Bayesian Nonparametric Modeling of Latent Structures

by

Zhengming Xing

Department of Electrical and Computer Engineering
Duke University

Date: _____

Approved:

_____
Lawrence Carin, Supervisor

_____
David Dunson

_____
Robert Calderbank

_____
Guillermo Sapiro

_____
Sunshine Hillygus

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Electrical and Computer Engineering in the Graduate School of Duke University
2015

Abstract

Bayesian Nonparametric Modeling of Latent Structures

by

Zhengming Xing

Department of Electrical and Computer Engineering
Duke University

Date: _____

Approved:

_____
Lawrence Carin, Supervisor

_____
David Dunson

_____
Robert Calderbank

_____
Guillermo Sapiro

_____
Sunshine Hillygus

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Electrical and Computer Engineering
in the Graduate School of Duke University
2015

Copyright © 2015 by Zhengming Xing
All rights reserved except the rights granted by the
Creative Commons Attribution-Noncommercial Licence

Abstract

In this information-explosive era, unprecedented amount of data has been collected in diverse fields such as social networks, infectious disease and political science. This high dimensional, complex and heterogeneous data imposes tremendous challenges on traditional statistical models. Bayesian nonparametric methods address these challenges by providing models that can fit the data with growing complexity. In this thesis, we design novel Bayesian nonparametric models on datasets from three different fields: hyperspectral images analysis, infectious disease and political behaviors.

We first consider analysis of noisy and incomplete hyperspectral imagery, with the objective of removing the noise and inferring the missing data. The noise statistics may be wavelength-dependent, and the fraction of data missing (at random) may be substantial, including potentially entire spectral bands, offering the potential to significantly reduce the quantity of data that need be measured. We achieve this objective by employing Bayesian dictionary learning model, considering two distinct means of imposing sparse dictionary usage and drawing the dictionary elements from a Gaussian process prior, imposing structure on the wavelength dependence of the dictionary elements.

In the second application area, a Bayesian statistical model is developed for analysis of the time-evolving properties of infectious disease, with a particular focus on viruses. The model employs a latent semi-Markovian state process, and the state-

transition statistics are driven by three terms: (*i*) a general time-evolving trend of the overall population, (*ii*) a semi-periodic term that accounts for effects caused by the days of the week, and (*iii*) a regression term that relates the probability of infection to covariates (here, specifically, to the Google Flu Trends data).

In the third application area, extensive information on 3 million randomly sampled United States citizens is used to construct a statistical model of constituent preferences for each U.S. congressional district. This model is linked to the legislative voting record of the legislator from each district, yielding an integrated model for constituency data, legislative roll-call votes, and the text of the legislation. The model is used to examine the extent to which legislators' voting records are aligned with constituent preferences, and the implications of that alignment (or lack thereof) on subsequent election outcomes. The analysis is based on a Bayesian nonparametric formalism, with fast inference via a stochastic variational Bayesian analysis.

To my family

Contents

List of Tables

List of Figures

Acknowledgements

First of all, I would like to express my gratitude to my advisor Professor Lawrence Carin, for his valuable guidance, trust, patience and encouragement during my Ph.D study. Without his tremendous support, this work would not be possible finished. His influence on my thought and working attitude will continue beyond the completion of this thesis.

I would like to thank my committee members including Professor Robert Calderbank, David Dunson, Guillermo Sapiro and Sunshine Hillygus for their time and advices. I would also like to thank Dr. Xuejun liao, Professor Rebecca Willett, Mauro Maggioni and Dr. Bradley Nicholson for their help during my Ph.D study.

My thanks and appreciation also go to all my former and current colleagues in our research group: Dr. Mingyuan Zhou, Dr. Lingbo Li, Dr. Minhua Chen, Dr. Xianxing Zhang, Dr. Lihan He, Dr. Qi An, Dr. Chunping Wang, Dr. Eric Wang, Dr. Haojun Chen, Dr. Lu Ren, Dr. Chenhui Cai, Dr.John Paisley, Dr. Miao Liu, Dr. Bo Chen, Dr.Jianbo Yang, Dr. Xin Yuan, Dr. Liming Wang, Dr.Haichao Zhang, Dr. Esther Salazar, Shaobo Han, Wenzhao Lian, Yingjian Wang, Changwei Hu, David Carlson, Kyle Ulrich, Zhe Gan, Yunchen Pu and Zhao Song. I really enjoy the time spent with them and feel so lucky to work in such dynamic research environment.

Last but not least, I would like to thank my parent for their love and support.

1

Introduction

With the development of social networks and mobile devices, unprecedentedly rich and complex amount of data has been collected in various fields. While the massive data undoubtedly challenges the traditional way of storing, transferring and analyzing data, it also provides opportunities for innovative statistic analysis. For example, Google Flu Trends provided the estimation of flu activities by aggregating the Google search queries. Nielsen updated their TV popularity metrics with the consideration of social activities in Twitter. Nate Silver successfully predicted the 2012 General election result for all 50 states by modeling political poll results and demographic information.

In this dissertation, we design statistical models for several novel datasets, such as a Duke-student-study dataset that contains three years of self-reported flu-related symptoms data for nearly one thousand students; a Catalist dataset which contains 3 million American voters' information ranging from religion, party affiliation to financial status. The challenges of designing statistical models for these complex data include accounting for dependencies within the dataset, utilizing side information, jointly analyzing the data collected from different sources and scaling up the inference

algorithms. Bayesian nonparametric methods address these challenges by allowing the complexity of the model to grow with the size of the data. The dependence (e.g. spatial, temporal dependence) and side information can be introduced to the nonparametric Bayesian priors by assigning a dependent random probability measure [DP08], [RDCD11]. Integrating the Bayesian nonparametric prior with a proper generative model, we may jointly model data from multiple sources. The stochastic variational inference [HBWP13] method provides an efficient way of scaling up the inference algorithms for Bayesian nonparametric models.

The remaining of this chapter is organized as following, a brief review of several widely used nonparametric priors is given in Section 1.1 and a summary of factor analysis and latent Dirichlet allocation are presented in Section 1.2 and Section 1.3, respectively. The organization of this thesis is provided in Section 1.4.

1.1 Nonparametric Prior

In this section, we will give a brief review of several class of most widely used nonparametric priors in machine learning, including Gaussian Process, Dirichlet Process and Beta Process.

1.1.1 Gaussian process

A Gaussian process [RW06] is a collection of random variables, any finite collection of which follows a multivariate Gaussian distribution. A Gaussian process can be fully specified by it mean and covariance function. Assume we have a collection of random function variable $f(x)$ indexed by argument (input) $\mathbf{x}$, the Gaussian process can be represented as following

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\boldsymbol{x}), k(\boldsymbol{x})) \tag{1.1}$$

where $m(\boldsymbol{x})$ is the mean function and $k(\boldsymbol{x})$ is the covariance function. There exists multiple ways of defining these functions and can be found in [RW06]. Given a finite set of input (training data) $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^N$, we may calculated the mean $\boldsymbol{m}$ and covariance $\boldsymbol{\Sigma}$, which define a joint distribution over the function values $f(\boldsymbol{x})$.

$$f(\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{\Sigma}) \tag{1.2}$$

where $\boldsymbol{m} = [m(\mathbf{x}_1), ..., m(\mathbf{x}_N)]^T$ and $\boldsymbol{f} = [f(\mathbf{x}_1), ..., f(\mathbf{x}_N)]^T$. Given a test data $\boldsymbol{x}_{N+1}$, we may calculated the predict value based on conditional distribution $p(f(\mathbf{x}_{n+1})|\mathcal{D})$, which is

$$\begin{aligned} f(\boldsymbol{x}_{N+1}|\mathcal{D}) &\sim \mathcal{N}(m_{\mathcal{D}}, k_{\mathcal{D}}) \\ m_{\mathcal{D}}(\boldsymbol{x}_{N+1}) &= m(\boldsymbol{x}_{N+1}) + \Sigma(\mathbf{X}, \boldsymbol{x}_{N+1})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{f} - \boldsymbol{m}) \\ k_{\mathcal{D}}(\boldsymbol{x}_{N+1}, \boldsymbol{x}_{N+1}) &= k(\boldsymbol{x}_{N+1}, \boldsymbol{x}_{N+1}) - \Sigma(\mathbf{X}, \boldsymbol{x}_{N+1})^T \boldsymbol{\Sigma}^{-1} \Sigma(\mathbf{X}, \boldsymbol{x}_{N+1}) \end{aligned} \tag{1.3}$$

where $\Sigma(\mathbf{X}, \boldsymbol{x}_{N+1})$ is a vector of covariance between training case and $\boldsymbol{x}_{N+1}$.

Through this thesis, for simplicity, we assume the mean function to be zero and covariance function has the following form

$$k(\mathbf{x}_n, \mathbf{x}_m) = \zeta_1 \exp(||\mathbf{x}_n - \mathbf{x}_m||_2/\zeta_2) \tag{1.4}$$

where ζ_1 and ζ_2 are the hyperparameters and we may place gamma priors on them. Gaussian process models can be used to formulate a Bayesian framework for regression and classification. In Chapter 2, the Gaussian process is employed as the prior for dictionary elements, resulting in wavelength dependent structure in the dictionary elements.

1.1.2 Dirichlet process

Similar to Gaussian process, Dirichlet process (DP) is a distribution over distributions. The finite dimensional marginal distribution of Dirichlet process is Dirichlet

distributed. Specifically, let H be a probability measure over some measurable space S and α be a positive concentration number. G is a draw from DP(αH), if

$$(G(A_1), ..., G(A_n)) \sim \text{Dir}(\alpha H(A_1), ..., \alpha H(A_n) \tag{1.5}$$

for any finite measurable partition A_1,...,A_n of S. Given a measurable set A, the mean of the DP is $E(G(A)) = H(A)$ and the variance of DP is $var(G(A)) = H(A)(1 - H(A))/(\alpha + 1)$. As can be seen, a larger α will result a smaller variance, which implies the DP will concentrate more of its mass around the mean.

A stick breaking construction is a more explicit representation of a DP, which can be expressed as follows [Set91],

$$\begin{aligned} \beta_k &\sim \text{Beta}(1, \alpha) \ \ , \theta_k^* \sim H \\ \pi_k = \beta_k \prod_{l=1}^{k-1} (1 - \beta_l) \ \ &, G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}^* \end{aligned} \tag{1.6}$$

where $\delta_{\theta_k}^*$ is an atom at θ_k^*. Note the weights $\{\pi_k\}_{l=1}^{\infty}$ are breaking a unit length "stick " with the control of parameter α. If α is large, each β_k draw from Beta$(1, \alpha)$ will be relatively small, thus we will obtain more short sticks. On the other hand, if α is small, we will have a few large sticks with the remaining very small. The stick breaking distribution over $\boldsymbol{\pi}$ is sometimes written as $\boldsymbol{\pi} \sim \text{GEM}(\alpha)$.

The most common application of Dirichlet process prior is in clustering data using mixture model. Given a set of data $\{\boldsymbol{x}_i\}_i^N$, the DP mixture model can be represented as following

$$\begin{aligned} \boldsymbol{\pi} &\sim \text{GEM}(\alpha) \ , \theta_k^* \sim H \\ z_i \sim \text{Mult}(\pi) \ , &\boldsymbol{x}_i \sim F(\theta_{z_i^*}) \end{aligned} \tag{1.7}$$

where z_i is the cluster assignment variables. $F(\theta_{z_i}^*)$ is the distribution with parameter $\theta_{z_i}^*$ over data in cluster k. H is the prior over cluster parameters.

The Dirichlet process has provided a solution for clustering data in single group. [TJBB06] has extend the DP to the hierarchical Dirichlet process (HDP) for clustering the data from multiple groups. Suppose we have observations $\{\boldsymbol{x}_{ij}\}$, which denotes the ith observation in group j. Our objective is to model each group with a mixture model and then link these mixture models. The HDP achieved this by defining a set of probability measure G_j, one for each group, and a global random probability measure G_0. The mathematical formulation is as follows,

$$G_0 \sim DP(\gamma, H)\ , G_j \sim DP(\alpha_0, G_0) \tag{1.8}$$

where γ and α_0 is the concentration parameters. H is the base probability measure.

The HDP has been employed for various of applications, such as topic modeling [WMM09], and for a hidden Markov model [FSJW08]. In Chapter 4, a matrix completion model integrated with HDP will be used to capture the statistical characterization of people living in each US congressional district.

1.1.3 Beta process

The beta process was defined by [Hjo90] for application in survival analysis. [TJ07b] modified the definition of the beta process with a complete random measure. Assume B_0 is a finite and continuous base measure on some measurable space Ω and c is a positive concentration function over Ω, the Levy measure of the beta process is

$$\gamma(d\omega, dp) = c(\omega)p^{-1}(1-p)^{c(\omega)-1}dpB_0(d\omega) \tag{1.9}$$

and we denote the beta process as $BP(c, B_0)$. Like the Dirichlet process, the beta process can be written in set function form,

$$B = \sum_i p_i \delta_{\omega_i} \tag{1.10}$$

where $(\omega_i, p_i) \in \Omega \times [0, 1]$ is a set of points from a Poisson process with base measure γ. δ_{ω_i} is a point mass at location ω. [PC09b] further extended [TJ07b]'s definition

to two parameter setting, which this thesis is based upon. Let a and b be two scale parameters and $\{B_1, ..., B_k\}$ be K equal measured partition region of space ω. The set function form of beta process is as following,

$$H = \sum_{k=1}^{K} \pi_k \delta_{\psi_k} \ , \pi_k \sim \text{Beta}(a/K, b(K-1)/K) \ \ , \psi_k \sim B_\psi \tag{1.11}$$

π_k serves as a new measure on Ω which parameterize a Bernoulli process. Bernoulli process is conjugate prior of beta process. A draw from Bernoulli process $X_i \sim BeP(H)$ can be denoted as the measure $X_i(\omega) = \sum_k z_{ik} \delta_{\omega_k}(\omega)$, where $z_{ik} \sim \text{beta}(\pi_k)$.

The beta-Bernoulli process provides a convenient prior for binary vectors $\boldsymbol{z}_i \in \{0,1\}^K$. Specifically, the model is as following

$$z_{ik} \sim \text{Bernoulli}(\pi_k), \ \pi_k \sim \text{Beta}(a/K, b(K-1)/K) \tag{1.12}$$

If we let $K \to \infty$ and integrate out π_k. The draws of $\{\boldsymbol{z}_i\}_{i=1}^N$ may constituted as follows. For each $\boldsymbol{z}_i$, draw $c_i \sim \text{Poisson}(\frac{a}{b+i-1})$ and define $C_i = \sum_{j=1}^{i} c_j$ with $C_0 = 0$. Let z_{ik} represent the kth component of $\boldsymbol{z}_i$, and $z_{ik} = 0$ for $k > C_i$. For $k = 1, ..., C_{i-1}$, $z_{ik} \sim \text{Bernoulli}(\frac{n_{ik}}{b+i-1})$, where $n_{ik} = \sum_{j=1}^{i-1} z_{jk}$. For $k = C_{i-1}+1, ..., C_i$, we set $z_{ik} = 0$. Notice with increasing i, $\frac{a}{b+i-1}$ becomes small, and it is probable that c_i will be small. Hence, with increasing i, the number of new non-zero components of z_i diminishes. In Chapter 2, we employ Beta-Bernoulli process as our prior on sparse code to infer the proper number of dictionary atoms.

1.2 Factor analysis

Factor analysis is a statistical method which aims at representing large set of correlated variables with a small uncorrelated set of latent variables. Factor analysis is widely used in social science, gene expression analysis and chemistry. In this thesis, factor analysis models serve as a building block for the proposed model in Chapter 2

and Chapter 4. Following is a brief review of Factor analysis model. Let us assume a data matrix $\mathbf{Y} \in \mathbb{R}^{P \times N}$ and each column is denoted as $\boldsymbol{y}_i$. The factor analysis model is as follows,

$$\boldsymbol{y}_i = \mathbf{D}\boldsymbol{s}_i + \boldsymbol{\epsilon}_i \tag{1.13}$$

where $\mathbf{D} \in \mathbb{R}^{P \times K}$ is the factor loading matrix and the kth column of $\mathbf{D}$ is drawn $\boldsymbol{d}_k \sim \mathcal{N}(0, \mathbf{I})$. $\boldsymbol{s}_i \in \mathbb{R}^K$ is the factor score and drawn $\boldsymbol{s}_i \sim \mathcal{N}(0, \beta^{-1})$. $\boldsymbol{\epsilon}_i$ is the noise term and usually assumed Gaussian. K is the number of factors and usually much less that P and N. For applications like gene expression analysis, we may employ sparse priors such as the Laplace prior on factor loading matrix $\mathbf{D}$. Specifically, let d_{pk} denote the element in the pth row and kth column of $\mathbf{D}$, the hierarchical representation of Laplace prior is $d_{pk} \sim \mathcal{N}(0, \alpha_{pk}^{-1}),\ \alpha_{pk} \sim \text{InvGamma}(a_0, b_0)$.

Factor analysis models are also used to solve the problem of matrix factorization where the substantial amount of data in matrix $\mathbf{Y}$ are missing. Moreover, the factor analysis model is not limited in real value matrices. With proper link functions, we can extend the similar model construction to binary and ordered categorical data.

1.3 Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) [BNJ03] is a generative model widely used in exploring the latent semantics of a corpus of documents. LDA assumes each document is a mixture of latent topics, represented by a distribution over words. Suppose we have N documents in a corpus, the vocabulary size is V, K is the number of topics and V_n is the number of words in document n. The generative process of LDA is summarized as following,

- For each topic $k = 1 :, .., K$, draw topic distribution $\boldsymbol{\phi}_k \sim \text{Dir}(\zeta)$
- For document n

- Draw distribution over topic $\boldsymbol{\theta}_n \sim \text{Dir}(\gamma)$
- For the ith word
  - Draw topic indicator $z_{in} \sim \text{mult}(\boldsymbol{\theta}_n)$
  - Draw words $w_{in} \sim \text{mult}(\boldsymbol{\phi}_{z_{in}})$

According to the above generative process, the joint distribution is written as

$$p(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi}|\zeta, \gamma) = \prod_{k=1}^{K} \text{Dir}(\boldsymbol{\phi}_k|\gamma) \prod_{n=1}^{N} \text{Dir}(\boldsymbol{\theta}_n|\gamma) \prod_{n=1}^{N} \prod_{i=1}^{V_n} \theta_{n,z_{in}} \phi_{w_{in},z_{in}} \tag{1.14}$$

By integrating out $\boldsymbol{z}$, LDA is a version of probabilistic principle component analysis with Dirichlet prior assigned on each principle component. LDA is also close related with Poisson factor analysis and detail discussion can be found in [ZHDC11].

1.4 Thesis organization

The remaining chapters are organized as following:

In Chapter 2, we consider analysis of noisy and incomplete hyperspectral imagery, with the objective of removing the noise and inferring the missing data. The noise statistics may be wavelength-dependent, and the fraction of data missing (at random) may be substantial, including potentially entire bands, offering the potential to significantly reduce the quantity of data that need be measured. To achieve this objective, the imagery is divided into contiguous three-dimensional (3D) spatio-spectral blocks, of spatial dimension much less than the image dimension. It is assumed that each such 3D block may be represented as a linear combination of dictionary elements of the same dimension, plus noise, and the dictionary elements are learned *in situ* based on the observed data (no *a priori* training). The number of dictionary elements needed for representation of any particular block is typically small relative to the block dimensions, and all the image blocks are processed jointly ("collaboratively") to infer the underlying dictionary. We address dictionary learning from a

Bayesian perspective, considering two distinct means of imposing sparse dictionary usage. These models allow inference of the number of dictionary elements needed as well as the underlying wavelength-dependent noise statistics. It is demonstrated that drawing the dictionary elements from a Gaussian process prior, imposing structure on the wavelength dependence of the dictionary elements, yields significant advantages, relative to the more-conventional approach of using an i.i.d. Gaussian prior for the dictionary elements; this advantage is particularly evident in the presence of noise. The framework is demonstrated by processing hyperspectral imagery with a significant number of voxels missing uniformly at random, with imagery at specific wavelengths missing entirely, and in the presence of substantial additive noise.

In Chapter 3, a Bayesian statistical model is developed for analysis of the time-evolving properties of infectious disease, with a particular focus on viruses. The model employs a latent semi-Markovian state process, and the state-transition statistics are driven by three terms: (*i*) a general time-evolving trend of the overall population, (*ii*) a semi-periodic term that accounts for effects caused by the days of the week, and (*iii*) a regression term that relates the probability of infection to covariates (here, specifically, to the Google Flu Trends data). Computations are performed using Markov Chain Monte Carlo sampling. Results are presented using a novel data set: daily self-reported symptom scores from hundreds of Duke University undergraduate students, collected over three academic years. The illnesses associated with these students are (imperfectly) labeled using real-time (RT) polymerase chain reaction (PCR) testing for several viruses, and gene-expression data were also analyzed. The statistical analysis is performed on the daily, self-reported symptom scores, and the RT PCR and gene-expression data are employed for analysis and interpretation of the model results.

In Chapter 4, extensive information on 3 million randomly sampled United States citizens is used to construct a statistical model of constituent preferences for each

U.S. congressional district. This model is linked to the legislative voting record of the legislator from each district, yielding an integrated model for constituency data, legislative roll-call votes, and the text of the legislation. The model is used to examine the extent to which legislators' voting records are aligned with constituent preferences, and the implications of that alignment (or lack thereof) on subsequent election outcomes. The analysis is based on a Bayesian nonparametric formalism, with fast inference via a stochastic variational Bayesian analysis.

2

Dictionary learning for noisy and incomplete hyperspectral Images

2.1 Introduction

Hyperspectral imagery (HSI) is of significant importance for many remote-sensing applications [DE09, ZG08, ZBGS08, DCSVRC08, MSB07, LWB90, IC00]. When performing such sensing, one often encounters imperfections in the data. For example, some of the voxels in the datacube may be missing, data from entire spectral bands may be missing, and the data are often contaminated with noise. The "cleaning up" of such realistic data often constitutes the first step in HSI analysis. In this paper we address these problems by utilizing new technology being developed in the field of dictionary learning for image analysis. Such dictionary-learning approaches exploit the fact that typical natural imagery may be (blockwise) expanded in terms of a linear combination of dictionary elements [AEB06, EA06, MES08, MBPS09, MBP$^+$08, MSE08, MBP$^+$09, RPCL06, DCS09, RBL$^+$07, BDE07, YWHM09]. The low-dimensional nature of such representations makes them appropriate for addressing image imperfections of the type discussed above. Additionally, as elucidated

below, one may *exploit* the ability to mitigate such imperfections to simplify hyperspectral measurements, reducing the quantity of data that need be measured in the first place (*e.g.*, *purposefully* introducing missing data within the measurement process, with the missing data recovered subsequently in the analysis).

There has been significant recent interest in sparse image representations, in the context of denoising and interpolation [AEB06, EA06, MES08, MBPS09, MBP+08, MSE08, MBP+09, RPCL06], compressive sensing (CS) [CT06, DCS09], and classification [WYG+09]. These applications exploit the fact that images may be sparsely represented in an appropriate dictionary. Recent research has demonstrated the significant utility of learning an often over-complete dictionary matched to the signals of interest (*e.g.*, images) [AEB06, EA06, MES08, MBPS09, MBP+08, MSE08, MBP+09, RPCL06, DCS09, RBL+07, BDE07, YWHM09], which should be contrasted with using orthonormal expansions like the discrete cosine transform or wavelets. To the authors' knowledge, none of this previous dictionary-learning research has focused on HSI data, which manifests challenges, for example, with regard to image dimensionality.

In addition, most of the methods for learning dictionaries are based on solving an optimization problem [AEB06, EA06, MES08, MBPS09, MBP+08, MSE08, MBP+09], in which one seeks to match the dictionary to the imagery of interest, while simultaneously encouraging a sparse representation. These methods have demonstrated state-of-the-art performance for denoising, super-resolution, interpolation, and inpainting. However, such methods typically assume one has access to the noise/residual variance, the size of the dictionary is set *a priori* or fixed via cross-validation, and a single ("point") estimate is learned. In HSI applications the noise variance may vary as a function of wavelength, and the wavelength-dependent noise statistics must be inferred in the analysis.

Dictionary learning has recently been cast as a factor-analysis[1] problem, with the factor loadings corresponding to the dictionary elements. The beta process (BP) [PC09a, ZCP+09, TJ07a] and the Indian buffet process (IBP) [GG05, KG07] are non-parametric Bayesian methods well matched to estimation in factor analysis, allowing one to infer the number of factors (dictionary elements) based on the data itself. Further, one may place a prior on the noise or residual variance, with this inferred from the processed data as well [PC09a, ZCP+09]. In this paper we extend this concept by making the noise statistics a function of the wavelength. An approximation to the full posterior density function of the model parameters may be manifested via Gibbs sampling, yielding an ensemble of dictionary representations. It has recently been demonstrated that an ensemble of solutions is often better than a single "best" solution [EY10] (an ensemble of multiple solutions captures uncertainty in the inference, for example based on limited data). We also compare the BP-based Bayesian construction to a generalized version of Bayesian Lasso [PC08], which allows further linkage of the Bayesian approach to previous optimization-based dictionary-learning methods [AEB06, EA06, MES08, MBPS09, MBP+08, MSE08, MBP+09].

The HSI problem has unique characteristics that should be accounted for when performing dictionary learning. One typically deals with datacubes with over 100 spectral bands, and it is expected that the image associated with most materials will be a relatively smooth function of wavelength. In the aforementioned Bayesian dictionary learning approaches, the components of the dictionary are typically drawn i.i.d. from a Gaussian distribution [ZCP+09], with this corresponding to an ℓ_2 regularizer in optimization-based approaches [MBPS09, MBP+08], as illustrated below. A contribution of this paper involves drawing the components of the dictionary from a Gaussian *process* (GP) [RW06], which allows one to impose a preference for

[1] In factor analysis data are represented as a linear combination of learned basis vectors; the basis vectors are termed "factor loadings" and the weights in the superposition are called "factor scores" [Wes03]

dictionaries with smoothness as a function of wavelength; related smoothness constraints have been considered with non-negative matrix facortorization [Hoy04]. The Bayesian formalism employed in this paper allows one to infer the GP parameters in a data-adaptive manner.

The inference of dictionary elements for representation of HSI data may be related to previous HSI research on endmembers estimation, with which HSI data have been linearly expanded [ZG08, ZG07]. One distinction between the proposed model and much of the endmember research is that we model local spatial information within the dictionary, in addition to the spectral information addressed by most previous endmember research. Further, the dictionary elements are inferred in the presence of significant corruption to the datacube, including missing and noisy data (there might not be any "natural" endmembers in this corrupted data). Nevertheless, there are close connections between dictionary learning and endmember analysis for HSI; specifically, the idea of sparseness has been utilized widely for learning the dictionary elements [AEB06, EA06, MES08, MBPS09, MBP$^+$08, MSE08, MBP$^+$09], as well as in recent endmember research [ZG07]. However, the explicit form of the sparseness promotion employed here is distinct from that employed previously in endmember research. For example, the beta process, when coupled with a Bernoulli process, imposes a self-consistency of the dictionary usage across the image. Additionally, within the GP we impose a prior belief about smoothness of the dictionary elements as a function of wavelength.

We demonstrate that typical spatio-spectral blocks of HSI data may be represented as a linear combination of a small number of dictionary components, much like hyperspectral signatures are typically represented in terms of a small number of endmembers. Let $\boldsymbol{x}_i \in \mathbb{R}^P$ represent the ith block of data (unwrapped into a vector), with $P = n_x \cdot n_y \cdot n_\lambda$; n_x and n_y define the number of pixels in each spatial direction within the block, and n_λ is the number of wavelengths (typically $n_x = n_y = 2$ or

4, and n_λ is 100 or $\boldsymbol{x}$,$\boldsymbol{x}$ more). If $\boldsymbol{x}_i$ can be represented as a linear combination of a small number of dictionary elements, then $\boldsymbol{x}_i = \mathbf{D}\boldsymbol{s}_i$ (this will be made more formal below), where the columns of $\mathbf{D} \in \mathbb{R}^{P \times K}$ represent K dictionary elements and $\boldsymbol{s}_i \in \mathbb{R}^K$ is a sparse weight vector. Compressive sensing [CT06] theory suggests that if $\mathbf{D}$ is known, one may infer $\boldsymbol{s}_i$ and hence $\boldsymbol{x}_i$ by measuring $\boldsymbol{y}_i = \boldsymbol{\Phi}_i \boldsymbol{x}_i$, where $\boldsymbol{y}_i \in \mathbb{R}^m$ and $\boldsymbol{\Phi}_i \in \mathbb{R}^{m \times P}$, for $m \ll P$. The theory dictates that it is necessary for the rows of $\boldsymbol{\Phi}_i$ to be incoherent with the columns of $\mathbf{D}$. If the columns of $\mathbf{D}$ are "spread out" over their P dimensions, then this desired incoherence may be achieved by defining the rows of $\boldsymbol{\Phi}_i$ by randomly selecting m rows from the $P \times P$ identity matrix; this corresponds to measuring a subset of the voxels in $\boldsymbol{x}_i$, selected uniformly at random.

For the hyperspectral data of interest here, it is expected that the columns of $\mathbf{D}$ will be spread out in $\mathbb{R}^P$, as the spectral signature of most materials has non-zero contribution at most wavelengths of interest. In this paper we assume access to randomly (down) sampled components of the vectors $\{\boldsymbol{x}_i\}_{i=1,N}$, and using all of these subsampled vectors ("collaboratively") we infer both $\mathbf{D}$ and $\{\boldsymbol{s}_i\}_{i=1,N}$, with this performed in the presence of additive noise of unknown variance. Since we must infer $\mathbf{D}$ as well as $\{\boldsymbol{s}_i\}_{i=1,N}$, using subsampled versions of $\{\boldsymbol{x}_i\}_{i=1,N}$, this research is closely related to matrix completion [CR08] and collaborative filtering [Mar03]. A distinction with most matrix-completion research is that here the support of each sparse vector $\boldsymbol{s}_i$ may be different, while in matrix-completion theory [CR08] one typically assumes that all $\{\boldsymbol{s}_i\}_{i=1,N}$ have the same support; this assumes the $\{\boldsymbol{x}_i\}_{i=1,N}$ live in a *linear* subspace of $\mathbb{R}^P$ (low-rank assumption). Here the support of the sparse vectors $\{\boldsymbol{s}_i\}_{i=1,N}$ need not be the same, and in this sense the $\{\boldsymbol{x}_i\}_{i=1,N}$ live in a *nonlinear* subspace of $\mathbb{R}^P$ (*e.g.*, a union of subspaces, manifold, etc.), a much more realistic assumption for real HSI. This paper addresses this mathematical problem for the practical challenge of analyzing subsampled hyperspectral imagery; there is also interest in future research on extending the linear-subspace theory [CR08] to

the nonlinear case considered here.

The idea of significantly down-sampling hyperspectral data has at least three potential applications: (*i*) it may be used to process hyperspectral imagery with noisy and missing data, including the case for which (portions of) entire spectral bands may be missing; (*ii*) it may be used to accelerate computations and analyses based on hyperspectral data (*e.g.*, inferring $\{\boldsymbol{s}_i\}_{i=1,N}$ for subsequent material classification [CXG+10]) based on significantly downsampled $\{\boldsymbol{x}_i\}_{i=1,N}$, even when the entire datacube is measured; and (*iii*) it may be used to significantly reduce the quantity of data measured by a hyperspectral camera, or pulled off such. The latter application is similar to compressive sensing [CT06, DCS09], in which the quantity of data that is measured or sampled is significantly less than that of conventional sensors. However, in compressive sensing there has been much research on developing projection/measurement matrices $\boldsymbol{\Phi}$ in which the matrix elements are draws from a subGaussian distribution [CT06, DCS09]. The practical implementation of such projection matrices is often difficult. The proposed measurements are also random, corresponding to random selection of rows of the $P \times P$ identity matrix. However, the proposed compressive measurements in this work may in principle be implemented by modifying *existing* hyperspectral cameras (*e.g.*, one may either "turn off" sensors at a large fraction of the voxels, or simply don't read such data off the camera). We also note that while the specific examples shown considered electro-optic HSI systems, the basic framework may be applied to many other wideband sensing systems, and the focus of this paper is on the underlying statistics and mathematical modeling.

The remainder of this Chapter is organized as follows. In Section 2.2 we provide details on the problems under study, and describe the proposed dictionary learning framework for HSI data. Bayesian inference is performed with a Gibbs sampler, as discussed in Section 2.3, and several example results are presented in Section 3.5

based on real HSI data. Conclusions and directions for future research are discussed in Section 2.5.

2.2 Bayesian Dictionary Learning Framework

Assume a hyperspectral image (HSI) is measured, and that it is partitioned into contiguous sets of voxels, with the ith set denoted $\boldsymbol{x}_i \in \mathbb{R}^{n_x \times n_y \times n_\lambda}$, where n_x and n_y represent the number of pixels in the two spatial dimensions, and n_λ represents the number of sensor wavelengths. In previous endmember research [ZG07] one typically assumes $n_x = n_y = 1$ and consequently the signal is analyzed as a function of wavelength alone, but for our applications we have found improved performance if spatial extent is accounted for. We will also assume that there may be missing components of $\boldsymbol{x}_i$, with the missing values to be inferred in the analysis. The model is fit for voxels for which data are available, and based on the inferred model (discussed further below) the missing values are imputed. For a given image we assume a set of blocks, $\{\boldsymbol{x}_i\}_{i=1,N}$, manifested by potentially considering all possible sets of (possibly overlapping) blocks. In the subsequent discussion, the $\boldsymbol{x}_i$ will be assumed represented as an "unwrapped" vector $\boldsymbol{x}_i \in \mathbb{R}^P$, with $P = n_x \cdot n_y \cdot n_\lambda$.

2.2.1 Factor modeling for dictionary learning

The factor model for each $\boldsymbol{x}_i$ is represented as

$$\boldsymbol{x}_i = \mathbf{D}\boldsymbol{s}_i + \boldsymbol{\epsilon}_i \tag{2.1}$$

where $\mathbf{D} \in \mathbb{R}^{P \times K}$ has columns that define dictionary elements, $\boldsymbol{s}_i \in \mathbb{R}^K$, and $\boldsymbol{\epsilon}_i \in \mathbb{R}^P$ represents noise (or model residual). Note that the dictionary $\mathbf{D}$ is shared across all vector $\{\boldsymbol{x}_i\}_{i=1,N}$, and the factor score $\boldsymbol{s}_i$ is meant to be a sparse, and therefore only a subset of the dictionary elements (columns) are used to represent any particular $\boldsymbol{x}_i$. Our objective is to infer the dictionary $\mathbf{D}$ based upon all $\{\boldsymbol{x}_i\}_{i=1,N}$, and the number

of employed dictionary elements (used columns of $\mathbf{D}$ for representation of $\{\boldsymbol{x}_i\}_{i=1,N}$) is anticipated to be small relative to N (and small relative to P for large n_λ). Once $\mathbf{D}$ is so learned, it may be used via the model to impute missing data, and the $\boldsymbol{\epsilon}_i$ may be subtracted out, to remove noise.

We constitute such a model in a Bayesian setting, and therefore priors are placed on the columns of $\mathbf{D}$, on the sparse vectors $\boldsymbol{s}_i$, and on the noise $\boldsymbol{\epsilon}_i$. Concerning the prior for $\boldsymbol{\epsilon}_i \in \mathbb{R}^P$, we assume the jth component of $\boldsymbol{\epsilon}_i$ may be drawn from the prior

$$\epsilon_{ij} \sim \mathcal{N}(0, \alpha^{-1}) \ , \quad \alpha \sim \text{Gamma}(a_0, b_0) \tag{2.2}$$

The parameters (a_0, b_0) are termed "hyper-parameters", and the gamma probability density function is represented $\text{Gamma}(\alpha; a_0, b_0) = c\alpha^{a_0-1}\exp(-b_0\alpha)$, with $c = \beta^{a_0}/\Gamma(a_0)$, where is $\Gamma(\cdot)$ is a gamma function; the gamma distribution is a "conjugate" prior for α, in that given observed data drawn from the associated Gaussian distribution, the posterior of α is also gamma distributed, with updated hyperparameters [BS09]. The fact that such conjugate priors only require one to update hyperparameters significantly simplifies inference, as discussed further in Section 2.3.

Note that this prior is on the *marginal* probability for each component of the noise, while the estimated posterior distribution does not assume the noise components are independent, and the full noise statistics are inferred (approximately). An important aspect of using such Bayesian constructions is that the noise statistics may be inferred (in terms of a posterior distribution on α), and need not be known *a priori*; most previous research on dictionary learning has assumed that the noise variance is known [EA06, MBPS09, MBP$^+$08, MSE08, MBP$^+$09]. Additionally, in (2.2) a single noise precision α is assumed associated with each $\boldsymbol{\epsilon}_i$; here we also consider the case for which a separate α_λ is assumed for the data at wavelength λ, with a separate gamma prior of the form above employed for each wavelength. This model is appropriate for

the realistic case in which the noise variance is a function of wavelength.

2.2.2 Shrinkage sparseness priors and Bayesian Lasso

There are multiple ways one may impose a desire for a sparse factor score $\boldsymbol{s}_i$, and we consider two methods in this paper. The first method is based on a Bayesian form of Lasso [Tib94]; we consider this construction with the goal of making connections with previous research on sparse dictionary learning. The Bayesian Lasso model was first developed in [PC08]. In this model we utilize the relationship

$$\frac{\sqrt{\gamma\alpha}}{2}\exp(-\sqrt{\gamma\alpha}|s|) = \int_0^\infty \mathcal{N}(s;0,(\alpha\xi)^{-1})\text{InvGa}(\xi;1,\gamma/2)d\xi \tag{2.3}$$

where InvGa$(\cdot)$ represents the inverse-gamma distribution, with $\text{InvGa}(\xi;a,b) = \frac{b^a}{\Gamma(a)}\xi^{-a-1}\exp(-b/\xi)$. Assuming for a moment that the dictionary $\mathbf{D}$ is known, we may represent a draw of the data block $\boldsymbol{x}_i$ in the following manner:

$$\begin{aligned}
\boldsymbol{x}_i &\sim \mathcal{N}(\mathbf{D}\boldsymbol{s}_i, \alpha^{-1}\mathbf{I}_P) \\
s_{ik} &\sim \mathcal{N}(0, \alpha^{-1}\xi_{ik}^{-1}) \\
\alpha &\sim \text{Gamma}(a_0, b_0) \\
\xi_{ik} &\sim \text{InvGa}(1, \gamma_{ik}/2) \\
\gamma_{ik} &\sim \text{Gamma}(a_1, b_1)
\end{aligned} \tag{2.4}$$

where $\mathbf{I}_P$ is the $P \times P$ identity matrix. Below we provide intuition for this model construction by relating it to previous optimization-based approaches.

