Analysis of Multiple Curves and Peaks with applications in molecular biology

Nicoleta Serban

Thesis Proposal, Carnegie Mellon University

1 Introduction

We have witnessed rapid development in our understanding of molecular biology in the past decades. Its breadth and profound applications have made it one of the most active area of modern science. In my thesis, I will analyze two such applications. They involve principled techniques applied to structured datasets obtained from complex experiments.

The first application is determination of protein structures using nuclear magnetic resonance (NMR). One challenge to NMR protein structure determination is that only small proteins can be analyzed. On the other hand, most of the interesting proteins are very complex molecules with very intricate functions and thus relatively few protein structures can be solved using NMR. An obstacle is that large molecules require estimating a large number of resonance frequencies (the main parameters) from already noisy NMR spectroscopy data. Thus, improving the process of 3D protein structure estimation will help recover the structure of a larger number of proteins and their roles in the biological processes and human evolution. The endpoint would be a better understanding of unsolved diseases, of cell development, of our evolution, etc.

The second application is understanding the expression of a large number of sequences of DNA under different experimental conditions. One byproduct of understanding gene expression is discovering new genes and their roles in biological pathways. Another byproduct is inferring gene regulation under environmental changes of the cells. The statistical challenge in gene expression analysis using microarray technology is that we study a large pool of DNA sequences simultaneously. In our application, the expression of the DNA sequences is observed over time where each gene exhibits a different expression profile. Some of the gene expression profiles are overexpressed, some of them are underexpressed and many of them are constantly expressed. Another statistical challenge in microarray data is that the expression of the DNA sequences is contaminated by significant experimental noise.

My proposal is divided in two parts describing the statistical approaches for the applications briefly introduced above. More detailed description of the background of the two applications will be part of my thesis.

2 Peak Identification and Estimation with application to protein structure

The first statistical problem presented in my proposal is motivated by an NMR (nuclear magnetic resonance) experiment performed at Carnegie Mellon University. The NMR technique is used to determine 3D protein structures. The statistical problem is to identify and estimate peaks. We apply the peak identification and estimation method to two-dimensional frequency data after Fourier transformation of the NMR signals. The two-dimensional frequency data consist of very sharp peaks, some of the peaks are isolated, some of them are close together, and some of them are partially or totally overlapping. The peak distribution over the two frequencies is not uniform. Also the peaks have different heights and shapes due to different amplitudes and decay times. Some of the peaks will have flat tops due to limited digital resolution, and some of them will be skewed due to misphasing. In addition to the true peaks, we also expect artifactual peaks. A perspective plot of the intensities on a subset of 100 by 50 design points is in Figure 1.

One might imagine the peak analysis to be an easy problem: just find a data-driven threshold for the noise level and then apply a mixture model to the data above the noise level. There are several difficulties that make the peak analysis still be in the spot light of many researchers.

How do we estimate the noise level? The first difficulty is identifying a data-driven treshold for the noise level. Most of the solutions are based on visually-chosen thresholds.

How can we identify the significant modes? The second difficulty is identifying a large number of modes. The problem of identifying the location of the significant modes and their number has been mostly introduced under the density estimation framework and most of the approaches deal with smooth 1-dimensional density functions (see Minotte, Scott (1993), Minotte (1997), Chaudhuri, Marron(1997), Davies and Kovac (2004)). These methods cannot be easily extended to the regression problem and/or to bivariate predictors. A statistical method for estimating the lower bound of the number of extremes is proposed in Davies and Kovac (2001) for the regression problem (see also the references therein). They also provide intervals for the location of each extreme. Even though the authors don't make any smoothness assumptions, this method is developed only for univariate predictors.

What is the number of components in the mixture model? The third difficulty is that we expect overlapped peaks, peaks that may be a mixture of components, and peaks that are split due to noise. In our application, it is important to account for these problematic peaks. There are a few proposals for modality tests and for detecting the presence of mixture within the density problem (see Hartigan, Hartigan (1985), Muller, Sawitzki (1991), Roeder (1994), Walther (2004)). To the best of my knowledge, there are no such proposals for the regression problem.

Another issue is that we can only estimate a lower bound for the number of components since there will be peaks below the noise level. A limitation for EM type estimation is that we need the number of components and not a lower bound.

What would the model function be? The fourth difficulty surrounds the model assumptions. We can use prior information and assume a parametric model function or if we don't know anything about the peaks' shape we may assume a nonparametric model function. Both alternatives have their limitations.

Asymptotics? Most of the statistical properties of the methods referred above are based on large N, the sample size. We will see in the text that we assume small σ, the error standard deviation, rather than large sample size.

To the best of my knowledge, there hasen't been proposed a technique that incorporates all the difficulties above in the statistical literature, but there have been a few proposals for complete systems for peak analysis in 2D NMR data. One very complex system is AUTOPSY (Koradi et al 1998). A downside of this technique is the large number of unjustified "empirical factors" and tuning parameters the user needs to input. It is also quite complex, consisting of a very large number of steps. There are other software packages that have incorporated automated peak picking procedures which are not as complex as AUTOPSY (see Neidig et al (1995), Kleywegt, Boelens, Kaptein (1990)). For related work on determination of protein structures with NMR spectroscopy see Gronwald, Kalbitzer (2004) and the references therein.

There is also a large literature for peak detection rising from astrophysics, proteomics, and other imaging data. These data display particular difficulties in the peak shapes, peak overlapping, etc.

2.1 Method

Our aim is to develop an automated statistical method for peak identification and estimation which overcomes many of the difficulties mentioned above. In this section we propose such a method. The technique consists of a series of steps summarized below.

