HHS Public Access

Author manuscript

Biometrics. Author manuscript; available in PMC 2020 March 13.

Published in final edited form as:

Biometrics. 2020 March ; 76(1): 87–97. doi:10.1111/biom.13151.

A Bayesian Approach to Joint Modeling of Matrix-valued Imaging Data and Treatment Outcome with Applications to Depression Studies

Bei Jiang[1,*], **Eva Petkova**[2,3,**], **Thaddeus Tarpey**[2,***], **R. Todd Ogden**[4,****]

[1]Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, AB T6G 2E8, Canada.

[2]Department of Population Health, New York University, New York, NY 10016, USA.

[3]Department of Child and Adolescent Psychiatry, New York University, New York, NY 10016, USA.

[4]Department of Biostatistics, Columbia University, New York, NY 10032, USA.

Summary:

In this paper we propose a unified Bayesian joint modeling framework for studying association between a binary treatment outcome and a baseline matrix-valued predictor. Specifically, a joint modeling approach relating an outcome to a matrix-valued predictor through a probabilistic formulation of multilinear principal component analysis (MPCA) is developed. This framework establishes a theoretical relationship between the outcome and the matrix-valued predictor although the predictor is not explicitly expressed in the model. Simulation studies are provided showing that the proposed method is superior or competitive to other methods, such as a two-stage approach and a classical principal component regression (PCR) in terms of both prediction accuracy and estimation of association; its advantage is most notable when the sample size is small and the dimensionality in the imaging covariate is large. Finally, our proposed joint modeling approach is shown to be a very promising tool in an application exploring the association between baseline EEG data and a favorable response to treatment in a depression treatment study by achieving a substantial improvement in prediction accuracy in comparison to competing methods.

Keywords

Antidepressant response; Dimension reduction; Major depression disorder; Matrix-valued imaging data; Multilinear principal component analysis; Regularization

[*] bei1@ualberta.ca. [**] Eva.Petkova@nyulangone.org. [***] Thaddeus.Tarpey@nyulangone.org. [****] to166@cumc.columbia.edu.

Supporting Information

Web Appendices, Tables, and Figures referenced in Sections 1–4 are available with this paper at the Biometrics website on Wiley Online Library. The C++ codes (with R interface) to implement our joint modeling approach and a sample simulated dataset are also available on Github at https://github.com/beijiang6.

1 Introduction

Although antidepressants have been commonly prescribed for the treatment of depression since they were first introduced into clinical practice in the 1950s, they are not effective for many patients (Holsboer, 2008). Current clinical practice typically involves a long trial and error process to find an effective antidepressant treatment, and during this time the patient may be non-functioning, have poor quality of life, and be at risk for suicide. Therefore, identification of pre-treatment biomarkers that predict response to a specific antidepressant is of great clinical interest. Standard biomarkers all have low predictive utility, but more recently, the use of biomarkers derived from neuroimaging technologies, such as magnetic resonance imaging (MRI), functional MRI and electroencephalography (EEG), has shown promise for predicting patient response to antidepressants (Wade and Iosifescu, 2016).

Among these neuroimaging modalities, the use of EEG based biomarkers has been extensively investigated in depression studies due to its wide availability, non-invasive nature and low cost (Dunlop and Mayberg, 2014). EEG records brain electrical activity from electrodes placed on the scalp with high temporal resolution (in milliseconds) during a task-related activity or at rest without stimuli. To address the relatively poor spatial resolution, the use of current source density (CSD) analysis of EEG has become popular due to its ability to estimate the source of the current generating the measured potentials (Kamarajan et al., 2015). In our motivating study, the investigators hypothesize, based on both theoretical considerations and empirical evidence, that the CSD based EEG measures of power spectra in the theta (4 – 7 Hz) and alpha (7 – 15 Hz) frequency bands (i.e., 45 unique frequencies given a resolution of 0.25 Hz; see Web Appendix A for details) at a total of 14 electrode locations in the posterior brain region, are related to whether a patient will respond favorably to an antidepressant from the class of selective serotonin reuptake inhibitors (SSRI). These CSD-based EEG measures for each subject are represented as a 14×45 matrix, with the rows and columns corresponding to electrodes and theta/alpha frequency bands respectively. Our task is to relate these matrix-valued EEG measures to patients' response to antidepressants.

Research dealing with high dimensional matrix-valued variables (also known as matrix-variates) has attracted considerable attention in the past ten years. There is a vast literature on dimension reduction techniques for matrix-variates (e.g., Li et al. 2010; Ding and Cook 2014; Xue and Yin 2014, 2015; Virta et al. 2017). In regression settings, existing methods for modeling matrix-variates as covariates focus on incorporating various regularization schemes that also preserve the inherent matrix nature of the covariate, for example, by imposing a low rank bilinear structure to the associated regression coefficients for the matrix-valued covariate (Hung and Wang 2013, Zhou et al. 2013, Hoff 2015, Jiang et al. 2017), applying a particular structured lasso penalty (Zhao and Leng 2014), applying a nuclear norm penalty (Zhou and Li 2014), and utilizing the envelope concept originally proposed in Cook et al. (2010) to achieve supervised sufficient dimension reduction for tensor-variates (e.g., Li and Zhang 2017; Zhang and Li 2017; Ding and Cook 2018).

Classical principal component analysis (PCA), a ubiquitous tool for dimension reduction, has been extended to matrix-variates and, more generally, multi-dimensional arrays (Lu et al. 2008). Termed "multilinear principal component analysis" (MPCA), this procedure has been

successfully used in many applications (e.g., Lu et al. (2011)). However, earlier work do not develop the MPCA within an estimation-inference framework (e.g., Hung et al. (2012)).

In this paper, we propose a Bayesian formulation of a probabilistic MPCA model and utilize a joint modeling framework to simultaneously relate MPCA features to a binary response. Specifically, the contributions of this manuscript are:

1. Our Bayesian MPCA model can be considered as a multilinear extension of the popular probabilistic PCA model (Tipping and Bishop, 1999). A similar model was previously studied in Ding and Cook (2014) under the name of dimension folding PCA.
2. A theoretically structured framework is developed relating a binary outcome to the original matrix-valued covariate using a joint modeling approach, where regularization for the regression with high-dimensional matrix-valued covariates is naturally achieved. Note that our framework is closely related to the supervised sufficient dimension reduction methods for matrix-variates (e.g., Li et al. 2010, Ding and Cook 2014, Virta et al. 2017). However, the advantage of our framework is that it allows uncertainties in the estimation of both the MPCA model and the outcome model to be naturally quantified by Bayesian credible intervals.
3. Previous approaches to applying MPCA in a regression context with matrix-valued predictors (Hung and Wang, 2013; Jiang et al., 2017) have first performed MPCA on the predictors and then regressed the outcome on the extracted features. In contrast, here we explicitly account for the estimation error in the MPCA (avoiding attenuation bias) and relate the MPCA features to the outcome variable in one modeling step.

