MODEL-BASED LONGITUDINAL CLUSTERING
WITH VARYING CLUSTER ASSIGNMENTS

Daniel K. Sewell[1], Yuguo Chen[2], William Bernhard[2] and Tracy Sulkin[2]

[1]*University of Iowa and* [2]*University of Illinois, Urbana-Champaign*

Abstract: It is often of interest to perform clustering on longitudinal data, yet it is difficult to formulate an intuitive model for which estimation is computationally feasible. We propose a model-based clustering method for clustering objects that are observed over time. The proposed model can be viewed as an extension of the normal mixture model for clustering to longitudinal data. While existing models only account for clustering effects, we propose modeling the distribution of the observed values of each object as a blending of a cluster effect and an individual effect, hence also giving an estimate of how much the behavior of an object is determined by the cluster to which it belongs. Further, it is important to detect how explanatory variables affect the clustering. An advantage of our method is that it can handle multiple explanatory variables of any type through a linear modeling of the cluster transition probabilities. We implement the generalized EM algorithm using several recursive relationships to greatly decrease the computational cost. The accuracy of our estimation method is illustrated in a simulation study, and U.S. Congressional data is analyzed.

Key words and phrases: Cluster analysis, EM algorithm, multinomial logistic regression, normal mixture models, time series.

1. Introduction

Numerous applications exist in which it is of interest to partition the data into homogeneous groups, or clusters. With longitudinal data in which we observe many objects over a series of time points, one often wishes to better understand how and why these objects are grouped and how they transition from one group to another over time. Longitudinal clustering is an important topic in such fields as genomics, clinical research, political science, etc.

Much of the previous work on clustering longitudinal data focuses on curve clustering, also called trajectory clustering. Curve clustering methods view the time series as curves or sometimes as random functions, and intend to give cluster assignments of the objects based on shape similarity. For example, Ray and Mallick (2006) used Bayesian wavelets to model the curves, and Luan and Li (2003) used B-splines to model the cluster means; often this type of longitudinal

clustering is achieved by regression-based approaches (see, e.g., Lou, Jiang, and Keng (1993); Gaffney and Smyth (1999)).

More related to our work is latent transition analysis (LTA) (see, e.g., Graham et al. (1991); Collins and Wugalter (1992); Collins et al. (1994)), which is commonly used in the social sciences. LTA assumes that observed categorical responses are noisy measurements of sequential latent stages; i.e., we can imperfectly measure over time a subject's latent stage (at each time point). Vermunt, Langeheine, and Bockenholt (1999) and Chung, Park, and Lanza (2005) were able to incorporate extra covariates in their model to help estimate the latent stages. However, these methods are focused on categorical panel data and sequential stages hypothesized a priori which are strongly associated with the measured categorical variables.

We are concerned with clustering vectors of continuous data observed over time, and where the clusters are based on locations rather than the shape of the trajectories, i.e., we cluster based on the objects' observed values rather than the shape of the objects' time series. A simple algorithm for this context is the extension of the k-means method for longitudinal data (Genolini and Falissard (2010)). This does not, however, account for the correlation between each object at different time points, but simply vectorizes each object's time series, and then performs k-means based on the Euclidean distance (or some other metric) of the vectorized time series. Further, this method requires the objects' time series to all be of the same length, which may not be the case in practice. A model based approach which accounts for correlation across time points was given by De la Cruz-Mesía, Quintana, and Marshall (2008), who extended the model consisting of mixture of normal distributions (see Banfield and Raftery (1993)) to univariate time series data. Further work in the context of mixture of normal distributions was done by McNicholas and Murphy (2010), who derived estimates for specific covariance constraints corresponding to univariate time series data. Anderlucci and Viroli (2014) built on these works to handle multivariate data.

Scott, James, and Sugar (2005) used a hidden Markov model to estimate the cluster assignments which were allowed to differ for each object at each time point, estimation done using Markov chain Monte Carlo methods. While a vast improvement, this work still assumes local independence, ignoring individual effects that may be important in modeling the data, and is not flexible enough to handle multiple explanatory variables of varying types.

Here we consider data consisting of multiple variables (the response variables) at many (and possibly at a varying number of) time points from many objects. We wish to cluster these objects at each time point based on the similarity of their responses. The number of clusters is assumed to be fixed, and the clusters themselves are assumed to be static in the sense that the structures

that define each cluster are constant over time. These defining structures will be described in Section 2. We propose a model-based clustering algorithm for longitudinal data that allows cluster assignments to change over time and borrows information across each object's time series when making these assignments, thus allowing the researcher to better understand how these objects transition from one group to another. We incorporate temporal dependence into the model by modeling each object's current location as a blending of the current (unobserved) cluster assignment and the object's previous values. Our model allows explanatory variables of any form to be incorporated into the model in order to explain the clustering by incorporating multinomial logistic regression into the model. This differs from such other clustering models that incorporate covariates as those of De la Cruz-Mesía, Quintana, and Marshall (2008) and Anderlucci and Viroli (2014), in that the group memberships of the objects are predicted by the covariates. This results in clustering being performed purely on the response variables while simultaneously learning how the covariates explain the clustering results. Estimation is accomplished by using a generalized EM algorithm. The computational cost of a straightforward implementation of this algorithm can be, however, prohibitively high. To address this, we derive several recursive relationships which greatly reduce the computational cost from exponential to linear with respect to the number of time points.

The rest of the paper is organized as follows. Section 2 describes the proposed model. Section 3 outlines the details of the generalized EM algorithm. Section 4 provides the results of a simulation study. Section 5 gives the analysis of longitudinal data collected on Democrats serving in the U.S. House of Representatives. Section 6 gives a brief discussion.

2. Models

We first present in Section 2.1 the basic model for clustering longitudinal data. We then extend this model in Section 2.2 to model how covariate information may explain the clusters of the response variables. One can then partition the data into clusters and discover how covariates can explain the clustering assignments.

2.1. Non-regressed cluster probabilities

Denote the value of the ith object at time t as $\mathbf{X}_{it}$, and let $\mathcal{X}_i = (\mathbf{X}'_{i1}, \ldots, \mathbf{X}'_{iT_i})'$, for $i = 1, \ldots, n$, denote the data to be clustered, where n is the number of distinct objects, T_i is the length of the ith time series, and the dimension of each $\mathbf{X}_{it}$ is p. We assume that each object is initially assigned to one of K clusters according to the $1 \times K$ probability vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ (K is a fixed positive integer assumed to be known). The notation to be used throughout lets

Z_{it} be the ith object's cluster assignment at time t. Hence $\mathbb{P}(Z_{i1} = k) = \alpha_k$. For subsequent time points, these objects are assigned to a cluster according to the $K \times K$ transition matrix $\boldsymbol{\beta}$, $\mathbb{P}(Z_{it} = k | Z_{i(t-1)} = h) = \beta_{hk}$.

Hidden Markov models are in wide use in many areas (Frühwirth-Schnatter (2006)). Such usefulness motivates us to borrow principles from this class of models. As will be seen, each cluster in our context is defined by a specific static probability distribution. A hidden Markov model applied to our setting implies that if we knew which cluster a particular object belonged to at each time point, the observed temporal observations from the object would look like random samples from the clusters' corresponding distributions. In practice one would expect these observations to still be dependent in some way, even after accounting for the clustering. For example, in the context of clustering human behaviors, we might expect a person's future behavior to depend on that person's past behavior as well as on the influence exerted on him/her by the group to which he/she belongs. Then the current expected behavior of a person should be some blending of previous behavior (individual effect) and the overall expected behavior from someone coming from the group to which the person belongs (cluster effect). This more general type of dependency is known as a Markov switching model.

We make two assumptions about the dependence structure of the model: given all past cluster assignments, the current cluster assignment depends only on the cluster assignment at the previous time point; given the entire history of both the cluster assignments and the object's values, the current value of the object depends only on the value of the object at the previous time point as well as the current cluster assignment. These requirements amount to the dependency assumptions

$$Z_{it}|Z_{i1}, \ldots, Z_{i(t-1)} \overset{\mathcal{D}}{=} Z_{it}|Z_{i(t-1)} \triangleq \beta_{Z_{i(t-1)}Z_{it}}, \tag{2.1}$$

(hence the cluster assignments follow a Markov process) and

$$\mathbf{X}_{it}|\mathbf{X}_{i1}, \ldots, \mathbf{X}_{i(t-1)}, Z_{i1}, \ldots, Z_{iT_i} \overset{\mathcal{D}}{=} \mathbf{X}_{it}|\mathbf{X}_{i(t-1)}, Z_{it}. \tag{2.2}$$

A widely used clustering method is to fit a normal mixture model, and the commonly used k-means is a special case of this. This approach performs well in a wide range of settings and so we make the distributional assumptions that

$$\pi(\mathbf{X}_{i1}|Z_{i1}) = N(\mathbf{X}_{i1}|\boldsymbol{\mu}_{Z_{i1}}, \Sigma_{Z_{i1}}), \tag{2.3}$$

$$\pi(\mathbf{X}_{it}|\mathbf{X}_{i(t-1)}, Z_{it}) = N(\mathbf{X}_{it}|\lambda\boldsymbol{\mu}_{Z_{it}} + (1 - \lambda)\mathbf{X}_{i(t-1)}, \Sigma_{Z_{it}}), \tag{2.4}$$

where $\boldsymbol{\mu}_k$ and Σ_k are the mean vector and covariance matrix for the kth cluster for $k = 1, \ldots, K$, $\lambda \in (0, 1)$, and $N(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$ denotes the multivariate normal density with mean $\boldsymbol{\mu}$ and covariance matrix Σ. The $\boldsymbol{\mu}_k$'s and Σ_k's are the temporally

constant structures which define each cluster. The transition distribution in (2.4) blends the role of the current cluster with the individual effect. That is, in this framework one can look at the distribution of the current position of the ith individual as being influenced by where the individual has been, and by a sense of belonging to a particular cluster. This model can be thought of as extending the normal mixture model clustering to longitudinal data.

Assuming independence between the objects to be clustered, the complete-data likelihood (the distribution of the observed data and unobserved cluster assignments) can be written as

$$
\prod_{i=1}^{n} \pi(\mathbf{X}_{i1}, \ldots, \mathbf{X}_{iT_i}, Z_{i1}, \ldots Z_{iT_i})
$$

$$
= \prod_{i=1}^{n} \alpha_{Z_{i1}} \pi(\mathbf{X}_{i1}|Z_{i1}) \prod_{t=2}^{T_i} \beta_{Z_{i(t-1)}Z_{it}} \pi(\mathbf{X}_{it}|\mathbf{X}_{i(t-1)}, Z_{it}). \qquad (2.5)
$$

We can then write the marginal distribution of the data as

$$
\prod_{i=1}^{n} \pi(\mathbf{X}_{i1}, \ldots, \mathbf{X}_{iT_i}) = \prod_{i=1}^{n} \sum_{Z_{i1}=1}^{K} \Big(\alpha_{Z_{i1}} \pi(\mathbf{X}_{i1}|Z_{i1}) \sum_{Z_{i2}=1}^{K} \Big(\beta_{Z_{i1}Z_{i2}} \pi(\mathbf{X}_{i2}|\mathbf{X}_{i1}, Z_{i2})
$$

$$
\cdots \sum_{Z_{iT_i}=1}^{K} \Big(\beta_{Z_{i(T_i-1)}Z_{iT_i}} \pi(\mathbf{X}_{iT_i}|\mathbf{X}_{i(T_i-1)}, Z_{iT_i}) \Big) \cdots \Big). \quad (2.6)
$$

2.2. Regressed cluster probabilities

Suppose there is a set of explanatory variables that may influence how the objects are clustered. That is, we are not interested in clustering the objects based on the values of these explanatory variables, but rather we are interested in how these variables explain the clustering results. Let $\mathbf{w}_{it}$ denote a vector of length d corresponding to these explanatory variables for the ith object at time t. Then the model described in Section 2.1 can be slightly modified to allow $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to be functions of the explanatory variables, $\mathbf{w}_{it}$, and unknown parameters, denoted as $\delta_{\ell k}$ and $\boldsymbol{\gamma}_{\ell k}$ for $\ell = 0, \ldots, K$ and $k = 1, \ldots, K$. Mimicking the multinomial logistic regression model, for $k = 1, \ldots, K$ we let

$$
\log \left(\frac{\alpha_k(\mathbf{w}_{it})}{\alpha_K(\mathbf{w}_{it})} \right) = \delta_{0k} + \mathbf{w}'_{it} \boldsymbol{\gamma}_{0k}, \qquad (2.7)
$$

where we fix $\delta_{0K} = 0$ and $\boldsymbol{\gamma}_{0K} = \mathbf{0}$ for identifiability; similarly for each $h = 1, \ldots, K$, for $k = 1, \ldots, K$ we let

$$
\log \left(\frac{\beta_{hk}(\mathbf{w}_{it})}{\beta_{hK}(\mathbf{w}_{it})} \right) = \delta_{hk} + \mathbf{w}'_{it} \boldsymbol{\gamma}_{hk}, \qquad (2.8)
$$

where again we fix $\delta_{hK} = 0$ and $\boldsymbol{\gamma}_{hK} = \mathbf{0}$ for $h = 1, \ldots, K$. Then we can write the initial and transition cluster probabilities, respectively, as

$$\alpha_k(\mathbf{w}_{it}) = \frac{\exp(\delta_{0k} + \mathbf{w}'_{it}\boldsymbol{\gamma}_{0k})}{1 + \sum_{\ell=1}^{K-1} \exp(\delta_{0\ell} + \mathbf{w}'_{it}\boldsymbol{\gamma}_{0\ell})}, \tag{2.9}$$

$$\beta_{hk}(\mathbf{w}_{it}) = \frac{\exp(\delta_{hk} + \mathbf{w}'_{it}\boldsymbol{\gamma}_{hk})}{1 + \sum_{\ell=1}^{K-1} \exp(\delta_{h\ell} + \mathbf{w}'_{it}\boldsymbol{\gamma}_{h\ell})}, \tag{2.10}$$

for $h, k = 1, \ldots, K$. It is easy to see that setting $\boldsymbol{\gamma}_{0k} = \boldsymbol{\gamma}_{hk} = \mathbf{0}$ for all h, k yields a model equivalent to the non-regressed transition model of Section 2.1.

