Matrix Variate Logistic Regression Analysis

Hung Hung[a*]and Chen-Chien Wang[b]

[a]Institute of Epidemiology and Preventive Medicine, National Taiwan University

[b]Institute of Statistical Science, Academia Sinica

Abstract

Logistic regression has been widely applied in the field of biostatistics for a long time. It aims to model the conditional success probability of an event of interest as the logit function of a linear combination of covariates, for the sake of further interpretation of covariates and prediction of new observation. In some applications, however, covariates of interest have a natural structure, such as being a matrix, at the time of being collected. The rows and columns of the covariate matrix would have different meanings, and they must contain useful information regarding the response. If we simply stack X as a vector and fit the conventional logistic regression model, we may discard relevant information and may also suffer the problem of inefficiency in estimating parameters. Motivated from this reason, we propose in this paper the matrix variate logistic (MV-logistic) regression model. The most important feature of our model is that it retains the inherent structure of the covariate matrix. Another advantage is the parsimony of parameters needed. These features lead to a good performance of MV-logistic regression in many situations. Simulation studies and a data example, the EEG data, demonstrate the usefulness of the proposed method.

Keywords and phrases: Asymptotic theory, Logistic regression, Matrix variate covariates, Regularization, Tensor object.

[*] Corresponding author. *Email address*: hhung@ntu.edu.tw

1 Introduction

Logistic regression has been widely applied in the field of biostatistics for a long time. It aims to model the logit transformation of conditional success probability of an event of interest as a linear combination of covariates. After obtaining the data, maximum likelihood estimates (MLE) can then be used to construct subsequent statistical analysis such as prediction and interpretation. We refer the readers to MuCulloch, Searle, and Neuhaus (2008) for details and further extensions of logistic regression.

In some applications, covariates of interest have a natural matrix structure at the time of being collected. For example, research interest focuses on predicting the disease status of a future patient, based on his/her image of certain organ. In this situation, we actually obtain the data set of the form $\{(Y_i, X_i)\}_{i=1}^n$ which are random copies of (Y, X), where Y is a binary random variable with value 1 indicating diseased and 0 otherwise, and X is a $p \times q$ covariate matrix representing the corresponding image of the i-th patient. As another example, the EEG data set which will be analyzed in this article later, concerns the relationship between the genetic predisposition and alcoholism. In this study, $Y = 1$ and $Y = 0$ represent alcoholic group and control group, respectively, and the covariate X is a 256×64 matrix, where its (i, j)-th element $X_{(i,j)}$ is the measurement of voltage value at channel of electrode i and time point j. To fit logistic regression model, one usually stacks X column by column as a pq-vector, say $\mathsf{vec}(X)$, and subsequent statistical analysis follows without any difference. However, it does not seem to be wise to ignore the inherent matrix structure of X. More specifically, the rows and columns of X must contain information regarding Y and we should incorporate this information into statistical analysis. Another drawback of stacking X as a long vector is that the number of parameters usually becomes extremely large in comparison with the sample size. For example, we will have $1 + 64 \times 256 = 16385$ parameters in the EEG data when we fit logistic regression model by using $\mathsf{vec}(X)$ directly. As high dimensionality makes statistical inference procedure instable, it becomes urgent to seek a more efficient method in dealing with matrix covariate.

Motivated from this observation, when the covariates of interest have a natural matrix structure, in contrast to the conventional logistic regression model, we propose the matrix variate logistic (MV-logistic) regression model

$$P(Y=1|X) = \frac{\exp(\gamma + \alpha^T X\beta)}{1+\exp(\gamma + \alpha^T X\beta)}, \tag{1}$$

where $\alpha = (\alpha_1, \cdots, \alpha_p)^T$ and $\beta = (\beta_1, \cdots, \beta_q)^T$ are the coefficients for rows and columns of X, and γ is the intercept term, which preserves the matrix structure of X. Obviously, those parameters in model (1) possess their own meanings of interpretations. In the EEG data, for instance, α is the effect of different time points, and β is the effect of different channels. By fitting MV-logistic regression model, we are able to extract those column (row) information contained in X, which are ignored in the conventional logistic regression model. Note that under model (1), (α, β) can only be identified up to scale, since $(c^{-1}\alpha, c\beta)$ will result in the same model for any constant c. For the sake of identifiability, without loss of generality, we assume $\alpha_1 = 1$ in the rest of discussion. Denote the rest of parameters in α to be α^*, i.e., $\alpha = (1, \alpha^{*T})^T$, and $\theta = (\gamma, \alpha^*, \beta)$ be the parameters of interest. We thus have $p+q$ free parameters contained in θ. One can see that a merit of model (1) is the parsimony of parameters used. If we fit the conventional logistic regression model for $(Y, \mathsf{vec}(X))$, we require $pq+1$ parameters while it is only $p+q$ in MV-logistic regression model. Thus, when model (1) is correctly specified, we can expect an efficiency gain in estimating parameters.

Adoption of model (1) is equivalent to modeling the covariate-specific odds ratio R_{ij} of $X_{(i,j)}$ (with increment 1) while keeping the rest covariates fixed as

$$\ln(R_{ij}) = \alpha_i \beta_j \Leftrightarrow R_{ij} = \{\exp(\beta_j)\}^{\alpha_i}. \tag{2}$$

Take EEG data to exemplify again, relation (2) implies that each channel has its own baseline odds ratio $\exp(\beta_j)$. Depending on the time of being measured, the odds ratio is further modified by taking a power of α_i. A value of α_i greater than one then indicates the i-th time point has larger influence in comparison with the baseline time point, i.e., the first time point (since we set $\alpha_1 = 1$).

