Smoothing Dissimilarities for Cluster Analysis: Binary Data and Functional Data

David B. Hitchcock

University of South Carolina

Department of Statistics

Joint work with Zhimin Chen

University of South Carolina

Current and Future Trends

in Nonparametrics Conference

Columbia, SC

October 12, 2007

Outline

1. Introduction: Cluster Analysis and Dissimilarities
2. Background: Functional Data and Smoothing for Cluster Analysis
3. The Case of Binary Data and Smoothed Dissimilarities
4. Simulation Study
5. Example with Test Item Response Data

A Motivating Binary Data Example

- Data set: The binary measurement (1 = correct, 0 = incorrect) of the responses to 60 ACT (multiple-choice) test questions by a sample of students
- Goal: Determine whether 300 **students** fall into natural groups, based on test item responses
- Another Possible Goal: Determine whether 60 **items** fall into natural groups, based on students' responses
- Exploratory data analytic tool of cluster analysis can be used to answer the statistical questions posed in this example.

Cluster Analysis and Dissimilarities

- In cluster analysis, we partition N objects into k groups, often based on the pairwise dissimilarities among the objects.
- If observed data contain random variation, the pairwise distances will contain random error.
- The closer the dissimilarities in the data are to the "true," underlying dissimilarities between the systematic components of the data, the better the clustering result will be in showing the "true" clustering structure.
- With noisy data, smoothing before clustering likely to produce more accurate clusters.

Simulated Example

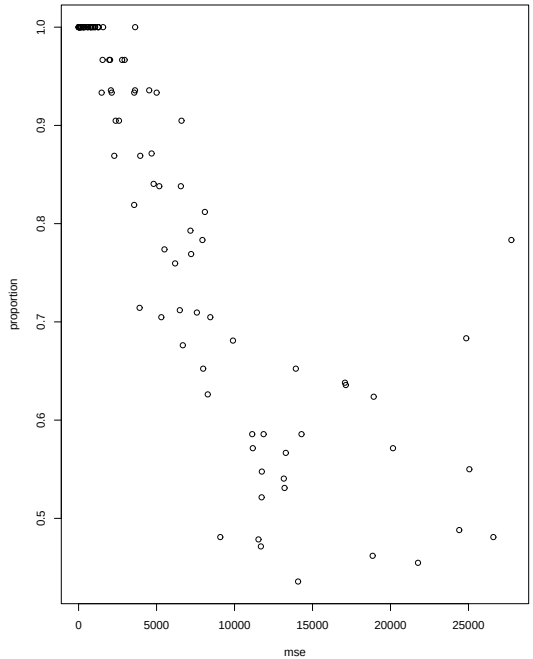

Figure 1: Proportion of correct groupings vs. MSE of dissimilarities.

- Plot: proportion of objects correctly grouped (by K-medoids algorithm) against mean squared discrepancy between "observed" dissimilarities and underlying dissimilarities.
- Trend: Negative association indicates that as dissimilarities get farther from the "truth," clustering success decreases.

Background Research

FUNCTIONAL DATA:

Possible model for discretized curves

- Functional observations are conceptually curves on domain $[0, T]$, but observed via vectors of n discrete measurements

$$\mathbf{y}_{ij} = \mu_i(t_j) + \boldsymbol{\epsilon}_{ij}, i = 1, \ldots, N, j = 1, \ldots, n$$

(Errors could be independent or dependent across measurement points)

- Dissimilarity metric for functional data: squared L_2 distance (or approximate version for discretized data)

$$\int_0^T [y_i(t) - y_j(t)]^2 \, dt,$$

Linear Smoothers and Cluster Analysis

- QUESTION: Will using a smoothed version of the data $\mathbf{S}\mathbf{y}_i$ result in better **clustering of the underlying curves** than using the raw observed data?
- Focus on **basis function** smoothing methods characterized by smoothing matrix $\mathbf{S}$ (symmetric, idempotent, rank r). (Examples: regression splines, Fourier series)
- In particular: $\mathbf{S}$ is, in our case, a B-spline smoother. For a cubic spline basis with m knots, $r = rank(\mathbf{S}) = m + 4$.
- We will investigate varying choices of m and thus r.

- We also use a James-Stein-type shrinkage smoother which is a weighted average of the observed data vector $\mathbf{y}$ and a linear smooth $\mathbf{Sy}$:

$$\mathbf{Sy}_i + \left(1 - \frac{n-r-2}{||\mathbf{y}_i - \mathbf{Sy}_i||^2}\right)_+ (\mathbf{y}_i - \mathbf{Sy}_i)$$

  for $i = 1, \ldots, N$, where $r = rank(\mathbf{S})$.