The model of Section 2.1 then changes in that we are conditioning on the explanatory variables $\mathbf{w}_{it}$. Specifically we have that the conditional distribution of the cluster assignments changes from (2.1) to

$$Z_{it}|Z_{i1}, \ldots, Z_{i(t-1)}, \mathbf{w}_{i1}, \ldots, \mathbf{w}_{iT_i} \overset{\mathcal{D}}{=} Z_{it}|Z_{i(t-1)}, \mathbf{w}_{it} \triangleq \beta_{Z_{i(t-1)}Z_{it}}(\mathbf{w}_{it}), \tag{2.11}$$

while (2.2) does not change. The complete-data likelihood is

$$\prod_{i=1}^{n} \pi(\mathbf{X}_{i1}, \ldots, \mathbf{X}_{iT_i}, Z_{i1}, \ldots Z_{iT_i}|\mathbf{w}_{i1}, \ldots, \mathbf{w}_{iT_i})$$

$$= \prod_{i=1}^{n} \alpha_{Z_{i1}}(\mathbf{w}_{i1})\pi(\mathbf{X}_{i1}|Z_{i1}) \prod_{t=2}^{T_i} \beta_{Z_{i(t-1)}Z_{it}}(\mathbf{w}_{it})\pi(\mathbf{X}_{it}|\mathbf{X}_{i(t-1)}, Z_{it}), \tag{2.12}$$

and the marginal likelihood is

$$\prod_{i=1}^{n} \pi(\mathbf{X}_{i1}, \ldots, \mathbf{X}_{iT_i}|\mathbf{w}_{i1}, \ldots, \mathbf{w}_{iT_i})$$

$$= \prod_{i=1}^{n} \sum_{Z_{i1}=1}^{K} \left(\alpha_{Z_{i1}}(\mathbf{w}_{i1})\pi(\mathbf{X}_{i1}|Z_{i1}) \sum_{Z_{i2}=1}^{K} \left(\beta_{Z_{i1}Z_{i2}}(\mathbf{w}_{i2})\pi(\mathbf{X}_{i2}|\mathbf{X}_{i1}, Z_{i2}) \right. \right.$$

$$\left. \left. \cdots \sum_{Z_{iT_i}=1}^{K} \left(\beta_{Z_{i(T_i-1)}Z_{iT_i}}(\mathbf{w}_{iT_i})\pi(\mathbf{X}_{iT_i}|\mathbf{X}_{i(T_i-1)}, Z_{iT_i}) \right) \cdots \right) \right). \tag{2.13}$$

2.3. Computational issues

The estimation procedure to be outlined in Section 3 is an iterative procedure to maximize the likelihood, the marginal distribution of the data (2.6), and hence we must be able to compute the likelihood to know if we have reached convergence. In computing the likelihood we run into a problem — the number of terms to be summed grows exponentially with T_i. To make this less computationally expensive, we can consider two recursive relationships. First consider the

joint density $\pi(\mathbf{X}_1, \ldots, \mathbf{X}_t, Z_t)$ (where the subscript i is suppressed for ease of notation where obvious). It can be shown that $\pi(Z_t|Z_{t-1}, \mathbf{X}_1, \ldots, \mathbf{X}_{t-1}) = \beta_{Z_{t-1}Z_t}$ (see Appendix A). Using this, we see that

$$\pi(\mathbf{X}_1, \ldots, \mathbf{X}_t, Z_t)$$
$$= \sum_{Z_{t-1}=1}^{K} \pi(\mathbf{X}_t|\mathbf{X}_{t-1}, Z_t)\pi(Z_{t-1}, Z_t|\mathbf{X}_1, \ldots, \mathbf{X}_{t-1})\pi(\mathbf{X}_1, \ldots, \mathbf{X}_{t-1})$$
$$= \pi(\mathbf{X}_1, \ldots, \mathbf{X}_{t-1}) \sum_{Z_{t-1}=1}^{K} \pi(Z_{t-1}|\mathbf{X}_1, \ldots, \mathbf{X}_{t-1})\beta_{Z_{t-1}Z_t}\pi(\mathbf{X}_t|\mathbf{X}_{t-1}, Z_t). \quad (2.14)$$

Computing the marginal likelihood $\pi(\mathbf{X}_1, \ldots, \mathbf{X}_T)$ can be accomplished by using this recursion in two steps, first obtaining $\pi(Z_{t-1}|\mathbf{X}_1, \ldots, \mathbf{X}_{t-1})$ and then $\pi(\mathbf{X}_1, \ldots, \mathbf{X}_t, Z_t)$. More specifically, we first compute and normalize

$$\pi(Z_1|\mathbf{X}_1) \propto \pi(\mathbf{X}_1|Z_1)\alpha_{Z_1}. \quad (2.15)$$

Then for $t = 3, \ldots, T$, from (2.14) we can compute and normalize

$$\pi(Z_{t-1}|\mathbf{X}_1, \ldots, \mathbf{X}_{t-1})$$
$$\propto \sum_{Z_{t-2}=1}^{K} \pi(Z_{t-2}|\mathbf{X}_1, \ldots, \mathbf{X}_{t-2})\beta_{Z_{t-2}Z_{t-1}}\pi(\mathbf{X}_{t-1}|\mathbf{X}_{t-2}, Z_{t-1}). \quad (2.16)$$

Second, we can compute $\pi(\mathbf{X}_1) = \sum_{Z_1=1}^{K} \pi(\mathbf{X}_1|Z_1)\alpha_{Z_1}$, and then use the normalized results from (2.15) and (2.16) to compute, for $t = 2, \ldots, T$,

$$\pi(\mathbf{X}_1, \ldots, \mathbf{X}_t)$$
$$= \pi(\mathbf{X}_1, \ldots, \mathbf{X}_{t-1}) \sum_{Z_t=1}^{K} \sum_{Z_{t-1}=1}^{K} \pi(Z_{t-1}|\mathbf{X}_1, \ldots, \mathbf{X}_{t-1})\beta_{Z_{t-1}Z_t}\pi(\mathbf{X}_t|\mathbf{X}_{t-1}, Z_t). (2.17)$$

By utilizing these recursive relationships, the number of terms required to be summed in the computation of the likelihood density grows linearly, rather than exponentially, in time. Regarding the context of regressed cluster assignment probabilities, the previous work can still be applied simply by exchanging $\alpha_{Z_{i1}}$ for $\alpha_{Z_{i1}(\mathbf{w}_{i1})}$ and $\beta_{Z_{i(t-1)}Z_{it}}$ for $\beta_{Z_{i(t-1)}Z_{it}}(\mathbf{w}_{it})$ in equations (2.14) to (2.17).

3. Estimation

We aim to find the maximum likelihood estimators (MLE's) for the model parameters, henceforth denoted as Θ. In the case of non-regressed cluster probabilities,

$$\Theta = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \Sigma_1, \ldots, \Sigma_K, \lambda, \alpha_1, \ldots, \alpha_K, \beta_{11}, \ldots, \beta_{KK}\},$$

and for the case of regressed cluster probabilities,

$$\Theta = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \Sigma_1, \ldots, \Sigma_K, \lambda, \delta_{01}, \ldots, \delta_{K(K-1)}, \boldsymbol{\gamma}_{01}, \ldots, \boldsymbol{\gamma}_{K(K-1)}\}.$$

To find the MLE's, we employ the generalized EM algorithm (Dempster, Laird, and Rubin (1977); Wu (1983)) to obtain parameter estimates, as well as cluster assignment probabilities. Heretofore the notation for probability densities have not included the parameter set Θ, this parameter set being implicit in the density itself; henceforth the dependency on the parameter Θ needs to be explicitly written as $\pi(\cdot|\Theta)$. We first derive the solutions for the non-regressed case. Letting $\widehat{\Theta}$ denote the current estimate of Θ, we iteratively find

$$Q(\Theta, \widehat{\Theta}) = \mathbb{E}\left(\log(L^c)|\mathcal{X}_1, \ldots, \mathcal{X}_n, \widehat{\Theta}\right), \tag{3.1}$$

(E step) where L^c is the complete-data likelihood, as given in (2.5), and find Θ^*, where $Q(\Theta^*, \widehat{\Theta}) \geq Q(\widehat{\Theta}, \widehat{\Theta})$ (M step), subsequently setting $\widehat{\Theta} = \Theta^*$. Now $Q(\Theta, \widehat{\Theta})$ can, with some algebra (see Appendix B), be written in the tractable form

$$\begin{aligned}
Q(\Theta, \widehat{\Theta}) = &\sum_{i=1}^{n} \sum_{\ell_1=1}^{K} \cdots \sum_{\ell_{T_i}=1}^{K} \left[\log(\alpha_{\ell_1}) + \sum_{t=2}^{T_i} \log(\beta_{\ell_{t-1}\ell_t}) \right] \\
&\cdot \mathbb{P}(Z_{i1} = \ell_1|\mathcal{X}_i, \widehat{\Theta}) \prod_{s=2}^{T_i} \mathbb{P}(Z_{is} = \ell_s|Z_{i(s-1)} = \ell_{s-1}, \mathcal{X}_i, \widehat{\Theta}) \\
&+ \sum_{i=1}^{n} \sum_{\ell_1=1}^{K} \cdots \sum_{\ell_{T_i}=1}^{K} \left[\log(\pi(\mathbf{X}_{i1}|Z_{i1} = \ell_1, \Theta)) \right. \\
&\left. + \sum_{t=2}^{T_i} \log(\pi(\mathbf{X}_{it}|\mathbf{X}_{i(t-1)}, Z_{it} = \ell_t, \Theta)) \right] \\
&\cdot \mathbb{P}(Z_{i1} = \ell_1|\mathcal{X}_i, \widehat{\Theta}) \prod_{s=2}^{T_i} \mathbb{P}(Z_{is} = \ell_s|Z_{i(s-1)} = \ell_{s-1}, \mathcal{X}_i, \widehat{\Theta}). \tag{3.2}
\end{aligned}$$

In computing a valid Θ^* such that $Q(\Theta^*, \widehat{\Theta}) \geq Q(\widehat{\Theta}, \widehat{\Theta})$, we wish the number of terms to be summed to grow linearly with respect to time, and not exponentially. To this end, we utilize recursive relationships to compute the conditional distributions $\mathbb{P}(Z_{it}|Z_{i(t-1)}, \mathcal{X}_i, \widehat{\Theta})$ and the marginal distributions $\mathbb{P}(Z_{it}|\mathcal{X}_i, \widehat{\Theta})$. Again leaving off the subscript i where obvious, it can be shown (see Appendix C) that for $2 \leq t \leq T$,

$$\mathbb{P}(Z_t|Z_{t-1}, \mathcal{X}, \Theta) \propto q(Z_t|Z_{t-1}), \tag{3.3}$$

where

$$q(Z_T|Z_{T-1}) = \beta_{Z_{T-1}Z_T}\pi(\mathbf{X}_T|\mathbf{X}_{T-1}, Z_T, \Theta), \tag{3.4}$$

and, for $2 \leq t < T$,

$$q(Z_t|Z_{t-1}) = \beta_{Z_{t-1}Z_t}\pi(\mathbf{X}_t|\mathbf{X}_{t-1}, Z_t, \Theta) \sum_{\ell_{t+1}=1}^{K} q(Z_{t+1} = \ell_{t+1}|Z_t). \tag{3.5}$$

We then need only normalize to obtain the conditional distributions. To obtain the marginals, we start by computing and normalizing

$$\mathbb{P}(Z_1|\mathcal{X}, \Theta) \propto \alpha_{Z_1}\pi(\mathbf{X}_1|Z_1, \Theta) \sum_{\ell_2=1}^{K} q(Z_2 = \ell_2|Z_1), \tag{3.6}$$

and recursively obtain

$$\mathbb{P}(Z_t|\mathcal{X}, \Theta) = \sum_{\ell_{t-1}=1}^{K} \mathbb{P}(Z_{t-1} = \ell_{t-1}|\mathcal{X}, \Theta)\mathbb{P}(Z_t|Z_{t-1} = \ell_{t-1}, \mathcal{X}, \Theta). \tag{3.7}$$

This can be derived in a similar fashion as that found in Appendix C.

