Journal of Multivariate Analysis 180 (2020) 104668

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Dimensionality reduction for binary data through the projection of natural parameters

Andrew J. Landgraf [a], Yoonkyung Lee [b,*]

[a] Root Insurance, 80 E Rich St Suite 500, Columbus, OH 43215, USA
[b] Department of Statistics, The Ohio State University, Columbus, OH 43210, USA

ARTICLE INFO

Article history:
Received 18 January 2020
Received in revised form 5 August 2020
Accepted 9 August 2020
Available online 19 August 2020

AMS 2010 subject classifications:
62H25
62J12

Keywords:
Binary data
Exponential family
Logistic PCA
Principal component analysis

ABSTRACT

Principal component analysis (PCA) for binary data, known as logistic PCA, has become a popular alternative to dimensionality reduction of binary data. It is motivated as an extension of ordinary PCA by means of a matrix factorization, akin to the singular value decomposition, that maximizes the Bernoulli log-likelihood. We propose a new formulation of logistic PCA which extends Pearson's formulation of a low dimensional data representation with minimum error to binary data. Our formulation does not require a matrix factorization, as previous methods do, but instead looks for projections of the natural parameters from the saturated model. Due to this difference, the number of parameters does not grow with the number of observations and the principal component scores on new data can be computed with simple matrix multiplication. We derive explicit solutions for data matrices of special structure and provide a computationally efficient algorithm for solving for the principal component loadings. Through simulation experiments and an analysis of medical diagnoses data, we compare our formulation of logistic PCA to the previous formulation as well as ordinary PCA to demonstrate its benefits.

1. Introduction

Principal component analysis (PCA) is perhaps the most popular dimensionality reduction technique (see [9] for example). It is useful for data compression, visualization, and feature discovery. PCA can be motivated either by maximizing the variance of linear combinations of the variables [5] or by minimizing the reconstruction error of a lower dimensional projection of the cases [16]. There is an implicit connection between standard PCA and the Gaussian distribution in Pearson's formulation. [21] also showed that PCA provides the maximum likelihood estimate for a factor model, where the data are assumed to be Gaussian.

Although PCA is commonly used for dimensionality reduction for various types of data in practice, the fact that PCA finds a low-rank subspace by implicitly minimizing the reconstruction error under the squared error loss renders direct application of PCA to non-Gaussian data such as binary responses or counts conceptually unappealing. Moreover, the probabilistic interpretation of PCA with normal likelihood in [21] suggests the possibility of proper likelihood-based loss functions in defining the best subspace of a given rank for other types of data. With this motivation, Collins et al. [1] proposed a generalization of PCA to exponential family data using the generalized linear model (GLM) framework, and [13,19], and [20] examined similar generalizations for binary data in particular, using the Bernoulli likelihood, which

* Corresponding author.
E-mail address: yklee@stat.osu.edu (Y. Lee).

https://doi.org/10.1016/j.jmva.2020.104668
0047-259X/

is referred to as logistic PCA. Generalized PCA estimates the natural parameters of a data matrix in a lower dimensional subspace by minimizing the negative log-likelihood under an exponential family distribution. In the Gaussian case, generalized PCA is shown to be equivalent to the truncated singular value decomposition (SVD). This formulation for low-rank matrix factorizations of the natural parameters has been extended further by considering general loss functions and additional penalties on the row and column factors [22].

In this paper, we argue that Collins et al.'s logistic PCA [1] is more closely related to SVD than PCA because it aims at a low-rank factorization of the natural parameters matrix. Consequently, each case has its own incidental parameter, and the total number of parameters increases with the number of cases. The drawback of the formulation becomes apparent when it comes to prediction. To apply logistic PCA to new data, one needs to carry out another optimization, which is prone to overfit. This is in contrast with standard PCA where the principal component scores for the new data are simply given by linear combinations of the observed values of the variables.

Retaining the structure of standard PCA, we generalize PCA in such a way that the principal component scores are linear functions of the data. This is done by interpreting Pearson's formulation [16] in a slightly different manner. A projection of the data with minimum reconstruction error under squared loss can be viewed alternatively as a projection of the natural parameters of a saturated model with minimum deviance for Gaussian data. This alternative interpretation allows a coherent generalization of standard PCA to exponential family distributions. When the distribution is Gaussian, this generalization simplifies to standard PCA. Due to the prevalence of binary data and for simplicity of exposition of the new generalization of PCA, we focus on logistic PCA in this paper.

Our formulation has several benefits over the formulation in [1]. The number of parameters does not increase with the number of observations, the principal component scores are easily interpretable as linear functions of the data, and applying principal components to a new set of data only requires a matrix multiplication. Furthermore, while very little is known about solutions to the formulation in [1], some explicit solutions to our formulation can be derived for binary data matrices with special structures.

Computationally, our formulation of logistic PCA, like the previous one, leads to a non-convex problem. We derive a practical algorithm for generating solutions. It involves iterative spectral decompositions, using the majorization–minimization (MM) algorithm (see [6]). Using the MM algorithm, we apply our formulation of logistic PCA to several datasets, examining the advantages and trade-offs with existing methods.

The rest of the paper is organized as follows. Section 2 gives background on PCA and logistic PCA. In Section 3, we introduce our formulation of logistic PCA and qualitatively compare it to the previous formulation. In Section 4, we derive the first-order optimality conditions for logistic PCA solutions and characterize explicit solutions that satisfy the conditions for data matrices with special structures. In Section 5, we derive an algorithm for logistic PCA. Section 6 shows the potential benefits of our formulation via data analyses with simulated and real data. Finally, Section 7 concludes the paper with a discussion and further extensions of the proposed method. Proofs and additional comments on selection of the number of principal components are in Section 8.

2. Background

Pearson [16] considered a geometric problem of finding an optimal representation of multivariate data in a low dimension with respect to mean squared error. Assume that data consist of $\mathbf{x}_i \in \mathbb{R}^d$, $i \in \{1, \ldots, n\}$. To project the original d-dimensional data into lower dimensions, say, $k < d$, we represent each $\mathbf{x}$ by $\boldsymbol{\mu} + \mathbf{U}\mathbf{U}^\top(\mathbf{x} - \boldsymbol{\mu})$, where $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\mathbf{U}$ is a $d \times k$ matrix with orthonormal columns such that $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_k$. Pearson [16] showed that the minimum of the mean squared error of the k-dimensional representation, or equivalently, the total squared error

$$\sum_{i=1}^{n} \|\mathbf{x}_i - \boldsymbol{\mu} - \mathbf{U}\mathbf{U}^\top(\mathbf{x}_i - \boldsymbol{\mu})\|^2 = \|\mathbf{X} - \mathbf{1}_n\boldsymbol{\mu}^\top - (\mathbf{X} - \mathbf{1}_n\boldsymbol{\mu}^\top)\mathbf{U}\mathbf{U}^\top\|_F^2 \tag{1}$$

is attained when $\boldsymbol{\mu}$ equals the sample mean and $\mathbf{U}$ is a matrix with columns containing the first k eigenvectors of the sample covariance matrix. Here $\mathbf{X}$ is the $n \times d$ matrix with $\mathbf{x}_i^\top$ in the ith row, $\mathbf{1}_n$ is the vector of n ones, and $\|\cdot\|_F$ is the Frobenius norm.

The MSE criterion is closely linked to a Gaussian assumption. Borrowing the characterization of PCA in [1], suppose that $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\theta}_i, \mathbf{I}_d)$, and $\boldsymbol{\theta}_i$ are constrained to lie in a k-dimensional subspace. That is, $\boldsymbol{\theta}_i$ are in the span of a k-dimensional orthonormal basis $\mathbf{b}_\ell \in \mathbb{R}^d$, $\ell \in \{1, \ldots, k\}$ so that $\boldsymbol{\theta}_i = \sum_{\ell=1}^{k} a_{i\ell}\mathbf{b}_\ell$ for some $a_{i\ell}$. In this case, the negative log-likelihood is proportional to

$$\sum_{i=1}^{n} \|\mathbf{x}_i - \boldsymbol{\theta}_i\|^2 = \|\mathbf{X} - \mathbf{A}\mathbf{B}^\top\|_F^2, \tag{2}$$

where $\mathbf{A}$ is an $n \times k$ matrix with elements $a_{i\ell}$ and $\mathbf{B}$ is a $d \times k$ matrix made up of the k basis vectors. It is well known (see [2]) that this objective function is minimized by the rank k truncated singular value decomposition (SVD) of $\mathbf{X}$, where $\mathbf{B}$ consists of the first k right singular vectors and $\mathbf{A}$ consists of the first k left singular vectors scaled by the first k singular values or, equivalently, $\mathbf{A} = \mathbf{X}\mathbf{B}$. The fact that the columns in $\mathbf{B}$ span the same subspace as the first k eigenvectors of $\mathbf{X}^\top\mathbf{X}$

yields the equivalence in the solutions of PCA and SVD. Note that $\mathbf{x}_i$ in (2) is assumed to be mean-centered. Otherwise, $\mathbf{x}_i$ is to be replaced with $\mathbf{x}_i - \boldsymbol{\mu}$ as in (1).

Using the equivalence and the alternative formulation of PCA through the mean space approximation under a Gaussian distribution, [1] extended PCA to exponential family data. The extension is similar to the way that generalized linear models (GLMs) [15] extend linear regression to response variables that are non-Gaussian. Such an extension for non-Gaussian data would be useful for handling binary, count, or non-negative data that abound in practice. Given a sample of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from an exponential family distribution with corresponding natural parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$, generalized PCA finds a k-dimensional subspace for the natural parameters by minimizing the negative log-likelihood. This is done by approximating the $n \times d$ matrix of the natural parameters $\boldsymbol{\Theta} = [\theta_{ij}]$ by factorization of the form $\boldsymbol{\Theta} = \mathbf{A}\mathbf{B}^\top$, where $\mathbf{A}$ and $\mathbf{B}$ are of rank k.

Based on this generalization of PCA, [13] and [19] have developed algorithms specifically for binary data. [19] also extended the specification of the natural parameter space in [1] by introducing variable main effects or biases, $\boldsymbol{\mu}$, so that $\boldsymbol{\Theta} = \mathbf{1}_n\boldsymbol{\mu}^\top + \mathbf{A}\mathbf{B}^\top$. In an extension of probabilistic PCA [21], [20] proposed a factor model for binary data with a similar structure, where the case-specific parameters in $\mathbf{A}$ are assumed to come from a k dimensional standard normal distribution. The marginal distribution of $\mathbf{X}$ is then used to define a likelihood of $\mathbf{B}$ and $\boldsymbol{\mu}$ and is maximized with respect to $\mathbf{B}$ and $\boldsymbol{\mu}$. Like other methods, estimating principal component scores on new data requires additional computation to solve for the parameters on the new cases.

3. New formulation of logistic PCA

We propose a new formulation of generalized PCA and demonstrate its conceptual and computational advantages over the current formulation.

3.1. New interpretation of PCA

For the new formulation, we begin with a new interpretation of standard PCA. Using key concepts in GLMs, we interpret PCA as a technique for low dimensional projection of the natural parameters of the saturated model, which are the same as the data under the normal likelihood.

