Canonical k-Means Clustering for Functional Data

Thaddeus Tarpey[1] and Eva Petkova[2]

[1] Department Mathematics and Statistics, Wright State University, Dayton, Ohio 45435, thaddeus.tarpey@wright.edu.

[2] Department of Child and Adolescent Psychiatry, New York University, New York, NY 10016-6023

Abstract

Cluster analysis is a powerful tool for discovering sources of heterogeneity in data. However, clinically interesting sources of heterogeneity, such as placebo-effects or specific drug effects may be swamped out by other sources of variability in the data which can cause a distribution to deviate from normality. This paper proposes linearly transforming the data before clustering. An example of this is a canonical type transformation to maximize between cluster variability relative to within cluster variability.

Keywords: B-splines, independent component analysis, k-means clustering, placebo effect.

1 Introduction

In functional data analysis, each individual data point produces a curve and the shape of these curves can provide useful insight into the nature of the problem being studied. For example, longitudinal outcome trajectories from a clinical trial shed light into the nature of response to treatment (e.g. non-response, specific drug response, placebo response, or combinations of these effects, etc.). Clustering the curves can allow for the identification of prototypical response trajectories and hence identify clinically meaningful but yet distinct types of outcomes to treatment. One of the driving motivations for this work is to identify response trajectory shapes that can distinguish between subjects that respond due to specific effects of an active treatment (e.g. a drug) from subjects who respond primarily due to nonspecific effects of treatment (which we shall call "placebo" effects).

The problem of clustering functional data has received a lot of attention in recent years (Lipkovich *et al.*, 2008; Luschgy and Pagés, 2002; Abraham *et al.*, 2003; Tarpey and Kinateder, 2003) A closely related approach is to fit finite mixture models to functional data or growth mixture models (GMM) for longitudinal data (e.g. Muthén and Shedden, 1999; James and Sugar, 2003). This paper focuses on non-hierarchical clustering methods, in particular, variants of the well-known k-means algorithm (e.g. Hartigan and Wong, 1979).

A major shortcoming with k-mean-type algorithms is that the resulting partitions of the data are often driven by the directions of primary variability in the data, regardless of whether or not primary directions of variability correspond to distinct sub-populations or continuous latent variables (e.g. degree of placebo response) that result in clinically interesting heterogeneity in the population of interest. The existence of distinct sub-populations and/or continuous latent variables will cause a distribution to deviate from multivariate normality. Additionally, if the variability of these effects, which presumably are of primary interest from a clinical perspective, are swamped by other sources of variability in the data (e.g. degree of baseline disease severity among subjects), then clustering algorithms used to discover this heterogeneity will likely miss these effects. For instance, the variability due to the specific chemical effect of an active medication may be quite small in magnitude due to other non-specific effects of treatment (e.g. Petkova *et al.*, 2009).

The approach described in this paper is to linearly transform the functional data to maximize the between cluster variability relative to the within cluster variability. Closely related to this approach are projection pursuit clustering methods (Bock, 1987; Bolton and Krzanowski, 2003). More recently, Yatracos (2013) proposed a hierarchial clustering approach also based on projections that successively maximize a component of the variance for one-dimensional projections of the data.

Consider a linear model for fitting a curve to an outcome vector $\boldsymbol{y}_i$ using a design matrix of basis functions $\boldsymbol{X}_i$:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i. \tag{1}$$

The functional data can be partitioned by clustering the estimated coefficients $\hat{\boldsymbol{\beta}}_i$ (Tarpey and Kinateder, 2003). However, consider a nonsingular matrix $\boldsymbol{A}$. Then the model in (1) is identical to

$$\boldsymbol{y}_i = [\boldsymbol{X}_i\boldsymbol{A}^{-1}][\boldsymbol{A}\boldsymbol{\beta}_i] + \boldsymbol{\epsilon}_i = \boldsymbol{Z}_i\boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_i, \tag{2}$$

which can be regarded a modification of the basis representation of the functional observations. As noted by Tarpey (2007), k-means clustering the coefficients $\hat{\boldsymbol{\beta}}_i$ from (1) can lead to quite different results than clustering the coefficients $\hat{\boldsymbol{\alpha}}_i$ in (2) even though both models produce identical fits. The goal of clustering then is to find a basis, or given a basis, a linear transformation $\boldsymbol{A}$ that will steer the k-means algorithm in a direction that will discover true clusters or clinically relevant partitions of the data.

