Markov Models for Discrete Time Series Data

Evan Munro

August 28, 2018

1 Introduction

The paper introduces two new methods for modeling time-dependent data with discrete outcomes. The first is a version of Hamilton (1989)'s markov switching model for categorical response variables. The model assumes that there are a set of hidden states in the economy that follow a markov 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. This construction can also be considered a form of a hierarchical latent variable model; it is a dynamic version of LDA (Blei *et al.* , 2003), or a simplified Dynamic LDA (Blei & Lafferty, 2006). Hierarchical latent variable models are increasingly being used in economics to estimate structural models of complex datasets; SHOPPER (Ruiz *et al.* , 2017) uses such a model to estimate an interpretable framework for consumer choice in supermarkets.

This work is also related to the computer science and psychology literature on methods for summarizing categorical data using common states, or clusters. The first is K-Modes (Huang, 1998), which 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 second is Grade of Membership (GoM) (Erosheva *et al.* , 2007), which is a hierarchical latent variable model for categorical survey data, where each state involves a multinomial distribution over question responses for each question. Each individual is modeled as a mixture over states.

The application section of this paper models macroeconomic survey data, since the response data for popular macroeconomic surveys are publicly available and widely referenced in the press. Furthermore, the current popular methods for summarizing survey data into continuous measures, such as PCA (Filmer & Pritchett, 2001) and averaging z-scores, involve assumptions and transformations that are unlikely to hold even for ordinal ordered responses, and are not able to handle missing data or unordered categorical outcomes. The approaches presented here easily handle missing data and unordered discrete outcomes. Our approach assumes that the states are hidden sentiments in the economy which generate different probabilities over survey response permutations. Depending on economic news and other factors, different sentiments are prevalent at different times, and result in fluctuations in the number of survey respondents that answer certain response permutations in each month. The markov-switching model assumes that there is a single sentiment in each

month, and provides estimates of the distribution over survey responses that each sentiment involves, as well as the probability that the economy is in such a state in each month. The discrete state space model assumes that at each month there is a mixture of sentiments; based on the mixture, different respondents choose survey response permutations according to the multinomial distributions for different sentiments.

2 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 $\{1, \ldots, V\}$ and there are $N = \sum_{t=1}^{T} N_t$ total data points, with N_t representing the number of outcome data points at each time t (for many applications, there are different numbers of data points at each time period). The full dataset is represented by $\mathbf{x}$.

2.1 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 $P_j \sim Dir(\alpha)$, and the state-specific means $\mu \sim N(a_0, A_0)$.

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 original version. The joint likelihood of the model is as follows:

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

The model is evaluated via Gibbs Sampling. 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. $p_{dir}(x; \alpha)$ refers to the Dirichlet density function for probability vector q and parameters α.

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:

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

3 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 a new model, 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 as follows.

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 *et al.* , 2010) to draw g_t. SGLD is a form of gradient descent, adding Gaussian noise at each step, which Welling *et al.* (2010) 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

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

$$v1 = v0 + T$$

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

4 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, listed below. Two subindexes are also reported. The questions along with possible responses are as follows:

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), 9 (N/A)
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), 9 (N/A)
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) - 5 (bad times), 8 (don't know), 9 (N/A)
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) - 5 (bad times), 8 (don't know), 9 (N/A)

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), 9 (N/A)

Under the regular procedure for calculating the ICS, data that is missing or incomplete is imputed or 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, unlike averaging approaches which assume that the distances are equal for each consecutive potential response [How do I explain this better?].

Mapping to Model Framework

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 11111, which in term corresponds to the survey response of an optimistic respondent who answers "better" or "good" to all of the questions.

A Note on Identification

To deal with the label-swapping issue that is inherent in both models we have described, 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 55555 on state 1. 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.

5 Application

5.1 Markov-Switching Model

The below table gives the top 5 survey permutations and their probabilities for each state in a model estimated with $K = 2$. The estimate of the transition probabilities is $P(S_{t+1} =$

$1|S_t = 1) = 0.89$ and $P(S_{t+1} = 2|S_t = 2) = 0.80$.

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

The below figure shows the probability of being in the more pessimistic state. The Michigan ICS is also plotted and recessions shaded.

