Integrating Fuzziness into Vector Autoregressive (VAR)-Based Clustering

Jonathan J. Park, Sy-Miin Chow, & Peter C. M. Molenaar
The Pennsylvania State University

Modeling Dynamic Processes

Vector Autoregression (VAR) models provide an intuitive framework from which we may model relations among a set of variables through time. The VAR of lag order 1 may be represented as:

$$\boldsymbol{\eta}(t) = \boldsymbol{\mu} + \boldsymbol{\Phi}^{*}\boldsymbol{\eta}(t-1) + \zeta^{*}(t) \tag{1}$$

$$\zeta^{*}(t) \sim N(0, \boldsymbol{\Psi}^{*}) \tag{2}$$

Clustering on Dynamic Models

VAR-based clustering methods have seen significant strides in recent years due to calls for methods geared towards identifying nomothetic trends in idiographic processes.
VAR-based clustering methods may be utilized when within-sample heterogeneity is expected to affect the quality of parameter estimates obtained from fitting VAR models to groups of subjects.
Clustering methods for VAR models have primarily focused on "hard" methods/algorithms where all subjects belong to one cluster or another (i.e., **A** or **B**). A popular hard clustering method is the k-means clustering algorithm which identifies k clusters of subjects by minimizing this objective function:

$$J(X) = \sum_{j=1}^{C}\sum_{i=1}^{N_j} \|\mathbf{x}_i - \bar{\mathbf{x}}_j\|^2 \tag{3}$$

Clusters which force individuals into bins may increase the heterogeneity within clusters and thus counteract the benefits of clustering to begin with. Fuzzy clustering methods have been proposed as an alternative method which allows for subjects to belong to all clusters by varying *degrees*. The fuzzy c-means algorithm is the fuzzy extension of the k-means algorithm and optimizes the following objective function:

$$J(U, X) = \sum_{j=1}^{C}\sum_{i=1}^{N} \mu_{ij}^{m} \|\mathbf{x}_i - \bar{\mathbf{x}}_j\|^2 \tag{4}$$

where **U** is the membership degree matrix that relates how strongly a subject belongs to any of the c-clusters and is calculated as:

$$\mu_{ij}^{m} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\|\mathbf{x}_i - \bar{\mathbf{x}}_j\|}{\|\mathbf{x}_i - \bar{\mathbf{x}}_k\|}\right)^{\frac{2}{m-1}}} \tag{5}$$

Our research questions were thus:

- What is the effect on parameters estimates when fuzzy subjects are erroneously included in group-level models?
- What conditions affect the our ability to confidently classify subjects as "fuzzy" or not?

Methods & Design

A Monte Carlo simulation was conducted across 100 replications to address the two hypotheses originally posited.

Design Parameters	Values
Sample Size (N)	30 ($n_{sg1} = 10; n_{sg2} = 10; n_{fg} = 10$)
Times (T)	250, 500
Subgroup Distances	1, 3, 9

Data were generated using two 10-variate VAR(1) models. Subjects were assigned to one of 3 conditions: Subgroup 1, Subgroup 2, and a fuzzy class. Subjects in the fuzzy class were simulated such that they were drawn from a 50% mixture distribution of the prior two subgroup models.

