots, m. \tag{1}$$

where $\mathbb{E}(\epsilon_{ij}) = 0$. Thus, Y_{ij} is the j^{th} observation on the i^{th} curve. In the examples of interest, N and m are both large but N is typically much larger than m. In the microarray setting, Y_{ij} is the log gene expression of gene i at time t_j.

We assume that the curves f_i belong to a Sobolev space $\mathcal{F} \equiv \mathcal{F}_\beta(c)$ of unknown order β and unknown radius c. See the appendix for a formal definition of $\mathcal{F}$. Let $\psi_1, \psi_2, \ldots$ be an orthonormal basis for $\mathcal{F}$ and write

$$f_i(t) = \sum_{j=1}^{\infty} \theta_{ij} \psi_j(t) \tag{2}$$

where

$$\theta_{ij} = \int f_i(t) \psi_j(t)\, dt. \tag{3}$$

We estimate f_i by

$$\widehat{f}_i^J(t) = \sum_{j=1}^{J} \widehat{\theta}_{ij} \psi_j(t) \tag{4}$$

where the estimates $\widehat{\theta}_{ij}$ and the choice of smoothing parameter J are described below. We call a curve f_i *null* or *inactive* if f_i is constant as a

function of t. Otherwise, f_i is *non-null* or *active*. Let $\mathcal{A}$ denote the set of active curves.

Let $\theta_i = (\theta_{i1}, \theta_{i2}, \ldots,)$ be the vector of coefficients for curve f_i. We will view the (θ_i, σ_i)'s as random draw from some distribution $\mathbb{P}$. We assume that $\mathbb{P}$ has compact support. In a slight abuse of notation, we also use $\mathbb{P}$ to denote the marginal law of the θ_i's.

3 Clusters of Curves

Since our goal is to cluster the curves, we need a measure of the efficacy of a set of clusters. Let $\mathcal{C} = \{f_1, \ldots, f_N\}$ denote a finite set of curves. A clustering algorithm may be viewed as a map

$$T : \mathcal{C} \times \mathcal{C} \to \{0, 1\}$$

where

$$T(f, g) = \begin{cases} 1 & \text{if } f \text{ and } g \text{ are in the same cluster} \\ 0 & \text{otherwise.} \end{cases}$$

The cluster map T induces a partition $\{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$ of $\mathcal{C}$ where two curves f and g are in the same partition element if and only if $T(f, g) = 1$. The numbering of the partition elements is arbitrary. Generally, one uses an algorithm that can produce k clusters for any given k. Thus, let us write T_k for the cluster map that yields k clusters. For example, T_k might be the output of the k-means clustering algorithm.

We address two different questions for the efficacy of the clusters. The first is: how good are the estimated clusters? The second is: how close is the clustering using the estimated curves $\widehat{\mathcal{C}} = \{\widehat{f}_1, \ldots, \widehat{f}_N\}$ to the clustering using the true curves $\mathcal{C} = \{f_1, \ldots, f_N\}$? The first concerns cluster quality; the second concerns estimation error. We will define two parameters, Ω and η associated with these questions.

3.1 Cluster Quality

Regarding cluster quality, many such criteria have been proposed. We shall use the following. Suppose that $\mathcal{C}_1, \ldots, \mathcal{C}_k$ are clusters, that is, they form a

partition of $\mathcal{C}$. Define the cluster *purity*

$$\Omega = \min_{1 \le j \le k} \min_{f,g \in \mathcal{C}_j} \rho(f,g)$$

where

$$\rho(f,g) = \frac{\int (f(x) - \overline{f})(g(x) - \overline{g}) dx}{\sqrt{\int (f(x) - \overline{f})^2 dx \int (g(x) - \overline{g})^2 dx}}$$

and $\overline{f} = \int f(x)\, dx$.

Thus, $\rho(f,g)$ is the Pearson correlation between the curves f and g and Ω measures the worst pairwise correlation over all the clusters. Note that $-1 \le \Omega \le 1$ and $\Omega = 1$ if and only all the curves in each cluster are proportional to each other. Write $\Omega(k)$ to indicate the dependence on the number of clusters.

In the Fourier domain we can rewrite Ω as follows. Let $f = \sum_j a_j \psi_j$ and $g = \sum_j b_j \psi_j$. From $a = (a_0, a_1, \ldots)$ define a new vector $\widetilde{a} = (\widetilde{a}_1, \widetilde{a}_2, \ldots)$ obtained by discarding a_0 and normalizing:

$$\widetilde{a}_j = \frac{a_j}{\sqrt{\sum_{j=1}^{\infty} a_j^2}}, \quad j \ge 1. \tag{5}$$

Define $\widetilde{b} = (\widetilde{b}_1, \widetilde{b}_2, \ldots)$ similarly. Then,

$$\rho(f,g) = 1 - \frac{|| \widetilde{a} - \widetilde{b}||^2}{2}. \tag{6}$$

Hence, correlation clustering in function space is equivalent to Euclidean clustering in the Fourier domain, after the transformation $a \mapsto \widetilde{a}$.

Generally, $\Omega(k)$ will increase as k increases. We will examine Ω as a function of k. If we want to choose one value of k, we take the smallest k such that $\Omega \ge 1 - \epsilon$ for some user-specified ϵ. This gives the smallest number of clusters that guarantees that all curves within a cluster are $(1-\epsilon)$-similar.

3.2 Estimation Error

Regarding estimation error, we proceed as follows. Let $\mathcal{C} = \{f_1, \ldots, f_n\}$ denote the true curves and let $\widehat{\mathcal{C}} = \{\widehat{f}_1, \ldots, \widehat{f}_n\}$ denote the estimated curves.

Let T and $\widehat{T}$ denote the corresponding clustering maps. Various methods have been proposed to compare two clusterings. See, for example, Rand (1971) and Fowlkes and Mallows (1983), Meilă (2000).

We define the *misclustering* rate for K clusters $\eta(K)$ by

$$\eta(K) = \frac{1}{\binom{N}{2}} \sum_{r<s} I\Big(T_K(f_r, f_s) \neq \widehat{T}_K(\widehat{f}_r, \widehat{f}_s)\Big). \tag{7}$$

Thus, η is the fraction of all pairs which are either incorrectly put in the same cluster or are incorrectly put in separate clusters. We write $\eta(K)$ to indicate the dependence on the number of clusters K. The misclustering rate can be expressed as

$$\eta = 1 - \mathcal{R}(T, \widehat{T})$$

where $\mathcal{R}$ is the Rand index (Rand, 1971).

3.3 k-means clustering

Let us briefly review some facts about k-means clustering. Let $\theta_1, \ldots, \theta_N \sim \mathbb{P}$ where each θ_i is a vector in $\mathbb{R}^d$. The k-means algorithm searches for the k vectors $a = \{a_1, \ldots, a_k\}$ that minimize

$$\frac{1}{n} \sum_{i=1}^{n} \min_{1 \leq j \leq k} ||\theta_i - a_j||^2.$$

This is equivalent to minimizing

$$W(a, \mathbb{P}_N) = \int \min_{a \in \mathcal{A}} ||\theta - a||^2 d\mathbb{P}_N(\theta)$$

over all possible choices of sets $\mathcal{A}$ containing k or fewer points, where $\mathbb{P}_N$ is the empirical measure putting mass $1/N$ on each θ_i. The centers a determine a *tessellation* $\{\mathbb{A}_1, \ldots, \mathbb{A}_k\}$ where $\theta \in \mathbb{A}_j$ if θ is closer to a_j than any other center.

Pollard (1981) shows, under weak conditions, that the minimizer $a = (a_1, \ldots, a_k)$ converges almost surely to the population minimizer $\overline{a}$ of $W(a, \mathbb{P})$. Also, Pollard (1982) shows that

$$\sqrt{N}(a - \overline{a}) \rightsquigarrow N(0, S)$$

for some $kd \times kd$ non-singular matrix S.

4 CATS

Our strategy for analyzing data of this form involves a series of steps summarized below; see also Figure 1.

