Clustering for sparsely sampled functional data

GARETH M. JAMES and CATHERINE A. SUGAR
Marshall School of Business, University of Southern California, Los Angeles, California 90089-0809

Abstract

We develop a flexible model-based procedure for clustering functional data. The technique can be applied to all types of curve data but is particularly useful when individuals are observed at a sparse set of time points. In addition to producing final cluster assignments, the procedure generates predictions and confidence intervals for missing portions of curves. Our approach also provides many useful tools for evaluating the resulting models. Clustering can be assessed visually via low dimensional representations of the curves, and the regions of greatest separation between clusters can be determined using a discriminant function. Finally, we extend the model to handle multiple functional and finite dimensional covariates and show how it can be applied to standard finite dimensional clustering problems involving missing data.

Some key words: Functional clustering; Discriminant functions; Curve estimation; High dimensional data.

1 Introduction

Cluster analysis is the art of identifying groups in data. It can be thought of as the dual of discriminant analysis, the key distinction being that in cluster analysis the group labels are not known *a priori*. There are many clustering methods, ranging from heuristic approaches such as k-means (Hartigan and Wong, 1978) and linkage analysis (Kaufman and Rousseeuw, 1990) to more formal model-based procedures (Banfield and Raftery, 1993). In this paper we present a model-based approach for clustering functional data. The method is particularly effective when the observations are sparse, irregularly spaced, or occur at different time points for each subject.

1.1 Model-based clustering

In model-based clustering it is assumed that the observations $\mathbf{x}_1,\ldots,\mathbf{x}_n$ are generated according to a mixture distribution with G components. Let $f_k(\mathbf{x}|\theta_k)$ be the density corresponding to the kth cluster, parameterized by θ_k, and let $\mathbf{z}_i = (z_{i1},\ldots,z_{iG})$ be the cluster membership vector for the ith observation where $z_{ik} = 1$ if $\mathbf{x}_i$ is a member of the kth cluster and 0 otherwise. The $\mathbf{z}_i$'s are unknown and are generally treated in one of two ways. In the "classification likelihood" approach, the $\mathbf{z}_i$'s are viewed as parameters and the model is fit by maximizing the likelihood

$$L_C(\theta_1,\ldots,\theta_G;\mathbf{z}_1,\ldots,\mathbf{z}_n|\mathbf{x}_1,\ldots\mathbf{x}_n) = \prod_{i=1}^{n} f_{\mathbf{z}_i}(\mathbf{x}_i|\theta_{\mathbf{z}_i}). \tag{1}$$

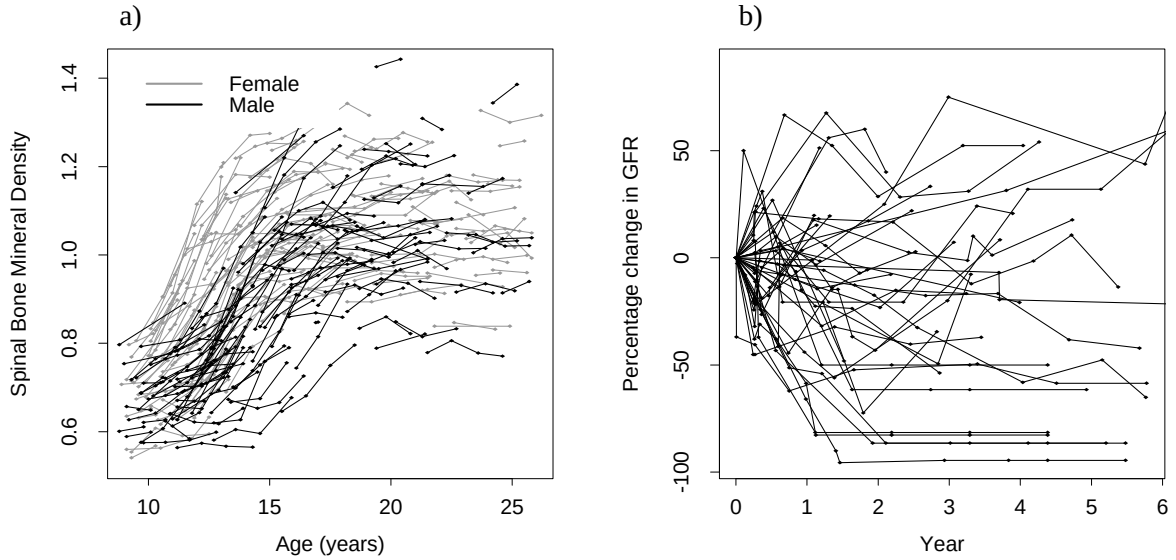

Figure 1: *(a) Measurements of spinal bone mineral density (g/cm^2) for males (black) and females (grey) at various ages, $n = 280$. (b) Percentage change in GFR scores for* 39 *patients with membranous nephropathy.*

When $f_k(\mathbf{x}|\theta_k)$ is multivariate normal with the identity covariance matrix this approach produces the k-means solution. Alternatively, the cluster memberships may be treated as missing data where $\mathbf{z}_i$ is multinomial with parameters $(\pi_1,\ldots,\pi_G)$ and π_k is the probability that an observation belongs to the kth cluster. Then the parameters are estimated by maximizing

$$L_M(\theta_1,\ldots,\theta_G;\pi_1,\ldots,\pi_G|\mathbf{x}_1,\ldots,\mathbf{x}_n) = \prod_{i=1}^{n}\sum_{k=1}^{G}\pi_k f_k(\mathbf{x}_i|\theta_k). \tag{2}$$

This is known as the "mixture likelihood" approach. Here too a multivariate normal distribution with mean μ_k and variance Σ_k has been used successfully in a wide range of applications (Banfield and Raftery, 1993; Celeux and Govaert, 1995; Dasgupta and Raftery, 1998). In both approaches it is generally necessary to use an iterative procedure, such as EM, to estimate the various parameters. The key difference between the classification and mixture approaches is that in the former each point is assigned to a unique cluster, while in the latter each point is assigned a probability of originating from each cluster and so influences all the parameter estimates.

1.2 Classical approaches to clustering functional data

Although normal mixture models are fairly straightforward to use in finite dimensional clustering problems, they are less easy to apply to infinite dimensional data such as curves. Most existing approaches for clustering functional data can be categorized as either *regularization* methods, *filtering* methods or hybrids of the two. Regularization methods work by discretization of the time interval. The resulting data vectors are autocorrelated and high dimensional, leading to unstable estimates of the within cluster covariance matrices unless some form of regularization constraint is imposed (DiPillo, 1976; Friedman, 1989; Banfield and Raftery, 1993; Hastie *et al.*, 1995). Filtering methods project each curve onto a finite dimensional basis,

$\{\phi_1(x), \ldots, \phi_p(x)\}$, and cluster the resulting basis coefficients.

The regularization and filtering approaches can work well when every curve has been observed over the same fine grid of points. However, they break down if, as is often the case in practice, the individual curves are sparsely sampled. Consider the following two examples. The first, illustrated in Figure 1(a), consists of measurements of spinal bone mineral density for 280 males and females taken at various ages and is a subset of the data presented in Bachrach *et al.* (1999). Even though, in aggregate, there are 860 observations taken over a period of almost two decades, there are only 2-4 measurements for each individual covering no more than a few years. None-the-less, we are interested in clustering this data to identify different patterns of growth. For instance, Figure 1(a) suggests that there may be differentiation based on gender, especially in the early years. The second data set, illustrated in Figure 1(b), shows percentage changes in glomular filtration rate (GFR) over a 6 year period, for a group of patients with membranous nephropathy, an auto-immune disease of the kidney. GFR is a standard measure of kidney function. Clinical observations suggest that patients fall into three categories, those that remain relatively stable in terms of GFR, those that decline slowly over time and those that deteriorate rapidly. People in the latter group need aggressive treatment so it is desirable to make an early prediction of cluster membership.

The regularization method can not be applied to these data sets because the curves are sampled at different times. The filtering method also has several problems. First, the variance of the estimated basis coefficients is different for each individual because the curves are measured at different time points. More weight should be placed on the more accurately estimated basis coefficients which standard filtering does not allow. More importantly, for sparse data sets many of the basis coefficients would have infinite variance, making it impossible to produce reasonable estimates. In the spinal bone density data set there are so few observations that it is not possible to fit a separate curve for each individual using any reasonable common basis. For data sets of this type a new approach is necessary.

1.3 An alternative functional clustering approach

In this paper we introduce a general approach to clustering functional data that incorporates the best properties of the regularization and filtering methods while avoiding their most serious drawbacks. As with the filtering approach we convert the original infinite dimensional problem into a finite dimensional one using basis functions. However, instead of treating the basis coefficients as parameters and fitting a separate spline curve for each individual, we use a random effects model for the coefficients. This allows us to borrow strength across curves, producing far superior results no matter how sparsely or irregularly the individual curves are sampled, provided that the total number of observations is large enough. Furthermore, it automatically weights the estimated spline coefficients according to their variances and is highly efficient because it requires fitting few parameters. Finally, it can be used to produce estimates of individual curves that are optimal in terms of mean squared error.

