LETTER **Communicated by Pearl Chiu**

Infinite Continuous Feature Model for Psychiatric Comorbidity Analysis

Isabel Valera
ivalera@mpi-sws.org
Max Planck Institute for Software Systems, 67663 Kaiserslautern, Germany

Francisco J. R. Ruiz
franrruiz@columbia.edu
Department of Signal Processing and Communications, University Carlos III in Madrid, 28911 Leganes, Spain; Gregorio Marañón Health Research Institute, 28007 Madrid, Spain; and Department of Computer Science, Columbia University, New York, NY 10027, U.S.A.

Pablo M. Olmos
olmos@tsc.uc3m.es
Department of Signal Processing and Communications, University Carlos III in Madrid, 28911 Leganes, Madrid, and Gregorio Marañón Health Research Institute, 28007 Madrid, Spain

Carlos Blanco
cblanco@nyspi.columbia.edu
Department of Psychiatry, New York State Psychiatric Institute, Columbia University, New York, NY 10032, U.S.A.

Fernando Perez-Cruz
Fernando.Perez-Cruz@Alcatel-Lucent.com
Department of Signal Processing and Communications, University Carlos III in Madrid, 28911 Leganes, Madrid; Gregorio Marañón Health Research Institute, 28007 Madrid, Spain; and Bell Labs, Alcatel-Lucent, New Providence, NJ 07974, U. S. A.

We aim at finding the comorbidity patterns of substance abuse, mood and personality disorders using the diagnoses from the National Epidemiologic Survey on Alcohol and Related Conditions database. To this end, we propose a novel Bayesian nonparametric latent feature model for categorical observations, based on the Indian buffet process, in which the latent variables can take values between 0 and 1. The proposed model has several interesting features for modeling psychiatric disorders. First,

I. Valera and F. J. R. Ruiz contributed equally to this letter.

Neural Computation **28, 354–381** (2016)
doi:10.1162/NECO_a_00805

the latent features might be off, which allows distinguishing between the subjects who suffer a condition and those who do not. Second, the active latent features take positive values, which allows modeling the extent to which the patient has that condition. We also develop a new Markov chain Monte Carlo inference algorithm for our model that makes use of a nested expectation propagation procedure.

1 Introduction

Clinical experience and several studies suggest that some psychiatric disorders may be more closely related to one another as indicated by the frequency of their cooccurrence, which may have etiological and treatment implications. These studies suggest that understanding the underlying interrelationships among psychiatric disorders can be useful for improving the diagnostic classification system and guiding treatment approaches for each disorder (Blanco et al., 2013). Moreover, the disorders are not thought to be on/off diagnostics, but rather manifestations or indicators of underlying continuous variables that represent predispositions to certain types of psychopathology. Motivated by this relevance, in this letter, we aim at finding the latent structure behind a database of psychiatric disorders. In particular, making use of the data from the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC), we focus on the analysis of 20 common psychiatric disorders, including substance use, mood, and personality disorders.[1] Our goal is to find comorbidity patterns in the database, allowing us to seek hidden causes behind the disorders and provide an individual risk characterization for each subject. We develop a tool that can be applied for personalized medicine, since it can detect subjects with higher risk of suffering from psychiatric disorders. Indeed, comorbidity scores are relevant in clinical practice to determine how aggressively to treat a condition.

For that purpose, we rely on latent feature modeling, which allows us to seek hidden causes and compact in a few features the immense redundant information in the observed data. In this context, factor analysis is probably the most commonly used approach for latent feature modeling (Loehlin, 1986). However, as detailed in section 1.1, it has several limitations in the modeling of psychiatric data. In this work, we alternatively propose a Bayesian nonparametric (BNP) latent variable model for categorical observations based on the Indian buffet process (IBP) (Griffiths &

[1]The NESARC database contains the responses of a representative sample of the U.S. population to a survey with questions related to the background of participants, alcohol and other drug consumption, and mental disorders. The first wave of NESARC sampled the adult U.S. population with over 43,000 respondents who answered almost 3,000 questions. Public use data are available for this wave of data collection. See http://aspe.hhs.gov/hsp/06/catalog-ai-an-na/nesarc.htm.

Ghahramani, 2011), but instead of on/off latent features, we extend the IBP model to allow for real-valued latent variables. In our model, latent variables can be interpreted as latent disorders; once a subject has a latent variable or disorder active, its value indicates the severity with what the subject suffers from it. We limit the latent variables to be between 0 and 1, which also helps to interpret the latent variables as a belief in the subject having a latent disorder. A continuous latent variable model allows us to distinguish between two subjects with the same hidden causes. In other words, different suffering levels (severity factors) of a disorder might lead to different treatments. Then, in contrast to a model with on/off latent variables, our model allows understanding the degree of the disorder and opens avenues for personalized medicine approaches.

We propose a spike and slab prior for the severity factors to readily account for subjects who do not have the disorder (spike component) and allows assigning a degree of severity for an active latent feature (slab component). We introduce gaussian weighting matrices to link the latent features and the observations through a multinomial probit likelihood, and we add a bias term, which plays the role of a latent feature that is always active. For a categorical observation space in a latent feature model without bias term, the observations for a subject with no active latent features are assumed independent and equally likely, which does not sound like an appealing outcome in our application. We also force the bias term to model the general population that does not have any latent disorder, which allows directly interpreting the active latent variables as latent features describing disorders. Additionally, this definition for the bias term decreases the runtime complexity of our inference algorithm.

We introduce a novel Markov chain Monte Carlo (MCMC) inference algorithm to sample the latent variables after collapsing the weighting matrices. Due to the nonconjugacy of the likelihood model, the marginal likelihood is approximated using a nested expectation propagation (EP) algorithm (Riihimäki, Jylänki, & Vehtari, 2013). To our knowledge, EP has never been used as a subroutine of an MCMC algorithm despite the fact that it has been proven to be more accurate than other methods, like the Laplace approximation (Cseke & Heskes, 2011; Kuss & Rasmussen, 2005). The proposed nested EP can be efficiently run within the MCMC algorithm because its complexity scales linearly with the number of observations, in contrast to its cubic complexity when applied to gaussian processes for multiclass classification (Riihimäki et al., 2013).

The main contributions of this letter are twofold. First, we develop a new latent variable model based on the IBP that presents the necessary features to analyze psychiatric data and an efficient MCMC inference algorithm. Second, we apply the proposed model to provide a thorough analysis of the relation among psychiatric disorders, obtaining results that are in agreement with previous work and, more important, new insights that would not be possible to obtain without the proposed model.

1.1 Related Work. Similar studies on comorbidity among psychiatric disorders have been carried out by Blanco et al. (2013) and (Ruiz, Valera, Blanco, & Perez-Cruz, 2014). Blanco et al. (2013) resort to factor analysis to study the latent relationship among psychiatric disorders. However, factor analysis has three main limitations when applied to a psychiatric database: the number of factors must be chosen in advance, the observations are assumed to be gaussian distributed, and it provides a nonsparse latent representation. Having to specify the number of factors in advance is a nuisance, especially when little or no prior knowledge about the expected number of factors that best explain the data is available. Moreover, the gaussian assumption cannot properly fit the observations provided by the clinical practitioners because the outcomes are categorical in nature (they indicate whether a subject has a disorder). Finally, since in factor analysis each subject is represented by nonsparse factors, they are more difficult to interpret because the observed disorders are assumed to be influenced by all latent factors for all subjects.

More recently, Ruiz, Valera, Blanco, and Perez-Cruz (2014) applied an IBP for categorical observations, which provides a sparse representation of the latent factors but also allows the number of factors to grow with the number of available observations, avoiding the need to prespecify its value. However, in this model, the latent factors are assumed to be binary, which implies that all the subjects with a latent feature active present with the same severity from the latent disorder. As a consequence, it does not allow personalized medicine approaches, as there is no individual-specific measurement of the degree of severity or risk.

Our model extends the IBP for categorical observations introduced by Ruiz, Valera, Blanco, and Perez-Cruz (2012) and Ruiz et al. (2014). In both works, the authors consider binary-valued latent features and apply an MCMC algorithm that resorts to the Laplace approximation to integrate out the weighting matrices that link the latent features and the observations (through a multinomial-logit likelihood). We show in section 4 that our approach not only provides more interpretable results thanks to the severity factors, but are also more accurate due to the EP approximation.

The combination of the IBP with continuous latent variables has been proposed by Knowles and Ghahramani (2011) for a BNP independent component analysis (ICA). In this model, the prior for the latent continuous variable and IBP matrices are conjugated with a gaussian likelihood, which significantly differs from our proposal.

In this letter, we combine the advantages of factor analysis and the IBP to propose a model that (1) allows for a potentially unbounded number of latent features, which can be interpreted as latent disorders; (2) represents each subject with a sparse latent feature vector whose elements indicate whether a latent disorder is active in a subject and, when active, their values indicate the degree of suffering; and (3) provides interpretable comorbidity patterns among the psychiatric disorders. Moreover, the proposed

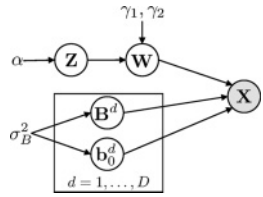


Figure 1: Graphical model.

EP approximation provides more accurate results than the previously used Laplace approximation.

2 Model Description

Latent modeling allows us to seek hidden causes and compact in a few features the immense redundant information in the observed data. The most common nonparametric tool for latent feature modeling is the IBP. The IBP is a prior distribution over binary matrices in which the number of columns (features) K is potentially infinite and the number of nonzero features in each row is distributed as mean-α Poisson (Griffiths & Ghahramani, 2011). Given a finite number of data points N (rows), it ensures that the number of nonzero columns K_+ is finite with probability one. In our application, the rows of the IBP matrix represent subjects, whereas the columns represent latent features, and the matrix represents the features that are active for each particular subject. Let $\mathbf{Z}$ be a random $N \times K$ binary matrix distributed following an IBP; $\mathbf{Z} \sim \text{IBP}(\alpha)$. The nth row of $\mathbf{Z}$, denoted by $\mathbf{z}_{n\cdot}$, represents the vector of latent features of the nth subject, and every entry of $\mathbf{Z}$ is denoted by z_{nk}. Hence, each element $z_{nk} \in \{0, 1\}$ indicates whether the kth feature contributes to the nth data point.

Here, we consider the model in Figure 1, in which $\mathbf{Z} \sim \text{IBP}(\alpha)$ and $\mathbf{X}$ is an $N \times D$ matrix that contains the D-dimensional row observation vectors, denoted by $\mathbf{x}_{n\cdot} = [x_{n1}, \ldots, x_{nD}]$, where $x_{nd} \in \{\mathcal{X}_1, \ldots, \mathcal{X}_R\}$, being $\mathcal{X}_r$ the possible outcomes. In our application at hand, $\mathbf{x}_{n\cdot}$ contains the psychiatric disorders for subject n, and $R = 2$ because disorders can be either present or not present. Note that in other applications, the number of outcomes R may be different for each dimension d, but we drop the dependence on d for notational simplicity.