- Estimator gives more weight to $\mathbf{y}_i$ when $||\mathbf{y}_i - \mathbf{Sy}_i||^2$ is large and more weight to $\mathbf{Sy}_i$ when $||\mathbf{y}_i - \mathbf{Sy}_i||^2$ is small.

- $||\cdot||$ denotes the Euclidean norm, $(\cdot)_+$ denotes the positive part.

The Cluster Analysis Problem: Simulation Results

- Simulated data set: $N = 40$ sample (discretized, $n = 50$) noisy curves were generated from 4 distinct signal curves.
- Random noise was added: (1) independent $N(0, \sigma^2)$ errors with varying σ^2 and (2) Ornstein-Uhlenbeck errors with varying σ^2 and $\beta = 1$.
- The noisy curves were smoothed in two ways: (1) using a cubic B-spline basis smoother, and (2) using the B-spline smoother, shrunk with the James-Stein adjustment.
- The unsmoothed data and both sets of smoothed data were clustered using the K-medoids algorithm.
- The resulting clusterings were judged based on the Rand statistic: the proportion of pairs of objects correctly grouped (either together or apart, depending on the “truth” for each pair).

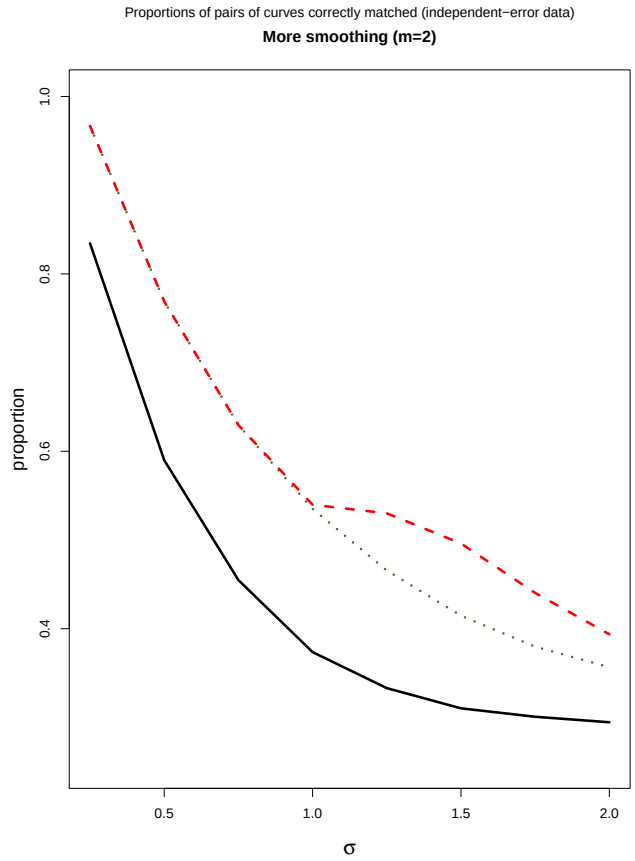

Figure 2: Solid line: Proportions based on observed data. Dotted line: Proportions based on B-spline approach. Dashed line: Proportions based on James-Stein approach.

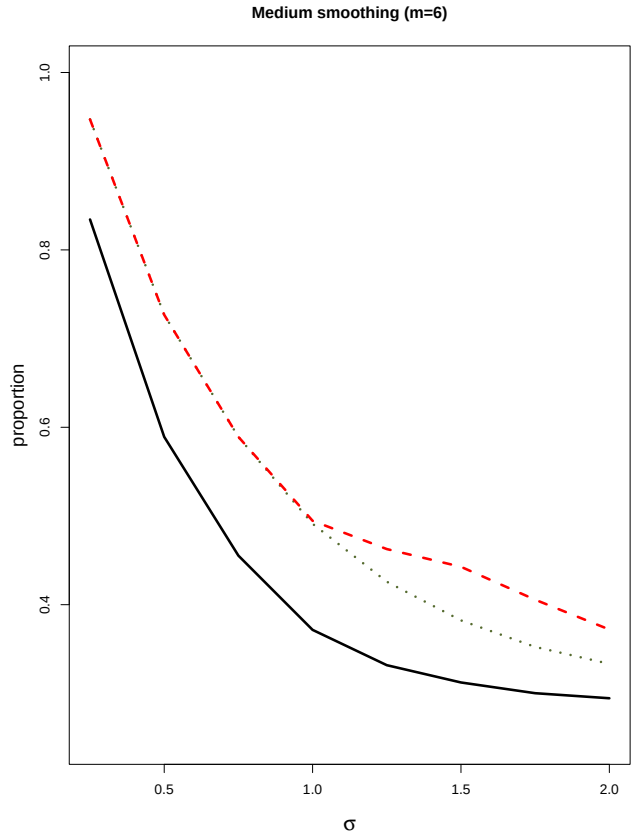

Figure 3: Solid line: Proportions based on observed data. Dotted line: Proportions based on B-spline approach. Dashed line: Proportions based on James-Stein approach.

