Markov Models for Multinomial Time Series

Evan Munro and Serena Ng

December 22, 2018

Abstract

Existing methods for extracting summary indexes from multivariate discrete data are generally ad-hoc, often treating discrete variables as if they were continuous. We introduce two Bayesian models for discrete time series data, which are the discrete versions of a basic Markov switching model and state space model. We then estimate the models on sets of variables from the Michigan Survey of Consumers and the Gallup Poll Social Series. We show that the index constructed using the Michigan data is highly correlated to the established index constructed with an ad-hoc method, but has additional economic interpretability. Then, we estimate the model on GPSS data, which does not have an established index, and extract economic insights that are not as clear when studying the data on a question-by-question basis. For example, we document how environmental concern fluctuates with financial insecurity, dropping steeply during the 2008-2009 recession and not recovering until after 2015.

1 Introduction

As data collection has become cheaper through collection, both traditional sources of discrete data, such as survey data responses, and non-traditional sources of discrete data, such as unorganized text data, network data, shopping basket data, and location data, have grown. The established methods for extracting low-dimensional factors from multivariate discrete data generally involve treating discrete data as if it was continuous. For ordered discrete multivariate data, typical methods involve assigning numerical indices to each ordered outcome and taking an average, either of each variable directly (or z-scores) (Bloom & Reenen, 2010), or of the relative percentage of respondents selecting the “high” vs the “low” outcome (Ludvigson, 2004). Methods involving averaging, however, rely on the assumption that the distance between each ordered categorical outcome is equal, which may not be an accurate assumption. Furthermore, missing or don’t know responses, which are very common in survey data and could contain important information about uncertainty, must be dropped or imputed.

For extracting information from sets of unordered categorical variables, many applied economists either analyze each variable separately, or extract an index using the Filmer-Pritchett method, which is a version of PCA for discrete data (Filmer & Pritchett, 2001). The Filmer-Pritchett method, however, involves converting each variable with multiple outcomes to a set of binary variables, and factorizing a matrix of binary responses. However, 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 (Kolenikov & Angeles, 2009).

This paper makes two major contributions. The first is showing how Bayesian text analysis algorithms popularized in the CS literature can be used to extract interpretable factors from multivariate categorical survey data, without treating discrete outcomes as if they were continuous. The second is how to introduce dynamics for time-dependent survey data with categorical outcomes, using two frameworks that applied econometricians are familiar with. The first method introduced is a version of Hamilton (1989)'s markov switching model for categorical response variables. The model assumes that there is a single hidden state at each point in time in the economy that generates outcome variables; this state is one of a finite set of states and follows a markov-switching process. Each hidden state involves a different multinomial distribution over categorical outcomes. 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 are a mixture of states that generate the outcome variables. The mixture probabilities follow an random walk process and each state corresponds to a different multinomial distribution over categorical outcomes. In the application section, categorical outcomes in the model correspond to survey respondents selecting permutations over survey responses in each month or year.

It is important to have methods specific to discrete data not only to avoid unwarranted assumptions, but also to avoid losing important economic information that are contained in discrete data that are not well-represented by a continuous variable. The index of Michigan Consumer sentiment, for example, is constructed from five questions from the survey that have ordered responses based on views about current and future economic conditions. A series of studies have not found that the index provides little to no forecasting power for consumer spending patterns beyond what is found in aggregate statistics on actual consumer purchases (Ludvigson, 2004). As a sanity check for the new methods, in the application section we show that the extracted index from the Michigan data closely correlates with the existing ad-hoc index, but provides additional information on which questions are driving fluctuations in sentiment. This allows us to distinguish what sort of information the discrete data does provide beyond what is contained in series published on real personal consumption expenditure, for example.

Unlike consumer spending, there are many important economic variables, however, that are not well-measured by spending and other aggregate economic indicators. For example, fluctuations in consumers' environmental sentiments are challenging to measure when the "environment" is not a tangible good, many recreation activities are free and most pollution controls are regulation rather than market-based.

In this case qualitative information provided by ordered and unordered categorical variables may provide important insights about aggregate preferences that are not easily quantified otherwise. For the second application, we extract indices from unordered categorical variables from Gallup Poll Social Series (GPSS) data, which do not have established indices published, and show how the resulting indices provide quantitative insights into consumer's preferences about non-tangible goods around recessions.

2 Related Work

There is a related statistics literature which has derived markov-switching models and state space models for discrete time series, especially poisson time series (Davis *et al.* , 2016). The models described in this paper are a Bayesian approach to these models for categorical time series with a Gaussian process for the latent variables. There is also related work developing dynamic models for discrete data with alternatives to Gaussian processes for the latent variables for improved sampling speed and convergence (Bradley *et al.* , 2018), (Linderman *et al.* , 2015). Given the economic survey data explored in this paper was not of particularly large size or dimension, we did not explore alternatives to the Gaussian process for dynamics.

In computer science, there is a class of related models analysis of large corpuses of text. Latent Dirichlet Allocation (LDA) is a popular model for estimating topics that make up documents in a corpus (Blei *et al.* , 2003). Basic LDA does not have a concept of time-dependence in topic proportions. The model presented adds dynamics in topic proportions to basic LDA. It can also be considered a simplified version of Dynamic LDA, which allows time evolution in both the composition of topics and the topic proportions. Another related model is Grade of Membership (GoM) (Erosheva *et al.* , 2007), which is a hierarchical latent variable model for categorical survey data without dynamics. In contrast to this paper, in GoM, responses to questions in a survey are assumed independent given an individual's mixture over states. Hierarchical latent variable models have also been used in the applied microeconomics literature to successfully model consumer choice of shopping baskets

(Ruiz *et al.* , 2018) and restaurants (Athey *et al.* , 2018).