We additionally propose an $N \times K$ severity matrix $\mathbf{W}$, where each element $w_{nk} \in [0, 1]$ represents how much the nth observation is influenced by the kth latent feature. Thus, w_{nk} indicates how much the kth latent factor influences the psychiatric disorders that the nth subject has. Similar to the model by Knowles and Ghahramani (2011), we propose a spike and slab prior over each w_{nk},

$$p(w_{nk}|\gamma_1, \gamma_2, z_{nk}) = (1 - z_{nk})\delta_0(w_{nk}) + z_{nk}\text{Beta}(w_{nk}|\gamma_1, \gamma_2), \tag{2.1}$$

where $\delta_i(\cdot)$ is the Kronecker delta function (mass point) at i, and γ_1 and γ_2 are hyperparameters of the model. We choose the beta distribution because it enforces the severity variables w_{nk} to be between 0 and 1.

We introduce the $K \times R$ matrices $\mathbf{B}^d$ to model the probability distribution over $\mathbf{X}$, such that $\mathbf{B}^d$ links the latent features with the dth column of the observation matrix $\mathbf{X}$, denoted by $\mathbf{x}_{\cdot d}$, similar to the standard gaussian observation model by Griffiths and Ghahramani (2011). The rth column of matrix $\mathbf{B}^d$ is denoted by $\mathbf{b}^d_{\cdot r}$. We also define the length-R row vectors $\mathbf{b}^d_0$, which model the bias term in the distribution over $\mathbf{x}_{\cdot d}$. This bias term is unnecessary from a modeling point of view, but it simplifies the interpretability of the resulting generative model. Without a bias term, a data point (subject in our application) with no active latent features would present outputs that are independent and uniformly distributed. In our data (and it may be common for other applications), most of the elements in the observation matrix $\mathbf{X}$ indicate the absence of the disorders; they represent the "normal behavior," as most of the subjects do not have any disorder. Hence, we use the bias term to model these subjects; consequently, the resulting active features can be directly interpreted as latent features that deviate from the norm (i.e., latent disorders). We denote by b^d_{0r} the rth element of vector $\mathbf{b}^d_0$. We assume a gaussian prior for the weighting vectors $\mathbf{b}^d_{\cdot r}$ with zero mean and covariance matrix $\mathbf{\Sigma}_b = \sigma_B^2 \mathbf{I}$ and, similarly, $b^d_{0r} \sim \mathcal{N}(b^d_{0r}|0, \sigma_B^2)$.

The probability of each element x_{nd} taking value in the set $\{\mathcal{X}_1, \ldots, \mathcal{X}_R\}$ follows a multinomial probit model (Riihimäki et al., 2013),

$$p(x_{nd} = \mathcal{X}_r | \mathbf{w}_{n\cdot}, \mathbf{B}^d, \mathbf{b}^d_0) = \mathbb{E}_{p(u_{nd})}\left[\prod_{\substack{r'=1 \\ r' \neq r}}^{R} \Phi(u_{nd} + (b^d_{0r} - b^d_{0r'}) + \mathbf{w}_{n\cdot}(\mathbf{b}^d_{\cdot r} - \mathbf{b}^d_{r'}))\right], \tag{2.2}$$

where $\mathbf{w}_{n\cdot}$ stands for the nth row of matrix $\mathbf{W}$, the auxiliary variable u_{nd} is distributed as $p(u_{nd}) = \mathcal{N}(u_{nd}|0, 1)$, $\mathbb{E}_{p(u_{nd})}[\cdot]$ denotes expectation with respect to the distribution $p(u_{nd})$, and Φ denotes the cumulative density function of the standard normal distribution. We use the multinomial probit likelihood because it is amenable to an EP inference algorithm (Girolami & Zhong, 2007).

This model assumes that the observations x_{nd} are independent given the severity matrix $\mathbf{W}$, the weighting matrices $\mathbf{B}^d$, and the vectors $\mathbf{b}^d_0$. Therefore, the likelihood can be factorized as

$$p(\mathbf{X}|\mathbf{W}, \mathbf{B}^1, \ldots, \mathbf{B}^D, \mathbf{b}^1_0, \ldots, \mathbf{b}^D_0) = \prod_{n=1}^{N}\prod_{d=1}^{D} p(x_{nd}|\mathbf{w}_{n\cdot}, \mathbf{B}^d, \mathbf{b}^d_0). \tag{2.3}$$

In our specific application, each observation $x_{nd} \in \{\text{yes}, \text{no}\}$ indicates whether subject n has the disorder d. Under our model, we assume that

the information in the data set $\mathbf{X}$ can be summarized with a smaller set of K latent disorders, such that z_{nk} indicates whether subject n has the latent disorder k and w_{nk} can be interpreted as a belief in the subject's suffering. Additionally, matrices $\mathbf{B}^d$ measure the influence of each latent disorder in the observed disorders, and the bias terms $\mathbf{b}_0^d$ model the population without the latent disorders.

3 Inference

The inference procedure involves obtaining the posterior distribution of matrices $\mathbf{Z}$, $\mathbf{W}$ and $\mathbf{B}^1, \ldots, \mathbf{B}^D$, and vectors $\mathbf{b}_0^1, \ldots, \mathbf{b}_0^D$, which is intractable. We rely on MCMC methods, which have been broadly applied in models involving the IBP (Griffiths and Ghahramani, 2011; Williamson, Wang, Heller, & Blei, 2010), to approximate it. Specifically, we propose an inference algorithm based on Metropolis-Hastings (MH) steps (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953; Hastings, 1970), in which we jointly sample z_{nk} and w_{nk} having marginalized the matrices $\mathbf{B}^d$ and vectors $\mathbf{b}_0^d$. Since the posterior of $\mathbf{B}^d$ and $\mathbf{b}_0^d$ is intractable, we derive a nested EP algortihm (Riihimäki et al., 2013) to approximately integrate out $\mathbf{B}^d$ and $\mathbf{b}_0^d$ in order to obtain the marginal likelihood $p(\mathbf{X}|\mathbf{W})$. The proposed nested EP algorithm avoids both numerical quadratures and independence assumptions among the columns of $\mathbf{B}^d$. We could sample instead from the full joint posterior, but the high dimensionality of our parameter space causes strong dependence among hyperparameters and latent values, resulting in a slow mixing of the Markov chains and hence requiring thousands of posterior draws (Riihimäki et al., 2013).

Our algorithm proceeds iteratively as follows. For each observation $n = 1, \ldots, N$:

- Step 1: Jointly sample z_{nk} and w_{nk} for $k = 1, \ldots, K_+$, where K_+ is the number of active latent features.
- Step 2: Consider adding new latent features for the nth observation, updating K_+ if necessary.

For conciseness, we drop the dependence of the hyperparameters in the notation throughout the rest of the letter.

In step 1, we rely on MH proposing to move from an initial pair (z_{nk}, w_{nk}) to (z_{nk}^*, w_{nk}^*) (jumping from matrices $\mathbf{Z}$ and $\mathbf{W}$ to $\mathbf{Z}^*$ and $\mathbf{W}^*$). Our proposal distribution is

$$q_1(z_{nk}^*, w_{nk}^* | z_{nk}, w_{nk}) = \begin{cases} \delta_1(z_{nk}^*) p(w_{nk}^* | z_{nk}^* = 1), & \text{if } z_{nk} = 0, \\ \frac{1}{2}\delta_0(z_{nk}^*)\delta_0(w_{nk}^*) + \frac{1}{2}\delta_1(z_{nk}^*) p(w_{nk}^* | z_{nk}^* = 1), & \text{if } z_{nk} \neq 0, \end{cases} \tag{3.1}$$

that is, if $z_{nk} = 0$ we propose to move to $z^*_{nk} = 1$ with w^*_{nk} sampled from equation 2.1. Otherwise, either a move to $z^*_{nk} = 0$ or to $z^*_{nk} = 1$ (with a value of w^*_{nk} drawn from the prior) is proposed with equal probability. The acceptance probability for the MH step is given by

$$\min\left(1, \frac{p(\mathbf{X}|\mathbf{W}^*)p([\mathbf{Z}^*])p(w^*_{nk}|z^*_{nk})}{p(\mathbf{X}|\mathbf{W})p([\mathbf{Z}])p(w_{nk}|z_{nk})} \frac{q_1(z_{nk}, w_{nk}|z^*_{nk}, w^*_{nk})}{q_1(z^*_{nk}, w^*_{nk}|z_{nk}, w_{nk})}\right), \tag{3.2}$$

where

$$\frac{p([\mathbf{Z}^*])}{p([\mathbf{Z}])} = \begin{cases} 1, & \text{if } z_{nk} = z^*_{nk}, \\ m_{\neg kn}/(N - m_{\neg kn}), & \text{if } z^*_{nk} = 1, z_{nk} = 0, \\ (N - m_{\neg kn})/m_{\neg kn}, & \text{if } z^*_{nk} = 0, z_{nk} = 1, \end{cases} \tag{3.3}$$

being $m_{\neg kn}$ the number of data points (excluding n) that have an active kth feature, namely, $m_{\neg kn} = \sum_{n' \neq n} z_{n'k}$. The distribution $p(w_{nk}|z_{nk})$ is given in equation 2.1 and, as previously stated, the probabilities $p(\mathbf{X}|\mathbf{W})$ are obtained using the nested EP algorithm detailed in the appendix, where it is also shown that the nested EP presents linear complexity with the number of observations.

For step 2, we need to define κ_n as the number of columns of $\mathbf{Z}$ that are active only in the nth row: $\kappa_n = \sum_{k=1}^{\infty} z_{nk} \prod_{n' \neq n}(1 - z_{n'k})$. Note that, after performing step 1, the initial value of κ_n is 0 due to the form of equations 3.2 and 3.3. The new value κ^*_n is sampled with an MH step. We include as part of the proposal the corresponding new values of the severity matrix, a $1 \times \kappa^*_n$ vector denoted by $\boldsymbol{\omega}^*_n$. Therefore, we propose to jump from an initial value of κ_n and $\boldsymbol{\omega}_n$ to κ^*_n and $\boldsymbol{\omega}^*_n$, where the latter variables are drawn from the proposal distribution,

$$q_2(\kappa^*_n, \boldsymbol{\omega}^*_n) = q_2(\kappa^*_n) q_2(\boldsymbol{\omega}^*_n|\kappa^*_n). \tag{3.4}$$

We make $q_2(\boldsymbol{\omega}^*_n|\kappa^*_n)$ equal to the prior, $q_2(\boldsymbol{\omega}^*_n|\kappa^*_n) = \prod_{k'=1}^{\kappa^*_n} p(\omega^*_{nk'}|z^*_{nk'} = 1)$, and $q_2(\kappa^*_n)$ is chosen following Knowles and Ghahramani (2011):

$$q_2(\kappa^*_n) = (1 - \pi)\text{Poisson}\left(\kappa^*_n|\alpha\lambda/N\right) + \pi\delta_1(\kappa^*_n), \tag{3.5}$$

where we set $\lambda = N/2$ and $\pi = 0.2$. The move in step 2 is accepted with probability

$$\min\left(1, \frac{p(\mathbf{X}|\mathbf{W}^*)}{p(\mathbf{X}|\mathbf{W})} \frac{(\alpha/N)^{\kappa^*_n}}{\kappa^*_n!} \frac{q_2(\kappa_n)}{q_2(\kappa^*_n)}\right). \tag{3.6}$$

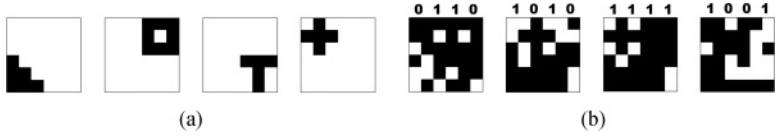


Figure 2: Toy example 1. (a) Base images. (b) Four observation examples. The numbers above each figure indicate which features are present in that image.