RMSEs Across Time and Subgroup Distance

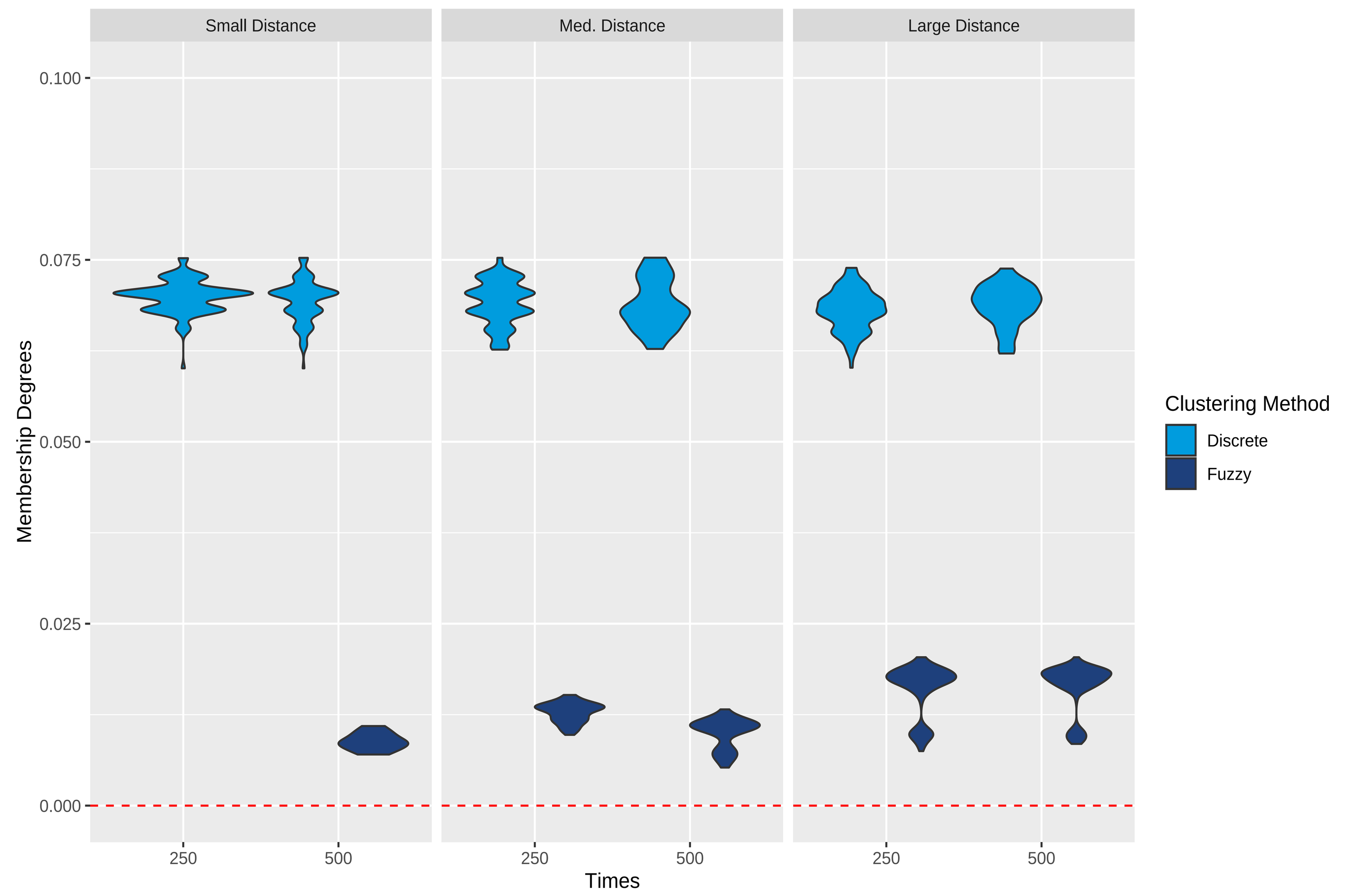

Figure 1: RMSEs as a function of Time and Subgroup Distances

Conclusions

- What is the effect on parameter estimates when fuzzy subjects are erroneously included in group-level models?
  - Group-level models formed by hard clustering had greater biases and RMSEs than those generated with fuzzy methods
  - Biases and RMSEs generally improved with more time-points but did not seem to improve with greater subgroup distance
- What conditions affect our ability to classify subjects as "fuzzy" or not
  - Greater subgroup separation–defined by unique cross-regression coefficients–increased our ability to clearly separate distinct subjects from fuzzy ones
  - More time-points was associated with better ability to separate distinct subjects from fuzzy ones