This work is also related to the computer science and psychology literature on methods for summarizing categorical data using clustering. K-Modes (Huang, 1998) is a version of k-means for categorical data that uses modes, rather than means for clusters and updates cluster modes using the frequency of responses in the dataset.

The application section is related to literature which has related economic phenomena like recessions to trends in consumer sentiment derived from survey data. Scruggs & 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 & Simon (2010) uses Gallup survey data on wellbeing to examine the relationship between the aggregate unemployment rate and health outcomes.

Dominitz & 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.

3 Model

In this section I first introduce the familiar form of the model for continuous-valued outcomes, before describing the new models for categorical outcomes. The following notation will remain the same throughout the section. The number of states is K, the number of possible discrete values that the categorical outcome variable can take is V and the total number of time periods is T. β is the $K \times V$ matrix that describes the K state-specific multinomial distributions over categorical outcomes. Each row, β_k, represents the probability vector for the k-th state-specific multinomial distribution and each row sums to 1.

$x_{t,n}$ are the outcome data; each takes a value in $\{1, \ldots, V\}$, where V is the number of survey response permutations. If, for example, there are 5 questions with 3 possible responses for each, then $V = 5^3$. There are $N = \sum\limits_{t=1}^{T} N_t$ total data points, with N_t representing the number of survey respondents at each time t (for many applications, there are different numbers of respondents at each time period). The full dataset is represented by $\mathbf{x}$.

3.1 Markov-Switching Model

The bayesian form of Hamilton's markov switching model with mean switching and unit variance is as follows. In this section, we follow Hamilton's notation for the state variable and use S to represent the T-length state vector taking values in $\{1, \ldots, K\}$ for each S_t. y_t represents some real-valued outcome variable.

$$y_t = \mu(S_t) + \epsilon_t, \qquad \epsilon_t \sim N(0, 1)$$

S_t follows a discrete-state markov process with $K \times K$ transition matrix P.

$$P(S_t = i | S_{t-1} = j) = P_{ij}, \qquad \sum_{j=1}^{K} P_{ij} = 1$$

The prior distributions on the parameters are as follows: Each column of the transition matrix $P_j \sim Dirichlet(\alpha)$, and the state-specific means have prior $\mu \sim Normal(a_0, A_0)$.

In the discrete form of the model each outcome variable follows a state-specific multinomial distribution.

$$x_{t,n} \sim Mult(\beta_{S_t})$$

. S_t continues to follow a discrete-state markov-process with transition matrix P.

$$P(S_t = i | S_{t-1} = j) = P_{ij}, \qquad \sum_{j=1}^{K} P_{ij} = 1$$

The state-specific means are replaced by the state specific multinomial distributions. Each row of β has prior $\beta_j \sim Dir(\eta)$ and each column of the transition matrix has prior $P_j \sim Dir(\alpha)$, as in the mean-switching version. In the sections that follow I use $p_{mult}(x;q)$ to refer to the multinomial density function for vector x and probability vector q. $p_{dir}(q;\alpha)$ refers to the Dirichlet density function for probability vector q and Dirichlet parameters α. The joint likelihood of the markov-switching model for categorical data is as follows:

$$Pr(x, S, \beta, P; \eta, \alpha) = \prod_{j=1}^{K} p_{dir}(P_j; \alpha) \prod_{j=1}^{K} p_{dir}(\beta_k; \eta) \prod_{t=1}^{T} p(S_t|S_{t-1}) \prod_{t=1}^{T} p_{mult}(x_t; \beta_{S_t})$$

The posterior distribution of the model parameters are estimated via Gibbs Sampling and the steps are described in Appendix A.

3.2 State Space Model

In the previous section, we restrict each time period to being in a single state. A more flexible formulation is to assume that each time period involves a unique mixture over states, so the state-vector is real-valued instead of discrete-valued. The linear state space model is as follows, for $n \times 1$ vector y_t, g_t a K-length continuous valued vector of states, A the $K \times K$ transition matrix, C the volatility matrix, and G the $n \times K$ output matrix.

$$y_t = Gg_t$$

$$g_t = Ag_{t-1} + Cw_t$$

$$g_0 \sim N(\mu_0, \sigma_0)$$

w_t is a vector of normally distributed variables, $w_t \sim N(0, I)$.

Latent Dirichlet Allocation is a hierarchical latent variable model that was originally designed for text data. It models documents as a mixture of states (topics), where each topic involves a distribution over possible words. In LDA, $x_{t,n}$ is an index of a word in $v = 1, \ldots, V$, from document t, where V indexes a list of all possible words in the document. $z_{t,n}$ is the topic assignment for each word. θ_t is a K-length state vector, which describes a mixture over the K multinomial distributions β_k for document t.

$$w_{t,n} \sim \text{Mult}(\beta_{z_{t,n}})$$
$$z_{t,n} \sim \text{Mult}(\theta_t)$$
$$\theta_t \sim \text{Dir}(\alpha)$$
$$\beta_k \sim \text{Dir}(\eta)$$

LDA assumes that given the mixtures over topics θ_t, the T documents are independent. However, the document-specific mixtures θ_t are independent Dirichlet-distributed. This independence is unlikely to hold in settings where there is time dependence in the states.