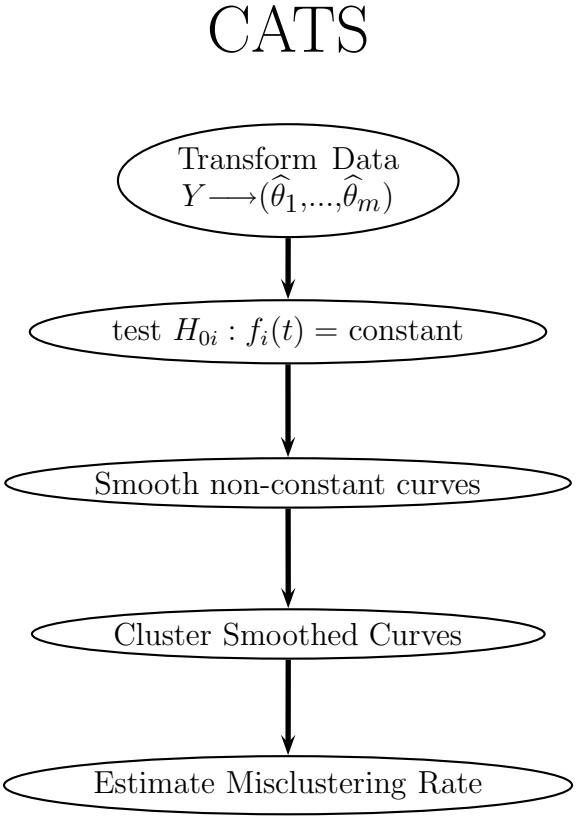

Figure 1: Data analysis strategy.

4.1 Transforming the Data

Without loss of generality, assume that all time points lie in $[0, 1]$. We transform the data into the Fourier domain as follows. Let

$$\phi_0(t) \equiv 1, \quad \text{and} \quad \phi_j(t) = \sqrt{2}\cos(j\pi t), j \geq 1$$

denote the cosine basis. Define the $m \times (k+1)$ matrix

$$\Phi = \begin{pmatrix} \phi_0(t_1) & \phi_1(t_1) & \cdots & \phi_k(t_1) \\ \phi_0(t_2) & \phi_1(t_2) & \cdots & \phi_k(t_2) \\ \vdots & \vdots & \vdots & \vdots \\ \phi_0(t_m) & \phi_1(t_m) & \cdots & \phi_k(t_m) \end{pmatrix}.$$

Now perform a Gram-Schmidt orthogonalization on the columns of Φ to make the columns orthogonal. Denote the new matrix by Ψ. Let $\widehat{\theta}_i = (\widehat{\theta}_{i1}, \ldots, \widehat{\theta}_{im})$

$$\widehat{\theta}_{ir} = \frac{1}{m} \sum_{j=1}^{m} \psi_{rj} Y_{ij}.$$

Under weak conditions, we have that $\widehat{\theta}_i \approx N(\theta_i, m^{-1}\Sigma_i)$ where Σ_i is diagonal with (j,j) element σ_j^2.

4.2 Smoothing

The function $\widehat{f}_i^J(t) = \sum_{j=0}^{J} \widehat{\theta}_{ij}\psi_{jt}$ is the smoothed version of the i^{th} profile. The parameter J controls the amount of smoothing. The optimal amount of smoothing will vary from curve to curve. Rather than trying to find an optimal amount of smoothing for each curve, instead we will find a single smoothing parameter that does reasonably well for all the curves. We consider two approaches.

Approach 1: Minimum Regret. Let

$$R_i(J) = \mathbb{E}\Big(\int (\widehat{f}_i^J(t) - f_i(t))^2 dt\Big)$$

denote the risk of the estimate of f_i. We will estimate the risk function for each curve. Then we choose J to minimize the total regret, the risk minus the minimum risk for each curve. Here are the steps.

Step 1. Variance Estimation. We estimate the variance σ_i^2 for gene i using the high component variance estimator:

$$\widehat{\sigma}^2 = \frac{1}{m-J} \sum_{i=J+1}^{m} Z_j^2$$

which is an asymptotic consistent estimator.

Step 2. Risk estimation. Define

$$\widehat{R}_i(J) = \frac{J\widehat{\sigma}_i^2}{m} + \sum_{j=J+1}^{m} \left(\widehat{\theta}_{ij}^2 - \frac{\widehat{\sigma}_i^2}{m} \right)_+ \tag{8}$$

which is a uniformly consistent estimate of the risk(see Beran and Dümbgen, 1998).

Step 3. Regret estimation. Define the regret

$$\widehat{r}_i(J) = \widehat{R}_i(J) - \min_{1 \le k \le m} \widehat{R}_i(J) \tag{9}$$

which measures how much risk is sacrificed for curve i if smoothing parameter J is used. Define the total regret

$$t(J) = \sum_{i=1}^{n} \widehat{r}_i(J). \tag{10}$$

Step 4. Smoothing parameter. Define

$$\widehat{J} = \text{argmin}_J t(J).$$

Approach 2: Multiscale Smoothing. Rather than searching for an optimal amount of smoothing, we can instead consider all values of J simultaneously and choose the one that leads to the most efficacious clustering. More precisely, we consider all the estimates $\widehat{f}^J$ or $1 \le J \le J_m$ where $J_m = o(m)$. We recommend the value $J_m = \sqrt{m}$. This leads to confidence balls for the curves of size $O(m^{-1/4})$ which is the smallest possible in a nonparametric sense (Li, 1989). Note that $\widehat{f}_i^J$ is actually estimating

$$f_i^J(t) = \sum_{j=0}^{J} \theta_{ij} \phi_j(t).$$

We can think of $f_i^J(t)$ as the smoothed version of the true curve. When J is small, we give up high resolution information about f_i but we can estimate

$f_i^J(t)$ accurately. We will probably not discover many clusters when J is small since there is not much shape information in f_i^J. As we increase J we can potentially discover more shape information leading to more refined clusters. However, the confidence sets for f_i^J get larger as J increases. In Section 4.6 we define the cluster uncertainty η which will be a function of the number of clusters k and the resolution level J. We will then produce confidence intervals for $\eta(k, J)$ and $\Omega(k, J)$ and plot these as functions of k for each J.

4.3 Screening Out Flat Curves

In this section, we explain how to test

$$H_{0i} : f_i(t) = c_i$$

or some constant c_i. If H_{0i} is true then $\sum_{j=1}^m \theta_i^2 = 0$. This suggests the test statistic

$$T_i = \sum_{j=1}^m \widehat{\theta}_{i,j}^2.$$

We reject the null hypothesis for large value of T_i. We estimate the p-value by permuting Y_i, Y_i^b for $b = 1, \ldots, B$ (in our simulation we usually take $B = 100,000$), and computing for each permutation the test statistic, $T(Y_i^b)$:

$$\widehat{P}_i = \frac{1}{B} \sum_{b=1}^B I(T(Y_i^b) \geq T(Y_i)).$$

To correct for the multiplicity problem we use the Benjamini-Hochberg (1995) method. Let $P_{(1)}, \ldots, P_{(n)}$ denote the ordered p-values. We reject H_{0i} if $P_i \leq T$ where $T = P_{(j)}$ and

$$j = \max \left\{ i : \ P_{(i)} \leq \frac{i\alpha}{n} \right\}. \tag{11}$$

This method controls the expected fraction of false discoveries to be less than or equal to α. See Benjamini and Hochberg (1995).

The FDR procedure assumes independent test statistics and the gene expression levels tend to be correlated. However, as shown in Storey and

Tibshirani (2002), the method works well even in the presence of dependence in clusters. Finally, we set $\widehat{\mathcal{A}} = \{i : \widehat{P}_i \leq T\}$. In our analysis, the significance level is $\alpha = .05$.

4.4 Confidence Ball for f_i

We use the method in Beran and Dümbgen (1998) for constructing a confidence ball $\mathbb{B}_i$ for f_i. Fix $\alpha > 0$ and let

$$\mathbb{B}_i = \Big\{(\theta_{i1}, \ldots, \theta_{im}) : \sum_{j=1}^{m} (\theta_{ij} - \widehat{\theta}_{ij})^2 \leq s_i^2\Big\} \tag{12}$$

where

$$s_i^2 = \frac{z_{\frac{\alpha}{N}} \widehat{\tau}_i}{\sqrt{m}} + \widehat{R}_i,$$

z_α is the α quantile of the standard normal and $\widehat{\tau}_i$ is given in the Appendix. The corresponding confidence ball for f_i is $\{\sum_{j=1}^{m} \theta_{ij}\psi_j(x) : \ \theta \in \mathbb{B}_i\}$. For notational convenience, the confidence ball for f_i will also be denoted by $\mathbb{B}_i$.

