Statistica Sinica **29** (2019), 1025-1046
doi:https://doi.org/10.5705/ss.202016.0445

A LOW RANK-BASED ESTIMATION-TESTING PROCEDURE FOR MATRIX-COVARIATE REGRESSION

Hung Hung and Zhi-Yu Jou

National Taiwan University

Abstract: Matrix-covariate is now frequently encountered in many biomedical researches. It is common to fit conventional statistical models by vectorizing matrix-covariate. This strategy results in a large number of parameters, while the available sample size is relatively too small to have reliable analysis results. To overcome the problem of high-dimensionality in hypothesis testing, a variance-component test has been proposed with superior detection power, but it is not straightforward to provide estimates of effect size. In this work, we overcome the problem of high-dimensionality by utilizing the inherent structure of the matrix-covariate. One advantage of our method is that estimation and hypothesis testing can be conducted simultaneously, as in the conventional case, while the estimation efficiency and detection power can be largely improved. Another merit is that, unlike existing methods, the proposed method avoids the problem of choosing identifiability constraints for the model parameters. Our method is applied to test the significance of gene-gene interactions in the PSQI data, and to test the association between electroencephalography and the alcoholic status in the EEG data.

Key words and phrases: High-dimensionality, hypothesis testing, low-rank, matrix-covariate, tensor.

1. Introduction

Techniques of detecting the association between the candidate covariate and the response variable are widely applied in many applications. Recently, matrix-covariate is more frequently encountered in biomedical researches, wherein the research interest focuses on the association between the response variable $Y \in \mathbb{R}$ and the $p \times q$ matrix-covariate $\boldsymbol{M} \in \mathbb{R}^{p\times q}$, possibly after adjusting for the effects of certain confounding factors $Z \in \mathbb{R}^m$. Using the Electroencephalography (EEG) data as our motivating example, Y is the binary alcoholic status, and $\boldsymbol{M}$ is a 256×64 matrix with its (j,k)-th element $\boldsymbol{M}(j,k)$ being the voltage value of the k-th electrode measured at the j-th time point. Figure 1 displays the voltage values $\boldsymbol{M}$ of one randomly selected subject from both the alcoholic group $(Y=1)$ and the control group $(Y=0)$. It can be seen that $\boldsymbol{M}|Y=1$ tends to be larger

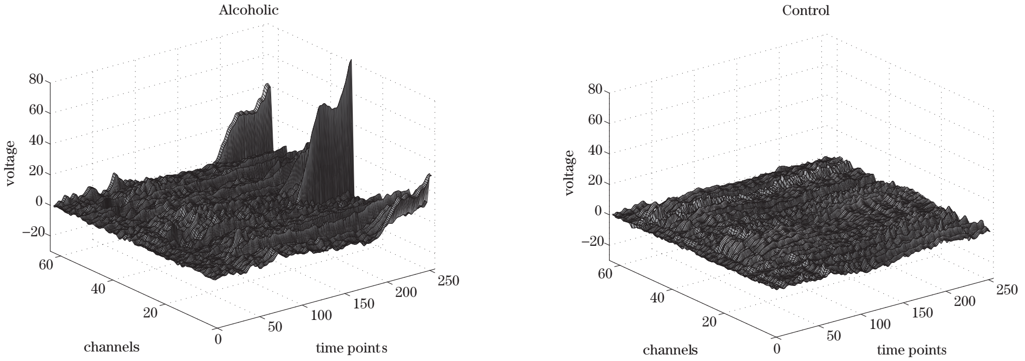

Figure 1. The voltage values of 64 channels at 256 time points from two randomly selected subjects. The left panel is from the alcoholic group (Y = 1), and the right panel is from the control group (Y = 0).

than $\boldsymbol{M}|Y = 0$ for some (j, k) values, indicating a strong association between Y and $\boldsymbol{M}$, and it is of interest to investigate this association in a quantitative manner. Sometimes $\boldsymbol{M}$ is not directly observed but is induced from the original covariates. The Pittsburgh Sleep Quality Index (PSQI) data collects for each subject the PSQI score (Y) and the multiple measurements of the genes ROR (with 19 markers) and NR1D1 (with 4 markers). Past study has found that ROR and NR1D1 cannot explain the variation of PSQI score well, and has suggested fitting the PSQI score on the gene-gene interactions (G×G) between ROR and NR1D1. Let $G = (g_1, \dots g_p)^\top$ denote the p-genetic markers (e.g., ROR) and let $E = (e_1, \dots e_q)^\top$ denote another q-genetic markers (e.g., NR1D1). It is equivalent to investigate if Y is associated with the pq products $\{g_j e_k : 1 \leq j \leq p, 1 \leq k \leq q\}$, or equivalently, with the matrix-covariate $\boldsymbol{M} = GE^\top$ by noting that $\boldsymbol{M}(j, k) = g_j e_k$, after adjusting for the effects of $Z = (G^\top, E^\top)^\top$. See Section 5 for more details of the EEG and PSQI data sets.

The research interests of these motivating examples all focus on the association between Y and the matrix-covariate $\boldsymbol{M}$, and two questions are commonly raised by practitioners:

(Q1) Does Y associate with $\boldsymbol{M}$?

(Q2) How does $\boldsymbol{M}$ affect Y?

To answer (Q1), a commonly used method is to fit a GLM for Y on each element of $\boldsymbol{M}$ separately, and then use the minimum p-value of the pq marginal tests as the test statistic (denoted by min-test). The min-test, however, can produce biased results due to the ignorance of the joint effects of $\boldsymbol{M}$. Joint inference is thus preferable, which fits the GLM (McCullagh and Nelder (1989))

$$Y|(Z,\boldsymbol{M}) \sim \text{Normal}(E(Y|Z,\boldsymbol{M}),\sigma^2) \tag{1.1}$$

for continuous Y with a common variance σ^2, or

$$Y|(Z,\boldsymbol{M}) \sim \text{Bernoulli}(E(Y|Z,\boldsymbol{M})) \tag{1.2}$$

for binary Y, and assumes that

$$g\{E(Y|Z,\boldsymbol{M})\} = \gamma + \xi^\top Z + \sum_{j,k} \eta_{jk}\boldsymbol{M}(j,k), \tag{1.3}$$

where g is the link function, γ is the intercept term, ξ is the effect of Z, and η_{jk} is the effect of $\boldsymbol{M}(j,k)$. Following convention, we adopt the identity link $g(u)=u$ for model (1.1), and the logit link $g(u)=\ln(u/(1-u))$ for model (1.2). Based on (1.3), answering (Q1)-(Q2) relies on estimating η_{jk}'s and testing the overall hypothesis

$$H_0 : \eta_{jk} = 0 \;\; \forall \;\; (j,k). \tag{1.4}$$

Although the joint method overcomes the problem of bias, it suffers from high-dimensionality, since the number of parameters for η_{jk}'s, say pq, can be large in comparison with the sample size n. In the EEG data, for example, there are $256\times 64 = 16{,}384$ parameters for the η_{jk}'s, while the available sample size is 122.

To overcome this problem in testing (1.4), Lin et al. (2013) apply the variance-component test (Lin (1997)) to propose GESAT. The authors extend model (1.3) to the generalized linear mixed model (GLMM) by assuming that the η_{jk}'s independently follow an arbitrary distribution with zero mean and a common variance τ^2. As a result, testing (1.4) is equivalent to testing $H_0 : \tau^2 = 0$ under the GLMM, and the test statistic of GESAT is

$$T_{\text{gesat}} = \left\| \sum_{i=1}^{n} (Y_i - \widetilde{\gamma} - \widetilde{\xi}^\top Z_i)\text{vec}(\boldsymbol{M}_i) \right\|^2 , \tag{1.5}$$

where $\{(Y_i, Z_i, \boldsymbol{M}_i)\}_{i=1}^n$ is a random sample from $(Y, Z, \boldsymbol{M})$, $\text{vec}(\boldsymbol{M}_i)$ is the pq-vector from stacking the columns of $\boldsymbol{M}_i$, and $(\widetilde{\gamma},\widetilde{\xi})$ is the restricted MLE of (γ,ξ) under (1.4). GESAT has the advantage of fast computation and is shown to be locally most powerful (Goeman, van de Geer and van Houwelingen (2006)). GESAT, however, aims to answering (Q1), but is not straightforward to provide estimates of the effect sizes of $\boldsymbol{M}$, (Q2). A naive solution is to fit model (1.3) to obtain the estimates of the η_{jk}'s, but this again suffers from high-dimensionality, and there is no guarantee that the estimates coincide with the conclusion from

GESAT.

