*Submitted to the Annals of Applied Statistics*

# LATENT CLASS MODELING USING MATRIX COVARIATES WITH APPLICATION TO IDENTIFYING EARLY PLACEBO RESPONDERS BASED ON EEG SIGNALS*

By Bei Jiang[†], Eva Petkova[‡,§], Thaddeus Tarpey[¶] and R. Todd Ogden[‖]

*University of Alberta[†], New York University[‡], Nathan S. Kline Institute for Psychiatric Research[§], Wright State University[¶] and Columbia University[‖]*

Latent class models are widely used to identify unobserved subgroups (i.e., latent classes) based upon one or more manifest variables. The probability of belonging to each subgroup is typically modeled as a function of a set of measured covariates. In this paper, we extend existing latent class models to incorporate matrix covariates. This research is motivated by a randomized placebo-controlled depression clinical trial. One study goal is to identify a subgroup of subjects who experience symptoms improvement early on during antidepressant treatment, which is considered to be an indication of a placebo rather than a true pharmacological response. We want to relate the likelihood of belonging to this subgroup of early responders to baseline electroencephalography (EEG) measurement that takes the form of a matrix. The proposed method is built upon a low rank Candecomp/Parafac (CP) decomposition of the target coefficient matrix through low-dimensional latent variables, which effectively reduces the model dimensionality. We adopt a Bayesian hierarchical modeling approach to estimate the latent variables, which allows a flexible way to incorporate prior knowledge about covariate effect heterogeneity and offers a data-driven method of regularization. Simulation studies suggest that the proposed method is robust against potentially misspecified rank in the CP decomposition. With the motivating example we show how the proposed method can be applied to extract valuable information from baseline EEG measurements that explains the likelihood of belonging to the early responder subgroup, helping to identify placebo responders and suggesting new targets for the study of placebo response.

**1. Introduction.** Placebo responses to antidepressant treatment (also known as non-specific response, i.e., an improvement in symptoms that is

*Supported in part by Grant 5R01MH099003 and Grant U01MH092221 from the US National Institute of Mental Health, and a Discovery Grant from Natural Sciences and Engineering Research Council of Canada.

*Keywords and phrases:* Candecomp/Parafac (CP) matrix decomposition, Bayesian hierarchical modeling, data-driven regularization, major depression, placebo effect

not due to the effect of the active chemicals in the drug) is highly prevalent. Patients who have responded to such non-specific aspects of the treatment are called placebo responders. Clearly, there could be placebo responders among both placebo and drug treated patients. For example, it is widely accepted that antidepressants from the class of the selective serotonin reuptake inhibitors (SSRIs) do not begin to exert their effect until at least two weeks of treatment, during which time serotonin levels can accumulate in the brain and exert a therapeutic effect (e.g., Quitkin et al. 1991; Stewart et al. 1998; Sonawalla and Rosenbaum 2002). Therefore, an early improvement experienced among drug treated patients is an indication of a placebo (i.e., non-specific) response rather than a true drug response.

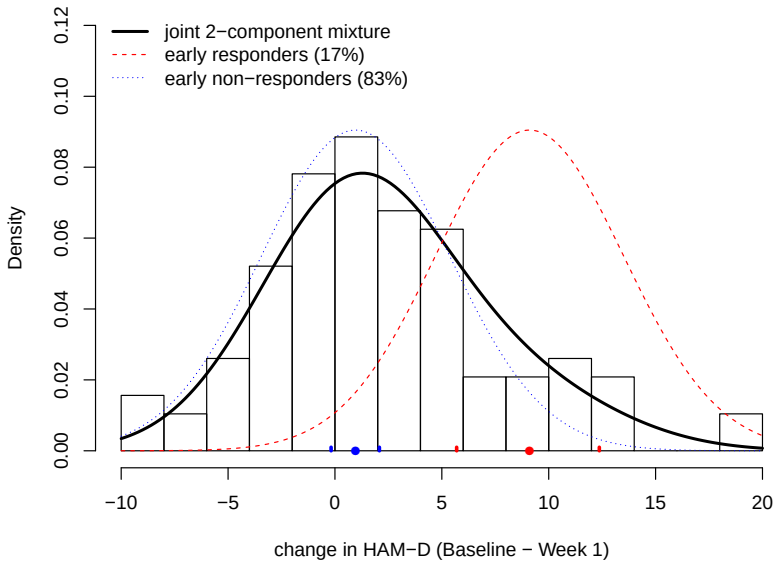



FIG 1. *Histogram of the change in HAM-D (baseline - week 1) showing the amount of improvement in depression symptoms after 1 week for both drug and placebo treated patients. The solid curve is the posterior density estimate of this distribution under the joint 2-component mixture model specified in (1) and (2) (and the dashed and dotted curves are the estimated component densities). The marked points on the x-axis are the posterior estimates of the subgroup means, along with the lower and upper bounds of the corresponding 95% credible intervals.*

Figure 1 shows a histogram of the change in the Hamilton Depression (HAM-D) scale (baseline - week 1) for the first 96 subjects with major depressive disorder (MDD) from an ongoing randomized placebo controlled clinical trial of sertraline. The HAM-D is a clinical measure designed to rate severity of depression, where higher scores indicate more severe depression;

and therefore a positive change in HAM-D (baseline - week 1) would indicate improvement in symptoms. Although HAM-D scores are bounded and discrete (we used the 17-item scale), they are typically modeled adequately as continuous variables (e.g., Bonate and Howard (2011)). The figure includes both placebo and sertraline treated patients. Patients' levels of depression were assessed at baseline and continued to be monitored after randomization, including at 1 week. The pattern in the distribution of the change in HAM-D, as suggested in Figure 1, is consistent with previous findings on early placebo response (e.g., Tarpey, Yun and Petkova 2008), indicating that these patients may possibly cluster into two clinically distinct groups: a small proportion of patients who experience an early improvement in symptoms (i.e., early responders), while the majority of patients are not early responders (i.e., early non-responders), who remain unimproved or in some cases got worse during early treatment. As discussed in Section 4, accounting for such heterogeneity in treatment response leads to a better model fit. The solid curve in Figure 1 represents the 2-component mixture model corresponding to the early placebo responder and non-responder subgroups estimated using the models specified in (1) and (2) and the dashed and dotted curves are the corresponding component densities).

Much research has focused on the identification of placebo responders and discovery of patients' characteristics that could be related to placebo response (e.g., Joyce and Paykel 1989; Tarpey, Petkova and Ogden 2003; Elliott et al. 2005; Muthén and Brown 2009; Petkova, Tarpey and Govindarajulu 2009; Tarpey and Petkova 2010). However, the typically measured clinical phenotypes, such as symptom severity and treatment history, have shown low predictive power (Leuchter et al. 2002; Phillips et al. 2015). One goal of the motivating study is to explore the predictive ability of baseline neuroimaging phenotypes, such as the brain activity measured through electroencephalography (EEG), for identification of early responders, who have improved due to non-specific placebo effects. Although EEG data are regarded as having relatively low spatial resolution, compared to data from other imaging modalities, EEG has found extensive use in depression studies, in part due to its non-invasive nature and cost-effectiveness. As commented by Holsboer (2008), "Studies that investigate the use of EEG as a tool to make predictions about whether patients will respond favourably to a given antidepressant have a long tradition". For example, a number of previous studies have indicated that pre-treatment EEG predicts response to active antidepressant treatment (e.g., Bruder et al. 2001, Bruder et al. 2008, Holsboer 2008, Tenke et al. 2011, Khodayari-Rostamabad et al. 2010, Tenke et al. 2011, Mumtaz et al. 2015, Patel, Khalaf and Aizenstein 2016,

Wade and Iosifescu 2016). However, the capability of EEG in differentiating patients who may have an early response due to non-specific placebo effect is unknown (Wade and Iosifescu 2016). Such knowledge is useful in clinical practices as it could guide clinicians in deciding which patients should receive an antidepressant and which are likely to improve without active drug. Also it could potentially lead to improvements and new developments in precision medicine for treating MDD and allow a sharper focus on the specific effects of active drug.

This problem can be naturally formulated as a latent class model (e.g., Lazarsfeld and Henry 1968; MacCutcheon 1987; Clogg 1995; Collins and Lanza 2013), which is often referred to as a mixture of experts model in the machine learning literature (e.g., Jacobs et al. 1991; Jordan and Jacobs 1994; Gormley and Murphy 2011; White and Murphy 2016). Specifically, a mixture distribution is postulated for the observed change in HAM-D scores to classify subjects into two subgroups corresponding to two unobserved latent classes: early responders and non-responders. Additionally, the latent class model can be used for prediction of the probability of being in the early responder subgroup as a function of the covariates of interest, including baseline EEG measurements. Latent class models and their extensions have been successfully used in various applications to accommodate heterogeneity in the outcome and to simultaneously characterize the latent class memberships through its association with explanatory variables (e.g., Bandeen-Roche et al. 1997; Muthén and Shedden 1999; Elliott 2007; Muthén and Brown 2009). In a more recent example, Shen and He (2015) used a logistic-normal mixture model to identify a subgroup of patients who benefited from an enhanced treatment effect in a randomized clinical trial and related baseline covariates of interest to the probability of being in this subgroup.

In our motivating dataset, each subject's EEG data takes the form of a $14 \times 45$ matrix. This EEG data matrix contains the current source density (CSD) amplitude spectrum values ($\mu V/m^2$) (Nunez and Srinivasan 2006) at a total of 14 electrodes located in brain's posterior (occipital and parietal) regions, crossed with 45 frequency ranges within the theta (4 - 7 Hz) and alpha (7 - 15 Hz) frequency bands (leading to a total of 45 frequencies, given a 0.25 Hz frequency resolution). The CSD measures of EEG are the widely preferred method for sharpening the spatial resolution of EEG data and thus improving interpretability (e.g., Tenke et al. 2011 and Kamarajan et al. 2015). The CSD measures at the 14 posterior brain region electrodes over the theta/alpha frequency bands have been previously reported to be related to antidepressant response (Bruder et al. 2001, Bruder et al. 2008

and Tenke et al. 2011) and hypothesized by the investigators in this study to be capable of differentiating patients who may have an early treatment response due to non-specific placebo effects. However, common practice in the EEG literature is to use low-dimensional EEG summaries, such as the mean over a small number of frequency bands. This practice potentially leads to an important loss of information. Instead, we propose to directly model the matrix-valued EEG data as predictors. To effectively exploit the information embedded in these EEG measures that relates to the subgroup membership, we consider a Bayesian hierarchical approach that utilizes the powerful Candecomp/Parafac (CP) decomposition (Kolda and Bader 2009). In particular, a CP decomposition imposes a special low rank structure on the target regression coefficient matrix that explicitly captures the bilinear row and column effects of the matrix covariate, and greatly reduces model dimensionality. In the case of EEG data, different electrodes and different frequencies could contribute to both the variability in the EEG signals and their effects on the likelihood for belonging to the early responder group; CP decomposition of these signals models the bilinear two-way interaction effects between electrodes and frequencies.

