Journal of Machine Learning Research 15 (2014) 1215-1247 Submitted 1/13; Revised 12/13; Published 4/14

Bayesian Nonparametric Comorbidity Analysis of Psychiatric Disorders

Francisco J. R. Ruiz[*] FRANRRUIZ@TSC.UC3M.ES
Isabel Valera[*] IVALERA@TSC.UC3M.ES
Department of Signal Processing and Communications
University Carlos III in Madrid
Avda. de la Universidad, 30
28911 Leganés (Madrid, Spain)

Carlos Blanco CBLANCO@NYSPI.COLUMBIA.EDU
Department of Psychiatry, New York State Psychiatric Institute
Columbia University
1051 Riverside Drive, Unit #69
New York, NY 10032 (United States of America)

Fernando Perez-Cruz FERNANDO@TSC.UC3M.ES
Department of Signal Processing and Communications
University Carlos III in Madrid
Avda. de la Universidad, 30
28911 Leganés (Madrid, Spain)

Editor: Athanasios Kottas

Abstract

The analysis of comorbidity is an open and complex research field in the branch of psychiatry, where clinical experience and several studies suggest that the relation among the psychiatric disorders may have etiological and treatment implications. In this paper, we are interested in applying latent feature modeling to find the latent structure behind the psychiatric disorders that can help to examine and explain the relationships among them. To this end, we use the large amount of information collected in the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC) database and propose to model these data using a nonparametric latent model based on the Indian Buffet Process (IBP). Due to the discrete nature of the data, we first need to adapt the observation model for discrete random variables. We propose a generative model in which the observations are drawn from a multinomial-logit distribution given the IBP matrix. The implementation of an efficient Gibbs sampler is accomplished using the Laplace approximation, which allows integrating out the weighting factors of the multinomial-logit likelihood model. We also provide a variational inference algorithm for this model, which provides a complementary (and less expensive in terms of computational complexity) alternative to the Gibbs sampler allowing us to deal with a larger number of data. Finally, we use the model to analyze comorbidity among the psychiatric disorders diagnosed by experts from the NESARC database.

Keywords: Bayesian nonparametrics, Indian buffet process, categorical observations, multinomial-logit function, Laplace approximation, variational inference

[*]. Both authors contributed equally.

1. Introduction

Health care increasingly needs to address the management of individuals with multiple coexisting diseases, who are now the norm, rather than the exception. In the United States, about 80% of Medicare spending is devoted to patients with four or more chronic conditions, with costs growing as the number of chronic conditions increases (Wolff et al., 2002). This explains the growing interest of researchers in the impact of comorbidity on a range of outcomes, such as mortality, health-related quality of life, functioning, and quality of health care. However, attempts to study the impact of comorbidity are complicated by the lack of consensus about how to define and measure it (Valderas et al., 2009).

Comorbidity becomes particularly relevant in psychiatry, where clinical experience and several studies suggest that the relation among the psychiatric disorders may have etiological and treatment implications. Several studies have focused on the search of the underlying interrelationships among psychiatric disorders, which can be useful to analyze the structure of the diagnostic classification system, and guide treatment approaches for each disorder (Blanco et al., 2013). Krueger (1999) found that 10 psychiatric disorders (available in the National Comorbidity Survey) can be explained by only two correlated factors, one corresponding to internalizing disorders and the other to externalizing disorders. The existence of the internalizing and the externalizing factors was also confirmed by Kotov et al. (2011). More recently, Blanco et al. (2013) have used factor analysis to find the latent feature structure under 20 common psychiatric disorders, drawing on data from the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC). In particular, the authors found that three correlated factors, one related to externalizing, and the other two to internalizing disorders, characterized well the underlying structure of these 20 diagnoses. From a statistical point of view, the main limitation of this study lies on the use of factor analysis, which assumes that the number of factors is known and that the observations are Gaussian distributed. However, the latter assumption does not fit the observed data, since they are discrete in nature.

In order to avoid the model selection step needed to infer the number of factors in factor analysis, we can resort to Bayesian nonparametric tools, which allow an open-ended number of degrees of freedom in a model (Jordan, 2010). In this paper, we apply the Indian Buffet Process (IBP) (Griffiths and Ghahramani, 2011), because it allows us to infer which latent features influence the observations and how many features there are. We adapt the observation model for discrete random variables, as the discrete nature of the data does not allow using the standard Gaussian observation model. There are several options for modeling discrete outputs given the hidden latent features, like a Dirichlet distribution or sampling from the features, but we opted for the generative model partially introduced by Ruiz et al. (2012), in which the observations are drawn from a multinomial-logit distribution, because it resembles the standard Gaussian observation model, as the observation probability distribution depends on the IBP matrix weighted by some factors.

The IBP model combined with discrete observations has already been tackled in several related works. Williamson et al. (2010) propose a model that combines properties from both the hierarchical Dirichlet process (HDP) and the IBP, called IBP compound Dirichlet (ICD) process. They apply the ICD to focused topic modeling, where the instances are documents and the observations are words from a finite vocabulary, and focus on decoupling the prevalence of a topic in a document and its prevalence in all documents. Despite the discrete nature of the observations under this model, these assumptions are not appropriate for observations such as the set of possible diagnoses or responses to the questions from the NESARC database, since categorical observations can only take values from a finite set where elements do not present any particular ordering. Titsias (2007) introduced the infinite gamma-Poisson process as a prior probability distribution over non-negative integer valued matrices with a potentially infinite number of columns, and he applied it to topic modeling of images. In this model, each (discrete) component in the observation vector of an instance depends only on one of the active latent features of that object, randomly drawn from a multinomial distribution. Therefore, different components of the observation vector might be equally distributed. Our model is more flexible in the sense that it allows different probability distributions for every component in the observation vector, which is accomplished by weighting differently the latent variables. Furthermore, a preliminary version of this model has been successfully applied to identify the factors that model the risk of suicide attempts (Ruiz et al., 2012).

The rest of the paper is organized as follows. In Section 2, we review the IBP model and the basic Gibbs sampling inference for the IBP, and set the notation used throughout the paper. In Section 3, we propose the generative model which combines the IBP with discrete observations generated from a multinomial-logit distribution. In this section, we focus on the inference based on the Gibbs sampler, where we make use of the Laplace approximation to integrate out the random weighting factors in the observation model. In Section 4, we develop a variational inference algorithm that presents lower computational complexity than the Gibbs sampler. In Section 5, we validate our model on synthetic data and apply it over the real data extracted from the NESARC database. Finally, Section 6 is devoted to the conclusions.

2. The Indian Buffet Process

Unsupervised learning aims to recover the latent structure responsible for generating the observed properties of a set of objects. In latent feature modeling, the properties of each object can be represented by an unobservable vector of latent features, and the observations are generated from a distribution determined by those latent feature values. Typically, we have access to the set of observations and the main goal of latent feature modeling is to find out the latent variables that represent the data.

The most common nonparametric tool for latent feature modeling is the Indian Buffet Process (IBP). The IBP places a prior distribution over binary matrices, in which the number of rows is finite but the number of columns (features) K is potentially unbounded, that is, $K \to \infty$. This distribution is invariant to the ordering of the features and can be derived by taking the limit of a properly defined distribution over $N \times K$ binary matrices as K tends to infinity (Griffiths and Ghahramani, 2011), similarly to the derivation of the Chinese restaurant process as the limit of a Dirichlet-multinomial model (Aldous, 1985). However, given a finite number of data points N, it ensures that the number of non-zero columns, namely, K_+, is finite with probability one.

Let $\mathbf{Z}$ be a random $N \times K$ binary matrix distributed following an IBP, i.e., $\mathbf{Z} \sim \mathrm{IBP}(\alpha)$, where α is the concentration parameter of the process, which controls the number of non-zero columns K_+. The n^{th} row of $\mathbf{Z}$, denoted by $\mathbf{z}_{n\bullet}$, represents the vector of latent features of the n^{th} data point, and every entry nk is denoted by z_{nk}. Note that each element $z_{nk} \in \{0, 1\}$ indicates whether the k^{th} feature contributes to the n^{th} data point. Since only the K_+ non-zero columns of $\mathbf{Z}$ contain the features of interest, and due to the exchangeability property of the features under the IBP prior, they are usually grouped in the left hand side of the

$$\mathbf{Z} = \left[\underbrace{\underbrace{\begin{matrix} z_{11} & z_{12} & \cdots & z_{1K_+} \\ z_{21} & z_{22} & \cdots & z_{2K_+} \\ \vdots & \vdots & \ddots & \vdots \\ z_{N1} & z_{N2} & \cdots & z_{NK_+} \end{matrix}}_{K_+ \text{ non-zero columns}} \begin{matrix} 0 & 0 & \cdots \\ 0 & 0 & \cdots \\ \vdots & \vdots & \ddots \\ 0 & 0 & \cdots \end{matrix}}_{K \text{ columns (features)}} \right] N \text{ data points}$$

Figure 1: Illustration of an IBP matrix.

2.1 The Stick-Breaking Construction

The stick-breaking construction of the IBP is an equivalent representation of the IBP prior, useful for inference algorithms other than Gibbs sampling, such as slice sampling or variational inference algorithms (Teh et al., 2007; Doshi-Velez et al., 2009).

In this representation, the probability of each latent feature being active is represented explicitly by a random variable. In particular, the probability of feature z_{nk} taking value 1 is denoted by ω_k, that is,

$$z_{nk} \sim \text{Bernouilli}(\omega_k).$$

Since this probability does not depend on n, the stick-breaking representation explicitly shows that the ordering of the data does not affect the distribution.

The probabilities ω_k are, in turn, generated by first drawing a sequence of independent random variables $v_1, v_2, \ldots$ from a beta distribution of the form

$$v_k \sim \text{Beta}(\alpha, 1).$$

Given the sequence of variables $v_1, v_2, \ldots$, the probability ω_1 is assigned to v_1, and each subsequent ω_k is obtained as

$$\omega_k = \prod_{i=1}^{k} v_i,$$

resulting in a decreasing sequence of probabilities ω_k. Specifically, the expected probability of feature z_{nk} being active decreases exponentially with the index k.

This construction can be understood with the stick-breaking process illustrated in Figure 2. Starting with a stick of length 1, at each iteration $k = 1, 2, \ldots$, a piece is broken off at a point v_k relative to the current length of the stick. The variable ω_k corresponds to the length of the stick just broken off, and the other piece of the stick is discarded.

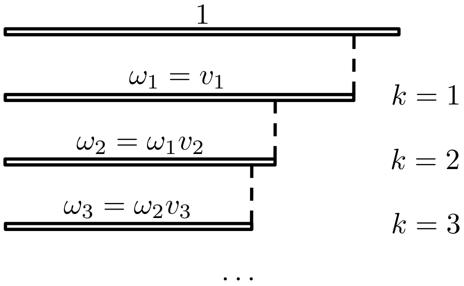


Figure 2: Illustration of the stick-breaking construction of the IBP.

2.2 Inference

Markov Chain Monte Carlo (MCMC) methods have been broadly applied to infer the latent structure $\mathbf{Z}$ from a given observation matrix $\mathbf{X}$ (see, e.g., in Griffiths and Ghahramani (2011); Williamson et al. (2010); Van Gael et al. (2009); Titsias (2007)), being Gibbs sampling the standard method of choice. This algorithm iteratively samples the value of each element z_{nk} given the remaining variables, that is, it samples from

$$p(z_{nk} = 1|\mathbf{X}, \mathbf{Z}_{\neg nk}) \propto p(\mathbf{X}|\mathbf{Z}) p(z_{nk} = 1|\mathbf{Z}_{\neg nk}), \tag{1}$$

where $\mathbf{Z}_{\neg nk}$ denotes all the entries of $\mathbf{Z}$ other than z_{nk}. The conditional distribution $p(z_{nk} = 1|\mathbf{Z}_{\neg nk})$ can be readily derived from the exchangeable IBP and can be written as

$$p(z_{nk} = 1|\mathbf{Z}_{\neg nk}) = \frac{m_{-n,k}}{N},$$

where $m_{-n,k}$ is the number of data points with feature k, not including n, i.e., $m_{-n,k} = \sum_{i \neq n} z_{ik}$. For each data point n, after having sampled all elements z_{nk} for the K_+ non-zero columns in $\mathbf{Z}$, the algorithm samples from a distribution (where the prior is a Poisson distribution with mean α/N) a number of new features necessary to explain that data point.

Although MCMC methods perform exact inference, they typically suffer from high computational complexity. To solve this limitation, variational inference algorithms can be applied instead at a lower computational cost, at the expense of performing approximate inference (Jordan et al., 1999). A variational inference algorithm for the IBP under the standard Gaussian observation model is presented by Doshi-Velez et al. (2009). This algorithm makes use of the stick breaking construction of the IBP, summarized above.

3. Observation Model

Unlike the standard Gaussian observation model, let us consider discrete observations, that is, each element $x_{nd} \in \{1, \dots, R_d\}$, where this finite set contains the indexes to all the possible values of x_{nd}. For simplicity and without loss of generality, we consider that $R_d = R$, but the following results can be readily extended to a different cardinality per input dimension, as well as mixing continuous variables with discrete variables, since given the latent feature matrix $\mathbf{Z}$ the columns of $\mathbf{X}$ are assumed to be independent.

We introduce the $K \times R$ matrices $\mathbf{B}^d$ and the length-R row vectors $\mathbf{b}_0^d$ to model the probability distribution over $\mathbf{X}$, such that $\mathbf{B}^d$ links the latent features with the d^{th} column of the observation matrix $\mathbf{X}$, denoted by $\mathbf{x}_{\bullet d}$, and $\mathbf{b}_0^d$ is included to model the bias term in the distribution over the data points. This bias term plays the role of a latent variable that is always active. For a categorical observation space, if we do not have a bias term and all latent variables are inactive, the model assumes that all the outcomes are independent and equally likely, which is not a suitable assumption in most cases. In our application, the bias term is used to model the people that do not suffer from any disorder and it captures the baseline diagnosis in the general population. Additionally, this bias term simplifies the inference since the latent features of those subjects that are not diagnosed any disorder do not need to be sampled.