Another set of methods overcome the problem of high-dimensionality by utilizing the matrix structure of $\boldsymbol{M}$. The main idea is to re-express (1.3) as

$$g\{E(Y|Z,\boldsymbol{M})\} = \gamma + \xi^\top Z + \mathrm{vec}(\boldsymbol{\eta})^\top \mathrm{vec}(\boldsymbol{M}), \tag{1.6}$$

where $\boldsymbol{\eta}$ is the $p \times q$ matrix with the (j,k)-th element being η_{jk}. Under (1.6), answering (Q1)-(Q2) relies on estimating the matrix $\boldsymbol{\eta}$ and testing

$$H_0 : \boldsymbol{\eta} = \mathbf{0}_{p\times q}. \tag{1.7}$$

Of course there is no difference between models (1.3) and (1.6), but model (1.6) that preserves the matrix structure of $\boldsymbol{M}$ provides a way to more efficiently estimate $\boldsymbol{\eta}$, via utilizing the matrix structure of $\boldsymbol{\eta}$. Hung and Wang (2013) propose the *matrix-variate logistic (MV-logistic) regression* for binary Y, by fitting (1.6) with the rank-1 constraint $\boldsymbol{\eta} = \boldsymbol{A}\boldsymbol{B}^\top$ for some $\boldsymbol{A} \in \mathbb{R}^p$ and $\boldsymbol{B} \in \mathbb{R}^q$. Zhou, Li and Zhu (2013) propose *tensor regression* by fitting (1.6) with the rank-r constraint $\boldsymbol{\eta} = \boldsymbol{A}\boldsymbol{B}^\top$ for some $\boldsymbol{A} \in \mathbb{R}^{p\times r}$, $\boldsymbol{B} \in \mathbb{R}^{q\times r}$, and $r \geq 1$. The main advantage of these methods is that one only requires $(p+q)r$ parameters to model $\boldsymbol{\eta}$ instead of pq, and an efficiency gain in estimating $\boldsymbol{\eta}$ is reasonably expected. Tensor regression, however, has the drawback that $(\boldsymbol{A},\boldsymbol{B})$ are not identifiable. This can be seen from $\boldsymbol{A}\boldsymbol{B}^\top = \boldsymbol{A}\boldsymbol{C}\boldsymbol{C}^{-1}\boldsymbol{B}^\top$ for any nonsingular $\boldsymbol{C} \in \mathbb{R}^{r\times r}$. To avoid this, Hung and Wang (2013) and Zhou, Li and Zhu (2013) impose extra constraints on $(\boldsymbol{A},\boldsymbol{B})$ so that conventional MLE arguments can be applied to develop asymptotic properties. Since the identifiability constraints are not unique, different choices of the constraints produce different analysis results, and this limits the applicabilities of the methods. Without using a parsimonious parameterization of $\boldsymbol{\eta}$, Zhou and Li (2014) propose *regularized matrix regression* by estimating $\boldsymbol{\eta}$ with the penalized MLE, where the imposed penalty function depends on $\boldsymbol{\eta}$ only through its singular values. Although regularized matrix regression avoids the problem of choosing identifiability constraints, it requires pq parameters to model $\boldsymbol{\eta}$, which can make the model fitting less efficient, especially for the case of large (p,q). Moreover, the asymptotic properties of the resulting estimator are difficult to derive. These methods all focus on the estimation of $\boldsymbol{\eta}$, (Q2), and extensions of these methods to testing (1.7), (Q1), are not discussed.

The aim of this study is to propose a unified inference procedure for (Q1)-(Q2) that adapts to the high-dimensional setting, while overcoming these drawbacks of existing methods. Our proposal follows tensor regression by considering the joint model (1.6) with the rank-r constraint $\boldsymbol{\eta} = \boldsymbol{A}\boldsymbol{B}^\top$. One contribution of this work is the development of an asymptotic property that is invariant to the choice of the identifiability constraint on $(\boldsymbol{A}, \boldsymbol{B})$, an essential difference to the works of Hung and Wang (2013) and Zhou, Li and Zhu (2013). Based on this asymptotic property, we propose a low rank-based test statistic for (1.7). The test statistic is also invariant to the choice of the identifiability constraint and, hence, is more applicable in applications. Comparing with GESAT, our method cannot only improve the detection power by utilizing the matrix structure of $\boldsymbol{\eta}$, but can provide estimate of $\boldsymbol{\eta}$ at the same time.

Some notation is set out here for ease of reference. For any $p \times q$ matrix $\boldsymbol{M}$, $\text{vec}(\boldsymbol{M})$ is the pq-vector achieved by stacking the columns of $\boldsymbol{M}$, $\boldsymbol{K}_{p,q}$ is the commutation matrix (Henderson and Searle (1979)) satisfying $\boldsymbol{K}_{p,q}\text{vec}(\boldsymbol{M}) = \text{vec}(\boldsymbol{M}^\top)$, and $\|\boldsymbol{M}\|_F$ denotes the Frobenius norm of $\boldsymbol{M}$. $\otimes$ stands for the Kronecker product, and $\|\cdot\|$ stands for the Euclidean norm of a vector. α denotes the significance level.

The rest of this paper is organized as follows. Our inference procedure for $\boldsymbol{\eta}$ is developed in Section 2, based on which the test statistics for (1.7) are proposed in Section 3. Section 4 conducts numerical studies to evaluate the performances of our proposal, and Section 5 conducts analyses for the PSQI and EEG data sets. The paper ends with a discussion in Section 6.

Remark 1. The idea of treating G $\times$ G as a matrix $\boldsymbol{M} = GE^\top$ is motivated from Hung et al. (2016). The authors also develop asymptotic property that is invariant to the choice of the identifiability constrain, but their result can only be applied under the normal model (1.1). Our asymptotic property extends the result of Hung et al. (2016) to the more general GLM that includes models (1.1)-(1.2) as special cases. Moreover, we aim to develop test statistic for (1.7), and this problem is not investigated in Hung et al. (2016)

2. The Low-Rank Inference Procedure for $\boldsymbol{\eta}$

2.1. Model specification

The rationale behind our proposal is based on the assumption that *most of the row and/or column attributes of $\boldsymbol{M}$ play no role in explaining Y*. In the EEG data, it is common that only a small portion of probes (the column attribute of $\boldsymbol{M}$) are relevant to the alcoholic status. In the PSQI data, it is believed that only a few elements of ROR and NR1D1 (the row and column attributes of $\boldsymbol{M}$) are influential to the PSQI score. This implies that most of the rows and/or columns of $\boldsymbol{\eta}$ are zeros, $\boldsymbol{\eta}$ is a low-rank matrix. It is thus reasonable to consider the rank-r GLM (Hung and Wang (2013); Zhou, Li and Zhu (2013)),

$$g\{E(Y|Z,\boldsymbol{M})\} = \gamma + \xi^\top Z + \mathrm{vec}(\boldsymbol{\eta})^\top \mathrm{vec}(\boldsymbol{M}) \quad \text{with} \quad \boldsymbol{\eta} = \boldsymbol{A}\boldsymbol{B}^\top, \tag{2.1}$$

where $\boldsymbol{A} \in \mathbb{R}^{p\times r}$, $\boldsymbol{B} \in \mathbb{R}^{q\times r}$, and $1 \le r \le \min\{p,q\}$ such that $\mathrm{rank}(\boldsymbol{\eta}) = r$. Here $(\boldsymbol{A},\boldsymbol{B})$ are not identifiable. Without imposing extra constraints on $(\boldsymbol{A},\boldsymbol{B})$, we leave $(\boldsymbol{A},\boldsymbol{B})$ arbitrary to develop our inference procedure for $\boldsymbol{\eta}$. This is achievable since we are interested in $\boldsymbol{\eta}$ instead of $(\boldsymbol{A},\boldsymbol{B})$, and only the identifiability of $\boldsymbol{A}\boldsymbol{B}^\top$ is required. We will see in Section 2.3 that the developed asymptotic properties depend on $(\boldsymbol{A},\boldsymbol{B})$ only through $\mathrm{span}(\boldsymbol{A})$ and $\mathrm{span}(\boldsymbol{B})$. As a result, we can use the convenient parameterization $\boldsymbol{\eta} = \boldsymbol{A}\boldsymbol{B}^\top$ to make inference about $\boldsymbol{\eta}$ without imposing any constraint on $(\boldsymbol{A},\boldsymbol{B})$, which makes our method more applicable in practice.