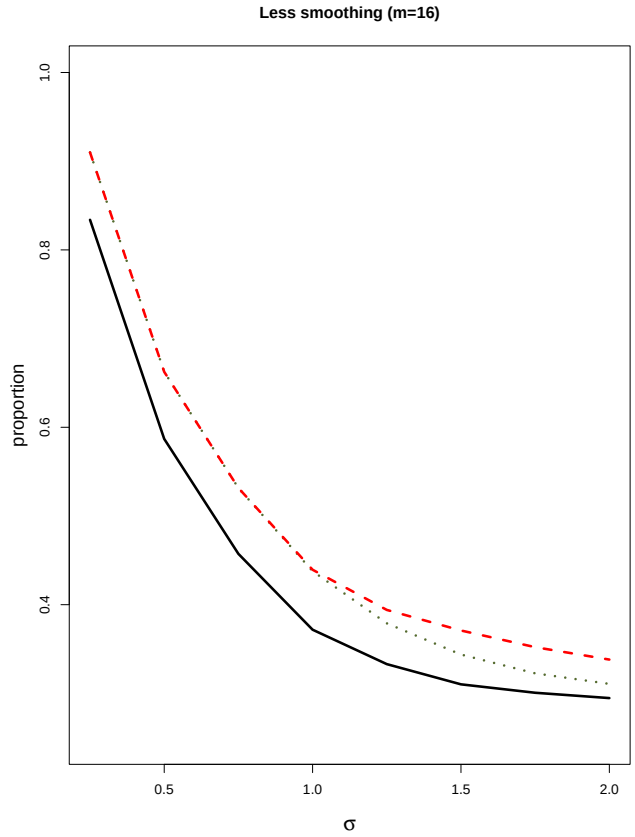

Figure 4: Solid line: Proportions based on observed data. Dotted line: Proportions based on B-spline approach. Dashed line: Proportions based on James-Stein approach.

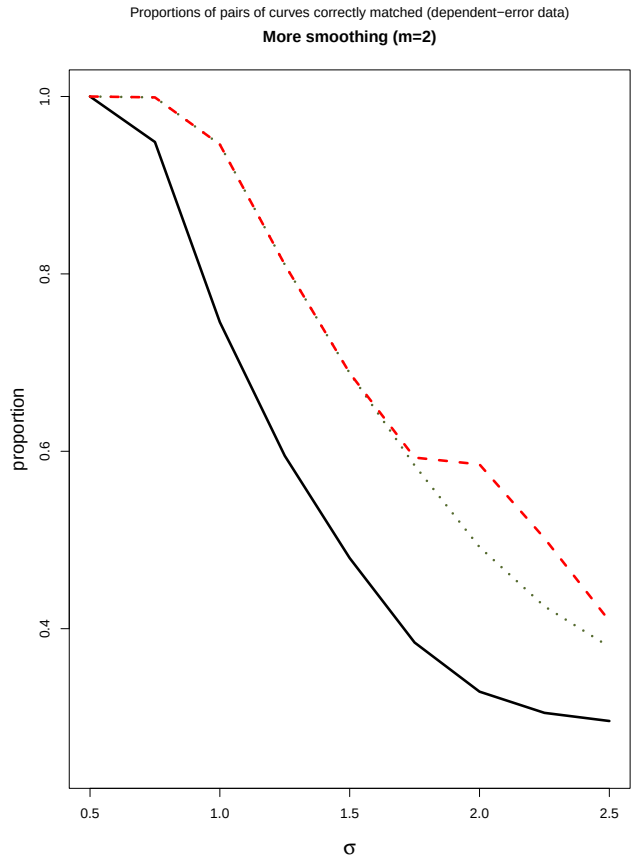

Figure 5: Solid line: Proportions based on observed data. Dotted line: Proportions based on B-spline approach. Dashed line: Proportions based on James-Stein approach.