Remark 1.1. *Although in the population level there is no difference for which α_i is set to be one, it is a crucial step in practical implementation. Intuitively, if the true value of α_i is near zero, setting $\alpha_i = 1$ will magnify a negligible error to an enormous one, which* *leads to unstable, unreliable, or even divergent numerical results. Here we provide an easy guidance to select the baseline effect. Let ρ_{ij} be the sample correlation coefficient between $X_{(i,j)}$ and Y. We then set $\alpha_i = 1$ if $i = \arg\max_k\{\sum_{j=1}^q |\rho_{kj}| : k = 1, \cdots, p\}$. The intuition is that we choose the one as the baseline which is the most likely to be correlated with the response Y. We find in our numerical studies this simple approach makes the estimation procedure stable and converges quickly.*

Remark 1.2. *Standardization of covariates before entering statistical analysis is usually applied in practice. In logistic regression model, this preprocessing will not affect the final result except the change of scale of parameter estimates, since the standard deviation of each covariate can be absorbed into the corresponding parameter. As to the case of MV-logistic regression model, however, we have pq covariates but only $p+q$ free parameters and, hence, it is generally impossible to absorb those standard deviations into less parameters. In summary, standardization of covariates will result in a different model. We will investigate this issue through EEG data in Section 4.*

2 Statistical Inference Procedure

Some notations are defined here for the ease of reference. Remember the parameters of interest is $\theta = (\gamma, \alpha^*, \beta)$. Denote the covariate-specific probability $P(Y_i = 1|X_i)$ by

$$\pi_i = \pi(\theta|X_i) = \frac{\exp(\gamma + \alpha^T X_i \beta)}{1 + \exp(\gamma + \alpha^T X_i \beta)}. \tag{3}$$

Let $\mathbf{Y} = (Y_1, \cdots, Y_n)^T$ and $\mathbf{X}(\theta) = [\ \mathbf{X}_1(\theta), \cdots, \mathbf{X}_n(\theta)\ ]^T$ be an $n \times pq$ matrix, where $\mathbf{X}_i(\theta) = (1, \beta^T X_i^T C_p, \alpha^T X_i)^T$ with $C_p = \partial\alpha/\partial\alpha^* = [\mathbf{0}_{(p-1)\times 1}, \mathbf{I}_{p-1}]^T$ and $\otimes$ being the Kronecker product. Note that the form of C_p will be changed according to which $\alpha_i = 1$ is assumed. Let also $\mathbf{\Pi}(\theta) = (\pi_1, \cdots, \pi_n)^T$ and $\boldsymbol{V}(\theta) = \mathrm{diag}(v_1, \cdots, v_n)$, where $v_i = \pi_i(1-\pi_i)$ be the conditional variance of Y_i given X_i. For any function g, we denote $g^{(i)}$ to be the i-th derivative of g with respect to its argument.

2.1 Estimation

The estimation of θ mainly relies on maximum likelihood method. Given the data set $\{(Y_i, X_i)\}_{i=1}^n$, the log-likelihood function of θ is derived to be

$$\begin{aligned} \ell(\theta) &= \sum_i Y_i \ln(\pi_i) + (1 - Y_i)\ln(1 - \pi_i) \\ &= \sum_i Y_i(\gamma + \alpha^T X_i \beta) - \ln(1 + \exp(\gamma + \alpha^T X_i \beta)), \end{aligned} \tag{4}$$

and θ can be estimated by the maximizer of $\ell(\theta)$.

In modern research of biostatistics, however, an important issue is the number of covariates could be large in comparison with the sample size. This will make the estimation procedure unstable, or traditional methodologies may even fail. As mentioned in the previous section, the number of free parameters in model (1) is $p + q$, while it is $pq + 1$ in the conventional logistic regression model. The MV-logistic regression then suffers less severity from the problem of high dimensionality. However, this can not entirely avoid the problem of instability. As we have seen in the EEG data set the covariate X is a 256×64 matrix (i.e., 320 parameters in MV-logistic regression), while there are only 122 observations (note also that there are $1 + 256 \times 64 = 16385$ parameters in conventional logistic regression). To overcome the difficulty of high dimensionality, Le Cessie and Van Houwelingen (1992) proposed the penalized logistic regression method. This motivates us to further consider the penalized MV-logistic regression. In particular, we attempt to estimate θ by maximizing the penalized log-likelihood function

$$\ell_\lambda(\theta) \;=\; \ell(\theta) - \lambda J(\theta), \tag{5}$$

where $J(\cdot) \geq 0$ is a twice continuously differentiable penalty function of θ, and $\lambda \geq 0$ is the regularization parameter that controls the amount of shrinkage. With a specific penalty function $J(\cdot)$, we then propose to estimate θ by

$$\hat{\theta}_\lambda = \arg\max_\theta \ell_\lambda(\theta). \tag{6}$$

There are many choices of $J(\cdot)$ depending on different research purposes, wherein $J(\theta) = \|\theta\|^2$ and $J(\theta) = \|\alpha^*\|^2 + \|\beta\|^2$ are the most widely applied ones. The main difference between these two choices is whether to put the penalty on the intercept term γ or not. The regularization parameter λ should also be determined in practice. In this article, we consider to select λ through maximizing the cross-validated classification accuracy. There are other selection criteria and can be found in Le Cessie and Van Houwelingen (1992).