4 Experiments

4.1 Experiments on Synthetic Data. In order to evaluate the performance of our model and inference algorithm, we generate two synthetic data sets and perform comparisons between a latent feature model with (1) on/off hidden variables and inference based on Gibbs sampling and the Laplace approximation (denoted by "On/Off+Lap.") (Ruiz et al., 2012); (2) on/off hidden variables and inference based on Gibbs sampling and the nested EP approximation described in section 4.2 ("On/Off+EP"); and (3) continuous hidden variables in $[0, 1]$ and inference based on MH steps and the nested EP approximation—the algorithm in section 3 ("Sev.+EP"). In the three cases, we remove the bias term because it is not needed for the synthetic data sets considered.

We generate binary-valued observation matrices $\mathbf{X}$, with $N = 100$ black-and-white images with dimensionality $D = 36$, that are built differently for each of the two data sets.

Toy example 1 replicates the synthetic experiment by Ruiz et al. (2012), in which each observation $\mathbf{x}_{n\cdot}$ is a combination of four latent black-and-white base images that can be present or absent with probability 0.5 independent of each other: $z_{nk} = 1$ with probability 0.5. Each white pixel in the composite image becomes black with probability 0.5, while black pixels remain black. We plot in Figure 2 the four base images and four observation examples.

Toy example 2 is similar to example 1, but we introduce a latent auxiliary matrix $\mathbf{A}$ to generate observations. As before, we assume four latent features that become active with probability 0.5, but we also generate a $N \times 4$ matrix $\mathbf{A}$, whose elements a_{nk} are beta$(2, 1)$ distributed. In this setup, we divide each image into four disjoint regions of nine pixels, each modeled by one of the latent features. Each of the nine pixels in observation n corresponding to feature k is set to black with probability $0.5 + 0.5a_{nk}$ if $z_{nk} = 1$, or with probability 0.5 otherwise.

In order to compare the three methods, we average over five independent realizations of the two synthetic data sets the following scores:

- Approximate marginal log-likelihood $p(\mathbf{X}|\mathbf{Z})$ or $p(\mathbf{X}|\mathbf{W})$ (Log-lik).
- Kullback-Leibler divergence (D_{KL}) between the true and the inferred probability of the observation matrix, that is, between the inferred

Table 1: Results for Toy Example 1.

	On/Off+Laplace	On/Off+EP	Sev.+EP
Log-lik	−1,943	−2,001	−1,948
D_{KL}	497.15	354.92	347.11

Table 2: Results for Toy Example 2.

	On/Off+Laplace	On/Off+EP	Sev.+EP
Log-lik	−2,122	−2,233	−2,151
D_{KL}	524.16	372.10	353.15

probability of the observations and the underlying true generative process described above. We compute the inferred probability using the mean of the approximate posterior of $\mathbf{B}^d$ and the sample of the latent feature matrix $\mathbf{Z}$ (or $\mathbf{W}$, if available).

In Tables 1 and 2, we show the results for the two synthetic data sets. Note that the obtained values of the average log likelihood are similar for the three considered methods (no significant statistical differences are found) in both examples. However, we can observe significant differences in terms of the Kullback-Leibler divergence, for which the model with severity factors combined with the EP inference provides the best results in both examples. In example 1, since the generative model considers binary latent variables (instead of continuous), both the "On/Off+EP" and the "Sev.+EP" methods provide similar results.

We have shown that the model with severity factors (combined with EP inference) better captures the true underlying probability of the observations given the latent variables. This improvement in the performance becomes more relevant in real applications such as the analysis in next section, where we are interested in studying the probability of subjects suffering from psychiatric disorders.

4.2 Experiments on Real Data

4.2.1 Database Description. The NESARC is a survey designed and conducted by the National Institute on Alcohol Abuse and Alcoholism (NIAAA) that samples the adult U.S. population; it has over 43,000 subjects. The NESARC includes almost 3000 questions not only on alcohol use and abuse, but also on a wide range of physical and psychiatric disorders, as well as significant background of its participants.

Table 3: Empirical Probabilities of Possessing at Least One Latent Feature, Extracted Directly from the Inferred IBP Matrix $\mathbf{Z}$.

Active Feature	Feature 1	Feature 2	Feature 3
Empirical Probability	0.0594	0.0239	0.0201

Based on information collected in the first wave of the NESARC, a set of preestablished and reliable diagnostic algorithms was applied to each subject to determine the presence or absence of 20 psychiatric disorders (Blanco et al., 2013). These 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]).

In this study, we apply the model in section 2 taking the diagnoses of the 20 psychiatric disorders for all the subjects in the NESARC as input data. This study provides an alternative to the factor analysis approach by Blanco et al. (2013) and to the IBP with categorical observations by Ruiz et al. (2014), who use the Laplace approximation to obtain the marginal likelihood. We compare our approach to their methods to show that our model not only provides results in agreement with previous studies, but also more interpretable results that allow us to obtain new insights on the comorbidity patterns among psychiatric disorders.

4.2.2 Experimental Setup. For the following experimental results, we set $\alpha = 1$, $\sigma_B^2 = 1$, $\gamma_1 = 2$, and $\gamma_2 = 1$ and run our inference algorithm. In order to speed up the inference procedure, we do not sample the rows of $\mathbf{W}$ corresponding to subjects who suffer from at most 1 out of the 20 disorders but instead fix these latent features to zero. The idea is that the $\mathbf{b}_0^d$ terms must capture the general population, and we use the active components of the matrix $\mathbf{W}$ to characterize the disorders. Besides speeding up the algorithm, this modification ensures that the active latent features increase the probability of suffering from the disorders and can be interpreted as latent disorders, which helps interpreting the obtained results.

4.2.3 Results. Similar to previous studies (Blanco et al., 2013; Ruiz et al., 2014), we find that we need three latent features to describe the data. In Table 3, we show the empirical probability of possessing each of the inferred latent features: the number of subjects in the database who possess each latent feature divided by the total number of subjects. Additionally, we plot

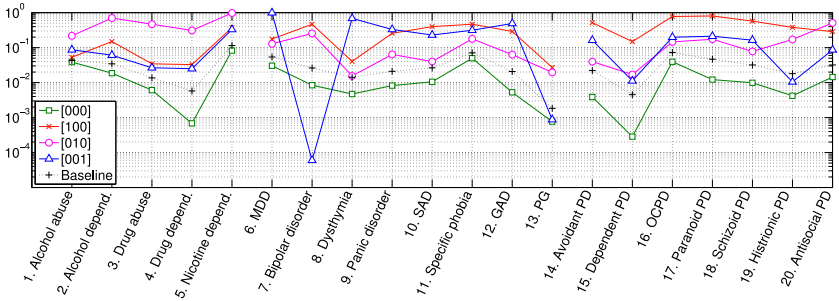


Figure 3: Probabilities of suffering from the 20 considered disorders for the latent feature vectors $\mathbf{w}_{n\cdot}$ shown in the legend. These probabilities have been obtained using the mean of the approximate posterior of the matrices $\mathbf{B}^d$.

in Figure 3 the approximate posterior probability of suffering from each of the considered disorders when only one of the latent features is active (assuming severity factors equal to one) and when none of them is active. As expected, for subjects with no active latent feature, the probability of having any disorder is below the baseline level (defined as the empirical probability of having each disorder in the full database).

We can interpret each of the obtained latent features from the analysis of Figure 3. Feature 1 (pattern [100]) increases the probability of having all disorders, except alcohol abuse, and thus seems to represent a general psychopathology factor, although it may particularly increase the risk of personality disorders (disorders 14–20). Feature 2 (pattern [010]) models substance use disorders and antisocial personality disorder, which is consistent with the externalizing factor identified in previous studies of the structure of psychiatric disorders (Krueger, 1999; Kendler, Prescott, Myers, & Neale, 2003; Vollebergh et al., 2001; Blanco et al., 2013). Feature 3 (pattern [001]) models mood or anxiety disorders and thus seems to represent the internalizing factor also identified in previous studies. Note that the probability of bipolar disorder presents a significantly different behavior, since major depression (MDD) and dysthymia are mutually exclusive with bipolar disorder.

In addition to the hidden relation among the disorders, our model also provides an individual-specific severity term that can be interpreted as our belief that the subject is suffering a latent disorder. We find that more than 80% of the subjects with active features have a severity factor above 0.5, and around 50% of them have a severity value greater that 0.75. The histograms for w_{n1}, w_{n2}, and w_{n3} are shown in Figure 4.

Additionally, we plot in Figure 5 the posterior probability of suffering from each of the disorders when only feature 1 is active, for any value of the severity w_{n1}. (Similar plots, for features 2 and 3, are provided in Figures 6

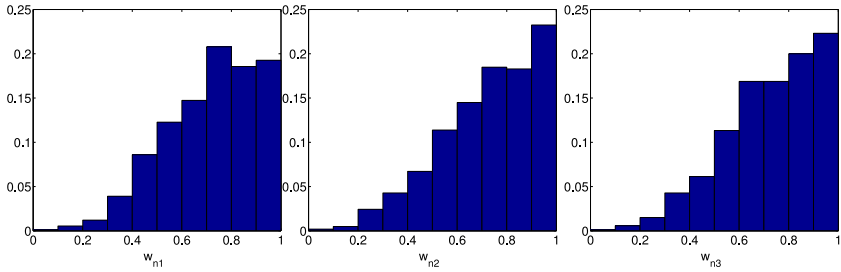


Figure 4: Normalized histograms of w_{n1}, w_{n2}, and w_{n3} (assuming that $z_{n1} = 1$, $z_{n2} = 1$ and $z_{n3} = 1$, respectively).