Theorem 1 follows directly from the theorems of Beran and Dümbgen.

Theorem 1 *Let $\mathcal{F}_\beta(c)$ denote a Sobolev space of order β and radius c. Then, for any $\beta > 1/2$ and any $c > 0$,*

$$\liminf_{N\to\infty} \sup_{f_1,\ldots,f_N \in \mathcal{F}_\beta(c)} \mathbb{P}\Big(f_i \in \mathbb{B}_i \text{ for all } i = 1, \ldots, N\Big) \geq 1 - \alpha.$$

Recall the mapping $\theta \mapsto \widetilde{\theta}$. The function $f(\theta) = \frac{\theta}{||\theta||}$. is continuous.

The set

$$\widetilde{\mathbb{B}}_i = \{\widetilde{\theta} : \theta_i \in \mathbb{B}_i\} \tag{13}$$

is a confidence set for $\widetilde{\theta}$ and is a compact set ($\widetilde{\mathbb{B}}_i = f(\mathbb{B}_i)$).

The confidence ball for the multiscale method is slightly different. Here we need a confidence ball for $(\theta_{i1}, \ldots, \theta_{iJ})$ which is somewhat simpler. Since, $\widehat{\theta}_{ij} \approx N(\theta_{ij}, \sigma_i^2/m)$, we have that

$$\sum_{j=1}^{J} (\theta_{ij} - \widehat{\theta}_{ij})^2 \approx \frac{\sigma_i^2}{m} \chi_J^2.$$

Hence,

$$\mathbb{B}_i^J = \left\{ (\theta_1, \ldots, \theta_J) : \sum_{j=1}^{J} (\theta_j - \widehat{\theta}_j)^2 \leq \frac{\widehat{\sigma}_i^2 \chi^2_{J,\alpha'}}{m} \right\} \tag{14}$$

is an approximate $1 - \alpha'$ confidence set for $(\theta_{i1}, \ldots, \theta_{iJ})$. We take $\alpha' = \alpha/(NJ_m)$ to ensure that the coverage is uniform over curves i and scales J.

We can use the estimated confidence balls $\mathbb{B}_i$ to screen out the remaining false positives after screening out the flat curves using hypothesis testing. If $(0, 0, \ldots) \in \mathbb{B}_i$ then f_i is a false positive.

4.5 Clustering

We want to identify curves with similar shape. In the microarray setting for example, genes with similar expression profiles are co-expressed gene. Co-expressed genes are likely to be co-regulated and hence co-expression can suggest functional pathways and interactions between genes.

For $r, s \in \widehat{\mathcal{A}}$ define

$$d(r, s) = \sum_{i=1}^{J} (\widetilde{\theta}_{rj} - \widetilde{\theta}_{sj})^2$$

where J is the smoothing parameter and $\widehat{\theta}_r$ and $\widehat{\theta}_s$ are the cosine transforms for the curves Y_r and, respectively, Y_s and $\widehat{\theta} \mapsto \widetilde{\theta}$ is the transform described in Section 3.1.

Then we apply the k-means clustering algorithm with the distance defined above. (Any other clustering method could be used.) We estimate the number of clusters using the gap method of Tibshirani, Walther, and Hastie (2000). Specifically, they propose testing under the null hypothesis whether the number of clusters is 1 versus the alternative hypothesis that the number of clusters is greater than 1. The null distribution, called the reference distribution, is the uniform distribution under which a clustering method would provide only one cluster. The test statistic is called the "gap" statistic. Tibshirani *et al.* propose taking the null distribution to be a uniform distribution on the hyper-rectangle over the range of the observed data.

We can also infer the number of clusters using the curves for $\Omega(j, K)$ and $\eta(J, K)$. We expect to see either a unimodal pattern with the mode at the true number of clusters or the curves will flat out by true cluster number.

4.6 Estimating the Misclustering Rate

In the following, we provide a confidence interval for the misclustering rate

$$\eta = \frac{1}{\binom{N}{2}} \sum_{r<s} I\Big(T(f_r, f_s) \neq \widehat{T}(\widehat{f}_r, \widehat{f}_s)\Big). \tag{15}$$

Theorem 2 *Suppose that $\mathbb{B}_i$ is a $1-(\alpha/N)$ confidence set for f_i. Let*

$$\overline{\eta} = \frac{1}{\binom{N}{2}} \sum_{r<s} \max_{f\in\mathbb{B}_r, g\in\mathbb{B}_s} I\Big(T(f, g) \neq T(\widehat{f}_r, \widehat{f}_s)\Big). \tag{16}$$

Then,

$$\mathbb{P}(\eta \in [0, \overline{\eta}]) \geq 1 - \alpha. \tag{17}$$

Computing (16) can be computationally demanding. A simplification occurs with k-means clustering. Recall that k-means clustering produces a set of cluster centers $a_1, \ldots, a_k$. This, in turn, produces the Voronoi tessellation $\{\mathbb{A}_1, \ldots, \mathbb{A}_k\}$ where $f \in \mathbb{A}_j$ if f is closer to a_j than any other cluster center. In this case, $T(f, g) = 1$ if and only if f and g belong to the same member of the tessellation.

Theorem 3 *Assume the conditions of the main theorem in Pollard (1982). Let $\{\mathbb{A}_1, \ldots, \mathbb{A}_k\}$ be the tessellation based on the true curves and let $\{\widehat{\mathbb{A}}_1, \ldots, \widehat{\mathbb{A}}_k\}$ be the tessellation from the estimated curves. Let $j(i)$ denote the index of the tessellation element containing $\widehat{f}_i$. Then*

$$\eta \leq \frac{1}{N} \sum_{i=1}^{N} I\Big(f_i \notin \widehat{\mathbb{A}}_{j(i)}\Big) \left(1 + \frac{1}{N-1} \sum_{i=1}^{N} \Big(1 - I\Big(f_i \notin \widehat{\mathbb{A}}_{j(i)}\Big)\Big)\right) + O_P\left(\frac{1}{\sqrt{m}}\right). \tag{18}$$

Outline of Proof. Let $\theta = (\theta_1, \ldots, \theta_N)$ denote the true coefficient vectors and let $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_N)$ denote the estimated coefficient vectors. Let

$a = (a_1, \ldots, a_k)$ denote the cluster centers based on θ and let $\widehat{a} = (\widehat{a}_1, \ldots, \widehat{a}_k)$ denote the cluster centers based on $\widehat{\theta}$. Let μ denote the population minimizer of $\int \min_{a \in \mathcal{A}} ||\theta - a||^2 d\mathbb{P}(\theta)$. Thus

$$a = \text{argmin}_u \int m_a(\theta) d\mathbb{P}_N(\theta)$$

where

$$m_a(\theta) = \min_{1 \leq j \leq k} ||\theta - a_j||^2.$$

Hence, a is an M-estimator. Using Pollard (1992) or Theorem 5.23 of van der Vaart (1998), we have that

$$\int m_a(\theta) d\mathbb{P}(\theta) = \int m_\mu(\theta) d\mathbb{P}(\theta) + \frac{1}{2}(a-\mu)^T V_\mu (a-\mu) + o(||a-\mu||^2)$$

where V_μ is the second derivative of the map $a \mapsto R(a) \equiv \int m_a(\theta) d\mathbb{P}(\theta)$.

