

Clustering Smoothed Functional Data
David Hitchcock
Department of Statistics, University of Florida

- Often, the variables being measured on each observation are part of an underlying time process.
- These data, in the terminology of Ramsay and Silverman (1997), are called **functional data**.
- Idea: the underlying data are intrinsically curves; the observed vector \mathbf{x} is merely a discretized representation of the functional datum $x(t)$.
- In the analysis of functional data, the observed vector is typically converted to a curve via a smoothing method.
- In clustering functional data, it may be useful to smooth \mathbf{x} and cluster the smoothed curves rather than the observed data.

Bias-Variance Tradeoff

- As a functional measurement is smoothed more, its variance decreases, but it becomes more biased.
- Our hope: When clustering (less noisy) smoothed data, functional data which truly belong to the same cluster should appear more similar when represented as smooth curves.
- This leads to smaller within-cluster variability and a more apparent clustering structure. But if the model for the curves is misspecified, a bias is introduced.

Figure 1: Harmonic regression of the log expression ratio of a yeast gene

- If a deterministic method is used to cluster the smoothed data, we must choose a dissimilarity measure for two curves $x_i(t)$ and $x_j(t)$, e.g., the L_2 norm over one cell cycle

$$\| x_i(t) - x_j(t) \| = \left[\int_{t=0}^T (x_i(t) - x_j(t))^2 dt \right]^{1/2}.$$

- It can be shown that (if the basis expansion is properly orthonormalized), the L_2 distance between two (mean-centered) curves is proportional to the Euclidean distance between the coefficients of the basis functions.

- So we can use

$$d(x_i(t), x_j(t)) \propto [(\hat{\beta}_{1i} - \hat{\beta}_{1j})^2 + \dots + (\hat{\beta}_{pi} - \hat{\beta}_{pj})^2]^{1/2}.$$

- Then a standard cluster analysis method (like PAM or a linkage method) may be invoked, using the estimated coefficients of the data curves as input.

Simulation Study for Clustering Smoothed Functional Data

- Sixty functional observations (consisting of 20 measurements over time) were generated according to four different second-order Fourier curves, plus a random noise term.
- The observations were grouped into four clusters via two methods: a PAM clustering of the **raw data**, and a PAM clustering of the estimated coefficients of a first-order Fourier model (**clustering the smoothed data**).
- Since the fitted model was intentionally misspecified, it contained a built-in bias. Also, the noise term contributed to a variance in the raw data.

Simulation Results: Clustering With and Without Smoothing

std. dev. σ:	0.2	0.4	0.6	0.8	1.0
bias = 0.06	.979	.986	1.074	1.220	1.317
bias = 0.12	.856	.963	1.041	1.206	1.263
bias = 0.20	.782	.893	1.001	1.185	1.304
bias = 0.25	.767	.846	.964	1.142	1.240
bias = 0.30	.770	.795	.922	1.147	1.238

Table 1: Relative Performance of Clustering: With Smoothing versus Without Smoothing.

- Numbers in the table are ratios of the proportion of pairs of objects correctly grouped by the smoothing method, versus by the raw-data method.
- Ratios greater than 1 indicate a superior performance by the smoothing method.
- Conclusion: Smoothing the data is most beneficial when the observed data are noisy, especially when the chosen model has low bias.