Smoothing Dissimilarities to Cluster Binary Data

David B. Hitchcock

University of South Carolina

Department of Statistics*

Zhimin Chen

Morehouse School of Medicine

Cardiovascular Research Institute

March 10, 2008

Abstract

Cluster analysis attempts to group data objects into homogeneous clusters on the basis of the pairwise dissimilarities among the objects. When the data contain noise, we might consider performing a smoothing operation, either on the data themselves or on the dissimilarities, before implementing the clustering algorithm. Possible benefits to such pre-smoothing are discussed in the context of binary data. We suggest a method for cluster analysis of binary data based on "smoothed" dissimilarities. The smoothing method presented borrows

*Corresponding author: David B. Hitchcock, Department of Statistics, University of South Carolina, Columbia, SC 29208 (email: hitchcock@stat.sc.edu) (Phone: 803-777-5346).

ideas from shrinkage estimation of cell probabilities. Some simulation results are given showing that improvement in the accuracy of the clustering result is obtained via smoothing, especially in the case in which the observed data contain substantial noise. The method is illustrated with an example involving binary test item response data.

KEY WORDS: Cluster analysis; Contingency table; Matching coefficient; Shrinkage; Stein estimation; ACT.

1 Introduction

Cluster analysis is the statistical technique of separating objects, or observations, into homogeneous groups on the basis of (typically multivariate) data for several variables. We often picture the variables as continuous, but there is a substantial literature about clustering objects based on binary data (e.g., Everitt, Landau and Leese, 2001; Kaufman and Rousseeuw, 1990).

When the data contain some type of noise (whether measurement error or merely unexplained variability), it is intuitive that smoothing the data, when done properly, may better recapture the underlying process generating the data. Often individual data values contain substantial noise and thus are less trustworthy to reflect the process we hope to understand. Smoothing methods attempt to reduce this noise by balancing the information in individual data points with information in the data set as a whole, or by shrinking data values toward some assumed structural model. Cluster analysis itself may be viewed as a type of smoothing, in the sense of being a technique to obtain a less complex structure from noisy data. However, standard clustering methods can be sensitive to outliers that could exist when we directly cluster observed data. Therefore clustering a smoothed version of the data may be preferable to clustering the observed (unsmoothed) data.

In certain situations the idea of smoothing is natural. For example, Hitchcock, Booth and Casella (2007) showed that a shrinkage method of smoothing could aid in the clustering of functional data (data arising as curves). With binary data, the idea of "smoothing" seems less natural than with functional data, but the concept of shrinkage will be an important one in the methods discussed here.

A common method for clustering binary data objects is to define pairwise dissimilarities among the objects, each of which is typically a function of the number of matches (or mismatches) among the p binary variables measured on the pair of objects. A "match" occurs when, for a certain variable, both objects share the same value (both 0 or both 1). For any pair of objects a 2×2 table of matches and mismatches may be constructed. Our smoothing method will fundamentally use this table.

In Section 2 we will formally define the dissimilarities for a set of binary data and introduce a clustering method based on a smoothed version of this collection of dissimilarities. Section 3 describes a simulation study to determine the effect of this smoothing method on the accuracy of the cluster analysis. In Section 4, we apply the method to a real data set involving test item responses, and Section 5 is a conclusion.

2 Method

In this section we present a method of clustering binary data objects based on dissimilarities that are "smoothed" via a shrinkage technique. As a motivation for this approach, consider the following hypothetical example. A class of schoolchildren are given a series of tests, each of which entails performing some physical task (e.g., doing a pull-up, jumping over a bar, etc.). The data point observed on each child for each task is binary (0/1) according to whether the task was successfully completed. The goal is to group the set of children into clusters based on the set of binary data. Note such binary

	$\mathbf{Y}_{k'}$		
$\mathbf{Y}_k$	0	1	Totals
0	a	b	$a+b$
1	c	d	$c+d$
Totals	$a+c$	$b+d$	$P=a+b+c+d$

Table 1: Table listing number of matches and mismatches for a pair of objects $\mathbf{Y}_k$ and $\mathbf{Y}_{k'}$. The number (among the P variables) of variables for which $\mathbf{Y}_k = \mathbf{Y}_{k'} = 0$ is a; the number of variables for which $\mathbf{Y}_k = 0$ and $\mathbf{Y}_{k'} = 1$ is b; the number of variables for which $\mathbf{Y}_k = 1$ and $\mathbf{Y}_{k'} = 0$ is c; the number of variables for which $\mathbf{Y}_k = \mathbf{Y}_{k'} = 1$ is d.

observations are imperfect: They do not account for how close a child came to accomplishing the task, nor the ease with which it was accomplished. A child's binary score on a task may not measure his or her underlying ability on that task; using the observed binary data may lead to deceptive results if there is enough "noise" in the measurements. We suggest that a smoothing procedure may yield dissimilarities that are better inputs to a clustering algorithm in many cases.

2.1 Dissimilarities for a Binary Data Set

For any two objects (represented by the binary data vectors $\mathbf{Y}_k, \mathbf{Y}_{k'}$), consider the 2×2 table of matches and mismatches shown in Table 1.

A variety of measures of similarity or of distance between a pair of objects may be calculated from the elements of this 2×2 table; see Finch (2005) for a discussion of several such metrics. A very common measure of similarity between the two objects is the *simple matching coefficient*

$$s_{kk'} = (a+d)/P,$$

where $a + d$ is the number of variables on which the two objects match and P is the total number of binary variables (Sokal and Michener, 1958). Then the corresponding dissimilarity measure is

$$d_{kk'} = 1 - s_{kk'} = (b + c)/P. \tag{1}$$

This choice of dissimilarity measure implicitly assumes that, for any pair of objects, matches on 0 and matches on 1 are equally informative in the cluster analysis. In some analyses, this is not the case, and the similarity coefficient may weight a and d unequally (Johnson and Wichern, 2002). For this study, we will focus on the simple measure that weights both types of matches equally. According to Hands and Everitt (1987), this simple matching coefficient is the choice most often used in practice to cluster binary data.

Once a $n \times n$ dissimilarity matrix $\mathbf{D}$ containing all pairwise dissimilarities is constructed, standard clustering methods (such as hierarchical linkage methods or partitioning methods such as K-means) can be used to group the objects. Note that in the clustering process, these pairwise dissimilarities play an analogous role to distance measures (such as Euclidean distance) that are commonly employed in cluster analysis with continuous data.

While the standard approach uses the observed dissimilarities, we propose to use “smoothed” dissimilarities: in particular, dissimilarities based on a smoothed version of the 2 × 2 table. Since the underlying data process contains random noise, we assume that this smoothing will reduce the noise and produce dissimilarities more reflective of the true discrepancies among the signal components of the various objects.

The problem of smoothing entries in a 2 × 2 table is quite well studied, in the very different context of categorical data analysis. A widely used approach in estimating the cell probabilities for a two-way contingency table is to shrink the observed cell proportions toward some model-based probability estimate. We will borrow this approach in our dissimilarity estimation problem.