Although motivated by modeling EEG data, the proposed framework is applicable to matrix-valued predictors that occur in other contexts, e.g., 2D images/objects and spatial-temporal data that can arise in biomedical studies and computer vision applications (see Lu et al., 2013 and references therein).

2 Model and methods

In this section, we present the joint modeling framework for the baseline matrix-valued covariate and the binary treatment outcome. For each subject i, $i = 1, \ldots, n$, let $\boldsymbol{x}_i \in \mathcal{R}^{p \times q}$ be the $p \times q$ matrix-valued covariate, z_i be the vector of scalar covariates, and o_i be the binary indicator outcome (e.g., an indicator for responding favorably or not to treatment).

2.1 Model for the baseline matrix-valued covariate

Our proposed model for the baseline matrix-valued covariate $\boldsymbol{x}_i$ utilizes multilinear principal component analysis (MPCA) (Hung et al., 2012) to represent $\boldsymbol{x}_i$, aiming to find a low dimensional representation of $\boldsymbol{x}_i$ while preserving the inherent matrix structure in $\boldsymbol{x}_i$. Specifically, this MPCA formulation for $\boldsymbol{x}_i$ takes the form

$$\boldsymbol{x}_i = \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{u}_i\boldsymbol{B}^T + \boldsymbol{\epsilon}_i, \quad \text{where} \quad \text{vec}(\boldsymbol{\epsilon}_i) \sim \mathcal{N}\left(\boldsymbol{0}, \phi^{-1}\boldsymbol{I}_{pq\times pq}\right). \tag{1}$$

Here (i) $\boldsymbol{\mu} \in \mathcal{R}^{p\times q}$ is the population mean, (ii) $\boldsymbol{u}_i \in \mathcal{R}^{p_0\times q_0}$ is a matrix of latent features obtained by mapping $\boldsymbol{x}_i$ through a multilinear projection with two projection matrices $\boldsymbol{A} = \left(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{p_0}\right) \in \mathcal{R}^{p\times p_0}$ and $\boldsymbol{B} = \left(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_{q_0}\right) \in \mathcal{R}^{q\times q_0}$,, such that $\boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{I}_{p0\times p0}$, $\boldsymbol{B}^\top\boldsymbol{B} = \boldsymbol{I}_{q0\times q0}$, and $\text{vec}(\boldsymbol{u}_i) \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I}_{p_0q_0\times p_0q_0}\right)$, and (iii) $\boldsymbol{\epsilon}_i$ is a matrix of noise terms that are independent of $\boldsymbol{u}_i$ and vec(·) denotes the vectorization operator. Similar to the probabilistic formulation of PCA (e.g., Tipping and Bishop 1999), the motivation here is that, with $p_0 < p$ and $q_0 < q$, the latent matrix $\boldsymbol{u}_i$ will offer a more parsimonious representation of the original high-dimensional matrix $\boldsymbol{x}_i$. Model (1) can be equivalently written as

$$\text{vec}(\boldsymbol{x}_i) = \text{vec}(\boldsymbol{\mu}) + (\boldsymbol{B} \otimes \boldsymbol{A})\text{vec}(\boldsymbol{u}_i) + \text{vec}(\boldsymbol{\epsilon}_i), \tag{2}$$

where ⊗ denotes the Kronecker product. This formulation of the model implies a conditional normal distribution for $\boldsymbol{u}_i$ given $\boldsymbol{x}_i$ and the model parameters; this closed-form full conditional posterior distribution will be used in the Gibbs sampling.

Remark 1. In our formulation of the MPCA model, we have assumed the variables in latent $\boldsymbol{u}_i$ are uncorrelated in a similar spirit as the orthogonal factors assumption in factor analysis. Alternatively, we could allow the latent features in $\boldsymbol{u}_i$ to be correlated, similar to oblique factor models (e.g., Lawley and Maxwell 1962). However, this generalization adds complexity and effort for estimation. From the Bayesian point of view, the model assumption on $\boldsymbol{u}_i$ also serves as a prior and it is very common to assume independent priors on unknown quantities. Additionally, our simulation studies in Section 3 show that the model performs equally well regardless of whether the latent features are correlated or not.

2.2 Model for the binary treatment response

Instead of directly regressing o_i on the original $\boldsymbol{x}_i$, we relate the likelihood of $o_i = 1$ to the latent features $\boldsymbol{u}_i$ along with a vector of scalar covariates $\boldsymbol{z}_i$, through a probit model.

Specifically, assuming that the variation due to the random noise $\boldsymbol{\epsilon}_i$ provides no information about o_i, the model is

$$\Phi^{-1}[\Pr\{o_i = 1\}] = \psi + \gamma^\top\boldsymbol{z}_i + \boldsymbol{\theta}^\top\text{vec}(\boldsymbol{u}_i), \tag{3}$$

where Φ(·) denotes the cumulative distribution function for a standard normal distribution. The proposed modeling framework also provides regularization since the the low-dimensional latent features $\boldsymbol{u}_i$ enter the model instead of the original $\boldsymbol{x}_i$ thereby imposing a structure on the associated coefficients for the original $\boldsymbol{x}_i$, leading to a potentially substantial dimension reduction from the original pq parameters for $\boldsymbol{x}_i$ to p_0q_0 parameters for $\boldsymbol{u}_i$.

Remark 2. In contrast to jointly modeling the matrix covariate $\boldsymbol{x}_i$ and outcome o_i, a two-stage modeling approach can be considered as follows: Stage 1 − perform MPCA on $\boldsymbol{x}_i$, such that

$\boldsymbol{x}_i$ can be approximated by lower dimensional features $\hat{\boldsymbol{u}}_i^* \approx \boldsymbol{U}^\top \boldsymbol{x}_i \boldsymbol{V} \in \mathscr{R}^{p_0 \times q_0}$ with $p_0 < p$ and $q_0 < q$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are the eigenvector matrices such that $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}_{p0 \times p0}$ and $\boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}_{q0 \times q0}$; Stage 2 − fit a regression model for o_i with the extracted low dimensional features $\hat{\boldsymbol{u}}_i^*$ as well as other covariates as predictors. Then the desired coefficient matrix for the original matrix covariate $\boldsymbol{x}_i$ can be recovered from $\boldsymbol{\beta}^* = \boldsymbol{U}\boldsymbol{\theta}^*\boldsymbol{V}^\top$, where $\boldsymbol{\theta}^*$ is the coefficient matrix for $\hat{\boldsymbol{u}}_i^*$ in the second-stage model. However, due to potential estimation errors in extracting features $\hat{\boldsymbol{u}}_i^*$, a two-stage modeling approach is subject to attenuation bias and could be less efficient than a joint modeling approach. This is a phenomenon that has been noted in the joint modeling literature (e.g., Ibrahim et al., 2010) and is illustrated in our simulation study in Section 3. Please refer to Web Appendix B for additional remarks on comparison with other related methods.