To obtain $\hat{\theta}_\lambda$, the iterative Newton method can be applied. The gradient of $\ell_\lambda(\theta)$ (with respect to θ) is calculated to be

$$\ell_\lambda^{(1)}(\theta) = \ell^{(1)}(\theta) - \lambda J^{(1)}(\theta), \tag{7}$$

where $\ell^{(1)}(\theta) = \mathbf{X}(\theta)^T\{\mathbf{Y} - \mathbf{\Pi}(\theta)\}$. Moreover, the Hessian matrix of $\ell_\lambda(\theta)$ is derived to be

$$\ell_\lambda^{(2)}(\theta) = -H_\lambda(\theta) + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & \sum_{i=1}^n C_p^T X_i(Y_i - \pi_i) \\ 0 & \sum_{i=1}^n X_i^T C_p(Y_i - \pi_i) & 0 \end{bmatrix}, \tag{8}$$

where

$$H_\lambda(\theta) = \mathbf{X}(\theta)^T \boldsymbol{V}(\theta)\mathbf{X}(\theta) + \lambda J^{(2)}(\theta). \tag{9}$$

As suggested by Green (1984), we will ignore the last term of (8) since its expectation is zero. Finally, $\hat{\theta}_\lambda$ can be obtained through iterating

$$\theta_{(k+1)} = \theta_{(k)} + \left\{H_\lambda(\theta_{(k)})\right\}^{-1} \ell_\lambda^{(1)}(\theta_{(k)}),\ k = 0, 1, 2, \cdots, \tag{10}$$

until there is no significant difference between $\theta_{(k+1)}$ and $\theta_{(k)}$, and report $\theta_{(k+1)}$ as the final estimate. A zero initial $\theta_{(0)}$ is suggested and performs well in our numerical studies.

Remark 2.1. *We suggest to use a larger penalty (e.g., 10λ) in the first iteration to obtain an initial value for the subsequent iterations. It is found in our numerical studies that this procedure fastens the convergence of iterative Newton method.*

2.2 Asymptotic Properties

Consistency and asymptotic normality of $\hat{\theta}_\lambda$ can be derived through usual arguments of MLE. These asymptotic properties will be helpful in making statistical inference about θ and its related quantities such as $\pi(\theta|X_i)$. The result is summarized in the following theorem.

Theorem 2.2. *Assume the validity of model (1) and the regularity conditions of likelihood function. Assume also the information matrix* $I(\theta) = E[\mathbf{X}_i(\theta)v_i\mathbf{X}_i(\theta)^T]$ *is nonsingular. Then, for any fixed* λ*, we have*

$$\sqrt{n}\left(\hat{\theta}_\lambda - \theta\right) \xrightarrow{d} N\left(0, \Sigma(\theta)\right)$$

as n *goes to infinity, where* $\Sigma(\theta) = \{I(\theta)\}^{-1}$*.*

Proof. By taking Taylor's expansion of $\ell_\lambda^{(1)}(\hat{\theta}_\lambda)$ around the true θ, we have

$$0 = \frac{1}{n}\ell_\lambda^{(1)}(\hat{\theta}_\lambda) = \frac{\sqrt{n}}{n}\ell_\lambda^{(1)}(\theta) + \frac{1}{n}\ell_\lambda^{(2)}(\theta)\{\sqrt{n}(\hat{\theta}_\lambda - \theta)\} + o_p(\sqrt{n}\|\hat{\theta}_\lambda - \theta\|). \tag{11}$$

First observe that by central limit theorem and Slutsky's theorem we have

$$\begin{aligned} \frac{\sqrt{n}}{n}\ell_\lambda^{(1)}(\theta) &= \frac{\sqrt{n}}{n}\ell^{(1)}(\theta) + \frac{\sqrt{n}}{n}\lambda J^{(1)}(\theta) \\ &\xrightarrow{d} N(0, I(\theta)). \end{aligned} \tag{12}$$

Since $E[\frac{1}{n}\sum_{i=1}^n X_i^T C_p(Y_i - \pi_i)] = 0$, by (8) and the law of large number we have

$$-\frac{1}{n}\ell_\lambda^{(2)}(\theta) = \frac{1}{n}H_\lambda(\theta) + o_p(1) \xrightarrow{p} I(\theta). \tag{13}$$

The boundedness of $\|\frac{-1}{n}\ell_\lambda^{(2)}(\theta)\|$ from (13) and (11)-(12) then imply $\|\hat{\theta}_\lambda - \theta\| = O_p(n^{-1/2})$. The proof is now completed by using (11)-(13) and Slutsky's theorem again. □

From Theorem 2.2, $\hat{\theta}_\lambda$ is proven to be a consistent estimator of θ. It also enables us to construct confidence region of θ, provided we have an estimate of $\Sigma(\theta)$. This can be done by the usual empirical estimator with the unknown θ being replaced by $\hat{\theta}_\lambda$. In particular, define

$$\hat{\Sigma}(\theta) = \left(\frac{1}{n}H_\lambda(\theta)\right)^{-1}\left(\frac{1}{n}\mathbf{X}(\theta)^T\boldsymbol{V}(\theta)\mathbf{X}(\theta)\right)\left(\frac{1}{n}H_\lambda(\theta)\right)^{-1}. \tag{14}$$

We propose to estimate the asymptotic covariance matrix $\Sigma(\theta)$ by $\hat{\Sigma}(\hat{\theta}_\lambda)$. For any $0 < a < 1$, an approximated $100(1-a)\%$ confidence interval of θ_i, the i-th element of θ, is then constructed to be

$$\left(\hat{\theta}_{\lambda,i} - z_{\frac{a}{2}}\frac{[\hat{\Sigma}(\hat{\theta}_\lambda)]_i}{\sqrt{n}},\ \hat{\theta}_{\lambda,i} + z_{\frac{a}{2}}\frac{[\hat{\Sigma}(\hat{\theta}_\lambda)]_i}{\sqrt{n}}\right), \tag{15}$$

