The Annals of Applied Statistics
2010, Vol. 0, No. 00, 1–26
DOI: 10.1214/10-AOAS327

SPARSE LOGISTIC PRINCIPAL COMPONENTS ANALYSIS FOR BINARY DATA

BY SEOKHO LEE, JIANHUA Z. HUANG[1] AND JIANHUA HU[2]

Harvard School of Public Health, Texas A&M University and University of Texas M. D. Anderson Cancer Center

We develop a new principal components analysis (PCA) type dimension reduction method for binary data. Different from the standard PCA which is defined on the observed data, the proposed PCA is defined on the logit transform of the success probabilities of the binary observations. Sparsity is introduced to the principal component (PC) loading vectors for enhanced interpretability and more stable extraction of the principal components. Our sparse PCA is formulated as solving an optimization problem with a criterion function motivated from a penalized Bernoulli likelihood. A Majorization–Minimization algorithm is developed to efficiently solve the optimization problem. The effectiveness of the proposed sparse logistic PCA method is illustrated by application to a single nucleotide polymorphism data set and a simulation study.

1. Introduction. Principal components analysis (PCA) is a widely used method for dimensionality reduction, feature extraction and visualization of multivariate data. Several sparse PCA methods have recently been introduced to improve the standard PCA [e.g., Jolliffe, Trendafilov and Uddine (2003); Zou, Hastie and Tibshirani (2006); Shen and Huang (2008)]. By requiring the principal component loading vectors to be sparse, sparse PCA methods yield PCs that are more easily interpretable. Sparsity also regularizes the extraction of PCs and thus makes the extraction more stable. Such stability is much desired when the dimension is high, especially in the so-called high-dimension low-sample-size settings. As extensions of the standard PCA, however, these sparse PCA methods are mostly suitable to variables of a continuous type, they are not generally appropriate for other data types such as binary data or counts. Although the basic objective of PCA, or its sparse version, can be achieved regardless of the nature of the original

Received January 2009; revised January 2010.

[1]Supported in part by Grants from the National Science Foundation (DMS-06-06580, DMS-09-07170), the National Cancer Institute (CA57030), the Virtual Center for Collaboration between Statisticians in the US and China, and King Abdullah University of Science and Technology (KAUST, Award KUS-CI-016-04).

[2]Supported in part by Grants from the National Science Foundation (DMS-07-06818) and the National Institute of Health (R01-RGM080503A, R21-CA129671).

Key words and phrases. Binary data, dimension reduction, MM algorithm, LASSO, PCA, regularization, sparsity.

variable, it is true that variances and covariances have especial relevance for multivariate Gaussian variables, and that linear functions of binary variables are less readily interpretable than linear functions of continuous variables [Jolliffe (2002)]. The goal of this paper is to develop a sparse PCA method for binary data.

There are two commonly used definitions of PCA that give rise to the same result. PCA can be defined by finding the orthogonal projection of the data onto a low dimensional linear subspace such that the variance of the projected data is maximized [Hotelling (1933)]. Alternatively, PCA can also be defined by finding the linear projection that minimizes the mean squared distance between the data points and their projections [Pearson (1901)]. Shen and Huang (2008) developed their sparse PCA method following the viewpoint of Pearson. Suppose $\mathbf{y}_1, \ldots, \mathbf{y}_n \in \mathbb{R}^d$ are the n data points and consider a k-dimensional ($k < d$) linear manifold spanned by a bases $\tilde{\mathbf{b}}_1, \ldots, \tilde{\mathbf{b}}_k$ with a shift vector $\boldsymbol{\mu}$. According to Pearson, the PCA minimizes the following reconstruction error,

$$\sum_{i=1}^{n} \|\mathbf{y}_i - (\boldsymbol{\mu} + a_{i1}\tilde{\mathbf{b}}_1 + \cdots + a_{ik}\tilde{\mathbf{b}}_k)\|^2, \tag{1.1}$$

subject to the constraint that $\mathbf{A} = (a_{ij})$ has orthonormal columns. Usually the variables presented in $\mathbf{y}_i$ are scaled so that they have the same order of magnitude. Note that (1.1) is a least squares regression if a_{ik}'s were known. In light of this connection to regression and borrowing the idea from LASSO [Tibshirani (1996)], Shen and Huang (2008) proposed to add an L_1 penalty $\|\tilde{\mathbf{b}}_1\|_1 + \cdots + \|\tilde{\mathbf{b}}_k\|_1$ to the reconstruction error (1.1) to obtain sparse loading vectors $\tilde{\mathbf{b}}_1, \ldots, \tilde{\mathbf{b}}_k$. Since the reconstruction error (1.1) can be viewed as the negative log likelihood up to a constant for the Gaussian distributions with mean vectors $\boldsymbol{\theta}_i = \boldsymbol{\mu} + a_{i1}\tilde{\mathbf{b}}_1 + \cdots + a_{ik}\tilde{\mathbf{b}}_k$ for $i = 1, \ldots, n$ and identity covariance, the method of Shen and Huang can be interpreted as a penalized likelihood approach for the sparse PCA. The key idea of the current paper is to replace the Gaussian likelihood by the Bernoulli likelihood where $\boldsymbol{\theta}_i$ will be the logit transform of the success probabilities. We refer to the proposed PCA method as sparse logistic PCA. The relationship of the proposed sparse logistic PCA to the sparse PCA of Shen and Huang is analogous to the relationship between logistic and linear LASSO regression.

We develop an iterative weighted least squares algorithm to perform the proposed sparse logistic PCA. Since the log Bernoulli likelihood is not quadratic and the L_1 penalty function is nondifferentiable, the optimization problem defining the sparse logistic PCA is not straightforward to solve. Our algorithm applies the general idea of optimization transfer or Majorization–Minimization (MM) algorithm [Lange, Hunter and Yang (2000); Hunter and Lange (2004)]. By iteratively replacing the complex objective function with suitably defined quadratic surrogates, each step of our algorithm solves a weighted least squares problem and has closed form. The algorithm is easy to implement and guaranteed at each iteration to improve the penalized PCA log-likelihood. We show that the same MM algorithm

is applicable when there are missing data. We also develop a method for choosing the penalty parameters and for choosing the number of important principal components. PCA of binary data using Bernoulli likelihood has previously been studied by Collins, Dasgupta and Schapire (2001), Schein, Saul and Ungar (2003) and de Leeuw (2006), but none of these works considered sparse loading vectors. As we demonstrate using simulation and real data, sparsity can enhance interpretation of results and improve the stability and accuracy of the extracted principal components.

Other approaches of sparse PCA are not as easily extendible to binary data. Jolliffe, Trendafilov and Uddine (2003) modified the defining maximum variance problem of the standard PCA by applying an L_1-norm constraint on the PC loading vectors to obtain PCA with sparse loadings. Its use of sample variance makes it unappealing for binary data. Zou, Hastie and Tibshirani (2006) rewrote PCA as a regression-type optimization problem and then applied the LASSO penalty [Tibshirani (1996)] to obtain sparse loadings. However, since the data appear both as regressors and responses in their regression-type problem, the connection of their approach to the penalized likelihood is not as natural as Shen and Huang (2008).

The rest of this article is organized as follows. In Section 2 we introduce the optimization problem that yields the sparse logistic PCA and provides methods for tuning parameter selection. Section 3 applies the sparse logistic PCA to a single nucleotide polymorphism data set and compares it with the nonsparse version of logistic PCA. Section 4 presents a Majorization–Minimization algorithm for efficient computation of the sparse logistic PCA and Section 5 discusses how to handle missing data. Results of a simulation study are given in Section 6. Section 7 concludes the paper with some discussion. An Appendix contains proofs of theorems.

2. Sparse logistic PCA with penalized likelihood.

2.1. *Penalized Bernoulli likelihood.* Consider the $n \times d$ binary data matrix $\mathbf{Y} = (y_{ij})$, each row of which represents a vector of observations from binary variables. We assume that entries of $\mathbf{Y}$ are realizations of mutually independent random variables and that y_{ij} follows the Bernoulli distribution with success probability π_{ij}. Let $\theta_{ij} = \log\{\pi_{ij}/(1-\pi_{ij})\}$ be the logit transformation of π_{ij}. Define the inverse logit transformation $\pi(\theta) = \{1 + \exp(-\theta)\}^{-1}$. Then the success probabilities can be represented using the canonical parameters as $\pi_{ij} = \pi(\theta_{ij})$. The individual data generating probability becomes

$$\Pr(Y_{ij} = y_{ij}) = \pi(\theta_{ij})^{y_{ij}}\{1 - \pi(\theta_{ij})\}^{1-y_{ij}} = \pi(q_{ij}\theta_{ij}),$$

with $q_{ij} = 2y_{ij} - 1$ since $\pi(-\theta) = 1 - \pi(\theta)$. This representation leads to the compact form of the log likelihood as

$$\ell = \sum_{i=1}^{n}\sum_{j=1}^{d} \log \pi(q_{ij}\theta_{ij}). \tag{2.1}$$

Note that the Bernoulli distributions are in the exponential family and θ_{ij} are the corresponding canonical parameters.

To build a probabilistic model for principal components analysis of binary data, the d-dimensional canonical parameter vectors $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{id})^T$ are constrained to reside in a low dimensional manifold of $\mathbb{R}^d$ with the dimensionality k. (The choice of k will be discussed later in Section 2.3.) Specifically, we assume that, for some vectors $\boldsymbol{\mu}, \tilde{\mathbf{b}}_1, \ldots, \tilde{\mathbf{b}}_k \in \mathbb{R}^d$, the vector of canonical parameters satisfies $\boldsymbol{\theta}_i = \boldsymbol{\mu} + a_{i1}\tilde{\mathbf{b}}_1 + \cdots + a_{ik}\tilde{\mathbf{b}}_k$ for $i = 1, \ldots, n$. We call $\tilde{\mathbf{b}}_1, \ldots, \tilde{\mathbf{b}}_k$ the principal component loading vectors and the coefficients $\mathbf{a}_i = (a_{i1}, \ldots, a_{ik})^T$ the principal component scores (PC scores) for the ith observation. Geometrically, the vectors of canonical parameters $\boldsymbol{\theta}_i$ are projected onto the k-dimensional manifold which is the affine subspace spanned by k PC loading vectors and translated by the intercept vector $\boldsymbol{\mu}$. In matrix form, the canonical parameter matrix $\boldsymbol{\Theta} = (\theta_{ij}) = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)^T$ is represented as

$$\boldsymbol{\Theta} = \mathbf{1}_n \otimes \boldsymbol{\mu}^T + \mathbf{A}\mathbf{B}^T, \tag{2.2}$$

where $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_n)^T$ is the $n \times k$ principal component score matrix and $\mathbf{B} = (\tilde{\mathbf{b}}_1, \ldots, \tilde{\mathbf{b}}_k)$ is the $p \times k$ principal component loading matrix. For identifiability purpose, we require that $\mathbf{A}$ has orthonormal columns.