Hence, we assume that the probability of each element x_{nd} taking value r $(r = 1, \dots, R)$, denoted by π_{nd}^r, is given by the multiple-logistic function, i.e.,

$$\pi_{nd}^r = p(x_{nd} = r | \mathbf{z}_{n\bullet}, \mathbf{B}^d, \mathbf{b}_0^d) = \frac{\exp\left(\mathbf{z}_{n\bullet}\mathbf{b}_{\bullet r}^d + b_{0r}^d\right)}{\sum\limits_{r'=1}^{R} \exp\left(\mathbf{z}_{n\bullet}\mathbf{b}_{\bullet r'}^d + b_{0r'}^d\right)}, \tag{2}$$

where $\mathbf{b}_{\bullet r}^d$ denotes the r^{th} column of $\mathbf{B}^d$ and b_{0r}^d denotes the r^{th} element of vector $\mathbf{b}_0^d$. Note that the matrices $\mathbf{B}^d$ are used to weight differently the contribution of each latent feature to every component d

and Ghahramani (20

with zero mean and

distributed with zero

in Figure 3.

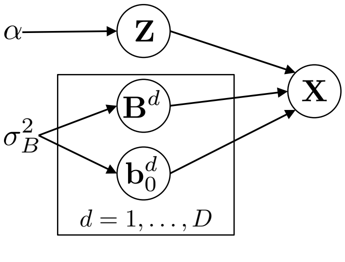


Figure 3: Graphical probabilistic model of the IBP with discrete observations.

The choice of the observation model in Eq. (2), which combines the multiple-logistic function with Gaussian parameters, is based on the fact that it induces dependencies among the probabilities π_{nd}^r that cannot be captured with other distributions, such as the Dirichlet distribution (Blei and Lafferty, 2007). Furthermore, this multinomial-logistic normal distribution has been widely used to define probability distributions over discrete random variables (Williams and Barber, 1998; Blei and Lafferty, 2007).

We consider that elements x_{nd} are independent given the latent feature matrix $\mathbf{Z}$, the weighting matrices $\mathbf{B}^d$ and the weighting vectors $\mathbf{b}_0^d$. Then, the likelihood for any matrix $\mathbf{X}$ can be expressed as

$$p(\mathbf{X}|\mathbf{Z}, \mathbf{B}^1, \ldots, \mathbf{B}^D, \mathbf{b}_0^1, \ldots, \mathbf{b}_0^D) = \prod_{n=1}^{N}\prod_{d=1}^{D} p(x_{nd}|\mathbf{z}_{n\bullet}, \mathbf{B}^d, \mathbf{b}_0^d) = \prod_{n=1}^{N}\prod_{d=1}^{D} \pi_{nd}^{x_{nd}}. \tag{3}$$

3.1 Laplace Approximation for Gibbs Sampling Inference

In Section 2, the (heuristic) Gibbs sampling algorithm for posterior inference over the latent variables of the IBP, detailed in Griffiths and Ghahramani (2011), has been briefly reviewed. To sample from Eq. (1), we need to integrate out $\mathbf{B}^d$ and $\mathbf{b}_0^d$ in (3), as sequentially sampling from the posterior distribution of these variables is intractable, for which an approximation is required. We rely on the Laplace approximation to integrate out the parameters $\mathbf{B}^d$ and $\mathbf{b}_0^d$ for simplicity and ease of implementation. We first consider the finite form of the proposed model, where K is bounded.

We can simplify the notation in Eqs. 2 and 3 by considering an extended latent feature matrix $\mathbf{Z}$ of size $N \times (K+1)$, in which the elements of the first column are equal to one, and D extended weighting matrices $\mathbf{B}^d$ of size $(K+1) \times R$, in which the first row equals the vector $\mathbf{b}_0^d$. With these definitions, Eq. (2) can be rewritten as

$$\pi_{nd}^r = p(x_{nd} = r|\mathbf{z}_{n\bullet}, \mathbf{B}^d) = \frac{\exp\left(\mathbf{z}_{n\bullet}\mathbf{b}_{\bullet r}^d\right)}{\sum_{r'=1}^{R} \exp\left(\mathbf{z}_{n\bullet}\mathbf{b}_{\bullet r'}^d\right)}.$$

Unless otherwise specified, we use the simplified notation throughout this section. For this reason, the index k over the latent variables takes the values in $\{0, 1, \ldots, K\}$, with $z_{n0} = 1$ for all n.

Recall that our model assumes independence among the observations given the hidden latent variables. Then, the posterior $p(\mathbf{B}^1, \ldots, \mathbf{B}^D|\mathbf{X}, \mathbf{Z})$ factorizes as

$$p(\mathbf{B}^1, \ldots, \mathbf{B}^D|\mathbf{X}, \mathbf{Z}) = \prod_{d=1}^{D} p(\mathbf{B}^d|\mathbf{x}_{\bullet d}, \mathbf{Z}) = \prod_{d=1}^{D} \frac{p(\mathbf{x}_{\bullet d}|\mathbf{B}^d, \mathbf{Z})p(\mathbf{B}^d)}{p(\mathbf{x}_{\bullet d}|\mathbf{Z})}.$$

Hence, we only need to deal with each term $p(\mathbf{B}^d|\mathbf{x}_{\bullet d}, \mathbf{Z})$ individually. The marginal likelihood $p(\mathbf{x}_{\bullet d}|\mathbf{Z})$, which we are interested in, can be obtained as

$$p(\mathbf{x}_{\bullet d}|\mathbf{Z}) = \int p(\mathbf{x}_{\bullet d}|\mathbf{B}^d, \mathbf{Z})p(\mathbf{B}^d)d\mathbf{B}^d. \tag{4}$$

Although the prior $p(\mathbf{B}^d)$ is Gaussian, due to the non-conjugacy with the likelihood term, the computation of this integral, as well as the computation of the posterior $p(\mathbf{B}^d|\mathbf{x}_{\bullet d}, \mathbf{Z})$, turns out to be intractable.

Following a similar procedure as in Gaussian processes for multiclass classification (Williams and Barber, 1998), we approximate the posterior $p(\mathbf{B}^d|\mathbf{x}_{\bullet d}, \mathbf{Z})$ as a Gaussian distribution using Laplace's method. In order to obtain the parameters of the Gaussian distribution, we define $f(\mathbf{B}^d)$ as the un-normalized log-posterior of $p(\mathbf{B}^d|\mathbf{x}_{\bullet d}, \mathbf{Z})$, i.e.,

$$f(\mathbf{B}^d) = \log p(\mathbf{x}_{\bullet d}|\mathbf{B}^d, \mathbf{Z}) + \log p(\mathbf{B}^d). \tag{5}$$

As proven in Appendix A, the function $f(\mathbf{B}^d)$ is a strictly concave function of $\mathbf{B}^d$ and therefore it has a unique maximum, which is reached at $\mathbf{B}^d_{\text{MAP}}$, denoted by the subscript 'MAP' (*maximum a posteriori*) because it coincides with the mean of the Gaussian distribution in the Laplace's approximation. We resort to Newton's method to compute $\mathbf{B}^d_{\text{MAP}}$.

We stack the columns of $\mathbf{B}^d$ into $\boldsymbol{\beta}^d$, i.e., $\boldsymbol{\beta}^d = \mathbf{B}^d(:)$ for avid Matlab users. The posterior $p(\mathbf{B}^d|\mathbf{x}_{\bullet d}, \mathbf{Z})$ can be approximated as

$$p(\boldsymbol{\beta}^d|\mathbf{x}_{\bullet d}, \mathbf{Z}) \approx \mathcal{N}\left(\boldsymbol{\beta}^d \middle| \boldsymbol{\beta}^d_{\text{MAP}}, (-\nabla\nabla f)|_{\boldsymbol{\beta}^d_{\text{MAP}}}\right),$$

where $\boldsymbol{\beta}^d_{\text{MAP}}$ contains all the columns of $\mathbf{B}^d_{\text{MAP}}$ stacked into a vector and $\nabla\nabla f$ is the Hessian of $f(\boldsymbol{\beta}^d)$. Hence, by taking the second-order Taylor series expansion of $f(\boldsymbol{\beta}^d)$ around its maximum, the computation of the marginal likelihood in (4) results in a Gaussian integral, whose solution can be expressed as

$$\begin{aligned}\log p(\mathbf{x}_{\bullet d}|\mathbf{Z}) &\approx -\frac{1}{2\sigma_B^2}\operatorname{trace}\left\{(\mathbf{B}^d_{\text{MAP}})^\top \mathbf{B}^d_{\text{MAP}}\right\} \\ &- \frac{1}{2}\log\left|\mathbf{I}_{R(K+1)} + \sigma_B^2 \sum_{n=1}^{N}\left(\operatorname{diag}(\widehat{\boldsymbol{\pi}}_{nd}) - (\widehat{\boldsymbol{\pi}}_{nd})^\top \widehat{\boldsymbol{\pi}}_{nd}\right) \otimes (\mathbf{z}_{n\bullet}^\top \mathbf{z}_{n\bullet})\right| + \log p(\mathbf{x}_{\bullet d}|\mathbf{B}^d_{\text{MAP}}, \mathbf{Z}),\end{aligned} \tag{6}$$

where $\widehat{\boldsymbol{\pi}}_{nd}$ is the vector $\boldsymbol{\pi}_{nd} = \left[\pi^1_{nd}, \pi^2_{nd}, \ldots, \pi^R_{nd}\right]$ evaluated at $\mathbf{B}^d = \mathbf{B}^d_{\text{MAP}}$, and $\operatorname{diag}(\widehat{\boldsymbol{\pi}}_{nd})$ is a diagonal matrix with the values of $\widehat{\boldsymbol{\pi}}_{nd}$ as its diagonal elements.

Similarly as in Griffiths and Ghahramani (2011), it is straightforward to prove that the limit of Eq. (6) is well-defined if $\mathbf{Z}$ has an unbounded number of columns, that is, as $K \to \infty$. The resulting expression for the marginal likelihood $p(\mathbf{x}_{\bullet d}|\mathbf{Z})$ can be readily obtained from Eq. (6) by replacing K by K_+, $\mathbf{Z}$ by the submatrix containing only the non-zero columns of $\mathbf{Z}$, and $\mathbf{B}^d_{\text{MAP}}$ by the submatrix containing the $K_+ + 1$ corresponding rows.

3.2 Speeding Up the Matrix Inversion

In this section, we propose a method that reduces the complexity of computing the inverse of the Hessian for Newton's method (as well as its determinant) from $\mathcal{O}(R^3K_+^3 + NR^2K_+^2)$ to $\mathcal{O}(RK_+^3 + NR^2K_+^2)$, effectively accelerating the inference procedure for large values of R.

Let us denote with $\mathbf{Z}$ the matrix that contains only the $K_+ + 1$ non-zero columns of the extended full IBP matrix. The inverse of the Hessian for Newton's method, as well as its determinant in (6), can be efficiently carried out if we rearrange the inverse of $\nabla\nabla f$ as follows:

$$(-\nabla\nabla f)^{-1} = \left(\mathbf{D} - \sum_{n=1}^{N}\mathbf{v}_n\mathbf{v}_n^\top\right)^{-1},$$

where $\mathbf{v}_n = (\boldsymbol{\pi}_{nd})^\top \otimes \mathbf{z}_{n\bullet}^\top$ and $\mathbf{D}$ is a block-diagonal matrix, in which each diagonal submatrix is given by

$$\mathbf{D}_r = \frac{1}{\sigma_B^2}\mathbf{I}_{K_+ + 1} + \mathbf{Z}^\top \operatorname{diag}\left(\boldsymbol{\pi}_{\bullet d}^r\right)\mathbf{Z}, \tag{7}$$

with $\boldsymbol{\pi}_{\bullet d}^r = \left[\begin{array}{c} \pi_{1d}^r, \ldots, \pi_{Nd}^r \end{array}\right]^\top$. Since $\mathbf{v}_n\mathbf{v}_n^\top$ is a rank-one matrix, we can apply the Woodbury identity (Woodbury, 1949) N times to invert the matrix $-\nabla\nabla f$, similar to the RLS (Recursive Least Squares) updates (Haykin, 2002). At each iteration $n = 1, \ldots, N$, we compute

$$(\mathbf{D}^{(n)})^{-1} = \left(\mathbf{D}^{(n-1)} - \mathbf{v}_n\mathbf{v}_n^\top\right)^{-1} = (\mathbf{D}^{(n-1)})^{-1} + \frac{(\mathbf{D}^{(n-1)})^{-1}\mathbf{v}_n\mathbf{v}_n^\top(\mathbf{D}^{(n-1)})^{-1}}{1 - \mathbf{v}_n^\top(\mathbf{D}^{(n-1)})^{-1}\mathbf{v}_n}. \tag{8}$$

For the first iteration, we define $\mathbf{D}^{(0)}$ as the block-diagonal matrix $\mathbf{D}$, whose inverse matrix involves computing the R matrix inversions of size $(K_+ + 1) \times (K_+ + 1)$ of the matrices in (7), which can be efficiently solved applying the Matrix Inversion Lemma. After N iterations of (8), it turns out that $(-\nabla\nabla f)^{-1} = (\mathbf{D}^{(N)})^{-1}$.

In practice, there is no need to iterate over all observations, since all subjects sharing the same latent feature vector $\mathbf{z}_{n\bullet}$ and observation x_{nd} can be grouped together, therefore requiring (at most) $R2^{K_+}$ iterations instead of N. In our applications, it provides significant savings in run-time complexity, since $R2^{K_+} \ll N$.

For the determinant in (6), similar recursions can be applied using the Matrix Determinant Lemma (Harville, 1997), which states that $|\mathbf{D} + \mathbf{v}\mathbf{u}^\top| = (1 + \mathbf{v}^\top\mathbf{D}\mathbf{u})|\mathbf{D}|$, and $|\mathbf{D}^{(0)}| = \prod_{r=1}^R |\mathbf{D}_r|$.

4. Variational Inference

Variational inference provides a complementary (and less expensive in terms of computational complexity) alternative to MCMC methods as a general source of approximation methods for inference in large-scale statistical models (Jordan et al., 1999). In this section, we adapt the infinite variational approach for the linear-Gaussian model with respect to a full IBP prior introduced by Doshi-Velez et al. (2009) to the model proposed in Section 3. This approach assumes the (truncated) stick-breaking construction for the IBP in Section 2.1, which bounds the number of columns of the IBP matrix by a finite (but large enough) value, K. Then, in the truncated stick-breaking process, $\omega_k = \prod_{i=1}^k v_i$ for $k \leq K$ and zero otherwise.