Note that in [PC08] the authors considered a simpler model, in which the parameter γ_{ik} is replaced by a k-independent γ_i. The model in (2.4) may be viewed as a generalization of that in [PC08], with component-dependent hyperparameters. Similar component-dependent shrinkage has been utilized in the relevance-vector machine (RVM) [Tip01], which employs a Student-t rather than a Laplace sparseness-promoting prior; in this case ξ_{ik} is drawn from a gamma distribution rather than

an inverse-gamma, but otherwise (2.4) is equivalent to the RVM model. The latent variable ξ_{ik} controls whether the kth component of $\boldsymbol{s}_i$ has significant amplitude: if ξ_{ik} is large then the kth component of $\boldsymbol{s}_i$ is negligible (leading to approximately sparse vectors).

We initially consider the simplest model for the kth column of $\mathbf{D}$, $\boldsymbol{d}_k$, such that we may complete the connection of the above model to previous dictionary-learning approaches. Specifically, consider

$$\boldsymbol{d}_k \sim \mathcal{N}(0, \frac{1}{P}\mathbf{I}_P) \tag{2.5}$$

If we integrate out ξ_{ik} from the hierarchy in (2.4) using (2.3), the above model (applied jointly to all data $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1,N}$) has a log posterior density function that satisfies

$$\begin{aligned} &-\log p(\alpha, \{\boldsymbol{s}_i, \boldsymbol{\gamma}_i\}_{i=1,N}|\mathcal{D}) = \\ &\frac{\alpha}{2}\sum_{i=1}^{N}\|\boldsymbol{x}_i - \mathbf{D}\boldsymbol{s}_i\|_2^2 + \sum_{i=1}^{N}\sum_{k=1}^{K}\sqrt{\alpha\gamma_{ik}}|s_{ik}| + \frac{P}{2}\sum_{k=1}^{K}\|\boldsymbol{d}_k\|_2^2 + f(\alpha, \{\boldsymbol{\gamma}_i\}_{i=1,N}) \end{aligned} \tag{2.6}$$

where $f(\alpha, \{\boldsymbol{\gamma}_i\}_{i=1,N})$ is a function that captures regularization placed on α and $\{\boldsymbol{\gamma}_i\}_{i=1,N}$ by the respective gamma priors (as well as other constants). The function $f(\alpha, \{\boldsymbol{\gamma}_i\}_{i=1,N})$ essentially constrains the Lagrange multipliers (defined by α in the first term and $\sqrt{\alpha\gamma_{ik}}$ in the second) in a regularization-based solution. It is important to recognize that while the hierarchical form of the Bayesian model reflected in (2.4) and (2.5) looks somewhat unusual to those unfamiliar with such methods, the log of the posterior in the simplified model corresponds almost exactly to the form of models widely used in optimization-based inference of the model parameters [AEB06, EA06, MES08, MBPS09, MBP$^+$08, MSE08, MBP$^+$09], and it is also closely related to optimization approaches applied to learning endmembers and related HSI research [ZG07, ZG08, ZBGS08, DCSVRC08]. Specifically, the first term to the right

of the equal sign in (2.6) corresponds to the ℓ_2 error between the data $\mathcal{D}$ and the model (which results from the Gaussian assumption on the noise and residual), the second term is a generalized ℓ_1 (Lasso) sparsifying regularizer on the dictionary weights, and the third term is a widely used smoothness term applied to the columns of the dictionary. Concerning the generalized Lasso term, with separate weights $\sqrt{\alpha\gamma_{ik}}$ on each term $|s_{ik}|$, a similar approach has been employed in *adaptive* Lasso [Zou06].

With modern computers and numerical methods like Gibbs sampling (discussed further below), we may approximate the full posterior density function on model parameters, as opposed to a single "point" solution that maximizes (2.6) with respect to the dictionary weights and model parameters. In this context, we note that each consecutive density function in the hierarchical model (2.4) is in the conjugate-exponential family, and therefore all Gibbs update equations are analytic (see [PC08] for a closely related model); recall from above that conjugate priors yield updates that simplify correspond to refinements of model hyperparameters, with this performed sequentially in a Gibbs sampler.

In addition to the generalized Bayesian Lasso model in (2.4), we also considered the original such model [PC08], in which instead of k-dependent γ_{ik} and ξ_{ik}, we consider k-independent γ_i and ξ_i (this also corresponds to the traditional Lasso model [Tib94] when viewed from a MAP perspective, as in (2.6)). We found that this form of the Bayesian Lasso does *not* yield sparse representations in general, based upon a Gibbs sampler implementation. This can be understood by examining (2.4) with $\xi_{ik} \rightarrow \xi_i$, which implies that $s_{ik} \sim \mathcal{N}(0, (\alpha\xi_i)^{-1})$ for all components k; if ξ_i is large then all s_{ik} will tend to be small, while otherwise $\boldsymbol{s}_i$ will tend not to be sparse. With the generalized Bayesian Lasso, the k-dependent ξ_{ik} allows sparseness to be manifested by favoring many of the ξ_{ik} to be large (as a function of component k), but not all of them. Finally, we also considered drawing ξ_{ik} from a gamma rather than an inverse-gamma prior, thereby manifesting a fully Bayesian implementation

of the RVM [Tip01] model. We found that this RVM-like construction yields results almost identical to the generalized Bayesian Lasso model in (2.4); in the former the shrinkage prior is a Student-t [BS09], while in the latter it is the "double"-exponential in (2.3), and each encourages sparse dictionary weights.

2.2.3 Beta-Bernoulli sparseness priors

The generalized Bayesian Lasso construction discussed above imposes that $\{\boldsymbol{s}_i\}_{i=1,N}$ should be sparse, but it does not impose further structure (such as that the $\{\boldsymbol{s}_i\}_{i=1,N}$ should have self-consistency in which dictionary elements are used across the data $\{\boldsymbol{x}_i\}_{i=1,N}$). Further, the shrinkage prior does not impose explicit sparseness on $\boldsymbol{s}_i$, only that many of its components should be very small (but not exactly zero). Finally, the model does not allow one to directly impose a belief about the number of columns of $\mathbf{D}$ that will actually be used to represent the data (*i.e.*, although $\mathbf{D} \in \mathbb{R}^{P\times K}$, we generally set K to a large value, with the goal of automatically inferring the size of the dictionary actually used in the model). To address these goals, researchers have recently developed an Indian buffet process (IBP) [GG05], which may be represented in terms of the beta and Bernoulli processes [TJ07a]; this construction explicitly imposes sparseness. When presenting results, we make comparisons between the beta-Bernoulli method of this section and the generalized Bayesian Lasso model discussed in Section 2.2.2; these are alternative means of constituting the sparse $\{\boldsymbol{s}_i\}_{i=1,N}$.

In this construction the factor scores are represented as

$$\boldsymbol{s}_i = \boldsymbol{w}_i \circ \boldsymbol{z}_i \tag{2.7}$$

$$\boldsymbol{w}_i \sim \mathcal{N}(0, \gamma_w^{-1}\mathbf{I}_k) \tag{2.8}$$

where $\boldsymbol{w}_i \in \mathbb{R}^K$, $\boldsymbol{z}_i \in \{0,1\}^K$, and $\circ$ represents the pointwise (Hadamard) vector product. The sparse binary vectors $\{\boldsymbol{z}_i\}_{i=1,N}$ are constructed via the following beta-

Bernoulli process

$$z_{ik} \sim \text{Bernoulli}(\pi_k) \tag{2.9}$$

$$\pi_k \sim \text{Beta}(a_3/K, b_3(K-1)/K) \tag{2.10}$$

The Bernoulli distribution simply yields a $z_{ik} = 1$ with probability π_k, and $z_{ik} = 0$ with probability $1 - \pi_k$; the beta distribution is a prior on a continuous real random between $(0, 1)$, and is represented as $\text{Beta}(\pi; a, b) = c\pi^{a-1}(1-\pi)^{b-1}$, where $a > 0$, $b > 0$ and $c = \Gamma(a+b)/(\Gamma(a)\Gamma(b))$. In the limit $K \rightarrow \infty$ this construction reduces to a generalization of the Indian buffet process [TJ07a, PC09a]. In practice we truncate K, and the number of non-zero components of each $\boldsymbol{z}_i$ is a random number drawn from $\text{Binomial}(K, a_3K/(a_3 + b_3(K-1)))$, and in the limit $K \rightarrow \infty$ this reduces to $\text{Poisson}(a_3/b_3)$. We may therefore explicitly impose a prior belief on the number of dictionary elements used for each $\boldsymbol{x}_i$ (*i.e.*, the number of non-zero components in $\boldsymbol{s}_i$).

An important aspect of the above beta-Bernoulli construction is that the set of probabilities $\{\pi_k\}_{k=1,K}$ are shared for all $\{\boldsymbol{z}_i\}_{i=1,N}$, which implies that if a particular π_k is large (near one) then the associated dictionary element $\boldsymbol{d}_k$ is likely to be used to represent many of the vectors $\{\boldsymbol{x}_i\}_{i=1,N}$. Similarly, if π_k is small, then associated dictionary element is unlikely to be used across $\{\boldsymbol{x}_i\}_{i=1,N}$. Hence, the model imposes a self-consistency in the use of dictionary elements, which is well matched to the properties of many natural images [BCMS09]. This is a key property of this sparseness construction, which is not accounted for in the Bayesian Lasso model in (2.4).

2.2.4 Gaussian process for dictionary elements

The prior on the dictionary elements presented in (2.5) was considered primarily to make linkages to previous sparse dictionary-learning research, where this prior manifests a smoothness constraint from a maximum *a posteriori* (MAP) perspective.

However, in the context of HSI data, we have further prior information that should be exploited. Specifically, in many cases the signature of materials is a smooth function of wavelength (at least for fine wavelength sampling, and hence large n_λ). To impose this prior knowledge more explicitly, rather than drawing the components of $\boldsymbol{d}_k$ i.i.d. from a normal distribution as in (2.5), we draw $\boldsymbol{d}_k$ from a Gaussian *process* (GP) [RW06].

For the GP construction, let $\lambda_1, \ldots, \lambda_{n_\lambda}$ represent the sensor wavelengths, in increasing order. We wish to impose that for a given spatial location, the correlation between the signal at λ_j and $\lambda_{j'}$ increases with decreasing $|\lambda_j - \lambda_{j'}|$. The GP is a natural way to do this. Specifically, for each spatial location the wavelength-dependent components of each $\boldsymbol{d}_k$ are drawn from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}(j, j') = \boldsymbol{\Sigma}(j', j) \geqslant 0$ represents the correlation between the signal at wavelengths λ_j and $\lambda_{j'}$. As is customary in GP analysis, we assume the covariance matrix has the form

$$\boldsymbol{\Sigma}(j, j') = \zeta_1 \exp[-|\lambda_j - \lambda_{j'}|/\zeta_2] \tag{2.11}$$

Separate gamma priors may be placed on both ζ_1 and ζ_2, although in the experiments we simply set ζ_2 to promote a high probability of smoothness between consecutive wavelengths (we could also place a hyper-prior on ζ_2, but doing so one must employ Metropolis-Hastings sampling [Has70], as there is no analytic Gibbs update equation in this case); a gamma prior is placed on ζ_1, allowing inference of an approximate posterior distribution on this parameter. As is well known, the GP construction does not require uniform sampling of wavelength, and once inference is performed using the available data, it may be used to impute signal values at any other wavelengths (to infer the image at wavelengths for which no data are measured).

This GP-based construction is examined within the dictionary learning applied to HSI data, and it is compared to performance based upon the more-typical i.i.d. normal construction in (2.5). The GP prior for the dictionary will be employed both

in the Bayesian Lasso sparseness construction of Section 2.2.2 and the beta-Bernoulli construction of Section 2.2.3.

2.3 Computational Details

We use Gibbs sampling for the model inference. Samples from the posterior distribution of each random variable are approximated by iteratively sampling from the conditional distributions, given all the other random variables. For the beta-Bernoulli model with GP, the full likelihood is represented as

$$
\begin{aligned}
&P(\mathbf{Y}, \mathbf{D}, \mathbf{W}, \mathbf{Z}, \boldsymbol{\pi}, \gamma_w, \alpha, \zeta_1) = \\
&\prod_{i=1}^{N} \mathcal{N}(\boldsymbol{y}_i; \boldsymbol{\Phi}_i \mathbf{D}(\boldsymbol{w}_i \circ \boldsymbol{z}_i), \alpha^{-1} \boldsymbol{\Phi}_i^T \boldsymbol{\Phi}_i) \mathrm{Gamma}(\alpha; a_0, b_0) \\
&\prod_{k=1}^{K} \mathcal{N}(\boldsymbol{d}_k; 0, P^{-1} \boldsymbol{\Sigma}) \mathrm{Gamma}(\zeta_1; a_4, b_4) \\
&\prod_{i=1}^{N} \prod_{k=1}^{K} \mathrm{Bernoulli}(z_{ik}; \pi_k) \mathrm{Beta}(\pi_k; a_3/K, b_3(K-1)/K) \\
&\prod_{i=1}^{N} \mathcal{N}(\boldsymbol{w}_i; \mathbf{0}, \gamma_w^{-1} \mathbf{I}_K) \mathrm{Gamma}(\gamma_w; a_2, b_2);
\end{aligned}
$$

where $\boldsymbol{y}_i = \boldsymbol{\Phi}_i \boldsymbol{x}_i$; if there are n_i observed voxels from $\boldsymbol{x}_i$, then $\boldsymbol{\Phi}_i \in \{0,1\}^{n_i \times P}$, where the rows of $\boldsymbol{\Phi}_i$ are all zero except a single one, corresponding to which voxels are observed. At each iteration, the samples are drawn from the following conditional distributions.

Sampling $\boldsymbol{d}_k$:

$$
p(\boldsymbol{d}_k|-) = \mathcal{N}(\boldsymbol{\mu}_{d_k}, \boldsymbol{\Sigma}_{d_k})
$$

where the covariance $\boldsymbol{\Sigma}_{d_k}$and mean $\boldsymbol{\mu}_{d_k}$ can be expressed as

$$\boldsymbol{\Sigma}_{d_k} = (P\boldsymbol{\Sigma} + \alpha \sum_{i=1}^{N} z_{ik}^2 w_{ik}^2 \boldsymbol{\Phi}_i^T \boldsymbol{\Phi}_i)^{-1}$$
$$\boldsymbol{\mu}_{d_k} = \alpha \boldsymbol{\Sigma}_{d_k} \sum_{i=1}^{N} z_{ik} w_{ik} \widetilde{\boldsymbol{x}}_i^{-k}$$

where $\widetilde{\boldsymbol{x}}_i^{-k} = \boldsymbol{\Phi}_i^T \boldsymbol{y}_i - \boldsymbol{\Phi}_i^T \boldsymbol{\Phi}_i \mathbf{D}(\boldsymbol{w}_i \circ \boldsymbol{z}_i) + \boldsymbol{\Phi}_i^T \boldsymbol{\Phi}_i \boldsymbol{d}_k (w_{ik} z_{ik})$. In this and the notation below, $p(\boldsymbol{d}_k|-)$ is the probability of $\boldsymbol{d}_k$ conditioned on all other parameters being fixed to the last value in the sequence of Gibbs update equations.

Sampling z_{ik} and w_{ik}:

$$p(z_{ik}|-) = \text{Bernoulli}(\tfrac{\widetilde{\pi_k}}{\widetilde{\pi_k}+1-\pi_k})$$
$$p(w_{ik}|-) = (1 - z_{ik})\mathcal{N}(0, \gamma_w^{-1}) + z_{ik}\mathcal{N}(\mu_{w_{ik}}, \Sigma_{w_{ik}})$$

where

$$\widetilde{\pi_k} = \pi_k \exp\left(-\tfrac{\alpha}{2} w_{ik}^2 \boldsymbol{d}_k^T \boldsymbol{\Phi}_i^T \boldsymbol{\Phi}_i \boldsymbol{d}_k - 2 w_{ik} \boldsymbol{d}_k^T \widetilde{\boldsymbol{x}}_i^{-k}\right)$$
$$\boldsymbol{\Sigma}_{w_{ik}} = (\gamma_w + \alpha \boldsymbol{d}_k^T \boldsymbol{\Phi}_i^T \boldsymbol{\Phi}_i \boldsymbol{d}_k)^{-1}$$
$$\boldsymbol{\mu}_{w_{ik}} = \alpha \Sigma_{w_{ik}} \boldsymbol{d}_k^T \boldsymbol{\Phi}_i^T \boldsymbol{\Phi}_i \widetilde{\boldsymbol{x}}_i^{-k}$$

Sampling π_k:

$$p(\pi_k|-) = \text{Beta}(a_3/K + \textstyle\sum_{i=1}^{N} z_{ik}, b_3(K-1)/K + N - \sum_{i=1}^{N} z_{ik})$$

Sampling γ_w:

$$p(\gamma_w|-) = \text{Gamma}(a_2 + KN/2, b_2 + \textstyle\sum_{i=1}^{N} \boldsymbol{w}_i^T \boldsymbol{w}_i/2)$$

Sampling α:

$$p(\alpha|-) = \text{Gamma}(a_0 + \tfrac{1}{2}\textstyle\sum_{i=1}^{N} |\boldsymbol{\Phi}_i||_{l_0}, b_0 + \tfrac{1}{2}\sum_{i=1}^{N} ||\boldsymbol{\Phi}_i^T \boldsymbol{y}_i - \boldsymbol{\Phi}_i^T \boldsymbol{\Phi}_i \mathbf{D}(\boldsymbol{w}_i \circ \boldsymbol{z}_i)||_{l_2})$$

where$||.||_{\ell_0}$ denotes the ℓ_0 norm and $\|\cdot\|_{l_2}$ denotes the ℓ_2 norm.

Sampling ζ_1:

$$p(\zeta_1|-) = \text{Gamma}(a_4 + \frac{PK}{2}, b_4 + \frac{\sum_{k=1}^{K} \boldsymbol{d}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{d}_k}{2})$$

For the shrinkage-based model (Section 2.2.2), we obtain the conditional distribution of each random variable via a similar procedure. In this case, we note that the conditional distribution of ξ_{ik} is the inverse-Gaussian distribution with parameters $\lambda' = \gamma_k$ and $\mu' = \sqrt{\frac{\gamma_k}{w_i k^2}}$. The inverse-Gaussian density function is given by

$$f(x) = \sqrt{\frac{\lambda'}{2\pi}} x^{-3/2} exp(-\frac{\lambda'(x-\mu')^2}{2(\mu')^2 x}) \ ; \ x > 0$$

The sampling method for the inverse-Gaussian distribution is discussed in [V.S93].

2.4 Examples Using Measured HSI Data

2.4.1 Data considered and model parameter settings

The results presented below are based on analysis of two real hyperspectral data sets, one termed Urban and the other AP Hill. The Urban scene was taken with the Hyperspectral Digital Collection Experiment (HyDICE) sensor over Copperas Cove, Texas; the data are publicly available at `http://www.agc.army.mil/hypercube/`. The APHill scene was taken with the Hyperspectral Mapper (HyMAP) over Virginia (with permission from the US Army Engineer Research and Development Center, Topographic Engineering Center, Fort Belvoir, VA). The Urban data consists of 162 spectral wavelengths and 150×150 spatial pixels, and the AP Hill data has 106 spectral bands and 300×300 spatial pixels. Both datasets have complete datacubes, and in the experiments below we perform analysis based on downsampled versions of each. Note that for each datacube we have removed water-absorption bands (this is how the data were provided to the authors for analysis).

Concerning parameter settings, the gamma priors on the precision of the noise were set as $a_0 = b_0 = 10^{-6}$, with these same hyperparameters used in all results. For the shrinkage model, the hyperparameters $a_1 = b_1 = 10^{-6}$. For the beta-Bernoulli model, we set $a_3 = 128$ and $b_3 = N/4$. In all experiments the truncation level (for the shrinkage and beta-Bernoulli model) was set at $K = 128$. For the Gaussian process, the gamma prior on ζ_1 was set to have parameters $a_4 = b_4 = 10^{-6}$. While there may appear to be a relatively large number of model parameters, these are all set in a "standard" way (*e.g.*, all gamma priors are set with the same hyperparameters, as in [Tip01]), and there has been no tuning of any parameters. We found the beta-Bernoulli model to be particularly insensitive to the truncation level K, with almost identical results manifested for $K = 256$.

Both the shrinkage factor analysis (SFA) of Section 2.2.2 and the beta-process factor analysis (BPFA) model of Section 2.2.3 may be implemented with Gibbs sampling, with analytic update equations, as summarized briefly in Section 2.3. For the results presented below we employed 100 burn-in iterations and 100 collection samples; while this number of samples is clearly insufficient to accurately estimate the full posterior distribution on all model parameters, it has in practice proven sufficient for estimation of mean parameters and the associated mean hyperspectral image. Specifically, the collection samples may be used to provide a mean estimate of the underlying hyperspectral data, while also providing "error bars" (*e.g.*, standard deviation). When presenting inferred images below, we present the mean inferred image. All computations were performed on a desktop computer: Intel Core$^{\rm TM}$, 2 Duo 2.8G CPU, and 3GB RAM. For analysis of the Urban data (all 150×150 spatial pixels, and 162 wavelengths), based upon 2% of the data cube selected uniformly at random, each Gibbs iteration of the GP-based BPFA model required about 10 seconds, while each iteration of the SFA model required about 80 seconds. Note that the BPFA

model has at least two advantages: (i) it is highly insensitive to the truncation level K, and (ii) it is considerably faster than the SFA model. This computational acceleration is manifested because the beta-Bernoulli construction imposes that many of the factor scores are exactly zero, and therefore when the binary indicator $z_{ik} = 0$, one need not update the associated w_{ik}. By contrast, the shrinkage prior imposes that many of the factor scores are small, but not exactly zero, and therefore without setting an (arbitrary) threshold, one must always update all of the factor scores at each Gibbs iteration. This appears to be the main advantage of the BPFA framework *vis-a-vis* SFA, as the accuracy of the results from the two models are often similar, as discussed below.

2.4.2 Recovery of missing voxels

The first experiments consider the Urban and AP Hill data, and we assume observation of 2% of the hyperspectral datacube, with observed voxels selected uniformly at random (98% of the datacube, selected uniformly at random, is either not measured or simply not used in the analysis). Results below are shown for one example such draw of observed voxels, but in the context of numerous such draws highly similar results were observed. In Figures 2.1 and 2.2 are shown recovered images at spectral bands 20 and 100 for the Urban data, and in Figures 2.3 and 2.4 the same is done for the AP Hill data. These results are based upon utilizing image blocks $\boldsymbol{x}_i$ with 4×4 spatial support. Results are shown for the beta-Bernoulli and shrinkage-based factor analysis models (BPFA and SFA, respectively), with and without the Gaussian process (GP) used as a prior for the factor loadings. When GP is not employed, the components of the factor loading are drawn i.i.d. from a normal distribution, as in (2.5). The missing voxels are inferred at all spectral bands simultaneously, and here we only show results at two of the spectral bands, for visualization.

While the results in Figures 2.1-2.4 appear good based upon each of the methods

considered, closer inspection is required to assess modeling quality. In Figure 2.5 we show results for the Urban data, in which we present the results for an entire spectral signature at a representative spatial location. Results are shown with BPFA and GP-BPFA, based upon analysis with 2×2 spatial blocks, and 4×4 blocks. Note that the block size is a modeling/analysis choice, and a given block size is employed on the same (downsampled) data. Specifically, in Figure 2.5 we consider 5% of the datacube selected uniformly at random, and the entire downsampled datacube is analyzed jointly (although we only show spectra at one spatial location). There is a tradeoff in selecting the spatial support of the blocks. In most previous endmember research [ZG08, ZG07], investigators have not considered spatial information at all. By considering 2×2 or 4×4 data blocks $\boldsymbol{x}_i$, there is an opportunity to also employ spatial information in the modeling. However, if the spatial block size becomes too large, there is a danger of increased spectral-signature contamination (mixing/blurring), as a result of containing many material types within the same block $\boldsymbol{x}_i$. As illustrated in the below results, for the data considered we have found 4×4 spatial blocks to provide a good compromise.

In Figure 2.5 we plot the mean inferred spectra, as well as error bars reflective of one standard deviation (estimated from the Gibbs collection samples). The GP tends to yield tighter standard deviations, and the results based upon 4×4 blocks appear to be most accurate. Note that the BPFA results (without GP) are based upon 2×2 blocks, and these results manifest high variability as a function of wavelength, particularly about spectral band 80. These qualitative observations are now made quantitative.

In Table 2.1 we summarize PSNR values computed on the entire inferred datacube, for the Urban data, with no additional additive noise (additive noise is considered below). Similar results were observed for the AP Hill data, and are omitted for brevity. In Table 2.1 we consider observing 2% and 5% of the datacube, uniformly

at random. The results in this table reflect one draw of the observed data, but in considering many such draws, all inferences were consistent with this table. When observing only 2% of the datacube, there is a clear advantage to employing larger spatial blocks, the 4×4 blocks performing often significantly better than the 2×2 blocks. When observing more of the voxels, 5%, the advantage of the 4×4 blocks is present, but not as marked (note of course that the best block size depends also on the sensor spatial resolution). When observing only 2% of the datacube, there is an advantage of the GP when employed within the BPFA model. As more data are observed (5%), the necessity of the GP is less apparent in the case of no additive noise. This is expected, as the GP imposes smoothness as function of wavelength, and this prior information is of particular importance when the observed data are limited. However, when the quantity of observed data is larger, the imposition of smoothness may not be as necessary, and may even be detrimental, if the data alone are sufficient to infer appropriate factor loadings (analogous to spatial-spectral endmembers). The BPFA and SFA yield comparable results, although we have found the BPFA less sensitive to the truncation level K. In fact, it is possible that the SFA results may be improved further by tuning K, but it is anticipated that such tuning will be inappropriate in practice. Additionally, the BPFA has a significant computational advantage, as discussed above.

Note from Table 2.1 that based upon only 5% of the datacube, the 4×4 spatial blocks yield PSNR values of roughly 40 dB, which is consistent with the quality of traditional coding algorithms. The significant advantage of the method developed here is that the datacube has been reconstructed in a manner which may not require one to measure all the voxels in the first place (most traditional compression algorithms first assume access to the entire datacube).

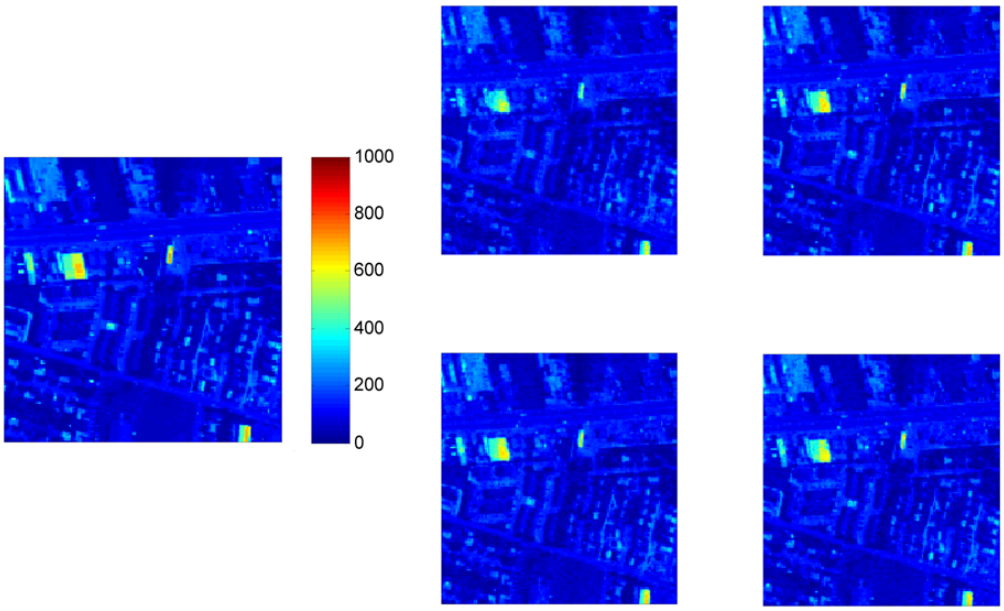


FIGURE 2.1: Recovery of Urban hyperspectral data (normalized reflectance), based upon measuring 2% of the datacube, with voxels selected uniformly at random. The analysis is performed using 4×4 spatial blocks, and all 162 spectral bands. These results are for spectral band 20, although all spectral bands are recovered simultaneously. The same color scale is used in all images, and the total datacube is of dimension $150 \times 150 \times 162$. Results are shown for the beta-process based factor analysis (FA) model (BPFA) and for the shrinkage-based FA model (SFA), with and without a Gaussian process (GP) employed for the factor loadings. Left column: Original image for band 20 at top, and at bottom the observed data from spectral band 20 used in the analysis (unobserved pixels are here set to zero for visualization; we used similar downsampled data of this type from all spectral bands within the joint analysis). Right two columns, clockwise from top-center image: BPFA, GP-BPFA, GP-SFA, SFA.

2.4.3 Missing spectral bands

One may wish to make an inference of the spectral signature at a wavelength that was not actually measured by the sensor. This objective may be manifested if the sensor fails at a wavelength or set of wavelengths. This objective may also be of interest to make interpolations of the datacube, at wavelengths for which the sensor

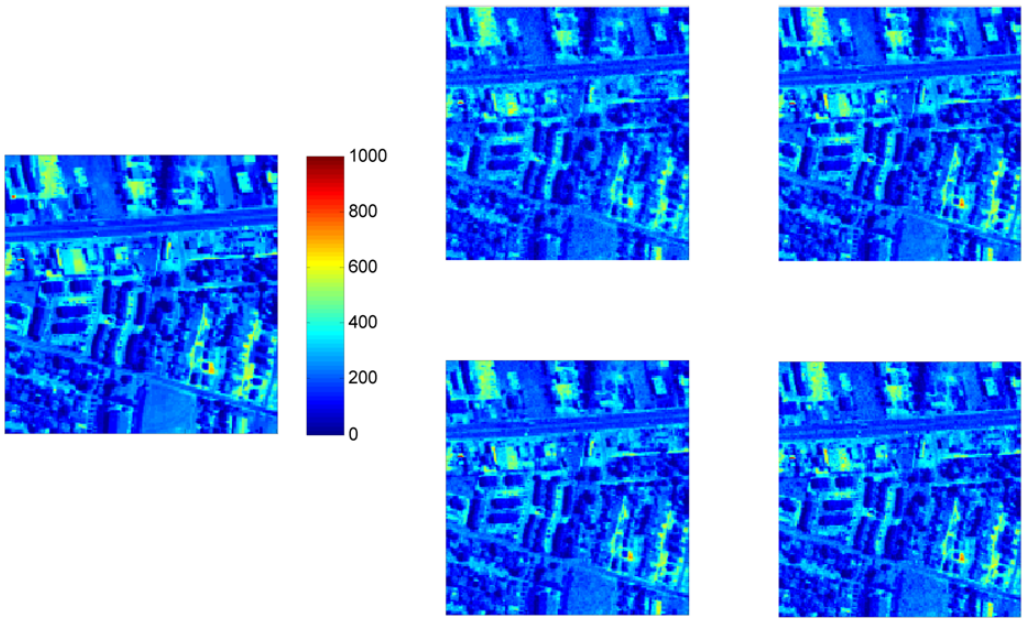


FIGURE 2.2: Recovery of Urban hyperspectral data (normalized reflectance), based upon measuring 2% of the datacube, with voxels selected uniformly at random. The analysis is performed using 2×2 spatial blocks, and all 162 spectral bands. These results are for spectral band 100, although all spectral bands are recovered simultaneously. The subfigures are presented as in Figure 2.1.

was simply not designed to sample. We would also like to make inferences about such missing spectra using a significantly downsampled hyperspectral datacube.

To examine this problem, we consider the Urban hyperspectral data, and select 16 of the 162 spectral bands at random. These 16 spectral bands are removed entirely, and a GP-BPFA analysis is performed using 5% of the remaining voxels, with those selected uniformly at random. The goal, essentially, is to interpolate for the missing 16 spectral bands, in the presence of massive downsampling of the datacube. Note that in this case we *must* use the GP-based formulation (the inferred GP covariance matrix, which is a continuous function of wavelength, may be used to interpolate any spectral band). Illustrative results are shown in Figure 2.6, in which the inferences

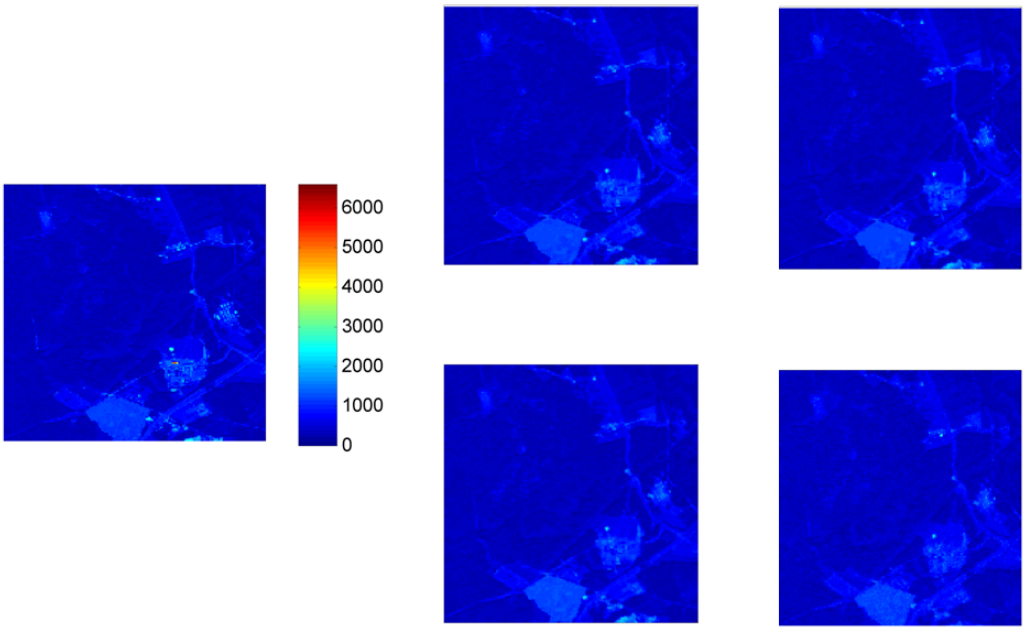


FIGURE 2.3: Recovery of AP Hill hyperspectral data (normalized reflectance), based upon measuring 2% of the datacube, with voxels selected uniformly at random. The analysis is performed using 2×2 spatial blocks, and all 106 spectral bands. These results are for spectral band 20, although all spectral bands are recovered simultaneously. The subfigures are presented as in Figure 2.1.

for 2 of the 16 missing spectral bands are depicted. These results are based upon 4×4 spatial blocks, and the PSNR across all 16 missing bands is 38.3 dB.

2.4.4 Denoising

It is anticipated that in practice hyperspectral data will be noisy. In fact, the Urban and AP Hill data considered above are almost certainly undermined by sensor noise, although in the above reconstructions the evaluation of model performance was based upon the assumption that the original hyperspectral datacubes were noise-free. We examine the noise robustness of the proposed models by now adding i.i.d. Gaussian noise to the data. We initially consider the case for which the noise variance is the

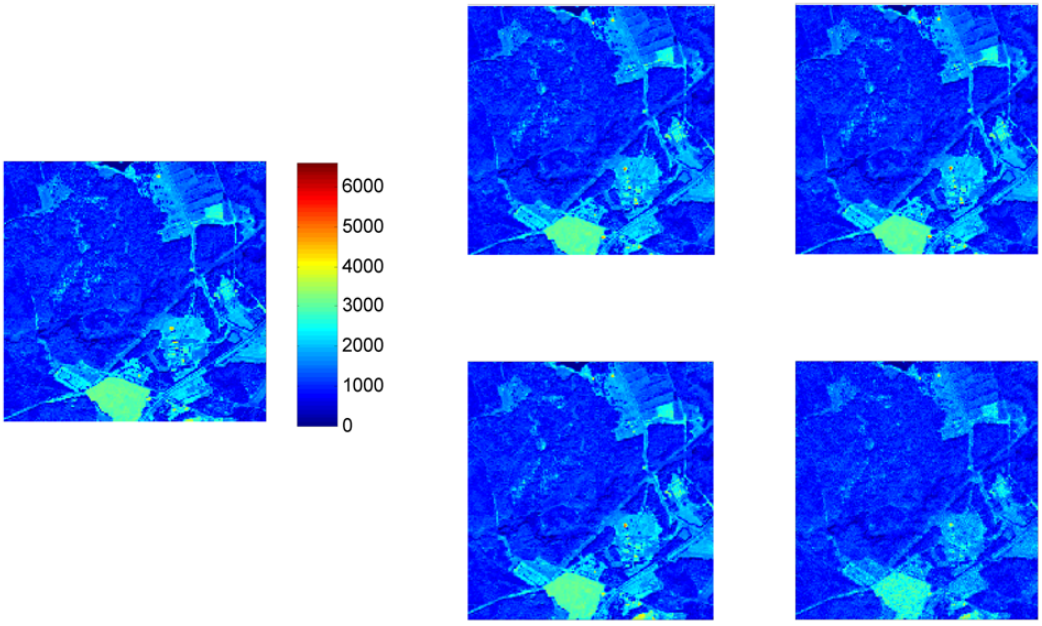


FIGURE 2.4: Recovery of AP Hill hyperspectral data (normalized reflectance), based upon measuring 2% of the datacube, with voxels selected uniformly at random. The analysis is performed using 2×2 spatial blocks, and all 106 spectral bands. These results are for spectral band 100, although all spectral bands are recovered simultaneously. The subfigures are presented as in Figure 2.1.

same at each spectral band, but then we consider the more-realistic case for which the noise variance is wavelength dependent. Additionally, it is possible that the noise variance may vary as a function of spatial position, although that is not considered within these examples. We note, however, that although the proposed models assume an i.i.d Gaussian *prior*, with potentially wavelength-dependent variance, if spatial dependence is also manifested in the actual noise statistics, this should be approximated via the posterior density function on the noise statistics, which need not be Gaussian or stationary.