2.2 Possible Choices for the Model-based Estimators

For any particular pair of binary objects, let $\boldsymbol{\pi} = (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})'$ denote the set of true probabilities of a value falling in the respective cell in the 2×2 table in Table 1. (For example, in the cluster analysis setting, π_{11} represents the probability that the pair of objects both have a value of zero for a particular variable.) Then let $\tilde{\boldsymbol{\pi}} = (\tilde{\pi}_{11}, \tilde{\pi}_{12}, \tilde{\pi}_{21}, \tilde{\pi}_{22})'$ be a set of cell-probability estimates based on some model.

We now suggest possible models to obtain the $\{\tilde{\pi}_{ij}\}$ values. If the investigator has no prior knowledge whatsoever about the clustering structure among the binary objects, a logical choice is some type of default or noninformative model. For example, a simple default model might assign equal probabilities to a value falling in each of the four cells such that $\tilde{\pi}_{ij} = 0.25$, $i = 1, 2$; $j = 1, 2$. A model assuming independence between rows and columns might assign $\tilde{\pi}_{ij} = \hat{\pi}_{i+}\hat{\pi}_{+j}$, $i = 1, 2$, $j = 1, 2$, where a $+$ indicates summation over the subscript's set of values. For example, $\hat{\pi}_{i+} = \sum_j \hat{\pi}_{ij}$, where the $\{\hat{\pi}_{ij}\}$ represent the observed cell proportions.

On the other hand, a model of dependence might assume two objects are, say, four times as likely to "match" on a binary variable than to have a mismatch, in which case we would assign $\tilde{\boldsymbol{\pi}} = (\tilde{\pi}_{11}, \tilde{\pi}_{12}, \tilde{\pi}_{21}, \tilde{\pi}_{22})' = (0.4, 0.1, 0.1, 0.4)'$.

Furthermore, if we have some prior knowledge of the clustering structure of the objects, we could vary the sets of $\{\tilde{\pi}_{ij}\}$ values across object pairs. For example, for a pair of objects strongly suspected to belong to the same cluster, we could assign $\tilde{\boldsymbol{\pi}} = (0.45, 0.05, 0.05, 0.45)'$; for a pair of objects suspected to belong to different clusters, we could assign $\tilde{\boldsymbol{\pi}} = (0.2, 0.3, 0.3, 0.2)'$.

In the next section we present a shrinkage method that yields a "smoothed" result that is robust with respect to a possibly misguided choice of model-based estimates.

2.3 A Shrinkage-type Smoother for the 2×2 Table

A general form of shrinkage estimator π^*_{ij} of each cell probability is written as (see, e.g., Albert, 1987) a weighted average of the observed cell proportion $\hat{\pi}_{ij}$ and some model-based estimate $\tilde{\pi}_{ij}$:

$$\pi^*_{ij} = (1-\lambda)\hat{\pi}_{ij} + \lambda\tilde{\pi}_{ij}. \qquad (2)$$

Here, λ represents a smoothing parameter that often depends on the sample cell counts. Among the first to derive such an estimator were Fienberg and Holland (1973), who showed that by placing a Dirichlet prior (having means $\{\gamma_{ij}\}$) on the set of cell probabilities, a Bayes estimator resembling (2) could be derived:

$$\frac{P}{P+\kappa}\hat{\pi}_{ij} + \frac{\kappa}{P+\kappa}\gamma_{ij}.$$

They showed that choosing

$$\kappa = \frac{1-\sum \pi^2_{ij}}{\sum(\gamma_{ij}-\pi_{ij})^2}$$

minimized a total mean squared error criterion, and used the sample proportions $\{\hat{\pi}_{ij}\}$ to estimate the unknown $\{\pi_{ij}\}$. Since one must choose the prior parameters $\{\gamma_{ij}\}$, a natural (empirical Bayesian) approach suggested by Fienberg and Holland is to use some model-based estimates $\{\tilde{\pi}_{ij}\}$. The resulting estimators of $\{\pi_{ij}\}$ have the form of (2) with

$$\lambda = \frac{\hat{\kappa}}{P+\hat{\kappa}}$$

with

$$\hat{\kappa} = \frac{1-\sum \hat{\pi}^2_{ij}}{\sum(\tilde{\pi}_{ij}-\hat{\pi}_{ij})^2}.$$

When applying this techique in smoothing the dissimilarities, we will focus on the Fienberg-Holland approach, which provides an objective choice of the smoothing parameter, but we note that κ (and thus λ) could be chosen subjectively. In a 2×2 table, the Fienberg-Holland estimate of κ is

$$\hat{\kappa} = \frac{[1 - (\hat{\pi}_{11}^2 + \hat{\pi}_{12}^2 + \hat{\pi}_{21}^2 + \hat{\pi}_{22}^2)]}{(\tilde{\pi}_{11} - \hat{\pi}_{11})^2 + (\tilde{\pi}_{12} - \hat{\pi}_{12})^2 + (\tilde{\pi}_{21} - \hat{\pi}_{21})^2 + (\tilde{\pi}_{22} - \hat{\pi}_{22})^2}. \qquad (3)$$

Recall that the observed cell count for cell (i, j) (yielding the "observed" dissimilarities) is simply $\hat{\pi}_{ij}P$; recall P is the total number of binary variables (equivalently, the sum of the cell counts). Therefore a set of "smoothed" cell counts may be obtained by replacing $\hat{\pi}_{ij}$ by π^*_{ij}. Thus

$$a_{smooth} = \pi^*_{11}P = \left[\left(\frac{P}{P+\hat{\kappa}}\right)\hat{\pi}_{11} + \left(\frac{\hat{\kappa}}{P+\hat{\kappa}}\right)\tilde{\pi}_{11}\right]P,$$

with the other smoothed cell counts b_{smooth}, c_{smooth}, and d_{smooth} defined analogously for the $(1,2)$, $(2,1)$, and $(2,2)$ cells. The estimate (3) inherently safeguards against poorly chosen model-based estimates $\{\tilde{\pi}_{ij}\}$: If the observed proportions greatly contradict the $\{\tilde{\pi}_{ij}\}$, then the denominator of (3) will be very large, leading to $\{\pi^*_{ij}\}$ heavily weighting the $\{\hat{\pi}_{ij}\}$ values.

The smoothed dissimilarity may be calculated following, for example, formula (1), such that for objects k and k'

$$d_{kk'}^{smooth} = (b_{smooth} + c_{smooth})/P$$

(although any dissimilarity measure based on that "smoothed" 2×2 table could be used).

3 Simulation Study

In this section we describe a simulation study to measure the effect of smoothing the dissimilarities on the accuracy of the clustering of a binary data set. For a simulated data set of n objects (i.e., individuals), generated from a built-in clustering structure, we will measure "accuracy" via the statistic proposed by Rand (1971). For any partitioning of the objects, the Rand statistic gives the proportion of pairs of objects that are correctly placed either together or apart (depending on how the true structure places the pair of objects). Thus the Rand statistic is

$$\frac{m_1 + m_2}{n(n-1)/2}$$

where m_1 is the count of pairs of objects coming from the same subpopulation that are (correctly) placed in the same cluster and m_2 is the count of pairs of objects coming from different subpopulations that are (correctly) placed in different clusters. Thus the Rand statistic measures the concordance between the true clustering structure and the partition yielded by a clustering algorithm.

