Exploring the Mind: Integrating Questionnaires and fMRI

Esther Salazar[1] ESTHER.SALAZAR@DUKE.EDU
Ryan Bogdan[2] RBOGDAN@ARTSCI.WUSTL.EDU
Adam Gorka[3] ADAM.GORKA@DUKE.EDU
Ahmad R. Hariri[3,4] AHMAD.HARIRI@DUKE.EDU
Lawrence Carin[1] LCARIN@DUKE.EDU

[1]Department of Electrical & Computer Engineering, Duke University, Durham, NC, USA
[2]Department of Psychology, Washington University, St. Louis, MO, USA
[3]Department of Psychology & Neuroscience and [4]Institute for Genome Sciences & Policy, Duke University, Durham, NC, USA

Abstract

A new model is developed for joint analysis of ordered, categorical, real and count data. The ordered and categorical data are answers to questionnaires, the (word) count data correspond to the text questions from the questionnaires, and the real data correspond to fMRI responses for each subject. The Bayesian model employs the von Mises distribution in a novel manner to infer sparse graphical models jointly across people, questions, fMRI stimuli and brain region, with this integrated within a new matrix factorization based on latent binary features. The model is compared with simpler alternatives on two real datasets. We also demonstrate the ability to predict the response of the brain to visual stimuli (as measured by fMRI), based on knowledge of how the associated person answered classical questionnaires.

1. Introduction

This paper is motivated by analysis of heterogeneous data, such as ordered, categorical, real and count data. Such data are of interest, for example, in cognitive and brain science, in which subjects may answer questionnaires, and also (separately) undergo fMRI interrogation, with fMRI data measured as a function of visual stimulus. In this paper we focus on fMRI data, but the same approach may be applied to electroencephalography and other brain-imaging modalities. The fMRI data are typically real valued. One also has access to the text of the questions (word-count data), which ideally should be leveraged, to inform statistical relationships between the questions.

It is anticipated that the manner in which multiple people answer a given question, or how the brains of multiple people respond to a given stimuli, are *not* independent. While each individual is unique, there are typically statistical relationships between people and their brains, which we wish to infer. Further, for a particular person, the answers to multiple questions are typically not independent, and the fMRI responses to different stimuli are also typically not independent. We wish to infer these statistical relationships, *jointly* across people, questions, brain regions, stimuli, and the text of the questions.

The mapping of text questions to ordered/categorical answers from multiple people has been considered in the analysis of roll-call data (Zhang & Carin, 2012; Gerrish & Blei, 2011). Concerning the aforementioned borrowing of strength, one of the key contributions of this work concerns joint inference of the (sparse) graphical interrelationships between the people and between the questions, with this not addressed in the above roll-call studies.

The inferred graphs significantly generalize recent research on such methods as graphical lasso (Friedman et al., 2008) and related Bayesian models (Yoshida & West, 2010), in that we consider graph learning within the context of heterogenous data, and multiple sparse graphs are learned at once. Such joint analysis of the hierarchical relationships between multiple data axes is related to biclustering/co-clustering (Kriegel et al., 2009). However, here we generalize these relationships to sparse graphs (in co-clustering a tree construction

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

is *assumed*), and the analysis is performed here within a generative statistical model.

The joint analysis of ordered, categorical and real data was considered in (Salazar et al., 2012), but that work did not exploit text (count data), and it did not infer sparse graphical interrelationships. We demonstrate quantitatively that leveraging the text and sparse graphical models yields better predictive accuracy. The topics learned from the text also provide important interpretative value. In this paper we model the text with a state-of-the-art focused topic model (Williamson et al., 2010), while also accounting for available metadata (*e.g.*, labels on the types of questions).

A contribution of this paper concerns the joint analysis of how people answer questionnaires and how their brain responds to external stimuli (here visual), the latter measured via fMRI. Researchers within the machine learning community have analyzed fMRI and EEG data, with such goals as predicting what finite set of objects an individual is thinking of based upon fMRI/EEG data (Mitchell et al., 2008; Fyshe et al., 2012). In this paper we ask a novel and practical question, which to our knowledge has not been considered previously: can one predict the fMRI response (here from the amygdala) to external stimuli, based upon knowledge of how the subject answers a questionnaire?

To integrate the multiple forms of data, we generalize the binary matrix factorization introduced in (Meeds et al., 2007) and further developed in (Salazar et al., 2012). Specifically, the latent binary features of each data axis are modeled via a probit model, and the latent real variables in the probit are modeled via a multivariate normal distribution, with sparse precision matrix. We perform a detail quantitative analysis of the proposed model, and demonstrate significant performance gains relative to an Indian buffet process (IBP) model of the latent binary features (Meeds et al., 2007) and to a generalized Tucker model, which employs no latent binary features (Xu et al., 2012). The sparse inverse covariance matrix is implemented in a novel manner, coupling a covariance decomposition in (Yoshida & West, 2010) with a new utilization of the matrix von Mises-Fisher distribution (Hoff, 2009).

2. Modeling Framework

2.1. Notation

Let $\boldsymbol{Y}^r = \{y^r_{ij}\}$ denote the $N \times P_1$ matrix of real responses, where people are indexed $i = 1, \ldots, N$, and components of the real vector of data are indexed $j = 1, \ldots, P_1$. Let $\tilde{\boldsymbol{Y}}^o = \{\tilde{y}^o_{ij}\}$ denote an $N \times P_2$ matrix of ordered responses for the same N subjects, with $j = 1 \ldots, P_2$. We also employ a distinct probit model for the categorical data, along the lines in (Salazar et al., 2012); here the discussion focuses on the real and ordered parts of the data for conciseness, as this is sufficient to elucidate the driving components of the model.

