Statistics and Its Interface Volume 0 (0000) 1–15

Stratified Psychiatry via Convexity-Based Clustering with Applications Towards Moderator Analysis

Thaddeus Tarpey*, and Eva Petkova† and Liangyu Zhu‡

Understanding heterogeneity in phenotypical characteristics, symptoms manifestations and response to treatment of subjects with psychiatric illnesses is a continuing challenge in mental health research. A long-standing goal of medical studies is to identify groups of subjects characterized with a particular trait or quality and to distinguish them from other subjects in a clinically relevant way. This paper develops and illustrates a novel approach to this problem based on a method of optimal-partitioning (clustering) of functional data. The proposed method allows for the simultaneous clustering of different populations (e.g., symptoms of drug and placebo treated patients) in order to identify prototypical outcome profiles that are distinct from one or the other treatment and outcome profiles common to the different treatments. The clustering results are used to discover potential treatment effect modifiers (i.e., moderators), in particular, moderators of specific drug effects and placebo response. A depression clinical trial is used to illustrate the method.

Keywords and phrases: longitudinal data analysis, mixed models, partitioning, personalized medicine, placebo response.

1. INTRODUCTION

Diagnosing and determining effective treatments for mental illnesses has been and remains a very difficult problem. There is a broadening recognition that the promise of tackling these problems may lie with *stratified psychiatry:* a one-size-fits-all strategy is insufficient while application of a personalized medicine tailored to individuals is not feasible on a large scale. In between these two extremes is stratified medicine with the goal to "... stratify a broad-illness phenotype into a finite number of treatment-relevant subgroups" [10, page 3].

A natural approach to these problems from a statistical perspective is to use methods associated with optimal stratification (or clustering), whereby the goal is to estimate a partition of a population or distribution into homogeneous subgroups, an approach that has a long history in statistics [e.g. 3].

The problem addressed in this paper is to determine an optimum partitioning of a population that is comprised of two or more well-defined sub-populations. In particular, the method focuses on how to partition outcome data from two or more treatments with the acknowledgement that there will be substantial overlap between the outcomes in the different treatments, but also that there may be sets of outcomes that are typical for one of the treatments but not to the others. The motivation comes from randomized clinical trials where it is useful to identify outcomes that are specific to only one treatment. In a trial comparing an active treatment to a placebo, there will often be drug and placebo treated subjects with similar outcomes. However, if there are specific drug effects, one can expect there would exist areas in the outcome space that are primarily populated by drug-treated subjects. Thus, the goal of this paper is to determine a stratification procedure that optimally distinguishes specific (drug) outcomes from non-specific (placebo) outcomes. Another potential application is the problem of making a diagnosis in situations where a clear demarcation does not exist between different illnesses. For instance, it is difficult and controversial to classify a child with Attention Deficit Hyperactivity Disorder (ADHD) or Autism Spectrum Disorder (ASD) if the child exhibits symptoms common to both illnesses [e.g., 6, 1].

Clustering algorithms can be used to estimate an optimal stratification by partitioning a data set into non-overlapping strata. Perhaps the most used algorithm for clustering is the k-means algorithm [e.g. 7, 11]. The utility of the k-means algorithm is that it determines a partition resulting in boundaries in the outcome space that can be used from a clinical perspective for making treatment decisions. However, when data is available from two or more treatment arms, it is not immediately clear how one would implement traditional clustering algorithms. One can cluster the data from each arm separately, but a single partitioning of the pooled data may be desired. A natural approach to the problem is to employ a k-means clustering algorithm on data pooled across different treatments [e.g. 16]. This approach can be improved because it does not directly address the problem

*Corresponding author, Professor at Wright State University
†Associate Professor at New York University
‡Graduate Student at North Carolina State University

of finding strata in the partition that are homogeneous as possible with respect to the different treatment groups.

The k-means algorithm is a special case of *convexity-based* clustering [2]. This paper examines a convexity-based clustering approach where the objective function is given in terms of a likelihood ratio that can be used to partition the outcomes pooled across treatment arms. Convexity-based clustering incorporates the strengths of classical discriminant analysis (which is a supervised learning method, since the treatment labels are known) and applies these strengths to the unsupervised learning problem of cluster analysis. Convexity-based clustering is reviewed in Section 2. In a supervised learning setting, training data is available from distinct groups with labels indicating group membership, which can be employed for estimating a discriminant function and this function is then used to classify future unlabeled observations to one group or the other. In Section 3, the convexity-based clustering is utilized to generalize discriminant analysis to the realm of unsupervised learning. Section 4 provides a one-dimensional example to illustrate convexity-based clustering. The convexity-based clustering is then applied to data from a depression clinical trial in Section 5. The results of the convexity-based clustering are employed to evaluate baseline predictors as moderators of treatment effect in Section 6 and the paper is concluded in Section 7.

2. CONVEXITY-BASED CLUSTERING

This section reviews the basics of convexity-based clustering presented in the work of [2]. The convexity-based clustering represents generalization of the well-known k-means clustering [e.g. 8, 9, 11], which is reviewed first. Given a random variable $\boldsymbol{X}$, the goal is to partition the support $\mathcal{X}$ of $\boldsymbol{X}$ in terms of a given optimality criterion. Here, the support of $\boldsymbol{X}$ can be arbitrary (e.g. $\Re$, $\Re^p$, the set of square-integrable functions on an interval C, $L^2[C]$, etc.).

2.1 k-Means Clustering

k-means clustering is one of the best known non-hierarchical clustering algorithms. Given a data set and an initial set of k cluster means, the basic algorithm forms clusters by assigning each data point to the cluster to which the point is closest, where closeness is typically measured as the Euclidean distance to the cluster mean. Once data points have been assigned to clusters, the cluster means are updated by computing the mean of the points assigned to each cluster. The algorithm iterates between assigning points to clusters based on nearest cluster distance and updating the cluster means until convergence.

The k-means algorithm is an example of a *self-consistency algorithm* [12], which in principal can be applied not only to an empirical distribution defined by a data set, but also directly to theoretical probability distributions [15]. Since the k-means algorithm iterates by assigning points based on a minimal distance to cluster means, the optimality criterion for the algorithm is to form a partition of the support of $\boldsymbol{X}$ that minimizes the within cluster variances, i.e. to find most homogeneous clusters. The goal of k-means clustering is to find a set of k distinct points $\{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_k\}$ in $\mathcal{X}$ that minimize

$$E[\min_j \|\boldsymbol{X} - \boldsymbol{\xi}_j\|^2], \tag{1}$$

over all sets of k distinct points. Such a set of points is called the *k principal points* of $\boldsymbol{X}$ [4]. For $\mathcal{X} = \Re^p$, any set of k points determines a partition of $\Re^p$, say $\{B_1, \ldots, B_k\}$, where $\boldsymbol{x} \in B_j$ if $\|\boldsymbol{x} - \boldsymbol{\xi}_j\|^2 < \|\boldsymbol{x} - \boldsymbol{\xi}_h\|^2, h \neq j$. It is easy to see that the solution to the k-means problem requires the cluster means $\boldsymbol{\xi}_j$ to be the centroids over the respective sets forming the partition [e.g., 5]:

$$\boldsymbol{\xi}_j = E[\boldsymbol{X}|\boldsymbol{X} \in B_j].$$

From this it follows that the k-means criterion is equivalent to finding a partition that maximizes

$$\sum_{j=1}^{k} P(B_j)\|E[\boldsymbol{X}|\boldsymbol{X} \in B_j]\|^2, \tag{2}$$

for a given k, where $P(B_j)$ is just $P(\boldsymbol{X} \in B_j)$. If we let $\phi(\boldsymbol{x})$ denote the convex function $\phi(\boldsymbol{x}) = \|\boldsymbol{x}\|^2$, then (2) can be written

$$\sum_{j=1}^{k} P(B_j)\phi(E[\boldsymbol{X}|\boldsymbol{X} \in B_j]). \tag{3}$$

The idea behind convexity-based clustering is to allow ϕ to be other convex functions besides the squared norm function.