Smoothing. We consider data of the form

$$Z_{ij} = f(x_i, y_j) + \sigma\epsilon_{ij}, \quad i = 1, \ldots n, \ j = 1, \ldots, m \tag{1}$$

where $\mathbb{E}(\epsilon_{ij}) = 0$. In the examples of interest, n and m are both large and not necessarily equal (in our dataset, $n = 512$ and $m = 256$). The design points are equally spaced: $(x_i, y_j) = (i/n, j/m)$.

We assume that the function f displays spatial inhomogeneity. That is, the noise is spread fairly uniformly through all data, whereas the signal is quite sparse, being concentrated in some region of the 2D space of (x_i, y_j). This suggests a spatially adaptive smoothing method. One smoothing technique is tensor-product wavelet estimation with

block thresholding. We extend the 1D block thresholding (Cai 1999, Cai, Silverman 2001) to 2D wavelet estimation. A block is a rectangular region around a point (x_i, y_j). The blocks can be disjoint or overlap.

Local Maxima identification. Our goal is to estimate the location of the L components, $\mu_l,\ l = 1, \ldots, L$. In the transformed NMR data, each component corresponds to a frequency line (each frequency line corresponds to one sinusoidal component in the time domain data).

A first task is to identify local maxima/peaks in the frequency data. Under the assumption of stationary Gaussian random field noise, Worsley (1995) estimates the number of local maxima using the Hadwiger characteristic of connected sets. We propose a simpler method that will provide us with a lower bound for the number of peaks as well as initial values for the peak location.

The noise level is estimated using the same concept as in soft- and hard-thresholding (Donoho 1995; Donoho, Johnstone 1995). We estimate the noise level (see Figure 3) by:

$$\widehat{T} = \widehat{\sigma}\sqrt{2\log\left(n \times m\right)}$$

where $\widehat{\sigma}$ is the median absolute deviation of the wavelet coefficients at the finest scale of resolution divided by 0.6745 (MAD variance estimator).

We find the local maxima as follows:

1. Arrange the Z_{ij} larger than the noise level, $\widehat{T}$, in increasing order, $Z_{(ij)}$. Denote $(x_{(i)}, y_{(j)})$ the design point in the 2D frequency coordinates of $Z_{(ij)}$, $i = 1, \ldots n$, $j = 1, \ldots m$.

2. Start with the maximum value $Z_{(nm)}$ and put $(x_{(n)}, y_{(m)})$ in a list of accepted peaks, $\mathcal{L}_P$, and also in a list of design points already visited, $\mathcal{L}_C$. Next examine $Z_{(nm-1)}$ and so on, in decreasing order.

3. For each such $Z_{(ij)} \geq \widehat{T}$, check if its design point $(x_{(i)}, y_{(j)})$ is within a distance of radius $r = 1$ from any one in the list of design points already visited, $\mathcal{L}_C$.

   - If $(x_{(i)}, y_{(j)})$ is not within a distance of radius r from each of the design points in the list $\mathcal{L}_C$ then declare $Z_{(ij)}$ a local maximum, and put $(x_{(i)}, y_{(j)})$ in the peak list, $\mathcal{L}_P$, as well as the list of design points already visited, $\mathcal{L}_C$.
   - If $(x_{(i)}, y_{(j)})$ is within a distance of radius r from at least one of the design points in the list then $Z_{(ij)}$ is not a local maximum, and only put $(x_{(i)}, y_{(j)})$ in the list of design points already considered, $\mathcal{L}_C$.

   Then consider the next $Z_{(ij-1)} \geq \widehat{T}$, and repeat.

The algorithm above provides a list of local maxima, $\mathcal{L}_P$, and their locations in the 2D frequency coordinates.

In the following, a component is defined by a local maximum or a peak. However, a peak or local maximum may consist of more than one component as we will discuss later in the text. For now, we assume that each local maximum represents one component only.

Peak-Parameter Estimation. At this step we take the locations of the identified peaks and estimate the parameters corresponding to each of them. The center of a peak needs not fall exactly at the location we identified at the previous step. We estimate the peak location as well as its width and its amplitude.

We begin with the assumption:

$$f(x) = \sum_{l=1}^{L} A_l g\left(\sigma_l^{-1}(x - \mu_l)\right). \tag{2}$$

where g is known symmetric function.

This assumption is relevant for NMR data since the NMR signals are sums of sinusoids. Each component in the frequency domain model (2) corresponds to one sinusoidal component in the time domain data.

The shape of peaks is expected to be fairly symmetric and the peak data are additive. Hence, we assume a parametric additive model of the form:

$$\mathbb{E}[Z|X,Y] = \sum_{k=1}^{L} f_k(X,Y)$$

$$f_k(x,y) = H_k e^{-\left(\frac{(x-\mu_{xk})^2}{(2\sigma_{xk}^2)} + \frac{(y-\mu_{yk})^2}{(2\sigma_{yk}^2)}\right)}$$

where L is the number of components. H_k is the amplitude, $\mu_k = (\mu_{xk}, \mu_{yk})$ is the peak location, and $\sigma_k = \sigma_{xk}, \sigma_{yk})$ controls the width of the peak in each of the two dimensions. Denote the parameters for peak k: $\theta_k = (H_k, \mu_k, \sigma_k)$. For each peak, there are 5 parameters to be estimated. Thus for a large number of peaks ($L = 163$ in our NMR data) the dimension of the parameter space is large (about 800 for our data). We overcome the problem of estimating a large number of parameters by using a modified Gauss-Seidel-Newton algorithm (also called backfitting):

1. Initialize: First we need to obtain initial values close to the true parameters. For example, the initial value for μ_k is the location of the local maximum obtained in the previous section, and the height is the intensity value at that location.

