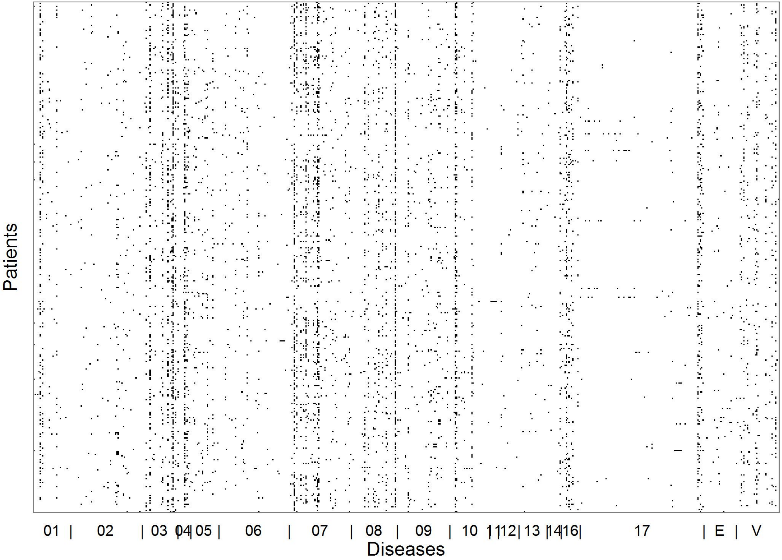
A Characterization of Latent Factors for Binary Data with Application to a Study of the Comorbidity in ICU Patients

Yoonkyung Lee*
Department of Statistics
The Ohio State University
*joint work with Andrew Landgraf

May 26-27, 2017
The Spring Conference of the Korean Statistical Society
Sookmyung Women's University
Seoul, Korea

Example: Patient-Diagnosis Matrix



Main Questions

- How to characterize common factors underlying a set of binary variables?
- Principal components are often used to define latent factors for continuous variables.

  Can we apply the same idea to binary data?
- Any implicit link between PCA and Gaussian distributions?
- How to extend PCA to non-Gaussian data?

Generalization of PCA

Collins et al. (2001), *A generalization of principal components analysis to the exponential family*

- Draws on the ideas from the exponential family and generalized linear models.

- For Gaussian data, assume that $x_i \sim N_p(\theta_i, \boldsymbol{I}_p)$ and $\theta_i \in \mathbb{R}^p$ lies in a k dimensional subspace:

$$\text{for a basis } \{b_\ell\}_{\ell=1}^k, \quad \theta_i = \sum_{\ell=1}^{k} a_{i\ell} b_\ell = B_{(p\times k)} a_i$$

- To find $\Theta = [\theta_{ij}]$, maximize the log likelihood or equivalently minimize the negative log likelihood (or deviance):

$$\sum_{i=1}^{n} \|x_i - \theta_i\|^2 = \|X - \Theta\|_F^2 = \|X - AB^\top\|_F^2$$

Generalization of PCA

- According to Eckart-Young theorem, the best rank-k approximation of $X(= U_{n\times p}D_{p\times p}V_{p\times p}^{\top})$ is given by the rank-k truncated singular value decomposition $\underbrace{U_k D_k}_{A} \underbrace{V_k^{\top}}_{B^{\top}}$.
- For exponential family data, factorize the matrix of natural parameter values Θ as $AB^{\top}$ with rank-k matrices $A_{n\times k}$ and $B_{p\times k}$ (of orthogonal columns) by maximizing the log likelihood.
- For binary data $X = [x_{ij}]$ with $P = [p_{ij}]$, "logistic PCA" looks for a factorization of $\Theta = \left[\log \frac{p_{ij}}{1-p_{ij}}\right] = AB^{\top}$ that maximizes

$$\ell(X;\Theta) = \sum_{i,j}\left\{x_{ij}(a_i^{\top}b_{j*}) - \log(1+\exp(a_i^{\top}b_{j*}))\right\}$$

subject to $B^{\top}B = \boldsymbol{I}_k$.