2.2 A General Convexity-Based Clustering

Let $\phi(\boldsymbol{x})$ denote a convex function and define the *tangent support plane* at point $\boldsymbol{z}$ by

$$t(\boldsymbol{x}; \boldsymbol{z}) = \phi(\boldsymbol{z}) + \nabla\phi(\boldsymbol{z})'(\boldsymbol{x} - \boldsymbol{z}), \tag{4}$$

where ∇ is the gradient operator. Given a partition $\{B_1, \ldots, B_k\}$ and a set of support points $\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k\}$, the volume between ϕ and the approximating support hyperplanes can be computed as

$$\sum_{j=1}^{k} \int_{B_j} (\phi(\boldsymbol{x}) - t(\boldsymbol{x}; \boldsymbol{z}_j))\, dF(\boldsymbol{x}), \tag{5}$$

where F is the distribution function under consideration. To avoid complications, assume F is continuous and ϕ is differentiable everywhere. The goal is to find an optimal partition and a set of support points that provides the best approximation by minimizing the volume (5) between the surface defined by $\phi(\boldsymbol{x})$ and the piecewise function defined by the

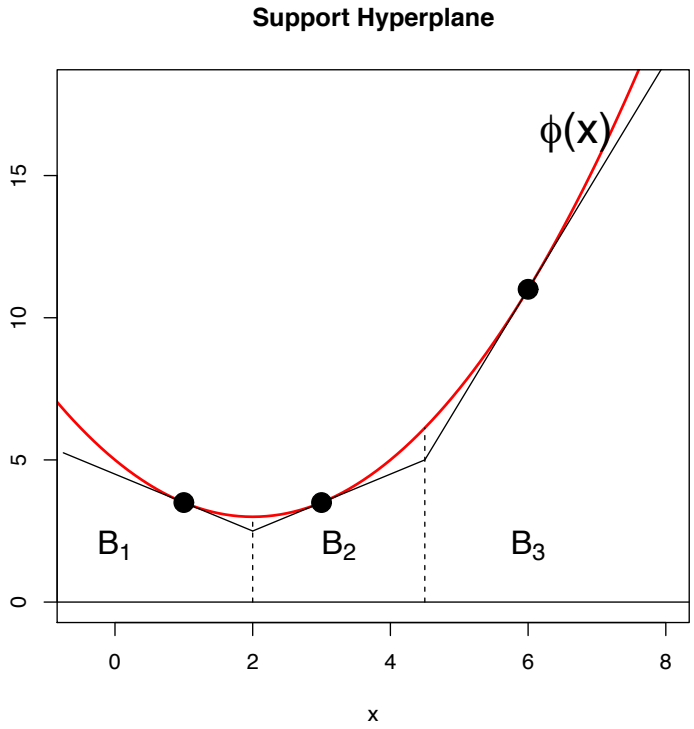

Figure 1. An illustration of a tangent support plane using $k = 3$ points for a convex function $\phi(x)$.

support tangent hyperplanes over the partition $B_1, \ldots, B_k$. The optimal partition that minimizes this volume will equivalently maximize [see 2, Corollary 2.2]

$$\sum_{j=1}^{k} P(B_j)\phi(E[\boldsymbol{X}|\boldsymbol{X} \in B_j]), \tag{6}$$

which is the same criterion as the k-means algorithm (3). The notion of a tangent support plane for a convex function ϕ is illustrated in Figure 1 with $k = 3$ support points.

[2] (Theorem 2.1) showed that given a partition $\{B_1, \ldots, B_k\}$, the minimal volume problem and equivalently the problem of maximizing (6), the support points $\boldsymbol{z}_j$ must satisfy

support point就是中心

$$\boldsymbol{z}_j = E[\boldsymbol{X}|\boldsymbol{X} \in B_j]. \tag{7}$$

Alternatively, given support points $\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k\}$, in order to maximize (6), the partition must satisfy

$$B_j = \{\boldsymbol{x} \in \mathcal{X} : t(\boldsymbol{x}; \boldsymbol{z}_j) = \max_h t(\boldsymbol{x}; \boldsymbol{z}_h)\}. \tag{8}$$

Equations (7) and (8) suggest an iterative algorithm for *convexity-based clustering*: given an initial set of support points, use (8) to determine the corresponding partition; given the partition, update the support points using (7); iterate between these two steps until convergence.

A useful generalization of the convexity-based clustering algorithm considered by [2] that we shall implement, is to consider a pre-specified function $\lambda : \mathcal{X} \to \Re^q$ instead of $\boldsymbol{x}$ directly. The support points and the sets in the partition in this general framework are given by

$$w_j = E[\lambda(\boldsymbol{x})|\boldsymbol{x} \in B_j]$$

and

$$D_j = \{\lambda(\boldsymbol{x}) : t(\lambda(\boldsymbol{x}); w_j) = \max_h t(\lambda(\boldsymbol{x}); w_h)\}.$$

This induces a partitioning of $\mathcal{X}$ as

$$B_j = \lambda^{-1}(D_j) = \{\boldsymbol{x} \in \mathcal{X} : \lambda(\boldsymbol{x}) \in D_j\}.$$

The partitioning algorithm described above can be formulated in this more general setting as follows:

Generalized Convexity-Based Clustering Algorithm

0. Start with an initial partition $B_1, \ldots, B_k$ of $\mathcal{X}$
1. Calculate the support points

   $$w_j = E[\lambda(\boldsymbol{x})|\boldsymbol{x} \in B_j]. \tag{9}$$

2. Determine a minimum support plane partition

   $$D_j = \{\lambda(\boldsymbol{x}) : t(\lambda(\boldsymbol{x}); w_j) = \max_h t(\lambda(\boldsymbol{x}); w_h)\} \tag{10}$$

   for $j = 1, \ldots, k$.
3. Update the partition by $B_j \leftarrow \lambda^{-1}(D_j)$.
4. Repeat steps 1 – 3 until a convergence criterion is met.

3. CONVEXITY-BASED CLUSTERING: FROM DISCRIMINANT ANALYSIS TO PARTITIONING

Consider the case where a population consists of T sub-populations or classes. For the sake of discussion (and for the illustrations below) we shall focus on the case of $T = 2$ with two sub-populations, say I and II. In the supervised learning setting of discriminant analysis, each observation in the data comes with a class label indicating to which sub-population the observation belongs. Using this information, a discriminant function can be defined for classifying new unlabeled observations to one or the other class. Suppose the densities in each sub-population are f_1 and f_2 with prior probabilities π_1 and π_2. Then the optimal rule for classification in terms of minimizing the probability of misclassification is Bayes' rule, where an observation $\boldsymbol{x}$ is classified to the population I if

$$\frac{\pi_1 f_1(\boldsymbol{x})}{\pi_1 f_1(\boldsymbol{x}) + \pi_2 f_2(\boldsymbol{x})} > \frac{\pi_2 f_2(\boldsymbol{x})}{\pi_1 f_1(\boldsymbol{x}) + \pi_2 f_2(\boldsymbol{x})}, \tag{11}$$

and it is classified to population II otherwise. If f_1 and f_2 are multivariate normal densities, then Bayes' rule coincides

with Fisher's linear discriminant function when each group has the same covariance matrix, and it coincides with a quadratic discriminant function if the covariance matrices differ.

