Dynamic Hierarchical Latent Variable Models for Categorical Survey Data

Evan Munro * Serena Ng †

March 29, 2019

Abstract

Existing methods that construct summary indices from multivariate discrete data often treat discrete variables as if they were continuous. We adapt statistical methods used for text and genomic analysis to estimate latent variables from a cross-section of categorical survey data. The indices differ according to the conditional independence assumptions made on the joint distribution of survey responses. Advantages of these approaches over existing methods are shown in an example that estimates wealth indices from data for Latin America and the Caribbean. In the case when a series of time dependent survey responses is available, we introduce dynamics by considering discrete versions of the basic Markov switching model and state space model. The proposed methods work well when applied to the Michigan survey on consumer confidence, and the Gallup Poll data on environmental sentiment.

JEL Classification:
Keywords:

*Graduate School of Business, Stanford University. Email: munro@stanford.edu,
†Department of Economics, Columbia University, and NBER. Email: serena.ng@columbia.edu

1 Introduction

A variable is continuous if it can take on an infinite number of values between any two points in a range. This contrasts with categorical or discrete data which can only take on distinct number of values. According to the statistical package SPSS, there are three types of categorical variables: nominal variables, which represent unordered categories (such as zip code and NAICS codes), ordinal data (such as satisfaction ratios of excellent/good/bad), and numerical (count) data, which represent ordered data such as age. Nominal and ordinal data are said to be non-metrical because the distance between two categories have no interpretation.

Economic well being is naturally measured by income and wealth, and continuous data for such variables are readily available for most developed countries. But there is a lack of reliable income data in the developing world. Proxy data that is easier to collect, such as ownership of assets and housing characteristics, is available. As a result, non-metrical data are often the only type of data with which we can create indices of economic well-being. Many researchers use survey data on household consumption from the Demographic and Health Surveys (DHS), or the Living Standard Measurement Surveys (LSMS), to extract a wealth index that proxies for income for each household in the sampl. And regardless of stages of economic development, some concepts simply do not have a metrical representation. For example, the sentiment towards 'the environment' is difficult to measure because it is not a tangible good. It is common for studies of health and socioeconomic status to involve some type of non-metrical data. As we now have the capability to conduct surveys online and text data can be easily scraped, non-metrical data are increasingly available. See, for example, Bloom and 'van Reenen' (2007) and Baker *et al.* (2016).

While there exist many dimension reduction tools for analyzing continuous data and their cardinal features, methods for analyzing non-metrical data and their qualitative features are more limited and generally not as well understood. As will be seen from our review of these methods below, existing methods either treat discrete data as if they were continuous, or ignore the distance between the ordered outcomes, or convert the multinomial outcomes into distinct binary responses, which induces spurious correlation in the data.

This paper makes two major contributions to dimension reduction analysis of categorical data. The first is to adapt the widely popular Latent Dirichlet Allocation (LDA) Bayesian approach developed in Blei *et al.* (2003) for topic modeling to study a cross-section of multivariate ordered or unordered categorical survey responses. The LDA setup allows us to analyze the data as discrete instead of treating them as if they were continuous. We also relate LDA to Grade of Membership (GoM), a statistical approach developed in Erosheva *et al.* (2007) to analyze disability survey data. Both LDA and GoM are Bayesian hierarchical latent variable models for discrete data. We show that they impose different conditional independence assumptions on the joint distribution of the data and, as a result, recover different parameters of interest. While the GoM assumptions permit estimation of individual-level indices, the LDA assumptions permit estimation of group-level indices. Both GoM and LDA are more flexible versions of a third approach known as latent class modeling. We use the DHS data for Latin America and the Caribbean to illustrate the insights that these probabilistic methods can deliver, and contrast them with different variations of principal components.

The second contribution is to introduce dynamics into the latent variable models, making it possible to analyze time-dependent categorical data. Depending on the desired trade-off between model flexibility and computational complexity, dynamics can be added using a discrete-valued latent variable or a continuous-valued latent variable. Two methods are considered. The first is a version of Hamilton (1989)'s Markov-switching model for categorical response variables, and has been suggested in the statistics literature by MacDonald and Zucchini (1997). The model assumes that there is a single hidden state at each point in time that generates outcome variables; this state is one of a finite set of states and follows a Markov-switching process. The second is a version of a linear state space model for categorical response variables. The model assumes that at each point in time, there is a mixture of states that generate the outcome variables. The mixture probabilities follow a random walk process and each state corresponds to a different multinomial distribution over categorical outcomes. As a sanity check for the new dynamic methods, we show in the application section that the extracted index from the Michigan data closely correlates with the published index of consumer sentiment, but the mapping from response patterns to index for the probabilistic model is fully parametrized and interpretable. We also extract indices from categorical variables from Gallup Poll Social Series (GPSS) data. We show that the resulting indices provide useful insights into consumer's preferences towards non-tangible goods around recessions.

2 Dimension Reduction of Categorical Data: Static Case

Most surveys collect non-metrical data in its raw form, but the series that we use in economic analysis are continuous variables transformed from the raw data in a non-trivial way. For example, instead of the responses to the five survey questions posed by researchers at the University of Michigan about current and future economic conditions, we typically analyze the Index of Consumer Sentiment. For each of the five questions, a relative score is calculated as the difference between the percentage of respondents giving favorable and unfavorable responses. The five components are then combined with equal weights. As Ludvigson (2004) points out, the index aggregates responses to disparate questions with qualitative response categories, and in practice, the index has little to no forecasting power for consumer spending patterns beyond what is found in aggregate statistics on actual consumer purchases. But this does not preclude the possibility that questions underlying the index have useful information. One reason why the five survey responses are not further analyzed is that there are few tools that specifically compress information in discrete data. Such methods are useful, since the discrete nature of the data requires different modeling choices.

Suppose that we have a dataset with $i = 1, \dots, N$ survey respondents, who each respond to $j = 1, \dots, J$ survey questions with categorical outcomes. Each survey question has possible outcomes indexed by $l_j = 1, \dots, L_j$. A missing or don't know response is included as response L_j for each question. The number of choices can vary across survey questions. Survey respondents may be grouped into $c = 1, \ldots, C$ groups, in which case N_c denotes the number of respondents in group c. Groups can be formed in a variety of ways: respondents who are of the same sex, the same country, or who responded to a survey at the same time, for example. The raw dataset X is an $N \times J$ matrix where $X_{ij} \in \{1, \ldots, L_j\}$ corresponds to the index of the outcome that survey respondent i chose for question j. For models where respondents are assigned to groups, we sometimes add a third index c to denote the group membership of individual i, and refer to responses X_{cij}. The goal is to recover a $K-1$-dimensional real-valued index g_i that is a low dimensional representation of X_i, the vector of survey responses for individual i.

2.1 Frequentist Methods

Suppose $K = 2$ and we want to assign a single index to each individual based on their survey responses. This has been implemented in several ways. Though these dimension reduction methods for categorical data are popular and simple, they each involve significant limitations.

Simple Averaging It is always possible to construct a common factor by averaging, see Pesaran (2006). For example, an individual-level index can be created by averaging the individual z-score for each of the responses. For ordered multivariate discrete survey data, the average is taken over the numerical value that each ordered outcome is assigned. For example, Bloom and Reenen (2010) defines scores of 1 to 5 for each of the 18 survey questions on management practices. A firm has 'good management' if the average over the 18 scores is high. Ludvigson (2004), on the other hand, uses the relative percentage of respondents selecting the "high" vs. the "low" ordered outcomes. These averaging methods rely on the assumption that the distance between each ordered categorical outcome is equal, which may not be an accurate assumption.

PCA One alternative is to treat ordered discrete data as if they were continuous and apply PCA to the scaled X directly. This relies on assuming that the ordinal data has constant distance between different categories, which would be a questionable assumption for many survey datasets. Vyas and Kumaranayake (2006) discusses the advantages and limitations of using PCA to construct indices of socio-economic status.

The Filmer-Pritchett Method For unordered categorical variables, a popular method is to extract an index using the Filmer-Pritchett method, which is a version of PCA for discrete data. The article Filmer and Pritchett (2001) has over 5000 citations and is widely used to construct well-being indices for developing countries. The method involves converting each variable with multiple outcomes to a set of binary variables, and factorizing a matrix of binary responses. However, as shown in Kolenikov and Angeles (2009), converting multinomial outcomes to binary variables in this way introduces spurious negative correlations within the multiple columns that are mapped from a single question. The factors extracted from matrix factorization, then, are not an optimal low-dimensional representation of the data.

Practitioners generally drop all rows of X that contain missing data. Each categorical variable j is transformed into j^* binary variables, where $j^* = L_j - 1$ and $\tilde{J} = \sum_{j=1}^{J} j^*$. Each column is normalized by the its mean and standard deviation to derive a transformed $N \times \tilde{J}$ matrix of data denoted $\tilde{X}$. Let S be the $\tilde{J} \times \tilde{J}$ sample covariance matrix of $\tilde{X}$. The first k-principal components are

$$Z = \tilde{X} A$$

where A is a $\tilde{J} \times k$ orthonormal matrix consisting of the eigenvectors corresponding to the k largest eigenvalues of S. When applied to the DHS data, the first column of Z may be used as the estimated wealth index $g_i^{(FP)}$.

The Filmer-Pritchett index is thus constructed as a weighted sum of indicators for each of the possible responses in the survey, corresponding to indicators for asset ownership, for example, which explains the maximum variance in the data. A non-trivial issue arises because the binary variables are derived from categorical variables, As a consequence, there will be spurious negative correlations between variables from the same categorical variable. The directions of maximum variance may be related to the spurious negative correlations in the augmented data matrix rather than in directions that correspond to differences in wealth between households in the survey. Kolenikov and Angeles (2009) and Ng (2015) have pointed out these issues and examined alternative principal components and factor analysis methods for extracting latent variables from both continuous and categorical data, although neither found clear evidence that a single method is best.

Analysis of Polychoric Correlations Yet another alternative is to replace S in PCA above by a polychoric correlation matrix. A polychoric correlation coefficient is a measure of association for ordinal variables proposed by Karl Pearson. The idea is that if two variables underlying the ordinal data are joint normal, its contingency table can be seen as discretization of the joint distribution. The polychoric correlation is then the linear correlation of the assumed joint normal distribution.

For structural equation modeling, a parametric model for ordinal survey data is needed. Lee *et al.* (1990b) assumes that X_j, the data for survey question j, is generated from an underlying continuous variable w_j with thresholds corresponding to each of the outcomes L_j. Applying PCA to the polychoric covariance matrix gives a household wealth index. Since the polychoric correlations matrix is based on latent continuous variables by assumption, it is also possible to use use maximum likelihood with a number of identifying restrictions, as in Lee *et al.* (1990a). The approach of polychoric correlation has a rich theoretical foundation. However, it is suitable for ordinal data only, and cannot handle unordered categorical variables.

Missing responses or 'don't know' responses are very common in survey data, and refusing to answer certain questions could be relevant to characterizing a household. The methods described in this subsection requires dropping all respondents that have missing data or imputing their responses. This can involve a large loss of information and decrease in sample size.

3 Bayesian Analysis

The Bayesian approach to categorical data that we follow treats the survey response outcome data as random variables, and estimates directly the joint distribution of survey responses $p(X_1, \ldots, X_N)$. In order to make the joint distribution tractable, some simplifying conditional independence assumptions must be made. We derive those independence assumptions by adding some latent variables, and specifying an interpretable process that includes the latent variables and generates the discrete data. In some literature, this approach is referred to as specifying a generative model, since the process generating the full joint distribution of the data is estimated. It is also referred to as a probabilistic graphical model. The model is probabilistic because each of the latent variables and outcome variables are specified as random; they are graphical because the conditional independence assumptions between variables can be represented in a directed acyclic graph.