2. Backfitting: For $j = 1, \ldots, p, 1, \ldots, p, \ldots$ regress the partial residuals:

$$E_i^{(j)} = Z_i - \left(\sum_{k=1}^{j-1} f_k(X_i, Y_i|\theta_k^{(j+1)}) + \sum_{k=j+1}^{p} f_k(X_i, Y_i|\theta_k^{(j)})\right)$$

against the predictor $f_j(X_i, Y_i)$ with $i = 1, \ldots, N$ and estimate the parameter θ_j conditional on the other parameters using a nonlinear least squares estimation algorithm. Thus the parameters of one peak are estimated while all the others are fixed. We cycle over all components for a few times until convergence.

We estimate the parameters by minimizing the square loss $\mathbb{E}(Z - f(X, Y))^2$ for a subspace of functions in L_2. An advantage of using least squares loss over some other losses (e.g. L2E) is that we don't need to make any assumption about the distribution of the errors. By using backfitting method with good initial values we estimate the peak parameters in a very localized manner, thus robust to outliers. However, the robustness to outliers may fail when estimating peaks that overlap.

2.2 Discussion

One standard approach to protein structure determination is Nuclear Magnetic Resonance (NMR). NMR lies at the interface between biology, chemistry and physics. However, for an NMR experiment to reach the final objective (i.e. identify protein bonds), a biologist/chemist/physicist will have to analyze spectroscopy data. These are signal data over a large number of time points in one or more dimensions. Parametrically, the 2D NMR signal data can be described as a sum of exponentially decaying sinusoids plus noise:

$$S(t_{1k}, t_{2j}) = \sum_{l=1}^{L} A_l e^{i\phi_l} e^{-t_{1k}/\tau_{1l}} e^{-t_{2j}/\tau_{2l}} e^{it_{1k}w_{1l}} e^{it_{2j}w_{2l}} + \epsilon_{kj} \tag{3}$$

here ω_{1l}, ω_{2l} are the resonance frequencies of one sinusoid, A_l is the amplitude, τ_{1l}, τ_{2l} are the decay times, and ϕ_l is the phase at time 0. Each sinusoid corresponds to a single nuclear resonance. The parameters of major interest are the sinusoidal frequencies ω_{1l} and ω_{2l}.

For 1D NMR signals, there are a few statistical approaches to fitting the parametric model (estimating the parameters for each resonance/sinusoid in the time domain). A common one is estimation of a the nonlinear regression model. A recent one is using the filter diagonalization method (FWD) (Hu et al 1998). However, this is not the standard approach to estimating the NMR spectroscopy parameters since we need the number of resonance frequencies, L. Also, these techniques are not yet extended to 2D NMR signals.

The standard approach to parameter estimation is a two-step technique: transformation into frequency domain (estimating the power spectral density) and peak identification/estimation in the frequency domain. The transformation step is quite complex and involves a few other sub-steps such as apodization (signal convolution with a window function) and/or zero padding, phase correction (correction for the phase ϕ_l at time $t = 0$), and baseline correction. We processed our spectroscopy data with the widely used FELIX program.

We apply our smoothing and peak identification technique to the frequency domain data. The perspective plot of a subset of the smooth intensity data is in Figure 2. The image of all intensity data after smoothing and thresholding is in Figure 4.

3 Cluster Analysis of Gene Expression Profiles

Our second application presented in this proposal is motivated by a genetic microarray experiment conducted at the University of Pittsburgh. This experiment provided expression profiles for 5355 DNA sequences over 15 time points. The primary goal is to analyze the similarity of these expression profiles and cluster them by shape. This problem is challenging because of the small number of time points but large number of expression profiles, the small signal-to-noise ratio, and the large number of flat profiles (constantly expressed genes). We also expect artifactual signals which are due to the experimental error rather than inherent signal. We present such artifactual signals due to cross-hybridization in Handley et al (2003).

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). However, none of these approaches provides estimates of uncertainty, which is particularly relevant given the highly noisy character of microarray data.

3.1 Method

We propose a technique for nonparametrically estimating and clustering a large number of 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 added by 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).

We consider data of the form,

$$Y_{ij} = f_i(t_{ij}) + \sigma_i \epsilon_{ij}, \quad i = 1, \ldots, N, \quad j = 1, \ldots, m. \tag{4}$$

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 radius c.

Our method involves a series of steps briefly described below.

Data transformation. Let $\psi_1, \psi_2, \ldots$ be an orthonormal basis for $\mathcal{F}$ and write

$$f_i(t) = \sum_{j=1}^{\infty} \theta_{ij}\psi_j(t) \text{ where } \theta_{ij} = \int f_i(t)\psi_j(t)\,dt. \tag{5}$$

We estimate f_i by $\widehat{f}_i^J(t) = \sum_{j=1}^{J} \widehat{\theta}_{ij}\psi_j(t)$. We transform the data into the Fourier domain using the cosine basis:

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

where $\Psi = \{\psi_{rj}\}_{r,j}$ is the Gram-Schmidt orthogonalized matrix of $\Phi = \{\phi_r(t_j)\}_{r=1,\ldots,m,\ j=1,\ldots,m}$, $\phi_0(t) = 1,\ \phi_r(t) = \sqrt{2}\cos(2\pi r)$.

Smoothing. We propose two different approaches for estimating J. One approach is based on minimizing the regret function:

$$\widehat{r}_i(J) = \widehat{R}_i(J) - \min_{1\le k\le m} \widehat{R}_k(J) \text{ with } \widehat{R}_i(J) = \frac{J\widehat{\sigma}_i^2}{m} + \sum_{j=J+1}^{m}\left(\widehat{\theta}_{ij}^2 - \frac{\widehat{\sigma}_i^2}{m}\right)_+ \tag{6}$$

which measures how much risk is sacrificed for curve i if smoothing parameter J is used. We want the smoothing parameter to be the same for all curves even though the minimum regret will be attained at different values of J. To find the optimal amount of smoothing simultaneous for all curves, we minimize the total regret function:

$$t(J) = \sum_{i=1}^{n} \widehat{r}_i(J). \tag{7}$$

The second approach is to consider values of J simultaneously and choose the one that leads to the most efficacious clustering (the multiscale approach).