Write

$$f(\boldsymbol{x}) = \pi_1 f_1(\boldsymbol{x}) + \pi_2 f_2(\boldsymbol{x}) \tag{12}$$

for the mixture density in the denominator of (11). If the right hand side of (11) is subtracted from the left hand side, then the magnitude of this squared difference for an observation $\boldsymbol{x}$ is a measure of the strength in how well the observation can be classified to one or the other sub-populations:

$$\left(\frac{\pi_1 f_1(\boldsymbol{x})}{f(\boldsymbol{x})} - \frac{\pi_2 f_2(\boldsymbol{x})}{f(\boldsymbol{x})} \right)^2.$$

This expression can be written as

$$\phi(\lambda(\boldsymbol{x}))$$

where ϕ is the convex function

$$\phi(\lambda) = (1 - 2\lambda)^2 \tag{13}$$

and

$$\lambda(\boldsymbol{x}) = \frac{\pi_2 f_2(\boldsymbol{x})}{f(\boldsymbol{x})}, \tag{14}$$

which is just the posterior probability that an observation $\boldsymbol{x}$ belongs to population II. The criterion to be maximized by the k-means algorithm (3) in this setting is

$$\mathcal{C} = \sum_{j=1}^{k} P(B_j) \phi(E[\lambda(\boldsymbol{X}) | \boldsymbol{X} \in B_j]), \tag{15}$$

where the expectation is taken with respect to the mixture density (12). Let

$$P_1(B_j) = \int_{B_j} f_1(\boldsymbol{x}) d\boldsymbol{x}$$

denote the probability a random observation from the population I lies in B_j (similarly for P_2) and let

$$P(B_j) = \pi_1 P_1(B_j) + \pi_2 P_2(B_j).$$

Then the criterion (15) to be maximized becomes

$$\begin{aligned}
\mathcal{C} &= \sum_{j=1}^{k} P(B_j) \phi(E[\lambda(\boldsymbol{X}) | \boldsymbol{X} \in B_j]) \\
&= \sum_{j=1}^{k} P(B_j) (1 - 2E[\lambda(\boldsymbol{X}) | \boldsymbol{X} \in B_j])^2 \\
&= \sum_{j=1}^{k} P(B_j) \left(1 - \frac{2}{P(B_j)} \int_{B_j} \lambda(\boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x} \right)^2 \\
&= \sum_{j=1}^{k} P(B_j) \left(1 - \frac{2}{P(B_j)} \int_{B_j} \pi_2 f_2(\boldsymbol{x}) d\boldsymbol{x} \right)^2 \\
&= \sum_{j=1}^{k} \frac{(\pi_1 P_1(B_j) - \pi_2 P_2(B_j))^2}{P(B_j)}. \qquad (16)
\end{aligned}$$

Thus, the goal of the convexity-based clustering criterion is to find a partition with strata B_j that are maximally homogeneous with respect to one or the other sub-populations I and II. Equation (16) is similar to the ϕ-divergence between distributions used to maximize the power of a goodness-of-fit test of the null hypothesis that a distribution equals some specified distribution (see [2] for references).

The convexity-based clustering described here will be particularly suitable when the two sub-populations overlap substantially. If the two populations do not overlap (or overlap just negligibly), then the convexity-based clustering will not be very advantageous. For instance, if there exists a partition $\{B_1, \ldots, B_k\}$ where either $P_1(B_j) = 0$ or $P_2(B_j) = 0$ for every j, then the support of the two sub-populations do not overlap at all and an optimal partition can be obtained simply using $k = 2$ with strata $\{f_1(\boldsymbol{x}) > 0\}$ and $\{f_2(\boldsymbol{x}) > 0\}$ and the criterion (16) obtains the maximum value of 1.

The generalized convexity-based clustering algorithm described in the previous section can be derived by computing the tangent function $t(\lambda; w)$ for (13). In particular, given an ordered set $w_1 < \cdots < w_k$, or support points, it follows after some straightforward algebra that the sets D_j given by (10) are determined by the midpoints of the w_j. Using the known treatment labels (populations I and II), the convexity-based clustering algorithm for generalizing the usual discriminant analysis becomes:

Semi-Superivsed Discriminant Clustering Algorithm

0. Start with an initial partition $B_1, \ldots, B_k$ of $\mathcal{X}$.
1. Calculate the support points (9) as

$$w_j = \frac{\pi_2 P_2(B_j)}{P(B_j)}. \tag{17}$$

2. Determine a minimum support plane partition

$$D_j = \{\lambda \in \Re : \|\lambda - w_j\| < \|\lambda - w_h\|, h \neq j\} \tag{18}$$

for $j = 1, \ldots, k$,.

3. Update the partition by $B_j \leftarrow \lambda^{-1}(D_j)$, whereby if

$$\lambda(\boldsymbol{x}) = \frac{\pi_2 f_2(\boldsymbol{x})}{f(\boldsymbol{x})} \in D_j, \text{ then } \boldsymbol{x} \rightarrow B_j.$$

4. Repeat steps 1 – 3 until a convergence criterion is met.

This algorithm can be naturally generalized to accommodate $T > 2$ groups or treatments. In this more general setting, λ and w become vector-valued and the convex function ϕ generalizes to

$$\boldsymbol{\phi}(x_1, \ldots, x_T) = \sum_{h=1}^{T} (1 - 2x_h)^2.$$

Additionally, using a similar derivation as above, the sets D_j in (18) in the case of more than two groups define a Voronoi partition in $\Re^{T-1}$ in this more general setting.

3.1 Implementing the Algorithm

Applying the semi-supervised discriminant clustering described above requires estimating the densities f_1 and f_2. A straightforward approach is to assume some parametric family for the densities, say multivariate normal, and then estimate the parameters of the densities using maximum likelihood. Alternatively, nonparametric estimators of the densities could be used. The prior probabilities π_1 and π_2 can be estimated as sample proportions, if the sample sizes in each sub-population are random, or values for these priors can be substituted if they are known.

The initial partition can be achieved by simply applying the k-means algorithm on the pooled data from both sub-populations. Once the densities are estimated, the function λ from (14) can be computed for each data observation $\boldsymbol{x}_i$ as

$$\hat{\lambda}_i := \hat{\lambda}(\boldsymbol{x}_i) = \frac{\hat{\pi}_2 \hat{f}_2(\boldsymbol{x}_i)}{\hat{f}(\boldsymbol{x}_i)}. \tag{19}$$

A straightforward way of computing the w_j in (17) of step 1 of the Semi-Superivsed Discriminant Clustering algorithm is via Monte Carlo simulation: if data can be simulated from sub-populations I and II, say once the parameters of these two distributions have been estimated, then the probabilities in (17) can be well-approximated by sample means from a very large data set simulated from the two sub-populations. This procedure is illustrated in Section 5.

4. ONE-DIMENSIONAL ILLUSTRATIONS

To shed light on the mechanics behind convexity-based clustering, here we present simple 1-dimensional illustrations, where the algorithm is applied to a population defined by two univariate normal distributions. These illustrations use the convex function given by (13) in order to generalize Bayes' classification rule from the realm of supervised to unsupervised learning.

