Biostatistics (2013), **14**, 1, *pp.* 189–202
doi:10.1093/biostatistics/kxs023
Advance Access publication on July 2, 2012

Matrix variate logistic regression model with application to EEG data

HUNG HUNG*

Institute of Epidemiology and Preventive Medicine, National Taiwan University, Taipei, Taiwan
hhung@ntu.edu.tw

CHEN-CHIEN WANG

Institute of Statistical Science, Academia Sinica, Taipei, Taiwan

Summary

Logistic regression has been widely applied in the field of biomedical research for a long time. In some applications, the covariates of interest have a natural structure, such as that of a matrix, at the time of collection. The rows and columns of the covariate matrix then have certain physical meanings, and they must contain useful information regarding the response. If we simply stack the covariate matrix as a vector and fit a conventional logistic regression model, relevant information can be lost, and the problem of inefficiency will arise. Motivated from these reasons, we propose in this paper the matrix variate logistic (MV-logistic) regression model. The advantages of the MV-logistic regression model include the preservation of the inherent matrix structure of covariates and the parsimony of parameters needed. In the EEG Database Data Set, we successfully extract the structural effects of covariate matrix, and a high classification accuracy is achieved.

Keywords: Asymptotic theory; Logistic regression; Matrix covariate; Regularization; Tensor objects.

1. Introduction

The logistic regression model has been widely applied for a long time. It aims to model the logit transformation of the conditional probability of interest as a linear combination of covariates. When the number of covariates exceeds the available sample size, penalized logistic regression (Lee and Silvapulle, 1988; Le Cessie and Van Houwelingen, 1992) is used instead to avoid the problem of high dimensionality. Recently, penalized logistic regression has been applied to processes in the field of biomedical research, such as cancer classification, risk factor selection, and gene interaction detection, etc. See Zhu and Hastie (2004) and Park and Hastie (2008) among others. We refer the readers to McCulloch *and others* (2008) and Hastie *and others* (2009) for details and further extensions of logistic regression model.

In some applications, covariates of interest have a natural matrix structure at the time of collection. For example, the Electroencephalography (EEG) Database Data Set, which will be analyzed in this article later, concerns the relationship between the genetic predisposition and alcoholism. Here, the observed data are of

*To whom correspondence should be addressed.

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 alcoholic group and 0 otherwise, and X is a 256×64 matrix with its (i, j)th element $X_{(i,j)}$ being the measurement of voltage value at time point i and channel of electrode j. With a matrix covariate, one usually stacks X column by column as a pq-vector, say $\mathsf{vec}(X)$, and the subsequent statistical analysis proceeds in the usual way. Specifically, the conventional logistic regression model admits that each $X_{(i,j)}$ possesses its own effect ξ_{ij}. Motivated from the matrix structure of X, it is natural to store ξ_{ij}'s into a $p \times q$ parameter matrix ξ. Then, the conventional logistic regression model takes the formulation

$$\text{logit}\{P(Y = 1|X)\} = \gamma + \sum_{i,j} \xi_{ij} X_{(i,j)} = \gamma + \mathsf{vec}(\xi)^{\mathrm{T}} \mathsf{vec}(X), \tag{1.1}$$

where γ is the intercept term. It is obvious that model (1.1) does not consider the inherent matrix structure of X, as the corresponding parameter matrix ξ enters model (1.1) only through $\mathsf{vec}(\xi)$ while its matrix structure is ignored. As X is a matrix at the time of collection, it is reasonable for ξ to possess certain structure. For example, those $X_{(i,j)}$'s in a common region or in the same column (row) of X may have similar effects. Directly fitting (1.1) without considering the structural relationships among ξ_{ij}'s will cause the problem of inefficiency. Moreover, by ignoring the inherent matrix structure, one can hardly identify the row and column effects from their joint effect matrix ξ. Another problem of model (1.1) is that the number of parameters usually becomes extremely large. For instance, we have $1 + 64 \times 256 = 16\,385$ parameters for the EEG Database Data Set when fitting model (1.1), while the available sample size is 122 only. The primary focus of this research is thus on overcoming the aforementioned problems via incorporating the hidden structural information of X into statistical model building.

The rest of this article is organized as follows. Section 2 introduces the matrix variate logistic regression model and its statistical meaning. Section 3 deals with the asymptotic properties of our estimators and the implementation algorithm. The proposed method is further evaluated through two simulation studies in Section 4 and the EEG Database Data Set in Section 5. Extension and connection to tensor learning are discussed in Section 6.

2. Matrix variate logistic regression model

2.1 *Model specification*

To incorporate the structural information into modeling, as motivated from the matrix structure of X, we propose the matrix variate logistic (MV-logistic) regression model

$$\text{logit}\{P(Y = 1|X)\} = \gamma + \alpha^{\mathrm{T}} X \beta, \tag{2.1}$$

where $\alpha = (\alpha_1, \ldots, \alpha_p)^{\mathrm{T}}$ and $\beta = (\beta_1, \ldots, \beta_q)^{\mathrm{T}}$ are the row and column parameter vectors, respectively, and γ is the intercept term. As the matrix structure of X is preserved in model (2.1), these parameters have their own physical meanings. In the EEG Database Data Set, for instance, α is interpreted as the time effect and β as the channel effect. By fitting the MV-logistic regression model, we are able to extract the column (row) information of X, which will provide further insights into the relationship between Y and X. Note that under model (2.1), (α, β) can be identified only up to a scale, since $(c^{-1}\alpha, c\beta)$ will result in the same model for any constant $c \neq 0$. For the sake of identifiability, without loss of generality, we assume that $\alpha_1 = 1$ in our derivation (see Remark 2.1 for further discussion). Denote the rest of the parameters in α by α^*, i.e. $\alpha = (1, \alpha^{*\mathrm{T}})^{\mathrm{T}}$ and $\theta = (\gamma, \alpha^{*\mathrm{T}}, \beta^{\mathrm{T}})^{\mathrm{T}}$ are the parameters of interest. We thus have $p + q$ free parameters contained in θ, while it is $pq + 1$ in the conventional logistic regression model (1.1). One can see that a merit of model (2.1) is the parsimony of parameters used. Thus, when model (2.1) is correctly specified, an efficiency gain is reasonably expected.