Previously, when states were discrete, we modeled using a discrete markov process. Now that the states are mixtures and are real-valued, we model using a linear markov process. Our model combines the dynamics of a linear state space model with the hierarchical bayesian nature of latent dirichlet allocation to form the discrete state space model.

The model is as follows. $z_{t,n}$ is a state assignment in $\{1, \ldots, K\}$ for each data point $x_{t,n}$. g_t is a K-length real-valued state vector.

$$x_{t,n} \sim \text{Mult}(\beta_{z_{t,n}})$$
$$z_{t,n} \sim \text{Mult}(\theta(g_t))$$
$$g_t = g_{t-1} + w_t, \qquad w_t \sim N(0, \sigma I)$$
$$\sigma \sim \text{IGamma}(v0, s0)$$
$$\beta_k \sim \text{Dir}(\eta)$$

where $\theta(\cdot)$ is the softmax function, $\theta(y) = \frac{exp(y)}{\sum_{k=1}^{K} exp(y_k)}$ for K-length vector y, which transforms g_t into a vector of proportions between 0 and 1. g_t follows the dynamics of a restricted state space model with $A = I$ and $C = diag(\sigma)$. The gibbs-sampling steps necessary to estimate the model are in Appendix A.

4 Data

4.1 Michigan Survey Data

The raw data contains the survey responses for 500 telephoned respondents in continental U.S. (excluding Hawaii and Alaska) each month of the year. For each month's sample, an independent cross-section sample of households is drawn, and each are often 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.

The raw data on the survey respondents for every month from January 1978 to November 2017 is publicly available. 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.

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

The data is mapped to the model framework as follows. Each month t has N_t survey respondents. $x_{t,n}$ is the response for the n-th survey respondent for month t and corresponds to an index $v \in \{1, \ldots, V\}$. Each index v corresponds to a permutation of potential survey responses. So, for the Michigan data, $x_{t,n} = 1$ corresponds to a survey respondent answering 11111 to the survey, which means the respondent is an optimist who answers "better" or "good" to all of the questions on the U.S. economy.

4.2 Gallup Poll Social Series

The raw 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 repeated each year since 2001. The most recent data available is from 2016. 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; each outcome corresponds to a permutation of potential survey responses.

5 Application

First, we estimate a discrete state space model and markov-switching model on responses to the Michigan Survey of Consumers. We show that the estimated index from the discrete state space 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 discrete state space model on GSPSS responses 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.

5.1 Michigan Consumer Sentiment Indices

The Michigan 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_i 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, and add 2.0 to correct for sample design changes in the 1950s.

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

I estimate a discrete state space model on the Michigan Consumer survey data and compare the estimated weight g_{1t} on the positive sentiment to the ICS, rescaled to match the range and variance of the new index, which is between 0 and 1. To deal with the label-swapping identification issue that is inherent in both new models, we identify the "positive" sentiment using the prior distribution on β. In a two-state model, we do this by decreasing the prior probability of the extreme outcome on state 1. For the Michigan data, this corresponds, for example, to 55555, which is corresponds to an individual who selects the most negative responses for each of the five questions making up the ICS. This can be considered as the discrete and Bayesian analogy to the lower triangular restriction that is placed on factor models for identification purposes. The states are then labelled by ordering them based on the posterior probability of the extreme outcome; the one involving a higher probability of responding 55555 is the more "negative" sentiment and the other is the more "positive" sentiment.