We target a method that can produce a sparse loading matrix, a loading matrix with many zero elements. A sparse loading matrix implies variable selection in principal components analysis, since each principal component only involves those variables corresponding to the nonzero elements of the loading vector. We propose to perform variable selection using the penalized likelihood with a sparsity inducing penalty. Let $\mathbf{b}_j^T$ denote the jth row of $\mathbf{B}$. Then (2.2) implies that $\theta_{ij} = \mu_j + \mathbf{a}_i^T\mathbf{b}_j$ where μ_j is the jth element of $\boldsymbol{\mu}$. The log likelihood can be written as

$$\ell(\boldsymbol{\mu}, \mathbf{A}, \mathbf{B}) = \sum_{j=1}^{d}\sum_{i=1}^{n} \log \pi\{q_{ij}(\mu_j + \mathbf{a}_i^T\mathbf{b}_j)\}. \tag{2.3}$$

If $\mathbf{a}_i$ were observable, (2.3) is the log likelihood for d logistic regressions

$$\operatorname{logit} P(Y_{ij} = 1) = \mu_j + \mathbf{a}_i^T\mathbf{b}_j.$$

This connection with logistic regression suggests use of the L_1 penalty to get a sparse loading matrix, as in the LASSO regression [Tibshirani (1996)].

Specifically, consider the penalty

$$P_{\boldsymbol{\lambda}}(\mathbf{B}) = \sum_{l=1}^{k} \lambda_l \|\tilde{\mathbf{b}}_l\|_1 = \lambda_1 \sum_{j=1}^{d} |b_{j1}| + \cdots + \lambda_k \sum_{j=1}^{d} |b_{jk}|, \tag{2.4}$$

where λ_l are regularization parameters whose selection will be discussed later. We obtain sparse principal components by maximizing the following penalized log

likelihood:

$$f(\boldsymbol{\mu}, \mathbf{A}, \mathbf{B}) = \ell(\boldsymbol{\mu}, \mathbf{A}, \mathbf{B}) - nP_{\lambda}(\mathbf{B}), \tag{2.5}$$

subject to the constraint that $\mathbf{A}$ has orthonormal columns. Note that $\mathbf{B}$ enters the likelihood together with $\mathbf{A}$ through $\mathbf{AB}^T$ and so $\mathbf{B}$ can be arbitrarily small by just increasing the magnitude of $\mathbf{A}$ and not changing the likelihood. The orthonormal constraint on $\mathbf{A}$ prevents elements of $\mathbf{A}$ becoming arbitrarily large and thus validates our use of the L_1 penalty on $\mathbf{B}$.

The sparse principal components can be equivalently formulated as minimizing the following criterion function:

$$S(\boldsymbol{\mu}, \mathbf{A}, \mathbf{B}) = -\ell(\boldsymbol{\mu}, \mathbf{A}, \mathbf{B}) + nP_{\lambda}(\mathbf{B}), \tag{2.6}$$

subject to the constraint that $\mathbf{A}$ has orthonormal columns. In (2.6) the negative log likelihood can be interpreted as a loss function and the L_1 penalties increase the loss for nonzero elements of $\mathbf{B}$ according to their magnitude. This penalized loss interpretation is also appealing in the sense that the independent Bernoulli trials assumption for obtaining the likelihood (2.3) need not be a realistic representation of the actual data generating process but rather a device for generating a suitable loss function. Since the L_1 penalties regularize the loss minimization, the sparse logistic PCA is sometimes also referred to as the regularized logistic PCA. We shall focus on the minimization problem (2.6) for the rest of the paper. A computational algorithm for solving the minimization problem is presented in Section 4.

The effectiveness of the proposed sparse logistic PCA is illustrated in Figure 1 using a rank-one model (i.e., $k = 1$). While the sparse logistic PCA can recover the original loading vector well, the nonregularized logistic PCA gives more noisy results. A systematic simulation study is reported in Section 6.

2.2. *Choosing the penalty parameters.* Although different penalty parameters can be used for different PC loading vectors for maximal flexibility of the methodology, we consider using only a single penalty parameter λ for all PC loadings. This simplification substantially reduces the computation time, especially when k is large. Note that a larger value of λ will lead to a smaller number of nonzeros in the loading matrix $\mathbf{B}$ and reduced model complexity, but the reduced model complexity is usually associated with less good fit of the model. To compromise the goodness of fit and model complexity, for fixed k, we choose λ by minimizing the following BIC criterion:

$$\mathrm{BIC}(\lambda) = -2\ell(\boldsymbol{\mu}, \mathbf{A}, \mathbf{B}) + \log n \times m(\lambda), \tag{2.7}$$

where $m(\lambda)$ is a measure of the degrees of freedom. Note that Zou, Hastie and Tibshirani (2007) showed that the number of nonzero coefficients is an unbiased estimate of the degrees of freedom for the LASSO regression. The degrees of freedom $m(\lambda)$ used in (2.7) is defined as $m(\lambda) = d + nk + |\mathcal{B}(\lambda)|$, where d is the length of the vector $\boldsymbol{\mu}$, nk is the total number of elements of $\mathbf{A}$, and $|\mathcal{B}(\lambda)|$ is the cardinality of the index set $\mathcal{B}(\lambda)$ of the nonzero loadings in $\mathbf{B}$ when the penalty parameter is λ. We use a grid search to find the optimal λ that minimizes the BIC.

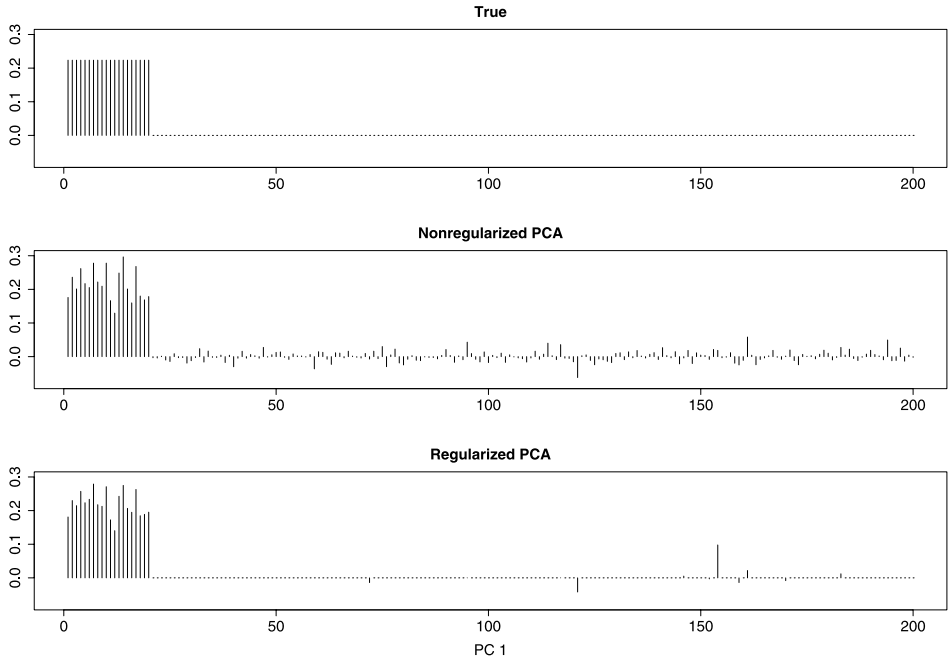

FIG. 1. *A simulated data set with* $n = 100$, $d = 200$, *and* $k = 1$. *Top, middle and bottom panels show respectively the true loadings, loadings from the nonregularized logistic PCA and from the regularized logistic PCA. The penalty parameter is selected using the BIC.*

2.3. *Determining the dimensionality of the subspace.* The BIC criterion defined in (2.7) can also be used to select a suitable "k." A two-dimensional grid search can be used to find the minimizer of the BIC with respect to both k and λ. To expedite computation, we implement the following strategy: First fix k at a reasonable large value and select a good λ, then using this λ we refine the choice of k and, finally, we refine λ with the refined k. When optimizing with respect to λ, a coarse grid can be used in the first step and a finer grid in the second step. Our simulation study showed that this strategy works reasonably well (see Section 6.3).

REMARK 1. In classical multivariate analysis, the percentage of total variance explained by the principal components provides an intuitive measure that can be used for subjectively choosing the appropriate number of principal components. Zou, Hastie and Tibshirani (2006) and Shen and Huang (2008) extended it to sparse PCA by modifying the definition of variance explained by the PCs. Since there is no clear definition of total variance for the binary data, extension of the notion of "percentage of variance explained" to logistic PCA is an interesting but unsolved problem.

3. Application to single nucleotide polymorphism data. Association studies based on high-throughput single nucleotide polymorphism (SNP) data [Brookes (1999); Kwok et al. (1996)] have become a popular way to detect genomic regions associated with human complex diseases. A SNP is a single base pair position in genomic DNA at which the sequence (alleles) variation occurs between members of a species, wherein the least frequent allele has an abundance of 1% or greater. A crucial issue in association studies is population stratification detection [Hao et al. (2004)], which is to determine whether a population is homogeneous or has hidden structures within it. With the presence of population stratification, the naive case-control approach not accounting for this factor would yield biased results [Ewens and Spielman (1995)] and, therefore, draw inaccurate scientific conclusions. See Liang and Kelemen (2008) for an extensive discussion of statistical methods and difficulties for SNP data analysis.

The proposed sparse logistic PCA method can be used for population stratification detection. For the purpose of demonstration, we use the SNP data set available in the International HapMap project [The International HapMap Consortium (2005)]. It consists of 3 different ethnic populations of 90 Caucasians (Utah residents with ancestry from northern and western Europe; CEO), 90 Africans (Yoruba in Ibadan, Nigeria; YRI) and 90 Asians (45 Han Chinese in Beijing, China; CHB and 45 Japanese in Tokyo, Japan; JPT). Our task is to detect this three-subpopulation structure using the SNP data on the 270 subjects. At many SNP locations, heterozygosity distribution and allele frequency are known to be different among populations and could confound the effect of the risk of disease. To account for this factor, Serre et al. (2008) selected 1536 SNPs with similar heterozygosity distribution and allele frequency. The locations of these SNPs cover all the chromosomes except for the sex-determining chromosome. Among these 1536 SNPs, 1392 are shared by three ethnic groups, which are used in our analysis. We coded 0 for the most prevalent homogeneous base pair (wild-type) and 1 for others (mutant), resulting in a 270×1392 binary matrix. This data matrix has 2.37% missing entries.