3.1 Setup of Simulations

We first discuss the rationale for our mechanism of generating the simulated data sets. We assume that our data consist of n objects, on which P binary variables are measured. As is common for binary data, we assume a latent continuous (specifically, normal) process, in which the binary observations are generated based on cutpoints. For example, assume $\mathbf{Y}_1^*, \dots, \mathbf{Y}_n^*$ are independent multivariate normal random vectors, having possibly different mean vectors:

$$\mathbf{Y}_i^* \sim N_P(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}),\ i = 1, \dots, n,\ j = 1, \dots, C,\ C < n.$$

Here C has a natural meaning as the true number of clusters in the data set. Then the binary random vectors $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ are generated by dichotomizing the normal values. Let the $n \times P$ data matrix $\mathbf{Y}^*$ have the rows $\mathbf{Y}_1^*, \dots, \mathbf{Y}_n^*$. For each element Y_{ip}^* in $\mathbf{Y}^*$, let

$$Y_{ip} = \begin{cases} 1 & \text{if } Y_{ip}^* \geq \xi_p, \\ 0 & \text{if } Y_{ip}^* < \xi_p, \end{cases}$$

where ξ_p is the cutpoint for the p-th latent variable. Then $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ are the rows of the matrix $\mathbf{Y}$ having elements $\{Y_{ip}\}$.

By assuming a noisy latent continuous structure, we end up with binary observations whose values may not reflect the true locations of the processes generating the data. The greater the amount of noise we incorporate in the continuous process, the less trustworthy are the resulting observed binary values. We now specify how the simulations were conducted.

For each data set in the simulation study, we generated a sample of $n = 50$ objects, each with $P = 8$ binary variables measured on it, from three subpopulations. The binary observations were obtained by generating (latent) multivariate $N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}), j = 1, 2, 3,$ vectors $\mathbf{Y}_1^*, \ldots, \mathbf{Y}_{50}^*$ and dichotomizing each entry in these vectors:

$$Y_i = \begin{cases} 1 & \text{if } Y_i^* \geq 0, \\ 0 & \text{if } Y_i^* < 0. \end{cases}$$

The three subpopulation mean vectors were generated randomly: $\boldsymbol{\mu}_1 \sim N(-\boldsymbol{\delta}, \mathbf{I}_P), \boldsymbol{\mu}_2 \sim N(\mathbf{0}, \mathbf{I}_P), \boldsymbol{\mu}_3 \sim N(\boldsymbol{\delta}, \mathbf{I}_P)$, where $\boldsymbol{\delta}$ equals a positive scalar δ times a vector of ones. (Once generated, $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\boldsymbol{\mu}_3$ were taken to be fixed and kept constant across simulated data sets.) Note that for larger δ, the cluster centers tend to be more dispersed, creating greater cluster separation. The within-cluster dispersion was controlled by $\boldsymbol{\Sigma}$; for most of the simulations, we chose $\boldsymbol{\Sigma} = \sigma \mathbf{I}_P$, with smaller σ yielding tighter clusters.

Once a $n \times P$ matrix of multivariate normal data was generated, the sample of binary data was obtained by transforming each nonnegative data value to 1 and each negative value to 0. For each of 5000 generated data sets, the clustering was carried out on this binary data set using, first, the standard methods based on unsmoothed dissimilarities and, second, methods based on smoothed dissimilarities discussed in this paper. The performance was judged by a comparison of the clustering results to the known three-cluster structure, as measured by the average (across data sets) Rand statistic value.

3.2 Parameter Settings and Results

We used a variety of different settings for the simulations. We allowed δ (measuring cluster center separation) to take the values $0.5, 1, 3$, or 5. We allowed σ (measuring within-cluster data dispersion) to take the values $1, 5, 10$, or 15. We examined the performance of the smoothed-dissimilarity method with three different choices of model for the cell probabilities for the 2×2 table of matches/mismatches: the independence model ($\tilde{\boldsymbol{\pi}} = \{\tilde{\pi}_{ij}\} = \hat{\pi}_{i+}\hat{\pi}_{+j}$), the equal-probability (EP) model $\tilde{\boldsymbol{\pi}} = (0.25, 0.25, 0.25, 0.25)'$, and the high-probability-of-match (HPM) model $\tilde{\boldsymbol{\pi}} = (0.4, 0.1, 0.1, 0.4)'$. Finally, for each simulated data set, the "best" 3-cluster partition of the objects was found using two distinctly different clustering algorithms: The first was a hierarchical clustering method, *average linkage*, implemented by the `R` function `hclust` (R Development Core Team, 2006). The other was a partitioning method, *K-medoids* (a robust analogue of the K-means algorithm), implemented by the `R` function `pam` (Kaufman and Rousseeuw, 1987) in the `cluster` package.

In all cases, the value of κ was determined using the objective Fienberg-Holland approach described above.

Note that, since we average across the 5000 data sets, the Monte Carlo standard error for each of these proportions (the Rand indices) is at most 0.007 (when the true Rand proportion is 0.5). The Monte Carlo standard error is probably much smaller in many cases presented here. Using this standard error, we may at least loosely judge which differences in Rand statistics are real and which are due to chance.

For the average linkage clustering method, the simulation results are given in Table 2 and presented graphically in Figure 1. The smoothing approach (in particular, smoothing toward the independence or equal-probability models) shows significantly better results in many situations. The biggest advantage

Table 2: Average Rand statistic values for the (average linkage) clusterings of the simulated binary data, based on the observed dissimilarities and three different "smoothed" dissimilarities.

Average Linkage Method

δ	$\sigma = 1$	$\sigma = 5$	$\sigma = 10$	$\sigma = 15$
0.5	0.8509 (o)	0.8599 (o)	0.7460 (o)	0.7576 (o)
	0.8945 (s/i)	0.9097 (s/i)	0.8400 (s/i)	0.8330 (s/i)
	0.9094 (s/E)	0.9159 (s/E)	0.8416 (s/E)	0.8295 (s/E)
	0.8204 (s/H)	0.8302 (s/H)	0.7292 (s/H)	0.7372 (s/H)
1	0.9380 (o)	0.8089 (o)	0.7601 (o)	0.7657 (o)
	0.9409 (s/i)	0.8477 (s/i)	0.8594 (s/i)	0.8500 (s/i)
	0.9497 (s/E)	0.8666 (s/E)	0.8576 (s/E)	0.8454 (s/E)
	0.9312 (s/H)	0.8040 (s/H)	0.7377 (s/H)	0.7476 (s/H)
2	0.9816 (o)	0.8925 (o)	0.8390 (o)	0.8211 (o)
	0.9848 (s/i)	0.9199 (s/i)	0.8990 (s/i)	0.8830 (s/i)
	0.9885 (s/E)	0.9383 (s/E)	0.8981 (s/E)	0.8782 (s/E)
	0.9813 (s/H)	0.8544 (s/H)	0.8116 (s/H)	0.8015 (s/H)
3	0.9627 (o)	0.9180 (o)	0.8803 (o)	0.8388 (o)
	0.9678 (s/i)	0.9255 (s/i)	0.9162 (s/i)	0.8933 (s/i)
	0.9811 (s/E)	0.9475 (s/E)	0.9302 (s/E)	0.8981 (s/E)
	0.9717 (s/H)	0.9283 (s/H)	0.8459 (s/H)	0.8173 (s/H)

NOTE: Each value is the average (across 5000 data sets) Rand statistic for the clusterings produced from an average linkage algorithm based on (top within each cell) the observed dissimilarities (o); (second within cell) the smoothed dissimilarities based on the independence model (s/i); (third within cell) the smoothed dissimilarities based on the equal-probability model (s/E); (bottom within cell) the smoothed dissimilarities based on the high-probability-of-match model (s/H).