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,

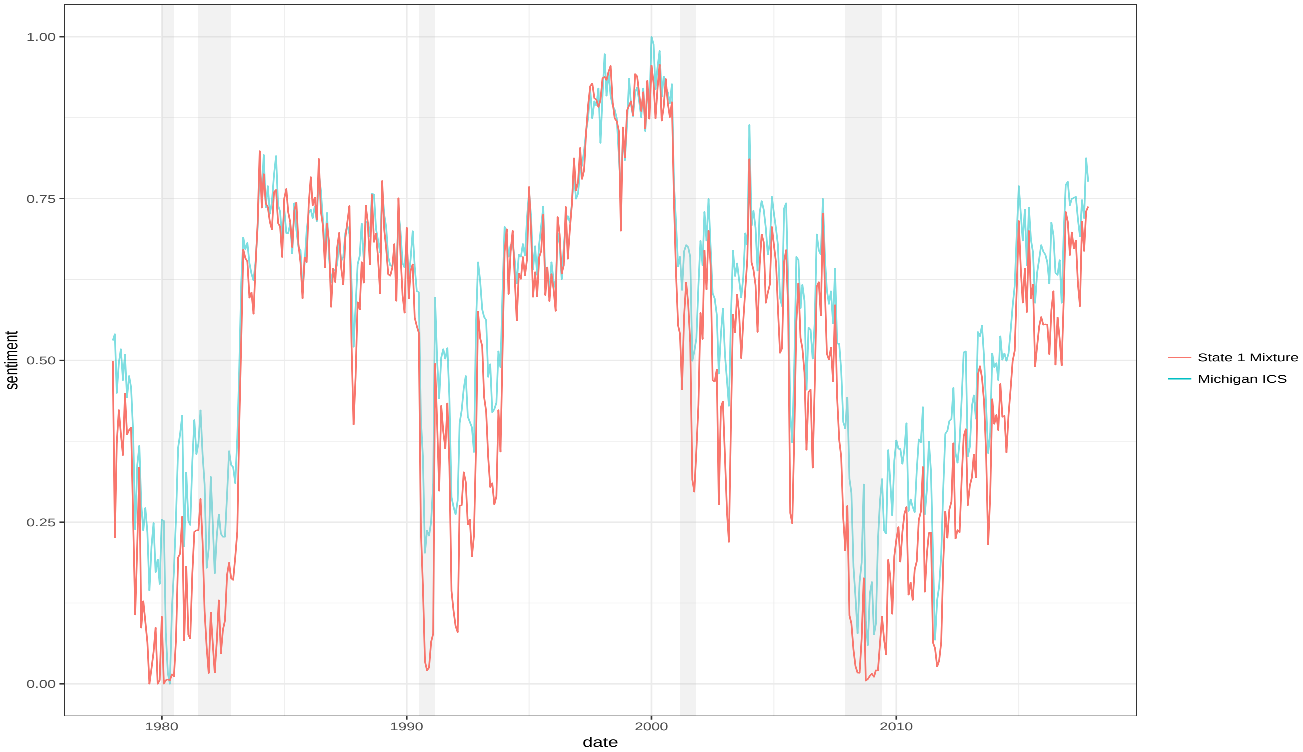

Figure 1: Probability of State 1 vs. ICS

as seen in Figure 3. The discrete state space model index 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. The discrete state space model, however, specifies a probability model that generates the index, which allows additional interpretation of the index. Table 1 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 State 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

State 1		**State 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 1: Model 1: Top 5 State-Specific Multinomial Probabilities

higher weight on State 2, however, there is a higher probability of many respondents selecting all pessimistic answers to the survey (55555).

We also estimate a discrete markov-switching model with $K = 2$ on the Michigan Consumer Survey data with two states. The latent variable S_t is now discrete and has dynamics corresponding to a transition matrix, compared to the discrete state space model g_t which is continuous and has random walk dynamics. The below table gives the top 5 survey permutations and their probabilities for each of the estimated states in the model.

State 1		**State 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 2: Model 1: Top 5 State-Specific Multinomial Probabilities

The discrete state space model has a more complex and computationally intensive estimation procedure than the discrete markov switching model. Examining the difference between Table 2 and Table 1 indicates one advantage of allowing there to be a mixture of states in each period rather than a single state in each period. In both tables, State 1 corresponds to more positive sentiment towards current and future economy, while State 2 corresponds to a more negative sentiment. However, in all periods, there are many respondents responding “11111” to the survey. The State 2, in the markov-switching model, then, 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 states in the discrete state space model allows a cleaner separation between the optimistic group of respondents and pessimistic group of respondents. In the discrete state space model, there always non-zero weight on State 1, so State 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 states are highly persistent: the estimate of the transition probabilities between states is $P(S_{t+1} = 1|S_t = 1) = 0.89$ and $P(S_{t+1} = 2|S_t = 2) = 0.80$. The below figure shows the probability of being in State 2, the more pessimistic state. The Michigan ICS is also plotted and recessions shaded.