Screening out flat curves. We remove those curves (expression profiles) which are constant over time using simultaneous hypothesis testing. The null hypothesis for gene expression profile i is: $H_{0i} : f_i(t) = c_i$ for some constant c_i. This suggests the test statistic

$$T_i = \sum_{j=2}^{m} \widehat{\theta}_{ij}^2.$$

We reject the null hypothesis for large value of T_i. To correct for the multiplicity problem we use the Benjamini-Hochberg (1995) FDR method. We obtain a set of non-constant expression profiles $\widehat{\mathcal{A}}$. The curves in $\widehat{\mathcal{A}}$ will be clustered with the technique described below.

Confidence Set for f_i. For the minimum regret smoothing approach, 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

$$\mathbb{B}_i = \Big\{(\theta_{i1}, \ldots, \theta_{im}) : \sum_{j=1}^{m} (\theta_{ij} - \widehat{\theta}_{ij})^2 \leq s_i^2\Big\} \text{ where } s_i^2 = \frac{z_{\frac{\alpha}{N}} \widehat{\tau}_i}{\sqrt{m}} + \widehat{R}_i, \tag{8}$$

where z_α is the α quantile of the standard normal and $\widehat{\tau}_i$ is the estimate of the asymptotic variance of

$$\sqrt{n}\left(\sum_{j=1}^{m} (\theta_{ij} - \widehat{\theta}_{ij})^2 - \widehat{R}_i\right).$$

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 $c > 0$,*

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

Clustering. We want to identify curves with similar shape. This suggests using Pearson correlation as the similarity measure.

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.

We cluster the standardized curve coefficients

$$\widetilde{\theta}_j = \frac{\theta_j}{\sqrt{\sum_{j=2}^{\infty} \theta_j^2}}, \quad j \geq 2$$

since the correlation between two curves can be expressed as:

$$\rho(f_i, f_j) = 1 - \frac{||\ \ \widetilde{\theta}_i - \widetilde{\theta}_j||^2}{2}. \tag{9}$$

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

We cluster the standardized estimated coefficients using the k-means clustering algorithm with the Euclidean distance. Any other distance-based clustering method could be used.

We estimate the number of clusters using the gap method of Tibshirani, Walther, and Hastie (2000). We can also infer the number of clusters using the clustering estimation error introduced next.

Estimating the clustering error rate. 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 $T(f, g) = 1$.

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. We define the *clustering estimation error* rate for K clusters $\eta(K)$ by

$$\eta(K) = \frac{1}{\binom{N}{2}} \sum_{r<s} I\Big(T_K(f_r, f_s) \neq \widehat{T}_K(\widehat{f}_r, \widehat{f}_s)\Big). \tag{10}$$

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 clustering estimation error rate can be expressed as $\eta = 1 - \mathcal{R}(T, \widehat{T})$ where $\mathcal{R}$ is the Rand index (Rand, 1971).

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. Similarly, $\widehat{T}(\widehat{f}, \widehat{g}) = 1$ if and only if $\widehat{f}$ and $\widehat{g}$ belong to the same member of the tessellation of the estimated curves $\{\widehat{\mathbb{A}}_1, \ldots, \widehat{\mathbb{A}}_k\}$. Based on these facts, we estimate an asymptotic upper bound for the clustering estimation error using only the tesselation of the estimated curves.

Theorem **2** *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*

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

The asymptotic upper bound in the above theorem is computationally feasible. We estimate $I\Big(f_i \notin \widehat{\mathbb{A}}_{j(i)}\Big) = I\Big(\mathbb{B}_i \cap \widehat{\mathbb{A}}_r \neq \emptyset \ \text{ for some } r \neq j(i)\Big)$. Replacing the estimate of $I\Big(f_i \notin \widehat{\mathbb{A}}_{j(i)}\Big)$ we obtain the upper bound in the following theorem.

Theorem **3** *Let*

$$\overline{\eta} = \sum_{i=1}^{N} I\Big(\mathbb{B}_i \cap \widehat{\mathbb{A}}_r \neq \emptyset \ \text{ for some } r \neq j(i)\Big) \ \textit{and} \ \ \widehat{\eta} = \frac{\bar{\eta}}{N}\left(1 + \frac{N - \bar{\eta}}{N - 1}\right).$$

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

3.2 Discussion

To evaluate the validity of our clustering method, we generated synthetic datasets for different number of time points, $m = 15, 25, 50$, for different levels of signal-to-noise ratio and accounting for a small time-varying variance.

With a small smoothing parameter, the nonparametric test proves to be powerful in identifying constant curves. This step is important, because a large number of noisy flat curves affects the clustering. We've tried a few other tests but the nonparametric test presented in this paper proved to be the most powerful at a small number of design points.

The last step is the inference of the cluster estimation error. We estimated the clustering error rate based on the fraction of all pairs put incorrectly in the same cluster or put incorrectly in different cluster. To the best of our knowledge, this approach to cluster estimation has not been considered previously. Note that the cluster error rate does not tell us how many clusters there are, but rather, how many we can actually estimate.

For the gene expression data, we identified two clusters of gene expression profiles which showed a change in expression over treatment times. The confidence balls of the non-constant gene expression profiles are in Figure 6.

Besides characterizing the uncertainty of the clustering, we can use the estimated upper bound for the clustering error rate to infer the smoothing parameter and the number of clusters. The estimated clustering error bound is about 0.1 for 2 clusters which shows that the error due to the fact we are clustering the estimated rather than the true curves is low (see Figure 7).