The hyperparameters of the model are contained in the set $\mathcal{H} = \{\alpha, \sigma_B^2\}$ and, similarly, $\Psi = \{\mathbf{Z}, \mathbf{B}^1, \ldots, \mathbf{B}^D, \mathbf{b}_0^1, \ldots, \mathbf{b}_0^D, v_1, \ldots, v_K\}$ denotes the set of unobserved variables in the model. Under the truncated stick-breaking construction for the IBP, the joint probability distribution over all the variables $p(\Psi, \mathbf{X}|\mathcal{H})$ can be factorized as

$$\begin{aligned} p(\Psi, \mathbf{X}|\mathcal{H}) = &\prod_{k=1}^K \left(p(v_k|\alpha)\prod_{n=1}^N p(z_{nk}|\{v_i\}_{i=1}^k)\right)\prod_{d=1}^D\left(p(\mathbf{b}_0^d|\sigma_B^2)\prod_{k=1}^K p(\mathbf{b}_{k\bullet}^d|\sigma_B^2)\right) \\ &\times \prod_{n=1}^N\prod_{d=1}^D p(x_{nd}|\mathbf{z}_{n\bullet}, \mathbf{B}^d, \mathbf{b}_0^d), \end{aligned}$$

where $\mathbf{b}^d_{k\bullet}$ is the k^{th} row of matrix $\mathbf{B}^d$.

We approximate $p(\Psi|\mathbf{X},\mathcal{H})$ with the variational distribution $q(\Psi)$ given by

$$q(\Psi)=\prod_{k=1}^{K}\left(q(v_k|\tau_{k1},\tau_{k2})\prod_{n=1}^{N}q(z_{nk}|\nu_{nk})\right)\prod_{k=0}^{K}\prod_{r=1}^{R}\prod_{d=1}^{D}q(b^d_{kr}|\phi^d_{kr},(\sigma^d_{kr})^2),$$

where the elements of matrix $\mathbf{B}^d$ are denoted by b^d_{kr}, and

$$\begin{aligned}&q(v_k|\tau_{k1},\tau_{k2})=\mathrm{Beta}(\tau_{k1},\tau_{k2}),\\&q(b^d_{kr}|\phi^d_{kr},(\sigma^d_{kr})^2)=\mathcal{N}(\phi^d_{kr},(\sigma^d_{kr})^2),\\&q(z_{nk}|\nu_{nk})=\mathrm{Bernoulli}(\nu_{nk}).\end{aligned}$$

Inference involves optimizing the variational parameters of $q(\Psi)$ to minimize the Kullback-Leibler divergence from $q(\Psi)$ to $p(\Psi|\mathbf{X},\mathcal{H})$, i.e., $D_{KL}(q||p)$. This optimization is equivalent to maximizing a lower bound on the evidence $p(\mathbf{X}|\mathcal{H})$, since

$$\begin{aligned}\log p(\mathbf{X}|\mathcal{H})&=\mathbb{E}_q\left[\log p(\Psi,\mathbf{X}|\mathcal{H})\right]+H[q]+D_{KL}(q||p)\\&\geqslant\mathbb{E}_q\left[\log p(\Psi,\mathbf{X}|\mathcal{H})\right]+H[q],\end{aligned}\tag{9}$$

where $\mathbb{E}_q[\cdot]$ denotes the expectation with respect to the distribution $q(\Psi)$, $H[q]$ is the entropy of distribution $q(\Psi)$ and

$$\begin{aligned}\mathbb{E}_q\left[\log p(\Psi,\mathbf{X}|\mathcal{H})\right]=&\sum_{k=1}^{K}\mathbb{E}_q\left[\log p(v_k|\alpha)\right]+\sum_{d=1}^{D}\sum_{k=1}^{K}\mathbb{E}_q\left[\log p(\mathbf{b}^d_{k\bullet}|\sigma^2_B)\right]+\sum_{d=1}^{D}\mathbb{E}_q\left[\log p(\mathbf{b}^d_0|\sigma^2_B)\right]\\&+\sum_{k=1}^{K}\sum_{n=1}^{N}\mathbb{E}_q\left[\log p(z_{nk}|\{v_i\}_{i=1}^k)\right]+\sum_{n=1}^{N}\sum_{d=1}^{D}\mathbb{E}_q\left[\log p(x_{nd}|\mathbf{z}_{n\bullet},\mathbf{B}^d,\mathbf{b}^d_0)\right].\end{aligned}\tag{10}$$

The derivation of the lower bound in (9) is straightforward, with the exception of the terms $\mathbb{E}_q\left[\log p(z_{nk}|\{v_i\}_{i=1}^k)\right]$ and $\mathbb{E}_q\left[\log p(x_{nd}|\mathbf{z}_{n\bullet},\mathbf{B}^d,\mathbf{b}^d_0)\right]$ in (10), which have no closed-form solution, so we instead need to bound them. Deriving these bounds leads to a new bound $\mathcal{L}(\mathcal{H},\mathcal{H}_q)$, which can be obtained in closed-form, such that $\log p(\mathbf{X}|\mathcal{H})\geq\mathcal{L}(\mathcal{H},\mathcal{H}_q)$, being $\mathcal{H}_q$ the full set of variational parameters. The final expression for $\mathcal{L}(\mathcal{H},\mathcal{H}_q)$, as well as the details on the derivation of the bound, are provided in Appendix B.

In order to maximize the lower bound $\mathcal{L}(\mathcal{H},\mathcal{H}_q)$, we need to optimize with respect to the value of the variational parameters. To this end, we can iteratively maximize the bound with respect to each variational parameter by taking the derivative of $\mathcal{L}(\mathcal{H},\mathcal{H}_q)$ and setting it to zero. This procedure readily leads to the following fixed-point equations:

1. For the variational Beta distribution $q(v_k|\tau_{k1},\tau_{k2})$,

$$\begin{aligned}\tau_{k1}&=\alpha+\sum_{m=k}^{K}\left(\sum_{n=1}^{N}\nu_{nm}\right)+\sum_{m=k+1}^{K}\left(N-\sum_{n=1}^{N}\nu_{nm}\right)\left(\sum_{i=k+1}^{m}\lambda_{mi}\right),\\\tau_{k2}&=1+\sum_{m=k}^{K}\left(N-\sum_{n=1}^{N}\nu_{nm}\right)\lambda_{mk}.\end{aligned}$$

2. For the Bernoulli distribution $q(z_{nk}|\nu_{nk})$,

$$\nu_{nk} = \frac{1}{1+\exp(-A_{nk})},$$

where

$$\begin{aligned} A_{nk} = &\sum_{i=1}^{k} [\psi(\tau_{i1}) - \psi(\tau_{i1}+\tau_{i2})] - \left[\sum_{m=1}^{k} \lambda_{km}\psi(\tau_{m2}) + \sum_{m=1}^{k-1}\left(\sum_{n=m+1}^{k}\lambda_{kn}\right)\psi(\tau_{m1}) \right.\\ &\left. - \sum_{m=1}^{k}\left(\sum_{n=m}^{k}\lambda_{kn}\right)\psi(\tau_{m1}+\tau_{m2}) - \sum_{m=1}^{k}\lambda_{km}\log(\lambda_{km})\right] \\ &+ \sum_{d=1}^{D}\left(\phi^d_{kx_{nd}} - \xi_{nd}\sum_{r=1}^{R}\left[\exp\left(\phi^d_{0r} + \frac{1}{2}(\sigma^d_{0r})^2\right)\left(1-\exp\left(\phi^d_{kr}+\frac{1}{2}(\sigma^d_{kr})^2\right)\right) \times \right.\right. \\ &\left.\left. \times \prod_{k'\neq k}\left(1-\nu_{nk'}+\nu_{nk'}\exp\left(\phi^d_{k'r}+\frac{1}{2}(\sigma^d_{k'r})^2\right)\right)\right]\right), \end{aligned}$$

and $\psi(\cdot)$ stands for the digamma function (Abramowitz and Stegun, 1972, p. 258–259).

3. For the feature assignments, which are Bernoulli distributed given the feature probabilities, we have lower bounded $\mathbb{E}_q\left[\log p(z_{nk}|\{v_i\}_{i=1}^{k})\right]$ by using the multinomial approach in Doshi-Velez et al. (2009) (see Appendix B for further details). This approximation introduces the auxiliary multinomial distribution $\boldsymbol{\lambda}_k = [\lambda_{k1}, \dots, \lambda_{kk}]$, where each λ_{ki} can be updated as

$$\lambda_{ki} \propto \exp\left(\psi(\tau_{i2}) + \sum_{m=1}^{i-1}\psi(\tau_{m1}) - \sum_{m=1}^{i}\psi(\tau_{m1}+\tau_{m2})\right),$$

where the proportionality ensures that $\boldsymbol{\lambda}_k$ is a valid distribution.

4. The maximization with respect to the variational parameters ϕ^d_{kr}, ϕ^d_{0r}, $(\sigma^d_{kr})^2$, and $(\sigma^d_{0r})^2$ has no analytical solution, and therefore, we need to resort to a numerical method to find the maximum, such as Newton's method or conjugate gradient algorithm, for which the first and the second derivatives[1] (given in Appendix C) are required.

5. Finally, we lower bound the likelihood term $\mathbb{E}_q\left[\log p(x_{nd}|\mathbf{z}_{n\bullet}, \mathbf{B}^d, \mathbf{b}^d_0)\right]$ by resorting to a first-order Taylor series expansion around the auxiliary variables ξ^{-1}_{nd} for $n = 1, \dots, N$ and $d = 1, \dots, D$ (see Appendix B for further details), which are optimized by the expression

$$\xi_{nd} = \left[\sum_{r=1}^{R}\exp\left(\phi^d_{0r}+\frac{1}{2}(\sigma^d_{0r})^2\right)\prod_{k=1}^{K}\left(1-\nu_{nk}+\nu_{nk}\exp\left(\phi^d_{kr}+\frac{1}{2}(\sigma^d_{kr})^2\right)\right)\right]^{-1}.$$

[1] Note that the second derivatives are strictly negative and, therefore, the maximum with respect to each parameter is unique.

5. Experiments

In this section, we first use a toy example to show how our model with discrete observations works and then we turn to two experiments over the NESARC database.

5.1 Inference over Synthetic Images

We generate an illustrative example inspired by the example in Griffiths and Ghahramani (2011) to show that the proposed model works as expected. We define four base black-and-white images, shown in Figure 4a, that can be present with probability 0.3, independently of the others. These base images are combined to create a binary composite image. We also multiply each white pixel independently with equiprobable binary noise, hence each white pixel in the composite image can be turned black 50% of the times, while black pixels always remain black. We generate 200 observations to learn the IBP model (several examples can be found in Figure 4c). The Gibbs sampler has been initialized with $K_+ = 2$, setting each $z_{nk} = 1$ with probability $1/2$, and setting the hyperparameters to $\alpha = 0.5$ and $\sigma_B^2 = 1$.

After 350 iterations, the Gibbs sampler returns four latent features. Each of the four features recovers one of the base images with a different ordering, which is inconsequential. In Figure 4b, we have plotted the posterior probability for each pixel being white, when only one of the components is active. As expected, the black pixels are known to be black (almost zero probability of being white) and the white pixels have about a 50/50 chance of being black or white, due to the multiplicative noise. The Gibbs sampler has used as many as eleven hidden features, as shown in Figure 4e, but after less than 50 iterations, the first four features represent the base images and the others just lock on to a noise pattern, which eventually fades away.

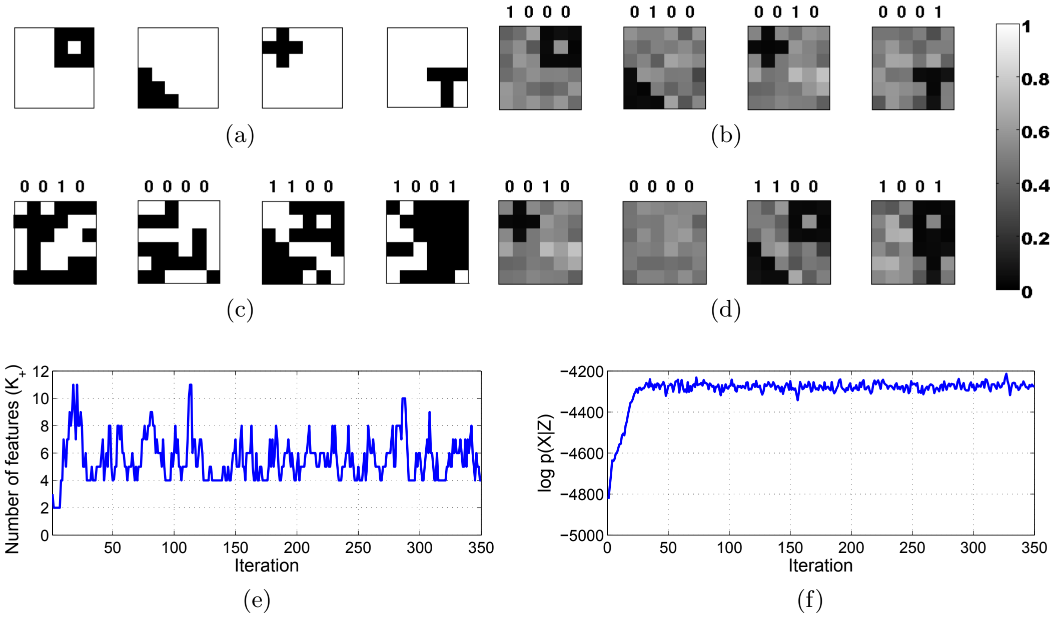
In Figure 4d, we depict the posterior probability of pixels being white for the four images in Figure 4c, given the inferred latent feature vectors for these observations. Note that the model behaves as expected and properly captures the generative process, even for those observations which do not possess any latent features, for which the vectors $\mathbf{b}_0^d$ do not provide significant information about the black-or-white probabilities.

5.2 Comorbidity Analysis of Psychiatric Disorders

In the present study, our objective is to provide an alternative to the factor analysis approach used by Blanco et al. (2013) with the IBP for discrete observations introduced in the present paper. We build an unsupervised model taking the 20 disorders used by Blanco et al. (2013) as input data, drawn from the NESARC data.