In Table 2.2 we present results for which the noise standard deviation at each spectral band is either 5, 15, 25, 35 or 50. These results are for the Urban data,

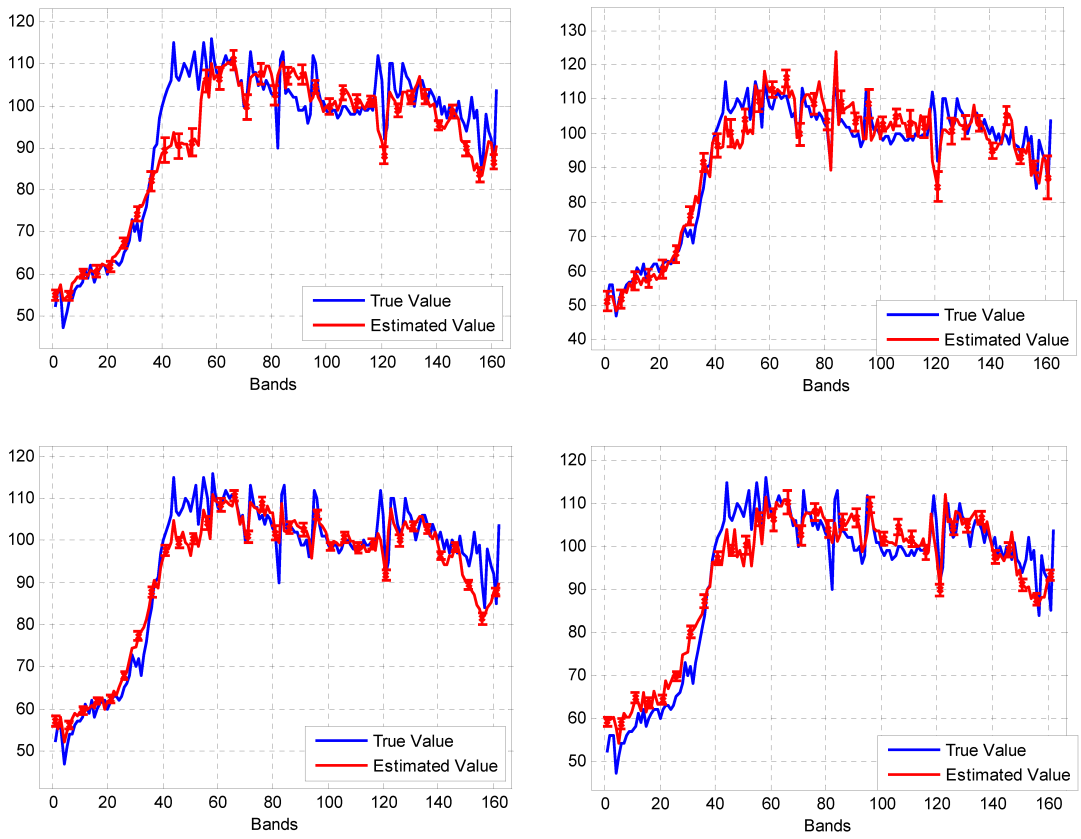


FIGURE 2.5: Representative wavelength-dependent signature (normalized reflectance) at one spatial location, for the Urban hyperspectral data. The top row is based upon recovery using 2×2 spatial patches, and the bottom row uses 4×4 spatial patches. In all cases the same data were used for analysis, based upon selecting 5% of the voxels in the datacube uniformly at random. The left column corresponds to results based upon GP-BPFA, and the right column is BPFA.

with similar results manifested for the AP Hill data, omitted for brevity. All of these results employ the GP, as this was found to be essential for the noisy data. Specifically, the imposition of smoothness in the factor loadings across wavelength plays an important role in mitigating noise. In Table 2.2 we show results based upon observing 2% and 5% of the datacube, uniformly at random, and these results are based upon analyzing 4×4 spatial blocks; the larger spatial blocks, relative to 2×2, also played an important role in enhancing robustness to noise. In these results we present the PSNR value, for GP-BPFA and GP-SFA, as a function of the noise

Table 2.1: Accuracy of recovered datacube (PSNR, in dB), based on observing 2% and 5% of the voxels, selected uniformly at random. Results are shown for the Urban data, considering analysis with 2×2 and 4×4 spatial blocks. Results are shown for the beta-process based factor analysis (FA) model (BPFA) and for the shrinkage-based FA model (SFA), in each case with and without a Gaussian process (GP) employed for the factor loadings. There is no additive noise in this case (processing original datacube).

	2×2, 2%	4×4, 2%	2×2, 5%	4×4, 5%
BPFA	26.4	30.4	38.9	39.4
GP-BPFA	31.4	33.1	39.5	40.2
SFA	30.6	31.3	39.8	41.3
GP-SFA	29.4	33.5	38.0	41.2

standard deviation, and we also present the mean estimate for the noise standard deviation. Note that for standard deviations in excess of 5 the models infer the underlying noise standard deviation with high accuracy.

It is more realistic to expect the noise standard deviation to vary as a function of wavelength. To examine this case we again considered the Urban data, but now the noise standard deviation for each wavelength is drawn from Gamma$(75, 1/3)$, which has 25 for a mean and a 8.3 variance. We again only consider BPFA and SFA with GP, as the GP prior on the factor loadings was found to be essential to achieving noise robustness in this case. The results in Table 2.3 consider 2×2 and 4×4 spatial blocks in the $\boldsymbol{x}_i$, and we consider cases for which 5% to 20% of the datacube is observed, uniformly at random. The results of GP-BPFA and GP-SFA are comparable, but as discussed above the former has significant advantages with regard to computational speed and robustness to setting the truncation K. The 4×4 blocks provide typically 1 dB better PSNR values relative to 2×2, and hence such block sizes are recommended.

In addition to estimating the underlying datacube, and hence denoising, the model also infers the underlying noise statistics. In Figure 2.7 we present the true

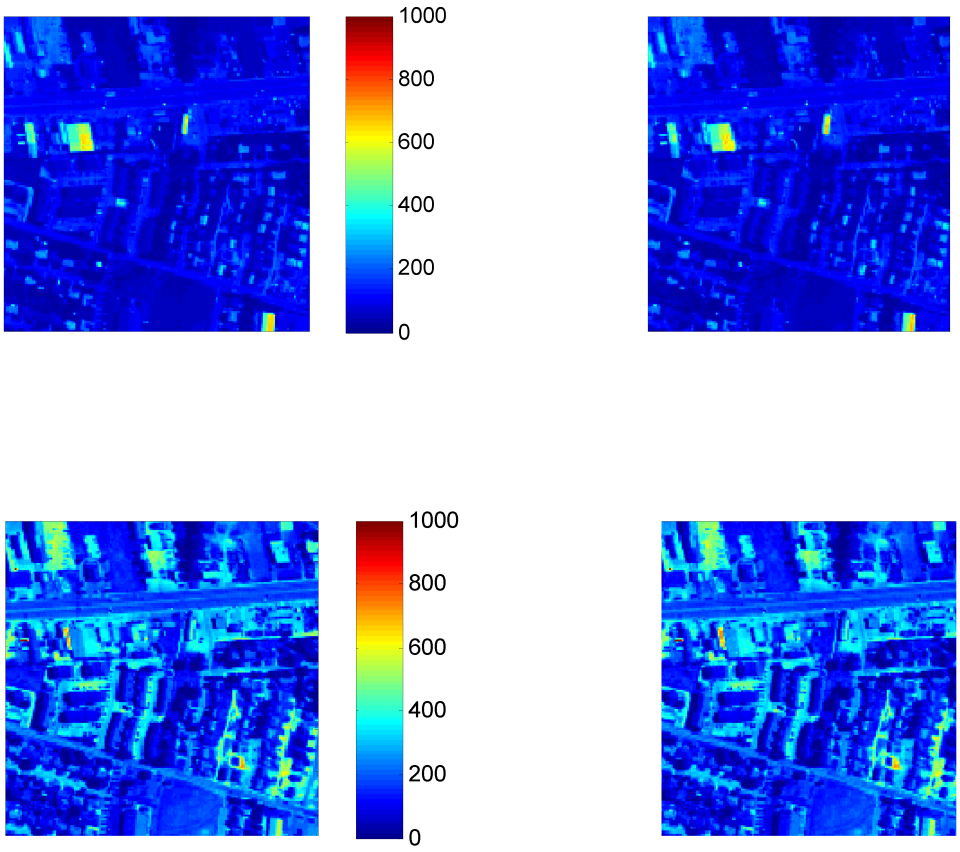


FIGURE 2.6: Recovery of missing spectral bands from the Urban hyperspectral data. Of the 162 spectral bands, the data for 16 of the bands are removed entirely; of the remaining 146 bands, 5% of the voxels are sampled, selected uniformly at random. These figures present example recovery of the images at 2 of the 16 wavelengths for which data were missing entirely, based upon processing with 4×4 spatial blocks, and using the beta-process factor analysis model with a Gaussian process on the factor loadings. The left column corresponds to the original imagery at these two example wavelengths, and the right correspond to the recovered images (PSNR 38.3 dB for the recovered bands). The color scale on the right images is the same as that for the left.

and inferred noise standard deviation for the GP-BPFA; the results for GP-SFA are very similar, and are omitted for brevity. Results are shown for 4×4 blocks. The model infers mean wavelength-dependent noise standard deviations, and via the collection samples we also present standard deviations on the estimates. Both the GP-BPFA and GP-SFA perform this task well, and we underscore that without the GP both models failed in this task with 2% observations.

Table 2.2: Accuracy of recovered datacube (PSNR), based on observing 2% and 5% of the voxels, selected uniformly at random. Results are shown for the Urban data, considering analysis of 4×4 spatial blocks. Results are shown for the beta-process based factor analysis (FA) model (BPFA) and for the shrinkage-based FA model (SFA), in each case with a Gaussian process (GP) employed for the factor loadings. Results are shown for noise standard deviations of 5, 15, 25, 35 and 50, where the PSNR is shown, as well as the inferred noise standard deviation. The same noise standard deviation is employed at all spectral bands. The first number is the inferred noise standard deviation, and the second is the associated PSNR.

	5	15	25	35	50
GP-BPFA, 2%	10.4, 33.0	17.3, 31.9	26.4, 30.1	36.0, 29.9	51.2, 28.7
GP-SFA, 2%	8.5, 33.2	14.3, 31.9	21.8, 30.6	29.7, 29.5	41.5, 28.1
GP-BPFA, 5%	9.2, 39.1	16.8, 36.8	26.0, 34.9	35.7, 33.4	50.5, 31.8
GP-SFA, 5%	6.0, 40.1	14.3, 36.8	23.4, 34.6	32.8, 33.0	46.9, 31.3

Table 2.3: Accuracy of recovered datacube (PSNR), based on observing 2% through 20% of the voxels, selected uniformly at random. Results are shown for the Urban data, considering analysis with 2×2 and 4×4 spatial blocks. Results are shown for the beta-process based factor analysis (FA) model (BPFA) and for the shrinkage-based FA model (SFA), in each case with a Gaussian process (GP) employed for the factor loadings. The noise standard deviation at each spectral band is drawn from Gamma$(75, 1/3)$.

	2×2, 5%	2×2, 10%	2×2, 15%	2×2, 20%	4×4, 5%	4×4, 10%	4×4, 15%	4×4, 20%
GP-BPFA	32.2	35.9	37.6	38.6	33.4	36.7	38.3	39.3
GP-SFA	32.8	35.8	37.3	38.3	34.7	37.4	38.9	39.9

2.4.5 Related Algorithms

Assuming there are only missing voxels and no noise in the data, we can consider the recovery of these missing voxels as a matrix completion problem. Given a HSI image $\mathbf{I} \in \mathbb{R}^{N_x \times N_y \times N_\lambda}$, we can unwrap each $n_x \times n_y \times n_\lambda$ spatial block (with overlapping) into a $P = n_x n_y n_\lambda$ demential vector, and construct a spatial block matrix $\mathbf{X} \in \mathbb{R}^{P \times N}$, where $N = (N_x - n_x + 1)(N_y - n_y + 1)$. For example, when $n_x = n_y = 2$, we have $P = 648$ and $N = 22,201$ for the Urban data.

To complete a matrix with missing data, we may assume that the original fully observed matrix satisfy the low rank assumption and use low rank matrix completion

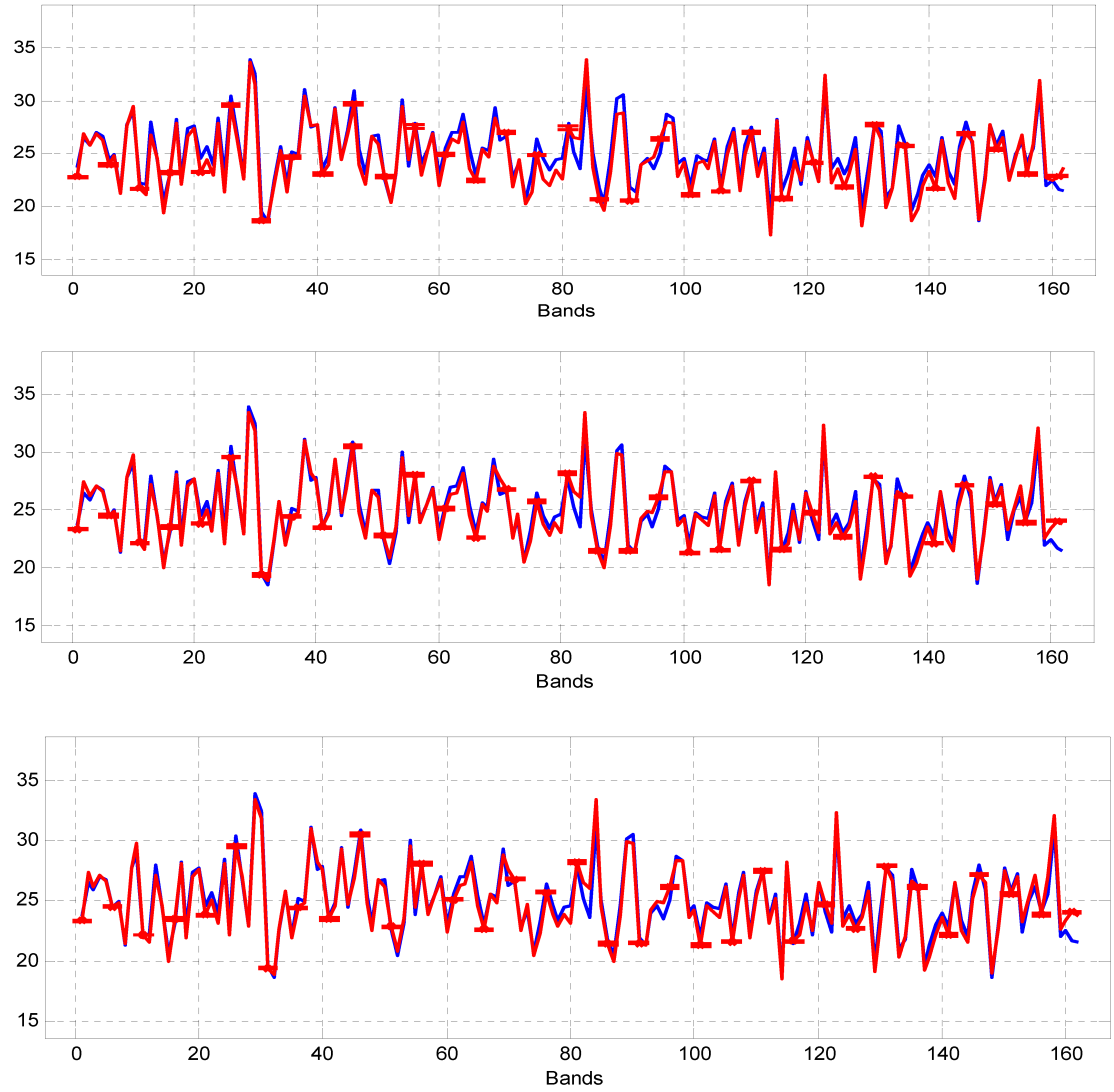


FIGURE 2.7: True (blue) and estimated (red) noise variance, as a function of spectral band, using GP-BPFA. Results are shown for 4×4 spatial patches, and from top to bottom 10%, 15% and 20% of the voxels are observed, selected uniformly at random. Results are for the Urban hyperspectral data. The error bars on the inferred results correspond to one standard deviation, as computed from the posterior density function; only a subset of the error bars are shown, to enhance readability. The noise variance at each spectral band is drawn from Gamma$(75, 1/3)$.

algorithms [JEZ10, LCM10], which minimize the $L2$ error between the observations and estimations under the nuclear norm penalty. We test two state-of-the-art low rank matrix completion algorithms with matlab code available online: the singular

value thresholding (SVT) algorithm proposed in [JEZ10] and the augmented lagrange multiplier (ALM) algorithm proposed in [LCM10]. We find that even after careful parameter tunings, the SVT code[2] fails to yield reasonable results for all the cases we consider for both HSI images. We show the results of ALM[3] for the Urban data in Table 2.4. Although ALM works for both HSI images, almost always for all the cases we have tested so far, it gives an estimation of rank one, which suggests that ALM is essentially substituting weighted average for missing data. Thus it is not surprising that it does not provide good reconstruction and does not improve as either the spatial block size increases or the observed data ratio increases. For comparison, we note that BPFA already provides the PSNRs of 26.4 and 30.4 with 2% observation with 2×2 and 4×4 spatial block sizes, respectively. These results suggest that the low rank assumption (a single linear subspace) is restrictive to recover missing voxels in HSI images, and the sparse representation assumption (union-of-subspaces) of our algorithms is much more realistic. We may give an explanation that the sparse representation allows the discovery of dictionary atoms that are sparsely used. These dictionary atoms may correspond to signals with small singular values under singular value decompositions, and therefore they are likely to be neglected by algorithms based on the methods of thresholding singular values. In this sense, the union of subspace assumption could be a much better choice than the low rank assumption.

Note that the KSVD algorithm is a pioneer to exploit sparse representation for data reconstruction and demonstrates state-of-art results in gray-scale image restorations [EA06]. By introducing weighed rank-one approximation and additional constraints to reduce color artifacts, it has been extended for RGB color images (three spectral bands) [MES08]. We closely follow the extension from gray-scale to color images described in [MES08] to extent KSVD for HSI images. There are several

[2] http://www-stat.stanford.edu/∼candes/svt/

[3] http://perception.csl.uiuc.edu/matrix-rank/Files/inexact_alm_mc.zip

Table 2.4: Accuracy of recovered datacube (PSNR), based on observing 2% through 10% of the voxels, selected uniformly at random. Results are shown for the Urban data, considering analysis with 2×2 and 4×4 spatial blocks. Results are shown for the augmented lagrange multiplier (ALM) algorithm and for the KSVD algorithm.

	2×2, 2%	4×4, 2%	2×2, 5%	4×4, 5%	2×2, 10%	4×4, 10%
ALM	23.3	21.91	23.19	21.93	23.25	21.94
K-SVD	14.58	15.78	17.73	20.26	23.32	25.67

parameters need to be carefully tuned, such as the sparsity level (the number of dictionary elements used by each spatial block vector) and the dictionary size. KSVD also requires good initialization, which we find crucial for HSI images. We initialize the KSVD dictionary elements by randomly sample spatial blocks from the original complete HSI image (this maybe impossible in practise and is only considered for the comparison purpose). The results of KSVD after careful parameter tunings are shown in Table 2.4. We find that KSVD improves as the spatial block sizes increases from 2×2 to 4×4 and the observed data ratio increases, but it performs much worse than the Bayesian algorithms we considered. The reason may be that our Bayesian algorithms are not sensitive to initialization (random initializations work well) and the truncation level of the dictionary size, and the number of dictionary atoms used by each spatial block are automatically inferred.

When there are noise in the HSI images, the low-rank assumption based SVT and ALM and the sparse assumption based KSVD will be even less attractive compared to the proposed Bayesian algorithms, since they usually require the knowledge of noise variance while the proposed Bayesian algorithms automatically infer the noise level. Furthermore, when there are entire spectral bands missing, GP-BPFA and GP-SFA could produce good estimation while SVT, ALM and KSVD are all guaranteed to fail.

2.5 Conclusions

Sparsity is playing an increasing role in many image processing problems [AEB06, EA06, MES08, MBPS09, MBP$^+$08, MSE08, MBP$^+$09, RPCL06]. In the analysis of hyperspectral imagery, researchers have used sparsity for endmember research [ZG07]. In this paper we employed sparsity in a manner analogous to that utilized in previous endmember research, albeit here in a Bayesian manner. However, unlike in previous endmember studies, which were based on the spectral signature alone, in the analysis considered here the dictionary elements (analogous to endmembers) are learned while taking into account both spatial and spectral information.

A unique aspect of the work presented here is that rather than analyzing the entire datacube directly, we have processed a significantly downsampled version. Specifically, we have performed the analysis based on observing a small fraction of the voxels, selected uniformly at random. It was demonstrated that one may accurately recover the missing data, even in the presence of substantial wavelength-dependent noise.

There has been very little previous research in which the potential to massively down-sample a hyperspectral datacube has been considered. In addition, the manner in which that analysis has been performed here is also unique. Specifically, we have considered a fully Bayesian formulation, with previous techniques (applied to grey-scale or RGB imagery) based upon optimization approaches [AEB06, EA06, MES08, MBPS09, MBP$^+$08, MSE08, MBP$^+$09]. The Bayesian analysis yields "error bars" on all model parameters, of interest when one may desire a measure of confidence in the inferred missing data. Additionally, the Bayesian approach is well suited to analysis of noisy data, particularly when the noise statistics (*e.g.*, variance) is unknown and may be a function of wavelength.

In the context of denoising, particularly with wavelength-dependent noise vari-

ance, it has been found that imposition of smoothness on the factor loadings (as a function of wavelength) is critical to achieving accurate results. A Gaussian process [RW06] has been used to impose smoothness on the factor loadings, this having not been considered previously, for simpler image-processing tasks based on grey-scale or RGB imagery [ZCP+09] (where the number of wavelengths is much less than that in hyperspectral data, and hence such smoothness constraints are unnecessary).

The Bayesian analysis has been performed using two constructions, one based upon use of shrinkage priors and the other based on a beta-Bernoulli construction. The former is related to previous research on Lasso [Tib94], while the latter is related to the Indian buffet process [GG05]. The shrinkage construction has the advantage of close relationships with previous optimization-based approaches, with that linkage made explicit. It was found that the shrinkage and beta-Bernoulli approaches yielded similar results, with the latter much less sensitive to the truncation level on the number of factors. However, the beta-Bernoulli construction is significantly more efficient computationally, and therefore this is deemed to be the favored approach.

There are several directions of interest for future research. First, in all examples the analysis has been employed with no *a priori* training data. Specifically, learning of the dictionary and of the missing values has been performed only based upon the (significantly downsampled) data under test. While this is a good illustration of the power of the models, in practice one would expect to have available a database of potential signatures (not necessarily complete, but still providing useful prior information). It is of interest to combine such prior knowledge with the *in situ* dictionary-learning approach developed here. Imposition of such prior knowledge is anticipated to substantially improve modeling performance.

A second clear direction of future research concerns examination of material classification based upon hyperspectral datacubes recovered from massively downsampled measurements. This line of research is critical, as the principal objective of hyper-

spectral measurements concerns material characterization. Based upon the quality of the recovered data, as discussed in this paper, it is anticipated that high-quality material characterization will be achieved. Preliminary research in this direction is encouraging [CXG+10].

3

Bayesian modeling of temporal properties of infectious disease in a college student population

3.1 Introduction

There has been significant interest in the analysis of community-to-individual and individual-to-individual transfer of virus [BKG$^+$11, CCV$^+$04, CO07, JJVF12, OBB$^+$00, YHD$^+$10]. Many of these studies have been concerned with infection transfer in households [CCV$^+$04, CO07, OBB$^+$00], in confined spaces like elementary schools [YHD$^+$10], as well as transfer among domestic animal populations [JJVF12]. The modeling may take different forms, depending upon the data considered and questions being asked. There has been a significant interest in influenza, and in such studies the interest is typically on influenza-like illness (ILI). For example, based on the incidences of ILI, one may be interested in the analysis of large-scale dynamics of epidemic propagation [BKG$^+$11, Het00, DPH12], where in this case the data may be counts of space-time ILI events. There are other studies for which the modeling is performed at or near the level of the symptoms or biomarkers, which are noisy and often imperfect [JJVF12]. The studies are complicated by missing and incomplete

data, and an unknown number of competing pathogens [YHD$^+$10].

Infection dynamics are complex, and therefore the power and flexibility of Bayesian models are attractive [BKG$^+$11, CCV$^+$04, CO07, YHD$^+$10]; we employ such a modeling approach in this study. Most of these models assume that a given individual is in a particular state of health, such as susceptible (S), exposed (E), infective (I) and recovered (R); an individual in state I is infectious, in that they are capable of transmitting the virus. The sequence of states considered in such a model defines its character, for example susceptible-infective-recovered (SIR) models are widely considered [Het00], and once in the R state individuals are often assumed removed from the population from the standpoint of infection transfer (because of acquired immunity, or because of death; in some settings R represents "removed" rather than recovered).

When dealing with many competing pathogens [YHD$^+$10], such as distinct viruses characteristic of common/typical colds, even after an individual recovers from one virus, they may soon be susceptible to another. That is the case considered in this paper, which motivates a SIS model. We effectively ignore the short time period that may exist between state R and the return to S; during this short time period there may be some cross-immunity between pathogens [YHD$^+$10]. Additionally, we do not explicitly model the distinction between states E and I, as these states are not distinguishable with the data considered. With the observed data under consideration, only symptoms allow distinction of states, and therefore we assume state S is one characterized by no or minimal symptoms, and state I is one in which symptoms are observed (there are complications with this definition of state I, as discussed below).

Propagation of influenza and influenza-like viruses has been considered within school settings, for example the Pittsburgh Influenza Prevention Project (PIPP) considered data from ten public elementary schools in the city of Pittsburgh [SCS$^+$11].

Studies have also been conducted concerning influenza propagation in families, including data from France from over 300 families [CSL+02]. In analyzing such data [YHD+10, CCV+04] state-based models like those discussed above are typically employed, and two forms of dynamics are often considered for the probability of transitioning from S to E, or directly from S to I. One is based upon community-to-person contacts, associated with interactions outside close contacts, and the other is associated with person-to-person transfer among the close contacts. The person-to-person transfer is employed to model interactions between individuals in confined or intimate settings, such as the aforementioned elementary schools or households. In these settings the symptoms themselves are not modeled. Rather, it is assumed that some other mechanism is available to determine an individuals health state, and that this is done separately. For example, in [CCV+04] clinical influenza was defined as the presence of fever or feverishness, or at least two of the following signs: sore throat, headache, stiffness or myalgias, fatigue, cough, nasal congestion or rhinorrhea or sneezing. Similarly, in [BKG+11], the data modeled were defined cases of illness, with symptoms themselves not modeled.

There are potential pitfalls associated with attempting to model person-to-person transfer, when this mechanism is tied to symptoms. It has been observed that individuals may be infected with a virus but display no symptoms [PM09]; additionally, for those who do ultimately have symptoms, pathogen transfer may occur in the presymptomatic state [PM09]. In other words, even though the absence of symptoms from an individual may indicate that she is in state S, she in fact may be in state I, and she may transfer/shed virus. The effectiveness of pathogen transfer from asymptomatic shedders is not well understood. Additionally, the data of interest for person-to-person transfer may be incomplete, in that it only accounts for a subset of close contacts. In [CCV+04] the authors modeled infection transfer within elementary schools, but not within the households of the students; in [YHD+10]

the authors modeled person-to-person transfer within households, but not among other close contacts outside the households. For these reasons, and because of the characteristics of the data considered here (detailed in Section 3.2), we only model community-to-person transfer. This allows for the possible transfer of pathogen from an asymptomatic (but virus shedding) member of the community to a member of our study. However, an asymptomatic member of our study will still be deemed in state S by our model, based on symptoms, even though they may be in state I and asymptomatically shedding virus; this issue is examined in detail when presenting results.

Within the proposed SIS model, we assume that the probability of transferring from state S to state I is time-dependent. Further, we assume that different individuals have distinct degrees of susceptibility to common viruses, and this is modeled as well (*i.e.*, there is a person-dependent character to the degree of susceptibility, and hence to the characteristics of state S). A unique aspect of this study is that the modeling is performed directly at the level of observed symptoms, rather than using pre-specified means of defining whether one is in the S or I state. Specifically, in most of the above studies the state S/I of the individual was assumed observed, and the goal was to infer the statistics of the state dynamics (*e.g.*, the probability of transiting from S to I, and the duration of being in state I). In this study the symptoms are the observed data, and the state S/I is treated as being *latent*, and to be inferred. As discussed in Section 3.2, we also have access to (imperfect) labels on the health of the individual at a given time, based upon real-time (RT) polymerase chain reaction (PCR) testing and gene-expression data. We compare the model-inferred state of the individual to the state based upon RT PCR and gene expression. This comparison provides insights into such mechanisms as the aforementioned asymptomatic virus shedders, as well as individuals who are symptomatic but not in state I as defined by RT PCR.

The time dependence in the probability of transiting from S to I captures time variation in the viruses present at a given time, as well as time-dependent dynamics of human interaction (*e.g.*, the mixing of a new set of people may increase virus transfer [SWZS11, Kak07]). A unique characteristic of the data considered here is that it is collected daily, for an entire university academic year; further, we have data from three full academic years. By comparison, the data in [CCV$^+$04] only existed for 15 days after a household index case. The long time scale, and the daily sampling, introduce interesting phenomena that have not been investigated previously, to our knowledge (in [BKG$^+$11] weekly sampling of symptoms was performed). Specifically, how one reports symptoms may be linked to their mood, which may vary with the day of the week. For example, it has been demonstrated that the way in which individuals rate music is linked to the day (and even time) of reporting [JT11, SC12]. Since we are analyzing symptom data, we must consider biases in data reporting that may occur based upon the day of reporting. The degree of missingness tends to also be linked to the day of the week. This therefore motivates employing a semi-periodic, or weekly effect in the probability of transiting from state S to I. This is a novel characteristic of the model developed here, in which we generalize the use of models that employ seasonal terms [WH89] (here they become weekly, and they are modeling a latent process).

Covariates may be available that can be employed to impact the probability of transiting from state S to I. Given the very long longitudinal length of our data, we may consider new forms of covariates, becoming available from a web-centric world. We consider the Google Flu Trends data [GMP$^+$09]. These covariates are constituted by region-specific web searches of words linked to ILI, and specifically here we employ the Google Flu Trends for Durham, NC, the city in which Duke University resides. Other recent statistical analyses have modeled the space-time properties of such Google data alone [DLP12, FD11], but here we are focused on

observed symptoms and the time-dependent Google data is a covariate. In [DLP12] the focus is on modeling an epidemic like influenza, and therefore they employ a susceptible-exposed-infected-recovered (SEIR) model; once in the recovered state, an individual is effectively removed from the pandemic dynamics because of immunity. Here we are interested in modeling long-term illness dynamics from common viruses (producng ILI), in addition to influenza, and therefore removing an individual from the population upon recovery is not appropriate (in the data we observe some people with repeated ILI). This may motivate a SEIS model, but for simplicity we consider a SIS model [Het00].

3.2 Motivating Data and Questions

3.2.1 Self-reported daily symptom data

Self-reported symptom-score data were collected from undergraduate students at Duke University, following guidelines specified by the Duke Institutional Review Board (IRB). Data were collected daily during the 2009-2010, 2010-2011 and 2011-2012 academic school years; in each year, data were collected from the beginning of September until May, using a web-based tool. For each of these collection periods, respectively 246, 378 and 242 students participated. The total number of days in which data were recorded were respectively 222, 214, and 227 over each collection period. The 2009-2010 collection period coincided with the novel H1N1 pandemic [TKR09].

The students' reported symptoms were routinely monitored by Duke University health professionals. When a student was deemed – from reported symptoms – to likely be sick with an infectious disease (*e.g.*, virus), the student was contacted and nasal and blood samples were taken. These are termed index cases. Further, each student provided a list of close contacts (other students they interacted with frequently). Blood samples were then collected daily for a week on these close contacts,

with the hope that we may observe the transfer of infectious disease (and to analyze that in the context of the blood samples).

3.2.2 Virus identification and gene-expression data

For the students from whom samples were collected, RT PCR testing was done for a set of viruses. The particular viruses for which a RT PCR test was available were: Rhinovirus, Coxsackie, Echovirus, Coronavirus (229E, HKU1, NL63, OC43), Parainfluenzavirus (1, 2, 3, 4), RSV A/B, Influenza A, Influenza B, Metapneumovirus (A & B), and Adenovirus E & B, and Bocavirus (platform used: Qiagen ResPlex II V2.0). Therefore, *if* one of these viruses was responsible for the student's illness, and *if* the virus was present in the collected sample, and *if* the RT PCR test worked properly, then the virus type responsible for illness can be detected. However, a negative RT PCR result does not necessarily imply that the student is not sick with a virus, as there may have been a poor sample (in which markers of the virus were not present), a non-tested (within the RT PCR library) virus may have been responsible for illness, and the RT PCR test is itself imperfect. In the context of this study, across the three years, 897 viral etiology results based on RT PCR were constituted.

In addition to the aforementioned RT PCR tests, we also used the available blood samples to perform gene-expression analysis. Let $\boldsymbol{x}_q \in \mathbb{R}^G$ represent the expression data for subject q, for G genes. We performed sparse Bayesian factor analysis on the set of data $\mathbf{X} \in \mathbb{R}^{G \times Q}$, where column q of $\mathbf{X}$ corresponds to $\boldsymbol{x}_q$. Details on the factor analysis method may be found in [CCL$^+$08, CCP$^+$10, CZW$^+$11]. In [ZCL$^+$09, CCP$^+$10, CZW$^+$11] it was demonstrated that one of the factors in such an analysis may be linked to the host response to virus, and in multiple experiments this signature has been found invariant to the particular type of virus studied. Therefore, for the Q samples defining $\mathbf{X}$, we used this factor analysis to define which of these subjects appear to be infected by a virus (by the presence of an elevated form

of a specific factor [ZCL$^+$09, CCP$^+$10, CZW$^+$11]). We emphasize that this test is also imperfect, but it provides more generality than the RT PCR tests, which are constrained to specific types of viruses. In these experiments $Q = 34$ and $G = 22277$.

Based upon the RT PCR and gene-expression tests outlined above, we can (imperfectly) label each of the subjects for whom samples were collected as being sick with a virus or not, at a given point in time; these tests are used to assess the quality of the model developed in Sections 3.3 and 3.4 to detect sickness based on the symptom scores alone (with state of health defined by the inferred latent state, S or I). It is important to emphasize that the labels we will use to assess performance are imperfect, in the sense that the RT PCR and gene-expression tests are only testing for the presence of a virus (and even these tests are not perfect). It is possible that an individual may be sick for another reason (*e.g.*, due to allergies or bacteria); in this case the virus-driven labels may indicate that the student is not sick, while the symptoms indicate otherwise (our symptom-based declaration of health or sick may indicate state I, while the virus-driven labels may indicate S). These issues will be revisited when presenting results.

3.2.3 Impact of form of data on the developed model

This paper is principally directed toward analyzing the self-reported symptom-score data, with a focus on community-to-person transmission of pathogens. In each student dorm (living facility), roughly 10% to 20% of the students participated in the study, and therefore the close contacts are very sparsely sampled. Further, many of the students spend most of their time outside the dorm, and interact infrequently with many members of the same dorm. It was therefore deemed inappropriate to try to model person-to-person pathogen transfer. We also considered developing dorm-dependent models for pathogen transfer, but the dynamics across the different dorms (*e.g.*, fraction of students sick at any given time) did not vary substantially,

and therefore it was deemed most appropriate to develop a single community-to-person model for all students who participated in the study.

However, as detailed below, the close contacts constitute a separate set of data (with associated "truth" for the presence/absence of virus), within the context of the imperfections of the RT PCR and gene-expression data. We therefore use these data for model testing.

3.2.4 Questions to be examined in this study

- We have access to data over three academic years, with one year corresponding to the presence of novel H1N1 virus. During that year there was heightened awareness on campus about non-pharmacological ways to reduce virus transmission, with many highly visible reminders (*e.g.*, students were prominently reminded about hand washing, use of disinfectants, not touching eyes and nose, etc.). Disinfectant soap was widely accessible throughout the campus, at locations in which people congregate. We wish to examine how this heightened awareness affected the time-dependent hazard of community-to-person transmission of pathogens, relative to the other two years of the study, in which pathogen transmission was far less of a focus.

- The students who participated in this study primarily lived on a separate campus dedicated for first-year students. Therefore, most of the students were Freshman, and at the beginning of the academic year most of these students were coming together, and living in close proximity (in dorms), for the first time. We wish to examine the impact of this new mixing of people on the time-dependent hazard of community-to-person transmission of pathogens.

- The Duke University campus resides within the surrounding city of Durham, NC. We wish to examine how the time-dependent hazard of community-to-

person transmission of pathogens of Duke students relates to such metrics as Google Flu Trends. Specifically, we wish to examine the extent to which Google Flu Trends for Durham, NC predicts the hazard of pathogen transmission on the Duke campus.

- We have access to which dorm room each student lived in. Based upon the symptom scores, the model predicts whether each student is infected at a given time. While we do not explicitly model person-to-person transmission within the model (for reasons stated above), we may use model predictions on the state of health to examine whether someone getting infected at time t within a given dorm raises (or lowers) the incidence of infection of other students in the study who lived in the same dorm (and more specifically, on the same dorm floor). For example, an infected neighbor in a dorm may *heighten* awareness of the danger of pathogen transfer, yielding phenomenon like that associated with the exposure to novel H1N1 (heightened awareness, and hence precautions). We examine this issue in detail, as a function of the type of virus associated with each index case (with virus type imperfectly determined via RT PCR).

- We examine and analyze real-world characteristics of daily self-reported symptom data. This includes day-of-the-week dependent phenomenon in the data (weekly, semi-periodic effects), and connections to data missingness.

- We examine the utility of using symptoms alone for classification of the latent state S/I, with comparisons to RT PCR. This is of clinical relevance, as clinicians typically make a diagnosis based directly on symptoms. We also examine the presence of non-symptomatic individuals who are shedding the virus. Further, we examine cases for which symptoms are clear, but extensive RT PCR testing is negative.

3.3 Basic Modeling Setup

3.3.1 Observed symptoms and the latent state of health

Assume access to self-reported data from N individuals, provided daily over multiple months. The data correspond to the strength of various infectious-disease-related symptoms, reported separately by each of the N students. Eight symptoms are recorded: nasal discharge, nasal congestion, sneezing, cough, malaise, throat discomfort, fever and headache. Each of the eight symptoms is reported on an ordinal scale, from 0 to 4, with 0 being no symptoms, and 4 "maximum" symptoms. Before the study each of the students is instructed on how to connect perceived symptoms to this scale. Nevertheless, there is clearly subjectivity to the mapping from perceived symptoms to ordinal data, and this subjectivity should be accounted for in the statistical analysis. We note that such subjectivity is always present when individuals report symptom severity to a doctor or nurse.

Let $\boldsymbol{y}_{nt} \in \{0, \ldots, M\}^J$ represent the J symptom scores reported by individual $n \in \{1, \ldots, N\}$ on day t, where for our study $J = 8$ and $M = 4$; we use generalized notation because the basic modeling strategy may be applied to other types of related data. It is assumed that, at a given time, individual n is either in an infective state I or in susceptible state S. When in state S, the student is not currently sick from a virus, and therefore does not display ILI symptoms; however, the student is assumed susceptible to virus infection. When in state S, different individuals may have distinct levels of susceptibility to virus-borne illness, and this is accounted for in the model. We have tied state I to symptoms, as is common [BKG$^+$11, CCV$^+$04, CO07, JJVF12, OBB$^+$00, YHD$^+$10]. However, there are asymptomatic individuals who shed virus [PM09] and hence are in the infective state I; these individuals are identified and discussed when presenting results.