The functional clustering model and an EM style fitting procedure are presented in Section 2. The model is extremely flexible and many standard clustering tools can be easily implemented with it. In Section 3 we demonstrate how to obtain low-dimensional representations of the curves so that the clusters may be assessed visually. We also show how to compute estimates, confidence intervals and prediction intervals for individual curves. Model selection techniques, such as methods for selecting the number of clusters and the spline basis, are described in Section 4. Section 5 shows how the model can be generalized to include multiple functional and finite dimensional variables. Finally, Section 6 discusses how this model can be used to cluster standard high dimensional data with missing values.

2 Modeling functional data

2.1 A general functional model

Since most functional data is longitudinal we adopt the convention of parameterizing our models in terms of t, time. However, our approach applies equally well in other contexts. Let $g(t)$ be the curve of a randomly chosen individual. We will assume that $g(t)$ follows a Gaussian process. If $g(t)$ is a member of the kth cluster we write its expected value and covariance as

$$E\{g(t)\} = \mu_k(t), \quad Cov\{g(t), g(t')\} = \omega_k(t,t')$$

In practice we do not observe $g(t)$ perfectly nor do we observe it at all time points. Let $\mathbf{Y}$ be the vector of observed values of $g(t)$ at times $t_1, \ldots, t_n$. We assume that the measurement errors are independent and normally distributed with mean zero and constant variance σ^2 so that

$$\mathbf{Y} \sim N(\mathbf{M}_k, \Omega_k + \sigma^2 I)$$

where

$$\mathbf{M}_k = \begin{bmatrix} \mu_k(t_1) \\ \mu_k(t_2) \\ \vdots \\ \mu_k(t_n) \end{bmatrix}, \qquad \Omega_k = \begin{bmatrix} \omega_k(t_1,t_1) & \omega_k(t_1,t_2) & \cdots & \omega_k(t_1,t_n) \\ \omega_k(t_2,t_1) & \omega_k(t_2,t_2) & \cdots & \omega_k(t_2,t_n) \\ \vdots & \vdots & \ddots & \vdots \\ \omega_k(t_n,t_1) & \omega_k(t_n,t_2) & \cdots & \omega_k(t_n,t_n) \end{bmatrix} \tag{3}$$

The regularization and filtering approaches can both be viewed as methods for estimating the parameters in (3). The regularization approach obtains estimates of $\mu_k(t)$ and $\omega_k(t,t')$ on a fine lattice of time points. Generally no assumptions are made about the functional form of $\mu_k(t)$ but some restrictions are placed on the structure of $\omega_k(t,t')$. In the filtering method $g(t)$ is represented in terms of a p-dimensional set of basis functions $\phi(t) = (\phi_1(t), \ldots, \phi_p(t))$, i.e. $g(t) = \phi(t)\eta$. The η's are estimated separately for each individual using least squares. The estimated coefficient vectors are then clustered and the resulting cluster means are multiplied by $\phi(t)$ to obtain estimates of the $\mu_k(t)$'s. Estimates of the $\omega_k(t,t')$'s are obtained in a similar manner.

2.2 The functional clustering model

We now present a version of the general functional model that is appropriate for clustering all types of functional data. Let $g_i(t)$ be the true value for the ith individual or curve at time t, and let $\mathbf{g}_i, \mathbf{Y}_i$ and ε_i be, respectively, the corresponding vectors of true values, observed values and measurement errors at times $t_{i1}, \ldots, t_{in_i}$. Then

$$\mathbf{Y}_i = \mathbf{g}_i + \varepsilon_i, \quad i = 1, \ldots, n$$

where n is the number of individuals. The measurement errors are assumed to have mean zero and to be uncorrelated with each other and $\mathbf{g}_i$. Note that this involves an implicit assumption that the unobserved time points are missing at random. Since there are a finite number of observations it is necessary to impose some structure on the individual curves. Like the filtering approach, our method models $g_i(t)$ using basis functions. We chose natural cubic splines because they have desirable mathematical properties, are easy to implement, and require a relatively minimal number of parametric assumptions (de Boor, 1978; Green and Silverman, 1994). We let

$$g_i(t) = \mathbf{s}(t)^T \eta_i$$

where $\mathbf{s}(t)$ is a p-dimensional spline basis vector and η_i is a vector of spline coefficients. The η_i's are modeled using a Gaussian distribution,

$$\eta_i = \mu_{\mathbf{z}_i} + \gamma_i, \quad \gamma \sim N(\mathbf{0}, \Gamma), \tag{4}$$

where $\mathbf{z}_i$ denotes the unknown cluster membership.

There is a further parameterization of the cluster means that will prove useful for producing low-dimensional representations of the curves. Note that μ_k can be rewritten as

$$\mu_k = \lambda_0 + \Lambda\alpha_k, \tag{5}$$

where λ_0 and α_k are respectively p- and h-dimensional vectors, and Λ is a $p \times h$ matrix with $h \leq \min(p, G-1)$. When $h = G-1$, (5) involves no loss of generality while $h < G-1$ implies that the means lie in a restricted subspace. With this formulation the functional clustering model (FCM) can be written as

$$\mathbf{Y}_i = S_i(\lambda_0 + \Lambda\alpha_{\mathbf{z}_i} + \gamma_i) + \varepsilon_i, \quad i = 1, \ldots n, \tag{6}$$

$$\varepsilon_i \sim N(\mathbf{0}, R), \quad \gamma_i \sim N(\mathbf{0}, \Gamma),$$

where $S_i = (\mathbf{s}(t_{i1}), \ldots, \mathbf{s}(t_{in_i}))^T$ is the spline basis matrix for the ith curve. As with finite dimensional models, two different forms of the FCM can be obtained depending on whether the $\mathbf{z}_i$'s are treated as parameters or missing data. There are many possible forms for R and Γ, the covariances of the ε_i's, and the γ_i's. For now we use $R = \sigma^2 I$ and a common Γ for all clusters because we are interested in sparse data sets for which it is desirable to use a small number of parameters. Other choices are explored in Section 5.2. Note that λ_0, Λ and α_k are confounded if no constraints are imposed. Therefore we require that

$$\sum_k \alpha_k = 0 \tag{7}$$

$$\text{and} \qquad \Lambda^T S^T \Sigma^{-1} S \Lambda = I \tag{8}$$

where S is the basis matrix evaluated over a fine lattice of time points that encompasses the full range of the data and $\Sigma = \sigma^2 I + S\Gamma S^T$. The restriction in (7) means that $\mathbf{s}(t)^T\lambda_0$ may be interpreted as the overall mean curve. There are many possible constraints that could be placed on Λ. The reason for the particular form used in (8) will become apparent in Section 3.1.

Notice that our functional clustering model (6) is a special case of the model of Section 2.1 with $\mu_k(t) = \mathbf{s}(t)^T\mu_k = \mathbf{s}(t)^T(\lambda_0 + \Lambda\alpha_k)$ and $\omega_k(t,t') = \mathbf{s}(t)^T\Gamma\mathbf{s}(t')$. Although similar in structure, the FCM differs from the filtering approach in two key respects. The first difference is in the handling of the basis coefficients. In the filtering approach the η_i's are treated as parameters or fixed effects and are estimated directly using only the values obtained from that individual. In the FCM the η_i's are treated as random effects and need not be estimated directly. This allows strength to be borrowed across curves, providing superior results for data containing a large number of sparsely sampled curves. The second major difference in the FCM is the additional parameterization of the cluster means using (5). This formulation has two advantages. First, allowing $h < G-1$ reduces the number of parameters to be estimated which can result in a superior fit for sparse data. Second, as we show in Section 3.1, this parameterization leads to a simple low-dimensional representation of the individual curves that allows for graphical assessment of clustering.

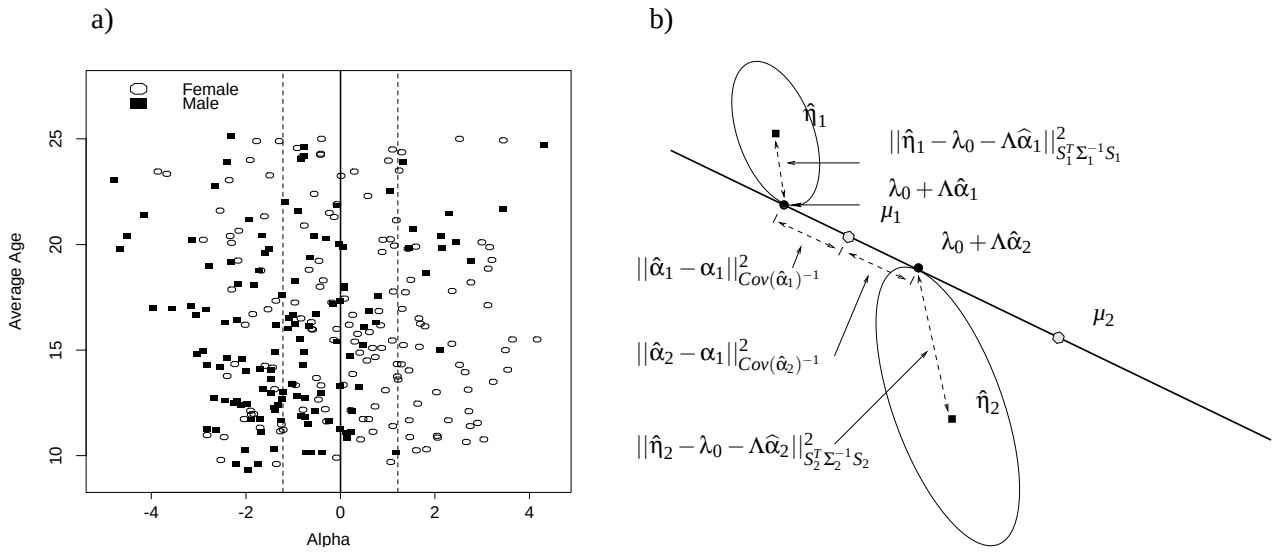

Figure 2: *a) A linear discriminant plot for the bone mineral density data. b) An illustration of the decomposition of the distance between two curves and two cluster centers.*

2.3 Fitting the model

Fitting the FCM involves estimating $\lambda_0, \Lambda, \alpha_k, \Gamma$ and σ^2. This is achieved by maximizing either the classification likelihood given by (1) or the mixture likelihood given by (2), noting that under (6), conditional on the *i*th curve belonging to the *k*th cluster,

$$\mathbf{Y}_i \sim N(S_i(\lambda_0 + \Lambda\alpha_k), \Sigma_i) \tag{9}$$

where $\Sigma_i = \sigma^2 I + S_i \Gamma S_i^T$. In both cases this involves an iterative procedure. Curves are first either assigned to a cluster (classification) or assigned a probability of belonging to a cluster (mixture). Then the parameters are estimated given the current assignments and the process is repeated. Details of the algorithm are provided in Appendix A.