To see why conditional independence assumptions are necessary when modeling the full joint distribution of data, imagine a survey dataset with $N = 100$ respondents, $J = 5$ questions and $L_j = 5$ possible responses to each question. For each individual, $\prod_{j=1}^{J} L_j - 1 = 5^5 - 1$ parameters are required to capture all possible dependencies between questions. For a dataset with $N = 100$ individuals, in order to capture all possible dependencies between responses for all individuals, we require a number of parameters in the order of 100^{5^5}, which is an infeasibly large number. As a result, the approaches described augment the joint distribution of observed responses with a set of random latent variables, each with their own prior distribution. A series of conditional independence assumptions involving the latent variables and the observed data are made, which allows factorization and estimation of the joint distribution of the data and latent variables. The estimated latent variables provide interpretable low-dimensional summaries of the observed data and replace the PCA-based indexes described earlier in this section. Bayesian hierarchical latent variable models have appeared recently in the economics literature, including in analysis of consumer choice in Ruiz *et al.* (2018); Athey *et al.* (2018), CEO management practices in Bandiera *et al.* (2017), and text analysis of central bank communication in Hansen *et al.* (2018).

For the remainder of this section, we introduce the three models in Table 1, which each assume multinomial data is generated from a mixture of multinomial distributions. The main difference between the models is where in the hierarchy of aggregate, group, and individual the mixture is located. Each corresponds to a different set of conditional independence assumptions involving the data and the specified latent variables.

3.1 Grade of Membership Model

Grade of Membership, used by Erosheva *et al.* (2007) to analyze disability survey data, is a hierarchical latent variable model for categorical survey data that is a structural alternative to PCA-based methods. The latent variables and hyperparameters of the model with K hidden profiles and their dimensions are in Table 1. There are two sets of Dirichlet hyperparameters, $\alpha \in \mathbb{R}^k$ and $\eta_j \in \mathbb{R}^{L_j}$, for the multinomial latent variables in the model. There are J multinomial distributions $\beta_{jk} \in \Delta^{L_j - 1}$

	GoM	LDA	Latent Class
Mixture Level	Individual	Group	Aggregate
Dynamic Version		SS-M	MS-M
Profile Latent Variable (β)	$K \times J \times L_j$	$K \times P$	$K \times P$
Assignment Latent Variable (z)	$N \times J$	$N \times 1$	$C \times 1$
Mixture Latent Variable (g)	$N \times K$	$C \times K$	$K \times 1$
Mixture Hyperparameter (α)	$N \times K$	$C \times K$	$K \times 1$
Profile Hyperparameter (η)	$K \times J \times L_j$	$K \times P$	$K \times P$

Table 1: Hierarchical Latent Variable Models

for each profile k giving profile-specific distributions over responses. There is the mixture parameter for each individual, $g_i \in \Delta^{K-1}$, which provides a multinomial distribution over K profiles for each individual. z is an indicator variable giving the assigned profile $z_{ij} \in \{1, \ldots, K\}$ for each individual i and each question j.

The full specification of the GoM model assumes that the discrete outcome data X_{ij} is generated as follows:

$$X_{ij} | z_{ij}, \beta_j \sim \text{Multinomial}(\beta_{j,z_{ij}})$$

$$z_{ij} | g_i \sim \text{Multinomial}(g_i)$$

$$g_i \sim \text{Dirichlet}(\alpha)$$

$$\beta_{jk} \sim \text{Dirichlet}(\eta_j)$$

The model described relies on a set of conditional independence assumptions to factorize the joint distribution. These are listed below.

GoM 1 ***: Conditional Independence of Questions*** *The joint probability of assigning question j to profile k and of individual i selecting response l_j is independent of that individual's other responses, given the profiles and the individual-level mixture over profiles. In addition, responses are independent of individual-level mixtures given the profile assignments, and assignments are independent of profiles given the individual-level mixtures.*

$$Pr(X_i, z_i | g_i, \beta) = \prod_{j=1}^{J} Pr(X_{ij}, z_{ij} | g_i, \beta)$$

$$Pr(X_{ij} | g_i, z_{ij}, \beta) = Pr(X_{ij} | z_{ij}, \beta)$$

$$Pr(z_{ij} | g_i, \beta) = Pr(z_{ij} | g_i)$$

GoM 2 ***: Conditional Independence of Individuals****: Conditional on individual i's mixture over profiles g_i, the probability of a certain response for individual i is independent of other indi-*

viduals' responses.

$$Pr(X, z|g, \beta) = \prod_{i=1}^{N} Pr(X_i, z_i|g_i, \beta)$$

GoM 3 ***: Independence of Profiles and Mixtures (GoM)*** *Profiles and mixtures are independent of each other. In addition,* g_i *is independent of* g_r *for* $r \neq i$ *and* β_{jk} *is independent of* β_{mf} *for* $f \neq k$ *and* $m \neq j$*.*

$$Pr(g) = \prod_{i=1}^{N} Pr(g_i)$$

$$Pr(\beta) = \prod_{k=1}^{K} \prod_{j=1}^{J} Pr(\beta_{kj})$$

With these assumptions, the GoM joint likelihood factorizes as follows: First, the factorize the joint distribution into marginal and conditional distributions.

$$Pr(\beta, g, z, X) = Pr(\beta, g)Pr(X, z|\beta, g)$$

Then, apply GoM 3 and GoM 2:

$$Pr(\beta, g, z, X) = \prod_{k=1}^{K} \prod_{j=1}^{J} Pr(\beta_{jk}) \prod_{i=1}^{N} Pr(g_i) \prod_{i=1}^{N} Pr(X_i, z_i|\beta, g_i)$$

Then, apply GoM 1 and factorize the joint distribution of X_{ij} and z_{ij} into marginal and conditional distributions:

$$Pr(\beta, g, Z, X) = \prod_{k=1}^{K} \prod_{j=1}^{J} Pr(\beta_{jk}) \prod_{i=1}^{N} Pr(g_i) \prod_{i=1}^{N} \prod_{j=1}^{J} Pr(X_{ij}|\beta, z_{ij})Pr(z_{ij}|\beta, g_i)$$

Lastly, include the explicit probabilities from the model specification, which indicates multinomial conditional distributions for the outcome variables and the assignments. $Pr(X_{ij}|\beta, z_{ij}) = \beta_{j,z_{ij},X_{ij}}$ and $Pr(Z_{ij}|\beta, g_i) = g_{i,z_{ij}}$.

$$Pr(\beta, g, Z, X) = \prod_{k=1}^{K} \prod_{j=1}^{J} Pr(\beta_{jk}) \prod_{i=1}^{N} Pr(g_i) \prod_{i=1}^{N} \prod_{j=1}^{J} g_{i,z_{ij}} \beta_{j,z_{ij},X_{ij}}$$

The probabilities of the latent variables β and G depend on the hyperparameters α and η and the Dirichlet density function. For example, for g_i:

$$p(g_i) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} g_{i1}^{\alpha_1 - 1} \dots g_{iK}^{\alpha_K - 1}$$

The assumption that the outcome variables are multinomial is robust to both ordered and unordered categorical data. Missing data is simply included as a possible categorical outcome. GoM assumes a more natural underlying parametric model for discrete data than assuming categorical variables are normally distributed or derived from thresholded continuous variables. For example, in the household survey data, a researcher could assume that there are two extreme types of households in a developing country. One is a poor household, profile 1, which has high probability of having pit toilets, low probability of owning a bicycle or car, and, low probability of having a savings account. The other is a rich household, profile 2, which has a high probability of having a flush toilet, car, and a savings account. Individual households are modeled as a mixture of these two extreme types; a wealthy household has g_{i2} close to 1, and a poor household has g_{i2} close to 0. A middle income household might have a mixture closer to the center of the simplex. This approach allows for infinite heterogeneity in household characteristics (since there are infinite points defined on the simplex over extreme profiles), while maintaining an interpretable and low-dimensional underlying structure (since mixtures of only K profiles are assumed to generate all household responses). The mixture weight g_{i2} of a household on the wealthy profile can be considered to be a probability estimate of an individual-level wealth index, and used in further economic analysis. This is a Bayesian parametric alternative to existing PCA-based methods for creating individual-level wealth indexes.

GoM can be estimated using gibbs sampling, or a fast variational inference procedure using the R package `mixedMem`. Due to the conjugacy of the Multinomial and Dirichlet distributions, the sampling procedure is straightforward. See Erosheva *et al.* (2007) for the full details of the sampling procedure for GoM with an added prior distribution on α.

3.2 Latent Dirichlet Allocation Model

GoM, PCA, and polychoric correlation methods produce a decomposition of the $N \times J$ raw survey data X into individual-level weights on profiles g_i and profile-specific weights on responses β_k. In some cases, though, applied economists may not require an individual-level index, but are instead interested in an group-level index. For example, Bloom and Reenen (2010) explain cross-country differences in productivity using a country-wide index of management ability, which is an average of firm-level management score in each country. The firm-level score is derived from averaging ordinal outcome variables from a firm management survey. It is always possible to estimate a group level index simply by averaging an individual-level index. However, if a group level index is the target, the reduction in parameters from estimating a group rather than individual-index directly can be significant, and can be used to offset the increase in parameters that can result from modifying some of GoM's problematic independence assumptions.

GoM 1, conditional question independence, is not likely to hold in most survey data. There is likely to be much more complex dependence between an individual's responses to different survey questions than can be captured by the individual-specific mixtures g_i. One way to loosen GoM 1 is to model the probability $P(X_{i1}, X_{i2}, \ldots, X_{iJ})$ directly, rather than factorizing the joint distribution into a set of J multinomial distributions conditional on the latent variable g. Each row X_i of the $N \times J$ matrix corresponds to one permutation p of all possible survey response permutations indexed by $\{1, \ldots, P\}$, where $P = \prod_{j=1}^{J} L_j$. The multinomial distribution over outcomes $\beta_k \in \Delta^{P-1}$, is P-dimensional. This is in contrast to GoM, where each profile has a separate multinomial distribution for each question j over L_j possible responses.

Assuming that the low dimensions latent variable that determines the structure of outcomes is group-level rather than individual-level means that g is now $C \times K$ rather than $N \times K$. We map each individual i to group c and assign a second index to each row X_{ci} to indicate which group an individual is assigned to. Even for a moderate number of survey questions J, the number of permutations P can be quite large. However, the model considered is actually a reformulated Latent Dirichlet Allocation Blei *et al.* (2003), designed for the sparse, high-dimensional settings of text analysis, and able to handle large P and large C. The full specification of the model is:

$$X_{ci} | z_{ci}, \beta \sim \text{Multinomial}(\beta_{z_{ci}})$$

$$z_{ci} | g_c \sim \text{Multinomial}(g_c)$$

$$g_c \sim \text{Dirichlet}(\alpha)$$

$$\beta_k \sim \text{Dirichlet}(\eta)$$

The conditional independence assumptions for the model are below. GoM 1 is now eliminated entirely, since we directly estimate the probability of each possible survey response permutation for an individual. GoM 2 has now been strengthened so that group-level mixtures capture all relationships between individual's responses: the response of an individual is now independent of the response of other individuals given the group-level mixture assignments, rather than individual-level mixture assignments. GoM 3 is maintained in an appropriately modified form.