Adoption of model (2.1) is equivalent to modeling the covariate-specific odds ratio R_{ij} of $X_{(i,j)}$ while keeping the rest of the covariates fixed as

$$\log(R_{ij}) = \alpha_i \beta_j \quad \Leftrightarrow \quad R_{ij} = \{\exp(\beta_j)\}^{\alpha_i}. \tag{2.2}$$

Thus, the effect of $X_{(i,j)}$ depends on the product $\alpha_i \beta_j$ instead of on α_i or β_j solely. A positive $\alpha_i \beta_j$ implies $R_{ij} > 1$. Note that since we set $\alpha_1 = 1$, if $\beta_j > 0$ (< 0), then $\alpha_i > 1$ (< 1) indicates $R_{ij} > R_{1j}$. Take the EEG Database Data Set again for example, relation (2.2) implies that each channel has its own baseline odds ratio $\exp(\beta_j)$. Depending on the measured time point, say i, the odds ratio is further modified by taking a power of α_i.

Remark 2.1. Although at the population level there is no difference for which α_i is set to one, it is crucial to practical implementation. If α_i is near zero, setting $\alpha_i = 1$ will lead to unstable 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 $i = \text{argmax}_{k \in \{1,\ldots,p\}} \sum_{j=1}^{q} |\rho_{kj}|$. The intuition is to choose the one as the baseline that is the most likely to be correlated with the response. We find in our numerical studies that this simple approach performs well.

Remark 2.2. Standardization of covariates will not affect the final result of the logistic regression model except a change of scale in parameter estimates, since the standard deviation of each covariate can be absorbed into the corresponding parameter. In the 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 fewer parameters. In summary, standardization of covariates will result in a different MV-logistic regression model. We will further discuss this issue in Section 5.

Remark 2.3. The proposed MV-logistic regression model (2.1) is closely related to tensor learning in the field of computer science. Detailed discussions and comparisons are placed in Section 6.

2.2 *Statistical meaning of MV-logistic regression model*

This subsection is devoted to the investigation of the statistical meaning of the MV-logistic regression model. We will assume the validity of the general model (1.1), (γ_0, ξ_0) being the true value of (γ, ξ). Under this situation, one will see that fitting the MV-logistic regression model actually aims to estimate *the best rank-1 approximation* of ξ_0, in the sense of minimum Kullback–Leibler divergence (KL-divergence), or equivalently, maximum likelihood. This observation supports the applicability of MV-logistic regression model in practice.

First observe that, as $\alpha^{\mathrm{T}} X \beta = \mathsf{vec}(\alpha\beta^{\mathrm{T}})^{\mathrm{T}} \mathsf{vec}(X)$, MV-logistic regression model (2.1) is equivalent to the conventional model (1.1) with the constraint $\xi = \alpha\beta^{\mathrm{T}}$. Thus, MV-logistic regression utilizes the matrix structure of ξ and approximates it by a rank-1 matrix $\alpha\beta^{\mathrm{T}}$ in model fitting. Secondly, it is known that maximizing the likelihood function is equivalent to minimizing the KL-divergence (Bickel and Doksum, 2001). Let $f(y|X; \gamma, \xi)$ be the conditional distribution function of Y given X under model (1.1). The KL-divergence between (γ_0, ξ_0) and any (γ, ξ) is defined as

$$KL(\gamma_0, \xi_0 \| \gamma, \xi) = E_X \left\{ \int \left(\log \frac{f(y|X; \gamma_0, \xi_0)}{f(y|X; \gamma, \xi)} \right) f(y|X; \gamma_0, \xi_0) \, \mathrm{d}y \right\}, \tag{2.3}$$

where $E_X(\cdot)$ takes expectation with respect to X. Combining these two facts, at the population level, fitting MV-logistic regression to estimate (γ, α, β) is equivalent to searching the minimizer of the minimization problem

$$\min_{\gamma,\alpha,\beta} KL(\gamma_0, \xi_0 \| \gamma, \alpha\beta^{\mathrm{T}}). \tag{2.4}$$

Let (α_0, β_0) be the minimizer of (α, β) in (2.4). MV-logistic regression then aims to search the matrix $\alpha_0\beta_0^{\mathrm{T}}$, which is called the best rank-1 approximation of the true parameter matrix ξ_0. Notice that the optimality between ξ_0 and $\alpha_0\beta_0^{\mathrm{T}}$ here is not measured by the usual Frobenius norm, but the KL-divergence (2.3) instead (see Remark 2.4 for more explications). If $\xi_0 = \alpha_0\beta_0^{\mathrm{T}}$, MV-logistic regression must be more efficient than conventional logistic regression in estimating ξ_0 by standard MLE arguments. Even if ξ_0 does not equal $\alpha_0\beta_0^{\mathrm{T}}$, from the view point of the best rank-1 approximation of ξ_0, we are still in favor of MV-logistic regression, especially when the sample size is relatively small in comparison with the number of covariates. In fact, there is a trade-off between "correctness of model specification" and "efficiency of estimation". With limited sample size available, instead of unstably estimating the full parameter matrix ξ_0 by conventional approach, the MV-logistic regression model aims to more efficiently estimate the best rank-1 approximation of ξ_0.

Clearly, the performance of the MV-logistic regression model relies on "how well the true parameter matrix ξ_0 can be approximated by a rank-1 matrix $\alpha_0\beta_0^{\mathrm{T}}$". As demonstrated in Section 5 that MV-logistic regression outperforms the conventional approach in the EEG Database Data Set, we believe that this condition for ξ_0 is not restrictive for the following two reasons. First, it is reasonable to assume that most covariates have effect sizes near zero in modern biomedical research and, hence, ξ_0 is plausible to be a low-rank matrix. The second reason is based on the characteristic of the underlying study. In the EEG study, for instance, measurements at the same time points (channels) would have similar effects. These facts reflect the potential of ξ_0 to be well explained by low-rank approximation and, hence, the validity of the MV-logistic regression model. The robustness of the MV-logistic regression model will be studied further in Section 4.2.

Remark 2.4. Given the matrix ξ_0, the weighted rank-1 approximation problem of ξ_0 with the weight matrix W (Lu *and others*, 1997; Manton *and others*, 2003) is to find the minimizer (α_0, β_0) of the minimization problem $\min_{\alpha,\beta} \|\xi_0 - \alpha\beta^{\mathrm{T}}\|_W^2$, where $\|\cdot\|_W^2 = \mathsf{vec}(\cdot)^{\mathrm{T}} W \mathsf{vec}(\cdot)$. The resulting matrix $\alpha_0\beta_0^{\mathrm{T}}$ is called the best rank-1 approximation of ξ_0 in the sense of minimum $\|\cdot\|_W$ norm. When $W = I_{pq}$, the $\|\cdot\|_W$ norm reduces to the Frobenius norm, and $\alpha_0\beta_0^{\mathrm{T}}$ can be obtained by singular value decomposition (SVD) of ξ_0. Parallel to this idea, fitting the MV-logistic regression model (with (2.4) being the estimation criterion) can be treated as finding the best rank-1 approximation of ξ_0 in the sense of minimum KL-divergence (2.3).