where $z_{\frac{a}{2}}$ is the $1-\frac{a}{2}$ quantile of standard normal and $[\hat{\Sigma}(\hat{\theta}_\lambda)]_i$ denotes the i-th diagonal element of $\hat{\Sigma}(\hat{\theta}_\lambda)$. We would also be interested in making statistical inference about the success probability $\pi(\theta|x)$ for any given $p \times q$ matrix x. This quantity can be straightforwardly estimated by $\pi(\hat{\theta}_\lambda|x)$. A discrimination rule is then to classify a subject with matrix covariate x to the "diseased" group if $\pi(\hat{\theta}_\lambda|x) > 0.5$ and the "non-diseased" group otherwise. Moreover, by applying delta method and the result of Theorem 2.2, we deduce that

$$\sqrt{n}\left\{\ln\left(\frac{\pi(\hat{\theta}_\lambda|x)}{1-\pi(\hat{\theta}_\lambda|x)}\right)-\ln\left(\frac{\pi(\theta|x)}{1-\pi(\theta|x)}\right)\right\}\xrightarrow{d} N\left(0,\sigma_\pi^2(\theta|x)\right), \tag{16}$$

where

$$\sigma_\pi^2(\theta|x)=\boldsymbol{x}(\theta)^T\Sigma(\theta)\boldsymbol{x}(\theta) \tag{17}$$

and $\boldsymbol{x}(\theta)=(1,\beta^T x^T C_p,\alpha^T x)^T$. The asymptotic variance $\sigma_\pi^2(\theta|x)$ can be estimated by $\hat{\sigma}_\pi^2(\hat{\theta}_\lambda|x)$, where

$$\hat{\sigma}_\pi^2(\theta|x)=\boldsymbol{x}(\theta)^T\hat{\Sigma}(\theta)\boldsymbol{x}(\theta). \tag{18}$$

An approximated $100(1-a)\%$ confidence interval of $\pi(\theta|x)$ is then constructed to be

$$\left(\frac{\exp(\hat{\gamma}+\hat{\alpha}^T x\hat{\beta}-z_{\frac{a}{2}}\frac{\hat{\sigma}_\pi(\hat{\theta}_\lambda|x)}{\sqrt{n}})}{1+\exp(\hat{\gamma}+\hat{\alpha}^T x\hat{\beta}-z_{\frac{a}{2}}\frac{\hat{\sigma}_\pi(\hat{\theta}_\lambda|x)}{\sqrt{n}})},\frac{\exp(\hat{\gamma}+\hat{\alpha}^T x\hat{\beta}+z_{\frac{a}{2}}\frac{\hat{\sigma}_\pi(\hat{\theta}_\lambda|x)}{\sqrt{n}})}{1+\exp(\hat{\gamma}+\hat{\alpha}^T x\hat{\beta}+z_{\frac{a}{2}}\frac{\hat{\sigma}_\pi(\hat{\theta}_\lambda|x)}{\sqrt{n}})}\right), \tag{19}$$

which is guaranteed to be a subinterval of $[0,1]$.

3 Simulation Studies

In this section the proposed method is evaluated through simulation studies under two different settings. Let the parameters in model (1) be set as $\alpha=(1,0.5,-0.5\mathbf{1}_{p-2}^T)^T$, $\beta=(1,0.5,1,-1,-\mathbf{1}_{q-4})^T$, and $\gamma=1$, where $\mathbf{1}_a$ is the $a\times 1$ vector of ones. A common value of $\lambda=0.5$ is used throughout the simulations. Simulation results are reported with $(p,q)=(10,12)$ and 500 replicates.

3.1 Simulation with tensor structure

In the first simulation study, we evaluate the proposed method under the validity of model (1). We first generate X such that $\mathsf{vec}(X)$ follows a pq-variate normal distribution with mean zero and covariance matrix I_{pq}. Conditional on X, Y is generated from model (1) with the specified θ. We then obtain $\hat{\theta}_\lambda$ and related quantities by the proposed method. The averages of $\hat{\theta}_\lambda$ and standard errors from (14) are provided in Table 1 for $n = 150, 300$. It can be seen that the biases of $\hat{\theta}_\lambda$ is small in comparison with standard deviations, and decrease as the sample size increases. The standard deviations are also well estimated by our empirical estimators, the diagonal elements of $\hat{\Sigma}(\hat{\theta}_\lambda)$. These observations also validates our Theorem 2.2 and the proposed estimators $\hat{\Sigma}(\hat{\theta}_\lambda)$.

To further demonstrate the superiority of matrix-variate logistic regression, we also fit the conventional penalized logistic regression (Le Cessie and Van Houwelingen, 1992) to obtain estimate of (γ, η) with $\eta = (\beta \otimes \alpha)$ as if we ignore the matrix structure of X, and denote it by $(\tilde{\gamma}, \tilde{\eta})$. It is then compared with $(\hat{\gamma}, \hat{\eta})$, the estimates obtained from MV-logistic regression. In each simulation run, we generate another independent data set under the same setting with the same sample size n, and calculate the classification accuracies by using $\hat{\eta}^T\mathsf{vec}(X)$ and $\tilde{\eta}^T\mathsf{vec}(X)$ as predictors. Table 2 provides the averages of similarities $\hat{\eta}^T\tilde{\eta}/(\|\hat{\eta}\|\,\|\tilde{\eta}\|)$ and the classification accuracies. It is obviously detected that the MV-logistic regression produces larger similarities and classification accuracies.

3.2 Simulation without tensor structure

An important issue in practice is the validity of MV-logistic regression. When the covariate of interest X has a natural matrix structure, it is reasonable to apply model (1) for subsequent analysis. The parameters involved in the model also have the corresponding physical meanings. As demonstrated in the previous numerical studies, MV-logistic regression definitely outperforms the conventional logistic regression under model (1). Note that one advantage of MV-logistic regression is the parsimony of parameter used, since the columns (rows) of X share the same parameter α (β). In practice, however, this assumption could be violated even when X has a natural matrix structure. For example, one of the columns (rows) of X could possess a different effect. Here we will evaluate the performance of MV-logistic regression through simulations when the underlying distribution model departures from model (1).