3 Functional clustering tools

Next we discuss three important ways in which our procedure can be used to study clustering of functional data. Section 3.1 describes how to obtain low-dimensional plots of curve data sets, enabling one to visually assess clustering. In Section 3.2 we show how to construct discriminant functions to identify the regions of greatest separation between clusters. Finally, in Section 3.3 we develop optimal methods for estimating the entire curve for an individual, along with pointwise confidence and prediction intervals.

3.1 Low-dimensional graphical representations

One of the chief difficulties in high-dimensional clustering is visualization of the data. Plotting functional data is easier because of the continuity of the dimensions. However, it can still be hard to see the clusters since variations in shape and the location of time-points make it difficult to assess the relative distances between curves. These problems are exacerbated when the curves are fragmentary, as in Figure 1(a). In this section we develop a set of graphical tools for use with functional data. Our method is based on projecting the curves into a low-dimensional space so that they can be plotted as points, making it much easier to detect the presence of clusters.

Figure 2(a) shows the bone mineral curves projected onto a one-dimensional space. The horizontal axis represents the projected curve, $\widehat{\alpha}_i$, while the vertical axis gives the average age of observation for each individual. Points to the left of zero are assigned to cluster 1 and the remainder to cluster 2. Squares represent males and circles females. The dotted lines at α_1 and α_2 correspond to the projected cluster centers. Notice that while there is a significant overlap, most males belong to cluster 1 and most females to cluster 2 even though the model was fit without using gender labels. The plot shows that the clustering separates the genders most strongly for those younger than 16 years. In fact, 74% of such individuals matched the majority gender of their cluster compared with only 57% of those older than 16. This is because girls typically begin their growth spurt before boys.

Figure 2(b) illustrates the procedure by which the $\widehat{\alpha}_i$'s are derived using a two cluster, two curve example. First, $\mathbf{Y}_i$ is projected onto the p-dimensional spline basis to get

$$\hat{\eta}_i = (S_i^T \Sigma_i^{-1} S_i)^{-1} S_i^T \Sigma_i^{-1} \mathbf{Y}_i. \tag{10}$$

Second, $\hat{\eta}_i$ is projected onto the h-dimensional space spanned by the means μ_k to get $\lambda_0 + \Lambda\widehat{\alpha}_i$ where

$$\widehat{\alpha}_i = \left(\Lambda^T {S_i}^T \Sigma_i^{-1} S_i \Lambda\right)^{-1} \Lambda^T {S_i}^T \Sigma_i^{-1} S_i (\hat{\eta}_i - \lambda_0). \tag{11}$$

Thus, $\widehat{\alpha}_i$ is the h-dimensional projection of $\mathbf{Y}_i$ onto the mean space after centering. Notice that in this example $\hat{\eta}_2$ is closest to μ_2 in Euclidean distance but after projection onto the mean space it is closest to μ_1 and will be assigned to cluster 1.

Theorem 1 shows that there is a direct relationship between the posterior probability of the ith curve belonging to the kth cluster and the squared distance between $\widehat{\alpha}_i$ and α_k.

Theorem 1 *For $\mathbf{Y}_i$ drawn from the FCM*

$$\log P(z_{ik} = 1|\mathbf{Y}_i) = C(\mathbf{Y}_i) + \log(\pi_k) - \frac{1}{2}||\widehat{\alpha}_i - \alpha_k||^2_{Cov(\widehat{\alpha}_i)^{-1}}$$

where $C(\mathbf{Y}_i)$ is a constant with respect to k and

$$Cov(\widehat{\alpha}_i) = \left(\Lambda^T {S_i}^T \Sigma_i^{-1} S_i \Lambda\right)^{-1}. \tag{12}$$

Hence,

$$\arg\max_k P(z_{ik} = 1|\mathbf{Y}_i) = \arg\min_k \left(||\widehat{\alpha}_i - \alpha_k||^2_{Cov(\widehat{\alpha}_i)^{-1}} - 2\log\pi_k\right) \tag{13}$$

A proof of this result can be found in Appendix B. From (13) and Bayes rule we note that cluster assignments based on the $\widehat{\alpha}_i$'s will minimize the expected number of misassignments. Thus no clustering information is lost through the projection of $\mathbf{Y}_i$ onto the lower dimensional space. We call the $\widehat{\alpha}_i$'s functional linear discriminants because they are exact analogues of the low-dimensional representations used to visualize data in linear discriminant analysis (LDA). In the finite-dimensional setting the linear discriminants all have identity covariance so separation between classes can be assessed visually using the Euclidean distance metric. In the functional clustering setting $Cov(\widehat{\alpha}_i)$ is given by (12). When all curves are measured at the same time points constraint (8) will guarantee $Cov(\widehat{\alpha}_i) = I$ for all i, again allowing the Euclidean metric to be used. When curves are measured at different time points it is not possible to impose a constraint that will simultaneously cause $Cov(\widehat{\alpha}_i) = I$ for all i. However, when the cluster means lie in a one dimensional subspace ($h = 1$), assuming equal priors, (13) simplifies to

$$\arg\min_k \frac{1}{Var(\hat{\alpha}_i)}(\widehat{\alpha}_i - \alpha_k)^2 = \arg\min_k (\widehat{\alpha}_i - \alpha_k)^2,$$

which yields the same assignments as if the $\widehat{\alpha}_i$'s all had the same variance. In this situation it is useful to

plot the functional linear discriminants versus their standard deviations to indicate not only to which cluster each point belongs but also the level of accuracy with which it has been observed. Note that for a two cluster model h must be 1. However, it will often be reasonable to assume the means lie approximately in one dimension even when there are more than two clusters.

Linear discriminant plots have other useful features. Note that the functional linear discriminant for a curve observed over the entire grid of time points used to form S will have identity covariance. Thus, the Euclidean distance between the α_k's gives the number of standard deviations separating the cluster means for a fully observed curve. The degree to which the variance for an individual curve is greater than 1 indicates how much discriminatory power has been lost due to taking observations at only a subset of time points. This has implications for experimental design in that it suggests how to achieve minimum variance, and hence optimal cluster separation, with a fixed number of time points. For instance the cluster means in Figure 2(a) are 2.4 standard deviations apart, indicating that the groups can be fairly well separated if curves are measured at all time points. The overlap between the two groups is due to the extreme sparsity of sampling, resulting in the $\widehat{\alpha}_i$'s having standard deviations up to 2.05.

Plots for the membranous nephropathy data, given in Figure 3, provide an example in which the differing covariances of the $\widehat{\alpha}_i$'s must be taken into account more carefully. Nephrologists' experiences suggest that patients with this disease fall into three groups, either faring well, deteriorating gradually or collapsing quickly. Hence we fit a three cluster model whose mean curves are shown in Figure 3(a). The issue of identifying the optimal number of clusters is addressed more formally in Section 4.1. With three clusters the means must lie in a plane. Figure 3(b) shows a two-dimensional linear discriminant plot with solid circles indicating cluster centers. To circumvent the problem caused by the unequal covariances, we use different symbols for members of different clusters. Note that while most patients fall in the cluster corresponding to their closest cluster in Euclidean distance, there are several that do not. In this example the cluster centers lie essentially on a straight line so it is sufficient to fit a one-dimensional model ($h = 1$). The corresponding plots are shown in Figure 3(c) and (d). The basic shapes of the mean curves are reassuringly similar to those in 3(a), but are physiologically more sensible in the right tail. Figure 3(d) plots one dimensional $\widehat{\alpha}_i$'s versus their standard deviations. We see that the cluster on the right is very tight while the other two are not as well separated. Figures 3(e) and (f) show respectively the overall mean curve, $\mathbf{s}(t)^T\lambda_0$ and the function $\mathbf{s}(t)^T\Lambda$. The latter, when multiplied by α_k, gives the distance between $\mu_k(t)$ and the overall mean curve. From Figure 3(e) we see that on average the patients showed a decline in renal function. The primary distinction lies in the speed of the deterioration. For example, the fact that Figure 3(f) shows a sharp decline in the first two years indicates that patients in the third cluster, which has a highly positive α_3, experience a much sharper initial drop than average. In fact all patients in cluster 3 eventually required dialysis.