Detailed description of the method and results are presented in Serban and Wasserman(2004).

4 Proposed Work

My thesis proposal consists of two statistical problems. My proposed work focuses on peak identification and estimation. If time allows, I will further investigate other ideas

related to the analysis of multiple curves. Below there is an outline of my proposed work in prioritized order.

1. To complete our technique for peak identification and estimation we need to estimate a lower bound for the number of components (see Section 4.1) and to estimate peak location (see Section 4.2).

2. Peak analysis is common in other science areas such as proteomics and astrophysics. Ideally, we would like to validate and/or extend our technique across different applications to understand the generality of our method. We would also like to compare the performance of other methods related to peak analysis, assuming software is available, with our technique.

3. There are a few other directions we can take if the time allows (see Section 4.3). They are important problems within the NMR framework.

4. We have completed our technique for cluster analysis of multiple curves. We still have a few ideas to follow if we complete our work on peak analysis (see Section 4.4).

4.1 Lower Bound for Number of Components

The direction we want to take next is estimation of a lower bound for the number of components within the additive model. We can estimate a lower bound for the number of local maxima using the algorithm proposed in Section 2. However, the number of local maxima is not equal to the number of components. We need to account for two scenarios. First, two local maxima may correspond to only one component. Second, one local maximum may consist of more than one component. See Figure 5. We define the two problems within a more general framework.

Separation of peaks. First, we want to test whether two local maxima are correctly separated using our method or any other method. We propose a test under the model:

$$Z_i = f(x_i) + \sigma\epsilon_i, \quad i = 1, \ldots, N \tag{11}$$

where $f \in \mathcal{F}$, $\mathcal{F}$ is some nonparametric class of function and $x_i, i = 1, \ldots, n$ are univariate design points.

We begin by assuming that the centers of the two local maxima are fixed, known and the observed data points between the two local maxima are $Z_i,\ i = n, n+1, \ldots, m$ with $m \leq N,\ n \geq 1, m - n \geq 3$. The separation test becomes:

$$H_0 :\ f(x) \textit{ monotone} \text{ between the design points } x_n \text{ and } x_m$$
$$H_A :\ f(x) \textit{ convex} \text{ between the design points } x_n \text{ and } x_m$$

where H_0 is rejected if the two peaks represent two components in the additive model.

The challenging problem is to develop an asymptotic powerful test under the assumption of $\sigma \longrightarrow 0$ and N small.

The test can be easily extended to 2-dimensional peaks.

Detecting mixtures of components. Second, we want to design a test for identifying peaks that may be a mixture of more than one component and to estimate the number of components. We will need a powerful test under the assumption of $\sigma \longrightarrow 0$ and N small. The test hypotheses are:

$$H_0 : \ f(x) = A_1 g(\sigma_1^{-1}(x - \mu_1)), \ H_A : \ f(x) = A_1 g(\sigma_1^{-1}(x - \mu_1)) + \ldots + A_k g(\sigma_k^{-1}(x - \mu_k))$$

where g is a parametric function, which is inferred from the shape of isolated peaks, and the locations of the modes are unknown.

The number of components. After identifying the local maxima, we can further apply the separation test to any two close peaks in order to detect components which are incorrectly separated due to noise. Further, we apply the test that accounts for the number of components to all the identified peaks. Finally, we can estimate a lower bound of the number of components based on the number of local maxima, the number of peaks that are incorrectly separated, and the number of components of each of the identified peaks.

4.2 Parameters Estimation

After separating all the peaks and accounting for all the components, the next step is to estimate the peaks' parameters such as their locations, widths, and heights.

Under a parametric model, we can use prior information regarding the peaks' shape to define the additive model function in (2). Further, we can apply the backfitting algorithm to estimate the model parameters. For NMR frequency data, it is reasonable to assume that all peaks have a similar shape given by g in (2). For other types of applications, the assumption on the shape should be reconsidered. One way to relax the shape restriction is to assume a nonparametric model.

Under a nonparametric model, the location, width and height parameters cannot be estimated from the model function. However, we can infer them from the smoothed function.

For the beginning, we will focus on parametric versions of the model function. If time allows, we will also explore nonparametric model functions.

4.3 Other Considerations

Noise peaks. An important problem to be examined is that of differentiating the signal peaks from the noise peaks. Artifactual peaks could be produced by solvent lines or

poor data processing. An approach to this problem is to classify the peaks as signal or artifactual according to the shape and linewidth since we expect a small proportion of artifactual peaks and similar shape with similar linewidth for all the components. Classification approaches have been already proposed in the literature (see Carrara et al. 1993, Schulte et al. 1997), but they have their limitations.

Data processing. NMR data processing involves a series of steps. Before Fourier transformation, the signal is convoluted with a window function (this is called the *apodization* step) to enhance the sensitivity or the resolution. After apodization, NMR signals are transformed into Fourier space. In the Fourier space, the frequency data is phase corrected (to correct for ψ_l in (3)) and baseline corrected (correct the baseline distortions due to delay in detection, to the band limit of the signal, etc.). Phase correction and baseline correction are still open problems even though there are several proposals for each of them.

3D NMR data. Currently, NMR experiments for protein structure determination can also include 3 or more dimensions. One challenging problem is to extend our technique to more than two dimensions.

Analysis of time domain data. A different approach to parameter estimation is to fit the parametric time domain model (3). By fitting this model, we avoid processing steps altogether, which introduces different sources of error. This is a different statistical problem by itself and there is quite an extensive literature related to it. For that, we may not achieve substantial progress on this problem within our timeline but it is worth consideration.