2.3 Identifiability

The MPCA submodel for the matrix-valued covariate $\boldsymbol{x}_i$ (i.e., model (1)) is identifiable only up to orthogonal rotations; that is, for any orthonormal matrices $\boldsymbol{P} \in \mathscr{R}^{p0 \times p0}$ and $\boldsymbol{Q} \in \mathscr{R}^{q0 \times q0}$ with $\boldsymbol{P}\boldsymbol{P}^\top = \boldsymbol{I}_{p0 \times p0}$ and $\boldsymbol{Q}\boldsymbol{Q}^\top = \boldsymbol{I}_{q0 \times q0}$, letting $\widetilde{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{P}$, $\widetilde{\boldsymbol{B}} = \boldsymbol{B}\boldsymbol{Q}$ and $\widetilde{\boldsymbol{u}}_i = \boldsymbol{P}^\top \boldsymbol{u}_i \boldsymbol{Q}$ leads to an identical likelihood, because mpw $\boldsymbol{A}\boldsymbol{u}_i\boldsymbol{B}^\top = \widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{u}}_i\widetilde{\boldsymbol{B}}^\top$. As a consequence, letting $\widetilde{\boldsymbol{\theta}} = \left(\boldsymbol{Q}^\top \otimes \boldsymbol{P}^\top\right)\boldsymbol{\theta}$ leads to $\widetilde{\boldsymbol{\theta}}^\top \mathrm{vec}(\widetilde{\boldsymbol{u}}) = \boldsymbol{\theta}^\top \mathrm{vec}(\boldsymbol{u}_i)$, and hence an identical likelihood in submodel (3) for the outcome too. In our analysis, such rotation invariance can be neglected, because the goal is not to interpret individual features in $\boldsymbol{u}_i$, but rather, to recover and make inference about the regression of the binary outcome o_i on the matrix-valued covariate $\boldsymbol{x}_i$.

In Web Appendix C, we show that the joint models (1) and (3) imply the following regression model for o_i given the original matrix-valued covariate $\boldsymbol{x}_i$, the scalar covariates $\boldsymbol{z}_i$, and the model parameters

$$\Phi^{-1}[\Pr\{o_i = 1\}] = \varphi + \boldsymbol{\alpha}^\top \boldsymbol{z}_i + \boldsymbol{\beta}^\top \mathrm{vec}(\boldsymbol{x}_i), \tag{4}$$

where $\varphi = \left(1 + \boldsymbol{\theta}^\top \boldsymbol{C}\boldsymbol{\theta}\right)^{-\frac{1}{2}}\left[\gamma_0 - \phi\boldsymbol{\theta}^\top \boldsymbol{C}(\boldsymbol{B} \otimes \boldsymbol{A})^\top \mathrm{vec}(\boldsymbol{\mu})\right]$, $\boldsymbol{\alpha} = \left(1 + \boldsymbol{\theta}^\top \boldsymbol{C}\boldsymbol{\theta}\right)^{-\frac{1}{2}}\boldsymbol{\gamma}$ and $\boldsymbol{\beta} = \phi(1 + \boldsymbol{\theta}^\top \boldsymbol{C}\boldsymbol{\theta})^{-\frac{1}{2}}(\boldsymbol{B} \otimes \boldsymbol{A})\boldsymbol{C}\boldsymbol{\theta}$. It can be shown by simple algebra that $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ remain unchanged by substituting $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{\theta}$ by their rotated counterparts (introduced above) $\widetilde{\boldsymbol{A}}$, $\widetilde{\boldsymbol{B}}$ and $\widetilde{\boldsymbol{\theta}}$ respectively. Therefore, this implied regression model of o_i on $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ is identifiable and provides a unique correspondence between the low-dimensional coefficients $\boldsymbol{\theta}$ for the latent features $\boldsymbol{u}_i$ and the high-dimensional coefficients $\boldsymbol{\beta}$ for the matrix-valued covariate $\boldsymbol{x}_i$.

2.4 Prior distributions

In this section we present our choices of prior distributions for the parameters in our previously described joint models (1) and (3). In model (1), $\boldsymbol{A} \in \mathscr{R}_{p \times p0}$ and $\boldsymbol{B} \in \mathscr{R}_{q \times q0}$ are orthonormal and hence belong to Stiefel manifolds, i.e., the spaces of orthonormal matrices, denoted by $\mathscr{V}_{p0,\,p}$ and $\mathscr{V}_{q0,\,q}$. It is natural to adopt uniform distributions on $\mathscr{V}_{p0,\,p}$ and $\mathscr{V}_{q0,\,q}$.

as priors for $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively. There exists a conditional representation of the uniform distribution on Stiefel manifolds (see Hoff (2007) for more discussion), which will facilitate the Gibbs sampling of the columns of $\boldsymbol{A}$ and $\boldsymbol{B}$ from their full conditional posterior distributions. For other model parameters, we adopt the commonly used non-informative conjugate priors: a flat prior for $\boldsymbol{\mu}$, i.e., $\pi(\boldsymbol{\mu}) \propto 1$, $\phi^{-1} \sim \text{Gamma}(a_0, b_0)$, $\psi \sim \mathcal{N}(0, \tau_0)$, $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{0}, \tau_0 \boldsymbol{I})$, and $\gamma \sim \mathcal{N}(\boldsymbol{0}, \tau_0 \boldsymbol{I})$, where $\tau_0 = 10$ and $a_0 = b_0 = 0.1$. We note that in the simulation and real example sections, the results were insensitive to the choice of these hyperparameters. Figure 1 presents a graphical representation of the hierarchical structure of the joint models (1) and (3), along with prior distributions for model parameters. The details to derive posterior distributions and Gibbs sampling algorithm were provided in the Web Appendix D.

2.5 Prediction and selection of dimensionality

For a future sample with baseline covariates $\boldsymbol{x}_{new}$ and $\boldsymbol{z}_{new}$, the posterior predictive probability of responding favorably to the treatment, i.e., $o_{new} = 1$, is given by,

$$\begin{aligned} &\Pr(o_{new} = 1 \,|\, \boldsymbol{x}_{new}, \boldsymbol{z}_{new}, \boldsymbol{o}, \boldsymbol{x}, \boldsymbol{z}) \\ &= \int f(\boldsymbol{v}, \boldsymbol{u}_{new}, \boldsymbol{u} \,|\, \boldsymbol{o}, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{x}_{new}, \boldsymbol{z}_{new}) \times \Pr(o_{new} = 1 \,|\, \boldsymbol{x}_{new}, \boldsymbol{z}_{new}, \boldsymbol{v}, \boldsymbol{u}_{new}, \boldsymbol{u}) d\boldsymbol{v} d\boldsymbol{u}_{new} d\boldsymbol{u} \\ &= \int f(\boldsymbol{v}, \boldsymbol{u}_{new}, \boldsymbol{u} \,|\, \boldsymbol{o}, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{x}_{new}, \boldsymbol{z}_{new}) \times \Phi\left(\psi + \boldsymbol{z}_{new}^{\top} \boldsymbol{\gamma} + \text{vec}(\boldsymbol{u}_{new})^{\top} \boldsymbol{\theta}\right) d\boldsymbol{v} d\boldsymbol{u}_{new} d\boldsymbol{u} \end{aligned} \tag{5}$$