In practice the questions can be partitioned into Q mutually exclusive sets (questionnaire types). For $q = 1, \ldots, Q$, let $\mathcal{I}_q$ denote the index set containing all questions in the qth questionnaire, *i.e.*, $\mathcal{I}_1 = \{1, \ldots, J_1\}$, $\mathcal{I}_2 = \{J_1 + 1, \ldots, J_1 + J_2\}$ and so on, where $\sum_{q=1}^{Q} |\mathcal{I}_q| = P_2$ and $|\mathcal{I}_q|$ denotes the cardinality of the set $\mathcal{I}_q$. Following this assumption, we have $\tilde{\boldsymbol{Y}}^o = (\tilde{\boldsymbol{Y}}^o_1, \ldots, \tilde{\boldsymbol{Y}}^o_Q)$ where $\tilde{\boldsymbol{Y}}^o_q = \{\tilde{y}^o_{ij}\}$, for $j \in \mathcal{I}_q$, denotes the $N \times J_q$ matrix of ordered responses for the qth questionnaire. Finally, assuming that questions in the qth questionnaire have $L_q + 1$ possible answers, we have that $\tilde{y}^o_{ij} \in \{0, \ldots, L_q\}$. We assume all questions in a given questionnaire have the same number of possible answers, for notational simplicity.

2.2. Ordered probit model

Let $\boldsymbol{Y}^o = \{y^o_{ij}\} \in \mathbb{R}^{N \times P_2}$ be a *latent* response matrix, where y^o_{ij} denotes a Gaussian random variable with mean μ^o_{ij} and unit variance. For $j \in \mathcal{I}_q$, we assume there are $L_q - 1$ ordered cut-points $\boldsymbol{c}^{(q)} = (c^{(q)}_1, \ldots, c^{(q)}_{L_q - 1})$ where $0 \leq c^{(q)}_1 \leq c^{(q)}_2 \leq \ldots \leq c^{(q)}_{L_q - 1}$. Then, for $i = 1, \ldots, N$, $j \in \mathcal{I}_q$ and for every questionnaire q, the ordered probit model is defined by $\tilde{y}^o_{ij} = h \in \{0, \ldots, L_q\}$ if $c^{(q)}_{h-1} \leq y^o_{ij} < c^{(q)}_h$, where $c^{(q)}_{-1} = -\infty$, $c^{(q)}_0 = 0$ and $c^{(q)}_{L_q} = \infty$, and $y^o_{ij} = \mu^o_{ij} + \epsilon^o_{ij}$, with $\epsilon^o_{ij} \sim \mathcal{N}(0, 1)$. Consequently, $\Pr(\tilde{y}^o_{ij} \leq j) = \Pr(y^o_{ij} < c^{(q)}_j) = \phi(c^{(q)}_j - \mu^o_{ij})$ where ϕ is the standard normal distribution function. The parameters of this model are the mean component μ^o_{ij} and the cut-points $\boldsymbol{c}^{(q)}$. As considered in Albert & Chib (1997) and Albert & Chib (2001), we transform $\boldsymbol{c}^{(q)}$ into a real-valued vector $\boldsymbol{\alpha}^{(q)} = (\alpha^{(q)}_1, \ldots, \alpha^{(q)}_{L_q - 1})$ such that $\alpha^{(q)}_1 = \log c^{(q)}_1$ and $\alpha^{(q)}_j = \log(c^{(q)}_j - c^{(q)}_{j-1})$ for $2 \leq j \leq L_q - 1$ with inverse map given by $c^{(q)}_j = \sum_{i=1}^{j} \exp(\alpha^{(q)}_i)$ for $1 \leq j \leq L_q - 1$. We consider a multivariate normal prior distribution for $\boldsymbol{\alpha}^{(q)}$ to simplify posterior inference.

For categorical observations (no preferred ordering in the answers), we employ the probit construction in Salazar et al. (2012), to which the reader is referred for details.

2.3. Binary & low-rank matrix factorization

We are now interested in performing decompositions of the real matrices $\boldsymbol{Y}^r$ and $\boldsymbol{Y}^o$, with a similar decomposition performed for the categorical data. Consider the decompositions

$$\boldsymbol{Y}^r = \boldsymbol{L}\boldsymbol{M}^r\boldsymbol{R}_r^T + \boldsymbol{E}^r \ , \quad \boldsymbol{Y}^o = \boldsymbol{L}\boldsymbol{M}^o\boldsymbol{R}_o^T + \boldsymbol{E}^o \quad (1)$$

where $\boldsymbol{L} \in \{0,1\}^{N\times K}$, $\boldsymbol{R}_r \in \{0,1\}^{P_1\times K}$, $\boldsymbol{R}_o \in \{0,1\}^{P_2\times K}$, $\boldsymbol{M}^r \in \mathbb{R}^{K\times K}$, $\boldsymbol{M}^o \in \mathbb{R}^{K\times K}$, $\boldsymbol{E}^r \in \mathbb{R}^{N\times P_1}$ and $\boldsymbol{E}^o \in \mathbb{R}^{N\times P_2}$; note that $\boldsymbol{L}$ is shared for modeling $\boldsymbol{Y}^r$ and $\boldsymbol{Y}^o$, because the same N subjects are responsible for the real, ordered and categorical data. Each component of $\boldsymbol{E}^o$ is drawn i.i.d. from $\mathcal{N}(0,1)$, and each component of $\boldsymbol{E}^r$ is drawn i.i.d. from $\mathcal{N}(0,\alpha_r^{-1})$, with a flat/broad gamma prior placed on α_r.

For either $a = r$ or $a = o$, we employ a decomposition

$$\boldsymbol{M}^a = \sum_{k=1}^{K} \lambda_k^a \boldsymbol{u}_{a,k}\boldsymbol{v}_{a,k}^T, \quad (2)$$

where $\lambda_k^a \in \mathbb{R}$, $\boldsymbol{u}_{a,k} \in \mathbb{R}^K$ and $\boldsymbol{v}_{a,k} \in \mathbb{R}^K$. We employ priors $\boldsymbol{u}_{a,k} \sim \mathcal{N}(0, \boldsymbol{I}_K)$ and $\boldsymbol{v}_{a,k} \sim \mathcal{N}(0, \boldsymbol{I}_K)$.