One advantage of model (2.1) is the parsimony of parameters. The conventional model (1.3) requires $1 + m + pq$ parameters. Since one only requires $(p+q-r)r$ parameters to specify a rank-r $p \times q$ matrix, the effective number of parameters in model (2.1) is

$$s_r = 1 + m + (p+q-r)r. \tag{2.2}$$

Comparing s_r with $1+m+pq$, we expect an efficiency gain by fitting model (2.1) when r is small. We note that the effective number of parameters s_r is only used to have insight about the advantage of model (2.1). The development of the inference procedures for $\boldsymbol{\eta}$ is still based on the convenient parameterization $\boldsymbol{\eta} = \boldsymbol{A}\boldsymbol{B}^\top$ without imposing any constraint on $(\boldsymbol{A},\boldsymbol{B})$.

2.2. Estimation and implementation

Let the data $\{(Y_i, Z_i, \boldsymbol{M}_i)\}_{i=1}^n$ be random copies of $(Y, Z, \boldsymbol{M})$, and let the vector of covariates be $X_i = (1, Z_i^\top, \mathrm{vec}(\boldsymbol{M}_i)^\top)^\top$. Let

$$\theta = (\gamma, \xi^\top, \mathrm{vec}(\boldsymbol{A})^\top, \mathrm{vec}(\boldsymbol{B})^\top)^\top \tag{2.3}$$

be the parameters of model (2.1), and take the induced parameters of interest to be

$$\beta = \beta(\theta) = (\gamma, \xi^\top, \mathrm{vec}(\boldsymbol{A}\boldsymbol{B}^\top)^\top)^\top, \tag{2.4}$$

consisting of the intercept, the effect of Z, and the effect of $\boldsymbol{M}$.

The log-likelihood function of θ is $(n/\sigma^2)\ell(\theta) - (n/2)\ln(2\pi\sigma^2)$, with

$$\ell(\theta) = -\frac{1}{2n}\sum_{i=1}^n \{Y_i - \beta(\theta)^\top X_i\}^2 \tag{2.5}$$

under the normal model (1.1), and is

$$\ell(\theta) = \frac{1}{n}\sum_{i=1}^{n} Y_i\{\beta(\theta)^\top X_i\} - \ln(1 + \exp(\beta(\theta)^\top X_i)) \tag{2.6}$$

under the logistic model (1.2). To stabilize the estimation process, a widely used strategy is to employ the penalized MLE. A natural choice of the penalty is $\|\boldsymbol{A}\boldsymbol{B}^\top\|_F^2$, but it may cause instability in implementation. Since $\|\boldsymbol{A}\boldsymbol{B}^\top\|_F \le \|\boldsymbol{A}\|_F\|\boldsymbol{B}\|_F$, we propose to estimate θ by

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmax}}\left\{\ell(\theta) - \frac{1}{2}\lambda\|\boldsymbol{A}\|_F^2\|\boldsymbol{B}\|_F^2\right\}, \tag{2.7}$$

where $\lambda \ge 0$ controls the effect of penalty. As will be shown in (2.9)-(2.10), the merit of using $\|\boldsymbol{A}\|_F^2\|\boldsymbol{B}\|_F^2$ is that (2.7) can be obtained by iteratively solving a conventional L_2-penalized MLE problem where existing algorithm can be directly applied (Le Cessie and Van Houwelingen (1992)). With $\hat{\theta}$ being obtained, β is estimated by

$$\widehat{\beta} = \beta(\widehat{\theta}) = (\widehat{\gamma}, \widehat{\xi}^\top, \mathrm{vec}(\widehat{\boldsymbol{\eta}})^\top)^\top \quad \text{with} \quad \widehat{\boldsymbol{\eta}} = \widehat{\boldsymbol{A}}\widehat{\boldsymbol{B}}^\top. \tag{2.8}$$

Obviously, the performance of $\widehat{\beta}$ depends on the values of (r, λ). A simple idea is to select both (r, λ) by cross-validation. Here $\widehat{\beta}$ is a continuous function of λ, but such continuity does not hold for the discrete parameter r. This indicates that $\widehat{\beta}$ is more sensitive to the selection of r. We propose a more stable selection criterion for r via the number of effective parameters s_r. First observe that an over-specification of r does not affect the consistency of $\widehat{\beta}$, while an under-specification of r can make $\widehat{\beta}$ biased. This suggests that a large r should be used to ensure the validity of $\widehat{\beta}$. On the other hand, we prefer a small r such that the rank-r GLM can be well fitted with sample size n. Consequently, we are motivated to select r as large as possible while keeping s_r relatively smaller than n (e.g., $n/s_r \ge 3$). We thus suggest selecting a large r so that the rank-r GLM is more plausibly correct, while preserving the estimation efficiency of $\widehat{\beta}$ with limited sample size. With a fixed r, λ is selected by cross-validation.

To implement our method, we use the alternating method to solve (2.7). Model (2.1) can be expressed as

$$g\{E(Y|Z)\} = \gamma + \xi^\top Z + \mathrm{vec}(\boldsymbol{A})^\top\mathrm{vec}(\boldsymbol{M}\boldsymbol{B}) \tag{2.9}$$

$$= \gamma + \xi^\top Z + \mathrm{vec}(\boldsymbol{B})^\top\mathrm{vec}(\boldsymbol{M}^\top\boldsymbol{A}). \tag{2.10}$$

Observe that (2.9) is the GLM with parameters $(\gamma, \xi^\top, \mathrm{vec}(\boldsymbol{A})^\top)^\top$ and data $(Y, Z, \boldsymbol{M}\boldsymbol{B})$. Thus, when $\boldsymbol{B}$ is fixed, maximizing (2.7) is the conventional L_2-penalized MLE problem under (2.9) with penalty $(1/2)\lambda_{\boldsymbol{B}}$ for $\|\boldsymbol{A}\|_F^2$, where $\lambda_{\boldsymbol{B}} = \lambda\|\boldsymbol{B}\|_F^2$. The case for fixed $\boldsymbol{A}$ is made similar by using the data $(Y, Z, \boldsymbol{M}^\top\boldsymbol{A})$

to fit model (2.10) with parameters $(\gamma, \xi^\top, \mathrm{vec}(\boldsymbol{B})^\top)^\top$ and penalty $(1/2)\lambda_{\boldsymbol{A}}$ for $\|\boldsymbol{B}\|_F^2$, where $\lambda_{\boldsymbol{A}} = \lambda\|\boldsymbol{A}\|_F^2$. We then iterate the roles of $\boldsymbol{A}$ and $\boldsymbol{B}$ until convergence. The implementation algorithm is given below.

Alternating Algorithm

1. Given an initial value $\boldsymbol{B}_{(0)}$, $k = 0, 1, 2, \ldots$, do Steps 2-4.
2. Given $\boldsymbol{B}_{(k)}$, obtain the L_2-penalized MLE $(\gamma^*_{(k+1)}, \xi^*_{(k+1)}, \boldsymbol{A}_{(k+1)})$ from fitting (2.9) on the data $\{(Y_i, Z_i, \boldsymbol{M}_i\boldsymbol{B}_{(k)})\}_{i=1}^n$ with the penalty $(1/2)\lambda_{\boldsymbol{B}_{(k)}}$ for $\|\boldsymbol{A}\|_F^2$.
3. Given $\boldsymbol{A}_{(k+1)}$, obtain the L_2-penalized MLE $(\gamma_{(k+1)}, \xi_{(k+1)}, \boldsymbol{B}_{(k+1)})$ from fitting (2.10) on the data $\{(Y_i, Z_i, \boldsymbol{M}_i^\top\boldsymbol{A}_{(k+1)})\}_{i=1}^n$ with the penalty $(1/2)\lambda_{\boldsymbol{A}_{(k+1)}}$ for $\|\boldsymbol{B}\|_F^2$.
4. Let $\theta_{(k+1)} = (\gamma_{(k+1)}, \xi_{(k+1)}^\top, \mathrm{vec}(\boldsymbol{A}_{(k+1)})^\top, \mathrm{vec}(\boldsymbol{B}_{(k+1)})^\top)^\top$. Repeat the procedure until the convergence of $\beta(\theta_{(k+1)})$. Output $\widehat{\beta} = \beta(\theta_{(\infty)})$ and $\widehat{\theta} = \theta_{(\infty)}$.