.

where $\boldsymbol{v}$ includes all model parameters. The integral in expression (5) can be approximated by the Monte Carlo estimate $\sum_{m=1}^{M} \Phi\left(\psi^{(m)} + \boldsymbol{z}_{new}^{\top} \boldsymbol{\gamma}^{(m)} + \text{vec}(\boldsymbol{u}_{new})^{\top} \boldsymbol{\theta}^{(m)}\right) / M$, where $\boldsymbol{\gamma}^{(m)}$, $\boldsymbol{\theta}^{(m)}$ and $\boldsymbol{u}^{(m)}_{new}$ are the posterior samples based on the data $\{\boldsymbol{o}, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{x}_{new}, \boldsymbol{z}_{new}\}$. The full conditional posterior distribution for $\boldsymbol{u}_{new}$ is now given by $[\text{vec}(\boldsymbol{u}_{new}) | \cdot] \sim \mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Psi}})$, where $\widetilde{\Psi} = (1 + \phi)^{-1} \boldsymbol{I}, \widetilde{\boldsymbol{\mu}} = \phi \widetilde{\boldsymbol{\Psi}} (\boldsymbol{B} \otimes \boldsymbol{A})^{\top} (\text{vec}(\boldsymbol{x}_{new}) - \text{vec}(\boldsymbol{\mu}))$. For other parameters, the posterior samples can be obtained similarly using Gibbs sampling as described in Web Appendix D.

To select the dimensionality (p_0, q_0) in our probabilistic MPCA submodel (1), we consider the out-of-sample prediction accuracy as measured by the area under the curve (AUC) of the receiver operating characteristic (ROC), since our focus is on the predictive power of the model. Specifically, we fit candidate models with different choices of (p_0, q_0) and select a specific (p_0, q_0) that leads to the highest out-of-sample or cross-validation AUC; more details can be found in Web Appendix E.

3 Simulation study

In this section, we evaluate the performance of our proposed method on simulated datasets with controlled low-dimensional structure (i.e., with p_0 and q_0 fixed), and focus on the following two aspects: i) estimation of the coefficient $\boldsymbol{\beta}$ for the matrix-valued covariate $\boldsymbol{x}_i \in \mathcal{R}^{p \times q}$ in model (4), and ii) out-of-sample prediction accuracy of the binary outcome.

3.1 Simulation setup

To minimize Monte Carlo errors, we consider a nested simulation design. Specifically, each dataset was simulated from the following procedure:

1. Let $(p_0, q_0) = (2, 2)$ and the elements in $\text{vec}(\boldsymbol{u}_i)$ are generated independently from uniform distribution on $[-1, 1]$.
2. Let $o_i \sim \text{Bernoulli}(\Phi[0.2z_i + \mathbf{1}^\top \text{vec}(\boldsymbol{u}_i)])$ with $\mathbf{1} = (1, \ldots, 1)^\top$ and $z_i \sim \text{uniform}(-0.5, 0.5)$.
3. For each simulated set of $\{(o_i, z_i, \boldsymbol{u}_i) : i = 1, \ldots, n\}$ from 1) and 2), we simulate additional sets of $\{\boldsymbol{x}_i : i = 1, \ldots, n\}$ for different choices of (p, q) as follows,
   a. Let $\boldsymbol{A} \sim \text{uniform}\left(\mathscr{V}_{p, p0}\right)$ and $\boldsymbol{B} \sim \text{uniform}\left(\mathscr{V}_{q, q_0}\right)$.
   b. Let $\boldsymbol{x}_i = \boldsymbol{A}\boldsymbol{u}_i\boldsymbol{B}^\top + \boldsymbol{\epsilon}_i$ with $\text{vec}(\epsilon_i) \sim \mathcal{N}\left(\mathbf{0}, 0.2^2 \boldsymbol{I}_{pq \times pq}\right)$.

Using the above procedure, for each sample size $n \in \{50, 100, 200\}$, we simulate two hundred sets of $\{(o_i, z_i, \boldsymbol{u}_i) : i = 1, \ldots, n\}$ and three sets of $\{A, B\}$, corresponding to three choices of $(p, q) \in \{(25, 25), (30, 30), (35, 35)\}$ respectively. Therefore, for each sample size n, each of the simulated sets of $\{(o_i, z_i, \boldsymbol{u}_i) : i = 1, \ldots, n\}$ are common in all three final datasets of $\{(o_i, z_i, \boldsymbol{u}_i, \boldsymbol{x}_i) : i = 1, \ldots, n\}$ under the three different (p, q) settings. Because of such nested simulation designs, the datasets simulated with the same sample size n (irrespective of the setting of (p, q) for $\boldsymbol{x}_i$) share the same sets of the latent features $\boldsymbol{u}_i$ and therefore share the same true prediction accuracy (and hence the same true AUC values obtained based on the outcome submodel (3) with true parameter values). We repeat the simulation by generating $\text{vec}(\boldsymbol{u}_i)$ from correlated uniform distributions on $[-1, 1]$. Specifically, the elements in $\text{vec}(\boldsymbol{u}_i)$ were sequentially generated such that the s^{th} and $(s + 1)^{th}$ entries have a correlation of 0.3.

For each simulated dataset, we obtain the posterior samples of all model parameters by retaining every 5^{th} draw from 25,000 iterations after a burn-in period of 5,000 iterations. Under each simulation scenario, when assessing the performance with respect to the estimation of the associated coefficient for $\boldsymbol{x}_i$, we focus on the mean squared error (MSE). For the s^{th} simulated dataset, it is defined as $\text{MSE}^{(s)} = (pq)^{-1}\left\|\hat{\boldsymbol{\beta}}^{(s)} - \boldsymbol{\beta}^{(s)}\right\|_F^2$, where $\|\cdot\|^2_F$ is the Frobenius norm, $\boldsymbol{\beta}^{(s)}$ is the true coefficient matrix and $\hat{\boldsymbol{\beta}}^{(s)}$ is its posterior mean estimate. To access the out-of-sample predictive accuracy measured by AUC, for each simulated dataset, $\{(o_i, z_i, \boldsymbol{x}_i) : i = 1, \ldots, n\}$, we also simulate an additional testing dataset of size $\tilde{n} = 25, \left\{\left(o_i^{tst}, z_i^{tst}, \boldsymbol{x}_i^{tst}\right) : i = 1, \ldots, \tilde{n}\right\}$. The out-of-sample AUC is then obtained based on $p\left(o_i^{tst} = 1 \mid z^{tst}, \boldsymbol{x}^{tst}, \boldsymbol{o}, \boldsymbol{z}, \boldsymbol{x}\right)$, $i = 1, \ldots, \tilde{n}$ using the procedure as described in Section 2.5. We also compare our joint modeling approach with the two-stage modeling approach (described in Remark 2) and the classical PCR approach (described in Web Remark B1 in Web Appendix B) when all three methods have the same number of reduced dimensions, i.e., the same number of extracted features when predicting the outcome. For the two-stage approach, an MPCA procedure is first performed on the matrix-valued covariate using the rTensor package in R software (R Core Team, 2018) and then the p_0q_0 extracted MPCA features are

used as predictors in the second-stage probit regression model for the outcome. To compare with PCR, we follow Hung et al. (2012) to first transform the matrix-valued covariate into a vector, and then extracted the leading p_0q_0 principal components as the predictors in the probit regression model for the outcome.