LDA 1 ***: Conditional Independence of Individuals Given Group Structure*** *Conditional on group c's mixture over profiles g_c, the probability of a certain response and assignment for individual i in group c is independent of other individuals' responses. In addition, responses are independent of group-level mixtures given the profile assignments and assignments are independent of profiles given the group-level mixtures.*

$$Pr(X, z | g, \beta) = \prod_{i=1}^{N} Pr(X_{ci}, z_{ci} | g_c, \beta)$$

$$Pr(X_{ci} | g_c, z_{ci}, \beta) = Pr(X_{ci} | z_{ci}, \beta)$$

$$Pr(z_{ci} | g_c, \beta) = Pr(z_{ci} | g_{ci})$$

LDA 2 ***: Independence of Profiles and Mixtures*** *g_c is independent of g_d for $c \neq d$ and β_k is independent of β_f for $f \neq k$.*

The joint likelihood for LDA is factorized using the conditional independence assumptions:

$$p(\beta, g, Z, X) = \prod_{k=1}^{K} p(\beta_k) \prod_{c=1}^{C} p(g_c) \prod_{i=1}^{N} g_{c,z_{ci}} \beta_{z_{ci},X_{ci}}$$

The estimate of g_c provides an index for group c. For the DHS survey example, an economist could assume that there are two profiles of household wealth: one with a higher probability on bundles of purchases of higher quality household assets and the other with higher probability on permutations of responses with low asset ownership. Due to country-specific economic institutions, there is a country-specific probability that a household is assigned to the profile correlated with poverty. The weight on the profile with high asset ownership can be interpreted as a country-specific wealth index.

The Gibbs sampling procedure for estimating LDA is described in detail in Griffiths and Steyvers (2004). Estimation procedures are available in R, for example using the package `topicmodels` or `lda`. As in GoM, the conjugacy of the Dirichlet and multinomial distributions results in conjugate posterior distributions and computationally efficient sampling.

3.3 Latent Class Models

In the preceding section, we characterized GoM as an individual-level mixture model and LDA as an group-level mixture model. It is also worth pointing out the relationship of these latent variable models to a Bayesian latent class model. A Bayesian latent class model assumes outcomes in a dataset come from a finite set of profiles, where each individual in a group is assigned to a single profile. A profile corresponds to a multinomial distribution over outcomes.

Below is the specification of a latent class model:

$$X_{ci}|\beta, z_c \sim \text{Multinomial}(\beta_{z_c})$$

$$z_c|g \sim \text{Multinomial}(g)$$

$$g \sim \text{Dirichlet}(\alpha)$$

$$\beta_k \sim \text{Dirichlet}(\eta)$$

In a latent class model, the mixture g is aggregate-level, so is K-dimensional, rather than $N \times K$ as in GoM. All individuals in a group are assigned to the same profile, whereas in LDA individuals in a group can be assigned to different profiles, and in GoM each individual is assigned to a mixture of profiles. The conditional independence assumption of assignments for individuals in a latent class model are strict compared to the corresponding assumptions in GoM and LDA.

Latent Class 1 ***: Conditional Independence of Assignments Given Mixture****: Conditional on the aggregate mixture over profiles g, the profile assignment of group c, z_c, is independent of the* *profile assignment of group d, z_d.*

$$Pr(z|g, \beta) = \prod_{c=1}^{C} Pr(z_c|g)$$

Latent Class 2 ***: Conditional Independence of Individuals Given Assignments****: Conditional on an individual's assignment based on their group membership, individual responses are independent. Furthermore, responses are independent of the aggregate-level mixture given the profile assignments.*

$$Pr(X|g, z, \beta) = \prod_{i=1}^{N} Pr(X_{ci}, |z_c, \beta)$$

Latent Class 3 ***: Independence of Profiles*** *β_k is independent of β_f for $f \neq k$.*

The resulting joint likelihood of the model is

$$p(\beta, g, z, X) = Pr(g) \prod_{k=1}^{K} Pr(\beta_k) \prod_{c=1}^{C} p(z_c|g) \prod_{i=1}^{N} \beta_{z_c, X_{ci}}$$

The model assumes homogeneity within groups in a way that is not likely to hold in most survey data where groups contain more complex forms of heterogeneity. As a result, the interpretation of g is more limited than g_i or g_c in GoM and LDA: the model simply estimates the predominant patterns in the data at a aggregate-level, rather than an individual or group level profiles that can be interpreted as an index.

So far, in this section we have showed that GoM is a parametric alternative to Filmer-Pritchett and Polychoric correlation-based PCA methods that does not involve unwarranted transformations of the data, and allows ordered, unordered, and missing responses in categorical data. Furthermore, GoM is part of a class of hierarchical latent variable models with mixture weights at the individual, group (LDA), or aggregate-level (Latent Class). Every Bayesian hierarchical latent variable model relies on a set of conditional independence assumptions, some of which are strong. What remains is a short discussion of identification in these models before we discuss introducing dynamics in the mixture weights for time series survey data where some of the GoM and LDA independence assumptions are not likely to hold.

3.4 Identification of Static Hierarchical Latent Variable Models

In the section on GoM, we described a hypothetical "rich" and "poor" profile in the context of household wealth index estimation based on household survey data. Without prior restrictions, the posterior likelihood of a model with Profile 1 as the "poor" profile and Profile 2 as the "rich" profile is the same as a model with the labels swapped. To avoid this label-swapping issue, we use the prior distribution of β_k to label each profile before estimation. For two profiles, each profile is assigned a permutation of survey responses that correspond to an extreme response: for example, in the DHS example, owning every asset. Then, the hyperparameter η is adjusted so that β_{kj} has high prior probability of being close to zero for the extreme response assigned to the other profile. This can be considered as the discrete and Bayesian analogy to the lower triangular restriction that is placed on factor models for identification purposes.

In addition to the label-swapping issue, there is also the potential issue of the posterior likelihood being flat for multiple values of the model parameters, which corresponds to a frequentist notion of parameter identification. If there are multiple parameter values for which the posterior likelihood is similar, the sampling procedure may converge to different values for parameters depending on the starting point of the algorithm, and interpretation of the results is difficult. In a Bayesian framework the shape of the posterior distribution depends both on the model specified and restrictions implemented, as well as the data available. So, assessing the flatness of the peak in the posterior distribution of the parameters is best done using typical Bayesian methods for checking posterior convergence, see Gelman *et al.* (2013). For example, running chains from multiple starting values and ensuring that posterior estimates converge to similar values.

We have found that for small values of K, and with the label-swapping restrictions on the prior parameters described above, that the parameter posterior distributions generally appear to be single-peaked with consistent estimated mean values for the parameters across model runs. For large values of K, more prior restrictions on the profiles may be required to avoid flat areas in posterior distributions.

4 Application to DHS Surveys

Using the static models described, we explore the low-dimensional structure of a dataset of 187,616 survey responses from the Demographic and Health Surveys in 5 Caribbean and Latin American countries from 2009-2012. Young (2012) shows that aggregate economic measures from developing countries from data sources like the Penn World Tables are based on very little hard data. The DHS survey data, which includes millions of survey responses on demographics, health, and economic outcomes from dozens of developing world households since the 1990s, has become a crucial source of data for deriving consumption and income estimates for households in the developing world. The World Bank publishes wealth indices derived from Filmer-Pritchett type PCA on a variety of asset indicators from the survey; a variety of other researchers have derived their own PCA-based socioeconomic status indicators from the data, for example, in Jean *et al.* (2016) to train machine learning models to associate local income measures with satellite images.

The survey data used in our example include variables on water quality and house floor quality (categorized from 1 to 4), toilet quality (scored from 1 to 5) and binary variables on ownership of electricity, radio, tv, fridge, motorbike, car and phone. The mean response for each of these variables for each country is in Table 2. Colombia is the wealthiest country by most asset-based measures, and Nicaragua and Haiti are the poorest.

Most economists would agree that the asset ownership indicators are correlated with measures of household wealth and expenditure, but there is little agreement on how best to extract measures

Country	water	toilet	floor	electric	radio	tv	fridge	motorbike	car	phone
Colombia (CO)	3.50	4.66	2.86	0.95	0.72	0.87	0.71	0.22	0.10	0.90
Guyana (GY)	2.78	3.95	2.92	0.71	0.55	0.73	0.55	0.09	0.14	0.80
Haiti (HT)	2.52	2.46	2.15	0.30	0.51	0.23	0.08	0.07	0.04	0.75
Nicaragua (NI)	3.24	1.75	1.52	0.23	0.53	0.19	0.06	0.15	0.04	0.56
Peru (PE)	3.45	3.44	2.24	0.82	0.84	0.73	0.33	0.13	0.09	0.73

Table 2: Mean Response for Each Country

of wellbeing from noisy and heterogeneous survey response data. Existing methods in the literature that will be used as benchmark are

i PCA on the discrete data (PCA)

ii the Filmer-Pritchett method (FPRIT)

iii PCA with polychoric correlation (PCHOR)

iv mean of individual Z-scores across all responses (ZSCORE)

The first three are application of principal components to different transformations of the data. They may be thought of as sophisticated weighted averages of the data. Method four takes a simple average of the standardized data. The three PCA-based methods estimate a very similar first principal component with positive loadings on all variables. The benefits of these four approaches are their simplicity; they are easy to compute, and a higher score means that an individual has higher than average asset ownership in the pooled sample. Table 3 provides a table of loadings for each of the PCA methods and the amount of variance in the data explained by the first principal component. For Filmer-Pritchett, where categorical variables are split into multiple binary variables, only the loading for the binary variable representing the highest quality of asset ownership is shown.

	PCA	FPRIT	PCHOR
water_qual	0.28	0.34	0.63
toilet_qual	0.40	0.38	0.81
floor_qual	0.36	0.10	0.74
electric	0.39	0.37	0.90
radio	0.18	0.16	0.40
tv	0.41	0.38	0.94
fridge	0.36	0.33	0.90
motorbike	0.13	0.11	0.35
car	0.16	0.14	0.58
phone	0.32	0.28	0.76
% Var.	**39 %**	**24 %**	**53 %**

Table 3: First Principal Component Loadings for PCA-Based Methods

The interpretation is simple but limited: a positive household index means that individual has higher weighted asset ownership than the average household in the sample. A single component explains 55% or less of the variance in the data for all three methods. Increasing the dimension of the index by adding principal components increases the proportion of variance explained, but at the cost of losing clear interpretability of the index. Furthermore, it is not clear how to evaluate sampling uncertainty around these indices.

The next two subsections consider the probabilistic approaches considered earlier.

4.1 Household-Level Index by GoM

The GoM model extracts an individual-level index from the raw response data pooled across countries, where the data is not standardized as in PCA-based methods. Each individual is modeled as a mixture, g_i, over profiles, which each involve a separate set of multinomial distributions, β_k, over responses for each question in the survey. Figure 1 shows the multinomial distributions over each question for each of the two profiles estimated by GoM with $K = 2$. The interpretation of the weights profiles for GoM is more detailed than a weighted average of the responses to the survey. A household with the estimated g_{i2} close to one, has high weight on the multinomial distribution in the second column in the figure. This indicates that the home likely has high water quality and has a small but positive probability of having a motorcycle or a car. A household with a weight close to 0 on Profile 2 indicates that while the home might have a radio, it is not likely to have good floor quality or fridge. A medium income household with good water and floor quality but no fridge or phone might have a weight on Profile 2 closer to 0.5.

In Figure 2, the histograms for the household-level indices for Nicaragua and Guyana are plotted for both PCA on the untransformed responses and GoM. The first principal component in Nicaragua has a right skew, and in Guyana has a left skew; for Guyana, many individuals have average or above average asset ownership, and in Nicaragua many individuals are very poor. For GoM, there is a similar shape to the distribution of household indices. Guyana has a GDP/capita in the middle of the countries in the sample: most individuals are assigned to profile 2, but there is a significant proportion who are poorer and have a higher weight on Profile 1. In Nicaragua, the poorest country in the western hemisphere, most individuals have 0 weight on the wealthier profile, but there is a small proportion of households who are categorized as wealthy.

GoM is more computationally expensive, and doesn't give as smooth of an index as PCA does, since for GoM many household weights are concentrated at 0 and 1. As shown in Figure 1, the posterior estimate of g_{ik} on the profile corresponding to higher asset ownership, like the PCA methods, does provide a good indicator of the wealth of a household. It is also important to point out that g_{ik} is the posterior estimate of a probability model, and has standard errors that can be used to conduct inference: for example, it can be used to test whether the latent variable estimate for household i compared to household j is equal or not. Furthermore, the GoM index retains interpretability as K increases, whereas interpretability becomes difficult for PCA as the number of components increase.