Moreover,

$$a = \mu + \frac{1}{N} S \sum_{i=1}^{N} Y_i + o_P\left(\frac{1}{\sqrt{N}}\right)$$

where $S = -V^{-1}$ and $Y_i = \dot{m}_\mu(\theta_i)$. Recall that $\widehat{\theta}_i \approx N(\theta_i, m^{-1}\Sigma_i)$. Treating this approximation as exact, we can view $\widehat{\theta}_i$ as a sample from $\mathbb{P}_m = \mathbb{P} \oplus Q_m$ where $\mathbb{P}_m(\widehat{\theta} \in B) = \int \mathbb{Q}(\widehat{\theta} \in B|\theta) d\mathbb{P}(\theta)$ and $\mathbb{Q}_m$ denotes the $N(\theta_i, m^{-1}\Sigma_i)$. With $R_m(a) = \int m_a(\widehat{\theta}) d\mathbb{P}(\widehat{\theta})$, it is easy to see that $\mu_m \equiv \text{argmin} R_m(a) = \mu + O(m^{-1/2})$. Also,

$$\widehat{a} = \mu_m + \frac{S_m}{N} \sum_{i=1}^{N} \widehat{Y}_i + o_P\left(\frac{1}{\sqrt{N}}\right)$$

where $S_m = S + O(m^{-1/2})$ and $\widehat{Y}_i = Y_i + O_P(m^{-1/2})$. It follows that

$$||a - \widehat{a}|| = O_P\left(\frac{1}{\sqrt{m}}\right).$$

Recall that $\mathbb{P}$ has support on a compact set. Restricted to this compact set, $d_H(\mathbb{A}_j \Delta \widehat{\mathbb{A}}_j)$ is a continuous function of $\mu - \widehat{a}$, where Δ denotes symmetric set difference and d_H is the Hausdorff distance. It then follows that

$$d_H(\mathbb{A}_j \Delta \widehat{\mathbb{A}}_j) = O_P(1/\sqrt{m}). \tag{19}$$

Let $\mathbb{A}(f)$ denote the member of the tessellation that contains f and similarly for $\widehat{\mathbb{A}}(\widehat{f})$. Now,

$$T(f,g) = \sum_{j=1}^{k} I(f \in \mathbb{A}_j) I(g \in \mathbb{A}_j) \quad \text{and} \quad \widehat{T}(\widehat{f},\widehat{g}) = \sum_{j=1}^{k} I(\widehat{f} \in \widehat{\mathbb{A}}_j) I(\widehat{g} \in \widehat{\mathbb{A}}_j).$$

Let us define

$$T(\widehat{f},\widehat{g}) \equiv \sum_{j=1}^{k} I(\widehat{f} \in \mathbb{A}_j) I(\widehat{g} \in \mathbb{A}_j).$$

It then follows from (19) that

$$\widehat{T}(\widehat{f},\widehat{g}) = T(\widehat{f},\widehat{g}) + O_P\left(\frac{1}{\sqrt{m}}\right).$$

Thus,

$$I\Big(T(f_i,f_j) \neq \widehat{T}(\widehat{f}_i,\widehat{g}_i)\Big) = S(i,j) + O_P\left(\frac{1}{\sqrt{m}}\right)$$

where $S(i,j) = I(T(f_i,f_j) \neq T(\widehat{f}_i,\widehat{g}_i))$. Let $\mathcal{C} = \{i : \ \mathbb{B}_i \not\subset \widehat{\mathbb{A}}(i)\}$ (with the cardinal $|\mathcal{C}|$). Note that

$$S(i,j) \leq \max(I(i \in \mathcal{C}), I(j \in \mathcal{C})).$$

Thus,

$$\begin{aligned}
\eta &= \frac{1}{\binom{N}{2}} \sum_{r<s} I\Big(T(f_r,f_s) \neq \widehat{T}(\widehat{f}_r,\widehat{f}_s)\Big) \\
&= \frac{1}{\binom{N}{2}} \sum_{r<s} S(i,j) + O_P\left(\frac{1}{\sqrt{m}}\right) \\
&= \frac{1}{N(N-1)} \sum_{r=1}^{N} \sum_{s \neq r} S(i,j) + O_P\left(\frac{1}{\sqrt{m}}\right) \\
&\leq \frac{1}{N(N-1)} \sum_{r \in \mathcal{C}} (N-1) + \frac{1}{N(N-1)} \sum_{r \in \mathcal{C}^c} \sum_{s \neq r} I(s \in \mathcal{C}) + O_P\left(\frac{1}{\sqrt{m}}\right) \\
&= \frac{|\mathcal{C}|(N-1)}{N(N-1)} + \frac{1}{N(N-1)}(N - |\mathcal{C}|)|\mathcal{C}| + O_P\left(\frac{1}{\sqrt{m}}\right) \\
&= \frac{|\mathcal{C}|}{N}\left(1 + \frac{N-|\mathcal{C}|}{N-1}\right) O_P\left(\frac{1}{\sqrt{m}}\right). \quad \blacksquare
\end{aligned}$$

Theorem **4** *Let*

$$\overline{\eta} = \sum_{i=1}^{N} I\Big(\mathbb{B}_i \cap \widehat{\mathbb{A}}_r \neq \emptyset \ \text{ for some } r \neq j(i)\Big) \tag{20}$$

$$\widehat{\eta} = \frac{\bar{\eta}}{N}\left(1 + \frac{N - \bar{\eta}}{N-1}\right)$$

Then, $[0, \widehat{\eta}]$ *is an approximate* $1-\alpha$ *confidence interval for* η*.*

In the next section we describe an algorithm for computing $\overline{\eta}$.

4.7 Algorithm for Computing $\overline{\eta}$

To compute $\overline{\eta}$ we need to compute (20) which requires evaluating

$$\delta_i \equiv I\Big(\mathbb{B}_i \cap \widehat{\mathbb{A}}_r \neq \emptyset \ \text{ for some } r \neq j(i)\Big)$$

where $\mathbb{B}_i$ is the confidence ball and $j(i)$ is the cluster index for the profile f_i.

The following algorithm can be used.

1. For $r \neq j(i)$ do:
   (a) Let $\mathcal{H}$ be the hyperplane that bisects the line joining $\mathbb{C}_{j(i)}$, the center of the cluster $j(i)$, and $\mathbb{C}_r$, the center of an arbitrary cluster $r \neq j(i)$. Thus we construct the hyperplane, $\mathcal{H}$, which bisects the segment joining $\mathbb{C}_{j(i)}$ and a different cluster center $\mathbb{C}_r$.
   (b) Let $\mathbb{B}_i$ be the confidence set of the profile f_i. Define $\widetilde{c}_i^{\min} = f(c_i^{\min})$ $(f(\theta) = \theta/||\theta||)$the closest point in $\widetilde{\mathbb{B}}_i$ to the hyperplane of bisection.
   (c) If $d(C_{j(i)}, \widetilde{c}_i^{\min})) > d(C_r, \widetilde{c}_i^{\min}))$ set $\delta_{ir} = 1$. Otherwise set $\delta_{ir} = 0$.
2. Set $\delta_i = \max_{r \neq j(i)} \delta_{ir}$.

We present an analytic solution to $\widetilde{c}_i^{\min}$ in the appendix. Having the coordinates of the closest point in $\mathbb{B}_i$ to the bisection hyperplane it is not difficult to compute the distances $d(C_{j(i)}, \widetilde{c}_i^{\min}))$, $d(C_r, \widetilde{c}_i^{\min}))$ and thus δ_i.

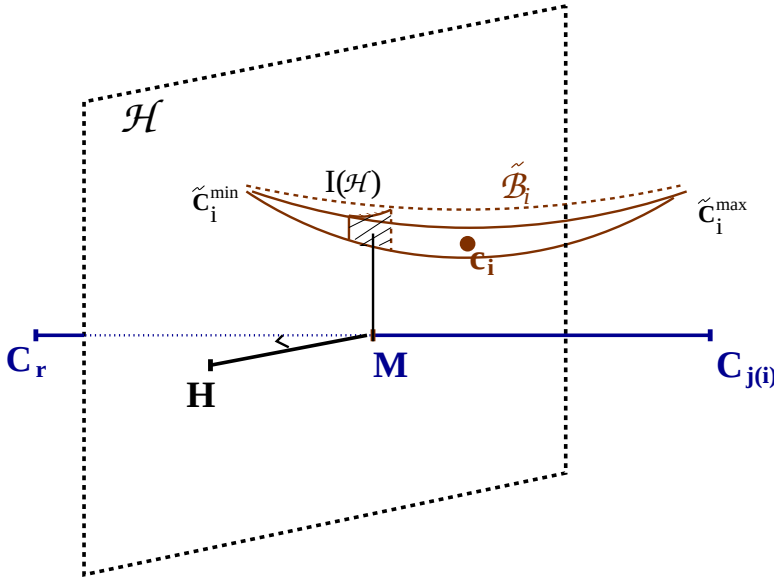

Figure 2: Minimum Distance. This figure is a 3D version of our algorithm. C_r and $C_{j(i)}$ are cluster centers and c_i is a confidence set center corresponding to curve i. Note that the confidence set is not a ball. In this case, the curve is misclustered or $\delta_i = 1$