Table 3: Average Rand statistic values for the (K-medoids) clusterings of the simulated binary data, based on the observed dissimilarities and three different "smoothed" dissimilarities.

	K-medoids Method			
δ	$\sigma = 1$	$\sigma = 5$	$\sigma = 10$	$\sigma = 15$
0.5	0.7663 (o)	0.5776 (o)	0.5687 (o)	0.5672 (o)
	0.7575 (s/i)	0.5603 (s/i)	0.5631 (s/i)	0.5637 (s/i)
	0.7549 (s/E)	0.5749 (s/E)	0.5695 (s/E)	0.5693 (s/E)
	0.7670 (s/H)	0.5746 (s/H)	0.5659 (s/H)	0.5644 (s/H)
1	0.7734 (o)	0.6536 (o)	0.6334 (o)	0.5665 (o)
	0.7508 (s/i)	0.6764 (s/i)	0.6521 (s/i)	0.5690 (s/i)
	0.7790 (s/E)	0.6645 (s/E)	0.6448 (s/E)	0.5706 (s/E)
	0.7692 (s/H)	0.6383 (s/H)	0.6155 (s/H)	0.5624 (s/H)
2	0.9432 (o)	0.7068 (o)	0.6817 (o)	0.6230 (o)
	0.8730 (s/i)	0.7195 (s/i)	0.6979 (s/i)	0.6457 (s/i)
	0.9039 (s/E)	0.7117 (s/E)	0.6880 (s/E)	0.6382 (s/E)
	0.9393 (s/H)	0.6979 (s/H)	0.6606 (s/H)	0.6077 (s/H)
3	0.8894 (o)	0.8007 (o)	0.7848 (o)	0.7477 (o)
	0.8174 (s/i)	0.7794 (s/i)	0.7722 (s/i)	0.7493 (s/i)
	0.8610 (s/E)	0.7721 (s/E)	0.7594 (s/E)	0.7408 (s/E)
	0.9183 (s/H)	0.8347 (s/H)	0.7829 (s/H)	0.7445 (s/H)

NOTE: Each value is the average (across 5000 data sets) Rand statistic for the clusterings produced from a K-medoids algorithm based on (top within each cell) the observed dissimilarities (o); (second within cell) the smoothed dissimilarities based on the independence model (s/i); (third within cell) the smoothed dissimilarities based on the equal-probability model (s/E); (bottom within cell) the smoothed dissimilarities based on the high-probability-of-match model (s/H).

Table 4: Average Rand statistic values for the clusterings of the simulated binary data with correlations.

Average Linkage Method

Correlation Structure	$\sigma = 1$	$\sigma = 5$	$\sigma = 10$	$\sigma = 15$
A	0.8505 (o)	0.7728 (o)	0.7358 (o)	0.7260 (o)
	0.8364 (s/i)	0.7939 (s/i)	0.7538 (s/i)	0.7574 (s/i)
	0.8968 (s/E)	0.8164 (s/E)	0.7794 (s/E)	0.7704 (s/E)
	0.8446 (s/H)	0.7565 (s/H)	0.7227 (s/H)	0.7052 (s/H)
B	0.9063 (o)	0.7636 (o)	0.7615 (o)	0.7669 (o)
	0.9066 (s/i)	0.8445 (s/i)	0.8263 (s/i)	0.8361 (s/i)
	0.9382 (s/E)	0.8483 (s/E)	0.8272 (s/E)	0.8375 (s/E)
	0.9024 (s/H)	0.7415 (s/H)	0.7396 (s/H)	0.7485 (s/H)

K-medoids Method

Correlation Structure	$\sigma = 1$	$\sigma = 5$	$\sigma = 10$	$\sigma = 15$
A	0.7856 (o)	0.5992 (o)	0.5998 (o)	0.5698 (o)
	0.7153 (s/i)	0.5830 (s/i)	0.5871 (s/i)	0.5561 (s/i)
	0.7529 (s/E)	0.5937 (s/E)	0.5945 (s/E)	0.5661 (s/E)
	0.7727 (s/H)	0.5975 (s/H)	0.5962 (s/H)	0.5687 (s/H)
B	0.7684 (o)	0.6021 (o)	0.6054 (o)	0.5786 (o)
	0.6779 (s/i)	0.5907 (s/i)	0.6078 (s/i)	0.5735 (s/i)
	0.7180 (s/E)	0.5989 (s/E)	0.6101 (s/E)	0.5804 (s/E)
	0.7819 (s/H)	0.5994 (s/H)	0.5965 (s/H)	0.5735 (s/H)

NOTE: Each value is the average (across 5000 data sets) Rand statistic for the clusterings produced from a K-medoids algorithm based on (top within each cell) the observed dissimilarities (o); and the smoothed dissimilarities based on: (second within cell) the independence model (s/i); (third within cell) the equal-probability model (s/E); (bottom within cell) the high-probability-of-match model (s/H).

for the smoothing approach is when σ is relatively large (the right sides of the plots in Figure 1). This corresponds to data having high within-cluster variability, the most difficult situation in which to get an accurate partition. It is for this sort of data for which we intuitively expect the smoothing to be of the greatest benefit.

For the K-medoids method, the simulation results are given in Table 3 and presented graphically in Figure 2. The comparison among methods is not especially conclusive when the K-medoids algorithm is used. For most situations, the various methods produce similarly accurate partitions. For $\delta = 1$ (medium spacing between clusters), smoothing toward the independence or EP model does slightly better. On the other hand, for $\delta = 3$ and $\sigma = 1$ (large dispersion between clusters, small dispersion within clusters), using the observed dissimilarities does notably better than smoothing toward the independence or EP model. Smoothing the dissimilarities toward the HPM model does the best overall in that situation, though. This latter situation is the easiest situation in which to get an accurate clustering. When there are tight clusters that are spaced far apart, smoothing toward a "default" model seems unnecessary.

In examining the Rand proportions, we see that the proportions based on the average linkage algorithm are generally somewhat higher than those based on the K-medoids method. Although we do not wish to definitively compare the different types of clustering algorithms here, we do note that the average linkage approach with any of the dissimilarity methods performed better than the K-medoids method with any of the dissimilarity methods, as least according to the Rand index.

The simulation results may provide guidelines for the admittedly tricky question of choosing the model toward which to smooth the dissimilarities. In the absence of any prior knowledge about the individual objects, we could smooth toward a noninformative model (independence or equal-probability) when we expect the data to be relatively noisy — that is, when we expect the within-cluster variability to be large relative to the spacing between cluster centers. If the clusters are tight and well-spaced, smoothing the dissimilarities toward a high-probability-of-match model seems to be a good strategy. (With binary data, such a prior judgment might be difficult to make based on initial data examination; perhaps a plot of principal component scores from an exploratory principal component analysis could lend some insight.) If we have prior expectations about how individual objects should be grouped, of course, we could subjectively choose different $\tilde{\boldsymbol{\pi}}$ values for different object pairs.