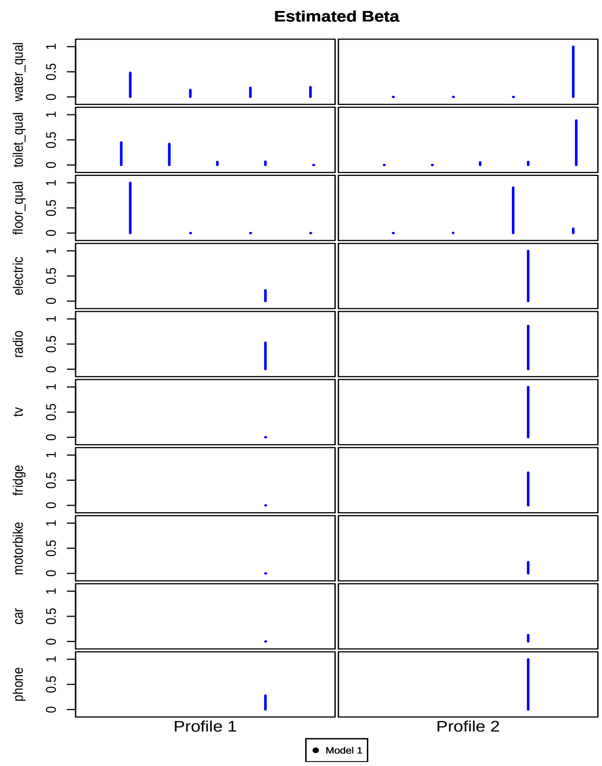

Figure 1: GoM Estimated Profile-Specific Distributions with $K = 2$

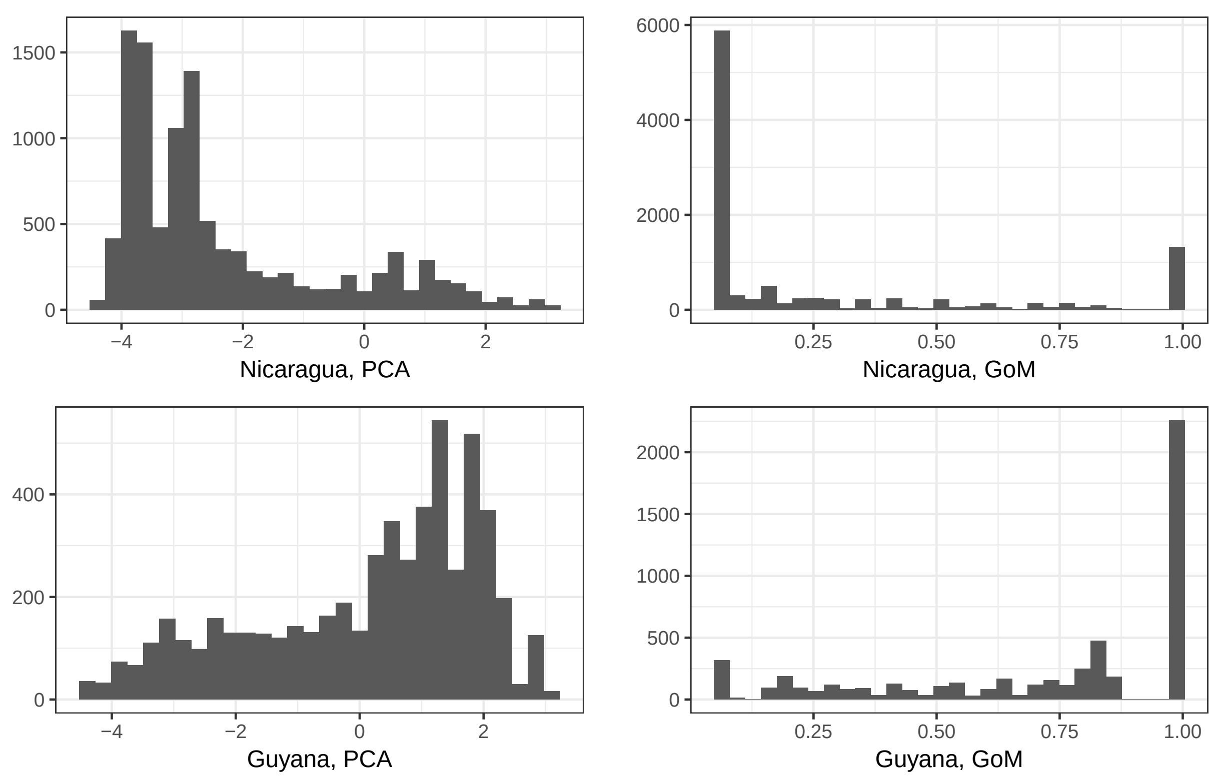

Figure 2: Distribution of Household-Level Indices for GoM vs. PCA

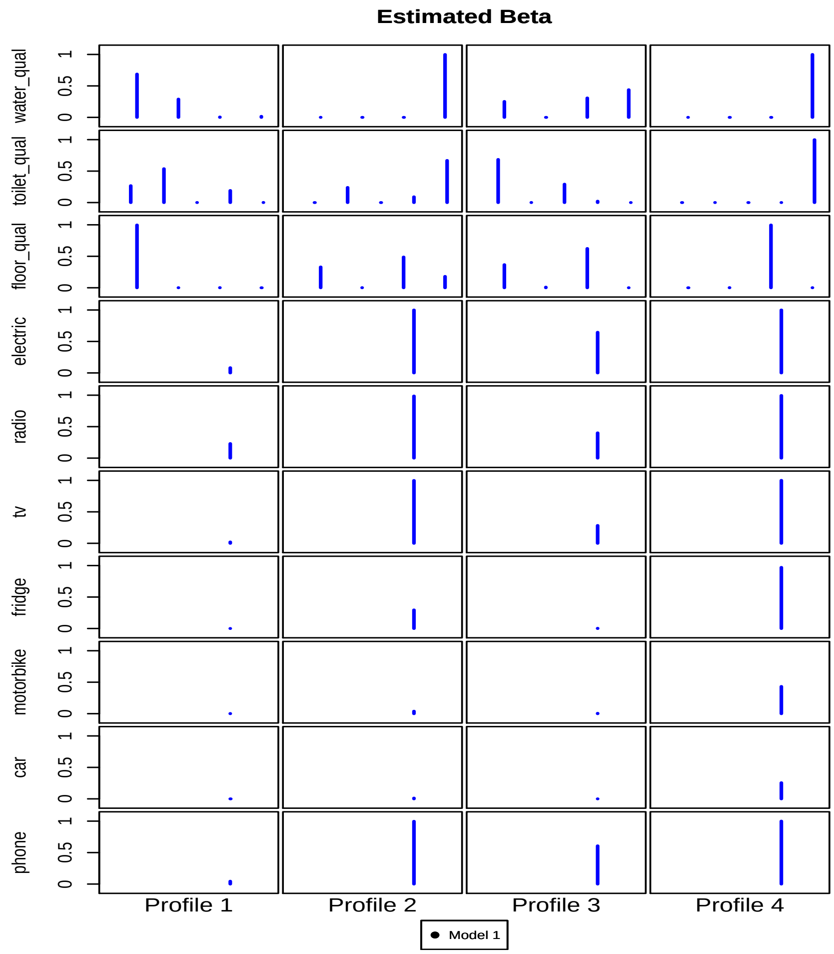

Figure 3: GoM Estimated Profile-Specific Distributions with $K = 4$

To show this explicitly, Figure 3 plots the posterior estimate of β_{jk}, the profile-specific multinomial distribution over question responses, for a model with $K = 4$ profiles and the same 10 questions. The rich heterogeneity of households in the datasets is even more apparent when estimating the model with four profiles. Profile 2 and Profile 4 are the wealthier profiles. A household with a high weight on profile 4 has a high chance of having high water, toilet, and floor quality, and owning all assets, except a car. Profile 2 has high quality water infrastructure and likely owns cheap assets, but does not have a high probability of owning the more expensive assets surveyed. Profile 3 is a rising income household with mixed quality toilet and water and moderate probability of owning cheaper assets. Profile 4 corresponds to a poor household with no assets and the lowest quality infrastructure. The estimated GoM household-specific indices retain interpretability as the latent variable increases in dimension.

4.2 Country-Level Index by LDA and LCA

The GoM index indicates which countries have a higher average weight on the profile corresponding to higher asset ownership. The PCA and z-score indices indicate which countries have higher average asset ownership relative to other countries in the sample. But these descriptive indices may not always be easy to interpret. The LDA also gives country-level indices, but permits interpretation by imposing a probabilistic structure.

With a representative sample, or appropriate sample weights, household data can be aggregated into country-level statistics, which can be useful for countries where there is little GDP and investment data necessary to back out aggregate household consumption. The five household-level indices discussed in the previous section, for example, can be averaged by country to derive a country-level wealth index. Those are in Table 4, along with the estimated country-level indices g_c or z_c from LDA and LCA models estimated on the survey response data with $K = 2$.

Country	PCA	FPRIT	PCHOR	ZSCORE	GOM	LDA	LCA
Colombia	1.10	1.20	0.62	1.02	0.86	1.00	1.00
Guyana	0.13	-0.28	0.05	0.30	0.71	0.55	0.00
Haiti	-1.87	-2.42	-1.16	-0.81	0.35	0.01	0.00
Nicaragua	-2.33	-2.48	-1.30	-0.86	0.27	0.01	0.00
Peru	-0.08	-0.03	-0.03	0.35	0.66	0.27	0.00

Table 4: Country-Level Socioeconomic Status Indicators

Though all methods indicate that Haiti and Nicaragua are the poorest countries, and Colombia is the richest, the LDA index identifies the predominant sets of bundles owned by individuals in the sample, and classifies each country as a mixture over two different multinomial distributions over bundles. Profile 1, identified as the wealthy profile, is the one that has a lower probability on owning no assets.

LCA assigns each country to a single profile that is a probability distribution over bundles of assets. The latent class model isolates Colombia as the wealthy country but lumps together all the rest in the same class. This model is limiting, since even among this small sample of countries, there is more complex heterogeneity than LCA can capture with a finite number of classes. Allowing there to be a country-specific mixture, as in LDA, shows that Guyana and Peru to have a significant proportion of individuals who own more expensive bundles, and have a very different household wealth distribution compared to Haiti and Nicaragua.

LDA differs from GoM by relaxing GoM 1, in return for strengthening GoM 2. Strengthening GoM 2 reduces the complexity of heterogeneity modeled at the household level: individuals are assigned to only one of two profiles, whereas in GoM individual households are assigned to a mixture over profiles. Relaxing GoM 1 results in profiles β_k that can model more complex correlations in question responses compared to GoM. For example, for a specific survey response r^*:

$$P(r^*) = P(\text{water} = 3, \text{toilet} = 3, \text{floor} = 3, \text{elec} = 1,$$
$$\text{radio} = 1, \text{tv} = 0, \text{fridge} = 1, \text{motorbike} = 0, \text{car} = 0, \text{phone} = 1)$$

For GoM, the probability of this response, for individual i, conditional on the assignment of each individual's question response to a profiles, z_{ij} for $j = 1, \ldots, 10$, factorizes:

$$P(r^*|z_i) = P(\text{water} = 3|z_{i1})P(\text{toilet} = 3|z_{i2})P(\text{floor} = 3|z_{i3})P(\text{elec} = 1|z_{i4})P(\text{radio} = 1|z_{i5})$$
$$P(\text{tv} = 0|z_{i6})P(\text{fridge} = 1|z_{i7})P(\text{motorbike} = 0|z_{i8})P(\text{car} = 0|z_{i9})P(\text{phone} = 1|z_{i10})$$

where each probability is given by the profile specific multinomial distributions over question responses, $\beta_{z_{ij},j}$.