For each iteration, the maximum log-likelihood attained in either Step 2 or Step 3 of Alternating Algorithm cannot decrease. To see this, let $f(\gamma, \xi, \boldsymbol{A}, \boldsymbol{B}) = \ell(\theta) - (\frac{\lambda}{2})\|\boldsymbol{A}\|_F^2\|\boldsymbol{B}\|_F^2$ be the objective function being maximized. We must have

$$\begin{aligned} f(\gamma_{(k)}, \xi_{(k)}, \boldsymbol{A}_{(k)}, \boldsymbol{B}_{(k)}) &\overset{\text{Step 2}}{\leq} f(\gamma^*_{(k+1)}, \xi^*_{(k+1)}, \boldsymbol{A}_{(k+1)}, \boldsymbol{B}_{(k)}) \\ &\overset{\text{Step 3}}{\leq} f(\gamma_{(k+1)}, \xi_{(k+1)}, \boldsymbol{A}_{(k+1)}, \boldsymbol{B}_{(k+1)}). \end{aligned}$$

Since f is bounded by 0, the algorithm must converge to a local maximum. The algorithm depends on the selection of an initial value $\boldsymbol{B}_{(0)}$ and suggest choosing $\boldsymbol{B}_{(0)}$ as the leading r right singular vectors of $\widetilde{\boldsymbol{\eta}}$, where $\widetilde{\boldsymbol{\eta}}$ is the L_2-penalized MLE of $\boldsymbol{\eta}$ under the conventional model (1.6). We also suggest using multiple random initial values to find the global maximum.

2.3. Asymptotic property

We now proceed to derive the asymptotic property of $\widehat{\beta}$. Let $\beta_0 = (\gamma_0, \xi_0^\top, \mathrm{vec}(\boldsymbol{\eta}_0)^\top)^\top$ be the true value of β under model (2.1), and assume the existence of a regular point (Shapiro (1986)) θ_0 in the parameter space of θ such that $\beta_0 = \beta(\theta_0)$. Take

$$\boldsymbol{\Delta}(\theta) = \frac{\partial\beta(\theta)}{\partial\theta} = \begin{bmatrix} \boldsymbol{I}_{m+1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (\boldsymbol{B} \otimes \boldsymbol{I}_p) & (\boldsymbol{I}_q \otimes \boldsymbol{A})\boldsymbol{K}_{q,r} \end{bmatrix} \tag{2.11}$$

and let $\boldsymbol{\Delta}_0 = \boldsymbol{\Delta}(\theta_0)$. We state asymptotic property of $\widehat{\beta}$, its proof is deferred to the Supplementary Materials.

Theorem 1. *Assume the validity of model* (2.1) *with* $\beta_0 = \beta(\theta_0)$ *and* $\text{rank}(\boldsymbol{\Delta}_0) = s_r$*. If* $\lambda = o_p(n^{-1/2})$*, then, with fixed* (p, q) *and as* $n \to \infty$*,* $\sqrt{n}(\widehat{\beta} - \beta_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_0)$ *with the asymptotic covariance matrix* $\boldsymbol{\Sigma}_0 = \boldsymbol{\Delta}_0(\boldsymbol{\Delta}_0^\top \boldsymbol{V}_0 \boldsymbol{\Delta}_0)^+ \boldsymbol{\Delta}_0^\top$*, where*

(a) $\boldsymbol{V}_0 = \sigma^{-2} E(X_i X_i^\top)$ *for the normal model* (1.1),

(b) $\boldsymbol{V}_0 = E(\nu_i(\theta_0) X_i X_i^\top)$ *with* $\nu_i(\theta) = \exp(\beta(\theta)^\top X_i)/\{1 + \exp(\beta(\theta)^\top X_i)\}^2$ *for the logistic model* (1.2).

The asymptotic property of $\hat{\beta}$ is the core for developing our test statistics in Section 3. We propose estimating the asymptotic covariance matrix $\boldsymbol{\Sigma}_0$ by the sandwich-type estimator

$$\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Delta}} \left\{ \widehat{\boldsymbol{\Delta}}^\top (\widehat{\boldsymbol{V}} + \lambda \boldsymbol{D}) \widehat{\boldsymbol{\Delta}} \right\}^+ \widehat{\boldsymbol{\Delta}}^\top \widehat{\boldsymbol{V}} \widehat{\boldsymbol{\Delta}} \left\{ \widehat{\boldsymbol{\Delta}}^\top (\widehat{\boldsymbol{V}} + \lambda \boldsymbol{D}) \widehat{\boldsymbol{\Delta}} \right\}^+ \widehat{\boldsymbol{\Delta}}^\top, \qquad (2.12)$$

where $\widehat{\boldsymbol{\Delta}} = \boldsymbol{\Delta}(\widehat{\theta})$, $\widehat{\boldsymbol{V}} = \hat{\sigma}^{-2} \cdot (1/n) \sum_{i=1}^n X_i X_i^\top$ with $\hat{\sigma}^2 = \{1/(n - s_r)\} \sum_{i=1}^n (Y_i - \widehat{\beta}^\top X_i)^2$ for the case of (1.1) and $\widehat{\boldsymbol{V}} = (1/n) \sum_{i=1}^n \nu_i(\hat{\theta}) X_i X_i^\top$ for the case of (1.2). Here $\boldsymbol{D}$ is a block-diagonal matrix with diagonal elements $(\mathbf{0}_{1+m}, \|\widehat{\boldsymbol{B}}\|_F^2 \boldsymbol{I}_{pr}, \|\widehat{\boldsymbol{A}}\|_F^2 \boldsymbol{I}_{qr})$.

3. Detecting the Significance of $\boldsymbol{\eta}$

3.1. The low rank-based test statistic

This section develops test statistics for the null hypothesis (1.7). A natural strategy for testing (1.7) via model (2.1) is to use the Wald-test statistic

$$T_{\text{wald}} = \text{vec}(\widehat{\boldsymbol{\eta}})^\top \left\{ [\widehat{\boldsymbol{\Sigma}}]_{\boldsymbol{\eta}} / n \right\}^+ \text{vec}(\widehat{\boldsymbol{\eta}}), \qquad (3.1)$$

where $[\widehat{\boldsymbol{\Sigma}}]_{\boldsymbol{\eta}}$ is the sub-matrix of $\widehat{\boldsymbol{\Sigma}}$ that corresponds to the asymptotic covariance matrix of $\text{vec}(\widehat{\boldsymbol{\eta}})$ in Theorem 1. Here $[\widehat{\boldsymbol{\Sigma}}]_{\boldsymbol{\eta}}$ is singular due to the over-parameterization of $\boldsymbol{\eta} = \boldsymbol{A}\boldsymbol{B}^\top$, and the Moore-Penrose generalized inverse is used. Observe that T_{wald} is a weighted sum over the differences $(\widehat{\boldsymbol{\eta}} - \mathbf{0})$. Thus, T_{wald} can be less powerful in testing (1.7) when $\boldsymbol{\eta}$ is sparse, as the contribution of differences can be averaged out during summation. A better strategy for testing (1.7) with sparse $\boldsymbol{\eta}$ is to use the test statistic

$$T_{\max} = \max_{l \in \{m+2, \ldots, 1+m+pq\}} \frac{\widehat{\beta}_l^2}{[\widehat{\boldsymbol{\Sigma}}]_{\beta_l} / n}, \qquad (3.2)$$

where the maximum is taken over the estimates of $\boldsymbol{\eta}$. $T_{\max}$ is a generalization of the min-test, in the sense that $T_{\max}$ further considers the joint effects among $\boldsymbol{M}$. As opposed to $(T_{\text{wald}}, T_{\max})$ that utilize the low-rank structure of $\boldsymbol{\eta}$, T_{gesat} uses the technique of variance-component test to overcome the problem of high-dimensionality. It is known that T_{gesat} is locally most powerful, and T_{gesat} is thus expected to have superior performance when $\boldsymbol{\eta}$ has weak effect. However, there is no guarantee for its performance otherwise, and ignoring the low-rank structure of $\boldsymbol{\eta}$, plausibly true in many applications, may also decrease the detection power of T_{gesat}.

In all, $(T_{\text{wald}}, T_{\max}, T_{\text{gesat}})$ have their own merits in testing (1.7), depending on the underlying characteristic of $\boldsymbol{\eta}$: T_{wald} is preferred when $\boldsymbol{\eta}$ has dense effect, $T_{\max}$ is preferred when $\boldsymbol{\eta}$ has sparse effect, and T_{gesat} is preferred when $\boldsymbol{\eta}$ has weak effect. Ideally, one should choose the test statistic according to the alternative hypothesis, which is rarely known a priori. A reasonable strategy then is to combine $(T_{\text{wald}}, T_{\max}, T_{\text{gesat}})$ to adapt to various situations. There exist many combination methods based on the p-values, but they do not apply in our case since the limiting distribution of $T_{\max}$ is not easy to derive. Alternatively, we propose to use the product

$$T = T_{\text{wald}} \cdot T_{\max} \cdot T_{\text{gesat}} \tag{3.3}$$

as the test statistic, where a large value of T indicates a rejection of (1.7). The advantage of the multiplicative combination is that T is less affected by the scales of $(T_{\text{wald}}, T_{\max}, T_{\text{gesat}})$. We emphasize that T is developed based on the asymptotic property in Theorem 1 and, hence, is also invariant to the choice of the identifiability constraint of $(\boldsymbol{A}, \boldsymbol{B})$.