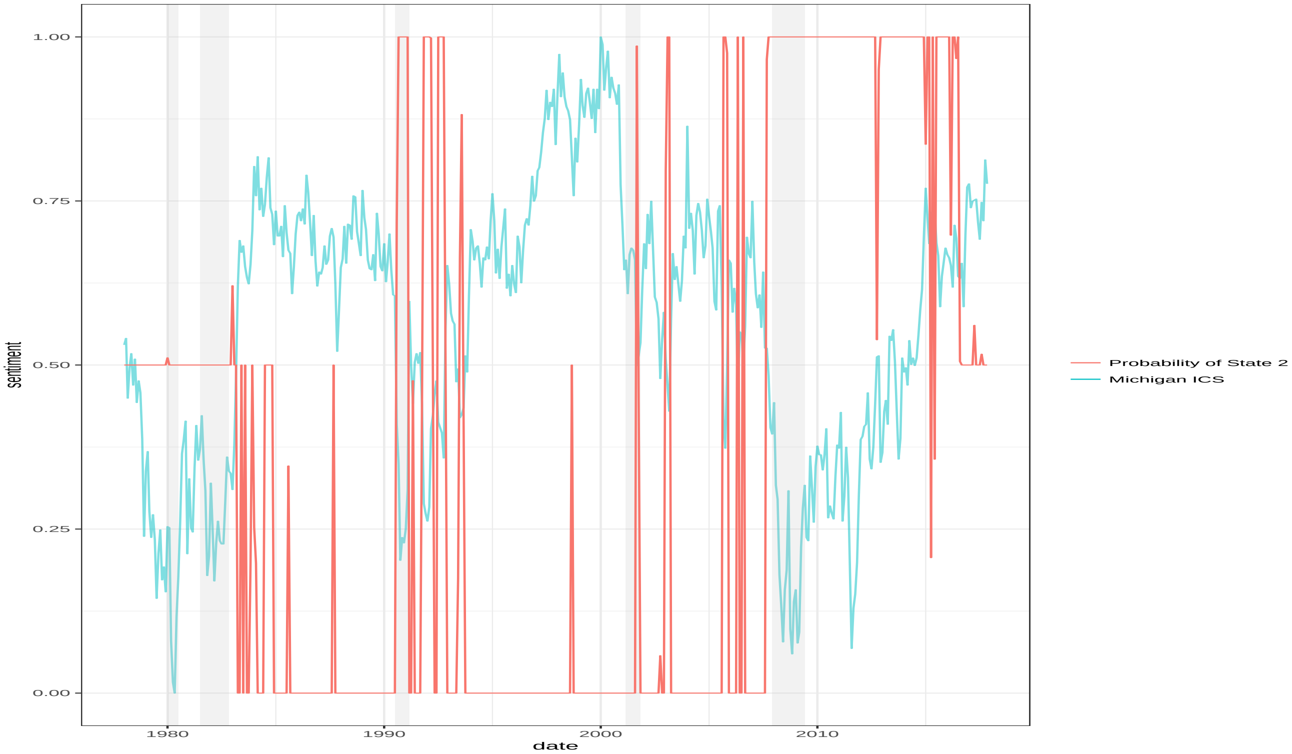

Figure 2: Probability of State 2

The probability of State 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 apparent when examining trends in Personal Consumption Expenditure below.

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 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 state in the model that involves a multinomial distribution over survey response permutations. This motivates extending these methods to quantify

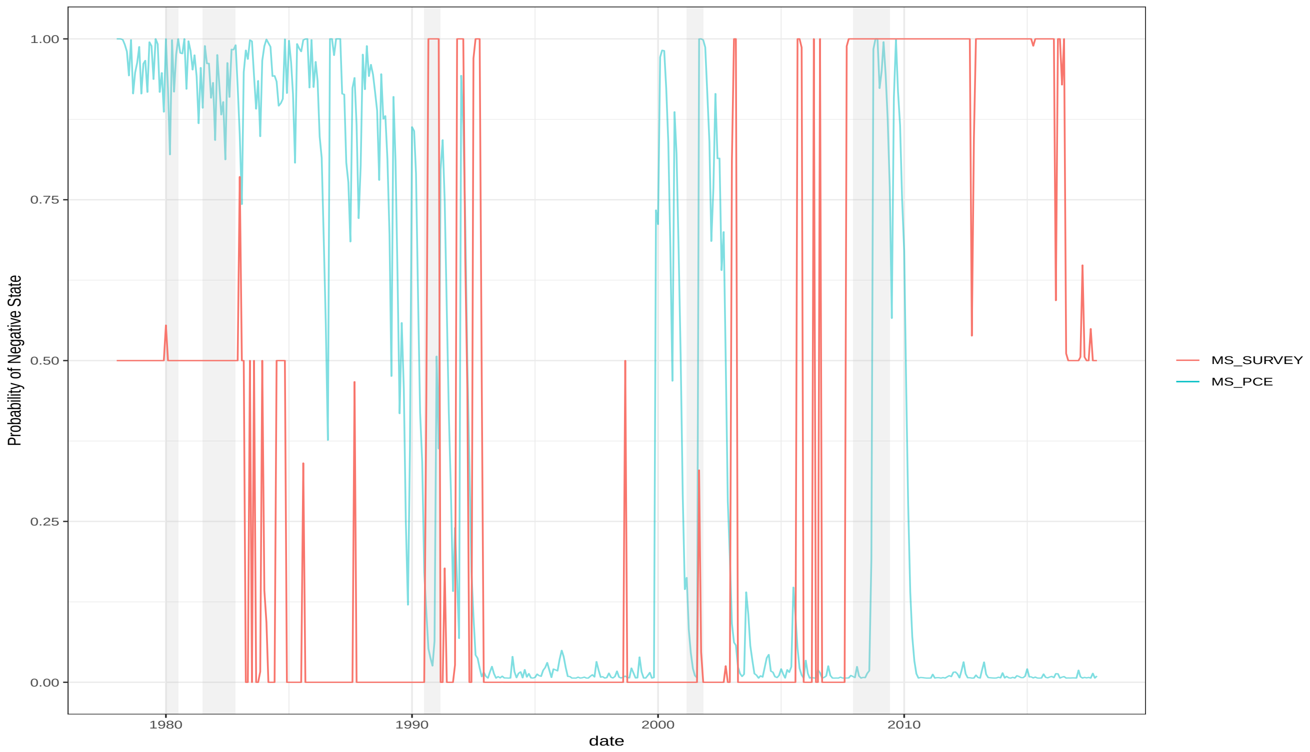

Figure 3: Probabilities of Negative State in DiscreteMS vs. MS on PCE

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.

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

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 of values. 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 the discrete state space model with two states, 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 state 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 5, we plot the weight on the state 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 state corresponding to traditional values slowly decreases from over 90 percent in the early 2000s to under 85 percent in 2016, without showing any break in the trend during the recession.

Estimated sentiment based on satisfaction with personal health and the U.S. health care system, though, does bottom out in 2009, as shown in Figure 6. 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.