Drawbacks of the Matrix Factorization Formulation

- Involves estimation of both case-specific (or row-specific) scores A and variable-specific (or column-specific) factors B: more of extension of SVD than PCA.
- The number of parameters increases with the number of observations.
- The scores of generalized PC for new data involve additional optimization while PC scores for standard PCA are simple linear combinations of the data.

Alternative Interpretation of Standard PCA

- Assuming that data are centered, minimize

$$\sum_{i=1}^{n} \|x_i - VV^\top x_i\|^2 = \|X - XVV^\top\|_F^2$$

$$\text{subject to } V^\top V = \boldsymbol{I}_k.$$

- $XVV^\top$ can be viewed as a rank-k projection of the matrix of natural parameters ("means" in this case) of the saturated model $\tilde{\Theta}$ (best possible fit) for Gaussian data.

- Standard PCA finds the best rank-k projection of $\tilde{\Theta}$ by minimizing the deviance under Gaussian distribution.

Natural Parameters of the Saturated Model

- For an exponential family distribution with natural parameter θ and pdf

$$f(x|\theta) = \exp\left(\theta x - b(\theta) + c(x)\right),$$

$E(X) = b'(\theta)$ and the canonical link function is the inverse of b'.

	θ	$b(\theta)$	canonical link
$N(\mu, 1)$	μ	$\theta^2/2$	identity
Bernoulli(p)	logit(p)	$\log(1 + \exp(\theta))$	logit
Poisson(λ)	$\log(\lambda)$	$\exp(\theta)$	log

- Take $\tilde{\Theta} = [\text{canonical link}(x_{ij})]$.

New Formulation of Logistic PCA

Landgraf and Lee (2015), *Dimensionality Reduction for Binary Data through the Projection of Natural Parameters*

- Given $x_{ij} \sim \mathit{Bernoulli}(p_{ij})$, the natural parameter (logit p_{ij}) of the saturated model is

$$\tilde{\theta}_{ij} = \text{logit}(x_{ij}) = \infty \times (2x_{ij} - 1)$$

  We will approximate $\tilde{\theta}_{ij} \approx m \times (2x_{ij} - 1)$ for large $m > 0$.

- Project $\tilde{\Theta}$ to a k-dimensional subspace by using the deviance $D(X; \Theta) = -2\{\ell(X; \Theta) - \ell(X; \tilde{\Theta})\}$ as a loss:

$$\min_{V \in \mathbb{R}^{p \times k}} D(X; \underbrace{\tilde{\Theta} V V^\top}_{\hat{\Theta}}) = -2 \sum_{i,j} \left\{ x_{ij} \hat{\theta}_{ij} - \log(1 + \exp(\hat{\theta}_{ij})) \right\}$$

$$\text{subject to } V^\top V = \boldsymbol{I}_k$$

Logistic PCA vs Logistic SVD

- The previous logistic SVD (matrix factorization) gives an approximation of logit P:

$$\hat{\Theta}_{LSVD} = AB^\top$$

- Alternatively, our logistic PCA gives

$$\hat{\Theta}_{LPCA} = \underbrace{\tilde{\Theta} V}_{A} V^\top,$$

which has much fewer parameters.

- Computation of PC scores on new data only requires matrix multiplication for logistic PCA while logistic SVD requires fitting k-dimensional logistic regression for each new observation.

- Logistic SVD with additional A is prone to overfit.

Geometry of Logistic PCA

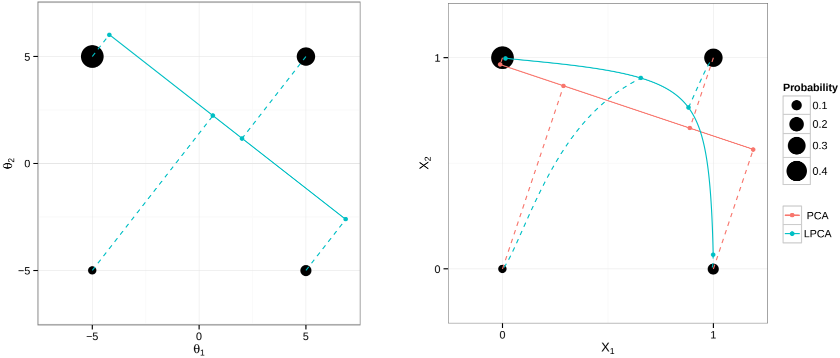


Figure: Logistic PCA projection in the natural parameter space with $m = 5$ (left) and in the probability space (right) compared to the PCA projection