As proposed in (Bhattacharya & Dunson, 2011), we use the following shrinkage prior for λ_k^a: $\lambda_k^a|\tilde{\tau}_k \sim \mathcal{N}(0, \tilde{\tau}_k^{-1})$, $\tilde{\tau}_k = \prod_{h=1}^{k} \tilde{\delta}_h$, $\tilde{\delta}_h|a_1 \sim \text{Gamma}(a_1, 1)$, where $a_1 > 1$ and $\tilde{\tau}_k$ tends stochastically towards infinity as k goes to infinity, shrinking λ_k^a toward zero with increasing k. Therefore, the representation in (2), with λ_k^a so defined, encourages that $\boldsymbol{M}^a$ be near low-rank (*near* because the singular values diminishes quickly with increasing k, but do not go to exactly zero). This is analogous to the nuclear norm in related low-rank research (Candes & Recht, 2008).

2.4. Correlated binary features

We consider construction of the binary matrix $\boldsymbol{L}$, with identical constructions employed for $\boldsymbol{R}_o$ and $\boldsymbol{R}_r$. Our goal is to infer a sparse graphical model between the N subjects characterized by $\boldsymbol{L}$. Letting $\boldsymbol{L} = \{l_{ik}\}$, we employ

$$l_{ik}|\eta_{ik} = \begin{cases} 1 & \text{if } \eta_{ik} > 0 \\ 0 & \text{otherwise} \end{cases} \ , \quad \boldsymbol{\eta}_k \sim \mathcal{N}(0, \boldsymbol{\Sigma}_L) \quad (3)$$

with $\boldsymbol{\eta}_k = (\eta_{1k}, \ldots, \eta_{Nk})^T \in \mathbb{R}^N$. Matrix $\boldsymbol{\Sigma}_L \in \mathbb{R}^{N\times N}$ imposes an underlying covariance structure on $\boldsymbol{l}_k = (l_{1k}, \ldots, l_{Nk})^T$ for $k = 1, \ldots, K$, and here we wish to impose that the precision matrix $\boldsymbol{\Sigma}_L^{-1}$ is sparse, allowing inference of a sparse dependency graph between the N subjects (Friedman et al., 2008; Yoshida & West, 2010). Extending the construction in (Yoshida & West, 2010), we employ

$$\boldsymbol{\eta}_k = \boldsymbol{\Psi}^{1/2}\boldsymbol{\Phi}_B\boldsymbol{f}_k + \boldsymbol{\epsilon}_k, \quad \boldsymbol{\Phi}_B = \boldsymbol{\Phi} \circ \boldsymbol{B},$$

$$\boldsymbol{f}_k \sim \mathcal{N}(0, \boldsymbol{\Delta}), \quad \text{and} \quad \boldsymbol{\epsilon}_k \sim \mathcal{N}(0, \boldsymbol{\Psi}), \quad (4)$$

where $\boldsymbol{\Delta} = \text{diag}(\delta_1, \ldots, \delta_K)$, $\boldsymbol{\Psi} = \text{diag}(\psi_1, \ldots, \psi_N)$, $\boldsymbol{\Psi}^{1/2}\boldsymbol{\Phi}_B \in \mathbb{R}^{N\times K}$ represents the factor loading matrix, $\boldsymbol{B} = \{b_{ij}\} \in \{0,1\}^{N\times K}$ is a binary matrix that defines the sparsity pattern of the factor loading matrix, and $\boldsymbol{\Phi}_B \in \mathbb{R}^{N\times K}$ is orthogonal (*i.e.*, $\boldsymbol{\Phi}_B^T\boldsymbol{\Phi}_B = \boldsymbol{I}_K$) with $\circ$ representing the element-wise product. Consequently, the covariance matrix, $\boldsymbol{\Sigma}_L$, and the corresponding precision matrix, $\boldsymbol{\Sigma}_L^{-1}$, are given by

$$\begin{aligned} \boldsymbol{\Sigma}_L &= \boldsymbol{\Psi}^{1/2}\left(\boldsymbol{I}_N + \boldsymbol{\Phi}_B\boldsymbol{\Delta}\boldsymbol{\Phi}_B^T\right)\boldsymbol{\Psi}^{1/2}, \\ \boldsymbol{\Sigma}_L^{-1} &= \boldsymbol{\Psi}^{-1/2}\left(\boldsymbol{I}_N - \boldsymbol{\Phi}_B\boldsymbol{T}\boldsymbol{\Phi}_B^T\right)\boldsymbol{\Psi}^{-1/2}, \end{aligned}$$

where $\boldsymbol{T} = \text{diag}(\tau_1, \ldots, \tau_K)$ and $\tau_k = \delta_k/(1+\delta_k)$ for $k = 1, \ldots, K$. Note that the sparse loading matrix, $\boldsymbol{\Psi}^{1/2}\boldsymbol{\Phi}_B$, induces some zero elements in the covariance matrix. However, an important property of model (4) is that the orthogonal property of the sparse loading matrix also induces off diagonal zeros in the precision matrix $\boldsymbol{\Sigma}_L^{-1}$ which defines a conditional independence or Gaussian graphical models. In particular, under this construction, the location of zero entries (sparse structure) in the covariance and precision matrices are exactly the same. Here, $\boldsymbol{\Psi}$ is fixed to be the identity to allow a simple identification strategy.