3.2 Discriminant functions

In the membranous nephropathy example we saw that plots like Figure 3(f) provide useful information about the traits that distinguish one cluster from another. In this section we more formally present a set of curves that identify the dimensions, or equivalently time points, of maximum discrimination between clusters and trace their connection to classical discriminant functions. Intuitively, the dimensions with largest average separation relative to their variability will provide the greatest discrimination. Average separation can be determined by examining $S\Lambda$ while variability is calculated using the covariance matrix, $\Sigma = S\Gamma S^T + \sigma^2 I$. These two quantities can work in opposite directions, making it difficult to identify the regions of greatest discrimination. Consider, for example, Figure 4 which illustrates the covariance and correlation functions for the bone mineral density data. From Figure 4(a) it is clear that the relationship between a person's bone mineral density before and after puberty is weak but the measurements after puberty are strongly correlated with each other. Figure 4(b) has a sharp peak in the early puberty years corresponding to the period of

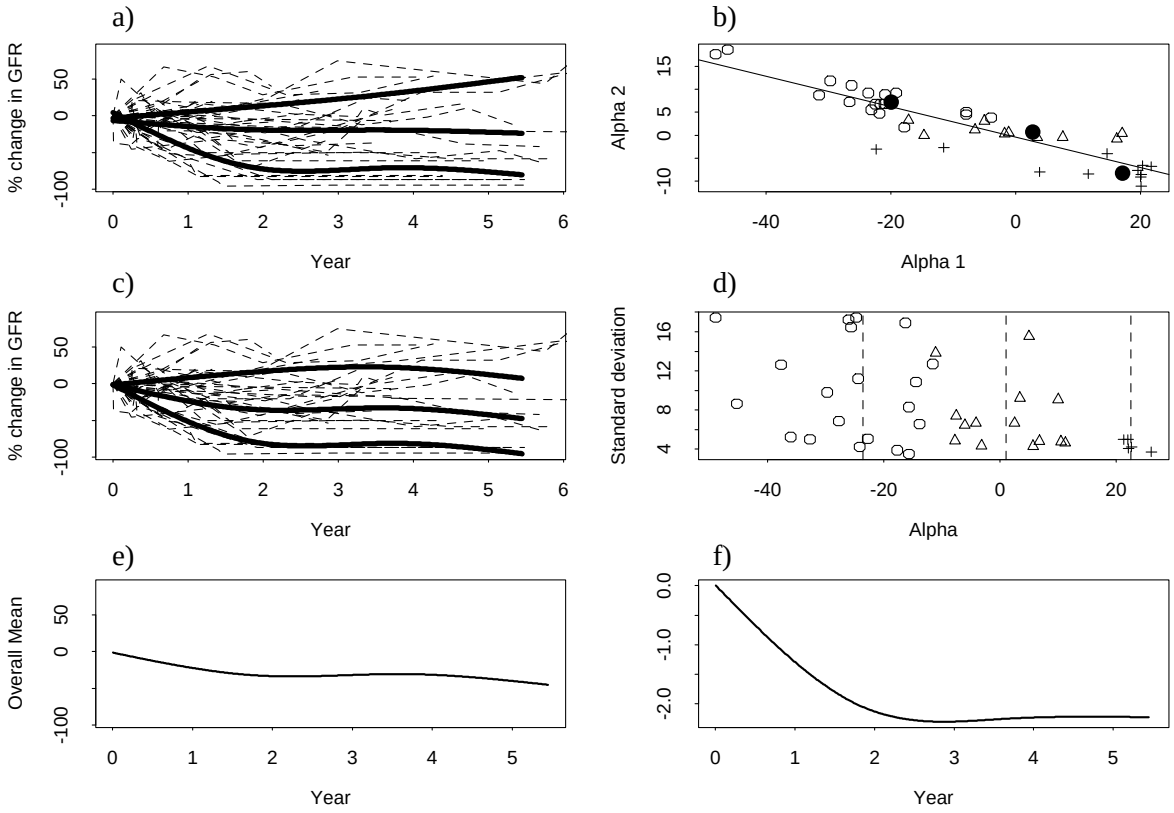

Figure 3: *Assessment plots for the membranous nephropathy data. The cluster mean curves and linear discriminant plots for a fit with $h = 2$ are shown in (a) and (b). The equivalent plots for a fit with $h = 1$ are given in (c) and (d). Finally, (e) shows the overall mean curve and (f) the characteristic pattern of deviations about the overall mean.*

greatest variability. However, this is also the period of greatest distance between the cluster mean curves.

The dimensions of maximum discrimination must also be the ones that are most important in determining cluster assignment. When observations are made at all time points, the spline basis matrix is S, and equations (11) and (13) imply that curves should be assigned based solely on the Euclidean distance between $\widehat{\alpha} = \Lambda^T S^T \Sigma^{-1}(\mathbf{Y} - S\lambda_0)$ and the α_k's. Thus

$$\Lambda^T S^T \Sigma^{-1} \tag{14}$$

gives the optimal weights to apply to each dimension for determining cluster membership. Dimensions with low weights contain little information about cluster membership and therefore do little to distinguish among groups, while dimensions with large weights have high discriminatory power. Notice that this set of weights fits with the intuitive notion that dimensions with high discrimination should have large average separation, $S\Lambda$, relative to their variability, Σ.

When the α_k's are one dimensional, $\Lambda^T S^T \Sigma^{-1}$ is a vector and the weights can be plotted as a single curve, as illustrated by Figure 5 for the bone density and membranous nephropathy examples. For the bone mineral data the highest absolute weights occur in the puberty years, confirming our earlier interpretation from the linear discriminant plot, Figure 2(a). For the membranous nephropathy data most of the discrimination between clusters occurs in the early and late stages of disease. The difference between patients in the later time periods is not surprising. However, the discriminatory power of the early periods is encouraging since one of the primary goals of this study was to predict disease progression based on entry characteristics.

For a two cluster model the vector $\Lambda^T S^T \Sigma^{-1}$ is equivalent to the classical discriminant function for a

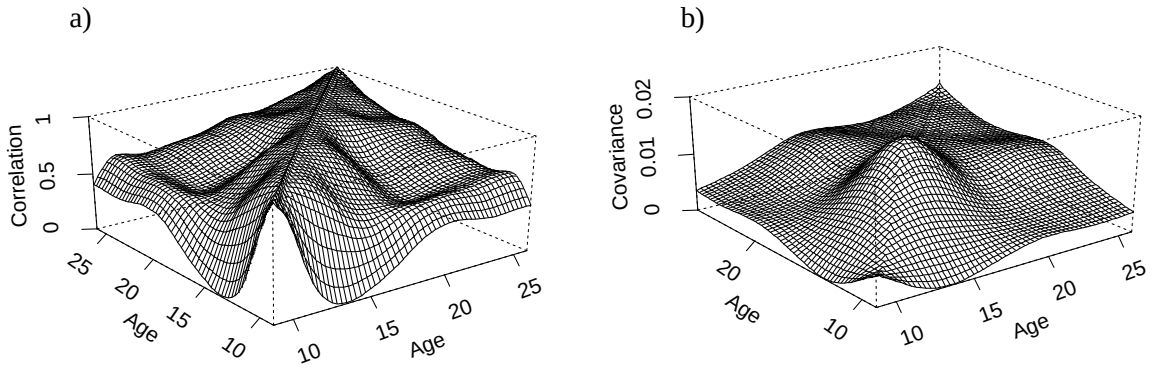

Figure 4: *Estimated (a) correlation and (b) covariance of $g_i(t_1)$ with $g_i(t_2)$.*

Gaussian mixture. Recall that for a Gaussian mixture the weights placed on each dimension by the discriminant function are

$$(\mu_1 - \mu_2)^T \Sigma^{-1} \tag{15}$$

where Σ is the within group covariance matrix. In the FCM the kth cluster mean is $S(\lambda_0 + \Lambda\alpha_k)$ so (15) becomes

$$(S\Lambda\alpha_1 - S\Lambda\alpha_2)^T \Sigma^{-1} = (\alpha_1 - \alpha_2)^T \Lambda^T S^T \Sigma^{-1} \tag{16}$$

which is equal to (14) up to the multiplicative term $\alpha_1 - \alpha_2$. In a two cluster model the α_k's are scalars and so do not effect the relative weight placed on each dimension. In fact as long as the α_k's are one dimensional (14) and (16) will give the same relative weighting for any pair of clusters. In general, (14) will produce h distinct sets of weights where h is the dimension of the α_k's.