We first generate $\mathsf{vec}(X)$ from a 120-variate standard normal distribution. In each simulation run, we also independently generate a random vector δ from a 120-variate normal distribution with mean zero and covariance matrix $\sigma^2 I_{120}$. Then, conditional on X, Y is generated from a Bernoulli distribution with

$$P(Y=1|X)=\frac{\exp(\gamma+\eta^{*T}\mathsf{vec}(X))}{1+\exp(\gamma+\eta^{*T}\mathsf{vec}(X))}, \tag{20}$$

where $\eta^* = \eta + \delta$ and $\eta = (\beta \otimes \alpha)$. It is obvious that with an extra term δ, η^* does not possess a tensor structure and thus model (1) is incorrect. The amount of σ controls the magnitude of departure of (20) from (1). We then fit conventional logistic regression to obtain $\tilde{\eta}^*$. We also fit model (1) to obtain $\hat{\eta}^*$. Of course in this situation evaluations of $\hat{\theta}_\lambda$ are meaningless. Instead, we will compare the classification accuracy. That is, both $\tilde{\eta}^*$ and $\hat{\eta}^*$ are applied to another independently generated data set and the classification accuracies are calculated. The averages of classification accuracies and standard deviations are placed in Table 3 for $\sigma = 0.1, 0.3, 0.5$ and $n = 150$.

It is detected that MV-logistic regression still outperforms conventional logistic regression with moderate size of σ. The classification accuracy of MV-logistic regression decreases as σ increases, and is roughly the same with the one of conventional logistic regression when $\sigma = 0.5$. Our simulation results indicate that MV-logistic regression has certain robustness against the violation of model specification, and can be treated as a good "working" model even if the underlying distribution do not obey model (1).

4 The EEG Data

The EEG data is analyzed in this section to demonstrate the usefulness of MV-logistic regression. The data consists of 122 subjects, wherein 77 of them belong to the group of alcoholism ($Y_i = 1$) and the rest 45 subjects are in the control group ($Y_i = 0$). Each subject completed a total of 120 trials under three different conditions (single stimulus, two matched stimuli, and two unmatched stimuli). In each trial, measurements from 64 electrodes placed on subject's scalp at 256 time points are collected, which resulted in a 256×64 covariate matrix. It is interested to distinguish two types of subjects based on the collected matrix covariates. The data set can be downloaded from the web site of UCI Machine Learning Repository (*http://archive.ics.uci.edu/ml/datasets/EEG+Database*).

The EEG data was recently analyzed by Li, Kim, and Altman (2010), whose main purpose focused on dimension reduction. Here we adopt a similar strategy for data preprocessing. In particular, we consider partial data set under the condition of single stimulus only, and the averaged matrix covariates over different trials of the same subject, denoted by X_i^*, will be considered in our analysis. Note that with 256×64 covariate matrix, we will have 320 free parameters which still largely excesses the number of sample size 122. To make the analysis more reliable, before fitting MV-logistic regression, an unsupervised dimension reduction method, the generalize low rank approximations (GLRAM) of Ye (2005), is performed to reduce the dimension of X_i^*. GLRAM is an extension of principal component analysis to matrix object, which aims to find orthogonal bases $A \in \mathbb{R}^{p \times p_0}$ and $B \in \mathbb{R}^{q \times q_0}$ with $p_0 < p$ and $q_0 < q$ such that X_i^* is well explained by the lower dimensional transformation $A^T X_i^* B$. Hung, Wu, Tu, and Huang (2011) also developed statistical justification of GLRAM. Detailed analysis procedure is listed below.

1. Apply GLRAM to find A and B under $(p_0, q_0) = (15, 15)$. Define $\hat{X}_i^* = A^T X_i^* B$.

2. Standardize each element of $\hat{X}_i^*$ to obtain the covariate matrix X_i.

3. Fit MV-logistic regression with $\{(Y_i, X_i)\}_{i=1}^{122}$. We apply the rule suggested in Remark 1.1 to set $\alpha_3 = 1$ and denote the rest α_i's as α^*.

To estimate θ, we adopt the penalty function $J(\theta) = \|\theta\|^2$. The penalty $\lambda = 24$ is chosen so that the leave-one-out classification accuracy is maximized. The resulting estimates of α and β are provided in Figure 1. The corresponding 95% confidence intervals constructed from (15) are also reported. Note that the estimates of α^* are all smaller than 1, which indicates a smaller effect in comparison with the third time point. This is expectable since we set $\alpha_3 = 1$ according to Remark 1.1. As many estimates of α_i's and β_j's are significantly different from zero, channels and measurement times surely play important roles in distinguishing alcoholic and control groups. Figure 2 (a) provides the estimated success probability of each subject (by using the rest 121 subjects) as well as the 95% confidence interval from (19), and Figure 2 (b) gives the estimated density functions of $\hat{\alpha}^T X_i \hat{\beta}$ for two types of subjects. An obvious separation of two groups is detected in these figures which demonstrates the usefulness of our method in classification.