The model parameters to be inferred are $\boldsymbol{\Phi}$, $\boldsymbol{\Delta}$ and $\boldsymbol{B}$. Following the Bayesian approach, we must now place priors on these parameters that may reflect a *priori* knowledge. Specifically, the joint prior distribution is $p(\boldsymbol{\Phi}, \boldsymbol{\Delta}, \boldsymbol{B}) = p(\boldsymbol{\Phi}|\boldsymbol{B})p(\boldsymbol{\Delta})p(\boldsymbol{B})$. Note that the prior distribution for $\boldsymbol{\Phi}$ depends on the configuration of $\boldsymbol{B}$, i.e. the prior, if available, is defined over the non-zero factor loadings. However, here we consider a uniform prior distribution on $\boldsymbol{\Phi}$. As pointed out in (Yoshida & West, 2010), this prior involves a uniform density on the hypersphere defined by the orthogonality constraint that is conditioned by setting some elements of $\boldsymbol{\Phi}$ to zero (this is done considering the location of zero elements in $\boldsymbol{B}$). Moreover, the uniform prior distribution for $\boldsymbol{\Phi}$ imply that

$$p(\boldsymbol{\Phi}_B|\boldsymbol{\eta}^*, \boldsymbol{F}) \propto \text{etr}\left((\boldsymbol{\eta}^*\boldsymbol{F})^T\boldsymbol{\Phi}_B\right),$$

where $\boldsymbol{\eta}^* = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_K)^T$, $\boldsymbol{F} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_K)^T$ and $\text{etr}(\cdot)$ denotes the exponential trace function. The above representation is equivalent to the matrix von Mises-Fisher distribution. Therefore, posterior inference for $\boldsymbol{\Phi}_B$ is obtained by iteratively sampling from this distribution (Hoff, 2009).

Concerning the sparse structure $\boldsymbol{B}$, we consider independent beta priors on the binary variates b_{ij}, i.e., $p(b_{ij}) = \text{Beta}(a, b)$. Finally, for each element of $\boldsymbol{\Delta}$, δ_k $(k = 1, \ldots, K)$, we consider an inverse gamma prior such that $\delta_k \sim \text{IG}(c, d)$.

2.5. Focused topic modeling of questions

In Section 2.4 we imposed a prior on the binary matrix $\boldsymbol{R}_o$, and a similar binary feature matrix is manifested for the categorical questions. We now integrate the latent binary features of the questions to a topic model of the text, using a generalization of the focused topic model (FTM) first developed in (Williamson et al., 2010).

Let $\boldsymbol{r}_j^o$ represent the jth row of $\boldsymbol{R}_o$, corresponding to the jth ordered question. Constitute a K-dimensional positive, real vector $\boldsymbol{\phi}$, with component k drawn $\phi_k \sim \text{Gamma}(\gamma, 1)$. Then the generative process for drawing the words for question j is: (1) Draw topic proportions for the question $\boldsymbol{\theta}_j = (\theta_{j1}, \ldots, \theta_{jK}) \sim \text{Dir}(\boldsymbol{r}_j^o \circ \boldsymbol{\phi})$. (2) For the nth word in question j, draw the topic index $z_{jn} \sim \text{Mult}(1, \boldsymbol{\theta}_j)$, and then draw word $w_{jn} \sim \text{Mult}(1, \boldsymbol{\beta}_{z_{jn}})$, where $\boldsymbol{\beta}_k \sim \text{Dir}(\boldsymbol{\eta})$ is a distribution over words for topic k (we typically set $\boldsymbol{\eta} = (1/W, \ldots, 1/W)^T$, where the vocabulary is of size W).

In the above construction the binary feature vector $\boldsymbol{r}_j^o$ associated with question j was latent. In practice, we may have multiple types of questionnaires, such as the Q questionnaires discussed in Section 2.1. To account for these known labels, we append a Q-dimensional binary vector to $\boldsymbol{r}_j^o$, which is all zeros with a single one; this imposes that one topic distribution $\boldsymbol{\beta}_k$ is explicitly associated with each of the questionnaire types (the remaining topics are shared across all questionnaires). Through this construction we also learn an associated word distribution (topic) with each of the latent binary features, evincing understanding to the latent binary features characteristic of the questions.

2.6. Other modeling choices & related work

We consider other options for modeling the real matrices $\boldsymbol{Y}^r$ and $\boldsymbol{Y}^o$, and perform comparisons in Section 4.1. The basic form of the binary matrix factorization in (1) was proposed in (Meeds et al., 2007), in which $\boldsymbol{L}$, $\boldsymbol{R}_o$ and $\boldsymbol{R}_r$ were constituted via the Indian buffet process (IBP), which assumes that the order of the rows and columns is exchangeable. Additionally, in (Meeds et al., 2007) the matrix $\boldsymbol{M}^a$ was *not* modeled as low-rank. When presenting comparison results, the model in (Meeds et al., 2007) is denoted *IBP*. We examine the importance of imposing low-rank on $\boldsymbol{M}^a$, while retaining an IBP on $\boldsymbol{L}$, $\boldsymbol{R}_o$ and $\boldsymbol{R}_r$; this comparison model is termed *IBP+Low Rank*. We wished to examine the importance of imposing that $\boldsymbol{L}$, $\boldsymbol{R}_o$ and $\boldsymbol{R}_r$ are binary. When these matrices are modeled as real, we effectively have a Tucker-like model (Xu et al., 2012). We consider such a construction, in which $\boldsymbol{\eta}_k$ is used to directly model the columns of $\boldsymbol{L}$, $\boldsymbol{R}_o$ and $\boldsymbol{R}_r$, *without* the probit model. This is referred to as *Tucker+Sparse Precision*. In this context, to examine the importance of a sparse $\boldsymbol{\Sigma}_L^{-1}$, we consider the same Tucker representation, but the matrix $\boldsymbol{\Sigma}_L$ is modeled as low-rank and sparse, as in (Salazar et al., 2012); this is referred to below as *Tucker+Low Rank*. To the authors' knowledge the *Tucker+Sparse Precision* and *Tucker+Low Rank* models are new. When comparing to the models above *without* the binary representations for $\boldsymbol{L}$, $\boldsymbol{R}_o$ and $\boldsymbol{R}_r$, we do not consider the text of the questions, because the integration of the topic model in Section 2.5 requires the binary decomposition for linkage to the topics.