Remark 2. Besides the overall hypothesis (1.7), one can also identify the significant covariates $\boldsymbol{M}(j,k)$'s by considering the pq individual hypotheses

$$H_0^{(jk)} : \eta_{jk} = 0, \quad 1 \leq j \leq p, \quad 1 \leq k \leq q. \tag{3.4}$$

Let ρ_{jk} be the p-value of $\widehat{\beta}_l^2/([\widehat{\boldsymbol{\Sigma}}]_{\beta_l}/n)$, where l is such that $\beta_l = \eta_{jk}$. The identified significant covariates are $\{\boldsymbol{M}(j,k) : \rho_{jk} < \alpha/(pq), 1 \leq j \leq p, 1 \leq k \leq q\}$ with the family-wise error rate being controlled at α by Bonferroni correction.

3.2. Calculation of p-value

Since the null distribution of T is not straightforward to derive, we propose to use parametric bootstrap to obtain its p-value. The idea of the parametric bootstrap, as summarized below, is to generate the null data from model (2.1) given $(\gamma, \xi, \boldsymbol{\eta}) = (\widetilde{\gamma}, \widetilde{\xi}, 0)$, where $(\widetilde{\gamma}, \widetilde{\xi})$ are the restricted MLE of (γ, ξ) under H_0

(Bůžková, Lumely and Rice (2011)).

Parametric bootstrap test

1. Conditional on $(Z_i, \boldsymbol{M}_i)$, generate $Y_i^{(b)}$ from model (2.1) with $(\gamma, \xi, \boldsymbol{\eta}) = (\widetilde{\gamma}, \widetilde{\xi}, 0)$ (and with $\sigma^2 = [1/\{n-(m+1)\}] \sum_{i=1}^n (Y_i - \widetilde{\gamma} - \widetilde{\xi}^\top Z_i)^2$ for normal model). Obtain the test statistic $T^{(b)}$ by fitting model (2.1) using $\{(Y_i^{(b)}, Z_i, \boldsymbol{M}_i)\}_{i=1}^n$.
2. Obtain the p-value of T as $(1/b') \sum_{b=1}^{b'} I(T^{(b)} > T)$ for a large number b'.

Although the parametric bootstrap procedure can be applied in various situations, its performance depends on the randomness of $(\widetilde{\gamma}, \widetilde{\xi})$. As a result, using the parametric bootstrap can only control the type-I error asymptotically. Alternatively, an exact test can be constructed when model (2.1) reduces to

$$g\{E(Y|\boldsymbol{M})\} = \gamma + \mathrm{vec}(\boldsymbol{\eta})^\top \mathrm{vec}(\boldsymbol{M}) \quad \text{with} \quad \boldsymbol{\eta} = \boldsymbol{A}\boldsymbol{B}^\top. \tag{3.5}$$

In this situation, the null data can be generated simply by randomly permuting Y_i to destroy its connection with $\boldsymbol{M}_i$.

Permutation test

1. Generate $\{Y_i^{(b)}\}_{i=1}^n$ by randomly permutating $\{Y_i\}_{i=1}^n$. Obtain the test statistic $T^{(b)}$ by fitting model (3.5) using $\{(Y_i^{(b)}, \boldsymbol{M}_i)\}_{i=1}^n$.
2. Obtain the p-value of T as $(1/b') \sum_{b=1}^{b'} I(T^{(b)} > T)$ for a large number b'.

The permutation test cannot be applied in the presence of Z, as permuting Y destroys not only its connection with $\boldsymbol{M}$ but also its connection with Z, which makes the resulting p-value biased. Both re-sampling procedures are suggested to obtain the p-values, according to the underlying data structure. Moreover, these procedures can also be used to obtain the p-values of $(T_{\mathrm{wald}}, T_{\max})$, and the p-values ρ_{jk}'s of the individual tests discussed in Remark 2.

4. Simulation Studies

4.1. Simulation settings

Simulation studies were conducted to evaluate our method, where we used the PSQI and EEG data sets to generate the simulation data:

(PSQI) Let $Z = (G^\top, E^\top)^\top$ and $\boldsymbol{M} = GE^\top$, where $G \in \mathbb{R}^{15}$ and $E \in \mathbb{R}^7$ randomly generated from ROR and NR1D1 of the PSQI data with $n = 400$. Given $(Z, \boldsymbol{M})$, Y was generated from the normal model (1.1) and (2.1) with $\gamma = 10$.

(EEG) The matrix $\boldsymbol{M} \in \mathbb{R}^{6\times 6}$ was generated by randomly selecting six rows and six columns of the EEG signals with $n = 150$. Given $\boldsymbol{M}$, Y was generated from the logistic model (1.2) and (3.5) with $\gamma = 0$.

We considered three simulations with different specifications of $(\xi, \boldsymbol{\eta})$. The first study evaluated the asymptotic property of $\widehat{\beta}$ established in Theorem 1, the second evaluated the performance of the proposed test method T, and the third evaluated the performance of T with larger rank$(\boldsymbol{\eta})$ or (p, q) values. Simulation results from fitting the rank-r GLM with $r = 3$ are reported with 500 replicates.

4.2. Evaluation of $\hat{\boldsymbol{\beta}}$

In this simulation study we set

$$\boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\eta}_{11} & \mathbf{0}_{2\times(q-1)} \\ \mathbf{0}_{(p-2)\times 1} & \mathbf{0}_{(p-2)\times(q-1)} \end{bmatrix} \quad \text{with} \quad \boldsymbol{\eta}_{11} = \frac{1}{\sqrt{2}} \cdot \mathbf{1}_2,$$

and set $\xi = (\xi_G^\top, \mathbf{0}_{p-5}^\top, \xi_E^\top, \mathbf{0}_{q-3}^\top)^\top$ with $\xi_G = (1/(10\sqrt{5}))\mathbf{1}_5$ and $\xi_E = (1/(10\sqrt{3}))$ $\mathbf{1}_3$ for the PSQI-simulation. Simulation results are reported in Table 1, which provides the means and standard deviations (SD) of $\widehat{\beta}$, the standard errors (SE) from the means of the diagonal elements of $\widehat{\boldsymbol{\Sigma}}$ corresponding to $(\gamma, \xi_G, \xi_E, \boldsymbol{\eta}_{11})$, and the averaged mean squared error $\text{AMSE}_\beta = E\|\widehat{\beta} - \beta\|^2/(1+m+pq)$. We also report the averaged mean squared errors $\text{AMSE}_\xi = E\|\widehat{\xi} - \xi\|^2/m$ and $\text{AMSE}_{\boldsymbol{\eta}} = E\|\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}\|^2/(pq)$ to summarize the performance of $\widehat{\beta}$ corresponding to the vector and matrix parameters $(\xi, \boldsymbol{\eta})$.

Biases of $\hat{\beta}$ arise under both PSQI and EEG settings, but they are relatively small in comparison with the corresponding SDs. This, together with the small values of AMSE_β, AMSE_ξ, and $\text{AMSE}_{\boldsymbol{\eta}}$, suggests that $\widehat{\beta}$ is a consistent estimator. Moreover, the similar values of SDs and SEs support the validity of $\widehat{\boldsymbol{\Sigma}}$. In this simulation, the true rank of $\boldsymbol{\eta}$ is 1, but the specified rank of model (2.1) is 3, indicating that over-specification of the rank parameter r does not affect the validity of Theorem 1. Recall that it only requires s_r parameters in model (2.1), instead of $1 + m + pq$ parameters in the conventional model (1.6). Consequently, the asymptotic results of Theorem 1 are more plausibly true even with limited sample size, as shown in Table 1. In summary, our simulation studies demonstrate

Table 1. The means (Mean) and standard deviations (SD) of $\hat{\beta}$, and the standard errors (SE) from the diagonal elements of $\widehat{\boldsymbol{\Sigma}}$. The last three rows give the means and standard deviations of AMSE_{β}, AMSE_{ξ}, and $\mathrm{AMSE}_{\boldsymbol{\eta}}$.

	PSQI				EEG			
	True	Mean	SD	SE	True	Mean	SD	SE
γ	10.000	10.003	0.063	0.059	0.000	−0.023	0.202	0.201
ξ_G	0.045	0.038	0.110	0.102				
	0.045	0.039	0.107	0.101				
	0.045	0.044	0.107	0.101				
	0.045	0.045	0.101	0.101				
	0.045	0.043	0.109	0.099				
ξ_E	0.058	0.053	0.089	0.081				
	0.058	0.060	0.085	0.080				
	0.058	0.057	0.086	0.079				
$\boldsymbol{\eta}_{11}$	0.707	0.635	0.119	0.109	0.707	0.604	0.220	0.202
	0.707	0.639	0.121	0.108	0.707	0.555	0.223	0.207
AMSE_{β}		0.008	0.002			0.030	0.010	
AMSE_{ξ}		0.010	0.005					
$\mathrm{AMSE}_{\boldsymbol{\eta}}$		0.007	0.002			0.030	0.010	

the validity and applicability of the proposed $(\widehat{\beta}, \widehat{\boldsymbol{\Sigma}})$.