3. STATISTICAL INFERENCE PROCEDURE

Some notation is defined here for ease of reference. For any function g, $g^{(k)}$ represents the kth derivative of g with respect to its argument. The parameters of interest are $\theta = (\gamma, \alpha^{*\mathrm{T}}, \beta^{\mathrm{T}})^{\mathrm{T}}$. Denote $P(Y_i = 1|X_i)$ by $\pi_i = \pi(\theta|X_i) = \exp(\gamma + \alpha^{\mathrm{T}}X_i\beta)\{1 + \exp(\gamma + \alpha^{\mathrm{T}}X_i\beta)\}^{-1}$. Define $\Pi(\theta) = (\pi_1, \ldots, \pi_n)^{\mathrm{T}}$ and $V(\theta) = \mathrm{diag}(v_1, \ldots, v_n)$, where $v_i = \pi_i(1 - \pi_i)$ is the conditional variance of Y_i given X_i. Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ and $\mathbf{X}(\theta) = [\mathbf{X}_1(\theta), \ldots, \mathbf{X}_n(\theta)]^{\mathrm{T}}$ be an $n \times pq$ matrix, where $\mathbf{X}_i(\theta) = (1, \beta^{\mathrm{T}}X_i^{\mathrm{T}}C, \alpha^{\mathrm{T}}X_i)^{\mathrm{T}}$, $C = \partial\alpha/\partial\alpha^* = [\mathbf{0}_{p-1}, \mathbf{I}_{p-1}]^{\mathrm{T}}$, $\mathbf{0}_a$ is the a-vector of zeros, and $\mathbf{I}_a$ is the $a \times a$ identity matrix. One can treat $\mathbf{X}_i(\theta)$ as the working covariates of the ith subject, which will be used in the development of our method. Note that C depends on which element of α is set to one. As $\alpha_1 = 1$ throughout the paper, the notation C is used for simplicity.

3.1 *Estimation and implementation*

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

$$\ell(\theta) = \sum_{i=1}^{n} Y_i(\gamma + \alpha^{\mathrm{T}} X_i \beta) - \log\{1 + \exp(\gamma + \alpha^{\mathrm{T}} X_i \beta)\}, \tag{3.1}$$

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

In modern research in biostatistics, however, an important issue is that 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 (2.1) is $p + q$, while it is $pq + 1$ in the conventional logistic regression model. MV-logistic regression then suffers less severity from high dimensionality, but still cannot avoid it entirely. As we have seen in the EEG Database Data Set, the covariate X is a 256×64 matrix (i.e. 320 parameters in the MV-logistic regression model), while there are only 122 observations. To overcome the difficulty of high dimensionality, Le Cessie and Van Houwelingen (1992) proposed the penalized logistic regression method. This motivates us to consider the penalized MV-logistic regression. Define the penalized log likelihood function to be

$$\ell_\lambda(\theta) = \ell(\theta) - \lambda J(\theta), \tag{3.2}$$

where $J(\theta) \geqslant 0$ is a twice continuously differentiable penalty function of θ and $\lambda \geqslant 0$ is the regularization parameter. We then propose to estimate θ by

$$\hat{\theta}_\lambda = \underset{\theta}{\operatorname{argmax}}\, \ell_\lambda(\theta). \tag{3.3}$$

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

Interpretations of the elements of $\hat{\theta}_\lambda$ have been introduced in Section 2.1. Let $\hat{\theta}_\lambda = (\hat{\gamma}, \hat{\alpha}^{*\mathrm{T}}, \hat{\beta}^{\mathrm{T}})^{\mathrm{T}}$ and $\hat{\alpha} = (1, \hat{\alpha}^{*\mathrm{T}})^{\mathrm{T}}$. The odds ratio R_{ij} in (2.2) is estimated by $\exp(\hat{\alpha}_i \hat{\beta}_j)$. We would also be interested in the success probability $\pi(\theta|x)$ for any given $p \times q$ matrix x, which can be estimated by $\pi(\hat{\theta}_\lambda|x)$. A discrimination rule is then to classify a subject with matrix covariate x as the "diseased" group if $\pi(\hat{\theta}_\lambda|x) > 0.5$ and as the "non-diseased" group otherwise. Detailed inference procedures about θ and $\pi(\theta|x)$ will be discussed in the next subsection.

We close this subsection by introducing the iterative Newton method to obtain $\hat{\theta}_\lambda$. The gradient (with respect to θ) of $\ell_\lambda(\theta)$ in (3.2) is calculated to be

$$\ell_\lambda^{(1)}(\theta) = \ell^{(1)}(\theta) - \lambda J^{(1)}(\theta), \tag{3.4}$$

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

$$\ell_\lambda^{(2)}(\theta) = -H_\lambda(\theta) + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & \sum_{i=1}^{n} C^{\mathrm{T}} X_i (Y_i - \pi_i) \\ 0 & \sum_{i=1}^{n} X_i^{\mathrm{T}} C (Y_i - \pi_i) & 0 \end{bmatrix}, \tag{3.5}$$

where

$$H_\lambda(\theta) = \mathbf{X}(\theta)^{\mathrm{T}} \boldsymbol{V}(\theta)\mathbf{X}(\theta) + \lambda J^{(2)}(\theta). \tag{3.6}$$

As suggested by Green (1984), we will ignore the last term of (3.5), since its expectation is zero. Finally, $\hat{\theta}_\lambda$ can be obtained through iterating

$$\theta_{(k+1)} = \theta_{(k)} + \{H_\lambda(\theta_{(k)})\}^{-1} \ell_\lambda^{(1)}(\theta_{(k)}), \quad k = 0, 1, 2, \ldots, \tag{3.7}$$

until there is no significant difference between $\theta_{(k+1)}$ and $\theta_{(k)}$, and outputs $\hat{\theta}_\lambda = \theta_{(k+1)}$. A zero initial $\theta_{(0)}$ is suggested and performs well in our numerical studies.