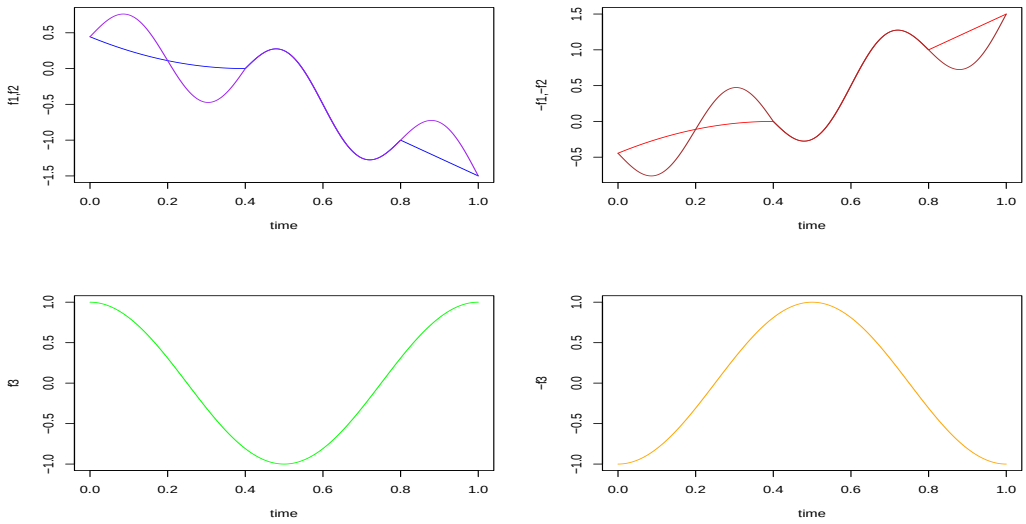

Figure 3: The left panel consists of the three curves: f_1 and f_2 (the same pattern, in the same figure), and f_3. The plots on the right are the curves with negative means $(-1)f_1, (-1)f_2, (-1)f_3$.

5 Example: Synthetic Data

We generate synthetic data according to the regression model:

$$Y_j = f(t_j) + \sigma\epsilon_j$$

with $j = 1, \ldots, m$. We want to use these synthetic data to evaluate the methods introduced in this paper. Subsequently, we need f to take various shapes. The regression functions for f are:

$$\mathbf{f_1}(\mathbf{t}) = I_{\{t\in S\}}(t)\left(\left(\tfrac{2-5t}{2}\right) \wedge \left(\left(\tfrac{5t-2}{3}\right)^2 + \sin\tfrac{5\pi t}{2}\right)\right) + I_{\{t\in S^c\}}(t)\left(\left(\tfrac{2-5t}{2}\right) \wedge \left(\tfrac{5t-2}{3}\right)^2\right)$$

$$\mathbf{f_2}(\mathbf{t}) = \left(\tfrac{2-5t}{2}\right) \wedge \left(\left(\tfrac{5t-2}{3}\right)^2 + \sin\tfrac{5\pi t}{2}\right)$$

$$\mathbf{f_3}(\mathbf{t}) = \cos(2\pi t) \quad \text{where} \quad S = \left(\tfrac{2}{5}, \tfrac{4}{5}\right)$$

The first two curves have similar pattern with small perturbation at the

smoothing parameter	J = 3	J = 4	J = 5	J = 6	J=25
# of true positives	578	589	589	586	1
# of false positives	8	20	17	14	0
# of false positives with $(0, 0, \ldots) \in \mathbb{B}$	3	4	4	4	0

Table 1: Number of rejection according to FDR for different values of smooth parameter - synthetic data. The number of non-constant curves is 600 with different noise levels. The last line is the number of false positives that can be identified using their confidence sets: if the confidence set contains $(0, \ldots, 0)$ then it is a constant curve.

early and late time points (see 3, left upper plot). We also include in the analysis the reversed (negated) curves (see figure 3, right plots).

The synthetic data contain 150 curves for each of the 4 patterns on different scale. For simplicity, we take the noise ϵ_{ij} being normally distributed. Among the 2000 curves in the synthetic data, 1400 are constant curves ($Y_{ij} = N_m(0, \sigma_i)$). The 600 non-constant curves are defined over $m = 25$ design points ($Y_{ij} = \mu_i f(t_j) + N_m(\mu_i, \sigma_i I_m)$ with f taking one of the three shapes f_1, f_2 or f_3). The standard deviation is between $\sigma \in (0.2, 0.5)$ We believe that these synthetic data are fairly representative of amount of structure and complexity of a dataset provided by a microarray experiment. However, the generated curves are independent which is not the case of the expression profiles from microarray experiments. The reason we don't consider "loose dependence" for the synthetic curves is that the testing procedure complemented with FDR correction for multiple inference can be equally applied under cluster dependence and independence as long as the number of curves is large.

Testing for non-constant curves. A first step is to apply the smoothing algorithm presented in Section 4.2. Applying the minimum regret approach to the synthetic data, the smoothing parameter is estimated to be $\widehat{J} = 4$.

The error rate for the multiple hypothesis testing is controlled at the level of significance $\alpha = .05$. The number of rejections differs with J, the smoothing parameter (see Table 5). The larger number of true positives are

for $J = 4$ and $J = 5$.

It is worth mentioning that we've implemented several forms of *runs test*. The nonparametric test outperforms all of them. Usually, the runs tests recognize nonrandom patterns with a large number of design points (≥ 35).

Clustering. We generate data from 4 different curve patterns (see Figure 3). The gap method applied to the transformed data of the non-constant curves (see Section 4.5) estimates the number of clusters to be $\widehat{K} = 4$. Thus gap method in the context of our algorithm doesn't account for the small perturbation when the number of design points is as small as 25.

Misclustering. The misclustering bound is asymptotic. To evaluate the validity of this result, we consider the same curves in figure 3 over $m = 100$ design points rather than $m = 25$. A synthetic dataset with 600 non-constant and 10 flat curves is generated according to the 4 patterns in figure 3. The estimated balls of the 610 profiles are in figure 4. The radius of these ball is obtained using χ^2 approximation. The ones centered around 0 are the 10 flat curves. Thus we screen out those curves f_i with $(0, 0, \ldots) \in \mathbb{B}_i$.

We compute the estimated misclustering rate for smoothing parameter $J = 3, 4, 5, 6$ and for the number of clusters $K = 2, \ldots, 8$ (see figure 5). The true misclustering rate is 0 for $K = 2, 3, 4, 5, 6$ clusters. The number of clusters inferred from the misclustering curves is $\widehat{K} = 4$.

6 Example: Adipose Cell Experiment

In this section we apply CATS to data from an experiment on adipose cells. We examined data from spotted cDNA microarrays (Research Genetics, Carlsbad, California) experiment. The experiment was completed in February, 2002 at the University of Pittsburgh (Peters *et al.*). The spotted cDNA microarray experiment consists of a time-sequenced sampling of differential expression in mRNA from (3T3L1 cultured) adipose cells originally obtained from mice. These cells were treated with a drug, *troglitazone*, which is a member of a family of drugs known as thiazolidendiones (TZD's). In our experiment, the drug treatment of the cells lasted for different periods of time ranging from 0 hours to 24 hours.

The data consist of 15 measurements of mouse adipose cells at different

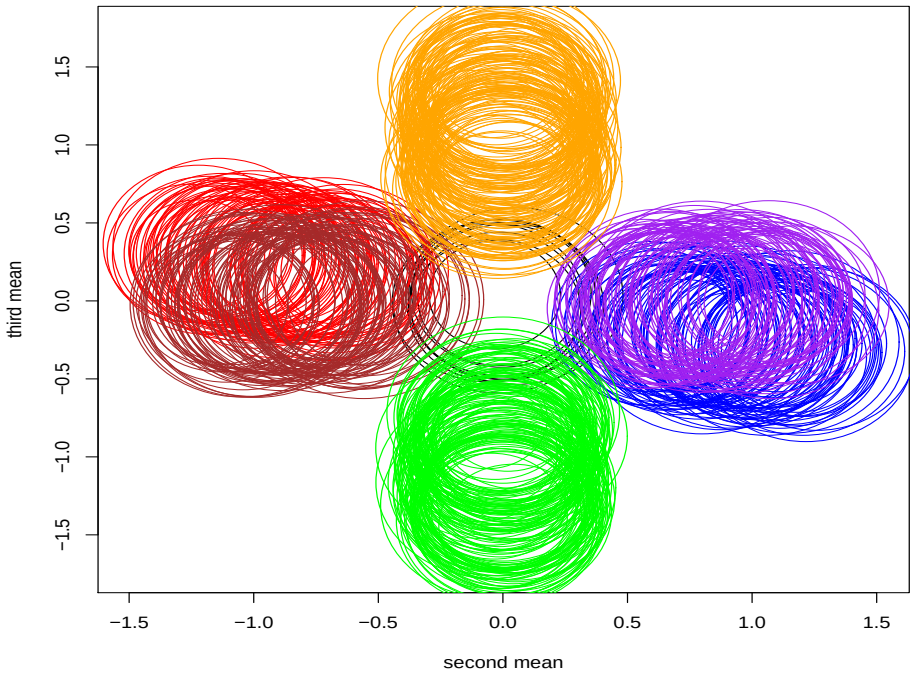

Figure 4: 2D confidence balls for 610 synthetic curves. Each circle represents the confidence ball in 2D of one curve with radius computed using χ^2 approximation. Each color is associated with a different curve pattern/cluster. The circles (black color) in the middle correspond to constant curves. Note that the misclustering rate is based on the confidence sets of $\widetilde{\theta}$ as defined in equations (5) and (13) and not on these 2D balls.