4.4 Further Ideas on Analysis of Multiple Curves

Clustering confidence sets. An extension to our clustering analysis is using the confidence sets of the true curves rather than their estimated curves and clustering based on the distances between the confidence sets. A popular measure of the closeness between two sets is the Hausdorff distance. However, in the context of our problem, we propose using the maximal distance between two sets:

$$M(\mathbb{B}_1, \mathbb{B}_2) = \sup_{f_1 \in \mathbb{B}_1, f_2 \in \mathbb{B}_2} \rho(f_1, f_2)$$

where $\rho(f_1, f_2)$ is the distance between any two objects f_1 and f_2 (in our example is the correlation coefficient). This distance has the property in the theorem below.

Theorem 4 *Let $\mathcal{F}_\beta(c)$ denote a Sobolev space of order β and radius c. Then, for any $\beta > 1/2$ and any $c > 0$,*

$$\liminf_{N \to \infty} \sup_{f_1, \ldots, f_N \in \mathcal{F}_\beta(c)} \mathbb{P}\Big(M(\mathbb{B}_i, \mathbb{B}_j) \geq d(f_i, f_j),\ \forall i, j = 1, \ldots, N \Big) \geq 1 - \alpha.$$

Both the Hausdorff distance and the maximal measure can be computed when the measure of similarity between two curves is the correlation or the Euclidean distance.

Next we want to use a clustering algorithm which requires only the distance matrix for the objects, i.e. the matrix of all pairwise distances. Our distance matrix is given by the maximal distance $D = \{d_{ij} = M(\mathbb{B}_i, \mathbb{B}_j)\}$. One such algorithm is the single-linkage tree, which is the oldest clustering algorithm based on the distance matrix.

We would like to further develop this confidence-set clustering in the context of clustering a large number of curves.

Optimal smoothing. The current smoothing procedures are developed in general contexts. Such an example is the smoothing by minimum regret used in our technique. A different approach to the estimation of the optimal smoothing in the context of clustering is to define a distance between the true tessellation and the estimated tesselation and minimize this distance:

$$\min d(\mathcal{T}_k, \widehat{\mathcal{T}}_k).$$

Such a technique allows us to identify the smoothing level at which the true clustering is the closest to the estimated clustering.

One challenge is to define the distance to be minimized in order to have a feasible computational problem. This is one research idea we will further examine.

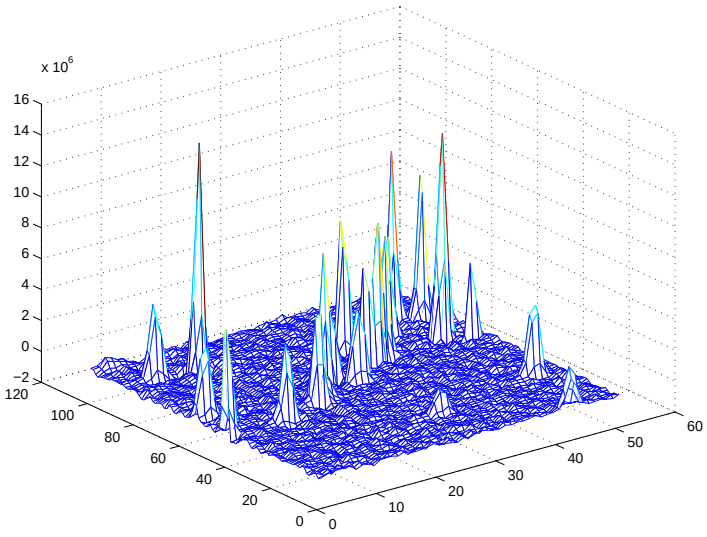

Figure 1: Perspective plot of intensities on a subset of the design points (100×50).