We employ an ordinal probit model to link the J observed symptoms to the latent

state. Specifically, consider $\boldsymbol{r}_{it} \in \mathbb{R}^J$, drawn conditioned on the latent state as

$$\boldsymbol{r}_{nt}|z_{nt} \sim \mathcal{N}(\boldsymbol{\mu}_{z_{nt}}, \boldsymbol{\Sigma}_{z_{nt}}^{-1}) \tag{3.1}$$

where $z_{nt} = S$ or $z_{nt} = I$. Let r_{njt} represent the jth component (symptom) of $\boldsymbol{r}_{nt}$, and y_{njt} similarly represent the jth component of $\boldsymbol{y}_{nt}$. The mapping from real r_{njt} to ordinal y_{njt} is manifested via a traditional probit model as

$$y_{njt} = m \;\text{ if }\; \tau_{j,m-1} < r_{njt} \leqslant \tau_{j,m} \tag{3.2}$$

where each $\tau_{j,m} \in \mathbb{R}$, $\tau_{j,m-1} < \tau_{j,m}$, $\tau_{j,-1} = -\infty$, $\tau_{j,0} = 0$, and $\tau_{j,M} = \infty$. We wish to infer $\{\tau_{j,1}, \ldots, \tau_{j,M-1}\}$, with this performed by considering an improper uniform prior on $\tau_{j,1} < \cdots < \tau_{j,M-1}$ [OD04]. Uniform improper priors result in proper posterior distributions under mild conditions, as detailed in [OD04], yielding practically useful sufficient conditions that are met in our study. As discussed in the Appendix A, for identifiability purposes the covariance matrices $\boldsymbol{\Sigma}_S^{-1}$ and $\boldsymbol{\Sigma}_I^{-1}$ are restricted to correspond to correlation matrices [CD01], with diagonal elements all equal to one.

Note that we assume that the statistics of the symptoms for the infective individuals, characterized by $\mathcal{N}(\boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I^{-1})$, are independent of the length of time in which the individual has been in state I. This is a modeling simplification, and one may also link the symptom statistics to the length of time the subject has been in the infective state. The variability in the symptom scores within a given state, characterized by $\boldsymbol{\mu}_I$ and $\boldsymbol{\Sigma}_I$, account for variability in how a given individual maps perceived symptom strength to ordinal values. Further, $\boldsymbol{\Sigma}_I$ accounts for variability in symptom strength across an extended period of infection (typically from weak, to strong, and back to weak symptoms over the period of infection).

3.3.2 Semi-Markov latent-state dynamics

The probability of individual n transiting from a state of susceptibility at time $t-1$, $z_{n,t-1} = S$, to a state of illness/infection at time t, $z_{n,t} = I$, is modeled as

$$p(z_{n,t} = I | z_{n,t-1} = S) = \Phi(\gamma_{nt}) \tag{3.3}$$

where $\Phi(x) = \int_{-\infty}^{x} d\eta \mathcal{N}(\eta; 0, 1)$ is a cumulative distribution function, with $\mathcal{N}(\eta; 0, 1)$ a normal distribution function for variable η, characterized by zero mean and unit variance (probit transition statistics). The model of the time-evolving variable $\gamma_{nt} \in \mathbb{R}$ is discussed in Section 3.3.3. Related forms of time and covariate dependent probabilities of community-to-person transfer has been considered in [YHD$^+$10]; however, in that work, and much of the literature, the state S or I was assumed observed, where here the state is latent and is to be inferred upon the observed symptoms.

If individual n transits from $z_{n,t-1} = S$ to $z_{n,t} = I$, then it is assumed that $z_{n,t+d} = I$ for $0 \leqslant d \leqslant D_{nt}$, where D_{nt} is a random variable defining the number of days of infection. We employ the model

$$D_{nt} = c + \hat{D}_{nt} \ , \quad \hat{D}_{nt} \sim \text{Pois}(\lambda_n) \tag{3.4}$$

where $c > 0$ is a minimum number of days infected, and the rate parameter λ_n is assumed drawn from a gamma distribution. We discuss setting c when presenting experimental results; the imposition of a lower bound c on the number of days of being infected (*i.e.*, in the state I) helps distinguish isolated days when one may not feel well, for various reasons, from actual extended periods of infection.

In [BKG$^+$11] the length of time D_{nt} in state I was a real random variable, and was drawn from a gamma distribution. Here we observe discrete temporal data (days), and employ Poisson random variables for the length of time in state S; the lower bound c assures that we are not undermined by draws from Pois(λ_n) that could be equal to zero.

3.3.3 Modeling the time-dependent probability of becoming infected

The time-evolving parameter γ_{nt}, in concert with a probit link function, defines the probability with which one transits from a susceptible to infective state. We model this time-evolving parameter via four terms:

$$\gamma_{nt} = a_n + \sum_{i=1}^{3} \gamma_t^{(i)} \tag{3.5}$$

with $\gamma_t^{(1)}$ modeling the general trend within the population to become infected, $\gamma_t^{(2)}$ is associated with periodic (weekly) effects characterizing unique aspects of the day of the week, and $\gamma_t^{(3)}$ is a regression term. Concerning $\gamma_t^{(3)}$, we specifically perform regression to the Google flu-trends data. Note that $\{\gamma_t^{(i)}\}_{i=1,3}$ are independent of the individual index n, and are therefore shared across the population. The term $a_n \in \mathbb{R}$ is an individual-dependent tendency to get infected, which we place a normal prior on (when a_n is large and positive the nth individual has a heightened susceptibility toward illness, with the opposite true when a_n is negative with large magnitude).

The model in [YHD+10] also imposed covariate-dependent state-transition statistics. However, the length of the data considered in that study, and in most of the literature, precluded the need to consider semi-periodic terms. Further, most such models are not performed at the level of symptoms, and therefore they do not have to address semi-periodic missing data phenomenon, and other characteristics of the symptoms.

General-trend term

An autoregressive model is employed for $\gamma_t^{(1)}$:

$$\gamma_t^{(1)} \sim \mathcal{N}(\omega \gamma_{t-1}^{(1)}, \beta^{-1}) \tag{3.6}$$

where a gamma prior is placed on β and $\omega \in (0, 1)$ is drawn from a truncated normal distribution, $\omega \sim \mathcal{N}_{(0,1)}(\mu_\omega, \sigma_\omega)$. This imposes that the time dependence of the

general trend toward illness varies smoothly.

Weekly or periodic term

The observed data are characterized by clear dependencies on the day of the week on which symptoms are reported. The day of the week may impact general feelings of well being (Monday vs. Friday), and certain portions of the week may be characterized by heightened student workload, stress, and lack of sleep/exercise. A seven-day semi-periodic term is therefore employed to model $\gamma_t^{(2)}$. This term builds upon modeling strategies discussed in [WH89] (Chapter 8.6); for completeness, we here provide some details.

Using notation from [WH89], the unique terms of a periodic function may be represented in terms of φ_j, for $j = 0, \ldots, p-1$, where p is the period of the weekly/repeating term (for our problem $p = 7$, for the days of the week). Basic Fourier analysis dictates that where the components of the two-dimensional vector $\boldsymbol{\epsilon}_r(j)$ are drawn $\boldsymbol{\epsilon}_r(j) \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\theta r}^{-1})$, and $\nu_r(j) \sim \mathcal{N}(0, \zeta_{Sr}^{-1})$, with a gamma prior placed on ζ_{Sr} and a Wishart prior on $\boldsymbol{\Sigma}_{\theta r}$. The term $\boldsymbol{\epsilon}_r(j)$ models noise in the Fourier components over a given time period, and $\nu_r(j)$ represents measurement noise.

A prior is placed on $\boldsymbol{\theta}_r(0)$, corresponding to the Fourier components over the first week of data, and then (**??**) is repeated cyclically over the multiple weeks, through sequential draws of $\{\nu_r(j)\}$ and $\{\boldsymbol{\epsilon}_r(j)\}$. Note that with the zero mean priors on $\{\nu_r(j)\}$ and $\{\boldsymbol{\epsilon}_r(j)\}$, conditioned on $\boldsymbol{\theta}_r(0)$, the expectation of (**??**) corresponds to (**??**). A zero-mean normal prior is placed on $\boldsymbol{\theta}_r(0)$, for each r. With $S_r(j)$ so drawn, one may superpose the Fourier components to constitute $\gamma_t^{(2)}$; for the weekly data under consideration, there are $h = 3$ Fourier components, in addition to the mean a_0.

Regression term

Assume that we have access to a time-dependent covariate f_t, which in our problem corresponds to the Google Flu Trends [GMP+09] data for the region in which the individuals under study reside. The regression term is modeled as

$$\gamma_t^{(3)} \sim \mathcal{N}(\xi f_t, \alpha_f^{-1}) \tag{3.7}$$

where a zero-mean normal prior is placed on ξ and a gamma prior is placed on α_f.

3.4 Additional Model Considerations

In the previous section it was assumed that the parameters λ_n and a_n were drawn i.i.d., with the former controlling the length of time individual n tends to be in an infective state, and the latter controlling the tendency of individual n to get infected. The parameter a_n has the impact of controlling the degree to which one is susceptible to virus, and hence to transition from state S to I (large a_n implies higher susceptibility).

It is anticipated that individuals may cluster in terms of their (*e.g.*, genetic or behavioral) tendency to get infected, and in the length with which they stay infected. It is desirable to account for this in the model (it allows sharing of statistical strength between individuals). Additionally, for the dataset that motivates this paper, we have access to the residence location of each student, and therefore it is possible to use this spatial information as a covariate. For example, one may consider the spatial location of each student when modeling the time-dependent tendency to get infected, via including spatial information in $\gamma_t^{(1)}$, for example. Other modeling issues discussed below include consideration of missing data, and the joint modeling of data from multiple years.

3.4.1 Clustering tendency toward infection, and length of infection

A natural means of clustering λ_n and a_n is to employ a Dirichlet process, with which the number of clusters may be inferred nonparametrically. Specifically, we draw

$$\lambda_n \sim G_\lambda \ , \quad G_\lambda \sim \mathrm{DP}(\alpha_{0\lambda} G_{0\lambda}) \tag{3.8}$$

$$a_n \sim G_a \ , \quad G_a \sim \mathrm{DP}(\alpha_{0a} G_{0a}) \tag{3.9}$$

where the base measures $G_{0\lambda}$ and G_{0a} correspond, respectively, to gamma and normal distributions. Gamma priors are placed on the DP parameters $\alpha_{0\lambda}$ and α_{0a}.

3.4.2 Spatial covariates

As indicated above, for the motivating data, we have knowledge of the residence location (dorm room) of each individual (student), and therefore it is possible to exploit spatial information when modeling the general trend toward being infected, reflected in $\gamma_t^{(1)}$. One could also consider utilizing spatial information when modeling the weekly (semi-periodic) term $\gamma_t^{(2)}$ and the regression term $\gamma_t^{(3)}$, but spatial dependencies for these terms are less well motivated.

In our numerical experiments, we considered assigning a separate $\gamma_t^{(1)}$ for each floor of a dorm. In this case all students on a given floor shared the same floor-dependent variant of $\gamma_t^{(1)}$ (*i.e.*, rather than sharing a single $\gamma_t^{(1)}$ across all students, a separate such term was employed for each door floor). We also considered assigning a separate term of the form $\gamma_t^{(1)}$ to each dorm (*i.e.*, all residents in a given dorm, independent of floor, shared the same $\gamma_t^{(1)}$). In our experiments, we found that such added modeling complexity did not improve the predictive performance of the model, and in some cases reduced performance (since the students were spatially segregated in these tests, fewer students were associated with a particular floor/dorm-dependent $\gamma_t^{(1)}$, and therefore statistical strength was diffused). There did not appear to be clear situations for which a given dorm or specific dorm floor had a greater tendency

toward health (state S) or sickness (state I) than the general population. A potential reason for this is that students spend a significant portion of their time away from their dorm, in classes and other activities, mixing with the general population.

For these reasons, for the results below we do not explicitly leverage spatial covariates for student dorm rooms within the model. However, when presenting results we will examine some of the inferred parameters in the context of student residency location.

3.4.3 Missing data

There is a substantial quantity of missing data in self-reported studies, and it is anticipated that the missingness is *not* manifested uniformly at random. It is likely that individuals are less likely to pay attention to reporting symptoms when they are feeling well, with greater attention paid during the time of actual illness. If data are missing from individual n on day t, the "observations" are denoted $\boldsymbol{y}_{nt} = \varnothing$. The probability of the null observation in states S is defined as $\eta \in (0, 1)$, and the probability of a null observation in state I is $\rho \in (0, 1)$. We now consider the case of missing data a null observation, and the observation probability for symptom j, individual n and day t is generalized to [YK03b]

$$y_{njt}|S \sim [\eta\delta_{\varnothing} + (1-\eta)\textstyle\sum_{k=0}^{M} p(y_{njt} = k|S)\delta_k] \tag{3.10}$$

$$y_{njt}|I \sim [\rho\delta_{\varnothing} + (1-\rho)\textstyle\sum_{k=0}^{M} p(y_{njt} = k|I)\delta_k] \tag{3.11}$$

where $p(y_{njt} = k|S)$ and $p(y_{njt} = k|I)$ are the observation probabilities from Section 3.3.1 (assuming symptoms are not missing); the symbol δ_k is a unit measure concentrated at the point k. It is assumed that η and ρ are drawn from uniform priors over [0,1].

3.4.4 Modeling multiple years of data

The experiments detailed in Sections 3.2 and 3.5 correspond to (ideally) daily student recording of symptom scores, over an entire academic year; imperfections in this process naturally manifest missing data. Data of this type were collected over three academic years. It is desirable to analyze all of these data jointly, to achieve maximal statistical strength in the results. However, because of the influx of new (freshman) students, and the exit/graduation of others (seniors), the explicit set of students considered on consecutive years is largely distinct. Additionally, each year is characterized (for example) by a distinct respiratory viral illness season, and this must be accounted for when deciding which components of the model to share between multiple years. For example, one of the years during which we collected data corresponded to the presence of an unusual (and potentially dangerous) H1N1 flu, which had characteristics (*e.g.*, time of arrival) distinct from typical flu seasons.

So motivated, in the experiments that follow, the explicit $\{\gamma_t^{(1)}, \gamma_t^{(2)}, \gamma_t^{(3)})\}$ are modeled as being distinct among the three years of data. However, the priors on parameters with which these time-dependent functions are constituted are shared across years. Specifically, for the AR(1) model of $\gamma_t^{(1)}$, the priors for ω and β are shared across the multiple years of data. For $\gamma_t^{(2)}$, the parameters $\boldsymbol{\Sigma}_{\theta r}$ and ζ_{Sr} are shared across years, as is the prior on $\boldsymbol{\theta}_r(0)$. Finally, for $\gamma_t^{(3)}$, the priors for ξ and α_f are shared across the multiple years.

Concerning λ_n and a_n, the DP-drawn priors G_λ and G_a are shared across the multiple years, and therefore the clustering of types of people (by susceptibility toward illness, and length of illness) is performed jointly across the multiple years. Finally, concerning the observed symptoms, the parameters $\{\boldsymbol{\mu}_S, \boldsymbol{\mu}_I, \boldsymbol{\Sigma}_S, \boldsymbol{\Sigma}_I\}$ are shared across the multiple years, as are the ordinal probit cut points $\{\tau_{j,1}, \ldots, \tau_{j,M-1}\}_{j=1,J}$, and η and ρ (for missing data).

3.5 Results

The modeling software was implemented in MATALB™. On a laptop with a 2.7 GHz dual core CPU, each Gibbs iteration takes about 30 seconds, to process all three years of data. We considered 7000 MCMC samples, with the first 2000 discarded as burn-in.

3.5.1 Symptom correlation

The inferred correlation matrix $\mathbf{\Sigma}_I^{-1}$ for the infective state is shown in Figure 3.1, where here we present the maximum *a posterior* MCMC collection sample. As expected, all symptoms are relatively highly correlated within the infective state, with minimum correlation between any two symptoms in excess of 0.65 . Note that nasal discharge and nasal congestion are particularly highly correlated, as are throat discomfort, malaise and cough.

We note at this point how we help the model distinguish between the states S and I. Based upon the model construction above, the only way states S and I are distinguished is via a requirement that an individual remain in the I state for a minimum of c days. Based on expertise of the infectious disease medical doctors who are co-authors on this study, we set $c = 3$ days, consistent with the minimum length of time one would be anticipated to manifest symptoms due to infectious disease of the type associated with common viruses. In addition, recall from Section 3.2.2 that a subset of the subjects within the study were confirmed via RT PCR and/or gene-expression analysis to be ill due to a virus. A small subset of these infective individuals had their data removed from the subsequent analysis, and the correlation between the symptoms of this subset of confirmed cases were used to set the hyperparameters in the prior for $\mathbf{\Sigma}_I$. This setting of the model parameters significantly distinguished the S and I states, yielding interpretable results.

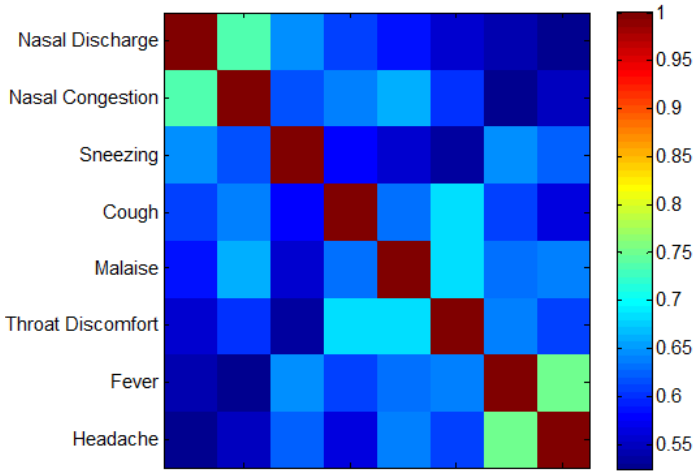


FIGURE 3.1: Inferred correlation matrix for infective state I, $\mathbf{\Sigma}_I^{-1}$, with the approximate MAP solution depicted, corresponding to the maximum *a posterior* collection sample.

3.5.2 Example student trajectories and inferred diagnoses

For each individual in the study, as a function of time (day), we infer the probability that the student is in state I (*i.e.*, that they are sick). To demonstrate this, and to give a sense of the reported symptom scores, in Figure 3.2 self-reported data and the inferred probability of being in state I are depicted for four example individuals; the order of the symptoms (1-8, top to bottom) in Figure 3.2 is consistent with the order of the symptoms in the correlation matrix of Figure 3.1. In Figure 3.2, based upon averaging across all MCMC collection samples, we plot the probability the individual is in the infective state I, for each day. The results in Figure 3.2 are based upon a joint analysis of all self-reported symptom-score data, across all three years.

The inference of the state of health of the subjects in Figure 3.2 is illustrative of model prediction over all time, the results of which provide *interpretative* value for analysis of infectious disease. However, in a clinical setting one would like to make a prediction about the health of an individual based on all symptoms up to the current

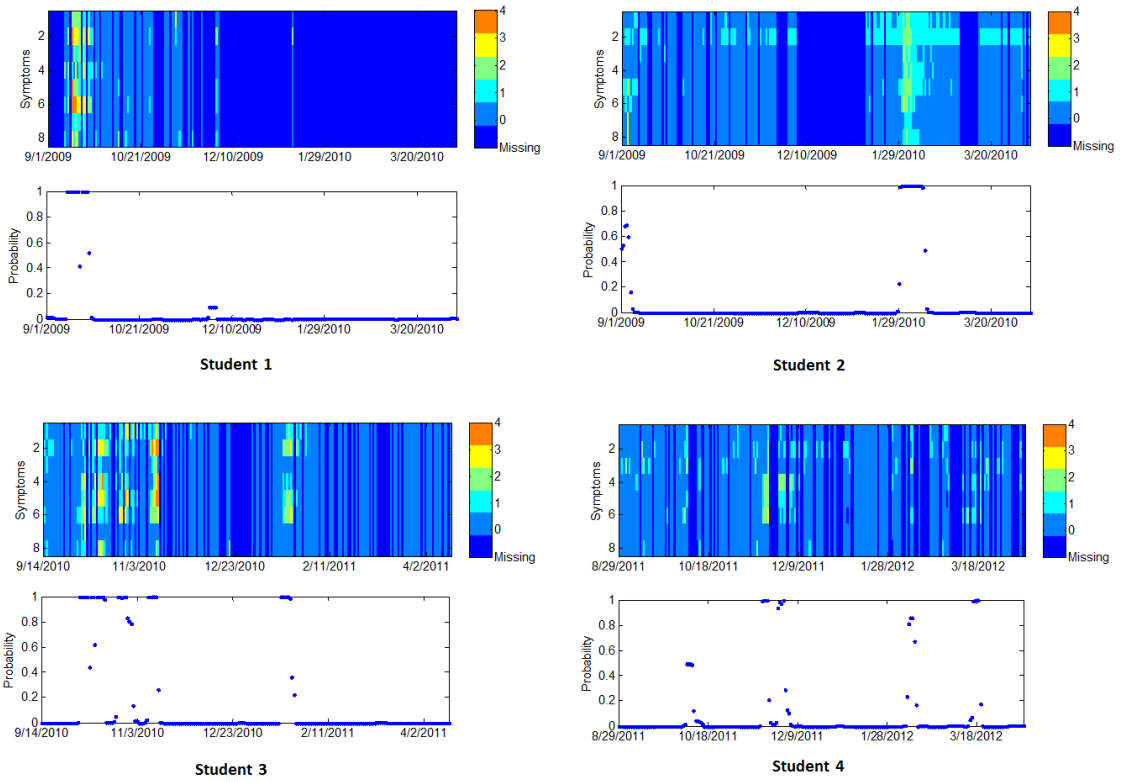


FIGURE 3.2: State of health of four students. For each student, self-reported symptom scores are shown in the top figure. Different colors denote different scores (missing, 0, 1, 2, 3, 4). The probability that a student is in an infective state I at a given time is presented in the bottom subfigure for each of the four students.

point in time (not based on all data, even into the future). We utilize the model for this practical purpose in Section 3.5.6.

3.5.3 Characteristics of missing data

In Section 3.4.3 we proposed a model for the missing data. Specifically, it was assume that if a student is in the susceptible state, S, they do not report symptoms (which are likely negligible) with probability η, thereby manifesting missing data. By contrast, when in the infective state I, it is anticipated that one may be more likely to report symptom scores (which are non-negligible, by definition); the probability of not reporting when in state I is represented by ρ (see Section 3.4.3). Within the

Table 3.1: Summary on properties of student reporting frequency and associated reported symptom scores. The average symptom score reported is the average of the *sum* of the scores for eight symptoms.

Missingness	0-20%	20-40%	40-60%	60-80%	80-100%
# Students	22	158	303	178	205
Avg. Symp. Score	2.67	2.22	2.64	3.27	4.33
Avg. Sick Prob.	0.12	0.11	0.15	0.2	0.32

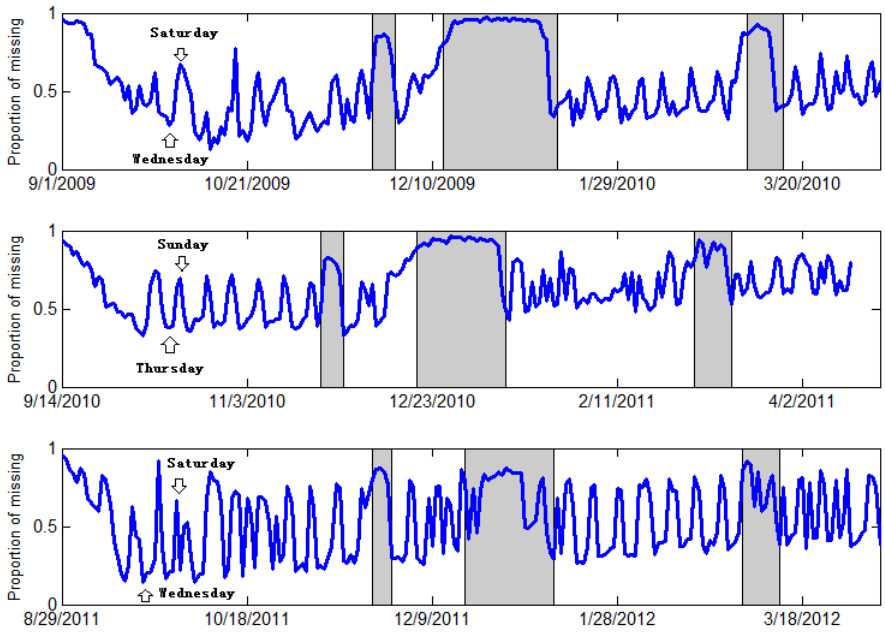


FIGURE 3.3: Fraction of missing data over days. From top to bottom are the results for academic year 2009-2010, 2010-2011 and 2011-2012. The gray bars reflect, from left-to-right, Thanksgiving break, Winter/Christmas break, and Spring break.

analysis, we inferred a mean $\eta = 0.65$, with standard deviation of η equal to 0.01 (reflecting the uncertainty in this parameter from the approximate posterior); the inferred mean for ρ was 0.28, with standard deviation 0.03. Hence, the model infers that when a student is in state S (healthy), a student doesn't report any symptoms 65% of the time, while when in state I (sick) the students don't report symptoms

28% of the time.

In Table 3.1 we show data on the characteristics of student reporting and associated symptom scores. In this table is depicted the percentage of days students didn't report data, and the number of such students in each class of missingness. Note that the largest group of students, with 303 members, did not report symptoms on 40-60% of the days. For each class of missingness, we also report the average reported symptom score, recalling that the values were 0 to 4, with 4 the largest/strongest symptom (eight different symptoms are considered). Note that the average symptom score is particularly large for those students who report data infrequently, and the probability that students are in state I when reporting is heightened for the group that rarely reports. The data in Table 3.1 motivates the model in Section 3.4.3, in which the degree of missingness is assumed to be linked to the latent state of health.

In Figure 3.3 we show the fraction of students who do not report symptoms (fraction of missing data), as a function of day for each of the three years of the study. There is clearly a weekly semi-periodic effect, which has motivated the term $\gamma_t^{(2)}$ in the model. This is discussed further in Section 3.5.7 below.

3.5.4 Virus infection probability over time

We examine the probability of being in the infective state for students living in proximity to infective individuals. As mentioned in Section 3.2.2, RT PCR test results are available for 897 samples (each from an individual student, and infection case), for the set of viruses discussed in Section 3.2.2. If a positive RT PCR-based virus detection occurred for a given student, that student was deemed to be in an infective state (note that, with RT PCR, most sources of error occur with false negatives, so a positive RT PCR test does have a high chance of actually correspond to someone infected with a virus – this is discussed further below).

We wish to examine the probability of whether a student is in state I, relative to

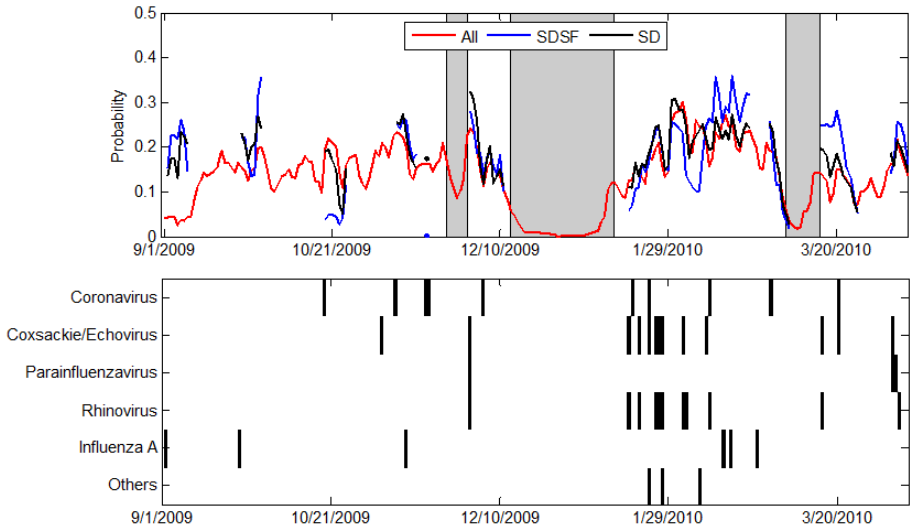


FIGURE 3.4: Top figure: Probability of being in the infective state I on a given day, for academic year 2009-2010. "All" refers to the average across all the students. "SDSF" refers to the average of students living in the same dorm and same floor with infective individuals. "SD" refers to the average of students living in the same dorm with infective individuals. The vertical gray bars represent, from left-to-right, Thanksgiving break, inter-semester (Winter/Christmas) break, and Spring break. Bottom figure: RT PCR test results, black line denotes at that time certain type of virus was detected.

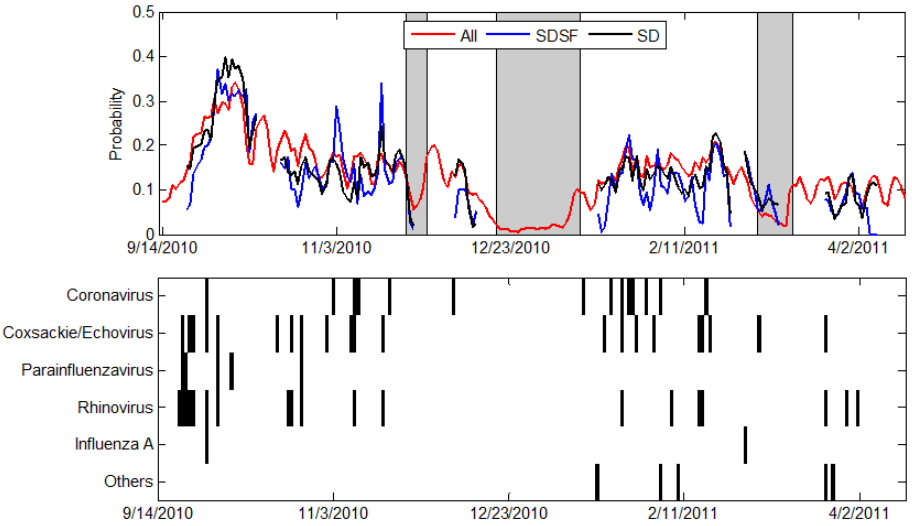


FIGURE 3.5: As in Figure 3.4, for academic year 2010-2011.

that student's living conditions with respect to another student who had a positive RT PCR test. Specifically, assume that a given student has a positive RT PCR test.

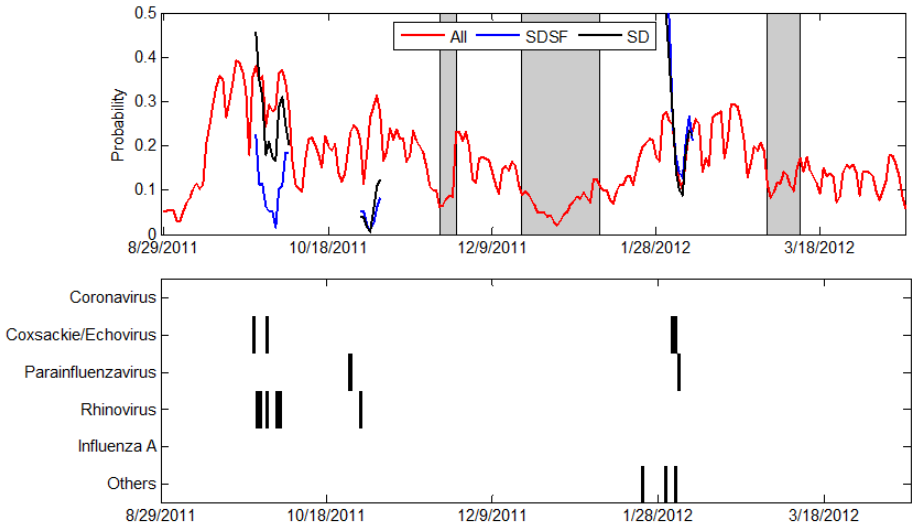


FIGURE 3.6: As in Figure 3.4, for academic year 2011-2012.

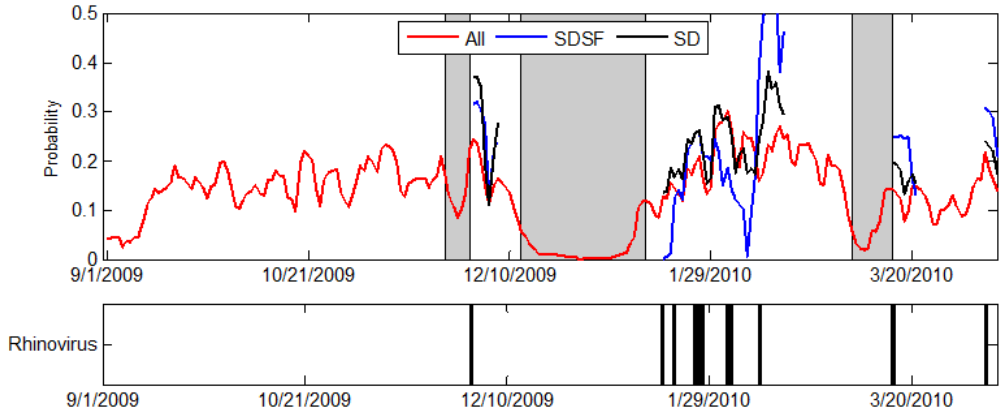


FIGURE 3.7: Top figure: Probability of being in the infective state I on a given day, for academic year 2009-2010. "All" refers to the average across all the students. "SDSF" refers to the average of students living in the same dorm same floor with an infective individual. "SD" refers to the average of students living in the same dorm with an infective individual. The SD and SDSF cases are only for confirmed cases of Rhinovirus. The vertical gray bars represent, from left-to-right, Thanksgiving break, inter-semester (Winter/Christmas) break, and Spring break. Bottom figure: RT PCR test results, black line denotes at that time Rhinovirus was detected.

Over a period of a week after that positive RT PCR test, we examine the probability of being in an infective state for all students who shared a dorm with the student confirmed by RT PCR as being in state I. We also examined the probability of

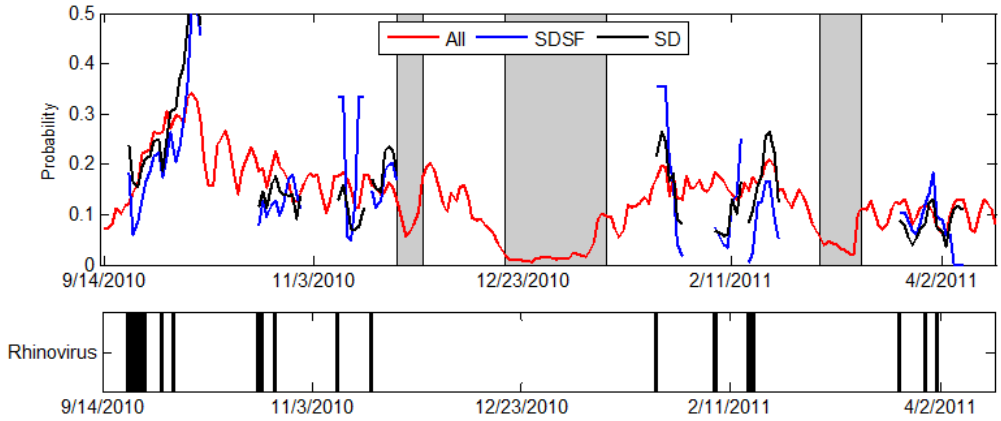


FIGURE 3.8: As in Figure 3.7, for academic year 2010-2011.

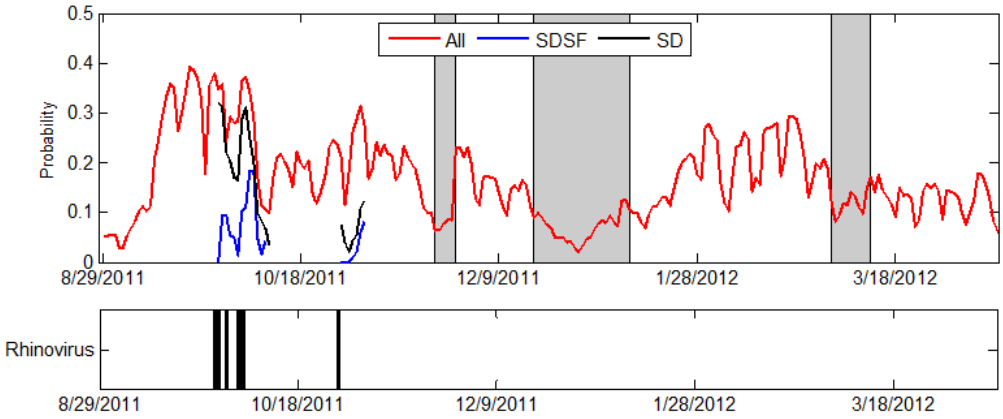


FIGURE 3.9: As in Figure 3.7, for academic year 2011-2012.

being in state I for all students on the same dorm floor (not just the same dorm) of a student confirmed by RT PCR as being infected, again for a week after RT PCR confirmation. When multiple instances overlap in time (multiple positive RT PCR tests), average results are presented across those multiple instances.

To be precise, let $\mathcal{X}$ represent a particular set of students (*e.g.*, a set of students in the same dorm of a RT PCR-confirmed infective student, or a set of students on the same dorm floor of a RT PCR-confirmed infective student). Let $|\mathcal{X}|$ represent the number of individuals in this set. Then we are interested in computing $S_{\mathcal{X}} = \frac{1}{|\mathcal{X}|} \sum_{n \in \mathcal{X}} p(z_{nt} = I | \mathbf{y}_{nt})$, where $p(z_{nt} = I | \mathbf{y}_{nt})$ is computed from our model.

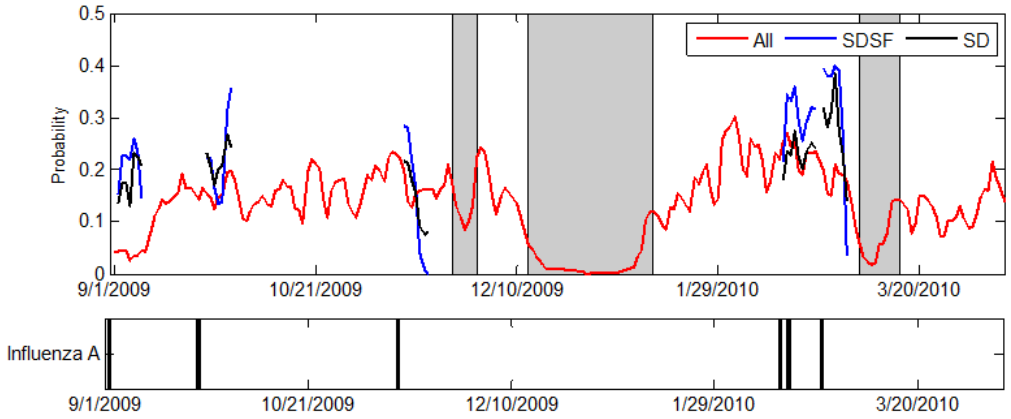


FIGURE 3.10: Top figure: Probability of being in the infective state on a given day, for academic year 2009-2010. "All" refers to the average across all the students. "SDSF" refers to the average of students living in the same dorm and same floor with an infective individual. "SD" refers to the average of students living in the same dorm with an infective individual. The SD and SDSF cases are only for confirmed cases of Influenza A. The vertical gray bars represent, from left-to-right, Thanksgiving break, inter-semester (Winter/Christmas) break, and Spring break. Bottom figure: RT PCR test results, black line denotes at that time Influenza A was detected.