In the left panel of Figure 2, the convexity-based clustering algorithm was applied on a two component univariate normal mixture (with equal mixing weights) with densities

$$f_1(x) \sim N(0, 1) \text{ and } f_2 \sim N(2, 1)$$

which are plotted as dashed and dotted curves respectively and the mixture density is potted as a solid gray curve. A partition was formed using $k = 3$. In this case, the two populations have the same variance, but different locations. The three convexity-based cluster means are denoted by diamonds at the base of the density plot and a color-coding is provided at the top to show the $k = 3$ strata. For the sake of a comparison, the $k = 3$ principal points of the mixture distribution [4] are also shown (denoted by the solid circles at the bottom). The principal points are the population quantities that the cluster means from the k-means algorithm are estimating. The partitioning that the principal points determine is not concerned with the purity of the strata with respect to the mixing components. In this illustration, however, the convexity-based cluster means and the principal points are very similar to one another.

The right panel of Figure 2 shows another univariate illustration with $k = 4$ for a population defined by a 2-component normal mixture of densities $N(0, 1)$ and $N(0, 2)$. In this case, the two populations share a common mean ($\mu = 0$) but they differ in spread ($\sigma^2 = 1$ versus $\sigma^2 = 2$). The $k = 4$ strata are denoted by the color-coded bar along the top of the density curves. In this case, the convexity-based cluster means (diamonds) and the principal points (solid circles) do not coincide at all, as was the case in the previous illustration: here the convexity-based cluster means all coincide at the origin (the common mean) whereas the principal points are spread out. The strata formed by the convexity-based clustering are not even convex sets whereas clusters formed by k-means clustering will always be convex sets. Initially, this may seem to be a weakness of the convexity-based clustering approach and in some applications, non-convex clusters may not be useful. However, in the right panel of Figure 2, note that the blue-colored stratum covers the extreme left and right of the mixture distribution and is comprised mostly of the more variable sub-population. The black-colored stratum is mostly populated by the less variable sub-population, and corresponds to the region where the two populations most heavily overlap.

5. EXAMPLE: IDENTIFYING SPECIFIC DRUG RESPONDERS IN TREATING DEPRESSION

The convexity-based clustering procedure was applied to a 6-week longitudinal depression study where subjects were

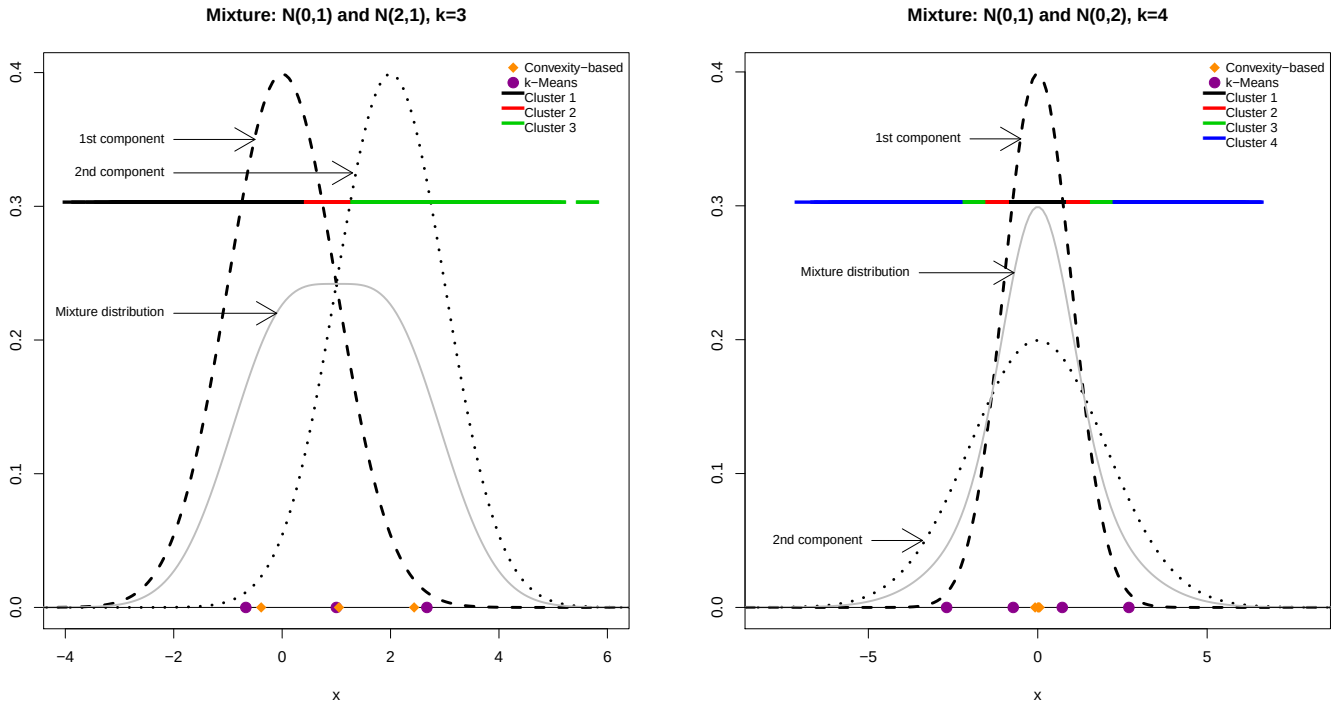

Figure 2. A univariate illustration of the convexity-based clustering. In the left panel the convexity-based clustering is applied to a two component mixture of $N(0,1)$ and $N(2,1)$ with equal mixing weights using $k = 3$. In the right panel, the algorithm is applied using $k = 4$ to an equal mixture of $N(0,1)$ and $N(0,2)$ distributions. The convexity-based cluster means are denoted by diamonds and for comparison, k-means cluster means are denoted by solid circles.

randomized to be treated with Fluoxetine or a placebo. The outcome (recorded at each weekly visit) was severity of depression assessed with the Hamilton Rating Scale for Depression (HRSD) or sometimes known simply as the HAM-D. Lower scores on this scale indicate lower levels of depression. A mixed-effects model was fit separately to the data from both arms using orthogonal-quadratic polynomials, with random subject intercept, linear and quadratic terms.

Other basis functions could have been used to fit the longitudinal trajectories (e.g. Fourier basis functions or B-splines). We choose to use orthogonal quadratic polynomials for the following reasons. (i) Most clustering algorithms are based on a minimal Euclidean distance. When the data are curves in function space, the L^2 distance between curves corresponds to the usual Euclidean distance between regression coefficients when an orthonormal basis is used to fit the curves. Additionally, differences in clustering that occur due to the choice of the basis functions used to represent the curves are minimized when using an orthogonal basis function representation [14]. (ii) Quadratic functions are easily interpretable and provide a good fit to the data over this relatively short longitudinal evaluation period (6 weeks). In particular, with orthogonal quadratic polynomials, the coefficient of the linear polynomial corresponds to the average quadratic slope of the parabola, which is an overall measure of improvement throughout the trial [13, 18]. Also, the coefficient of the quadratic polynomial is a simple measure of the trajectory's curvature which has important interpretations in clinical settings, particularly when modeling placebo response.