The NESARC database was designed to estimate the prevalence of psychiatric disorders, as well as their associated features and level of disability. The NESARC had two waves of interviews (first wave in 2001-2002 and second wave in 2004-2005). For the following experimental results, we only use the data from the first wave, for which 43,093 people were selected to represent the U.S. population of 18 years of age and older. Through 2,991 entries, the NESARC collects data on the background of participants, alcohol and other drug use and use disorders, and other mental disorders. Public use data are currently available for this wave of data collection.[2]

[2] See http://aspe.hhs.gov/hsp/06/catalog-ai-an-na/nesarc.htm



(a) (b)

(c) (d)

(e) (f)

Figure 4: Experimental results of the infinite binary multinomial-logistic model over the image data set. (a) The four base images used to generate the 200 observations. (b) Probability of each pixel being white, when a single feature is active (ordered to match the images on the left), computed using the matrices $\mathbf{B}^d_{\text{MAP}}$. (c) Four data points generated as described in the text. The numbers above each figure indicate which features are present in that image. (d) Probabilities of each pixel being white after 350 iterations of the Gibbs sampler inferred for the four data points on (c). The numbers above each figure show the inferred value of $\mathbf{z}_{n\bullet}$ for these data points. (e) The number of latent features K_+ and (f) the approximate log of $p(\mathbf{X}|\mathbf{Z})$ over the 200 iterations of the Gibbs sampler.

The 20 disorders include substance use disorders (alcohol abuse and dependence, drug abuse and dependence and nicotine dependence), mood disorders (major depressive disorder (MDD), bipolar disorder and dysthymia), anxiety disorders (panic disorder, social anxiety disorder (SAD), specific phobia and generalized anxiety disorder (GAD)), pathological gambling (PG) and seven personality disorders (avoidant, dependent, obsessive-compulsive (OC), paranoid, schizoid, histrionic and antisocial personality disorders (PDs)).

We run the Gibbs sampler over $3,500$ randomly chosen subjects out of the 43,093 participants in the survey, having initialized the sampler with an active feature, i.e., $K_+ = 1$, having set $z_{nk} = 1$ randomly with probability 0.5, and fixing $\alpha = 1$ and $\sigma^2_B = 1$. After convergence, we run an additional Gibbs sampler with 10 iterations for each of the remaining subjects in the database, restricted to their latent features (that is, we fix the latent features learned for the $3,500$ subjects to sample the feature vector of each subject). Then, we run additional iterations of the Gibbs sampler over the whole database, finally obtaining three latent features. In order to speed up the sampling procedure, we do not sample the

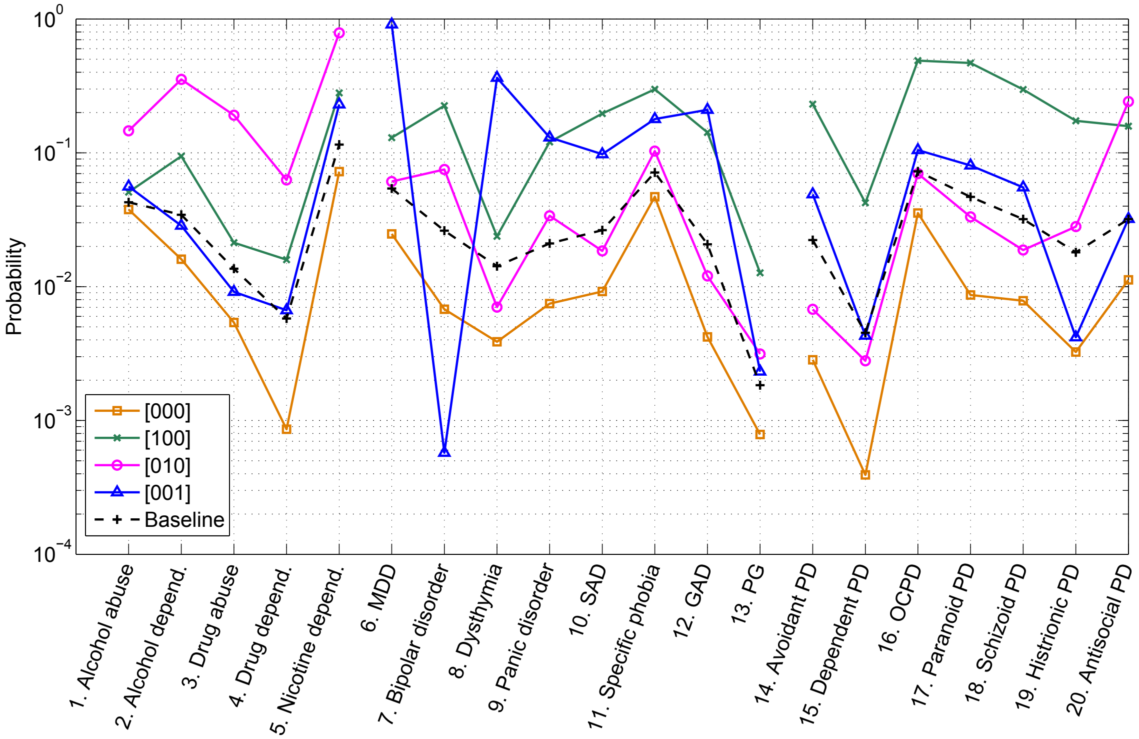


Figure 5: Probabilities of suffering from the 20 considered disorders. These probabilities have been obtained using the matrices $\mathbf{B}^d_{\text{MAP}}$, when none or a single latent feature is active. The legend shows the latent feature vector corresponding to each curve. The baseline has been obtained taking into account the 43, 093 subjects in the database.

rows of $\mathbf{Z}$ corresponding to those subjects who do not suffer from any of the 20 disorders, but instead fix these latent features to zero. The idea is that the $\mathbf{b}^d_0$ terms must capture the general population that does not suffer from any psychiatric disorder, and we use the active components of the matrix $\mathbf{Z}$ to characterize the disorders.

To examine the three latent features, we plot in Figure 5 the posterior probability of having each of the considered disorders, when none or one of the latent features is active. As expected, for those subjects who do not possess any feature, the probability of having any of the disorders is below the baseline level (due to the contribution of the vectors $\mathbf{b}^d_0$), defined as the empirical probability in the full sample, that is, taking into account the 43, 093 participants. Feature 1 increases the probability of having all the disorders, and thus seems to represent a general psychopathology factor, although it may particularly increase the risk of personality disorders. Feature 2 models substance use disorders and antisocial personality disorder, consistent with the externalizing factor identified in previous studies of the structure of psychiatric disorders (Krueger, 1999; Kendler et al., 2003; Vollebergh et al., 2001; Blanco et al., 2013). Feature 3 models mood or anxiety disorders, and thus seems to represent the internalizing factor also identified in previous studies.

Feature vector	1 x x	x 1 x	x x 1
Empirical Probability	0.0748	0.0330	0.0227

(a)

Feature vector	1 1 x	1 x 1	x 1 1
Empirical Probability	0.0028	0.0012	0.0009
Product Probability	0.0025	0.0017	0.0007

(b)

Table 1: Probabilities of possessing at least (a) one latent feature, or (b) two latent features, as given in the patterns shown in the heading rows. The symbol 'x' denotes either 0 or 1. The 'empirical probability' rows contain the probabilities extracted directly from the inferred IBP matrix $\mathbf{Z}$, while the 'product probability' row shows the product of the corresponding two latent feature probabilities given in (a).

Thus, in accord to previous results from the studies on the latent structure of the comorbidity of psychiatric disorders, detailed in Section 1, we find that the patterns of comorbidity of common psychiatric disorders can be well described by a small number of latent features. In addition, nosologically related disorders, such as social anxiety disorder and avoidant personality disorder, tend to be modeled by similar features. As found in previous results (Blanco et al., 2013), no disorder is perfectly aligned along one single latent feature, therefore suggesting that disorders can develop through multiple etiological paths. For instance, the risk of nicotine dependence may be particularly high in individuals with a propensity towards externalization or internalization.

In Table 1a, we first show the empirical probability of possessing each latent feature, that is, the number of subjects in the database that possess each latent feature divided by the total number of subjects. We also show in Table 1b the probability of possessing at least two features as the product of the probabilities in Table 1a (Product Probability), and also the empirical probability. We include Table 1 to show that the three features are nearly independent of one another, since the probability of possessing any two particular features is close to the product of the probabilities of possessing them individually. The differences in Table 1b are not statistically significant. Then, besides explicitly capturing the probability of each disorder, our model also provides a way to measure independence among the latent features. Note that although the proposed model assumes that the latent features are independent *a priori*, we could have found that the empirical probability does not correspond to the product one. Therefore, the independence among the three latent features follows the model's assumption and, from a psychiatric perspective, it also shows that the three factors (internalizing, externalizing and general psychopathology factor) are independent one another, that is, suffering from one group of disorders does not imply an increased probability of suffering from any other group of disorders.

Finally, we remark that we have also applied the variational inference algorithm to study the comorbidity patterns of psychiatric disorders but, since both algorithms (the variational and the Gibbs sampler) infer the same three latent features, we only plot the results for the

Gibbs sampling algorithm in this section and apply the variational inference algorithm in next section.

5.3 Comorbidity Analysis of Personality Disorders

In order to identify the seven personality disorders studied in the previous section, psychiatrists have established specific diagnostic criteria for each of them. These criteria correspond to affirmative responses to one or several questions in the NESARC survey and this correspondence is shown in Appendix D. Then, there exists a set of criteria to identify if a subject presents any of the following personality disorders: avoidant, dependent, obsessive-compulsive, paranoid, schizoid, histrionic and antisocial. In the present analysis, we consider as input data the fulfillment of the 52 criteria (i.e., $R = 2$) corresponding to all the disorders for the 43,093 subjects and we apply the variational inference algorithm truncated to $K = 25$ features, as detailed in Section 4, to find the latent structure of the data.

In order to properly initialize the huge amount of variational parameters, we have previously run six Gibbs samplers over the data but taking only the criteria corresponding to the avoidant PD and another PD (that is, the seven criteria for the avoidant PD and the seven for the dependent PD, the criteria for the avoidant PD with the eight for the OCPD, etc.) for $10,000$ randomly chosen subjects. After running the six Gibbs samplers, we obtain 18 latent features that we group in a unique matrix $\mathbf{Z}$ to obtain the weighting matrices $\mathbf{B}^d_{\text{MAP}}$, which are used to initialize some parameters ν_{nk} and ϕ^d_{kr}. We do this because the variational algorithm is sensitive to the starting point and a random initialization would not produce good solutions.

We run enough iterations of the variational algorithm to ensure convergence of the variational lower bound (the lower bound at each iteration is shown in Figure 6). We construct a binary matrix $\mathbf{Z}$ by setting each element $z_{nk} = 1$ if $\nu_{nk} > 0.5$. We flip (changing zeros by ones, and vice versa) those features possessed by more than 80% of the subjects, obtaining only 10 latent features possessed by more than 50 subjects among the $43,093$ in the database and then recomputing the weighting matrices. In Table 2, we show the probability of occurrence of each feature (top row), as well as the probability of having active only one single feature (bottom row). We also show the 'empirical' and the 'product' probabilities of possessing at least two latent features in Table 3, and the probabilities of possessing at least two features given that one of them is active in Table 4.

Features	1	2	3	4	5	6	7	8	9	10
Total	43.45	19.01	15.28	13.99	11.76	8.97	7.54	6.91	1.86	1.43
Single feature	13.48	3.62	2.22	1.34	2.27	0.49	0.76	1.07	0	0

Table 2: Probabilities (%) of possessing (top row) at least one latent feature, or (bottom row) a single feature.

In Figure 7, we plot the probability of meeting each criterion in the general population (dashed line) and the probability of meeting each criterion for those subjects that do not have any active feature in our model (solid line). There are $15,185$ subjects (35.2% of the

Features	1	2	3	4	5	6	7	8	9	10
1	🟧	9.92	8.96	8.48	5.67	7.22	4.92	3.85	1.46	1.42
2	8.26	🟧	4.43	4.54	3.67	1.90	1.43	2.08	0.71	0.21
3	6.64	2.90	🟧	3.29	2.18	3.00	2.02	1.58	0.54	0.20
4	6.08	2.66	2.14	🟧	2.79	1.91	2.39	1.40	1.25	0.03
5	5.11	2.23	1.80	1.65	🟧	1.31	1.35	0.85	0.57	0.00
6	3.90	1.71	1.37	1.26	1.05	🟧	1.10	0.80	0.44	0.14
7	3.28	1.43	1.15	1.06	0.89	0.68	🟧	0.65	0.28	0.00
8	3.00	1.31	1.06	0.97	0.81	0.62	0.52	🟧	0.51	0.07
9	0.81	0.35	0.28	0.26	0.22	0.17	0.14	0.13	🟧	0.00
10	0.62	0.27	0.22	0.20	0.17	0.13	0.11	0.10	0.03	🟧

Table 3: Probabilities (%) of possessing at least two latent features. The elements above the diagonal correspond to the 'empirical probability', that is, extracted directly from the inferred IBP matrix $\mathbf{Z}$, and the elements below the diagonal correspond to the 'product probability' of the corresponding two latent feature probabilities given in the first row of Table 2.

population) which do not present any active feature, and for these people the probability of meeting any criterion is reduced significantly.