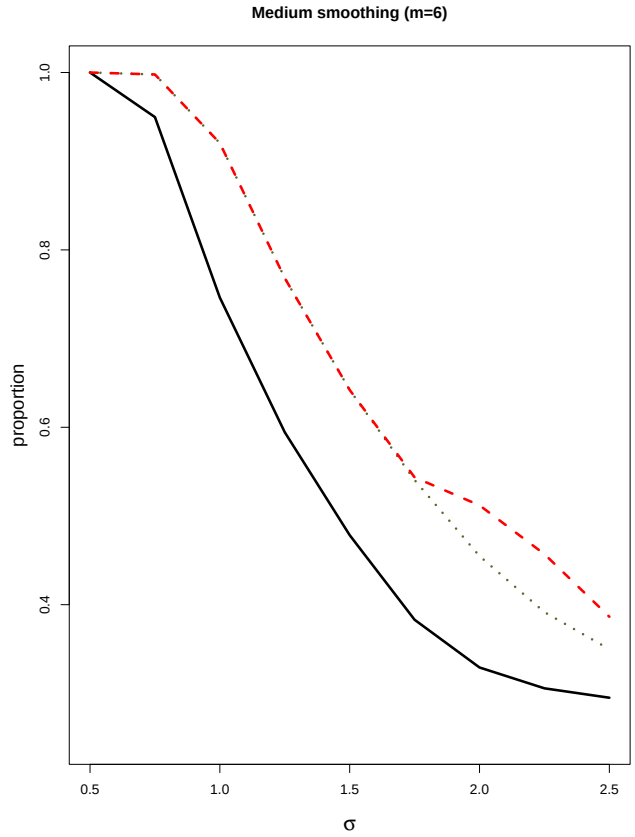

Figure 6: Solid line: Proportions based on observed data. Dotted line: Proportions based on B-spline approach. Dashed line: Proportions based on James-Stein approach.

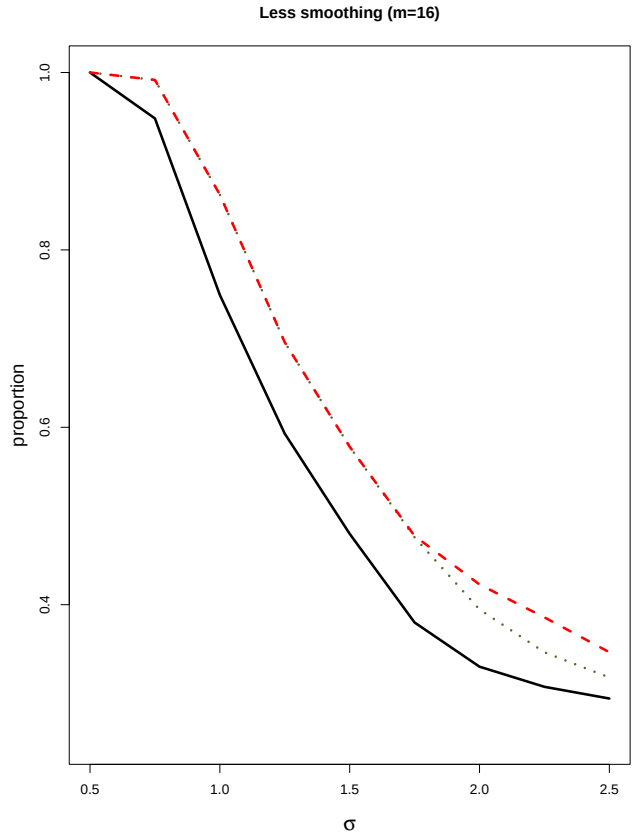

Figure 7: Solid line: Proportions based on observed data. Dotted line: Proportions based on B-spline approach. Dashed line: Proportions based on James-Stein approach.

New Situation:

Binary Data

- With binary data, the notion of "smoothing" is a less obvious one.
- Suppose underlying data come from a (noisy) continuous distribution, with "cut-points" generating the observed binary data.
- Assume data consist of n objects, on which P binary variables are measured:
- Assume a latent continuous (possibly normal) process (e.g., $\mathbf{Y}_i^* \sim N_P(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$, $i = 1, \ldots, n$, $j = 1, \ldots, C$, $C < n$.)
- Here C = the true number of clusters in the data set.
- Then the binary random vectors $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ are generated by dichotomizing the normal values.
- Clustering the observed binary data may not recover the separation between the groups in the underlying data.
- How to "smooth" binary data for cluster analysis?

Dissimilarities for Binary Data

- Cluster analysis of binary data typically based on dissimilarities that are a function of matches/mismatches for two objects

Table 1: Table listing number of matches and mismatches for a pair of objects.

Y	0	1	Totals
0	a	b	$a+b$
1	c	d	$c+d$
Totals	$a+c$	$b+d$	$P=a+b+c+d$

- Simple (mis)matching coefficient: $(b+c)/P$, where $P = a+b+c+d =$ number of variables
- Could shrink the values in this 2×2 table toward some model.
- Goal: "Smoothed" dissimilarities are more reflective of the true discrepancies among the signal components of the objects.

Shrinkage Estimation for Cell Probabilities in 2×2 tables

- Many authors have suggested estimators of π_{ij} of the form:

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

- Often result from Bayesian framework (putting Dirichlet prior on the set of probabilities $\pi_{11}, \ldots, \pi_{22}$). $\Rightarrow$ Resulting posterior Bayes estimates have this form (Fienberg and Holland 1973; Albert 1987).