The proposed model (with binary factorization) is denoted *BMF+Sparse Precision*, for binary matrix factorization (BMF). In this context, to examine the importance of the sparse precision matrix, we consider all aspects of the proposed model unchanged, but now $\boldsymbol{\Sigma}_L$ is modeled as sparse and low-rank (Salazar et al., 2012); this is referred to as *BMF+Low Rank*.

Finally, the recent work of (Hahn et al., 2012) directly considered a real factor-model representation of $\boldsymbol{Y}^o$ (not Tucker or BMF), where the factor scores and loadings are real, and the factor loadings are sparse; this model is referred to as *Sparse Matrix Factorization*.

3. Posterior inference

Bayesian model inference is performed via a Markov chain Monte Carlo (MCMC) algorithm which involves Gibbs sampling and Metropolis-Hastings steps for a subset of the parameters. Specifically,

- For the ordered probit model, we use the algorithm proposed in (Albert & Chib, 1997) for the cumulative model for ordinal responses independently for every questionnaire. As pointed out in Section 2.2, the transformed cut-point vector $\boldsymbol{\alpha}^{(q)}$ follows a multivariate normal prior distribution and the full conditional posterior does not have a close form, therefore a Metropolis-Hastings step is employed. The other parameters are drawn via Gibbs sampling.

- Updating the binary features $\boldsymbol{L}$ and $\boldsymbol{R}_r$ is straightforward and tractable since each elements of the matrix are updated from Bernoulli distributions with updated probabilities. Also, the real value vectors $\boldsymbol{u}_{a,k}$ and $\boldsymbol{v}_{a,k}$

(for either $a = r$ or $a = o$) are updated from Gaussian distributions (see Salazar et al., 2012, for more details).

• In order to sample the binary matrix $\boldsymbol{R}_o$ and the FTM parameters, we use the algorithms proposed in (Zhang & Carin, 2012) and (Williamson et al., 2010), respectively.

• For orthogonal matrix $\boldsymbol{\Phi}_B$, as considered in (Yoshida & West, 2010), we modified the orthogonality condition and assume that $\boldsymbol{\Phi}^T\boldsymbol{\Phi} = \boldsymbol{I}_K$ with $b_{ij} = 1$ for all i and j. Then, we updated the matrix $\boldsymbol{\Phi}$ using a Gibbs sampling scheme proposed in (Hoff, 2009). The algorithm use Gibbs sampling to construct a dependent Markov chain from the full conditional distribution which converges in distribution to the von Mises distribution. Finally, we update the elements of the sparse binary matrix $\boldsymbol{B}$ using Bernoulli distributions with updated probabilities derived from the full conditional posterior distribution.

The complete software package will be released upon publication.

4. Applications

4.1. Senate voting data

To compare components of the proposed model to the alternatives in Section 2.6, we first test performance on the US Senate voting dataset. This consists of a binary vote matrix, $\tilde{\boldsymbol{Y}}^o \in \{0,1\}^{102\times 657}$, from the U.S. Senate during the 110th Congress (January 2007 to January 2009). In this application, we only use the binary observations without the use of associated legislative text, to help elucidate the power of the proposed matrix representation (and the real matrix decompositions are not appropriate for the integration to the topic model, as in Section 2.5).

We perform analysis with $K = 50$ (the inferred rank of $\boldsymbol{M}^o$ was less than 5, and larger values of K yielded very similar results). We considered 10,000 MCMC iterations, with a burn-in of 2,000; we collected every fourth sample, yielding 2,000 collection samples. The posterior results indicate that there are approximately 15 latent binary features for senators and 20 latent binary features for legislation.

From the learned sparse precision matrix of the legislators, we are able to construct a network between the senators (we simultaneously learn such a sparse network for the legislation, which is omitted for brevity). We use a common technique for visualizing social networks called *spring-embedding* (Eades, 1984; Fruchterman & Reingold, 1991). In particular, we use the Kamada-Kawai algorithm (Kamada & Kawai, 1989) implemented in the open source program `Pajek`[1]. Figure 1, left panel, shows the learned connectivity network for senators derived from the precision matrix.

Note that Republican and Democrat senators form two principal clusters, with strong ties within both groups. Arlen Specter is depicted as a Democrat (blue circle), but his network is highly linked to Republicans (red circles). Further, the two Independents are closely linked with Democrats. Senator Specter was a Republican senator for over two decades, before switching in this legislative session to a Democrat, and the two Independents caucused with the Democrats.

Figure 1, right panel, shows the average fraction of correct predictions for different percentage (40%, 50%, 60%, 70%) of missingness; comparisons are made to the models summarized in Section 2.6. The missing values were selected uniformly at random, and the mean and standard deviation (error bars) are shown, based on 15 runs. A clear advantage is manifested by imposing low-rank on $\boldsymbol{M}^a$, particularly comparing *IBP* and *IBP+Low Rank*. For both the real (Tucker) and binary (BMF) models for $\boldsymbol{L}$, $\boldsymbol{R}_o$ and $\boldsymbol{R}_r$, imposing correlation between the rows and columns yields significant performance gains. Generally the sparse graphical relationship between the rows and columns yields better performance relative to the low-rank covariance structure, and therefore it appears that the sparse graphical representation may be better matched to the (real-world) data considered (we see this below as well for the neuroscience application). Overall, the proposed model, with binary factorization and sparse precision matrix, yields the best quantitative performance, against a wide range of alternative models.