3.2 Simulation results

Table 1 summarizes the out-of-sample AUC values in predicting the binary outcome across 200 simulations under each of the scenarios described above. The overall prediction accuracy improves as the sample size n increases for all three methods. We further have the following main observations:

- Due to vectorization in PCR, the PCR approach provides the smallest out-of-sample AUC values in all scenarios, while both our joint modeling and two-stage approaches provide much better out-of-sample prediction performance.
- Both our joint modeling and two-stage approaches are little affected by increasing the dimension of the covariate (p, q). Even under the most difficult scenario with $n = 50$ and $(p, q) = (35, 35)$, the out-of-sample AUC values are close to the truth, where the small changes in the AUC values for different (p, q) are likely due to the uncertainty associated with estimating AUC. Our joint modeling approach achieves further improved prediction performance in comparison to the two-stage modeling approach, especially in the small sample size case with $n = 50$. However, increasing p and q drastically deteriorates the prediction performance of PCR approach. As the sample size increases (in our simulation, to $n = 200$), differences in the prediction performance of the two-stage and joint modeling approaches tend to diminish. This is not surprising since the impact of estimation error declines for both methods as the sample size increases.
- Our joint modeling approach performs equally well regardless of whether the latent features $\boldsymbol{u}_i$ are generated from independent or correlated uniform distributions, suggesting some robustness to violation of the distribution assumption on $\boldsymbol{u}_i$ in model (1).

The advantage of our joint modeling approach is most notably shown in the estimation of the coefficients $\boldsymbol{\beta}$ for $\boldsymbol{x}_i$. Table 2 provides a comparison of the averages of the MSEs in estimating $\boldsymbol{\beta}$ for these three approaches. Note that, because the values of the coefficients for $\boldsymbol{u}_i$ in the probit model are fixed for different choices of (p, q) in the simulation to evaluate the impact on the prediction accuracy by increasing the values of (p, q), the corresponding true values for $\boldsymbol{\beta}$ are therefore of different magnitudes for different choices of (p, q), and hence the MSE's for these $\boldsymbol{\beta}$'s cannot be directly compared. Our main observations are as follows:

- The two-stage approach fails to carry over the uncertainty of estimating $\boldsymbol{u}_i^*$ in the first-stage MPCA to the second stage model. As expected, the joint modeling approach provides a reduction in MSE in comparison to the two-stage approach, and such a reduction is substantial for the smaller sample size ($n = 50$). These

higher MSEs in the two-stage approach are due to the estimation errors in extracting features $\hat{\boldsymbol{u}}_i^*$ in the first-stage MPCA procedure. Note that in the scenario with $n = 200$, such an impact due to estimation errors becomes almost negligible.

- Due to the fact that PCR has to rely on reshaping the matrix into a very long vector before extracting lower dimensional features, it is not surprising that PCR leads to the highest MSEs in all scenarios.

These comparisons demonstrate that the MPCA-based approach is a very promising tool for feature extraction for matrix-valued covariate relative to classical PCA-based approach and importantly, that efficiency can be further improved by the joint modeling approach in comparison to the two-stage modeling approach.

Remark 3. Additional simulations (presented in Web Appendix F) also show that our joint modeling approach can outperform the tensor regression approach proposed by Zhou et al. (2013) in both prediction and estimation. Also, it is important to note that our approach is not sensitive to model misspecification of the assumed association structure between the outcome and the matrix-valued covariate.

4 Application to predict antidepressant response using EEG data

In this section we analyze the motivating dataset described in Section 1 using the proposed joint modeling approach. In this dataset, the subjects were 58 patients with major depressive disorder (MDD) treated with a standard Selective Serotonin Reuptake Inhibitor (SSRI) for 8 weeks. The primary goal is to elucidate biological mechanisms of response to the SSRI, thus only subjects who underwent the entire 8 weeks of treatment per protocol, are included in the analysis and all of them had their depression symptoms assessed at treatment end. Patients' depression symptom severity was assessed using the 17-item Hamilton Depression Rating Scale (HAMD-17). The HAMD-17 scores range from 0 to 52, where higher scores indicate more severe depression. The outcome is response to the treatment, denoted by o_i, defined as a dichotomous indicator of a 50% or more reduction of symptoms after 8 weeks of treatment compared to baseline (Israel, 2006). Of main interest is to predict SSRI response using baseline EEG biomarkers. Recall that these baseline EEG biomarkers take the form of 14×45 matrices, denoted by $\boldsymbol{x}_i$, containing the CSD amplitude spectrum values ($\mu V/m^2$) at the 14 electrodes located in posterior brain regions, including occipital and parietal lobes, measured at the theta (4 – 7 Hz) and alpha (7 – 15 Hz) frequency bands (leading to a total of 45 frequencies with a given 0.25 Hz frequency resolution). We have followed the EEG literature to normalize these EEG CSD measures by taking log-transformations. Seven of the patients who completed treatment as per protocol had missing EEG data due to technical EEG recording issues that were independent of patient-related characteristics. Of the remaining 51 subjects with EEG data, 26 (51%) were classified as SSRI responders ($o_i = 1$) and 25 (49%) were classified as SSRI non-responders ($o_i = 0$). We also adjust for gender (1 for female; 0 for male) and depression chronicity (1 for being depressed for at least 24 months in the past 4 to 5 years; 0 otherwise), denoted by $\boldsymbol{z}_i$, in the outcome model (3). To select the dimensions (p_0, q_0) for the latent $\boldsymbol{u}_i$ in our model (1), we performed 10-fold cross-validation (CV) by randomly dividing the 51 subjects into 10

subsets of about equal size; this 10-fold CV procedure is then repeated ten times. For all models fitted here, we ran three MCMC chains of 65,000 iterations with the initial 15,000 iterations discarded as burn-in samples on a high performance computing cluster, and retained every 5^{th} sample, leading to 10,000 posterior samples used in the analysis. The mixing and convergence of the chains for fitting these models was very clean. To illustrate, we computed the Gelman-Rubin convergence diagnostics (Gelman et al., 1992) for our best fitting model, where all univariate potential scale reduction factors (PSRFs) are close to 1 and the multivariate PSRF is 1.035, suggesting that non-convergence was not detected; Web Figure 1 (in Web Appendix G) shows trace plots based on three MCMC chains for 28 randomly selected coefficients in $\boldsymbol{\beta}$ under this best fitting model. The running time of our algorithm depends on (p_0, q_0); this best fitting model takes 6.9 mins to obtain 10,000 draws on a 1.2 GHz Dual-Core Intel Core M MacBook with 8 GB Memory.