To elaborate on this perspective, recall that for a data matrix $\mathbf{X}$, the $d \times k$ matrix with the first k principal component loading vectors minimizes the squared error in (1). The principal component loadings are given by the first k eigenvectors of the sample covariance matrix. To draw the connection to the Gaussian model and GLMs, suppose that mean-centered $\mathbf{x}_i$ are normally distributed with known variance: $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\theta}_i, \mathbf{I}_d)$. The normal deviance for mean $\boldsymbol{\Theta}$ with data $\mathbf{X}$ is

$$-2\left[\ln p(\mathbf{X}; \boldsymbol{\Theta}) - \ln p(\mathbf{X}; \tilde{\boldsymbol{\Theta}})\right] = -2\sum_{i=1}^{n}\left(\ln p(\mathbf{x}_i; \boldsymbol{\theta}_i) - \ln p(\mathbf{x}_i; \tilde{\boldsymbol{\theta}}_i)\right),$$

which is simplified to $\sum_{i=1}^{n} \|\mathbf{x}_i - \boldsymbol{\theta}_i\|^2$. With the identity function as the canonical link for a normal distribution, the natural parameter θ_{ij} is the mean itself in this case, and the saturated model, which is the best possible fit to the data defined in terms of deviance, is the model with natural parameter $\tilde{\theta}_{ij} = x_{ij}$. In other words, for Gaussian data, the natural parameters of the saturated model are equal to the data, $\tilde{\boldsymbol{\Theta}} = \mathbf{X}$. This implies that a data projection, $\mathbf{X}\mathbf{U}\mathbf{U}^\top$, is also interpreted as the projection of the mean parameters of the saturated model, $\tilde{\boldsymbol{\Theta}}\mathbf{U}\mathbf{U}^\top$. Hence, standard PCA minimizing the squared error in (1) can be viewed as a technique for minimizing the normal deviance by projecting the natural parameters of the saturated model into a lower dimensional space.

3.2. Alternative formulation to logistic PCA

When the data are binary, assume instead that x_{ij} are from Bernoulli(p_{ij}). The natural parameter for the Bernoulli distribution is $\theta_{ij} = \text{logit}\, p_{ij}$. Let $\tilde{\theta}_{ij}$ represent the natural parameter of the saturated model. The saturated model occurs when $p_{ij} = x_{ij}$, which means that

$$\tilde{\theta}_{ij} = \begin{cases} -\infty & \text{if } x_{ij} = 0 \\ \infty & \text{if } x_{ij} = 1. \end{cases}$$

To apply an equivalent principal component analysis to binary data, we need to minimize the Bernoulli deviance by projecting the natural parameters of the saturated model onto a k-dimensional space. For convenience, define $q_{ij} = 2x_{ij} - 1$, which converts the binary variable from taking values in $\{0, 1\}$ to $\{-1, 1\}$. Let $\mathbf{Q} = 2\mathbf{X} - \mathbf{1}_n\mathbf{1}_d^\top$ be the matrix with elements q_{ij}. For practical purposes, we introduce a positive constant m for numerical approximation of the saturated model parameter $\tilde{\theta}_{ij}$ by $m \cdot q_{ij}$, and therefore approximate the matrix of natural parameters for the saturated model $\tilde{\boldsymbol{\Theta}}$ by $m\mathbf{Q}$. The choice of m results in the value of the sigmoid function at m, $\exp(m)/(1 + \exp(m))$, as approximation of the saturated probability parameter $p = 1$. For example, $m = 1$ corresponds to $p \approx 0.7311$, $m = 3$ to 0.9526, $m = 5$ to 0.9933, and $m = 10$ to 0.9999, which suggests that m does not need to be very large for accurate approximation of the saturated

model parameters in practice. Taking m as a tunable parameter, we show how the choice of m affects the analysis in Section 6.1.

Let $D(\mathbf{X}; \Theta)$ denote the deviance of estimated natural parameter matrix Θ with the data matrix $\mathbf{X}$. As in standard PCA, the natural parameters are estimated with a matrix of the form $\Theta = \mathbf{1}_n\boldsymbol{\mu}^\top + (\tilde{\Theta} - \mathbf{1}_n\boldsymbol{\mu}^\top)\mathbf{U}\mathbf{U}^\top$. The objective function to minimize is the Bernoulli deviance,

$$\begin{aligned} D(\mathbf{X}; \mathbf{1}_n\boldsymbol{\mu}^\top + (\tilde{\Theta} - \mathbf{1}_n\boldsymbol{\mu}^\top)\mathbf{U}\mathbf{U}^\top) &= -2\left(\ln p(\mathbf{X}; \mathbf{1}_n\boldsymbol{\mu}^\top + (\tilde{\Theta} - \mathbf{1}_n\boldsymbol{\mu}^\top)\mathbf{U}\mathbf{U}^\top) - \ln p(\mathbf{X}; \tilde{\Theta})\right) \\ &= -2\langle \mathbf{X}, \mathbf{1}_n\boldsymbol{\mu}^\top + (\tilde{\Theta} - \mathbf{1}_n\boldsymbol{\mu}^\top)\mathbf{U}\mathbf{U}^\top\rangle + 2\sum_{i=1}^{n}\sum_{j=1}^{d} \ln\left(1 + \exp(\mu_j + [\mathbf{U}\mathbf{U}^\top(\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\mu})]_j)\right), \end{aligned} \tag{3}$$

subject to $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_k$, where $\langle \mathbf{A}, \mathbf{B}\rangle = \mathrm{tr}(\mathbf{A}^\top\mathbf{B})$ is the trace inner product. Using appropriate deviance, this methodology can be extended to any distribution in the exponential family. See [10] for extension.

3.3. Comparison to previous techniques

The main advantage of the proposed formulation is that we only solve for the principal component loadings and not simultaneously for the principal component scores. The previous method for logistic PCA posits that the logit of the probability matrix, logit $\mathbf{P}$, can be represented by a low-rank matrix factorization $\Theta = \mathbf{A}\mathbf{B}^\top$, assuming $\boldsymbol{\mu} = \mathbf{0}$ here to simplify exposition. Our formulation, on the other hand, assumes the logit of the probability matrix has the form $\Theta = \tilde{\Theta}\mathbf{U}\mathbf{U}^\top$.

To highlight the difference between the two formulations, we will call the previous formulation logistic SVD (LSVD) and our formulation logistic PCA (LPCA). The $d \times k$ principal component loading matrices for LSVD and LPCA are $\mathbf{B}$ and $\mathbf{U}$, respectively, and the $n \times k$ matrices of principal component scores are $\mathbf{A}$ and $\tilde{\Theta}\mathbf{U}$. The loading matrices are comparable, but the score matrices take different forms. The form of $\tilde{\Theta}\mathbf{U}$, along with m, can act as an implicit regularizer. While there is no restriction on how large the elements of $\mathbf{A}$ can be, the elements of $\tilde{\Theta}\mathbf{U}$ are bound between $-m\sqrt{d}$ and $m\sqrt{d}$, for instance.

We illustrate a number of advantages of the alternative formulation. Conceptually, when the main effects are not included in logistic SVD, the cases and variables are treated interchangeably. That is, an analysis of $\mathbf{X}$ will produce the same low-rank fitted matrix as an analysis of $\mathbf{X}^\top$, and the loadings of $\mathbf{X}$ will equal the scaled scores of $\mathbf{X}^\top$ and vice versa. Logistic PCA, however, will very likely have a different solution for the respective loadings and scores. Since the intent of PCA is to reduce the data dimensionality using the relation among the variables primarily, we believe that it is desirable to maintain the inherent difference between cases and variables in the analysis.

Another difference is that the proposed formulation allows quick and easy evaluation of principal component scores for new data. Let $\mathbf{x}^* \in \{0, 1\}^d$ be a new observation. We wish to calculate the principal component scores for this new data point assuming that the loadings have already been estimated using another dataset. Let $\hat{\mathbf{B}}$ and $\hat{\mathbf{U}}$ be the principal component loadings estimated from the LSVD and LPCA formulations, respectively. To determine the new principal component scores, $\mathbf{a}^* \in \mathbb{R}^k$, for LSVD, one needs to find

$$\mathbf{a}^* = \underset{\mathbf{a}\in\mathbb{R}^k}{\arg\min} \sum_{j=1}^{d}\left[-x_j^*(\mathbf{a}^\top\hat{\mathbf{b}}_{j\cdot}) + \ln\left(1 + \exp(\mathbf{a}^\top\hat{\mathbf{b}}_{j\cdot})\right)\right],$$

where $\hat{\mathbf{b}}_{j\cdot}$ is the jth row of $\hat{\mathbf{B}}$. This is equivalent to a logistic regression problem, where the d-dimensional response vector is $\mathbf{x}^*$ and the $d \times k$ design matrix is $\hat{\mathbf{B}}$. The $\mathbf{a}^*$ can be viewed as the coefficient vector that maximizes the likelihood defined through $\mathbf{x}^*$.

By contrast, the LPCA formulation only requires a matrix multiplication for the new principal component scores: $\hat{\mathbf{U}}^\top\tilde{\boldsymbol{\theta}}^*$, where $\tilde{\boldsymbol{\theta}}^* := m(2\mathbf{x}^* - 1) := m\mathbf{q}^*$ is taken as the approximate natural parameters for $\mathbf{x}^*$ under the saturated model. This process is analogous to computing the principal component scores for new data in standard PCA, where $\mathbf{x}^*$ itself acts as $\tilde{\boldsymbol{\theta}}^*$, a set of the natural parameters for the saturated model. Further, predicting a low-rank estimate of the natural parameters on a set of new data only requires calculating

$$\hat{\Theta}^* = \tilde{\Theta}^*\hat{\mathbf{U}}\hat{\mathbf{U}}^\top. \tag{4}$$

Quick evaluation of principal component scores for new data can be particularly useful in a number of situations. For example, principal component regression (or classification) (see [4], §3.5) can be extended to logistic PCA when all the covariates are binary. If it is necessary to make predictions for a large amount of new data or predictions are required in real-time, the LSVD method may be too slow. Further, our proposed method will be much more efficient when the number of principal components to retain is chosen by cross validation [9, §6.1.5]. In this case, we select the number of components that best reconstruct the dataset on held-out observations, and the cross validation requires applying logistic PCA to new data repeatedly.

Another major difference between this formulation and the previous one is that the alternative formulation entails much fewer parameters. In particular, LSVD has $kn - k(k-1)/2$ additional parameters in the $\mathbf{A}$ matrix, if the columns are constrained to be orthogonal to each other. This additional number of incidental parameters could potentially be very large and be ripe for over-fitting. As [23] discussed, logistic SVD can be viewed as an estimation method for a factor model. Instead of marginalizing over the case-specific factors, which are latent variables in factor analysis, logistic SVD takes a degenerate approach to computing point estimates for them. Since the number of latent factors is proportional to the number of observations, overfitting can easily occur.

The alternative formulation of logistic PCA does have an additional parameter, m, that previous formulations do not. We treat m as a tuning parameter. As m gets larger, the estimated probabilities defined through the elements of $\tilde{\Theta}$ will be close to 0 or 1. Conversely, if m is small, the probability estimates will be close to 0.5. If the user has domain knowledge of the range of the likely probabilities, they can use this to guide their choice of m. We have found that cross validation is an effective way to choose m. Simulations in Section 6.1 show the potential benefits of properly choosing m.

Because LSVD has many more parameters than LPCA given a rank, LSVD is guaranteed to have a lower in-sample deviance. Despite this, the simulations in Section 6.1 show that LPCA can do just as well or better at estimating the true probabilities if m and k are chosen properly.

4. Logistic PCA for patterned data

The properties of standard PCA solutions are well understood algebraically. Under certain assumptions, the solutions are explicitly known. For example, when the variables are uncorrelated, the loadings are the standard bases and the principal components are ordered from highest variance to lowest. By contrast, not much is known about the solutions of logistic PCA or logistic SVD. To obtain analogous results for patterned data with logistic PCA, we derive necessary conditions for the solutions first, and find solutions that satisfy (or nearly satisfy) these optimality conditions under different sets of assumptions on data matrices. These results help us gain a better understanding of logistic PCA.