The choice of $(p_0, q_0) = (15, 15)$ in Step 1 is the same with the data preprocessing step of Li, Kim, and Altman (2010). By fitting MV-logistic regression, we correctly classify 105 subjects (through leave-one-out classification procedure) under this choice, while the best result of Li, Kim, and Altman (2010) from dimension folding (a dimension reduction technique that preserves the matrix structure of covariates) followed by quadratic discriminant analysis gives 97. Though the purpose of Li, Kim, and Altman (2010) is focused on dimension reduction, it is not our aim here to compare the classification accuracy, but only want to reveal the usefulness of MV-logistic regression. We think the reasons of a better performance for MV-logistic regression are twofold. First, we adopt a different data preprocessing technique GLRAM in Step 1, where Li, Kim, and Altman (2010) use a version of $(2D)^2$PCA (Zhang and Zhou, 2005). Hung, Wu, Tu, and Huang (2011) prove that GLRAM is asymptotically more efficient than $(2D)^2$PCA in extracting bases and, hence, it is reasonable for GLRAM to produce better result. Second, standardization in Step 2 makes the EEG data more suitable to fit MV-logistic regression model. Without standardization, MV-logistic regression cannot produce such a high classification accuracy. This also reflects that standardization is an important issue before fitting models which preserves the matrix structure of covariates. We remind the reader again that standardization of covariates will result in a different MV-logistic regression model.

We also compare our method with the penalized logistic regression of Le Cessie and Van Houwelingen (1992) under various combinations of (p_0, q_0), and the leave-one-out classification accuracies are summarized in Table 4. It is obvious that MV-logistic regression outperforms conventional logistic regression in every setting, which indicates that the superiority of our method does not contribute from any specific choice of (p_0, q_0). Moreover, since the same data set is used for two methods, it implies the good performance of MV-logistic regression is not due to GLRAM, but the nature of MV-logistic regression. As mentioned previously, MV-logistic regression preserves the matrix information of X, requires less parameters in model fitting, and an efficiency gain is expected. We can also detect this fact in Table 4. Obviously, as the number of p_0q_0 increases, the classification accuracy of conventional logistic regression decays rapidly, while those of MV-logistic regression roughly remain constant.

5 Extension to Multi-Class Response

In this section we briefly illustrate the extension of MV-logistic regression to the case of H classes with $H \geq 2$. Let $Y_i^* \in \{1, \cdots, H\}$ be the random variable indicating which class the i-th subject belongs to. We can equivalently code Y_i^* as $Y_i = (Y_{1i}, \cdots, Y_{H-1,i})^T$, where Y_{hi} is the random variable of the i-th subject with value one indicating the subject belongs to category h, $h = 1, \cdots, H-1$, and $\{Y_{hi} = 0, h = 1, \cdots H-1\}$ means the i-th subject belongs to category H. Consider the model

$$\ln \left\{ \frac{P(Y_i = h|X)}{P(Y_i = H|X)} \right\} = \gamma_h + \alpha_h^T X \beta_h, \; h = 1, \cdots H-1, \tag{21}$$

with $\alpha_h^T = (1, \alpha_h^{*T})^T$ for the sake of identifiability. The parameter of interest is $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_{H-1})$ with $\theta_h = (\gamma_h, \alpha_h^*, \beta_h)$. This gives the covariate-specific probability $P(Y_i = h|X_i)$ for $h = 1, \cdots, H-1$ to be

$$\pi_{hi} = \pi_h(\boldsymbol{\theta}|X_i) = \frac{\exp(\gamma_h + \alpha_h^T X_i \beta_h)}{1 + \sum_{h=1}^{H-1} \exp(\gamma_h + \alpha_h^T X_i \beta_h)}, \tag{22}$$

and for $h = H$ to be

$$\pi_{Hi} = \pi_H(\boldsymbol{\theta}|X_i) = \frac{1}{1 + \sum_{h=1}^{H-1} \exp(\gamma_h + \alpha_h^T X_i \beta_h)}. \tag{23}$$

Based on the data $\{(X_i, Y_i)\}_{i=1}^n$, the log-likelihood function of θ is given by

$$\ell^*(\boldsymbol{\theta}) = \sum_{i=1}^n \left\{ \sum_{h=1}^{H-1} Y_{hi}(\gamma_h + \alpha_h^T X_i \beta_h) - \ln \left(1 + \sum_{h=1}^{H-1} \exp(\gamma_h + \alpha_h^T X_i \beta_h) \right) \right\}. \tag{24}$$

The penalized log likelihood is given by $\ell_\lambda^*(\boldsymbol{\theta}) = \ell^*(\boldsymbol{\theta}) - \lambda J(\boldsymbol{\theta})$ and we can estimate $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}_\lambda = \arg\max_{\boldsymbol{\theta}} \ell_\lambda^*(\boldsymbol{\theta})$.

To apply the Newton method to estimate $\boldsymbol{\theta}$, we need to calculate the gradient vector and Hessian matrix of $\ell^*_\lambda(\boldsymbol{\theta})$. Let $\mathbf{Y}_h = (Y_{h1}, \cdots, Y_{hn})^T$, $\mathbf{Y}^* = (\mathbf{Y}_1^T, \cdots, \mathbf{Y}_{H-1}^T)^T$, $\mathbf{\Pi}_h(\boldsymbol{\theta}) = (\pi_{h1}, \cdots, \pi_{hn})^T$, and $\mathbf{\Pi}^*(\boldsymbol{\theta}) = (\mathbf{\Pi}_1(\boldsymbol{\theta})^T, \cdots, \mathbf{\Pi}_{H-1}(\boldsymbol{\theta})^T)^T$ for $h = 1, \cdots, H-1$. Let also $\mathbf{X}^*(\boldsymbol{\theta}) = \text{diag}(\mathbf{X}(\theta_1), \cdots, \mathbf{X}(\theta_{H-1}))$, where $\mathbf{X}(\cdot)$ is defined as before. Then, we have

$$\ell^{*(1)}_\lambda(\boldsymbol{\theta}) = \mathbf{X}^*(\boldsymbol{\theta})^T \{\mathbf{Y}^* - \mathbf{\Pi}^*(\boldsymbol{\theta})\} - \lambda J^{(1)}(\boldsymbol{\theta}). \tag{25}$$