Recent related work that also explores low rank CP decomposition in regression problems with multidimensional covariates includes Hung and Wang 2013 and Zhou, Li and Zhu 2013. Specifically, Hung and Wang (2013) considered logistic regression for matrix covariates with the rank in CP decomposition fixed at one; more generally, Zhou, Li and Zhu (2013) proposed a new class of generalized linear models (GLMs) for array covariates of arbitrary order. Both papers focused on penalized maximum likelihood estimation methods. In contrast, we adopt a hierarchical approach in formulating the CP decomposition and employ Bayesian methods for parameter estimation. Our approach is new and is characterized by the following novel features:

1. It allows for the incorporation of prior knowledge on covariate effect heterogeneity by using postulated prior distributions on the latent variables associated with the electrodes and the frequencies.
2. It provides a method of regularization, with the amount of shrinkage being determined in a data-driven fashion.
3. The credible intervals for all the elements in the resulting regression coefficient matrix through CP decomposition can be obtained straightforwardly, as a natural consequence of applying Bayesian methods; construction of such confidence intervals are not discussed in Hung and Wang 2013 and Zhou, Li and Zhu 2013.

The remainder of the paper is organized as follows. Section 2 presents the proposed hierarchical models, the Bayesian method for estimation, and the choice of rank in the CP decomposition. The performances of the proposed method are evaluated through two simulation studies in Section 3, when the rank is correctly assumed or misspecified. In Section 4, we apply the proposed method to our motivating study to explore the association between the baseline characteristics, including matrix EEG measurements and the likelihood of being in an early responder subgroup. We conclude with a discussion in Section 5.

**2. The hierarchical Bayesian modeling and estimation.** In this section, we present the model for the observed clinical outcome and baseline EEG measurements. First, we assume a 2-mixture latent class model to reflect the widely held theory in psychiatry that there will be early responders and non-responders to antidepressant treatment. Then, the binary subgroup indicators are modeled via a hierarchical probit model as a function of the baseline EEG measurements and other covariates of interest. We choose a probit link as it is frequently used in practice and can lead to closed-form full conditional posterior distributions in the Gibbs sampler (discussed in Section 2.5). For more general link functions, please refer to Kim, Chen and Dey (2008).

2.1. *Model for the observed outcome.* For each subject $i = 1, ..., n$, let $y_i$ denote the observed clinical outcome, where higher $y_i$ values indicate greater clinical improvement. We consider the following model for $y_i$,

$$y_i = \eta_0 + \eta_1 \gamma_i + \epsilon_i, \ \epsilon_i \overset{iid}{\sim} N(0, \sigma^2) \tag{1}$$

where $\gamma_i$ is an indicator variable with $\gamma_i = 1$ indicating an early responder and $\gamma_i = 0$ indicating not an early responder who does not demonstrate non-specific effects. We constrain $\eta_0 + \eta_1 > 0$ so that the early responder subgroup consists of subjects who experience improved symptoms and hence positive clinical outcome values. As discussed in the Introduction section, because early improvement of depressive symptoms within one week of treatment is believed to be a non-specific placebo response rather than due to medication effect (e.g., Quitkin et al. 1991; Stewart et al. 1998; Sonawalla and Rosenbaum 2002), we do not include a treatment indicator variable in model (1).

2.2. *Model for the latent class indicator with matrix covariates.* For given positive integers $p$ and $q$, $\mathbb{R}^{p \times q}$ denotes the space of all matrices of dimension

$p \times q$. For each subject $i$, let $\boldsymbol{x}_i \in \mathbb{R}^{p\times q}$ denote the matrix covariate and $\boldsymbol{z}_i$ is a vector that contains all scalar covariates for subject $i$. To relate the covariates $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ to the likelihood of being an early responder, i.e., $\gamma_i = 1$, we consider the following probit model for the latent class indicator:

$$\Phi^{-1}[\Pr\{\gamma_i = 1\}] = \boldsymbol{\theta}^\top \boldsymbol{z}_i + \langle \boldsymbol{\Theta}, \boldsymbol{x}_i \rangle, \tag{2}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of a standard normal distribution, $\boldsymbol{\Theta} \in \mathbb{R}^{p\times q}$ denotes the target coefficient matrix for $\boldsymbol{x}_i$, and their inner product is defined as $\langle \boldsymbol{\Theta}, \boldsymbol{x}_i \rangle = \text{vec}(\boldsymbol{\Theta})^\top \text{vec}(\boldsymbol{x}_i)$. Instead of focusing on estimating the entire matrix $\boldsymbol{\Theta}$, we assume a low-dimensional structure on $\boldsymbol{\Theta}$ through CP decomposition (Kolda and Bader 2009). Specifically, we represent the target coefficient matrix $\boldsymbol{\Theta}$ by a sum of $R$ outer products of two non-zero column vectors such that $R < \min(p, q)$; that is, we can express $\boldsymbol{\Theta} = \sum_{r=1}^{R} \boldsymbol{\alpha}_r \boldsymbol{\beta}_r^\top$, where $\boldsymbol{\alpha}_r = (\alpha_{1r}, \ldots, \alpha_{pr})^\top \in \mathbb{R}^p$ and $\boldsymbol{\beta}_r = (\beta_{1r}, \ldots, \beta_{qr})^\top \in \mathbb{R}^q$, $r = 1, \ldots, R$. Further, letting $\boldsymbol{A} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_R] \in \mathbb{R}^{p\times R}$ and $\boldsymbol{B} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_R] \in \mathbb{R}^{q\times R}$, we can rewrite $\boldsymbol{\Theta} = \boldsymbol{A}\boldsymbol{B}^\top$. Under this setup, model (2) can be rewritten as,

$$\Phi^{-1}[\Pr\{\gamma_i = 1\}] = \boldsymbol{\theta}^\top \boldsymbol{z}_i + \left\langle \boldsymbol{A}\boldsymbol{B}^\top, \boldsymbol{x}_i \right\rangle, \tag{3}$$

The task is to estimate the two low-dimensional matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, leading to $R(p + q)$ parameters, instead of the total $pq$ matrix parameters in the unconstrained $\boldsymbol{\Theta}$. In the case of a rank-one (i.e., $R = 1$) CP decomposition, model (3) is reduced to $\Phi^{-1}[p\{\gamma_i = 1\}] = \boldsymbol{\theta}^\top \boldsymbol{z}_i + \boldsymbol{\alpha}_1^\top \boldsymbol{x}_i \boldsymbol{\beta}_1$. In contrast to a variable selection approach that forces some elements in $\boldsymbol{\Theta}$ to be zero, the proposed CP decomposition approach provides regularization by imposing sparsity on the total number of rank one matrices to express $\boldsymbol{\Theta}$, leading to a low rank approximation. Therefore, the proposed approach could potentially outperform a simple variable selection approach when the true effect signal in $\boldsymbol{\Theta}$ can be well approximated by a low rank structure. Note that $\boldsymbol{A}\boldsymbol{B}^\top = \boldsymbol{A}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{-1}\boldsymbol{B}^\top$ for any non-singular matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{R\times R}$, $\boldsymbol{A}$ and $\boldsymbol{B}$ are not individually identifiable and therefore lack interpretability. However, $\boldsymbol{\Theta} = \boldsymbol{A}\boldsymbol{B}^\top$ as a whole is identifiable and therefore good mixing and convergence can be achieved for all parameters in $\boldsymbol{\Theta}$. We defer the discussion of the selection of rank $R$ in Section 2.6.

Further, we can re-express $\boldsymbol{A}$ and $\boldsymbol{B}$ with respect to their row vectors. Specifically, let $\tilde{\boldsymbol{\alpha}}_j^\top = (\alpha_{j1}, \ldots, \alpha_{jR})$ denote the $j^{th}$ row of $\boldsymbol{A}$, $j = 1, \ldots, p$ and $\tilde{\boldsymbol{\beta}}_k^\top = (\beta_{k1}, \ldots, \beta_{kR})$ denote the $k^{th}$ row of $\boldsymbol{B}$, $k = 1, \ldots, q$, we can rewrite $\boldsymbol{A} = [\tilde{\boldsymbol{\alpha}}_1, \ldots, \tilde{\boldsymbol{\alpha}}_p]^\top$, $\boldsymbol{B} = [\tilde{\boldsymbol{\beta}}_1, \ldots, \tilde{\boldsymbol{\beta}}_q]^\top$. Then, $\tilde{\boldsymbol{\alpha}}_j$ and $\tilde{\boldsymbol{\beta}}_k$ can be interpreted as representing the effects due to the $j$-th row and $k$-th column

component of the matrix covariate $\boldsymbol{x}_i$, respectively; and the CP decomposition $\boldsymbol{\Theta} = \boldsymbol{A}\boldsymbol{B}^\top$, or its $(j,k)$th element $\Theta_{jk} = \langle \tilde{\boldsymbol{\alpha}}_j, \tilde{\boldsymbol{\beta}}_k \rangle = \sum_{r=1}^{R} \alpha_{jr}\beta_{kr}$ is equivalent to modeling the bilinear two-way interaction effects between the row and column components of the matrix covariate.