Using the method of Lagrange multipliers, we find that the value of α_k that maximizes $Q(\Theta, \widehat{\Theta})$ is

$$\alpha_k^* = \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}(Z_{i1} = k|\mathcal{X}_i, \widehat{\Theta}), \tag{3.8}$$

and similarly the optimizing value of β_{hk} is

$$\beta_{hk}^* = \frac{\sum_{i=1}^{n} \sum_{t=2}^{T_i} \mathbb{P}(Z_{i(t-1)} = h|\mathcal{X}_i, \widehat{\Theta})\mathbb{P}(Z_{it} = k|Z_{i(t-1)} = h, \mathcal{X}_i, \widehat{\Theta})}{\sum_{i=1}^{n} \sum_{t=2}^{T_i} \mathbb{P}(Z_{i(t-1)} = h|\mathcal{X}_i, \widehat{\Theta})}. \tag{3.9}$$

To update λ, $\boldsymbol{\mu}_k$, and Σ_k, $k = 1, \ldots, K$, we employ a coordinate ascent method. We initialize $\lambda^* = \hat{\lambda}$, $\boldsymbol{\mu}_k^* = \hat{\boldsymbol{\mu}}_k$, and $\Sigma_k^* = \widehat{\Sigma}_k$, and then find

$$\lambda^* = \underset{\lambda}{\operatorname{argmax}} \ Q(\{\lambda, \Theta^* \setminus \{\lambda^*\}\}, \widehat{\Theta})$$

$$= \frac{\sum_{i=1}^{n} \sum_{t=2}^{T_i} \sum_{\ell_t=1}^{K} \mathbb{P}(Z_{it} = \ell_t|\mathcal{X}_i, \widehat{\Theta})(\mathbf{X}_{it} - \mathbf{X}_{i(t-1)})'\Sigma_k^{-1}(\boldsymbol{\mu}_k - \mathbf{X}_{i(t-1)})}{\sum_{i=1}^{n} \sum_{t=2}^{T_i} \sum_{\ell_t=1}^{K} \mathbb{P}(Z_{it} = \ell_t|\mathcal{X}_i, \widehat{\Theta})(\boldsymbol{\mu}_k - \mathbf{X}_{i(t-1)})'\Sigma_k^{-1}(\boldsymbol{\mu}_k - \mathbf{X}_{i(t-1)})}, \tag{3.10}$$

$$\boldsymbol{\mu}_k^* = \underset{\boldsymbol{\mu}_k}{\operatorname{argmax}} \ Q(\{\boldsymbol{\mu}_k, \Theta^* \setminus \{\boldsymbol{\mu}_k^*\}\}, \widehat{\Theta})$$

$$= \frac{\sum_{i=1}^{n}\{\mathbb{P}(Z_{i1} = k|\mathcal{X}_i, \widehat{\Theta})\mathbf{X}_{i1} + \lambda \sum_{t=2}^{T_i} \mathbb{P}(Z_{it} = k|\mathcal{X}_i, \widehat{\Theta})(\mathbf{X}_{it} - (1-\lambda)\mathbf{X}_{i(t-1)})\}}{\sum_{i=1}^{n}\{\mathbb{P}(Z_{i1} = k|\mathcal{X}_i, \widehat{\Theta}) + \lambda^2 \sum_{t=2}^{T_i} \mathbb{P}(Z_{it} = k|\mathcal{X}_i, \widehat{\Theta})\}},$$

$$\Sigma_k^* = \underset{\Sigma_k}{\operatorname{argmax}} \, Q(\{\Sigma_k, \Theta^* \setminus \{\Sigma_k^*\}\}, \widehat{\Theta}) \tag{3.11}$$

$$= \frac{\sum_{i=1}^n \{\mathbb{P}(Z_{i1}=k|\mathcal{X}_i, \widehat{\Theta})(\mathbf{X}_{i1}-\boldsymbol{\mu}_k)(\mathbf{X}_{i1}-\boldsymbol{\mu}_k)' + \sum_{t=2}^{T_i} \mathbb{P}(Z_{it}=k|\mathcal{X}_i, \widehat{\Theta})\mathbf{H}_{it}\mathbf{H}_{it}'\}}{\sum_{i=1}^n \left\{ \mathbb{P}(Z_{i1}=k|\mathcal{X}_i, \widehat{\Theta}) + \sum_{t=2}^{T_i} \mathbb{P}(Z_{it}=k|\mathcal{X}_i, \widehat{\Theta}) \right\}} \tag{3.12}$$

for $k = 1, \ldots, K$, where $\mathbf{H}_{it} = \mathbf{X}_{it} - \lambda \boldsymbol{\mu}_k - (1-\lambda)\mathbf{X}_{i(t-1)}$. See Appendix D for proofs of these solutions.

In the case of regressed cluster probabilities, the complete-data likelihood in $Q(\Theta, \widehat{\Theta})$ is that found in (2.12). To obtain the correct updates for λ, $\boldsymbol{\mu}_k$, and Σ_k, $k = 1, \ldots, K$, simply replace $\mathbb{P}(Z_{it}|\mathcal{X}_i, \Theta)$ above with $\mathbb{P}(Z_{it}|\mathcal{X}_i, \Theta, \mathbf{w}_{i1}, \ldots, \mathbf{w}_{iT_i})$; these latter distributions are easily computed by first replacing α_k with $\alpha_k(\mathbf{w}_{i1})$ in (3.6) and β_{hk} with $\beta_{hk}(\mathbf{w}_{it})$ in (3.4) and (3.5), and then proceeding with the recursive relationships previously outlined. To obtain updates for δ_{0k}, $\boldsymbol{\gamma}_{0k}$, δ_{hk}, and $\boldsymbol{\gamma}_{hk}$, $h = 1, \ldots, K$, $k = 1, \ldots, K-1$, one can use a numerical optimization method (we implemented the quasi-Newton BFGS method) to find

$$\underset{\{\delta_{0k}, \boldsymbol{\gamma}_{0k} : k=1,\ldots,K\}}{\operatorname{argmax}} = \sum_{i=1}^n \sum_{k=1}^K \mathbb{P}(Z_{i1} = k|\mathcal{X}_i, \widehat{\Theta}) \log(\alpha_k(\mathbf{w}_{i1})), \tag{3.13}$$

$$\underset{\{\delta_{hk}, \boldsymbol{\gamma}_{hk} : k=1,\ldots,K\}}{\operatorname{argmax}}$$

$$= \sum_{i=1}^n \sum_{t=2}^{T_i} \sum_{k=1}^K \mathbb{P}(Z_{i(t-1)} = h|\mathcal{X}_i, \widehat{\Theta})\mathbb{P}(Z_{it} = k|Z_{i(t-1)} = h, \mathcal{X}_i, \widehat{\Theta}) \log(\beta_{hk}(\mathbf{w}_{it})), \tag{3.14}$$

for $h = 1, \ldots, K$. Since (3.13) and (3.14) are both concave, finding the global maximum at each M-step is assured upon convergence of the BFGS algorithm.

4. Simulation Study

We simulated two illustrative data sets, one without and one with regressed cluster probabilities, where p, the dimension of each $\mathbf{X}_{it}$, is two. We then simulated 100 data sets where $p = 5$, 50 without and 50 with regressed cluster probabilities (thus a total of 102 simulated data sets). The resulting clustering was compared in the variation of information (VI) (Meilă (2003)) and in the corrected Rand index (CRI) (Hubert and Arabie (1985)). The VI, a true metric on clusterings, yields a value of 0 for identical clusterings; higher values indicate more disparate clustering assignments. The corrected Rand index, adjusted to account for chance, yields a value of 1 for perfect cluster agreement and has an expected value of 0 for random cluster assignments. The results are both in the text as well as in Table 1.

Each simulated data set (when $p = 2$ and when $p = 5$) without regressed cluster probabilities was achieved according to the following. The number of objects n was 100, the maximum number of time steps $\max_i\{T_i\}$ was 10, and the number of clusters K was 5. To check whether our method can accurately group the objects, the simulations consisted of a wide range of possible parameter values. We obtained the simulated data in the following way. The quantities T_i were uniformly sampled from 1 to 10; λ was drawn from a beta distribution with both parameters equal to 10 (the mean of the distribution is 0.5); $\boldsymbol{\alpha}$ was drawn from a Dirichlet distribution with parameters all equalling 10; $\boldsymbol{\mu}_k$'s were drawn from a multivariate normal distribution with mean $\mathbf{0}$ and diagonal covariance matrix with diagonal entries equal to 20; Σ_k's were drawn from an inverse Wishart distribution with $p + 4$ degrees of freedom and diagonal scale matrix whose diagonal entries were 3. The hth row of β was set to be proportional to

$$\left(\frac{1}{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_h\|}, \dots, \frac{1}{\|\boldsymbol{\mu}_{h-1} - \boldsymbol{\mu}_h\|}, 15 \times \max_{k \neq h} \left\{ \frac{1}{\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_h\|} \right\}, \frac{1}{\|\boldsymbol{\mu}_{h+1} - \boldsymbol{\mu}_h\|}, \right.$$
$$\left. \dots, \frac{1}{\|\boldsymbol{\mu}_K - \boldsymbol{\mu}_h\|} \right).$$

This formulation of β was chosen to put most of the probability on remaining in the same cluster and to force the probability of jumping to a different cluster to decrease as the distance to that different cluster increased. The Z_{it}'s were then generated according to $\boldsymbol{\alpha}$ and β, and the $\mathcal{X}_i$'s were generated according to (2.3) and (2.4).

The simulations with regressed cluster probabilities were generated in a similar manner, but we allowed $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to depend on a positive valued covariate ($w_{it} > 0$) which affects the probability of belonging to cluster one. A high value of the covariate implied a tendency to belong to cluster one while a low value of the covariate implied a tendency to avoid belonging to cluster one. Beyond this there was no discriminating influence from the covariate on the probabilities of belonging or transitioning to the other clusters. Just as in the simulation without a covariate, there was also an inherent tendency to remain in the same cluster rather than switch (though this could, for objects not in cluster one, potentially be countered by extremely large covariate values). This was accomplished in the following way. For $i = 1, \dots, n$, the explanatory variable w_{i1} ($d = 1$) was drawn from a chi-squared distribution with five degrees of freedom, and for $t = 2, \dots, T_i$, w_{it} was drawn from a normal distribution with mean $w_{i(t-1)}$ and standard deviation 0.5. In this way the objects' covariate values had a certain degree of smoothness over time. The parameters γ_{0k} and, for $h = 1, \dots, K$, γ_{hk} were set to be $\log(1.5)$ if $k = 1$, and 0 otherwise. We set δ_{0k} and, for $h = 1, \dots, K$, δ_{hk} equal to $-5\log(1.5)$ if $k = 1$, and 0 otherwise; in addition, $\log(15)$ was added to δ_{hh},

Table 1. From the simulations run without and with regressed cluster probabilities, the estimated clustering results are compared to the true cluster assignments for both k-means and our proposed approach using both variation of information (VI) and the corrected Rand index (CRI). Means over 50 simulations are given with medians in parentheses.

	Without regressed cluster probabilities		With regressed cluster probabilities	
	Proposed method	k-means	Proposed method	k-means
VI	0.0180 (0)	0.5516 (0.5175)	0.0531 (0)	0.7413 (0.7429)
CRI	0.9953 (1)	0.8426 (0.8609)	0.9780 (1)	0.7729 (0.7953)

$h < K$, and $-\log(15)$ was added to δ_{Kk}, $k < K - 1$. This leads to the situation described above where there is an increased chance of staying in the same cluster than switching to another cluster; further, if w_{it} is 5 (the mean of w_{it}) there is no particular tendency or avoidance of belonging to cluster one, if $w_{it} > 5$ there is an increased probability of belonging to cluster one, and if $w_{it} < 5$ there is a decreased probability of belonging to cluster one. The solid line curves in Figure 4 graphically demonstrate the initial cluster probabilities and transitional cluster probabilities for varying values of the covariate w_{it}.