- We use shrinkage estimates of the form:

$$\pi^*_{ij} = \frac{P}{\kappa + P}\hat{\pi}_{ij} + \frac{\kappa}{\kappa + P}\tilde{\pi}_{ij}$$

- κ is a parameter controlling the amount of smoothing

- Could choose κ subjectively, or Fienberg and Holland's Empirical Bayes approach provides a data-driven way estimate κ:

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

Shrinking the Dissimilarities among the Binary Data

Find estimators $\tilde{\pi}_{11}, \ldots, \tilde{\pi}_{22}$ using some model.

If investigator has no prior knowledge about clustering structure among binary objects, could choose some default/noninformative model:

- A model assuming independence within the 2×2 table:

  Example: $\tilde{\pi}_{ij} = \hat{\pi}_{i+}\hat{\pi}_{+j}, i = 1, 2, j = 1, 2$

  or

- a model assuming equal cell probabilities

  Example: $\tilde{\pi}_{ij} = 0.25, i = 1, 2; j = 1, 2.$

If investigator suspects a high likelihood of matches or mismatches between a pair of objects, could choose a more subjective model:

- a set of user-specified probabilities
  Example: $\tilde{\boldsymbol{\pi}} = (\tilde{\pi}_{11}, \tilde{\pi}_{12}, \tilde{\pi}_{21}, \tilde{\pi}_{22})' = (0.4, 0.1, 0.1, 0.4)'$.

If we have prior knowledge of the clustering structure of the objects, we could vary the $\tilde{\boldsymbol{\pi}}$ values across object pairs.

- Example: For two objects strongly suspected to belong to same cluster, assign $\tilde{\boldsymbol{\pi}} = (0.45, 0.05, 0.05, 0.45)'$.
- Example: For two objects suspected to belong to different clusters, assign $\tilde{\boldsymbol{\pi}} = (0.2, 0.3, 0.3, 0.2)'$.

- "Smoothed" cell probabilities π^*_{ij} are the linear combination of the observed proportions and the model-based probability estimates.

- "Smoothed" cell count for (1,1) cell:

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

  where $P = a + b + c + d$.

- Other smoothed cell counts defined similarly:
  $b_{smooth} = P\pi^*_{12}, c_{smooth} = P\pi^*_{21}, d_{smooth} = P\pi^*_{22}.$

- After values in the 2×2 table (for each pair of objects) are smoothed, calculate the smoothed dissimilarities by:

$$d_{ij}^{smooth} = \frac{b_{smooth} + c_{smooth}}{a_{smooth} + b_{smooth} + c_{smooth} + d_{smooth}}$$

- Use these smoothed dissimilarities as the inputs to a standard clustering algorithm.

Simulation Study

- Consider a sample of 50 objects generated from three subpopulations.
- Generate 3 clusters of multivariate ($P = 8$) normal latent data Y^*, adding built-in Gaussian noise (cluster sizes 20, 15, and 15).
- Variety of mean structures (cluster centers close together and far apart) and choices of within-cluster dispersion levels.
- Generated binary data Y by dichotomizing the latent normal data: $Y_i = 1$ if $Y_i^* \geq 0$; $Y_i = 0$ if $Y_i^* < 0$, for each data value and for each of the P variables.
- Result: Simulated 50 individuals with 8 binary variables measured on each of them.

Comparing the Two Methods

- Clustered the 50 individuals into 3 clusters via average linkage clustering and K-medoids clustering.
- Clustering was performed based on (1) the observed (unsmoothed) dissimilarities and (2) the smoothed dissimilarities.
- For smoothed dissimilarities, tried shrinking toward three types of model:
  - the independence model
  - the equal-probability model
  - the high-probability-of-match model.
- Used Rand statistic to judge resulting clusterings and determine which method more accurately partitioned the data into the true underlying clusters.
- Result: Using smoothed dissimilarities led to notable improvement in certain cases.
- Best results when noise level in data was large.