4.3. Evaluation of T

This section evaluates the performance of T in testing (1.7). To make the comparisons more informative, we considered two types of $\boldsymbol{\eta}$ with different effect sizes c.

(S1) $\boldsymbol{\eta}$ has zero effects except for 2 randomly selected elements, with values given by cU with $U \in \mathbb{R}^2$ generated from the unit sphere.

(S2) $\boldsymbol{\eta}$ has zero effects except for a 5×2 sub-matrix, where values are given by cU with $U \in \mathbb{R}^{10}$ generated from the unit sphere.

These settings give $\mathrm{rank}(\boldsymbol{\eta}) = 2$. We set $\xi = (\xi_G^\top, \mathbf{0}_{p-5}^\top, \xi_E^\top, \mathbf{0}_{q-3}^\top)^\top$ with $\xi_G = (c/(10\sqrt{5}))\mathbf{1}_5$ and $\xi_E = (c/(10\sqrt{3}))\mathbf{1}_3$ for the PSQI-simulation. To implement T, we used 10-fold cross-validation to select $\lambda \in \{((p+q-r)r)/(\sqrt{n}\log(n)), ((p+q-r)r)/n, ((p+q-r)r)/n^{3/2}\}$ such that the condition $\lambda = o_p(n^{-1/2})$ is satisfied, where $(p+q-r)r$ is the number of parameters used to specify $\boldsymbol{\eta}$ with rank r. Besides T, we also implemented $(T_{\mathrm{wald}}, T_{\mathrm{max}}, T_{\mathrm{gesat}})$ for comparisons. Here $(T_{\mathrm{wald}}, T_{\mathrm{max}})$ were constructed from the rank-r GLM solely, while T_{gesat} was from the variance-component test without using the matrix structure of $\boldsymbol{\eta}$.

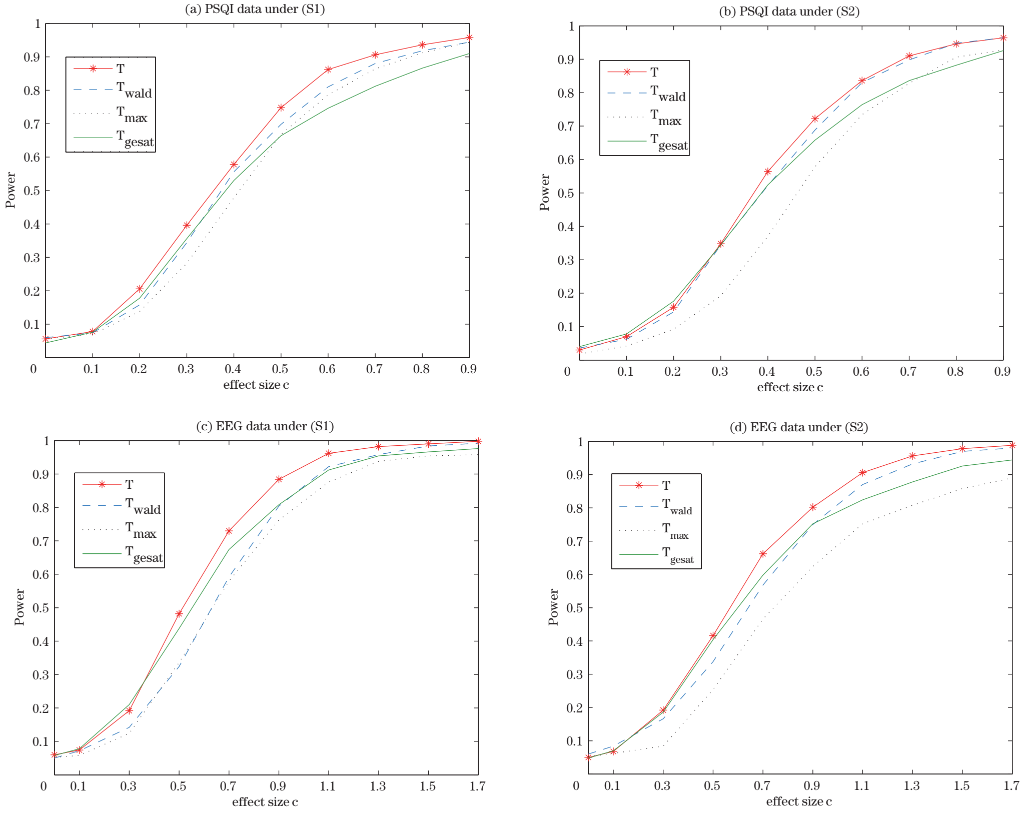

Figure 2. The power functions of T, T_{wald}, $T_{\max}$, and T_{gesat} under the sparse $\boldsymbol{\eta}$ (S1) and the low-rank $\boldsymbol{\eta}$ (S2) at different effect sizes c. (a) The case of PSQI under (S1). (b) The case of PSQI under (S2). (c) The case of EEG under (S1). (d) The case of EEG under (S2).

Comparing T with $(T_{\text{wald}}, T_{\max}, T_{\text{gesat}})$ can reveal possible drawbacks of different methods. For fair comparisons, we used the same re-sampling scheme to obtain the p-values of all methods, and their power functions at the significance level 0.05 over different effect sizes c are reported in Figur 2.

At $c = 0$, all methods control the type-I errors at 0.05, approximately, which suggests the validity of the re-sampling schemes (parametric bootstrap or permutation) to obtain valid p-values.

We next compared the detection powers of $(T_{\text{wald}}, T_{\max}, T_{\text{gesat}})$ at $c > 0$. For the case of small effect size c, T_{gesat} was detected to have better performance than $(T_{\text{wald}}, T_{\max})$. This is reasonable since T_{gesat} is the locally most powerful test. For the case of moderate effect size c, T_{wald} outperformed T_{gesat} in all settings, indicating gain from considering the low-rank structure of $\boldsymbol{\eta}$. As $T_{\max}$ is designed to test (1.7) with sparse $\boldsymbol{\eta}$, we detected a better performance of $T_{\max}$ under (S1) than under (S2). These observations reflect the fact that $(T_{\text{wald}}, T_{\max}, T_{\text{gesat}})$ have their own merits in testing (1.7), depending on the underlying characteristic of $\boldsymbol{\eta}$. Combing these methods, T showed the best performer in all situations. In particular, T had detection powers comparable to T_{gesat} for small c, and had higher detection powers than $(T_{\text{wald}}, T_{\max})$ when c is moderate to large. In summary, our simulation results demonstrate the applicability of T regardless of the form of $\boldsymbol{\eta}$.

4.4. Evaluation of T with larger rank($\boldsymbol{\eta}$) or (p, q)

We considered two extensions of the simulation studies to evaluate the performance of T.

In the first extension, we used the setting of (S2) in Section 4.3 except that the non-zero sub-matrix of $\boldsymbol{\eta}$ was of size 5×3 with $\text{rank}(\boldsymbol{\eta}) = 3$ or 5×4 with $\text{rank}(\boldsymbol{\eta}) = 4$. For the case of $\text{rank}(\boldsymbol{\eta}) = 4$, the fitted rank-3 GLM is not a correct model. Simulation results from both the PSQI and EEG data sets are placed in Figure 3. It can be seen that $T_{\max}$ and T_{gesat} have similar performances with the results of (S2) in Figure 2, while $T_{\max}$ is more sensitive to the increase of $\text{rank}(\boldsymbol{\eta})$. This is reasonable since $T_{\max}$ is specifically designed for the case of sparse $\boldsymbol{\eta}$. Despite the adverse influence of $T_{\max}$, the combined test statistic T is still found to be the best performer with the highest detection powers, even when the model rank is under-specified. This extended simulation demonstrates that T is able to adapt to various situations of $\boldsymbol{\eta}$.