For the illustrative simulation without regressed cluster probabilities ($p = 2$), the VI from using the proposed approach was 0.2679 and CRI was 0.9232. Simply using k-means and choosing the best k-means result from 15 starting positions (which was used to initialize the generalized EM algorithm), led to a VI of 1.031 and CRI of 0.6615. Figure 1(a) gives the plot of the log-likelihood over the iterations of the generalized EM algorithm of the proposed method, starting with the initialized values. We see that the algorithm converged within a small number of iterations. Figure 2 shows the simulated data, where the numbers correspond to the true cluster assignments, and the shapes correspond to the estimated hard cluster assignments. Regarding the 50 simulations without regressed cluster probabilities ($p = 5$), the mean (median) VI was 0.0180 (0) and CRI was 0.9953 (1). Using k-means, the mean (median) VI was 0.5516 (0.5175) and CRI was 0.8426 (0.8609). Using a UNIX machine with a 2.40 GHz processor, the mean (median) elapsed time for our proposed method was 0.5507 sec (0.5120 sec). These values of VI and CRI imply that our model performs extremely well, and much better than the naïve k-means approach.

For the illustrative simulation with regressed cluster probabilities ($p = 2$), the VI from using the proposed approach was 0.5080 and the CRI was 0.8232. Simply using k-means resulted in a VI of 1.342 and a CRI of 0.5147. Figure 1(b) gives the plot of the log-likelihood over the iterations of the generalized EM algorithm in the proposed method, starting with the initialized values. From this we again see that the algorithm reached convergence within a small number of iterations. Figure 3 shows the simulated data, where the numbers correspond

(a) Without regressed cluster probabilities
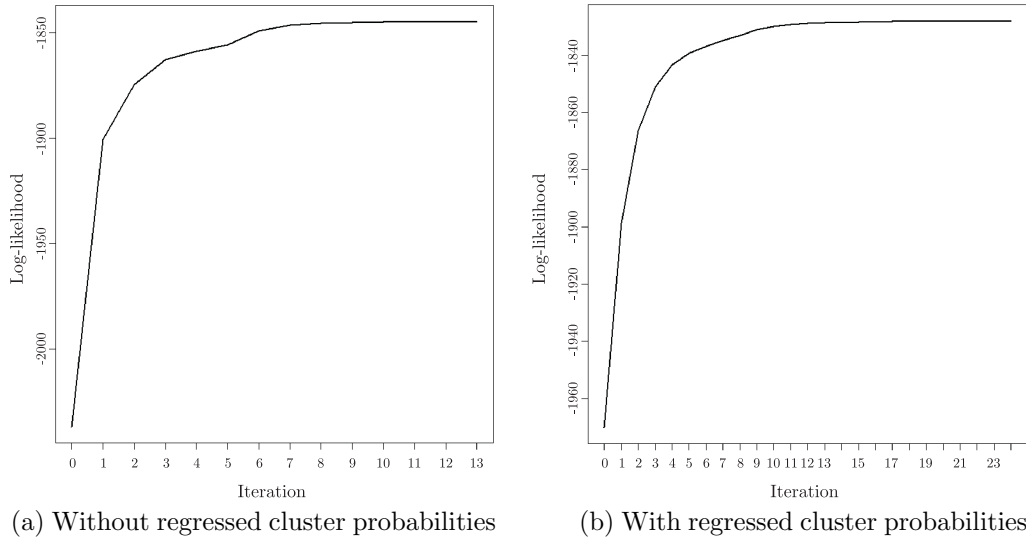
(b) With regressed cluster probabilities

Figure 1. Log-likelihood over the iterations for the two illustrative simulations.

to the true cluster assignments, and the shapes correspond to the estimated hard cluster assignments. Regarding the 50 simulations with regressed cluster probabilities ($p = 5$), the mean (median) VI was 0.0531 (0) and CRI was 0.9780 (1). Using k-means, the mean (median) VI was 0.7413 (0.7429) and CRI was 0.7729 (0.7953). The mean (median) elapsed time for our proposed method was 41.82 sec (35.41 sec). As in the case for non-regressed cluster probabilities, these values of VI and CRI imply that our model performs extremely well, and much better than the naïve k-means approach.

We can also investigate whether the model captured the effect of the explanatory variable. An effective visualization of this is to plot the cluster probabilities $\alpha_k(w_{i1})$ and $\beta_{hk}(w_{it})$ for $h, k = 1, \ldots, K$, over the range of w_{it}. This is more intuitive than trying to look at the individual δ_{hk}'s and γ_{hk}'s, allowing the user to see exactly how the explanatory variables affect the cluster probabilities. Figure 4 gives these plots, where each line corresponds to probabilities of belonging to a particular cluster. Here we can see that the estimated probability curves (solid lines) correspond very closely to the true probability curves (dashed lines), implying that the estimation method captured the effect of the explanatory variable effectively.

5. U.S. Congressional Data

We collected 15 variables measuring legislative activity for Democrats in the House of Representatives in the 101st-110th Congresses (1989-2008). Because
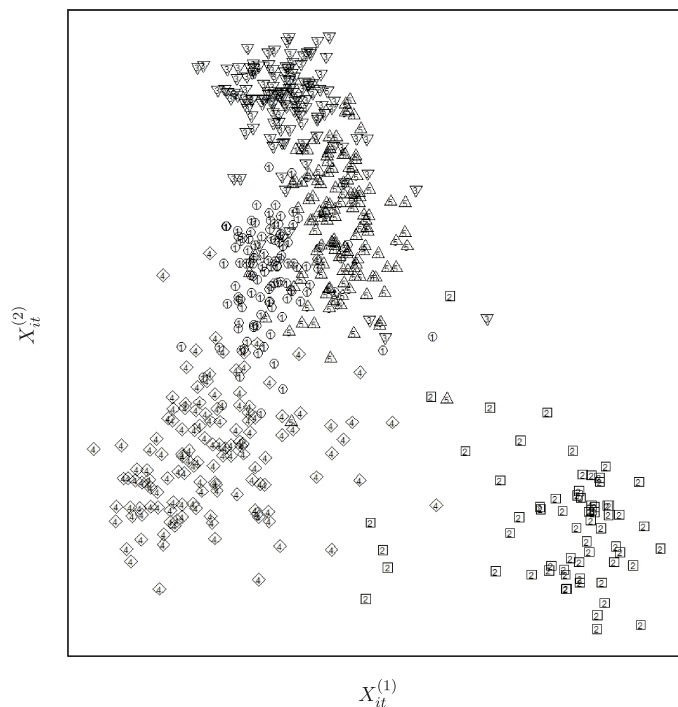
Figure 2. Plot of simulated data without regressed cluster probabilities, where the horizontal and vertical axes correspond to the two variables on which clustering is performed. Numbers correspond to true clustering, shapes correspond to estimated hard clustering assignments.

there are multiple variables for a single concept, we combined these indicators into eight indices — paying attention to the home district; showboating (e.g., making speeches and writing editorials); voting with the party on roll calls; giving campaign funds to the party; specializing in particular policy issues; building bipartisan coalitions; overall fundraising; and lawmaking (e.g., introducing legislation). See Appendix E for more details on these variables. Thus $\mathbf{X}_{it}$ is the eight-dimensional vector corresponding to the scores on these eight indices measured for the ith member of Congress (MC) at his or her t^{th} term. There were $n = 539$ unique MCs that served one or more terms over the 20 years included in the study.

We wished to determine the impact of the MCs' ideology on the MCs' behaviors. To this end we included a covariate for MC ideology to help predict initial cluster assignments for individual MCs, as well as transitions between clusters across time. Our measure of ideology was the common-space NOMINATE score. These scores are based on a multidimensional scaling algorithm for roll call voting developed by Poole and Rosenthal (1997) and are a function of how often each
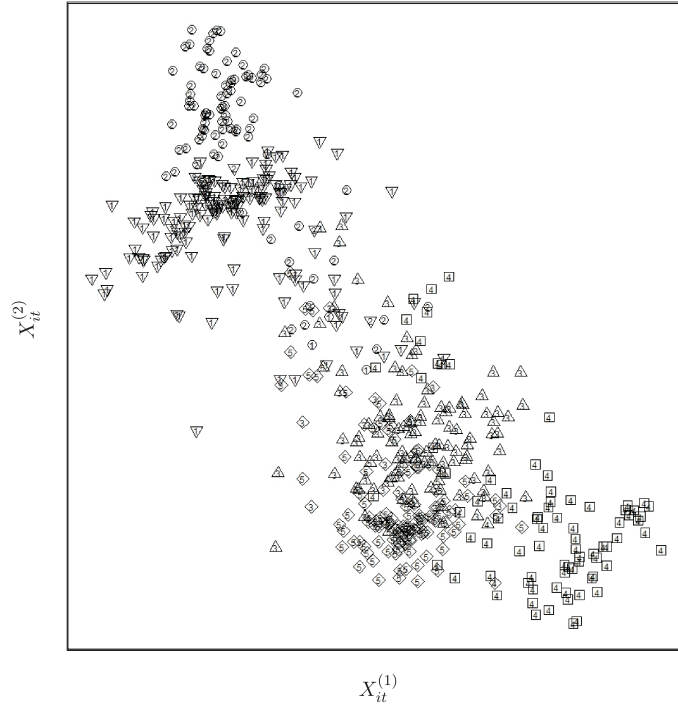
Figure 3. Plot of simulated data with regressed cluster probabilities, where the horizontal and vertical axes correspond to the two variables on which clustering is performed. Numbers correspond to true clustering, shapes correspond to estimated hard clustering assignments.

individual MC makes the same vote choice as every other MC. The algorithm allows all legislators to be arrayed on a scale of about -1 to 1, with negative scores indicating liberal ideology and positive scores indicating conservative ideology. Since the sample is solely Democrats, the scores are skewed toward the negative end of the scale, ranging from -0.725 to 0.190. In addition to this ideology variable, we also included time dummy variables in order to account for behavioral shifts across the entire party that occur during the various Congresses (i.e., each Congress has its own dummy variable). Lastly, for the 101st Congress we differentiated between true freshmen politicians as well as those MCs who have served previous terms not included in the data by allowing the true freshmen (for all Congresses) and the first observed MCs (of the 101st Congress) to have different initial clustering coefficients δ_{0k} and $\boldsymbol{\gamma}_{0k}$. This was accomplished by letting $\mathbf{w}_{i1} = (ideology, ideology \cdot 1_{\{\text{First Observed}\}}, 1_{\{t_1 < 101\}}, 1_{\{t_1 = 101\}}, \ldots, 1_{\{t_1 = 109\}})$, where $ideology$ is the NOMINATE score, $1_{\{\text{First Observed}\}}$ is 1 if not a true freshman and 0 otherwise, $1_{\{t_1 < 101\}}$ if the MC's first term was before the 101st