We applied the sparse logistic PCA to this SNP data set to explore variability among high dimensional SNP variables, using the computation algorithm given in Sections 4 and 5 below. The method described in Section 2.3 was used for model selection. Specifically, we initially fixed the reduced dimension to $k = 30$ and chose the penalty parameter λ among the rough grid of $0, 1.5^{-18}, 1.5^{-17}, \ldots, 1.5^{-10}$ using the BIC criterion defined in Section 2.3. Given the selected $\lambda = 1.5^{-16}$, the dimension k was refined by minimizing the BIC, giving $k = 10$. Finally, with $k = 10$, we refined λ by searching over the grid $0, 0.0005, 0.0010, 0.0015, \ldots, 0.0100$, resulting in $\lambda = 0.0015$. As a comparison, we also applied the nonregularized logistic PCA to the data, which corresponds to $\lambda = 0$ in our general formulation of regularized logistic PCA.

To examine which principal components represent the variability associated with three racial groups, we used a F-test where scores for each fixed PC is

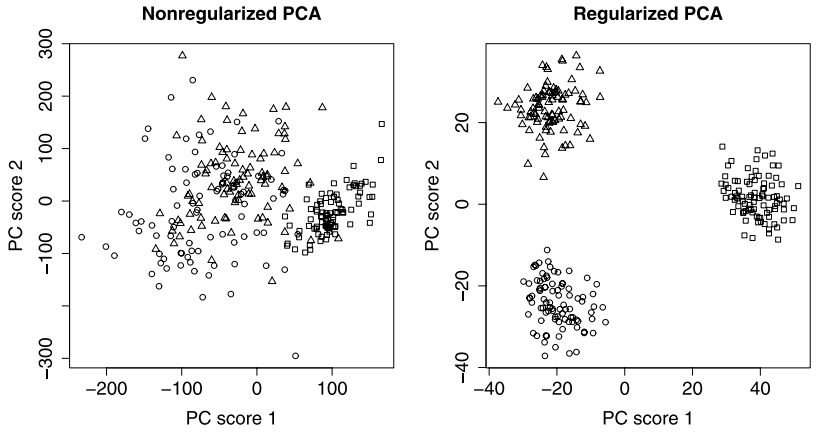

FIG. 2. *The scatterplots of the first two PC scores from the nonregularized* (*left*) *and regularized logistic PCA*. *Circle*, *rectangle and triangle represent Caucasian*, *African and Asian population respectively.*

regressed on the group dummy variables. For the sparse logistic PCA, only the first two PCs were highly significant with both p-values less than 0.0001 and the remaining eight PCs were not significant with large p-values (0.7681, 0.9109, 0.4764, 0.5523, 0.3376, 0.5415, 0.4480, 0.6441 for the third to the tenth PCs respectively). This result suggests that the sparse logistic PCA can effectively compress the racial group information into two leading PCs. Similar compression was not achieved by the nonregularized logistic PCA; the F-test was significant for all the first ten PCs with p-values <0.0001, <0.0001, 0.0002, 0.0001, <0.0001, <0.0001, <0.0001, 0.0028, <0.0001 and 0.0299 respectively.

Pairwise scatterplots were used to check clustering of subjects using the PC scores. Figure 2 shows the scatterplots of first 2 PC scores with and without regularization. The three ethnic groups are clearly separated by the regularized PCA but not by the nonregularized PCA. To verify that the group separation obtained is not because of luck, we permuted observations for each SNP and applied the sparse logistic PCA to the permuted data set; no clear clustering showed up in the PC scores.

The proposed sparse PCA method allows directly identifying the SNPs that contribute to the group separation. The selected model has 790 and 658 nonzero loadings (representing the SNPs) respectively for the first 2 PCs, among which 509 SNPs are shared. Therefore, 939 SNPs involved in the first 2 PC directions are claimed to be associated with the ethnic group effect. Our result suggests that the population stratification factor should be taken into consideration at these 939 SNP locations in the subsequent study of the association between SNPs and the disease phenotype to avoid biased conclusion. Although in light of our simulation results, some selected SNPs could be false positives, we believe that a large proportion of the selected SNPs are relevant in differentiation among the three racial groups,

because the studied SNPs were delicately selected to represent the most genetic diversity of the whole genome [Serre et al. (2008)] and the genetic differentiation is the greatest when defined on a continental basis, which is the case for our comparison between Caucasian, Asian and African [Risch et al. (2002)].

We further compared the regularized and nonregularized logistic PCA by assessing the variability of the probability estimates using the parametric bootstrap. For each method, we generated 100 bootstrapped data sets of binary matrices; each binary matrix has entries that are independently drawn from the Bernoulli distribution with success probability $\hat{\pi}_{ij}$ for the (i, j)th entry, where $\hat{\pi}_{ij}$ is the estimated probability. We then applied the method to these bootstrapped data sets to obtain 100 bootstrapped probabilities for each (i, j) combination and to construct a 90% variability interval using the 5% and 95% quantiles of the bootstrapped probabilities. These 90% variability intervals were plotted against the ordered $\hat{\pi}_{ij}$ to form a variability envelop. The variability envelop for the regularized PCA is narrower than that for the nonregularized PCA, indicating that regularization indeed reduces the variability of the probability estimates (Figure 3).

Our working model for the logistic PCA specified by (2.1) and (2.2) assumes that, conditional on the principal component scores, the observations are independent. Since there exists spatial dependency among SNPs, one may have concerns about the validity of our analysis results if the dependence is strong. In our data set, the 1536 SNPs were selected from the whole genome to capture most of the genetic diversity in population considering factors of physical distances, allele frequencies and linkage disequilibrium patterns. The selected SNPs are sufficiently well separated within each chromosome so that they can be representative of the whole genome [Serre et al. (?Serre07)]. Therefore, we expect that the spatial dependency in this data set should not be too serious to invalidate our results. To

<ref:Serre07?>

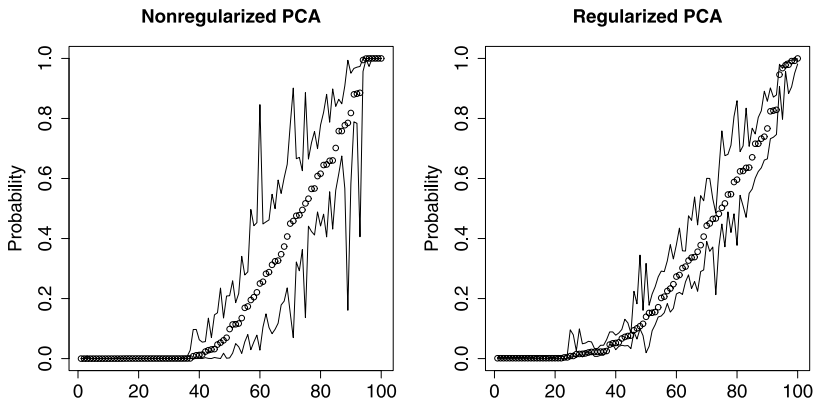

FIG. 3. *The SNP data*: 90% *bootstrap variability envelope* (*showed as lines*) *of the probability estimates*, *using* 100 *randomly selected SNPs. Circles are the estimated probabilities* $\hat{\pi}_{ij}$ *from the SNP data. Results are based on* 100 *bootstrap samples*.

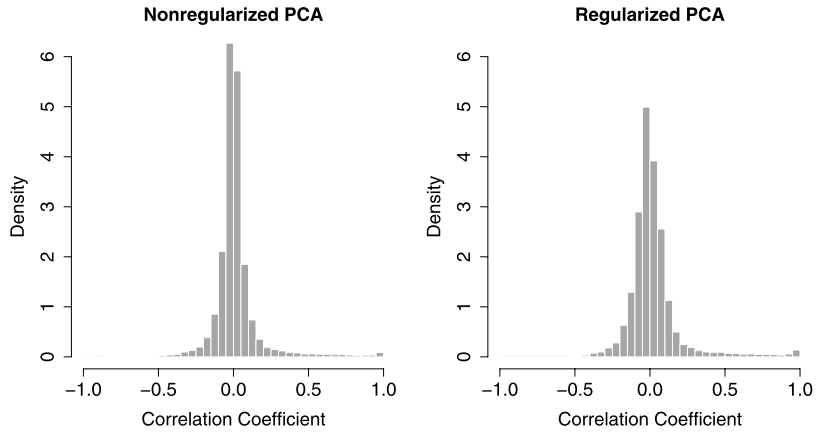

FIG. 4. *Histograms of pairwise correlations of Pearson's residuals from nonregularized* (*left*) *and regularized* (*right*) *logistic PCA*.

address this issue empirically, we first computed Pearson's residuals after fitting the models for the nonregularized and regularized logistic PCA, then calculated pairwise correlations of these Pearson's residuals for all SNP pairs for each chromosome. Figure 4 shows the histogram of the pairwise correlations for each model. For both models most pairwise correlations are close to zero, indicating that the SNPs are weakly correlated. We noticed that there exists a very small proportion of SNP pairs that are highly correlated. Examination of the physical locations revealed that those highly correlated SNP pairs consist of SNPs in close vicinity, indicating the imperfection of the initial SNP selection process.

4. Computational algorithm. We develop a Majorization–Minimization (MM) algorithm for minimizing (2.6), which iteratively minimizes a suitably defined quadratic upper bound of (2.6). Instead of directly dealing with the nonquadratic log likelihood and the nondifferentiable sparsity inducing L_1 penalty, the MM algorithm sequentially optimizes a quadratic surrogate objective function. A function $g(x|y)$ is said to majorize a function $f(x)$ at y if

$$g(x|y) \geq f(x) \qquad \text{for all } x \quad \text{and} \quad g(y|y) = f(y).$$

In the geometrical view the function surface $g(x|y)$ lies above the function $f(x)$ and is tangent to it at the point y so $g(x|y)$ becomes an upper bound of $f(x)$. To minimize $f(x)$, the MM algorithm starts from an initial guess $x^{(0)}$ of x, and iteratively minimizes $g(x|x^{(m)})$ until convergence, where $x^{(m)}$ is the estimate of x at the mth iteration. The MM algorithm decreases the objective function in each step and is guaranteed to converge to a local minimum of $f(x)$. When applying the MM algorithm, the majorizing function $g(x|y)$ is chosen such that it is easier to minimize than the original objective function $f(x)$. See Hunter and Lange (2004) for an introductory description of the MM algorithm.