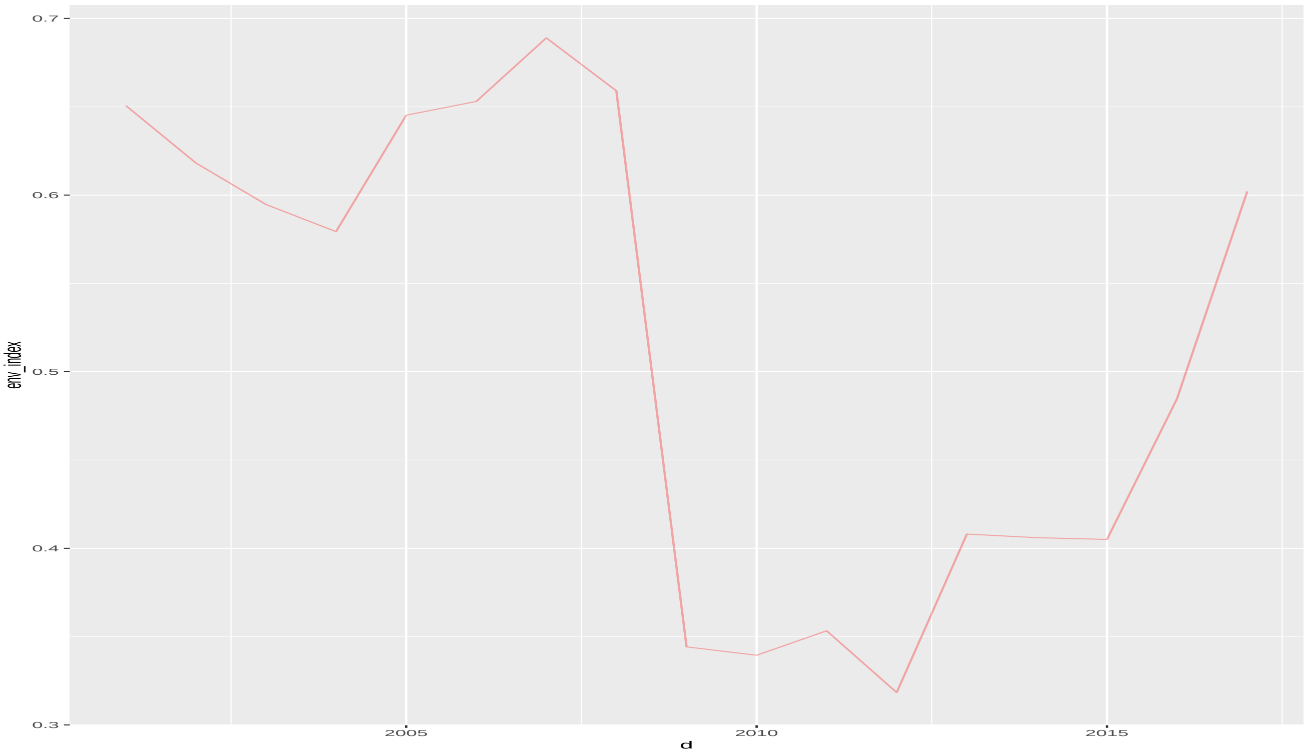

Figure 4: Index of Environmental Concerns

We find that the discrete state space model 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.

6 Conclusion

We derive Bayesian approaches to extracting interpretable discrete and continuous-valued indexes from high-dimensional multinomial time series. The models provide a structural parameteric form for extracting low-dimensional summaries of high-dimensional discrete data, in contrast to existing methods for summarizing categor-