Figure 3 shows a scatterplot of the estimated quadratic outcome trajectories for Fluoxetine-treated (left panel) and placebo-treated (right panel) subjects. Figure 4 shows the estimated coefficients (the linear and quadratic terms) for individual subjects in the Fluoxetine (black) and placebo (red) treated subjects along with contours of equal probability for each distribution. As these two figures show, there is a large degree of overlap between the parabolas of Fluoxetine- and placebo-treated subjects, which makes the problem of teasing out regions that are homogeneous with respect to the two treatments difficult.

The distributions of the coefficients for the symptom trajectories of subjects (intercepts, linear and quadratic terms) were estimated from mixed effects models separately for the Fluoxetine and placebo groups which were assumed to follow a trivariate normal distributions in the maximum likelihood estimation. As noted above, with orthogonal polynomials, the coefficient of the linear function corresponds to the "average quadratic slope" [13], which is an overall measure of improvement (or worsening) throughout the 6-week trial. The convex clustering was applied to the bivariate distribution defined by the (average) slope and quadratic polynomial (concavity) coefficients only since these two coefficients determine the shape of the trajectory whereas the intercept corresponds only to a vertical shift of the trajectories.

In order to apply the clustering algorithm, a large simulated data sample (of size 50,000 for each treatment) was obtained as follows. The mixed effects model can be expressed as

$$\boldsymbol{y}_i = \boldsymbol{X}_i(\boldsymbol{\beta} + \boldsymbol{b}_i) + \boldsymbol{\epsilon}, \tag{20}$$

where $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{D})$ is the vector of random effects, assumed independent of the error $\boldsymbol{\epsilon}_i$, and $\boldsymbol{X}_i$ is the design matrix for the orthogonal polynomials over time for the ith subject. The parameters of the distribution of the trajectories are specified by their coefficient distribution

$$\boldsymbol{\beta} + \boldsymbol{b}_i \sim N(\boldsymbol{\beta}, \boldsymbol{D}), \tag{21}$$

which can be estimated using maximum likelihood.

Once $\boldsymbol{\beta}$ and $\boldsymbol{D}$ were estimated for each treatment arm, Monte Carlo simulation was used to generate a very large sample to approximate the multivariate normal coefficient distribution (21). Details of this approach are given in [17]. This Monte Carlo sample can then be used to compute the probabilities $P_1(B_j)$ and $P_2(B_j)$ needed in the iterations of the clustering algorithm.

Various values of k were tried for the convexity-based clustering and for illustration, we present the results for $k = 4$ here. Figure 5 shows the estimated boundaries formed by a $k = 4$ semi-supervised discriminant clustering partition denoted by solid curves. (Results for $k = 3$ and $k > 4$ produced cluster partition patterns similar to those shown in Figure 5.) The k-means algorithm was used to initialize the algorithm and the algorithm essentially converged after only about 5 iterations. The $k = 4$ strata are denoted C_1 to C_4. The solid points on the figure correspond to the means of the coefficient distributions for the outcome trajectories under Fluoxetine (black) and placebo (red) treatment. It is interesting to note that the center curve (between C_2 and C_3) on Figure 5 coincides with the quadratic discriminant boundary between the placebo and Fluoxetine populations. As illustrated below, using $k = 4$ provides a more refined differentiation between placebo and drug-treated subjects than simply using a standard discriminant function. Estimating a 2-class discriminant function defined using the placebo and drug arm labels in the study is not very useful because the full extent of heterogeneity in the data is not captured completely by the two treatment arm labels, as can be seen in Figure 3.

Table 1 shows a breakdown in percentages of Fluoxetine- and placebo-treated subjects in the data that are classified to the $k = 4$ clusters from the convexity-based clustering (rounded to the nearest integer). The percentages also break down as to whether a person was classified as a treatment responder or non-responder by the end of the trial using the clinical global impression (CGI) scale for improvement. According to this criteria, a subject is rated as a responder if their CGI score was 1 (very much improved) or 2 (much improved). Higher scores on this scale range from 3 (minimally

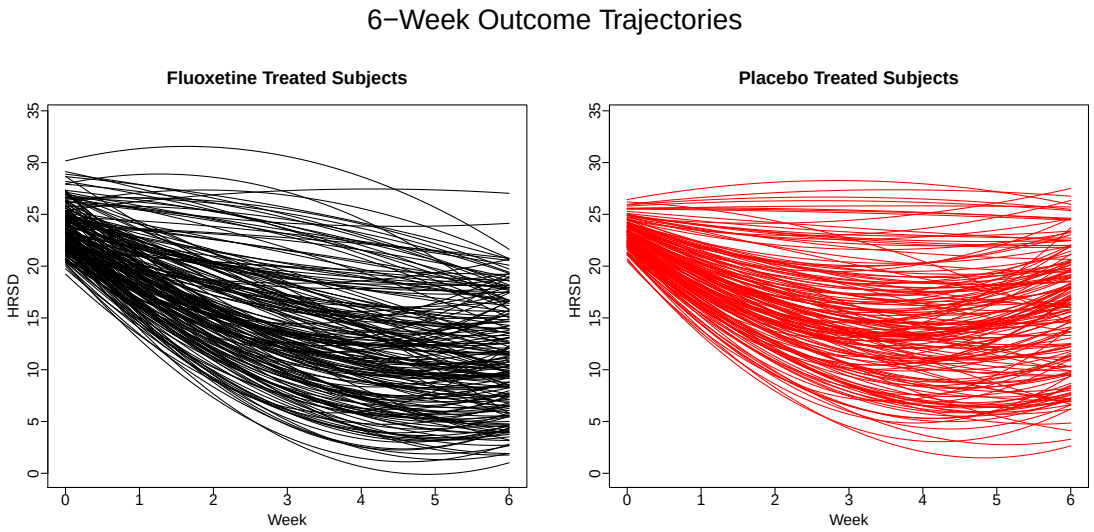

Figure 3. Quadratic outcome trajectories for Fluoxetine (left panel) and placebo treated (right panel) subjects from a 6-week study.

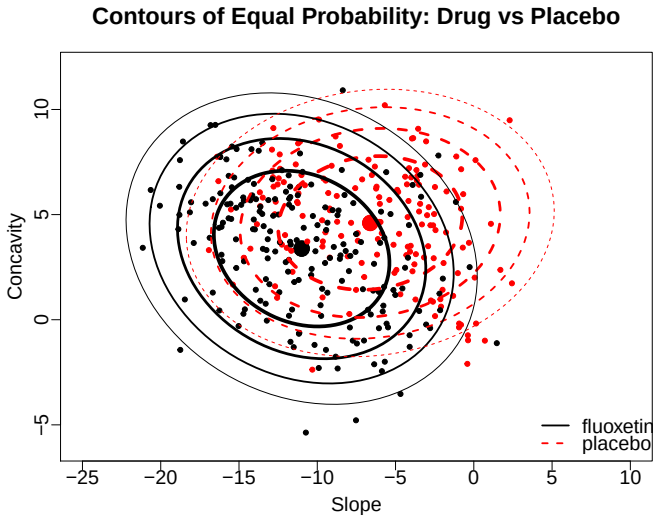

Figure 4. Contours of equal density for the joint distribution of the average slope and concavity coefficients for Fluoxetine and placebo treated subjects. The points (black for Fluoxetine and red for placebo) are the individual coefficient (estimated fixed effect plus predicted random effects).