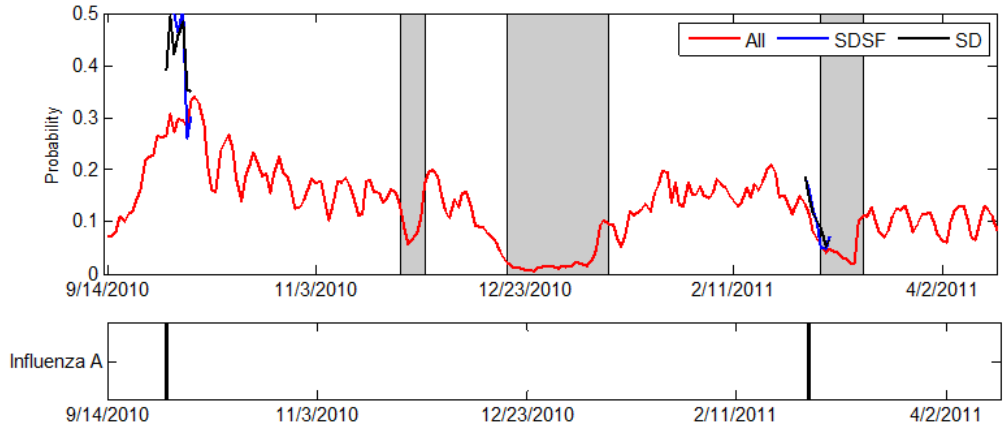


FIGURE 3.11: As in Figure 3.10, for academic year 2010-2011.

This provides a means of examining the inferred degree of enhanced probability of becoming ill with virus, given a nearby confirmed case (recognizing the imperfections in our $p(z_{nt} = I|\mathbf{y}_{nt})$, most notably that one may become sick for other reasons than virus transfer). Such that we have enough individuals in a given set to make this investigation meaningful, we only consider cases for which $|\mathcal{X}| \geqslant 5$; *e.g.*, when

examining propagation of infectious disease on a dorm floor, we only consider cases for which 5 or more students within the study live on the same floor.

To summarize the form of the results, in Figure 3.4 are shown results for the 2009-2010 academic year. On the bottom of Figure 3.4, a black bar represents the presence of a RT PCR-confirmed virus of noted type. At the top in Figure 3.4 is shown the average probability of being in the infective state, under three circumstances. In red are shown results for all students and all times, and therefore in this case $\mathcal{X}$ denotes the set of all students. The blue curve corresponds to the case for which $\mathcal{X}$ corresponds to the set of students from the same dorm and same floor (SDSF) of a RT PCR-confirmed case. Finally, for the black curve, $\mathcal{X}$ corresponds to the set of students in the same dorm (SD) of a RT PCR-conformed case. Unlike the case of all students (red), for the cases of SDSF and SD the curves are not shown at all times, because we only look within a window of seven days after a RT PCR-based detection, and in some cases there are no data (*e.g.*, a given RT PCR-confirmed student doesn't have a sufficient number of students in the study on the same floor, or there were no RT PCR-confirmed detections in the last seven days).

From Figure 3.4, we typically see the following trend. If a student gets sick (is in state I) within a given dorm, from one of the specified viruses, then over the proceeding seven days the average probability of students within the same dorm (SD) will have a heightened probability of being in the infective state as compared to the general student population. Moreover, if a student gets infected, within a week students on the same dorm floor (SDSF) typically have, on average, an even higher probability of being in the infective state. This is not always true, but it seems to be a fairly common situation.

In Figures 3.5 and 3.6 results are shown in this same format as in Figure 3.4, for the 2010-2011 and 2011-2012 academic years, respectively. Given the relatively small number of RT PCR-confirmed cases relative to the total population size, it is difficult

to make strong conclusions from Figures 3.4-3.6. There are periods in which being on the same dorm as an infective student clearly manifests increased probability of being in the infective state, over a subsequent 7 day period, but during other times this trend is not evident (*e.g.*, during the 2011-2012 academic year). One interpretation is that the presence of dorm colleagues who are sick may heighten attention to protecting oneself, through hand washing, etc. Therefore, from this perspective, the presence of a sick student may actually encourage more-healthful behavior in others.

To examine this issue from a finer perspective, we now examine these same types of curves, but for two specific viruses: Rhinovirus and Influenza A. Rhinovirus is associated with the "common cold," and therefore it is a virus that all students will come in contact with, in and outside their dorm. Therefore, in the case of Rhinovirus, the connection to the dorm, and who is infected there at a given time, may be more tenuous (students will come in contact with Rhinovirus and associated infective students in their classes, and other activities outside their dorm). Influenza A occurs more rarely, and therefore if someone is confirmed as infected with Influenza A, it is anticipated that students within the same dorm (SD) and same dorm floor (SDSF) may be at higher risk of infection.

In Figures 3.7-3.9 we show results like discussed above, but now for the SD and SDSF cases we only consider situations in which there was PCR-confirmed Rhinovirus-induced illness. For the case of Rhinovirus, we generally observe that if in the same dorm (SD) or on the same dorm floor (SDSF), when a given student is infected his/her dorm neighbors have a heightened probability of being in the infective state over the next week. However, there are cases for which this is not the case, which indicates that for Rhinovirus transmission activities outside the dorm may be as or more important that the degree of infection within the dorm.

In Figures 3.10-3.11 similar results are shown as above, but now only Influenza

A is considered for the SD and SDSF cases. There are fewer Influenza A cases than Rhinovirus, so conclusions must be drawn with care. Nevertheless, for the case of Influenza A, the SD probability of being infected within a week of a confirmed Influenza A case is heightened relative to the general population, and the SDSF is generally further heightened. Note that in many cases, after roughly 5 days from a confirmed Influenza A case, the SD/SDSF probability of being infected is *less than* that of the general population; right after the confirmed Influenza A case the SD/SDSF probability of being infected increases, but then it diminishes relative to the general student population (*e.g.*, see the case in November 2009, in Figure 3.10). This phenomenon may be attributed to acquired immunity, after being infected.

An interesting phenomenon is observed in Figures 3.7-3.9, when considering the probability of being in an infective state for all of the students (red curve). Note that when the students come together at the beginning of the school year, and after the long Winter break, a general increase in the probability of being in an infective state is observed. Note that at the beginning of the school year, this is particularly evident in the 2010-2011 year (Figure 3.8), and in 2011-2012 (Figure 3.9). Therefore, in Figure 3.8 and 3.9 the students are coming together for the first time at the beginning of the school year, from all over the United States, and from many other countries across the world. This phenomenon of increased probability of infection as students come together for the first time, or after extended break, may be associated with the general spread of infectious disease caused by a new mix of people, as been observed previously in the literature [SWZS11, Kak07].

Note that Figure 3.7 for 2009-2010 has temporal dependence (red curve) that is distinct from 2010-2011 and 2011-2012 (respectively Figures 3.7-3.8). This may be attributed to the fact that the 2009-2010 academic year was the year of the novel H1N1 virus, and significantly heightened on-campus attention to protecting oneself from virus transfer. These results seem to indicate that the heightened attention to

viral transfer manifested by the novel H1N1 virus had a significant impact in reducing the probability of students transiting from the S to I state (not just from H1N1 virus, but from all viruses), when compared to two years in which such attention to viruses was far more muted on campus.

3.5.5 Classification performance based on symptoms

In the above results, we considered the average probability that students were in state I at a particular point in time. We wish to now examine the accuracy of the prediction of infection, relative to an objective "truth." To do this, we considered all 897 individuals for whom RT PCR-based virus-identification was performed (for a subset of these, for which the RT PCR test was negative, confirmation gene-expression analysis was also performed).

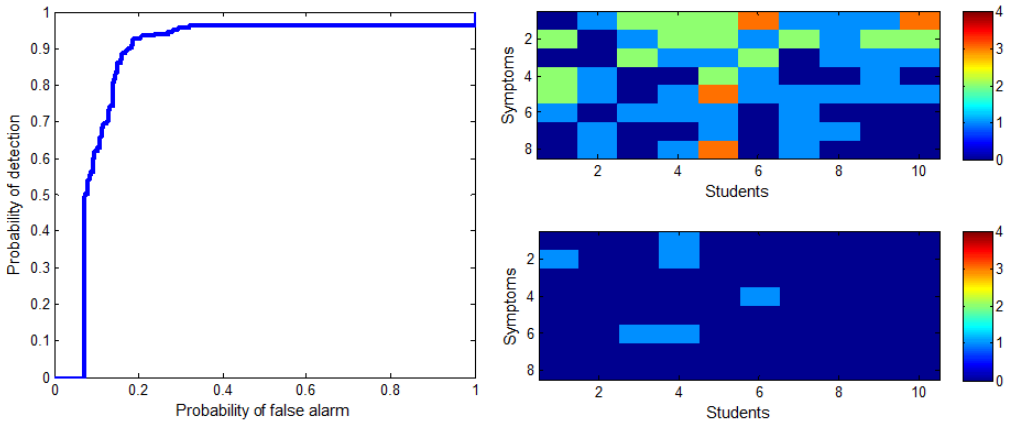


FIGURE 3.12: Left column: ROC curve. Right column: Top figure is the symptom scores of students who are healthy (in state S) but labeled infective (in state I) with high probability by the model. The bottom figure shows the symptom scores of students who are infective (in state I) but labeled healthy (in state S) by the model. The order of the symptoms (1-8, top to bottom) is consistent with the order of the symptoms in the correlation matrix of Figure 3.1.

In our study there were two reasons a given student could have a RT PCR test performed: (i) based upon their self-reported data, a doctor in (near) real-time

determined that they were infected, and therefore they were contacted for acquisition of a sample; (*ii*) a given individual was a close contact of a person who was sampled in the case of (*i*). Therefore, for the close contacts, the student may not be in an infective state at the time of sampling, either because there was no disease transfer, or because the onset of illness was manifested at a later time.

Based upon the symptoms, and the doctor-based diagnosis, all individuals in case (*i*) above are defined as being in the infective state, essentially by definition (these students were only contacted because their symptoms were deemed above a threshold of illness). For the close contacts, all individuals for whom the RT PCR test was negative were deemed to be in the *healthy* (susceptible, S) state, and all others were deemed to be in the infective state (the RT PCR test may miss some infected people, which the gene-expression analysis can pick up, and this issue is discussed when presenting results).

The receiver operating characteristic (ROC) curve is manifested by thresholding $p(z_{nt} = I|-)$, and is shown in the left column of Figure 3.12. Note that the model achieves a 90% detection rate at a false-alarm rate of 15%. However, the quality of the ROC is undermined by imperfections in the definition of "truth." People who are sick as a result of illnesses other than virus will deemed as healthy in the truth (negative RT PCR test), but in reality they are sick. Another source of errors are manifested by positive RT PCR tests for the presence of virus, but the individual shows no symptoms – these are termed "shedders" in the medical community [DEL07]. These individuals are carrying the virus, and shedding the virus, but they do not show any symptoms. The RT PCR test will deem these individuals as being in the infective state I, but from the standpoint of symptoms, which is what our analysis considers, these people are not infected (there are no symptoms present that would allow one to declare they are infected, based on symptoms alone).

In Figure 3.12, left, note (*a*) the presence of many false alarms before any detec-

tions are achieved (left-most part of the ROC), and (b) after a probability of false alarms of about 0.35, the detection probability is stuck at around 0.95, until the very rightmost part of the ROC. Concerning (a), on the top-right of Figure 3.12, we show the symptom scores for the ten students who characterize the individuals detected as being infected, but RT PCR deems as being healthy, or in state S (the false alarms at the beginning of the ROC). Based upon the symptoms (right in Figure 3.12), these students are almost certainly sick due to some cause other than virus, or because of limitations of the RT PCR test (*e.g.*, poor samples, or because the illness was caused by a virus other than that tested by the RT PCR).

At bottom-right in Figure 3.12 is shown the symptom scores of students who were deemed infective via RT PCR, but our model deemed as healthy, based upon the symptoms. We see that the symptoms of these students are indeed very mild, or absent. These individuals were likely carrying a virus that was tested, and that was detected via RT PCR. However, these individuals were likely recently infected with the virus, and still carrying it, but no longer infected. Alternatively, these individuals may have been asymptomatic shedders.

Gene expression data were available for 6 of the individuals considered at right in Figure 3.12. In all of these cases, the gene-expression analysis was able to confirm the labels inferred by our algorithm based on symptoms.

3.5.6 Online prediction of health

The results in Figure 3.12 on predicting the state of health were based on *all* of the self-reported data, at all times for which data were reported. Of course, in a clinical setting a clinician must predict the state of health only based upon symptoms up to the point at which a diagnosis is made. It is desirable to predictive probability that a particular student is in state I on day $t+1$, based on symptom scores up to day t.

Let $\mathbf{y}_1^t = \{\mathbf{y}_1, ..., \mathbf{y}_t\}$ represent the symptom scores up to time t for the student in

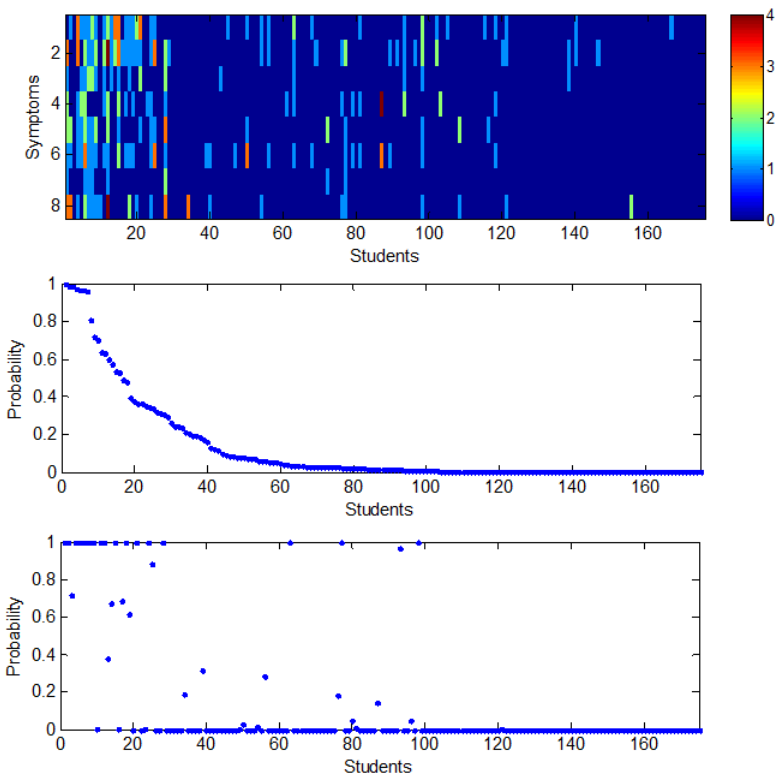


FIGURE 3.13: The top figure is the symptoms scores for students at time $t+1$ (can be considered as "truth"). The middle figure is $p(z_{nt+1} = I|\mathbf{y}_{n1}^t, -)$, the predictive probability that a given student is in the infective state at $t+1$. The bottom figure is the probability that students stay in infected at $t+1$ given all the data. The order of the symptoms (1-8, top to bottom) is consistent with the order of the symptoms in the correlation matrix of Figure 3.1.

question. The probability that the student is in state I on day $t+1$ may be expressed as

$$p(z_{t+1} = I|\mathbf{y}_1^t, \Omega) = \frac{\sum_{d_{t+1}=1}^{D_{max}} p(z_{t+1} = I, d_{t+1}, \mathbf{y}_1^t|\Omega)}{\sum_{d_{t+1}=1}^{D_{max}} p(z_{t+1} = I, d_{t+1}, \mathbf{y}_1^t, |\Omega) + p(z_{t+1} = S, \mathbf{y}_1^t|\Omega)} \tag{3.12}$$

where $p(z_{t+1} = I, d_{t+1}, \mathbf{y}_1^t|\Omega)$ represents the joint probability of data $\mathbf{y}_1^t$, that the student is in state I on day $t+1$, and that they are in day d_{t+1} of being infected; $d_{t+1} \in \{1, .., D_{max}\}$ is the number of days left in infective state at time $t+1$. $p(z_{t+1} = S, \mathbf{y}_1^t|\Omega)$ represents the joint probability of the data and being in the susceptible state S. In both cases, Ω represents model parameters learned from data up to day t. The

details for calculating $p(z_{t+1} = I, d_{t+1}|\mathbf{y}_1^t, \Omega)$ and $p(z_{t+1} = S, \mathbf{y}_1^t|\Omega)$ are provided in Appendix B.

In this experiment, the first two years of data, and the third year of data up to day $t = 140$ are employed to learn Ω (these are typical results for many values of t selected in Year 3). In Figure 3.13 are shown the model predictions for all students in Year 3 (2011-2012), where in Figure 3.13 the students are ordered from left to right from the most to least probable of being in state I on day $t+1$. At the top in Figure 3.13 are shown the symptoms reported on day $t+1$ (for those for whom scores were provided), and it is evident that the individuals who are deemed most likely to be in state I on day $t+1$ (based on data up to day t) tend to have the strongest symptoms on that day. In the middle in Figure 3.13 is shown the probability of being in an infective state on day $t+1$, based on data up to day t. Finally, the bottom part of Figure 3.13 shows the probability of being in an infective state on day $t+1$ based on all of the data. Note that there is generally good agreement (middle and bottom figures) on which students are most likely to be in the infective state on day $t+1$.

3.5.7 Breaking out model components

In Figure 3.14 are plotted the posterior mean of the general trend terms $\gamma_t^{(1)}$ for academic years $2009-2010$, $2010-2011$ and $2011-2012$; the error bars reflect one standard deviation (estimated from the Gibbs collection samples). The weekly parameter $\gamma_t^{(2)}$ is displayed in Figure 3.15. In this figure the weeks are identified, with the beginning of a week defined here as Monday. We observe the that the weekly pattern (impacting the probability of transiting from healthy to infective state) is typically peaked at either Wednesday or Thursday, and tends to be smaller around the weekend. This is possibly reflective of the fact that students are more likely to report symptoms during the school week than they are on the weekend, when they may be distracted by funner activities. Of course, another interpretation is that the

probability that the students will *feel* infected/sick is diminished during the weekend, relative to the middle of the week, when they may be under greater stress.

Recall Figure 3.3 from above, which depicts the degree of missingness on average as a function of days. By construction, heightened missingness is deemed associated with health, and weekends tend to be periods of high missingness. Whatever the cause of the weekly effects (student laziness/distraction or actual health), model interpretation may be improved by removing this effect. We consider this below.

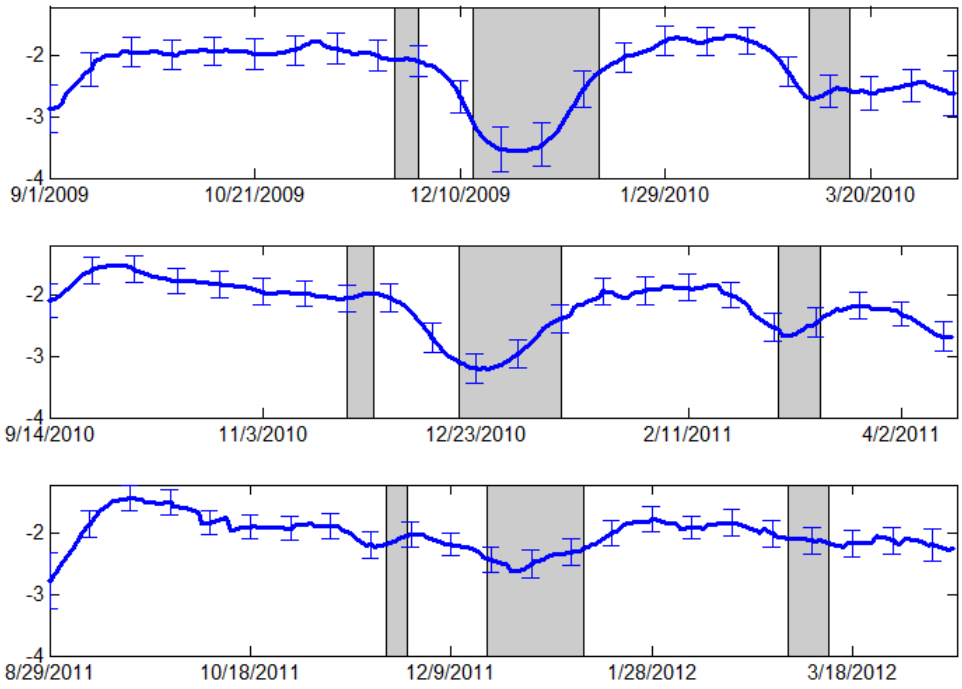


FIGURE 3.14: General trend term $\gamma_t^{(1)}$. From top to bottom are the results for academic year 2009-2010, 2010-2011 and 2011-2012. The error bars reflect one standard deviation. The gray bars reflect, from left-to-right, Thanksgiving break, Winter/Christmas break, and Spring break.

In Figure 3.16 we show $\gamma_t^{(3)}$ associated with the Google Flu Trend data. In this plot we show the mean and one standard deviation, again from posterior collection samples. The posterior distribution in this term is manifested by the posterior distri-

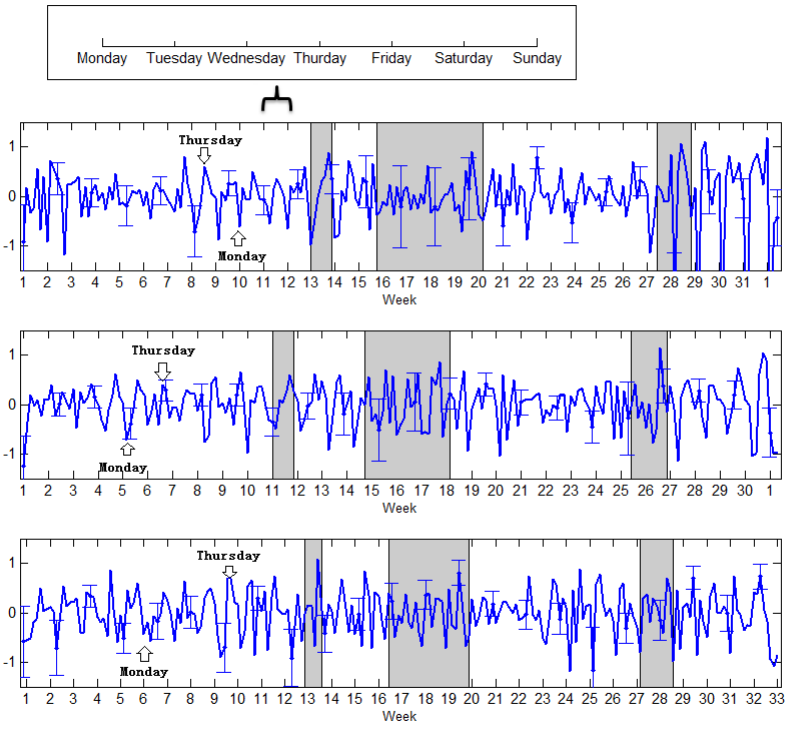


FIGURE 3.15: Weekly or semi-periodic term $\gamma_t^{(2)}$. From top to bottom are the results for academic year 2009-2010, 2010-2011 and 2011-2012. The gray bars reflect, from left-to-right, Thanksgiving break, Winter/Christmas break, and Spring break.

bution on ξ, as the total term is ξf_t, and $f(t)$ represents the deterministic/observed Google FluTrends (for the city of Durham, NC). Note that the contribution of the Google Flu Trend term is relatively small (large mass concentrated around zero, particularly for the first two years), which implies that the spread of infectious disease among students on the Duke University campus is a relatively isolated ecosystem, distinct from the city and community of Durham used here for f_t.

In Figure 3.17 we depict the inferred probability of transiting from state S to state I, as a function of day, for each of the three years of the study. The data were analyzed using all components of the model. However, after this analysis, to remove the effects of the weekly term, we show the model-inferred probability

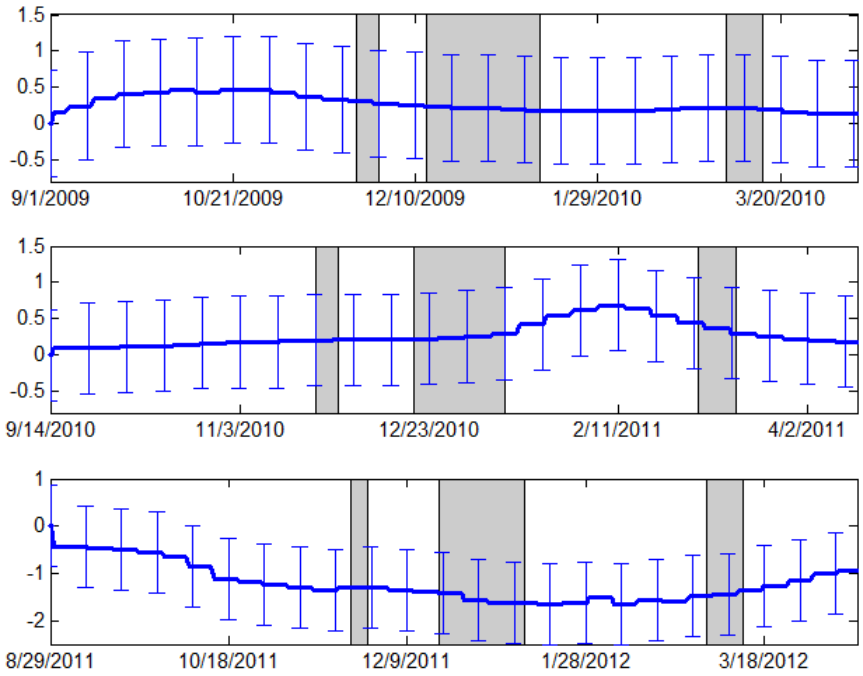


FIGURE 3.16: Google Flu Trends (for Durham, NC, USA) regression term $\gamma_t^{(3)}$. From top to bottom are the results for academic year 2009-2010, 2010-2011 and 2011-2012. The gray bars reflect, from left-to-right, Thanksgiving break, Winter/Christmas break, and Spring break.

of transiting from $S \to I$, and with the weekly term removed. It is interesting to examine the red curve in Figure 3.17, in which the weekly effects are removed. Recall that the 2010-2011 and 2011-2012 academic years were distinct from 2009-2010, as the latter was associated with the novel H1N1 virus. Note that at the beginning of the academic year in 2010-2011 and 2011-2012, there is a clear increased probability of getting infected within the first month or so the students are together, presumably a mixing effect [SWZS11, Kak07] caused by interactions of many people who have never met before, coming from all over the United States, and also from outside the United States. It appears that the heightened attention to viruses (from the alarm associated with novel H1N1) dampened this phenomenon in 2009-2010. During the first semester of 2009-2010, when there was so much attention to viruses on campus,

there is a noticeable decrease in the probability of transiting from state S to I, after the weekly effects are removed (relative to 2010-2011 and 2011-2012).

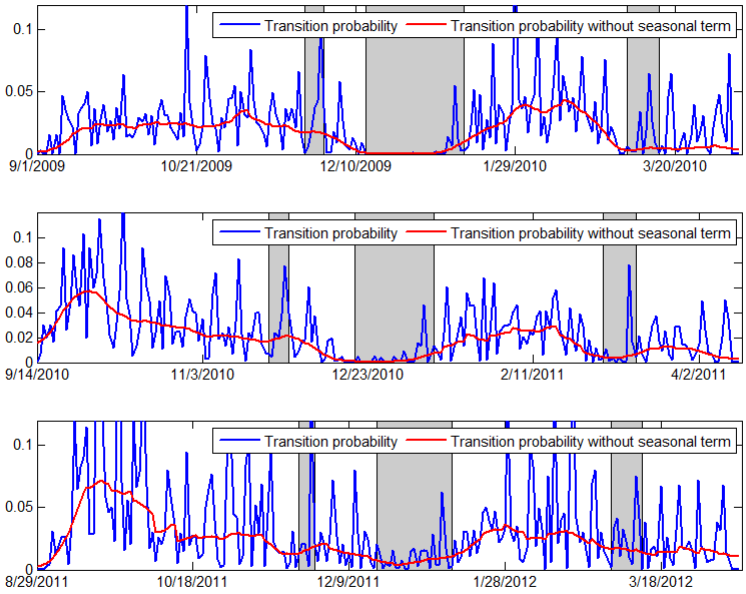


FIGURE 3.17: Probability of transiting from state S to state I. The blue curve represents the total probability, and the red curve represents the probability with the weekly term $\gamma_t^{(2)}$ removed.

3.6 Conclusions

A statistical model has been developed for analysis of the time-dependent symptom scores provided by a large group of undergraduate college students. Unlike almost all studies of data related to infection transfer, the model has operated directly on the observed symptoms, and the state of the students were assumed to be latent. The community-to-person mechanism for pathogen transfer has been modeled in terms of a SIS analysis, and computations have been performed using Bayesian (MCMC) methods. A detailed characterization of the data and the scientific questions that have motivated this study are discussed in Section 3.2; a comprehensive answering of

these questions with the available data has been provided in Section 3.5. For brevity, we do not repeat these details here. We note that these data are presented here for the first time, and were collected by the authors; all data will be made available to the research community.

There are further questions that may be examined with the collected data, and that are worthy of future study. The identification of the virus responsible for each illness has been (imperfectly) constituted via RT PCR, for a large set of common viruses considered. We have access to the dorm in which each individual resided. A more detailed analysis of pathogen transfer as a function of virus type can be examined. In this paper we have presented results in this direction, but more explicit modeling could be performed (not necessarily at the symptom level, but after the responsible virus has been identified by RT PCR).

The gene expression data from this study have only been employed here in a limited manner, as the focus has been on self-reported symptom scores. However, for the close contacts, we have daily gene expression data for a week. For close contacts who transited from state S to I, we have the opportunity to analyze the time trajectory of the gene expression data as the host responds to the (known) virus. We have performed work of this type for people enrolled in challenge studies (controlled experiments) [CZW$^+$11]; the data from this study offers the potential for similar studies on data from individuals who became ill in natural settings. We have preliminary results in this direction on these data, which are encouraging and will be presented elsewhere.

4

A big-data investigation of electoral representation

4.1 Introduction

One of the fundamental research topics in political science is the extent to which elected officials represent the preferences of the citizens who elect them. Although democratic theorists assume an electoral connection between representatives and their constituents, data limitations have historically made it difficult to empirically evaluate both legislators and the public within the same policy space. A long line of research has estimated the ideological preferences of legislators from their voting records, using an "ideal point" model [CJR04, PR85]. Such a model typically assumes each legislator and each piece of legislation can be represented by a point in a one-dimensional latent space. More recently [GB11, SCD$^+$12, SDC13, WLS$^+$10, WSDC13, ZC12] have offered approaches for incorporating information beyond roll-call votes. For example, in [GB11, WLS$^+$10] a latent factor model is proposed to jointly analyze the congressional votes and the legislative text. In [GB12] the authors improve the model by allowing the ideological position of legislators to vary on specific issues. Further, in [SDC13, WSDC13] a spatio-temporal model is pro-

posed, accounting for the time of the votes and the spatial location of the legislators' districts. However, these methods do not explicitly account for properties of the constituents living within a given electoral district.

Estimating the ideological preferences of a member's ideological district is far more difficult. Some researchers rely on crude proxies such as presidential vote share [CWBC02]. More recently, scholars have turned to public opinion polls – often pooling many different national surveys to increase sample sizes [BH10, Cli06, LP09]. For example, in [LP09] over 100 surveys are aggregated to estimate state-level ideological preferences. Unfortunately, these works are limited by the relatively small number of survey respondents, which causes inaccuracy in parameter estimation, while also hindering access to finer-scale (district level) constituency information.

Motivated by these challenges, we propose a new scalable Bayesian model to jointly analyze individual-level constituency information, congressional roll-call votes, and associated legislative text. For the constituent information, we leverage a random, de-identified sample of 3 million individuals from the political data vendor Catalist, which collects, maintains, and updates a database with political, demographic, and commercial characteristics on 280 million Americans. Matrix factorization [SM08] is integrated with the hierarchical Dirichlet process (HDP) [TJBB04], yielding a statistical characterization of people living within each US congressional district. Further, a topic model is employed on the text of the legislation. The inferred district-level feature vectors of the people living in each district and the topic distribution on a given piece of legislation are employed to infer roll-call votes. Within the model is a novel component that allows inference of the degree to which a given legislator votes in a manner aligned with the interests of his/her constituents. The inferred value of this parameter is examined in the context of the success of the legislator in the next election, yielding a new means to evaluate the relationship between legislative behavior, constituent preferences, and electoral outcomes. To

address the massive scale of the constituency data, stochastic variational Bayesian inference [BS12, HBWP13, WPB11] is utilized.

While the explicit data considered here are associated with politics, the basic model setup is more general. One may envision trying to assess whether specific individuals, from a region or group with particular demographics, will like/dislike given products. The binary legislative votes are analogous to like/dislike of particular products (here legislation), targeted toward specific people. The text of the legislation is like a document describing the product in question. Given a new product/legislation, with an associated text description, we wish to predict whether it will be liked/disliked by particular people (here, whether legislators will vote yes/no on a new piece of legislation).

4.2 Model construction

4.2.1 Data and notation

We jointly analyze congressional roll call votes and constituent information for the $J = 435$ congressional districts across the United States. Individual-level constituent information comes from Catalist, a political data vendor (`www.catalist.us`). An academic subscription provided a 1% random sample of their database (3 million cases) in 2012, and includes a wide range of demographic, political, and commercial characteristics about each individual. For each (anonymous) individual in the Catalist data, there is an associated vector of attributes, describing personal information, such as race, income, education level and voting-turnout history; these features are mixed, real and binary. Let $\mathbf{X}_j \in \mathbb{R}^{P^r \times N_j}$ denote real-valued attributes for individuals in district $j \in \{1, \ldots, J\}$, where N_j denotes the number of individuals from district j for whom we have Catalist data, and P^r represents the number of real attributes. Let $\mathbf{B}_j \in \{0, 1\}^{P^b \times N_j}$ denote the binary attributes for the same individuals. Additionally, we have a series of Congressional votes on pieces of legislation, for legislators

elected around the time the Catalist data were collected (we consider roll-call data in 2009-2011). Let $\mathbf{R} \in \{0,1\}^{J\times L}$ denote Congressional roll-call votes on bills reaching the House floor (there are 6% missing votes). Finally, for each piece of legislation, we have the associated text of the bill. The lth piece of legislation is denoted $\boldsymbol{w}_l$, where $\boldsymbol{w}_l \in \mathbb{Z}_+^V$ represents the count of each word in the text (a vector of nonnegative integers), where the vocabulary dimension is V.

4.2.2 Matrix factorization of constituent data

The matrix of real-valued individual-level data from people in district j is factorized as

$$\mathbf{X}_j = \mathbf{D}^r \mathbf{\Lambda}^r \mathbf{S}_j^r + \mathbf{E}_j^r, \tag{4.1}$$

where $\mathbf{D}^r \in \mathbb{R}^{P^r\times K^r}$, $\mathbf{S}_j^r \in \mathbb{R}^{K^r\times N_j}$, $\mathbf{\Lambda}^r = \text{diag}(\lambda_1^r, \ldots, \lambda_{K^r}^r)$, and $\mathbf{E}_j^r \in \mathbb{R}^{P^r\times N_j}$. Each column of $\mathbf{E}_j^r$ is drawn from $\mathcal{N}(0, \sigma_j^{-1}\mathbf{I})$ and a diffuse gamma prior is placed on σ_j, *i.e.*, $\text{Ga}(10^{-6}, 10^{-6})$. Note that $\mathbf{D}^r$ and $\mathbf{\Lambda}^r$ are shared for all districts j. Each column of $\mathbf{D}^r$ is drawn from $\mathcal{N}(0, \mathbf{I}_{P^r})$, where $\mathbf{I}_{P^r}$ is the $P^r \times P^r$ identity matrix. We wish to impose that $|\lambda_k^r|$ decreases as index k increases; hence, while we truncate the model to K^r factors, through the λ_k^r we infer the subset of factors that are needed to represent the data. To achieve this, we employ the multiplicative gamma process (MGP) proposed in [BD11]: $\lambda_k^r \sim \mathcal{N}(0, 1/\tau_k^r)$, $\tau_k^r \sim \prod_{h=1}^k \varphi_h^r$, and $\varphi_h^r \sim \text{Ga}(a_1, 1)$. By choosing $a_1 > 1$, $\mathbb{E}(\varphi_h^r) > 1$, encouraging τ_k^r to increase with k; this in turn results in increasing encouragement of shrinking the amplitude of λ_k^r as k increases.

For the observed matrix of binary data for people in district j, $\mathbf{B}_j$, we employ a probit model, and a latent $\tilde{\mathbf{B}}_j \in \mathbb{R}^{P^b\times N_j}$ [AC93a]. Let $\tilde{b}_{jpn}$ be element (p,n) in $\tilde{\mathbf{B}}_j$ and let b_{jpn} represent element (p,n) in $\mathbf{B}_j$; these are related via the probit link: $b_{jpn} = 0$ if $\tilde{b}_{jpn} + \epsilon_{jpn}^b \geqslant 0$, and $b_{jpn} = 1$ if $\tilde{b}_{jpn} + \epsilon_{jpn}^b < 0$, where $\epsilon_{jpn}^b \sim \mathcal{N}(0,1)$. We factorize the latent matrix as $\tilde{\mathbf{B}}_j = \mathbf{D}^b \mathbf{\Lambda}^b \mathbf{S}_j^b$, where $\mathbf{D}^b \in \mathbb{R}^{P^b\times K^b}$ and $\mathbf{S}_j^b \in \mathbb{R}^{K^b\times N_j}$. The columns of $\mathbf{D}^b$ are drawn with the same class prior as employed above for $\mathbf{D}^r$,

and the MPG prior is employed for $\mathbf{\Lambda}^b = \text{diag}(\lambda_1^b, \ldots, \lambda_{K^b}^b)$.