3.2 *Asymptotic properties*

Asymptotic properties of $\hat{\theta}_\lambda$ can be derived through usual arguments of MLE. The result is summarized in Theorem 3.1. For the proof see supplementary material available at *Biostatistics* online.

THEOREM 3.1 Assume model (2.1) and the regularity conditions (A)–(C) in Fan and Li (2001) for the likelihood function. Assume also that $\lambda/n = o(n^{-1/2})$ as n tends to infinity. Then, $\sqrt{n}(\hat{\theta}_\lambda - \theta)$ converges weakly to $N\{0, \Sigma(\theta)\}$, where $\Sigma(\theta) = \{I(\theta)\}^{-1}$ and $I(\theta) = E\{\mathbf{X}_i(\theta) v_i \mathbf{X}_i(\theta)^{\mathrm{T}}\}$.

From Theorem 3.1, $\hat{\theta}_\lambda$ is shown to be a consistent estimator for θ. It also enables us to construct a 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

$$\hat{\Sigma}(\theta) = \left(\frac{1}{n} H_\lambda(\theta)\right)^{-1} \left(\frac{1}{n}\mathbf{X}(\theta)^{\mathrm{T}} \boldsymbol{V}(\theta)\mathbf{X}(\theta)\right) \left(\frac{1}{n} H_\lambda(\theta)\right)^{-1}. \tag{3.8}$$

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

$$\left(\hat{\theta}_{\lambda,i} - z_{a/2}\frac{[\hat{\Sigma}(\hat{\theta}_\lambda)]_i}{\sqrt{n}},\ \hat{\theta}_{\lambda,i} + z_{a/2}\frac{[\hat{\Sigma}(\hat{\theta}_\lambda)]_i}{\sqrt{n}}\right), \tag{3.9}$$

where $z_{a/2}$ is the $1 - a/2$ quantile of standard normal and $[\hat{\Sigma}(\hat{\theta}_\lambda)]_i$ denotes the ith diagonal element of $\hat{\Sigma}(\hat{\theta}_\lambda)$. As to making inference about $\pi(\theta|x)$ for any $p \times q$ matrix x, by applying the delta method and the result of Theorem 3.1, we have

$$\sqrt{n}\left\{\log\left(\frac{\pi(\hat{\theta}_\lambda|x)}{1-\pi(\hat{\theta}_\lambda|x)}\right) - \log\left(\frac{\pi(\theta|x)}{1-\pi(\theta|x)}\right)\right\} \xrightarrow{d} N\{0, \sigma_\pi^2(\theta|x)\}, \tag{3.10}$$

where $\sigma_\pi^2(\theta|x) = \boldsymbol{x}(\theta)^{\mathrm{T}}\Sigma(\theta)\boldsymbol{x}(\theta)$ and $\boldsymbol{x}(\theta) = (1, \beta^{\mathrm{T}}x^{\mathrm{T}}C, \alpha^{\mathrm{T}}x)^{\mathrm{T}}$. The asymptotic variance $\sigma_\pi^2(\theta|x)$ can be estimated by $\hat{\sigma}_\pi^2(\hat{\theta}_\lambda|x)$, where $\hat{\sigma}_\pi^2(\theta|x) = \boldsymbol{x}(\theta)^{\mathrm{T}}\hat{\Sigma}(\theta)\boldsymbol{x}(\theta)$. An approximate $100(1-a)\%$ confidence interval of $\pi(\theta|x)$ is then constructed to be

$$\left(\frac{\exp\{\hat{\gamma} + \hat{\alpha}^{\mathrm{T}}x\hat{\beta} - z_{a/2}(\hat{\sigma}_\pi(\hat{\theta}_\lambda|x)/\sqrt{n})\}}{1+\exp\{\hat{\gamma} + \hat{\alpha}^{\mathrm{T}}x\hat{\beta} - z_{a/2}(\hat{\sigma}_\pi(\hat{\theta}_\lambda|x)/\sqrt{n})\}}, \frac{\exp\{\hat{\gamma} + \hat{\alpha}^{\mathrm{T}}x\hat{\beta} + z_{a/2}(\hat{\sigma}_\pi(\hat{\theta}_\lambda|x)/\sqrt{n})\}}{1+\exp\{\hat{\gamma} + \hat{\alpha}^{\mathrm{T}}x\hat{\beta} + z_{a/2}(\hat{\sigma}_\pi(\hat{\theta}_\lambda|x)/\sqrt{n})\}}\right) \tag{3.11}$$

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

4. Simulation studies

The proposed method is evaluated through simulations under two different settings. Let the parameters be set as $\gamma = 1$, $\alpha = (1, 0.5, -0.5\mathbf{1}_{p-2}^{\mathrm{T}})^{\mathrm{T}}$, and $\beta = (1, 0.5, 1, -\mathbf{1}_{q-3}^{\mathrm{T}})^{\mathrm{T}}$, where $\mathbf{1}_a$ is the a-vector of ones, and the penalty function $J(\theta) = (\|\alpha^*\|^2 + \|\beta\|^2)/2$ is considered. The tuning parameters for different logistic regression models are determined from an independent simulation by maximizing the classification accuracy. Simulation results are reported with 1000 replicates.

4.1 *Simulation under the MV-logistic regression model*

The first simulation study evaluates the proposed method under model (2.1) with $(p, q) = (12, 10)$. We first generate X such that $\mathsf{vec}(X)$ follows a pq-variate normal distribution with mean zero and covariance matrix $\mathbf{I}_{pq}$. Conditional on X, Y is generated from model (2.1) with the specified θ value. The averages of $\hat{\theta}_\lambda$ and standard errors from the diagonal elements of $\hat{\Sigma}(\hat{\theta}_\lambda)$ in (3.8) are provided in Table 1. For the case of small sample size $n = 150$, biases for $\hat{\theta}_\lambda$ are detected. The biases are mainly due to the penalty term $\lambda J(\theta)$, as is the case of ridge regression. In fact, from the proof of Theorem 3.1, the bias term is derived to be $\lambda n^{-1}\{I(\theta)J^{(1)}(\theta)\}$, which is of order n^{-1}. The biases, however, are all relatively small in comparison with the corresponding standard deviations of $\hat{\theta}_\lambda$, and decrease as the sample size increases. Moreover, all the standard deviations are well estimated by the diagonal elements of $\hat{\Sigma}(\hat{\theta}_\lambda)$. These observations validate Theorem 3.1 and the proposed empirical variance estimator.