To find a suitable majorizing function of (2.6), we treat the log likelihood term and the penalty term separately. For the log likelihood term, note that, for a given point y,

$$-\log\pi(x) \le -\log\pi(y) - \{1-\pi(y)\}(x-y) + \frac{2\pi(y)-1}{4y}(x-y)^2 \tag{4.1}$$

$$\le -\log\pi(y) - \{1-\pi(y)\}(x-y) + \frac{1}{8}(x-y)^2, \tag{4.2}$$

and the equalities hold when $x = y$ [Jaakkola and Jordan (2000); de Leeuw (2006)]. These inequalities provide quadratic upper bounds for the negative log inverse logit function at the tangent point y. We refer to the former bound as the tight bound, and the latter bound as the uniform bound since its curvature does not change with y. We pursue here the MM algorithm by using the uniform bound and leave the discussion of using the tight bound to the supplemental article [Lee, Huang and Hu (2010)]. Use of the tight bound usually leads to a smaller number of iterations of the algorithm but longer computation time because of the complexity involved in computing the bound. For the penalty term, the inequality

$$|x| \le \frac{x^2+y^2}{2|y|}, \qquad y \ne 0, \tag{4.3}$$

gives an upper bound for $|x|$ and the equality holds when $x = y$ [Hunter and Li (2005)]. Application of (4.2) and (4.3) yields a suitable majorizing function of (2.6) and thus an MM algorithm.

Now we present details of the MM algorithm via the uniform bound. Let $\mathbf{\Theta}^{(m)}$ be the estimate of $\mathbf{\Theta}$ obtained in the mth step of the algorithm, with the entries $\theta_{ij}^{(m)} = \mu_j^{(m)} + \mathbf{a}_i^{(m)T}\mathbf{b}_j^{(m)}$. By completing the square, the uniform bound (4.2) can be rewritten as

$$-\log\pi(x) \le -\log\pi(y) + \tfrac{1}{8}[x-y-4\{1-\pi(y)\}]^2. \tag{4.4}$$

Substituting x and y with $q_{ij}\theta_{ij}$ and $q_{ij}\theta_{ij}^{(m)}$ respectively in (4.4) and noticing that $q_{ij} = \pm 1$, we obtain

$$-\log\pi(q_{ij}\theta_{ij}) \le -\log\pi\bigl(q_{ij}\theta_{ij}^{(m)}\bigr) + w_{ij}^{(m)}\bigl(\theta_{ij} - x_{ij}^{(m)}\bigr)^2, \tag{4.5}$$

where $w_{ij}^{(m)} = 1/8$ and

$$x_{ij}^{(m)} = \theta_{ij}^{(m)} + 4q_{ij}\bigl\{1-\pi\bigl(q_{ij}\theta_{ij}^{(m)}\bigr)\bigr\}. \tag{4.6}$$

The superscript m of $w_{ij}^{(m)}$ and $x_{ij}^{(m)}$ indicates the dependence on $\mathbf{\Theta}^{(m)}$. Summing over all i, j of (4.5) and ignoring a constant term that does not depend on unknown parameters, we obtain the following quadratic upper bound of the negative log-likelihood:

$$\sum_{i=1}^{n}\sum_{j=1}^{d} w_{ij}^{(m)}\bigl(\theta_{ij} - x_{ij}^{(m)}\bigr)^2 = \sum_{i=1}^{n}\sum_{j=1}^{d} w_{ij}^{(m)}\bigl\{x_{ij}^{(m)} - (\mu_j + \mathbf{a}_i^T\mathbf{b}_j)\bigr\}^2. \tag{4.7}$$

On the other hand, (4.3) implies that the penalty $P_{\boldsymbol{\lambda}}(\mathbf{B})$ has the following quadratic upper bound:

$$P_{\boldsymbol{\lambda}}(\mathbf{B}) \le \lambda_1 \sum_{j=1}^{d} \frac{b_{j1}^2 + b_{j1}^{(m)2}}{2|b_{j1}^{(m)}|} + \cdots + \lambda_k \sum_{j=1}^{d} \frac{b_{jk}^2 + b_{jk}^{(m)2}}{2|b_{jk}^{(m)}|}. \tag{4.8}$$

Combining (4.7) and (4.8) yields the following quadratic upper bound (up to a constant) of the criterion function $S(\boldsymbol{\mu}, \mathbf{A}, \mathbf{B})$ defined in (2.6):

$$\begin{aligned} &g\big(\boldsymbol{\mu}, \mathbf{A}, \mathbf{B} | \boldsymbol{\mu}^{(m)}, \mathbf{A}^{(m)}, \mathbf{B}^{(m)}\big) \\ &\quad = \sum_{i=1}^{n} \sum_{j=1}^{d} \big[w_{ij}^{(m)} \big\{ x_{ij}^{(m)} - (\mu_j + \mathbf{a}_i^T \mathbf{b}_j) \big\}^2 + \mathbf{b}_j^T \mathbf{D}_{\boldsymbol{\lambda},j}^{(m)} \mathbf{b}_j \big], \end{aligned} \tag{4.9}$$

where $\mathbf{D}_{\boldsymbol{\lambda},j}^{(m)}$ is a diagonal matrix with diagonal elements $\lambda_l / \{2|b_{jl}^{(m)}|\}$ for $l = 1, \ldots, k$.

THEOREM 4.1. (i) *Up to a constant that depends on* $\boldsymbol{\mu}^{(m)}$, $\mathbf{A}^{(m)}$ *and* $\mathbf{B}^{(m)}$ *but not on* $\boldsymbol{\mu}$, $\mathbf{A}$ *and* $\mathbf{B}$, *the function* $g(\boldsymbol{\mu}, \mathbf{A}, \mathbf{B} | \boldsymbol{\mu}^{(m)}, \mathbf{A}^{(m)}, \mathbf{B}^{(m)})$ *defined in* (4.9) *majorizes* $S(\boldsymbol{\mu}, \mathbf{A}, \mathbf{B})$ *at* $(\boldsymbol{\mu}^{(m)}, \mathbf{A}^{(m)}, \mathbf{B}^{(m)})$.

(ii) *Let* $(\boldsymbol{\mu}^{(m)}, \mathbf{A}^{(m)}, \mathbf{B}^{(m)})$, $m = 1, 2, \ldots$, *be a sequence obtained by iteratively minimizing the majorizing function. Then* $S(\boldsymbol{\mu}^{(m)}, \mathbf{A}^{(m)}, \mathbf{B}^{(m)})$ *decreases as* m *gets larger and it converges to a local minimum of* $S(\boldsymbol{\mu}, \mathbf{A}, \mathbf{B})$ *as* m *goes to infinity.*

The majorizing function given in (4.9) is quadratic in each of $\boldsymbol{\mu}$, $\mathbf{A}$ and $\mathbf{B}$ when the other two are fixed and, thus, alternating minimization of (4.9) with respect to $\boldsymbol{\mu}$, $\mathbf{A}$ and $\mathbf{B}$ has closed-form solutions, which are given below. We now drop the superscript in $x_{ij}^{(m)}$ for notational convenience. Recall that $w_{ij}^{(m)} = 1/8$ is a constant. For fixed $\mathbf{A}$ and $\mathbf{B}$, set $x_{ij}^{\dagger} = x_{ij} - \mathbf{a}_i^T \mathbf{b}_j$, the optimal $\hat{\mu}_j$ is given by

$$\hat{\mu}_j = \arg\min_{\mu_j} \sum_{i=1}^{n} (x_{ij}^{\dagger} - \mu_j)^2 = \frac{1}{n} \sum_{i=1}^{n} x_{ij}^{\dagger}, \qquad j = 1, \ldots, d. \tag{4.10}$$

This leads to a simple matrix formula $\hat{\boldsymbol{\mu}} = \frac{1}{n} \mathbf{X}^{\dagger T} \mathbf{1}_n$, which is obtained by taking the column means of $\mathbf{X}^{\dagger} = (x_{ij}^{\dagger})$.

To update $\mathbf{A}$ and $\mathbf{B}$ for fixed $\boldsymbol{\mu}$, set $x_{ij}^{*} = x_{ij} - \mu_j$ or in matrix form, $\mathbf{X}^{*} = (x_{ij}^{*}) = \mathbf{X} - \mathbf{1}_n \otimes \boldsymbol{\mu}^T$. Denote the ith row vector of $\mathbf{X}^{*}$ as $\mathbf{x}_i^{*T}$. For fixed $\boldsymbol{\mu}$ and $\mathbf{B}$, the ith row of $\mathbf{A}$ is updated by minimizing with respect to $\mathbf{a}_i$ the sum of squares $\sum_{j=1}^{d} (x_{ij}^{*} - \mathbf{a}_i^T \mathbf{b}_j)^2 = (\mathbf{x}_i^{*} - \mathbf{B}\mathbf{a}_i)^T (\mathbf{x}_i^{*} - \mathbf{B}\mathbf{a}_i)$, which has a closed form solution

$$\hat{\mathbf{a}}_i = (\mathbf{B}^T \mathbf{B})^{-1} \mathbf{B}^T \mathbf{x}_i^{*}, \qquad i = 1, \ldots, n, \tag{4.11}$$

or $\hat{\mathbf{A}} = \mathbf{X}^{*} \mathbf{B} (\mathbf{B}^T \mathbf{B})^{-1}$ in matrix form. The columns of updated $\mathbf{A}$ can be made orthonormal by using the QR decomposition. Denote the jth column vector of $\mathbf{X}^{*}$

as $\tilde{\mathbf{x}}_j^*$. For fixed $\boldsymbol{\mu}$ and $\mathbf{A}$, the jth row of $\mathbf{B}$ is updated by solving the ridge regression problem that minimizes with respect to $\mathbf{b}_j$ the penalized sum of squares

$$\frac{1}{8}\sum_{i=1}^{n}(x_{ij}^* - \mathbf{a}_l^T\mathbf{b}_j)^2 + n\sum_{l=1}^{k}\lambda_l\frac{b_{jl}^2}{2|b_{jl}^{(m)}|}$$

$$= \frac{1}{8}(\tilde{\mathbf{x}}_j^* - \mathbf{A}\mathbf{b}_j)^T(\tilde{\mathbf{x}}_j^* - \mathbf{A}\mathbf{b}_j) + n\mathbf{b}_j^T\mathbf{D}_{\boldsymbol{\lambda},j}\mathbf{b}_j,$$

which has a closed form solution

$$\hat{\mathbf{b}}_j = (\mathbf{A}^T\mathbf{A} + 8n\mathbf{D}_{\boldsymbol{\lambda},j})^{-1}\mathbf{A}^T\tilde{\mathbf{x}}_j^*, \qquad j = 1,\ldots,d. \tag{4.12}$$

Since, during the iteration, $\mathbf{A}$ is made orthonormal, $\mathbf{A}^T\mathbf{A}$ becomes the identity matrix of size k. Therefore, since the matrices to be inverted are diagonal matrices, $\hat{\mathbf{b}}_j$ can be obtained by component-wise shrinkage

$$\hat{b}_{jl} = \frac{|b_{jl}^{(m)}|}{|b_{jl}^{(m)}| + 4n\lambda_l}\tilde{\mathbf{a}}_l^T\tilde{\mathbf{x}}_j^*, \qquad l = 1,\ldots,k,\ j = 1,\ldots,d,$$

where $\tilde{\mathbf{a}}_l$ is the lth column of $\mathbf{A}$.