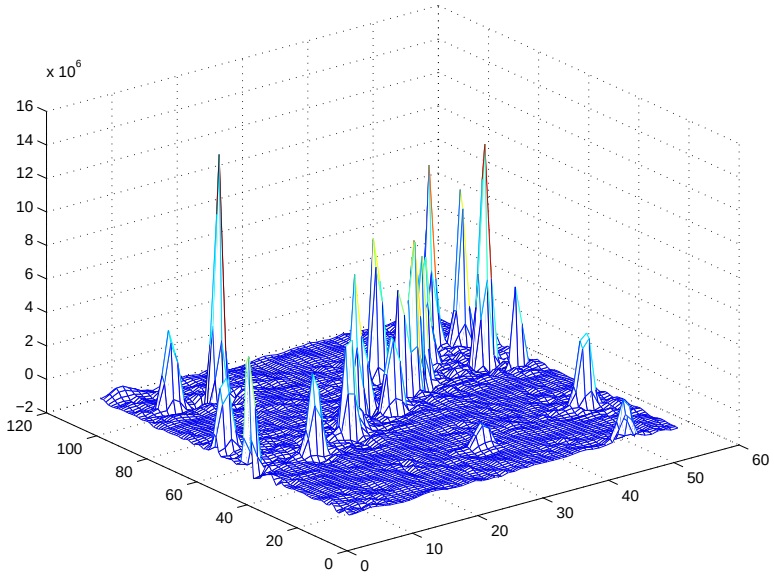

Figure 2: Smooth frequency data

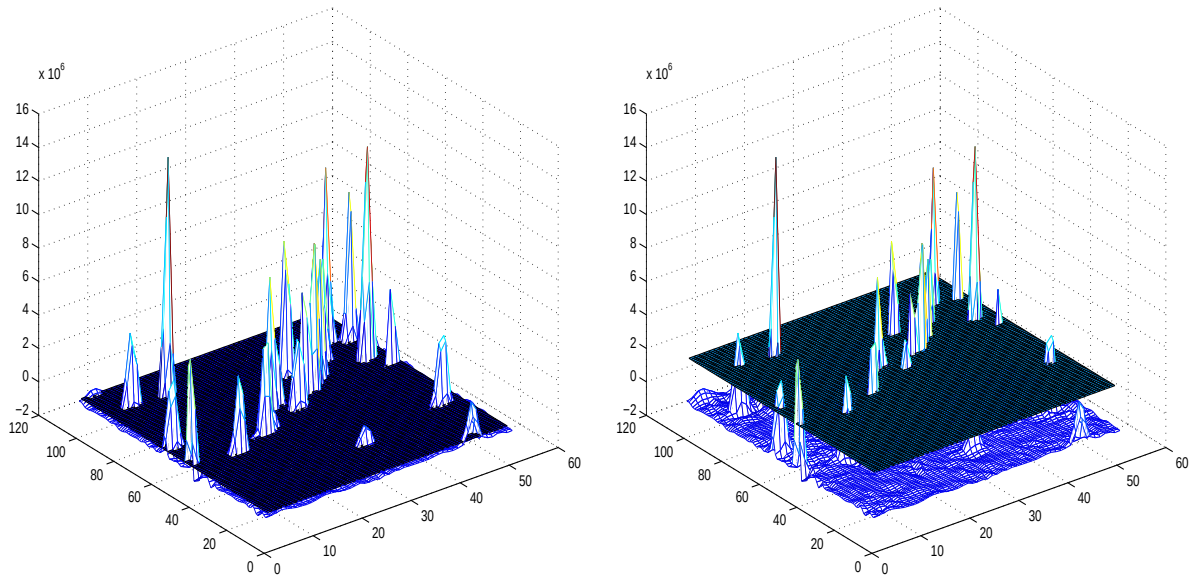

Figure 3: Left figure shows the estimated noise level ($\widehat{T} = \widehat{\sigma}\sqrt{2(n \times m)}$) and the right figure shows the threshold level equal to $10\widehat{T}$ of intensities on a subset of the design points (100×50).

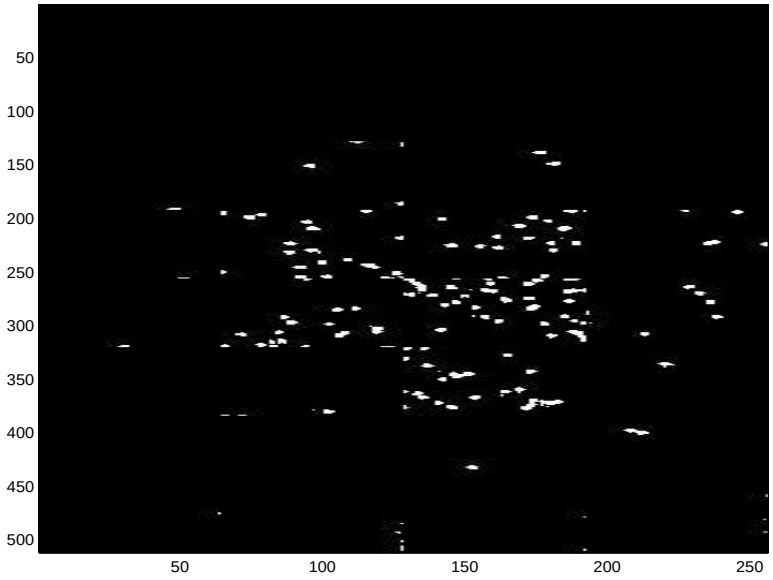

Figure 4: Image of all data above the noise level after smoothing and thresholding.

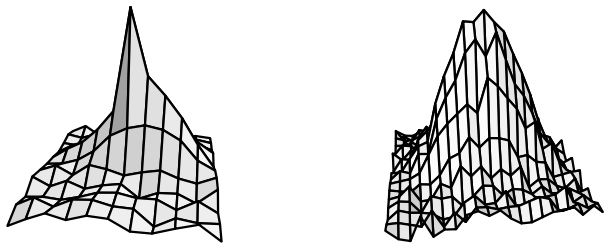

Figure 5: Left panel: Two components in one local maximum. Right panel: One component represented by two local maxima.

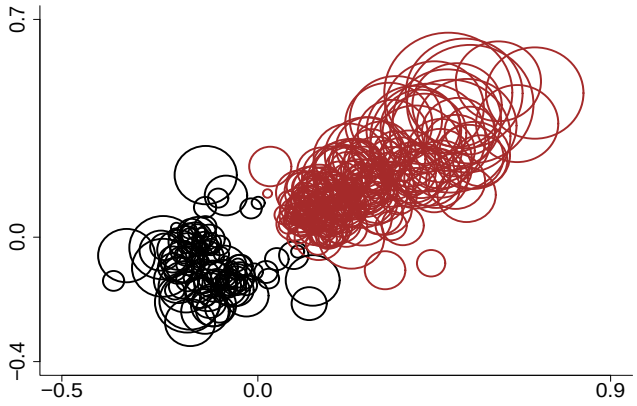

Figure 6: 2D balls for significant genes. Each circle represents the confidence ball of one gene with radius computed using χ^2 approximation. The red circles are in cluster 1 and the black circles are in cluster 2.

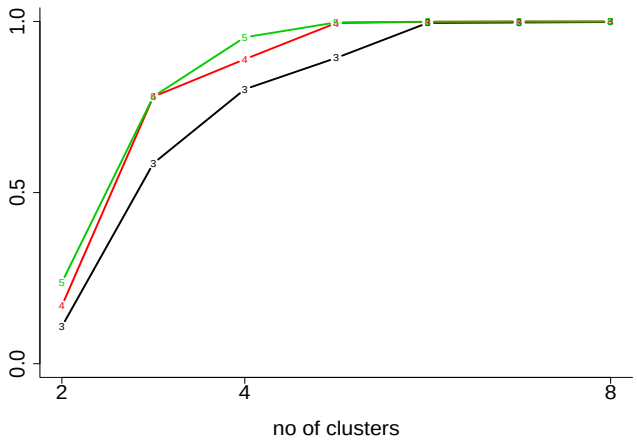

Figure 7: Upper bound for η - microarray data. Each curve represents the asymptotic upper bound for η for a given smooth parameter J over the number of clusters. For example, curve 3 in the figure consists of the estimated upper bound for $J = 3$.