(a) Initial Clustering

(b) From Cluster 1

(c) From Cluster 2

(d) From Cluster 3

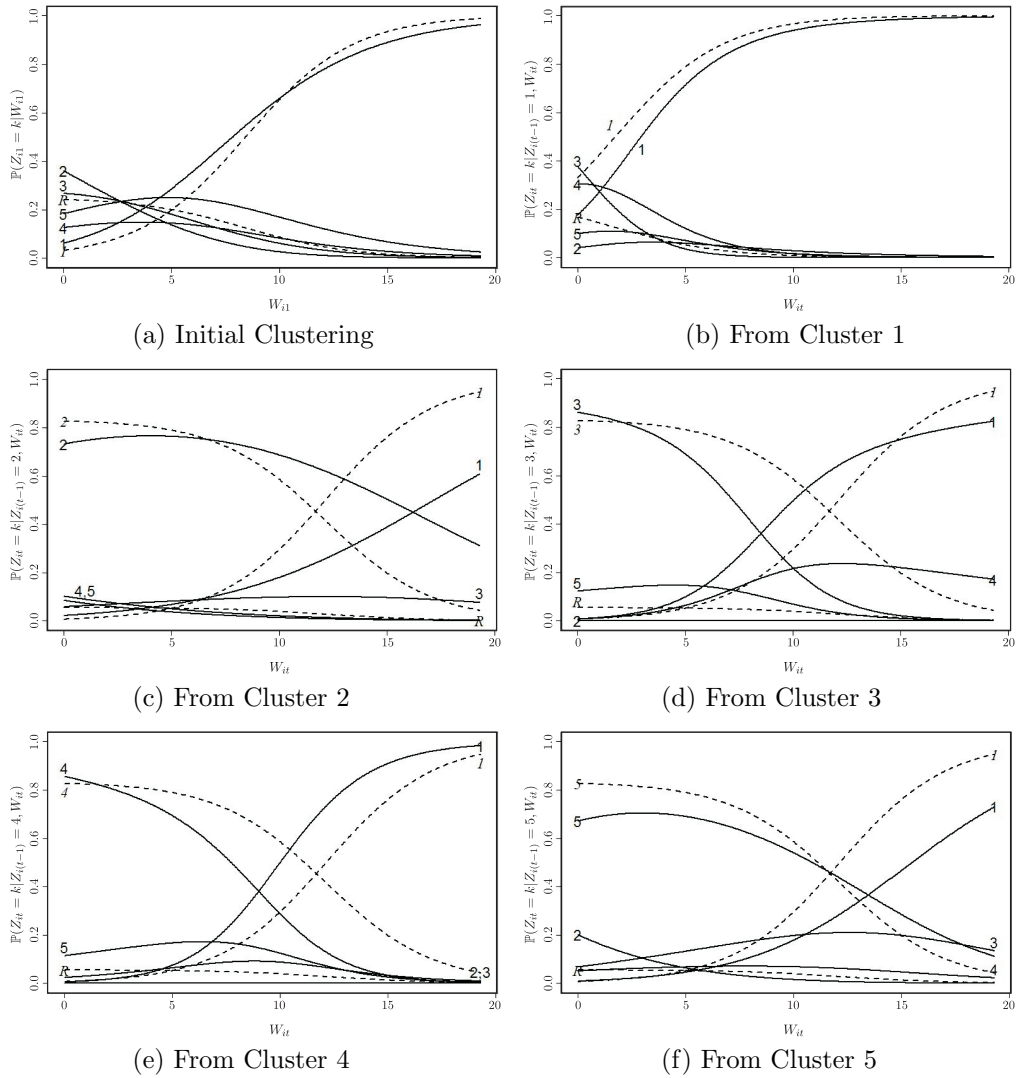(e) From Cluster 4

(f) From Cluster 5

Figure 4. Cluster probabilities for illustrative simulated data with regressed cluster probabilities. Solid lines are estimated probability curves, dashed lines are true curves. Each curve represents, for the varying values of the explanatory variable, the probability of belonging to the cluster whose number is adjacent to the curve. There are only two or three (depending on the figure) unique true probability curves which are labeled in italics, hence "R" indicates the probability curve corresponding to those remaining cluster numbers which are not unique, e.g., clusters 2 to 5 in the initial clustering plot.

Congress and $1_{\{t_1=s\}}$ is 1 if the ith MC's first term (of the study) was served

during the sth Congress and 0 otherwise, for $s = 101, 102, \ldots, 109$.

To determine the number of clusters we used the average silhouette statistic (Rousseeuw (1987)). We chose this rather than a statistic that is dependent on the number of model parameters, such as AIC or BIC, in order to avoid the number of covariates determining the number of clusters. That is, the number of clusters ought not to depend on the number of covariates used to explain the clustering. The results indicated four clusters of legislator types.

We applied our proposed method and obtained the following results. Figure 5 gives the cluster means $\boldsymbol{\mu}_k$. Cluster one, the largest cluster (50% of Z_{it}'s equal 1), consists of MCs that we call party soldiers. This group is marked by its very high score on party voting, but relatively low scores on lawmaking, showboating, and other activities. These are the rank-and-file legislators who toe the party line, but do not distinguish themselves in other ways. Cluster two, the second largest cluster comprising 26% of the observations, represents district advocates — these are legislators who devote their efforts to their home districts, but are not particularly active in the lawmaking process and are not committed partisans. Cluster three, with 16% of the observations, are the elites — MCs who are publicly visible, strongly support the party both in voting and giving of contributions, and who are very active in the lawmaking process. Across the sample period, the actual party leadership (e.g., Richard Gephardt and Nancy Pelosi) fall into this cluster. The final cluster, the "conscientious objectors," is the smallest, with just 8% of observations. These are MCs who are publicly visible (scoring high on the showboat index), are policy specialists, and are very bipartisan in their coalitions, and also have the lowest mean of the four clusters on party support in voting. These MCs chart their own courses, pursuing their policy goals with less regard for the party leadership's preferences.

The results are in line with intuitions about the relationship between ideology and legislative styles. For example, as shown in Figure 6, the probability of falling in the party soldier cluster decreases as the ideology score increases: strong liberals are more likely than their moderate counterparts to be party soldiers. On the other hand, the probability of being a conscientious objector increases as the ideology score increases: moderates are more likely than liberals to fall into this category. The probabilities associated with belonging to the district advocates or elites clusters are less linear. The likelihood of being in the elite category peaks just to the left of the mean ideology score — leaders tend to have more pragmatic outlooks, and, while solidly liberal, are seldom ideologically extreme. The probability of being a district advocate, on the other hand, is greatest to the right of the mean ideology score. MCs who focus on their districts often come from heterogeneous constituencies, and, as such, must devote special time and attention to cultivating the district in order to win reelection.
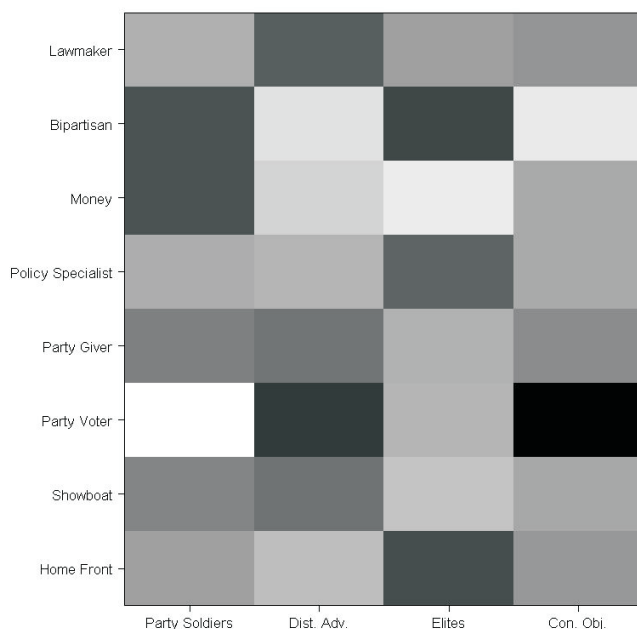
Figure 5. Cluster means for the Democrat MCs. The vertical axis corresponds to the behavioral variables on which the MCs were clustered, and the horizontal axis corresponds to the clusters. Lighter hues indicate higher values, darker hues indicate lower values.

Figure 7 demonstrates the relationship between the MCs' ideology and their probability of falling in a particular cluster at time $t+1$, given their cluster assignment at time t. These results largely correspond to intuition as well. For example, party soldiers at time t are more likely to remain as such at $t+1$ if they are very liberal. In contrast, district advocates, conscientious objectors, and elites are the most likely to maintain their cluster assignments across time if they are moderate, although the probability of remaining an elite drops off quickly if the MC is too moderate.

6. Discussion

We developed a clustering model for longitudinal data that accounts for time-varying cluster assignments, time series objects of varying length, and time-dependence via blending cluster effects with individual effects. We have shown how variables that explain the clustering can be integrated into the model, providing insight into the relationships existing between the clustering and the explanatory variables. We derived efficient methods for computing the likelihood,
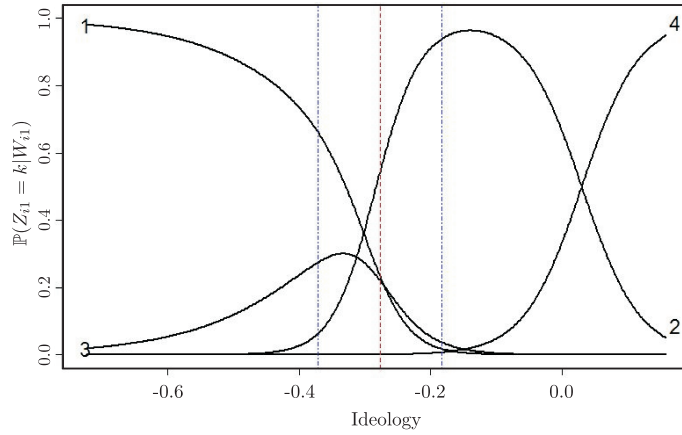
Figure 6. Initial cluster probabilities for true freshmen in the Congressional data. Each curve represents, for the varying values along the horizontal axis of the explanatory variable (ideology), the probability of belonging to the cluster whose number is adjacent to the curve. The horizontal axis spans the range of all MCs ideology; the dot-dashed lines give the 25% and 75% quantiles of ideology, and the long dashed line gives the mean ideology.

as well as efficient estimation methods using the generalized EM algorithm.

While the model given in Section 2.2 is completely identifiable with the constraints that $\delta_{\ell K} = 0$ and $\boldsymbol{\gamma}_{\ell K} = \mathbf{0}$ for $\ell = 0, \ldots, K$, the MLE's of the $\delta_{\ell k}$'s and $\boldsymbol{\gamma}_{\ell k}$'s may diverge and/or may not be unique if there are insufficient data points, especially if some cluster transitions are rare. A simple example of this has no observed transitions out of, say, cluster h. In such a case, the (possibly non-unique) MLE's $\hat{\delta}_{hh} \to \infty$, $\hat{\delta}_{hk} \to -\infty$ for $k \neq h$, and $\hat{\boldsymbol{\gamma}}_{hk}$, $k = 1, \ldots, K$, may take any finite values. As these computational issues are due to the true probability distributions of the cluster assignments not being adequately represented by the data, we would expect that these issues would disappear as $n \to \infty$. Future study on the existence and uniqueness of the MLE's would certainly be useful.

As mentioned in Section 2.1, the dependency structure of the observed values $\mathbf{X}_{it}$ follows a Markov switching model, where at time t all the previous information in $\mathcal{X}_i$ up to time $t-1$ is captured in $\mathbf{X}_{i(t-1)}$. There is, of course, a large number of possible dependency structures that could be represented in the data. For example, if there was some seasonality of length s in the individual effect, we might change (2.2) to

$$\mathbf{X}_{it}|\mathbf{X}_{i1}\ldots,\mathbf{X}_{i(t-1)}, Z_{i1},\ldots, Z_{iT_i} \stackrel{\mathcal{D}}{=} \mathbf{X}_{it}|\mathbf{X}_{i(t-1)}, \mathbf{X}_{i(t-s)}, Z_{it}, \qquad (6.1)$$

and change (2.4) to

$$\pi(\mathbf{X}_{it}|\mathbf{X}_{i(t-1)}, \mathbf{X}_{i(t-s)}, Z_{it}) = N(\mathbf{X}_{it}|(1-\lambda_1-\lambda_s)\boldsymbol{\mu}_{Z_{it}}+\lambda_1\mathbf{X}_{i(t-1)}+\lambda_s\mathbf{X}_{i(t-s)}, \Sigma_{Z_{it}}). \qquad (6.2)$$

(a) Transition from Party Soldier (cluster 1)   (b) Transition from District Advocate (cluster 2)

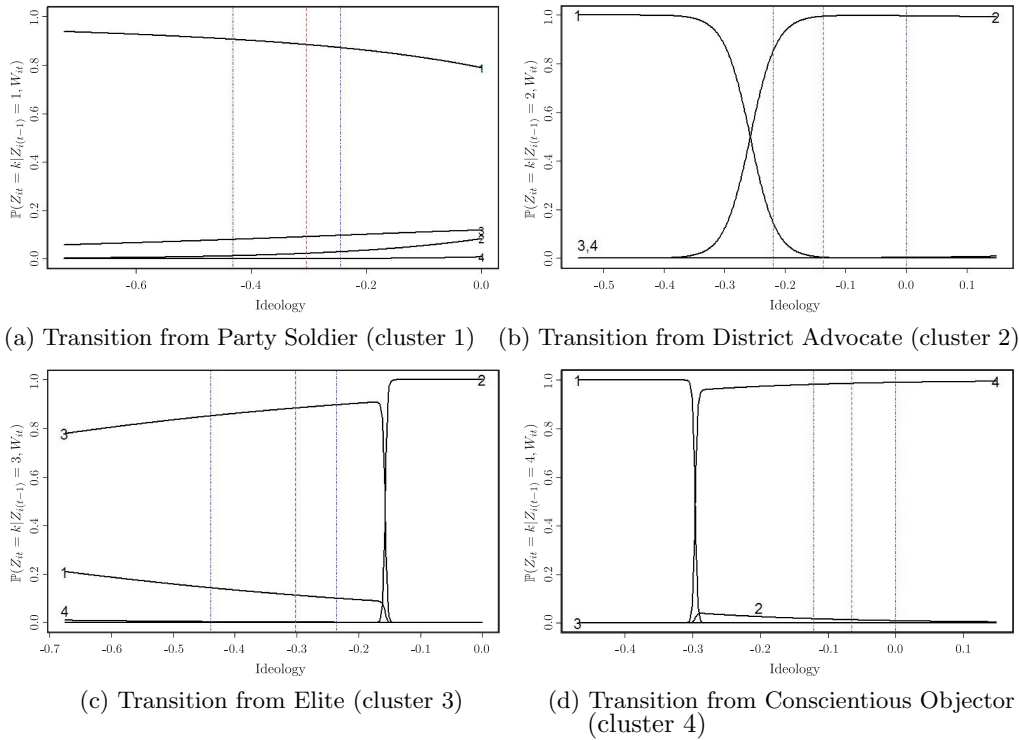(c) Transition from Elite (cluster 3)   (d) Transition from Conscientious Objector (cluster 4)

Figure 7. Transition cluster probabilities for Congressional data. Each curve represents, for the varying values along the horizontal axis of the explanatory variable (ideology), the probability of belonging to the cluster whose number is adjacent to the curve given that they currently belong to cluster: (a) 1 (party soldiers) (b) 2 (district advocates) (c) 3 (elites) (d) 4 (conscientious objectors). The horizontal axis spans the range of all MCs ideology; the dot-dashed lines give the 25% and 75% quantiles of ideologies of MCs who have belonged to the corresponding cluster, and the long dashed line gives the mean.