We have found results that are in accordance with previous studies and at the same time provide new information to understand personality disorders. Out of the 10 features, 6 of them directly describe personality disorders. Feature 1 increases the probability of fulfilling the criteria for OCPD, Feature 3 increases the probability of fulfilling the criteria for antisocial, Feature 4 increases the probability of fulfilling the criteria for paranoid, Feature 5 increases the probability of meeting the criteria for schizoid, Feature 8 increases the probability of fulfilling the criteria for histrionic and Feature 7 increases the probability of meeting the criteria for avoidant and dependent. In Figure 8, we plot the probability ratio between the probability of meeting each criterion when a single feature is active with respect to the probability of meeting each criterion in the general population (baseline in

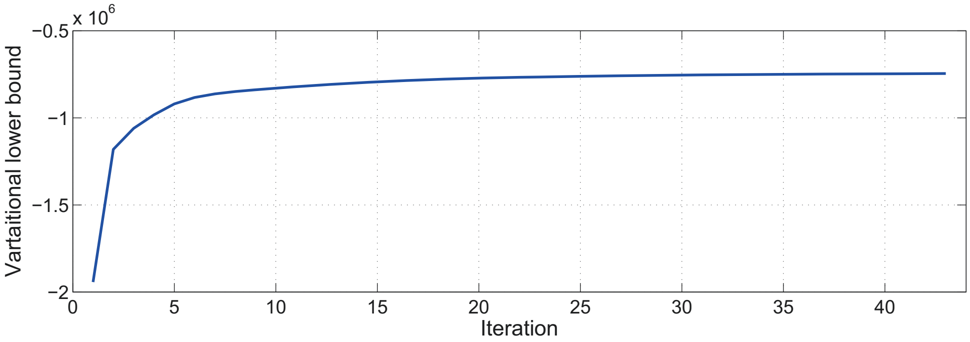


Figure 6: Variational lower bound $\mathcal{L}(\mathcal{H}, \mathcal{H}_q)$ at each iteration.

k_2 k_1	1	2	3	4	5	6	7	8	9	10
1	100	22.83	20.63	19.53	13.05	16.62	11.33	8.85	3.37	3.27
2	52.19	100	23.33	23.90	19.32	10.00	7.51	10.95	3.75	1.09
3	58.68	29.03	100	21.54	14.29	19.66	13.25	10.34	3.51	1.29
4	60.63	32.47	23.52	100	19.97	13.65	17.05	10.02	8.92	0.20
5	48.22	31.25	18.57	23.77	100	11.11	11.49	7.24	4.88	0.00
6	80.47	21.18	33.47	21.29	14.56	100	12.23	8.92	4.86	1.53
7	65.26	18.92	26.83	31.63	17.91	14.55	100	8.65	3.66	0.03
8	55.62	30.11	22.86	20.28	12.32	11.58	9.43	100	7.39	1.07
9	78.46	38.23	28.77	67.00	30.76	23.41	14.82	27.40	100	0.12
10	99.19	14.40	13.75	1.94	0.00	9.55	0.16	5.18	0.16	100

Table 4: Probabilities (%) of possessing at least features k_1 and k_2 given that k_1 is active, i.e., $\left(\sum_{n=1}^{N} z_{nk_1} z_{nk_2}\right) / \left(\sum_{n=1}^{N} z_{nk_1}\right)$.

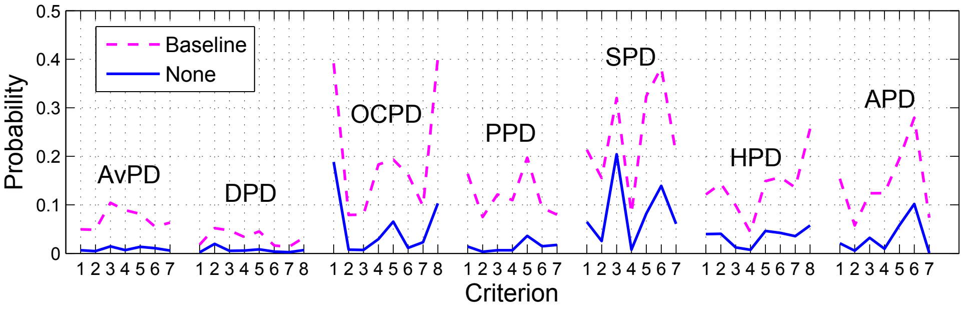


Figure 7: Probability of meeting each criterion. The probabilities when no latent feature is active (solid curve) have been obtained using the matrices $\mathbf{B}^{d}_{\text{MAP}}$, while the baseline (dashed curve) has been obtained taking into account the 43,093 subjects in the database.
(AvPD=Avoidant PD, DPD=Dependent PD, OCPD=Obsessive-compulsive PD, PPD=Paranoid PD, SPD=Schizoid PD, HPD=Histrionic PD, APD=Antisocial PD)

Figure 7). So, if the ratio is above one, it means that the feature increases the probability of meeting that criterion with respect to the general population. In all these plots, we also show the probability ratio between not having any active feature and the general population, which serves as a reference for a low probability of fulfilling a criterion. Note that the scale on the vertical axis may be different through all the figures for a better display. In Figure 8, we can see that only the criteria for one of the personality disorders is systematically above one, when one feature is active, except for Feature 7 that increases the probability for both avoidant and dependent. In the figure, we can also notice that when one feature is active

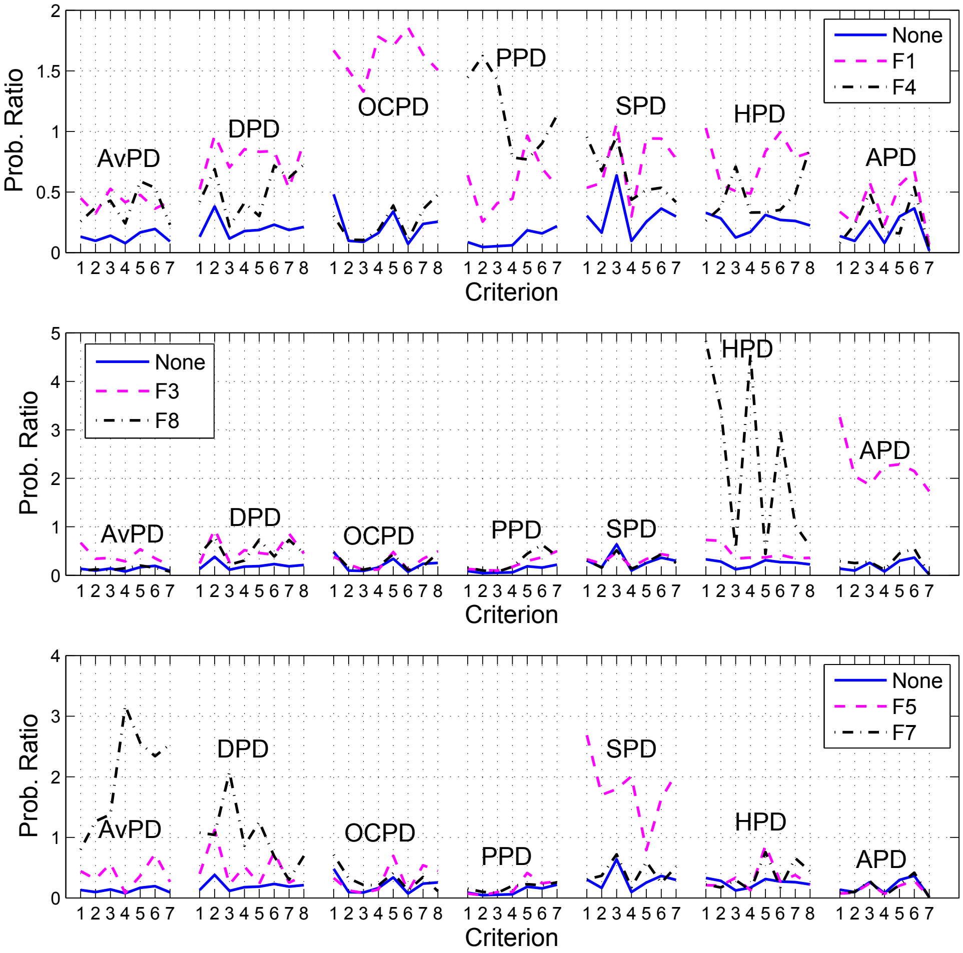


Figure 8: Probability ratio of meeting each criterion, with respect to the baseline. These probabilities have been obtained using the matrices $\mathbf{B}^d_{\text{MAP}}$, when none or a single feature is active (the legend shows the active latent features).

the probability of the criteria for the other disorders is above the probability for the subjects that do not have any active feature, although lower than the general population (above the solid line and below one). It partially shows the comorbidity pattern for each personality disorder. For example, Feature 1, besides increasing the probability of meeting the criteria for OCPD, also increases the probability of meeting criterion 3 for schizoid and criterion 1 for histrionic. It is also important to point out that Feature 8 increases significantly the probability of meeting criteria 1, 2, 4 and 6 for histrionic (and mildly for criterion 7), but it does not affect criteria 3, 5 and 8, although the probability of meeting these criteria are increased by Feature 4 (paranoid) and Feature 5 (schizoid). In a way, it indicates that criteria 3 and 8 are more related to paranoid disorder and criterion 5 to schizoid disorder.

As seen in Figure 9, Features 2 and 6 mainly reduce the probability of meeting the criteria for dependent PD. Feature 2 also reduces criteria 4-7 for avoidant and mildly increases

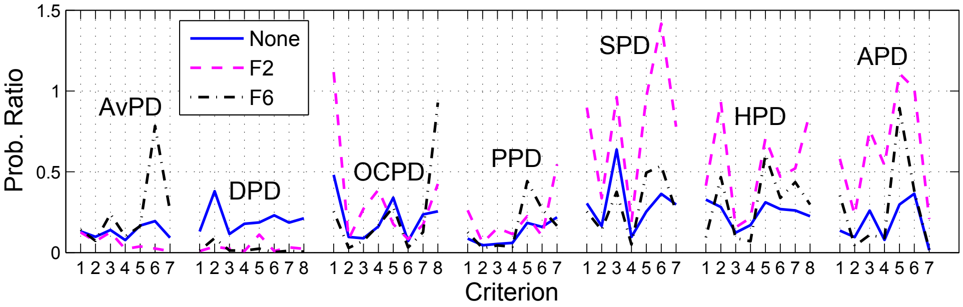


Figure 9: Probability ratio of meeting each criterion, with respect to the baseline. These probabilities have been obtained using the matrices $\mathbf{B}^d_{\text{MAP}}$, when none or a single feature is active (the legend shows the active latent features).

criterion 1 for OCPD, criterion 6 for schizoid and criteria 5 and 6 for antisocial. Feature 6 also reduces some criteria below the probability for the subjects with no active features. But for most of the criteria the probability ratio moves between one and the ratio for the subjects with no active feature. When these features appear by themselves, the subjects might be similar to the subjects without any active feature, they become relevant when they appear together with other features. These features are less likely to be isolated features than the previous ones, as reported in Table 2. For example, Feature 2 appears frequently with Features 1, 3, 4 and 5, as shown in Table 4, and the probability ratios are plotted in Figure 10 and compared to the probability ratio when each feature is not accompanied by Feature 2. We can see that when we add Feature 2 to Feature 1, the comorbidity pattern changes significantly and it results in subjects with higher probabilities of meeting the criteria for every other disorder except avoidant and dependent. Additionally, when we add Feature 2 to Feature 5, we can see that meeting the criteria for schizoid is even more probable, together with criterion 5 for histrionic.

Either Feature 1 or Features 1 and 3 typically accompany Feature 6, and Feature 6 is seldom seen by itself (see Tables 2 and 5). In Figure 11, we show the probability ratio when Feature 1 is active and when Features 1 and 3 are active, as reference, and when we add Feature 6 to them. Adding Feature 6 mainly reduces the probability of meeting the criteria for dependent. It is also relevant to point out that Features 1 and 3 increase the probability of meeting the criteria 5 and 6 for paranoid, while Feature 4 mainly increased the probability of meeting the criteria 1-4 for paranoid personality disorder, as shown in Figure 8.

Feature 9 is similar to Feature 7, as it captures an increase in the probability of meeting the criteria for avoidant and dependent, but it never appears isolated and most times it appears together with Features 1 and 4.

Feature 10 never appears isolated and it mainly appears only with Feature 1. This feature by itself only indicates that the probability of all the criteria should be much lower than the subjects with no active features, except for antisocial, which behaves as the subjects with no active features. When we add Feature 1 to Feature 10, we get that the probability of meeting the criteria for OCDP goes to that of the subject with no active features, as can

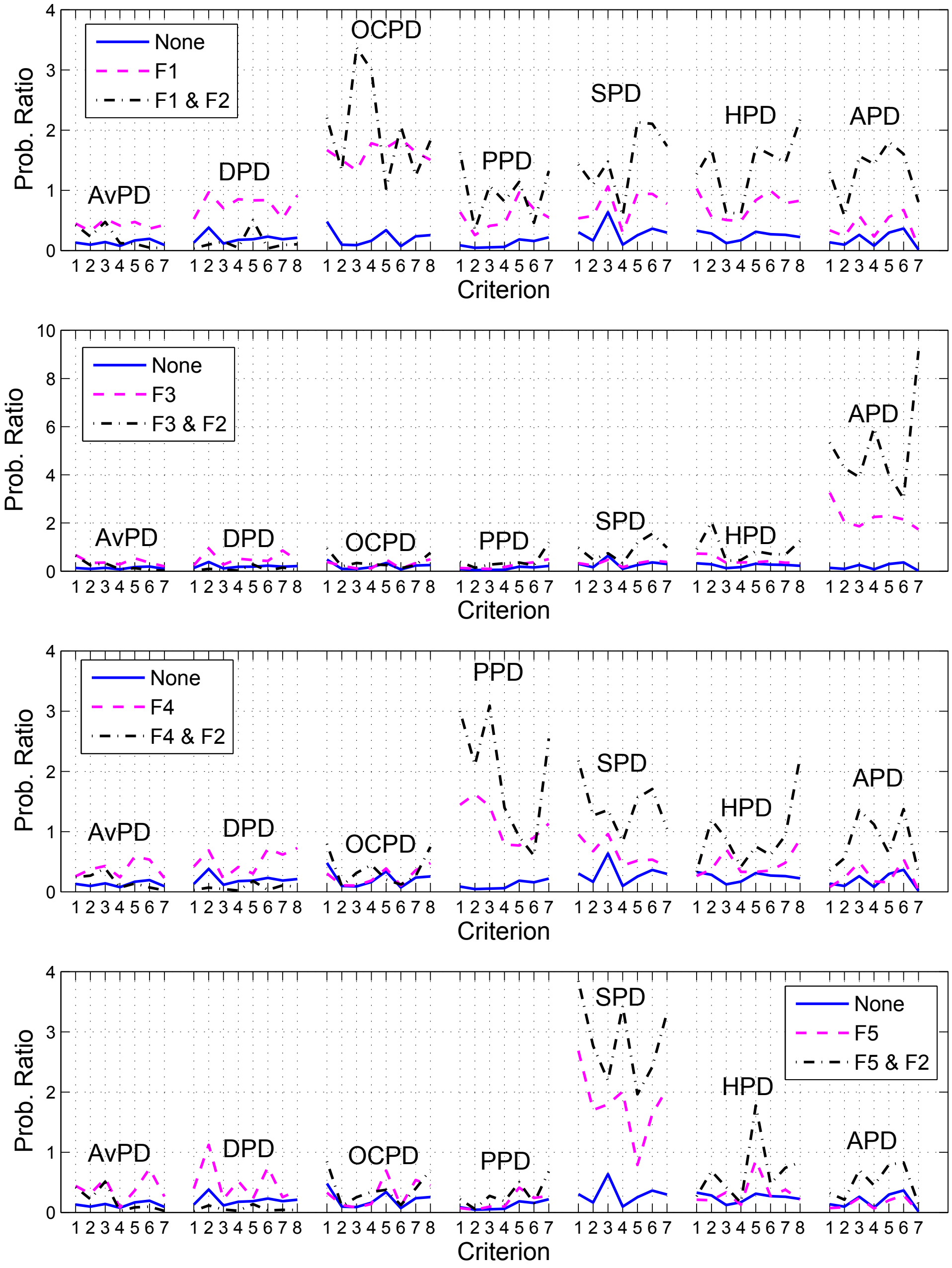


Figure 10: Probability ratio of meeting each criterion, with respect to the baseline. These probabilities have been obtained using the matrices $\mathbf{B}^{d}_{\text{MAP}}$, when none, a single or two features are active (the legend shows the active latent features).