Letting $\boldsymbol{\Sigma} = \sigma \mathbf{I}_P$ implies that the set of P normal latent variables are mutually independent. To investigate the situation when the latent variables are correlated, we performed some of the simulations under two other correlation structures. Under correlation structure A, $\boldsymbol{\Sigma}$ has elements

$$\{\sigma_{ij}\} = \begin{cases} \sigma & \text{if } i = j, \\ 0.25\sigma & \text{if } i \neq j. \end{cases}$$

Correlation structure A assumes equal (positive) correlation among all the pairs of latent variables. Correlation structure B is a compromise between structure A and the uncorrelated case; under B, $\boldsymbol{\Sigma}$ has elements

$$\{\sigma_{ij}\} = \begin{cases} \sigma & \text{if } i = j, \\ 0.25\sigma & \text{if } i, j \in \{6, 7, 8\}, \\ 0 & \text{otherwise.} \end{cases}$$

Correlation structure B assumes three of the eight latent variables are pairwise positively correlated, while the other pairs are uncorrelated.

Table 4 (and Figure 3) give the results for clustering (using both average-linkage and K-medoids) under correlation structures A and B. (For all such results, the between-cluster dispersion parameter δ equaled 1.) We see that again, the differences among methods is more notable when using the average linkage algorithm. The results from the independent-data simulations are basically maintained under these correlation structures: Particularly for the average linkage clustering, smoothing toward either the independence model or the equal-probability model produces the most accurate clustering result.

Since the method proposed here relies on a smoothed version of the dissimilarities rather than those computed from the observed data, it is intuitive that the method should perform best when the data contain a large amount of noise (high σ). In such cases, the clusters formed using the observed dissimilarities can be trusted less to reflect the true structure of the data, and smoothing toward a reasonable model may produce valuable gains.

The simulation results may indicate that (at least in the absence of prior knowledge), smoothing the dissimilarities toward the independence model is a valid strategy. This represents a reasonably noninformative choice of model, and the results yielded from smoothing toward the independence model seem better or comparable to other choices for many of the settings examined.

4 An Application to Binary Test Data

In this section we apply the smoothed-dissimilarity clustering method to a real data set, the ACT mathematics test results for 2115 male examinees, studied in Ramsay and Silverman (2002) and made available on Silverman's web site `http://www.stats.ox.ac.uk/~silverma/fdacasebook/testitems.html` in plain text form. The data are given as a matrix of zeroes and ones having 2115 rows (representing the examinees) and 60 columns (representing the test items). An observation $y_{ij} = 0$ indicates that student i answered item j incorrectly, while $y_{ij} = 1$ indicates a correct response. For the purpose of this example, we used a subset of the overall data: We selected every fortieth row to obtain a smaller data set of 53 examinees and 60 test items.

4.1 Clustering the Test Items

The nature of this data set leads to two possible clustering problems. We could cluster the male students into natural groups, on the basis of their responses to the 60 items. We might also wish to cluster the 60 items into groups. In this latter case, the 53 selected students' scores (right/wrong) on any particular item would be the P binary variables measured on it. The results of the cluster analysis of the items may be easier to understand for this particular data set: While the students are listed in random order, the ACT test questions are typically meant to be ordered from easiest to most difficult. The clustering of items should therefore follow roughly according to the ordering of the items.

We do not know *a priori* the correct number of clusters; we used both the hierarchical average linkage algorithm and the nonhierarchical K-medoids algorithm, trying a variety of choices of C. We examined groupings based on $C = 2, 3, 4$, or 5 clusters. We clustered the items based on (1) the observed dissimilarities and (2) the smoothed dissimilarities based on smoothing toward the independence model. For the average linkage result, we can examine the dendrograms showing how the items were partitioned (see Figures 4 and 5). For the K-medoids result, the `clusplot` function in R allows us to plot the scores for the first two principal components (PCs) of the data and to identify the clusters in terms of their scores on the first two PCs. The best separation, according to these tools, appears to result from choosing $C = 3$ clusters. This choice of C also produced a reasonably large value for the "average silhouette width" (0.111 for $C = 3$ using the observed dissimilarities and 0.101 using the smoothed dissimilarities), a measure of the "goodness-of-separation" of a clustering partition (Rousseeuw, 1987). The clusplots based on the observed dissimilarities and based on the smoothed dissimilarities for $C = 3$ are shown in Figures 6 and 7. (The clusplots for $C = 4$ or greater showed much less clear separation.)

For the 3-cluster partition, the separation among clusters is only slightly different using the method based on the smoothed dissimilarities compared to using the observed dissimilarities. We now highlight the differences in the clustering partitions, which are displayed in Table 5 for the average linkage results and Table 6 for the K-medoids results.

The partition followed the natural item ordering fairly well, especially with the average-linkage results. For example, the first 18 (apparently easiest?) items were placed in cluster 1 by the average linkage algorithm using the observed dissimilarities; the the first 21 items were placed in cluster 1 by the average linkage algorithm using the smoothed dissimilarities. Test items 34 and 35 were placed in different clusters by the average linkage algorithm depending on whether the observed dissimilarities or smoothed dissimilarities were used.

The clustering results from the K-medoids algorithm differed rather more substantially from the numerical ordering of the test items. Several items, such as items 11, 20, 35, 40, and 46, were placed in different clusters by the K-medoids algorithm depending on whether the observed dissimilarities or smoothed dissimilarities were used. While for this real data example, it is impossible to say whether using the observed dissimilarities or the smoothed dissimilarities produces a “better” partition of the test items, clearly there were at least a few differences.

5 Conclusion

We have introduced a novel method of smoothing the dissimilarities among binary data as a preliminary step to cluster analysis. This method, described in Section 2, borrows ideas developed for the shrinkage estimation of cell probabilities in contingency tables. The simulation study in Section 3

Table 5: Table indicating average linkage clustering of the 60 test items into three clusters, based on data from 53 male students.

	Based on Observed Dissimilarities
Cluster	Test Items
1	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
2	19 20 21 22 23 24 25 26 27 28 29 30 31 32 33
3	34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60
	Based on Dissimilarities Smoothed Toward Independence Model
Cluster	Test Items
1	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
2	22 23 24 25 26 27 28 29 30 31 32 33 34 35
3	36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60

Table 6: Table indicating K-medoids clustering of the 60 test items into three clusters, based on data from 53 male students.

	Based on Observed Dissimilarities
Cluster	Test Items
1	1 2 3 4 5 6 7 8 9 10 12 15 17 18 21 23 24 25 26 27 28 29 31 32 34 35 38
2	13 14 16 19 20 22 30 36 37 39 40 41 42 44 45 46 47 50 51 52
3	11 33 43 49 53 54 55 56 57 58 59 60
	Based on Dissimilarities Smoothed Toward Independence Model
Cluster	Test Items
1	1 2 3 4 5 6 7 8 9 10 12 15 17 18 20 21 23 24 25 26 27 28 29 31 32 34 38
2	11 13 14 16 19 22 30 35 36 37 39 41 42 44 45 47 50 51 52
3	33 40 43 46 49 53 54 55 56 57 58 59 60

indicates that the smoothing method most effectively improves clustering accuracy in the most difficult situation for clustering: when the within-cluster data variability is high and when the true clusters do not have a large amount of separation. In Section 4, we apply the method to a test item response data set, in order to cluster the test items based on 53 students' binary (correct/incorrect) results.