2 Canonical Transformation Clustering

Clustering functional data using the k-means algorithm will perform best if the linear transformations used to fit the curves stretch the data in a direction that corresponds to interesting sources of heterogeneity in the distribution, such existence of distinct clusters. Because the algorithm iterates by assigning points to the cluster whose center is closest, the optimization achieved by the algorithm is to find groupings that maximize the between group sum-of-squares relative to minimizing the within group sum-of-squares.

Consider an arbitrary partitioning of the underlying distribution into k strata. Let $\boldsymbol{\mu}_j$ and $\boldsymbol{\Psi}_j$ denote the mean and covariance matrix respectively of the random regression coefficients $\boldsymbol{\beta}_i$ for the jth stratum and let π_j denote the proportion of the population in the stratum, $j = 1, 2, \ldots, k$. The covariance matrix for the $\boldsymbol{\beta}_i$ can be decomposed as

$$\text{cov}(\boldsymbol{\beta}_i) = \boldsymbol{W} + \boldsymbol{B}, \tag{3}$$

where

$$\boldsymbol{W} = \sum_{j=1}^{k} \pi_j \boldsymbol{\Psi}_j \;\text{ and }\; \boldsymbol{B} = \sum_{j=1}^{k} \pi_j (\boldsymbol{\mu}_j - \boldsymbol{\mu})(\boldsymbol{\mu}_j - \boldsymbol{\mu})',$$

are the within cluster (or stratum) and the between cluster covariance matrices respectively and where $\boldsymbol{\mu} = \sum_{j=1}^{k} \pi_j \boldsymbol{\mu}_j$. From (3) one can see that in order to optimize the k-means clustering, a transformation should be used that minimizes the contribution of the within cluster variability while maximizing the between cluster variability. A canonical discriminant function is defined as "linear combinations of variables that best separate the mean vectors of two or more groups of multivariate observations relative to the within-group variance" (Rencher, 1993). In canonical discriminant analysis, transformations based on vectors $\boldsymbol{a}_j$ that successively maximize $(\boldsymbol{a}_j'\boldsymbol{B}\boldsymbol{a}_j)/(\boldsymbol{a}_j'\boldsymbol{W}\boldsymbol{a}_j)$ are used. The solution is to choose the $\boldsymbol{a}_j$

as the eigenvectors of $\boldsymbol{W}^{-1}\boldsymbol{B}$. A canonical transformation for clustering is now defined by first linearly transforming the regression coefficient vector into Fisher's canonical variates followed by a stretching of the coefficient distribution to accent the between cluster variability and minimize the within cluster variability. In particular, consider a linear transformation that simultaneously diagonalizes $\boldsymbol{W}$ and $\boldsymbol{B}$. Denote the spectral decomposition of $\boldsymbol{W}^{-1/2}\boldsymbol{B}\boldsymbol{W}^{-1/2}$ by $\boldsymbol{H}\boldsymbol{D}\boldsymbol{H}'$ where $\boldsymbol{H}$ is an orthogonal $p \times p$ matrix and $\boldsymbol{W}^{1/2}$ is the symmetric square root of $\boldsymbol{W}$. Let $\boldsymbol{\Gamma} = \boldsymbol{W}^{-1/2}\boldsymbol{H}$. Let the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ in $\boldsymbol{D}$ be arranged from largest to smallest down the diagonal. Then from (3), the covariance matrix of $\boldsymbol{\Gamma}'\boldsymbol{b}$ will be

$$\boldsymbol{I} + \boldsymbol{D}. \tag{4}$$

In order to accent the between cluster variability and diminish the contribution of the within cluster variability, one can further transform using a *canonical* transformation for clustering