4.2. fMRI data and behavioral questionnaires

We perform joint analysis of real, ordered, categorical, and word-count (text) data. Specifically, the real data is composed of $P_1 = 16$ measurements associated with fMRI data from the amygdala. The data are associated with each hemisphere (left/ right) and subregions (basolateral/central-medial); for each of these four amygdala regions the response is measured to four types of visual stimuli, meant to be associated with fear, anger, surprise and be neutral (see Figure 2). The four spatial regions and four stimuli are summarized in the 16-dimensional real vector. We target the amygdala for several reasons (Hariri, 2009). First, the amygdala is a critical neural hub for learning predictive links between stimuli that signal important changes in our environment and subsequently generat-

[1] See http://pajek.imfm.si/doku.php for more details

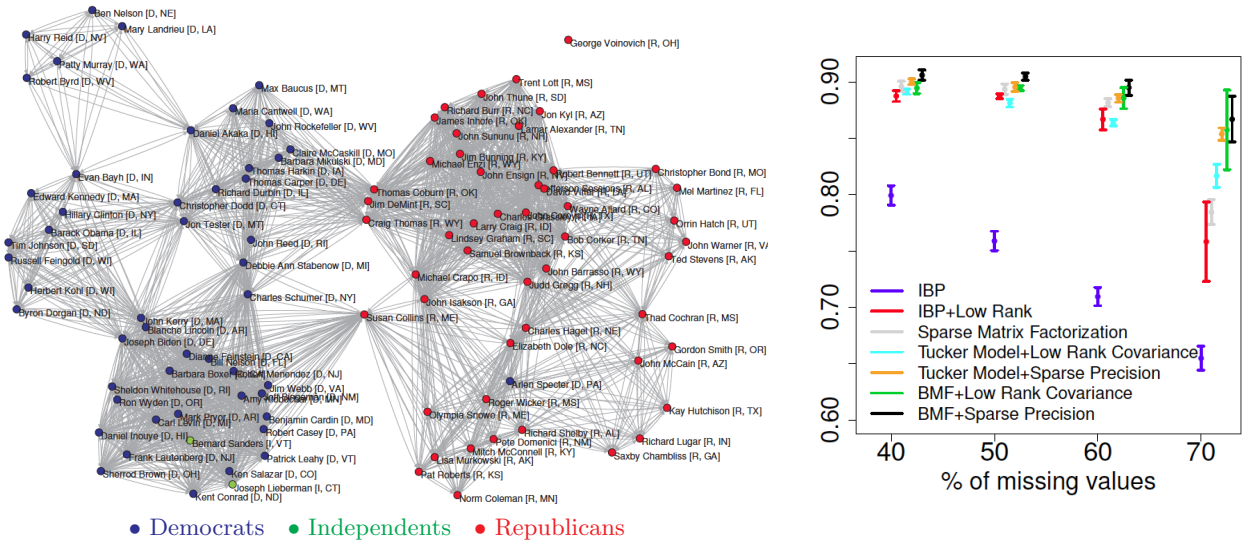

Figure 1. Voting data. Left panel: Learned connectivity network between senators using a sparse graphical factor model with probit link. The colored bullets represent the parties. Right panel: Average fraction of correct predictions as a function of the fraction of missing data (averaged over 15 runs). The results, using the proposed model, were compared with six related models. Error bars indicate the standard deviation around the mean. The results of the different models are shifted slightly horizontally with respect to each other, for better viewing.

ing adaptive behavioral and physiologic responses. A second related feature is a generally robust reactivity of the amygdala to visual cues signaling such changes that is readily measured by fMRI and exhibits considerable inter-individual variability. Third, amygdala reactivity as measured by fMRI is stable over time within individuals and thus represents a trait-like feature similar to those assessed by the questionnaires included in our analyses (Manuck et al., 2007). Finally, the amygdala is clearly important in the emergence of psychopathology, particularly mood and anxiety disorders, and variability in its reactivity is associated with individual differences in relative risk for dysfunction as well as response to common treatments. Thus, developing models that can predict amygdala reactivity with fidelity without the need for direct assays via neuroimaging has tremendous potential to advance ongoing efforts to better treat and even prevent mental illness at a population level.

The ordered/categorical data consist of answers to questions concerning different aspects of behavior, personality and life experiences (*e.g.* anxiety, interpersonal support, psychopathy, drug use, stress, NEO personality, mindfulness, insomnia, eating disorders, childhood trauma, depression and alcohol use).

The questions are grouped into $Q = 18$ widely used questionnaires, with $P_2 = 616$ questions in total. Data were collected from $N = 400$ people. Respondents

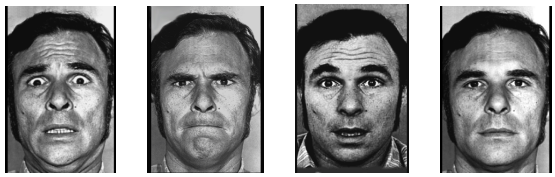

Figure 2. Prototypical examples of the 4 facial expressions used for visual stimuli.

ranged in age from 18 to 22 years old (college students). The percentage of missing values is approximately 2%. Additionally, we process the text of the questions. The joint data analysis is performed with $K = 50$, $a_1 = 3$ (shrinkage prior parameter), and $\gamma = 6$ (FTM hyperparameter). For the beta parameters, we set $a = b = 0.5$ which imposes a horseshoe-shaped prior. The MCMC algorithm was run for 10,000 iterations with a burn-in of 5,000 and then every second sample was collected, to yield 2,500 posterior samples.