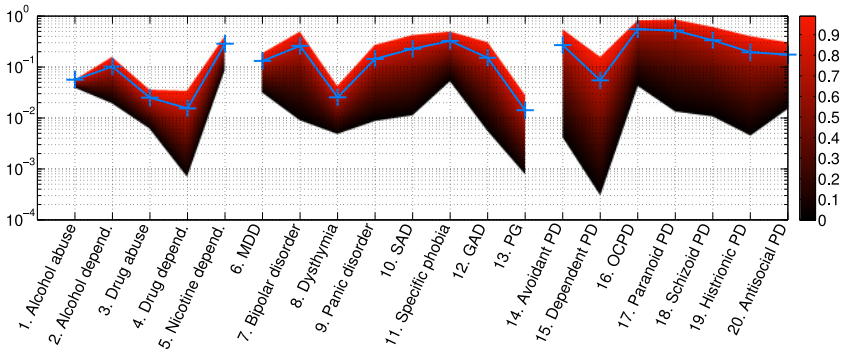


Figure 5: Probabilities of suffering from the 20 considered disorders when only feature 1 is active for any value of the severity w_{n1} (shown in the bar on the right). These probabilities have been obtained using the mean of the approximate posterior of the matrices $\mathbf{B}^d$. The solid line represents the empirical probabilities, obtained for subjects who have only feature 1 active.

and 7, respectively.) When the severity reaches 0 (depicted in black), feature 1 turns inactive, and therefore the corresponding probabilities coincide with the green line in Figure 3 (pattern [000]). As the severity approaches 1 (depicted in red), the corresponding probabilities coincide with the red line in Figure 3 (pattern [100]). The solid line in Figure 5 represents the empirical probability of suffering from each disorder, obtained for subjects who have only feature 1 active. We can see that although the probability of suffering from each disorder becomes higher when the inferred severity value increases, each disorder is affected differently by the value of the severity factor. For instance, the probability of suffering from OCPD goes from 0.04 in the general population to 0.8 for the subjects with a severity

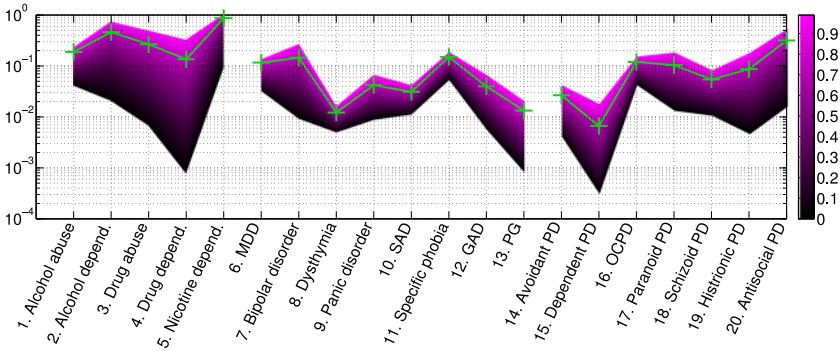


Figure 6: Probabilities of suffering from the 20 considered disorders when only feature 2 is active for any value of the severity w_{n2} (shown in the bar on the right). These probabilities have been obtained using the mean of the approximate posterior of the matrices $\mathbf{B}^d$. The solid line represents the empirical probabilities, obtained for subjects who have only feature 2 active.

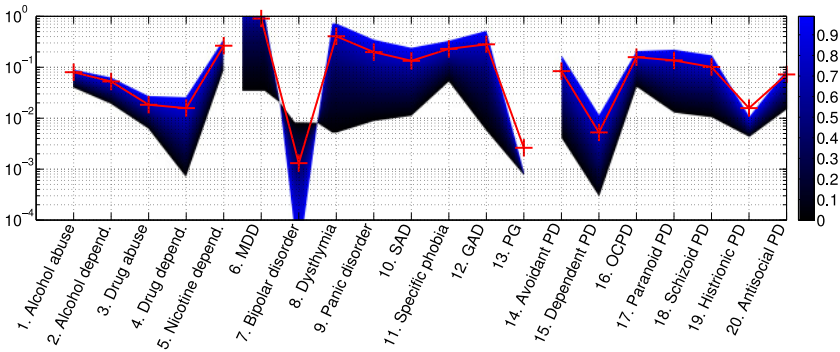
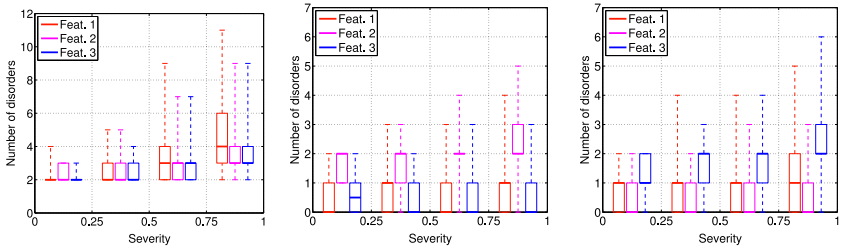


Figure 7: Probabilities of suffering from the 20 considered disorders when only feature 3 is active for any value of the severity w_{n3} (shown in the bar on the right). These probabilities have been obtained using the mean of the approximate posterior of the matrices $\mathbf{B}^d$. The solid line represents the empirical probabilities, obtained for subjects who have only feature 3 active.

factor for feature 1 near to one, while the probability for alcohol abuse changes from only 0.04 to 0.05.

To further analyze the impact of severity, we depict in Figure 8 the distribution of the number of disorders for subjects whose inferred severity is between the numbers shown in the horizontal axis. As expected, as the inferred severity increases, so does the number of disorders that a subject suffers. Figure 8a shows that feature 1 (general psychopathology factor) is



(a) 20 disorders.

(b) Externalizing disorders (disorders 1 to 5 in Fig. 3 and antisocial PD).

(c) Internalizing disorders (disorders 6 and 8-12 in Fig. 3).

Figure 8: Distribution of the number of disorders for subjects who have only active one latent feature (shown in legend), whose inferred severity is between the numbers shown in the horizontal axis. The thick line corresponds to the median, the edges of the box are the 25th and 75th percentiles, and the whiskers represent the most extreme values.

the feature with the highest impact on the average number of disorders. However, when we consider only a subset of the disorders (see Figures 8b and 8c), features 2 and 3 become more relevant. These subsets have been chosen to match the externalizing and internalizing factors, respectively. Hence, we can interpret each latent feature as a latent disorder (that groups several observed disorders) and its related severity factor as a belief in the suffering from this disorder. As a consequence, subjects with higher suffering of a latent disorder tend to suffer from several comorbid disorder in the same group of disorders.

4.2.4 Comparisons with Previous Approaches. In order to examine the effect of the severity in our IBP-based model, we plot in Figure 9 the results reported by Ruiz et al. (2014) for the probability of suffering from each disorder, denoted as "on/off" in the figure. When we compare these probabilities with our results in Figure 3 (which are replicated in Figure 9), we can observe that the inferred probabilities for our continuous latent feature model are more extreme than for the binary latent feature model. This is due to the fact that our model includes the individual-specific severity terms $\mathbf{w}_{n\cdot}$, which allow weighting the contribution of each latent feature specifically for each subject in the database, and therefore when we set $w_{nk} = 1$ as in Figure 9, we obtain the most severe contribution of latent feature k. The binary latent feature model by Ruiz et al. (2014) cannot capture this variation across the subjects, and hence the probability curves tend to be adjusted so as to explain all subjects having feature k active, regardless of their severity level. In other words, including the continuous latent variables (or severity terms) allows a wider dynamical range for the probability

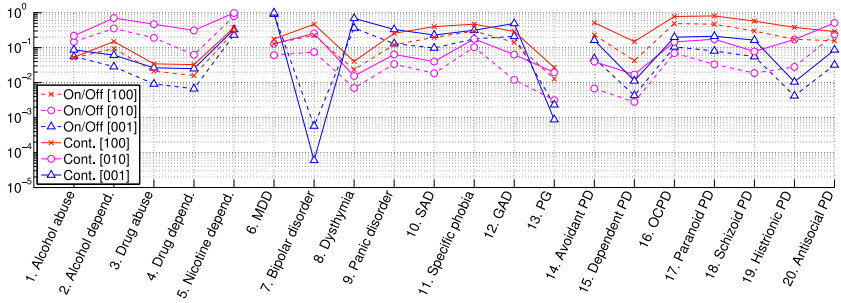


Figure 9: Comparison of IBP-based approaches. Probabilities of suffering from the 20 considered disorders for the latent feature vectors are shown in the legend for both the binary latent feature model (On/Off) and our continuous latent feature model (Cont.).

of suffering each disorder, being that dynamical range modulated by w_{nk}. If we had only on/off latent variables, we could not distinguish between subjects who suffer only minor disorders (lower probability or fewer of them) to those suffering major disorders (higher probability or many of them). Considering continuous values for the latent variables does not change the role of the three latent variables; the conclusions for the general population remain unchanged. However, it allows providing an individual characterization for each subject (see, e.g., Figure 8) that the previous model with on/off latent variables could not provide, increasing the applicability of the proposed model for personalized medicine.

Now we compare our results to those reported by Blanco et al. (2013), who apply factor analysis to study the comorbidity patterns among psychiatric disorders. In factor analysis, the observation matrix $\mathbf{X}$ can be expressed as $\mathbf{X} = \mathbf{LF} + \boldsymbol{\epsilon}$, where $\mathbf{L}$ is an $N \times K$ matrix ($K \ll N$), $\mathbf{F}$ is a $K \times D$ matrix, and $\boldsymbol{\epsilon}$ is additive gaussian noise. Therefore, factor analysis assumes that the observation matrix is gaussian, although the observations are actually of a categorical nature (a subject may either suffer from a disorder or not suffer). We replicate the experiments by Blanco et al. (2013) assuming $K = 3$ factors and plot the obtained factors in Figure 10. In this figure, we observe that although the obtained factors allow for a similar interpretation than under our IBP-based results in Figure 3, factor analysis does not provide a probabilistic interpretation of the latent factors, since they are not bounded.

Additionally, we plot in Figure 11 a normalized histogram for each column (factor) of the obtained matrix $\mathbf{L}$. Since factor analysis cannot provide a sparse representation of the data set, the resulting matrix $\mathbf{L}$ is dense, which implies that we cannot distinguish among subjects with or without active latent factors. As a consequence, in the histograms, we observe a peak around zero, and only around 23% of the subjects present absolute values

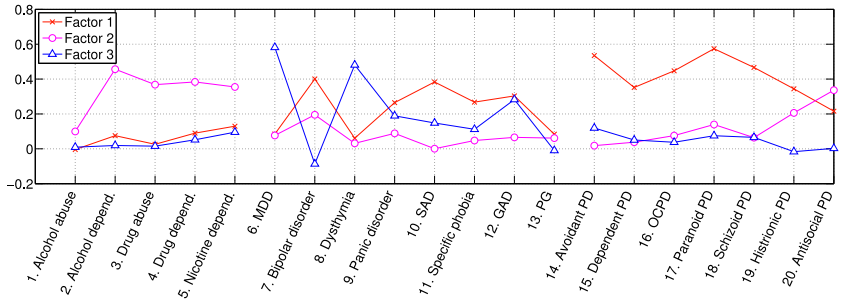


Figure 10: Representation of the factors **F** obtained with the factor analysis approach.