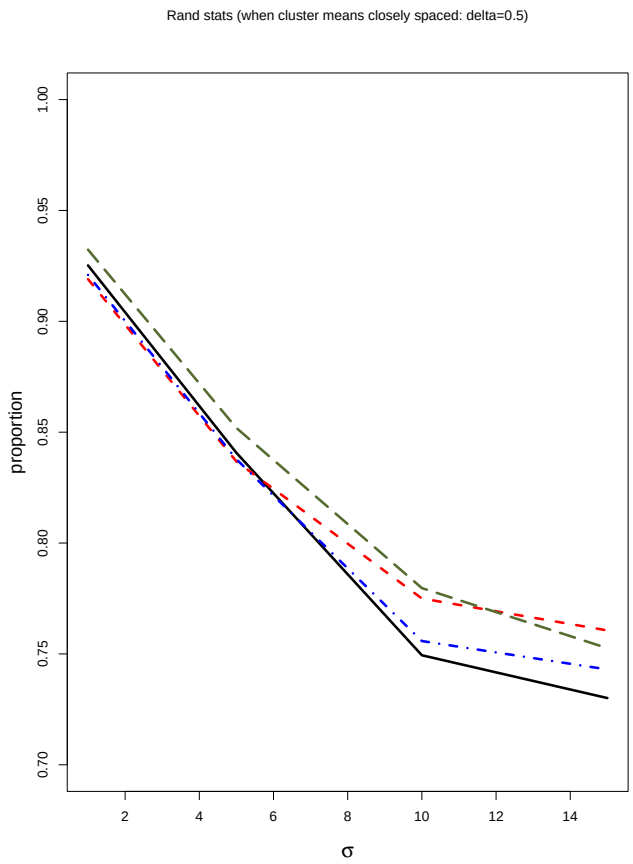

Figure 8: Rand Proportions, Average Linkage. Solid line: Based on observed dissimilarities. Dashed line: Based on smoothing toward independence model. Long-dash line: Based on smoothing toward equal-probability model. Dash-dotted line: Based on smoothing toward high-probability-of-match model.

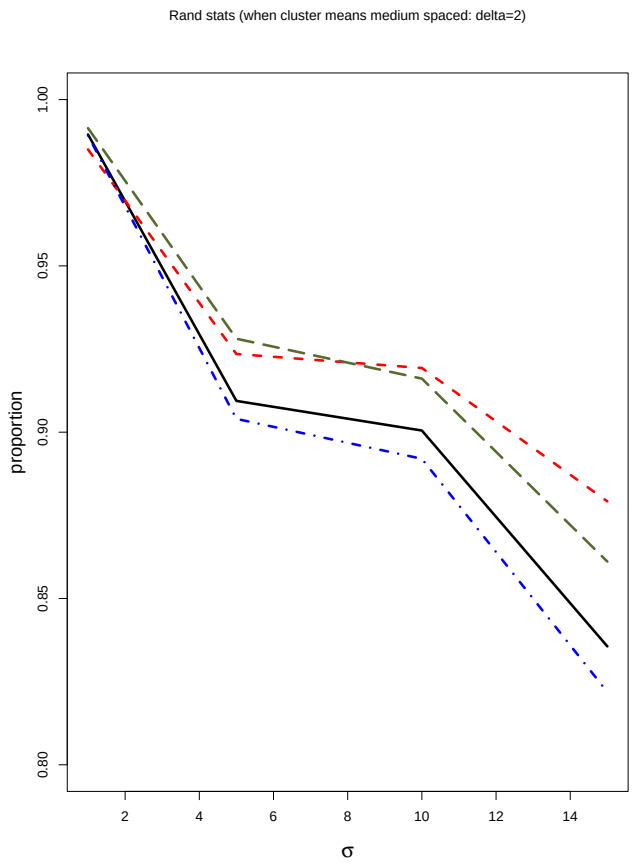

Figure 9: Rand Proportions, Average Linkage. Solid line: Based on observed dissimilarities. Dashed line: Based on smoothing toward independence model. Long-dash line: Based on smoothing toward equal-probability model. Dash-dotted line: Based on smoothing toward high-probability-of-match model.