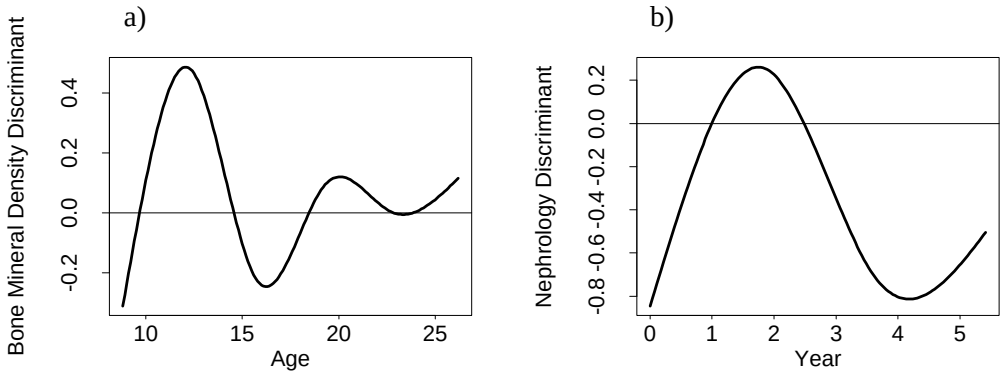

Figure 5: *Discriminant curves for (a) the bone mineral density data and (b) the membranous nephropathy data with $h = 1$.*

3.3 Curve estimation

Another major advantage of the functional clustering procedure is that it can accurately predict unobserved portions of $g_i(t)$, the true curve for the ith individual, even in situations where the regularization and filtering

methods break down. When using a basis representation a natural estimate for $g_i(t)$ is $\hat{g}_i(t) = \mathbf{s}(t)^T \hat{\eta}_i$, where $\hat{\eta}_i$ is a prediction for η_i. The filtering method takes $\hat{\eta}_{F_i} = (S_i^T S_i)^{-1} S_i Y_i$, provided that the inverse exists. Theorem 2 gives the optimal procedure for computing $\hat{\eta}_i$ under the FCM:

Theorem 2 *Under the FCM (6) the prediction of $g_i(t)$ with minimum mean squared error, $E_\eta(\hat{g}_i(t) - g_i(t))^2$, is $\hat{g}_i(t) = \mathbf{s}(t)^T E(\eta_i|\mathbf{Y}_i)$.*

A proof of this result can be found in Appendix C. When the mixture likelihood is used the $\mathbf{z}_i$'s are treated as missing data, yielding

$$\hat{\eta}_{M_i} = E(\eta_i|\mathbf{Y}_i) = \lambda_0 + \Lambda \sum_{k=1}^{G} \alpha_k \pi_{k|i} + \left(\sigma^2 \Gamma^{-1} + S_i^T S_i\right)^{-1} S_i^T \left(\mathbf{Y}_i - S_i \left(\lambda_0 + \Lambda \sum_{k=1}^{G} \alpha_k \pi_{k|i}\right)\right) \tag{17}$$

where

$$\pi_{k|i} = P(z_{ik} = 1|\mathbf{Y}_i) = \frac{f(y|z_{ik} = 1)\pi_k}{\sum_{j=1}^{G} f(y|z_{ij} = 1)\pi_j} \tag{18}$$

and $f(y|z_{ik} = 1)$ is given by (9). Alternatively, under the classification likelihood the $\mathbf{z}_i$'s are treated as parameters, yielding

$$\hat{\eta}_{C_i} = E(\eta_i|\mathbf{Y}_i) = \lambda_0 + \Lambda\alpha_{\mathbf{z}_i} + \left(\sigma^2 \Gamma^{-1} + S_i^T S_i\right)^{-1} S_i^T \left(\mathbf{Y}_i - S_i \left(\lambda_0 + \Lambda\alpha_{\mathbf{z}_i}\right)\right) \tag{19}$$

where $\mathbf{z}_i = \arg\max_k f(y|z_{ik} = 1)$. In general, the functional mixture approach produces significant improvements over the filtering method when σ^2 is very large, the components of Γ are very small, or $S_i^T S_i$ is close to singular. In fact, when $S_i^T S_i$ is singular the filtering approach breaks down completely while the functional clustering method can still produce reliable predictions.

It is also important to obtain a measure of the uncertainty in our predictions of the individual curves. We achieve this through pointwise confidence and prediction intervals. For example, using the mixture likelihood, the distribution of η_i given $\mathbf{Y}_i$ is a mixture of normals whose kth component has mixture probability $\pi_{k|i}$, and mean and covariance

$$E(\eta_i|\mathbf{Y}_i, z_{ik} = 1) = \lambda_0 + \Lambda\alpha_k + \left(\sigma^2 \Gamma^{-1} + S_i^T S_i\right)^{-1} S_i^T \left(\mathbf{Y}_i - S_i \left(\lambda_0 + \Lambda\alpha_k\right)\right),$$

and $Cov(\eta_i|\mathbf{Y}_i, z_{ik} = 1) = \left(\Gamma^{-1} + S_i^T S_i/\sigma^2\right)^{-1}$. Hence, conditional on $\mathbf{z}_i$ and $\mathbf{Y}_i$

$$g_i(t) \sim N\left[\mathbf{s}(t)^T E(\eta_i|\mathbf{Y}_i, z_{ik} = 1), \mathbf{s}(t)^T Cov(\eta_i|\mathbf{Y}_i, z_{ik} = 1)\mathbf{s}(t)\right].$$

Thus, if the cluster membership $\mathbf{z}_i$ were known then

$$\{c_{ik1}^\tau(t), c_{ik2}^\tau(t)\} = \mathbf{s}(t)^T E(\eta_i|\mathbf{Y}_i, z_{ik} = 1) \pm \Phi((1+\tau)/2)^{-1} \sqrt{\mathbf{s}(t)^T Cov(\eta_i|\mathbf{Y}_i, z_{ik} = 1)\mathbf{s}(t)},$$

where Φ is the standard normal cdf, would be a τ pointwise confidence interval for $g_i(t)$. Since the cluster memberships are unknown, a highly conservative approach would be to use the interval

$$\left\{\min_k c_{ik1}^\tau(t), \max_k c_{ik2}^\tau(t)\right\}. \tag{20}$$

The following two step procedure is superior. First, find the smallest collection of clusters with total probability at least τ_1 of having generated the curve in question. Second, construct τ_2 confidence intervals for $g_i(t)$ conditional on membership in each of these clusters and take the pointwise extremes. If the clusters

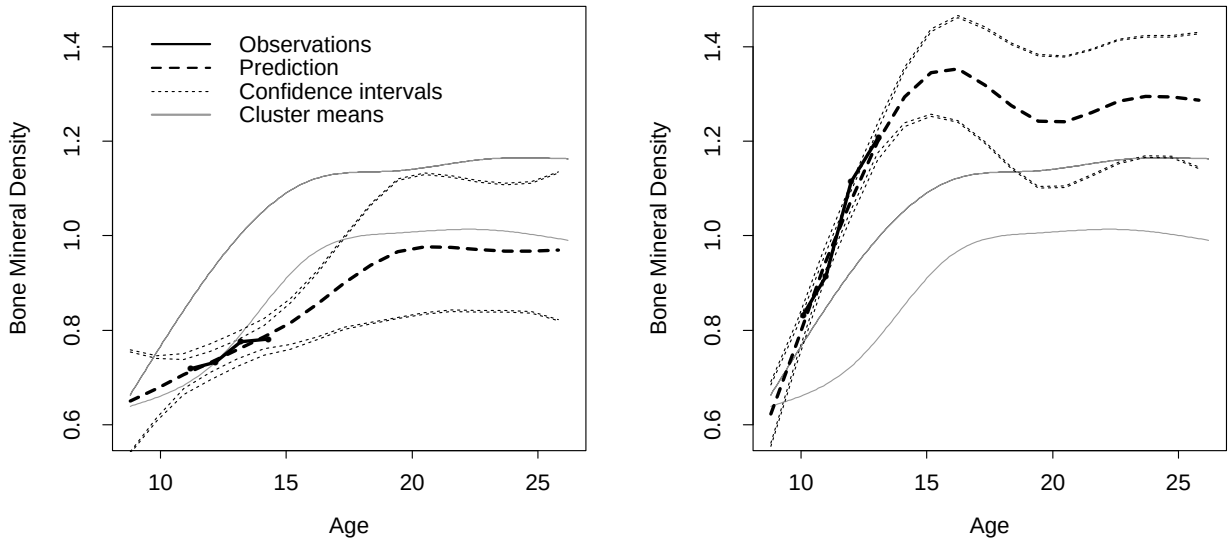

Figure 6: *Curve estimates, confidence intervals, and prediction intervals for two subjects from the bone density study.*

are ordered from largest to smallest $\pi_{k|i}$, this procedure produces a $\tau = \tau_1\tau_2$ pointwise confidence interval with bounds

$$\left\{ \min_{k:\sum_{j=1}^{k-1}\pi_{j|i}<\tau_1} c_{ik1}^{\tau_2}(t), \max_{k:\sum_{j=1}^{k-1}\pi_{j|i}<\tau_1} c_{ik2}^{\tau_2}(t) \right\}.$$

This interval will, in general, be narrower than that given by (20) because it ignores the clusters with low posterior probability which are also the ones furthest from the predicted curve. Figure 6 illustrates this approach for two subjects from the bone density study. For each plot, the two solid grey lines give the cluster mean curves, the curve fragment gives the observed values for a single individual, and the dashed line gives the corresponding prediction. The dotted lines represent 95% confidence and prediction intervals. Note that the confidence interval provides bounds for the underlying function $g_i(t)$ while the prediction interval bounds the observed value of $g_i(t)$. As usual, the prediction interval is produced by adding σ^2 to the variance used in the confidence interval.

4 Model selection

In Appendix A we outline an EM procedure for fitting the functional clustering model. However, there are several model selection questions that are worth discussing in greater detail. In particular one must choose how many clusters to fit, the number of knots to use in the spline basis and the dimension of the mean space.