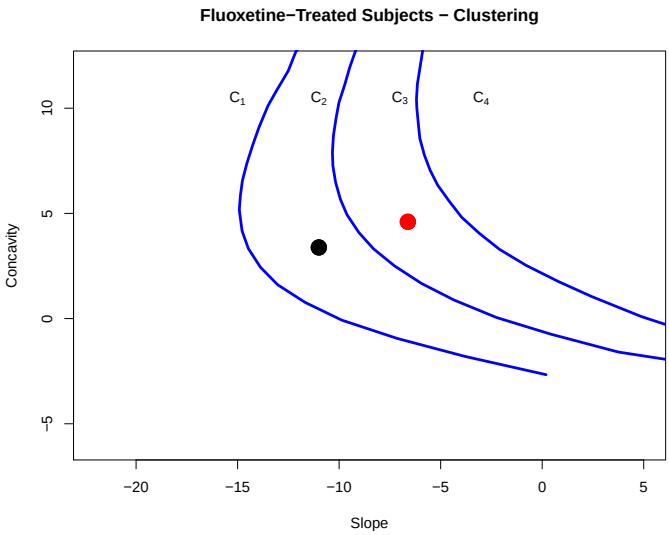

Figure 5. Convexity-based clustering partition for $k = 4$

Table 1. Percent of Subjects Classified to Each Cluster

	Fluoxetine, $n = 196$			Placebo, $n = 162$		
Cluster	% Responders	% Non-Responders	% Total	% Responders	% Non-Responders	%Total
1	29	6	35	5	0	5
2	31	14	45	24	4	28
3	4	12	16	10	31	41
4	0	4	4	0	26	26
Overall	64	36	100	39	61	100

improved) to 7 (very much worse) and subjects with such scores are rated as treatment non-responders.

As Table 1 shows, Cluster 1 consists primarily of Fluoxetine treated responders and very few placebo treated subjects. The placebo treated subjects that were classified to Cluster 1 were rated as responders. It is interesting to note that of the 11 Fluoxetine-treated subjects rated as non-responders and classified to Cluster 1, all but one were dropouts with only observations up to visit 3 or 4. Thus, Cluster 1 can be labeled as a "Drug Responder" cluster.

The primary motivation of the semi-supervised discriminant clustering in this example is to identify subjects who will benefit from the specific effects of the medication, not just non-specific placebo effects. If this can be accomplished, then treatment can be targeted towards patients who will benefit from it. If a particular treatment is not going to be very beneficial for a patient (for instance, if the patient's response to Fluoxetine is primarily a placebo response), then these patients can be switched to a different medication that may be more efficacious.

From the semi-supervised discriminant clustering results, we can infer that subjects classified to Cluster 1 tend to benefit from the specific effects of the drug. The reasoning here is that subjects in Cluster 1 are almost all responders treated with the active drug. Therefore, they must be mostly specific drug responders. This does not rule out the possibility their responses could also be partly due to placebo effects. However, if these responses were mostly due to placebo effects, then this cluster would be populated by placebo-treated responders as well, but very few placebo treated responders are in Cluster 1. The curve coefficients for drug-treated subjects classified to Cluster 1 are shown in Figure 6 (the 11 non-responders who were primarily dropouts are indicated by open circles in this plot); the corresponding outcome trajectories are shown in Figure 7.

Cluster 4 is populated overwhelmingly with placebo-treated non-responders and only a few drug-treated non-responders. Clearly, subjects classified to Cluster 4 experience very weak placebo effects, if any. However, Cluster 4 distinguishes between drug and placebo treated subjects since this cluster consists mostly of placebo-treated subjects. It stands to reason then, that most of these placebo-treated non-responders would have benefited from the specific effect of the drug had they been treated with the active medication.

Clusters 2 and 3 do not discriminate as well between drug and placebo treated subjects since both these clusters contain large percentages of subjects from both treatments. Cluster 2 is populated primarily by CGI-rated responders and obviously the placebo-treated subjects in Cluster 2 are placebo responders. The estimated trajectories for placebo-treated subjects in Cluster 2 are shown in Figure 8. The solid curves in this figure correspond to CGI-rated responders and hence, placebo-responders have curves that are characterized by immediate improvement from baseline that level off or even deteriorate by the end of the 6-week period. We note here the distribution of subjects in Clusters 2 and 3: among Fluoxetine-treated subjects 48% are in Cluster 2 and 14% are in Cluster 3; among placebo-treated subjects 30% are in Cluster 2 and 41% are in Cluster 3. One can infer that the drug was responsible for pushing subjects from Cluster 3, where the majority subjects are non-responders, to Cluster 2, where there is higher prevalence of treatment responders. Although it is difficult to distinguish between specific, non-specific and mixed (specific and non-specific) responders in Cluster 2, it is reasonable to conclude that some non-negligible proportion of them were helped by the drug.

If baseline covariates can be found that predict whether a subject will fall in Cluster 1 if treated with Fluoxetine, then these covariates can be used to predict who will benefit from the specific effects of the drug – this idea is explored in Section 6. Similarly if there are baseline covariates that can predict whether a subject will fall in Cluster 4 when treated with placebo, then these covariates can be used to identify subjects who are likely to benefit from the specific effects of the drug. If a set of covariates can successfully predict whether a subject will be in Cluster 2 if treated with placebo, then these covariates can identify placebo responders. Finally, if a set of covariates can predict that a subject will fall in Cluster 3, whether treated with a drug or placebo, then these covariates can help identify subjects that are non-responsive to both specific and non-specific effects of treatment. Of course, as with most all classification methods though, we expect that there will be misclassifications.

6. MODERATOR IMPORTANCE PLOTS

The semi-supervised discriminant clustering has allowed us to partition the set of observations (symptom trajectories over time) into k strata/clusters that are maximally homogeneous with respect to one or the other of two treatments (e.g., Fluoxetine or placebo). The partition obtained from the convexity-based clustering uses the objective function (14). We can plot the estimated function $\hat{\lambda}(\cdot)$ versus a baseline predictor to determine if the convexity-based classification depends on some baseline characteristic.

Figure 9 shows a plot of $\hat{\lambda}(\cdot)$ versus baseline Clinical Global Impression of severity (CGI-severity) for drug-treated subjects (black dots) and for placebo-treated subjects (red squares). CGI-severity takes values from 1 to 7 with higher values indicating more severe depression, specifically, 1 = normal (not ill), 2 = minimally ill, 3 = mildly ill, 4 = moderately ill, 5 = markedly ill, 6 = severely ill, 7 = extremely ill. (Only one person had a value less than 4 and only three had values equal to 7.) The horizontal lines in Figure 9 mark the cutoff values, see (17), which determine cluster membership for the observations. The large solid symbols are the mean values of λ at each unique CGI

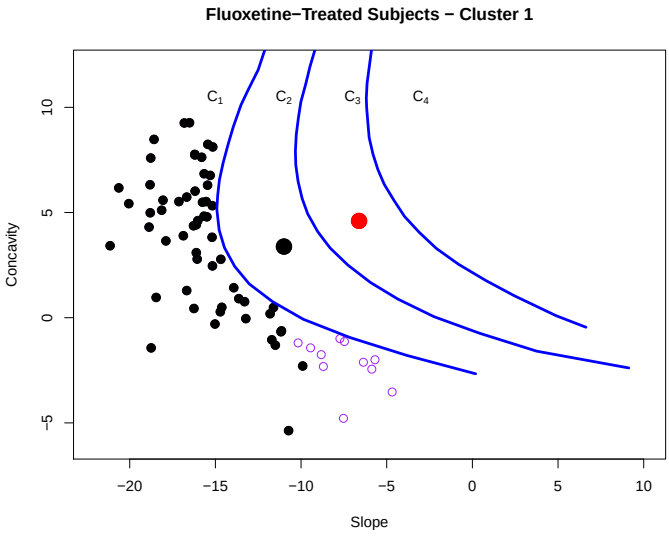

Figure 6. Drug-treated subjects classified to Cluster 1. Open circles correspond to drug-treated subjects who were rated as CGI non-responders (all but one of these subjects dropped out).