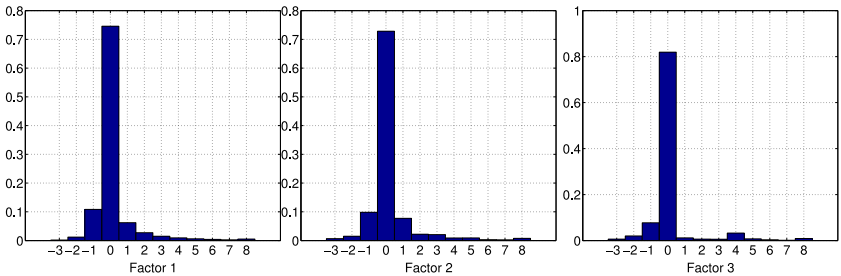
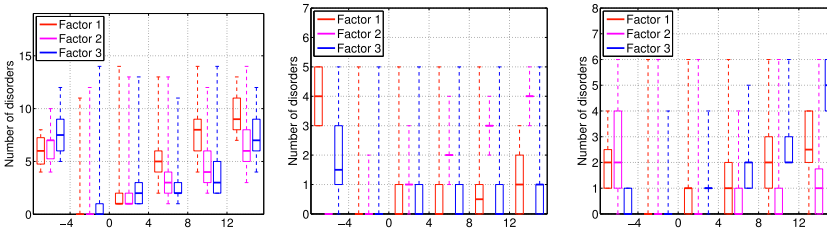


Figure 11: Normalized histograms of the factor values in **L** for the factor analysis approach.

greater than 0.5. We additionally plot in Figure 12 the average number of disorders that subjects with different values for the latent factors suffer from. In contrast to Figure 8, in this plot we represent all the subjects, since we cannot restrict our analysis to subjects who have only one active latent factor. This explains the behavior in Figure 12 when the factor values drop below -4. Because the latent factors are not bounded, they cannot be easily interpreted as a belief in the suffering from latent disorders.

4.2.5 Summary and Interpretation of the Results. Our model provides results that are consistent with previous studies on the latent structure of psychiatric disorders but also provide new insights. We find that the comorbidity patterns of common psychiatric disorders can be described by a small number of latent features, even though the model has enough a priori flexibility to account for a potentially unbounded number of features. In addition, nosologically related disorders, such as social anxiety disorder and avoidant personality disorder, tend to be modeled by similar



(a) 20 disorders.

(b) Externalizing disorders (disorders 1 to 5 in Fig. 3 and antisocial PD).

(c) Internalizing disorders (disorders 6 and 8-12 in Fig. 3).

Figure 12: Distribution of the number of disorders for the factor analysis approach for subjects who have latent factor values between the numbers shown on the horizontal axis. The thick line corresponds to the median, the edges of the box are the 25th and 75th percentiles, and the whiskers represent the most extreme values.

features. We also find that no disorder is perfectly aligned along a single latent feature, which suggests that disorders can develop through multiple etiological paths. For instance, the risk of nicotine dependence may be high in individuals with a propensity to externalization or internalization, as Blanco et al. (2013) suggested.

From the analysis of the figures, we can conclude that the probability of appearance of a disorder changes significantly when the value of the severity associated with that group of disorders changes. We also find that most of the subjects with active latent features suffer from three or more disorders, and in general, most of the disorders that a subject suffers belong to the group of disorders modeled by the same latent feature. Therefore, a subject with feature 2 (feature 3) active has a higher probability of suffering simultaneously from several externalizing (internalizing) disorders, especially if the corresponding severity value is high. In this way, we can understand the importance of the severity factors in the model, because they allow explaining the comorbidity among the disorders and also understanding the stress each subject suffers.

5 Conclusion

In this letter, we have proposed a new model that combines the IBP prior with continuous-valued severity factors to characterize categorical observations using the multinomial probit likelihood, and we have derived a nested EP approximation to integrate out the weighting factors, which allows us to efficiently run an MCMC sampler. The proposed model has several relevant properties that stand out when compared with previous approaches: (1) it addresses the modeling of categorical observations without assuming that

they are gaussian distributed; (2) it provides a sparse latent representation since the latent variables can be active or inactive; (3) the active latent variables are continuous valued, allowing us to distinguish different subjects that have the same active latent features; and (4) we do not need to prespecify the number of latent variables in advance, as the inference procedure can find it.

We have applied our model to the NESARC database to find the hidden features that characterize 20 common psychiatric disorders, finding that three latent features capture the comorbidity patterns. Hence, we have shown that the approach provides results over the NESARC database that concur with previous work and provide new information, such as the individual-specific severity terms for each latent feature. The severity terms can help to more deeply analyze the comorbidity patterns among the disorders and detect subjects with a higher risk of suffering from one or several disorders. Although our analysis focuses only on psychiatric disorders, the model is general in the sense that it can be used as a latent variable model for other applications with categorical data.

Appendix: Nested EP

In this section, we adapt the nested EP algorithm introduced by (Riihimäki et al., 2013) to approximate the marginal likelihood $p(\mathbf{X}|\mathbf{W})$. The proposed model assumes that the observations are independent given $\mathbf{W}$ and the weighting matrices and vectors. Then the posterior $p(\mathbf{B}^1, \ldots, \mathbf{B}^D|\mathbf{X}, \mathbf{W})$ factorizes as[2]

$$p(\mathbf{B}^1, \ldots, \mathbf{B}^D|\mathbf{X}, \mathbf{W}) = \prod_{d=1}^{D} \frac{p(\mathbf{B}^d)p(\mathbf{x}_{\cdot d}|\mathbf{W}, \mathbf{B}^d)}{p(\mathbf{x}_{\cdot d}|\mathbf{W})}. \tag{A.1}$$

The computation of the marginal likelihood, $p(\mathbf{x}_{\cdot d}|\mathbf{W}) = \int p(\mathbf{B}^d)p(\mathbf{x}_{\cdot d}|\mathbf{W}, \mathbf{B}^d)d\mathbf{B}^d$, is intractable, because the prior and likelihood are not conjugate. Therefore, we run D parallel nested EP algorithms to compute $p(\mathbf{x}_{\cdot d}|\mathbf{W})$, and the marginal likelihood $p(\mathbf{X}|\mathbf{W})$ is the product of the individual terms $p(\mathbf{x}_{\cdot d}|\mathbf{W})$ for $d = 1, \ldots, D$. In the description of the nested EP algorithm, we do not make explicit the dependence on d, unless necessary, to avoid cluttering of notation.

Besides EP, we could also approximate this posterior using multidimensional quadratures (Seeger & Jordan, 2004) or the Laplace approximation (Girolami & Zhong, 2007). We choose the nested EP algorithm because EP approaches are typically more accurate than the Laplace approximation and

[2] In what follows, the bias term is incorporated in $\mathbf{W}$ and $\mathbf{B}^d$; $\mathbf{W}$ stands for $[\mathbf{1}\ \mathbf{W}]$ and $\mathbf{B}^d$ denotes $[(\mathbf{b}_0^d)^\top\ (\mathbf{B}^d)^\top]^\top$.

computationally less demanding than numerical quadratures (Riihimäki et al., 2013). The nested EP consists of two loops, which we describe; they are summarized in algorithms 1 and 2. We also show in this section that the complexity of the nested EP is linear in the number of observations.

For convenience, we stack the columns of $\mathbf{B}^d$ into the vector $\boldsymbol{\beta}^d$—$\boldsymbol{\beta}^d = \mathbf{B}^d(:)$ in Matlab notation. Note that given $\mathbf{W}$, we need only to account for the parameters corresponding to the K_+ active features. To obtain the marginal likelihood, we need to approximate the posterior $p(\boldsymbol{\beta}^d|\mathbf{x}_{\cdot d}, \mathbf{W})$ with a tractable distribution. The likelihood $p(\mathbf{x}_{\cdot d}|\mathbf{W}, \boldsymbol{\beta}^d)$ contains a product of nonconjugate terms (sites) (Seeger, 2008), denoted by $t_n^d(\boldsymbol{\beta}^d) = p(x_{nd}|\mathbf{W}, \boldsymbol{\beta}^d)$, and hence the posterior can be expressed as

$$p(\boldsymbol{\beta}^d|\mathbf{x}_{\cdot d}, \mathbf{W}) = \frac{\mathcal{N}(\boldsymbol{\beta}^d|\mathbf{0}, \sigma_B^2\mathbf{I}) \prod_{n=1}^N t_n^d(\boldsymbol{\beta}^d)}{p(\mathbf{x}_{\cdot d}|\mathbf{W})}. \tag{A.2}$$

The EP approximation consists of replacing each site $t_n^d(\boldsymbol{\beta}^d)$ with a tractable term $\widetilde{t}_n^d(\boldsymbol{\beta}^d)$, resulting in an approximate distribution that we denote by $q_{\mathrm{EP}}(\boldsymbol{\beta}^d)$. We choose $\widetilde{t}_n^d(\boldsymbol{\beta}^d)$ to be a scaled gaussian with the $R(K_+ + 1) \times 1$ vector $\widetilde{\boldsymbol{\lambda}}_n$ and the $R(K_+ + 1) \times R(K_+ + 1)$ matrix $\widetilde{\boldsymbol{\Pi}}_n$ as natural parameters, and scaling constant $\widetilde{Z}_n$, $\widetilde{t}_n^d(\boldsymbol{\beta}^d) = \widetilde{Z}_n \mathcal{N}(\boldsymbol{\beta}^d|\widetilde{\boldsymbol{\Pi}}_n^{-1}\widetilde{\boldsymbol{\lambda}}_n, \widetilde{\boldsymbol{\Pi}}_n^{-1})$, yielding

$$\begin{aligned} q_{\mathrm{EP}}(\boldsymbol{\beta}^d) &= \mathcal{N}(\boldsymbol{\beta}^d|\boldsymbol{\Pi}_{\mathrm{EP}}^{-1}\boldsymbol{\lambda}_{\mathrm{EP}}, \boldsymbol{\Pi}_{\mathrm{EP}}^{-1}) \\ &= \frac{1}{Z_{\mathrm{EP}}} \mathcal{N}(\boldsymbol{\beta}^d|\mathbf{0}, \sigma_B^2\mathbf{I}) \prod_{n=1}^N \widetilde{Z}_n \mathcal{N}(\boldsymbol{\beta}^d|\widetilde{\boldsymbol{\Pi}}_n^{-1}\widetilde{\boldsymbol{\lambda}}_n, \widetilde{\boldsymbol{\Pi}}_n^{-1}), \end{aligned} \tag{A.3}$$

where $\boldsymbol{\lambda}_{\mathrm{EP}}$ and $\boldsymbol{\Pi}_{\mathrm{EP}}$ are the natural parameters of the gaussian distribution $q_{\mathrm{EP}}(\boldsymbol{\beta}^d)$. We choose $\widetilde{Z}_n$ following Seeger (2005) in order for Z_{EP} to become a good approximation of the marginal likelihood $p(\mathbf{x}_{\cdot d}|\mathbf{W})$.