*Remark 2.1.* Following Li, Kim and Altman (2010) and Hung and Wang (2013), a data preprocessing step to reduce the dimensionality of the original matrix covariate $\boldsymbol{x}_i$ can be considered before applying our proposed method. For example, when the original matrix covariate $\boldsymbol{x}_i$ can be well approximated by a lower dimensional matrix $\hat{\boldsymbol{x}}_i^* = \boldsymbol{U}^\top \boldsymbol{x}_i \boldsymbol{V} \in \mathcal{R}^{p_0 \times q_0}$ with $p_0 < p$ and $q_0 < q$ through Multilinear Principal Component Analysis (Lu, Plataniotis and Venetsanopoulos (2008)), where $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{p0}) \in \mathbb{R}^{p \times p_0}$ and $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{q0}) \in \mathbb{R}^{q \times q_0}$ are the eigenvector matrices such that $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_{p_0 \times p_0}$ and $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}_{q_0 \times q_0}$, the proposed method can be applied to model the lower dimensional $\hat{\boldsymbol{x}}_i^*$ with the associated coefficient matrix $\boldsymbol{\Theta}^*$ represented by $\boldsymbol{A}^*\boldsymbol{B}^{*\top}$. Finally, the desired coefficient matrix for the original matrix covariate $\boldsymbol{x}_i$ can then be recovered from $\boldsymbol{\Theta} = \boldsymbol{U}\boldsymbol{A}^*\boldsymbol{B}^{*\top}\boldsymbol{V}^\top$. More detailed discussion is given in Section 3.1 in Appendix B in the supplementary material. The intuition behind including an MPCA step in combination with our proposed approach is similar to conducting a principle component regression (PCR). By eliminating noisy and potentially irrelevant and redundant data features in the original space, the extracted MPCA features can be highly informative but take the form of a relatively low dimensional matrix and therefore some estimation efficiency gain would be expected when considering an MPCA step before applying our proposed approach. Further, as discussed in Section 3.1 in Appendix B in the supplementary material, an MPCA step would not artificially make the coefficient matrix follow the assumed low rank in our proposed method. Simulations (presented in Section 3.2 in Appendix B in the supplementary material) also show that the MPCA preprocessing step can improve the efficiency of the proposed estimation method, even when the original matrix predictor is not of extremely large dimensions. However, like PCA, MPCA might not be effective at all times in practice. While there exists no one universal solution for all applications, we stress that the utility of our proposed method for matrix covariates is not related to any data preprocessing step, although an effective dimension reduction preprocessing is likely to further improve efficiency.

2.3. *Specification of priors.* For the parameters in the model (1), we impose diffuse priors: $\eta_0 \sim N(0, \tau_0^2)$, $\eta_1|\eta_0 \sim N(0, \tau_0^2)\mathrm{I}(-\eta_0, \infty)$ with $\tau_0^2 = 100$ and $\sigma^2 \sim$ inverse gamma$(a_0, b_0)$ with $a_0 = b_0 = 0.01$.

For the row and column effect parameters in the CP decomposition in

model (3), we consider the following hierarchical priors,

$$\tilde{\boldsymbol{\alpha}}_1, \ldots, \tilde{\boldsymbol{\alpha}}_p \overset{iid}{\sim} \text{MVN}(\boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha); \text{ and } \tilde{\boldsymbol{\beta}}_1, \ldots, \tilde{\boldsymbol{\beta}}_q \overset{iid}{\sim} \text{MVN}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta). \tag{4}$$

In other words, the parameters representing the row effects, i.e., $\{\tilde{\boldsymbol{\alpha}}_j\}_{j=1}^p$ are assumed to come from the same underlying distribution, which allows borrowing information across different rows when estimating any individual parameter and therefore provides a data-driven method of regularization. The same is applied to the column effects, i.e., $\{\tilde{\boldsymbol{\beta}}_k\}_{k=1}^q$. To complete the specification of these hierarchical priors, we define the following hyper-priors,

$$\boldsymbol{\mu}_\alpha, \boldsymbol{\mu}_\beta \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_0); \text{ and } \boldsymbol{\Sigma}_\alpha, \boldsymbol{\Sigma}_\beta \sim \text{inverse Wishart}(\boldsymbol{S}_0, s_0) \tag{5}$$

For the hyper-parameters in these priors, we let $\boldsymbol{\Sigma}_0 = (9/4)\boldsymbol{I}$; and we assume a diffuse prior for $\boldsymbol{\Sigma}_\alpha$ and $\boldsymbol{\Sigma}_\beta$ with $\boldsymbol{S}_0 = 10\boldsymbol{I}$, and $s_0 = R + 1$. In the case of a rank-one (i.e., $R = 1$) CP decomposition, the parameters in (4) and (5) will be scalars; and accordingly the above Normal-Wishart priors can be replaced by the Normal-Gamma priors. Lastly, for the covariate effect parameters in the probit model, we specify a prior $\boldsymbol{\theta} \sim N(\mathbf{0}, \boldsymbol{V}_0)$, where $\boldsymbol{V}_0 = (9/4)\boldsymbol{I}$. This specification, along with $\boldsymbol{\Sigma}_0 = (9/4)\boldsymbol{I}$ specified above, are chosen in order to bound the probability that $\gamma_i = 1$ in model (2) to be away from 0 and 1, following the suggestion given by Garrett and Zeger (2000) among many others (e.g., Elliott et al. (2005); Neelon, O'Malley and Normand (2011); Jiang et al. (2015)). When a training dataset prior to the analysis of the current dataset is available, an alternative approach is to reset these hyper-parameters based upon the posterior distributions of the parameters using the training dataset.

2.4. *Hierarchical structure specification.* We let $\boldsymbol{\phi}$ include all parameters in models (1) and (3), $\boldsymbol{\phi} = (\eta_0, \eta_1, \sigma^2, \boldsymbol{\theta}, \boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha, \boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$. The unobserved latent variables are denoted by $\boldsymbol{\nu} = (\boldsymbol{\gamma}, \{\tilde{\boldsymbol{\alpha}}_j\}_{j=1}^p, \{\tilde{\boldsymbol{\beta}}_k\}_{k=1}^q)^\top$. The complete data likelihood of $\boldsymbol{\phi}$ (based on the complete data $(\boldsymbol{y}, \boldsymbol{\nu})$ ) is given by

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{\nu}|\boldsymbol{\phi}) = & \prod_{i=1}^n \left[ \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(y_i - \eta_0 - \eta_1\gamma_i)^2}{2\sigma^2} \right\} \left\{ \pi_i^{\mathrm{I}(\gamma_i=1)} (1-\pi_i)^{\mathrm{I}(\gamma_i=0)} \right\} \right] \\ & \times \prod_{j=1}^p \frac{1}{\sqrt{(2\pi)^R |\boldsymbol{\Sigma}_\alpha|}} \exp\left\{ -\frac{1}{2} (\tilde{\boldsymbol{\alpha}}_j - \boldsymbol{\mu}_\alpha)^\top \boldsymbol{\Sigma}_\alpha^{-1} (\tilde{\boldsymbol{\alpha}}_j - \boldsymbol{\mu}_\alpha) \right\} \\ & \times \prod_{k=1}^q \frac{1}{\sqrt{(2\pi)^R |\boldsymbol{\Sigma}_\beta|}} \exp\left\{ -\frac{1}{2} (\tilde{\boldsymbol{\beta}}_j - \boldsymbol{\mu}_\beta)^\top \boldsymbol{\Sigma}_\beta^{-1} (\tilde{\boldsymbol{\beta}}_j - \boldsymbol{\mu}_\beta) \right\} \end{aligned} \tag{6}$$

where $\pi_i = \Phi(\boldsymbol{\theta}^\top \boldsymbol{z}_i + \langle \boldsymbol{A}\boldsymbol{B}^\top, \boldsymbol{x}_i \rangle)$ with $\boldsymbol{A} = [\tilde{\boldsymbol{\alpha}}_1, \ldots, \tilde{\boldsymbol{\alpha}}_p]^\top$ and $\boldsymbol{B} = [\tilde{\boldsymbol{\beta}}_1, \ldots, \tilde{\boldsymbol{\beta}}_q]^\top$.

2.5. *Posterior computation.* First, note that $\left\langle \boldsymbol{A}\boldsymbol{B}^\top, \boldsymbol{x}_i \right\rangle$ in the probit model (3) can be rewritten as a linear function with respect to $\tilde{\boldsymbol{\alpha}}_1, \ldots, \tilde{\boldsymbol{\alpha}}_p$ or $\tilde{\boldsymbol{\beta}}_1, \ldots, \tilde{\boldsymbol{\beta}}_q$ as follows,

$$\left\langle \boldsymbol{A}\boldsymbol{B}^\top, \boldsymbol{x}_i \right\rangle = \sum_{j=1}^{p} \tilde{\boldsymbol{\alpha}}_j^\top \boldsymbol{u}_{ij} = \sum_{k=1}^{q} \tilde{\boldsymbol{\beta}}_k^\top \boldsymbol{v}_{ik} \tag{7}$$

where $\boldsymbol{u}_{ij}^\top$ denotes the $j^{th}$ row of $\boldsymbol{x}_i\boldsymbol{B} \in \mathbb{R}^{p\times R}$, $j = 1, \ldots, p$ and $\boldsymbol{v}_{ik}^\top$ denotes the $k^{th}$ row of $\boldsymbol{x}_i^\top \boldsymbol{A} \in \mathbb{R}^{q\times R}$, $k = 1, \ldots, q$. This suggests that $\{\tilde{\boldsymbol{\alpha}}_j\}_{j=1}^p$ and $\{\tilde{\boldsymbol{\beta}}_k^\top\}_{k=1}^q$ can be updated iteratively in a similar fashion as in a regular regression model. With the data augmentation algorithm of Albert and Chib (1993) for our binary probit model, the posterior computation becomes straightforward with Gibbs sampling. Specifically, we introduce a latent variable $w_i$ such that $\gamma_i = \mathrm{I}(w_i > 0)$ and $w_i \sim \mathrm{N}(\boldsymbol{\theta}^\top \boldsymbol{z}_i + \langle \boldsymbol{A}\boldsymbol{B}^\top, \boldsymbol{x}_i \rangle, 1)$. The detailed MCMC algorithm is given in Web Appendix A in the supplementary material.