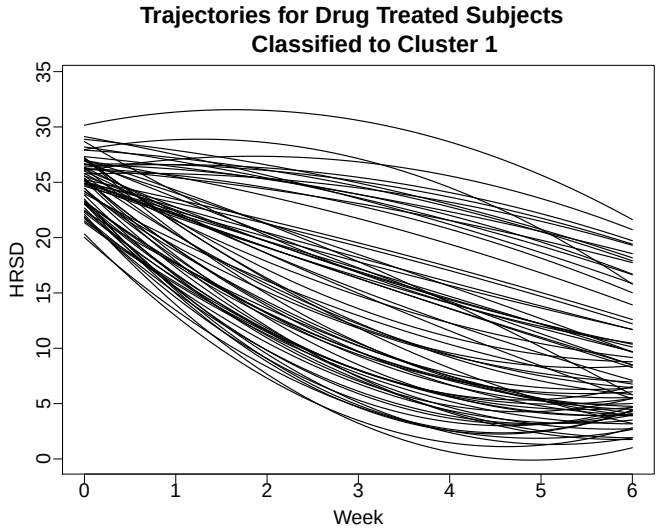

Figure 7. Estimated outcome trajectories for drug-treated subjects classified to Cluster 1.

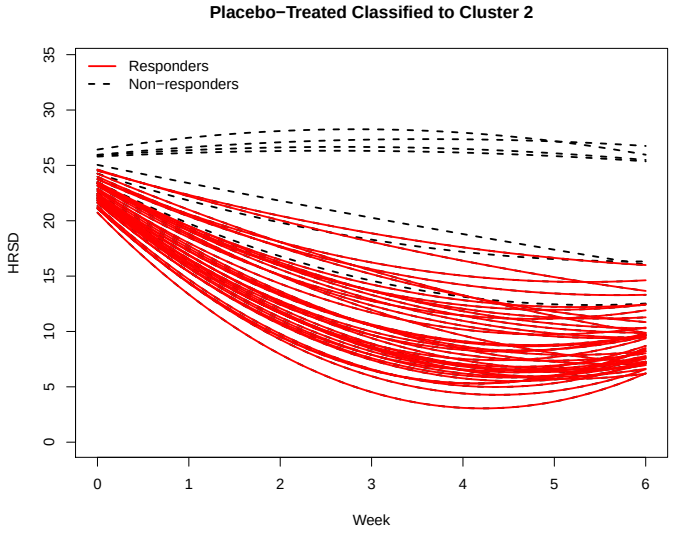

Figure 8. Estimated outcome trajectories for placebo-treated subjects classified to Cluster 2. CGI-rated responders are denoted by solid red curves.

value for the drug-treated subjects (circles) and the placebo-treated subjects (squares) respectively. As the figure shows, among drug-treated individuals, the more severely depressed a subject was at baseline the the more likely they were to be classified towards the specific drug responder Cluster 1. In particular, only 24% of the drug-treated subjects with a baseline CGI of 4 were classified to the drug-responder cluster, whereas 53% of drug-treated subjects with a baseline CGI of 5 were classified to the drug responder cluster and 40% of drug-treated subjects with a baseline CGI of 6 where classified to the drug-responder cluster. On the other hand, 31% of placebo-treated subjects with a low baseline CGI = 4 were classified to the non-responder Cluster 4, compared to only 14% of placebo-treated subjects with baseline CGI = 5 (only one out of the nine placebo-treated subject with baseline CGI of 6 were classified to Cluster 4). Interestingly, placebo-treated subjects who were moderately depressed at baseline (CGI=4) where more likely to fall into the non-responder cluster than more severely depressed placebo-treated subjects. Thus, there is modest evidence that baseline CGI severity is helpful in predicting whether or not a Fluoxetine-treated patient will respond due to specific effects of the drug.

A similar analysis was done using the baseline predictor age (in years) and the results are shown in Figure 10. The function $\hat{\lambda}$ plotted versus age was estimated using a penalized cubic spline for both drug and placebo-treated subjects and the estimated curves are shown in Figure 10: the top curve is for drug-treated subjects and the bottom curve is for placebo-treated subjects. Both curves in Figure 10 are basically flat indicating that age appears to have no substantial moderating effect in predicting cluster membership in this case.

7. DISCUSSION

Because of the difficulty in diagnosing many mental diseases and also the difficulty of distinguishing between different mental illness when making a diagnoses, unsupervised learning tools are paramount to better understand sources of heterogeneity in mental disporders. Cluster analysis has long been a mainstay of unsupervised learning. This paper has proposed a semi-supervised clustering algorithm that incorporates the powerful features of discriminant analysis from supervised learning.

A common complication that arises in applications, such as in the example in Section 5, is missing data (e.g. due to drop out). Estimating the longitudinal trajectories, the best linear unbiased predictors (BLUP) for the trajectories are obtainable under the assumption that the missing data are missing at random (MAR) conditional on the prior observations of the outcome. If there is a reason to suspect the MAR condition is violated, imputation methods based on more covariates than just the previous observations of the outcome can be used to produce multiple imputed data sets. Comparing the clustering results for the multiple imputed data sets would be a nice way of assessing the effect of the missing data on the clustering inferences.

The ultimate goal of the approach presented in this paper is to discover biomarkers for mental illnesses such as depression. That is, to use baseline measures, even including complicated modalities such as brain images, that can (hopefully) help in diagnosing mental illness and also be used to predict the optimal treatment for patients. In the example presented here there were only very limited number

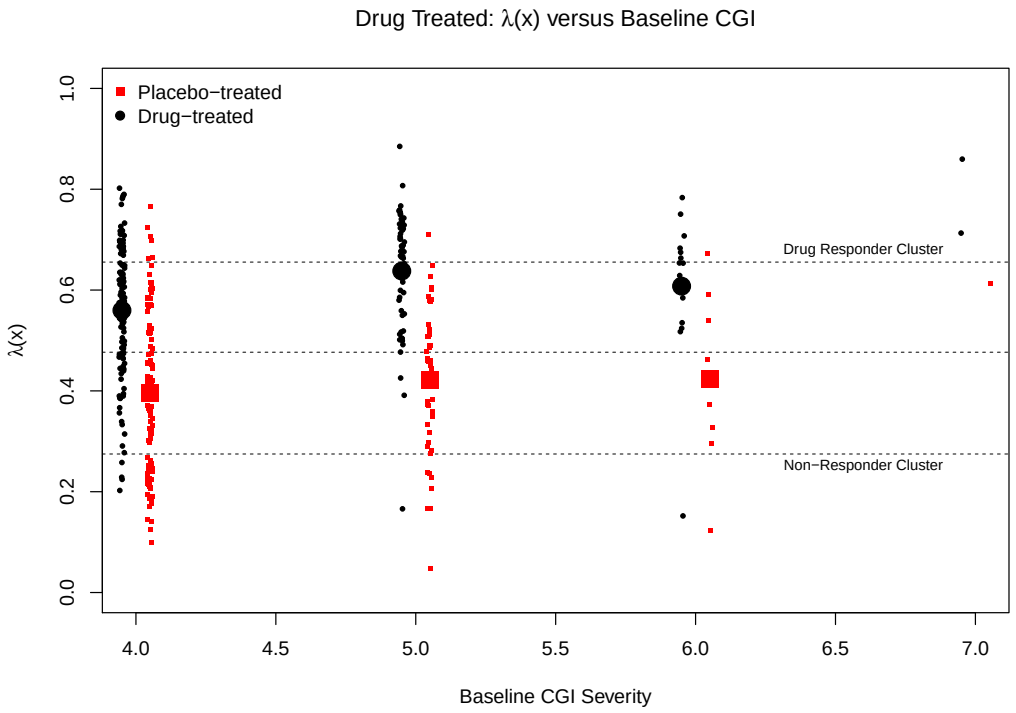

Figure 9. A moderator importance plot: a plot of $\hat{\lambda}(\cdot)$ versus baseline CGI severity for Fluoxetine-treated subjects. The horizontal lines demark the cut-offs for cluster membership.