The average of the ten repeated 10-fold cross-validation AUCs for fitting models under different choices of (p_0, q_0) are shown in Tables 3 (a)-(b); the model with $(p_0, q_0) = (4, 4)$ leads to the highest cross-validation AUC=0.697. In comparison, the highest cross-validation AUC by the two-stage, PCR and tensor regression (Zhou et al., 2013) approaches are 0.688, 0.611 and 0.685, respectively. These latent MPCA features extracted 83% of the major variation in the original EEG data, which is calculated as the posterior mean of $\sum_{i=1}^{n} \left\| \boldsymbol{A}^{\top}(\boldsymbol{x}_i - \boldsymbol{\mu})\boldsymbol{B} \right\|_F^2 / \sum_{i=1}^{n} \|\boldsymbol{x}_i - \overline{\boldsymbol{x}}\|_F^2$ (Hung et al., 2012), where $\overline{\boldsymbol{x}} = n^{-1}\sum_{i=1}^{n} \boldsymbol{x}_i$ In terms of predicting an SSRI responder, our results show that chronically depressed patients are less likely to respond to the antidepressant at the end of the 8^{th} week of treatment, with a marginally significant effect of $\hat{\gamma}_2 = -1.32$, (90% CI: −2.60, −0.12), while gender is not a contributing factor, with $\hat{\gamma}_1 = 0.01(90\%\text{Cl}: -0.97, 1.02)$. This finding agrees with the general findings in the depression literature that chronically depressed patients may often be less responsive to antidepressant treatment (e.g., Al-Harbi, 2012). Figure 2(a) shows the posterior density of the probability of being an SSRI responder for all subjects in the study under this best chosen model. The obvious separation in the SSRI responders and non-responders distributions respectively provides evidence about the effectiveness of our model in classifying SSRI responders and non-responders. Note that all three subfigures (a)-(c) in Figure 2 appear in color in the electronic version of this article, and any mention of color refers to that version. Figure 2(b) shows the posterior density of $\boldsymbol{\beta}^{\top}\text{vec}(\boldsymbol{x}_i)$ for the responders and non-responders respectively, which further illustrates the usefulness of baseline EEG measurements in distinguishing the two groups. For this best fitting model, the Hosmer-Lemeshow test with $g = 10$ produced a p-value of 0.9897 indicating no evident calibration problems with the model.

Figure 2(c) plots the heat map of the coefficients $\boldsymbol{\beta}$ for the matrix-valued covariate $\boldsymbol{x}_i$, where asterisks (*) are used to mark the significant effects with the 95% credible interval not covering zero. As illustrated in Section 2.3, the estimates for $\boldsymbol{\beta}$ can be obtained by mapping $\boldsymbol{\theta}$ for the latent features $\boldsymbol{u}_i \in \mathcal{R}^{p_0 \times q_0}$ to the original feature space for $\boldsymbol{x}_i \in \mathcal{R}^{p \times q}$ through model (4). Except for the three electrodes PO4, PO6 and OZ on the outer boundary of the posterior brain region, all other 11 electrodes jointly are shown to have played a significantly important role in predicting SSRI response and such effects concentrate on the overlapping

frequency range between theta and alpha bands (around 7 Hz) through most of the alpha frequency range. These results corroborate previous findings that baseline CSD measures at alpha or/and theta frequency bands are viable predictors of clinical response to SSRIs (e.g., Tenke et al., 2017).

5 Discussion

In this paper we have proposed a joint modeling framework to study the relationship between a matrix-valued covariate and a binary outcome. It consists of defining a Bayesian probabilistic MPCA model for the matrix-valued covariate, and an outcome model to simultaneously relate extracted MPCA features to the outcome. The probabilistic MPCA model is closely related to a latent factor model in formulation, but extends it to further take into account the inherent matrix structure in the covariate with the MPCA features modeled as latent variables. Our simulations show that our joint modeling approach achieves better performance in both prediction accuracy and estimation of association than the two-stage approach and the classical PCR approach, especially in difficult cases with small sample sizes and large dimensionality in the matrix-valued covariate. Because the estimation errors for both the two-stage and joint modeling approaches decline when the sample size increases, the differences in the predictive accuracy for both methods tend to diminish as seen in our simulation when $n = 200$. We also note that the two-stage approach can provide potential flexibility in relating MPCA features to the outcome, e.g., by using classification/regression trees approach or support vector machine method in the second prediction stage. While such extensions of the outcome submodel may not be always possible within the proposed framework, the joint modeling approach had the benefit of providing a framework to carry out valid statistical inference on the association parameters of interest. In the two-stage approach, we employ the unsupervised MPCA procedure implemented in the R package (rTensor). Alternatively, it is also worthwhile to consider a supervised sufficient dimension reduction approach (e.g., Li et al. (2010)) to extract low-dimensional features, which utilizes the outcome information and may potentially further improve prediction results.

When analyzing our motivating dataset, we found that neither the two-stage approach nor PCR approach could achieve better prediction accuracy than our joint modeling approach. Based on the joint modeling approach, our analysis indicates that EEG CSD measures at the majority of posterior brain regions (except the three electrodes located on the outer boundary of the posterior brain region) at the alpha frequency band are associated with SSRI response. This finding is consistent with previous findings in the field of major depression disorders that the EEG measures at the alpha and theta frequency band have been consistently reported to be associated with antidepressant treatment outcome (e.g., Olbrich and Arns, 2013).

This work can be extended in many ways. For example, the method can be readily extended to accommodate general K–dimensional arrays ($K > 2$) by replacing ($\boldsymbol{B} \otimes \boldsymbol{A}$) in expression (2) by ($\boldsymbol{A}_K \otimes \cdots \otimes \boldsymbol{A}_1$), where $\boldsymbol{A}_k$ is the associated multilinear projection matrix. An additional direction for future work is to consider some structured priors to encourage row-wise and column-wise shrinkage for matrix-valued data, or more generally shrinkage for each dimension for a general K–dimensional array data, leading to variable selection in the

original data space. Alternatively, incorporation of shrinkage or regularization on the coefficients for the latent MPCA features in the outcome model could be helpful, especially when these features are inherently of high dimension. For example, as suggested by one of the reviewers, we could consider incorporating a low-rank bilinear structure, where in model (3), instead of $\boldsymbol{\theta}^{\top}\mathrm{vec}(\boldsymbol{u}_i)$, we consider $\boldsymbol{\theta_1}^{\top}\boldsymbol{u}_i\boldsymbol{\theta_2}$ with $\theta_1 \in \mathscr{R}^{p_0 \times R}$ and $\theta_2 \in \mathscr{R}^{q_0 \times R}$. However, a preliminary analysis to incorporate such a structure in the outcome model for the extracted MPCA features from our EEG data did not further improve the prediction accuracy.