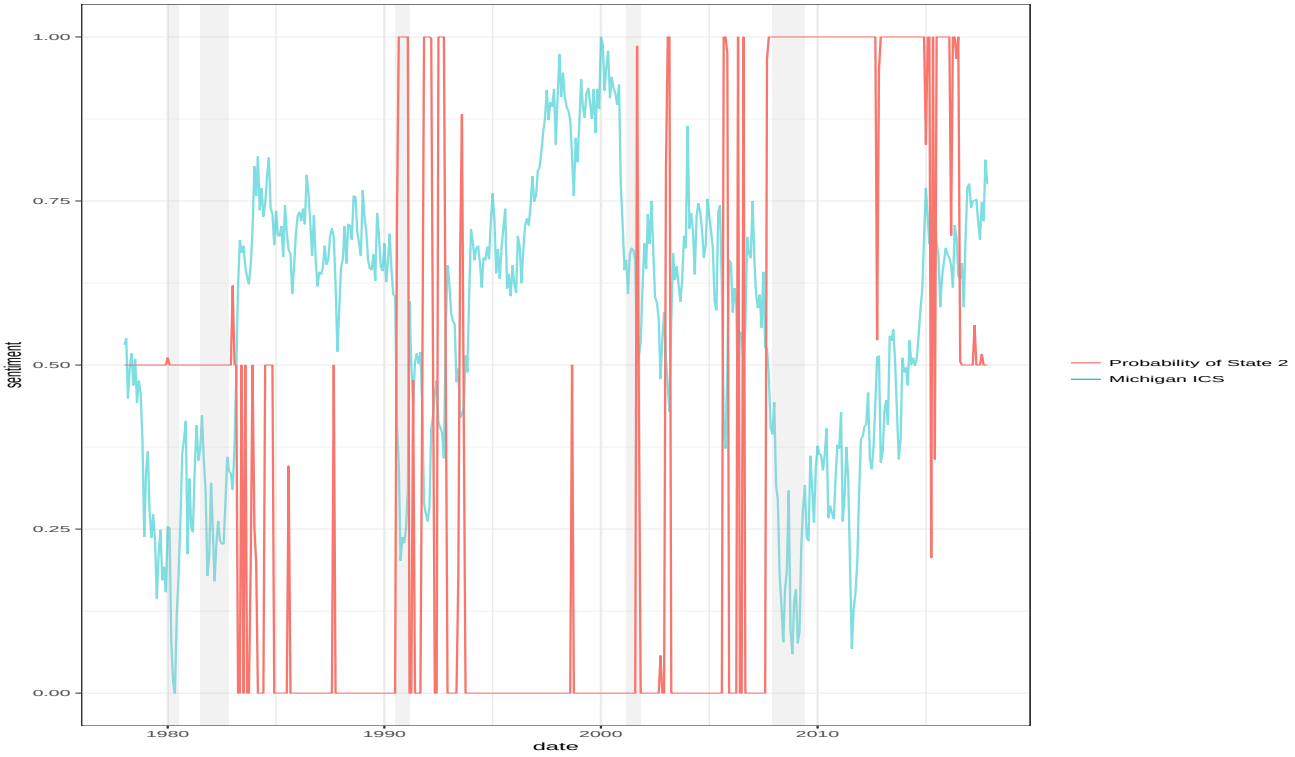

Figure 1: Probability of State 2

Points of interest:

- The probability of State 2 remains high for long after the 2008/2009 recession, but for other recessions drops quickly back to 0; showing how patterns of consumer confidence were different in this recession. This characteristic is much less obvious looking just at ICS, compared to the markov switching model.
- Transition probability matrix makes sense vs. other markov-switching models of macro phenomena

- There are certain periods that are not well-captured by either state (these are the periods near beginning and end of the period where the probability is exactly 0.5, meaning multinomial density of responses for probabilities in both states are close to 0)

5.2 State Space Model

For the discrete state space model, the state-specific multinomial distributions are quite different from the markov-switching model. Since the model is more flexible and allows two states to exist in one month, the negative state no longer contains high probability of positive response permutations. The below table gives the top 5 survey permutations and their probabilities for each state in a model estimated with $K = 2$.

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

The below figure shows the estimates for θ_{g_t} for State 1 along with the publicly available ICS rescaled between 0 and 1. The results are nearly identical to the results from the model without dynamics (see the ECB presentation slides).