4.2.3 Clustering the constituency latent features

Individual n sampled from district j is characterized by the nth column of $\mathbf{S}_j^r$ and $\mathbf{S}_j^b$. Assuming that people are likely clustered with respect to the attributes included in the Catalist database, we develop a joint mixture model for the columns of $\mathbf{S}_j^r$ and $\mathbf{S}_j^b$. Let $\boldsymbol{s}_{jn}^r$ and $\boldsymbol{s}_{jn}^b$ denote the nth columns of $\mathbf{S}_j^r$ and $\mathbf{S}_j^b$, respectively. We impose the following hierarchical Dirichlet process (HDP) [TJBB04] model:

$$\begin{gathered} \boldsymbol{s}_{jn}^r \sim f(\boldsymbol{\theta}_{jn}),\ \boldsymbol{s}_{jn}^b \sim f(\boldsymbol{\psi}_{jn}),\ \{\boldsymbol{\theta}_{jn}, \boldsymbol{\psi}_{jn}\} \sim G_j, \\ G_j \sim \text{DP}(\kappa, G_0),\ G_0 \sim \text{DP}(\kappa_0, H) \end{gathered} \tag{4.2}$$

where $H(\boldsymbol{\theta}, \boldsymbol{\psi}) = H_r(\boldsymbol{\theta}) H_b(\boldsymbol{\psi})$, and therefore $G_0 = \sum_t \nu_t \delta_{(\boldsymbol{\theta}_t^*, \boldsymbol{\psi}_t^*)}$, with $\nu_t > 0$, $\sum_t \nu_t = 1$ and $\delta_{(\boldsymbol{\theta}_t^*, \boldsymbol{\psi}_t^*)}$ a unit point measure concentrated at the pair $(\boldsymbol{\theta}_t^*, \boldsymbol{\psi}_t^*)$. The distribution $f(\cdot)$ here corresponds to multivariate Gaussian, and H_r and H_b are each Normal-Wishart distributions. Diffuse gamma priors are placed on κ and κ_0. We employ the stick-breaking representation [Set91] of the HDP developed in [TJBB04] and a point estimate of $\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots)^T$ [BS12, LPJK07] to simplify the variational derivations (discussed in Section 4.3). The number of components ("sticks") used to approximate G_0 and each of the G_j is truncated to T. Each district j is characterized by $G_j = \sum_{t=1}^T \pi_{jt} \delta_{(\boldsymbol{\theta}_t^*, \boldsymbol{\psi}_t^*)}$. The "atoms" $\{\boldsymbol{\theta}_t^*, \boldsymbol{\psi}_t^*\}$ are shared across all J districts, and hence the jth district is distinguished by the probability vector $\boldsymbol{\pi}_j = (\pi_{j1}, \ldots, \pi_{jT})^T$.

4.2.4 Modeling the text of legislation

Consider a corpus of L pieces of legislation, voted on during a Congressional session. A probability vector $\boldsymbol{\beta}_l$ is inferred to represent the lth piece of legislation. Specifically, we employ a basic topic model, latent Dirichlet allocation (LDA) [BNJ03] to model each of the L documents, from which we constitute $\boldsymbol{\beta}_l$, a probability vector over topics

(assumed here to be truncated to K topics). Topic $k \in \{1, \ldots, K\}$ is characterized by a V-dimensional probability vector $\boldsymbol{\phi}_k$, and a word from document/legislation l is associated with topic k with probability β_{lk}. If a word is drawn from topic k, the specific word is drawn Mult$(1, \boldsymbol{\phi}_k)$ [BNJ03].

The vote of the jth legislator on bill l is modeled in terms of $\boldsymbol{\pi}_j$ and $\boldsymbol{\beta}_l$, coupling the constituency data and the text of legislation to predict roll-call votes. Rather than predicting roll-call votes directly based on $\boldsymbol{\pi}_j$ and $\boldsymbol{\beta}_l$ (the doing of which significantly complicates inference), we introduce surrogates for $\boldsymbol{\pi}_j$ and $\boldsymbol{\beta}_l$ [BM07]. Specifically, individual $n \in \{1, \ldots, N_j\}$ in district j has an associated latent variable $c_{jn} \in \{1, \ldots, T\}$, identifying which model parameters $(\boldsymbol{\theta}^*_{c_{jn}}, \boldsymbol{\psi}^*_{c_{jn}})$ are used for his/her representation. This assigns individual n in district j to a cluster, with cluster t characterized by $(\boldsymbol{\theta}^*_t, \boldsymbol{\psi}^*_t)$. The VB analysis yields the expected probability of which of the T clusters person n in district j is associated with, this probability vector denoted $\tilde{\boldsymbol{\pi}}_{jn}$.

Similarly, we introduce latent variable $z_{il} \in \{1, ..., K\}$, assigning a topic to word i in document l. Within the VB inference of LDA, we manifest $\tilde{\boldsymbol{\beta}}_{li}$, the expected probability vector for which topic word i in document l is associated with. We predict the roll call vote associated with district j for legislation l in terms of the two probability vectors $\tilde{\boldsymbol{\pi}}_j = \frac{1}{N_j} \sum_{n=1}^{N_j} \tilde{\boldsymbol{\pi}}_{jn}$ and $\tilde{\boldsymbol{\beta}}_l = \frac{1}{W_l} \sum_{i=1}^{W_l} \tilde{\boldsymbol{\beta}}_{li}$, assuming W_l total words in document l.

4.2.5 Coupling constituency characteristics and legislative text: Roll-call analysis

Like for the binary attributes $\mathbf{B}_j$ discussed above, for the binary roll-call votes we assume a latent matrix $\tilde{\mathbf{R}} \in \mathbb{R}^{J \times L}$ which we factorize as $\tilde{\mathbf{R}} = \mathbf{D}^\ell \boldsymbol{\Lambda}^\ell \mathbf{S}^\ell + \mathbf{E}^\ell$. The MPG prior is imposed for the elements of the diagonal matrix $\boldsymbol{\Lambda}^\ell$.

Row j of $\mathbf{D}^\ell$, denoted by the column vector $\boldsymbol{d}^\ell_j$, is a feature vector associated with district j, from the standpoint of voting on legislation. The lth column of

$\mathbf{S}^\ell$, denoted by the column vector $\boldsymbol{s}_l^\ell$, is similarly a feature vector for legislation l (from the standpoint of how the text affects the voting). We connect the voting characteristics of the legislators from district j to the constituency characteristics of his/her district by modeling $\boldsymbol{d}_j$ in terms of $\tilde{\boldsymbol{\pi}}_j$. Similarly, we connect votes to the properties (text) of the legislation by modeling $\boldsymbol{s}_l^\ell$ in terms of $\tilde{\boldsymbol{\beta}}_l$. Specifically, we impose the models

$$\boldsymbol{d}_j^\ell = \mathbf{U}^d \tilde{\boldsymbol{\pi}}_j + \boldsymbol{d}_0^\ell + \boldsymbol{\xi}_j \ , \quad \boldsymbol{s}_l^\ell = \mathbf{U}^s \tilde{\boldsymbol{\beta}}_l + \boldsymbol{s}_0^\ell \ , \tag{4.3}$$

where $\mathbf{U}^d \in \mathbb{R}^{K^\ell \times T}$, $\boldsymbol{\xi}_j \in \mathbb{R}^{K^\ell}$, $\boldsymbol{d}_0^\ell \in \mathbb{R}^{K^\ell}$, $\mathbf{U}^s \in \mathbb{R}^{K^\ell \times K}$ and $\boldsymbol{s}_0^\ell \in \mathbb{R}^{K^\ell}$. The elements of $\mathbf{U}^d$, $\boldsymbol{d}_0^\ell$, $\mathbf{U}^s$ and $\boldsymbol{s}_0^\ell$ and are drawn i.i.d. from, respectively, $\mathcal{N}(0, \alpha_d^{-1})$, $\mathcal{N}(0, \alpha_{d0}^{-1})$, $\mathcal{N}(0, \alpha_s^{-1})$ and $\mathcal{N}(0, \alpha_{s0}^{-1})$, with diffuse gamma priors on α_d, α_{d0}, α_s and α_{s0}.

The vector $\boldsymbol{\xi}_j$ is employed to identify legislators who may be voting against the interests of their constituents, as defined by the attributes in the Catalist database. Since it is hoped that most of $\boldsymbol{d}_j^\ell$ is captured by these features, we impose a prior on $\boldsymbol{\xi}_j$ that encourages (near) sparsity. Therefore, we impose the hierarchical shrinkage prior $\xi_{jk} \sim \mathcal{N}(0, \alpha_{jk}^{-1})$, $\alpha_{jk} \sim \text{InvGa}(1, \gamma_{jk}/2)$, $\gamma_{jk} \sim \text{Ga}(10^{-6}, 10^{-6})$.

The matrix $\mathbf{E}^\ell \in \mathbb{R}^{J \times L}$ models "random effects." Let E_{jl}^ℓ represent component (j, l) of $\mathbf{E}^\ell$. We impose $E_{jl}^\ell = \delta_l + \delta_{jl}$, where δ_l is a random effect associated with legislation l and δ_{jl} is a random effect associated with the legislation-legislator pair. We further connect δ_l to the legislative text by modeling it in terms of $\tilde{\boldsymbol{\beta}}_l$: $\delta_l = \boldsymbol{w}^T \tilde{\boldsymbol{\beta}}_l + w_0$, where $\boldsymbol{w} \in \mathbb{R}^K$ and $w_0 \in \mathbb{R}$ are i.i.d draw from $\mathcal{N}(0, \alpha_w^{-1})$ and $\mathcal{N}(0, \alpha_{w0}^{-1})$. Diffuse gamma prior is placed on α_w and α_{w0}. There are ceremonial pieces of legislation, for which every legislator tends to vote "yes," and for such legislation δ_l tends to be large and positive. There are also pieces of legislation l for which the jth legislator may vote idiosyncratically, for which δ_{jl} may be large negative or positive (meaning that legislator votes uncharacteristically "no" or "yes," respectively). We don't assume a random effect δ_j, which would imply that the jth legislator tends to always vote one

way ("yes" or "no"), *independent* of the legislation.

We expect $\{\delta_{jl}\}$ to be sparse (or nearly sparse), and therefore on each we impose a shrinkage prior (in the same hierarchical manner discussed above for $\boldsymbol{\xi}_j$). We could impose similar random effects on the demographic data model, for representation of $\tilde{\mathbf{B}}_j$, but this proved unnecessary, as there we were model binary traits (*e.g.*, gender), rather than votes.

4.2.6 Model summary

Figure 4.1 provides a graphical representation of the model, with shaded and unshaded nodes indicating observed and latent variables, respectively. To assist with understanding the multiple components of the model, and their motivations, we provide an overarching summary below.

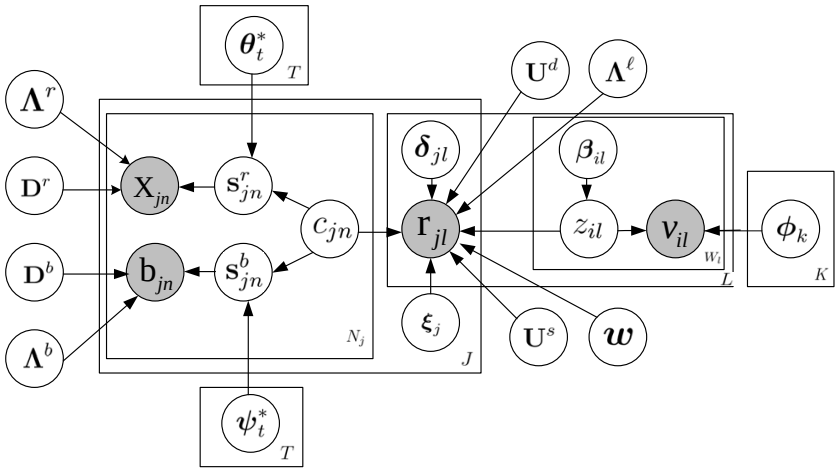


FIGURE 4.1: Graphical representation of the model.

The demographic data from distict j are represented by matrix factorizations (factor analysis), where column n of the factor-score matrices $\mathbf{S}_j^r$ (real data) and $\mathbf{S}_j^b$ (binary data) characterize person n in district j. For both matrix factorizations, the multiplicative gamma process is employed to encourage that only a relatively small number of factors are expected to define person choices.

We assume that the people (columns of $\mathbf{S}_j^r$ and $\mathbf{S}_j^b$) in each district will cluster into types of preferences. A truncated HDP is employed to infer this clustering. The probability of each of the T clusters is represented for district j by probability vector $\boldsymbol{\pi}_j$; $\{\boldsymbol{\theta}_t^*, \boldsymbol{\psi}_t^*\}_{t=1,T}$ represent the cluster-dependent parameters.

The vote $r_{jl} \in \{0, 1\}$ of congressman j on legislation l is characterized, via a probit matrix factorization, as an inner product between a feature vector for legislator j, $\boldsymbol{d}_j^\ell$, and a feature vector for legislation l, $\boldsymbol{s}_l^\ell$. To infer the relationship between how the congressman from district j votes relative to the interests of her/his constituents, we relate $\boldsymbol{d}_j^\ell$ to $\boldsymbol{\pi}_j$ via linear regression. We similarly wish to relate feature vector legislation $\boldsymbol{s}_l^\ell$ to the text of the associated legislation; in this case a regression is performed between $\boldsymbol{s}_l^\ell$ and $\boldsymbol{\beta}_l$, the latter the text-dependent distribution over topics (inferred here for simplicity via LDA, but any topic model may be used).

A key novelty of the model is a term $\boldsymbol{\xi}_j$, constituting a "random effect" in the regression between $\boldsymbol{\pi}_j$ and $\boldsymbol{d}_j^\ell$; $\boldsymbol{\xi}_j$ allows inference of the degree to which the congressman from district j appears to vote in a manner inconsistent with the preferences of her/his constituents. A random effect δ_l also allows identification of atypical legislation, linked to the text of the legislation via $\boldsymbol{\beta}_l$.

The regressions above were discussed in terms of $\boldsymbol{\pi}_j$ and $\boldsymbol{\beta}_l$. For technical reasons, discussed in the preceding sections, it is significantly more convenient to employ closely related surrogates $\tilde{\boldsymbol{\pi}}_j$ and $\tilde{\boldsymbol{\beta}}_l$; these are defined in terms of the relative counts of indicator variables c_{jn} and z_{il}, for person n in district j, and word i in document/legislation l.

4.3 Scaling Up: Variational Bayes and Stochastic Gradient Descent Inference

The Catalist data considers 2,969,925 people, and to handle data of this size we employ a mini-batch-based inference algorithm, stochastic variational Bayesian (VB)

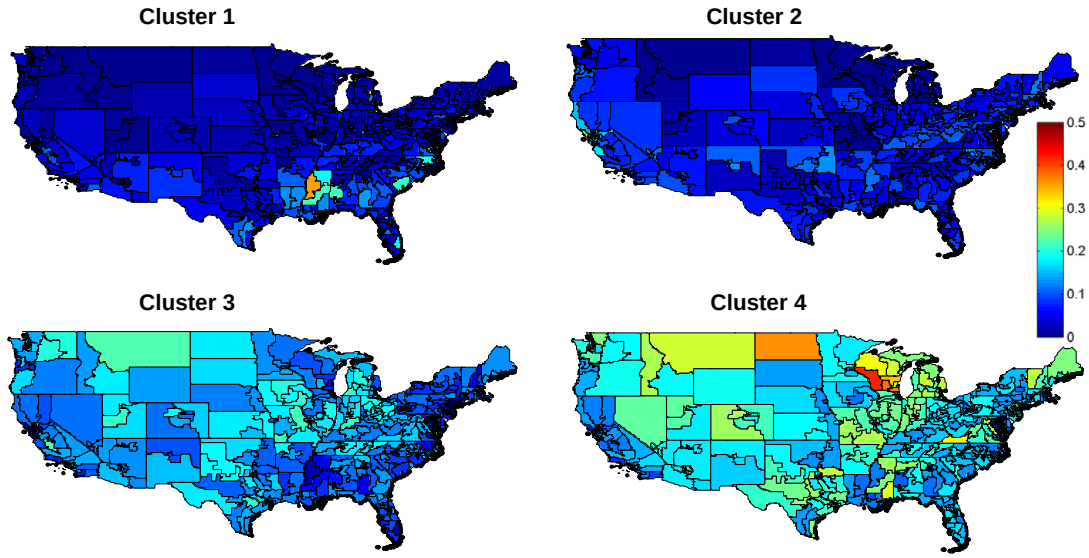


FIGURE 4.2: The expected probability of demographic clusters $\mathbb{E}[\pi_{tj}]$ $(t = 1, 2, 3, 4)$ for the 432 congressional districts across US (excluding Alaska and Hawaii).

Table 4.1: Center of clusters in original space $\{\Phi(\mathbb{E}[\mathbf{D}^b\Lambda^b\boldsymbol{\theta}^{\mu*}]), \mathbb{E}[\mathbf{D}^r\Lambda^r\boldsymbol{\psi}^{\mu*}]\}$. First 7 columns are the probability of answer "yes" for the corresponding attributes.

Cluster	Male	2006 Election	2008 Election	Black	Caucasian	Hispanic	Democrat	Republican	Age	Purchase Power
1	0.38	0.07	0.27	0.57	0.19	0.18	0.93	0.01	52	11509
2	0.39	0.63	0.87	0.29	0.55	0.08	0.93	0.04	49	76843
3	0.49	0.09	0.27	0.03	0.90	0.05	0.10	0.36	28	59999
4	0.49	0.07	0.22	0.04	0.88	0.06	0.11	0.34	48	74286

analysis [HBWP13, WPB11, BS12]. Unlike traditional VB inference [Bea03], which includes the whole dataset when updating the parameters, the stochastic variational inference method samples a subset of the data (mini-batch), and calculates a noisy natural gradient to optimize the variational objective function. Specifically, the individuals in the Catalist data are partitioned into $N^* = 15$ mini-batches, and each mini-batch contains individuals from all $J = 435$ congressional districts. The congressional votes and associate text are considered as a whole, since the size of that data is relatively small. The variational parameters specific to each individual mini-batch (in our case, the variational parameters associated with $\{\boldsymbol{s}^r_{jn}, \boldsymbol{s}^b_{jn}, c_{jn}\}$), are

called "local" parameters, denoted $\boldsymbol{\Theta}^l$. The remaining variational parameters, not specific to the mini-batch, are called "global" parameters, denoted $\boldsymbol{\Theta}^g$. At the hth iteration, the hth mini-batch is selected, and local variational parameters of the mini-batch $\boldsymbol{\Theta}^l$ are optimized; intermediate global parameters $\tilde{\boldsymbol{\Theta}}^g$ are then estimated with the most recent mini-batch. The new estimated global parameters are updated by computing the weighted average of previous value and $\tilde{\boldsymbol{\Theta}}^g$, $\boldsymbol{\Theta}^g \leftarrow (1-\omega_h)\boldsymbol{\Theta}^g + \omega_h \tilde{\boldsymbol{\Theta}}^g$, where $\omega_h \in (0,1)$ is the weight given to each new batch, and also called the learning rate. Following [HBWP13], we let $\omega_h = (a_3 + h)^{-b_3}$, where $b_3 \in (0.5, 1]$ controls the rate of decay of the contribution from old mini-batches and $a_3 \geqslant 0$ serves to slow down the decay rate for initial iterations. In the experiments, we set $a_3 = 1$ and $b_3 = 0.8$. One may employ the method proposed in [RWBX13] to adapt the learning step.

Details of the VB update equations are presented in the Supplementary Material. In the following, we examine two of the update equations, as they provide insight into how different parts of model relate to one another.

Variational Distribution for c_{jn}: The posterior-approximating distribution for the indicator variable c_{jn}, $q(c_{jn})$, is a categorical distribution with parameter $\tilde{\boldsymbol{\pi}}_{jn}$, the components of which satisfy $\tilde{\pi}_{jnt} \propto \exp\{\mathbb{E}[\log p(\boldsymbol{s}^r_{jn}|\boldsymbol{\theta}^*_t)] + \mathbb{E}[\log p(\boldsymbol{s}^b_{jn}|\boldsymbol{\psi}^*_t)] + \mathbb{E}[\log(\pi_{jt})] + \sum_{l=1}^{L} \mathbb{E}[\log p(\tilde{r}_{jl}|c_{jn} = t, -)]\}$. The term $\mathbb{E}[\log(\pi_{jt})]$ characterizes the clustering characteristics of district j, where $p(\boldsymbol{s}^r_{jn}|\boldsymbol{\theta}^*_t)$ and $p(\boldsymbol{s}^b_{jn}|\boldsymbol{\psi}^*_t)$ characterize the properties of cluster t. The term $p(\tilde{r}_{jl}|c_{jn} = t, -)$ characterizes the latent real matrix associated with the binary legislative votes of the representative from district j on all L pieces of legislation.

Variational Distribution for z_{il}: The approximating distribution for the latent topic associated with word i in legislation l, $q(z_{il})$, is a categorical distribution with parameter $\tilde{\boldsymbol{\beta}}_{il}$, and $\tilde{\beta}_{ilk} \propto \phi_{v_{il},k} \exp\{\mathbb{E}[\log\ \beta_{lk}] + \sum_{j=1}^{J} \mathbb{E}[\log p(\tilde{r}_{jl}|z_{il} = k, -)]\}$. Note

that this update equation is affected by the fit of the word to the topic (first term) plus the impact of that topic to the roll-call votes from the J legislators (second term).

4.4 Experimental Results

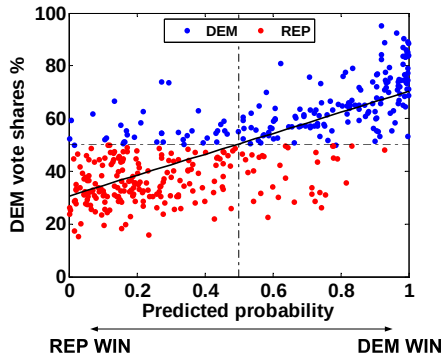
We employ the proposed model on the Catalist data discussed above ($P^r = 28$ and $P^b = 51$; 2,969,925 total people across the $J = 435$ Congressional districts, with typically 5,000 to 7,000 people from each district). The constituency characteristics are summarized in the Supplementary Material. The roll-call data are from the 111th US Congress (January 3, 2009 - January 3, 2011), consistent with the time period of the Catalist data. Roll-call votes on a total of $L = 802$ bills are considered. For the text of the bill, we follow the n-gram preprocessing procedure described in [GB11], and obtain a bag of words with vocabulary size $V = 4743$.

The election for the 112th Congress took place on November 2, 2010, and we used votes on bills in the 111th Congress that occurred before then to examine the party affiliation of each winner of that election, and examine the vote share, relative to the roll-call data.

For model initialization, we first consider each data source separately. For example, we take a subset of the Catalist data, to infer $\mathbf{D}^r$, $\mathbf{D}^b$, $\mathbf{\Lambda}^r$ and $\mathbf{\Lambda}^b$. Then K-means was performed on the learned latent features for the individuals, to initialize the HDP model. Similarly, LDA was first applied to the legislative text to infer initial topics. The results are repeatable for different related forms of this initialization.

We set $K = 30$, $T = 15$, $K^r = K^b = 20$, $K^l = 10$ and the MGP hyperparameter is $a_1 = 2$. The Catalist data are randomly partitioned into 15 mini-batches, each of size 197,995. We implemented the proposed model in MATLAB, and ran the code on a PC with 8 cores, 3.2GHz CPU, and 128 GB memory. We considered 40 VB iterations per mini batch, and the total computation time for these data was 16

FIGURE 4.3: Left column: Probability of Democratic win vs the vote share received for Democratic candidates. The solid line is a linear regression fit with vote share and predicted probability. Right column: Actual (empirical) probability of Democratic candidates win in each predicted probability bin.



Predicted Prob. bin of DEM win	Actual Prob. of DEM win
0-0.2	0.09
0.2-0.4	0.21
0.4-0.6	0.43
0.6-0.8	0.74
0.8-1	0.96

hours.

4.4.1 Inferred district-level characteristics and Congressional election results

Using the full model, we infer $\mathbb{E}[\boldsymbol{\pi}_j]$, the expected probability of demographic clusters for district j. The characteristics of cluster t may be interpreted by mapping the cluster center $\{\mathbb{E}[\boldsymbol{\theta}_t^{\mu*}], \mathbb{E}[\boldsymbol{\psi}_t^{\mu*}]\}$ back to the original data space $\{\mathbb{E}[\mathbf{D}^r\Lambda^r\boldsymbol{\theta}_t^{\mu*}], \Phi(\mathbb{E}[\mathbf{D}^b\Lambda^b\boldsymbol{\psi}_t^{\mu*}])\}$, where $\Phi(\cdot)$ is the cumulative probability function of standard normal (from the probit model). In Figure 4.2, we plot $\mathbb{E}[\pi_{tj}]$ of four example clusters, for 432 congressional districts (excluding Alaska and Hawaii). The corresponding $\{\mathbb{E}[\mathbf{D}^r\Lambda^r\boldsymbol{\theta}_t^{\mu*}], \Phi(\mathbb{E}[\mathbf{D}^b\Lambda^b\boldsymbol{\psi}_t^{\mu*}]\}$ are shown in Table 4.1 (this table provides mean values of a subset of Catalist parameters). From Table 4.1 and Figure 4.2, individuals in Clusters 1 and 2 are more likely to be Democrats. Cluster 1 seems to capture low-income Black and Hispanic Democrats, with poor turnout in the past election. In contrast, Cluster 2 is more likely to include high-income Democrats, with high turnout in previous elections. Cluster 1 is found to have high probability in many of the southern districts, especially these close to the border. Cluster 2 tends to appear in metropolitan areas, such as San Francisco, Los Angels, DC and New York. In a similar manner, Clusters 3 and 4 are more likely to include whites and Republicans (or undeclared voters).

Age and purchasing power (in U.S. dollars) seem to distinguish Clusters 3 and 4.

To further assess how well the latent estimates capture constituent preferences, we examine the ability of the model to predict the party affiliation of the district's House member, based on the constituent characteristics in the Catalist datafile. Specifically, we use $\mathbb{E}[\boldsymbol{\pi}_j]$ as a feature vector, and build a linear probit-regression classifier (similar results can be obtained with other probabilistic classifiers), where shrinkage is imposed on the regression weights, using the same prior as imposed in the full model on $\boldsymbol{\xi}_j$.

In Figure 4.3, we plot the probit-regression-based probability that a given district will select a Democratic legislator, and along the vertical axis is plotted the fraction of vote share received in the district for the Democratic candidate (in the 2010 election). We consider the 406 (of 435) districts for which there was a contested election, with two candidates. We partitioned the districts into 5 folds, and iteratively train on 4 folds and test on the rest. Note, for example, when the model predicted that the probability of a Democratic win was 0.5, the fraction of vote received on average was about 50%. In the table in Figure 4.3, we note that the predictions of the model are in close alignment with actual district-level voting. These results indicate that the characterization of people in each district based on the Catalist data is a good representation of voter preferences. This provides further insight into why the Catalist data are useful for inferring more-confident prediction of roll call votes based on held-out text of the legislation (see Table 4.2), and also why a legislator tends to perform poorly in the next election when her voting record is inconsistent with the district-level preferences

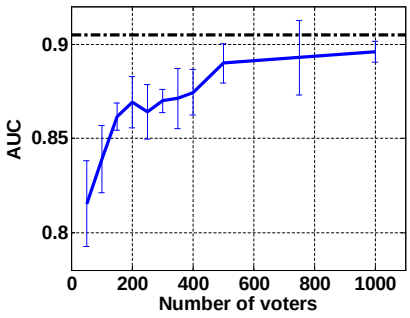


FIGURE 4.4: AUC versus number of voters in each districts. Black dash line corresponds to using all the data.

(reflected by large $\boldsymbol{\xi}_j$, as depicted in Figure 4.7).

It is of interest to examine the quality of the model as a function of the number of people per district we have demographic data from. Specifically, we train the whole model with a subset of randomly selected voters from Catalist dataset. We use $\mathbb{E}[\boldsymbol{\pi}_j]$ inferred from the subset as a feature vector, and perform the same prediction experiment discussed above. AUC (area under ROC curve) is employed as the metric for assessing the performance. In Figure 4.4, we plot the AUC as a function of average number of voters selected per district. The result is the average of 5 runs, and the error bar correspond to one standard deviation.

4.4.2 Insights on relationships between constituents and representatives

In the political science literature [CJR04] and in recent machine learning research [GB11], it has been assumed that the latent space of the legislators and legislation is one-dimensional based on roll call votes (*i.e.*, feature vectors like $\boldsymbol{d}_j^\ell$ and $\boldsymbol{s}_l^\ell$ are *assumed* to be one-dimensional). Via the MPG prior on $\boldsymbol{\Lambda}^\ell$, we may *infer* the dimensions of these vectors. In Figure 4.5, we depict $\mathbb{E}[\text{diag}(\boldsymbol{\Lambda}^\ell)]$, which indicates that there is indeed one dominant latent dimension, but also two additional weaker dimensions.

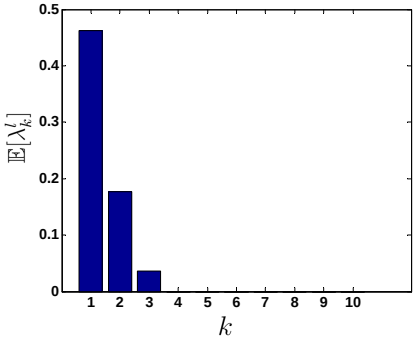
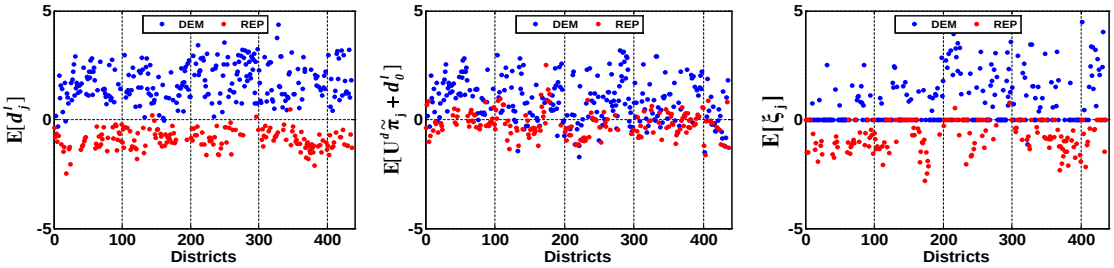


FIGURE 4.5: ($\mathbb{E}[\text{diag}(\boldsymbol{\Lambda}^\ell)]$.

To illustrate the connection of the dominant latent feature dimension to the characteristics of the representatives in each district, and to the characteristics of the people they represent, in Figure 4.6(a) we plot the principal dimension of $\boldsymbol{d}_j^\ell$ for each legislator, and note that Democrats tend to be positive in this dimension and Republicans negative. This result agrees with the ideal point obtained with the model in [CJR04, GB11].

In Figure 4.6(b) we plot the principal dimension of $\mathbb{E}[\mathbf{U}^d\tilde{\boldsymbol{\pi}}_j + \boldsymbol{d}_0^\ell]$, which within the model captures the roll-call-related preferences of the people who live in district j. Note that the Republican representatives (Figure 4.6(a)) appear to often be more negative in this dimension than their constituents (Figure 4.6(b)). Finally, in Figure 4.6(c) we plot $\mathbb{E}[\boldsymbol{\xi}_j]$ in the principal dimension. Recall that $\boldsymbol{\xi}_j$ in $\boldsymbol{d}_j^\ell = \mathbf{U}^d\tilde{\boldsymbol{\pi}}_j + \boldsymbol{d}_0^\ell + \boldsymbol{\xi}_j$ controls the degree to which the feature vector $\boldsymbol{d}_j^\ell$ associated with legislator j deviates from the characteristics of her constituents, reflected by $\tilde{\boldsymbol{\pi}}_j$. Moreover, a shrinkage prior was imposed on $\boldsymbol{\xi}_j$, and therefore large $|\boldsymbol{\xi}_j|$ is reflective of legislators who may be voting in a manner that is not well linked to the people who live in their district (from the standpoint of the Catalist data). Note that $\boldsymbol{\xi}_j$ tends to be sparse, implying that representatives typically vote in line with their constituents, but there are also often significant non-zero $\boldsymbol{\xi}_j$.



(a) (b) (c)

FIGURE 4.6: (a) : Principal dimension of $\mathbb{E}[\boldsymbol{d}_j^\ell]$. The horizontal axis is the index of districts (alphabetically ordered). (b): Principal dimension of $\mathbb{E}[\mathbf{U}^d\tilde{\boldsymbol{\pi}}_j + \boldsymbol{d}_0^\ell]$. (c): Principal dimension of $\mathbb{E}[\boldsymbol{\xi}_j]$.

We further examine the relationship between principal dimension of $\boldsymbol{\xi}_j$ (denoted ξ_{1j}) and the fraction of voter share for the jth legislator in the 2010 election. We focus on Democratic House members, as these were the ones for which there was significant turnover in that election. In Figure 4.7, we use box plots for two groups of Democratic representatives: those with $|\mathbb{E}[\xi_{1j}]| \geqslant 0.1$ and those with $|\mathbb{E}[\xi_{1j}]| < 0.1$.

The 0.1 threshold is illustrative, and many related small thresholds yield similar results. Members who voted in a way that the model infers as aligned with the interests of their constituents (small $|\mathbb{E}[\xi_{1j}]|$) on average received a 15% larger share of the election vote than those legislators with relatively large $|\mathbb{E}[\xi_{1j}]|$. Notice a small number of legislators with high value of $\mathbb{E}[\xi_{1j}]$ also receive high vote share. These representatives are mainly from the less competitive districts. For example, Nydia Velazquez (NY-12), one of the two outliers, was challenged only by a third party candidate.

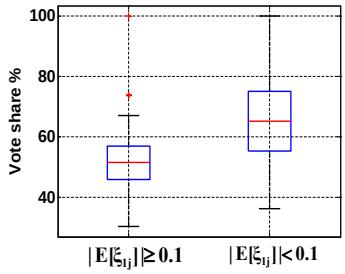


FIGURE 4.7: Vote share received for two groups of Democratic congressmen: those with $|\mathbb{E}[\xi_{1j}]| \geqslant 0.1$ and those with $|\mathbb{E}[\xi_{1j}]| < 0.1$.

4.4.3 Analysis of the legislative topics in latent space

We examine the relationship between the topics of the legislation and the latent space associated with the roll-call vote. Specifically, $\delta_l = \boldsymbol{w}^T \tilde{\boldsymbol{\beta}}_l + w_0$ is a random effect associated with legislation l, and note that it is directly linked to the topic distribution on the legislation $\tilde{\boldsymbol{\beta}}_l$. The feature vector associated with the legislation is $\boldsymbol{s}_l^\ell = \mathbf{U}^s \tilde{\boldsymbol{\beta}}_l + \boldsymbol{s}_0^\ell$, and we here consider $\mathbb{E}[\boldsymbol{s}_l]$ in the dominant (first) dimension, denoted $\mathbb{E}[s_{1l}]$. Based on Figure 4.6(a), positive values of $\mathbb{E}[s_{1l}]$ imply that the legislation is typically favored by Democrats, and negative values by Republicans.

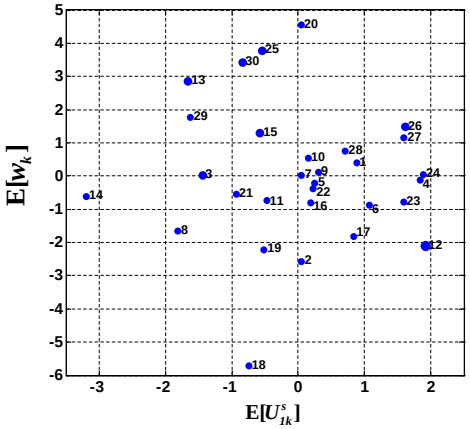
The kth component of the first row of $\mathbf{U}^s$, denoted U^s_{1k}, dictates the degree to which topic k contributes to s_{1l}. Further, component k of $\boldsymbol{w}$, w_k, dictates the degree to which topic k contributes to δ_l. Positive/negative values of U^s_{1k} correspond to topics favored by Democrats/Republicans, and positive/negative w_k correspond to topics that most congressman tend to vote "yes"/"no."

In Figure 4.8 we show the topics in the space $(\mathbb{E}[U^s_{1k}], \mathbb{E}[w_k])$, and also depict

Table 4.2: Comparison between proposed method and ideal point probit model from [GB11]. Shown are the number of votes in each probability bin, and the empirical probability of being correct in the prediction.

	Proposed model		Ideal point probit model	
Probit Confidence Bin	Votes in Confidence Bin	Empirical Probability	Votes in Confidence Bin	Empirical Probability
0.5-0.6	15821	0.54	16231	0.54
0.6-0.7	15241	0.63	17339	0.61
0.7-0.8	15170	0.7	17793	0.72
0.8-0.9	21756	0.81	22998	0.84
0.9-1	258367	0.98	251994	0.98
Pred. log-likelihood	-0.197		-0.204	

most-probable words associated with six example topics. During this time period, the Iraq and Afghanistan wars, which were started under a Republican president, tended to be aligned with the interest of the Republican Party (negative ($\mathbb{E}[U^s_{1k}]$); see Topic 3. By contrast, Topic 20, about health care, children and military veterans, tended to be favored irrespective of party (large positive $\mathbb{E}[w_k]$).



TOPIC 3	TOPIC 6	TOPIC 18
war	credit	related
military	loan	measure
force	income	recovery
freedom	tax	expense
international	insurance	requires
TOPIC 19	**TOPIC 20**	**TOPIC 26**
financial	child	community
transfer	care	American
property	medical	help
account	health care	opportunity
benefit	veteran	school

FIGURE 4.8: Left column: Regression weights of topics. Right column: Selected topics with the top-five most probable words shown.

4.4.4 Prediction based on legislative text

We consider prediction of the votes of each legislator on held-out legislation, where the votes are predicted entirely by the text of the held-out legislation (the topic model infers $\tilde{\boldsymbol{\beta}}_l$ for new legislation, from which $\boldsymbol{s}^\ell_l$ and δ_l are estimated, and used to predict the probability of a particular vote). This experiment serves as a measure of model fitness.

In [GB11] the authors developed a model like that in (4.3), except that they did not have access to district-level constituency characteristics, like the Catalist data considered here. Therefore, in [GB11] the authors used the model in (4.3), except that $\boldsymbol{d}_j^\ell$ was drawn i.i.d. from a symmetric multivariate Gaussian distribution, rather than being related to the constituency information (the latter implemented in the proposed model by relating $\boldsymbol{d}_j^\ell$ in (4.3) to $\boldsymbol{\pi}_j$). As the ideal point model in [CJR04, GB12], latent space s_l^l and the random effect δ_l are not associated with the legislative text, thus we cannot evaluate how well these two models predict votes for hold out legislations.