The computation in such a case should not be drastically different than in the simpler model. In general, if a more sophisticated Markov switching model is more appropriate to the context, it should be feasible to alter the model accordingly. See Frühwirth-Schnatter (2006) for more information on various Markov switching models.

Acknowledgements

and two referees for their constructive suggestions and comments.

Appendices

Appendix A. Proof of $\pi(Z_t|Z_{t-1}, \mathbf{X}_1, \ldots, \mathbf{X}_{t-1}) = \beta_{Z_{t-1}Z_t}$

We wish to find a tractable form of $\pi(Z_t|Z_{t-1}, \mathbf{X}_1, \ldots, \mathbf{X}_{t-1})$. Note that the subscript i is suppressed for ease of notation. The key equality to show is that $\pi(\mathbf{X}_1, \ldots, \mathbf{X}_{t-1}| Z_{t-1}, Z_t) = \pi(\mathbf{X}_1, \ldots, \mathbf{X}_{t-1}|Z_{t-1})$. This is shown by

$$
\begin{aligned}
&\pi(\mathbf{X}_1, \ldots, \mathbf{X}_{t-1}|Z_{t-1}, Z_t) \\
&= \sum_{Z_1, \ldots, Z_{t-2}} \frac{\pi(\mathbf{X}_1, \ldots, \mathbf{X}_{t-1}|Z_1, \ldots, Z_t)\pi(Z_1, \ldots, Z_t)}{\pi(Z_{t-1})\beta_{Z_{t-1}Z_t}} \\
&= \sum_{Z_1, \ldots, Z_{t-2}} \frac{\alpha_{Z_1}\pi(\mathbf{X}_1|Z_1)\beta_{Z_1Z_2}\pi(\mathbf{X}_2|\mathbf{X}_1, Z_2)\cdots\beta_{Z_{t-2}Z_{t-1}}\pi(\mathbf{X}_{t-1}|\mathbf{X}_{t-2}, Z_{t-1})}{\pi(Z_{t-1})} \\
&= \sum_{Z_1, \ldots, Z_{t-2}} \frac{\pi(\mathbf{X}_1, \ldots, \mathbf{X}_{t-1}|Z_1, \ldots, Z_{t-1})\pi(Z_1, \ldots, Z_{t-1})}{\pi(Z_{t-1})} \\
&= \sum_{Z_1, \ldots, Z_{t-2}} \frac{\pi(\mathbf{X}_1, \ldots, \mathbf{X}_{t-1}, Z_1, \ldots, Z_{t-1})}{\pi(Z_{t-1})} \\
&= \pi(\mathbf{X}_1, \ldots, \mathbf{X}_{t-1}|Z_{t-1}).
\end{aligned}
$$

From this it is straightforward to obtain the result in the following way

$$
\begin{aligned}
\pi(Z_t|Z_{t-1}, \mathbf{X}_1, \ldots, \mathbf{X}_{t-1}) &= \frac{\pi(\mathbf{X}_1, \ldots, \mathbf{X}_{t-1}, Z_t|Z_{t-1})}{\pi(\mathbf{X}_1, \ldots, \mathbf{X}_{t-1}|Z_{t-1})} \\
&= \frac{\pi(\mathbf{X}_1, \ldots, \mathbf{X}_{t-1}|Z_{t-1}, Z_t)\beta_{Z_{t-1}Z_t}}{\pi(\mathbf{X}_1, \ldots, \mathbf{X}_{t-1}|Z_{t-1})} = \beta_{Z_{t-1}Z_t}.
\end{aligned}
$$

Appendix B. Deriving the Tractable Form of $Q(\Theta, \widehat{\Theta})$

Let $\mathbf{Z}_i$ denote the latent cluster assignments $Z_{i1}, \ldots, Z_{iT_i}$. Then we have

$$
\begin{aligned}
Q(\Theta, \widehat{\Theta}) &= \sum_{\mathbf{Z}_1, \ldots, \mathbf{Z}_n} \sum_{i=1}^{n} \Bigg[\log(\alpha_{Z_{it}}\pi(\mathbf{X}_{i1}|Z_{i1}, \Theta)) \\
&\quad + \sum_{t=2}^{T_i} \log(\beta_{Z_{i(t-1)}Z_{it}}\pi(\mathbf{X}_{it}|\mathbf{X}_{i(t-1)}, Z_{it}, \Theta)) \Bigg] \prod_{j=1}^{n} \mathbb{P}(\mathbf{Z}_j|\mathcal{X}_j, \widehat{\Theta}) \\
&= \sum_{\mathbf{Z}_1, \ldots, \mathbf{Z}_n} \sum_{i=1}^{n} \sum_{\ell_1=1}^{K} \cdots \sum_{\ell_{T_i}=1}^{K} \mathbb{1}_{[Z_{i1}=\ell_1]} \cdots \mathbb{1}_{[Z_{iT_i}=\ell_T]}
\end{aligned}
$$

$$
\cdot \left[\log(\alpha_{\ell_1} \pi(\mathbf{X}_{i1} | Z_{i1} = \ell_1, \Theta)) \right.
$$

$$
\left. + \sum_{t=2}^{T_i} \log(\beta_{\ell_{t-1}\ell_t} \pi(\mathbf{X}_{it} | \mathbf{X}_{i(t-1)}, Z_{it} = \ell_t, \Theta)) \right] \prod_{j=1}^{n} \mathbb{P}(\mathbf{Z}_j | \mathcal{X}_j, \widehat{\Theta})
$$

$$
= \sum_{i=1}^{n} \sum_{\ell_1=1}^{K} \cdots \sum_{\ell_{T_i}=1}^{K} \left[\log(\alpha_{\ell_1} \pi(\mathbf{X}_{i1} | Z_{i1} = \ell_1, \Theta)) \right.
$$

$$
\left. + \sum_{t=2}^{T_i} \log(\beta_{\ell_{t-1}\ell_t} \pi(\mathbf{X}_{it} | \mathbf{X}_{i(t-1)}, Z_{it} = \ell_t, \Theta)) \right]
$$

$$
\cdot \sum_{\mathbf{Z}_1,\ldots,\mathbf{Z}_n} 1_{[Z_{i1}=\ell_1]} \cdots 1_{[Z_{iT_i}=\ell_{T_i}]} \prod_{j=1}^{n} \mathbb{P}(\mathbf{Z}_j | \mathcal{X}_j, \widehat{\Theta}). \tag{B.1}
$$

We can simplify this last expression by noting that, for a fixed i,

$$
\sum_{\mathbf{Z}_1,\ldots,\mathbf{Z}_n} 1_{[Z_{i1}=\ell_1]} \cdots 1_{[Z_{iT_i}=\ell_T]} \prod_{j=1}^{n} \mathbb{P}(\mathbf{Z}_j | \mathcal{X}_j, \widehat{\Theta})
$$

$$
= \left[\prod_{j \neq i} \sum_{\mathbf{Z}_j} \mathbb{P}(\mathbf{Z}_j | \mathcal{X}_j, \widehat{\Theta}) \right] \mathbb{P}(Z_{i1} = \ell_1, \ldots, Z_{iT_i} = \ell_{T_i} | X_i, \widehat{\Theta})
$$

$$
= \mathbb{P}(Z_{i1} = \ell_1 | \mathcal{X}_i, \widehat{\Theta}) \prod_{t=1}^{T_i} \mathbb{P}(Z_{it} = \ell_t | Z_{i(t-1)} = \ell_{t-1}, \mathcal{X}_i, \widehat{\Theta}). \tag{B.2}
$$

Thus we can rewrite Q in the more tractable form

$$
Q(\Theta, \widehat{\Theta}) = \sum_{i=1}^{n} \sum_{\ell_1=1}^{K} \cdots \sum_{\ell_{T_i}=1}^{K} \left[\log(\alpha_{\ell_1}) \right.
$$

$$
\left. + \sum_{t=2}^{T_i} \log(\beta_{\ell_{t-1}\ell_t}) \right] \mathbb{P}(Z_{i1} = \ell_1 | \mathcal{X}_i, \widehat{\Theta}) \prod_{s=2}^{T_i} \mathbb{P}(Z_{is} = \ell_s | Z_{i(s-1)} = \ell_{s-1}, \mathcal{X}_i, \widehat{\Theta})
$$

$$
+ \sum_{i=1}^{n} \sum_{\ell_1=1}^{K} \cdots \sum_{\ell_{T_i}=1}^{K} \left[\log(\pi(\mathbf{X}_{i1} | Z_{i1} = \ell_1, \Theta)) \right.
$$

$$
\left. + \sum_{t=2}^{T_i} \log(\pi(\mathbf{X}_{it} | \mathbf{X}_{i(t-1)}, Z_{it} = \ell_t, \Theta)) \right]
$$

$$
\cdot \mathbb{P}(Z_{i1} = \ell_1 | \mathcal{X}_i, \widehat{\Theta}) \prod_{s=2}^{T_i} \mathbb{P}(Z_{is} = \ell_s | Z_{i(s-1)} = \ell_{s-1}, \mathcal{X}_i, \widehat{\Theta}). \tag{B.3}
$$

Appendix C. Computing Marginal and Conditional Distributions of Z_{it}

As in the main text, the subscript i is suppressed. Note that the dependence on Θ has also been suppressed for ease of notation. We start by noticing that

$$\mathbb{P}(Z_T|Z_{T-1}, \mathcal{X}) \propto \pi(\mathcal{X}|Z_T, Z_{T-1})\mathbb{P}(Z_T|Z_{T-1}) \propto \beta_{Z_{T-1}Z_T}\pi(\mathbf{X}_T|\mathbf{X}_{T-1}, Z_T).$$
(C.1)

We can then find the iterative relationship, for $2 \leq t < T$,

$$
\begin{aligned}
\mathbb{P}(Z_t&|Z_{t-1}, \mathcal{X})\\
&\propto \pi(\mathbf{X}_t, \dots, \mathbf{X}_T|\mathbf{X}_1, \dots, \mathbf{X}_{t-1}, Z_t, Z_{t-1})\mathbb{P}(Z_t|Z_{t-1})\\
&\propto \beta_{Z_{t-1}Z_t} \sum_{\ell_{t+1}} \cdots \sum_{\ell_T} \pi(\mathbf{X}_t, \dots, \mathbf{X}_T, Z_{t+1} = \ell_{t+1}, \dots, Z_T = \ell_T|\mathbf{X}_{t-1}, Z_t)\\
&\propto \beta_{Z_{t-1}Z_t} \pi(\mathbf{X}_t|\mathbf{X}_{t-1}, Z_t) \sum_{\ell_{t+1}} \Big(\beta_{Z_t\ell_{t+1}}\pi(\mathbf{X}_{t+1}|\mathbf{X}_t, Z_{t+1} = \ell_{t+1}) \cdots \\
&\qquad \cdot \sum_{\ell_T} \Big(\beta_{\ell_{T-1}\ell_T}\pi(\mathbf{X}_T|\mathbf{X}_{T-1}, Z_T = \ell_T)\Big) \cdots \Big)\\
&\propto \beta_{Z_{t-1}Z_t}\pi(\mathbf{X}_t|\mathbf{X}_{t-1}, Z_t) \sum_{\ell_{t+1}} q(Z_{t+1} = \ell_{t+1}|Z_t).
\end{aligned}
$$
(C.2)