The MM algorithm will alternate between (4.10), (4.11) and (4.12) until convergence. The details are summarized in Algorithm 1. In this algorithm, k, the number of columns of $\mathbf{A}$ and $\mathbf{B}$, should be specified in advance. Different from the sequential extraction approach of Shen and Huang (2008), the matrices $\mathbf{A}$ and $\mathbf{B}$ obtained after applying Algorithm 1 depend on the value of k, but the results are reasonably stable when k is large enough. See Section 2.3 for discussion on choice of k. We use random initial values for $\boldsymbol{\mu}$, $\mathbf{A}$ and $\mathbf{B}$. As with any nonlinear optimization algorithms, our algorithm is not guaranteed to converge to a global minimum. We can follow the common practice to random start the algorithm several times and find the best solution. Our experience is that the algorithm with different initial values usually converges to the same solution (within the precision specified by the convergence criterion).

ALGORITHM 1 (Sparse logistic PCA algorithm I).

1. Initialize with $\boldsymbol{\mu}^{(1)} = (\mu_1^{(1)},\ldots,\mu_d^{(1)})^T$, $\mathbf{A}^{(1)} = (\mathbf{a}_1^{(1)},\ldots,\mathbf{a}_n^{(1)})^T$ and $\mathbf{B}^{(1)} = (\mathbf{b}_1^{(1)},\ldots,\mathbf{b}_d^{(1)})^T$. Set $m = 1$.
2. Compute $x_{ij}^{(m)}$ using (4.6) and set $\mathbf{X}^{(m)} = (x_{ij}^{(m)})$.
3. Set $\mathbf{X}^{(m)\dagger} = (x_{ij}^{(m)\dagger})$ with $x_{ij}^{(m)\dagger} = x_{ij}^{(m)} - \mathbf{a}_i^{(m)T}\mathbf{b}_j^{(m)}$. Update $\boldsymbol{\mu}$ using $\boldsymbol{\mu}^{(m+1)} = \frac{1}{n}\mathbf{X}^{(m)\dagger T}\mathbf{1}_n$.
4. Set $\mathbf{X}^{(m+1)*} = \mathbf{X}^{(m)} - \mathbf{1}_n \otimes \boldsymbol{\mu}^{(m+1)T}$.
5. Update $\mathbf{A}$ by $\mathbf{A}^{(m+1)} = \mathbf{X}^{(m+1)*}\mathbf{B}^{(m)}(\mathbf{B}^{(m)T}\mathbf{B}^{(m)})^{-1}$. Compute the QR decomposition $\mathbf{A}^{(m+1)} = \mathbf{QR}$ and then replace $\mathbf{A}^{(m+1)}$ by $\mathbf{Q}$.

6. Set $\mathbf{C}^{(m+1)} = (c_{jl}^{(m+1)}) = \mathbf{X}^{(m+1)*T}\mathbf{A}^{(m+1)}$. Update $\mathbf{B}$ by $\mathbf{B}^{(m+1)} = (b_{jl}^{(m+1)})$ where

$$b_{jl}^{(m+1)} = \frac{|b_{jl}^{(m)}|}{|b_{jl}^{(m)}| + 4n\lambda_l} c_{jl}^{(m+1)}, \qquad l = 1, \ldots, k,\ j = 1, \ldots, d.$$

7. Repeat steps 2 through 6 with m replaced by $m + 1$ until convergence.

REMARK 2. The orthogonalization in step 5 of Algorithm 1 does not change the descent property of the MM algorithm. Let $A^{(m+1)}$ be the optimizer before orthogonalization. Then $S(A^{(m+1)}, B^{(m)}) \le S(A^{(m)}, B^{(m)})$, where, for simplicity, μ is omitted from the objective function S. Let $A^{(m+1)} = \widetilde{A}^{(m+1)}R$ be the QR decomposition of $A^{(m+1)}$ and let $\widetilde{B}^{(m)} = B^{(m)}R^T$. Then $\widetilde{A}^{(m+1)}\widetilde{B}^{(m)T} = A^{(m+1)}B^{(m)T}$ and so $S(\widetilde{A}^{(m+1)}, \widetilde{B}^{(m)}) = S(A^{(m+1)}, B^{(m)})$. Consequently, $S(\widetilde{A}^{(m+1)}, \widetilde{B}^{(m)}) \le S(A^{(m)}, B^{(m)})$.

5. Handling missing data. Missing data are commonly encountered in real applications. In this section we extend our sparse logistic PCA method to cases when missing data are present.

Let $\mathcal{N} = \{(i, j) | y_{ij} \text{ is not observed}\}$ denote the index set for missing values. The sparse logistic PCA minimizes the following criterion function:

$$T(\boldsymbol{\mu}, \mathbf{A}, \mathbf{B}) = -\ell_{obs}(\boldsymbol{\mu}, \mathbf{A}, \mathbf{B}) + nP_{\boldsymbol{\lambda}}(\mathbf{B}), \tag{5.1}$$

where

$$\ell_{obs}(\boldsymbol{\mu}, \mathbf{A}, \mathbf{B}) = \sum_{(i,j)\notin\mathcal{N}} \sum \log \pi\{q_{ij}(\mu_j + \mathbf{a}_i^T \mathbf{b}_j)\} \tag{5.2}$$

can be interpreted as the observed data log likelihood for model (2.2). Similar to the nonmissing data case, direct minimization of (5.1) is not straightforward because the log likelihood term is not quadratic and the penalty term is nondifferentiable. Direct minimization of (5.1) is also complicated by the fact that the summation in the definition of the observed data log likelihood is not over a rectangular region. Again, we develop an iterative MM algorithm to solve the optimization problem. The strategy is to fill in the missing data with the fitted values based on the current parameter estimates, then proceed with the algorithm that assumes complete data, and iterate until convergence.

Define the working variables

$$z_{ij}^{(m)} = \begin{cases} x_{ij}^{(m)}, & (i, j) \notin \mathcal{N}, \\ \theta_{ij}^{(m)} = \mu_j^{(m)} + \mathbf{a}_i^{(m)T}\mathbf{b}_j^{(m)}, & (i, j) \in \mathcal{N}, \end{cases} \tag{5.3}$$

where $x_{ij}^{(m)}$ is defined in (4.6). Let

$$h(\boldsymbol{\mu}, \mathbf{A}, \mathbf{B} | \boldsymbol{\mu}^{(m)}, \mathbf{A}^{(m)}, \mathbf{B}^{(m)}) = \sum_{i=1}^{n} \sum_{j=1}^{d} [w_{ij}^{(m)} \{z_{ij}^{(m)} - (\mu_j + \mathbf{a}_i^T \mathbf{b}_j)\}^2 + \mathbf{b}_j^T \mathbf{D}_{\boldsymbol{\lambda},j}^{(m)} \mathbf{b}_j], \tag{5.4}$$

where $\mathbf{D}_{\boldsymbol{\lambda},j}^{(m)}$ are diagonal matrices with diagonal elements $\lambda_l / \{2|b_{jl}^{(m)}|\}$ for $l = 1, \ldots, k$. The following result extends Theorem 4.1 to the missing data case. The proof is given in the Appendix.

THEOREM 5.1. (i) *Up to a constant that depends on* $\boldsymbol{\mu}^{(m)}$, $\mathbf{A}^{(m)}$, *and* $\mathbf{B}^{(m)}$ *but not on* $\boldsymbol{\mu}$, $\mathbf{A}$, *and* $\mathbf{B}$, *the function* $h(\boldsymbol{\mu}, \mathbf{A}, \mathbf{B} | \boldsymbol{\mu}^{(m)}, \mathbf{A}^{(m)}, \mathbf{B}^{(m)})$ *defined in* (5.4) *majorizes* $T(\boldsymbol{\mu}, \mathbf{A}, \mathbf{B})$ *at* $(\boldsymbol{\mu}^{(m)}, \mathbf{A}^{(m)}, \mathbf{B}^{(m)})$.

(ii) *Let* $(\boldsymbol{\mu}^{(m)}, \mathbf{A}^{(m)}, \mathbf{B}^{(m)})$, $m = 1, 2, \ldots$, *be a sequence obtained by iteratively minimizing the majorizing function. Then* $T(\boldsymbol{\mu}^{(m)}, \mathbf{A}^{(m)}, \mathbf{B}^{(m)})$ *decreases as* m *gets larger and it converges to a local minimum of* $T(\boldsymbol{\mu}, \mathbf{A}, \mathbf{B})$ *as* m *goes to infinity.*

Note that the majorizing functions given in (5.4) have the same form as those given in (4.9) except that $x_{ij}^{(m)}$ in (4.9) is changed to $z_{ij}^{(m)}$ in (5.4). Thus, the computation algorithm developed in Section 4 is readily applicable in the missing data case with a simple replacement of $x_{ij}^{(m)}$ by $z_{ij}^{(m)}$. The working variable $z_{ij}^{(m)}$ in (5.4) is easily understood: It is the same as the nonmissing data case if y_{ij} is observable; otherwise, it is an imputed θ_{ij} value based on the reduced rank model (2.2) and the current guess of $\boldsymbol{\mu}$, $\mathbf{A}$ and $\mathbf{B}$.

6. Simulation study. In this section we demonstrate our sparse logistic PCA method using a simulation study. The method worked well in various settings that we tested, but here we only report results in a challenging case that the number of variables d is bigger than the sample size n.

6.1. *The signal-to-noise ratio.* To facilitate setting up simulation studies, we introduce a notion of signal-to-noise ratio for logistic PCA. In our logistic PCA model, the entries of the $n \times d$ data matrix are independent Bernoulli random variables with success probability $\pi_{ij} = \{1 + \exp(-\theta_{ij})\}^{-1}$ for the (i, j)th cell. The matrix of canonical parameters $\boldsymbol{\Theta} = (\theta_{ij})$ has a reduced rank representation $\boldsymbol{\Theta} = \mathbf{1} \otimes \boldsymbol{\mu}^T + \mathbf{A}\mathbf{B}^T$, where $\mathbf{A}$ is a $n \times k$ matrix of PC scores and $\mathbf{B}$ is a sparse $d \times k$ PC loading matrix. In our simulation study, elements of the lth column of $\mathbf{A}$ are independent draws from a zero-mean Gaussian distribution with variance σ_{al}^2, $1 \le l \le k$. The variance σ_{al}^2 measures the signal level of the lth PC. We set up the PC variances relative to a suitably defined baseline noise level.