4.1. First-order optimality conditions

To enforce orthonormality $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_k$, we add a constraint to the objective in (3) via the method of Lagrange multipliers. The Lagrangian is

$$L(\mathbf{U}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = D(\mathbf{X}; \mathbf{1}_n\boldsymbol{\mu}^\top + (\tilde{\Theta} - \mathbf{1}_n\boldsymbol{\mu}^\top)\mathbf{U}\mathbf{U}^\top) + \operatorname{tr}\left(\boldsymbol{\Lambda}(\mathbf{U}^\top\mathbf{U} - \mathbf{I}_k)\right),$$

where $\boldsymbol{\Lambda}$ is a $k \times k$ symmetric matrix of Lagrange multipliers [24].

Taking the gradient of the Lagrangian with respect to $\mathbf{U}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Lambda}$, and setting them equal to $\mathbf{0}$, we obtain the first-order optimality conditions for the solution of logistic PCA:

$$\left[\left(\mathbf{X} - \hat{\mathbf{P}}\right)^\top\left(\tilde{\Theta} - \mathbf{1}_n\boldsymbol{\mu}^\top\right) + \left(\tilde{\Theta} - \mathbf{1}_n\boldsymbol{\mu}^\top\right)^\top\left(\mathbf{X} - \hat{\mathbf{P}}\right)\right]\mathbf{U} = \mathbf{U}\boldsymbol{\Lambda} \tag{5}$$

$$\left(\mathbf{I}_d - \mathbf{U}\mathbf{U}^\top\right)\left(\mathbf{X} - \hat{\mathbf{P}}\right)^\top\mathbf{1}_n = \mathbf{0}_d \tag{6}$$

$$\mathbf{U}^\top\mathbf{U} = \mathbf{I}_k. \tag{7}$$

The matrix $\hat{\mathbf{P}}$ has the estimate of p_{ij} at $\mathbf{U}$ and $\boldsymbol{\mu}$, with $\hat{p}_{ij} = \sigma\left(\mu_j + [\mathbf{U}\mathbf{U}^\top(\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\mu})]_j\right)$ as its ijth element, where $\sigma(\theta) = \exp(\theta)/(1+\exp(\theta))$ is the inverse logit (or sigmoid) function. The details of the calculation can be found in Section 8.1. For the following sections, let

$$\mathbf{C}^m := \left(\mathbf{X} - \hat{\mathbf{P}}\right)^\top\left(\tilde{\Theta} - \mathbf{1}_n\boldsymbol{\mu}^\top\right) + \left(\tilde{\Theta} - \mathbf{1}_n\boldsymbol{\mu}^\top\right)^\top\left(\mathbf{X} - \hat{\mathbf{P}}\right),$$

which is labeled to explicitly state the dependence of $\tilde{\Theta}$ and $\hat{\mathbf{P}}$ on m.

Applying the Lagrangian method of multipliers to the standard PCA formulation (1) yields

$$\left[\left(\mathbf{X} - \mathbf{1}_n\boldsymbol{\mu}^\top\right)^\top\left(\mathbf{X} - \mathbf{1}_n\boldsymbol{\mu}^\top\right)\right]\mathbf{U} = \mathbf{U}\boldsymbol{\Lambda}$$

as part of the first-order optimality conditions, which is very similar to the form of Eq. (5). Unlike standard PCA, (5) is nonlinear in $\mathbf{U}$ and the solution is not known in closed form through an eigen-decomposition because the matrix in the left hand side of (5) depends on $\mathbf{U}$ through $\hat{\mathbf{P}}$. However, we can derive some explicit results for special cases using the optimality conditions.

4.2. Independence

There are a few natural extremes of data dependence that a given binary dataset can exhibit. On one end of the spectrum, all of the columns of $\mathbf{Q}$ can be in the span of a single vector such that $\mathbf{Q} = \mathbf{a}\mathbf{b}^\top$ with $\mathbf{a} \in \{\pm 1\}^n$ and $\mathbf{b} \in \{\pm 1\}^d$. Then projection of $\tilde{\Theta}$ along the direction of $\mathbf{u} = \mathbf{b}/\|\mathbf{b}\|$ gives $\hat{\Theta} = \tilde{\Theta}\mathbf{u}\mathbf{u}^\top = \tilde{\Theta}$. Thus with a rank-one approximation, the saturated model parameters can be reconstructed with arbitrary precision.

The opposite end of the spectrum is when the variables are independent of each other. In standard PCA, this implies that the covariance matrix is diagonal, so the principal component loadings are the standard basis vectors. Analogous to standard PCA results, we show below that, if the ℓth column of a dataset is uncorrelated with the other $d-1$ columns and its column mean is $1/2$, then the ℓth standard basis vector, $\mathbf{e}_\ell$, satisfies the first-order optimality conditions of logistic PCA with $k = 1$. If its column mean is not equal to $1/2$, $\mathbf{e}_\ell$ can nearly satisfy the optimality conditions with large enough m.

Let X_j be the length n vector of the jth column of $\mathbf{X}$ and $\bar{x}_j = X_j^\top \mathbf{1}_n/n$ be the corresponding column mean.

Theorem 1. *Assume that $X_\ell^\top X_j/n = \bar{x}_\ell \bar{x}_j$, for all $j \neq \ell$, i.e., the ℓth variable is uncorrelated with all other variables.*

(i) *If $\bar{x}_\ell = 1/2$, then $\mathbf{u} = \mathbf{e}_\ell$, the ℓth standard basis vector, satisfies the first-order optimality conditions of logistic PCA, regardless of m. That is,*

$$\mathbf{C}^m \mathbf{e}_\ell - \lambda_m \mathbf{e}_\ell = \mathbf{0}, \qquad \text{for some } \lambda_m.$$

(ii) *If $\bar{x}_\ell \neq 1/2$, then the first-order optimality conditions can be satisfied as close as desired with $\mathbf{u} = \mathbf{e}_\ell$ for m large enough. Formally, for any $\epsilon > 0$, there exists m_0 such that, for all $m > m_0$,*

$$\|\mathbf{C}^m \mathbf{e}_\ell - \lambda_m \mathbf{e}_\ell\|^2 < \epsilon, \qquad \text{with } \lambda_m = 0.$$

The proofs of Theorem 1 and subsequent theorems are given in Section 8.2. If multiple columns are uncorrelated with the remaining columns, this result easily generalizes to larger k. For example, if k columns are uncorrelated with all other columns, then a rank k solution comprising of the corresponding k standard basis vectors can be made arbitrarily close to (or exactly if the column means equal $1/2$) satisfying the necessary conditions (5)–(7). This leads to a natural question: when there are multiple candidate solutions, which one decreases the deviance the most?

Theorem 2. *For logistic PCA with $k = 1$, the standard basis vector which decreases deviance the most is the one corresponding to column with mean closest to $1/2$.*

This result corresponds to the ordering of variables by variance in standard PCA. The variables with means closest to $1/2$ have the largest variance for binary data. It is because the sample variance of X_j is given by $\bar{x}_j(1-\bar{x}_j)$ and the Bernoulli variance $p(1-p)$ decreases as p moves away from $1/2$. If the variables are independent, the deviance explained will be largest with a standard basis vector as principal component loadings corresponding to the variable with largest variance. Similar to the previous theorem, this theorem can be easily extended to k larger than 1. In this case, the loading matrix made up of the standard basis vectors corresponding to the k columns that are closest to $1/2$ will decrease the deviance the most out of all loading matrices that comprise of standard basis vectors.

4.3. Compound symmetry

With independence and perfect correlation being two extremes of the structure of the data, somewhere in the middle is compound symmetry. A covariance matrix Σ is compound symmetric if the diagonals are constant ($\Sigma_{jj} = c_1$ for all j) and the off-diagonals are all equal to each other ($\Sigma_{jk} = c_2$ for all $j \neq k$). The compound symmetry of Σ implies equal correlations among the variables. If Σ is compound symmetric, $\frac{1}{\sqrt{d}}\mathbf{1}_d$ is an eigenvector of the covariance matrix because $\left[(c_1 - c_2)\mathbf{I}_d + c_2 \mathbf{1}_d \mathbf{1}_d^\top\right] \mathbf{1}_d = \{c_1 + (d-1)c_2\}\mathbf{1}_d$. We show that, under more limiting conditions, $\frac{1}{\sqrt{d}}\mathbf{1}_d$ satisfies the optimality conditions for logistic PCA when $\mathbf{Q}^\top\mathbf{Q}$ is compound symmetric.

$\mathbf{Q}^\top\mathbf{Q}$ has a natural interpretation. The diagonals always equal n and the jkth off-diagonal is the number of records in which the jth and kth variables are the same minus the number of records in which the jth and kth variables differ. It measures how much the jth and kth variables agree with each other, and can range from $-n$ (total disagreement) to n (total agreement). $\mathbf{Q}^\top\mathbf{Q}$ is compound symmetric if all the bivariate agreements are the same.

Theorem 3. *Consider logistic PCA without main effects, where $\boldsymbol{\mu} = \mathbf{0}$. Assume that $\mathbf{Q}^\top\mathbf{Q}$ is compound symmetric. If β's exist such that the following condition is satisfied,*

$$\sigma\left(\frac{m}{d}\sum_{\ell \notin \{j,k\}} q_{i\ell}\right) = \frac{1}{2} + \sum_{\ell \notin \{j,k\}} q_{i\ell}\beta_{jk,\ell}, \text{ for all } j \neq k,\ j,k \in \{1,\ldots,d\}, i \in \{1,\ldots,n\} \tag{8}$$

then $\mathbf{u} = \frac{1}{\sqrt{d}}\mathbf{1}_d$ satisfies the first-order optimality conditions for logistic PCA, as characterized by Eqs. (5)–(7).

Eq. (8) is a condition that makes $\mathbf{C}^m$ with $\boldsymbol{\mu} = \mathbf{0}$ and $\mathbf{u} = \frac{1}{\sqrt{d}}\mathbf{1}_d$ compound symmetric. As a special case, if $X_1 = \cdots = X_d$, then the columns of $\mathbf{Q}$ are the same and the condition is easily met. Generally, when there are fewer columns, the condition is more likely to be met.

Corollary 1. *If $d \leq 4$ and $\mathbf{Q}^\top\mathbf{Q}$ is compound symmetric, then (8) is satisfied and therefore $\mathbf{u} = \frac{1}{\sqrt{d}}\mathbf{1}_d$ satisfies the first-order optimality conditions. When $d = 2$, $\mathbf{Q}^\top\mathbf{Q}$ is always compound symmetric.*

5. Computation

Optimizing for the principal component loadings of logistic PCA is difficult because of the non-convex objective function and the orthonormality constraint. We derive an algorithm for generating solutions which decreases the deviance at each iteration. One approach to minimizing the deviance is to iteratively minimize simpler objectives. Majorization–minimization [6] seeks to solve difficult optimization problems by majorizing the objective function, $L(\theta)$, with a simpler objective, $M(\theta|\theta^{(t)})$, and minimizing the majorizing function iteratively. The majorization function must be equal to or greater than the original objective for all inputs and equal to it at the current input value $\theta^{(t)}$: $L(\theta) \leq M(\theta|\theta^{(t)})$ for all θ, and $L(\theta^{(t)}) = M(\theta^{(t)}|\theta^{(t)})$.

The deviance of a single estimated natural parameter θ is quadratically approximated at $\theta^{(t)}$ by

$$\begin{aligned}
-2\ln p(x;\theta) &= -2x\theta + 2\ln(1+\exp(\theta)) \\
&\approx -2x\theta^{(t)} + 2\ln(1+\exp(\theta^{(t)})) + 2(\hat{p}^{(t)} - x)(\theta - \theta^{(t)}) + \hat{p}^{(t)}(1-\hat{p}^{(t)})(\theta-\theta^{(t)})^2 \\
&\leq -2x\theta^{(t)} + 2\ln(1+\exp(\theta^{(t)})) + 2(\hat{p}^{(t)} - x)(\theta - \theta^{(t)}) + \frac{1}{4}(\theta-\theta^{(t)})^2,
\end{aligned}$$

where $\hat{p}^{(t)} = \sigma(\theta^{(t)})$. The inequality is due to the variance of a Bernoulli random variable being bounded above by 1/4. [13] showed that the deviance itself is majorized by this same function.