2.6. *Rank selection.* We follow Zhou, Li and Zhu (2013) to formulate this task as a model selection problem. Given the hierarchical structure in our model, we adopt a more recent model selection criteria, Watanabe-Akaike information criterion (WAIC), as recommended by Gelman, Hwang and Vehtari (2014) for Bayesian hierarchical models. As a generalization of AIC (Akaike, 1974), WAIC was derived based on singular learning theory (Watanabe, 2010) as an asymptotically unbiased approximation to out of sample prediction error. Importantly, WAIC is straightforward to compute based on posterior draws without the need to adjust for the effective number of parameters in hierarchical models. Gelman, Hwang and Vehtari (2014) discussed the Bayesian aspects of model selection and concluded that while cross-validation is their preferred method, WAIC offers a computationally convenient alternative to it. In a Bayesian setting, another commonly used cross validation based criterion to assess the model's predictive performance is the logarithm of the pseudo marginal likelihood (LPML, see Geisser and Eddy, 1979; Gelfand and Dey, 1994). For a thorough discussion of Bayesian predictive model assessment methods, please see Vehtari and Ojanen (2012).

WAIC is defined based on the observed data ($\boldsymbol{y}$) likelihood, given all model parameters $\boldsymbol{\phi}$ and latent variables $\boldsymbol{\nu}$, denoted by $f(y_i \mid \boldsymbol{\nu}, \boldsymbol{\phi})$ and then adds

a penalty term to correct for model complexity,

$$\text{WAIC} = -2\sum_{i=1}^{n}\log\left[\mathrm{E}_{\boldsymbol{\nu},\boldsymbol{\phi}}\left\{f(y_i \mid \boldsymbol{\nu},\boldsymbol{\phi}) \mid \boldsymbol{y}\right\}\right] + 2p_{\text{WAIC}}$$

where, for our models $f(y_i \mid \boldsymbol{\nu},\boldsymbol{\phi}) = (2\pi\sigma^2)^{-1/2}\exp\left\{-(y_i-\eta_0-\eta_1\gamma_i)^2/2\sigma^2\right\}$. The penalty term $p_{\text{WAIC}}$ is defined as,

$$p_{\text{WAIC}} = 2\sum_{i=1}^{n}\left\{\log\left[\mathrm{E}_{\boldsymbol{\nu},\boldsymbol{\phi}}\left\{f(y_i \mid \boldsymbol{\nu},\boldsymbol{\phi}) \mid \boldsymbol{y}\right\}\right] - \mathrm{E}_{\boldsymbol{\nu},\boldsymbol{\phi}}\left\{\log f(y_i \mid \boldsymbol{\nu},\boldsymbol{\phi}) \mid \boldsymbol{y}\right\}\right\}.$$

As indicated by these expressions, WAIC can be obtained from its Monte Carlo estimate by averaging over posterior draws of $\boldsymbol{\nu}$ and $\boldsymbol{\phi}$.

2.7. *Prediction of future samples.* It is clinically useful to obtain the probability of being an early responder with associated prediction uncertainty for a future subject prior to treatment. Knowing a patient's likelihood to improve without an active chemical drug can guide an initial treatment decision, for example, replacing routine chemical drug treatment by a treatment with less severe side effects. Specifically, the prediction can be obtained as follows. For a future sample with baseline covariates $\{\boldsymbol{x}^{new}, \boldsymbol{z}^{new}\}$, the posterior predictive probability of being an early responder, i.e., $\gamma^{new} = 1$ is given by,

$$\begin{aligned} P_{new} &= \Pr(\gamma^{new}=1|\boldsymbol{x}^{new},\boldsymbol{z}^{new},\boldsymbol{y},\boldsymbol{x},\boldsymbol{z}) \\ &= \int \Pr(\gamma^{new}=1|\boldsymbol{x}^{new},\boldsymbol{z}^{new},\tilde{\boldsymbol{\alpha}},\tilde{\boldsymbol{\beta}},\boldsymbol{\phi}) f(\tilde{\boldsymbol{\alpha}},\tilde{\boldsymbol{\beta}},\boldsymbol{\phi}|\boldsymbol{y},\boldsymbol{x},\boldsymbol{z}) d\boldsymbol{\phi} d\tilde{\boldsymbol{\alpha}} d\tilde{\boldsymbol{\beta}} \\ &= \int \Phi(\boldsymbol{\theta}^\top \boldsymbol{z}^{new} + \langle \boldsymbol{A}\boldsymbol{B}^\top, \boldsymbol{x}^{new}\rangle) f(\tilde{\boldsymbol{\alpha}},\tilde{\boldsymbol{\beta}},\boldsymbol{\phi}|\boldsymbol{y},\boldsymbol{x},\boldsymbol{z}) d\boldsymbol{\phi} d\tilde{\boldsymbol{\alpha}} d\tilde{\boldsymbol{\beta}} \end{aligned} \tag{8}$$

where $\tilde{\boldsymbol{\alpha}}$, $\tilde{\boldsymbol{\beta}}$ and $\boldsymbol{\phi}$ are the vectorized versions of all $\tilde{\boldsymbol{\alpha}}_j^\top$, $j = 1,\ldots,p$, all $\tilde{\boldsymbol{\beta}}_k^\top$, $k = 1,\ldots,q$ and all model parameters, respectively.

With the MCMC posterior samples $\boldsymbol{\theta}^{(m)}$, $\{\tilde{\boldsymbol{\alpha}}_j^{(m)}\}_{j=1}^p$ and $\{\tilde{\boldsymbol{\beta}}_k^{(m)}\}_{k=1}^q$, $m = 1,\ldots,M$, conditional on the data $\{\boldsymbol{y},\boldsymbol{x},\boldsymbol{z}\}$, the quantity $P_{new}$ at the $m^{th}$ MCMC iteration is given by,

$$p_{new}^{(m)} = \Phi(\boldsymbol{\theta}^{(m)T}\boldsymbol{z}^{new} + \langle \boldsymbol{A}^{(m)}\boldsymbol{B}^{(m)T}, \boldsymbol{x}^{new}\rangle), \tag{9}$$

where $\boldsymbol{A}^{(m)} = \left[\tilde{\boldsymbol{\alpha}}_1^{(m)},\ldots,\tilde{\boldsymbol{\alpha}}_p^{(m)}\right]^\top$ and $\boldsymbol{B}^{(m)} = \left[\tilde{\boldsymbol{\beta}}_1^{(m)},\ldots,\tilde{\boldsymbol{\beta}}_q^{(m)}\right]^\top$. Then $P_{new}$ can be estimated by $M^{-1}\sum_{m=1}^{M}\Phi(\boldsymbol{\theta}^{(m)T}\boldsymbol{z}^{new} + \langle \boldsymbol{A}^{(m)}\boldsymbol{B}^{(m)T}, \boldsymbol{x}^{new}\rangle)$, and the associated uncertainty can be quantified by the corresponding credible interval.

**3. Simulations.** In this section, we describe several simulation studies to evaluate the performance of our proposed method, focusing on two aspects: 1) estimation of the coefficient matrix $\mathbf{\Theta}$ and 2) prediction accuracy of the latent class indicators for both the within and out-of samples. In our first study, we investigate how performances may be affected by different true rank values, dimensions of the matrix covariate and sample sizes, when the rank $R$ is correctly assumed and the two latent classes in the manifest model (1) are well separated. In our second study, we evaluate the robustness of our proposed method when the true rank of $\mathbf{\Theta}$ is equal to the assumed rank, and when it is not, under different degrees of overlap between the two latent classes in the manifest model (1).

For all simulation scenarios (for selected $p, q, R, \eta_0$ and $\eta_1$), the observed data $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$ and the latent class indicator $\gamma_i$, $i = 1, ..., n$, are generated as follows:

1. each element in $\boldsymbol{x}_i$, $\{\boldsymbol{x}_i\}_{j,k} \overset{iid}{\sim}$ uniform$(-1, 1)$, $j = 1, \ldots, p$ and $k = 1, \ldots, q$;
2. $\boldsymbol{z}_i = (1, z_{i1})^\top$ with $z_{i1} \sim$ uniform$(0, 1)$;
3. $\mathbf{\Theta}$ is generated as follows
   (a) let $\boldsymbol{\mu}_\alpha = \boldsymbol{\mu}_\beta = (0, \ldots, 0)^\top$ and $\mathbf{\Sigma}_\alpha = \mathbf{\Sigma}_\beta$ be diagonal with all diagonal elements equal to $0.5^2$; generate $\tilde{\boldsymbol{\alpha}}_j \overset{iid}{\sim} N(\boldsymbol{\mu}_\alpha, \mathbf{\Sigma}_\alpha)$, $j = 1, \ldots, p$ and $\tilde{\boldsymbol{\beta}}_k \overset{iid}{\sim} N(\boldsymbol{\mu}_\beta, \mathbf{\Sigma}_\beta)$, $k = 1, \ldots, q$;
   (b) set $\boldsymbol{A} = [\tilde{\boldsymbol{\alpha}}_1, \ldots, \tilde{\boldsymbol{\alpha}}_p]^\top$ and $\boldsymbol{B} = [\tilde{\boldsymbol{\beta}}_1, \ldots, \tilde{\boldsymbol{\beta}}_q]^\top$, then $\mathbf{\Theta} = \boldsymbol{A}\boldsymbol{B}^\top$.

   Note: when the true rank $R = 1$, the parameters in (a) are scalars and are generated from the corresponding univariate Normal distributions.
4. $\gamma_i$ is generated from model (3) given $\boldsymbol{x}_i$, $\boldsymbol{z}_i$ and $\mathbf{\Theta}$, where $\boldsymbol{\theta} = (0, 0.2)^\top$.
5. $y_i$ is generated from model (1) given $\gamma_i$, where $\eta_0 = 0$ and $\sigma = 0.2$; the value of $\eta_1$ is varied in different scenarios.

We have followed other applied work in the Bayesian literature to simulate $S = 100$ data sets corresponding to 100 draws of $\mathbf{\Theta}$ for each of the simulation scenarios. For each generated data set, we obtain the posterior samples of all model parameters using the Gibbs sampling algorithm described in the Section 2.5, retaining every $10^{th}$ draw from $150,000$ iterations after a burn-in period of $25,000$ iterations.

To assess the performance on the estimation of the coefficient matrix $\mathbf{\Theta}$, we obtain the overall mean squared error (MSE) based on the $S$ simulated

data sets, defined as follows,

$$\text{MSE} = \frac{1}{S}\sum_{s=1}^{S}\left\{\frac{1}{pq}\|\hat{\boldsymbol{\Theta}}^{(s)} - \boldsymbol{\Theta}^{(s)}\|_F^2\right\}$$

where $\|\cdot\|_F^2$ represents the Frobenius norm. $\boldsymbol{\Theta}^{(s)}$ is the true coefficient matrix from the $s^{th}$ simulated data set and $\hat{\boldsymbol{\Theta}}^{(s)}$ is its posterior mean estimate.