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

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

[5] Cai, T. (1999), "Adaptive Wavelet Estimation: A Block Thresholding and Oracle Inequality Approach", *Ann. of Statistics*, 27, 898-924.

[6] Cai, T., Silverman, B.W. (2001), "Incorporating Information on Neighboring Coefficients into Wavelet Estimation", *Sankhya*, 63, 127-148.

[7] Carrara, E.A., Pagliari, F., Nicolini, C. (1993), "Neural Networks for the Peak-Picking of Nuclear Magnetic Resonance Spectra", *Neural Networks*, 6, 1023-1032.

[8] haudhuri, P., Marron, J.S. (1999), "Sizer for exploration of structures in curves", J. Amer. Statist. Assoc., 94, 807-823.

[9] Davies, P.L., Kovac, A. (2001), *Local Extremes, Runs, Strings and Multiresolution*, Ann. Stat., 29, 1-65.

[10] Donoho, D.L. (1995), "De-Noising by Soft-Thresholding", *IEEE Transactions on Information Theory*, 41, 3, 613-627.

[11] Donoho, D.L., Johnstone, I.M. (1995), "Wavelet Shrinkage: Asymptotia?", *J. of Royal Society B*, 57, 2, 301-369.

[12] 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.

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

[14] 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.

[15] Gronwald, W., Kalbitzer, H.R. (2004), "Automated structure determination of proteins by NMR spectroscopy", *Prgress in NMR Spectroscopy*, 44, 33-96.

[16] Handley, D., Serban, N.,Peters, D., Doherty, R., Field, M., Wasserman, L., Spirtes, P., Scheines, R., Glymour, C. (2003), "Evidence of systematic Expressed Sequence Tag (EST) I.M.A.G.E. clone cross-hybridization on cDNA microarrays", *Genomics*.

[17] artigan, J.A. (1975), "Clustering Algorithms", John Wiley & Sons, Inc.

[18] artigan, J.A., Hartigan, P.M. (1985), "the Dip Test of Unimodality", Ann. Statist., 13, 70-84.

[19] 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.

[20] Hu,H., Van, Q. N., Mandelshtam, V.A., Shaka, A.J. (1998)," Reference Deconvolution, Phase Correction, and Line Listing of NMR Spectra by the 1D Filter Diagonalization Method", *Journal of Magnetic resonance*, 134, 76-87.

[21] Jennrich, R.I. (1969), "Asymptotic Properties of Non-linear Least Squares Estimators", *The Annals of Mathematical Statistics*, 40, 633-643.

[22] Kleywegt,G.J.,Boelens, R., Kaptein, R. (1990), "A versatile approach toward the partially automatic recognition of cross peaks in 2D NMR spectra", *J. Magn. Reson.* 88, 601–608.

[23] Koradi, R., Billeter, M., Engeli,, M., Güntert, P., Wüthrich, K. (1998), "Automated peak picking and peak integration in macromolecular NMR Spectra using AUTOPSY", *Journal of Magnetic Resonance*, 135, 288-297.

[24] cLachlan, G., Peel, D. (2000), "Finite Mixture Models", John Wiley& Sons.

[25] inotte, M.C. (1997), "Nonparametric testing of existence of modes", Ann. Statist., 25, 1646-1660.

[26] inotte, M.C., Scott, D.W. (1993), "The mode tree: a tool for visualization of nonparametric density features", J. Comput. Graph. Statist., 2, 51-68.

[27] uller, D.W., Sawitzki, G. (1991), "Excess Mass Estimates and Tests for Multimodality", J. Amer. Statist. Assoc., 86, 738-746.

[28] Neidig, K.-P., Geyer, M., Grler, A., Antz, C., Saffrich, R., Beneicke, W., and Kalbitzer, H.R. (1995), "AURELIA, a program for computer-aided analysis of multidimensional NMR spectra", *Journal of Biomolecular NMR* 6, 255-270.

[29] 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.

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

[31] Roeder, K. (1994), "A Graphical Technique for Determining the Number of Components in a Mixture of Normals", *Journal of the American Statistical Association*, 89, 487 - 495.

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

[33] Serban, N., Wasserman, L. (2004), "CATS: Cluster After Transformation and Smoothing", under revision *Journal of the American Statistical Association.*

[34] Schulte, A., Grler, A., Antz, C.; Neidig, K. P.; Kalbitzer, H.R. (1997), "Use of Global Symmetries in Automated Signal Class Recognition by a Bayesian Method", 129, pp. 165-172.

[35] Stoica, P., Moses, R. (1997), Introduction to Spectral Analysis, Prentice Hall.

[36] 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.

[37] Wakefield, J., Zhou, C., Self, S. (2002), "Modelling Gene Expression Data over Time: Curve Clustering with Informative Prior Distributions", *Bayesian Statistics 7, Proceedings of the Seventh Valencia International Meeting*, 2003.

[38] alther, G. (2004), "Multiscale Maximum Likelihood Analysis of a Semiparametric Model, with Applications", to appear in Ann. Statist.

[39] Worsley, K.J. (1995), *Estimating the Number of Peaks in a Random Field Using the Hadwiger Characteristic of Excursion Sets, with Applications to Medical Images*, Ann. Stat., 23, 640-669.

[40] Vrahatis, M.N., Magoulas, G.D., Plagianakos, V.P. (2002), "From linear to nonlinear iterative methods", *Applied Numerical Mathematics*, 45, 59-77.