EP chooses the parameters $\widetilde{\boldsymbol{\lambda}}_n$ and $\widetilde{\boldsymbol{\Pi}}_n$ by matching the moments of $p(\boldsymbol{\beta}^d|\mathbf{x}_{\cdot d}, \mathbf{W})$ and $q_{\mathrm{EP}}(\boldsymbol{\beta}^d)$, which is equivalent to minimizing the Kullback-Leibler divergence $D_{\mathrm{KL}}(p(\boldsymbol{\beta}^d|\mathbf{x}_{\cdot d}, \mathbf{W})||q_{\mathrm{EP}}(\boldsymbol{\beta}^d))$. This minimization is solved iteratively for $n = 1, \ldots, N$ (Minka, 2001; Seeger, 2008; Opper & Winther, 2005) (repeating until convergence) as follows:

1. Define the cavity distribution $q_{\neg n}(\boldsymbol{\beta}^d) \propto q_{\mathrm{EP}}(\boldsymbol{\beta}^d)/\widetilde{t}_n^d(\boldsymbol{\beta}^d)$, in which we have removed one approximate site. The natural parameters of the cavity distribution are $\boldsymbol{\Pi}_{\neg n} = \boldsymbol{\Pi}_{\mathrm{EP}} - \widetilde{\boldsymbol{\Pi}}_n$ and $\boldsymbol{\lambda}_{\neg n} = \boldsymbol{\lambda}_{\mathrm{EP}} - \widetilde{\boldsymbol{\lambda}}_n$.
2. Define the tilted distribution $\widehat{p}_n(\boldsymbol{\beta}^d) \propto q_{\neg n}(\boldsymbol{\beta}^d) t_n^d(\boldsymbol{\beta}^d)$ (which includes the true site) and minimize $D_{\mathrm{KL}}(\widehat{p}_n(\boldsymbol{\beta}^d)||q_{\mathrm{EP}}(\boldsymbol{\beta}^d))$ with respect to $q_{\mathrm{EP}}(\boldsymbol{\beta}^d)$.
3. Update the approximate site as $\widetilde{t}_n^d(\boldsymbol{\beta}^d) \propto q_{\mathrm{EP}}(\boldsymbol{\beta}^d)/q_{\neg n}(\boldsymbol{\beta}^d)$.

The standard EP algorithm solves step 3 by matching the moments between $\widehat{p}_n(\boldsymbol{\beta}^d)$ and $q_{\text{EP}}(\boldsymbol{\beta}^d)$, which is assumed to be tractable, but in this case, matching these moments is not tractable and we resort to another EP loop, the inner loop, and hence the name of the algorithm. The inner loop of the nested EP, detailed below, approximates the tilted distribution

$$\widehat{p}_n(\boldsymbol{\beta}^d) = \frac{1}{\widehat{Z}_n} q_{\neg n}(\boldsymbol{\beta}^d) t_n^d(\boldsymbol{\beta}^d) \tag{A.4}$$

by a gaussian distribution with natural parameters $\widehat{\boldsymbol{\lambda}}_n$ and $\widehat{\boldsymbol{\Pi}}_n$, which is similar to the EP algorithm resulting from a linear binary classifier with a multivariate gaussian prior and a probit likelihood function in the gaussian process setting (Qi, Minka, Picard, & Ghahramani, 2004). Now step 3 follows readily, since we can obtain the new natural parameters for the approximate site $\widetilde{t}_n^d(\boldsymbol{\beta}^d)$ as $\widetilde{\boldsymbol{\Pi}}_n^{new} = \widehat{\boldsymbol{\Pi}}_n - \boldsymbol{\Pi}_{\neg n}$ and $\widetilde{\boldsymbol{\lambda}}_n^{new} = \widehat{\boldsymbol{\lambda}}_n - \boldsymbol{\lambda}_{\neg n}$. A damping factor $\eta_{\text{O}} \in (0, 1]$ can be used in this step for numerical stability (Seeger, 2005).

The site parameters $\widetilde{t}_n^d(\boldsymbol{\beta}^d)$ can be updated in parallel for all n, recomputing the parameters of the posterior approximation $q_{\text{EP}}(\boldsymbol{\beta}^d)$ only once per iteration of the outer loop (Cseke & Heskes, 2011; Seeger, 2008). The approximate posterior parameters are $\boldsymbol{\Pi}_{\text{EP}} = \frac{1}{\sigma_B^2}\mathbf{I} + \sum_{n=1}^N \widetilde{\boldsymbol{\Pi}}_n$ and $\boldsymbol{\lambda}_{\text{EP}} = \sum_{n=1}^N \widetilde{\boldsymbol{\lambda}}_n$. After convergence, the marginal likelihood $p(\mathbf{x}_{\cdot d}|\mathbf{W})$ can be approximated following Seeger (2005) as

$$\begin{aligned}\log p(\mathbf{x}_{\cdot d}|\mathbf{W}) \approx \log Z_{\text{EP}} = &-\frac{1}{2}\log|\boldsymbol{\Pi}_{\text{EP}}| - \frac{KR}{2}\log\sigma_B^2 + \frac{1}{2}\boldsymbol{\lambda}_{\text{EP}}^\top \boldsymbol{\Pi}_{\text{EP}}^{-1}\boldsymbol{\lambda}_{\text{EP}} \\ &+ \sum_{n=1}^N \log \widetilde{Z}_n,\end{aligned} \tag{A.5}$$

where we choose

$$\begin{aligned}\log \widetilde{Z}_n = \log \widehat{Z}_n &+ \frac{1}{2}\boldsymbol{\lambda}_{\neg n}^\top \boldsymbol{\Pi}_{\neg n}^{-1}\boldsymbol{\lambda}_{\neg n} - \frac{1}{2}(\boldsymbol{\lambda}_{\neg n} + \widetilde{\boldsymbol{\lambda}}_n)^\top (\boldsymbol{\Pi}_{\neg n} + \widetilde{\boldsymbol{\Pi}}_n)^{-1}(\boldsymbol{\lambda}_{\neg n} + \widetilde{\boldsymbol{\lambda}}_n) \\ &+ \frac{1}{2}\log|\boldsymbol{\Pi}_{\neg n} + \widetilde{\boldsymbol{\Pi}}_n| - \frac{1}{2}\log|\boldsymbol{\Pi}_{\neg n}|.\end{aligned} \tag{A.6}$$

We can summarize the full outer loop as given in algorithm 1.

A.1 Inner Loop. The inner loop is an EP method that approximates by a gaussian the tilted distribution $\widehat{p}_n(\boldsymbol{\beta}^d)$, which can be expressed as

$$\widehat{p}_n(\boldsymbol{\beta}^d) = \frac{1}{\widehat{Z}_n} q_{\neg n}(\boldsymbol{\beta}^d) t_n^d(\boldsymbol{\beta}^d)$$

Algorithm 1: Outer Loop of the Nested EP Algorithm.

Input: $\mathbf{x}_{\cdot d}$, $\mathbf{W}$, σ_B^2 (optionally initial site parameters $\widetilde{\mathbf{\Pi}}_n^{ini}$, $\widetilde{\boldsymbol{\lambda}}_n^{ini}$, $\widetilde{\alpha}_{nr}^{ini}$, $\widetilde{\beta}_{nr}^{ini}$)

Output: $p(\mathbf{x}_{\cdot d}|\mathbf{W})$, $\mathbf{\Pi}_{\mathrm{EP}}$, $\boldsymbol{\lambda}_{\mathrm{EP}}$ (optionally site parameters $\widetilde{\mathbf{\Pi}}_n$, $\widetilde{\boldsymbol{\lambda}}_n$, $\widetilde{\alpha}_{nr}$, $\widetilde{\beta}_{nr}$)

initialize $\widetilde{\mathbf{\Pi}}_n \leftarrow \widetilde{\mathbf{\Pi}}_n^{ini}$, $\widetilde{\boldsymbol{\lambda}}_n \leftarrow \widetilde{\boldsymbol{\lambda}}_n^{ini}$ for $n = 1, \ldots, N$

initialize $\mathbf{\Pi}_{\mathrm{EP}} \leftarrow \frac{1}{\sigma_B^2}\mathbf{I} + \sum_{n=1}^{N} \widetilde{\mathbf{\Pi}}_n$, $\boldsymbol{\lambda}_{\mathrm{EP}} \leftarrow \sum_{n=1}^{N} \widetilde{\boldsymbol{\lambda}}_n$

repeat

for $n = 1, \ldots, N$ (in parallel) **do**

cavity evaluations: $\mathbf{\Pi}_{\neg n} \leftarrow \mathbf{\Pi}_{\mathrm{EP}} - \widetilde{\mathbf{\Pi}}_n$,

$\boldsymbol{\lambda}_{\neg n} \leftarrow \boldsymbol{\lambda}_{\mathrm{EP}} - \widetilde{\boldsymbol{\lambda}}_n$

tilted moments:

$[\widehat{\mathbf{\Pi}}_n, \widehat{\boldsymbol{\lambda}}_n, \widehat{Z}_n, \{\widetilde{\alpha}_{nr}, \widetilde{\beta}_{nr}\}] \leftarrow$ inner_loop

$(x_{nd}, \mathbf{w}_{n\cdot}, \mathbf{\Pi}_{\neg n}, \boldsymbol{\lambda}_{\neg n}, \{\widetilde{\alpha}_{nr}^{ini}, \widetilde{\beta}_{nr}^{ini}\})$

site updates: $\widetilde{\mathbf{\Pi}}_n \leftarrow \eta_{\mathrm{O}}(\widehat{\mathbf{\Pi}}_n - \mathbf{\Pi}_{\neg n}) + (1 - \eta_{\mathrm{O}})\widetilde{\mathbf{\Pi}}_n$,

$\widetilde{\boldsymbol{\lambda}}_n \leftarrow \eta_{\mathrm{O}}(\widehat{\boldsymbol{\lambda}}_n - \boldsymbol{\lambda}_{\neg n}) + (1 - \eta_{\mathrm{O}})\widetilde{\boldsymbol{\lambda}}_n$

end for

update $\mathbf{\Pi}_{\mathrm{EP}} \leftarrow \frac{1}{\sigma_B^2}\mathbf{I} + \sum_{n=1}^{N} \widetilde{\mathbf{\Pi}}_n$, $\boldsymbol{\lambda}_{\mathrm{EP}} \leftarrow \sum_{n=1}^{N} \widetilde{\boldsymbol{\lambda}}_n$

until stopping criterion

for $n = 1, \ldots, N$ (in parallel) **do**

compute $\log \widetilde{Z}_n$ from equation A.6

end for

compute $\log p(\mathbf{x}_{\cdot d}|\mathbf{W})$ from equation A.4

$$= \frac{1}{\widehat{Z}_n} \mathcal{N}(\boldsymbol{\beta}^d | \mathbf{\Pi}_{\neg n}^{-1} \boldsymbol{\lambda}_{\neg n}, \mathbf{\Pi}_{\neg n}^{-1}) \int \mathcal{N}(u_{nd}|0,1) \times \left(\prod_{\substack{r'=1 \\ r' \neq r}}^{R} \Phi(u_{nd} + \mathbf{w}_{n\cdot}(\mathbf{b}_{\cdot r}^d - \mathbf{b}_{r'}^d)) \right) du_{nd}, \tag{A.7}$$

with $x_{nd} = \mathcal{X}_r$.