The method presented in this paper may more accurately characterize the dissimilarities among noisy binary data by shrinking toward a particular smooth model. Furthermore, the nature of the shrinkage provides a safeguard such that the smoothing method for a reasonably well-chosen model will typically not produce significantly worse results compared to using the observed dissimilarities. The computationally straightforward nature of this smoothing method render it a viable option for investigators seeking to cluster binary data.

Acknowledgments

The authors are grateful to the associate editor and two anonymous referees for comments which resulted in an improvement of this paper.

References

[1] Albert, J. H., 1987. Empirical Bayes estimation in contingency tables. Comm. Statist. A—Theory Methods, 16, 2459-2485.

[2] Everitt, B., Landau, S. and Leese, M., 2001. Cluster Analysis. Edward Arnold Publishers, London.

[3] Fienberg, S. E., and Holland, P. W., 1973. Simultaneous estimation of multinomial cell probabilities. J. Amer. Statist. Assoc., 68, 683-691.

[4] Finch, H., 2005. Comparison of distance measures in cluster analysis with dichotomous data. J. Data Science, 3, 85-100.

[5] Hands, S. and Everitt, B., 1987. A Monte Carlo study of the recovery of cluster structure in binary data by hierarchical clustering techniques. Multivariate Behavioral Research, 22, 235-243.

[6] Hitchcock, D. B., Booth, J. G., and Casella, G., 2007. The effect of pre-smoothing functional data on cluster analysis. J. Stat. Comput. Simul., 77, 1043-1055.

[7] Johnson, R. A. and Wichern, D. W., 2002. Applied Multivariate Statistical Analysis. Prentice Hall, Upper Saddle River, N.J.

[8] Kaufman, L. and Rousseeuw, P. J., 1987. Clustering by Means of Medoids. In: Y. Dodge (Ed.), Statistical Data Analysis Based on the L_1 Norm, North Holland, Amsterdam, 405–416.

[9] Kaufman, L. and Rousseeuw, P. J., 1990. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, Inc., New York.

[10] Ramsay, J. O. and Silverman, B. W., 2002. Applied Functional Data Analysis: Methods and Case Studies. Springer-Verlag Inc., New York.

[11] Rand, W. M., 1971. Objective criteria for the evaluation of clustering methods. J. Amer. Statist. Assoc., 66, 846-850.

[12] R Development Core Team, 2006. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, (http://www.r-project.org/).

[13] Rousseeuw, P. J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math., 20, 53-65.

[14] Sokal, R. R., and Michener, C. D., 1958. A statistical method for evaluating systematic relationships. Univ. Kansas Sci. Bull., 38, 1409-1438.

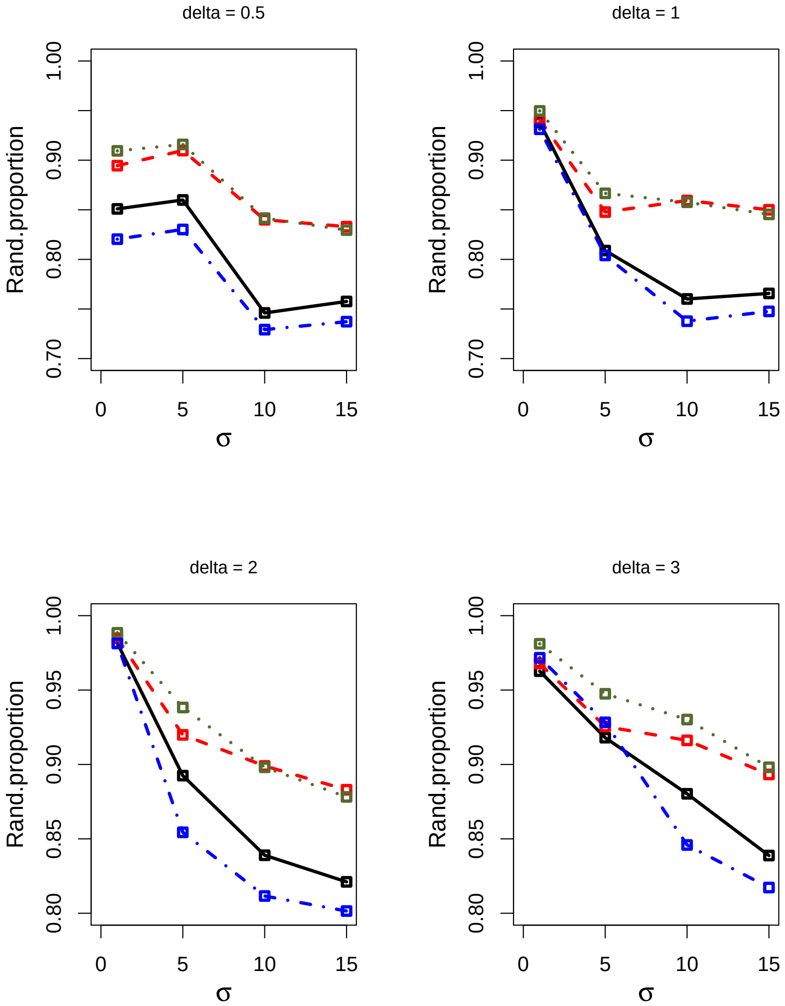

Figure 1: Rand proportions (averaged over 5000 simulated data sets) for average-linkage clusterings of the simulated data, based on the four different dissimilarity methods. Key: observed dissimilarities (solid line); smoothed dissimilarities based on the independence model (dashed line); smoothed dissimilarities based on the equal-probability model (dotted line); smoothed dissimilarities based on the high-probability-of-match model (dot-dash line)