of baseline measures and the moderator importance of the two baseline measures that were analyzed (baseline CGI-severity and age) were modest to non-existent. With the increasing availability of baseline measures such as brain images and genetic data, there is the hope of finding more powerful biosignatures of drug and placebo response.

ACKNOWLEDGEMENTS

We would like to thank the Editor and Associate Editor for considering our paper and two referees whose comments and suggestions led to a much improved paper. This work was supported by NIMH grant R01 MH099003. We would like to thank the Eli Lilly Company for providing the data used in this paper.

REFERENCES

[1] Adamo, N., Huo, L., S, S. A., Petkova, E., Castellanos, F. X., and Martino, A. D. (2013). Response time intra-subject variability: commonalities between children with autism spectrum disorders and children with adhd. *European Child and Adolescent Psychiatry* .

[2] Bock, H.-H. (2003). Convexity-based clustering criteria: theory, algorithms, and applications in statistics. *Statistical Methods & Applications* **12**:293–317.

[3] Dalenius, T. (1950). The problem of optimum stratification. *Skandinavisk Aktuarietidskrift* **33**:203–213.

[4] Flury, B. (1990). Principal points. *Biometrika* **77**:33–41.

[5] Flury, B. (1993). Estimation of principal points. *Applied Statistics* **42**:139–151.

[6] Grzadzinski, R., Di Martino, A., Brady, A., Mairena, M., O'Neale, M., Petkova, E., Lord, C., and Castellanos, F. (2011). Examining autistic traits in children with ADHD: Does the autism spectrum extend to ADHD? *Journal of Autism and Developmental Disorders* **41**:1178–1191.

[7] Hartigan, J. A. (1975). *Clustering Algorithms*. Wiley, New York.

[8] Hartigan, J. A. (1978). Asymptotic distributions for clustering criteria. *Annals of Statistics* **6**:117–131.

[9] Hartigan, J. A. and Wong, M. A. (1979). A K-means clustering algorithm. *Applied Statistics* **28**:100–108.

[10] Kapur, S., Phillips, A. G., and Insel, T. R. (2012). Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Molecular Psychiatry* **12**.

[11] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings 5th Berkeley Symposium on Mathematics, Statistics and Probability* **3**:281–297.

[12] Tarpey, T. (1999). Self–consistency and principal component analysis. *Journal of the American Statistical Association* **94**:456–467.

[13] Tarpey, T. (2003). Estimating the average slope. *Journal of Applied Statistics* **30**:389–395.

[14] Tarpey, T. (2007). Linear transformations and the k-means clustering algorithm: Applications to clustering curves. *The American Statistician* **61**:34–40.

[15] Tarpey, T. (2007). A parametric k-means algorithm. *Computational Statistics* **22**:71–89.

[16] Tarpey, T. and Petkova, E. (2010). Principal point classification: Applications to differentiating drug and placebo responses in longitudinal studies. *Journal of Statistical Planning and Inference* **140**:539–550.

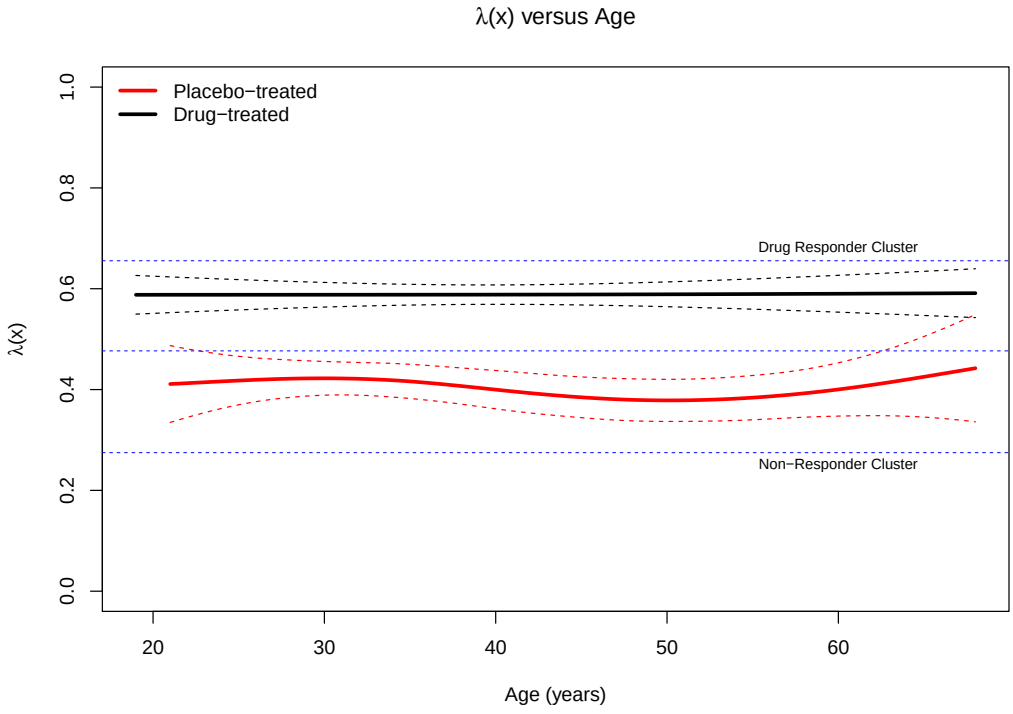

Figure 10. A moderator importance plot using the baseline predictor age.

[17] Tarpey, T., Petkova, E., Lu, Y., and Govindarajulu, U. (2010). Optimal partitioning for linear mixed effects models: Applications to identifying placebo responders. *Journal of the American Statistical Association* **105**:968–977.

[18] Tarpey, T., Petkova, E., and Ogden, R. T. (2003). Profiling placebo responders by self-consistent partitioning of functional data. *Journal of the American Statistical Association* **98**:850–858.

Thaddeus Tarpey
Department of Mathematics and Statistics,
Wright State University,
Dayton, Ohio 45435
E-mail address: thaddeus.tarpey@wright.edu

Eva Petkova
Department of Child and Adolescent Psychiatry,
New York University, New York, NY 10016-6023
and Nathan S. Kline Institute for Psychiatric Research,
Orangeburg, NY 10962
E-mail address: eva.petkova@nyumc.org

Liangyu Zhu
Department of Statistics, North Carolina State University
Rayleigh, NC 27695
E-mail address: lzhu12@ncsu.edu