Finally, we note that we select the number of latent features (p_0, q_0) in the MPCA model based on cross-validation, given our primary goals of analysis is association estimation and prediction of the outcome, rather than the more challenging task of estimating the true dimensions or interpreting latent features in an MPCA model. This choice for selecting number of latent features was also suggested by Tipping and Bishop (1999) in the framework of probabilistic PCA. In other contexts, the primary goal may be to estimate the unknown (p_0, q_0) in an MPCA framework and this would be an interesting goal for future investigation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

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

References

Al-Harbi KS (2012). Treatment-resistant depression: therapeutic trends, challenges, and future directions. Patient preference and adherence 6, 369. [PubMed: 22654508]

Cook RD, Li B, and Chiaromonte F (2010). Envelope models for parsimonious and efficient multivariate linear regression. Statistica Sinica pages 927–960.

Ding S and Cook RD (2014). Dimension folding PCA and PFC for matrix-valued predictors. Statistica Sinica 24, 463–492.

Ding S and Cook RD (2018). Matrix variate regressions and envelope models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80, 387–408.

Dunlop BW and Mayberg HS (2014). Neuroimaging-based biomarkers for treatment selection in major depressive disorder. Dialogues in clinical neuroscience 16, 479. [PubMed: 25733953]

Gelman A, Rubin DB, et al. (1992). Inference from iterative simulation using multiple sequences. Statistical science 7, 457–472.

Hoff PD (2007). Model averaging and dimension selection for the singular value decomposition. Journal of the American Statistical Association 102, 674–685.

Hoff PD (2015). Multilinear tensor regression for longitudinal relational data. The annals of applied statistics 9, 1169. [PubMed: 27458495]

Holsboer F (2008). How can we realize the promise of personalized antidepressant medicines? Nature Reviews Neuroscience 9, 638. [PubMed: 18628772]

Hung H and Wang C-C (2013). Matrix variate logistic regression model with application to EEG data. Biostatistics 14, 189–202. [PubMed: 22753784]

Hung H, Wu P, Tu I, and Huang S (2012). On multilinear principal component analysis of order-two tensors. Biometrika 99, 569–583.

Ibrahim JG, Chu H, and Chen LM (2010). Basic concepts and methods for joint models of longitudinal and survival data. Journal of Clinical Oncology 28, 2796. [PubMed: 20439643]

Israel JA (2006). Remission in depression: definition and initial treatment approaches. Journal of psychopharmacology 20, 5–10.

Jiang B, Petkova E, Tarpey T, and Ogden RT (2017). Latent class modeling using matrix covariates with application to identifying early placebo responders based on eeg signals. The annals of applied statistics 11, 1513–1536. [PubMed: 29152032]

Jolliffe IT (2002). Principal component analysis and factor analysis. Principal component analysis pages 150–166.

Kamarajan C, Pandey AK, Chorlian DB, 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. [PubMed: 25448264]

Kayser J and Tenke CE (2006). Principal components analysis of laplacian waveforms as a generic method for identifying ERP generator patterns: I. Evaluation with auditory oddball tasks. Clinical Neurophysiology 117, 348–368. [PubMed: 16356767]

Lawley DN and Maxwell AE (1962). Factor analysis as a statistical method. Journal of the Royal Statistical Society. Series D (The Statistician) 12, 209–229.

Li B, Kim MK, Altman N, et al. (2010). On dimension folding of matrix-or array-valued statistical objects. The Annals of Statistics 38, 1094–1121.

Li L and Zhang X (2017). Parsimonious tensor response regression. Journal of the American Statistical Association 112, 1131–1146.

Lu H, Plataniotis KN, and Venetsanopoulos A (2013). Multilinear subspace learning: dimensionality reduction of multidimensional data. Chapman and Hall/CRC.

Lu H, Plataniotis KN, and Venetsanopoulos AN (2008). MPCA: Multilinear principal component analysis of tensor objects. IEEE Transactions on Neural Networks 19, 18–39. [PubMed: 18269936]

Lu H, Plataniotis KN, and Venetsanopoulos AN (2011). A survey of multilinear subspace learning for tensor data. Pattern Recognition 44, 1540–1551.

Olbrich S and Arns M (2013). EEG biomarkers in major depressive disorder: discriminative power and prediction of treatment response. Intern. Rev. of Psychiatry 25, 604–618.

Ombao H, Lindquist M, Thompson W, and Aston J (2016). Handbook of Neuroimaging Data Analysis. CRC Press.

Petkova E, Ogden RT, Tarpey T, Ciarleglio A, Jiang B, Su Z, Carmody T, Adams P, Kraemer HC, Grannemann BD, et al. (2017). Statistical analysis plan for stage 1 EMBARC (establishing moderators and biosignatures of antidepressant response for clinical care) study. Contemporary Clinical Trials Communications 6, 22–30. [PubMed: 28670629]

R Core Team (2018). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

Svishcheva G, Belonogova N, and Axenovich T (2016). Some pitfalls in application of functional data analysis approach to association studies. Scientific reports 6, 23918. [PubMed: 27041739]

Tenke CE, Kayser J, Svob C, Miller L, Alvarenga JE, Abraham K, Warner V, Wickramaratne P, Weissman MM, and Bruder GE (2017). Association of posterior EEG alpha with prioritization of religion or spirituality: A replication and extension at 20-year follow-up. Biological psychology 124, 79–86. [PubMed: 28119066]

Tipping ME and Bishop CM (1999). Probabilistic principal component analysis. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61, 611–622.

Virta J, Li B, Nordhausen K, and Oja H (2017). Independent component analysis for tensor-valued data. Journal of Multivariate Analysis 162, 172–192.

Wade EC and Iosifescu DV (2016). Using electroencephalography for treatment guidance in major depressive disorder. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging 1, 411–422.

Wang J-L, Chiou J-M, and Müller H-G (2016). Functional data analysis. Annual Review of Statistics and Its Application 3, 257–295.

Xue Y and Yin X (2014). Sufficient dimension folding for regression mean function. Journal of Computational and Graphical Statistics 23, 1028–1043.

Xue Y and Yin X (2015). Sufficient dimension folding for a functional of conditional distribution of matrix-or array-valued objects. Journal of Nonparametric Statistics 27, 253–269.

Zhang X and Li L (2017). Tensor envelope partial least-squares regression. Technometrics 59, 426–436.

Zhao J and Leng C (2014). Structured lasso for regression with matrix covariates. Statistica Sinica pages 799–814.