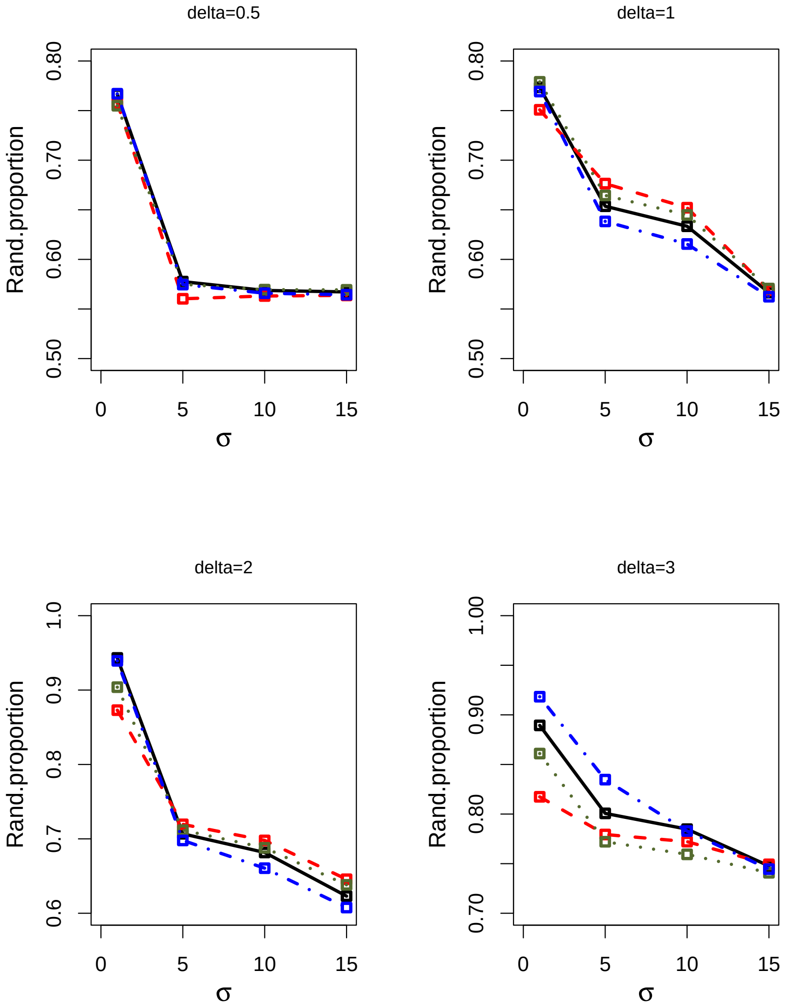

Figure 2: Rand proportions (averaged over 5000 simulated data sets) for K-medoids clusterings of the simulated data, based on the four different dissimilarity methods. Key: observed dissimilarities (solid line); smoothed dissimilarities based on the independence model (dashed line); smoothed dissimilarities based on the equal-probability model (dotted line); smoothed dissimilarities based on the high-probability-of-match model (dot-dash line)

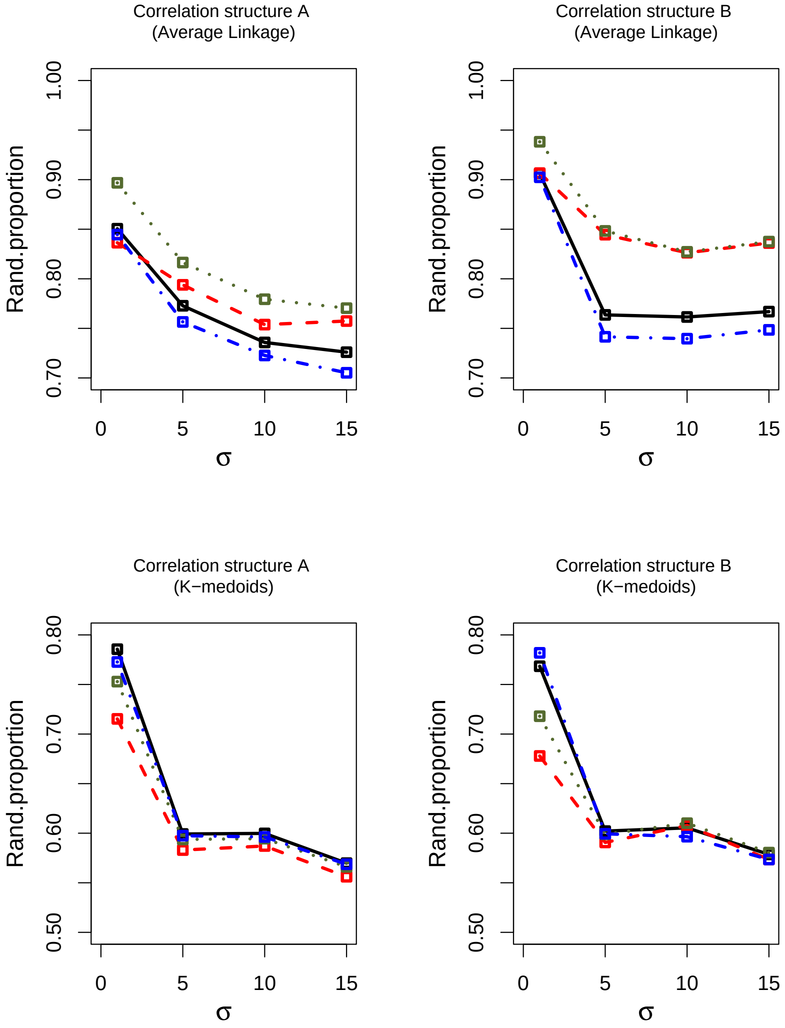

Figure 3: Rand proportions (averaged over 5000 simulated data sets) for clusterings of the simulated data with built-in correlation structures, based on the four different dissimilarity methods. Key: observed dissimilarities (solid line); smoothed dissimilarities based on the independence model (dashed line); smoothed dissimilarities based on the equal-probability model (dotted line); smoothed dissimilarities based on the high-probability-of-match model (dot-dash line)

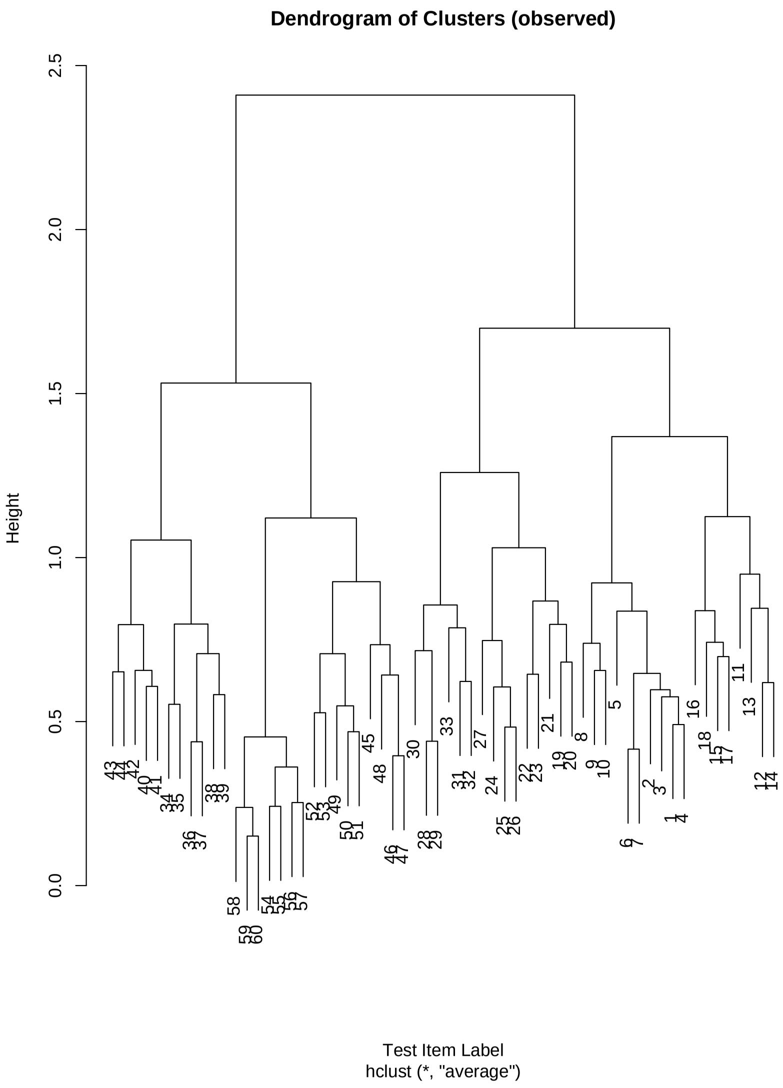

Figure 4: Dendrogram for the ACT test data, for the average-linkage clustering, based on the observed dissimilarities.