There are many performance measures to evaluate the prediction accuracy of binary classifications, including sensitivity, specificity, $F_1$-score (also known as F-score or F-measure), Matthews correlation coefficient (MCC, see Matthews, 1975) and area under the curve (AUC) of the receiver operating characteristic (ROC). Each measure has its own advantages and disadvantages under different situations, see Powers (2011) for an extensive discussion. In this paper, we consider the widely used AUC measure to quantify the accuracy of predicting the binary latent class indicators. Specifically, the posterior mean AUC is obtained by averaging the AUC values calculated across all MCMC iterations using the ROCR package in R (Sing et al. 2005). The reported AUCs are then the average posterior mean AUCs across $S$ simulated data sets. Specifically, for each simulated data set $\{\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}\}$ of size $n$ (with $n$ varying in different simulation scenarios), we also generate an additional validation data set of size $\tilde{n} = 50$ with baseline covariates $\{\boldsymbol{x}^{new}, \boldsymbol{z}^{new}\}$ to evaluate the out-of-sample predictive accuracy. The within sample AUC is obtained based on $p(\gamma_i = 1|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z})$, $i = 1, \ldots, n$; and the out-of-sample AUC is obtained based on $p(\gamma_i^{new} = 1|\boldsymbol{x}_i^{new}, \boldsymbol{z}_i^{new}, \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z})$, $i = 1, \ldots, \tilde{n}$, which can be computed from (8) as described in Section 2.7.

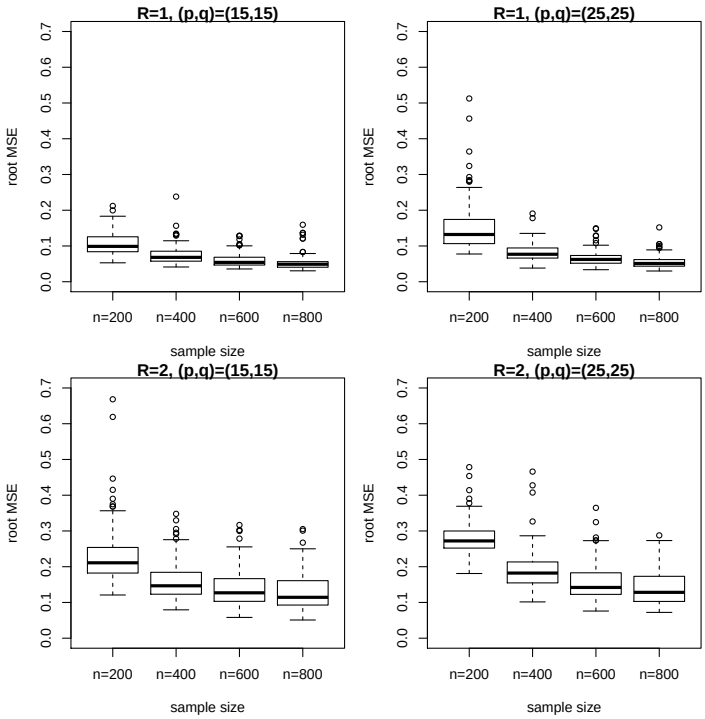
3.1. *Study 1.* In this section, we let $\eta_1 = 1.0$ in the manifest model (1) so that the two latent classes defined by $\gamma_i = 0$ and $\gamma_i = 1$ are well separated. Under this setup, we consider the rank $R \in \{1, 2\}$, the dimension of the matrix covariate $\boldsymbol{x}_i \in \mathbb{R}^{p\times q}$, $(p, q) \in \{(15, 15), (25, 25)\}$, and the sample size $n \in \{200, 400, 600, 800\}$, leading to a total of 16 scenarios. For each combination of the rank $R$ and the dimension $(p, q)$, we simulate 100 sets of the coefficient matrix $\boldsymbol{\Theta}$, based on which, we generate 100 data sets $\{(\boldsymbol{y_i}, \boldsymbol{x_i}, \boldsymbol{z_i}) : i = 1, \ldots, n\}$ for $n = 200, 400, 600$ and 800 respectively; that is, the 100 simulated sets of $\boldsymbol{\Theta}$ are common to all the 400 data sets under 4 different sample sizes.

Figure 2(a) shows the boxplots of the root MSEs of $\hat{\boldsymbol{\Theta}}$ across 100 simulations, for all 16 scenarios. Overall, we see that the estimation accuracy for $\hat{\boldsymbol{\Theta}}$ improves when the sample size increases. The results show that, relative to estimating a lower rank coefficient matrix, estimating a higher rank coefficient matrix requires a relatively larger sample size to achieve the same

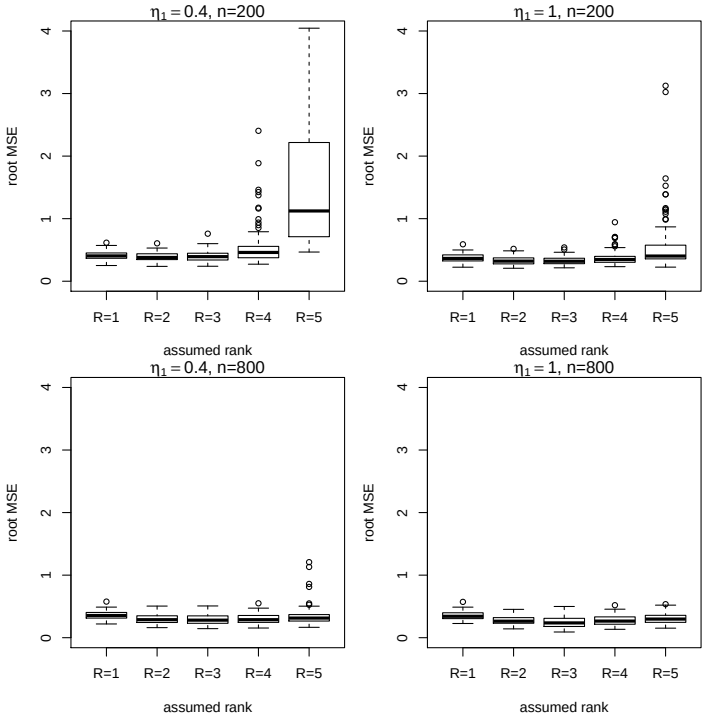
estimation accuracy as measured by the overall MSE of $\hat{\mathbf{\Theta}}$, in comparison to the estimation of a lower rank coefficient matrix. This is not surprising given the increase of the number of model parameters. However, increasing the dimensions of the matrix covariate from $(15, 15)$ to $(25, 25)$ results in very little deterioration in performance for $\hat{\mathbf{\Theta}}$, due to regularization imposed by the hierarchical modeling of the coefficient matrix. In fact, for the rank $R = 1$ case, the increases in the root MSE when $(p, q) = (25, 25)$ versus $(p, q) = (15, 15)$ are only 0.046, 0.008, 0.006, 0.002 for sample size $n = 200, 400, 600$ and 800, respectively; for rank $R = 2$ case, such increases in the root MSE are 0.041, 0.032, 0.020 and 0.012, respectively.

Next we turn our attention to evaluating the accuracy in predicting the latent class indicator $\gamma_i$ for both within-sample and out-of-sample. As shown in Table 1, the within sample AUC values are all $1's$ for all 16 simulation scenarios, suggesting perfect within sample prediction, partly due to fairly high separation in the two latent classes. The out-of-sample AUC values are all high and only slightly smaller than their within sample versions. This is because the out-of-sample prediction is solely dependent on the matrix covariate without relying on information from the clinical outcome and reflects how accurately the coefficient matrix can be estimated under different scenarios. Specifically, for fixed true rank $R$ and dimension $(p, q)$, the out-of-sample AUC values increase as the sample size increases; and for fixed sample size $n$ and dimension $(p, q)$, the out-of-sample AUC values for the true rank $R = 2$ cases are slightly smaller than the true rank $R = 1$ case. These results are consistent with the conclusions for the estimation of the matrix coefficient. More notably, by assuming a larger dimension $(p, q) = (25, 25)$ compared to dimension $(p, q) = (15, 15)$, the out-of-sample prediction accuracy is only reduced for the scenario when true rank $R = 2$ and sample size $n = 200$; the improved out-of-sample prediction accuracy under all other scenarios is likely due to more information brought in by assuming a larger dimensional matrix covariate with strong signals.

3.2. *Study 2.* In this section, we study the impact on the performance of our proposed method when the true rank of $\mathbf{\Theta}$ is either correctly specified or misspecified and when the two latent classes in the model (1) are either overlapping by letting $\eta_1 = 0.4$, or well separated by letting $\eta_1 = 1.0$. We consider two sample sizes $n \in \{200, 800\}$. For the matrix covariate $\boldsymbol{x}_i \in \mathbb{R}^{p \times q}$, we let $(p, q) = (15, 15)$, and the true rank $R = 3$ for the associated coefficient matrix $\mathbf{\Theta}$. We simulate 100 replicates of the coefficient matrix $\mathbf{\Theta}$ with $R = 3$ and $(p, q) = (15, 15)$, which are used in all the simulation scenarios to generate the data sets. Next, for each sample size $n$, we generate




(a) Study 1: the degree of overlapping between the two latent subgroups is fixed by letting $\eta_1 = 1.0$; true values for $R$, $(p, q)$, and $n$ vary under different simulation scenarios.




(b) Study 2: the true rank $R = 3$, $(p, q) = (15, 15)$, and $n = 200$ or 800; and $\eta_1 = 0.4$ and $\eta_1 = 1.0$ indicate high and low degrees of overlapping between the two latent subgroups, respectively. The models are fit with varying assumed rank values.