Zhou H and Li L (2014). Regularized matrix regression. Journal of the Royal Statistical Society 76, 463–483. [PubMed: 24648830]

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. [PubMed: 24791032]

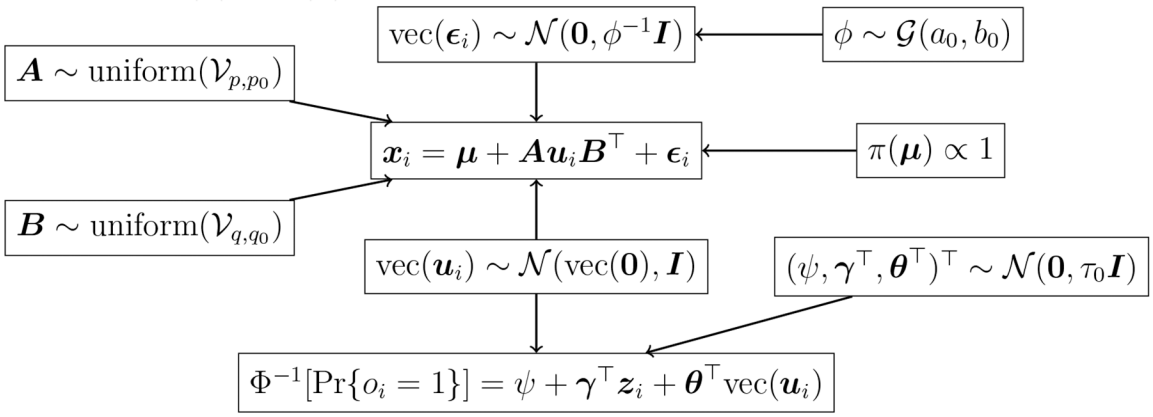
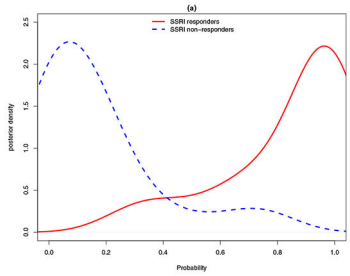
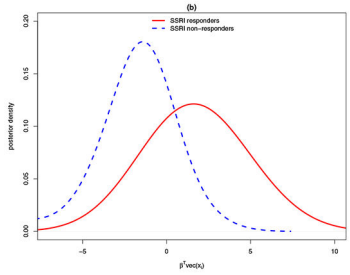


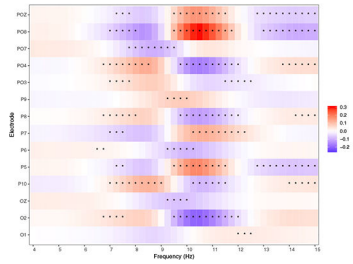
Figure 1.
A graphical representation of the hierarchical structure of the joint model that combines models (1) and (3).



(a) Posterior density estimate of the probability of response to SSRI treatment



(b) Posterior density estimate of $\beta^{\top}\mathrm{vec}(x_i)$.



(c) Heat map of the estimates of the coefficients β for the matrix-valued EEG data when applying our proposed method to predict antidepressant response using baseline EEG data (* indicates significance at the 0.05 level).

Figure 2.

Results of using EEG data to predict response to SSRI treatment. These figures appear in color in the electronic version of this article, and any mention of color refers to that version.

Table 1

The mean Area Under the ROC curves (AUC) for the 200 testing datasets in the simulation, under different simulation scenarios using three approaches (1) joint modeling, (2) two-stage modeling and (3) PCR (true AUC values refer to the mean AUC values obtained using the outcome submodel (3) with true parameter values for the 200 testing datasets.)

	(a) when elements in u_i are independent			**(b) when elements in u_i are correlated**		
	(p,q)=(25,25)	**(p,q)=(30,30)**	**(p,q)=(35,35)**	**(p,q)=(25,25)**	**(p,q)=(30,30)**	**(p,q)=(35,35)**
	n=50, true test AUC = 0.869			**n=50, true AUC = 0.904**		
joint	0.826	0.819	0.818	0.870	0.869	0.866
two-stage	0.822	0.814	0.814	0.864	0.862	0.861
PCR	0.751	0.721	0.688	0.838	0.824	0.793
	n=100, true test AUC = 0.864			**n=100, true test AUC = 0.901**		
joint	0.819	0.822	0.823	0.879	0.874	0.874
two-stage	0.818	0.821	0.822	0.878	0.872	0.873
PCR	0.797	0.783	0.770	0.872	0.863	0.857
	n=200, true test AUC = 0.854			**n=200, true test AUC = 0.902**		
joint	0.830	0.827	0.828	0.882	0.882	0.884
two-stage	0.829	0.828	0.827	0.882	0.881	0.884
PCR	0.822	0.814	0.810	0.880	0.876	0.874

Table 2

The average of the mean squared errors (MSEs) of the estimated coefficients $\widehat{\boldsymbol{\beta}} \in \mathcal{R}^{pq \times 1}$ across 200 datasets in the simulation (the displayed values are the actual values multiplied by 10^3), under different simulation scenarios using three approaches (1) joint modeling, (2) two-stage modeling and (3) PCR

	(a) when elements in u_i, are independent			**(b) when elements in u_i are correlated**		
	(p,q)=(25,25)	**(p,q)=(30,30)**	**(p,q)=(35,35)**	**(p,q)=(25,25)**	**(p,q)=(30,30)**	**(p,q)=(35,35)**
		n=50			**n=50**	
joint	1.235	0.904	0.682	1.350	0.982	0.770
two-stage	2.270	1.859	1.170	3.687	2.766	2.404
PCR	4.056	3.037	2.347	4.041	2.783	2.221
		n=100			**n=100**	
joint	0.640	0.489	0.359	0.713	0.529	0.381
two-stage	0.703	0.556	0.398	0.808	0.590	0.431
PCR	2.802	2.243	1.865	2.196	1.791	1.502
		n=200			**n=200**	
joint	0.289	0.227	0.176	0.338	0.255	0.185
two-stage	0.297	0.236	0.182	0.338	0.259	0.190
PCR	1.742	1.513	1.323	1.273	1.116	0.991

Table 3

Average of ten repeated 10-fold cross-validation AUCs for the prediction of antidepressant responders using EEG data under different models.

(a) Under different choices of MPCA dimensions (p_0; q_0) in (a1) our joint modeling approach and (a2) two-stage modeling approach.

	(a1) joint modeling					(a2) two-stage modeling				
p_0/q_0	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
2	0.457	0.552	0.604	0.570	0.570	0.450	0.558	0.613	0.571	0.570
3	0.655	0.676	0.639	0.646	0.576	0.650	**0.688**	0.632	0.648	0.562
4	0.598	0.649	0.654	**0.697**	0.591	0.599	0.642	0.646	0.660	0.586
5	0.573	0.590	0.631	0.690	0.596	0.567	0.618	0.644	0.670	0.514

(b) Under different choices of dimension r_0 in PCR approach.

r_0	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
	0.497	0.588	0.587	**0.611**	0.552	0.53	0.533	0.544	0.563	0.581	0.547	0.536