4.1 Choice of number of clusters

Most clustering procedures require one to choose the number of groups prior to fitting. This is one of the most difficult problems in cluster analysis. A popular choice in model-based clustering is to use Bayes factors (Kass and Raftery, 1995), which are difficult to calculate exactly but can be approximated using BIC (Schwarz, 1978). One disadvantage of this method is that it requires fitting the model for each potential

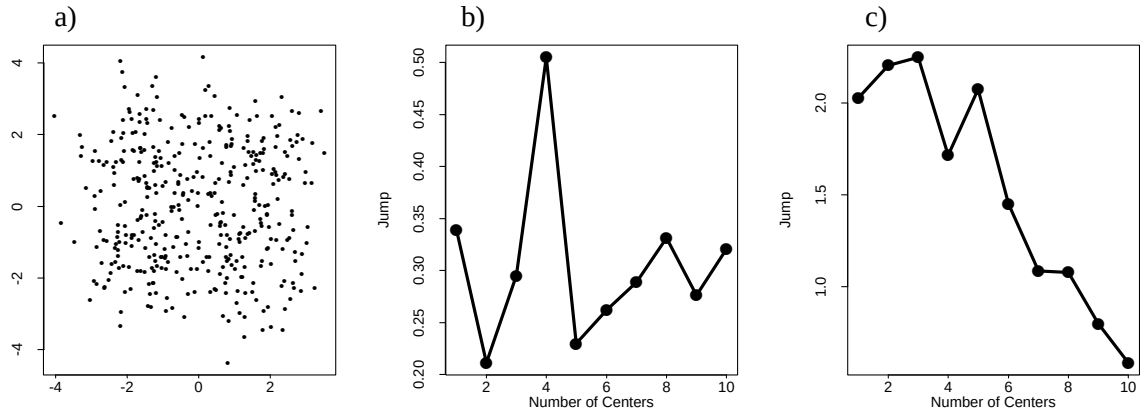

Figure 7: *(a) A mixture of four highly overlapping Gaussian clusters and (b) the associated jump plot. (c) The jump plot for the bone mineral density data.*

number of clusters. We utilize an alternative approach suggested by Sugar and James (2003). Their method is based on the "distortion function",

$$d_K = \frac{1}{p} \min_{\mathbf{c}_1,\ldots,\mathbf{c}_K} E(\eta_i - \mathbf{c}_{\mathbf{z}_i})^T \Gamma^{-1} (\eta_i - \mathbf{c}_{\mathbf{z}_i}) \tag{21}$$

where the η_i's are the spline coefficients in the functional clustering model. The distortion, d_K, is the average Mahalanobus distance between each η_i and its closest cluster center $\mathbf{c}_{\mathbf{z}_i}$. Consider Figure 7(a) which plots a mixture of four highly overlapping Gaussian clusters. From visual inspection it is not clear that the data consist of four clusters. However, Figure 7(b), which plots the jump $d_K^{-1} - d_{K-1}^{-1}$ from $K = 1$ to 10, shows a clear spike at $K = 4$. Sugar and James (2003) show, both theoretically and empirically, that for a large class of mixture distributions the largest jump will always correspond to the correct number of mixture components. Their key theorem is summarized below.

Theorem 3 *Suppose that the distribution of the η_i's is a mixture of G p-dimensional clusters with equal priors. Furthermore, assume that the clusters are identically distributed with covariance Γ_p and finite fourth moments in each dimension. Then, under suitable conditions, there exists a set of real valued numbers $Y > 0$ such that the jump $d_K^{-Y} - d_{K-1}^{-Y}$ will be maximized when $K = G$.*

The conditions under which this result will hold relate to the separation between cluster means relative to the entropy of each cluster. Further details as well as a proof of Theorem 3 can be found in Sugar and James (2003). It is clear from the simulation example that this approach can identify the correct number of mixture components even for highly overlapping clusters. Figure 7(c) shows the corresponding jump plot for the bone mineral density data. It appears that $1, 2, 3$ or 5 are all possible choices for the number of clusters and each has a reasonable interpretation. One would simply indicate that there is no strong clustering in the data. Two clusters correspond to a breakdown along gender lines. This data consisted of four ethnic groups, black, asian, white and hispanic, but the white and hispanic groups were indistinguishable (James and Hastie, 2001) so a three cluster fit may well correspond to the different ethnicities. Finally, the five cluster fit could indicate clustering into different gender-ethnicity combinations. Since the number of clusters is not clear cut, which of these four possible choices produced the largest jump depended on the exact choice of Y. Theorem 3 does not specify the optimal choice of Y. Sugar and James (2003) give results suggesting that one should set Y equal to half of the "effective" number of dimensions in the data. For the bone mineral example, the η_i's varied mostly along one dimension with a small amount of variability in a second dimension. Hence we

estimated the effective dimension to be 1.5 and set $Y = 0.75$. A similar approach applied to the membranous nephropathy data provided no statistical evidence of more than two distinct clusters. Thus the clinician's supposition of three clusters is not supported by the data. However, from Figure 3(d) there is evidence of at least two clusters.

In practice we estimate d_K by first predicting the η_i's via (17) and then using the k-means algorithm to approximate the distortion. Sugar and James (2003) note that the jump approach produces reasonable answers when one substitutes $\Gamma = I$ into (21) so we have followed this convention. The jump method has the advantage that it only requires the functional clustering procedure to be fit once. Since the k-means algorithm is significantly faster than the EM-based functional clustering procedure this produces a significant reduction in computation.

4.2 Other model selection problems

Another important issue is the selection of the spline basis. Most procedures use equally spaced knots which reduces the problem to one of selecting the correct number. One natural approach is to take the dimension of the basis, p, to corresponding to the largest cross-validated likelihood (James *et al.*, 2000). This works well but is generally computationally expensive. An alternative approach is to calculate the likelihood once for each value of p and apply a penalty term involving the number of parameters fit to the data. AIC and BIC are two such methods that have worked well on models of this type (Rice and Wu, 2001). In practice the final clustering appears to be fairly robust to any reasonable number of knots.

Finally one must choose h, the dimension of the cluster mean space. For the data sets illustrated in this article, the choice of h was not a serious problem because only a small number of clusters were involved. Recall that setting $h = G-1$ results in no restriction on the mean space. For the membranous nephropathy data with $G = 3$ clusters it was clear upon fitting the model with $h = 2$ that the α_k's lay approximately on a line implying that one should set $h = 1$. This approach can be applied in general by fitting the model with $h = G-1$, calculating the α_k's, testing whether the cluster centers appear to lie in a lower dimensional plane and then refiting the model with h set to this new dimension. Methods such as principal components analysis can be used to determine whether the means lie in a lower dimensional space.

5 Extensions of the functional clustering model

5.1 Incorporating multiple curves and covariates

In this section we extend our model to allow multiple functional variables per subject as well as finite dimensional covariates with possibly missing values. Let $\mathbf{Y}_{ij}$ represent the vector of observations of the jth curve for the ith individual at times $t_{ij1}, \ldots t_{ijn_{ij}}$ and let $\mathbf{X}_i$ be the vector of finite dimensional covariates. The functional clustering model of Section 2.2 generalizes to

$$\mathbf{Y}_{ij} = S_{ij}\eta_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, I\sigma_j^2), \quad j = 1, \ldots, J$$

where $\eta_{ij} = \mu_{\mathbf{z}_i j} + \gamma_{ij}$ and we allow a different variance for each curve's error vector. To model the finite dimensional covariates, $\mathbf{X}_i$, the spline basis is replaced by the identity, yielding

$$\mathbf{X}_i = I_{ix}\eta_{ix} + \varepsilon_{ix}, \quad \varepsilon_{ix} \sim N(0, I\sigma_x^2)$$

where $\eta_{ix} = \mu_{\mathbf{z}_i x} + \gamma_{ix}$. As before, we assume that $\gamma_i = (\gamma_{i1}, \ldots, \gamma_{iJ}, \gamma_{ix}) \sim N(0, \Gamma)$ and $\mu_k = (\mu_{k1}, \ldots, \mu_{kJ}, \mu_{kx}) = \lambda_0 + \Lambda\alpha_k$. Note that when we let $\mathbf{Y}_i = (\mathbf{Y}_{i1}, \ldots \mathbf{Y}_{iJ}, \mathbf{X}_i)$, $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iJ}, \varepsilon_{ix})$ and S_i be the block diagonal

matrix formed by $S_{i1},\ldots S_{iJ}$ and I_{ix}, then

$$\mathbf{Y}_i = S_i(\lambda_0 + \Lambda\alpha_{\mathbf{z}_i} + \gamma_i) + \varepsilon_i, \quad \varepsilon_i \sim N(\mathbf{0}, R), \quad \gamma_i \sim N(\mathbf{0}, \Gamma), \tag{22}$$

which is identical to the formulation of the standard functional clustering model. Hence (22) can be fit using the same EM procedure. This model is extremely flexible and can handle an arbitrary number of functional variables. In addition, covariates with missing observations can easily be included. For example, if the ith individual is missing the jth observation one would simply remove the jth row of I_{ix} before fitting the model. Despite the added complexity it is still possible to represent the individual data points in a low-dimensional subspace by taking advantage of the projection onto the space of cluster mean coefficients. Furthermore, standard high-dimensional clustering problems with missing data can simply be viewed as a special case of our model in which there are no functional covariates.