be seen in Figure 12. For us this is a spurious feature that is equivalent to not having any active feature and that the variational algorithm has not been able to eliminate. This is always a risk when working with flexible models, like BNP, in which a spurious component

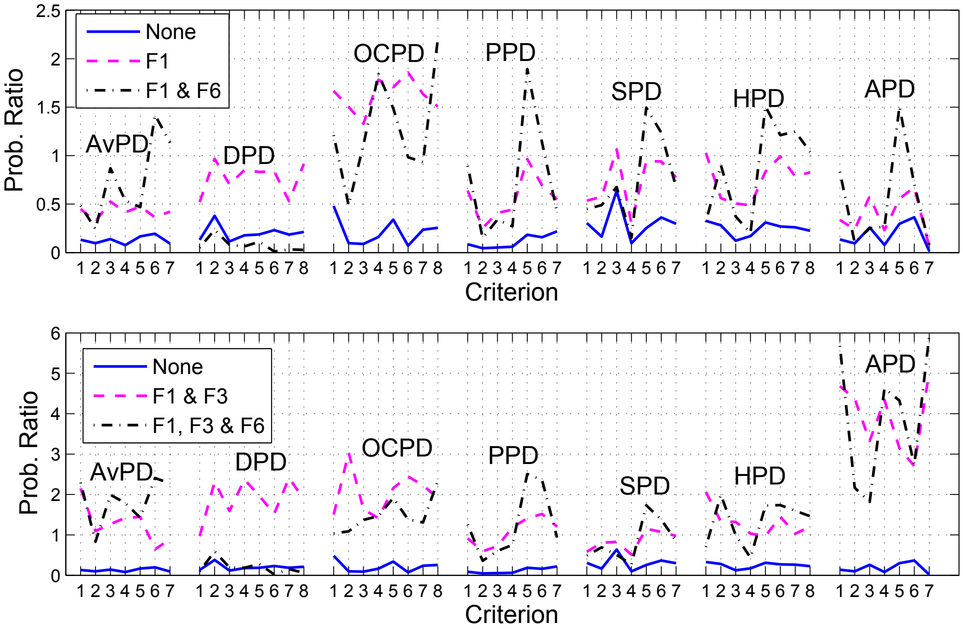


Figure 11: Probability ratio of meeting each criterion, with respect to the baseline. These probabilities have been obtained using the matrices $\mathbf{B}^d_{\text{MAP}}$, when none, a single or several features are active (the legend shows the active latent features).

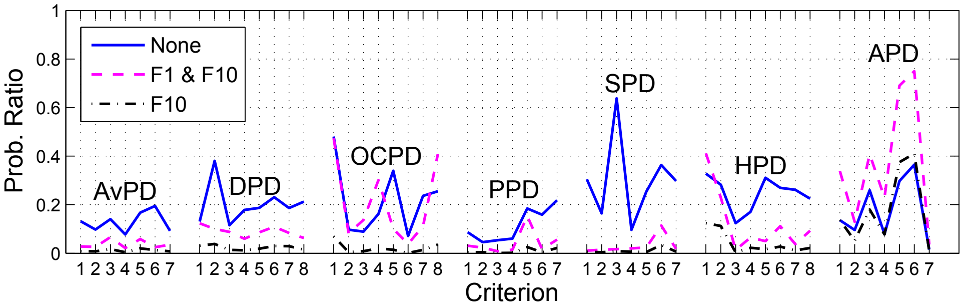


Figure 12: Probability ratio of meeting each criterion, with respect to the baseline. These probabilities have been obtained using the matrices $\mathbf{B}^d_{\text{MAP}}$, when none, a single or two features are active (the legend shows the active latent features).

might appear when it should not. These components can be eliminated by common sense in most cases or by further analysis by experts (psychiatric experts in our case). But it can also indicate an unknown component that can point towards a new research direction previously unknown, which is one of the attractive features of using generative models.

Besides the comorbidity patterns shown by the individual features that we have already reported, we can also see that almost all the features are positively correlated. In Table 3, we

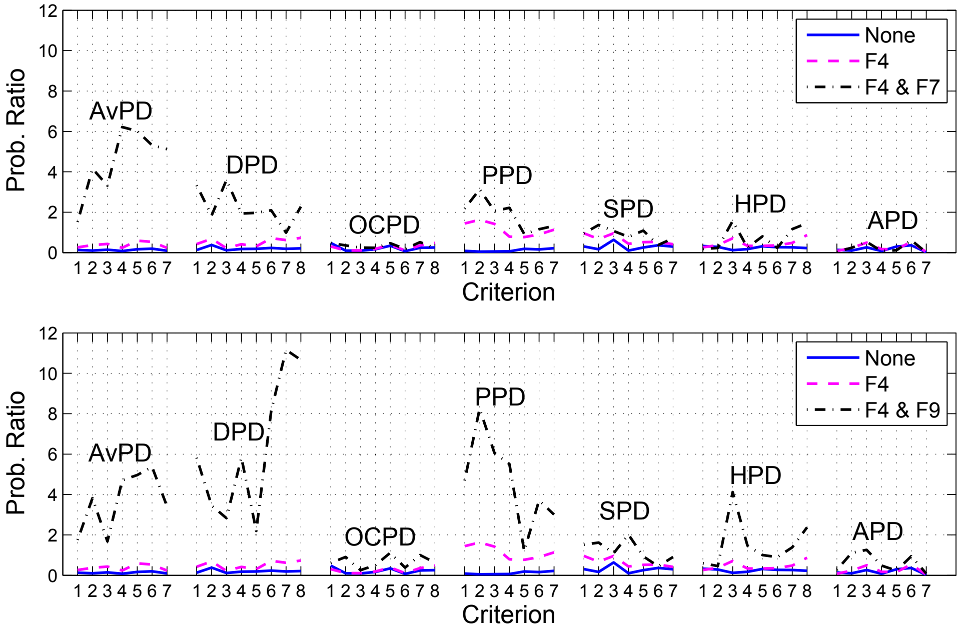


Figure 13: Probability ratio of meeting each criterion, with respect to the baseline. These probabilities have been obtained using the matrices $\mathbf{B}^d_{\mathrm{MAP}}$, when none, a single or two features are active (the legend shows the active latent features).

show the probability that any two features appear together (upper triangular sub-matrix) and the joint probability that we should observe if the features were independent (lower triangular sub-matrix). Ignoring Feature 10, all of the other features are positively correlated, except Features 2 and 7 and Features 8 and 5 that seem uncorrelated (the differences are not statistically significant). Most of the features are strongly correlated and the differences in Table 3 correspond to several standard deviations higher (between 3 and 42) that we should expect from independent random observations. For example, the correlation between Features 4 and 9 and Features 4 and 7 is quite high and both show subjects with higher probability of meeting the criteria for avoidant, dependent and paranoid. The difference between Features 7 and 9 is given by the criteria 1-4 for paranoid PD, that are significantly increased by Feature 9 and slightly by Feature 7, as it can be seen in Figure 13. Finally, it is worth mentioning that Feature 4 (paranoid) is the most highly correlated feature with all the others, so we can say that anyone suffering from paranoid PD has a higher comorbidity with any other personality disorder.

6. Conclusions

In this paper, we have proposed a new model that combines the IBP with discrete observations using the multinomial-logit distribution. We have used the Laplace approximation to integrate out the weighting factors, which allows us to efficiently run the Gibbs sampler.

	# Occurrences	**Features**									
		1	2	3	4	5	6	7	8	9	10
1	15185	0	0	0	0	0	0	0	0	0	0
2	5811	1	0	0	0	0	0	0	0	0	0
3	1561	0	1	0	0	0	0	0	0	0	0
4	1389	1	1	0	0	0	0	0	0	0	0
5	1021	1	0	1	0	0	0	0	0	0	0
6	977	0	0	0	0	1	0	0	0	0	0
7	958	1	0	0	0	0	1	0	0	0	0
8	956	0	0	1	0	0	0	0	0	0	0
9	946	1	0	0	1	0	0	0	0	0	0
10	687	1	0	0	0	1	0	0	0	0	0
11	576	0	0	0	1	0	0	0	0	0	0
12	553	1	0	0	0	0	0	1	0	0	0
13	495	0	1	0	0	1	0	0	0	0	0
14	486	1	0	0	0	0	0	0	1	0	0
15	460	0	0	0	0	0	0	0	1	0	0
16	451	0	1	1	0	0	0	0	0	0	0
17	438	1	0	0	0	0	0	0	0	0	1
18	414	1	0	1	0	0	1	0	0	0	0
19	385	0	1	0	1	0	0	0	0	0	0
20	370	1	1	0	1	0	0	0	0	0	0

Table 5: List of the 20 most common feature patterns.

We have also derived a variational inference algorithm, which allows dealing with larger databases and provides accurate results.

We have applied our model to the NESARC database to find out the hidden features that characterize the psychiatric disorders. First, we have used the Gibbs sampler to extract the latent structure behind 20 of the most common psychiatric disorders. As a result, we have found that the comorbidity patterns of these psychiatric disorders can be described by only three latent features, which mainly model the internalizing disorders, the externalizing disorders, and a general psychopathology factor. Additionally, we have applied the variational inference algorithm to analyze the relation among the 52 criteria defined by the psychiatrists to diagnose each of the seven personality disorders (that is, externalizing disorders). We have obtained that for most of the disorders, a latent feature appears to model all the criteria that characterize that particular disorder. In this experiment, we have also seen that avoidant and dependent PDs are jointly modeled by four features, and that paranoid disorder is the most highly correlated PD with all the others.

Acknowledgments

Francisco J. R. Ruiz is supported by an FPU fellowship from the Spanish Ministry of Education (AP2010-5333), Isabel Valera is supported by the *Plan Regional-Programas I+D* of *Comunidad de Madrid* (AGES-CM S2010/BMD-2422), Fernando Pérez-Cruz has been partially supported by a Salvador de Madariaga grant, and Carlos Blanco acknowledges NIH grants (DA019606 and DA023200) and the New York State Psychiatric Institute for their support. The authors also acknowledge the support of *Ministerio de Ciencia e Innovación* of Spain (projects DEIPRO TEC2009-14504-C02-00, ALCIT TEC2012-38800-C03-01, and program Consolider-Ingenio 2010 CSD2008-00010 COMONSENS). This work was also supported by the European Union 7th Framework Programme through the Marie Curie Initial Training Network "Machine Learning for Personalized Medicine" MLPM2012, Grant No. 316861.

Appendix A. Laplace Approximation Details

In this section we provide the necessary details for the implementation of the Laplace approximation proposed in Section 3.1. The expression in (5) can be rewritten as

$$\begin{aligned} f(\mathbf{B}^d) = \operatorname{trace}\left\{{\mathbf{M}^d}^\top \mathbf{B}^d\right\} - \sum_{n=1}^{N} \log\left(\sum_{r=1}^{R} \exp(\mathbf{z}_{n\bullet}\mathbf{b}^d_{\bullet r})\right) \\ - \frac{1}{2\sigma_B^2}\operatorname{trace}\left\{{\mathbf{B}^d}^\top \mathbf{B}^d\right\} - \frac{R(K+1)}{2}\log(2\pi\sigma_B^2), \end{aligned}$$

where $(\mathbf{M}^d)_{kr}$ counts the number of data points for which $x_{nd} = r$ and $z_{nk} = 1$, namely, $(\mathbf{M}^d)_{kr} = \sum_{n=1}^N \delta(x_{nd} = r) z_{nk}$, where $\delta(\cdot)$ is the Kronecker delta function. By definition, $(\mathbf{M}^d)_{0r} = \sum_{n=1}^N \delta(x_{nd} = r)$.

By defining $(\boldsymbol{\rho}^d)_{kr} = \displaystyle\sum_{n=1}^{N} z_{nk}\pi^r_{nd}$, the gradient of $f(\mathbf{B}^d)$ can be derived as

$$\nabla f = \mathbf{M}^d - \boldsymbol{\rho}^d - \frac{1}{\sigma_B^2}\mathbf{B}^d.$$

To compute the Hessian, it is easier to define the gradient ∇f as a vector, instead of a matrix, and hence we stack the columns of $\mathbf{B}^d$ into $\boldsymbol{\beta}^d$, i.e., $\boldsymbol{\beta}^d = \mathbf{B}^d(:)$ for avid Matlab users. The Hessian matrix can now be readily computed taking the derivatives of the gradient, yielding

$$\begin{aligned} \nabla\nabla f &= -\frac{1}{\sigma_B^2}\mathbf{I}_{R(K+1)} + \nabla\nabla \log p(\mathbf{x}_{\bullet d}|\boldsymbol{\beta}^d, \mathbf{Z}) \\ &= -\frac{1}{\sigma_B^2}\mathbf{I}_{R(K+1)} - \sum_{n=1}^{N}\Big(\operatorname{diag}(\boldsymbol{\pi}_{nd}) - (\boldsymbol{\pi}_{nd})^\top \boldsymbol{\pi}_{nd}\Big) \otimes (\mathbf{z}_{n\bullet}^\top \mathbf{z}_{n\bullet}), \end{aligned}$$

where $\operatorname{diag}(\boldsymbol{\pi}_{nd})$ is a diagonal matrix with the values of the vector $\boldsymbol{\pi}_{nd} = \left[\pi^1_{nd}, \pi^2_{nd}, \ldots, \pi^R_{nd}\right]$ as its diagonal elements.