Removing the marginalization with respect to the auxiliary variable u_{nd} and defining $\boldsymbol{\beta}_{\mathrm{I}}^d$ as the vector compound of $\boldsymbol{\beta}^d$ and u_{nd}, namely, $\boldsymbol{\beta}_{\mathrm{I}}^d = [(\boldsymbol{\beta}^d)^\top, u_{nd}]^\top$, we have the augmented tilted distribution,

$$\widehat{p}_n(\boldsymbol{\beta}_{\mathrm{I}}^d) = \frac{1}{\widehat{Z}_n}\mathcal{N}(\boldsymbol{\beta}_{\mathrm{I}}^d|\boldsymbol{\Pi}_{\mathrm{I}_n}^{-1}\boldsymbol{\lambda}_{\mathrm{I}_n}, \boldsymbol{\Pi}_{\mathrm{I}_n}^{-1})\prod_{\substack{r'=1\\ r'\neq r}}^{R}\Phi((\mathbf{h}_{nr'}^d)^\top\boldsymbol{\beta}_{\mathrm{I}}^d), \tag{A.8}$$

where we have defined $\boldsymbol{\Pi}_{\mathrm{I}n}$ as a block-diagonal matrix formed from $\boldsymbol{\Pi}_{\neg n}$ and 1, $\boldsymbol{\lambda}_{\mathrm{I}_n} = [\boldsymbol{\lambda}_{\neg n}^\top, 0]^\top$, and $\mathbf{h}_{nr'}^d = [(\mathbf{e}_r - \mathbf{e}_{r'})^\top \otimes \mathbf{w}_{n\cdot},\ 1]^\top$. Here, $\otimes$ denotes the Kronecker product, and $\mathbf{e}_r$ is the rth unit (column) vector of the R-dimensional standard basis. Note that we use the subscript I to denote the augmented variables that account for both $\boldsymbol{\beta}^d$ and u_{nd}. The normalization term $\widehat{Z}_n$ is the same for $\widehat{p}_n(\boldsymbol{\beta}^d)$ and for the augmented distribution $\widehat{p}_n(\boldsymbol{\beta}_{\mathrm{I}}^d)$, and it is defined as

$$\widehat{Z}_n = \int q_{\neg n}(\boldsymbol{\beta}^d)\mathcal{N}(u_{nd}|0,1)\prod_{r'\neq r}\Phi((\mathbf{h}_{nr'}^d)^\top\boldsymbol{\beta}_{\mathrm{I}}^d)d\boldsymbol{\beta}_{\mathrm{I}}^d. \tag{A.9}$$

Due to the multinomial probit model, equation A.8 contains a product of intractable functions of the scalar variables $s_{r'} = (\mathbf{h}_{nr'}^d)^\top\boldsymbol{\beta}_{\mathrm{I}}^d$, allowing us to apply a new inner EP loop, which is simpler than the outer loop since it involves only scalar operations. Hence, the augmented distribution in equation A.8 can be approximated by replacing each intractable term $\Phi(s_{r'})$ with a scaled univariate gaussian site function with natural parameters $\widetilde{\alpha}_{nr'}$ and $\widetilde{\beta}_{nr'}$, resulting in the approximate distribution

$$\begin{aligned} q_{\mathrm{I}_n}(\boldsymbol{\beta}_{\mathrm{I}}^d) &= \frac{1}{C_{\mathrm{I}_n}}\mathcal{N}(\boldsymbol{\beta}_{\mathrm{I}}^d|\boldsymbol{\Pi}_{\mathrm{I}_n}^{-1}\boldsymbol{\lambda}_{\mathrm{I}_n}, \boldsymbol{\Pi}_{\mathrm{I}_n}^{-1})\prod_{\substack{r'=1\\ r\neq r}}^{R}\widetilde{C}_{nr'}\mathcal{N}(s_{r'}|\widetilde{\alpha}_{nr'}^{-1}\widetilde{\beta}_{nr'}, \widetilde{\alpha}_{nr'}^{-1}) \\ &= \mathcal{N}(\boldsymbol{\beta}_{\mathrm{I}}^d|\widetilde{\boldsymbol{\Pi}}_{\mathrm{I}_n}^{-1}\widetilde{\boldsymbol{\lambda}}_{\mathrm{I}_n}, \widetilde{\boldsymbol{\Pi}}_{\mathrm{I}_n}^{-1}), \end{aligned} \tag{A.10}$$

where the normalization constant C_{I_n} approximates $\widehat{Z}_n$.

We start from $q_{nr'}(s_{r'}) = \mathcal{N}(s_{r'}|m_{nr'}, v_{nr'})$, being $m_{nr'} = (\mathbf{h}_{nr'}^d)^\top\widetilde{\boldsymbol{\Pi}}_{\mathrm{I}_n}^{-1}\widetilde{\boldsymbol{\lambda}}_{\mathrm{I}_n}$ and $v_{nr'} = (\mathbf{h}_{nr'}^d)^\top\widetilde{\boldsymbol{\Pi}}_{\mathrm{I}_n}^{-1}\mathbf{h}_{nr'}^d$. Then the cavity distribution $q_{n\neg r'}(s_{r'})$ can be written as

$$q_{n\neg r'}(s_{r'}) = \mathcal{N}(s_{r'}|m_{n\neg r'}, v_{n\neg r'}), \tag{A.11}$$

which has mean $m_{n\neg r'} = v_{n\neg r'}(m_{nr'}/v_{nr'} - \widetilde{\beta}_{nr'})$ and variance $v_{n\neg r'} = (1/v_{nr'} + \widetilde{\alpha}_{nr'})^{-1}$. The tilted distribution (including one true site),

$$\widehat{f}_{nr'}(s_{r'}) = \frac{1}{\widehat{C}_{nr'}} q_{n\neg r'}(s_{r'})\Phi(s_{r'}), \tag{A.12}$$

has mean $\widehat{m}_{nr'} = m_{n\neg r'} + \rho_{nr'} v_{n\neg r'}$, variance $\widehat{v}_{nr'} = v_{n\neg r'} - v^2_{n\neg r'}(\rho^2_{nr'} + \rho_{nr'} \frac{m_{n\neg r'}}{1+v_{n\neg r'}})$ and normalization constant $\widehat{C}_{nr'} = \Phi(\frac{m_{n\neg r'}}{\sqrt{1+v_{n\neg r'}}})$ (Qi et al., 2004), being

$$\rho_{nr'} = \frac{\mathcal{N}\left(\frac{m_{n\neg r'}}{\sqrt{1+v_{n\neg r'}}}|0,1\right)}{\Phi\left(\frac{m_{n\neg r'}}{\sqrt{1+v_{n\neg r'}}}\right)\sqrt{1+v_{n\neg r'}}}. \tag{A.13}$$

Finally, the site updates are computed as $\widetilde{\alpha}_{nr'} = 1/\widehat{v}_{nr'} - 1/v_{n\neg r'}$ and $\widetilde{\beta}_{nr'} = \widehat{m}_{nr'}/\widehat{v}_{nr'} - m_{n\neg r'}/v_{n\neg r'}$. Again, a damping factor $\eta_{\mathrm{I}} \in (0,1]$ can be used in this step. In this case, the site updates can be obtained in parallel for the different values of r', afterward recomputing the natural parameters of $q_{\mathrm{I}_n}(\boldsymbol{\beta}^d_{\mathrm{I}})$ as $\widetilde{\boldsymbol{\Pi}}_{\mathrm{I}_n} = \boldsymbol{\Pi}_{\mathrm{I}_n} + \sum_{r'\neq r} \widetilde{\alpha}_{nr'} \mathbf{h}^d_{nr'} (\mathbf{h}^d_{nr'})^\top$ and $\widetilde{\boldsymbol{\lambda}}_{\mathrm{I}_n} = \boldsymbol{\lambda}_{\mathrm{I}_n} + \sum_{r'\neq r} \widetilde{\beta}_{nr'} \mathbf{h}^d_{nr'}$.

The constants C_{I_n} (which approximates $\widehat{Z}_n$ in equation A.8) and $\widetilde{C}_{nr'}$ in equation A.10 can be computed after meeting the stopping criterion as

$$\begin{aligned}\log C_{\mathrm{I}_n} = &\sum_{r'=1r'\neq r}^{R} \left(\log \widetilde{C}_{nr'} + \frac{1}{2}\log(\widetilde{\alpha}_{nr'})\right) \\ &+ \frac{1}{2}\log(|\boldsymbol{\Pi}_{\mathrm{I}_n}| - |\widetilde{\boldsymbol{\Pi}}_{\mathrm{I}_n}|) + \frac{1}{2}\left(\widetilde{\boldsymbol{\lambda}}^\top_{\mathrm{I}_n} \widetilde{\boldsymbol{\Pi}}^{-1}_{\mathrm{I}_n} \widetilde{\boldsymbol{\lambda}}_{\mathrm{I}_n} - \boldsymbol{\lambda}^\top_{\mathrm{I}_n} \boldsymbol{\Pi}^{-1}_{\mathrm{I}_n} \boldsymbol{\lambda}_{\mathrm{I}_n}\right),\end{aligned} \tag{A.14}$$

and

$$\begin{aligned}\log \widetilde{C}_{nr'} = &\log \widehat{C}_{nr'} + \frac{1}{2}\log(v_{n\neg r'} + 1/\widetilde{\alpha}_{nr'}) \\ &+ \frac{1}{2}\left(\frac{m^2_{n\neg r'}}{v_{n\neg r'}} - \frac{\left(\frac{m_{n\neg r'}}{v_{n\neg r'}} + \widetilde{\beta}_{nr'}\right)^2}{1/v_{n\neg r'} + \widetilde{\alpha}_{nr'}}\right).\end{aligned} \tag{A.15}$$

Matrices $\widehat{\boldsymbol{\Pi}}_n$ and $\widehat{\boldsymbol{\lambda}}_n$ of the outer loop can be obtained from $\widetilde{\boldsymbol{\Pi}}_{\mathrm{I}_n}$ and $\widetilde{\boldsymbol{\lambda}}_{\mathrm{I}_n}$ after removing the effects of the auxiliary variable u_{nd}.

The complete algorithm is summarized in algorithm 2.

A.2 Computational Complexity. Although the nested EP is similar to the algorithm proposed by Riihimäki et al. (2013), the computational complexity is substantially different. The running time of the nested EP for our model is linear in the number of instances (N), while for the Gaussian

Algorithm 2: Inner Loop of the Nested EP Algorithm.