For LDA, on the other hand, the joint probability is given directly by β_{z_i,r^*}, where z_i is individual i's assignment to a profile. This means that for GoM, given an individual's mixture over profiles, the probability of an individual having water of a certain quality is independent of them having electricity. For LDA, even given an individual's assignment to a profile, the probability of an individual having water of a certain quality depends on whether the household has electricity or not. For economists characterizing wealth profiles in a country, with a specific intervention in mind, it may be useful to describe changes in conditional dependencies in asset ownership for different subgroups in a population.

GoM and LDA are hierarchical latent variable models that estimate similar wealth indices for households and countries as the existing methods. They are computationally expensive because of the probabilistic structure imposed. But it is because of the structure that clearer interpretation of the indices is possible even as the dimension of the latent variable increases.

5 Dynamic Models for Categorical Time Series

In the previous section, we described models for cross-sectional survey data, examples of which include development wealth surveys and management surveys. There is a large class of survey data, in macroeconomics and political economy, for example, where the same survey is run repeatedly at a constant interval, each time with a different sample of individuals. Measures derived from these surveys such as the confidence indices from the Michigan Survey of Consumers and political approval ratings from Gallup polls are frequently cited in the popular press. The models in the previous section are amenable to analysis of time series data; however, the addition of time dependence violates some of the conditional independence assumptions and must be appropriately modified. In this section, we continue to work with $N \times J$ survey response data X. However, we also assign

	MS-M	SS-M
Gaussian Version	Markov-Switching	Lin. State Space
Profile Parameter (β)	$K \times P$	$K \times P$
Assignment Parameter (z)	$T \times 1$	$N \times 1$
Mixture Parameter (g)	$K \times K$	$T \times K$
Mixture Hyperparameter (α)	$K \times K$	$T \times K$
Profile Hyperparameter (η)	$K \times P$	$K \times P$

Table 5: Dynamic Hierarchical Latent Variable Models

each individual i to the time period in which the response was observed, t. We denote X_t as the subset of X that includes only responses of individuals that were observed at time t.

We consider two ways of relaxing conditional independence: the first adds dynamics to the assignment parameters z in a dynamic version of the latent class model. This is the multinomial version of a Bayesian markov-switching model. The second method adds dynamics to the mixture parameters g_t in a dynamic version of LDA. This can be considered the multinomial version of a linear state space model.

There is a related statistics literature which has derived markov-switching models and state space models for discrete time series, mostly time series of counts, reviewed in Davis *et al.* (2016). The relationship between the models introduced here and markov switching and state space models for Gaussian data is in Table 5. Bradley *et al.* (2018) and Linderman *et al.* (2015) have related work developing dynamic versions of LDA with alternatives to a conditional lognormal distribution for the latent variables for improved sampling speed and convergence.

5.1 Multinomial Markov-Switching Model (MS-M)

For the first dynamic model introduced, as in the latent class model, each survey response is generated from a profile-specific multinomial distribution. The profile indicator for an individual, determined by the time at which they responded to the survey, follows a markov-switching process governed by an aggregate-level transition matrix g.

The model specification is as follows:

$$X_{ti}|\beta, z_t \sim \text{Multinomial}(\beta_{z_t})$$

$$z_t|z_{t-1}, g \sim \text{Multinomial}(g_{z_{t-1}})$$

$$g_k \sim \text{Dirichlet}(\alpha_k)$$

$$\beta_k \sim \text{Dirichlet}(\eta)$$

g is now a $K \times K$ where each column $g_k \in \Delta^{K-1}$ gives $Pr(z_t|z_{t-1} = k)$.

Assumption Latent Class 1 is replaced by MS-M 1. The other conditional independence assumptions are the same as in latent class models.

MS-M 1 ***: Markov Process of Assignments****: Conditional on the assignment of the responses in the previous period z_{t-1}, the profile assignment of responses in period t, z_t, is independent of those in other periods. In addition, assignments are independent of profiles given the transition matrix.*

$$Pr(z|g,\beta) = \prod_{t=1}^{T} Pr(z_t|z_{t-1}, g)$$

MS-M 2 ***: Conditional Independence of Individuals Given Assignments****: Conditional on an individual's assignment based on the time period in which the response was observed, individual responses are independent. Furthermore, responses are independent of the transition matrix given the profile assignments.*

$$Pr(X|g,z,\beta) = \prod_{i=1}^{N} Pr(X_{ti}, |z_t, \beta)$$

MS-M 3 ***: Independence of Profiles*** *β_k is independent of β_f for $f \neq k$.*

The likelihood of the markov-switching model for categorical data is as follows:

$$Pr(X,z,\beta,g) = \prod_{k=1}^{K} Pr(g_k) \prod_{k=1}^{K} Pr(\beta_k) \prod_{t=1}^{T} p(z_t|z_{t-1}, g) \prod_{i=1}^{N_t} \beta_{z_t, X_{ti}}$$

$Pr(\beta_k)$ and $Pr(g_k)$ are Dirichlet densities and $Pr(z_t|z_{t-1})$ is directly from the transition probability matrix. The posterior distribution of the model parameters are estimated via Gibbs Sampling using the below steps, which are described in full detail in Appendix A.

1. Generate z_t conditional on g, X, z_{t+1}, z_{t-1}, and β using a posterior multinomial distribution and Albert and Chib (1993)'s single-move sampling procedure.

2. Generate transition matrix g conditional on z using a posterior Dirichlet distribution.

3. Generate β conditional on X, z with a posterior Dirichlet distribution.

We call this model the multinomial markov switching model (MS-M), since it is equivalent to a Bayesian markov-switching model for Gaussian models with state-specific means replaced by state-specific multinomial distributions. The time dependence between responses at time t and responses at time $t-1$ is captured through the markov-switching process of z_t. It is computationally simple to estimate and has a convincing interpretation when analyzing discrete data where it is natural to assume a discrete-valued latent variable can influence responses patterns over time (for example, recessions can generate switching patterns in responses to consumer confidence surveys).

However, as in the latent class model, it is limiting in the form of heterogeneity that it can capture in responses for individuals across time. All responses at time t are assumed to come from the same profile indexed by z_t. This is unduly restrictive in many settings where it is more natural to think about evolution in response patterns as coming from proportions of respondents of different types changing, rather than nonlinear switching in the state of the entire survey sample. The next model addresses this by adding time dynamics to a time-specific mixture parameter g_t instead, which captures time dependence in responses while allowing a more flexible form of heterogeneity in responses in each time period.

5.2 Multinomial State Space Model (SS-M)

The second model introduced is a dynamic version of LDA. Like LDA, responses follow a profile-specific multinomial distribution and an individual's profile is generated from a group-specific mixture, where groups are indexed by time t rather than group c. The group, or time-specific mixture now follows a lognormal distribution. In static LDA $g_t \in \Delta^{K-1}$. In the dynamic version $g_t \in \mathbb{R}^K$, follows a random walk process and is converted to a vector of proportions $\psi(g_t) \in \Delta^{K-1}$ using the softmax function.

$$\psi(g_t) = \frac{\exp(g_t)}{\sum_{k=1}^{K} \exp(g_{tk})}$$

The model specification is:

$$X_{ti} | z_{ti} \sim \text{Multinomial}(\beta_{z_{ti}})$$

$$z_{ti} \sim \text{Multinomial}(\psi(g_t))$$

$$g_t = g_{t-1} + w_t, \qquad w_t \sim N(0, \sigma I)$$

$$\sigma \sim \text{InverseGamma}(v0, s0)$$

$$\beta_k \sim \text{Dirichlet}(\eta)$$

In the MS-M model, we added dynamics to the discrete-valued assignments z, and the dynamics were captured using a discrete markov process. In this model, the dynamics are added to the group-specific mixtures, and are represented by a linear markov process. The model is denoted the multinomial state space model, since it involves adding a multinomial outcome to the dynamics of a Bayesian linear state space model via a link function.

When individuals are grouped by time t, rather than by potentially independent group membership c, LDA 2, the independence of group mixtures no longer holds, and is adjusted. We maintain the assumption of conditional independence of individuals given the group structure.

SS-M 1 ***: Conditional Independence of Individuals Given Group Structure*** *Conditional on time t's mixture over profiles g_t, the probability of a certain response and assignment for individual i is independent of other individual's responses. In addition, responses are independent of time-specific mixtures given the profile assignments and assignments are independent of profiles* *given the time-specific mixtures.*

$$Pr(X, z|g, \beta) = \prod_{i=1}^{N} Pr(X_{ti}, z_{ti}|g_t, \beta)$$

$$Pr(X_{ti}|g_t, z_{ti}, \beta) = Pr(X_{ti}|z_{ti}, \beta)$$

$$Pr(z_{ti}|g_t, \beta) = Pr(z_{ti}|g_t)$$

SS-M 2 ***Markov Property of g_t and Independence of Profiles*** *β_k is independent of β_f for $f \neq k$. Given g_{t-1}, g_t is independent of g_s conditional on g_{s-1} for $s \neq t$.*

With these assumptions, the joint probability distribution of the SS-M model factorizes as follows:

$$p(\beta, g, Z, X) = \prod_{k=1}^{K} p(\beta_k) \prod_{t=1}^{T} p(g_t|g_{t-1}) \prod_{i=1}^{N_t} g_{t,z_{ti}} \beta_{z_{ti},X_{ti}}$$

The gibbs-sampling steps necessary to estimate the model are as follows, and are explained in detail in Appendix A:

1. Generate g_t conditional on σ, g_{t-1}, g_{t+1}, and z_t using Stochastic Gradient Langevin Dynamics
2. Generate z conditional on β, X, and g using a multinomial posterior
3. Generate β conditional on z using a Dirichlet posterior
4. Generate σ conditional on g using an Inverse-Gamma posterior

Step 1 is computationally intensive and can be difficult to tune. Dynamic LDA, from Blei and Lafferty (2006), is a similar model that allows evolution over time both in the profile mixtures g_t as well as the actual profiles β. In this paper, however, we are interested in identifying latent profiles in survey respondents that are constant over time; we assume that differences in patterns of survey responses over time is due to changes in prevalence of each latent profile, g_t, rather than changes in the latent profiles themselves.

In this section, we described a formulation where we assume that at each time period there is a mixture of dominant profiles. We will show that in many applications this assumption is effective in obtaining interpretable and economically meaningful latent states. For example, in the Michigan Survey of Consumers setting, it is natural to assume that there are always both optimistic and pessimistic people in the survey sample each month. Depending on general economic conditions and media reports, there are different proportions of optimistic people in each month. Being optimistic involves a different multinomial distribution over permutations of survey responses compared to being pessimistic.

6 Application of Dynamic Models

First, we estimate SS-M and MS-M on responses to the Michigan Survey of Consumers. We show that the estimated index from the SS-M model corresponds closely to the existing ICS, and that certain recessions are better characterized by switching in consumer sentiment than switching in real expenditure data. Second, we estimate SS-M on GPSS data on health, environment, and values and morals. During the 2008-2009 recession, which was characterized by a negative switch in consumer sentiment, concern towards the environment drops steeply, like the Michigan ICS and an index created from the GSPSS health survey. Sentiment toward traditional values, on the other hand, seems unaffected by financial insecurity.

This is related to existing work linking survey data to macroeconomic conditions. Scruggs and Benegal (2012) links poor economic conditions to the decline in public concern about the environment in the U.S. and EU during the Great Recession. Nicholson and Simon (2010) uses Gallup survey data on wellbeing to examine the relationship between the aggregate unemployment rate and health outcomes. Dominitz and Manski (2003) have examined the accuracy of Michigan survey data and concluded that asking questions about subjective probabilities is more accurate than the qualitative questions used as data in this paper. We take the stance that there is good quantitative information available in the patterns of aggregate qualitative responses, but show an improved method of extracting that information.