The Hessian matrix (after ignoring the zero expectation term as in (9)) is derived to be

$$H^*_\lambda(\boldsymbol{\theta}) = \mathbf{X}^*(\boldsymbol{\theta})^T \boldsymbol{V}^*(\boldsymbol{\theta}) \mathbf{X}^*(\boldsymbol{\theta}) + \lambda J^{(2)}(\boldsymbol{\theta}), \tag{26}$$

where $\boldsymbol{V}^*(\boldsymbol{\theta}) = [\boldsymbol{V}_{ij}(\boldsymbol{\theta})]$, $\boldsymbol{V}_{hh} = \text{diag}(\pi_{h1}(1-\pi_{h1}), \cdots, \pi_{hn}(1-\pi_{hn}))$, $h = 1, \cdots, H-1$, and $\boldsymbol{V}_{hk} = \boldsymbol{V}_{kh} = \text{diag}(-\pi_{h1}\pi_{k1}, \cdots, -\pi_{hn}\pi_{kn})$, $1 \leq h \neq k \leq H-1$. Finally, $\hat{\boldsymbol{\theta}}_\lambda$ is obtained by using (25) and (26) in the iteration equation (10).

By a similar proof of Theorem 2.2, we can also deduce that $\sqrt{n}(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta})$ converges weakly to a normal distribution with mean zero and covariance matrix $\Sigma^*(\boldsymbol{\theta}) = \{I^*(\boldsymbol{\theta})\}^{-1}$, where $I^*(\boldsymbol{\theta}) = E[\frac{1}{n}\mathbf{X}^*(\boldsymbol{\theta})^T \boldsymbol{V}^*(\boldsymbol{\theta}) \mathbf{X}^*(\boldsymbol{\theta})]$. Moreover, $\Sigma^*(\boldsymbol{\theta})$ can be estimated by $\hat{\Sigma}^*(\hat{\boldsymbol{\theta}}_\lambda)$, where

$$\hat{\Sigma}^*(\boldsymbol{\theta}) = \left(\frac{1}{n} H^*_\lambda(\boldsymbol{\theta})\right)^{-1} \left(\frac{1}{n}\mathbf{X}^*(\boldsymbol{\theta})^T \boldsymbol{V}^*(\boldsymbol{\theta}) \mathbf{X}^*(\boldsymbol{\theta})\right) \left(\frac{1}{n} H^*_\lambda(\boldsymbol{\theta})\right)^{-1}. \tag{27}$$

Subsequent statistical inference procedures are the same with what we have already established in previous sections.

6 Conclusions

Tensor objects are now frequently encountered in many applications, such as face recognition and image compression. One can also imagine a subject with p covariates measured on q time points at r different places is naturally stored as an order three tensor. In this paper we propose the MV-logistic regression model, when the covariates of interest have a natural matrix structure which is an order two tensor. The implementation and statistical inference procedure (based on the asymptotic normality of $\hat{\theta}_\lambda$) are also developed. The usefulness of our method is validated through simulation studies and EEG data set. We note that in the best case $(p_0, q_0) = (15, 15)$, through leave-one-out cross validation procedure, we successfully classify 105 of 122 subjects to the right group. One reason of the superiority for MV-logistic regression comes from the parsimony of parameter used. It is also found in our limited simulations that MV-logistic regression has certain robustness against the violation of model specification, and its nice performance is expected in many situations. Although we focus on matrix covariate in this article, the proposed method can be extended to tensor objects of higher order.

REFERENCES

Green, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives (with discussion). *J. R. Statist. Soc. B*, 46, 149-192.

Hung, H., Wu, P. S., Tu, I. P., and Huang, S. Y. (2011). On multilinear principal component analysis of order-two tensors, manuscript. arXiv:1104.5281v1.

Le Cessie, S. and Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Appl. Statist.*, 41, 191-201.

Li, B., Kim, M. K., and Altman, N. (2010). On dimension folding of matrix- or array-valued statistical objects. *The Annals of Statistics*, 38, 1094-1121.

MuCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models*, Wiley.

Ye, J. (2005). Generalized low rank approximations of matrices. *Machine Learning*, 61, 167-191.

Zhang, D. and Zhou, Z. H. (2005). $(2D)^2$PCA: Two-directional two-dimensional PCA for efficient face representation and recognition. *Neurocomputing*, 69, 224-231.

Table 1: Averages of $\hat{\theta}_\lambda$ (Mean), averages of $\hat{\Sigma}(\hat{\theta}_\lambda)$ (SE), and standard deviations of $\hat{\theta}_\lambda$ (SD) under model (1) for $n = 150, 300$.

		$n = 150$			$n = 300$		
	True	Mean	SD	SE	Mean	SD	SE
γ	1.000	1.149	0.432	0.447	1.084	0.301	0.292
α^*	0.500	0.547	0.164	0.149	0.521	0.098	0.091
	-0.500	-0.550	0.173	0.152	-0.518	0.094	0.092
	-0.500	-0.555	0.170	0.149	-0.514	0.096	0.091
	-0.500	-0.553	0.169	0.151	-0.521	0.097	0.091
	-0.500	-0.552	0.173	0.149	-0.516	0.101	0.092
	-0.500	-0.541	0.183	0.151	-0.524	0.103	0.093
	-0.500	-0.544	0.173	0.151	-0.514	0.095	0.091
	-0.500	-0.549	0.175	0.150	-0.518	0.101	0.091
	-0.500	-0.553	0.173	0.150	-0.515	0.097	0.091
	-0.500	-0.552	0.171	0.151	-0.518	0.099	0.092
	-0.500	-0.549	0.175	0.151	-0.519	0.098	0.092
β	1.000	1.167	0.298	0.292	1.105	0.206	0.202
	0.500	0.595	0.260	0.244	0.560	0.169	0.158
	1.000	1.165	0.291	0.292	1.114	0.213	0.203
	-1.000	-1.157	0.299	0.291	-1.110	0.216	0.203
	-1.000	-1.168	0.278	0.293	-1.103	0.207	0.202
	-1.000	-1.143	0.290	0.289	-1.097	0.213	0.203
	-1.000	-1.141	0.281	0.291	-1.102	0.206	0.203
	-1.000	-1.161	0.286	0.293	-1.101	0.206	0.203
	-1.000	-1.155	0.276	0.293	-1.108	0.200	0.202
	-1.000	-1.148	0.297	0.292	-1.109	0.215	0.203