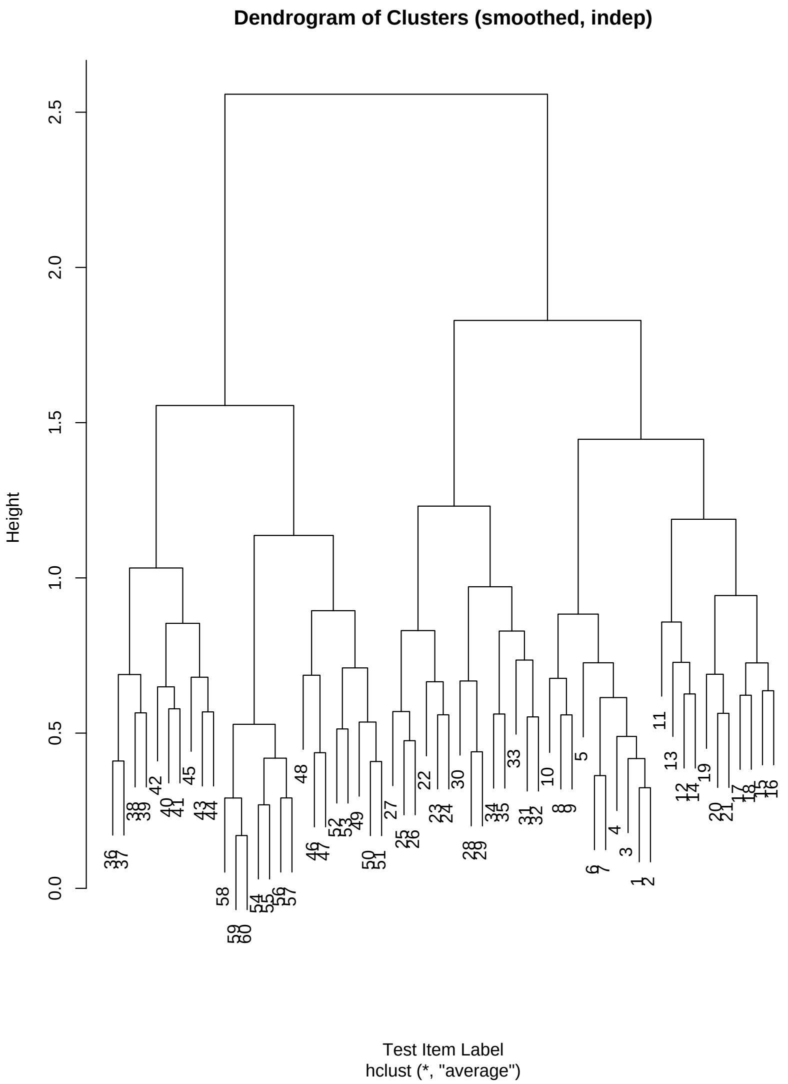

Figure 5: Dendrogram for the ACT test data, for the average-linkage clustering, based on the smoothed dissimilarities (smoothed toward independence model).

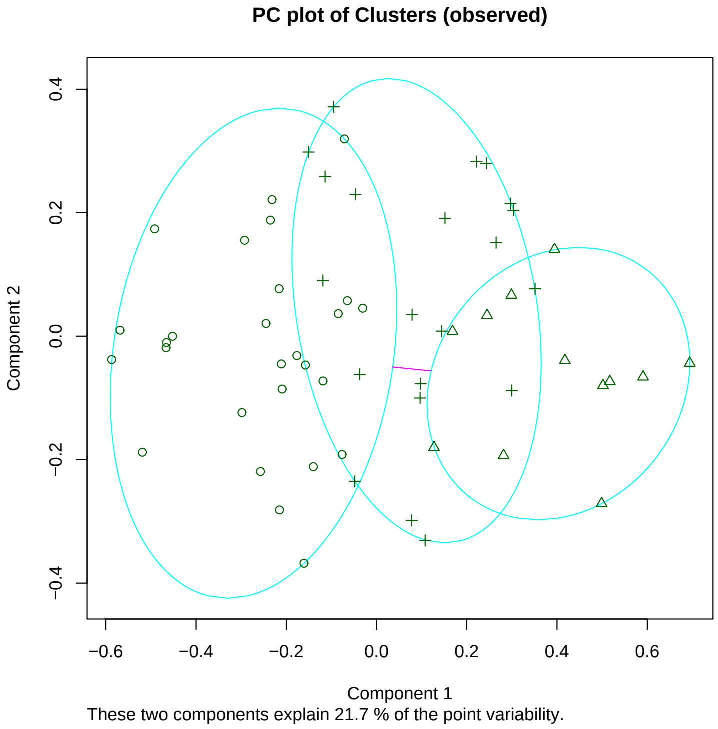

Figure 6: Plot of scores for first two principal components for the ACT test data, for the 3-cluster (K-medoids) partition, based on the observed dissimilarities.

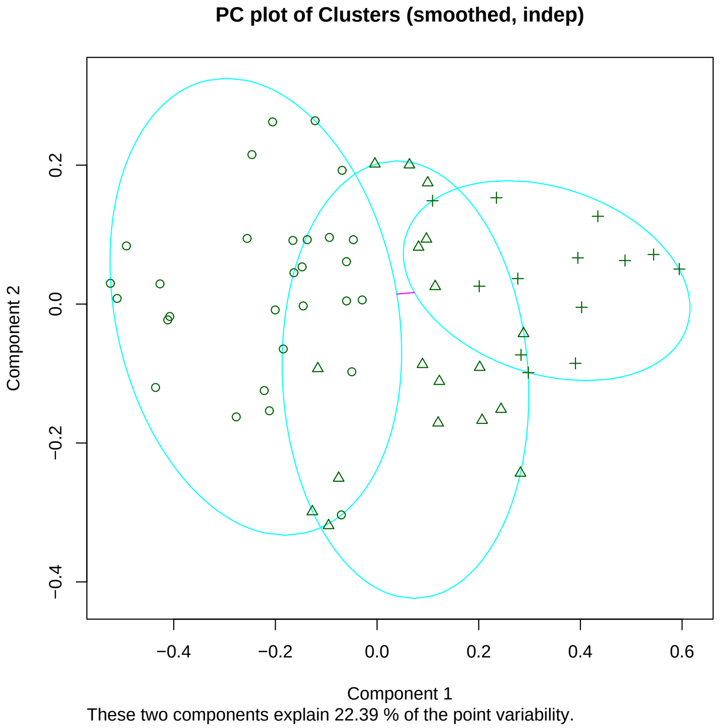

Figure 7: Plot of scores for first two principal components for the ACT test data, for the 3-cluster (K-medoids) partition, based on the smoothed dissimilarities.