Finally, note that, since $p(\mathbf{x}_{\bullet d}|\boldsymbol{\beta}^d, \mathbf{Z})$ is a log-concave function of $\boldsymbol{\beta}^d$ (Boyd and Vandenberghe, 2004, p. 87), $-\nabla\nabla f$ is a positive definite matrix, which guarantees that the maximum of $f(\boldsymbol{\beta}^d)$ is unique.

Appendix B. Lower Bound Derivation

In this section we derive the lower bound $\mathcal{L}(\mathcal{H}, \mathcal{H}_q)$ on the evidence $p(\mathbf{X}|\mathcal{H})$. From Eq. (9),

$$\begin{aligned}\log p(\mathbf{X}|\mathcal{H}) &= \mathbb{E}_q\left[\log p(\Psi, \mathbf{X}|\mathcal{H})\right] + H[q] + D_{KL}(q||p)\\ &\geqslant \mathbb{E}_q\left[\log p(\Psi, \mathbf{X}|\mathcal{H})\right] + H[q].\end{aligned}$$

The expectation $\mathbb{E}_q\left[\log p(\Psi, \mathbf{X}|\mathcal{H})\right]$ can be derived as

$$\begin{aligned}\mathbb{E}_q\left[\log p(\Psi, \mathbf{X}|\mathcal{H})\right] = &\sum_{k=1}^{K}\underbrace{\mathbb{E}_q\left[\log p(v_k|\alpha)\right]}_{1} + \sum_{d=1}^{D}\sum_{k=1}^{K}\underbrace{\mathbb{E}_q\left[\log p(\mathbf{b}_{k\bullet}^d|\sigma_B^2)\right]}_{2} + \sum_{d=1}^{D}\underbrace{\mathbb{E}_q\left[\log p(\mathbf{b}_0^d|\sigma_B^2)\right]}_{3}\\ &+ \sum_{k=1}^{K}\sum_{n=1}^{N}\underbrace{\mathbb{E}_q\left[\log p(z_{nk}|\{v_i\}_{i=1}^k)\right]}_{4} + \sum_{n=1}^{N}\sum_{d=1}^{D}\underbrace{\mathbb{E}_q\left[\log p(x_{nd}|\mathbf{z}_{n\bullet}, \mathbf{B}^d, \mathbf{b}_0^d)\right]}_{5},\end{aligned} \tag{11}$$

where each term can be computed as shown below:

1. For the Beta distribution over v_k,

$$\mathbb{E}_q\left[\log p(v_k|\alpha)\right] = \log(\alpha) + (\alpha - 1)\left[\psi(\tau_{k1}) - \psi(\tau_{k1} + \tau_{k2})\right].$$

2. For the Gaussian distribution over vectors $\mathbf{b}_{k\bullet}^d$,

$$\mathbb{E}_q\left[\log p(\mathbf{b}_{k\bullet}^d|\sigma_B^2)\right] = -\frac{R}{2}\log(2\pi\sigma_B^2) - \frac{1}{2\sigma_B^2}\left(\sum_{r=1}^{R}(\phi_{kr}^d)^2 + \sum_{r=1}^{R}\left(\sigma_{kr}^d\right)^2\right).$$

3. For the Gaussian distribution over $\mathbf{b}_0^d$,

$$\mathbb{E}_q\left[\log p(\mathbf{b}_0^d|\sigma_B^2)\right] = -\frac{R}{2}\log(2\pi\sigma_B^2) - \frac{1}{2\sigma_B^2}\left(\sum_{r=1}^{R}(\phi_{0r}^d)^2 + \sum_{r=1}^{R}(\sigma_{0r}^d)^2\right).$$

4. For the feature assignments, which are Bernoulli distributed given the feature probabilities, we have

$$\begin{aligned}\mathbb{E}_q\left[\log p(z_{nk}|\{v_i\}_{i=1}^k)\right] =&(1 - \nu_{nk})\mathbb{E}_q\left[\log\left(1 - \prod_{i=1}^{k} v_i\right)\right]\\ &+ \nu_{nk}\sum_{i=1}^{k}\left[\psi(\tau_{i1}) - \psi(\tau_{i1} + \tau_{i2})\right],\end{aligned}$$

where the expectation $\mathbb{E}_q\left[\log\left(1-\prod_{i=1}^{k} v_i\right)\right]$ has no closed-form solution. We can instead lower bound it by using the multinomial approach (Doshi-Velez et al., 2009). Under this approach, we introduce an auxiliary multinomial distribution $\boldsymbol{\lambda}_k = [\lambda_{k1}, \ldots, \lambda_{kk}]$ in the expectation and apply Jensen's inequality, yielding

$$\begin{aligned}\mathbb{E}_q\left[\log\left(1-\prod_{i=1}^{k} v_i\right)\right] \geq & \sum_{m=1}^{k} \lambda_{km}\psi(\tau_{m2}) + \sum_{m=1}^{k-1}\left(\sum_{n=m+1}^{k}\lambda_{kn}\right)\psi(\tau_{m1}) \\ & - \sum_{m=1}^{k}\left(\sum_{n=m}^{k}\lambda_{kn}\right)\psi(\tau_{m1}+\tau_{m2}) - \sum_{m=1}^{k}\lambda_{km}\log(\lambda_{km}),\end{aligned}$$

which holds for any distribution represented by the probabilities $\lambda_{k1}, \ldots, \lambda_{kk}$, for $1 \leq k \leq K$. Then,

$$\begin{aligned}\mathbb{E}_q\left[\log p(z_{nk}|\{v_i\}_{i=1}^{k})\right] \geq (1-\nu_{nk}) & \left[\sum_{m=1}^{k} \lambda_{km}\psi(\tau_{m2}) + \sum_{m=1}^{k-1}\left(\sum_{n=m+1}^{k}\lambda_{kn}\right)\psi(\tau_{m1})\right. \\ & \left. - \sum_{m=1}^{k}\left(\sum_{n=m}^{k}\lambda_{kn}\right)\psi(\tau_{m1}+\tau_{m2}) - \sum_{m=1}^{k}\lambda_{km}\log(\lambda_{km})\right] \\ & + \nu_{nk}\sum_{i=1}^{k}\left[\psi(\tau_{i1}) - \psi(\tau_{i1}+\tau_{i2})\right].\end{aligned}$$

5. For the likelihood term, we can write

$$\mathbb{E}_q\left[\log p(x_{nd}|\mathbf{z}_{n\bullet}, \mathbf{B}^d, \mathbf{b}_0^d)\right] = \phi^d_{0x_{nd}} + \sum_{k=1}^{K}\nu_{nk}\phi^d_{kx_{nd}} - \mathbb{E}_q\left[\log\left(\sum_{r=1}^{R}\exp(\mathbf{z}_{n\bullet}\mathbf{b}^d_{\bullet r} + b^d_{0r})\right)\right],$$

where the logarithm can be upper bounded by its first-order Taylor series expansion around the auxiliary variable ξ_{nd}^{-1} (for $n = 1, \ldots, N$ and $d = 1, \ldots, D$) (Blei and Lafferty, 2007; Bouchard, 2007), yielding

$$\log\left(\sum_{r=1}^{R}\exp(\mathbf{z}_{n\bullet}\mathbf{b}^d_{\bullet r} + b^d_{0r})\right) \leq \xi_{nd}\left(\sum_{r=1}^{R}\exp(\mathbf{z}_{n\bullet}\mathbf{b}^d_{\bullet r} + b^d_{0r})\right) - \log(\xi_{nd}) - 1.$$

The main advantage of this bound lies on the fact that it allows us to compute the expectation of the bound for the Gaussian distribution, since it involves the moment generating functions of the distributions $q(\mathbf{b}^d_{\bullet r})$ and $q(b^d_{0r})$. Then, we can lower bound the likelihood term as

$$\begin{aligned}\mathbb{E}_q\left[\log p(x_{nd}|\mathbf{z}_{n\bullet}, \mathbf{B}^d, \mathbf{b}_0^d)\right] \geq & \phi^d_{0x_{nd}} + \sum_{k=1}^{K}\nu_{nk}\phi^d_{kx_{nd}} + \log(\xi_{nd}) + 1 \\ - \xi_{nd}\sum_{r=1}^{R} & \left[\exp\left(\phi^d_{0r} + \frac{1}{2}(\sigma^d_{0r})^2\right)\prod_{k=1}^{K}\left(1-\nu_{nk}+\nu_{nk}\exp\left(\phi^d_{kr}+\frac{1}{2}(\sigma^d_{kr})^2\right)\right)\right].\end{aligned}$$

Substituting the previous results in (11), we obtain

$$\begin{aligned}
\mathbb{E}_q\left[\log p(\Psi, \mathbf{X}|\mathcal{H})\right] \geq{}& \sum_{k=1}^{K}\left[\log(\alpha) + (\alpha-1)\left(\psi(\tau_{k1}) - \psi(\tau_{k1}+\tau_{k2})\right)\right] \\
&- \frac{R(K+1)D}{2}\log(2\pi\sigma_B^2) - \frac{1}{2\sigma_B^2}\sum_{k=0}^{K}\sum_{d=1}^{D}\sum_{r=1}^{R}\left((\phi_{kr}^d)^2 + {(\sigma_{kr}^d)}^2\right) \\
&+ \sum_{n=1}^{N}\sum_{k=1}^{K}\Bigg[\nu_{nk}\sum_{i=1}^{k}\left[\psi(\tau_{i1}) - \psi(\tau_{i1}+\tau_{i2})\right] \\
&\qquad +(1-\nu_{nk})\Bigg(\sum_{m=1}^{k}\lambda_{km}\psi(\tau_{m2}) + \sum_{m=1}^{k-1}\left(\sum_{n=m+1}^{k}\lambda_{kn}\right)\psi(\tau_{m1}) \\
&\qquad -\sum_{m=1}^{k}\left(\sum_{n=m}^{k}\lambda_{kn}\right)\psi(\tau_{m1}+\tau_{m2}) - \sum_{m=1}^{k}\lambda_{km}\log(\lambda_{km})\Bigg)\Bigg] \\
&+ \sum_{n=1}^{N}\sum_{d=1}^{D}\Bigg[\phi_{0x_{nd}}^d + \sum_{k=1}^{K}\nu_{nk}\phi_{kx_{nd}}^d + \log(\xi_{nd}) + 1 \\
&-\xi_{nd}\sum_{r=1}^{R}\left[\exp\left(\phi_{0r}^d + \frac{1}{2}(\sigma_{0r}^d)^2\right)\prod_{k=1}^{K}\left(1-\nu_{nk}+\nu_{nk}\exp\left(\phi_{kr}^d + \frac{1}{2}{(\sigma_{kr}^d)}^2\right)\right)\right]\Bigg].
\end{aligned}$$

Additionally, the entropy of the distribution $q(\Psi)$ is given by

$$\begin{aligned}
&H[q] = \mathbb{E}_q\left[\log q(\Psi)\right] \\
&\quad = \sum_{k=1}^{K}\mathbb{E}_q\left[\log q(v_k|\tau_{k1},\tau_{k2})\right] + \sum_{d=1}^{D}\sum_{r=1}^{R}\sum_{k=0}^{K}\mathbb{E}_q\left[\log q(b_{kr}^d|\phi_{kr}^d, {(\sigma_{kr}^d)}^2)\right] + \sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}_q\left[\log q(z_{nk}|\nu_{nk})\right] \\
&\quad = \sum_{k=1}^{K}\left[\log\left(\frac{\Gamma(\tau_{k1})\Gamma(\tau_{k2})}{\Gamma(\tau_{k1}+\tau_{k2})}\right) - (\tau_{k1}-1)\psi(\tau_{k1}) - (\tau_{k2}-1)\psi(\tau_{k2}) + (\tau_{k1}+\tau_{k2}-2)\psi(\tau_{k1}+\tau_{k2})\right] \\
&\qquad + \sum_{d=1}^{D}\sum_{r=1}^{R}\sum_{k=0}^{K}\frac{1}{2}\log(2\pi e{(\sigma_{kr}^d)}^2) + \sum_{n=1}^{N}\sum_{k=1}^{K}\left[-\nu_{nk}\log(\nu_{nk}) - (1-\nu_{nk})\log(1-\nu_{nk})\right].
\end{aligned}$$

Finally, we obtain the lower bound on the evidence $p(\mathbf{X}|\mathcal{H})$ as

$$
\begin{aligned}
&\log p(\mathbf{X}|\mathcal{H}) \geq \mathbb{E}_q\left[\log p(\Psi, \mathbf{X}|\mathcal{H})\right] + H[q] \\
&\geq \sum_{k=1}^{K} \left[\log(\alpha) + (\alpha - 1)\left(\psi(\tau_{k1}) - \psi(\tau_{k1} + \tau_{k2})\right)\right] \\
&\quad - \frac{R(K+1)D}{2}\log(2\pi\sigma_B^2) - \frac{1}{2\sigma_B^2}\sum_{k=0}^{K}\sum_{d=1}^{D}\sum_{r=1}^{R}\left((\phi_{kr}^d)^2 + {(\sigma_{kr}^d)}^2\right) \\
&\quad + \sum_{n=1}^{N}\sum_{k=1}^{K}\Bigg[\nu_{nk}\sum_{i=1}^{k}\left[\psi(\tau_{i1}) - \psi(\tau_{i1} + \tau_{i2})\right] \\
&\qquad\qquad + (1 - \nu_{nk})\left(\sum_{m=1}^{k}\lambda_{km}\psi(\tau_{m2}) + \sum_{m=1}^{k-1}\left(\sum_{n=m+1}^{k}\lambda_{kn}\right)\psi(\tau_{m1})\right. \\
&\qquad\qquad \left. - \sum_{m=1}^{k}\left(\sum_{n=m}^{k}\lambda_{kn}\right)\psi(\tau_{m1} + \tau_{m2}) - \sum_{m=1}^{k}\lambda_{km}\log(\lambda_{km})\right)\Bigg] \\
&\quad + \sum_{n=1}^{N}\sum_{d=1}^{D}\Bigg[\phi_{0x_{nd}}^d + \sum_{k=1}^{K}\nu_{nk}\phi_{kx_{nd}}^d + \log(\xi_{nd}) + 1 \\
&\quad - \xi_{nd}\sum_{r=1}^{R}\left[\exp\left(\phi_{0r}^d + \frac{1}{2}(\sigma_{0r}^d)^2\right)\prod_{k=1}^{K}\left(1 - \nu_{nk} + \nu_{nk}\exp\left(\phi_{kr}^d + \frac{1}{2}{(\sigma_{kr}^d)}^2\right)\right)\right]\Bigg] \\
&\quad + \sum_{k=1}^{K}\left[\log\left(\frac{\Gamma(\tau_{k1})\Gamma(\tau_{k2})}{\Gamma(\tau_{k1} + \tau_{k2})}\right) - (\tau_{k1} - 1)\psi(\tau_{k1}) - (\tau_{k2} - 1)\psi(\tau_{k2}) + (\tau_{k1} + \tau_{k2} - 2)\psi(\tau_{k1} + \tau_{k2})\right] \\
&\quad + \sum_{d=1}^{D}\sum_{r=1}^{R}\sum_{k=0}^{K}\frac{1}{2}\log(2\pi e{(\sigma_{kr}^d)}^2) + \sum_{n=1}^{N}\sum_{k=1}^{K}\left[-\nu_{nk}\log(\nu_{nk}) - (1 - \nu_{nk})\log(1 - \nu_{nk})\right] \\
&= \mathcal{L}(\mathcal{H}, \mathcal{H}_q),
\end{aligned}
$$

where $\mathcal{H}_q = \{\tau_{k1}, \tau_{k2}, \lambda_{km}, \xi_{nd}, \nu_{nk}, \phi_{kr}^d, \phi_{0r}^d, {(\sigma_{kr}^d)}^2, (\sigma_{0r}^d)^2\}$ (for $k = 1, \ldots, K$, $m = 1, \ldots, k$, $d = 1, \ldots, D$, and $n = 1, \ldots, N$) represents the set of the variational parameters.