FIG 2. *Boxplots of the root mean squared errors (MSEs) of the coefficient matrix estimate* $\hat{\mathbf{\Theta}}$ *across 100 simulations, from study 1 and 2 respectively.*

100 replicates of $\{(\boldsymbol{x_i}, \boldsymbol{z_i}, \gamma_i) : i = 1, \ldots, n\}$, and based on which, we generate 100 replicates of the clinical outcome $\{y_i : i = 1, \ldots, n\}$ for $\eta_1 \in \{0.4, 1.0\}$ respectively; that is, the 100 simulated replicates of $\{(\boldsymbol{\Theta}, \boldsymbol{x}_i, \boldsymbol{z}_i, \gamma_i) : i = 1, \ldots, n\}$ are common to all the 200 data sets under 2 different degrees of overlapping between the two latent classes. For each simulated data set under all 4 simulation scenarios, we fit five models assuming the rank of $\boldsymbol{\Theta}$ being $R = 1$ to 5.

TABLE 1

*The mean Area Under the ROC curves (AUC) for both within and out-of-samples across 100 simulations, from study 1 and 2 respectively.*

(a) Study 1: the degree of overlapping between the two latent subgroups is fixed by letting $\eta_1 = 1.0$; true values for $R$, $(p, q)$, and $n$ vary under different simulation scenarios.

| | **true rank** $R = 1$ | | | | **true rank** $R = 2$ | | | |
|---|---|---|---|---|---|---|---|---|
| | $n = 200$ | $n = 400$ | $n = 600$ | n=800 | $n = 200$ | $n = 400$ | $n = 600$ | n=800 |
| | | | | **within sample AUC** | | | | |
| $(p, q) = (15, 15)$ | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| $(p, q) = (25, 25)$ | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | | | **out of sample AUC** | | | | |
| $(p, q) = (15, 15)$ | 0.86 | 0.90 | 0.90 | 0.91 | 0.80 | 0.88 | 0.91 | 0.92 |
| $(p, q) = (25, 25)$ | 0.87 | 0.93 | 0.94 | 0.95 | 0.74 | 0.88 | 0.92 | 0.94 |

(b) Study 2: the true rank $R = 3$, $(p, q) = (15, 15)$, and $n = 200$ or 800; and $\eta_1 = 0.4$ and $\eta_1 = 1.0$ indicate high and low degrees of overlapping between the two latent subgroups, respectively. The models are fit with varying assumed rank values.

| | $\eta_1 = 0.4$ | | | | | $\eta_1 = 1.0$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| assumed rank | R=1 | R=2 | R=3 | R=4 | R=5 | R=1 | R=2 | R=3 | R=4 | R=5 |
| | | | | | **within sample AUC** | | | | | |
| n=200 | 0.90 | 0.89 | 0.87 | 0.84 | 0.80 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| n=800 | 0.94 | 0.96 | 0.96 | 0.95 | 0.93 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | | | | **out of sample AUC** | | | | | |
| n=200 | 0.59 | 0.63 | 0.63 | 0.63 | 0.63 | 0.71 | 0.75 | 0.75 | 0.73 | 0.72 |
| n=800 | 0.80 | 0.84 | 0.84 | 0.81 | 0.79 | 0.82 | 0.88 | 0.88 | 0.86 | 0.84 |

Figure 2(b) summarizes the root MSEs of $\hat{\boldsymbol{\Theta}}$ across 100 simulations for each of the assumed models with varying ranks, under the 4 total simulation scenarios. When the sample size is large, the root MSE of $\hat{\boldsymbol{\Theta}}$ achieves the minimum at the true rank value 3, and slightly increases as the assumed rank is either increased beyond or decreased below this true rank value. This U-shaped trend in these root MSEs also suggests that our proposed hierarchical modeling approach for $\boldsymbol{\Theta}$ is robust to over-fitting regardless of the overlap between the latent classes. When the sample is small, such U-shaped trend in estimating $\boldsymbol{\Theta}$ is not as obvious, with similar performance at the true rank value or its adjacent rank values. For either sample size, as

indicated by these boxplots in Figure 2(b), more overlapping in the latent classes leads to slightly larger MSE of $\hat{\boldsymbol{\Theta}}$, due to difficulty in separating the two latent classes.

Since the true rank of $\boldsymbol{\Theta}$ is generally not known in practice, we next report the robustness of our proposed method in the case of incorrectly assuming the rank $R$. Table 1 reports the AUC values for both the within and out-of-samples under the high and low degrees of latent class overlapping scenarios, respectively for each sample size. In general, the predictive accuracy for both within and out-of-samples improves if the two latent classes are less overlapping. As expected, we see an indication for a U shape in these out-of-sample AUC values by fitting models with varying assumed ranks. However, when fitting a model with assumed rank being close enough to the true rank, the out-of-sample AUC values suggest little or no loss of predictive power under misspecified rank. In fact, under both simulation scenarios, assuming one rank lower than the true rank leads to the same out-of-sample AUC value as that by assuming the true rank. These investigations suggest that the hierarchical modeling approach for $\boldsymbol{\Theta}$ proposed here provides good robustness against misspecified rank regardless of how much the two latent classes overlap.

**4. Application to identify early responder subgroup using EEG data.** In this section, we present an analysis of the data introduced in Section 1. One study goal is to determine to what extent the resting state EEG alpha and theta power (i.e., indicating neural activity in the frequency ranges for alpha and theta waves) in the posterior region of brain under a closed eyes condition could help identify a potential early responder subgroup (which is believed to consist of subjects susceptible to non-specific placebo effects), given the outcome collected early in the course of treatment. Specifically, we use the model defined in (1) to describe the bimodal pattern in the change in HAM-D (baseline - week 1) (Figure 1), corresponding to two subgroups, defined by whether or not subjects demonstrate an early response, and the hierarchical model defined in (2) to relate the EEG measurements to the likelihood of responding early.

For the 96 study subjects we let $y_i$ denote the change in HAM-D (baseline - week 1), where a positive change indicates diminished symptom severity; and we let $\boldsymbol{x}_i^*$ denote the EEG measurement that takes the form of a $14 \times 45$ matrix. Before applying our proposed method, we performed the MPCA step as discussed in Section 2 to our EEG matrix covariate. This step plays an important role in removing some noisy and irrelevant information in our original EEG data, while attempts to directly apply the proposed method

to the original EEG data resulted in unstable estimate of the coefficient matrix. Specifically, MPCA procedure seeks to find two eigenvector matrices $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{p0}) \in \mathbb{R}^{p \times p_0}$ and $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{q0}) \in \mathbb{R}^{q \times q_0}$ respectively, such that they minimize the Frobenius norm loss $n^{-1} \sum_i^n \|\boldsymbol{x}_i^* - \hat{\boldsymbol{x}}_i^*\|_F^2$ between $\boldsymbol{x}_i^*$ and its lower-dimensional representation $\hat{\boldsymbol{x}}_i^* = \boldsymbol{U}^\top \boldsymbol{x}_i^* \boldsymbol{V} \in \mathbb{R}^{p_0 \times q_0}$, where $p_0 < p$ and $q_0 < q$. The explained proportion of total variation in $\boldsymbol{x}_i^*$ represented by $\hat{\boldsymbol{x}}_i^*$ is defined by $\sum_{i=1}^n \|\hat{\boldsymbol{x}}_i^* - \bar{\boldsymbol{x}}^*\|_F^2 / \sum_{i=1}^n \|\boldsymbol{x}_i^* - \bar{\boldsymbol{x}}^*\|_F^2$, where $\bar{\boldsymbol{x}}^* = n^{-1} \sum_{i=1}^n \boldsymbol{x}_i^*$. In our analysis, the MPCA dimensions $(p_0, q_0)$ was chosen via WAIC; see Table 2 for the illustration. When applying the MPCA method, we use the rTensor package in R software that implements the general algorithm by Lu, Plataniotis and Venetsanopoulos (2008) and consider $p_0 \in \{2, \ldots, 10\}$ and $q_0 \in \{2, \ldots, 6\}$. The resulting MPCA features (taking the form of a $p_0 \times q_0$ matrix) can be highly correlated and the range of their values is over different orders of magnitude, which may result in slow or no convergence in the MCMC algorithm. To prevent this from happening, we apply the common trick in applied regression by standardizing all $p_0 q_0$ elements in $\hat{\boldsymbol{x}}_i^*$ to have mean 0 and standard deviation 1. The resulting matrix covariate, denoted by $\boldsymbol{x}_i$, from this step is used in model (3). To emphasize the importance of regularization provided by our proposed method in our case, Web Figure 3 presents the trace plots of four random selected coefficients in our final model when fitting the total $p_0 q_0$ MPCA features directly without imposing any assumption, where convergence is not achieved. In contrast, as discussed below, applying our proposed method on the resulting MPCA extracted matrix covariate led to stable results and discovery of important EEG features.

We also adjust for additional baseline covariates in model (2) by letting $z_1$ =gender (1 for female; 0 for male) and $z_2$ =depression chronicity (1 for being depressed for 24 months or more in the past 4 to 5 years; 0 otherwise). For each combination of $(p_0, q_0)$, we fit models assuming the rank in model (2) both setting $R = 1$ and $R = 2$. For all these models considered here, we ran two MCMC chains of 175,000 iterations to reduce Monte Carlo errors, with the initial 25,000 iterations discarded as burn-in, and retained every $10^{th}$ draw to reduce autocorrelation. Convergence of the chains was assessed using the Gelman-Rubin statistic $\hat{R}$ (Gelman and Rubin 1992). The maximum value among all model parameters was less than 1.1, indicating convergence. To provide additional evidence for convergence, Web Figure 1 in the Web Appendix B in the supplementary material included the trace plots for 4 randomly selected coefficients in $\boldsymbol{\Theta}$ from our real data analysis.