Therefore, the deviance for the whole matrix is majorized by

$$\sum_{i,j}\left\{-2x_{ij}\theta_{ij}^{(t)} + 2\ln(1+\exp(\theta_{ij}^{(t)})) + 2(\hat{p}_{ij}^{(t)} - x_{ij})(\theta_{ij} - \theta_{ij}^{(t)}) + \frac{1}{4}(\theta_{ij}-\theta_{ij}^{(t)})^2\right\} = \frac{1}{4}\sum_{i,j}(\theta_{ij} - z_{ij}^{(t)})^2 + C,$$

where C is a constant that does not depend on θ_{ij}, and $z_{ij}^{(t)} = \theta_{ij}^{(t)} + 4[x_{ij} - \sigma(\theta_{ij}^{(t)})]$ are the working variables in the tth iteration. Further, let $\mathbf{Z}^{(t)}$ be a matrix whose ijth element equals $z_{ij}^{(t)}$. The working variables have a similar form to the so-called adjusted response used in the iteratively reweighted least squares algorithm for generalized linear models [15]. Instead of having weights equal to the estimated variance at the current estimates, we use the upper bound, which allows for minimization of the majorization function.

The logistic PCA objective function can be majorized around estimates of $\mathbf{U}^{(t)}$ and $\boldsymbol{\mu}^{(t)}$ as

$$D(\mathbf{X};\mathbf{1}_n\boldsymbol{\mu}^\top + (\tilde{\boldsymbol{\Theta}} - \mathbf{1}_n\boldsymbol{\mu}^\top)\mathbf{U}\mathbf{U}^\top) \leq \frac{1}{4}\sum_{i,j}\left(\mu_j + [\mathbf{U}\mathbf{U}^\top(\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\mu})]_j - z_{ij}^{(t)}\right)^2 + C,$$

and hence, the next iterates of $\mathbf{U}^{(t+1)}$ and $\boldsymbol{\mu}^{(t+1)}$ can be obtained by minimizing

$$\sum_{i,j}\left(\mu_j + [\mathbf{U}\mathbf{U}^\top(\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\mu})]_j - z_{ij}^{(t)}\right)^2 = \|\mathbf{1}_n\boldsymbol{\mu}^\top + (\tilde{\boldsymbol{\Theta}} - \mathbf{1}_n\boldsymbol{\mu}^\top)\mathbf{U}\mathbf{U}^\top - \mathbf{Z}^{(t)}\|_F^2. \tag{9}$$

Given initial estimates for $\mathbf{U}$ and $\boldsymbol{\mu}$, a solution can be found by iteratively minimizing Eq. (9), subject to orthonormality constraint $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_k$. With fixed $\boldsymbol{\mu}$, the minimizer of the majorizing function can be found by expanding Eq. (9). Letting $\tilde{\boldsymbol{\Theta}}_c := \tilde{\boldsymbol{\Theta}} - \mathbf{1}_n\boldsymbol{\mu}^\top$ and $\mathbf{Z}_c^{(t)} := \mathbf{Z}^{(t)} - \mathbf{1}_n\boldsymbol{\mu}^\top$, we see that

$$\begin{aligned}
&\arg\min_{\mathbf{U}^\top\mathbf{U}=\mathbf{I}_k} \|\tilde{\boldsymbol{\Theta}}_c\mathbf{U}\mathbf{U}^\top - \mathbf{Z}_c^{(t)}\|_F^2 = \operatorname{argmin}\ \left\{\operatorname{tr}\left(\mathbf{U}\mathbf{U}^\top\tilde{\boldsymbol{\Theta}}_c^\top\tilde{\boldsymbol{\Theta}}_c\mathbf{U}\mathbf{U}^\top\right) - \operatorname{tr}\left(\mathbf{U}\mathbf{U}^\top\tilde{\boldsymbol{\Theta}}_c^\top\mathbf{Z}_c^{(t)}\right) - \operatorname{tr}\left((\mathbf{Z}_c^{(t)})^\top\tilde{\boldsymbol{\Theta}}_c\mathbf{U}\mathbf{U}^\top\right)\right\} \\
&= \operatorname{argmin}\ \operatorname{tr}\left(\mathbf{U}^\top\left(\tilde{\boldsymbol{\Theta}}_c^\top\tilde{\boldsymbol{\Theta}}_c - \tilde{\boldsymbol{\Theta}}_c^\top\mathbf{Z}_c^{(t)} - (\mathbf{Z}_c^{(t)})^\top\tilde{\boldsymbol{\Theta}}_c\right)\mathbf{U}\right) = \operatorname{argmax}\ \operatorname{tr}\left(\mathbf{U}^\top\left(\tilde{\boldsymbol{\Theta}}_c^\top\mathbf{Z}_c^{(t)} + (\mathbf{Z}_c^{(t)})^\top\tilde{\boldsymbol{\Theta}}_c - \tilde{\boldsymbol{\Theta}}_c^\top\tilde{\boldsymbol{\Theta}}_c\right)\mathbf{U}\right).
\end{aligned}$$

The trace is maximized when $\mathbf{U}$ consists of the first k eigenvectors of the symmetric matrix $\tilde{\boldsymbol{\Theta}}_c^\top\mathbf{Z}_c^{(t)} + (\mathbf{Z}_c^{(t)})^\top\tilde{\boldsymbol{\Theta}}_c - \tilde{\boldsymbol{\Theta}}_c^\top\tilde{\boldsymbol{\Theta}}_c$ (see [3]).

With fixed $\mathbf{U}$, the minimization of the majorizing function with respect to $\boldsymbol{\mu}$ is a least squares problem, and the majorizing function is minimized by $\boldsymbol{\mu} = (\mathbf{Z}^{(t)} - \tilde{\boldsymbol{\Theta}}\mathbf{U}\mathbf{U}^\top)^\top\mathbf{1}_n/n$, which can be interpreted as the average differences between the projection of the uncentered saturated natural parameters and the current working variables. Algorithm 1 presents the MM algorithm for logistic PCA.

Since the optimization problem is non-convex, finding the global minimum is difficult in general. Due to the fact that the majorization function is above the deviance and tangent to it at the previous iteration, minimizing the majorization

input : Binary data matrix ($\mathbf{X}$), m, number of principal components (k), convergence criterion (ϵ)
output: $d \times k$ orthonormal matrix of principal component loadings ($\mathbf{U}$) and column main effects ($\boldsymbol{\mu}$)

Set $t = 0$ and initialize $\boldsymbol{\mu}^{(0)}$ and $\mathbf{U}^{(0)}$. We recommend setting $\mu_j^{(0)} = \text{logit}\, \bar{x}_j$ and setting $\mathbf{U}^{(0)}$ to the first k right singular vectors of $\mathbf{Q}$
repeat

1. $t \leftarrow t + 1$
2. Set the working variables
   $z_{ij}^{(t)} = \theta_{ij} + 4[x_{ij} - \sigma(\theta_{ij})]$ where $\theta_{ij} = \mu_j^{(t-1)} + [\mathbf{U}^{(t-1)}(\mathbf{U}^{(t-1)})^\top(\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\mu}^{(t-1)})]_j$
3. $\boldsymbol{\mu}^{(t)} = \frac{1}{n}(\mathbf{Z}^{(t)} - \tilde{\boldsymbol{\Theta}}\mathbf{U}^{(t-1)}(\mathbf{U}^{(t-1)})^\top)^\top \mathbf{1}_n$
4. Carry out the eigen decomposition of
$$\begin{aligned}&(\tilde{\boldsymbol{\Theta}} - \mathbf{1}_n(\boldsymbol{\mu}^{(t)})^\top)^\top(\mathbf{Z}^{(t)} - \mathbf{1}_n(\boldsymbol{\mu}^{(t)})^\top) + (\mathbf{Z}^{(t)} - \mathbf{1}_n(\boldsymbol{\mu}^{(t)})^\top)^\top(\tilde{\boldsymbol{\Theta}} - \mathbf{1}_n(\boldsymbol{\mu}^{(t)})^\top)\\ &\quad - (\tilde{\boldsymbol{\Theta}} - \mathbf{1}_n(\boldsymbol{\mu}^{(t)})^\top)^\top(\tilde{\boldsymbol{\Theta}} - \mathbf{1}_n(\boldsymbol{\mu}^{(t)})^\top)\\ &\quad = \mathbf{E}\boldsymbol{\Lambda}\mathbf{E}^\top\end{aligned}$$
   and set $\mathbf{U}^{(t)}$ to the first k eigenvectors in $\mathbf{E}$

until *Deviance converges*;

Algorithm 1: Majorization–minimization algorithm for logistic PCA.

function at each iteration must either decrease the deviance or cause no change at each iteration. While the majorization–minimization does not guarantee a global minimum generally, the quadratic majorization function and the smooth objective does guarantee finding a local minimum [11, Chapter 12].

An R [18] implementation of the MM algorithm for logistic PCA can be found at cran.r-project.org/web/packages/logisticPCA.

6. Numerical examples

In this section, we demonstrate how the logistic PCA formulation works on both simulated and real data. In particular, we numerically examine the differences between the previous formulation and the proposed formulation, the effect of m, and the effectiveness of the MM algorithm.

6.1. Simulation

For comparison of our formulation of logistic PCA (LPCA) with the previous formulation (LSVD), we simulate binary matrices from a family of multivariate Bernoulli mixtures that induce a low-rank structure in both the true probability matrix and the logit matrix. The components or clusters of the Bernoulli mixtures are determined by cluster-specific probability vectors.

Simulation setup

Let k indicate the number of mixture components or clusters of the probability vectors for d-variate binary random vectors making up an $n \times d$ data matrix $\mathbf{X}$. For a specified n, d, and k, let C_i be the cluster assignment for the ith observation, $i \in \{1, \ldots, n\}$, which takes one of k values $\{1, \ldots, k\}$ with equal probability. Further, we specify k true probability vectors, $\boldsymbol{p}_c \in [0, 1]^d$, for $c \in \{1, \ldots, k\}$. In our simulations, the probabilities are independently generated from a Beta(α, β) distribution and X_{ij} given $C_i = c$ are independently generated from Bernoulli(p_{cj}). For a Beta distribution, instead of the shape parameters α and β, we vary the mean $\bar{p} = \alpha/(\alpha + \beta)$ and concentration parameter $\phi = (\alpha + \beta)/2$, which is inversely related to the variance of $\bar{p}(1 - \bar{p})/(2\phi + 1)$. If we let $\mathbf{A}$ be an $n \times k$ indicator matrix with $A_{ic} = 1_{\{C_i = c\}}$ and $\mathbf{B}_*$ be a $d \times k$ matrix with the cth column equal to $\boldsymbol{p}_c$, then the true probability matrix is $\mathbf{P} = [p_{ij}] = \mathbf{A}\mathbf{B}_*^\top$, or equivalently, the logit matrix is $\boldsymbol{\Theta} = \text{logit}\, \mathbf{P} = \mathbf{A}\mathbf{B}^\top$ with $\mathbf{B} = \text{logit}\, \mathbf{B}_*$. The accuracy of the approximation of $\mathbf{P}$ through logistic PCA depends on the number of principal components considered and the value of parameter m. To reduce confusion between the true number of clusters k and the number of principal components considered, we will adopt the notation $\hat{k}$ for the latter. For all simulations in this section, we set $n = 100$ and $d = 50$. For simulations with $\bar{p} = 0.5$, we did not include the main effects $\boldsymbol{\mu}$ for either LPCA or LSVD to minimize the differences in implementations. The main effects are likely close to $\mathbf{0}$ when $\bar{p} = 0.5$.