5.2 Alternative covariance structures

To this point we have used a common covariance matrix, Γ, for all clusters and have taken R, the covariance of the ε_i's, to be a multiple of the identity matrix. These assumptions may be inappropriate for some data sets. For example, the differing variability in the $\widehat{\alpha}_i$'s of Figure 3(d), suggest that, for the membranous nephropathy data, a model that allows an alternative covariance structure may be more appropriate.

Any covariance matrix, Γ_k, may be reparameterized as $\Gamma_k = \Theta_k D_k \Theta_k^T$ where Θ_k is a matrix whose columns consist of the eigenvectors of Γ_k and D_k is a diagonal matrix whose elements are the eigenvalues. The standard FCM forces Θ_k and D_k to be the same for each cluster, in analogy with linear discriminant analysis (LDA). Allowing both Θ_k and D_k to vary over k, as in quadratic discriminant analysis (QDA), gives greater flexibility but, like QDA, this model tends to perform poorly unless all groups have a large number of observations (Wald and Kronmal, 1977). Numerous compromises between the LDA and QDA frameworks have been proposed. The simplest of these take Γ_k to be a multiple of the identity i.e. $\Gamma_k = \delta I$ or $\Gamma_k = \delta_k I$ (Ward, 1963; Banfield and Raftery, 1993; Celeux and Govaert, 1995). Although these models do not require large amounts of data, they are often overly restrictive because they assume independence between coordinates. As a compromise various authors have suggested classes of models in which Θ_k is allowed to vary over k but D_k remains fixed. Two examples are $\Gamma_k = \delta\Theta_k D\Theta_k^T$ and $\Gamma_k = \delta_k\Theta_k D\Theta_k^T$ (Murtagh and Raftery, 1984; Banfield and Raftery, 1993; Celeux and Govaert, 1995). With these structures, each cluster has the same shape but variable orientation. The use of shrinkage estimators is another strategy that has been highly successful in poorly-posed inverse problems (Titterington, 1985; O'Sullivan, 1985). For example, regularized discriminant analysis (Friedman, 1989) works by simultaneously shrinking the estimated covariance matrix towards both the identity matrix and towards the common sample covariance matrix.

Many covariance structures are also possible for the ε_i's. For example, one may believe that there is a linear relationship between time and variance and let R be diagonal with $var(\varepsilon_i(t)) = \beta_0 + \beta_1 t$. Alternatively, since the data being modeled are often time dependent, a correlation structure between the error terms may be appropriate. For example, one may choose to assume a constant correlation between adjacent error terms from the same individual i.e. $cor(\varepsilon_i(t), \varepsilon_i(t+1)) = \rho$. Such a structure seems more plausible for equally spaced time points and so may not be appropriate in settings such as the bone mineral density study. All of these approaches for modeling the covariance structures can easily be incorporated into the functional clustering procedure.

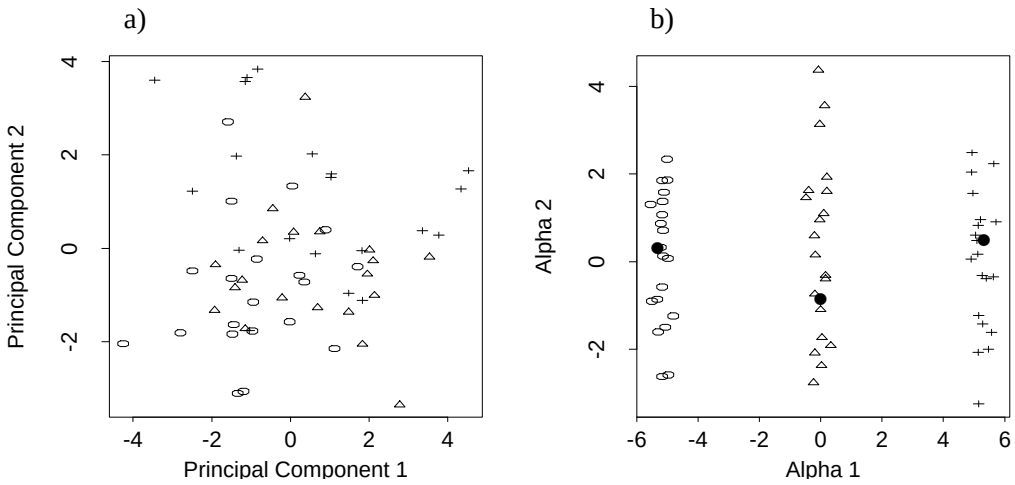

Figure 8: *Two dimensional representations of an eight dimensional data set using (a) principal components and (b) the functional clustering procedure.*

6 Discussion

The main goal of this paper was to develop a general procedure for clustering functional data. Our approach performs particularly well compared to other methods for sparsely sampled curves. However, the models presented in Sections 2 and 5 provide an extremely flexible framework that can easily be adapted to a variety of situations. For example we noted in Section 5.1 that replacing the spline basis matrix with the identity matrix yields a method for clustering finite dimensional data that can handle missing observations. Figure 8 provides a comparison of this approach with a more standard procedure. In high dimensional settings it is common to first reduce the number of dimensions, for instance by using principal components analysis (PCA), and then to cluster in the lower dimensional space. The danger is that one may inadvertently lose any discrimination between clusters in the process. Figure 8(a) provides just such an example. Here we have plotted the first two principal components of an eight dimensional simulated data set. The data were generated by sampling twenty observations from each of three clusters. The clusters had different means in the first dimension but were otherwise identically distributed. The second and third dimensions had a large degree of variability while the remaining five dimensions had comparatively little. When PCA is run the first two components explain over 90% of the variability. Generally this would be considered an adequate representation of the data. However, when these two components are plotted in Figure 8(a), with a different symbol for each cluster, it is clear that all discriminatory power has been lost. Alternatively, when the FCM is fit to this data and a two-dimensional linear discriminant plot is produced as in Figure 8(b) perfect separation is achieved. We did not incorporate missing values in this example because principal components can not readily be applied to such data. However, the FCM approach would have coped with ease by simply leaving out the appropriate rows of the identity matrix I_{ix}. As with most EM procedures, initialization of the algorithm is an important consideration. For the data of Figure 8 we found that the FCM procedure worked best when Γ was initialized with high variance in the second and third dimensions and low variance in other dimensions. This is not an unreasonable starting point since an examination of the sample covariance matrix for the raw data reveals that most of the variability is in these dimensions.

Acknowledgments

We would like to thank Brian Myers and members of his lab at the Stanford University School of Medicine for providing us with the Membranous Nephropathy data and the referees and editors for useful comments

and suggestions.

A The fitting algorithm

We first outline the procedure for fitting the mixture likelihood (2). The classification likelihood fitting procedure follows with only minor modifications. The standard approach to fitting a mixture likelihood is to treat the unknown cluster memberships $\mathbf{z}_i$ as missing data and to use the EM algorithm. Note that since the $\mathbf{z}_i$'s and γ_i's are assumed independent of one another the complete data distribution factors as $f(\mathbf{Y},\mathbf{z},\gamma) = f(\mathbf{Y}|\mathbf{z},\gamma)f(\mathbf{z})f(\gamma)$. Given that the $\mathbf{z}_i$'s are multinomial(π_k), the γ_i's are $N(0,\Gamma)$ and the $\mathbf{Y}_i$'s are conditional $N[S_i(\lambda_0+\Lambda\alpha_k+\gamma_i),\sigma^2]$ the complete data log likelihood, up to additive constants, is

$$\begin{aligned} l(\pi_k,\Gamma,\sigma^2,\lambda_0,\Lambda,\alpha_i) &= \sum_{i=i}^{n}\sum_{k=1}^{G} z_{ik}\log(\pi_k) && (23)\\ &- \frac{1}{2}\sum_{i=1}^{n}\left[\log|\Gamma| + \gamma_i^T\Gamma^{-1}\gamma_i\right] && (24)\\ &- \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{G} z_{ik}\left[n_i\log\sigma^2 + \frac{1}{\sigma^2}\left|\left|\mathbf{Y}_i - S_i\left(\lambda_0+\Lambda\alpha_k+\gamma_i\right)\right|\right|^2\right] && (25)\end{aligned}$$

The EM algorithm consists of iteratively maximizing the expected values of (23), (24) and (25) given $\mathbf{Y}_i$ and the current parameter estimates. Since all three parts involve separate parameters they can be maximized independently of each other. The expected value of (23) is maximized by setting

$$\pi_k = \frac{1}{n}\sum_{i=1}^{n}\pi_{k|i} \tag{26}$$

where $\pi_{k|i}$ is given by (18). Next the expected value of (24) is maximized by setting

$$\Gamma = \frac{1}{n}\sum_{i=1}^{n} E\left[\gamma_i\gamma_i^T\middle|\mathbf{Y}_i\right] = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{G}\pi_{k|i}E\left[\gamma_i\gamma_i^T\middle|\mathbf{Y}_i, z_{ik}=1\right] \tag{27}$$

which can be calculated using the fact that

$$\gamma_i|\mathbf{Y}_i, z_{ik}=1 \sim N\left((\sigma^2\Gamma^{-1}+S_i^TS_i)^{-1}S_i^T(\mathbf{Y}_i - S_i\lambda_0 - S_i\Lambda\alpha_i), (\Gamma^{-1}+S_i^TS_i/\sigma^2)^{-1}\right). \tag{28}$$