6.1 Michigan Consumer Sentiment Indices

The raw Michigan Survey of Consumers data contains the survey responses for 500 telephoned respondents in continental U.S. each month of the year. For each month's sample, an independent cross-section sample of households is drawn, and some are reinterviewed six months later. The total sample for any one survey is normally made up of about 60% new respondents, and 40% being interviewed for the second time. We don't model the repeated interviewing that occurs in the Michigan data.

The data for every month from January 1978 to November 2017 is public. The Index of Consumer Sentiment (ICS) is made up of five questions of the survey on the respondent's opinion about current and future economic conditions. The questions along with possible responses are in Appendix B. The ICS is constructed from the five questions listed in Appendix B based on the following ad-hoc procedure, published by the University of Michigan on their website:

1. Compute the relative scores X_j for each of the five questions. The relative scores are the percentage of respondents giving favorable replies minus the percent giving unfavorable replies, plus 100.

2. Round each relative score to the nearest whole number

3. Using the following formula, sum the relative scores, divide by the 1966 base period value,

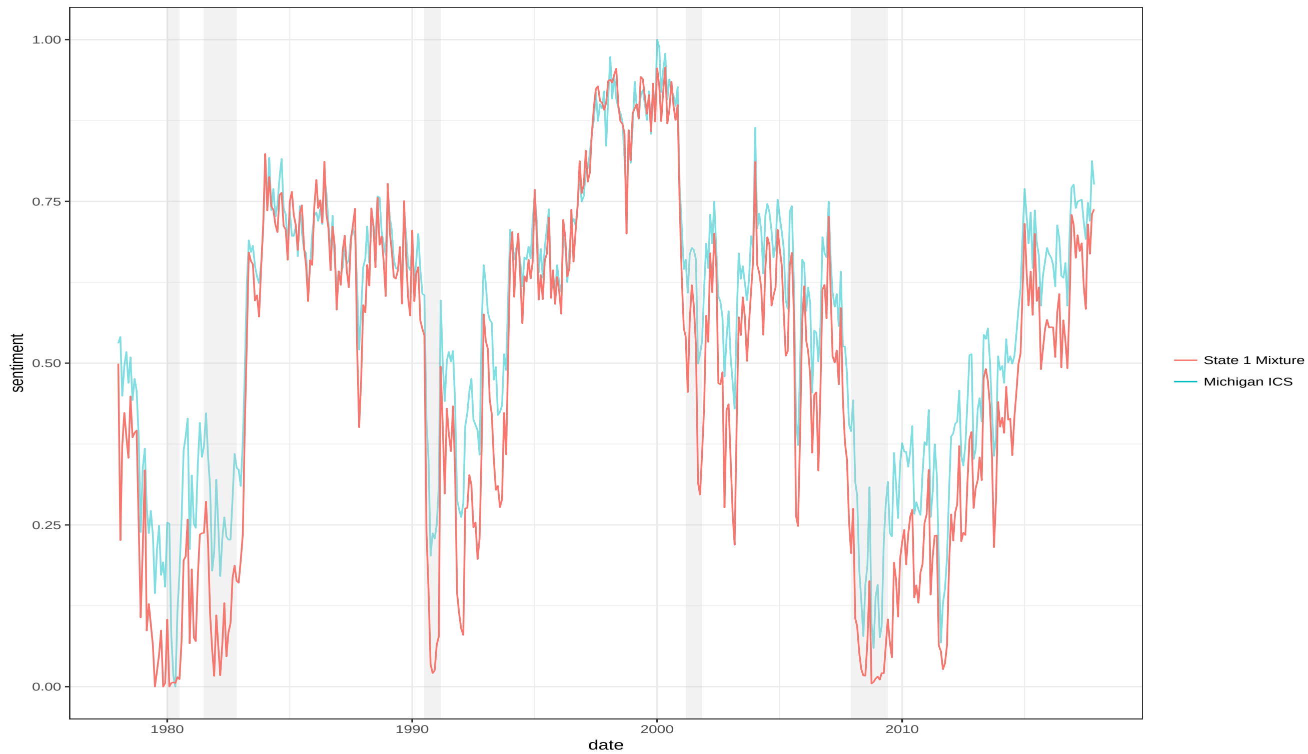

Figure 4: Probability of Profile 1 vs. ICS

and add 2.0 to correct for sample design changes in the 1950s.

$$ICS = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{6.7558} + 2.0$$

Under this procedure, responses that are missing or incomplete are dropped. With our procedure, it is not necessary to have adjustments for missing or incomplete data, as refusing to answer a question can be modeled as part of the permutations of potential survey responses. We also do not need to make any adjustment for outliers. Furthermore, for ordered variables, we make no assumption on if the distance between responses categorized as 1 and 2 is the same as the distance between 4 and 5.

Each month t has N_t survey respondents. X_{ti} is the response for the i-th survey respondent in month t and corresponds to a survey response permutation indexed by $p \in \{1, \ldots, P\}$. $X_{ti} = 11111$ corresponds to a survey respondent who answers "better" or "good" to all of the questions on the U.S. economy. We estimate SS-M on the Michigan Consumer survey data and compare the estimated weight g_{t1} on the more optimistic profile to the ICS, rescaled to match the range and variance of the new index, which is between 0 and 1. We identify the "positive" sentiment using the prior distribution on β, by decreasing the prior probability of responding 55555 for Profile 1.

Though the ad-hoc method used to create the ICS and the sampling procedure to estimate our index are very different, the trends in the two indices are very close, as seen in Figure 4. The index

Profile 1		**Profile 2**	
11111	11.1%	53551	3.9%
13111	8.4%	53555	3.5%
33111	5.1%	55555	3.0%
31111	3.1%	55551	2.8%
51111	2.4%	13551	2.8%

Table 6: SS-M: Top 5 Profile-Specific Multinomial Probabilities

Profile 1		**Profile 2**	
11111	6.8%	11111	3.6%
13111	5.1%	53551	3.3%
33111	3.1%	13111	3.0%
13551	2.0%	53555	2.8%
31111	2.0%	55555	2.5%

Table 7: MS-M: Top 5 Profile-Specific Multinomial Probabilities

based on SS-M tends to drop more during recessions. The Michigan data has ordered responses so it is relatively straightforward to come up with an ad-hoc method to combine the responses in each month using the percentage of favorable versus unfavorable responses. In this setting, both indexes capture similar low-dimensional representations of the five survey questions. SS-M, however, specifies a probability model that generates the index, which allows additional interpretation of the index. Table 6 contains the responses with the top 5 weights for each multinomial distribution parameter β_k corresponding to the two estimated states.

When g_t has a higher weight on Profile 1, then there is a higher probability of survey respondents in a month selecting all optimistic answers to the survey (11111), or a mix between indifferent and optimistic (for example, 33111). When there is a higher weight on Profile 2, however, there is a higher probability of many respondents selecting all pessimistic answers to the survey (55555).

We also estimate MS-M with $K = 2$ on the Michigan Consumer Survey data with two profiles. The latent variable z_t is discrete and has dynamics corresponding to a transition matrix. Table 7 gives the top 5 survey permutations and their probabilities for each of the estimated states in the model.

SS-M model has a more complex and computationally intensive estimation procedure than MS-M. Examining the difference between Table 7 and Table 6 indicates one advantage of allowing there to be a mixture of profiles in each period rather than a single state in each period. In both tables, Profile 1 corresponds to more positive sentiment towards current and future economy, while Profile 2 corresponds to a more negative sentiment. However, in all periods, there are many respondents responding "11111" to the survey. Profile 2, in the markov-switching model, while having more weight on pessimistic response permutations, also has weight on the common permutations that appear in every month, like 13111. The mixture of profiles in the SS-M allows a cleaner separation between the optimistic group of respondents and pessimistic group of respondents. In SS-M, there

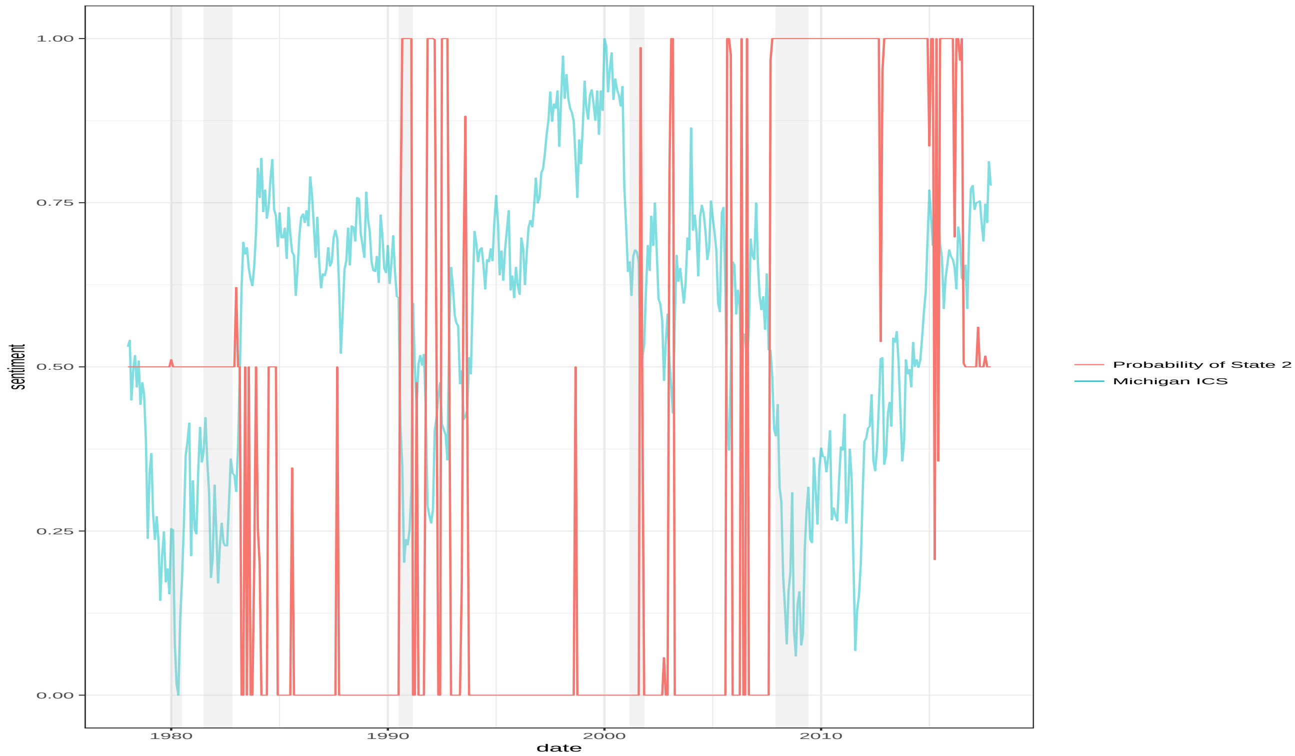

Figure 5: Probability of Profile 2

always non-zero weight on Profile 1, so Profile 2 corresponds more directly to the group of respondents that is concerned about the economy, which fluctuates with recessions and general economic conditions.

In the discrete markov switching model, the profiles, or states, are highly persistent: the estimate of the transition probabilities between states is $P(z_{t+1} = 1|z_t = 1) = 0.89$ and $P(z_{t+1} = 2|z_t = 2) = 0.80$. Figure 5 shows the probability of being in Profile 2, the more pessimistic state. The Michigan ICS is also plotted and recessions shaded.

The probability of Profile 2 remains high for long after the 2008-2009 recession, but for other recessions drops quickly back to 0. This shows how consumer confidence remained low for a long time after the recession was declared over, which is unique to the 2008-2009 recession compared to other recessions in the sample period. This characteristic of the 2008-2009 recession is not as apparent when examining trends in Personal Consumption Expenditure.