Appendix D. Deriving the Parameter Updates

D1. Update α

Letting λ_α be the Lagrange multiplier corresponding to the constraint $\sum_{\ell=1}^{K} \alpha_\ell = 1$, we have

$$
\begin{aligned}
\frac{\partial Q}{\partial \alpha_h} &= \frac{\partial}{\partial \alpha_h}\bigg\{ \sum_{i=1}^{n} \sum_{Z_{i1}=1}^{K} \cdots \sum_{Z_{iT_i}=1}^{K} \log(\alpha_{Z_{i1}})\mathbb{P}(Z_{i1}|\mathcal{X}_i, \widehat{\Theta}) \prod_{s=2}^{T_i} \mathbb{P}(Z_{is}|Z_{i(s-1)}, \mathcal{X}_i, \widehat{\Theta})\\
&\qquad - \lambda_\alpha \Big(\sum_{\ell=1}^{K} \alpha_\ell - 1 \Big) \bigg\}\\
&= \frac{1}{\alpha_h} \sum_{i=1}^{n} \mathbb{P}(Z_{i1} = h|\mathcal{X}_i, \widehat{\Theta}) \prod_{s=2}^{T_i} \sum_{Z_{is}=1}^{K} \mathbb{P}(Z_{is}|Z_{i(s-1)}, \mathcal{X}_i, \widehat{\Theta}) - \lambda_\alpha,
\end{aligned}
$$
(D.1)

hence

$$\hat{\alpha}_h = \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}(Z_{i1} = h|\mathcal{X}_i, \widehat{\Theta}).$$
(D.2)

D2. Update β_{hk}

Here we have K Lagrange multipliers corresponding to the constraint on each row of the transition matrix, specifically that $\sum_{k=1}^{K} \beta_{jk} = 1$ for $j = 1, \ldots, K$. Denoting these as λ_j, we have

$$\frac{\partial Q}{\partial \beta_{hk}} = \frac{\partial}{\partial \beta_{hk}} \Big\{ \sum_{i=1}^{n} \sum_{Z_{i1}=1}^{K} \cdots \sum_{Z_{iT_i}=1}^{K} \sum_{t=2}^{T_i} \log(\beta_{Z_{i(t-1)}Z_{it}}) \mathbb{P}(Z_{i1}|\mathcal{X}_i, \widehat{\Theta})$$

$$\prod_{s=2}^{T_i} \mathbb{P}(Z_{is}|Z_{i(s-1)}, \mathcal{X}_i, \widehat{\Theta}) - \sum_{j=1}^{K} \lambda_j \Big(\sum_{m=1}^{K} \beta_{jm} - 1 \Big) \Big\}. \qquad \text{(D.3)}$$

We can make this expression more tractable by noticing that, for any i and $t \geq 2$,

$$\sum_{Z_{i1}=1}^{K} \cdots \sum_{Z_{iT_i}=1}^{K} \log(\beta_{Z_{i(t-1)}Z_{it}}) \mathbb{P}(Z_{i1}|\mathcal{X}_i, \widehat{\Theta}) \prod_{s=2}^{T_i} \mathbb{P}(Z_{is}|Z_{i(s-1)}, \mathcal{X}_i, \widehat{\Theta})$$

$$= \sum_{Z_{i1}=1}^{K} \Big(\mathbb{P}(Z_{i1}|\mathcal{X}_i, \widehat{\Theta}) \sum_{Z_{i2}=1}^{K} \Big(\mathbb{P}(Z_{i2}|Z_{i1}, \mathcal{X}_i, \widehat{\Theta}) \cdots$$

$$\sum_{Z_{it}=1}^{K} \Big(\log(\beta_{Z_{i(t-1)}Z_{it}}) \mathbb{P}(Z_{it}|Z_{i(t-1)}, \mathcal{X}_i, \widehat{\Theta}) \cdots \sum_{Z_{iT_i}=1}^{K} \mathbb{P}(Z_{iT_i}|Z_{i(T_i-1)}, \mathcal{X}_i, \widehat{\Theta}) \Big) \cdots \Big)$$

$$= \sum_{Z_{i(t-1)}=1}^{K} \mathbb{P}(Z_{i(t-1)}|\mathcal{X}_i, \widehat{\Theta}) \sum_{Z_{it}=1}^{K} \log(\beta_{Z_{i(t-1)}Z_{it}}) \mathbb{P}(Z_{it}|Z_{i(t-1)}, \mathcal{X}_i, \widehat{\Theta}). \qquad \text{(D.4)}$$

Thus we can find that

$$\frac{\partial Q}{\partial \beta_{hk}} = \frac{1}{\beta_{hk}} \sum_{i=1}^{n} \sum_{t=2}^{T_i} \mathbb{P}(Z_{i(t-1)} = h|\mathcal{X}_i, \widehat{\Theta}) \mathbb{P}(Z_{it} = k|Z_{i(t-1)} = h, \mathcal{X}_i, \widehat{\Theta}) - \lambda_h \quad \text{(D.5)}$$

and hence

$$\hat{\beta}_{hk} = \frac{\sum_{i=1}^{n} \sum_{t=2}^{T_i} \mathbb{P}(Z_{i(t-1)} = h|\mathcal{X}_i, \widehat{\Theta}) \mathbb{P}(Z_{it} = k|Z_{i(t-1)} = h, \mathcal{X}_i, \widehat{\Theta})}{\sum_{i=1}^{n} \sum_{t=2}^{T_i} \mathbb{P}(Z_{i(t-1)} = h|\mathcal{X}_i, \widehat{\Theta})}. \qquad \text{(D.6)}$$

D3. Update λ, $\boldsymbol{\mu}_k$, and Σ_k

Before we proceed, we need a few preliminaries. The partial derivatives of (2.3) and (2.4) with respect to $\boldsymbol{\mu}_k$, $k = 1, \ldots, K$, are

$$\frac{\partial}{\partial \boldsymbol{\mu}_k} \log(\pi(\mathbf{X}_{i1}|k, \Theta)) = \frac{\partial}{\partial \boldsymbol{\mu}_k} \Big(-\frac{1}{2} \Big) [(\mathbf{X}_{i1} - \boldsymbol{\mu}_k)' \Sigma_k^{-1} (\mathbf{X}_{i1} - \boldsymbol{\mu}_k)]$$

$$= \Sigma_k^{-1}\mathbf{X}_{i1} - \Sigma_k^{-1}\boldsymbol{\mu}_k, \tag{D.7}$$

$$\frac{\partial}{\partial\boldsymbol{\mu}_k}\log(\pi(\mathbf{X}_{it}|\mathbf{X}_{i(t-1)},k,\Theta)) = \frac{\partial}{\partial\boldsymbol{\mu}_k}(-\frac{1}{2})\Big[(\mathbf{X}_{it} - (1-\lambda)\mathbf{X}_{i(t-1)} - \lambda\boldsymbol{\mu}_k)'$$

$$\Sigma_k^{-1}(\mathbf{X}_{it} - (1-\lambda)\mathbf{X}_{i(t-1)} - \lambda\boldsymbol{\mu}_k)\Big]$$

$$= \lambda\Sigma_k^{-1}(\mathbf{X}_{it} - (1-\lambda)\mathbf{X}_{i(t-1)}) - \lambda^2\Sigma_k^{-1}\boldsymbol{\mu}_k, \tag{D.8}$$

with respect to Σ_k^{-1}, $k = 1,\dots,K$, are

$$\frac{\partial}{\partial\Sigma_k^{-1}}\log(\pi(\mathbf{X}_{i1}|k,\Theta))$$

$$= \frac{\partial}{\partial\Sigma_k^{-1}}\left[\frac{1}{2}\log|\Sigma_k^{-1}| - \frac{1}{2}tr(\Sigma_k^{-1}(\mathbf{X}_{it} - \boldsymbol{\mu}_k)(\mathbf{X}_{it} - \boldsymbol{\mu}_k)')\right]$$

$$= \Sigma_k - \frac{1}{2}diag(\Sigma_k) - (\mathbf{X}_{it} - \boldsymbol{\mu}_k)(\mathbf{X}_{it} - \boldsymbol{\mu}_k)' + \frac{1}{2}diag((\mathbf{X}_{it} - \boldsymbol{\mu}_k)(\mathbf{X}_{it} - \boldsymbol{\mu}_k)')$$

$$= \Sigma_k \circ \left(\frac{1}{2}I_p\right) - (\mathbf{X}_{it} - \boldsymbol{\mu}_k)(\mathbf{X}_{it} - \boldsymbol{\mu}_k)' \circ \left(\frac{1}{2}I_p\right), \tag{D.9}$$

where $\circ$ is the Hadamard product, and similarly

$$\frac{\partial}{\partial\Sigma_k^{-1}}\log(\pi(\mathbf{X}_{it}|\mathbf{X}_{i(t-1)},k,\Theta)) \tag{D.10}$$

$$= \Sigma_k \circ \left(\frac{1}{2}I_p\right) - \left(\mathbf{X}_{it} - \lambda\boldsymbol{\mu}_k - (1-\lambda)\mathbf{X}_{i(t-1)}\right)\left(\mathbf{X}_{it} - \lambda\boldsymbol{\mu}_k - (1-\lambda)\mathbf{X}_{i(t-1)}\right)' \circ \left(\frac{1}{2}I_p\right),$$

and with respect to λ is

$$\frac{\partial}{\partial\lambda}\log(\pi(\mathbf{X}_{it}|\mathbf{X}_{i(t-1)},k,\Theta)) = \frac{\partial}{\partial\lambda}(-\frac{1}{2})\Big[(\mathbf{X}_{it} - \lambda(\boldsymbol{\mu}_k - \mathbf{X}_{i(t-1)}) - \mathbf{X}_{i(t-1)})'$$

$$\Sigma_k^{-1}(\mathbf{X}_{it} - \lambda(\boldsymbol{\mu}_k - \mathbf{X}_{i(t-1)}) - \mathbf{X}_{i(t-1)})\Big]$$

$$= (\mathbf{X}_{it} - \mathbf{X}_{i(t-1)})'\Sigma_k^{-1}(\boldsymbol{\mu}_k - \mathbf{X}_{i(t-1)})$$

$$- \lambda(\boldsymbol{\mu}_k - \mathbf{X}_{i(t-1)})'\Sigma_k^{-1}(\boldsymbol{\mu}_k - \mathbf{X}_{i(t-1)}). \tag{D.11}$$

To make the form of $Q(\Theta, \widehat{\Theta})$ more tractable, we notice that for any i,

$$\sum_{Z_{i1}=1}^{K}\cdots\sum_{Z_{iT_i}=1}^{K}\log(\pi(\mathbf{X}_{i1}|Z_{i1},\Theta))\mathbb{P}(Z_{i1}|\mathcal{X}_j,\widehat{\Theta})\prod_{s=2}^{T_i}\mathbb{P}(Z_{is}|Z_{i(s-1)},\mathcal{X}_i,\widehat{\Theta})$$

$$= \sum_{Z_{i1}=1}^{K}\log(\pi(\mathbf{X}_{i1}|Z_{i1},\Theta))\mathbb{P}(Z_{i1}|\mathcal{X}_j,\widehat{\Theta})\prod_{s=2}^{T_i}\sum_{Z_{is}=1}^{K}\mathbb{P}(Z_{is}|Z_{i(s-1)},\mathcal{X}_i,\widehat{\Theta})$$

$$= \sum_{Z_{i1}=1}^{K}\log(\pi(\mathbf{X}_{i1}|Z_{i1},\Theta))\mathbb{P}(Z_{i1}|\mathcal{X}_j,\widehat{\Theta}), \tag{D.12}$$

and for any i and $t \geq 2$,

$$\sum_{Z_{i1}=1}^{K} \cdots \sum_{Z_{iT_i}=1}^{K} \log(\pi(\mathbf{X}_{it}|\mathbf{X}_{i(t-1)}, Z_{it}, \Theta)) \mathbb{P}(Z_{i1}|\mathcal{X}_j, \widehat{\Theta}) \prod_{s=2}^{T_i} \mathbb{P}(Z_{is}|Z_{i(s-1)}, \mathcal{X}_i, \widehat{\Theta})$$

$$= \sum_{Z_{i1}=1}^{K} \cdots \sum_{Z_{it}=1}^{K} \mathbb{P}(Z_{i1}|\mathcal{X}_j, \widehat{\Theta}) \Big[\prod_{s=2}^{t-1} \mathbb{P}(Z_{is}|Z_{i(s-1)}, \mathcal{X}_i, \widehat{\Theta}) \Big] \log(\pi(\mathbf{X}_{it}|\mathbf{X}_{i(t-1)}, Z_{it}, \Theta))$$

$$\mathbb{P}(Z_{it}|Z_{i(t-1)}, \mathcal{X}_i, \widehat{\Theta}) \Big[\prod_{u=t+1}^{T_i} \sum_{Z_{iu}=1}^{K} \mathbb{P}(Z_{iu}|Z_{i(u-1)}, \mathcal{X}_i, \widehat{\Theta}) \Big]$$

$$= \sum_{Z_{it}=1}^{K} \log(\pi(\mathbf{X}_{it}|\mathbf{X}_{i(t-1)}, Z_{it}, \Theta)) \mathbb{P}(Z_{it}|\mathcal{X}_i, \widehat{\Theta}). \tag{D.13}$$