In the final step we maximize the expected value of (25). This involves an iterative procedure where λ_0 then α_k and finally the columns of Λ are repeatedly optimized while holding all other parameters fixed. First we set

$$\lambda_0 = \left(\sum_{i=1}^{n}S_i^TS_i\right)^{-1}\sum_{i=1}^{n}S_i^T\left(\mathbf{Y}_i - \sum_{k=1}^{G}\pi_{k|i}S_i\left(\Lambda\alpha_k+\hat{\gamma}_{ik}\right)\right) \tag{29}$$

where $\hat{\gamma}_{ik} = E\left[\gamma_i\middle|z_{ik}=1,\mathbf{Y}_i\right]$ which is calculated using (28). Next, the α_k's are calculated using

$$\alpha_k = \left(\sum_{i=1}^{n}\pi_{k|i}\Lambda^TS_i^TS_i\Lambda\right)^{-1}\sum_{i=1}^{n}\pi_{k|i}\Lambda^TS_i^T\left(\mathbf{Y}_i - S_i\lambda_0 - S_i\hat{\gamma}_{ik}\right). \tag{30}$$

Finally, each column of Λ is optimized holding all others fixed using

$$\lambda_m = \left(\sum_{i=1}^{n} \sum_{k=1}^{G} \pi_{k|i} \alpha_{km}^2 S_i^T S_i \right)^{-1} \sum_{i=1}^{n} \sum_{k=1}^{G} \pi_{k|i} \alpha_{km} S_i^T \left(\overline{\mathbf{Y}}_i - \sum_{l \neq m} \alpha_{kl} S_i \lambda_l - S_i \hat{\gamma}_{ik} \right) \tag{31}$$

where λ_m is the mth column of Λ, α_{km} is the mth component of α_k and $\overline{\mathbf{Y}}_i = \mathbf{Y}_i - S_i \lambda_0$. We iterate through (29), (30) and (31) until all parameters have converged which typically occurs rapidly. The final step is to set

$$\begin{aligned} \sigma^2 &= \frac{1}{\sum_{i=1}^{n} n_i} \sum_{i=1}^{n} \sum_{k=1}^{G} \pi_{k|i} E\left[(\overline{\mathbf{Y}}_i - S_i \Lambda \alpha_k - S_i \gamma_i)^T (\overline{\mathbf{Y}}_i - S_i \Lambda \alpha_k - S_i \gamma_i) | \mathbf{Y}_i, z_{ik} = 1 \right] \qquad (32) \\ &= \frac{1}{\sum_{i=1}^{n} n_i} \sum_{i=1}^{n} \sum_{k=1}^{G} \pi_{k|i} \Big[(\overline{\mathbf{Y}}_i - S_i \Lambda \alpha_k - S_i \hat{\gamma}_{ik})^T (\overline{\mathbf{Y}}_i - S_i \Lambda \alpha_k - S_i \hat{\gamma}_{ik}) \\ &\quad + tr\left(S_i Cov[\gamma_i | \mathbf{Y}_i, z_{ik} = 1] S_i^T \right) \Big] \end{aligned}$$

The algorithm iterates through (26),(27),(29), (30), (31) and (32) until all the parameters have converged.

When fitting the classification likelihood the only alteration to this procedure is that, for each i, $\pi_{k|i}$ is set to 1 if k equals

$$\arg\min_{k^*} ||\mathbf{Y}_i - S_i(\lambda_0 + \Lambda \alpha_{k^*})||_{(\sigma^2 I + S_i \Gamma S_i^T)^{-1}}$$

and 0 otherwise.

B Proof of Theorem 1

Recall that

$$\mathbf{Y}_i | z_{ik} = 1 \sim N\left[S_i \mu_k, \Sigma_i \right]$$

where $\mu_k = \lambda_0 + \Lambda \alpha_k$ and $\Sigma_i = \sigma^2 I + S_i^T \Gamma S_i$. Hence, using Bayes rule,

$$\log P(z_{ik} = 1 | \mathbf{Y}_i) = \log(\pi_k) - \frac{1}{2} ||\mathbf{Y}_i - S_i \mu_k||^2_{\Sigma_i^{-1}} + \text{constant}. \tag{33}$$

Furthermore, $||\mathbf{Y}_i - S_i \mu_k||^2_{\Sigma_i^{-1}}$ can be decomposed into three parts,

$$||\mathbf{Y}_i - S_i \mu_k||^2_{\Sigma_i^{-1}} = ||\mathbf{Y}_i - S_i \hat{\eta}_i||^2_{\Sigma_i^{-1}} + ||\hat{\eta}_i - \lambda_0 - \Lambda \widehat{\alpha}_i||^2_{Cov(\hat{\eta}_i)^{-1}} + ||\widehat{\alpha}_i - \alpha_k||^2_{Cov(\widehat{\alpha}_i)^{-1}}, \tag{34}$$

where $\hat{\eta}_i$ and $\widehat{\alpha}_i$ are given by (10) and (11) in the paper. The first term of (34) is the squared distance between the observed $\mathbf{Y}_i$ and its best cubic spline representation and serves as a measure of the adequacy of the spline basis. The second term gives the squared distance between the optimal spline coefficient vector $\hat{\eta}_i$ and its projection onto the subspace spanned by the cluster mean coefficients. The final term is the squared distance between this projected coefficient vector and μ_k or, equivalently, between $\widehat{\alpha}_i$ and α_k. All distances are measured relative to the appropriate covariances. Notice that the first two terms of (34) are constant with respect to k so the theorem is proved.

C Proof of Theorem 2

Let $\tilde{g}(t)$ be a predictor that depends on $g(t)$ only through $\mathbf{Y}$ and let $\hat{g}(t) = \mathbf{s}(t)^T E_\eta[\eta|\mathbf{Y}]$. Then

$$E_\eta[\tilde{g}(t) - g(t)]^2 = E_\mathbf{Y}[E_\eta[\tilde{g}(t) - g(t)]^2|\mathbf{Y}].$$

Note that

$$E_\eta[(\tilde{g}(t) - g(t))^2|\mathbf{Y}] = E_\eta[(\tilde{g}(t) - \hat{g}(t))^2|\mathbf{Y}] + E_\eta[(\hat{g}(t) - g(t))^2|\mathbf{Y}] + 2E_\eta[(\tilde{g}(t) - \hat{g}(t))(\hat{g}(t) - g(t))|\mathbf{Y}]$$

The cross-product term drops out because, conditional on $\mathbf{Y}$, $\tilde{g}(t) - \hat{g}(t)$ is a constant and the expected value of $\hat{g}(t) - g(t) = \mathbf{s}(t)^T E(\eta|\mathbf{Y}) - \mathbf{s}(t)^T\eta$ is zero. Hence

$$E_\eta[(\hat{g}(t) - g(t))^2] = E_\eta[\tilde{g}(t) - g(t)]^2 - E_\eta[(\tilde{g}(t) - \hat{g}(t))^2]$$

so $\mathbf{s}(t)^T E(\eta|\mathbf{Y})$ minimizes the mean squared error among all predictors that depend on η only through $\mathbf{Y}$.

References

Bachrach, L. K., Hastie, T. J., Wang, M. C., Narasimhan, B., and Marcus, R. (1999). Bone mineral acquisition in healthy Asian, Hispanic, Black and Caucasian youth; a longitudinal study. *Journal of Clinical Endocrinology & Metabolism* **84**, 4702–4712.

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-gaussian clustering. *Biometrics* **49**, 803–821.

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *The Journal of the Pattern Recognition Society* **28**, 781–793.

Dasgupta, A. and Raftery, A. E. (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association* **93**, 294–302.

de Boor, C. (1978). *A Practical Guide to Splines*. New York: Springer.

DiPillo, P. J. (1976). The application of bias to discriminant analysis. *Communications in Statistics, Part A - Theory and Methods* **A5**, 843–854.

Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association* **84**, 165–175.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models : A roughness penalty approach*. Chapman and Hall: London.

Hartigan, J. A. and Wong, M. A. (1978). Algorithm as 136 : A k-means clustering algorithm. *Applied Statistics* **28**, 100–108.

Hastie, T. J., Buja, A., and Tibshirani, R. J. (1995). Penalized discriminant analysis. *Annals of Statistics* **23**, 73–102.

James, G. M. and Hastie, T. J. (2001). Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society, Series B* **63**, 533–550.

James, G. M., Hastie, T. J., and Sugar, C. A. (2000). Principal component models for sparse functional data. *Biometrika* **87**, 587–602.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.

Kaufman, L. and Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley.

Murtagh, F. and Raftery, A. E. (1984). Fitting straight lines to point patterns. *Pattern Recognition* **17**, 479–483.

O'Sullivan, F. (1985). A statistical perspective on ill-posed inverse problems. *Statistical Science* **1**, 502–527.

Rice, J. A. and Wu, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **57**, 253–259.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464.

Sugar, C. A. and James, G. M. (2003). Finding the number of clusters in a data set : an information theoretic approach. *Journal of the American Statistical Association (Conditionally Accepted)* .

Titterington, D. M. (1985). Common structure of smoothing techniques in statistics. *International Statistical Review* **53**, 141–170.

Wald, P. W. and Kronmal, R. A. (1977). Discriminant functions when covariances are unequal and sample sizes moderate. *Biometrics* **33**, 479–484.

Ward, J. H. (1963). Hierarchical groupings to optimize an objective function. *Journal of the American Statistical Association* **58**, 234–244.