We estimate a standard markov mean-variance switching model on the log-difference of Personal Consumption Expenditure in the United States. The below figure shows the probability of being in the negative state for the model derived from survey data as well as the probability of the low mean state in the Gaussian markov-switching model derived from PCE data. As mentioned previously, the markov switching model on the survey data remains in the pessimistic state post-2008 recession until recently, but the model on the expenditure data classifies a return to growth state. The model estimated on PCE data tends to lag recessions and misses the 1990s recession. The model estimated

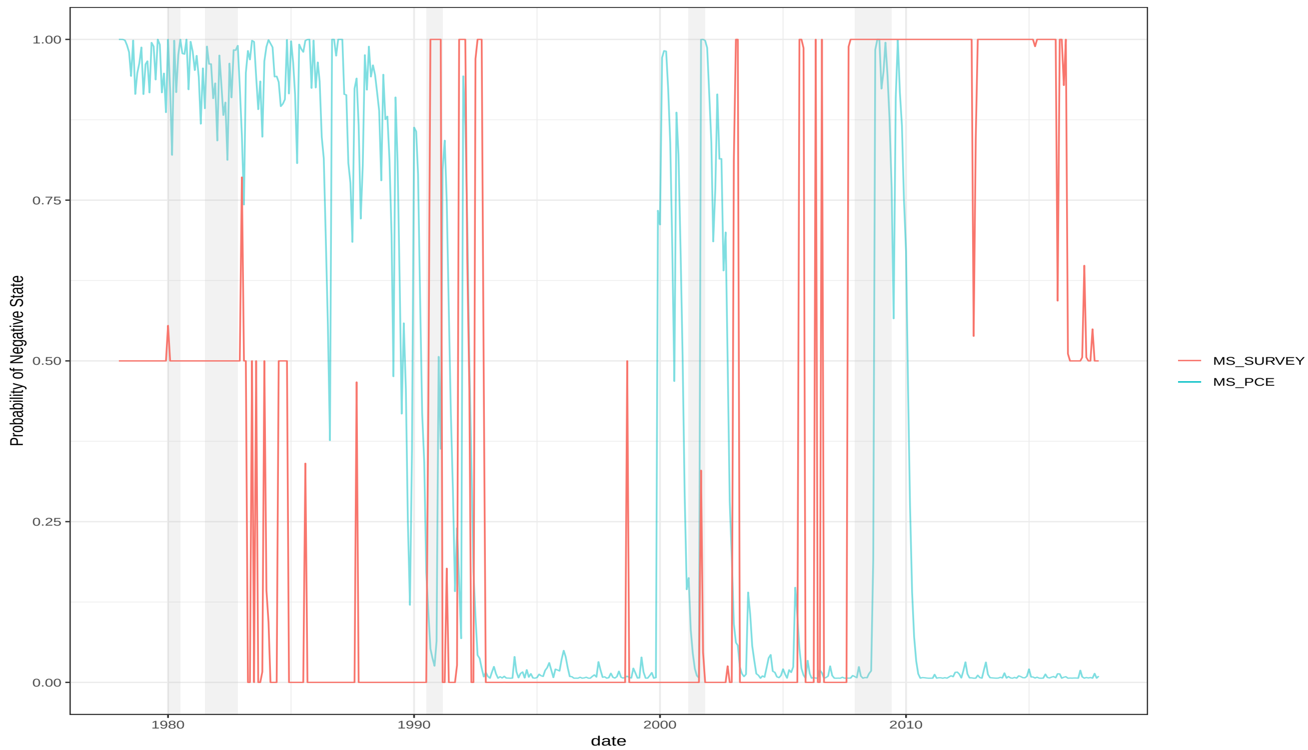

Figure 6: Probabilities of Negative State in MS-M vs. MS on PCE

on the discrete sentiment data leads the 2008 recession and captures the 1990s recession well, but misses the 2001 recession. Both models have some difficulty classifying the beginning of the period.

Most authors have found that the consumer sentiment data does not have predictive value for real economic data like consumption expenditure beyond other macroeconomic variables Ludvigson (2004). However, we have shown using a markov-switching model estimated directly on the survey data responses that, despite not having general independent predictive power, the survey data still contains distinct economic insights in the nature of certain recessions since 1980 compared to real economic data on consumer purchases.

In the Michigan ordinal survey data setting, where there is an existing ad-hoc sentiment index already constructed, we show that the latent variable g_t in the discrete space model can be interpreted as a sentiment index with a very similar trend to the existing ICS. Furthermore, both this continuous-valued index and the discrete-valued index from the markov-switching model have additional interpretation, since each corresponds to a profile in the model that involves a multinomial distribution over survey response permutations. This motivates extending these methods to quantify survey data that correspond to consumer characteristics like environmental concern and values where, unlike for consumer purchasing, there is little quantitative data to compare to.

6.2 Indexes from Gallup Data on Environment, Health, and Values

The raw Gallup Poll Social Series data contains the survey responses of a minimum of 1000 people in the U.S. each month. The core questions are devoted to a different topic each month, and are available from 2001 to 2016 through a subscription. The analysis in this section uses data from the survey on the environment conducted each March, the survey on health conducted each November, and the survey on values and morals done each May. The survey questions and responses used to create the environmental concerns index, health concerns index, and traditional values index are in Appendix B. The Gallup poll data is mapped into the model framework in the same way as for the Michigan data.

The GPSS data contains a variety of qualitative questions on important aspects of individual's economic condition that may not be well-measured in aggregate economic statistics. For example, consumer's concern towards the environment and their values are aspects of their preferences that affect their purchasing and other life decisions. Unlike consumption expenditure, there are not continuous-valued real economic statistics corresponding to environmental concern. The wealth of discrete data in survey series like Gallup contains important information about trends in these sentiments. However, the questions are not uniform. Some are ordered and others unordered, and there are differing scales and number of responses for each question. We show that the model presented in this paper can address these challenges and extract useful quantitative representations of health, environmental concern, and values during the last recession.

First, we estimate SS-M with two profiles, using the six questions in Appendix B on the respondents' views on current and future environmental conditions in the U.S. The weight on the first profile is plotted, which has higher weight on response permutations that correspond to the respondent thinking that the environment is poor, that global warming is already happening, that the threat of global warming is under exaggerated, and that environment should be prioritized over energy and economic security. The environmental concerns sentiment drops significantly during the recession when respondents were likely be distracted based on financial concerns, and was very slow to recover even to pre-2005 levels. Rather than examining trends in responses on a question by question basis, treating the responses as outcomes of a probability model allows estimation of a single factor corresponding to environmental concern.

We derive indices from sets of questions from a few other GPSS categories to compare. In Figure 7, we plot the weight on the profile that has higher probability on responses to the moral and values survey that correspond to traditional values (i.e. abortion/dealth penalty/suicide is immoral and religion is important). The weight on the profile corresponding to traditional values remains relatively stable around 80-85%, with a small increase in 20090.

Estimated sentiment based on satisfaction with personal health and the U.S. health care system, though it does not fluctuate as much as the sentiment towards the environment, does bottom out in 2009. During times when it is more difficult to afford health care, dissatisfaction with the health care system is higher. It is interesting that sentiment toward the environment follows trends in financial and health insecurity, and not the path of more fixed sentiments on values.

Figure 7: Indexes Derived from GPSS Data

We find that SS-M can be used on the 15 disparate questions listed in Appendix B to successfully extract three low-dimensional continuous-valued representations on environmental concerns, health concerns, and traditional values. As in the literature, we find that recessions correspond to decreases in environmental concern and health satisfaction.

7 Conclusion

We derive Bayesian approaches to extracting interpretable discrete and continuous-valued indexes from high-dimensional multinomial time series. The models provide a parametric form for extracting low-dimensional summaries of high-dimensional discrete data, in contrast to existing methods for summarizing categorical data which tend to use ad-hoc averaging methods or PCA. The models successfully extract interpretable indexes from ordered and unordered, categorical and binary, and static and time series survey data on economic well-being, values, health, and environmental sentiment.

The relation of these models to the extensive computer science literature on Bayesian hierarchical latent variable models suggests a variety of extensions for the models. For the survey data application, it is possible to introduce more complex forms of dynamics, or to directly model the relationship between responses and some outcome variable as in Mcauliffe and Blei (2008), or to integrate some of the latent variable estimates into causal analysis as suggested in Wang and Blei

(2018). It would also be straightforward to extend the models presented for survey data that had a mixture of continuous and discrete responses.

Along with developing survey data specific examples, this paper also describes the close relationship between of a variety of seemingly disparate models, developed originally for a variety of different types of data, showing that they are all similar hierarchical latent variable models with different conditional independence assumptions. Understanding these models more fully is important, as they have applications to estimating structural models on a variety of high-dimensional discrete economic data, including text data, network data, and consumer choice.

References

Albert, J. H. and Chib, S. (1993) Bayes Inference via Gibbs Sampling of Autoregressive Time Series Subject to Markov Mean and Variance Shifts, *Journal of Business & Economic Statistics*, **11**, 1–15.

Athey, S., Blei, D., Donnelly, R., Ruiz, F. and Schmidt, T. (2018) Estimating heterogeneous consumer preferences for restaurants and travel time using mobile location data, *AEA Papers and Proceedings*, **108**, 64–67.

Baker, S., Bloom, N. and Davis, S. J. (2016) Measuring economic policy uncertainty, *Quarterly Journal of Economics*, **131:4**, 1593–1636.

Bandiera, O., Hansen, S., Prat, A. and Sadun, R. (2017) Ceo behavior and firm performance, *NBER Working Paper Series*.

Bhadury, A., Chen, J., Zhu, J. and Liu, S. (2016) Scaling up Dynamic Topic Models.

Blei, D., Ng, A. Y. and Jordan, M. I. (2003) Latent Dirichlet Allocation, *Journal of Machine Learning Research*, **3**, 993–1022.

Blei, D. M. and Lafferty, J. D. (2006) Dynamic Topic Models, in *Proceedings of the 23rd international conference on Machine learning*, pp. 113–120.

Bloom, N. and Reenen, J. V. (2010) Why Do Management Practices Differ Across Firms and Countries?, *Journal of Economic Perspectives*, **24**, 203–224.

Bloom, N. and 'van Reenen', J. (2007) Measuring and explaining management practices across firms and countries, *Quarterly Journal of Economics*, **122:4**, 1341–1408.

Bradley, J. R., Holan, S. H. and Wikle, C. K. (2018) Computationally Efficient Multivariate Spatio-Temporal Models for High-Dimensional Count-Valued Data, *Bayesian Analysis*, **13**, 253–310.

Davis, R. A., Holan, S. H., Lund, R. and Ravishanker, N. (2016) *Handbook of discrete-valued time series*, CRC Press.

Dominitz, J. and Manski, C. F. (2003) How should we measure consumer confidence (sentiment)? evidence from the michigan survey of consumers, Working Paper 9926, National Bureau of Economic Research.

Erosheva, E. A., Fienberg, S. E. and Joutard, C. (2007) Describing Disability Through Individual-Level Mixture Models for Multivariate Binary Data, *The Annals of Applied Statistics*, **1**, 502–537.

Filmer, D. and Pritchett, L. H. (2001) Estimating Wealth Effects Without Expenditure Data - or Tears, *Demography*, **38**, 115–132.

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013) *Bayesian data analysis*, Chapman and Hall/CRC.

Griffiths, T. L. and Steyvers, M. (2004) Finding scientific topics, *Proceedings of the National Academy of Sciences*, **101**, 5228–5235.

Hamilton, J. D. (1989) A New Approach to the Economic Analysis of Nonstationary Time Series and the Business, *Econometrica*, **57**, 357–384.

Hansen, S., McMahon, M. and Prat, A. (2018) Transparency and Deliberation within the FOMC: A Computational Linguistics Approach, *Quarterly Journal of Economics*, pp. 801–870.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B. and Ermon, S. (2016) Combining satellite imagery and machine learning to predict poverty, *Science*, **353**, 790–794.

Kolenikov, S. and Angeles, G. (2009) Socioeconomic status measurement with discrete proxy variables: Is principal component analysis a reliable answer?, *Review of Income and Wealth*, **55**, 128–165.