$$\boldsymbol{C}\boldsymbol{\Gamma}'\boldsymbol{b} \tag{5}$$

where $\boldsymbol{C} = \text{diag}(c_1, c_2, \ldots, c_p)$ and the $c_j \geq 0$ are appropriately chosen constants. From (4), the covariance matrix for the canonically transformed coefficients in (5) is $\boldsymbol{C}^2 + \boldsymbol{C}^2\boldsymbol{D}$. Thus, choosing large values of c_j corresponding to eigenvalues in $\boldsymbol{D}$ greater than one inflates the between cluster variability relative to the within cluster variability of the canonically transformed coefficients and setting $c_j = 0$ for eigenvalues between zero and one minimizes the contribution of the within cluster variability. For instance, suppose the cluster means lie on a line. Then multiplying the positive eigenvalue λ_1 in $\boldsymbol{D}$ by a large value of c_1 transforms the coefficient distribution by stretching it in the direction of the line containing the cluster means. Consequently, the k-means algorithm will place cluster means along this line for large values of c_1. If the cluster means lie approximately in a q-dimensional plane, then one would choose $c_1, \ldots, c_q$ to be large and the remaining c_j to be small. The problem then is to determine the optimal settings for the c_j in order to optimize the k-means algorithm according to minimizing a mean squared error or a classification error rate.

3 Illustration

This section illustrates linear transformations of the functional data based on independent component analysis which attempts to find linear transformations via projections that deviate as much as possible from normality. Data from 12-week open-label acute phase of a depression discontinuation trial with $n = 429$ subjects will be used to illustrate the methods. The outcome is a subject's Hamilton Depression (Ham-D) score over the course of the 12 week treatment with prozac (weeks 0, 1, 2, 3, 4, 5, 6, 8, 10, 11, and 12) where lower scores correspond to lower levels of depression. B-splines (with a single knot) were used to fit curves to individuals' responses resulting in a $p = 5$ dimensional coefficient vector for each subject. A crude check of normality of the coefficient distribution was performed by running a Shapiro-Wilks test for normality for each of the five coefficient distribution and none of these tests returned a p-value below 0.05. Given the large sample size, this would provide some indication a multivariate normality assumption for the coefficient distribution is not unreasonable. However, an independent component analysis was performed using the fastICA algorithm in R (R Development Core Team, 2009) to estimate an ICA model (Hyvarinen and Oja, 2000), and the p-values for normality from the Shapiro-Wilks test for three of the five independent components were extremely small ($p < 0.00001$).

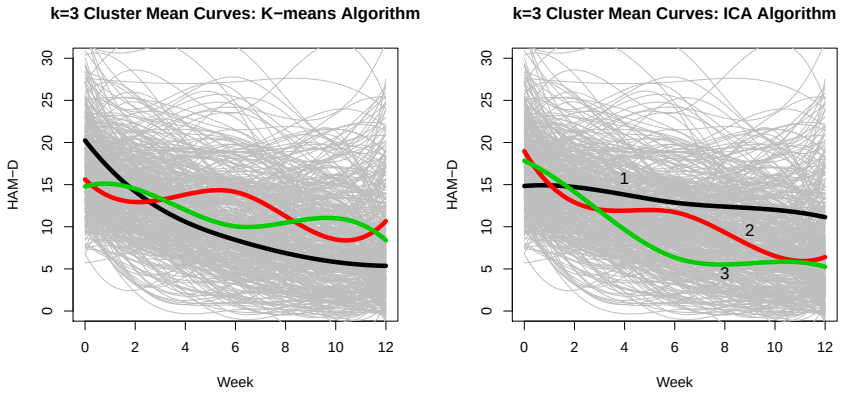

Figure 1: $k = 3$ cluster mean curves fit to the B-spline curves using the usual k-means algorithm (left panel) and using the ICA clustering algorithm in the right panel. Individual-level curves are plotted in grey.

Since ICA amounts to a linear transformation of the coefficient distribution, this distribution certainly appears to deviate from normality, which could be an indication of latent categorical or continuous variables creating heterogeneity within the trajectory outcome distribution. Figure 1 shows the results of a regular k-means clustering of the coefficients in the left panel and an ICA-based clustering in the right panel. For the ICA-based clustering, the 1st and 3rd independent components deviated strongly from normality (based on a Shaprio-Wilks test) and hence the k-means algorithm was constrained to run in the subspace spanned by these two independent components by inflating their variability (by a factor of 100) compared to the remaining independent components.