We define a baseline noise level for fixed n, d and k as follows. First we create a binary data matrix by generating $n \times d$ independent binary variables from

Bernoulli distribution with the success probability $1/2$. These binary variables are understood to come from the pure noise since they are generated without having any structure on the success probabilities. Then, we conduct a k-component logistic PCA without regularization and compute the average of the sample variances of the obtained k PC scores, which is denoted as σ_b^2. We repeat the above process of generating "pure noise" binary data matrices a large number of times (e.g., 100) and take the mean of σ_b^2 computed from these matrices as the baseline noise level.

With the notion of baseline noise level, we define the signal-to-noise ratio (SNR) for a PC as

$$\text{SNR} = \frac{\text{variance of PC scores}}{\text{baseline noise level}}. \tag{6.1}$$

In our simulation study we first compute the baseline noise level for a given combination of n, d and k, then use the above formula to specify the variances of PC scores based on the fixed values of SNR.

6.2. *Simulation setup*. We set the intrinsic dimension to be $k = 2$ and the number of rows of the data matrix to be $n = 100$. We varied the number of variables d and the signal-to-noise ratio SNR. We considered three choices of d: $d = 200$, $d = 500$ and $d = 1000$. The scores of the lth PC were randomly drawn from the $N(0, \sigma_{al}^2)$ distribution with $\sigma_{al}^2 = \text{SNR}_l \cdot (\text{baseline noise level})$, where SNR_l is the SNR for the lth PC. We considered two settings of SNR: $(3, 2)$ and $(5, 3)$. For example, when the SNR is $(3, 2)$, the variance of the first PC is 3 times the baseline noise level and the variance of the second PC is 2 times the baseline noise level. We construct two sparse PC loading vectors as follows: Let b_{j1} and b_{j2} denote correspondingly the components of the first and the second PC loading vectors. We let $b_{j1} = 1$ for $j = 1, \ldots, 20$, $b_{j2} = 1$ for $j = 21, \ldots, 40$, and the rest of b_{jl} are all taken to be 0. The mean vector $\boldsymbol{\mu}$ was set to be a vector of zeros.

6.3. *Simulation results*. Logistic PCA with and without sparsity inducing regularization was conducted on 100 simulated data sets for each setting. When applying the sparse logistic PCA algorithm, three choice of k were considered: k is fixed at the true value ($k = 2$), at a moderately large value ($k = 30$), and selected using the BIC. The penalty parameter was selected using the method described in Section 2.2.

To measure the closeness of the estimated PC loading matrix $\widehat{\mathbf{B}}$ and the true loading matrix $\mathbf{B}$, we use the principal angle between spaces spanned by $\widehat{\mathbf{B}}$ and $\mathbf{B}$. The principal angle measures the maximum angle between any two vectors on the spaces generated by the columns of $\widehat{\mathbf{B}}$ and $\mathbf{B}$. More precisely, it is defined by $\cos^{-1}(\rho) \times 180/\pi$, where ρ is the minimum eigenvalue of the matrix $\mathbf{Q}_{\widehat{\mathbf{B}}}^T \mathbf{Q}_{\mathbf{B}}$, where $\mathbf{Q}_{\widehat{\mathbf{B}}}$ and $\mathbf{Q}_{\mathbf{B}}$ are orthogonal basis matrices obtained by the QR decomposition of matrices $\widehat{\mathbf{B}}$ and $\mathbf{B}$, respectively [Golub and van Loan (1996)].

TABLE 1

The results of logistic PCA with and without sparsity inducing regularization, based on 100 *simulated data sets for each setting. The reported values are the mean* (*standard error*) *of the principal angle* (°) *between the estimated and the true PC loading matrices*

d	SNR	$k = 2$	$k = 30$	Selected k
200	SNR = (3, 2)			
	Nonregularized	12.532 (0.115)	35.725 (0.177)	–
	Regularized	5.860 (0.123)	10.125 (0.324)	5.816 (0.125)
	SNR = (5, 3)			
	Nonregularized	11.913 (0.122)	36.350 (0.189)	–
	Regularized	5.803 (0.128)	9.843 (0.321)	5.769 (0.127)
500	SNR = (3, 2)			
	Nonregularized	10.890 (0.095)	31.884 (0.188)	–
	Regularized	4.731 (0.115)	9.413 (0.282)	4.690 (0.101)
	SNR = (5, 3)			
	Nonregularized	10.166 (0.095)	31.941 (0.193)	–
	Regularized	4.729 (0.121)	9.242 (0.252)	4.544 (0.119)
1000	SNR = (3, 2)			
	Nonregularized	12.018 (0.167)	36.040 (0.181)	–
	Regularized	7.015 (0.486)	11.807 (0.433)	4.534 (0.141)
	SNR = (5, 3)			
	Nonregularized	11.370 (0.156)	36.144 (0.180)	–
	Regularized	6.767 (0.474)	10.825 (0.475)	4.196 (0.127)

The mean and standard deviation of principal angles for logistic PCA with and without regularization are presented in Table 1. Since smaller principal angles indicate better estimates of the PC loading matrix, the sparsity inducing regularization has a clear benefit—it can substantially reduce the principal angles. The benefit is even more profound when the number of PCs used in the program ($k = 30$) is larger than the true number that was used to generate the data ($k = 2$). The performance of sparse logistic PCA with selected k is similar to that when k is fixed at the true value. Frequencies of the selected k from 100 simulation data sets in each settings of Table 1 are shown in Table 2. When $d = 200$, the BIC finds well the true $k = 2$ but, as d gets larger, there is a trend that a slightly larger k is selected. The performance of using BIC to select k is considered as quite good, given that the sample size is only 100.

A useful feature of the sparse logistic PCA is its ability to select relevant variables when estimating the PC loading vectors. A zero loading of a variable on a PC means that the corresponding variable is not used when forming that PC, and a nonzero loading indicates a useful variable. Our experience with simulated data shows that nonzero loadings can almost always be identified by the method, but some identified nonzero loadings may correspond to irrelevant variables, cases of

TABLE 2
Frequencies of the selected k using the BIC

		Selected k						
d	SNR	1	2	3	4	5	6	7
200	(3, 2)	0	95	5	0	0	0	0
	(5, 3)	0	96	4	0	0	0	0
500	(3, 2)	1	58	37	4	0	0	0
	(5, 3)	0	60	36	3	1	0	0
1000	(3, 2)	3	34	36	15	10	1	1
	(5, 3)	2	31	47	15	4	1	0

false positives. Table 3 presents the percentages of false positives for various settings reported in Table 1. When d is 500 or 1000, the percentages of false positives are low, all below 20%. But when d is 200, the percentages of false positives are between 40% and 50%, suggesting big room for improvement in variable selection.

7. Discussion and extension. In this paper we propose a sparse PCA method for analyzing binary data by maximizing a penalized Bernoulli likelihood. The sparsity inducing L_1 penalty is used to acquire simple principal components for the sake of easy interpretation and stable estimation. The MM algorithm developed for implementation of our method provides a unified solution for dealing with (i) the nonquadratic likelihood, (ii) the nondifferentiable penalty function, and (iii) presence of missing data. Although the theoretical derivation is not straightforward, the steps of the algorithm are very simple—they are (weighted) penalized least squares with closed-form expressions.

TABLE 3
The results of logistic PCA with sparsity inducing regularization, based on 100 *simulated data sets for each setting in Table* 1. *The reported values are the mean* (*standard error*) *of the percentages of false positives. The description of results is in the text*

d	SNR	$k = 2$	$k = 30$	Selected k
200	(3, 2)	45.05(1.54)	41.51 (1.39)	44.94 (1.51)
	(5, 3)	48.16 (1.63)	40.53 (1.36)	48.26 (1.63)
500	(3, 2)	14.83 (0.74)	18.91 (0.51)	16.70 (0.72)
	(5, 3)	16.06 (0.68)	18.78 (0.42)	16.93 (0.68)
1000	(3, 2)	10.87 (0.75)	12.80 (0.73)	10.13 (0.60)
	(5, 3)	10.89 (0.70)	12.86 (0.73)	9.26 (0.50)

We have focused on the logit link so far, but other link functions can also be used. In particular, a slight modification of the proposed method can handle the probit link, where the success probabilities $\theta_{ij} = \Phi^{-1}(\pi_{ij})$ with $\Phi(\cdot)$ being the c.d.f. of the standard Gaussian distribution. The log likelihood function (2.3) of the reduced rank model is changed to

$$\ell(\boldsymbol{\mu}, \mathbf{A}, \mathbf{B}) = \sum_{j=1}^{d} \sum_{i=1}^{n} \log \Phi\{q_{ij}(\mu_j + \mathbf{a}_i^T \mathbf{b}_j)\}. \tag{7.1}$$

Instead of using the majorization in (4.2), we apply the following upper bound to majorize the negative log likelihood:

$$-\log \Phi(x) \le -\log \Phi(y) - \frac{\phi(y)}{\Phi(y)}(x - y) + \frac{1}{2}(x - y)^2, \tag{7.2}$$

where $\phi(\cdot)$ is the Gaussian density [Böhning (1999); de Leeuw (2006)]. Algorithm 1 still applies with appropriate changes to the definitions of the weights $w_{ij}^{(m)}$ and the working variables $x_{ij}^{(m)}$.

Our method can also be extended in a straightforward way to handle composite data which consists of both binary and continuous variables. While the binary variables are modeled with Bernoulli distributions, the continuous variables can be modeled with Gaussian distributions. Including some continuous variables corresponds to adding some negative Gaussian log likelihood terms to the log likelihood expression (2.3). Since the Gaussian log likelihood is quadratic, it blends in easily with the quadratic majorization used for the logistic PCA. Specifically, if the jth variable is of a continuous type, we assume $y_{ij} \sim N(\theta_{ij}, \sigma^2)$ with θ_{ij} satisfying (2.2), and simply let $x_{ij}^{(m)} = y_{ij}$ and $w_{ij}^{(m)} = 1/\sigma^2$ when forming the majorizing function (4.9). The residual variance σ^2 of fitting the continuous variables can be estimated using the residual sum of squares. Taking into account the fact that different weighting schemes are used for the binary variables and the continuous variables in the majorizing function, a slight modification of Algorithm 2 presented in the supplemental article [Lee, Huang and Hu (2010)] can be used for computation.