Lee, S.-Y., Bentler, P. and Poon, W.-Y. (1990a) Full Maximum Likelihood Analysis of Structural Equation Models with Polytomous Variables, *Statistics and Probability Letters*, **9**, 91–97.

Lee, S.-y., Poon, W.-y. and Bentler, P. M. (1990b) A Three-Stage Estimation Procedure for Structural Equation Models with Polytomous Variables, *Psychometrika*, **55**, 45–51.

Linderman, B. S. W., Johnson, M. J. and Adams, R. P. (2015) Dependent Multinomial Models Made Easy: Stick Breaking with the Polya-Gamma Augmentation.

Ludvigson, S. C. (2004) Consumer Confidence and Consumer Spending, *Journal of Economic Perspectives*, **18**, 29–50.

MacDonald, I. L. and Zucchini, W. (1997) *Hidden Markov and other models for discrete-valued time series*, vol. 110, CRC Press.

Mcauliffe, J. D. and Blei, D. M. (2008) Supervised topic models, in *Advances in neural information processing systems*, pp. 121–128.

Ng, S. (2015) Constructing Common Factors from Continuous and Categorical Data, *Econometric Reviews*, **34**, 1141–1171.

Nicholson, S. and Simon, K. (2010) How did the recession affect health and related activities of americans?, *Preliminary and Incomplete Draft*.

Pesaran, M. H. (2006) Estimation and inference in large heterogeneous panels with a multifactor error structure, *Econometrica*, **74**, 967–1012.

Ruiz, F. J. R., Athey, S. and Blei, D. M. (2018) SHOPPER: A Probabilistic Model of Consumer Choice with Substitutes and Complements.

Scruggs, L. and Benegal, S. (2012) Declining public concern about climate change: Can we blame the great recession?, *Global Environmental Change*, **22**, 505–515.

Vyas, S. and Kumaranayake, L. (2006) Constructing socio-economic status indices: How to use principal components, *Health and Policy Planning*, **21:6**, 459–468.

Wang, Y. and Blei, D. M. (2018) The blessings of multiple causes, *arXiv preprint arXiv:1805.06826*.

Welling, M. and Teh, Y. W. (2011) Bayesian Learning via Stochastic Gradient Langevin Dynamics, in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pp. 681–688.

Young, A. (2012) The African Growth Miracle, *Journal of Political Economy*, **120**, 696–739.

A Gibbs Sampling Steps

In this section, $p_{mult}(\cdot)$ is used to signify the multinomial density function, $p_{dir}(\cdot)$ the Dirichlet density function, and $p_{norm}(\cdot)$ the Normal density function.

A.1 Steps for Markov-Switching Model

Generating z_t conditional on g, X_t, z_{t+1}, z_{t-1}, and β

We follow the single-move gibbs sampling procedure of Albert and Chib (1993). We assume that $P(z_1|z_0) = \frac{1}{K}$ and $P(z_{T+1}|z_T) = \frac{1}{K}$.

For $t = 1, \ldots, T$, sample z_t from $\{1, \ldots, K\}$ from the posterior distribution, which is multinomial with

$$p(z_t = j|X_t, z_{t-1}, z_{t+1}, \beta, X_t) \propto p(z_t|z_{t-1}) p_{mult}(X_t; \beta_j) p(z_{t+1}|z_t)$$

The first and last terms are taken directly from the current estimate for the transition matrix: $p(z_t = j|z_{t-1} = k) = g_{jk}$.

Generating transition matrix g conditional on z

$$p(g_j|z) \propto p_{dir}(g_j; \alpha) p_{mult}(n_j; g_j)$$

where $n_{i,j} = \sum_{t=2}^{T} \mathbb{1}(z_t = i)\mathbb{1}(z_{t-1} = j)$

The posterior distribution for each column g_j in the transition matrix g is independent Dirichlet:

$$g_j \sim \text{Dir}(\alpha_1 + n_{1j}, \ldots, \alpha_K + n_{Kj})$$

Generating β conditional on X, z

$$p(\beta_k|G, X; \eta) \propto p_{dir}(\beta_k; \eta) p_{mult}(m_k; \beta_k)$$

$$m_{k,v} = \sum_{t=1}^{T} \sum_{i=1}^{N_t} \mathbb{1}(X_{ti} = p)\mathbb{1}(z_t = k)$$

The posterior distribution of the multinomial probabilities is independent Dirichlet for each profile k:

$$\beta_k \sim \text{Dir}(\eta_1 + m_{k1}, \ldots, \eta_P + m_{kP})$$

A.2 Steps for State Space Model

Generating g_t conditional on σ,g_{t-1}, g_{t+1}, and z_t

We adapt the method from Bhadury *et al.* (2016) for Dynamic LDA, and use Stochastic Gradient Langevin Dynamics Welling and Teh (2011) to draw g_t. SGLD is a form of gradient descent, adding Gaussian noise at each step, which Welling and Teh (2011) shows allows the method to generate samples from the true posterior without a Metropolis-Hastings test, as long as the shrinkage parameter ϵ_n fulfils certain conditions.

$$p(g_t|g_{t-1}, g_{t+1}, z_t) \propto p_{norm}(g_t; g_{t-1}, \sigma^2 I) p_{norm}(g_{t+1}; g_t, \sigma^2 I) \prod_{i=1}^{N_t} p_{mult}(z_{ti}; \theta(g_t))$$

In step s of the gibbs sampler, for each $k = 1, \ldots, K$,

$$\Delta g_{t,k}^{(s)} = \frac{\epsilon_s}{2} \nabla_{g_{t,k}} \log p(g_t^{(s-1)} | g_{t-1}^{(s)}, g_{t+1}^{(s-1)}, z_t^{(s-1)}) + \psi_i, \qquad \psi_i \sim N(0, \epsilon_s)$$

$$\nabla_{g_{tk}} p(g_t^{(s)} | g_{t-1}^{(s-1)}, g_{t+1}^{(s-1)}, z_t^{(s-1)}) = \frac{-1}{\sigma^2}(g_{k,t} - g_{k,t-1}) - \frac{1}{\sigma^2}(g_{k,t+1} - g_{kt}) + n_{tk} - N_t \psi(\theta_t)_k$$

$$n_{tk} = \sum_{i=1}^{N_t} \mathbb{1}(z_{ti} = k)$$

$\epsilon_s = a(b + s)^{-c}$ for gibbs sampling step s. We choose $a = 0.1$, $b = 1$ and $c = 0.5$ for our applications. One downside of this method is for each application having to tune those parameters to get a sequence of ϵ_s that allows for proper convergence.

Generating z_{ti} conditional on β,X, and g_t

The posterior distribution of z_{ti} is multinomial with probabilities:

$$p(z_{ti} = k | \beta, g_t) \propto \beta_{k,p} \theta(g_t)_k$$

for $X_{ti} = p$.

Generating β conditional on z

$$p(\beta|z) \propto p_{dir}(\beta; \eta) p_{mult}(z; \beta)$$

The posterior distribution of the multinomial probabilities is Dirichlet for each profile k:

$$\beta_k | z \sim \mathrm{Dir}(\eta_1 + m_{k,1}, \ldots, \eta_P + \mathrm{m}_{kP})$$
$$m_{kp} = \sum_{t=1}^{T} \sum_{i=1}^{N_t} \mathbb{1}(X_{ti} = p) \mathbb{1}(z_{ti} = k)$$

Generating σ conditional on g

$$\sigma \sim \mathrm{IGamma}(v1, s1)$$

$$v1 = v0 + T$$
$$s1 = s0 + \sum_{t=1}^{T} (g_t - g_{t-1})^2$$

B Survey Index Component Variables

B.1 Michigan Data

Below are the five questions used to create the Michigan indices.

1. Would you say that you are better off or worse off financially than you were a year ago?
    - (1) Better, (3) Same, (5) Worse, (8) Don't know or missing
2. Now looking ahead–do you think that a year from now you will be better off financially, or worse off, or just about the same as now?
    - (1) Better, (3) Same, (5) Worse, (8) Don't know or missing
3. Now turning to business conditions in the country as a whole–do you think that during the next twelve months we'll have good times financially, or bad times, or what?
    - (1) Good times, (2) Good with qualifications, (3) Pro-con, (4) Bad with qualifications, (5) Bad times, (8) Don't know or missing
4. Looking ahead, which would you say is more likely –that in the country as a whole we'll have continuous good times during the next five years or so, or that we will have period of widespread unemployment or depression, or what?
    - (1) Good times, (2) Good with qualifications, (3) Pro-con, (4) Bad with qualifications, (5) Bad times, (8) Don't know or missing
5. Generally speaking, do you think now is a good or bad time for people to buy major household items?
    - (1) Good, (3) Pro-con, (5) Bad, (8) Don't know or missing

B.2 Gallup Poll Social Survey

Environmental Concerns Index

Below are the six questions used to create the environmental concerns index.

1. How would you rate the overall quality of the environment in this country today ?
    - (1) Excellent, (2) Good, (3) Only fair, (4) Poor, (5) Don't know or missing
2. Right now, do you think the quality of the environment in the country as a whole is :
    - (1) Getting better, (2) Getting worse, (3) Same, (4) Don't know or missing
3. With which one of these statements about the environment and the economy do you most agree?
    - (1) Protect environment, even at risk of curbing economic growth, (2) Economic growth priority even if environment suffers to some extent, (3) Equal priority, (4) Don't know or missing

4. With which one of these statements about the environment and energy production do you most agree ?

    - (1) Protect environment, even at risk of limiting energy supplies which the U.S. produces, (2) Development of U.S. energy supplies – such as oil, gas and coal – should be given priority, even if the environment suffers to some extent, (3) Equal priority, (4) Other, don't know or missing

5. Which of the following statements reflects your view of when the effects of global warming will begin to happen?

    - (1) Already begun to happen, (2) Will start happening within a few years, (3) Will start happening within your lifetime, (4) Will not happen within your lifetime, but they will affect future generations, (5) Will never happen, (6) Don't know or missing

6. Thinking about what is said in the news, in your view is the seriousness of global warming:

    - (1) Generally exaggerated, (2) Generally correct, (3) Generally underestimated (4) Missing

Traditional Values Index

Below are the four questions used to create the traditional values index.

1. Regardless of whether or not you think it should be legal, please tell me whether you personally believe that in general abortion is:

    - (1) Morally acceptable, (2), Morally wrong, (3) Depends on the situation, (4) Not a moral issue, (5) Don't know or missing

2. Regardless of whether or not you think it should be legal, please tell me whether you personally believe that in general the death penalty is:

    - (1) Morally acceptable, (2), Morally wrong, (3) Depends on the situation, (4) Not a moral issue, (5) Don't know or missing

3. Regardless of whether or not you think it should be legal, please tell me whether you personally believe that in general suicide is:

    - (1) Morally acceptable, (2), Morally wrong, (3) Depends on the situation, (4) Not a moral issue, (5) Don't know or missing

4. How important would you say religion is in your own life?

    - (1) Very important, (2) Fairly important, (3) Not very important, (4) Don't know or missing

Health Concerns Index

Below are the five questions used to create the health concerns index.

1. How would you describe your own physical health at this time?
    - (1) Excellent, (2) Good, (3) Only fair, (4) Poor, (5) Don't know or missing
2. How would you describe your own mental health or emotional well-being at this time?
    - (1) Excellent, (2) Good, (3) Only fair, (4) Poor, (5) Don't know or missing
3. Are you generally satisfied or dissatisfied with the total cost of health care in this country?
    - (1) Satisfied, (2) Dissatisfied, (3) Don't know or missing
4. Overall, how would you rate the quality of health care in this country ?
    - (1) Excellent, (2) Good, (3) Only fair, (4) Poor, (5) Don't know or missing
5. Overall, how would you rate health care coverage in this country ?
    - (1) Excellent, (2) Good, (3) Only fair, (4) Poor, (5) Don't know or missing