Table 2: Average classification accuracies and similarities (standard deviations) under model (1) for $n = 150, 300$.

n		MV-logistic	Logistic
150	Similarity	0.950 (0.021)	0.716 (0.038)
	Accuracy	0.864 (0.031)	0.735 (0.039)
300	Similarity	0.981 (0.007)	0.854 (0.037)
	Accuracy	0.891 (0.027)	0.802 (0.034)

Table 3: Average classification accuracies (standard deviations) under model (20) for $\sigma = 0.1, 0.3, 0.5$ and $n = 150$.

σ	MV-logistic	Logistic
0.1	0.855 (0.032)	0.736 (0.037)
0.3	0.792 (0.038)	0.742 (0.037)
0.5	0.730 (0.048)	0.750 (0.039)

Table 4: The leave-one-out classification accuracy of MV-logistic/Logistic regression for EEG data under different combinations of (p_0, q_0).

	p_0			
q_0	15	30	60	120
15	0.861/0.803	0.844/0.779	0.844/0.689	0.803/0.566
20	0.836/0.795	0.828/0.762	0.853/0.648	0.787/0.549
30	0.844/0.787	0.828/0.754	0.828/0.615	0.812/0.549

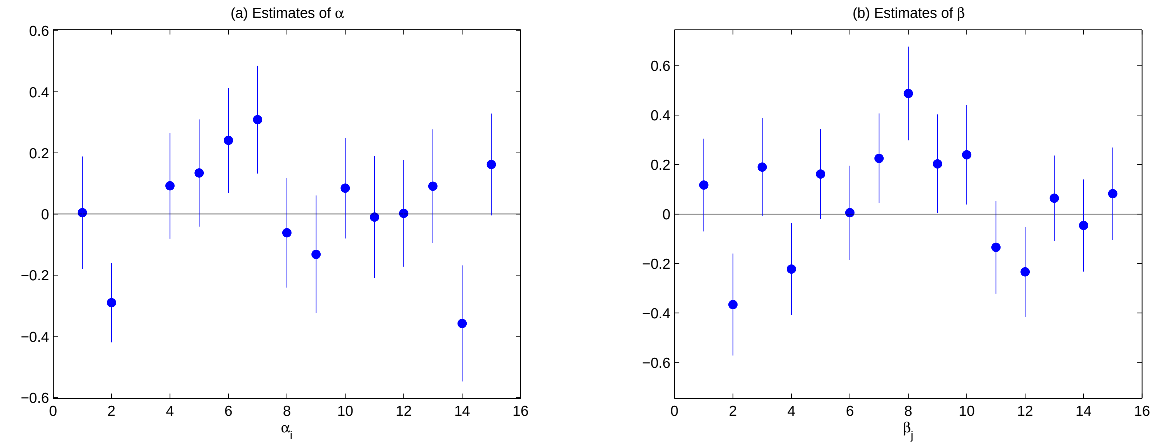

Figure 1: Estimates of α and β (the blue circles) with the corresponding 95% confidence intervals (the blue vertical lines) for EEG data under $(p_0, q_0) = (15, 15)$. Since we set $\alpha_3 = 1$, no estimate is provided for α_3 in (a).

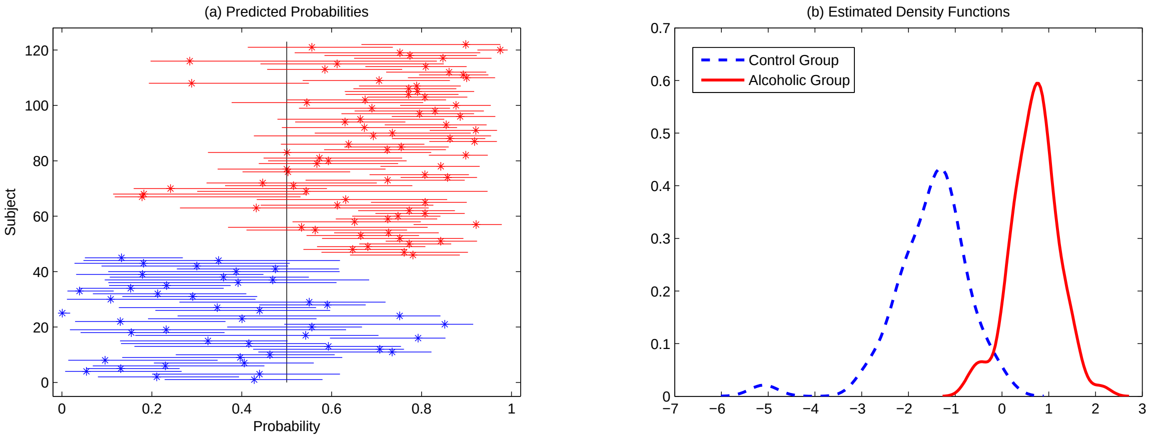

Figure 2: (a) Estimates of $\pi(\theta|X_i)$ (the symbol *) and the corresponding 95% confidence intervals (horizontal lines) for EEG data. Subjects 1-45 and 46-122 belong to the control and alcoholic groups, respectively. (b) Smoothing estimates of the density functions of $\hat{\alpha}^T X_i \hat{\beta}$ for control and alcoholic groups