APPENDIX

A.1. Proof of Theorem 4.1. We prove the results for both the tight and the uniform bound case. Applications of (4.1) and (4.2) yield the following majorizing functions of the negative log likelihood $-\ell(\boldsymbol{\mu}, \mathbf{A}, \mathbf{B})$:

$$\sum_{i=1}^{n} \sum_{j=1}^{d} \Big[-\log \pi\big(q_{ij}\theta_{ij}^{(m)}\big) - q_{ij}\big\{1 - \pi\big(q_{ij}\theta_{ij}^{(m)}\big)\big\}\big(\theta - \theta_{ij}^{(m)}\big) + \frac{2\pi\big(q_{ij}\theta_{ij}^{(m)}\big) - 1}{4q_{ij}\theta_{ij}^{(m)}}\big(\theta - \theta_{ij}^{(m)}\big)^2 \Big]$$

for the tight bound, and

$$\sum_{i=1}^{n}\sum_{j=1}^{d}\Big[-\log\pi\big(q_{ij}\theta_{ij}^{(m)}\big)-q_{ij}\{1-\pi\big(q_{ij}\theta_{ij}^{(m)}\big)\}(\theta-\theta_{ij}^{(m)})+\frac{1}{8}(\theta-\theta_{ij}^{(m)})^2\Big]$$

for the uniform bound. Note that

$$\{2\pi\big(q_{ij}\theta_{ij}^{(m)}\big)-1\}/\{4q_{ij}\theta_{ij}^{(m)}\}=\{2\pi\big(\theta_{ij}^{(m)}\big)-1\}/\{4\theta_{ij}^{(m)}\}$$

for $q_{ij}=\pm1$. By completing the squares and using the definitions of $x_{ij}^{(m)}$ and $w_{ij}^{(m)}$, these majorizing functions can be rewritten as

$$\begin{aligned}&-\tilde{\ell}\big(\boldsymbol{\mu},\mathbf{A},\mathbf{B}|\boldsymbol{\mu}^{(m)},\mathbf{A}^{(m)},\mathbf{B}^{(m)}\big)\\&\quad=-\ell\big(\boldsymbol{\Theta}^{(m)}\big)-2\sum_{i=1}^{n}\sum_{j=1}^{d}\{1-\pi\big(q_{ij}\theta_{ij}^{(m)}\big)\}^2+\sum_{i=1}^{n}\sum_{j=1}^{d}w_{ij}^{(m)}\big(\theta_{ij}-x_{ij}^{(m)}\big)^2.\end{aligned}$$

On the other hand, application of (4.3) yields the following majorizing function of $P_{\boldsymbol{\lambda}}(\mathbf{B})$:

$$\begin{aligned}\tilde{P}_{\boldsymbol{\lambda}}(\mathbf{B}|\mathbf{B}^{(m)})&=\lambda_1\sum_{j=1}^{d}\frac{b_{j1}^2+b_{j1}^{(m)2}}{2|b_{j1}^{(m)}|}+\cdots+\lambda_k\sum_{j=1}^{d}\frac{b_{jk}^2+b_{jk}^{(m)2}}{2|b_{jk}^{(m)}|}\\&=\sum_{j=1}^{d}\mathbf{b}_j^{(m)T}\mathbf{D}_{\boldsymbol{\lambda},j}^{(m)}\mathbf{b}_j^{(m)}+\sum_{j=1}^{d}\mathbf{b}_j^{T}\mathbf{D}_{\boldsymbol{\lambda},j}^{(m)}\mathbf{b}_j.\end{aligned}$$

Since the majorization relation between functions is closed under the formation of sums, $-\tilde{\ell}+n\tilde{P}_{\boldsymbol{\lambda}}(\mathbf{B}|\mathbf{B}^{(m)})$ majorizes $S(\boldsymbol{\mu},\mathbf{A},\mathbf{B})$ at $(\boldsymbol{\mu}^{(m)},\mathbf{A}^{(m)},\mathbf{B}^{(m)})$. Noticing that $-\tilde{\ell}+n\tilde{P}_{\boldsymbol{\lambda}}(\mathbf{B}|\mathbf{B}^{(m)})$ equals $g(\boldsymbol{\mu},\mathbf{A},\mathbf{B}|\boldsymbol{\mu}^{(m)},\mathbf{A}^{(m)},\mathbf{B}^{(m)})$ up to a constant independent of $(\boldsymbol{\mu},\mathbf{A},\mathbf{B})$, we complete the proof of part (i). Part (ii) of the theorem follows from the general property of the MM algorithm [Hunter and Lange (2004)]. □

A.2. Proof of Theorem 5.1. Note that the objective function to be minimized is the summation of two terms—the log likelihood term and the penalty term. Because the majorization property is closed under function summation, we deal with the two terms separately. We can find a majorization function of the penalty term as in Theorem 4.1. To find a majorization function of the log likelihood term, we apply the argument in the standard EM algorithm for handling missing data [Dempster, Laird and Rubin (1977)]. The complete data log likelihood is

$$\ell_{com}(\boldsymbol{\mu},\mathbf{A},\mathbf{B})=\sum_{(i,j)\notin\mathcal{N}}\sum\log\pi(q_{ij}\theta_{ij})+\sum_{(i,j)\in\mathcal{N}}\sum\log\pi(q_{ij}\theta_{ij}).$$

Its conditional expectation given the observed data and the current guess of the parameter values is

$$\begin{aligned} &Q(\boldsymbol{\mu}, \mathbf{A}, \mathbf{B} | \boldsymbol{\mu}^{(m)}, \mathbf{A}^{(m)}, \mathbf{B}^{(m)}) \\ &\quad = \sum_{(i,j)\notin\mathcal{N}} \sum \log \pi(q_{ij}\theta_{ij}) \\ &\qquad + \sum_{(i,j)\in\mathcal{N}} \sum E\big[\log \pi(q_{ij}\theta_{ij}) | \mathbf{Y}_o, \boldsymbol{\mu}^{(m)}, \mathbf{A}^{(m)}, \mathbf{B}^{(m)}\big], \end{aligned} \tag{A.1}$$

where $\mathbf{Y}_o$ denotes the observed data. By the standard EM theory,

$$\begin{aligned} -\tilde{\ell}_{obs}(\boldsymbol{\mu}, \mathbf{A}, \mathbf{B}) &\triangleq -Q(\boldsymbol{\mu}, \mathbf{A}, \mathbf{B} | \boldsymbol{\mu}^{(m)}, \mathbf{A}^{(m)}, \mathbf{B}^{(m)}) - \ell_{obs}(\boldsymbol{\mu}^{(m)}, \mathbf{A}^{(m)}, \mathbf{B}^{(m)}) \\ &\quad + Q(\boldsymbol{\mu}^{(m)}, \mathbf{A}^{(m)}, \mathbf{B}^{(m)} | \boldsymbol{\mu}^{(m)}, \mathbf{A}^{(m)}, \mathbf{B}^{(m)}) \end{aligned} \tag{A.2}$$

majorizes $-\ell_{obs}(\boldsymbol{\mu}, \mathbf{A}, \mathbf{B})$ at $(\boldsymbol{\mu}^{(m)}, \mathbf{A}^{(m)}, \mathbf{B}^{(m)})$, that is, $-\tilde{\ell}_{obs}(\boldsymbol{\mu}, \mathbf{A}, \mathbf{B}) \geq -\ell_{obs}(\boldsymbol{\mu}, \mathbf{A}, \mathbf{B})$, and the equality holds when $(\boldsymbol{\mu}, \mathbf{A}, \mathbf{B}) = (\boldsymbol{\mu}^{(m)}, \mathbf{A}^{(m)}, \mathbf{B}^{(m)})$.

Now we find a quadratic majorizing function of $-\tilde{\ell}_{obs}(\boldsymbol{\mu}, \mathbf{A}, \mathbf{B})$, which in turn majorizes $-\ell_{obs}(\boldsymbol{\mu}, \mathbf{A}, \mathbf{B})$ because of the transitivity of the majorization relation. We need only to find a quadratic majorization function of $-Q(\boldsymbol{\mu}, \mathbf{A}, \mathbf{B} | \boldsymbol{\mu}^{(m)}, \mathbf{A}^{(m)}, \mathbf{B}^{(m)})$, since it is the only term in the definition (A.2) of $-\tilde{\ell}_{obs}(\boldsymbol{\mu}, \mathbf{A}, \mathbf{B})$ that depends on the unknown parameters. According to (A.1), $-Q(\boldsymbol{\mu}, \mathbf{A}, \mathbf{B} | \boldsymbol{\mu}^{(m)}, \mathbf{A}^{(m)}, \mathbf{B}^{(m)})$ can be decomposed into two terms, one corresponding to observed data, the other corresponding to the missing data. The former term can been treated as in the proof of Theorem 4.1. When $(i, j) \notin \mathcal{N}$, $-\log \pi(q_{ij}\theta_{ij})$ is majorized by $w_{ij}^{(m)}(\theta_{ij} - x_{ij}^{(m)})^2$, up to a constant. To treat the latter term, note that, when $(i, j) \in \mathcal{N}$,

$$\begin{aligned} &E\big[\log \pi(q_{ij}\theta_{ij}) | \mathbf{Y}_o, \boldsymbol{\mu}^{(m)}, \mathbf{A}^{(m)}, \mathbf{B}^{(m)}\big] \\ &\quad = \pi\big(\theta_{ij}^{(m)}\big) \log \pi(\theta_{ij}) + \big\{1 - \pi\big(\theta_{ij}^{(m)}\big)\big\} \log\{1 - \pi(\theta_{ij})\} \\ &\quad = \sum_{q_{ij}=\pm 1} \pi\big(q_{ij}\theta_{ij}^{(m)}\big) \log \pi(q_{ij}\theta_{ij}), \end{aligned}$$

using the fact that the missing data are independent of the observed data, and that $1 - \pi(\theta) = \pi(-\theta)$. Then, by applying the inequalities (4.1) and (4.2) and using the definition of $w_{ij}^{(m)}$, we obtain that

$$\begin{aligned} &-E\big[\log \pi(q_{ij}\theta_{ij}) | \mathbf{Y}_o, \boldsymbol{\mu}^{(m)}, \mathbf{A}^{(m)}, \mathbf{B}^{(m)}\big] \\ &\quad \leq \sum_{q_{ij}=\pm 1} \pi\big(q_{ij}\theta_{ij}^{(m)}\big)\big[-\log \pi\big(\theta_{ij}^{(m)}\big) - \big\{1 - \pi\big(q_{ij}\theta_{ij}^{(m)}\big)\big\}\big\{q_{ij}\big(\theta_{ij} - \theta_{ij}^{(m)}\big)\big\} \\ &\qquad\qquad + w_{ij}^{(m)}\big\{\big(\theta_{ij} - \theta_{ij}^{(m)}\big)\big\}^2\big] \\ &\quad \leq C_m + w_{ij}^{(m)}\big\{\big(\theta_{ij} - \theta_{ij}^{(m)}\big)\big\}^2, \end{aligned}$$

where C_m is a constant independent of $\boldsymbol{\mu}$, $\mathbf{A}$ and $\mathbf{B}$. Combining the above results, we see that $-Q(\boldsymbol{\mu}, \mathbf{A}, \mathbf{B}|\boldsymbol{\mu}^{(m)}, \mathbf{A}^{(m)}, \mathbf{B}^{(m)})$ is up to a constant majorized by $\sum_{ij} w_{ij}^{(m)}\{(\theta_{ij} - z_{ij}^{(m)})\}^2$, where $z_{ij}^{(m)}$ equals $x_{ij}^{(m)}$ if $(i, j) \notin \mathcal{N}$, and $\theta_{ij}^{(m)}$ if $(i, j) \in \mathcal{N}$. The proof of part (i) is thus complete. Part (ii) of the theorem follows from the general result of the MM algorithm. □

Acknowledgments. We would like to thank Editor Michael Stein, an Associate Editor and two referees for helpful comments. We would also like to thank Lan Zhou for help in improving the writing of the paper.