As $\hat{\gamma} + \hat{\alpha}^{\mathrm{T}} X \hat{\beta} = \hat{\gamma} + \mathsf{vec}(\hat{\alpha}\hat{\beta}^{\mathrm{T}})^{\mathrm{T}}\mathsf{vec}(X)$ is the critical component used in prediction, we also report in Table 1 the averages of similarities $u^{\mathrm{T}}v/(\|u\| \cdot \|v\|)$ with $u = (\gamma, \mathsf{vec}(\alpha\beta^{\mathrm{T}})^{\mathrm{T}})^{\mathrm{T}}$ and $v = (\hat{\gamma}, \mathsf{vec}(\hat{\alpha}\hat{\beta}^{\mathrm{T}})^{\mathrm{T}})^{\mathrm{T}}$. Although biases arise especially for the case of $n = 150$, the similarities are not affected and have values very close to one in both cases. This means MV-logistic regression would have good performance in classification, since it is the direction of $(\gamma, \mathsf{vec}(\alpha\beta^{\mathrm{T}})^{\mathrm{T}})^{\mathrm{T}}$ that is relevant to classification. To demonstrate this fact, in each simulation replicate, we generate another independent data set to calculate the classification accuracy by using $\hat{\gamma} + \hat{\alpha}^{\mathrm{T}} X \hat{\beta}$. For comparison, we also fit the conventional logistic regression model (1.1) to obtain estimates of (γ, ξ) as if we ignore the relationship $\xi = \alpha\beta^{\mathrm{T}}$, and denote it by $(\tilde{\gamma}, \tilde{\xi})$. The classification accuracy obtained from $\tilde{\gamma} + \mathsf{vec}(\tilde{\xi})^{\mathrm{T}}\mathsf{vec}(X)$ is also calculated. The average classification accuracies and the winning proportions (the proportion of MV-logistic regression with higher classification accuracy over 1000 replicates) under $n = 150$ are provided in Table 2 (the row indexed by $\sigma = 0$). It is detected that MV-logistic regression produces more accurate results, and the superiority of MV-logistic regression becomes more obvious for larger (p, q) values.

4.2 *Simulation violating the MV-logistic regression model*

In the numerical study of Section 4.1, MV-logistic regression outperformed conventional logistic regression when data were generated from model (2.1). It is our purpose here to evaluate the performance of MV-logistic regression, when the underlying distribution departures from model (2.1). The same setting in Section 4.1 is used, except for each simulation replicate, Y is generated from the conventional logistic regression model (1.1) with $\gamma = 1$ and $\xi = \alpha\beta^{\mathrm{T}} + \delta$, where each element of the $p \times q$ matrix δ is also randomly drawn from a normal distribution with mean zero and variance σ^2. With an extra term δ, model (2.1) is violated, and the magnitude of violation is controlled by σ. Both models (1.1) and (2.1) are fitted to obtain the estimates $(\tilde{\gamma}, \tilde{\xi})$ and $(\hat{\gamma}, \hat{\xi})$ with $\hat{\xi} = \hat{\alpha}\hat{\beta}^{\mathrm{T}}$, respectively. We compare the classification accuracies of the predictors $\hat{\gamma} + \mathsf{vec}(\hat{\xi})^{\mathrm{T}}\mathsf{vec}(X)$ and $\tilde{\gamma} + \mathsf{vec}(\tilde{\xi})^{\mathrm{T}}\mathsf{vec}(X)$ applied on another independent data set, under different combinations of (p, q) and $\sigma = 0.1, 0.3, 0.5$. Note that $\sigma = 0$ means that MV-logistic regression is the true model. To see how these σ values affect the deviation from MV-logistic regression model, we also report the corresponding explained proportions ρ_σ defined below. Let $k = p \wedge q$ and $s_1 > \cdots > s_k > 0$

Table 1. *Averages of* $\hat{\theta}_\lambda$ (*Mean*), *averages of diagonal elements of* $\hat{\Sigma}(\hat{\theta}_\lambda)$ (*SE*), *and standard deviations of* $\hat{\theta}_\lambda$ (*SD*) *under model* (2.1) *with* $(p, q) = (12, 10)$. *The last row gives averages* (*standard deviations*) *of similarities* (*SIM*) *between* $(\gamma, \mathrm{vec}(\alpha\beta^{\mathrm{T}})^{\mathrm{T}})^{\mathrm{T}}$ *and* $(\hat{\gamma}, \mathrm{vec}(\hat{\alpha}\hat{\beta}^{\mathrm{T}})^{\mathrm{T}})^{\mathrm{T}}$

		$n = 150$			$n = 300$		
	True	Mean	SD	SE	Mean	SD	SE
γ	1.000	1.095	0.447	0.460	1.055	0.289	0.287
α^*	0.500	0.549	0.165	0.157	0.525	0.098	0.096
	−0.500	−0.543	0.169	0.156	−0.524	0.103	0.096
	−0.500	−0.558	0.162	0.158	−0.525	0.100	0.095
	−0.500	−0.557	0.170	0.157	−0.528	0.095	0.095
	−0.500	−0.560	0.168	0.157	−0.524	0.101	0.096
	−0.500	−0.550	0.165	0.156	−0.524	0.094	0.095
	−0.500	−0.557	0.168	0.157	−0.523	0.096	0.095
	−0.500	−0.546	0.161	0.157	−0.525	0.098	0.096
	−0.500	−0.557	0.164	0.157	−0.527	0.097	0.096
	−0.500	−0.555	0.163	0.157	−0.521	0.097	0.095
	−0.500	−0.554	0.159	0.156	−0.526	0.102	0.095
β	1.000	0.972	0.226	0.237	0.995	0.174	0.177
	0.500	0.485	0.206	0.207	0.506	0.141	0.143
	1.000	0.962	0.220	0.238	0.999	0.170	0.178
	−1.000	−0.970	0.222	0.238	−1.008	0.171	0.179
	−1.000	−0.974	0.219	0.238	−0.999	0.172	0.178
	−1.000	−0.969	0.222	0.238	−0.996	0.172	0.177
	−1.000	−0.962	0.215	0.238	−1.008	0.169	0.178
	−1.000	−0.968	0.223	0.238	−0.996	0.169	0.178
	−1.000	−0.966	0.220	0.238	−1.012	0.171	0.179
	−1.000	−0.959	0.216	0.237	−1.000	0.172	0.178
SIM		0.950	(0.021)		0.981	(0.007)	