Biases Across Time and Subgroup Distance

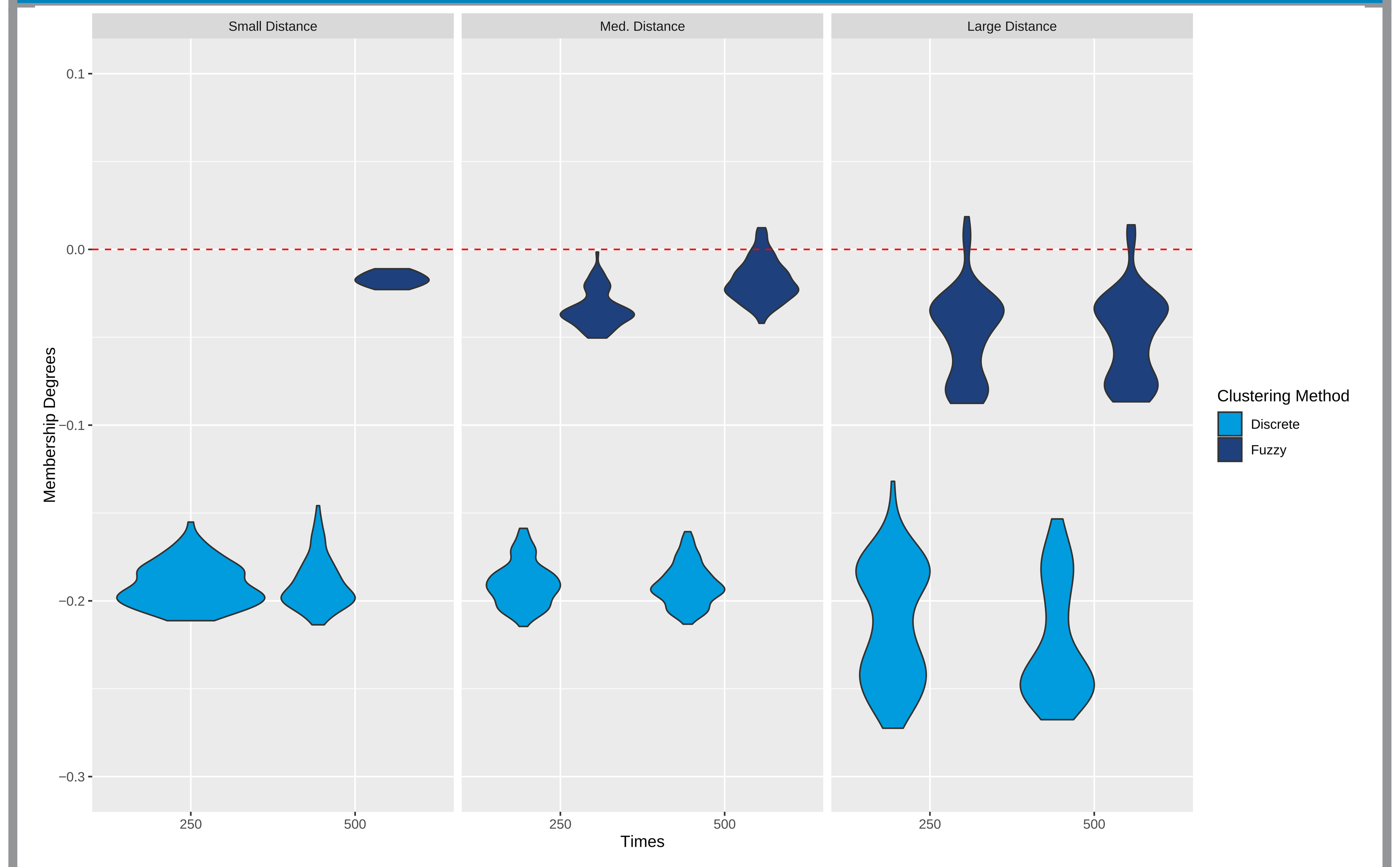

Figure 2: Biases by Time and Subgroup Distances

Separation of Fuzzy and Distinct Subjects

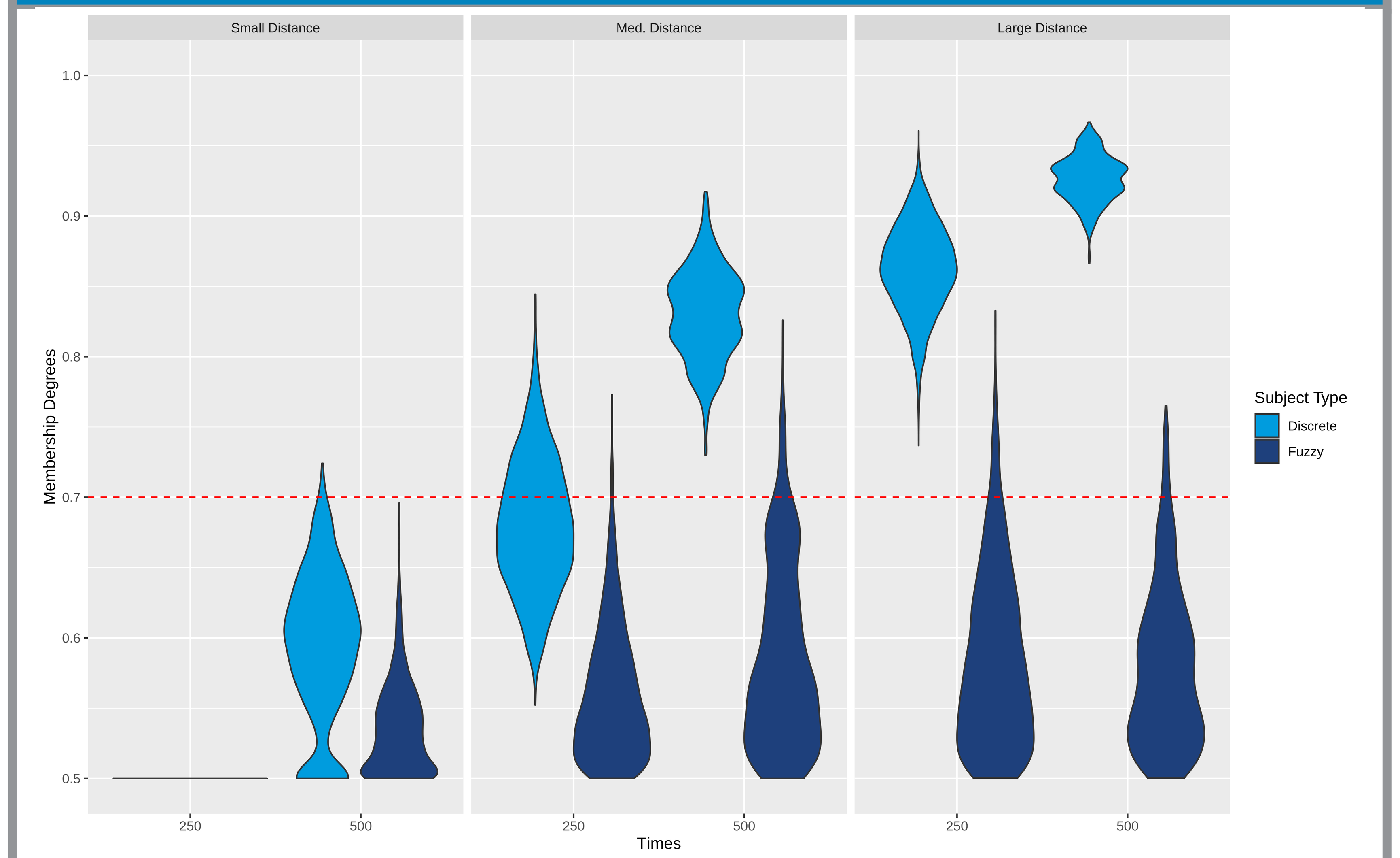

Figure 3: Membership Degrees by Time and Subgroup Distances

Acknowledgements

- Research reported in this publication was supported by the Intensive Longitudinal Health Behavior Cooperative Agreement Program funded by the National Institutes of Health under Award Number U24AA027684 and UL1TR002014-06; and National Science Foundation grant IGE-1806874.