The roll call votes and associated text of legislation of the 111th US House of Representatives are partitioned into 6 folds. We iteratively train the model using five folds, and test on the sixth. The presented result is an aggregation of all six folds. Prediction confidence [SDC13] and accuracy are employed as metrics. Specifically, for each held-out vote by legislator j on legislation l, the model yields a probability of "yes" , $p(r_{jl} = 1|-)$ and a probability of "no", $1-p(r_{jl} = 1|-)$. We take $max\{p(r_{jl} = 1|-), 1 - p(r_{jl} = 1|-)\}$ for each held out vote, irrespective of whether the actual prediction is "yes" or "no," and place them into the corresponding probability bin, with bins ranging from $[0.5 - 0.6]$ to $[0.9 - 1]$. We wish to examine whether the prediction confidence matches empirical results. For example, if we examine all votes for which the model predicts the vote with confidence in the range $[0.7 - 0.8)$, we would expect the model should be able to correctly predict the vote between 70%-80% of the time. For the test legislations, we also compute the predictive log-likelihood $\log p(r_{test}|r_{train})$, which averaged for all six folds. In Table 4.2, we compare the prediction confidence and of the proposed model and that in [GB11](probit link instead of logistic link). We observe that both models are "correct," in that the predicted confidence of the vote matches the empirical data (*e.g.*, for the proposed model, 258,367 of the held-out votes were predicted with a confidence of 0.9 to 1, and

the model was correct in its prediction 98% of the time). In comparing the proposed model and that in [GB11], note that the former places 6,000 more votes in the 0.9 to 1 confidence bin and the predictive log-likelihood also improved, suggesting that the use of constituency (Catalist) data yields more confident predictions in legislator votes, and that confidence is vindicated experimentally.

Improvement manifested by our model is most prominent for contested legislation and unusual districts. Specifically, most of the 6,000 votes discussed above are for closely contested bills (those receiving less than 400 yea votes, corresponding to 267 out of 802 bills). The congressmen for which the model provides most improvement in vote prediction are among Republicans in districts dominated by Democratic constituents, such as Ileana Ros-Lehtinen (FL-18) and Michael Castle (DE). Their district-level characteristics (larger proportion of Democratic voters) adjust the ideal points toward Democratics, yielding more-confident predictions.

4.5 Conclusions

Binary matrix factorization is employed for analysis of roll-call data, with latent features associated with legislation informed by a topic model of the legislative text, and the latent features of each legislator informed by a statistical model of the people living in their district. The model is employed in a new manner to uncover insights into the workings of electoral representation, based on large-scale data, here specific to the U.S. Congress. The model is shown to produce improved prediction of votes on held-out legislation based on the text of the legislations, and demonstrates the electoral consequences of legislators failing to represent the preferences of their constituents.

5

Conclusion and Future Directions

In this thesis, we employ Bayesian nonparametric methods to develop three Bayesian hierarchical models in three fields ,which are Hyperspectral image analysis , infectious disease and vote behavior. The key contribution of this thesis is summarized as following.

- We employed sparsity for hyperspectral image analysis in a Bayesian manner. Two constructions, one based upon use of shrinkage priors and the other based on a beta-Bernoulli construction, are proposed and compared. Moreover, unlike in previous endmember studies, which were based on the spectral signature alone, in the analysis considered here the dictionary elements (analogous to endmembers) are learned while taking into account both spatial and spectral information. Another unique aspect of the work presented here is that rather than analyzing the entire datacube directly, we have processed a significantly downsampled version. Specifically, we have performed the analysis based on observing a small fraction of the voxels, selected uniformly at random. It was demonstrated that one may accurately recover the missing data, even in the presence of substantial wavelength-dependent noise.

- We developed a statistical model for analysis of the time-dependent symptom scores provided by a large group of undergraduate college students. Unlike almost all studies of data related to infection transfer, the model has operated directly on the observed symptoms, and the state of the students were assumed to be latent. The community-to-person mechanism for pathogen transfer has been modeled in terms of a SIS analysis, and computations have been performed using Bayesian (MCMC) methods. A detailed characterization of the data and the scientific questions that have motivated this study are discussed; a comprehensive answering of these questions with the available data has been provided.

- We use congressional roll-call votes, legislative text, and individual-level constituency information about 300 million people to build a novel model to estimate the latent ideological preferences of both legislators and the districts they represent, allowing us to evaluate We would like to emphasis the key contribution of proposed model is to examine the extent to which legislators voting records are aligned with constituent preferences. We show that Democratic legislators who were more ideologically distant from their districts received a lower vote share in the subsequent election. While the current analysis (and planned extensions) are unique contributions to the field of political science, the model is one that has potential for application to other domains (e.g., consumer preferences).

These contribution motivate several directions for the future work.

- For the hyperspectral image analysis, first future direction is combining prior knowledge with the *in situ* dictionary-learning approach developed here. In all

examples the analysis has been employed with no *a priori* training data, while in practice one would expect to have available a database of potential signatures (not necessarily complete, but still providing useful prior information). Imposition of such prior knowledge is anticipated to substantially improve modeling performance. Second direction of future research concerns examination of material classification based upon hyperspectral datacubes recovered from massively downsampled measurements. This line of research is critical, as the principal objective of hyperspectral measurements concerns material characterization. Based upon the quality of the recovered data, as discussed in this paper, it is anticipated that high-quality material characterization will be achieved.

- For the proposed model on infectious disease, first future direction is to provide a more detailed analysis of pathogen transfer as a function of virus type can be examined. In this paper we have presented results in this direction, but more explicit modeling could be performed (not necessarily at the symptom level, but after the responsible virus has been identified by RT PCR). Second future direction is to analyze the time trajectory of the gene expression data as the host responds to the (known) virus. The gene expression data from this study have only been employed here in a limited manner, as the focus has been on self-reported symptom scores. The data from this study offers the potential for similar studies as [CZW$^+$11] on data from individuals who became ill in natural settings. We have preliminary results in this direction on these data, which are encouraging and will be presented elsewhere.

- For the analysis of vote behavior, one potential direction is to examine whether random effect term, which represent the ideologically distant between congressmen and their district is robust to consideration of factors like party, seniority,

campaign spending, and the like. Another direction will be extend the model beyond political science, such as online shopping dataset. We have mentioned this direction in the paper, it would be interesting to valid this direction with real dataset.

Appendix A

Appendix for Bayesian modeling of temporal properties of infectious disease in a college student population

A.1 MCMC Update Equations

The full posterior distribution can be approximate via Gibbs sampler, with Metropolis-Hastings updates for a subset of parameters. We briefly describe how to sample some of the most interesting parameters based on their conditional posterior distribution.

Sampling from the latent states

Sampling from the latent states z_{nt} is achieved by forward and backward sampling method [Mur02, YK03a, JW13, S.Y10]. Define the forward equation

$$\alpha_{nt}(m, d) = p(\mathbf{y}_{n1}^{t}, z_{nt} = m, d_{nt} = d), m \in \{S, I\}, d = 1, .., D_{max}$$

where $\mathbf{y}_{n1}^{t} = [\mathbf{y}_{n1}, ..., \mathbf{y}_{nt}]$ and d_{nt} is the number of days students n left in infective state I after day t. For the susceptible state S, d_{nt} is not necessary and omit for brevity. Let denotes $E_{nt} = \{z_{nt}, d_{nt}\}$, the forward equation can be calculated from the following induction function.

For the Markovian states $E_{nt} = \{S\}$, transition into states $\{S\}$ at time t takes place either from $\{I, 1\}$ or $\{S\}$ at time $t-1$

$$\alpha_{nt}(S) = (\alpha_{nt-1}(I,1) + \alpha_{nt-1}(S)p(z_{nt} = S|z_{nt-1} = S))p(y_{nt}|z_{nt} = S)$$

For the semi-Markovian state $E_{nt} = \{I, d\}$, transition into states $\{I, d\}$ at time t takes place either from $\{I, d+1\}$ or $\{S\}$ at time $t-1$

$$\alpha_{nt}(I,d) = (\alpha_{nt-1}(I,d+1) + \alpha_{nt-1}(S)p(z_{nt} = I|z_{nt-1} = S)p(d_{nt} = d))p(y_{nt}|z_{nt} = I)$$

Then we can sample E_{nt} (the state z_{nt} and the duration d_{nt}) from the backward sampling step. For $t = T$, sample E_{nt}

$$p(z_{nT} = I, d_{nT} = d|y_{n1}^T) = \frac{\alpha_{nT}(I,d)}{\alpha_{nT}(S) + \sum_{d=1}^{D_{max}} \alpha_{nT}(I,d)}$$

$$p(z_{nT} = S|y_{n1}^T) = \frac{\alpha_{nT}(S)}{\alpha_{nT}(S) + \sum_{d=1}^{D_{max}} \alpha_{nT}(I,d)}$$

For $t \in T-1, ..., 1$, sample E_{nt}

$$\begin{aligned}
& p(z_{nt} = I, d_{nt} = d|y_{n1}^t, E_{nt+1}) \\
= & \frac{\alpha_{nt}(I,d)p(z_{nt} = I, d_{nt} = d|E_{nt+1})}{\alpha_{nt}(S)p(z_{nt} = S|E_{nt+1}) + \sum_{d=1}^{D_{max}} \alpha_{nt}(I,d)p(z_{nt} = I, d_{nt} = d|E_{nt+1})} \\
& p(z_{nt} = S|y_{n1}^t, E_{nt+1}) \\
= & \frac{\alpha_{nt}(S)p(z_{nt} = S|E_{nt+1})}{\alpha_{nt}(S)p(z_{nt} = S|E_{nt+1}) + \sum_{d=1}^{D_{max}} \alpha_{nt}(I,d)p(z_{nt} = I, d_{nt} = d|E_{nt+1})}
\end{aligned}$$

The above method need to specify the maximum number of duration D_{max} in order to avoid infinite number of states E_{nt}. We may employ the beam sampling idea developed in [GSTG08, DWW12] to avoid setting D_{max}. The main idea of beam sampling is introducing auxiliary random variables $\mathbf{u}_{n1}^t$ for slice sampling. The forward

equation is modified as

$$\begin{aligned}\alpha^*_{int}(E_{nt}) &= p(E_{nt}, \mathbf{y}^t_{n1}, \mathbf{u}^t_{n1}) \\ &= \sum_{E_{nt-1}} I(0 < u_{nt} < p(E_{nt}|E_{nt-1}))\alpha^*_{nt-1}(E_{nt-1})p(y_{nt}|E_{nt})\end{aligned}$$

The backward sampling part is

$$p(E_{nt-1}|E_{nt}, \mathbf{y}^T_{n1}, u) \propto I(0 < u_{nt} < p(E_{nt}|E_{nt-1}))\alpha^*_{nt-1}(E_{nt-1})$$

where $I(.)$ is the indicator function. $I(g(u_{nt})) = 1$ if $g(u_n t)$ is true and $I(g(u_{nt})) = 0$ otherwise.

Sampling the correlation matrix

The parameter extension method introduced in [ZBB06] is employed to sample the correlation matrix $\mathbf{\Sigma}_I$. An unrestricted covariance matrix $\mathbf{\Sigma_1} \sim Wishart(m_1, \mathbf{V}_1)$, which can be decomposed as $\mathbf{\Sigma_1} = \mathbf{L}^{1/2}\mathbf{\Sigma}_I\mathbf{L}^{1/2}$, where $\boldsymbol{L}$ is the diagonal of the matrix with diagonal elements equivalent to the diagonal of $\mathbf{\Sigma_1}$. The prior for correlation matrix $\mathbf{\Sigma}_I$ is as following,

$$P(\mathbf{\Sigma}_I, \mathbf{L}) = Jacobian_{\mathbf{\Sigma}_1 \to (\mathbf{\Sigma}_I, \mathbf{L})} P(\mathbf{\Sigma}_1)$$

where $Jacobian_{\mathbf{\Sigma}_1 \to (\mathbf{\Sigma}_I, \mathbf{L})} = \prod_{i=1}^{J} q_i^{\frac{J-1}{2}}$ is the Jacobian transformation from $\mathbf{\Sigma_1}$ to $(\mathbf{L}, \mathbf{\Sigma}_I)$. Then the MH algorithm for sampling posterior distribution of $\mathbf{\Sigma}_I$ is as follows: at iteration t, generate the candidate values $\mathbf{\Sigma}^*_I$ from $\mathbf{\Sigma^*_1} = \mathbf{L}^{*1/2}\mathbf{\Sigma}^*_I\mathbf{L}^{*1/2} \sim Wish(m_1, \mathbf{V}_1)$, accept the value with probability $\alpha = min\{1, \frac{p(\mathbf{L}^*, \mathbf{\Sigma}^*_I|-)}{p(\mathbf{L}^t, \mathbf{\Sigma}^t_I|-)} \frac{q(\mathbf{\Sigma}^t_I|\mathbf{\Sigma}^*_I)}{q(\mathbf{\Sigma}^*_I|\mathbf{\Sigma}^t_I)}\}$, where $q(.|\mathbf{\Sigma}^t_I)$ is the proposal distribution given by product the jacobian term and Wishart density $Wishart(m_0, \mathbf{\Sigma}^t_I)$. Sampling $\mathbf{\Sigma_S}$ is performed using a similar procedure.

Sample γ_{nt}

Define $b_{nt} = 1$ if $z_{nt-1} = S$ and $z_{nt} = I$, $b_{nt} = 0$ if $z_{nt-1} = S$ and $z_{nt} = S$. b_{nt} is treat as missing data, if $z_{nt-1} = I$ and $z_{nt} = I$. We use q_{nt} to denote the missing data, with $q_{nt} = 0$ refers to missing and $q_{nt} = 1$ otherwise. We can sample γ_{nt} as following,

$$\gamma_{nt} \sim \mathcal{N}_{(0,+\infty)}(\textstyle\sum_{i=1}^{3} \gamma_t^{(i)} + a_n, 1), \text{ if } b_{nt} = 1$$
$$\gamma_{nt} \sim \mathcal{N}_{(-\infty,0)}(\textstyle\sum_{i=1}^{3} \gamma_t^{(i)} + a_n, 1), \text{ if } b_{nt} = 0$$

$\mathcal{N}_{(0,+\infty)}$ and $\mathcal{N}_{(-\infty,0)}$ are the truncated normal distributions with truncation level $(0, +\infty)$ and $(-\infty, 0)$.

Sample $\gamma_t^{(1)}$ and $\gamma_t^{(2)}$

Sampling $\gamma_t^{(1)}$ and $\gamma_t^{(2)}$ are achieved via forward filtering and backward sampling method [CK94, FS94]. Here we detail the update equations for sampling $\gamma_t^{(2)}$ and sampling $\gamma_t^{(1)}$ is performed in the similar way. Define $\mathbf{F} = [1, 0, 1, 0, 1, 0]^T$, $\mathbf{G} = \begin{pmatrix} \mathbf{J}(w) & 0 & 0 \\ 0 & \mathbf{J}(2w) & 0 \\ 0 & 0 & \mathbf{J}(3w) \end{pmatrix}$, then $\gamma_t^{(2)} = \mathbf{F}^T\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_t = \mathbf{G}\boldsymbol{\theta}_{t-1} + \boldsymbol{\epsilon}_t$. In the forward filtering step, assume $\boldsymbol{\theta}_0 \sim \mathcal{N}(\mathbf{m}_0^{(2)}, \mathbf{C}_0^{(2)})$, it can be shown the posterior at time t is

$$\boldsymbol{\theta}_t \sim \mathcal{N}(\mathbf{m}_t^{(2)}, \mathbf{C}_t^{(2)})$$

where $\mathbf{m}_t^{(2)} = \mathbf{C}_t^{(2)}(\sum_{n=1}^{N} \mathbf{F}\hat{\gamma}_{nt}^{(2)} q_{nt} + \mathbf{R}_t^{(2)-1}\mathbf{a}_t^{(2)})$, $\mathbf{C}_t^{(2)} = \mathbf{R}_t^{(2)} - \mathbf{A}_t^{(2)} Q_t^{(2)} \mathbf{A}_t^{(2)T}$, $\mathbf{A}_t^{(2)} = \sqrt{N_t}\mathbf{R}_t^{(2)}\mathbf{F}Q_t^{(2)-1}$, $Q_t^{(2)} = 1 + N_t\mathbf{F}^T R_t^{(2)}\mathbf{F}$, $\mathbf{a}_t^{(2)} = \mathbf{G}\mathbf{m}_{t-1}^{(2)}, \mathbf{R}_t^{(2)} = \mathbf{G}\mathbf{C}_{t-1}^{(2)}\mathbf{G}^T + \mathbf{W}_t$ and $\hat{\gamma}_{nt}^{(2)} = \gamma_{nt} - a_n - \gamma_t^{(1)} - \gamma_t^{(3)}$.

In the backward sampling step, first sample $\boldsymbol{\theta}_T \sim \mathcal{N}(\mathbf{m}_T^{(2)}, \mathbf{C}_T^{(2)})$ and then for day $T-1$ to day 1, sample

$$\boldsymbol{\theta}_t \sim \mathcal{N}(\hat{\mathbf{m}}_t^{(2)}, \hat{\mathbf{C}}_t^{(2)})$$

where $\hat{\mathbf{m}}_t^{(2)} = \mathbf{m}_t^{(2)} + \hat{\mathbf{B}}_t^{(2)}(\boldsymbol{\theta}_{t+1} - \mathbf{a}_{t+1}^{(2)})$, $\hat{\mathbf{C}}_t^{(2)} = \mathbf{C}_t^{(2)} - \hat{\mathbf{B}}_t^{(2)}\mathbf{R}_{t+1}^{(2)}\hat{\mathbf{B}}_t^{(2)T}$, $\hat{\mathbf{B}}_t^{(2)} = \mathbf{C}_t^{(2)}\mathbf{G}^T\mathbf{R}_{t+1}^{(2)-1}$. $\mathbf{W}_t$ is a block diagonal covariance matrix with each block equals to $\boldsymbol{\Sigma}_{\theta_r}$.

Sample r_{nt} and τ

Following the sampling algorithm in [AC93b], sample $\mathbf{r}_{nt} \sim \mathcal{N}_{(\boldsymbol{\tau}_{\mathbf{y}_{nt}-\mathbf{1}}, \boldsymbol{\tau}_{\mathbf{y}_{nt}})}(\boldsymbol{\mu}_{z_{nt}}, \boldsymbol{\Sigma}_{z_{nt}})$ where $\boldsymbol{\tau}_{\mathbf{y}_{nt}-\mathbf{1}}$ and $\boldsymbol{\tau}_{\mathbf{y}_{nt}}$ are the truncation level of the multivariate normal. For $m = 1, ..., M-1$, sample τ_{jm} from uniform distribution with interval $[max(max(r_{ntj} : y_{ntj} = m-1), \tau_{jm-1}), min(min(r_{ntj} : y_{ntj} = m, \tau_{jm+1}))]$.

A.2 Prediction

Denote the current data as $\mathbf{y}_{n1}^t = \{\mathbf{y}_{n1}, ..., \mathbf{y}_{nt}\}$, we may derive the one step predictive probability for student n as following,

$$p(z_{nt+1} = I | \mathbf{y}_{n1}^t, \Omega) = \frac{\sum_{d_{nt+1}=1}^{D_{max}} p(z_{nt+1} = I, d_{nt+1}, \mathbf{y}_{n1}^t | \Omega)}{\sum_{d_{nt+1}=1}^{D_{max}} p(z_{nt+1} = I, d_{nt+1}, \mathbf{y}_{n1}^t, |\Omega) + p(z_{nt+1} = S, \mathbf{y}_{n1}^t | \Omega)} \quad \text{(A.1)}$$

where Ω is the model parameter learned from current data $\mathbf{y}_{n1}^t$ and $d = 1, .., D_{max}$ is the number of days left in infective states at time $t+1$. If we define $\hat{\alpha}_{nt+1}(I, d) = p(z_{nt+1} = I, d_{nt+1} = d, \mathbf{y}_{n1}^t | \Omega)$ and $\hat{\alpha_{nt+1}}(S) = p(z_{nt+1} = S, \mathbf{y}_{n1}^t | \Omega)$, the induction equation for $\hat{\alpha_{nt+1}}$ can be derived.

Similar with deriving the forward induction function for α_{nt} in Appendix A, for the Markov states $\{S\}$, transition into $\{S\}$ at day $t+1$ can only take place from $\{S\}$ and $\{I, 1\}$ at time t.

$$\begin{aligned} \hat{\alpha}_{nt+1}(S) &= p(z_{nt+1} = S, \mathbf{y}_{n1}^t | \Omega) \\ &= \alpha_{nt}(I, 1) + P(z_{nt+1} = S | z_{nt} = S)\alpha_{nt}(S) \end{aligned}$$

For the semi-Markov states $\{I, d\}$, transition into $\{I, d\}$ takes place from $\{I, d+1\}$

and $\{S\}$.

$$\begin{aligned}\hat{\alpha}_{nt+1}(I,d) &= p(z_{nt+1}=I, d_{nt+1}=d, \mathbf{y}_{n1}^{t}|\Omega) \\ &= \alpha_{nt}(I,d+1) + \alpha_{nt}(S)p(z_{nt+1}=I|z_{nt}=S)p(d_{nt+1}=d)\end{aligned}$$

$\alpha_{nt}(I,1)$, $\alpha_{nt}(I,d+1)$ and $\alpha_{nt}(S)$ is obtained from training the model with current data. The one step forward prediction of transition probability $P(z_{nt+1}=I|z_{nt}=S)=\Phi(\hat{\gamma}_{nt+1})$ is obtained based on the properties of AR model.

$$\hat{\gamma}_{nt+1} \sim N(\mu_{\hat{\gamma}_{nt+1}}, \sigma_{\hat{\gamma}_{nt+1}})$$

where $\mu_{\hat{\gamma}_{nt+1}} = \mathbf{F}^T\mathbf{G}\mathbf{m}_t^{(2)} + \mu_{a_n} + \omega m_t^{(1)}$ and $\sigma_{\hat{\gamma}_{nt+1}} = \mathbf{F}^T(\mathbf{G}\mathbf{C}_t^{(2)}\mathbf{G}^T + \mathbf{W}_t)\mathbf{F} + \omega^2 C_t^{(1)} + \beta^{-1} + 1 + \sigma_{a_n}$, where $\mu_t^{(1)}$ and $C_t^{(1)}$ are the mean and variance obtained in the forward filtering step when sample $\gamma_t^{(1)}$. μ_{a_n} and σ_{a_n} are the posterior mean and variance of a_n. Notice for prediction, we do not take into account Google Flu Trend data.

Appendix B

Appendix for a big data investigation of electoral representative

B.1 Hierarchical representation of the model

$$\boldsymbol{x}_{jn} = \mathbf{D}^r \boldsymbol{\Lambda}^r \boldsymbol{s}^r_{jn} + \boldsymbol{\epsilon}^r_{jn} \qquad \boldsymbol{\epsilon}^r_{jn} \sim \mathcal{N}(0, \sigma_j^{-1}\mathbf{I}) \qquad \boldsymbol{s}^r_{jn} \sim \mathcal{N}(\boldsymbol{\theta}^*_{c_{jn}}, \mathbf{I})$$

$$\tilde{\boldsymbol{b}}_j = \mathbf{D}^b \boldsymbol{\Lambda}^b \boldsymbol{s}^b_{jn} + \boldsymbol{\epsilon}^b_{jn} \qquad b_{jnp} = 0, \text{ if } \tilde{b}_{jnp} > 0 \qquad b_{jnp} = 1, \text{ if } \tilde{b}_{jnp} < 0$$

$$\boldsymbol{s}^b_{jn} \sim \mathcal{N}(\boldsymbol{\psi}^*_{c_{jn}}, \mathbf{I}) \qquad c_{jn} \sim Cat(\boldsymbol{\pi}_j) \qquad \boldsymbol{\pi}_j \sim DP(\kappa\boldsymbol{\nu})$$

$$\nu_t = \nu'_t \prod_{i=1}^{t}(1-\nu'_t) \qquad \nu'_t \sim beta(1, \kappa_0) \qquad \boldsymbol{d}^r_p \sim \mathcal{N}(0, \mathbf{I})$$

$$\boldsymbol{d}^b_p \sim \mathcal{N}(0, \mathbf{I}) \qquad \boldsymbol{\epsilon}^b_{jn} \sim \mathcal{N}(0, \mathbf{I}) \qquad \boldsymbol{\psi}^*_t \sim \mathcal{N}(\mu^b_0, \boldsymbol{\Sigma}^b_0)$$

$$\boldsymbol{\theta}^*_t \sim \mathcal{N}(\boldsymbol{\mu}^r_0, \boldsymbol{\Sigma}^r_0) \qquad \lambda^r_k \sim \mathcal{N}(0, \tau_k^{r\,-1}), \qquad \tau^r_k = \prod_{h=1}^{k} \varphi^r_h$$

$$\varphi^r_h \sim Gamma(a_1, 1) \qquad \lambda^b_k \sim \mathcal{N}(0, \tau_k^{b\,-1}) \qquad \tau^b_k = \prod_{h=1}^{k} \varphi^b_h$$

$$\varphi^b_h \sim Gamma(a_1, 1) \qquad z_{il} \sim Cat(\boldsymbol{\beta}_l) \qquad \boldsymbol{\beta}_l \sim Dir(\eta_1)$$

$$v_{il} \sim Cat(\boldsymbol{\phi}_{z_{il}}) \qquad \boldsymbol{\phi}_k \sim Dir(\eta_2) \qquad \bar{z}_{lk} = \frac{1}{W_l}\sum_{i=1}^{W_l} I(z_{il} = k)$$

$$\bar{c}_{jt} = \frac{1}{N_j}\sum_{n=1}^{N_j} I(c_{jnt} = t) \qquad \tilde{r}_{jl} = \boldsymbol{d}^{\ell\,T}_j \boldsymbol{\Lambda}^l \boldsymbol{s}^\ell_l + \delta_l + \delta_{jl} + \epsilon^\ell_{jl} \qquad \epsilon^l_{jl} \sim \mathcal{N}(0, 1)$$

$$r_{jl} = 1, \text{ if } \tilde{r}_{jl} > 0 \qquad r_{jl} = 0, \text{ if } \tilde{r}_{jl} \leqslant 0 \qquad \boldsymbol{d}^\ell_j = \mathbf{U}^d \bar{\boldsymbol{c}}_j + \boldsymbol{d}^\ell_0 + \boldsymbol{\xi}_j$$

$$\boldsymbol{s}^\ell_l = \mathbf{U}^s \bar{\mathbf{z}}_l + s^\ell_0 \qquad \boldsymbol{u}^d_k \sim \mathcal{N}(0, \alpha_d^{-1}\mathbf{I}) \qquad \mathbf{u}^s_k \sim \mathcal{N}(0, \alpha_s^{-1}\mathbf{I})$$

$$\boldsymbol{\Lambda}^l = diag(\lambda^l_1, ..., \lambda^l_K) \qquad \lambda^l_k \sim \mathcal{N}(0, \tau_k^{l\,-1}) \qquad \tau^l_k = \prod_{h=1}^{k} \varphi^l_h$$

$$\varphi^l_h \sim Gamma(a_1, 1) \qquad \xi_{jk} \sim \mathcal{N}(0, \alpha_{jk}^{-1}) \qquad \alpha_{jk} \sim InvG(1, \gamma_{jk}/2)$$

$$\delta_l = \bar{\mathbf{z}}_l \mathbf{w} + w_0 \qquad \mathbf{w} \sim \mathcal{N}(0, \alpha_w^{-1}\mathbf{I}) \quad \delta_{il} \sim \mathcal{N}(0, \alpha_{jl}'^{-1})$$

$$\alpha_{jl}'^{-1} \sim InvG(1, \gamma_{jl}'/2)$$

$I(.)$ denotes the indicator function. $I(.) = 1$ if the inside condition holds, 0 otherwise. Diffuse gamma priors are placed on σ_j, γ_{jl}', γ_{jl}.

B.2 Inference

The posterior inference of the model is performed via Stochastic Variational Bayesian. Let $\mathcal{X} = \{\mathbf{X}_j, \mathbf{B}_j, \mathbf{R}\}$ denotes the training data, $\mathbf{\Gamma}$ denote all the hyper parameters and $\mathbf{\Theta}$ denotes all the latent variables. In our case, we use the following fully factorized variational distributions to approximate the posterior distribution.

$$q(\Theta) = \prod_{p=1}^{P^r} q(\boldsymbol{d}_p^r) \prod_{p=1}^{P^b} q(\boldsymbol{d}_p^b) \prod_{j=1}^{J} \prod_{n=1}^{N_j} q(\boldsymbol{s}_n^r) q(\boldsymbol{s}_n^b) q(c_{jn}) \prod_{k=1}^{K^r} q(\lambda_k^r) q(\varphi_k^r) \prod_{k=1}^{K^b} q(\lambda_k^b) q(\varphi_k^b)$$

$$\prod_{j=1}^{J} q(\boldsymbol{\pi}_j) \prod_{t=1}^{T} q(\boldsymbol{\theta}_t^*) q(\boldsymbol{\psi}_t^*) \prod_{k=1}^{K^\ell} q(\varphi_k^l) q(\lambda_k^l) q(\boldsymbol{u}_k^s) q(\boldsymbol{u}_k^d) \prod_{j=1}^{J} \prod_{k=1}^{K^\ell} q(\xi_{jk}) q(\alpha_{jk}) q(\gamma_{jk})$$

$$q(\boldsymbol{\nu}) q(\boldsymbol{w}) \prod_{l=1}^{L} \prod_{i=1}^{W^l} q(z_{il}) \prod_{g=1}^{K} q(\boldsymbol{\phi}_g) \prod_{l}^{L} q(\boldsymbol{\beta}_l) \prod_{j}^{J} \prod_{l=1}^{L} q(\delta_{jl}) q(\alpha_{jl}') q(\gamma_{jl}') \prod_{j=1}^{J} q(\sigma_j)$$

The evidence lower bound is as following.

$$\log(p(\mathcal{X}|\mathbf{\Gamma})) \geqslant \mathbb{E}[\log(p(\mathcal{X}, \mathbf{\Theta}, \mathbf{\Gamma}))] - \mathbb{E}[\log(q(\mathbf{\Theta}))]$$

we can expand the evidence lower bound with the given likelihoods and factorized variational distributions, the detail is omit for brevity.

B.2.1 Local parameters

We partition the Catalist data into N^* mini-batches and the roll call data and related legislative votes are considered as a whole. Each mini-batch contains voters from all the 435 districts, let us denote number of voters in district j within one mini-batch as

N'_j. The variational parameters specific to each batch are local parameters. In this model, the local parameters $\mathbf{\Theta}^l = \{\boldsymbol{\mu}_{s^r_{jn}}, \mathbf{\Sigma}_{s^r_{jn}}, \boldsymbol{\mu}_{s^r_{jn}}, \mathbf{\Sigma}_{s^r_{jn}}, \tilde{\boldsymbol{\pi}}_{jn}\}$ are the ones related with $\boldsymbol{s}^r_{jn}, \boldsymbol{s}^b_{jn}, c_{jn}$.

Update equations for $\boldsymbol{s}^r_{jn}$

$$q(\boldsymbol{s}^r_{jn}) \sim \mathcal{N}(\boldsymbol{\mu}_{s^r_{jn}}, \mathbf{\Sigma}_{s^r_{jn}})$$

where the mean and covariance matrix are as following

$$\mathbf{\Sigma}_{\boldsymbol{s}^r_{jn}} = (\mathbb{E}[\sigma_j]\mathbb{E}[\mathbf{\Lambda}^r \mathbf{D}^{r\,T} \mathbf{D}^r \mathbf{\Lambda}^r] + \mathbf{I})^{-1}$$

$$\boldsymbol{\mu}^r_{\boldsymbol{s}_{jn}} = \mathbf{\Sigma}_{\boldsymbol{s}^r_{jn}}(\mathbb{E}[\sigma_j]\mathbb{E}[\mathbf{\Lambda}^r \mathbf{D}^r]\boldsymbol{x}_{jn} + \textstyle\sum_{t=1}^{T} \tilde{\pi}_{jnt}\mathbb{E}[\boldsymbol{\theta}^*_t])$$

The related expectation is

$$\mathbb{E}[\mathbf{\Lambda}^r \mathbf{D}^r \mathbf{D}^r \mathbf{\Lambda}^r] = \textstyle\sum_{p=1}^{P^r}(\mathbb{E}[\boldsymbol{d}^r_p]\mathbb{E}[\boldsymbol{d}^r_p] + \mathbf{\Sigma}_{\boldsymbol{d}^r_p}) \odot (\mathbb{E}[\boldsymbol{\lambda}^r])\mathbb{E}[\boldsymbol{\lambda}^r]^T + Diag(\Sigma_{\lambda^r_1}, ..., \Sigma_{\lambda^r_K}))$$

Update equations for $\boldsymbol{s}^b_{jn}$

The update equations for $\boldsymbol{s}^b_{jn}$ is similar with $\boldsymbol{s}^r_{jn}$. We can obtain the updates by replacing the superscript r, $\mathbb{E}[\sigma_j]$ and $\boldsymbol{x}_{jn}$ with b,1 and $\mathbb{E}[\tilde{\boldsymbol{b}}_{jn}]$, respectively. The related expectation is as following,

$$\mathbb{E}[\tilde{b}_{jnp}] = \begin{cases} \mathbb{E}[\boldsymbol{d}^{b\,T}_p \mathbf{\Lambda}^b \boldsymbol{s}^b_{jn}] + \frac{\phi(\boldsymbol{d}^{b\,T}_p \mathbf{\Lambda}^b \boldsymbol{s}^b_{jn})}{1-\Phi(\boldsymbol{d}^{b\,T}_p \mathbf{\Lambda}^b \boldsymbol{s}^b_{jn})} \text{ if } b_{jnp} = 1 \\ \mathbb{E}[\boldsymbol{d}^{b\,T}_p \mathbf{\Lambda}^b \boldsymbol{s}^b_{jn}] - \frac{\phi(\boldsymbol{d}^{b\,T}_p \mathbf{\Lambda}^b \boldsymbol{s}^b_{jn})}{\Phi(\boldsymbol{d}^{b\,T}_p \mathbf{\Lambda}^b \boldsymbol{s}^b_{jn})}] \text{ if } b_{jnp} = 0 \end{cases}$$

Update equations for c_{jn}

$$q(c_{jn}) \sim Cat(\tilde{\boldsymbol{\pi}}_{jn})$$

The t dimension parameter $\tilde{\boldsymbol{\pi}}_{jn}$ is

$$\tilde{\pi}_{jnt} \propto \exp(\mathbb{E}[\log(\mathcal{N}(\boldsymbol{s}^r_{jn}|\boldsymbol{\theta}^*_t, \mathbf{I}))] + \mathbb{E}[\log(\mathcal{N}(\boldsymbol{s}^b_{jn}|\boldsymbol{\psi}^*_t, \mathbf{I}))] + \mathbb{E}[\log(\pi_{jt})] +$$

$$\textstyle\sum_{l=1}^{L} \frac{\mathbb{E}[\bar{r}^{(1)}_{jlnt} u^{(1)}_{tl}]}{N'_j} - \frac{\mathbb{E}[u^{(1)\,2}_{tl}]}{2N'^2_j})$$

The corresponding expectation and equations are as following

$$\bar{r}^{(1)}_{jlnt} = \tilde{r}_{jl} - \delta_l - \delta_{jl} - (\boldsymbol{d}^{\ell}_0 + \boldsymbol{\xi}_j)^T \boldsymbol{\Lambda}^{\ell} \boldsymbol{s}^{\ell}_l - \sum_{t' \neq t} \{[\boldsymbol{u}^{dT}_t \boldsymbol{\Lambda}^{\ell} \boldsymbol{s}^{\ell}_l][\frac{\sum_{n' \neq n} I(c_{jn'}=t)}{N'_j}]\}$$

$$u^{(1)}_{tl} = [\boldsymbol{u}^{dT}_t \boldsymbol{\Lambda}^{\ell} \boldsymbol{s}^{\ell}_l].$$

$$\mathbb{E}[\log(\mathcal{N}(\boldsymbol{s}^r_{jn}|\boldsymbol{\theta}^*_t, \mathbf{I}))] = -\frac{tr(\boldsymbol{\Sigma}_{\boldsymbol{s}^r_{jn}}) + \mathbb{E}[\boldsymbol{s}^{r\,T}_{jn}]\mathbb{E}[\boldsymbol{s}^r_{jn}] + tr(\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*_t}) + \boldsymbol{\theta}^{*\,T}_t \boldsymbol{\theta}^*_t - 2\mathbb{E}[\boldsymbol{s}^r_{jn}]^T \mathbb{E}[\boldsymbol{\theta}^*_t]}{2}$$

$$\mathbb{E}[\log(\mathcal{N}(\boldsymbol{s}^b_{jn}|\boldsymbol{\psi}^*_t, \mathbf{I}))] = -\frac{tr(\boldsymbol{\Sigma}_{\boldsymbol{s}^b_{jn}}) + \mathbb{E}[\boldsymbol{s}^{b\,T}_{jn}]\mathbb{E}[\boldsymbol{s}^b_{jn}] + tr(\boldsymbol{\Sigma}_{\boldsymbol{\psi}^*_t}) + \boldsymbol{\psi}^{*\,T}_t \boldsymbol{\psi}^*_t - 2\mathbb{E}[\boldsymbol{s}^b_{jn}]^T \mathbb{E}[\boldsymbol{\psi}^*_t]}{2}$$

$$\mathbb{E}[\log(\pi_{jt})] = \Psi(\theta_{\pi_j t}) - \Psi(\textstyle\sum_t \theta_{\pi_j t})$$

$\Psi(.)$ denotes the digamma function.

B.2.2 Global parameters

The remaining variational parameters are considered as global parameters $\boldsymbol{\Theta}^g$. We list the main update equations to calculate these intermediate global variational parameters $\tilde{\boldsymbol{\Theta}}^g$ as following.