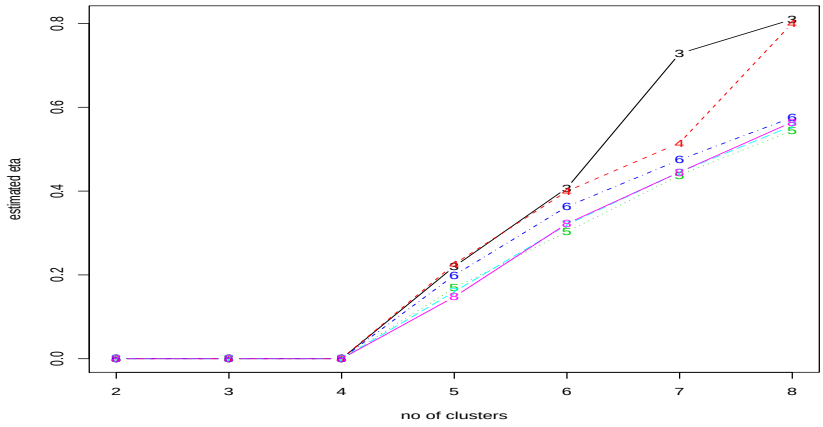

Figure 5: Upper bound for η - synthetic data. Each curve represents the asymptotic upper bound for η for a given smooth parameter J over the number of clusters. For example, curve 6 in the figure consists of the estimated upper bound for $J = 6$.

periods of time. For each measurement, target cDNA was obtained by mRNA extraction and reverse transcription (into complementary DNA). Then the cDNA targets were hybridized to microarrays. Each of the 15 hybridizations produced images, which were processed using the software package Pathways 3. The main quantity of interest reported by the image analysis methods is the intensity for each probe on each array.

After image processing, removing sources of experimental bias and variance, the gene expression data can be summarized by a matrix of intensities with 15 columns (corresponding to the number of arrays) and 3824 rows (corresponding to the number of probes). Each of the 3824 rows represents the expression profile over time of a DNA sequence.

The arrays need to be normalized to account for systematic differences between arrays. We use a global linear normalization that forces the log intensities to have median equal to zero at each array, making the median of the experiment array the same as that of the baseline array. This appears to perform well because we expect only a relative small proportion of the genes to vary significantly in expression between mRNA samples [24].

smoothing parameter	J = 3	J = 4	J=5	J=15
# of rejections	291	238	264	0

Table 2: Number of rejection according to FDR for different values of smooth parameter - microarray data.

Testing for active genes. Similarly to the analysis of synthetic data, we first screen out the inactive genes. Using minimum regret approach, the smoothing parameter of these data is $\widehat{J} = 3$.

We find 291 active genes among the 3824 DNA sequences for the estimated smoothing parameter $\widehat{J} = 3$.

Clustering. Next, we want to identify the cluster membership of the active genes in the microarray data. We estimate two global clusters for smoothing parameter $J = 3, 4, 5$. The separation between the two clusters when $J = 3$ is evident in Figure 6. The panel displays the second cosine transform component $(\widehat{\theta_2})$ vs. the third component $(\widehat{\theta_3})$ for the 291 significant expression profiles when the number of clusters is 2.

The average curves over time of the significant sequences in each cluster are shown in Figure 7. The sequences in the first cluster have a depression around 8 - 10 hours and the sequences in the second cluster have a rise around 6 - 10 hours. However, the second cluster selects the gene expression profiles with very low activity compared to the ones in the second cluster.

Misclustering. For these observed data we compute the approximate upper bound of the misclustering rate η for $J = 3, 4, 5$ as shown in figure 8. We infer from the misclustering rate curve for $J = 3$ that the number of clusters is $K = 2$. In fact there is a large jump for the estimated misclustering bound from $K = 2$ to $K = 3$.

7 Discussion

The methodologies introduced in this paper provide a means for screening out flat curves, a means for clustering non-constant curves and a means for quantifying the estimation error in the estimated clusters. These methods enable us to make inference on a large number of non-constant curves simul-

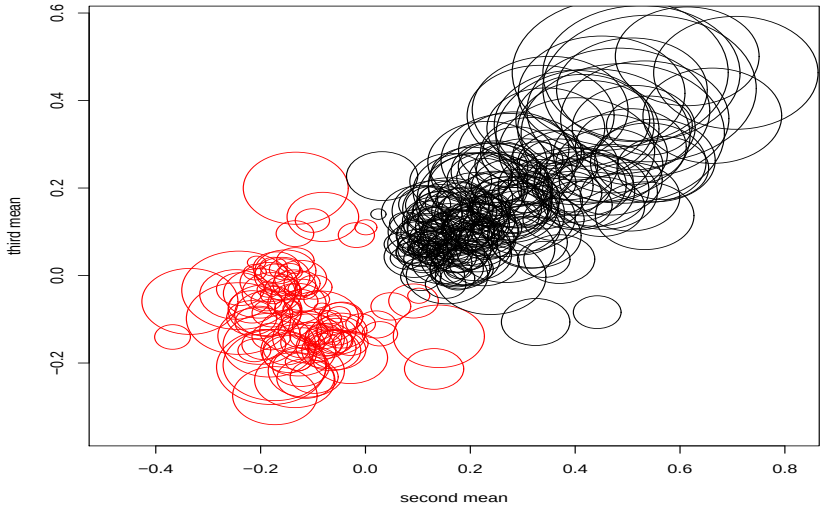

Figure 6: 2D balls for significant genes. Each circle represents the confidence ball of one gene with radius computed using χ^2 approximation. The red circles are in cluster 1 and the black circles are in cluster 2. Their average curves are in the next figure.

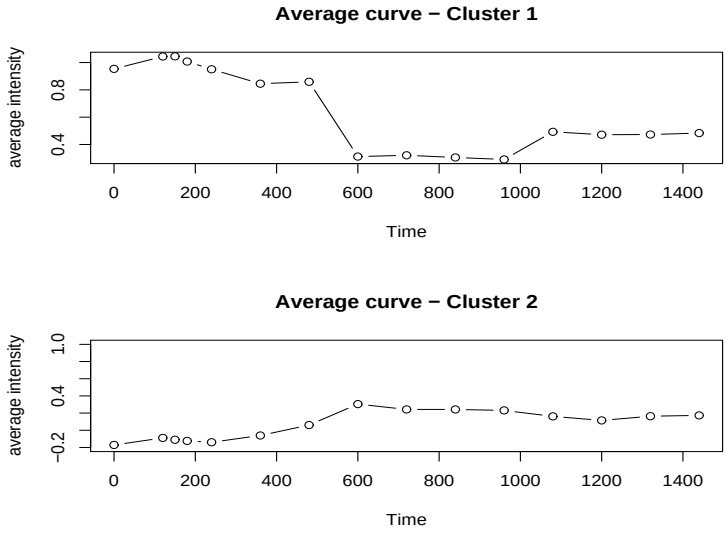

Figure 7: Average curves by cluster. In each plot, the average over all genes in one cluster is plotted over time.

taneously.

We propose a nonparametric test for filtering out the constant profiles. The test is applied after transforming the profile data using cosine basis. With a small smoothing parameter, the test proves to be powerful in identifying constant curves. This step is important, because a large number of flat curves with different degrees of random noise affects the clustering. In this way, we eliminate constant curves which could lead to misinterpretation in the reported results. We've tried a few other tests on different experimental data, and the nonparametric test presented in this paper proved to be the most powerful at a small number of design points. Significance over a small number of arrays is an important issue in the microarray experiments where their cost is highly expensive.