Table 2. *Averages of classification accuracies* (*CA*) *of MV-logistic/logistic regression and winning proportions* (*WP*) *of MV-logistic regression for different values of* σ, p, q *and* $n = 150$. ρ_σ *is the average of the explained proportions of the best rank-1 approximation of* ξ

(p, q)	(12, 10)		(20, 15)		(20, 30)	
σ	ρ_σ (%)	CA (WP)	ρ_σ (%)	CA (WP)	ρ_σ (%)	CA (WP)
0	100	0.867/0.739 (1.00)	100	0.866/0.670 (1.00)	100	0.839/0.622 (1.00)
0.1	69	0.856/0.741 (1.00)	63	0.848/0.668 (1.00)	58	0.821/0.622 (1.00)
0.3	44	0.794/0.745 (0.87)	36	0.765/0.669 (0.94)	32	0.720/0.623 (0.94)
0.5	33	0.727/0.751 (0.37)	26	0.677/0.673 (0.60)	22	0.631/0.621 (0.61)

be the k non-zero singular values of each generated ξ. Then ρ_σ is the average (over 1000 replicates) of $s_1/(\sum_{\ell=1}^{k} s_\ell)$. Here $0 < \rho_\sigma < 1$ measures how well ξ can be explained by its best rank-1 approximation. Small value of ρ_σ indicates severe violation of the MV-logistic regression model.

The analysis results with $n = 150$ are presented in Table 2. For every σ, conventional logistic regression is the correct model, and its classification accuracies are irrelevant to σ but will decay rapidly as pq increases. On the other hand, MV-logistic regression is less affected by (p, q) than conventional approach as it involves only $p + q$ parameters. The cost for its parsimony of parameters is that, for

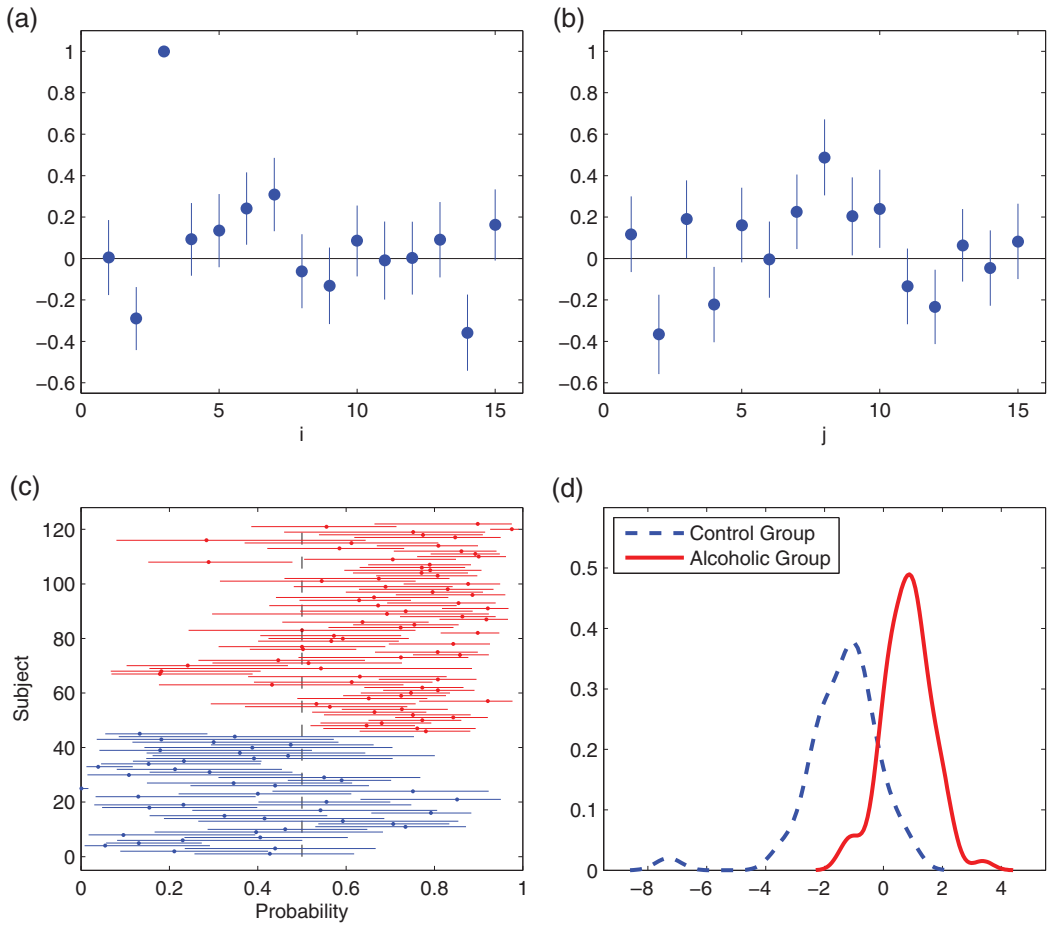

Fig. 1. Analysis results of the EEG Database Data Set with $(p_0, q_0) = (15, 15)$. (a)–(b) Estimates of α and β (the circles) with the 95% confidence intervals (the vertical bars). As we set $\alpha_3 = 1$, no confidence interval is provided for α_3. (c) Estimates of $\pi(\theta|X_i)$ (the symbol $*$) with 95% confidence intervals (the horizontal lines). Subjects 1–45 and 46–122 belong to the control and alcoholic groups, respectively. The vertical dash line indicates a probability of value 0.5. (d) Kernel density estimates of $\hat{\alpha}^{\mathrm{T}} X_i \hat{\beta}$ for control and alcoholic groups.

any fixed (p, q), its performance becomes worse for larger σ. Overall, for moderate deviations $\sigma \leqslant 0.3$, MV-logistic regression outperforms the conventional approach for every combination of (p, q). For $\sigma = 0.5$, MV-logistic regression has lower classification accuracy at $(p, q) = (12, 10)$. We note that in this case, $\rho_\sigma = 33\%$ implies a large deviation from the MV-logistic regression model, while conventional logistic regression is expected to have better performance as the number of its parameters is $121 < n = 150$. For larger (p, q) values, however, MV-logistic regression will be the winner even in the most extreme case of $(p, q) = (20, 30)$, where the explained proportion ρ_σ is 22% only. It indicates that when p and q are large, the gain in efficiency (from fitting the MV-logistic regression model) more easily exceeds the loss due to model misspecification. These observations show that the MV-logistic regression model has certain robustness against the violation of model specification, especially in the case of a large number of covariates.