At this point, a canonical clustering algorithm could be implemented to find optimal choices of the stretching coefficients c_3 and c_4 to optimize the k-means algorithm in this 2-dimensional subspace. The cluster means from the usual k-means algorithm (left panel) differ quite a bit from the ICA produced cluster means in the right panel. The cluster mean trajectories in the right panel of Figure 1 show one curve (black) corresponding to a steady improvement and eventual leveling off. The green curve shows an immediate improvement, perhaps due to initial placebo effects, followed by a stronger improvement, but then a deterioration in improvement perhaps indicating that initial benefits from the active drug are not sustained. On the other hand, the red curve would correspond to individuals showing an initial improvement that levels off and then additional improvement perhaps indicating that once the drug builds in the system, these individuals experience specific drug effects that improve mood.

4 Biosignatures for Treatment Response

One of the primary motivations for the clustering methodology is to develop biosignatures for treatment response. The idea here is once the clustering method has determined a partition of the distribution of response trajectories, then baseline covariates can be used to develop predictive models of whether or not subjects are likely to fall into one cluster or another based on these baseline measures. If particular clusters correspond predominately to placebo responders or specific drug responders and baseline measures, or some com-

bination of these measures, can be used to predict cluster membership, and hence act as biosigatures of treatment response. The ultimate goal is to implement this methodology to high dimensional predictors such as brain-imaging scans and genetic data.

5 Discussion

This paper has examined how modifying the basis functions used to fit functional data via linear transformations of the coefficient distribution can lead to cluster solutions that can discover interesting heterogeneity in functional data that a straightforward cluster analysis may miss. Anticipated improvements to this work are to utilize subject specific random effects in additional to fitting curves using penalized splines on both the fixed effect mean curve and the individual subject-specific random effect components of the curves (e.g. Chen and Wang, 2011). Another interesting question when clustering functional data to be explored is: How much structure in the curves does one need to extract in order to capture meaningful partitions when clustering curves? In other words, if the clusters of curves are distinguished simply by their intercepts, then we do not need to fit curves to the data at all, but just need to cluster the mean outcomes for each curve. Similarly, if the clusters are distinguished by their linear trends, regardless of the curvature in the trajectories, then we can obtain useful results by simply fitting straight lines to the data and clustering the lines.

Additional work will also address the question of finding linear transformations that maximize the R^2 for clustering where

$$R^2 = 1 - \frac{\text{within sum-of-squares}}{\text{total sum-of-squares}},$$

where the within and total sum-of-squares are from a k-means clustering. Preliminary results indicate that the optimal linear transformation is to simply project the data onto a one-dimensional subspace. The optimality of a one-dimensional projection likely follows due to the curse of dimensionality where points become further and further away from cluster centers as the dimension increases.

References

Abraham, C., Cornillon, P. A., Matzner-Lober, E., and Molinari, N. (2003). Unsupervised curve clustering using b-splines. *Scandinavian Journal of Statistics* pp. 581–595.

Bock, H. H. (1987). *On the interface between cluster analysis, principal component analysis, and multidimensional scaling*, pp. 17–34. D. Reidel Publishing Company.

Bolton, R. J. and Krzanowski, W. J. (2003). Projection pursuit clustering for exploratory data analysis. *Journal of Computational and Graphical Statistics* **12**:121–142.

Chen, H. and Wang, Y. (2011). A penalized spline approach to functional mixed effects model analysis. *Biometrics* **67**:861–870.

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

Hyvarinen, A. and Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks* **13**:411–430.

James, G. and Sugar, C. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association* **98**:397–408.

Lipkovich, I., Houston, J., and Ahl, J. (2008). Identifying patterns in treatment response prifiles in acute bipolar mania: a cluster analysis approach. *BMC Psychiatry* **8**:1–8.

Luschgy, H. and Pagés, G. (2002). Functional quantization of Gaussian processes. *Journal of Functional Analysis* **196**:486–531.

Muthén, B. and Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics* **55**:463–469.

Petkova, E., Tarpey, T., and Govindarajulu, U. (2009). Predicting potential placebo effect in drug treated subjects. *International Journal of Biostatistics* **5**.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rencher, A. C. (1993). Interpretation of canonical discriminant functions, canonical variates, and principal components. *The American Statistician* **46**:217.

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

Tarpey, T. and Kinateder, K. J. (2003). Clustering functional data. *Journal of Classification* **20**:93–114.

Yatracos, Y. G. (2013). Detecting clusters in the data from variance decompositions of its projections. *Journal of Classification* **30**:30–55.