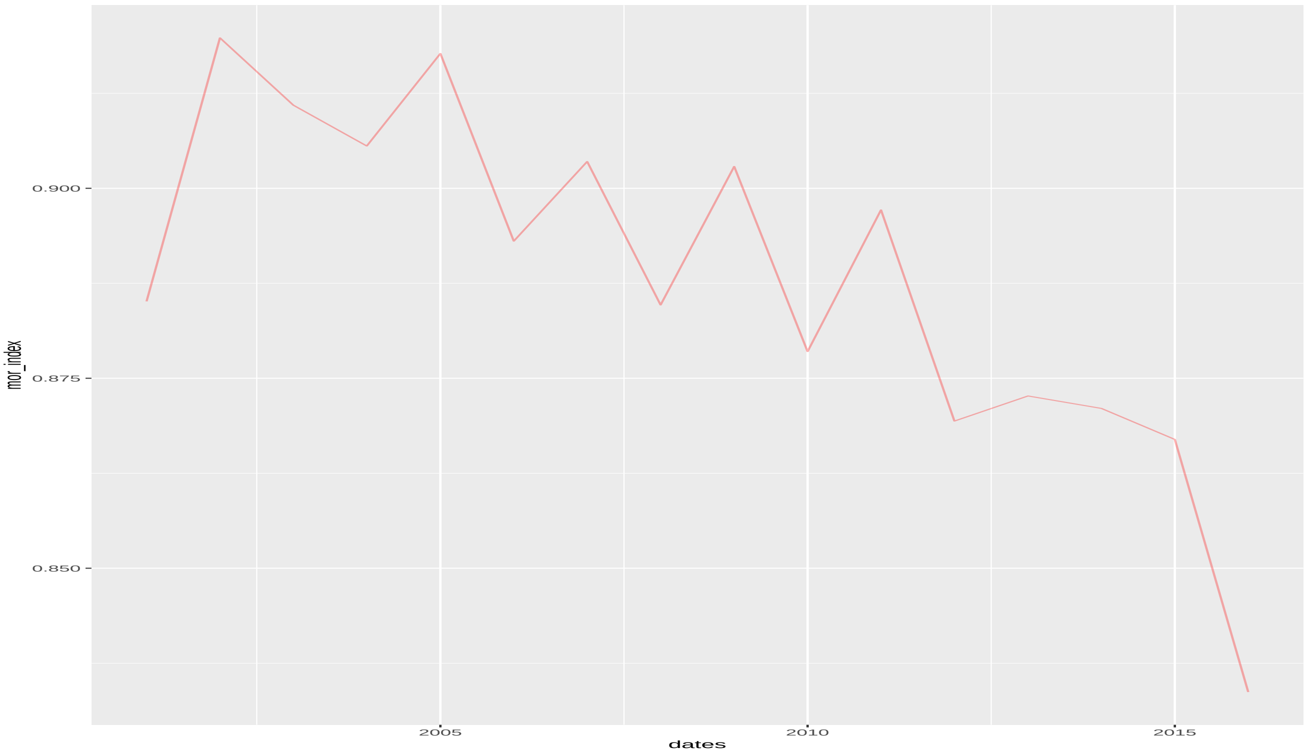

Figure 5: Index of Traditional Values

ical data which tend to use ad-hoc averaging methods or PCA. The models successfully extract interpretable indexes from a variety of ordered and unordered categorical time series data on consumer sentiment towards finance, health, environment, and traditional values.

The relation of these models to the extensive literature on hierarchical latent variable models suggests a variety of extensions for the models, including creating indexes in a supervised fashion based on the relationship between responses and some outcome variable.

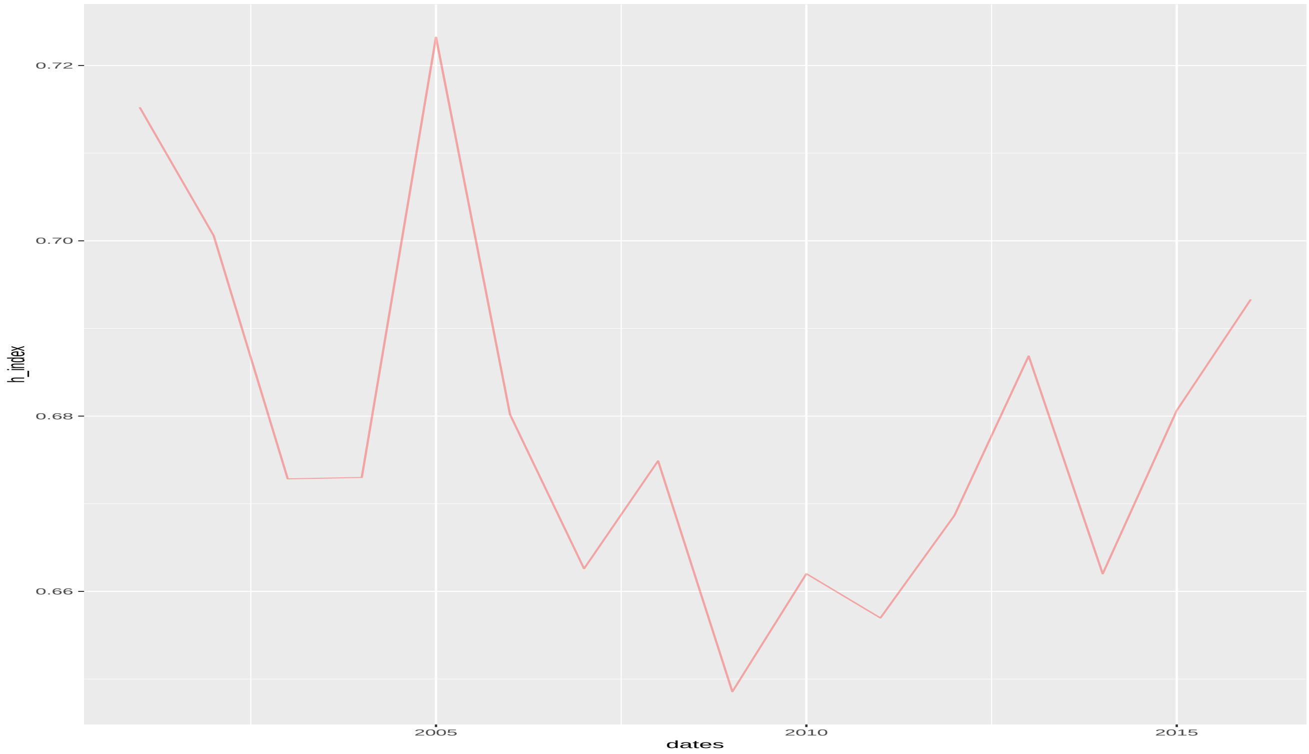

Figure 6: Index of Health Concerns

References

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

Athey, Susan, Blei, David, Donnelly, Robert, Ruiz, Francisco, & Schmidt, Tobias. 2018. Estimating Heterogeneous Consumer Preferences for Restaurants and Travel Time Using Mobile Location Data. *AEA Papers and Proceedings*, **108**, 64–67.

Bhadury, Arnab, Chen, Jianfei, Zhu, Jun, & Liu, Shixia. 2016. *Scaling up Dynamic Topic Models.*

Blei, David M, Ng, Andrew Y, & Jordan, Michael I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **3**, 993–1022.

Bloom, Nicholas, & Reenen, John Van. 2010. Why Do Management Practices Differ Across Firms and Countries? *Journal of Economic Perspectives*, **24**(1), 203–224.