Estimate of true probabilities

For the numerical study, we simulated 12 binary matrices from a variety of different cluster models with $\bar{p} = 0.5$. We varied the true number of clusters ($k \in \{2, 3, 5, 10\}$) and the concentration parameter for the probabilities ($\phi \in$

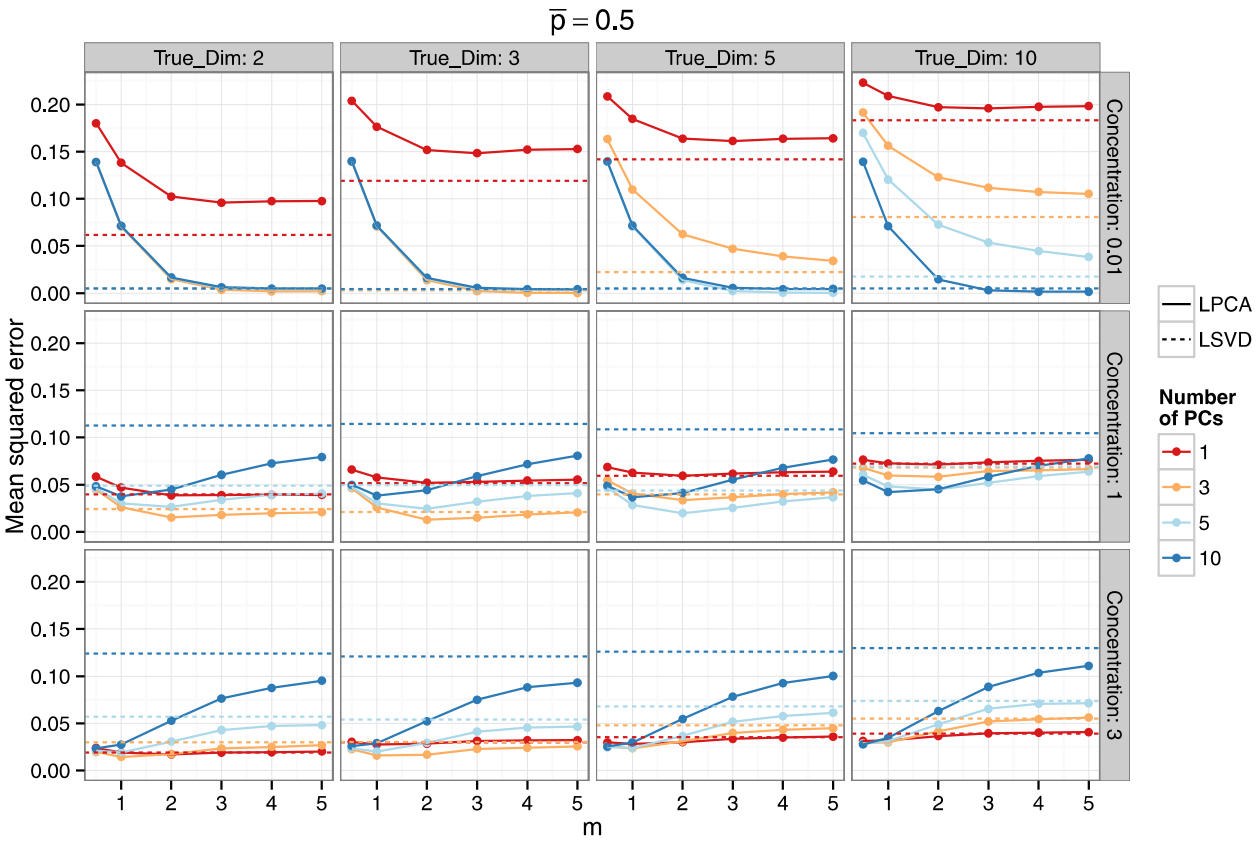

Fig. 1. Mean squared error of probability estimates derived from logistic PCA with varying m and logistic SVD in a simulation experiment where the rank of the true probability matrix ranges from 2 to 10 (from left to right) and the extent of concentration of the true probabilities varies from low to high (from top to bottom). The number of principal components ranges from 1 to 10 (from red to blue) for each combination of the true rank and concentration parameter. The solid lines are for logistic PCA (LPCA) while the dotted lines are for logistic SVD (LSVD). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$\{0.01, 1, 3\}$). For each of the 12 data matrices, we performed dimensionality reduction with LSVD and LPCA. For both, we varied the estimated number of components ($\hat{k} \in \{1, 3, 5, 10\}$) and for LPCA we also varied m from 0.5 to 5. Again, for LPCA, we used the MM algorithm and for LSVD we used the iterative SVD algorithm from [13].

After estimating the probability matrix as $\hat{\mathbf{P}}$, we compared it to the true probability matrix $\mathbf{P}$ by taking the element-wise mean squared error $\|\hat{\mathbf{P}} - \mathbf{P}\|_F^2/(nd)$. Fig. 1 shows the results. When the concentration parameter ϕ is low (the top row), most of the true probabilities are close to 0 or 1 with not much in between. In this situation, having the correct estimated dimension $\hat{k}$ is crucial for both methods and, in fact, having $\hat{k} > k$ does no harm. Further, for LPCA, higher m is better for estimating the probability matrix in this situation. Both of these results are in line with our expectations. When the rank of the estimate is allowed to be higher, the estimates are able to fit the data more closely in general. In doing so, the resulting estimated probabilities will be close to 0 and 1. As stated in Section 3.3, higher values of m enable estimates that are closer to 0 and 1 as well.

On the other hand, if the concentration is high, most of the true probabilities will be close to $\bar{p} = 0.5$. The bottom row of Fig. 1 shows the results when this is the case. The opposite conclusions are reached from this scenario. In general, having a lower rank or a lower m is better.

When the concentration is moderate ($\phi = 1$, the middle row of Fig. 1), the true probabilities are generated from a uniform distribution. For LPCA, having the correct estimated dimension is important, but so is m. For each of the different true dimensions, the lowest mean squared error is achieved when the estimated dimension matches the true dimension. This is not true for all values of m, as when $k = 10$, the estimate with $\hat{k} = 10$ is poor for large m. For each of the estimated ranks, there is a local minimum of MSE for choosing m. In contrast to the high and low concentration cases, the results for LSVD do not mirror those of LPCA for the same $\hat{k}$. Here, an estimated rank of $\hat{k} = 3$ is best for all of the true ranks.

Finally, for any of the given dataset simulations, there is a combination of m and $\hat{k}$ for which LPCA had a mean squared error as small as or smaller than any LSVD estimate. The challenge obviously is to determine the optimal combination data-adaptively in applications. We show that for each $\hat{k}$, using five-fold cross validation is an effective way to choose m. To accomplish this, we randomly split the rows into five groups, say, $G_1, \ldots, G_5$ with G_j for the jth group of row indices. For each of the groups, we perform LPCA with a given m on all of the data except the rows in that group. Let $\hat{\mathbf{U}}_m^{[-j]}$ denote

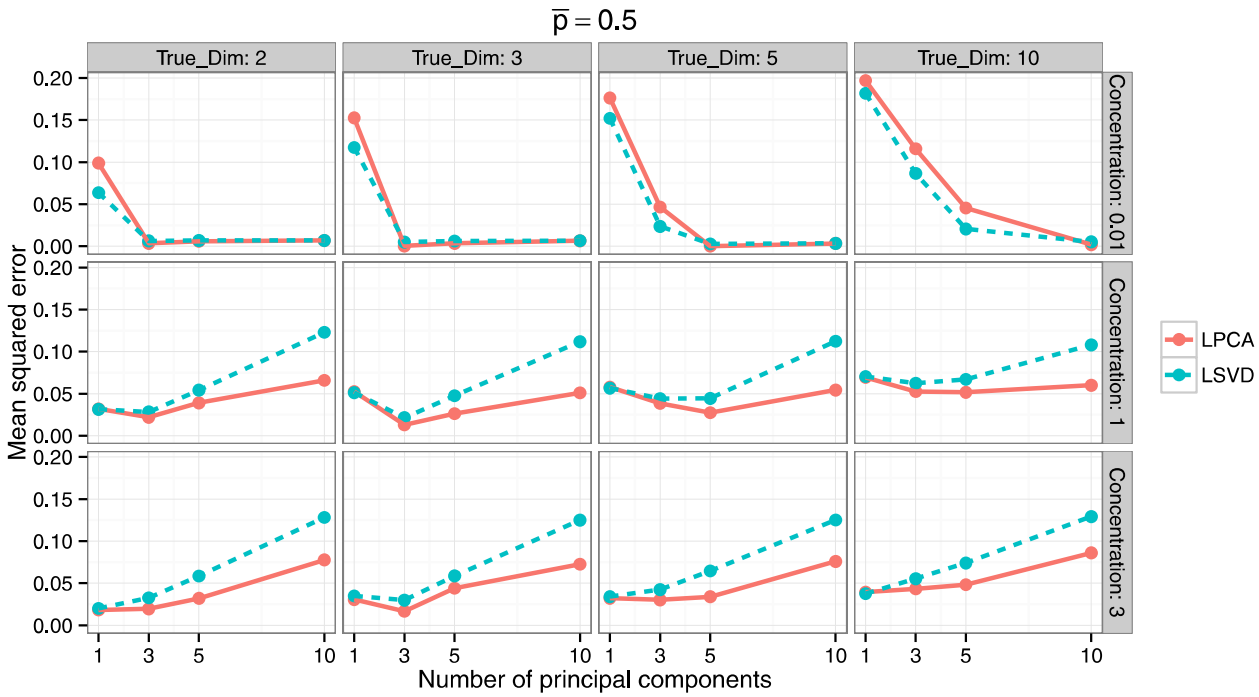

Fig. 2. Mean squared error of probability estimates derived from logistic PCA and logistic SVD in the simulation experiment where m in logistic PCA is chosen by cross validation. The rank of the true probability matrix ranges from 2 to 10 (from left to right) and the extent of concentration of the true probabilities varies from low to high (from top to bottom). The number of principal components ranges from 1 to 10 for each combination of the true rank and concentration parameter. The solid lines are for logistic PCA (LPCA) while the dotted lines are for logistic SVD (LSVD).

the estimated PC loadings with m when the data in G_j are held out. The superscript $[-j]$ is a shorthand for exclusion of the rows in G_j from the full data and likewise $[j]$ for subsetting the corresponding rows for validation. We then predict the natural parameters on the held-out group of rows, $\Theta^{[j]}$, with the given m and the fitted orthonormal matrix using (4): $\hat{\Theta}_m^{[j]} = \tilde{\Theta}^{[j]}\hat{\mathbf{U}}_m^{[-j]}\hat{\mathbf{U}}_m^{[-j]T}$ and record the predictive deviance, $D(\mathbf{X}^{[j]}; \hat{\Theta}_m^{[j]})$. After looping through all five groups, the m with the lowest predictive deviance ($\hat{m} = \arg\min_m \sum_{j=1}^{5} D(\mathbf{X}^{[j]}; \hat{\Theta}_m^{[j]})$) is used for LPCA with all the rows.

The resulting mean squared errors using the data-adaptively chosen m's are reported in Fig. 2. Using this strategy, the low-rank estimates from LPCA give more accurate estimates for the high and middle concentration scenarios, while LSVD has more accurate estimates for the low concentration scenario, unless $\hat{k} \geq k$. Standard techniques, such as those mentioned in Section 8.3, can be used to choose k for both LPCA and LSVD.