Table 2 presents the WAIC statistics for these models, where all these 2-component mixture models fit the data much better than a 1-component

TABLE 2
*WAIC from fitting different models for the prediction of the early responder subgroup using EEG data under different choices of MPCA dimensions in both row ($p_0$) and column ($q_0$) directions.*

| | **rank R=1** | | | | | **rank R=2** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $p_0$ / $q_0$ | 2 | 3 | 4 | 5 | 6 | 2 | 3 | 4 | 5 | 6 |
| 2 | 603.4 | 605.5 | 601.3 | 600.2 | 601.9 | 604.4 | 604.6 | 601.0 | 601.3 | 599.1 |
| 3 | 602.8 | 601.2 | 594.2 | 593.9 | 594.8 | 604.0 | 599.2 | 596.2 | 596.5 | 603.8 |
| 4 | 602.7 | 603.9 | 583.7 | 590.2 | 595.1 | 604.1 | 601.7 | 595.1 | 596.0 | 602.0 |
| 5 | 601.1 | 602.4 | 591.6 | 598.0 | 598.1 | 603.5 | 602.3 | 598.5 | 597.8 | 598.0 |
| 6 | 598.1 | 600.1 | 577.3 | 577.9 | 582.1 | 600.6 | 601.7 | 589.9 | 588.6 | 588.7 |
| 7 | 582.9 | 591.1 | 571.9 | 578.7 | 578.7 | 599.2 | 595.4 | 588.1 | 593.9 | 596.2 |
| 8 | 584.7 | 589.4 | 573.3 | 574.7 | 575.0 | 599.3 | 597.3 | 590.1 | 589.8 | 594.0 |
| 9 | 587.1 | 594.4 | **568.8** | 573.1 | 573.5 | 602.3 | 600.6 | 591.8 | 594.4 | 592.8 |
| 10 | 588.4 | 594.2 | 571.1 | 574.1 | 583.8 | 600.5 | 600.0 | 590.6 | 588.8 | 590.3 |

model (WAIC = 845.6 for 1 component). In particular, WAIC suggests that the model with $(p_0, q_0) = (9, 4)$, which extracted 98.6% of the major variation in our original EEG measurements, and rank $R = 1$ offers the best balance between goodness-of-fit and model complexity. Under this best fitting model, we classify subjects into one of the two subgroups based on the maximum posterior estimate of $p(\gamma_i|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z})$. Specifically, 16 subjects (17%) were classified to the early responder subgroup (the likely cause of the positive skewness seen in Figure 1), with the change in HAM-D (baseline - week1) centering at 9.10 (95% CI: 5.44, 12.37), while the other 80 subjects (83 %) were assigned to the other subgroup, with the change in HAM-D (baseline - week1) centering at 0.96 (95% CI: -0.18, 2.08). Note that this CI contains zero, which is consistent with what one would expect for subjects without a placebo response before a drug response begins to take effect. Figure 3 (a) shows the posterior density of the probability for subjects assigned to these two subgroups, respectively. The clear separation in these two distributions indicates that our model is effective in identifying these two postulated subgroups. Further, a chi-squared test indicates no significant difference in the proportion of early responders for the placebo arm and drug arm, with the proportion being 9/50=18% and 7/46=15% respectively. This provides supporting evidence that any improvement seen at week 1 is more likely due to non-specific placebo effects and that a specific drug effect is not evident by week 1.

In terms of predicting an early responder, our results suggest that chronically depressed patients are less likely to be early responders, with $\hat{\theta}_2 = -1.65$ (95% CI: -3.59, -0.12), while gender is not a contributing factor, with $\hat{\theta}_1 = -1.34$ (95% CI: -3.14, 0.02). Figure 3 (b) shows the posterior density of $\langle \boldsymbol{\Theta}, \boldsymbol{x}_i \rangle$ for subjects assigned to the early responder and non-responder subgroups, respectively. It clearly illustrates the usefulness of EEG measures

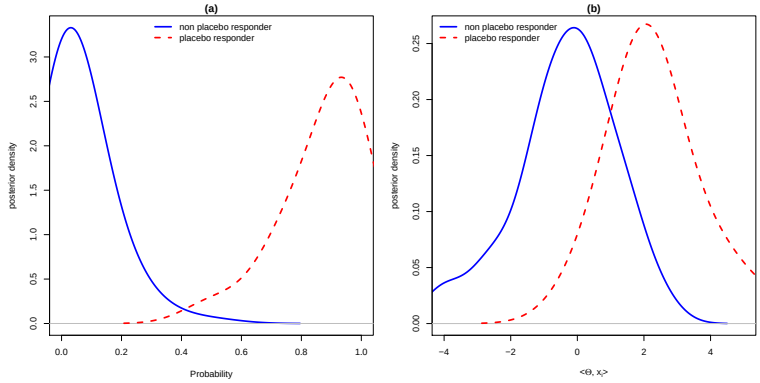



FIG 3. *Results of using EEG data to predict the early responder subgroup. Left: posterior density estimate of the probability of being in the early responder subgroup; Right: posterior density estimate of* $\langle \boldsymbol{\Theta}, \boldsymbol{x}_i \rangle$*, for the early responder and non-responder subgroups, respectively.*

in distinguishing the two subgroups.

Further, by mapping the coefficient matrix estimate $\hat{\boldsymbol{\Theta}} \in \mathbb{R}^{9\times 4}$ on the reduced feature space obtained by MPCA to the original space, we obtain the coefficient estimate at each combination of electrode locations crossed with frequency ranges. For the posterior estimates of $\Theta$ on the reduced MPCA feature space, please refer to Web Table 1 in the supplementary material. As shown in Figure 4, the heat maps for the coefficient matrix look very similar for ranks $R = 1$ and $R = 2$ models, which is consistent with our findings in the simulation studies. However, most of the estimates have much wider credible intervals (and hence are no longer statistically significant) under the assumption of $R = 2$. Based on our best fitting $R = 1$ model, the EEG CSD levels at the electrode locations "P7", "P9", "PO4" and "POZ" through most theta/alpha frequency ranges are found to play significantly important roles in predicting the membership in the early responder subgroup. This finding is consistent with the scientific hypothesis that EEG alpha and theta power recorded in the brain posterior region might be useful to identify potential early responders for patients with MDD and largely agrees with existing literature on EEG theta/alpha powers as predictors to antidepressant response (Wade and Iosifescu 2016). Additionally, we used a Bayes factor to compare a model with and without the EEG measurements. Using the approach of Chib (1995), the log marginal likelihood for the model using EEG measurements is estimated to be $-124$ compared to $-281$ in a model with no EEG measurements, producing a large Bayes factor, exp(157). This

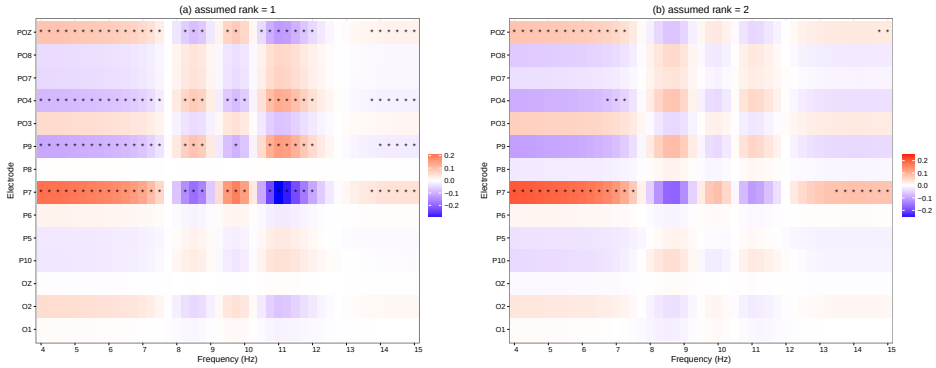



FIG 4. *Heat map of the estimated coefficient matrix for the prediction of the early responder subgroup using EEG data (* indicates significance at the 0.05 level): (a) rank $R = 1$ model; (b) rank $R = 2$ model.*

provides further evidence of the usefulness of EEG measurements for differentiating between early responders and early non-responders. In contrast to most common practices in the EEG literature, where the focus is on a univariate (i.e., scalar) predictor equal to the mean measure across one or a small subset of frequency bands, our approach can directly accommodate matrix-valued EEG data. In particular, our analysis found that theta/alpha power at "POZ" and "P7" have an inverse association with non-specific effect in comparison to the "PO4" and "P7" locations, further suggesting that the predictive values within different frequency ranges could be different. In fact, Ciarleglio et al. (2015) reported that EEG CSD measures at different frequency ranges within the theta/alpha band also predict differential response to treatment with sertraline versus placebo.

**5. Discussion.** In this paper, we have considered a regression extension of existing latent class models to incorporate matrix covariates as predictors of the latent subgroup membership. This research is motivated by a placebo controlled clinical trial to investigate whether the baseline matrix-valued EEG alpha and theta powers are associated with an early placebo responder subgroup inferred from the change in HAM-D scores (baseline - week1).

Our approach utilizes the powerful low-rank CP decomposition to achieve a bilinear representation of the target coefficient matrix. Specifically, such a CP matrix decomposition factorizes the coefficient matrix into row and column components and assumes a multiplicative form among them. For parameter estimation, we adopt a Bayesian hierarchical modeling approach that provides both a flexible way to incorporate a priori assumptions and a

data-driven method of regularization. Further, the simulation studies show that the proposed hierarchical approach is robust against rank misspecification in the sense that although the estimation of $\mathbf{\Theta}$ generally achieves the minimum MSE when the true rank is known, our approach leads to very stable estimates of $\mathbf{\Theta}$ across different choices of the assumed rank.

Using the proposed approach for our motivating data set, we are able to identify specific posterior regions at certain alpha and theta frequency ranges in the EEG CSD levels that are predictive of being in the early placebo responder subgroup. This finding raises hope for using EEG measures to differentiate potential early responders from non-responders in clinical practice to further guide the selection of effective treatment for patients with MDD. Although the proposed approach was motivated by modeling our matrix-valued EEG data (with electrode location and frequency range as its two dimensions) within the framework of latent class models, it can be broadly applied to any regression problem with covariates taking a natural matrix form. Also, the proposed approach readily extends to accommodate covariates that are multi-dimensional arrays in general regression settings. To extend the proposed low rank CP decomposition method by introducing cross-frequencies and cross-electrodes interactions would be interesting. For example, the EEG data could be also represented by three-way electrode-electrode-frequency arrays (also order-3 tensor) by mapping the scalp electrode locations to rectangular grids, and then one could study the two-way electrode-electrode interactions along with their interactions with different frequencies. This is a potentially promising direction for future extensions of our research.

We view our work as a first step toward a fully Bayesian treatment (i.e., also modeling the rank) of the CP decomposition problems in regression settings. Under the Bayesian hierarchical framework proposed in this paper, it is straightforward to infer the rank by further considering a prior for the rank $R$, e.g., a uniform prior with an upper bound. In this example, however, under the rank $R = 2$ model, the credible intervals of the coefficients' estimates, are much wider (resulting in the identification of only a very small number of statistically significant features, see (Figure 4 (b)). This suggests that at least in some cases, setting the rank $R$ as a parameter and estimating it, could diminish the ability to find important EEG features as the resulting estimates might need to be averaged across models with several different ranks. Future work should investigate the rank estimation thoroughly, where the implementation might not be trivial to move between models with parameter spaces of different dimensions (corresponding to different ranks). The improvement by averaging over models with different ranks might not

be always dramatic, given that our proposed method is not particularly sensitive to the choice of the rank as seen in our simulation studies.