Input: $x_{nd}, \mathbf{w}_{n\cdot}, \mathbf{\Pi}_{\neg n}, \boldsymbol{\lambda}_{\neg n}$ (optionally initial site parameters $\widetilde{\alpha}_{nr'}^{ini}, \widetilde{\beta}_{nr'}^{ini}$)

Output: $\widehat{\mathbf{\Pi}}_n, \widehat{\boldsymbol{\lambda}}_n, \widehat{Z}_n$ (optionally site parameters $\widetilde{\alpha}_{nr'}, \widetilde{\beta}_{nr'}$)

initialize $\widetilde{\alpha}_{nr'} \leftarrow \widetilde{\alpha}_{nr'}^{ini}, \widetilde{\beta}_{nr'} \leftarrow \widetilde{\beta}_{nr'}^{ini}$ for $r' = 1, \ldots, R$ (with $r' \neq r$ and $x_{nd} = \mathcal{X}_r$)

initialize $\mathbf{\Pi}_{\mathrm{I}_n}, \boldsymbol{\lambda}_{\mathrm{I}_n}$ from $\mathbf{\Pi}_{\neg n}, \boldsymbol{\lambda}_{\neg n}$

initialize $\widetilde{\mathbf{\Pi}}_{\mathrm{I}_n} \leftarrow \mathbf{\Pi}_{\mathrm{I}_n} + \sum_{r' \neq r} \widetilde{\alpha}_{nr'} \mathbf{h}_{nr'}^d (\mathbf{h}_{nr'}^d)^\top$, $\widetilde{\boldsymbol{\lambda}}_{\mathrm{I}_n} \leftarrow \boldsymbol{\lambda}_{\mathrm{I}_n} + \sum_{r' \neq r} \widetilde{\beta}_{nr'} \mathbf{h}_{nr'}^d$

repeat

for $r' = 1, \ldots, R$ with $r' \neq r$ (in parallel) **do**

marginal moments: $v_{nr'} \leftarrow (\mathbf{h}_{nr'}^d)^\top \widetilde{\mathbf{\Pi}}_{\mathrm{I}_n}^{-1} \mathbf{h}_{nr'}^d$, $m_{nr'} \leftarrow (\mathbf{h}_{nr'}^d)^\top \widetilde{\mathbf{\Pi}}_{\mathrm{I}_n}^{-1} \widetilde{\boldsymbol{\lambda}}_{\mathrm{I}_n}$

cavity evaluations: $v_{n\neg r'} \leftarrow (1/v_{nr'} + \widetilde{\alpha}_{nr'})^{-1}$, $m_{n\neg r'} \leftarrow v_{n\neg r'}(m_{nr'}/v_{nr'} - \widetilde{\beta}_{nr'})$

auxiliary variable: $\rho_{nr'} \leftarrow \mathcal{N}\left(\frac{m_{n\neg r'}}{\sqrt{1+v_{n\neg r'}}} | 0, 1\right) / \left(\Phi\left(\frac{m_{n\neg r'}}{\sqrt{1+v_{n\neg r'}}}\right)\sqrt{1+v_{n\neg r'}}\right)$

tilted moments:

$$\widehat{v}_{nr'} \leftarrow v_{n\neg r'} - v_{n\neg r'}^2 \left(\rho_{nr'}^2 + \rho_{nr'} \frac{m_{n\neg r'}}{1+v_{n\neg r'}}\right),$$

$$\widehat{m}_{nr'} \leftarrow m_{n\neg r'} + \rho_{nr'} v_{n\neg r'}, \widehat{C}_{nr'} \leftarrow \Phi\left(\frac{m_{n\neg r'}}{\sqrt{1+v_{n\neg r'}}}\right)$$

site updates: $\widetilde{\alpha}_{nr'} \leftarrow \eta_{\mathrm{I}}(1/\widehat{v}_{nr'} - 1/v_{n\neg r'}) + (1-\eta_{\mathrm{I}})\widetilde{\alpha}_{nr'}$,

$$\widetilde{\beta}_{nr'} \leftarrow \eta_{\mathrm{I}}(\widehat{m}_{nr'}/\widehat{v}_{nr'} - m_{n\neg r'}/v_{n\neg r'}) + (1-\eta_{\mathrm{I}})\widetilde{\beta}_{nr'}$$

end for

update $\widetilde{\mathbf{\Pi}}_{\mathrm{I}_n} \leftarrow \mathbf{\Pi}_{\mathrm{I}_n} + \sum_{r' \neq r} \widetilde{\alpha}_{nr'} \mathbf{h}_{nr'}^d (\mathbf{h}_{nr'}^d)^\top$, $\widetilde{\boldsymbol{\lambda}}_{\mathrm{I}_n} \leftarrow \boldsymbol{\lambda}_{\mathrm{I}_n} + \sum_{r' \neq r} \widetilde{\beta}_{nr'} \mathbf{h}_{nr'}^d$

until stopping criterion

for $r' = 1, \ldots, R$ with $r' \neq r$ (in parallel) **do**

compute $\log \widetilde{C}_{nr'}$ from equation A.15

end for

compute $\log \widehat{Z}_n$ from equation A.14, and $\widehat{\mathbf{\Pi}}_n, \widehat{\boldsymbol{\lambda}}_n$ from $\widetilde{\mathbf{\Pi}}_{\mathrm{I}_n}, \widetilde{\boldsymbol{\lambda}}_{\mathrm{I}_n}$

processes for multiclass classification, the computational complexity is cubic. The nested EP for our algorithm needs to integrate out $\boldsymbol{\beta}^d$, which is an $R(K_+ + 1)$-dimensional vector. Note that the outer loop of the proposed nested EP requires one loop in n and, since all the sites $t_n^d(\boldsymbol{\beta}^d)$ are functions of the same $R(K_+ + 1)$-dimensional random vector $\boldsymbol{\beta}^d$, no matrix inversion is

needed when we work with the natural parameters of the normal distributions. Each iteration of the inner loop, however, requires the inversion of a matrix of size $R(K_+ + 1) + 1$ (in practice, computed using the Cholesky decomposition), which has a complexity of $\mathcal{O}((R(K_+ + 1) + 1)^3)$. The overall complexity of the posterior approximation scales with $DN(R(K_+ + 1) + 1)^3$, because we iterate through the number of samples N and the dimensionality of the observation vector D. Evaluating the likelihood after convergence of the outer loop requires operations of matrices of size $R(K_+ + 1)$ within a loop in n, which leads to a complexity scaling with $N(R(K_+ + 1))^3$. Thus, the overall complexity of the full nested EP algorithm to evaluate the marginal likelihood $p(\mathbf{X}|\mathbf{W})$ is $\mathcal{O}(DN(R(K_+ + 1) + 1)^3)$. The EP procedure can be parallelized in the dimension of the observed instances (D) and in the number of instances N, providing significant savings in runtime complexity.

Furthermore, the site parameters of the inner loop can be stored after each inner EP run and used as starting parameters the next time the inner loop is called (Riihimäki et al., 2013). In addition, successive calls to the nested EP algorithm differ in just one element of w_{nk}, which allows reducing the number of outer loop iterations by storing the site parameters $\widetilde{\boldsymbol{\lambda}}_n$ and $\widetilde{\boldsymbol{\Pi}}_n$ after each nested EP run and continuing from the previous values in the next run. In the effort to add new features, the values of the old site parameters can still be used to build the new parameters (extended to account for the new features) $\widetilde{\boldsymbol{\lambda}}_n$ and $\widetilde{\boldsymbol{\Pi}}_n$.

Acknowledgments

I.V. is supported by the Humboldt Research Fellowship for Postdoctoral Researchers program and acknowledges the support of Plan Regional-Programas I+D of Comunidad de Madrid (AGES-CM S2010/BMD-2422). F.J.R.R. is supported by an FPU fellowship from the Spanish Ministry of Education (AP2010-5333). This work is also partially supported by Ministerio de Economía of Spain (projects COMONSENS, id. CSD2008-00010, and ALCIT, id. TEC2012-38800-C03-01); Comunidad de Madrid (project CASI-CAM-CM, id. S2013/ICE-2845); the Office of Naval Research (ONRN0014-11-1-0651); and the European Union 7th Framework Programme through the Marie Curie Initial Training Network Machine Learning for Personalized Medicine (MLPM2012, grant 316861).

References

Blanco, C., Krueger, R. F., Hasin, D. S., Liu, S. M., Wang, S., Kerridge, B. T., . . . Olfson, M. (2013). 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.

Cseke, B., & Heskes, T. (2011). Approximate marginals in latent gaussian models. *J. Mach. Learn. Res.*, *12*, 417–454.

Girolami, M., & Zhong, M. (2007). Data integration for classification problems employing gaussian process priors. In B. Schölkopf, J. Platt, & T. Hoffmann (Eds.), *Advances in neural information processing systems, 19* (pp. 465–472). Cambridge, MA: MIT Press.

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

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109.

Kendler, K. S., Prescott, C. A., Myers, J., & Neale, M. C. (2003). 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.

Knowles, D. A., & Ghahramani, Z. (2011). Nonparametric Bayesian sparse factor models with application to gene expression modelling. *Annals of Applied Statistics*, *5*(2B), 1534–1552.

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

Kuss, M., & Rasmussen, C. E. (2005). Assessing approximate inference for binary gaussian process classification. *Journal of Machine Learning Research*, *6*, 1679–1704.

Loehlin, J. C. (1986). *Latent variable models: An introduction to factor, path, and structural analysis*. Mahwah, NJ: Erlbaum.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, *21*(6), 1087–1092.

Minka, T. P. (2001). Expectation propagation for approximate Bayesian inference. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, (pp. 362–369). San Francisco: Morgan Kaufmann.

Opper, M., & Winther, O. (2005). Expectation consistent approximate inference. *J. Mach. Learn. Res.*, *6*, 2177–2204.

Qi, Y., Minka, T. P., Picard, R. W., & Ghahramani, Z. (2004). Predictive automatic relevance determination by expectation propagation. In *Proceedings of the Twenty-First International Conference on Machine Learning* (pp. 671–678). New York: ACM.

Riihimäki, J., Jylänki, P., & Vehtari, A. (2013). Nested expectation propagation for gaussian process classification with a multinomial probit likelihood. *Journal of Machine Learning Research*, *14*, 75–109.

Ruiz, F. J. R., Valera, I., Blanco, C., & Perez-Cruz, F. (2012). Bayesian nonparametric modeling of suicide attempts. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems, 25* (pp. 1862–1870). Red Hook, NY: Curran.

Ruiz, F. J. R., Valera, I., Blanco, C., & Perez-Cruz, F. (2014). Bayesian nonparametric comorbidity analysis of psychiatric disorders. *Journal of Machine Learning Research*, *15*, 1215–1247.

Seeger, M. W. (2005). *Expectation propagation for exponential families* (EPFL-Report 161464). Lausanne: Ecole Polytecnique Fédérale de Lausanne.

Seeger, M. W. (2008). Bayesian inference and optimal design for the sparse linear model. *J. Mach. Learn. Res.*, *9*, 759–813.

Seeger, M. W., & Jordan, M. I. (2004). *Sparse gaussian process classification with multiple classes* (Technical report). Berkeley: University of California, Berkeley.

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

Williamson, S., Wang, C., Heller, K. A., & Blei, D. M. (2010). The IBP compound Dirichlet process and its application to focused topic modeling. In *Proceedings of the 27th International Conference on Machine Learning* (pp. 1151–1158). Madison, WI: Omni Press.

Received April 10, 2015; accepted October 12, 2015.