With the above it is not difficult to find, for each distribution parameter, the value which maximizes $Q(\Theta, \widehat{\Theta})$ where the other parameters are fixed, hence finding the solutions to the coordinate ascent approach. The solutions are found as follows. The derivative of $Q(\Theta, \widehat{\Theta})$ with respect to λ is

$$\frac{\partial Q}{\partial \lambda} = \sum_{i=1}^{n} \sum_{t=2}^{T_i} \sum_{Z_{it}=1}^{K} \mathbb{P}(Z_{it}|\mathcal{X}_i, \widehat{\Theta}) \Big[(\mathbf{X}_{it} - \mathbf{X}_{i(t-1)})' \Sigma_k^{-1} (\boldsymbol{\mu}_k - \mathbf{X}_{i(t-1)})$$

$$- \lambda (\boldsymbol{\mu}_k - \mathbf{X}_{i(t-1)})' \Sigma_k^{-1} (\boldsymbol{\mu}_k - \mathbf{X}_{i(t-1)}) \Big]. \tag{D.14}$$

Hence the update for λ is

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} \sum_{t=2}^{T_i} \sum_{Z_{it}=1}^{K} \mathbb{P}(Z_{it}|\mathcal{X}_i, \widehat{\Theta})(\mathbf{X}_{it} - \mathbf{X}_{i(t-1)})' \Sigma_k^{-1} (\boldsymbol{\mu}_k - \mathbf{X}_{i(t-1)})}{\sum_{j=1}^{n} \sum_{s=2}^{T_j} \sum_{Z_{js}=1}^{K} \mathbb{P}(Z_{js}|\mathcal{X}_j, \widehat{\Theta})(\boldsymbol{\mu}_k - \mathbf{X}_{j(s-1)})' \Sigma_k^{-1} (\boldsymbol{\mu}_k - \mathbf{X}_{j(s-1)})}. \tag{D.15}$$

The derivative of $Q(\Theta, \widehat{\Theta})$ with respect to $\boldsymbol{\mu}_k$, $k = 1, \ldots, K$, is

$$\frac{\partial Q}{\partial \boldsymbol{\mu}_k} = \sum_{i=1}^{n} \Big\{ \mathbb{P}(Z_{i1} = k|\mathcal{X}_i, \widehat{\Theta}) \big[\Sigma_k^{-1} \mathbf{X}_{i1} - \Sigma_k^{-1} \boldsymbol{\mu}_k \big]$$

$$+ \sum_{t=2}^{T_i} \mathbb{P}(Z_{it} = k|\mathcal{X}_i, \widehat{\Theta}) \big[\lambda \Sigma_k^{-1} (\mathbf{X}_{it} - (1-\lambda)\mathbf{X}_{i(t-1)}) - \lambda^2 \Sigma_k^{-1} \boldsymbol{\mu}_k \big] \Big\}. \tag{D.16}$$

Hence the update for $\boldsymbol{\mu}_k$ is

$$\hat{\boldsymbol{\mu}}_k = \frac{\sum_{i=1}^{n} \Big\{ \mathbb{P}(Z_{i1} = k|\mathcal{X}_i, \widehat{\Theta})\mathbf{X}_{i1} + \lambda \sum_{t=2}^{T_i} \mathbb{P}(Z_{it} = k|\mathcal{X}_i, \widehat{\Theta})(\mathbf{X}_{it} - (1-\lambda)\mathbf{X}_{i(t-1)}) \Big\}}{\sum_{i=1}^{n} \Big\{ \mathbb{P}(Z_{i1} = k|\mathcal{X}_i, \widehat{\Theta}) + \lambda^2 \sum_{t=2}^{T_i} \mathbb{P}(Z_{it} = k|\mathcal{X}_i, \widehat{\Theta}) \Big\}}. \tag{D.17}$$

Let $\mathbf{A}_{ik} = \mathbf{X}_{i1} - \boldsymbol{\mu}_k$ and $\mathbf{B}_{itk} = \mathbf{X}_{it} - \lambda\boldsymbol{\mu}_k - (1-\lambda)\mathbf{X}_{i(t-1)}$. Then the derivative of $Q(\Theta, \widehat{\Theta})$ with respect to Σ_k^{-1}, $k = 1, \ldots, K$, is

$$\frac{\partial Q}{\partial \Sigma_k^{-1}} = \sum_{i=1}^{n} \left\{ \mathbb{P}(Z_{i1} = k | \mathcal{X}_i, \widehat{\Theta}) \left[\Sigma_k \circ \left(\frac{1}{2}I_p\right) - \mathbf{A}_{ik}\mathbf{A}'_{ik} \circ \left(\frac{1}{2}I_p\right) \right] \right.$$
$$\left. + \sum_{t=2}^{T_i} \mathbb{P}(Z_{it} = k | \mathcal{X}_i, \widehat{\Theta}) \left[\Sigma_k \circ \left(\frac{1}{2}I_p\right) - \mathbf{B}_{itk}\mathbf{B}'_{itk} \circ \left(\frac{1}{2}I_p\right) \right] \right\}. \quad \text{(D.18)}$$

Hence the update for Σ_k can be found as

$$\widehat{\Sigma}_k = \frac{\sum_{i=1}^{n} \left\{ \mathbb{P}(Z_{i1} = k | \mathcal{X}_i, \widehat{\Theta})\mathbf{A}_{ik}\mathbf{A}'_{ik} + \sum_{t=2}^{T_i} \mathbb{P}(Z_{it} = k | \mathcal{X}_i, \widehat{\Theta})\mathbf{B}_{itk}\mathbf{B}'_{itk} \right\}}{\sum_{i=1}^{n} \left\{ \mathbb{P}(Z_{i1} = k | \mathcal{X}_i, \widehat{\Theta}) + \sum_{t=2}^{T_i} \mathbb{P}(Z_{it} = k | \mathcal{X}_i, \widehat{\Theta}) \right\}}.$$

Appendix E. Details on Variables Measured on MCs

Our variables include the number of district offices operated by the MC, the proportion of legislative staff assigned to the district (rather than Capitol Hill), the number of bill introductions, cosponsorships, and amendments made by the MC, the number of one-minute speeches the MC gives on the House floor, the number of editorials and opinion pieces he or she writes for state and national papers, the total amount of campaign money he or she raises, the total amount of money he or she contributes to the party campaign committee (the Democratic Congressional Campaign Committee, which collects and redistributes funds for election campaigns), the total amount of money he or she contributes directly to colleagues, the percentage of the time he or she votes with the party on party votes (defined as votes where a majority of one party votes against a majority of the other), the percentage of the time he or she votes with the party leadership on leadership votes (defined as votes where the leadership of one party votes one way and the leadership of the other party the other), the number of issue areas (out of eighteen) in which he or she introduces legislative bills, the proportion of his or her cosponsorships that are bipartisan (i.e., for measures introduced by an MC from the Republican party), and the percentage of his or her introduced bills that are referred to a committee on which he or she sits.

Each index used in the clustering algorithm was constructed in the following way. First, each of the raw variables described above was standardized at each time point to eliminate global temporal patterns or shifts. Second, each variable was associated with one index (existing literature on congressional behavior pointed us to which variables should fall under each index). Finally, each index

was constructed by averaging the associated variables. As an example, suppose that there are three raw variables, $X_{it}^{(1)}$, $X_{it}^{(2)}$, and $X_{it}^{(3)}$, associated with index one. Then the first index X_{it1} (the first entry in $\mathbf{X}_{it}$) is computed as

$$X_{it1} = \frac{1}{3} \left[\frac{X_{it}^{(1)} - \bar{X}_t^{(1)}}{s_t^{(1)}} + \frac{X_{it}^{(2)} - \bar{X}_t^{(2)}}{s_t^{(2)}} + \frac{X_{it}^{(3)} - \bar{X}_t^{(3)}}{s_t^{(3)}} \right],$$

for each i and t, where $\bar{X}_t^{(\ell)}$ and $s_t^{(\ell)}$ are the sample mean and sample standard deviation respectively of $\{X_{it}^{(\ell)}\}_{i=1}^n$, $\ell = 1, 2, 3$.

References

Anderlucci, L. and Viroli, C. (2014). Covariance pattern mixture models for multivariate longitudinal data with application to the health and retirement study. arXiv preprint arXiv:1401.1301.

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49**, 803-821.

Chung, H., Park, Y. and Lanza, S. T. (2005). Latent transition analysis with covariates: Pubertal timing and substance use behaviours in adolescent females. *Statist. Medicine* **24**, 2895-2910.

Collins, L. M., Graham, J. W., Rousculp, S. S., Fidler, P. L., Pan, J. and Hansen, W. B. (1994). Latent transition analysis and how it can address prevention research questions. *NIDA Research Monograph* **142**, 81-111.

Collins, L. M. and Wugalter, S. E. (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Res.* **27**, 131-157.

De la Cruz-Mesía, R., Quintana, F. A. and Marshall, G. (2008). Model-based clustering for longitudinal data. *Comput. Statist. Data Anal.* **52**, 1441-1457.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39**, 1-38.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models: Modeling and Applications to Random Processes*. Springer.

Gaffney, S. and Smyth, P. (1999). Trajectory clustering with mixtures of regression models. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 63-72.

Genolini, C. and Falissard, B. (2010). KmL: k-means for longitudinal data. *Comput. Statist.* **25**, 317-328.

Graham, J. W., Collins, L. M., Wugalter, S. E., Chung, N. K. and Hansen, W. B. (1991). Modeling transitions in latent stage-sequential processes: a substance use prevention example. *J. Consulting and Clinical Psychology* **59**, 48-57.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *J. Classification* **2**, 193-218.

Lou, S., Jiang, J. and Keng, K. (1993). Clustering objects generated by linear regression models. *J. Amer. Statist. Assoc.* **88**, 1356-1362.

Luan, Y. and Li, H. (2003). Clustering of time-course gene expression data using a mixed-effects model with B-splines. *Bioinformatics* **19**, 474-482.

McNicholas, P. D. and Murphy, T. B. (2010). Model-based clustering of longitudinal data. *Canad. J. Statist.* **38**, 153-168.

Meilă, M. (2003). Comparing clusterings by the variation of information. In *Learning Theory and Kernel Machines*, 173-189. Springer.

Poole, K. T. and Rosenthal, H. (1997). *Congress: A Political-Economic History of Roll Call Voting*. Oxford University Press.

Ray, S. and Mallick, B. (2006). Functional clustering by Bayesian wavelet methods. *J. Roy. Statist. Soc. Ser. B* **68**, 305-332.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53-65.

Scott, S. L. James, G. M. and Sugar, C. A. (2005). Hidden Markov models for longitudinal comparisons. *J. Amer. Statist. Assoc.* **24**, 179-207.

Vermunt, J. K., Langeheine, R. and Bockenholt, U. (1999). Discrete-time discrete-state latent Markov models with time-constant and time-varying covariates. *J. Educational and Behavioral Statistics* **24**, 179-207.

Wu, C. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.* **11**, 95-103.

Department of Biostatistics, University of Iowa, 145 N. Riverside Drive, Iowa City, IA 52242, USA.

E-mail: daniel-sewell@uiowa.edu

Department of Statistics, University of Illinois at Urbana-Champaign, 725 South Wright Street, Champaign, IL 61820, USA.

E-mail: yuguo@illinois.edu

Department of Political Science, University of Illinois at Urbana-Champaign, 1407 W Gregory Drive, Urbana, IL 61801, USA.

E-mail: bernhard@illinois.edu

Department of Political Science, University of Illinois at Urbana-Champaign, 1407 W Gregory Drive, Urbana, IL 61801, USA.

E-mail: tsulkin@illinois.edu