Bradley, Jonathan R., Holan, Scott H., & Wikle, Christopher K. 2018. Computationally Efficient Multivariate Spatio-Temporal Models for High-Dimensional Count-Valued Data. *Bayesian Analysis*, **13**(1), 253–310.

Davis, Richard A, Holan, Scott H, Lund, Robert, & Ravishanker, Nalini. 2016. *Handbook of discrete-valued time series.* CRC Press.

Dominitz, Jeff, & Manski, Charles F. 2003 (August). *How Should We Measure Consumer Confidence (Sentiment)? Evidence from the Michigan Survey of Consumers.* Working Paper 9926. National Bureau of Economic Research.

Erosheva, Elena A, Fienberg, Stephen E, & Joutard, Cyrille. 2007. Describing Disability Through Individual-Level Mixture Models for Multivariate Binary Data. *The Annals of Applied Statistics*, **1**(2), 502–537.

Filmer, Deon, & Pritchett, Lant H. 2001. Estimating Wealth Effects Without Expenditure Data - or Tears. *Demography*, **38**(1), 115–132.

Hamilton, James D. 1989. A New Approach to the Economic Analysis of Nonstationary Time Series and the Business. *Econometrica*, **57**(2), 357–384.

Huang, Zhexue. 1998. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery*, **12**, 283–304.

Kolenikov, Stanislav, & Angeles, Gustavo. 2009. Socioeconomic Status Measurement With Discrete Proxy Variables: Is Prinicipal Component Analysis A Reliable Answer? *Review of Income and Wealth*, **55**(1), 128–165.

Linderman, By Scott W, Johnson, Matthew J, & Adams, Ryan P. 2015. *Dependent Multinomial Models Made Easy: Stick Breaking with the Polya-Gamma Augmentation.*

Ludvigson, Sydney C. 2004. Consumer Confidence and Consumer Spending. *Journal of Economic Perspectives*, **18**(2), 29–50.

Nicholson, Sean, & Simon, Kosali. 2010. How did the recession affect health and related activities of americans? *Preliminary and Incomplete Draft.*

Ruiz, Francisco J. R., Athey, Susan, & Blei, David M. 2018 (nov). *SHOPPER: A Probabilistic Model of Consumer Choice with Substitutes and Complements.*

Scruggs, Lyle, & Benegal, Salil. 2012. Declining public concern about climate change: Can we blame the great recession? *Global Environmental Change*, **22**(2), 505–515.

Welling, Max, & Teh, Yee Whye. 2011. Bayesian Learning via Stochastic Gradient Langevin Dynamics. *Pages 681–688 of: Proceedings of the 28th International Conference on International Conference on Machine Learning.*

A Gibbs Sampling Steps

A.1 Steps for Markov-Switching Model

Generating S_t conditional on P, x_t, S_{t+1}, S_{t-1}, and β

We follow the single-move gibbs sampling procedure of Albert & Chib (1993). We assume that $P(S_1|S_0) = \frac{1}{K}$ and $P(S_{T+1}|S_T) = \frac{1}{K}$.

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

$$p(S_t = i|x_t, S_{t-1}, S_{t+1}, \beta, x_t) \propto p(S_t|S_{t-1})p_{mult}(x_t; \beta_{S_t})p(S_{t+1}|S_t)$$

The first and last terms are taken directly from the current estimate for the transition matrix: $p(S_t = i|S_{t-1} = j) = P_{i,j}$.

Generating transition matrix P conditional on S

$$p(P_j|S) \propto p_{dir}(P_j; \alpha)p_{mult}(n_j; P_j)$$

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

The posterior distribution for each column P_j in the transition matrix P is independent Dirichlet:

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

Generating β conditional on x, S

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

$$m_{k,v} = \sum_{t=1}^{T}\sum_{n=1}^{N_t} \mathbb{1}(x_{t,n} = v)\mathbb{1}(S_t = k)$$

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

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

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 & Teh, 2011) to draw g_t. SGLD is a form of gradient descent, adding Gaussian noise at each step, which Welling & 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_i fulfils certain conditions.

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

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

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

$$\nabla_{g_{t,k}} p(g_t^{(i)}|g_{t-1}^{(i)}, g_{t+1}^{(i-1)}, z_t^{(i-1)}) = \frac{-1}{\sigma^2}(g_{k,t} - g_{k,t-1}) - \frac{1}{\sigma^2}(g_{k,t+1} - g_{k,t}) + n_{t,k} - N_t \theta(g_t)_k$$

$$n_{t,k} = \sum_{n=1}^{N_t} \mathbb{1}(z_{t,n} = k)$$

$\epsilon_i = a(b + i)^{-c}$ for gibbs sampling step i. 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_i that allows for proper convergence.

Generating $z_{t,n}$ conditional on β,x, and g_t

The posterior distribution of $z_{t,n}$ is multinomial with probabilities:

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

for $x_{t,n} = v$.

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 state k:

$$\beta_k|S \sim \text{Dir}(\eta_1 + m_{k,1}, \ldots, \eta_V + \text{m}_{k,V})$$

$$m_{k,v} = \sum_{t=1}^{T}\sum_{n=1}^{N_t} \mathbb{1}(x_{t,n} = v)\mathbb{1}(z_{t,n} = k)$$

Generating σ conditional on g

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

$$v1 = v0 + T$$

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

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