In the second extension, we used the same setting with PSQI-simulation in Section 4.3, except that $G = (G_1^\top, G_2^\top)^\top$ and $E = (E_1^\top, E_2^\top)^\top$, where $G_1 \in \mathbb{R}^{15}$ and $E_1 \in \mathbb{R}^7$ were randomly generated from ROR and NR1D1 of the PSQI data, and $G_2 \in \mathbb{R}^{15}$ and $E_2 \in \mathbb{R}^8$ were randomly generated as multivariate normal with zero mean and identity covariance matrix. In this case, the number of parameters for $\boldsymbol{\eta}$ in the conventional model (1.6) is $30 \times 15 = 450$, with the sample size 400. Simulation results from fitting the rank-3 GLM are placed in Figure 4. Generally, a similar conclusion as Section 4.3 can be made for the good performance of T. For $c = 0$, the type-I errors of all methods are well controlled at 0.05. We next compared the detection powers at $c > 0$. Here T_{gesat} is found to have the lowest detection powers under both (S1)-(S2), indicating that the performance of T_{gesat} can be questionable when $pq > n$. On the other hand, T still outperforms $(T_{\text{wald}}, T_{\max}, T_{\text{gesat}})$ with the highest detection powers

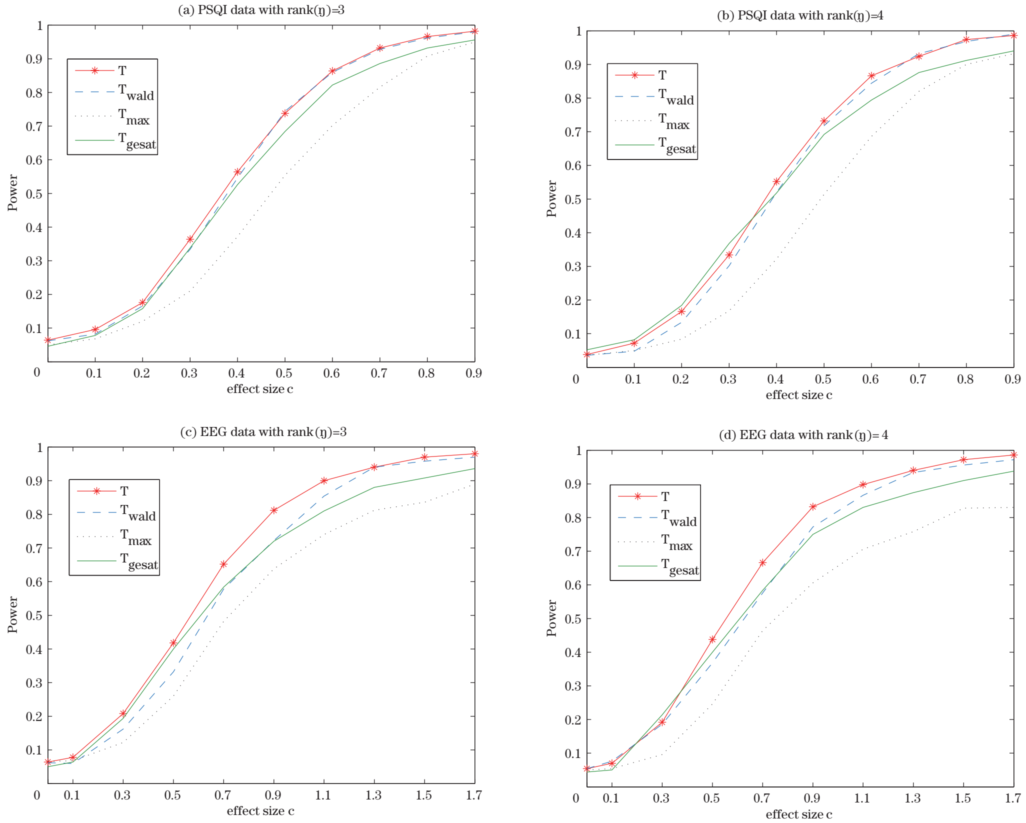

Figure 3. The power functions of T, T_{wald}, $T_{\max}$, and T_{gesat} under the low-rank $\boldsymbol{\eta}$ (S2) at different effect sizes c. (a) The case of PSQI with rank$(\boldsymbol{\eta}) = 3$. (b) The case of PSQI with rank$(\boldsymbol{\eta}) = 4$. (c) The case of EEG with rank$(\boldsymbol{\eta}) = 3$. (d) The case of EEG with rank$(\boldsymbol{\eta}) = 4$.

for all settings. This suggests that T is less affected by the problem of high-dimensionality, and is applicable even in the case of $pq > n$.

5. Data Analyses

5.1. The PSQI data

The Pittsburgh Sleep Quality Index (PSQI) data set (Lai et al. (2014)) includes 359 subjects with an average age of 41, range from 18 to 69. The participants consisted of 214 females and 145 males. For each subject, markers on 2 genes were collected: ROR with 19 markers (G) and NR1D1 with 4 markers (E). Also collected for each subject were the assessments of sleep quantity from Buysse, Reynolds and Monk. (1989), which consists of seven sores of PSQI: sleep

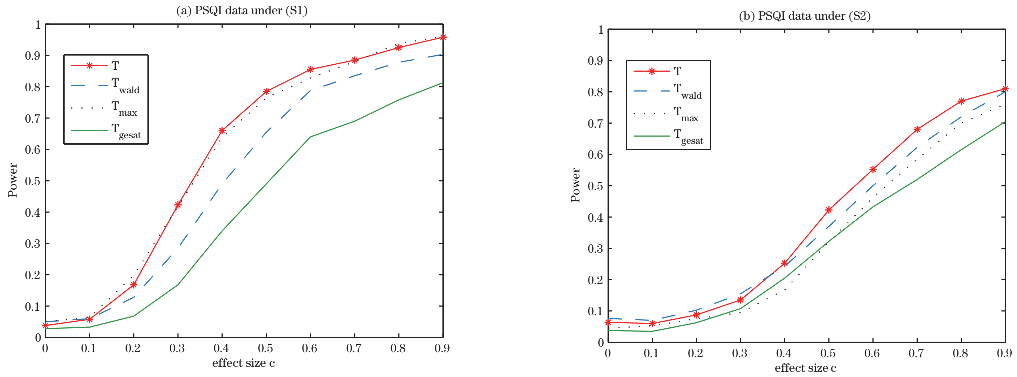

Figure 4. The power functions of T, T_{wald} , $T_{\max}$, and T_{gesat} under the sparse $\boldsymbol{\eta}$ (S1) and the low-rank $\boldsymbol{\eta}$ (S2) with $(p, q) = (30, 15)$ at different effect sizes c. (a) The case of PSQI under (S1). (b) The case of PSQI under (S2).

quality, sleep latency, sleep duration, habitual sleep efficiency, sleep disturbance, use of sleeping medication, and daytime dysfunction.

In our analysis, we considered the response Y to be the log-transformation of the sum of all PSQI scores (plus one to avoid taking logarithm of 0). Let $\boldsymbol{M} = GE^\top$ denote the 19×4 matrix of the interactions ROR×NR1D1, and let Z consist of ROR, NR1D1, Age, and Gender as possible confounding factors. Past study has found that ROR and NR1D1 cannot explain the variation of PSQI score well, and our interest focuses on whether the PSQI score is associated with ROR×NR1D1, after adjusting for the effects of Z. Fitting the normal model (2.1) with $r = 3$ gave the p-value (from the parametric bootstrap test) of T to be 0.008, while the p-value of T_{gesat} is 0.178. Thus, only the proposed low rank-based test method declares that ROR×NR1D1 is influential to the PSQI score. We further found that the p-value of $T_{\max}$ was smaller than 10^{-3}, while the p-value of T_{wald} was 0.145. This indicates that a sparse ROR×NR1D1 effect can be expected. The estimated effect sizes of ROR×NR1D1 are reported in Table 2, where the significant effects identified by individual tests (3.4) at the family-wise error rate 0.05 are marked in bold. In particular, the interactions rs11144047×rs12941497 and rs1327836×rs12941497 are identified, which confirms the spare effect of ROR×NR1D1. We note that rs11144047, rs1327836, and the interaction rs1327836×rs12941497 have been found to associate with bipolar disorder (Lai et al. (2015)), and bipolar disorder is known to associate with the PSQI score. Our analysis results not only support these findings in the literature, but also suggest that rs11144047 × rs12941497 is an influential

Table 2. The estimated effect sizes of ROR × NR1D1. The significant covariates identified by the individual tests (3.4) at the family-wise error rate 0.05 are marked in bold.