SUPPLEMENTARY MATERIAL

The MM algorithm for sparse logistic PCA using the tight bound (DOI: 10.1214/10-AOAS327SUPP; .pdf). We develop the MM algorithm for sparse logistic PCA using the tight majorizing bound. Comparison of the developed algorithm with the MM algorithm using the uniform bound in terms of computing time is also presented.

REFERENCES

BÖHNING, D. (1999). The lower bound method in probit regression. *Comput. Statist. Data Anal.* **30** 13–17. MR1681451

BROOKES, A. J. (1999). Review: The essence of SNPs. *Gene* **234** 177–186.

COLLINS, M., DASGUPTA, S. and SCHAPIRE, R. E. (2001). A generalization of principal component analysis to the exponential family. In *Advanced in Neural Information Processing System* **14**.

DE LEEUW, J. (2006). Principal component analysis of binary data by iterated singular value decomposition. *Comput. Statist. Data Anal.* **50** 21–39. MR2196220

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537

EWENS, W. J. and SPIELMAN, R. S. (1995). The transmission/disequilibrium test: History, subdivision, and admixture. *The American Journal of Human Genetics* **57** 455–464.

GOLUB, G. and VAN LOAN, C. (1996). *Matrix Computations*, 3rd ed. The Johns Hopkins Univ. Press. MR1417720

HAO, K., LI, C., ROSENOW, C. and WONG, W. H. (2004). Detect and adjust for population stratification in population-based association study using genomic control markers: An application of Affymetrix Genechip® Human Mapping 10K array. *European Journal of Human Genetics* **12** 1001–1006.

HOTELLING, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* **24** 417–441.

HUNTER, D. R. and LANGE, K. (2004). A tutorial on MM algorithms. *Amer. Statist.* **58** 30–37. MR2055509

HUNTER, D. R. and LI, R. (2005). Variable selection using MM algorithms. *Ann. Statist.* **33** 1617–1642. MR2166557

JAAKKOLA, T. S. and JORDAN, M. I. (2000). Bayesian parameter estimation via variational methods. *Statist. Comput.* **10** 25–37.

JOLLIFFE, I. T. (2002). *Principal Component Analysis*. Springer. MR2036084

JOLLIFFE, I. T., TRENDAFILOV, M. and UDDINE, M. (2003). A modified principal component technique based on the LASSO. *J. Comput. Graph. Statist.* **12** 531–547. MR2002634

KWOK, P. Y., DENG, Q., ZAKERI, H., TAYLOR, S. L. and NICKERSON, D. A. (1996). Increasing the information content of STS-based genome maps: Identifying polymorphisms in mapped STSs. *Genomics* **31** 123–126.

LANGE, K., HUNTER, D. R. and YANG, I. (2000). Optimization transfer using surrogate objective functions (with discussion). *J. Comput. Graph. Statist.* **9** 1–20. MR1819865

LEE, S., HUANG, J. Z. and HU, J. (2010). The MM algorithm for sparse logistic PCA using the tight bound: A supplementary note to "Sparse logistic principal components analysis for binary data." DOI: 10.1214/10-AOAS327SUPP.

LIANG, Y. and KELEMEN, A. (2008). Statistical advances and challenges for analyzing correlated high dimensional SNP data in genomic study for complex diseases. *Statistics Surveys* **2** 43–60. MR2520980

PEARSON, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh and Dublin Pholosophical Magazine and Journal of Science, Sixth Series* **2** 559–572.

RISCH, N., BURCHARD, E., ZIV, E. and TANG, H. (2002). Categorization of humans in biomedical research: Genes, race and disease. *Genome Biology* **3** comment 2007.1–2007.12.

SCHEIN, A. I., SAUL, L. K. and UNGAR, L. H. (2003). A generalized linear model for principal component analysis of binary data. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics* 14–21.

SERRE, D., MONTPETIT, A., PARÉ, G., ENGERT, J. G., YUSUF, S., KEAVNEY, B., HUDSON, K. J. and ANAND, S. (2008). Correction of population stratification in large multi-ethnic association studies. *PLoS ONE* **2** e1382.

SHEN, H. and HUANG, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *J. Multivariate Anal.* **99** 1015–1034. MR2419336

THE INTERNATIONAL HAPMAP CONSORTIUM (2005). A haplotype map of the human genome. *Nature* **437** 1299–1320.

TIBSHIRANI, R. J. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

ZOU, H., HASTIE, T. J. and TIBSHIRANI, R. J. (2006). Sparse principal component analysis. *J. Comput. Graph. Statist.* **15** 265–286. MR2252527

ZOU, H., HASTIE, T. J. and TIBSHIRANI, R. J. (2007). On the "Degrees of Freedom" of the LASSO. *Ann. Statist.* **35** 2173–2192. MR2363967

S. LEE
DEPARTMENT OF BIOSTATISTICS
HARVARD SCHOOL OF PUBLIC HEALTH
BOSTON, MASSACHUSETTS 02115
USA
E-MAIL: seokhol@hsph.harvard.edu

J. Z. HUANG
DEPARTMENT OF STATISTICS
TEXAS A&M UNIVERSITY
COLLEGE STATION, TEXAS 77843-3143
USA
E-MAIL: jianhua@stat.tamu.edu

J. HU
DEPARTMENT OF BIOSTATISTICS
DIVISION OF QUANTITATIVE SCIENCES
UNIVERSITY OF TEXAS M. D. ANDERSON CANCER CENTER
HOUSTON, TEXAS 77030-4009
USA
E-MAIL: jhu@mdanderson.org

THE LIST OF SOURCE ENTRIES RETRIEVED FROM MATHSCINET

The list of entries below corresponds to the Reference section of your article and was retrieved from MathSciNet applying an automated procedure. Please check the list and cross out those entries which lead to mistaken sources. Please update your references entries with the data from the corresponding sources, when applicable. More information can be found in the support page:
http://www.e-publications.org/ims/support/mrhelp.html.

BÖHNING, D. (1999). The lower bound method in probit regression. *Comput. Statist. Data Anal.* **30** 13–17. MR1681451 (2000a:62149)

Not Found!

Not Found!

DE LEEUW, J. (2006). Principal component analysis of binary data by iterated singular value decomposition. *Comput. Statist. Data Anal.* **50** 21–39. MR2196220

DEMPSTER, A. P., LAIRD, N. M., AND RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. With discussion. MR0501537 (58 #18858)

Not Found!

GOLUB, G. H. AND VAN LOAN, C. F. (1996). *Matrix computations*, Third ed. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD. MR1417720 (97g:65006)

Not Found!

Not Found!

HUNTER, D. R. AND LANGE, K. (2004). A tutorial on MM algorithms. *Amer. Statist.* **58** 30–37. MR2055509

HUNTER, D. R. AND LI, R. (2005). Variable selection using MM algorithms. *Ann. Statist.* **33** 1617–1642. MR2166557

Not Found!

JOLLIFFE, I. T. (2002). *Principal component analysis*, Second ed. Springer Series in Statistics. Springer-Verlag, New York. MR2036084 (2004k:62010)

JOLLIFFE, I. T., TRENDAFILOV, N. T., AND UDDIN, M. (2003). A modified principal component technique based on the LASSO. *J. Comput. Graph. Statist.* **12** 531–547. MR2002634

Not Found!

LANGE, K., HUNTER, D. R., AND YANG, I. (2000). Optimization transfer using surrogate objective functions. *J. Comput. Graph. Statist.* **9** 1–59. With discussion, and a rejoinder by Hunter and Lange. MR1819865 (2001k:62029)

Not Found!

LIANG, Y. AND KELEMEN, A. (2008). Statistical advances and challenges for analyzing correlated high dimensional SNP data in genomic study for complex diseases. *Stat. Surv.* **2** 43–60. MR2520980

Not Found!

Not Found!

Not Found!

Not Found!

SHEN, H. AND HUANG, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *J. Multivariate Anal.* **99** 1015–1034. MR2419336 (2009m:62184)

Not Found!

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242 (96j:62134)

ZOU, H., HASTIE, T., AND TIBSHIRANI, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Statist.* **15** 265–286. MR2252527

ZOU, H., HASTIE, T., AND TIBSHIRANI, R. (2007). On the "degrees of freedom" of the lasso. *Ann. Statist.* **35** 2173–2192. MR2363967 (2009d:62096)

META DATA IN THE PDF FILE

Following information will be included as pdf file Document Properties:

Title : Sparse logistic principal components analysis for binary data
Author : Seokho Lee, Jianhua Z. Huang, Jianhua Hu
Subject : The Annals of Applied Statistics , 2010, Vol.0, No.00, 1-26
Keywords: Binary data, dimension reduction, MM algorithm, LASSO, PCA, regularization, sparsity.

THE LIST OF URI ADRESSES

Listed below are all uri addresses found in your paper. The non-active uri addresses, if any, are indicated as ERROR. Please check and update the list where necessary. The e-mail addresses are not checked – they are listed just for your information. More information can be found in the support page:
http://www.e-publications.org/ims/support/urihelp.html.

200 http://www.imstat.org/aoas/ [2:pp.1,1] OK
200 http://dx.doi.org/10.1214/10-AOAS327 [2:pp.1,1] OK
200 http://www.imstat.org [2:pp.1,1] OK
200 http://lib.stat.cmu.edu/aoas/327/supplement.pdf [2:pp.22,22] OK
200 http://dx.doi.org/10.1214/10-AOAS327SUPP [2:pp.23,23] OK
--- mailto:seokhol@hsph.harvard.edu [2:pp.23,23] Check skip
--- mailto:jianhua@stat.tamu.edu [2:pp.23,23] Check skip
--- mailto:jhu@mdanderson.org [2:pp.23,23] Check skip