Update equations for $\boldsymbol{d}^r_p$

$$q(\boldsymbol{d}^r_p) \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{d}^r_p}, \boldsymbol{\Sigma}_{\boldsymbol{d}^r_p})$$

where the mean and covariance matrix are as following

$$\boldsymbol{\mu}_{\boldsymbol{d}^r_p} = \mathbb{E}[\sigma_j]\boldsymbol{\Sigma}_{\boldsymbol{d}^r_p}(\textstyle\sum_{j=1}^{J} \frac{N_j}{N'_j} \sum_{n=1}^{N'_j} \mathbb{E}[\boldsymbol{\Lambda}^r]\mathbb{E}[\boldsymbol{s}^r_{jn}]x_{jnp})$$

$$\boldsymbol{\Sigma}_{\boldsymbol{d}^r_p} = (\textstyle\sum_{j=1}^{J} \frac{N_j}{N'_j} \sum_{n}^{N'_j} \mathbb{E}[\boldsymbol{\Lambda}^r \boldsymbol{s}^r_{jn} \boldsymbol{s}^{r\,T}_{jn} \boldsymbol{\Lambda}^{r\,T}]\mathbb{E}[\sigma_j])^{-1}$$

The related expectation is

$$\mathbb{E}[\boldsymbol{\Lambda}^r \boldsymbol{s}^r_{jn} \boldsymbol{s}^{r\,T}_{jn} \boldsymbol{\Lambda}^{r\,T}] = (\boldsymbol{\Sigma}_{\boldsymbol{s}^r_{jn}} + \mathbb{E}[\boldsymbol{s}^r_{jn}]\mathbb{E}[\boldsymbol{s}^r_{jn}]) \odot (\mathbb{E}[\boldsymbol{\lambda}^r]\mathbb{E}[\boldsymbol{\lambda}^r]^T + Diag(\Sigma_{\lambda^r_1}, ..., \Sigma_{\lambda^r_K})).$$

where $\odot$ is the Hadamard product and $\boldsymbol{\lambda}^r = [\lambda^r_1, ..., \lambda^r_K]^T$

Update equations for λ^r_k

$$q(\lambda^r_k) \sim (\mu_{\lambda^r_k}, \Sigma_{\lambda^r_k})$$

The mean and variance are as following

$$\Sigma_{\lambda_k^r} = (\sum_{j=1}^{J} \frac{N_j}{N_j'} \sum_{n=1}^{N_j} \mathbb{E}[\sigma_j]\mathbb{E}[{\boldsymbol{d}_k^r}^T \boldsymbol{d}_k^r {s_{jnk}^r}^2] + \mathbb{E}[\tau_k^r])^{-1}$$

$$\mu_{\lambda_k^r} = \Sigma_{\lambda_k^r}(\sum_{j=1}^{J} \frac{N_j}{N_j'} \sum_{n=1}^{N_j} \mathbb{E}[s_{jnk}]\mathbb{E}[\boldsymbol{d}_k]^T \hat{\boldsymbol{x}}_{jn})$$

The related expectation and equations are

$$\mathbb{E}[{\boldsymbol{d}_k^r}^T \boldsymbol{d}_k^r {s_{jnk}^r}^2] = (\mathbb{E}[\boldsymbol{d}_k^r]^T\mathbb{E}[\boldsymbol{d}_k^r] + tr(\boldsymbol{\Sigma}_{d_k^r}))(\mathbb{E}[s_{jnk}^r]^2 + \Sigma_{s_{jnk}^r})$$

$$\hat{\boldsymbol{x}}_{jn} = \boldsymbol{x}_{jn} - \sum_{\bar{k}=1,\bar{k}\neq k}^{K^r} \mathbb{E}[\boldsymbol{d}_{\bar{k}}^r]\mathbb{E}[s_{jn\bar{k}}]\mathbb{E}[\lambda_{\bar{k}}^r]$$

Update equations of φ_h^r

$$q(\varphi_h^r) \sim Gamma(a_{\varphi_h^r}, b_{\varphi_h^r})$$

where the shape and scale parameters are as following

$$a_{\varphi_h^r} = a_1 + \frac{K^r - h + 1}{2}$$

$$b_{\varphi_h^r} = 1 + \sum_{k=h}^{K^r} \frac{\mathbb{E}[{\lambda_k^r}^2] \prod_{\bar{h}=1,\bar{h}\neq h}^{k} \mathbb{E}[\varphi_{\bar{h}}^r]}{2} I(k \geqslant h)$$

Update equations for $\boldsymbol{\theta}_t^*$

$$q(\boldsymbol{\theta}_t^*) \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}_t^*}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}_t^*})$$

where the mean and covariance matrix are

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}_t^*} = (\boldsymbol{\Sigma}_0^r + \sum_{j=1}^{J} \frac{N_j}{N_j'} \sum_{n=1}^{N_j'} \tilde{\pi}_{jnt}\mathbf{I})^{-1}$$

$$\mu_{\boldsymbol{\theta}_t^*} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}_t^*}(\boldsymbol{\mu}_0^r + \sum_{j=1}^{J} \frac{N_j}{N_j'} \sum_{n=1}^{N_j'} \tilde{\pi}_{jnt}\mathbb{E}[\boldsymbol{s}_{jn}^r])$$

Update equations for $\boldsymbol{d}_p^b, \lambda_k^b, \lambda_h^b, \boldsymbol{\psi}_t^*$

The update equations for $\boldsymbol{d}_p^b, \lambda_k^b, \lambda_h^b, \boldsymbol{\psi}_t^*$ are similar to $\boldsymbol{d}_p^r, \lambda_k^r, \lambda_h^r, \boldsymbol{\theta}_t^*$, respectively. We can obtain these update equations by replacing the superscript r, $\mathbb{E}[\sigma_j]$ and $\boldsymbol{x}_{jn}$ with b,1 and $\mathbb{E}[\tilde{\boldsymbol{b}}_{jn}]$, respectively.

Update equations for σ_j

$$q(\sigma_j) \sim Gamma(a_{\sigma_j}, b_{\sigma_j})$$

$$a_{\sigma_j} = a_0^r + P^r N_j/2$$

$$b_{\sigma_j} = b_0^r + \frac{N_j}{2N_j'} \sum_{n=1}^{N_j'} (\boldsymbol{x}_{jn}^T \boldsymbol{x}_{jn} + \mathbb{E}[\boldsymbol{s}_{jn}^{r\,T} \boldsymbol{\Lambda}^r \mathbf{D}^{r\,T} \mathbf{D}^r \boldsymbol{\Lambda}^r \boldsymbol{s}_{jn}^r] - 2\boldsymbol{x}_{jn}^T \mathbb{E}[\mathbf{D}^r] \mathbb{E}[\boldsymbol{\Lambda}^r] \mathbb{E}[\boldsymbol{s}_{jn}^r])$$

The related expectation is

$$\mathbb{E}[\boldsymbol{s}_{jn}^{r\,T} \boldsymbol{\Lambda}^r \mathbf{D}^{r\,T} \mathbf{D}^r \boldsymbol{\Lambda}^r \boldsymbol{s}_{jn}^r] =$$

$$tr((\sum_{p=1}^{P^r} (\mathbb{E}[\boldsymbol{d}_p^r] \mathbb{E}[\boldsymbol{d}_p^{r\,T}] + \boldsymbol{\Sigma}_{d_p^r}) \odot \mathbb{E}[\boldsymbol{\lambda}^r \boldsymbol{\lambda}^{r})^T]) (\mathbb{E}[\boldsymbol{s}_{jn}^r] \mathbb{E}[\boldsymbol{s}_{jn}^{r\,T}] + \boldsymbol{\Sigma}_{\boldsymbol{s}_{jn}^r}))$$

Update equations for $\boldsymbol{\pi}_j$

$$q(\boldsymbol{\pi}_j) \sim Dir(\theta_{\boldsymbol{\pi}_j})$$

$$\theta_{\boldsymbol{\pi}_j} = \kappa \boldsymbol{\nu} + \frac{N_j}{N_j'} \sum_{n=1}^{N_j'} \tilde{\boldsymbol{\pi}}_{jn}$$

Update equations for $\boldsymbol{\nu}$

We do a point estimate on $\boldsymbol{\nu}$ and q($\boldsymbol{\nu}$) is a degenerated distribution. The objective function of optimizing $\boldsymbol{\nu}$ is as following.

$$L(\boldsymbol{\nu}) = \log \mathrm{GEM}(\boldsymbol{\nu}; \kappa_0) + \sum_{j=1}^{J} \mathbb{E}[\log Dir(\boldsymbol{\pi}_j | \boldsymbol{\nu})]$$

where GEM($\boldsymbol{\nu}; \kappa_0$) refers to the stick breaking prior. The derivation of the gradient can be found in [**?**].

Update equations for $\boldsymbol{\xi}_k$

$$q(\boldsymbol{\xi}_k) \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\xi}_k}, \boldsymbol{\Sigma}_{\boldsymbol{\xi}_k})$$

where the mean and covariance is

$$\boldsymbol{\Sigma}_{\boldsymbol{\xi}_k} = (\sum_{l=1}^{L} (\mathbb{E}[\lambda_k^{l\,2}] \mathbb{E}[\bar{\boldsymbol{z}}_l^T \boldsymbol{u}_k^s \boldsymbol{u}_k^{s\,T} \bar{\boldsymbol{z}}_l]) + \mathbb{E}[\boldsymbol{\alpha}_k])^{-1}$$

$$\boldsymbol{\mu}_{\boldsymbol{\xi}_k} = \boldsymbol{\Sigma}_{\boldsymbol{\xi}_k} (\sum_{l=1}^{L} \mathbb{E}[\lambda_k^l \bar{\boldsymbol{z}}_l^T \boldsymbol{u}_k^s \hat{\boldsymbol{r}}_l])$$

where related equations are

$$\hat{\boldsymbol{r}}_l = \tilde{\boldsymbol{r}}'_l - \sum_{k=1}^{K^\ell} \lambda_k^l (\mathbf{C}\boldsymbol{u}_k^d + \boldsymbol{\xi}_k)\boldsymbol{u}_k^{s\,T}\bar{\boldsymbol{z}}_l + \lambda_k^l \boldsymbol{\xi}_k \boldsymbol{u}_k^{s\,T}\bar{\boldsymbol{z}}_l$$

$$\mathbf{C} = [\bar{\boldsymbol{c}}_1, ..., \bar{\boldsymbol{c}}_J]^T \text{ and } \mathbb{E}[\bar{\boldsymbol{c}}_j] = \frac{1}{N'_j}\sum_{n=1}^{N'_j} \tilde{\boldsymbol{\pi}}_{jn}$$

$$\tilde{\boldsymbol{r}}'_l = [\tilde{r}'_{j1}, ..., \tilde{r}'_{jL}]^T \text{ and } \tilde{r}'_{jl} = \tilde{r}_{jl} - \delta_l - \delta_{jl}$$

Update equations for $\boldsymbol{u}_k^d$

$$q(\boldsymbol{u}_k^d) \sim \mathcal{N}(\mu_{\boldsymbol{u}_k^d}, \boldsymbol{\Sigma}_{\boldsymbol{u}_k^d})$$

where the mean and covariance are

$$\boldsymbol{\Sigma}_{\boldsymbol{u}_k^d} = (\sum_{l=1}^{L} \mathbb{E}[\mathbf{C}^T\mathbf{C}\lambda_k^{l\,2}\bar{\boldsymbol{z}}_l^T\boldsymbol{u}_k^s\boldsymbol{u}_k^{s\,T}\bar{\boldsymbol{z}}_l] + \mathbb{E}[\alpha_d]\mathbf{I})^{-1}$$

$$\mu_{\boldsymbol{u}_k^d} = \boldsymbol{\Sigma}_{\boldsymbol{u}_k^d}(\sum_{l=1}^{L} \mathbb{E}[\lambda_k^l \bar{\boldsymbol{z}}_l^T \boldsymbol{u}_k^s \mathbf{C}^T \hat{\boldsymbol{r}}_l])$$

where the related equations are

$$\hat{\boldsymbol{r}}_l = \tilde{\boldsymbol{r}}'_l - \sum_{k=1}^{K} \lambda_k^l (\mathbf{C}\boldsymbol{u}_k^d + \boldsymbol{\xi}_k)\boldsymbol{u}_k^{s\,T}\bar{\boldsymbol{z}}_l + \lambda_k^l (\mathbf{C}\boldsymbol{u}_k^d)\boldsymbol{u}_k^{s\,T}\bar{\boldsymbol{z}}_l$$

$$\mathbb{E}[\bar{\boldsymbol{c}}_j\bar{\boldsymbol{c}}_j^T] = \frac{1}{N'^2_j}(\sum_{n=1}^{N'_j}\sum_{m\neq n} \tilde{\boldsymbol{\pi}}_{jn}\tilde{\boldsymbol{\pi}}_{jm}^T + \sum_{n=1}^{N'_j} \mathrm{diag}(\tilde{\boldsymbol{\pi}}_{jn}))$$

$$\mathbb{E}[\bar{\boldsymbol{z}}_l\bar{\boldsymbol{z}}_l^T] = \frac{1}{W_l^2}(\sum_{i=1}^{W_l}\sum_{m\neq i} \tilde{\boldsymbol{\beta}}_{il}\tilde{\boldsymbol{\beta}}_{ml}^T + \sum_{i=1}^{W_l} \mathrm{diag}(\tilde{\boldsymbol{\beta}}_{il}))$$

Update equations for $\boldsymbol{u}_k^s$

$$q(\boldsymbol{u}_k^s) \sim \mathcal{N}(\mu_{\boldsymbol{u}_k^s}, \boldsymbol{\Sigma}_{\boldsymbol{u}_k^s})$$

where the mean and covariance are

$$\boldsymbol{\Sigma}_{\boldsymbol{u}_k^s} = (\sum_{l=1}^{L} \mathbb{E}[\lambda_k^{l\,2}\bar{\boldsymbol{z}}_l(\mathbf{C}\boldsymbol{u}_k^d + \boldsymbol{\xi}_k)^T(\mathbf{C}\boldsymbol{u}_k^d + \boldsymbol{\xi}_k)\bar{\boldsymbol{z}}_l^T] + \mathbb{E}[\alpha_s]\mathbf{I})^{-1}$$

$$\mu_{\boldsymbol{u}_k^s} = \boldsymbol{\Sigma}_{\boldsymbol{u}_k^s}(\sum_{l=1}^{L} \mathbb{E}[\lambda_k^l \bar{\boldsymbol{z}}_l (\mathbf{C}\boldsymbol{u}_k^d + \boldsymbol{\xi}_k)^T \hat{\boldsymbol{r}}_l])$$

$\hat{\boldsymbol{r}}_l = \tilde{\boldsymbol{r}}'_l - \sum_{k'\neq k} \lambda_{k'}^l (\mathbf{C}\boldsymbol{u}_{k'}^d + \boldsymbol{\xi}_{k'})\boldsymbol{u}_{k'}^{s\,T}\bar{\boldsymbol{z}}_l$

Update equations for λ_k^l

$$q(\lambda_k^l) \sim \mathcal{N}(\mu_{\lambda_k^l}, \Sigma_{\lambda_k^l})$$

where the mean and variance are

$$\Sigma_{\lambda_k^l} = (\sum_{l=1}^{L} \mathbb{E}[\bar{\boldsymbol{z}}_l^T \boldsymbol{u}_k^s (\mathbf{C}\boldsymbol{u}_k^d + \boldsymbol{\xi}_k)^T (\mathbf{C}\boldsymbol{u}_k^d + \boldsymbol{\xi}_k) \boldsymbol{u}_k^{sT} \bar{\boldsymbol{z}}_l] + \mathbb{E}[\tau_k^l])$$

$$\mu_{\lambda_k^l} = \Sigma_{\lambda_k^l} (\sum_{l=1}^{L} \mathbb{E}[\hat{\boldsymbol{r}}_l^T (\mathbf{C}\boldsymbol{u}_k^d + \boldsymbol{\xi}_k) \boldsymbol{u}_k^{sT} \bar{\boldsymbol{z}}_l])$$

where $\hat{\boldsymbol{r}}_l = \tilde{\boldsymbol{r}}'_l - \sum_{k' \neq k} \lambda_{k'}^l (\mathbf{C}\boldsymbol{u}_{k'}^d + \boldsymbol{\xi}_{k'}) \boldsymbol{u}_{k'}^{sT} \bar{\boldsymbol{z}}_l$

Update equations for z_{il}

$$q(z_{il}) \sim Cat(\tilde{\boldsymbol{\beta}}_{il})$$

The parameter $\tilde{\boldsymbol{\beta}}_{il}$ is as following,

$$\tilde{\beta}_{ilk} \propto \exp\{\mathbb{E}[\log Cat(v_{il}|\boldsymbol{\phi}_k)] + \mathbb{E}[\log \beta_{lk}] + \sum_{j=1}^{J} (\frac{\mathbb{E}[\bar{r}_{jkl}^{(2)} u_{jk}^{(2)}]}{W_l} - \frac{\mathbb{E}[u_{jk}^{(2)\,2}]}{2W_l^2})$$

The related equations are as following,

$$\bar{r}_{jkil}^{(2)} = \tilde{r}'_{jl} - \sum_{k' \neq k} \{[\boldsymbol{d}_j^{\ell T} \boldsymbol{\Lambda}^\ell \boldsymbol{u}_{k'}^s + w_{k'}][\frac{\sum_{i' \neq i} I(z_{li'} = k')}{W_l}]\}$$

$$u_{jk}^{(2)} = [\boldsymbol{d}_j^{\ell T} \boldsymbol{\Lambda}^\ell \boldsymbol{u}_k^s + w_k].$$

Update equations for $\boldsymbol{\beta}_l$**,** $\boldsymbol{\phi}_k$

The update equation for $\boldsymbol{\beta}_l$ and $\boldsymbol{\phi}_k$ are same as the related parameter updates of latent Dirichlet allocation (LDA) and are omit for brevity.

After obtain all the intermediate global parameters $\tilde{\boldsymbol{\Theta}}^g$, we update the global parameters $\boldsymbol{\Theta}^g$ as following.

$$\boldsymbol{\Theta}^g \leftarrow (1 - \omega_h)\boldsymbol{\Theta}^g + \omega_h \tilde{\boldsymbol{\Theta}}^g$$

B.2.3 Algorithm

The stochastic variational Bayesian method for the proposed model is summarized in Algorithm 1.

Table B.1: Summary of Catalist attributes

Categories	Number of binary attributes	Number of real attributes	Description
Gender	1	0	male or female
Age	0	3	age, mean and standard deviation of age among house members
Finance	3	3	income; household value; information related with investment, bonds purchasing, credit card
Race	7	5	race includes Black, Caucasian, Hispanic, Asian etc.
Turnout	2	1	turnout in 2006 and 2008 general election; turnout rate of the household members
Party affiliation	9	2	Democrat, Republican and other independent party
Behavior	1	1	play golf or not; Internet usage
Children	4	1	have child between certain age or not
Religion	10	0	include Catholic, Protestant, Hindu, Muslim, Buddist etc.
Home	0	1	own or rent
Donation	3	0	donate to political, religious and environmental issues

Algorithm 1 Stochastic Variational Bayesian Analysis for Proposed Model

Partition $\mathbf{X}$ and $\mathbf{B}$ into N^* mini-batches.
Define local parameters $\mathbf{\Theta}^l$ and global parameters $\mathbf{\Theta}^g$.
Initialize $\mathbf{\Theta}^g$ by running model on a mini-batch
for $h = 1$ **to** N^* **do**
  $\omega_h = (a_3 + h)^{-b_3}$
  while stop criterion is not met **do**
    for $j = 1$ **to** J **do**
      for $n = 1$ **to** N_j^* **do**
        Estimate $\mathbf{\Theta}^l$
      end for
    end for
  end while
  Compute $\tilde{\mathbf{\Theta}}^g$
  Update $\mathbf{\Theta}^g \leftarrow (1 - \omega_h)\mathbf{\Theta}^g + \omega_h \tilde{\mathbf{\Theta}}^g$
end for

B.3 Catalist attributes

We summary the Catalist attributes used in the model in Table B.1.

Bibliography

[AC93a] J. H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *J. Am. Stat. Assoc.*, 1993.

[AC93b] J.H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.*, 88(422):669–679, June 1993.

[AEB06] M. Aharon, M. Elad, and A. M. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Processing*, 54:4311–4322, 2006.

[BCMS09] A. Buades, B. Coll, J.-M. Morel, and C. Sbert. Self-similarity driven color demosaicking. *IEEE Trans. Image Processing*, 18(6):1192–1202, 2009.

[BD11] A. Bhattacharya and D. B. Dunson. Sparse Bayesian infinite factor models. *Biometrika*, 2011.

[BDE07] A.M. Bruckstein, D.L. Donoho, and M. Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM review*, 51:34–81, 2007.

[Bea03] M. J. Beal. *Variational algorithms for approximate Bayesian inference*. PhD thesis, University of London, 2003.

[BH10] J. Bafumi and M. C. Herron. Leapfrog representation and extremism: A study of American voters and their members in congress. *Am. Polit. Sci. Rev.*, 2010.

[BKG+11] P.J. Birrella, G. Ketsetzisb, N.J. Gayc, B.S. Cooperd, A.M. Presanisa, R.J. Harrisb, A. Charlettb, X.-S. Zhangb, P.J. Whiteb, R.G. Pebody, and D. De Angelisa. Bayesian modeling to unmask and predict influenza A/H1N1pdm dynamics in London. *Proc. Nat. Acad. Sci.*, 108(45):18238–18243, November 2011.

[BM07] D. M. Blei and J. D. McAuliffe. Supervised topic models. In *NIPS*, 2007.

[BNJ03] D. M. Blei, A. Y. Ng, and M. I .Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 2003.

[BS09] José M Bernardo and Adrian FM Smith. *Bayesian theory*, volume 405. John Wiley & Sons, 2009.

[BS12] M. Bryant and E. B. Sudderth. Truly nonparametric online variational inference for hierarchical Dirichlet processes. In *NIPS*, 2012.

[CCL+08] C. Carvalho, J. Chang, J. Lucas, J.R. Nevins, Q. Wang, and M. West. High-dimensional sparse factor modelling: Applications in gene expression genomics. *J. Amer. Statist. Assoc.*, 103(484):1438–1456, December 2008.

[CCP+10] B. Chen, M. Chen, J. Paisley, A. Zaas, C. Woods, G.S. Ginsburg, A. Hero III, J. Lucas, D. Dunson, and L. Carin. Nonparametric Bayesian factor analysis: application to time-evolving viral gene-expression data. *BMC Bioinformatics*, 11(552), 2010.

[CCV+04] S. Cauchemez, F. Carrat, C. Viboud, A.J. Valleron, and P.Y. Boelle. A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. *Statist. Med.*, 23(22):3469–3487, 2004.

[CD01] M.H. Chen and D.K. Dey. *Generalized Linear Models: A Bayesian Prospective*. Springer, 2001.

[CJR04] J. Clinton, S. Jackman, and D. Rivers. The statistical analysis of roll call data. *Am. Polit. Sci. Rev.*, 2004.

[CK94] C.K. Carter and R. Kohn. On Gibbs sampling for state space models. *Biometrika*, 81(3), August 1994.

[Cli06] J. D. Clinton. Representation in Congress: constituents and roll calls in the 106th House. *J. Polit.*, 2006.

[CO07] D. Clancy and P.D. O'Neill. Exact Bayesian inference and model selection for stochastic models of epidemics among a community of households. *Scandinavian J. Stat.*, 34(2):259–274, June 2007.

[CR08] E.J. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. of Comput. Math.*, pages 717–772, 2008.

[CSL+02] F. Carrat, C. Sahler, M. Leruez, S. Roger, F. Freymuth, C. Le Gales, M. Bungener, B. Housset, M. Nicolas, and S. Rouzioux. Influenza burden of illness: estimates from a national prospective survey of household contacts in france. *Arch. Intern. Med.*, 162(16):1842–1848, September 2002.

[CT06] E. Candès and T. Tao. Near-optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Information Theory*, 52:5406–5425, 2006.

[CWBC02] Brandice Canes-Wrone, David W Brady, and John F Cogan. Out of step, out of office: Electoral accountability and house members' voting. *Am. Polit. Sci. Rev.*, 96(01):127–140, 2002.

[CXG+10] A. Castrodad, Z. Xing, J. Greer, E. Bosch, L. Carin, and G. Sapiro. Discriminative sparse representations in hyperspectral imagery. In *Proc. Int. Conf. Image Proc. (ICIP)*, 2010.

[CZW+11] M. Chen, A. Zaas, C. Woods, G.S. Ginsburg, J. Lucas, D. Dunson, and L. Carin. Predicting viral infection from high-dimensional biomarker trajectories. *J. Amer. Statist. Assoc.*, 106(496):1259–1279, 2011.

[DCS09] J.M. Duarte-Carvajalino and G. Sapiro. Learning to sense sparse signals: simultaneous sensing matrix and sparsifying dictionary optimization. *IEEE Transactions on Image Processing*, pages 1395–1408, 2009.

[DCSVRC08] J.M. Duarte-Carvajalino, G. Sapiro, M. Velez-Reyes, and P.E. Castillo. Multiscale representation and segmentation of hyperspectral imagery using geometric partial differential equations and algebraic multigrid methods. *IEEE Transactions on Geoscience and Remote Sensing*, 46:2418–2434, 2008.

[DE09] B. Demir and S. Erturk. Clustering based extraction of border training patterns for accurate SVM classification of hyperspectral images. *IEEE Geoscience and Remote Sensing Letters*, 6:840–844, 2009.

[DEL07] N. Dimmock, A. Easton, and K. Leppard. *Introduction to Modern Virology*. Blackwell, 2007.

[DLP12] V. Dukic, H.F. Lopes, and N.G. Polson. Tracking epidemics with Google Flu Trends data and a state-space SEIR model. *J. Amer. Statist. Assoc.*, 107(500):1410–1426, December 2012.

[DP08] David B Dunson and Ju-Hyun Park. Kernel stick-breaking processes. *Biometrika*, 95(2):307–323, 2008.

[DPH12] W. Dong, A. Pentland, and K.A. Heller. Graph-coupled HMMs for modeling the spread of infection. In *Uncert. Artificial Intell. (UAI)*, pages 227–236, 2012.

[DWW12] M. Dewar, C. Wiggins, and F. Wood. Inference in hidden Markov models with explicit state duration distributions. *IEEE Signal Process. Lett.*, 19(4):235–238, April 2012.

[EA06] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Trans. Image Processing*, 15:3736–3745, 2006.

[EY10] M. Elad and I. Yavneh. A weighted average of sparse representations is better than the sparsest one alone. *Preprint*, 2010.

[FD11] E. Fox and D.B. Dunson. Bayesian nonparametric covariance regression. *arXiv:1101.2017v2*, February 2011.

[FS94] S. Fruhwirth-Schnatter. Data augmentation and dynamic linear models. *J. Time Series Anal.*, 15(2):183–202, March 1994.

[FSJW08] Emily B Fox, Erik B Sudderth, Michael I Jordan, and Alan S Willsky. An hdp-hmm for systems with state persistence. In *Proceedings of the 25th international conference on Machine learning*, pages 312–319. ACM, 2008.

[GB11] S. Gerrish and D. M. Blei. Predicting legislative roll calls from text. In *ICML*, 2011.

[GB12] S. M. Gerrish and D. M. Blei. How they vote: Issue-adjusted models of legislative behavior. *Polit. Sci.*, 2012.

[GG05] T.L. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. In *Proc. Advances in Neural Information Processing Systems*, pages 475–482, 2005.

[GMP+09] J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S.Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457:1012–1014, February 2009.

[GSTG08] J.V. Gael, Y. Saatci, Y. Teh, and Z. Ghahramani. Beam sampling for the infinite hidden Markov model. In *International Conference on Machine Learning*, pages 1088–1095, New York,NY,USA, July 2008. ACM.

[Has70] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[HBWP13] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

[Het00] H.W. Hethcote. The mathematics of infectious diseases. *SIAM Rev.*, 42(4):599–653, 2000.

[Hjo90] Nils Lid Hjort. Nonparametric bayes estimators based on beta processes in models for life history data. *The Annals of Statistics*, pages 1259–1294, 1990.

[Hoy04] Patrik O Hoyer. Non-negative matrix factorization with sparseness constraints. *The Journal of Machine Learning Research*, 5:1457–1469, 2004.

[IC00] A. Ifarraguerri and C.-I. Chang. Unsupervised hyperspectral image analysis with projection pursuit. *IEEE Trans. Geosc. Remote Sensing*, 38:2529–2538, 2000.

[JEZ10] J.F.Cai, E.J.Candes, and Z.Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 2010.

[JJVF12] G. Jones, W. O. Johnson, W. D. Vink, and N. French. A framework for the joint modeling of longitudinal diagnostic outcome data and latent infection status: Application to investigating the temporal relationship between infection and disease. *Biometrics*, 68(2):371–379, June 2012.

[JT11] M. Jahrer and A. Toscher. Collaborative filtering ensemble. *KDD Cup Workshop*, 2011.

[JW13] M. Johnson and A. Willsky. Bayesian nonparametric hidden semi-Markov models. *J.Mach.Learn.Res.*, 14:673–701, February 2013.

[Kak07] V. Kak. Infections in confined spaces: cruise ships, military barracks, and college dormitories. *Infect. Dis. Clin. North Am.*, 21(3):773–784, September 2007.

[KG07] D. Knowles and Z. Ghahramani. Infinite sparse factor analysis and infinite independent components analysis. In *Proc. International Conference on Independent Component Analysis and Signal Separation*, 2007.

[LCM10] Zhouchen Lin, Minming Chen, and Yi Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*, 2010.

[LP09] J. R. Lax and J. H. Phillips. How should we estimate public opinion in the states? *Am. J. Polit. Sci.*, 2009.

[LPJK07] P. Liang, S. Petrov, M. I. Jordan, and D. Klein. The infinite PCFG using hierarchical Dirichlet processes. In *EMNLP-CoNLL*, 2007.

[LWB90] J. B. Lee, A.S. Woodyatt, and M. Berman. Enhancement of high spectral resolution remote-sensing data by a noise-adjusted principal components transform. *IEEE Trans. Geosc. Remote Sensing*, 28:295304, 1990.

[Mar03] B. Marlin. Modeling user rating profiles for collaborative filtering. In *Proc. Neural Information Processing Systems*, 2003.

[MBP+08] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised dictionary learning. In *Proc. Neural Information Processing Systems*, 2008.

[MBP+09] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Non-local sparse models for image restoration. In *Proc. International Conference on Computer Vision*, 2009.

[MBPS09] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *Proc. International Conference on Machine Learning*, 2009.

[MES08] J. Mairal, M. Elad, and G. Sapiro. Sparse representation for color image restoration. *IEEE Trans. Image Processing*, 17:53–69, 2008.

[MSB07] A. Mohan, G. Sapiro, and E. Bosch. Spatially coherent nonlinear dimensionality reduction and segmentation of hyperspectral images. *IEEE Trans. Geosc. Remote Sensing Letters*, 4:206–210, 2007.

[MSE08] J. Mairal, G. Sapiro, and M. Elad. Learning multiscale sparse representations for image and video restoration. *SIAM Multiscale Modeling and Simulation*, 7:214 – 241, 2008.

[Mur02] K. Murphy. Hidden semi-Markov models. Technical report, 2002.

[OBB+00] P.D. O'Neill, D.J. Balding, N.G. Becker, M. Eerola, and D. Mollison. Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo methods. *Appl. Statist.*, 49:517–542, 2000.

[OD04] S.M. O'Brien and D.B. Dunson. Bayesian multivariate logistic regression. *Biometrics*, 60(3):739–746, September 2004.

[PC08] T. Park and G. Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103:681–686, 2008.

[PC09a] J. Paisley and L. Carin. Nonparametric factor analysis with beta process priors. In *Proc. International Conference on Machine Learning*, 2009.

[PC09b] John Paisley and Lawrence Carin. Nonparametric factor analysis with beta process priors. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 777–784. ACM, 2009.

[PM09] E. Patrozou and L.A. Mermel. Does influenza transmission occur from asymptomatic infection or prior to symptom onset? *Public Health Rep.*, 124(2):193–196, March-April 2009.

[PR85] K. T. Poole and H. Rosenthal. A spatial model for legislative roll call analysis. *Am. J. Polit. Sci.*, 1985.

[RBL+07] R. Raina, A. Battle, H. Lee, B. Packer, and A.Y. Ng. Self-taught learning: transfer learning from unlabeled data. In *Proc. International Conference on Machine Learning*, 2007.

[RDCD11] Lu Ren, Lan Du, Lawrence Carin, and David Dunson. Logistic stick-breaking process. *The Journal of Machine Learning Research*, 12:203–239, 2011.

[RPCL06] M. Ranzato, C. Poultney, S. Chopra, and Y. Lecun. Efficient learning of sparse representations with an energy-based model. In *Proc. Neural Information Processing Systems*, 2006.

[RW06] C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[RWBX13] R. Ranganath, C. Wang, D. M. Blei, and E. Xing. An adaptive learning rate for stochastic variational inference. In *ICML*, 2013.

[SC12] J. Silva and L. Carin. Active learning for online bayesian matrix factorization. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 325–333, New York, NY, USA, 2012. ACM.

[SCD+12] E. Salazar, M. Cain, E. Darling, S. Mitroff, and L. Carin. Inferring latent structure from mixed real and categorical relational data. In *ICML*, 2012.

[SCS+11] S. Stebbins, D.A.T. Cummings, J.H. Stark, C. Vukotich, K. Mitruka, W. Thompson, C. Rinalso, L. Roth, M. Wagner, S.R. Wisniewski, V. Dato, H. Eng, and D.S. Burke. Reduction in the incidence of Influenza A but not Influenza B associated with use of hand sanitizer and cough hygiene in schools. *Pediatric Infectious Disease J.*, 30(11):921–926, November 2011.

[SDC13] E. Salazar, D. B. Dunson, and L. Carin. Analysis of space–time relational data with application to legislative voting. *Comput. Stat. Data An.*, 2013.

[Set91] J. Sethuraman. A constructive definition of Dirichlet priors. Technical report, DTIC Document, 1991.

[SM08] R. Salakhutdinov and A. Mnih. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *ICML*, 2008.

[SWZS11] Y. Sun, Z. Wang, Y. Zhang, and J. Sundell. In China, students in crowded dormitories with a low ventillation rate have more common colds: evidence for airborne transmission. *PLoS One*, 6(11), November 2011.

[S.Y10] S.Yu. Hidden semi-Markov model. *Artificial Intelligence*, 174(2):215–243, February 2010.

[Tib94] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[Tip01] M. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, June 2001.

[TJ07a] R. Thibaux and M.I. Jordan. Hierarchical beta processes and the indian buffet process. In *Proc. International Conference on Artificial Intelligence and Statistics*, 2007.

[TJ07b] Romain Thibaux and Michael I Jordan. Hierarchical beta processes and the indian buffet process. In *International conference on artificial intelligence and statistics*, pages 564–571, 2007.

[TJBB04] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *J. Am. Stat. Assoc.*, 2004.

[TJBB06] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476), 2006.

[TKR09] V. Trifonov, H. Khiabanian, and R. Rabadan. Geographic dependence, surveillance, and origins of the 2009 Influenza A (H1N1) virus. *New England J. Med.*, 361:115–119, 2009.

[V.S93] V.Seshadri. *The Inverse Gaussian Distribution: a case study in exponential families*. Oxford Science Publications, 1993.

[Wes03] Mike West. Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian Statistics*, 7:723–732, 2003.

[WH89] M. West and J. Harrison. *Bayesian Forecasting and Dynamic Models*. Springer, 1989.

[WLS+10] E. Wang, D. Liu, J. Silva, L. Carin, and D. B. Dunson. Joint analysis of time-evolving binary matrices and associated documents. In *NIPS*, 2010.

[WMM09] Hanna M Wallach, David Minmo, and Andrew McCallum. Rethinking lda: Why priors matter. 2009.

[WPB11] C. Wang, J. Paisley, and D. M. Blei. Online variational inference for the hierarchical Dirichlet process. In *AISTATS*, 2011.

[WSDC13] E. Wang, E. Salazar, D. B. Dunson, and L. Carin. Spatio-temporal modeling of legislation and votes. *Bayesian Anal.*, 2013.

[WYG+09] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Trans. Pattern Analysis Machine Intelligence*, 31:210–227, 2009.

[YHD+10] Y. Yang, M.E. Halloran, M.J. Daniels, I.M. Longini Jr., D.S. Burke, and D.A.T. Cummings. Modeling competing infectious pathogens from a bayesian perspective: Application to influenza studies with incomplete laboratory results. *J. Amer. Statist. Assoc.*, 105(492):1310–1322, December 2010.

[YK03a] S. Yu and H. Kobayashi. An efficient forward-backward algorithm for an explicit-duration hidden Markov model. *IEEE Signal Process. Lett.*, 10(1):11–14, January 2003.

[YK03b] S. Yu and H. Kobayashi. A hidden semi-Markov model with missing data and multiple observation sequences for mobility tracking. *Signal Process.*, 83(2):235–250, February 2003.

[YWHM09] J. Yang, J. Wright, T. Huang, and Y. Ma. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, 2009.

[ZBB06] X. Zhang, W. Boscardin, and T. Belin. Sampling correlation matrices in Bayesian models with correlated latent variables. *J. Comput. and Graph. Statist.*, 15(4):880–896, 2006.

[ZBGS08] A. Zare, J. Bolton, P. Gader, and M. Schatten. Vegetation mapping for landmine detection using long wave hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 46:172–178, 2008.

[ZC12] X. Zhang and L. Carin. Joint modeling of a matrix with associated text via latent binary features. In *NIPS*, 2012.

[ZCL+09] A.K. Zaas, M. Chen, J. Lucas, T. Veldman, A.O. Hero, J. Varkey, R. Turner, C. Oien, S. Kingsmore, L. Carin, C.W. Woods, and G.S. Ginsburg. Peripheral blood gene expression signatures characterize symptomatic respiratory viral infection. *Cell Host & Microbe*, 6(3):207–217, September 2009.

[ZCP+09] M. Zhou, H. Chen, J. Paisley, L. Ren, G. Sapiro, and L. Carin. Non-parametric bayesian dictionary learning for sparse image representations. In *Proc. Neural Information Processing Systems*, 2009.

[ZG07] A. Zare and P. Gader. Sparsity promoting iterated constrained endmember detection in hyperspectral imagery. *IEEE Trans. Geosci. Remote Sens.*, 45:446450, 2007.

[ZG08] A. Zare and P. Gader. Hyperspectral band selection and endmember detection using sparsity promoting priors. *IEEE Geoscience and Remote Sensing Letters*, 5:256–260, 2008.

[ZHDC11] Mingyuan Zhou, Lauren Hannah, David Dunson, and Lawrence Carin. Beta-negative binomial process and poisson factor analysis. *arXiv preprint arXiv:1112.3605*, 2011.

[Zou06] H. Zou. The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.

Biography

Zhengming Xing was born in August 10th, 1986 in Zhengzhou, China. He received his B.S degree from Civil Aviation University of China in 2008, majoring in Electrical Engineering. He began his studies at Duke University in August 2008 and received M.S degree in December,2011. He expected to receive his Ph.D degree in Electrical and Computer Engineering in September 2014.

Publications

- **Z. Xing**, B.Nicholson, M.Jimenez, T. Veldman, L. Hudson, J. Lucas, D. Dunson, A. K. Zaas, C. W. Woods, G. S. Ginsburg and L. Carin "Bayesian modeling of temporal properties of infectious disease in a college student population." *Journal of Applied Statistics*, 1-25, 2013.

- **Z. Xing**, M. Zhou, A. Castrodad, G. Sapiro and L. Carin "Dictionary learning for noisy and incomplete hyperspectral images ." *Siam Journal on Imaging Sciences*, 5(1):33–56, 2012.

- M. Zhou, H. Chen, J. Paisley, L. Ren, L.Li, **Z. Xing**, D. Dunson, G. Sapiro and L. Carin "Nonparametric Bayesian Dictionary Learning for Analysis of Noisy and Incomplete Images." *IEEE Transactions on Image Processing*, 21(1):130–144, 2012.

- A. Castrodad, **Z. Xing**, J. B. Greer, E. Bosch, L. Carin and G. Sapiro "Learning discriminative sparse representations for modeling, source separation, and

mapping of hyperspectral imagery." *IEEE Transactions on Geoscience and Remote Sensing*, 49(11):4263–4281, 2011.

- A. Castrodad, **Z. Xing**, J. Greer, E. Bosch, L. Carin and G. Sapiro "Discriminative sparse representations in hyperspectral imagery " *ICIP*, 1313-1316, 2011.