References

Albert, James H, & Chib, Siddhartha. 1993. *Bayes Inference via Gibbs Sampling of Autoregressive Time Series Subject to Markov Mean and Variance Shifts*. Tech. rept. 1.

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

Blei, David M, & Lafferty, John D. 2006. *Dynamic Topic Models*. Tech. rept.

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

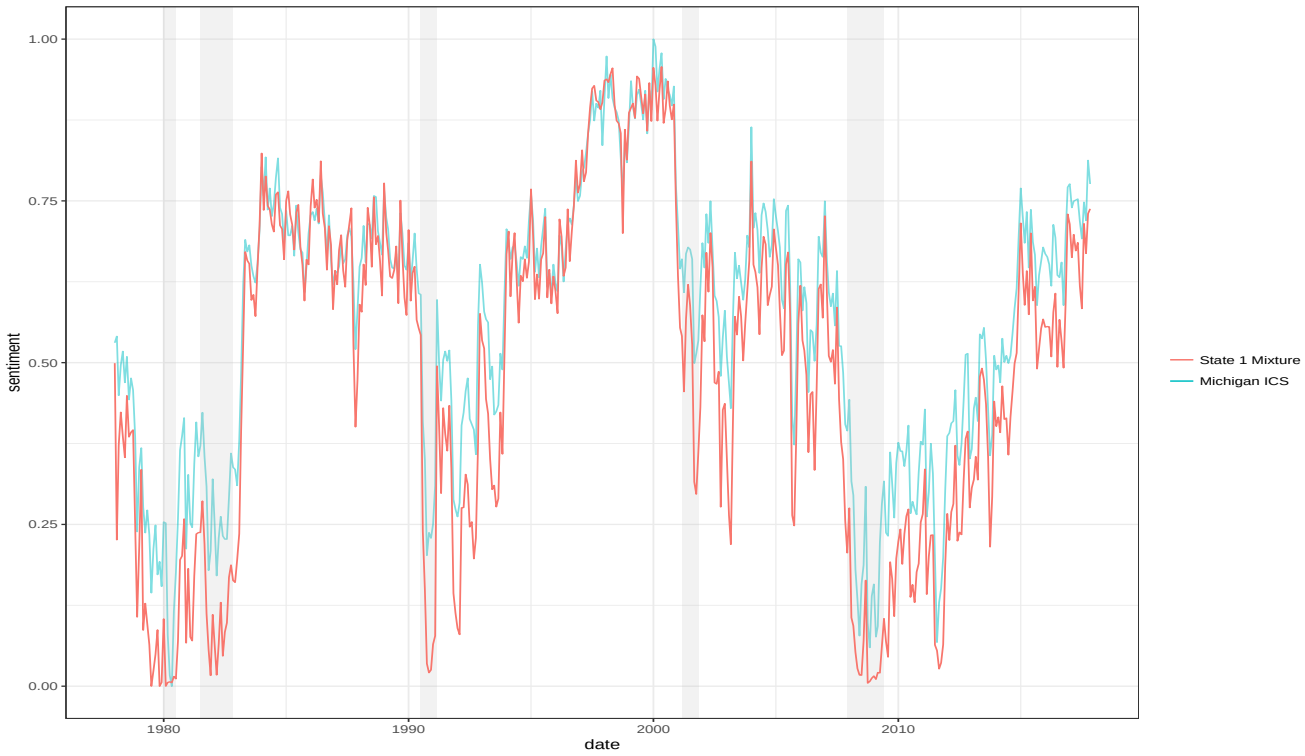

Figure 2: State 1 Mixture

Erosheva, Elena A, Fienberg, Stephen E, & Joutard, Cyrille. 2007. DESCRIBING DISABILITY THROUGH INDIVIDUAL-LEVEL MIXTURE MODELS FOR MULTIVARIATE BINARY DATA 1. **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*. Tech. rept.

Ruiz, Francisco J. R., Athey, Susan, & Blei, David M. 2017. SHOPPER: A Probabilistic Model of Consumer Choice with Substitutes and Complements. nov.

Welling, Max, Bren, D, & Teh, Yee Whye. 2010. *Bayesian Learning via Stochastic Gradient Langevin Dynamics*. Tech. rept.