5. THE EEG DATABASE DATA SET

In this section, we analyze the EEG Database Data Set to demonstrate the usefulness of MV-logistic regression. The data set consists of 122 subjects, wherein 77 of them belong to the group of alcoholism ($Y_i = 1$) and the remaining 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 results in a 256×64 covariate matrix. It is interesting to distinguish two types of subjects on the basis of 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 Database Data Set was recently analyzed by Li *and others* (2010), who mainly focused on dimension reduction. Here we adopt a similar strategy for data preprocessing. In particular, we consider partial data set of single-stimulus experimenters only, and the average matrix covariates over different trials of the same subject (denoted by X_i^*) will be considered in our analysis. This data setting is the same as the one used in Li *and others* (2010). Note that with a 256×64 covariate matrix, we have 320 free parameters, which is still in excess of the sample size of 122. Before fitting the MV-logistic regression model, the algorithm GLRAM (the Generalized Low Rank Approximations of Matrices) of Ye (2005) is performed to reduce the dimensionality of X_i^* first. GLRAM is an extension of principal component analysis to matrix objects, 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 $A^{\mathrm{T}} X_i^* B$. The detailed procedure is listed below.

(1) Apply GLRAM to find A and B under $(p_0, q_0) = (15, 15)$. Define $\hat{X}_i^* = A^{\mathrm{T}} X_i^* B$.
(2) Standardize each element of $\hat{X}_i^*$ to obtain the covariate matrix X_i.
(3) Fit MV-logistic regression model with $\{(Y_i, X_i)\}_{i=1}^{122}$. We apply the rule suggested in Remark 2.1 to set $\alpha_3 = 1$ and denote the remaining α_i's by α^*.

To estimate θ in Step 3, we adopt the penalty function $J(\theta) = \|\theta\|^2/2$. 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(a) and (b) with the corresponding 95% confidence intervals constructed from (3.9). 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. Observe that all the estimates of α^* are smaller than 1. Thus, for those channels with $\hat{\beta}_j > 0$, the effects of measurement times are all smaller than the third time point. In other words, most channels achieve the largest odds ratio at the third time point. For the rest of the channels with $\hat{\beta}_j < 0$, the largest effect happens at the 14th time point. Moreover, among all combinations of time points and channels, $X_{(3,8)}$ has the largest odds ratio (since $\hat{\alpha}_3 \hat{\beta}_8$ is the largest among all $\hat{\alpha}_i \hat{\beta}_j$) and would be critical in classifying alcoholic and control groups. Figure 1(c) provides the predicted probability of being alcoholism for every subject (by using the rest 121 subjects) as well as the 95% confidence interval from (3.11), and Figure 1(d) gives the kernel density estimates of $\hat{\alpha}^{\mathrm{T}} X_i \hat{\beta}$ for two types of subjects. An obvious separation of two groups is detected which demonstrates the usefulness of MV-logistic regression in classification.

The choice of $(p_0, q_0) = (15, 15)$ in Step 1 is the same as in the data preprocessing step of Li *and others* (2010). Under this choice, we correctly classify 105 of 122 subjects (through leave-one-out classification procedure) by fitting MV-logistic regression, while the best result of Li *and others* (2010) from dimension folding (a dimension reduction technique that preserves the matrix structure of covariates) followed by quadratic discriminant analysis gives 97. The explanations of a better performance for MV-logistic regression are twofold. First, we adopt a different data preprocessing technique GLRAM in Step 1, where Li *and others* (2010) use a version of $(2D)^2$PCA (Zhang and Zhou, 2005). Hung *and others* (2012) show

Table 3. *The leave-one-out classification accuracies of MV-logistic/logistic regression for the EEG Database Data Set under different combinations of* (p_0, q_0)

	q_0		
p_0	15	20	30
15	**0.861**/0.803	0.836/0.795	0.844/0.787
30	0.844/0.779	0.828/0.779	0.828/0.754
60	0.844/0.737	0.853/0.713	0.828/0.713

Table 4. *The leave-one-out classification accuracies of PCA followed by conventional logistic regression for the EEG Database Data Set under different choices of r*

r	1	2	3	4	5	6	7	8	9	10
Accuracy	0.631	0.648	0.730	**0.820**	**0.820**	0.803	0.787	0.779	0.787	0.795
r	20	30	40	50	60	70	80	90	100	120
Accuracy	0.746	0.779	0.795	0.754	0.738	0.746	0.713	0.705	0.631	0.492

that GLRAM is asymptotically more efficient than $(2D)^2$PCA in extracting bases, and hence it is reasonable for GLRAM to produce a better result. Second, standardization in Step 2 makes the EEG Database Data Set more suitable to fit the 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 the MV-logistic regression model. We remind the readers again that standardization of covariates will result in a different MV-logistic regression model.

For comparison, the p_0q_0 extracted covariates in Step 1 are also fitted with the conventional (penalized) logistic regression of Le Cessie and Van Houwelingen (1992), where the tuning parameter is selected in the same way such that the leave-one-out classification accuracy is maximized. Table 3 provides the analysis results of both methods under different choices of (p_0, q_0). One can see that MV-logistic regression uniformly outperforms conventional logistic regression. Moreover, the classification accuracy of the conventional approach decays rapidly as the numbers of p_0 and q_0 increase, while those of MV-logistic regression remain roughly constant. As mentioned previously, MV-logistic regression requires fewer parameters in model fitting and possesses certain robustness against model violation and, hence, an efficiency gain is reasonably expected.

Remark 5.1. We also compare our approach with the widely used procedure, principal component analysis (PCA) followed by logistic regression. In particular, PCA is applied to the vectorized covariates $\mathsf{vec}(X_i^*)$. The leading r principal components are then used to fit the conventional (penalized) logistic regression model, where the tuning parameter is also selected to maximize the leave-one-out classification accuracy (see Table 4 for the results). It can be seen that this widely applied approach cannot produce a classification accuracy higher than 0.820, while the best result from fitting MV-logistic regression is 0.861. It reveals the limitation of the conventional $\mathsf{vec}(X)$-based approach, which usually produces a large number of parameters and, hence, poor performance.