We also propose a clustering algorithm of the smoothed non-constant curves which is an alternative to clustering by correlation. The correlation coefficient of two curves can be expressed as the Euclidean distance of the normalized cosine transforms in the Fourier space. We use k-means in this article but any other clustering algorithm can be used after data transforma-

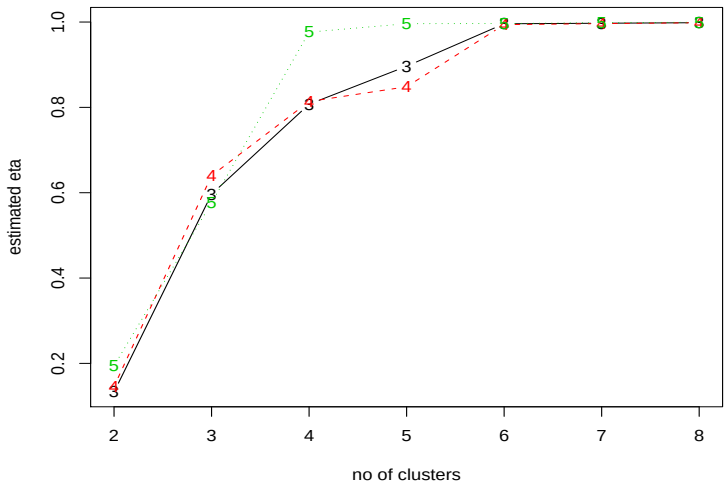

Figure 8: Upper bound for η - microarray data. Each curve represents the asymptotic upper bound for η for a given smooth parameter J over the number of clusters. For example, curve 3 in the figure consists of the estimated upper bound for $J = 3$.

tion and smoothing.

The last step is the inference on the cluster estimation. We consider a misclustering rate based on the fraction of all pairs put incorrectly in the same cluster or put incorrectly in different cluster. This misclustering rate was first introduced by Rand (1971). We find an asymptotic upper bound of this misclustering rate which can be approximated using 95% confidence balls estimated using the method in Beran & Dümbgen (1998). To our knowledge, this approach to cluster estimation has not been considered previously.

To evaluate the validity of our clustering method, we generated a synthetic dataset. We correctly estimate the true number of clusters and the true cluster membership. We use the approximate upper bound to make inference on the cluster error estimation as well as inference on the smoothing parameter and number of clusters.

For the gene expression data, we identified two clusters of gene expression profiles which showed a change in expression over treatment times. The estimated misclustering bound is about 0.1 which show that the our clustering method performs quite well on these data.

Appendix

Sobolev Ellipsoid

Let $\psi_1, \psi_2, \ldots$ be an orthonormal basis for L_2. The Sobolev ellipsoid $\mathcal{F}_\beta(c)$ or order β and radius c is

$$\left\{ f(x) = \sum_{j=1}^{\infty} \beta_j \psi_j(x) : \sum_{j=1}^{\infty} \beta_j^2 j^{2\beta} \leq c^2 \right\}.$$

τ^2 estimation

We estimate the confidence balls for the expression profiles in $\widehat{\mathcal{A}}$ under the normal means problem:

$$Z_j = \theta_j + \sigma\epsilon_j$$

with $j = 1, \ldots, m$ and $\epsilon_j \sim N(0, 1)$. Now let σ^2 be the variance of Z.

We estimate σ^2 with the high component variance estimator:

$$\widehat{\sigma}^2 = \frac{1}{m-J}\sum_{i=J+1}^{m} Z_j^2$$

which is an asymptotic consistent estimator of σ.

Define

$$\widehat{d} = \sqrt{m}(L(\widehat{\beta}Z, \theta) - \widehat{R}(\beta)) \tag{21}$$

where L is the mean square loss and $\widehat{R}(\beta)$ is the Stein's unbiased risk estimator (SURE). According to Beran & Dümbgen (1998) the asymptotic distribution for d is $N(0, \tau^2)$.

Substituting L and $\widehat{R}$ in (21) obtain

$$\widehat{d} = \sqrt{m}\left(\sum_{j=1}^{m}(\widehat{\beta}_j Z_j - \theta_j)^2 - \left(\sum_{j=1}^{m}\widehat{\sigma}^2\beta_j^2 + \sum_{j=1}^{m}\left(Z_j^2 - \widehat{\sigma}^2\right)(1-\beta_j)^2\right)\right).$$

We find that

$$\mathbb{V}(\widehat{d}) = \sum_{j=1}^{m}(2f_j-1)^2\left(1+\frac{1-2c_j}{m-J}\right)(4\theta_j^2\sigma^2+2\sigma^4)+4\sigma^2\sum_{j=1}^{m}f_j\left((1-f_j)+\frac{2(2f_j-1)c_j}{m-J}\right)\theta_j^2$$

where c_j is 1 for $j \geq J+1$ and 0 otherwise. Replace $\sigma^2 \leftarrow \widehat{\sigma}^2$ (where $\widehat{\sigma}^2$ is the high component variance of Z) and $\theta_j^2 \leftarrow (Z_j^2 - \widehat{\sigma}^2)_+$ to obtain the estimate $\widehat{\tau}^2$ of $\mathbb{V}(\widehat{d})$. When $f_j = 1$ for $j = 1, \ldots, k$ and 0 otherwise, the estimated variance becomes:

$$\widehat{tau}^2 = \sum_{j=1}^{m}\left(1+\frac{1-2c_j}{m-J}\right)(4(Z_j^2-\widehat{\sigma}^2)_+\widehat{\sigma}^2+2\widehat{\sigma}^4)+4\widehat{\sigma}^2\sum_{j=1}^{k}\frac{2c_j}{m-J}(Z_j^2-\widehat{\sigma}^2)_+.$$

The estimated variance of $\widehat{d}$ is different from the one presented in [5, 4] because it takes into account the dependence between Z and $\widehat{\sigma}$.

Analytic solution for $\overline{\eta}$

Let w and w_0 the weights in the equation of the hyperplane which bisects the segment joining $\mathbb{C}_r$ and $\mathbb{C}_{j(i)}$:

$$\mathcal{H} : h(x) = w^t x + w_0 = 0.$$

The bisection hyperplane is defined as follows. We know that the median point of the joining segment $\mathbb{C}_r\mathbb{C}_{j(i)}$ is in the hyperplane $\mathcal{H}$ and for any point H in the hyperplane, the line joining H and the median, M, is perpendicular on the line uniquely determined by $\mathbb{C}_r$ and $\mathbb{C}_{j(i)}$.

Denote the coordinates of an arbitrary point in $\mathcal{H}$: $h = (h_1, \ldots, h_k)$, the coordinates of $\mathbb{C}_r$, the arbitrary cluster center, $p = (p_1, \ldots, p_k)$, and the coordinates of $\mathbb{C}_{j(i)}$: $s = (s_1, \ldots, s_k)$.

We write that the joining segment HM is perpendicular on the line determined by $\mathbb{C}_{j(i)}$ and M:

$$\sum_{i=1}^{k} \left(h_i - \frac{p_i + s_i}{2} \right) \left(s_i - \frac{p_i + s_i}{2} \right) = 0.$$

It follows that

$$w_i = p_i - s_i \text{ for } i = 1, \ldots, k \text{ and } w_0 = \sum_{i=1}^{k} \frac{s_i^2 - p_i^2}{2}.$$

The confidence set is defined by:

$$\widetilde{\mathbb{B}}_i = \{\widetilde{\theta} : \widetilde{\theta} = f(\theta), \theta \in \mathbb{B}_i\}.$$

We check $\widetilde{\mathbb{B}}_i \cap \mathcal{H} \neq \emptyset$ by computing the minimum distance from the confidence set $\widetilde{\mathbb{B}}_i$ to the bisection hyperplane.

$$\min_{\theta_i \in \mathbb{B}_i} d(f(\theta_i), \mathcal{H}) = \min_{\theta_i \in \mathbb{B}_i} \left[\frac{< \theta_i, w >}{||\theta_i|| ||w||} + \frac{w_0}{||w||} \right].$$