New Formulation of Generalized PCA

Landgraf and Lee (2015), *Generalized PCA: Projection of Saturated Model Parameters*

- The idea can be applied to any exponential family distribution (e.g. Poisson, multinomial).
- Find the best rank-k projection of the matrix of natural parameters from the saturated model $\tilde{\Theta}$ by minimizing the appropriate deviance for the data:

$$\min_{V \in \mathbb{R}^{p \times k}} D(X; \tilde{\Theta} V V^\top)$$

$$\text{subject to } V^\top V = \boldsymbol{I}_k$$

- If desired, main effects μ can be added to the approximation of Θ:

$$\hat{\Theta} = \mathbf{1}\mu^\top + (\tilde{\Theta} - \mathbf{1}\mu^\top) V V^\top$$

MM Algorithm for Generalized PCA

- Majorize the objective function with a simpler objective at each iterate, and minimize the majorizing function. (Hunter and Lange, 2004)

- From the quadratic approximation of the deviance at $\Theta^{(t)}$, step t solution, and the fact that $p(1-p) \leq 1/4$,

$$
\begin{aligned}
& D(X; \mathbf{1}\mu^\top + (\tilde{\Theta} - \mathbf{1}\mu^\top)VV^\top) \\
\leq \quad & \frac{1}{4}\|\mathbf{1}\mu^\top + (\tilde{\Theta} - \mathbf{1}\mu^\top)VV^\top - Z^{(t+1)}\|_F^2 + C, \\
& \text{where } Z^{(t+1)} = \Theta^{(t)} + 4(X - \hat{P}^{(t)}).
\end{aligned}
$$

- Update Θ at step $(t+1)$:
  averaging for $\mu^{(t+1)}$ given $V^{(t)}$ and
  eigen-analysis of a $p \times p$ matrix for $V^{(t+1)}$ given $\mu^{(t+1)}$.

Medical Diagnosis Data

- Part of electronic health record data on 12,000 adult patients admitted to the intensive care units (ICU) in Ohio State University Medical Center from 2007 to 2010 (provided by S. Hyun)
- Patients are classified as having one or more diseases of over 800 disease categories from the International Classification of Diseases (ICD-9).
- Interested in characterizing the co-morbidity as latent factors, which can be used to define patient profiles for prediction of other clinical outcomes
- Analysis is based on a sample of 1,000 patients, which reduced the number of disease categories to 584.

Deviance Explained by Components

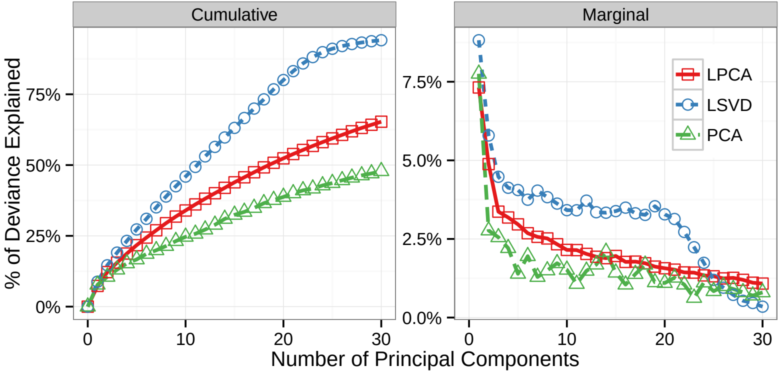


Figure: Cumulative and marginal percent of deviance explained by principal components of LPCA, LSVD, and PCA