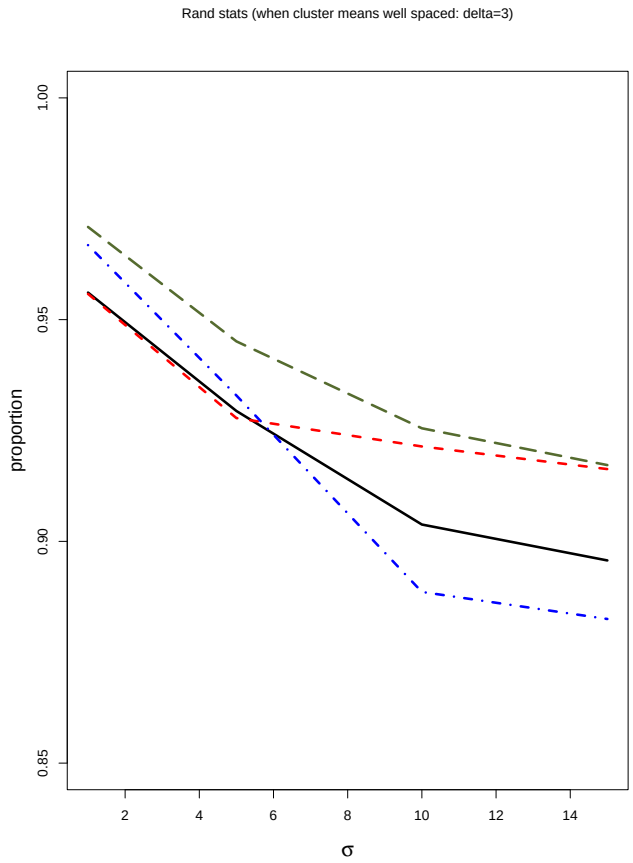

Figure 10: Rand Proportions, Average Linkage. Solid line: Based on observed dissimilarities. Dashed line: Based on smoothing toward independence model. Long-dash line: Based on smoothing toward equal-probability model. Dash-dotted line: Based on smoothing toward high-probability-of-match model.

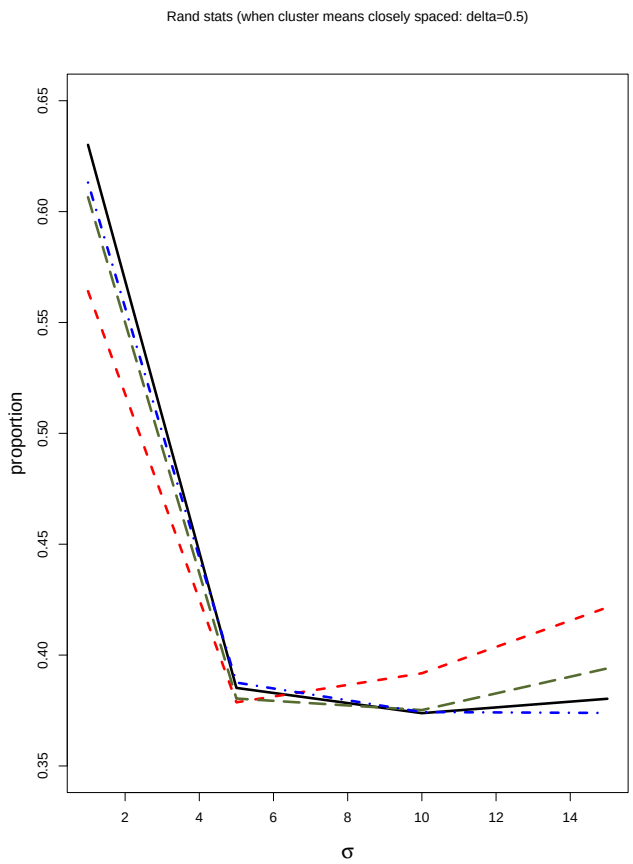

Figure 11: Rand Proportions, K-medoids Clustering. Solid line: Based on observed dissimilarities. Dashed line: Based on smoothing toward independence model. Long-dash line: Based on smoothing toward equal-probability model. Dash-dotted line: Based on smoothing toward high-probability-of-match model.

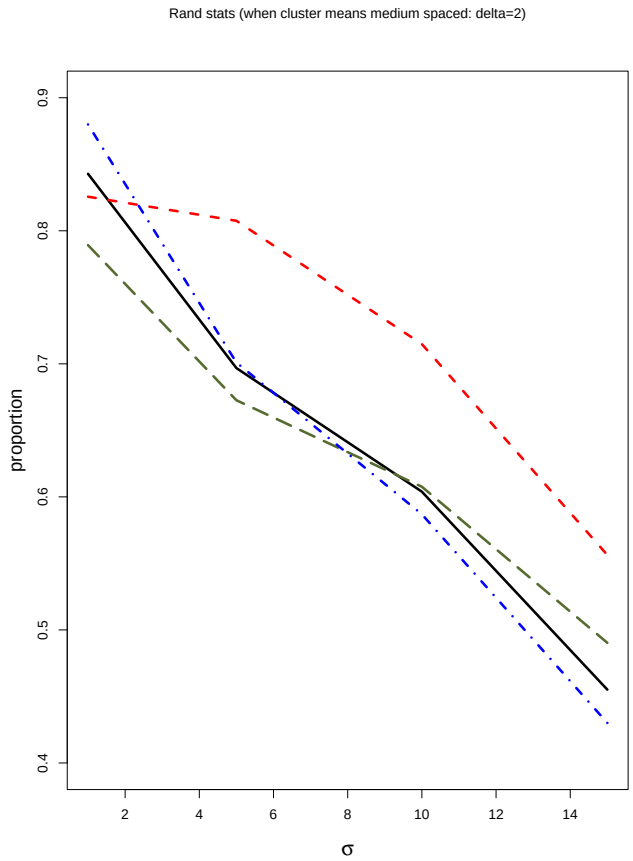

Figure 12: Rand Proportions, K-medoids Clustering. Solid line: Based on observed dissimilarities. Dashed line: Based on smoothing toward independence model. Long-dash line: Based on smoothing toward equal-probability model. Dash-dotted line: Based on smoothing toward high-probability-of-match model.