6.2. Data analysis

We present an application of logistic PCA to patient-diagnosis data, which are part of the electronic health records data on 12,000 adult patients admitted to the intensive care units at Ohio State University's Medical Center from 2007 to 2010. Patients can be admitted to an intensive care unit (ICU) for a wide variety of reasons, some of them frequently co-occurring. While in the ICU, patients are diagnosed with one or more medical conditions of over 800 disease categories from the International Classification of Diseases (ICD-9). It is often of interest to practitioners to study the comorbidity in patients, that is, what medical conditions patients are diagnosed with simultaneously. The latent factor view of principal components seems to be appropriate for describing the concept of comorbidity through a lower dimensional approximation of the disease probabilities or their transformations. Such a representation could reveal a common underlying structure capturing simultaneous existence of multiple medical conditions.

We analyzed a random sample of 1,000 patients. There were 584 ICD-9 codes that had at least one of the randomly-selected patients assigned to them. These patient-diagnosis data were organized in a binary matrix, $\mathbf{X}$, where x_{ij} is 1 if patient i has disease j and 0 otherwise. The proportions of disease occurrences ranged from 0.001 to 0.488 (the maximum corresponding to acute respiratory failure) with a median of 0.005 and the third quartile equals 0.017, meaning that most of the disease categories were rare. For comparison, we also applied logistic SVD and standard PCA to the data.

To decide on the number of principal components, we calculated the cumulative percent of deviance and the marginal percent of deviance explained by each of the three methods. The cumulative percent of deviance explained by principal components is defined analogously to the cumulative percent of variance explained in standard PCA, and the marginal percent of deviance explained by each component is defined similarly. See the definitions in Section 8.3 for more details. Fig. 3 illustrates the change in the percent of deviance explained as the number of components increases. For logistic

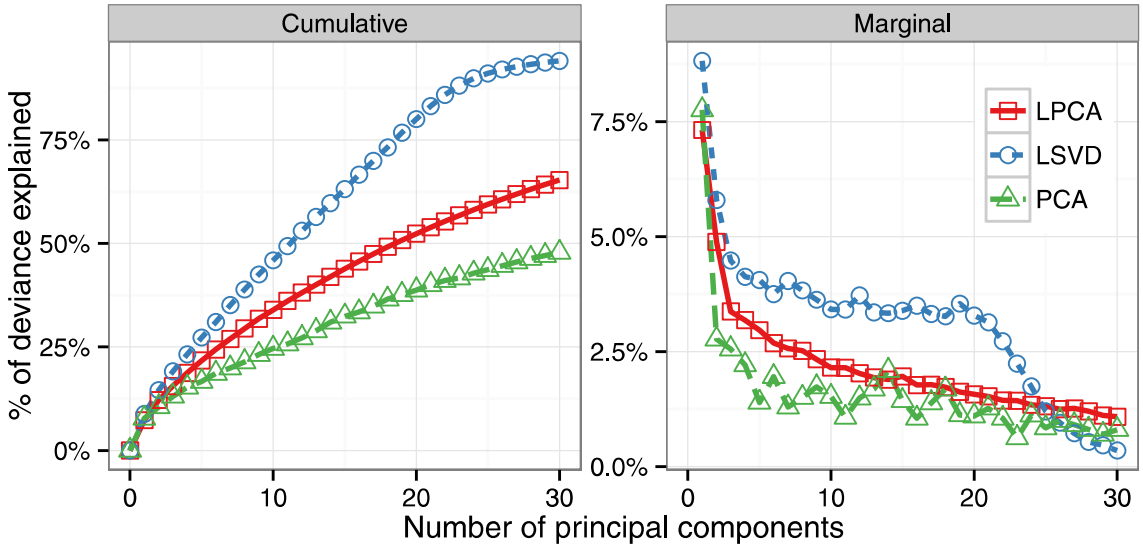

Fig. 3. Cumulative and marginal percent of deviance explained by principal components of logistic PCA (LPCA), logistic SVD (LSVD), and standard PCA for the patient-diagnosis data.

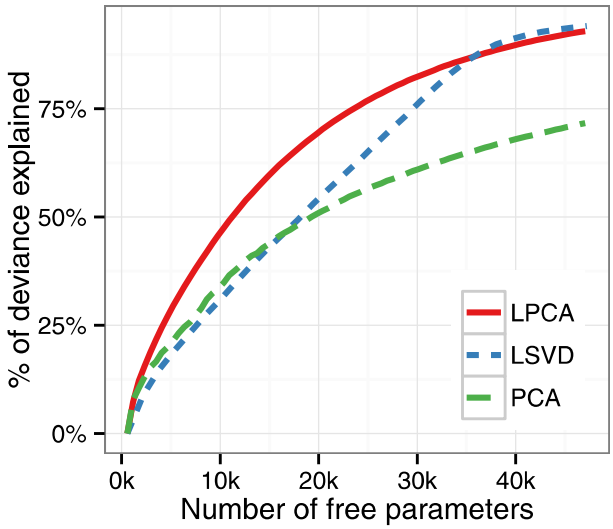

Fig. 4. Cumulative percent of deviance explained by principal components of logistic PCA (LPCA), logistic SVD (LSVD), and standard PCA plotted against the number of free parameters for the patient-diagnosis data.

PCA, m was chosen by five-fold cross validation. In order to have the same m for all k considered, we performed cross validation with $k = 15$, for which $m = 8$ had the lowest cross validation deviance. For standard PCA, we calculated the Bernoulli deviance using the reconstructed values as probability estimates. Since the reconstructed entries in the matrix could be outside the range of 0 to 1, we truncated them to be within the range $[10^{-10}, 1 - 10^{-10}]$.

If we were using LSVD, one reasonable choice for the number of components would be 24 because that is where the cumulative percent of deviance explained begins to level off and the cumulative percent of deviance explained is quite high, at 90%. The marginal percent of deviance plot suggests that the marginal contributions level off after the second component for both LPCA and LSVD. To have a manageable number of components to analyze, we may use two components.

For this dataset, LPCA fits the data significantly better than PCA with the same number of parameters per component added. Also, LSVD has higher percent of deviance explained than LPCA after the first few components. However, it is not a completely fair comparison between LSVD and LPCA, because LSVD has extra parameters in $\mathbf{A}$ to better fit the data. To illustrate this, we also plotted the percent of deviance explained as a function of the number of free parameters in Fig. 4. When indexed by the number of free parameters, LPCA looks quite favorable compared to LSVD. In fact, there is not a large difference between standard PCA and LSVD for smaller numbers of parameters.

We look at the principal component loadings of LPCA with two components, as chosen by the scree plot, to attempt to interpret the comorbidity in patients. The first component has high loadings for acute kidney failure and acute respiratory

failure, among others, which are common serious conditions that cause a patient to be admitted into the ICU. The second component has high loadings for diseases of the circulatory system such as aneurysm of coronary vessels and systolic heart failure. Subject-matter experts at Ohio State University stated that the diseases with high loadings were ones that they have observed co-occurring relatively often [7]. These findings indicate that the principal components have a meaningful interpretation related to the diseases patients have at the ICU.

7. Discussion

Previous formulations of logistic PCA have extended the singular value decomposition to binary data. Our formulation more consistently extends PCA to binary data by finding projections of the natural parameters of the saturated model. Our method produces principal components which are linear combinations of the data and can be quickly calculated on new data. We have given an MM algorithm for minimizing the deviance and shown how it performs on both simulated and real data.

Further, the formulation proposed in this paper can be extended to other members of the exponential family. Using the appropriate deviance and natural parameters from the saturated model, the formulation can naturally be applied to many types of data.

When $d \gg n$, standard PCA can be inconsistent and it has been shown that adding sparsity constraints to PCA can induce consistency [8]. Sparse loadings have the additional benefit of easier interpretation. [12] extended LSVD by adding an L_1 penalty to the loadings matrix $\mathbf{B}$. Our formulation can be extended in the same way, but further research is needed to find the best way to solve for the loadings.

8. Proofs and comments on number of principal components

This section includes derivations of (5) and (6), proofs of the theorems, and additional comments on selection of the number of principal components.

8.1. Calculation of gradient for logistic PCA

The gradient of the deviance in (3) with respect to $\mathbf{U}$ can be seen from the steps below.

$$\frac{1}{2}\frac{\partial D}{\partial \mathbf{U}} = -\frac{\partial}{\partial \mathbf{U}}\text{tr}\left(\mathbf{X}^\top\left(\mathbf{1}_n\boldsymbol{\mu}^\top + (\tilde{\boldsymbol{\Theta}} - \mathbf{1}_n\boldsymbol{\mu}^\top)\mathbf{U}\mathbf{U}^\top\right)\right) + \frac{\partial}{\partial \mathbf{U}}\sum_{i,j}\ln\left(1+\exp\left(\mu_j + [\mathbf{U}\mathbf{U}^\top(\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\mu})]_j\right)\right)$$

By standard matrix derivative rules (see, for example, [17]),

$$\frac{\partial}{\partial \mathbf{U}}\text{tr}\left(\mathbf{X}^\top(\tilde{\boldsymbol{\Theta}} - \mathbf{1}_n\boldsymbol{\mu}^\top)\mathbf{U}\mathbf{U}^\top\right) = \left(\mathbf{X}^\top(\tilde{\boldsymbol{\Theta}} - \mathbf{1}_n\boldsymbol{\mu}^\top) + (\tilde{\boldsymbol{\Theta}} - \mathbf{1}_n\boldsymbol{\mu}^\top)^\top\mathbf{X}\right)\mathbf{U}.$$

To take the derivative of the second piece, letting $\hat{\theta}_{ij} = \mu_j + [\mathbf{U}\mathbf{U}^\top(\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\mu})]_j$, note that

$$\frac{\partial}{\partial u_{k\ell}}\sum_{i,j}\ln\left(1+\exp(\hat{\theta}_{ij})\right) = \sum_{i,j}\frac{\exp(\hat{\theta}_{ij})}{1+\exp(\hat{\theta}_{ij})}\frac{\partial \hat{\theta}_{ij}}{\partial u_{k\ell}} = \sum_{i,j}\hat{p}_{ij}\frac{\partial \hat{\theta}_{ij}}{\partial u_{k\ell}}.$$

Since

$$\frac{\partial[\mathbf{U}\mathbf{U}^\top(\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\mu})]_j}{\partial u_{k\ell}} = \begin{cases}(\tilde{\theta}_{ik} - \mu_k)u_{j\ell} & \text{if } k \neq j\\ (\tilde{\theta}_{ik} - \mu_k)u_{j\ell} + (\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\mu})^\top U_\ell & \text{if } k = j,\end{cases}$$

that makes

$$\frac{\partial}{\partial u_{k\ell}}\sum_{i,j}\ln\left(1+\exp(\hat{\theta}_{ij})\right) = \sum_{i,j}\hat{p}_{ij}(\tilde{\theta}_{ik} - \mu_k)u_{j\ell} + \sum_i \hat{p}_{ik}(\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\mu})^\top U_\ell = (\tilde{\Theta}_k - \mathbf{1}_n\mu_k)^\top\hat{\mathbf{P}}U_\ell + \hat{P}_k^\top(\tilde{\boldsymbol{\Theta}} - \mathbf{1}_n\boldsymbol{\mu}^\top)U_\ell.$$

In matrix notation,

$$\frac{\partial}{\partial \mathbf{U}}\sum_{i,j}\ln\left(1+\exp\left(\mu_j + [\mathbf{U}\mathbf{U}^\top(\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\mu})]_j\right)\right) = \left(\hat{\mathbf{P}}^\top(\tilde{\boldsymbol{\Theta}} - \mathbf{1}_n\boldsymbol{\mu}^\top) + (\tilde{\boldsymbol{\Theta}} - \mathbf{1}_n\boldsymbol{\mu}^\top)^\top\hat{\mathbf{P}}\right)\mathbf{U},$$

and the result in (5) follows.