Appendix C. Derivatives for Newton's Method

- For the parameters of the Gaussian distribution $q(b_{kr}^d|\phi_{kr}^d, {(\sigma_{kr}^d)}^2)$ for $k = 1, \ldots, K$,

$$
\begin{aligned}
\frac{\partial}{\partial\phi_{kr}^d}\mathcal{L}(\mathcal{H}, \mathcal{H}_q) = -\frac{1}{\sigma_B^2}\phi_{kr}^d + \sum_{n=1}^{N}\Bigg[&\nu_{nk}\delta(x_{nd} = r) - \nu_{nk}\xi_{nd}\exp\left(\phi_{0r}^d + \frac{1}{2}(\sigma_{0r}^d)^2\right)\exp\left(\phi_{kr}^d + \frac{1}{2}{(\sigma_{kr}^d)}^2\right) \\
&\times\prod_{k' \neq k}\left(1 - \nu_{nk'} + \nu_{nk'}\exp\left(\phi_{k'r}^d + \frac{1}{2}(\sigma_{k'r}^d)^2\right)\right)\Bigg].
\end{aligned}
$$

$$\frac{\partial^2}{\partial(\phi^d_{kr})^2}\mathcal{L}(\mathcal{H},\mathcal{H}_q) = -\frac{1}{\sigma_B^2} - \sum_{n=1}^{N}\Bigg[\nu_{nk}\xi_{nd}\exp\left(\phi^d_{0r}+\frac{1}{2}(\sigma^d_{0r})^2\right)\exp\left(\phi^d_{kr}+\frac{1}{2}(\sigma^d_{kr})^2\right) \\ \times\prod_{k'\neq k}\left(1-\nu_{nk'}+\nu_{nk'}\exp\left(\phi^d_{k'r}+\frac{1}{2}(\sigma^d_{k'r})^2\right)\right)\Bigg].$$

$$\frac{\partial}{\partial(\sigma^d_{kr})^2}\mathcal{L}(\mathcal{H},\mathcal{H}_q) = -\frac{1}{2\sigma_B^2}+\frac{1}{2}(\sigma^d_{kr})^{-2} - \frac{1}{2}\sum_{n=1}^{N}\Bigg[\nu_{nk}\xi_{nd}\exp\left(\phi^d_{0r}+\frac{1}{2}(\sigma^d_{0r})^2\right)\exp\left(\phi^d_{kr}+\frac{1}{2}(\sigma^d_{kr})^2\right) \\ \times\prod_{k'\neq k}\left(1-\nu_{nk'}+\nu_{nk'}\exp\left(\phi^d_{k'r}+\frac{1}{2}(\sigma^d_{k'r})^2\right)\right)\Bigg].$$

$$\frac{\partial^2}{(\partial(\sigma^d_{kr})^2)^2}\mathcal{L}(\mathcal{H},\mathcal{H}_q) = -\frac{1}{2}(\sigma^d_{kr})^{-4} - \frac{1}{4}\sum_{n=1}^{N}\Bigg[\nu_{nk}\xi_{nd}\exp\left(\phi^d_{0r}+\frac{1}{2}(\sigma^d_{0r})^2\right)\exp\left(\phi^d_{kr}+\frac{1}{2}(\sigma^d_{kr})^2\right) \\ \times\prod_{k'\neq k}\left(1-\nu_{nk'}+\nu_{nk'}\exp\left(\phi^d_{k'r}+\frac{1}{2}(\sigma^d_{k'r})^2\right)\right)\Bigg].$$

- For the parameters of the Gaussian distribution $q(b^d_{0r}|\phi^d_{0r},(\sigma^d_{0r})^2)$,

$$\frac{\partial}{\partial\phi^d_{0r}}\mathcal{L}(\mathcal{H},\mathcal{H}_q) \\ = -\frac{1}{\sigma_B^2}\phi^d_{0r} + \sum_{n=1}^{N}\left[\delta(x_{nd}=r) - \xi_{nd}\exp\left(\phi^d_{0r}+\frac{1}{2}(\sigma^d_{0r})^2\right)\prod_{k=1}^{K}\left(1-\nu_{nk}+\nu_{nk}\exp\left(\phi^d_{kr}+\frac{1}{2}(\sigma^d_{kr})^2\right)\right)\right].$$

$$\frac{\partial^2}{(\partial\phi^d_{0r})^2}\mathcal{L}(\mathcal{H},\mathcal{H}_q) \\ = -\frac{1}{\sigma_B^2} - \sum_{n=1}^{N}\left[\xi_{nd}\exp\left(\phi^d_{0r}+\frac{1}{2}(\sigma^d_{0r})^2\right)\prod_{k=1}^{K}\left(1-\nu_{nk}+\nu_{nk}\exp\left(\phi^d_{kr}+\frac{1}{2}(\sigma^d_{kr})^2\right)\right)\right].$$

$$\frac{\partial}{\partial(\sigma^d_{0r})^2}\mathcal{L}(\mathcal{H},\mathcal{H}_q) \\ = -\frac{1}{2\sigma_B^2}+\frac{1}{2}(\sigma^d_{0r})^{-2} - \frac{1}{2}\sum_{n=1}^{N}\left[\xi_{nd}\exp\left(\phi^d_{0r}+\frac{1}{2}(\sigma^d_{0r})^2\right)\prod_{k=1}^{K}\left(1-\nu_{nk}+\nu_{nk}\exp\left(\phi^d_{kr}+\frac{1}{2}(\sigma^d_{kr})^2\right)\right)\right].$$

$$\frac{\partial^2}{(\partial(\sigma^d_{0r})^2)^2}\mathcal{L}(\mathcal{H},\mathcal{H}_q) \\ = -\frac{1}{2}(\sigma^d_{0r})^{-4} - \frac{1}{4}\sum_{n=1}^{N}\left[\xi_{nd}\exp\left(\phi^d_{0r}+\frac{1}{2}(\sigma^d_{0r})^2\right)\prod_{k=1}^{K}\left(1-\nu_{nk}+\nu_{nk}\exp\left(\phi^d_{kr}+\frac{1}{2}(\sigma^d_{kr})^2\right)\right)\right].$$

Appendix D. Correspondence Between Criteria and Questions in NESARC

Question Code	Personality disorder and criterion
S10Q1A1-S10Q1B7	Avoidant (1 question for each diagnostic criterion)
S10Q1A8-S10Q1B15	Dependent (1 question for each diagnostic criterion)
S10Q1A16-S10Q1B17	OCPD criterion 1
S10Q1A18-S10Q1B23	OCPD criteria 2-7
S10Q1A24-S10Q1B25	OCPD criterion 8
S10Q1A26-S10Q1B29	Paranoid criteria 1-4
S10Q1A30-S10Q1A31	Paranoid criterion 5
S10Q1A32-S10Q1B33	Paranoid criteria 6-7
S10Q1A45-S10Q1B46	Schizoid criterion 1
S10Q1A47-S10Q1B48	Schizoid criteria 2-3
S10Q1A50-S10Q1B50	Schizoid criterion 4
S10Q1A43-S10Q1B43	Schizoid criterion 5
S10Q1A51-S10Q1B52	Schizoid criterion 6
S10Q1A49-S10Q1B49 or S10Q1A53-S10Q1B53	Schizoid criterion 7
S10Q1A54-S10Q1B54 or S10Q1A56-S10Q1B56	Histrionic criterion 1
S10Q1A58-S10Q1B58 or S10Q1A60-S10Q1B60	Histrionic criterion 2
S10Q1A55-S10Q1B55	Histrionic criterion 3
S10Q1A61-S10Q1B61	Histrionic criterion 4
S10Q1A64-S10Q1B64	Histrionic criterion 5
S10Q1A59-S10Q1B59 or S10Q1A62-S10Q1B62	Histrionic criterion 6
S10Q1A63-S10Q1B63	Histrionic criterion 7
S10Q1A57-S10Q1B57	Histrionic criterion 8
S11Q1A20-S11Q1A25	Antisocial, criterion 1
S11Q1A11- S11Q1A13	Antisocial, criterion 2
S11Q1A8- S11Q1A10	Antisocial, criterion 3
S11Q1A17- S11Q1A18	Antisocial, criterion 4
S11Q1A26- S11Q1A33	Antisocial, criterion 4
S11Q1A14- S11Q1A16	Antisocial, criterion 5
S11Q1A6 and S11Q1A19	Antisocial, criterion 6
S11Q8A-B	Antisocial, criterion 7

Table 6: Correspondence between the criteria for each personality disorder and questions in NESARC.

References

M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* Dover Publications, New York, 1972.

D. Aldous. Exchangeability and related topics. In *École d'été de Probabilités de Saint-Flour, XIII—1983*, pages 1–198. Springer, Berlin, 1985.

C. Blanco, R. F. Krueger, D. S. Hasin, S. M. Liu, S. Wang, B. T. Kerridge, T. Saha, and M. Olfson. Mapping common psychiatric disorders: Structure and predictive validity in the National Epidemiologic Survey on Alcohol and Related Conditions. *Journal of the American Medical Association Psychiatry*, 70(2):199–208, 2013.

D. M. Blei and J. D. Lafferty. A correlated topic model of Science. *Annals of Applied Statistics*, 1(1):17–35, August 2007.

G. Bouchard. Efficient bounds for the softmax and applications to approximate inference in hybrid models. *Advances in Neural Information Processing Systems (NIPS), Workshop on Approximate Inference in Hybrid Models*, 2007.

S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, March 2004.

F. Doshi-Velez, K. T. Miller, J. Van Gael, and Y. W. Teh. Variational inference for the Indian buffet process, 2009.

T. L. Griffiths and Z. Ghahramani. The Indian buffet process: an introduction and review. *Journal of Machine Learning Research*, 12:1185–1224, 2011.

D. A. Harville. *Matrix Algebra From a Statistician's Perspective.* Springer-Verlag, 1997.

S. Haykin. *Adaptive Filter Theory.* Prentice Hall, 2002.

M. I. Jordan. Hierarchical models, nested models and completely random measures. In M.-H. Chen, D. Dey, P. Mueller, D. Sun, and K. Ye, editors, *Frontiers of Statistical Decision Making and Bayesian Analysis: In Honor of James O. Berger*. Springer, New York, (NY), 2010.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, November 1999.

K. S. Kendler, C. A. Prescott, J. Myers, and M. C. Neale. The structure of genetic and environmental risk factors for common psychiatric and substance use disorders in men and women. *Archives of General Psychiatry*, 60(9):929–937, Sep 2003.

R. Kotov, C. J. Ruggero, R. F. Krueger, D. Watson, Q. Yuan, and M. Zimmerman. New dimensions in the quantitative classification of mental illness. *Archives of General Psychiatry*, 68(10):1003–1011, 2011.

R. F. Krueger. The structure of common mental disorders. *Archives of General Psychiatry*, 56(10):921–926, 1999.

F. J. R. Ruiz, I. Valera, C. Blanco, and F. Perez-Cruz. Bayesian nonparametric modeling of suicide attempts. *Advances in Neural Information Processing Systems (NIPS)*, 25: 1862–1870, 2012.

Y. W. Teh, D. Görür, and Z. Ghahramani. Stick-breaking construction for the Indian buffet process. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 11, 2007.

M. Titsias. The infinite gamma-Poisson feature model. *Advances in Neural Information Processing Systems (NIPS)*, 19, 2007.

J. M. Valderas, B. Starfield, B. Sibbald, C. Salisbury, and M. Roland. Defining comorbidity: implications for understanding health and health services. *Annals of Family Medicine*, 7 (4):357–363, 2009.

J. Van Gael, Y. W. Teh, and Z. Ghahramani. The infinite factorial hidden Markov model. In *Advances in Neural Information Processing Systems (NIPS)*, volume 21, 2009.

W. A. Vollebergh, J. Ledema, R.V. Bijl, R. de Graaf, F. Smit, and J. Ormel. The structure and stability of common mental disorders: the NEMESIS study. *Archives of General Psychiatry*, 58(6):597–603, Jun 2001.

C. K. I. Williams and D. Barber. Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:1342–1351, 1998.

S. Williamson, C. Wang, K. A. Heller, and D. M. Blei. The IBP compound Dirichlet process and its application to focused topic modeling. In *International Conference on Machine Learning (ICML)*, pages 1151–1158, 2010.

J. L. Wolff, B. Starfield, and G. Anderson. Prevalence, expenditures, and complications of multiple chronic conditions in the elderly. *Archives of Internal Medicine*, 162(20): 2269–2276, 2002.

M. A. Woodbury. The stability of out-input matrices. *Mathematical Reviews*, 1949.