Deviance Explained by Parameters

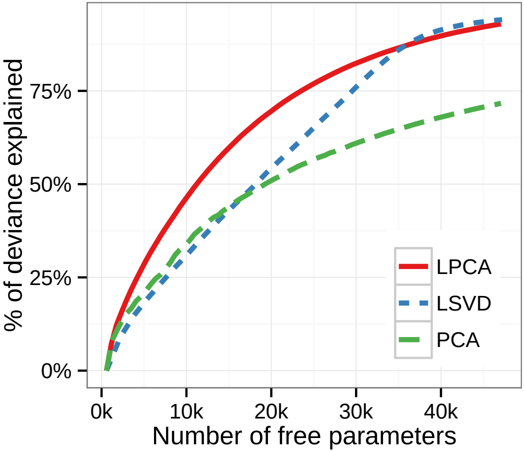


Figure: Cumulative percent of deviance explained by principal components of LPCA, LSVD, and PCA versus the number of free parameters

Predictive Deviance

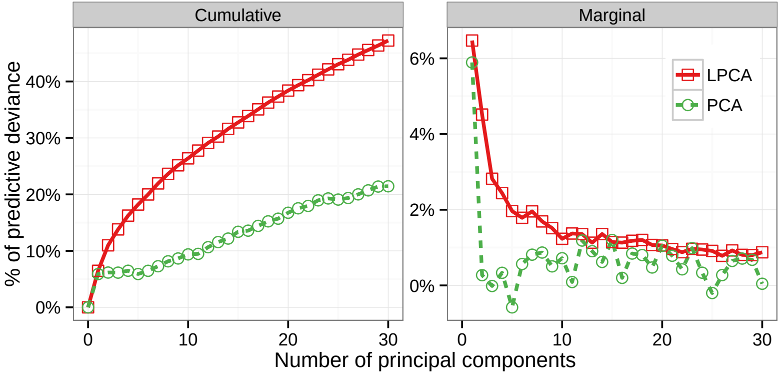


Figure: Cumulative and marginal percent of predictive deviance over test data (1,000 patients) by the principal components of LPCA and PCA

Interpretation of Loadings

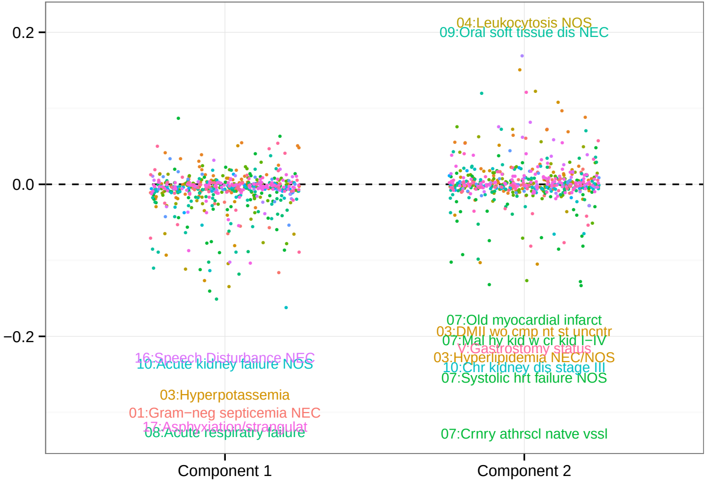


Figure: The first component is characterized by common serious conditions that bring patients to ICU, and the second component is dominated by diseases of the circulatory system (07's).

Concluding Remarks

- We have generalized PCA via projections of the natural parameters of the saturated model using the generalized linear model framework.
- We have extended generalized PCA to handle differential case weights, missing data, and variable normalization.
- Further extensions are possible with other constraints than rank for desirable properties (e.g. sparsity) on the loadings and predictive formulations.
- R package, logisticPCA is available at CRAN and generalizedPCA is available at GitHub.

Acknowledgments

Andrew Landgraf
@ Battelle Memorial Institute

Sookyung Hyun and Cheryl Newton
@ College of Nursing, OSU

DMS-15-13566

References

A. J. Landgraf and Y. Lee.
Dimensionality reduction for binary data through the projection of natural parameters.
Technical Report 890, Department of Statistics, The Ohio State University, 2015.
Also available at arXiv:1510.06112.

A. J. Landgraf and Y. Lee.
Generalized principal component analysis: Projection of saturated model parameters.
Technical Report 892, Department of Statistics, The Ohio State University, 2015.