Figure 3 shows the learned connectivity network for the amygdala fMRI measurements (top panel) and for the questions (bottom panel); a similar sparse graphical interrelationship is also inferred for the people (omitted), with all graphs learned simultaneously. The observed pattern of subregion- and expression-specific amygdala reactivity is consistent with both theoreti-

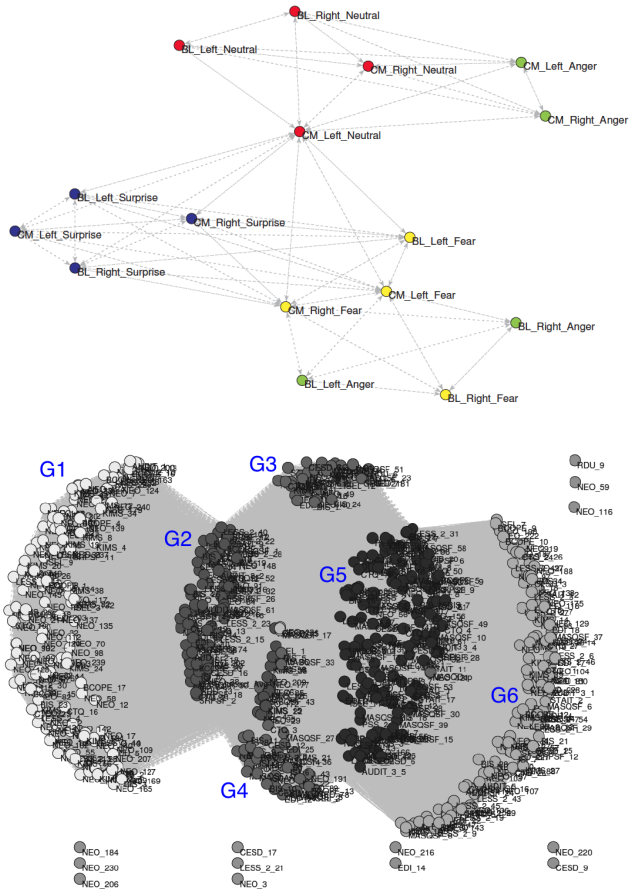

Figure 3. Top panel: Learned connectivity network for amygdala fMRI measurements. Bottom panel: Learned connectivity network for questions. G1, ..., G6 indicate groups of questions in the network.

cal models and empirical data revealing a sensitivity of the amygdala to specific feature-based cues that vary in their intensity across categorical facial expressions (Ahs et al., 2013). Specifically, amygdala reactivity in the basolateral and central-medial subregions in both the left and right hemispheres observes a correspondence with greater displacement of the eyebrows in neutral and surprised expressions relative to fearful and angry expressions (along a vertical axis in the top panel of Figure 3) as well as with elevation of the upper eyelid revealing more of the eye whites together with lowering of the brow in neutral and angry expressions relative to fearful and surprised expressions (along a horizontal axis in the top panel of Figure 3). A novel feature of our model not anticipated by prior research is the subregional segregation of reactivity to anger with the central-medial subregions in both hemispheres aligning with broader reactivity to neutral expressions and that of the basolateral subregions in both hemispheres with reactivity to fear. This subregional expression-specific pattern will be important to explore in future work.

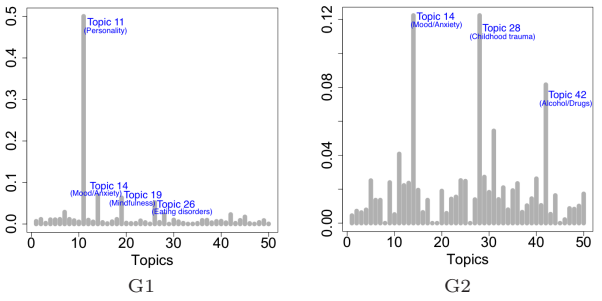

Figure 4. Example average topic distributions for two of the clusters in Figure 3.

Table 1. Six example learned topics from the questionnaire data, with the top-ten most probable words. These topics are associated with particular classes of questionnaires, identified within parentheses.

Topic 11 (Personality)	Topic 14 (Mood/Anxiety)	Topic 19 (Mindfulness)
FEEL	AFRAID	PROBLEM
WORRIER	SHAKY	FEELING
DOMINANT	FAMILY	HELP
PEOPLE	FEELINGS	FIND
CRAFTY	HURT	MYSELF
CHATTING	WORTHLESS	DAYDREAMING
SELDOM	SELF-CONFIDENCE	CRITICIZE
THINK	FELT	EMOTIONS
EMOTIONALLY	BELIEFS	PERSON
MYSELF	HAPPY	WORK

Topic 26 (Eating disorders)	Topic 28 (Childhood Trauma)	Topic 42 (Alcohol/drugs)
FOOD	FAMILY	DRINK
WEIGHT	STUPID	ALCOHOL
FEEL	PARENTS	REBELLIOUS
MYSELF	SEXUAL	DRUGS
THINK	EMOTIONALLY	OFTEN
EAT	TEACHER	RULES
NERVOUS	DRUNK	GUILTY
UPSET	EAT	SAD
SIZE	PROTECT	FELT
STUFF	PEOPLE	MYSELF

The connectivity network for questions (Figure 3, bottom panel), manifested via the sparse graphical model prior, reveals six groups of questions (denoted by G1, ..., G6) and thirteen independent questions. We note that each group is composed of questions from different questionnaires. In order to interpret each group, we analyzed the learned topics associated with each of them. Considering the most likely collection sample (for illustrative purposes), we computed the average distribution over topics for every question in a group. Figure 4 shows the results for two example

groups (other groups omitted in this figure for brevity). Some topics are widely used in some groups but less used in others. For instance, "Topic 11", related to personality traits or facets is used in groups G1 and G5. On the other hand, "Topic 14", related with mood and anxiety symptoms, is widely used in the groups G2 and G4. For additional interpretations, Table 1 shows six learned topics with the top-ten most probable words within each topic (note that the topic number is arbitrary, and happened to be associated with the most-probable collection sample). The topics in Table 1 corresponds to those associated with particular survey types (recall Section 2.5), with the survey noted in parentheses.