Because the minimum will be on the envelope, we solve

$$\min \frac{< \theta_i, w >}{||\theta_i|| ||w||}, \text{ with } ||\theta_i - \widehat{\theta}_i|| - r_i = 0. \tag{22}$$

The equivalent geometry problem to the optimization problem (22) is the following. We want to find the maximum angle to the origin between a fixed point W in the space and points on the envelope of the hypersphere $\mathbb{B}_i$. See figure 9 for a 3D description. The problem reduces to finding the maximum

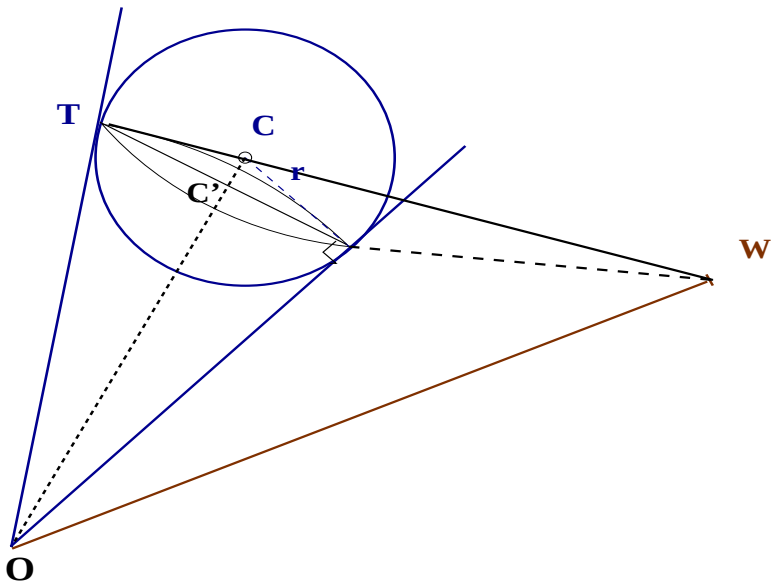

Figure 9: Maximum angle. We want to maximize the angle $\widehat{WOT}$ with T on the circle of center C and radius r. This angle gives the solution to the optimization problem we solve for $\bar{\eta}$ as defined in the text.

angle, $\widehat{TOW}$ where T falls on the envelope of a hypersphere. The problem (22) is easier to solve with the following constrains:

$$\min_{t_i} \frac{< t_i, w >}{||t_i||||w||} \quad \text{with } ||t_i - \widehat{\theta}_i|| = r_i, \ < t_i, (t_i - \widehat{\theta}_i) >= 0. \tag{23}$$

with the last equality due to the tangent in T from the origin. We rewrite again the minimization problem as:

$$\min_{t_i} < t_i, w > \quad \text{with } ||t_i||^2 = ||\widehat{\theta}_i||^2 - r_i^2, \ < t_i, \widehat{\theta}_i >= ||\widehat{\theta}_i||^2 - r_i^2. \tag{24}$$

We solve this optimization problem using Lagrange's theorem:

$$\begin{aligned} &\nabla f(t) = \mu \nabla g(t) + \lambda \nabla h(t) \\ &f(t) =< t, w >= \textstyle\sum_{j=1}^{J} t_j w_j \\ &g(t) = ||t||^2 - (||\widehat{\theta}_i||^2 - r_i^2) = \textstyle\sum_{j=1}^{J} t_j^2 - C \\ &h(t) =< t, \widehat{\theta}_i > -(||\widehat{\theta}_i||^2 - r_i^2) = \textstyle\sum_{j=1}^{J} t_j \widehat{\theta}_{ij} - C \end{aligned} \tag{25}$$

One solution to the problem gives the coordinates of $c_i^{\min}$ with the minimum distance $d(f(c_i^{\min}), \mathcal{H})$.

The algorithm is different for the case when the origin is in the ball $\mathbb{B}_i$. For this case, $C_i = ||\widehat{\theta}_i||^2 - r_i^2 \leq 0$ and the coordinates of the maximum angle satisfies:

$$\begin{array}{l} < t, w >= -||t||||w|| \Longleftrightarrow t = (-a)w \text{ with } a > 0 \\ ||t - \widehat{\theta}_i|| = r_i^2. \end{array} \tag{26}$$

References

[1] Benjamini, Y., Hochberg, Y. (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing", *Journal of Royal Statistical Society, B*, 57, 1.

[2] Bar-Joseph, Z., Gerber, G., Gifford, D.K., Jaakkola, T.S. (2002). "A new approach to analyzing gene expression time series data", *Proceedings of the 6th Annual International Conference on RECOMB*, pp 39-48.

[3] Ben-Dorr, A, Shamir, R. and Yakhimi, Z. (1999). "Clustering gene expression patterns", *J. of Computational Biology.*

[4] Beran, R. (2000), "REACT Scatterplot Smoothers: Superefficiency through basis economy", *Journal of the American Statistical Association*, 95, #449, pp 155-171.

[5] Beran, R., Dúmbgen, L. (1998), "Modulation of estimators and confidence sets", *Annals of Statistics*, 26, 5, pp 1826-1856.

[6] Duda, R.O., Hart, P.E. (1973). "Pattern classification and scene analysis", John Wiley & Sons, NY.

[7] Dudoit, S., Yang, Y.H., Callow, M.J., Speed, T.P. (Aug 2000). "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments", technical report.

[8] Efron, B., Storey, J. D., Tibshirani, R.(July 2001). "Microarrays, Empirical Bayes Methods, and False Discovery Rates", *Journal of the American Statistical Association*, 96.

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

[10] Fowlkes, E. B., Mallows, C. L. (1983), "A method for comparing two hierarchical clusterings", *Journal of the American Statistical Association*, 78 (383), pp. 553-569.

[11] Fridlyand, J., Dudoit, S. (Sept 2001), "Applications of resampling methods to estimate the number of clusters and to improve the accuracy of a clustering method", technical report # 600.

[12] Handley, D., Serban, N., Peters, D., O'Doherty, R., Field, M., Wasserman, L., Spirtes, P., Scheines, R., Glymour, C. (2003),"Evidence of cross-hybridization artifact in expressed sequence tags (ESTs) on cDNA microarrays", *Genetics*, in press.

[13] Hastie, T., Tibshirani, R., Eisen, M., Alizadeh, A., Levy, R., Staudt, L., Chan, W., Botstein, D., Brown, P. (2000), "'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns", *Genome Biology*, I(2):research0003.1-0003.21.

[14] Li, Ker-Chau (1989), "Honest confidence regions for nonparametric regression", *Annals of Statistics*, 3, pp 1001-1008.

[15] Meilă, Marina (2002), " Comparing clusterings", University of Washington, Statistics Technical Report 418.

[16] Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R., Tsui,K.W. (2001), "On differential Variability of Expression Ratios: Improving Statistical Inference about Gene Expression Changes from Microarray Data", *Journal of Computational Biology*, 8, # 1, pp. 37-52.

[17] Pollard, D. (1981), "A central limit theorem for k-means clustering", *Annals of Probability*, 1,pp. 919-926.

[18] Pollard, D. (1982), "Strong consistency of k-means clustering", *Annals of Statistics*, 1,pp. 135-140.

[19] Rand, W. M. (1971), "Objective criteria for the evaluation of clustering methods".*Journal of the American Statistical Association* , 66, pp. 846-850.

[20] Serban, N., Wasserman, L. (2003), "Identifying genes altered by a drug in temporal microarray data: A case study", JSM 2003.

[21] Storey, J. D. and Tibshirani, R.(2002), "Estimating FDR under Dependence with Applications to DNA microarrays", technical report.

[22] Tibshirani, R., Hastie, T., Narasimhan, B., Eisen, M., Sherlock, G., Brown, P., Botstein, D. (2001)."Exploratory screening of genes and clusters from microarray experiments", technical report.

[23] Tibshirani, R., Walther, G., Hastie, T. (Dec 2000), "Estimating the number of clusters in a dataset via the Gap statistic". Technical report, published in *JRSSB*,2000.

[24] Yang, Y. H., Dudoit, S., Luu, P., and Speed, T. P., "Normalization for cDNA Microarray Data", technical report.

[25] Yeung, K.Y., Murua, A., Raftery, A., Ruzzo, W.L. (2001). "Model-Based Clustering and Data Transformations for Gene Expression Data", technical report.

[26] van der Vaart, A.W. (1998). "Asymptotic statistics", Cambridge University Press.