6. Conclusions

In this work, we propose the MV-logistic regression model, when the covariates have a natural matrix structure. Its performance is validated through simulation studies and the EEG Database Data Set. The

superiority of the MV-logistic regression model comes from the fact that it requires fewer parameters in model fitting and aims to estimate the best rank-1 approximation of the true parameter matrix. It is also found in our simulation studies that MV-logistic regression has certain robustness against the violation of model specification. Thus, MV-logistic regression model can be used as a good "working" model, especially when the sample size is small. Although we focus on binary response Y, extension of MV-logistic regression model to the case of multiple classes is straightforward (see supplementary material available at *Biostatistics* online).

Besides the EEG example, tensor objects are frequently encountered in many applications, such as mammography images, ultrasound images, and magnetic resonance imaging (MRI) images, which are examples of order-two tensors (matrices). Recently, tensor learning has attracted the attention of statisticians, although it has been considered in the field of computer science and shown to be effective through numerical studies. For order-two tensor X, the main idea of a rank-1 tensor learning is to incorporate X into the learning process through $\alpha^{\mathrm{T}} X\beta$, instead of $\sum_{ij} \xi_{ij} X_{(i,j)}$. Many learning methods have been generalized to adapt to tensor inputs. For unsupervised tensor learning, for example, GLRAM (Ye, 2005) is a tensorial version of principal component analysis (PCA). For supervised tensor learning, the two-dimensional linear discriminant analysis (2D-LDA) (Ye *and others*, 2004) and support tensor machines (STM) (Cai *and others*, 2006) are tensorial versions of linear discriminant analysis (LDA) and support vector machines (SVM), respectively. In view of this point, our MV-logistic regression model can be treated as the tensorial extension of the conventional logistic regression model. We refer the readers to Tao *and others* (2007) for an introduction of supervised tensor learning methods.

Existing works of tensor learning methods, however, focus mainly on extending traditional approaches to incorporating tensorial covariate, but deal less with the corresponding physical interpretations and statistical properties. An important contribution of this study is that, besides model building, we also provide the corresponding statistical justifications. Firstly, we show that the rationale behind the MV-logistic regression model is to search the best rank-1 approximation of the true parameter matrix in the sense of minimum KL-divergence. This fact ensures the tensor model to maintain its performance while using fewer parameters in model fitting. Secondly, we derive the asymptotic normality of the proposed estimators, which facilitates the subsequent statistical inference procedures. To the best of our knowledge, these points were not mentioned in tensor learning literature.

In this paper, the discussion of the MV-logistic regression model is focused only on the case of rank-1 approximation owing to the simplicity in implementation and theoretical development, and we find this simple rank-1 model is suitable for the EEG Database Data Set. Of course we may encounter situations where rank-1 approximation is not adequate. It is therefore natural to extend this rank-1 structure to a more general rank-r setting. In particular, for any positive integer $r \leqslant p \wedge q$, consider the model

$$\mathrm{logit}\{P(Y=1|X)\} = \gamma + \mathsf{vec}(AB^{\mathrm{T}})^{\mathrm{T}}\mathsf{vec}(X), \tag{6.1}$$

where $A \in \mathbb{R}^{p\times r}$ and $B \in \mathbb{R}^{q\times r}$ are the row and column parameter matrices, respectively. Note that model (6.1) can be treated as the conventional logistic regression model (1.1) with the constraint $\xi = AB^{\mathrm{T}}$, and will reduce to our MV-logistic regression model (2.1) when $r = 1$. For the general rank-r setting, however, some issues need to be further studied, such as the problem of identifiability of parameters in (A, B) and the corresponding asymptotic properties. Moreover, the algorithm developed in this paper cannot be directly applied and also need to be revised to adapt to model (6.1). We believe that model (6.1) has its wide applicability in practice and is of great interest for future studies.

SOFTWARE

Software in the form of Matlab code is available on request from the corresponding author.

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

We thank the editor, associate editor, and two anonymous referees for valuable comments that substantially improved the paper, and thank Dr Su-Yun Huang from Academia Sinica for providing helpful ideas. The research of Hung Hung is supported by the National Science Council of Taiwan (NSC 100-2118-M-002-002-). *Conflict of Interest*: None declared.

References

Bickel, P. J. and Doksum, K. A. (2001). *Mathematical Statistics: Basic Ideas and Selected Topics Vol. I*, 2nd edition. Englewood Cliffs, NJ: Prentice-Hall.

Cai, D., He, X. and Han, J. (2006). Learning with tensor representation. Technical report, Computer Science Department, UIUC, UIUCDCS-R-2006-2716.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

Green, P. J. (1984). Iteratively reweighted least squares for likelihood estimation, and some robust and resistant alternatives (with discussion). *Journal of the Royal Statistical Society* **46**, 149–192.

Hastie, T., Tibishirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edition. Springer Series in Statistics. Berlin: Springer.

Hung, H., Wu, P. S., Tu, I. P. and Huang, S. Y. (2012). On multilinear principal component analysis of order-two tensors. *Biometrika*, in press, doi: 10.1093/biomet/ass019.

Le Cessie, S. and Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Journal of the Royal Statistical Society* **41**, 191–201.

Lee, A. and Silvapulle, M. (1988). Ridge estimators in logistic regression. *Communications in Statistics, Simulation and Computation* **17**, 1231–1257.

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.

Lu, W. S., Pei, S. C. and Wang, P. H. (1997). Weighted low-rank approximation of general complex matrices and its application in the design of 2-D digital filters. *IEEE Transactions on Circuits and Systems-I* **44**, 650–655.

Manton, J. H., Mahony, R. and Hua, Y. (2003). The geometry of weighted low-rank approximations. *IEEE Transactions on Signal Processing* **51**, 500–514.

McCulloch, C. E., Searle, S. R. and Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models*. New York, NY: Wiley.

Park, M. Y. and Hastie, T. (2008). Penalized logistic regression for detecting gene interactions. *Biostatistics* **9**, 30–50.

Tao, D., Li, X., Hu, W. and Maybank, S. J. (2007). Supervised tensor learning. *Knowledge and Information Systems* **13**, 1–42.

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

Ye, J., Janardan, R. and Li, Q. (2004). Two-dimensional linear discriminant analysis. *Advances in Neural Information* **17**, 1569–1576.

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.

ZHU, J. AND HASTIE, T. (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics* **5**, 427–443.

[*Received May 18, 2011; revised May 28, 2012; accepted for publication May 28, 2012*]