The gradient of the deviance with respect to $\boldsymbol{\mu}$ in (6) is derived as follows.

$$\frac{1}{2}\frac{\partial D}{\partial \boldsymbol{\mu}} = -\frac{\partial}{\partial \boldsymbol{\mu}}\text{tr}\left(\mathbf{X}^\top\mathbf{1}_n\boldsymbol{\mu}^\top\left(\mathbf{I}_d - \mathbf{U}\mathbf{U}^\top\right)\right) + \frac{\partial}{\partial \boldsymbol{\mu}}\sum_{i,j}\ln\left(1+\exp\left(\mu_j + \left[\mathbf{U}\mathbf{U}^\top(\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\mu})\right]_j\right)\right)$$

Using standard vector differentiation, $\frac{\partial}{\partial \boldsymbol{\mu}} \operatorname{tr}\left(\mathbf{X}^\top \mathbf{1}_n \boldsymbol{\mu}^\top \left(\mathbf{I}_d - \mathbf{U}\mathbf{U}^\top\right)\right) = (\mathbf{I}_d - \mathbf{U}\mathbf{U}^\top)\mathbf{X}^\top \mathbf{1}_n$ and

$$\frac{\partial}{\partial \boldsymbol{\mu}} \sum_{i,j} \ln\left(1 + \exp\left(\mu_j + \left[\mathbf{U}\mathbf{U}^\top(\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\mu})\right]_j\right)\right) = \sum_{i,j} \hat{p}_{ij}\left(\mathbf{e}_j - \mathbf{u}_{j\cdot}^\top \mathbf{U}^\top\right) = (\mathbf{I}_d - \mathbf{U}\mathbf{U}^\top)\hat{\mathbf{P}}^\top \mathbf{1}_n,$$

where $\mathbf{e}_j$ is a length d standard basis vector with 1 in the jth position.

8.2. Proof of theorems

Proof of Theorem 1.

(i) When $\mathbf{u} = \mathbf{e}_\ell$, the deviance as a function of $\boldsymbol{\mu}$ is simplified to

$$D(\mathbf{X}; \mathbf{1}_n\boldsymbol{\mu}^\top + (\tilde{\boldsymbol{\Theta}} - \mathbf{1}_n\boldsymbol{\mu}^\top)\mathbf{e}_\ell\mathbf{e}_\ell^\top) = 2\left[n\sum_{j=1, j\neq\ell}^{d}\left\{-\mu_j\bar{x}_j + \ln\left(1+\exp(\mu_j)\right)\right\} + \sum_{i=1}^{n}\left\{-x_{i\ell}\tilde{\theta}_{i\ell} + \ln\left(1+\exp(\tilde{\theta}_{i\ell})\right)\right\}\right], \tag{10}$$

which decouples to $(d-1)$ functions of μ_j that can be minimized separately. Thus, the solutions for the main effects $\boldsymbol{\mu}$ are known analytically. For $j \neq \ell$, $\hat{\mu}_j = \operatorname{logit} \bar{x}_j$. $\hat{\mu}_\ell$ is undefined because μ_ℓ has no effect on the deviance. We will let $\hat{\mu}_\ell = \operatorname{logit} \bar{x}_\ell$ for simplicity, although any constant would work. In this case, the estimated natural parameters are

$$\hat{\theta}_{ij} = \hat{\mu}_j + \delta_{j\ell}(\tilde{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\mu}})^\top \mathbf{e}_\ell = \begin{cases} \hat{\mu}_j & \text{if } j \neq \ell \\ \tilde{\theta}_{i\ell} & \text{if } j = \ell, \end{cases}$$

where $\delta_{j\ell}$ is the Kronecker delta.
Since $\mathbf{e}_\ell$ is a standard basis vector, $[\mathbf{C}^m\mathbf{e}_\ell]_j = c_{j\ell}^m$, where

$$c_{j\ell}^m = (X_j - \hat{P}_j)^\top(\tilde{\Theta}_\ell - \mathbf{1}_n\hat{\mu}_\ell) + (X_\ell - \hat{P}_j)^\top(\tilde{\Theta}_j - \mathbf{1}_n\hat{\mu}_j). \tag{11}$$

We will show that $c_{j\ell}^m$ is equal to 0 when $j \neq \ell$ and $\bar{x}_\ell = 1/2$. Looking at the first part of the summation in (11),

$$\begin{aligned}(X_j - \hat{P}_j)^\top(\tilde{\Theta}_\ell - \mathbf{1}_n\hat{\mu}_\ell) &= (X_j - \mathbf{1}_n\sigma(\hat{\mu}_j))^\top(mQ_\ell - \mathbf{1}_n\hat{\mu}_\ell) = (X_j - \mathbf{1}_n\bar{x}_j)^\top(m(2X_\ell - \mathbf{1}_n) - \mathbf{1}_n\hat{\mu}_\ell) \\ &= m\left[2X_j^\top X_\ell - 2n\bar{x}_j\bar{x}_\ell\right] = m\left[2n\bar{x}_j\bar{x}_\ell - 2n\bar{x}_j\bar{x}_\ell\right] = 0.\end{aligned}$$

The last line is due to the assumption that X_j and X_ℓ are uncorrelated.
From the fact that $(x_{i\ell} - \sigma(mq_{i\ell}))(mq_{ij} - \hat{\mu}_j) = (mq_{ij}q_{i\ell} - q_{i\ell}\hat{\mu}_j)/(1+\exp(m))$, the second part of $c_{j\ell}^m$ in (11) is

$$\begin{aligned}(X_\ell - \hat{P}_\ell)^\top(\tilde{\Theta}_j - \mathbf{1}_n\hat{\mu}_j) &= (X_\ell - \sigma(mQ_\ell))^\top(mQ_j - \mathbf{1}_n\hat{\mu}_j) = \sum_{i=1}^{n}(x_{i\ell} - \sigma(mq_{i\ell}))(mq_{ij} - \hat{\mu}_j) \\ &= \sum_{i=1}^{n}\frac{mq_{ij}q_{i\ell} - q_{i\ell}\hat{\mu}_j}{1+\exp(m)} = n\bar{Q}_\ell\frac{m\bar{Q}_j - \hat{\mu}_j}{1+\exp(m)}.\end{aligned}$$

If $\bar{Q}_\ell = 0$, or equivalently $\bar{x}_\ell = 1/2$, then this is exactly zero for all m.
When $j = \ell$, also using the fact that $q_{i\ell}^2 = 1$,

$$c_{\ell\ell}^m = 2(X_\ell - \hat{P}_\ell)^\top(\tilde{\Theta}_\ell - \mathbf{1}_n\hat{\mu}_\ell) = 2\sum_{i=1}^{n}\frac{mq_{i\ell}^2 - q_{i\ell}\hat{\mu}_\ell}{1+\exp(m)} = 2n\frac{m - \hat{\mu}_\ell\bar{Q}_\ell}{1+\exp(m)}.$$

When $\bar{x}_\ell = 1/2$, $\mathbf{C}^m\mathbf{e}_\ell = \lambda_m\mathbf{e}_\ell$, for all m, where $\lambda_m = 2nm/(1+\exp(m))$. With $\boldsymbol{\mu} = \hat{\boldsymbol{\mu}}$ and $\mathbf{u} = \mathbf{e}_\ell$, the first-order optimality conditions (5)–(7) are exactly satisfied.

(ii) When $\bar{x}_\ell \neq 1/2$, the first part of (11) still equals zero, but the other part does not. The squared norm of Eq. (11) equals

$$\|\mathbf{C}^m\mathbf{e}_\ell\|^2 = \left(2n\frac{m - \hat{\mu}_\ell\bar{Q}_\ell}{1+\exp(m)}\right)^2 + \sum_{j:j\neq\ell}\left(n\bar{Q}_\ell\frac{m\bar{Q}_j - \hat{\mu}_j}{1+\exp(m)}\right)^2,$$

which can be made as small as we desire by increasing m.

Proof of Theorem 2. If $\mathbf{u} = \mathbf{e}_\ell$ and $\hat{\mu}_j = \operatorname{logit}\bar{x}_j$ for $j \neq \ell$, then the deviance of the ℓth column does not depend on the column mean $\bar{x}_\ell$ as in (10) and is given by $-2\sum_{i=1}^{n}\ln\sigma(q_{i\ell}\tilde{\theta}_{i\ell}) = -2n\ln\sigma(m)$. Thus, the deviance depends on the other

$d-1$ columns of the dataset. The deviance of the jth column ($j \neq \ell$) with $\hat{\mu}_j = \text{logit}\,\bar{x}_j$ is

$$-2n\left(\bar{x}_j \ln \bar{x}_j + (1-\bar{x}_j)\ln(1-\bar{x}_j)\right),$$

which is maximized at $\bar{x}_j = 1/2$ and decreases as $\bar{x}_j$ moves away from $1/2$. Therefore, choosing ℓ with the column mean closest to $1/2$ will result in a fit with the lowest deviance.

Proof of Theorem 3. The first-order optimality condition for the loading vector with $k = 1$ and no main effects is $\mathbf{C}^m\mathbf{u} = \lambda_m\mathbf{u}$, where $\mathbf{C}^m := (\mathbf{X}-\hat{\mathbf{P}})^\top\tilde{\boldsymbol{\Theta}} + \tilde{\boldsymbol{\Theta}}^\top(\mathbf{X}-\hat{\mathbf{P}})$.

We will show that, with $\mathbf{u} = \frac{1}{\sqrt{d}}\mathbf{1}_d$ and the conditions listed in the theorem, $\mathbf{C}^m$ is compound symmetric, which in turn implies that $\mathbf{u}$ satisfies the first-order optimality conditions.

One useful implication of $\mathbf{u} \propto \mathbf{1}_d$ is that for each i, $\hat{p}_{ij} = \hat{p}_{ik}$ for all j, k because $\hat{\theta}_{ij} = u_j(\mathbf{u}^\top\tilde{\boldsymbol{\theta}}_i) = (m/d)\sum_{\ell=1}^d q_{i\ell}$. We will therefore let $\hat{\mathbf{p}}$ be the column vector with ith element equal to the common $\hat{p}_{ij}$ for all j.

First, we show that $c^m_{jj} = c^m_{kk}$, for all j, k. $c^m_{jj} = c^m_{kk}$ if and only if $X_j^\top Q_j - \hat{P}_j^\top Q_j = X_k^\top Q_k - \hat{P}_k^\top Q_k$. This condition, in turn, is equivalent to $\hat{\mathbf{p}}^\top(Q_j - Q_k) = (1/2)\mathbf{1}_n^\top(Q_j - Q_k)$ when $\hat{P}_j = \hat{P}_k = \hat{\mathbf{p}}$ with the fact that $Q_j^\top Q_j = n$ for all j. Focusing on the left hand side,

$$\hat{\mathbf{p}}^\top(Q_j - Q_k) = \sum_{i=1}^n \hat{p}_i(q_{ij} - q_{ik}) = \sum_{i:q_{ij}\neq q_{ik}} \hat{p}_i(q_{ij} - q_{ik}).$$

The second equality is due to the summation only being non-zero when $q_{ij} \neq q_{ik}$. If $q_{ij} \neq q_{ik}$, then $q_{ij} = -q_{ik}$ and thus $\sum_{\ell=1}^d q_{i\ell} = \sum_{\ell\notin\{j,k\}} q_{i\ell}$. If $q_{ij} = q_{ik}$, $\hat{p}_i(q_{ij} - q_{ik})$ will equal 0 regardless of $\hat{p}_i$. Therefore, we can state

$$\sum_{i:q_{ij}\neq q_{ik}} \hat{p}_i(q_{ij} - q_{ik}) = \sum_{i=1}^n \sigma\left(\frac{m}{d}\sum_{\ell\notin\{j,k\}} q_{i\ell}\right)(q_{ij} - q_{ik}),$$

and from the condition in (8),

$$\hat{\mathbf{p}}^\top(Q_j - Q_k) = \left(\frac{1}{2}\mathbf{1}_n + \sum_{\ell\notin\{j,k\}} Q_\ell\beta_{jk,\ell}\right)^\top (Q_j - Q_k) = \frac{1}{2}\mathbf{1}_n^\top(Q_j - Q_k).$$

The last equality is due to the compound symmetry of $\mathbf{Q}^\top\mathbf{Q}$, since all the off-diagonal elements are equal to each other. This proves that $c^m_{jj} = c^m_{kk}$.