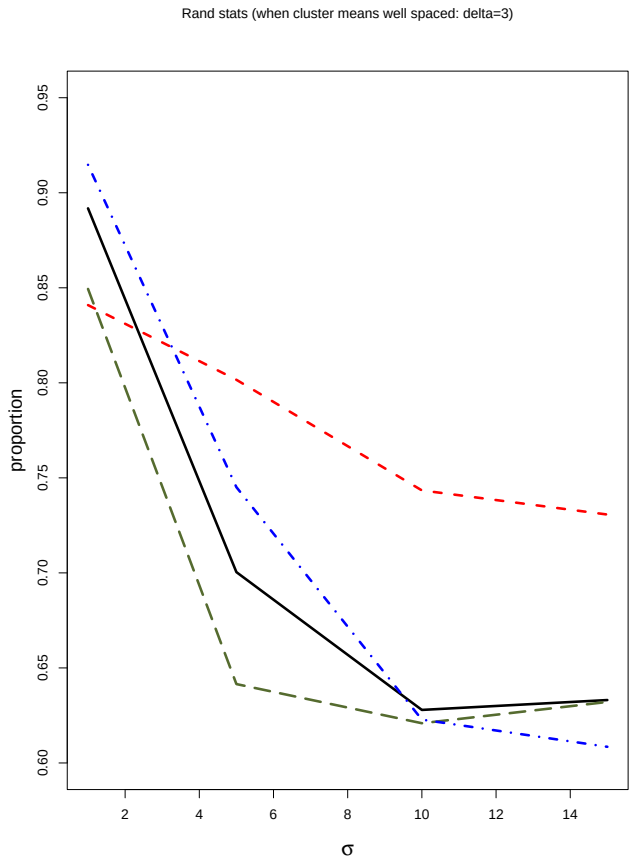

Figure 13: Rand Proportions, K-medoids Clustering. Solid line: Based on observed dissimilarities. Dashed line: Based on smoothing toward independence model. Long-dash line: Based on smoothing toward equal-probability model. Dash-dotted line: Based on smoothing toward high-probability-of-match model.

Example: Test Item Response Data Set

- Ramsay and Silverman (2002) presented ACT mathematics test results for 2115 male examinees.
- Data matrix (containing zeroes and ones) has 2115 students and 60 test items.
- Randomly selected 300 of the male students from the sample of 2115.
- An observation $y_{ij} = 0$ indicates that student i answered item j incorrectly, while $y_{ij} = 1$ indicates a correct response.

Example: Clustering the Test Items

- We treat the 60 test items as the observations (and the 300 students' responses as the binary variables for each item).
- Clustering solution would here place the test items into natural groups.
- Since ACT test questions are typically ordered from easiest to most difficult, clustering structure of items should resemble ordering of the items.
- The test items seemed to cluster best into 3 or 4 groups.
- The partition followed natural item ordering fairly well, with a few exceptions.

Table 2: Table indicating clustering of the 60 test items into four clusters, based on data from 300 randomly selected male students.

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

Table 3: Table indicating clustering of the 60 test items into four clusters, based on data from 300 randomly selected female students.

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

Conclusion

- There is justification for smoothing the data (or using a smoothed version of the dissimilarities) before engaging in techniques such as cluster analysis.
- The largest amount of improvement occurs especially when the data contain a high level of noise.

Future Research

- Show theoretically that the smoothed dissimilarities better estimate the true distance between cluster centers than the unsmoothed dissimilarities.
- Improve strategy for deciding what model to smooth toward.

References

- Albert, J. H. (1987), “Empirical Bayes Estimation in Contingency Tables”, *Communications in Statistics A, (Theory and Methods),* 16, 2459-2485.
- Fienberg, S. E., and Holland, P. W. (1973), “ Simultaneous Estimation of Multinomial Cell Probabilities”, *Journal of the American Statistical Association,* 68, 683-691.
- Hitchcock, D. B., Booth, J. G., and Casella, G. (2007), “The Effect of Pre-smoothing Functional Data on Cluster Analysis”, to appear in *Journal of Statistical Computation and Simulation*.
- Hitchcock, D. B. and Chen, Z. (2007), “Smoothing Dissimilarities to Cluster Binary Data”, submitted for publication.
- Ramsay, J. O. and Silverman, B. W. (2002), *Applied Functional Data Analysis: Methods and Case Studies*, New York, Springer-Verlag Inc.