		NR1D1			
		rs2314339	rs2071427	rs2269457	rs12941497
RORA	rs809736	−0.007	−0.026	−0.019	0.047
	rs4774388	0.001	0.014	0.016	−0.034
RORB	rs10491929	0.002	0.004	0.002	−0.005
	rs17611535	0.001	0.015	0.017	−0.036
	rs10217594	−0.004	0.007	0.016	−0.030
	rs7037043	−0.008	−0.023	−0.014	0.036
	rs2025882	−0.001	−0.004	−0.003	0.006
	rs7022435	−0.001	−0.005	−0.004	0.010
	rs3750420	0.001	0.001	0.000	−0.001
	rs1013078	−0.004	−0.027	−0.026	0.059
	rs2273975	−0.004	0.012	0.023	−0.044
	rs3903529	0.003	−0.004	−0.011	0.020
	rs11144041	0.004	0.006	0.000	−0.003
	rs7021908	−0.002	0.000	0.004	−0.006
	rs7865407	−0.003	−0.012	−0.009	0.022
	rs11144047	−0.007	−0.040	−0.036	**0.083**
	rs1327836	0.005	0.054	0.058	**−0.125**
	rs11144064	−0.001	−0.005	−0.004	0.010
	rs4098048	0.003	0.000	−0.005	0.008

interaction to bipolar disorder that is missed by the conventional analysis.

5.2. The EEG data

The EEG data can be obtained from the *UCI Machine Learning Repository*. There were 77 subjects in alcoholic group ($Y = 1$) and 45 subjects in control group ($Y = 0$). The voltage values of 64 channels at 256 time points were also collected for multiple trials, and we followed Hung and Wang (2013) in using the means over all trials to summarize the data. This gives, for each of 122 subjects, 64×256 measurements. The EEG data has been analyzed by Hung and Wang (2013) and Zhou and Li (2014), and they report the effects of the EEG signals. They also report the classification accuracies from their models to support a close connection between the EEG signals and alcoholic status. Hung and Wang (2013) and Zhou and Li (2014) do not test whether the EEG signals are truly associated with the alcoholic status via the overall hypothesis (1.7), nor did they identify the significant covariates via the individual hypothesis (3.4). The aim of our analysis is to answer these questions via the proposed test method.

In our analysis, we pre-processed the original covariates by averaging the

Table 3. The significant covariates of the EEG signals identified by the individual tests (3.4) at the family-wise error rate 0.05.

Region	Channel	Time	Effect Size
Parietal	20 (CP6)	11	-0.039
	20 (CP6)	18	0.033
	22 (CP2)	11	-0.027
	50 (CP4)	11	-0.027
	24 (P4)	11	-0.039
	26 (P8)	11	-0.036
	27 (P7)	11	-0.034
	27 (P7)	12	-0.022
	51 (P5)	11	-0.029
	52 (P6)	11	-0.032
Occipital	28 (PO2)	11	-0.024
	55 (PO7)	11	-0.030
	56 (PO8)	11	-0.029
	31 (O1)	11	-0.026
Central	42 (C6)	11	-0.024
Temporal	46 (TP8)	11	-0.027

voltage values for every eight time points. They gave, for each subject, a 32×64 matrix-covariate $\boldsymbol{M}$ (after component-wisely standardization such that $\boldsymbol{M}(j,k)$ has mean 0 and variance 1), where $\boldsymbol{M}(j,k)$ represents the average voltage value of the k-th channel over the time period $[8(j-1)+1, 8j]$. By fitting the logistic model (2.1) with $r = 1$ on $(Y, \boldsymbol{M})$, we found the p-value of T (from permutation test) to be smaller than 10^{-3}, indicating a strong association between the EEG signals and the alcoholic status. The significant covariates in $\boldsymbol{M}$ identified by the individual tests (3.4) at the family-wise error rate 0.05 are listed in Table 3. One can see that the regions Parietal and Occipital, which control the functions of sensation and vision, respectively, play important roles as to alcoholic status. Moreover, most of the identified channels have significant influence at the 11-th time period, indicating an early reaction of the brain to the stimuli. Our analysis not only provides evidence to support the conclusions of Hung and Wang (2013) and Zhou and Li (2014), but also suggests possible brain regions and reaction time periods for further investigations. Our result was obtained without imposing any identifiability constraint.

6. Conclusion

We propose novel methods to test the significance of matrix-covariate, and to identify the significant covariates of matrix-covariate. The rationale of our proposal is to utilize the matrix structure of $\boldsymbol{\eta}$ to achieve a parsimonious parameterization and, hence, a higher detection power than conventional methods. Our proposal differs from existing methods as no identifiability constraint is required when making inference about $\boldsymbol{\eta}$, and our method can provide estimates of $\boldsymbol{\eta}$ at the same time.

We discuss some extensions and limitations of this work.

1. We have developed our method under the normal model and logistic model, due to the data structures of our motivating examples (binary Y in the EEG data and continuous Y in the PSQI data). Our method can be extended to Poisson regression for count Y: $Y|(Z, \boldsymbol{M}) \sim \text{Poisson}(E(Y|Z, \boldsymbol{M}))$ with the link function $g(u) = \ln u$. In this case, Theorem 1 is still valid with $\boldsymbol{V}_0 = E(\exp(\beta(\theta_0)^\top X_i)X_i X_i^\top)$, and the testing procedures for (1.7) and (3.4) can be directly applied.

2. The matrix-covariate discussed in this work is an order-two tensor, which motivates the matrix structure of the parameter $\boldsymbol{\eta}$. Tensor-covariate can now be found in many applications. For example, p covariates measured at q time points under k environments corresponds to an order-three tensor (with dimension $p \times q \times k$) for each subject. Statistical inference procedures for GLM with tensor-covariate have been developed in Zhou, Li and Zhu (2013), with the focus on the estimation of the effect size of the corresponding tensor parameter. When the research aim is to test the existence of an association between the response and tensor-covariate, our low rank-based test methods can be extended, provided that a version of Theorem 1 is developed for tensor-covariate. Another issue is the "low-rank parameterization" for the tensor parameter. For the case of order-two tensor, the representation is unique, but is not for higher order tensors. It is of interest to study the effects of different low-rank parameterizations to detection power.

3. In Theorem 1, we have established the asymptotic property of $\widehat{\beta}$ with fixed (p, q) and diverging n. A similar result can be problematic when both (p, q) are larger than n. It is thus of interest to extend Theorem 1 to the case of diverging (n, p, q). This issue is beyond the scope of this work.

Supplementary Materials

The online supplementary material contains the proof of Theorem 1.

Acknowledgments

The authors thank Dr. Po-Hsiu Kuo from National Taiwan University who provided the PSQI data for our analysis.

References

Buysse, D. J., Reynolds, C. F. and Monk, T. H. (1989). The Pittsburgh sleep quality index: A new instrument for psychiatric practice and research. *Psychiatry Research* **28**, 193–213.

Bůžková, P., Lumely, T. and Rice., K. (2011). Permutation and parametric bootstrap tests for gene-gene and gene-environment interactions. *Annals of Human Genetics* **75**, 36–45.

Goeman, J. J., van de Geer, S. A. and van Houwelingen, H. C. (2006). Testing against a high dimensional alternative. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **68** 477–493.

Henderson, H. V. and Searle, S. R. (1979). Vec and vech operators for matrices, with some uses in Jacobians and multivariate statistics. *Canadian Journal of Statistics* **7**, 65–81.

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

Hung, H., Lin, Y. T., Wang, C. C., Chen, P., Huang, S. Y. and Tzeng, J. Y. (2016). Detection of gene-gene interactions using multistage sparse and low-rank regression. *Biometrics* **72**, 85–94.

Lai, Y. C., Huang, M. C., Chen, H. C., Lu, M. K., Chiu, Y. H., Shen, W. W., Lu, R. B. and Kuo, P. H. (2014). Familiality and clinical outcomes of sleep disturbance in major depressive and bipolar disorders. *Journal of Psychosomatic Research* **76**, 61–67.

Lai, Y. C., Kao, C. F., Lu, M. L., Chen, H. C., Chen, P. Y., Chen, C. H., Shen, W. W., Wu, J. Y., Lu, R. B. and Kuo, P. H. (2015). Investigation of associations between NR1D1, RORA and RORB genes and bipolar disorder. *PloS One* **10**, e0121245.

Le Cessie, S. and Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Applied Statistics*, **41**, 191–201.

Lin, X. (1997). Variance component testing in generalised linear models with random effects. *Biometrika* **84**, 309–326.

Lin, X., Lee, S., Christiani, D. C. and Lin, X. (2013). Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics* **14**, 667–681.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall/CRC: London.

Shapiro, A. (1986). Asymptotic theory of overparameterized structural models. *Journal of American Statistical Association* **81**, 142–149.

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.

Zhou, H. and Li, L. (2014). Regularized matrix regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 463–483.

Rm532, No.17, Xuzhou Rd, Taipei 100, Taiwan.
E-mail: hhung@ntu.edu.tw
Institute of Epidemiology and Preventive Medicine, National Taiwan University, Taiwan.
E-mail: misa7944@hotmail.com

(Received September 2016; accepted November 2017)