We will now show that $c^m_{jk} = c^m_{\ell r}$, as long as $j \neq k$ and $\ell \neq r$. An off-diagonal element of $\mathbf{C}^m$ is

$$c^m_{jk} = m(X_j - \hat{\mathbf{p}})^\top Q_k + m(X_k - \hat{\mathbf{p}})^\top Q_j = m\left[Q_j^\top Q_k + \frac{1}{2}\mathbf{1}_n^\top(Q_j + Q_k) - \hat{\mathbf{p}}^\top(Q_j + Q_k)\right].$$

Showing $c^m_{jk} = c^m_{\ell r}$ is equivalent to showing

$$Q_j^\top Q_k + \frac{1}{2}\mathbf{1}_n^\top(Q_j + Q_k) - \hat{\mathbf{p}}^\top(Q_j + Q_k) = Q_\ell^\top Q_r + \frac{1}{2}\mathbf{1}_n^\top(Q_\ell + Q_r) - \hat{\mathbf{p}}^\top(Q_\ell + Q_r)$$

or

$$\frac{1}{2}\mathbf{1}_n^\top(Q_j + Q_k) - \hat{\mathbf{p}}^\top(Q_j + Q_k) = \frac{1}{2}\mathbf{1}_n^\top(Q_\ell + Q_r) - \hat{\mathbf{p}}^\top(Q_\ell + Q_r),$$

where the terms cancel because $Q_j^\top Q_k = Q_\ell^\top Q_r$. Rearranging the terms,

$$\hat{\mathbf{p}}^\top(Q_j - Q_r) - \frac{1}{2}\mathbf{1}_n^\top(Q_j - Q_r) = \hat{\mathbf{p}}^\top(Q_\ell - Q_k) - \frac{1}{2}\mathbf{1}_n^\top(Q_\ell - Q_k),$$

where we can see that both sides equal 0 because, as we have already proven, $c^m_{jj} = c^m_{rr}$ and $c^m_{\ell\ell} = c^m_{kk}$.

Therefore, $\mathbf{C}^m$ is compound symmetric for all m. This implies that $\mathbf{u} = \frac{1}{\sqrt{d}}\mathbf{1}_d$ is an eigenvector of $\mathbf{C}^m$ and (in conjunction with the fact that $\mathbf{u}^\top\mathbf{u} = 1$) $\mathbf{u}$ satisfies the first-order optimality conditions ((5), (7)).

Proof of Corollary 1. For $d = 4$, without loss of generality, let $j = 1$ and $k = 2$.

$$\sigma\left(\frac{m}{4}(q_{i3} + q_{i4})\right) = \begin{cases} \sigma(m/2) & \text{if } q_{i3} = q_{i4} = 1, \\ \frac{1}{2} & \text{if } q_{i3} \neq q_{i4}, \\ \sigma(-m/2) & \text{if } q_{i3} = q_{i4} = -1. \end{cases}$$

Since $\sigma(-m/2) = 1 - \sigma(m/2)$,

$$\sigma\left(\frac{m}{4}(q_{i3} + q_{i4})\right) = \frac{1}{2} + q_{i3}\beta_{12,3} + q_{i4}\beta_{12,4},$$

where $\beta_{12,3} = \beta_{12,4} = \{\sigma(m/2) - 1/2\}/2$. Similarly it can be shown that (8) is satisfied for $d = 3$ and trivially for $d = 2$.

8.3. Number of principal components

Selecting the appropriate dimensions for effective data representation is a common issue for dimensionality reduction techniques. There has been relatively little discussion previously in the literature of how to select the number of PCs in logistic PCA. [12] derived a BIC heuristic to select the degree of sparsity for sparse logistic PCA and [14] proposed a Bayesian version of exponential family PCA with a prior on the loadings that controls the number of principal components. We propose a few methods for selection of dimensionality in logistic PCA, motivated by the current practices in standard PCA and the dual interpretation of squared error as the deviance for a Gaussian model.

One common approach in standard PCA is to look at the cumulative percent of the variance explained and select the number of components such that a chosen proportion, γ, is met or exceeded. Let $\hat{\mathbf{U}}_k$ be the rank k estimate of the principal component loadings. The criterion will choose a rank k model if k is the smallest integer such that

$$1 - \|\mathbf{X} - \{\mathbf{1}_n\hat{\boldsymbol{\mu}}^\top + (\mathbf{X} - \mathbf{1}_n\hat{\boldsymbol{\mu}}^\top)\hat{\mathbf{U}}_k\hat{\mathbf{U}}_k^\top\}\|_F^2 / \|\mathbf{X} - \mathbf{1}_n\hat{\boldsymbol{\mu}}^\top\|_F^2 > \gamma.$$

Similarly for logistic PCA, if $D(\mathbf{X}; \hat{\tilde{\boldsymbol{\Theta}}}_k)$ is the Bernoulli deviance of the rank-k principal component loadings, $\hat{\tilde{\boldsymbol{\Theta}}}_k = \mathbf{1}_n\hat{\boldsymbol{\mu}}^\top + (\tilde{\boldsymbol{\Theta}} - \mathbf{1}_n\hat{\boldsymbol{\mu}}^\top)\hat{\mathbf{U}}_k\hat{\mathbf{U}}_k^\top$, with the data $\mathbf{X}$, then we could choose the smallest integer k such that $1 - D(\mathbf{X}; \hat{\tilde{\boldsymbol{\Theta}}}_k)/D(\mathbf{X}; \mathbf{1}_n\hat{\boldsymbol{\mu}}^\top) > \gamma$. This criterion has a similar interpretation as in standard PCA that at least $100\gamma\%$ of the deviance is explained by k principal component loadings. Notice that, as expected, 100% of the deviance will be explained by d components because $\hat{\mathbf{U}}_d\hat{\mathbf{U}}_d^\top = \mathbf{I}_d$ and $D(\mathbf{X}; \tilde{\boldsymbol{\Theta}}) = 0$ by definition.

Another approach from standard PCA is to create a scree plot of the percent of variance explained by each component and look for an elbow in the plot. The same analogy can be made to logistic PCA, however with a modified definition of the percent of reduction in deviance for additional components. For logistic PCA, the principal component loadings matrices are not necessarily nested, meaning the first $k-1$ columns of $\hat{\mathbf{U}}_k$ do not necessarily equal $\hat{\mathbf{U}}_{k-1}$. For the reason, it would be more appropriate to define the marginal percentage of deviance explained by the additional kth component as $\{D(\mathbf{X}; \hat{\tilde{\boldsymbol{\Theta}}}_{k-1}) - D(\mathbf{X}; \hat{\tilde{\boldsymbol{\Theta}}}_k)\}/D(\mathbf{X}; \mathbf{1}_n\hat{\boldsymbol{\mu}}^\top)$.

CRediT authorship contribution statement

Andrew J. Landgraf: Conceptualization, Methodology, Investigation, Software, Formal analysis, Visualization, Writing - original draft, Writing - review & editing. **Yoonkyung Lee:** Conceptualization, Methodology, Investigation, Writing - review & editing, Supervision, Funding acquisition, Project administration.

Acknowledgments

We thank Vince Vu for his feedback on the earlier version of the work. We also thank Sookyung Hyun and Cheryl Newton at College of Nursing for providing the medical diagnoses data and valuable comments on a preliminary data analysis. We also sincerely thank the Editor, Dietrich von Rosen, for helpful comments that improved the presentation of the paper. This research was supported in part by National Science Foundation, USA grant DMS-15-13566.

References

[1] M. Collins, S. Dasgupta, R.E. Schapire, A generalization of principal components analysis to the exponential family, in: T. Dietterich, S. Becker, Z. Ghahramani (Eds.), Advances in Neural Information Processing Systems 14, MIT Press, Cambridge, MA, USA, 2001, pp. 617–624.
[2] C. Eckart, G. Young, The approximation of one matrix by another of lower rank, Psychometrika 1 (3) (1936) 211–218.
[3] K. Fan, On a theorem of Weyl concerning eigenvalues of linear transformations I, Proc. Natl. Acad. Sci. USA 35 (11) (1949) 652–655.
[4] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer, 2009.
[5] H. Hotelling, Analysis of a complex of statistical variables into principal components, J. Educ. Psychol. 24 (6) (1933) 417–441.
[6] D.R. Hunter, K. Lange, A tutorial on MM algorithms, Amer. Statist. 58 (1) (2004) 30–37.
[7] S. Hyun, C. Newton, personal communication, 2013.
[8] I.M. Johnstone, A.Y. Lu, On consistency and sparsity for principal components analysis in high dimensions, J. Amer. Statist. Assoc. 104 (486) (2009) 682–693.
[9] I. Jolliffe, Principal Component Analysis, Springer, 2002.
[10] A.J. Landgraf, Y. Lee, Generalized principal component analysis: Projection of saturated model parameters, Technometrics (2019) http://dx.doi.org/10.1080/00401706.2019.1668854, Advance online publication.
[11] K. Lange, Optimization, Springer, 2013.
[12] S. Lee, J.Z. Huang, J. Hu, Sparse logistic principal components analysis for binary data, Ann. Appl. Stat. 4 (3) (2010) 1579–1601.
[13] J. de Leeuw, Principal component analysis of binary data by iterated singular value decomposition, Comput. Statist. Data Anal. 50 (1) (2006) 21–39.

[14] J. Li, D. Tao, Simple exponential family PCA, in: Y.W. Teh, M. Titterington (Eds.), Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, in: Proceedings of Machine Learning Research, vol. 9, PMLR, Sardinia, Italy, 2010, pp. 453–460.
[15] P. McCullagh, J.A. Nelder, Generalized Linear Models, second ed., Chapman & Hall, London, 1989.
[16] K. Pearson, On lines and planes of closest fit to systems of points in space, Lond. Edinb. Dublin Phil. Mag. J. Sci. 2 (11) (1901) 559–572.
[17] K.B. Petersen, M.S. Pedersen, The Matrix Cookbook, Technical University of Denmark, Kongens Lyngby, Denmark, 2012, Available online at http://www2.compute.dtu.dk/pubdb/edoc/imm3274.pdf.
[18] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2019.
[19] A.I. Schein, L.K. Saul, L.H. Ungar, A generalized linear model for principal component analysis of binary data, in: C.M. Bishop, B.J. Frey (Eds.), Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, Vol. 38, Society for Artificial Intelligence and Statistics, New Jersey, USA, 2003.
[20] M.E. Tipping, Probabilistic visualisation of high-dimensional binary data, in: M. Kearns, S. Solla, D. Cohn (Eds.), Advances in Neural Information Processing Systems 11, MIT Press, Cambridge, MA, USA, 1998, pp. 592–598.
[21] M.E. Tipping, C.M. Bishop, Probabilistic principal component analysis, J. R. Stat. Soc. Ser. B Stat. Methodol. 61 (3) (1999) 611–622.
[22] M. Udell, C. Horn, R. Zadeh, S. Boyd, Generalized low rank models, Found. Trends Mach. Learn. 9 (1) (2016) 1–118.
[23] M. Welling, C. Chemudugunta, N. Sutter, Deterministic latent variable models and their pitfalls, in: M.J. Zaki, H. Park, C. Apte, K. Wang (Eds.), Proceedings of the 2008 SIAM International Conference on Data Mining, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2008, pp. 196–207.
[24] Z. Wen, W. Yin, A feasible method for optimization with orthogonality constraints, Math. Program. 142 (1–2) (2013) 397–434.