Immediate extensions of the proposed approach could be to consider more structured hierarchical priors for smoothness regularization. For instance, for the EEG data example without the MPCA preprocessing step, we could instead adopt a class of conditionally autoregressive (CAR) priors (Besag and Kooperberg 1995) for $\tilde{\boldsymbol{\alpha}}_j$ to reflect the spatial similarities among EEG signals at nearby electrode locations. Additionally, sparsity-inducing priors could be helpful to applications involving ultrahigh dimensional neuroimaging phenotypes that are in the form of multi-dimensional arrays.

**6. Supplementary Materials.** Web Appendices A and B referenced in Sections 2.5 and 4; C++/R codes to implement the Gibbs sampler for our proposed models are available with this paper at the journal website.

## References.


Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on* **19** 716–723.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association* **88** 669–679.

Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L. and Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association* **92** 1375–1386.

Besag, J. and Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika* **82** 733–746.

Bonate, P. L. and Howard, D. R. (2011). *Pharmacokinetics in drug development: advances and applications* **3**. Springer Science & Business Media.

Bruder, G. E., Stewart, J. W., Tenke, C. E., McGrath, P. J., Leite, P., Bhattacharya, N. and Quitkin, F. M. (2001). Electroencephalographic and perceptual asymmetry differences between responders and nonresponders to an SSRI antidepressant. *Biological psychiatry* **49** 416–425.

Bruder, G. E., Sedoruk, J. P., Stewart, J. W., McGrath, P. J., Quitkin, F. M. and Tenke, C. E. (2008). EEG alpha measures predict therapeutic response to an SSRI antidepressant: pre and post treatment findings. *Biological Psychiatry* **63** 1171.

Ciarleglio, A., Petkova, E., Ogden, R. T. and Tarpey, T. (2015). Treatment decisions based on scalar and functional baseline covariates. *Biometrics* **71** 884–894.

Clogg, C. C. (1995). Latent class models. In *Handbook of Statistical Modeling for the Social and Behavioral Sciences* 311–359. New York: Plenum Press.

Collins, L. M. and Lanza, S. T. (2013). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. Hoboken: John Wiley & Sons.

Elliott, M. R. (2007). Identifying latent clusters of variability in longitudinal data. *Biostatistics* **8** 756–771.

Elliott, M. R., Gallo, J. J., Ten Have, T. R., Bogner, H. R. and Katz, I. R. (2005). Using a Bayesian latent growth curve model to identify trajectories of positive affect and negative events following myocardial infarction. *Biostatistics* **6** 119.

Garrett, E. S. and Zeger, S. L. (2000). Latent class model diagnosis. *Biometrics* **56** 1055–1067.

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association* **74** 153–160.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)* 501–514.

Gelman, A., Hwang, J. and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing* **24** 997–1016.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science* 457–472.

Gormley, I. C. and Murphy, T. B. (2011). Mixture of experts modelling with social science applications. *Journal of Computational and Graphical Statistics* **19** 332–353.

Holsboer, F. (2008). How can we realize the promise of personalized antidepressant medicines? *Nature Reviews Neuroscience* **9** 638–646.

Hung, H. and Wang, C.-C. (2013). Matrix variate logistic regression model with application to EEG data. *Biostatistics* **14** 189–202.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J. and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural computation* **3** 79–87.

Jiang, B., Elliott, M. R., Sammel, M. D. and Wang, N. (2015). Joint modeling of cross-sectional health outcomes and longitudinal predictors via mixtures of means and variances. *Biometrics* **71** 487–497.

Jordan, M. I. and Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural computation* **6** 181–214.

Joyce, P. R. and Paykel, E. S. (1989). Predictors of drug response in depression. *Archives of General Psychiatry* **46** 89–99.

Kamarajan, C., Pandey, A. K., Chorlian, D. B. and Porjesz, B. (2015). The use of current source density as electrophysiological correlates in neuropsychiatric disorders: A review of human studies. *International Journal of Psychophysiology* **97** 310–322.

Khodayari-Rostamabad, A., Reilly, J. P., Hasey, G., MacCrimmon, D. et al. (2010). Using pre-treatment EEG data to predict response to SSRI treatment for MDD. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology* 6103–6106. IEEE.

Kim, S., Chen, M.-H. and Dey, D. K. (2008). Flexible generalized t-link models for binary response data. *Biometrika* **95** 93–106.

Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review* **51** 455–500.

Lazarsfeld, P. F. and Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.

Leuchter, A. F., Cook, I. A., Witte, E. A., Morgan, M. and Abrams, M. (2002). Changes in brain function of depressed subjects during treatment with placebo. *American Journal of Psychiatry* **159** 122–129.

Li, B., Kim, M. K. and Altman, N. (2010). On dimension folding of matrix-or array-valued statistical objects. *The Annals of Statistics* 1094–1121.

Lu, H., Plataniotis, K. N. and Venetsanopoulos, A. N. (2008). MPCA: Multilinear principal component analysis of tensor objects. *Neural Networks, IEEE Transactions on* **19** 18–39.

MacCutcheon, A. (1987). *Latent class analysis*. Thousand Oaks: Sage Publications.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure* **405**

442–451.

Mumtaz, W., Malik, A. S., Yasin, M. A. M. and Xia, L. (2015). Review on EEG and ERP predictive biomarkers for major depressive disorder. *Biomedical Signal Processing and Control* **22** 85–98.

Muthén, B. and Brown, H. C. (2009). Estimating drug effects in the presence of placebo response: causal inference using growth mixture modeling. *Statistics in Medicine* **28** 3363–3385.

Muthén, B. and Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics* **55** 463–469.

Neelon, B., O'Malley, A. J. and Normand, S.-L. T. (2011). A Bayesian Two-Part Latent Class Model for Longitudinal Medical Expenditure Data: Assessing the Impact of Mental Health and Substance Abuse Parity. *Biometrics* **67** 280–289.

Nunez, P. L. and Srinivasan, R. (2006). *Electric fields of the brain: the neurophysics of EEG*. Oxford University Press, USA.

Patel, M. J., Khalaf, A. and Aizenstein, H. J. (2016). Studying depression using imaging and machine learning methods. *NeuroImage: Clinical* **10** 115–123.

Petkova, E., Tarpey, T. and Govindarajulu, U. (2009). Predicting potential placebo effect in drug treated subjects. *The International Journal of Biostatistics* **5** Article 23.

Phillips, M. L., Chase, H. W., Sheline, Y. I., Etkin, A., Almeida, J. R., Deckersbach, T. and Trivedi, M. H. (2015). Identifying predictors, moderators, and mediators of antidepressant response in major depressive disorder: neuroimaging approaches. *American Journal of Psychiatry* **172** 124–138.

Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies* **2** 37-63.

Quitkin, F., McGrath, P., Rabkin, J., Stewart, J., Harrison, W., Ross, D., Tricamo, E., Fleiss, J., Markowitz, J. and Klein, D. (1991). Different types of placebo response in patients receiving antidepressants. *The American Journal of Psychiatry* **148** 197–203.

Shen, J. and He, X. (2015). Inference for Subgroup Analysis with a Structured Logistic-Normal Mixture Model. *Journal of the American Statistical Association* **110** 303–312.

Sonawalla, S. B. and Rosenbaum, J. F. (2002). Placebo response in depression. *Dialogues in Clinical Neuroscience* **4** 105.

Stewart, J. W., Quitkin, F. M., McGrath, P. J., Amsterdam, J., Fava, M., Fawcett, J., Reimherr, F., Rosenbaum, J., Beasley, C. and Roback, P. (1998). Use of pattern analysis to predict differential relapse of remitted patients with major depression during 1 year of treatment with fluoxetine or placebo. *Archives of General Psychiatry* **55** 334–343.

Tarpey, T., Petkova, E. and Ogden, R. T. (2003). Profiling placebo responders by self-consistent partitioning of functional data. *Journal of the American Statistical Association* **98** 850–858.

Tarpey, T. and Petkova, E. (2010). Latent regression analysis. *Statistical Modelling* **10** 133–158.

Tarpey, T., Yun, D. and Petkova, E. (2008). Model misspecification finite mixture or homogeneous? *Statistical modelling* **8** 199–218.

Tenke, C. E., Kayser, J., Manna, C. G., Fekri, S., Kroppmann, C. J., Schaller, J. D., Alschuler, D. M., Stewart, J. W., McGrath, P. J. and Bruder, G. E. (2011). Current source density measures of electroencephalographic alpha predict antidepressant treatment response. *Biological psychiatry* **70** 388–394.

Vehtari, A. and Ojanen, J. (2012). A survey of Bayesian predictive methods for model

assessment, selection and comparison. *Statistics Surveys* **6** 142–228.
Wade, E. C. and Iosifescu, D. V. (2016). Using EEG for Treatment Guidance in Major Depressive Disorder. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*.
Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *The Journal of Machine Learning Research* **11** 3571–3594.
White, A. and Murphy, T. B. (2016). Mixed-membership of experts stochastic blockmodel. *Network Science* **4** 48–80.
Zhou, H., Li, L. and Zhu, H. (2013). Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association* **108** 540–552.

Bei Jiang
632 Central Academic Building
Department of Mathematical and Statistical Sciences
University of Alberta
Edmonton, Alberta T6G1X3
Canada
E-mail: bei1@ualberta.ca

Thaddeus Tarpey
Department of Mathematics and Statistics
Wright State University
Dayton, Ohio 45435
USA
E-mail: thaddeus.tarpey@wright.edu

Eva Petkova
Department of Child and Adolescent Psychiatry
New York University of School of Medicine
1 Park Ave, 7th Floor
New York, New York 10016
USA
E-mail: eva.petkova@nyumc.org

R. Todd Ogden
Department of Biostatistics
Columbia University
722 W. 168th Street, 6th floor
New York, New York 10032
USA
E-mail: to166@cumc.columbia.edu