It is of interest to examine whether given the answers to the questionnaire, we can predict the corresponding 16-dimensional real vector associated with the amygdala data. In the training phase, we build a model using the full data from 80% of the subjects. The learned model is then employed to perform testing on the remaining 20% of data, where for these we only assume access to the questionnaires; the goal is to predict from the questionnaire data the associated 16-dimensional vector associated with the amygdala. We did this for 10 runs, and for each run the 20% of subjects for which amygdala data were held out were selected uniformly at random. When testing using the proposed model, we infer the latent binary features characteristic of the subject/person based only on the questionnaires, and then these binary features are employed within the learned model to predict the fMRI data.

Table 2. Average of the mean percent error (MPE) for Amygdala fMRI predictions for 20% of missing data (held-out data) averaged over 10 runs and standard deviations.

Model	Average MPE	Standard Deviation
Regression	66.7%	1.01%
BMF-Cov.(no text)	14.5%	0.83%
BMF-Prec.(no text)	14.2%	1.47%
BMF-Cov. (text)	12.6%	1.40%
BMF-Prec. (text)	11.3%	0.20%

To assess the prediction performance, we compare five models. The first is a regression model where the covariates are the ordered-categorical responses (imposing a sparseness prior on the regression weights); this is a simple baseline model for comparison. Using notation from Section 2.6, the other four models correspond to *BMF+Sparse Precision* and *BMF+Low rank*. We are interested in examining the utility of using text from the questions, so we use the BMF constructions (we could also use IBP, but the performance is markedly worse). When examining *BMF+Sparse Precision* and *BMF+Low rank*, we consider the models with and without use of the text, leading to four comparisons of this type. Table 2 shows the average mean percent error and the standard deviation for each model. The regression model based directly on the questionnaire answers yields performance that is markedly worse than that of variants of the proposed model. Using the text of the questions improves performance (helps learn the correlation structure between the questions), and the proposed *BMF+Sparse Precision* model with text yields the overall best performance.

5. Conclusions

A new model has been developed for joint analysis of ordered, categorical, real, and (word) count data. A key component of the model is development of a new framework for *jointly* learning sparse graphical models along multiple axes of the heterogeneous data. In the context of the motivating problem of integrating questionnaire and fMRI data, the sparse graphs are learned simultaneously for the people, questions, and components of the fMRI data. Comparisons have been made to numerous variations of the model, and to related published models. Encouraging results have been obtained on two real-world datasets, including the ability to predict the fMRI response of the amygdala, based on the answers to questionnaires considered by the subject. Comparisons on two real-world examples indicate that the sparse graphical model for the relationships between the rows and columns may be more consistent with real data than a low-rank and sparse covariance matrix.

References

Ahs, F., Davis, F.C., Gorka, A.X., and Hariri, A.R. Feature-based representations of emotional facial expressions in the human amygdala. *Social, Cognitive, and Affective Neuroscience*, 2013.

Albert, J. H. and Chib, S. Bayesian methods for cumulative, sequential and two-step ordinal data regression models. Technical report, Department of Mathematics and Statistics, Bowling Green State University, Ohio, 1997.

Albert, J. H. and Chib, S. Sequential ordinal modeling with applications to survival data. *Biometrics*, 57: 829–836, 2001.

Bhattacharya, A. and Dunson, D. B. Sparse bayesian

infinite factor models. *Biometrika*, 98(2):291–306, 2011.

Candes, E. J. and Recht, B. Exact matrix completion via convex optimization. *Found. of Comput. Math.*, 9:717–772, 2008.

Eades, P. A heuristic for graph drawing. *Congressus Numerantium*, 42:149–160, 1984.

Friedman, J., Hastie, T., and Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9:432–441, 2008.

Fruchterman, T. and Reingold, E. Graph drawing by force-directed placement. *Software: Practice and Experience*, 21(11):1129–1164, 1991.

Fyshe, A., Fox, E.B., Dunson, D.B., and Mitchell, T.M. Hierarchical latent dictionaries for models of brain activation. In *AISTATS*, 2012.

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

Hahn, P. R., Carvalho, C. M., and Scott, J. G. A sparse factor-analytic probit model for congressional voting patterns. *J. Royal Stat. Soc., C*, 61(4):619–635, 2012.

Hariri, A.R. The neurobiology of individual differences in complex behavioral traits. *Annual Review of Neurocience*, 32:225–247, 2009.

Hoff, P.D. Simulation of the matrix Bingham-von Mises-Fisher distribution, with applications to multivariate and relational data. *J. Comp. Graphical Stat.*, 18(2):438–456, 2009.

Kamada, T. and Kawai, S. An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31:7–15, 1989.

Kriegel, H.-P., Kroger, P., and Zimek, A. Clustering high dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Trans. Knowledge Discovery from Data*, 1:1–58, 2009.

Manuck, S.B., Brown, S.M., Forbes, E.E., and Hariri, A.R. Temporal stability of individual differences in amygdala reactivity. *Am. J. Psychiatry*, 164: 16131614, 2007.

Meeds, E., Ghahramani, Z., Neal, R., and Roweis, S. Modeling dyadic data with binary latent factors. In *NIPS*, 2007.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.M., Malave, V. L., Mason, R. A., and Just, M. A. Predicting human brain activity associated with the meanings of nouns. *Science*, 2008.

Salazar, E., Cain, M. S., Darling, E. F., Mitroff, S. R., and Carin, L. Inferring latent structure from mixed real and categorical relational data. In *ICML*, 2012.

Williamson, S., Wang, C., Heller, K. A., and Blei, D. M. The IBP compound Dirichlet process and its application to focused topic modeling. In *ICML*, 2010.

Xu, Z., Yan, F., and Qi, Y. Infinite tucker decomposition: nonparametric Bayesian models for multiway data analysis. In *ICML*, 2012.

Yoshida, R. and West, M. Bayesian learning in sparse graphical factor models via variational mean-field annealing. *JMLR*, 11:1771–1798, 2010.

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