Part III: Dimensionality reduction

Hotelling's principal component analysis (PCA)
to generalized PCA for non-Gaussian data

Hotelling, H. (1933), Analysis of a complex of statistical variables into principal components
Journal of Educational Psychology 24(6), 417-441

Pearson, K. (1901), On Lines and Planes of Closest Fit to Systems of Points in Space
Philosophical Magazine 2(11), 559-572.

Hotelling's Test Data

- Hotelling analyzed the correlations found in a sample of 140 seventh-grade children among numerous tests (from T. L. Kelley's study).

  Displayed below are the correlations among (1) reading speed, (2) reading power, (3) arithmetic speed and (4) arithmetic power:

$$\boldsymbol{R} = \begin{bmatrix} 1 & .698 & .264 & .081 \\ & 1 & -.061 & .092 \\ & & 1 & .594 \\ & & & 1 \end{bmatrix}$$

- Given correlated variables X_j, does there exist some more fundamental set of independent variables, perhaps fewer in number than the original X_j's, which determine the values of X_j's?

Principal Component Analysis (PCA)

PCA is concerned with explaining the variance-covariance structure of a set of correlated variables through a few *linear* combinations of these variables.

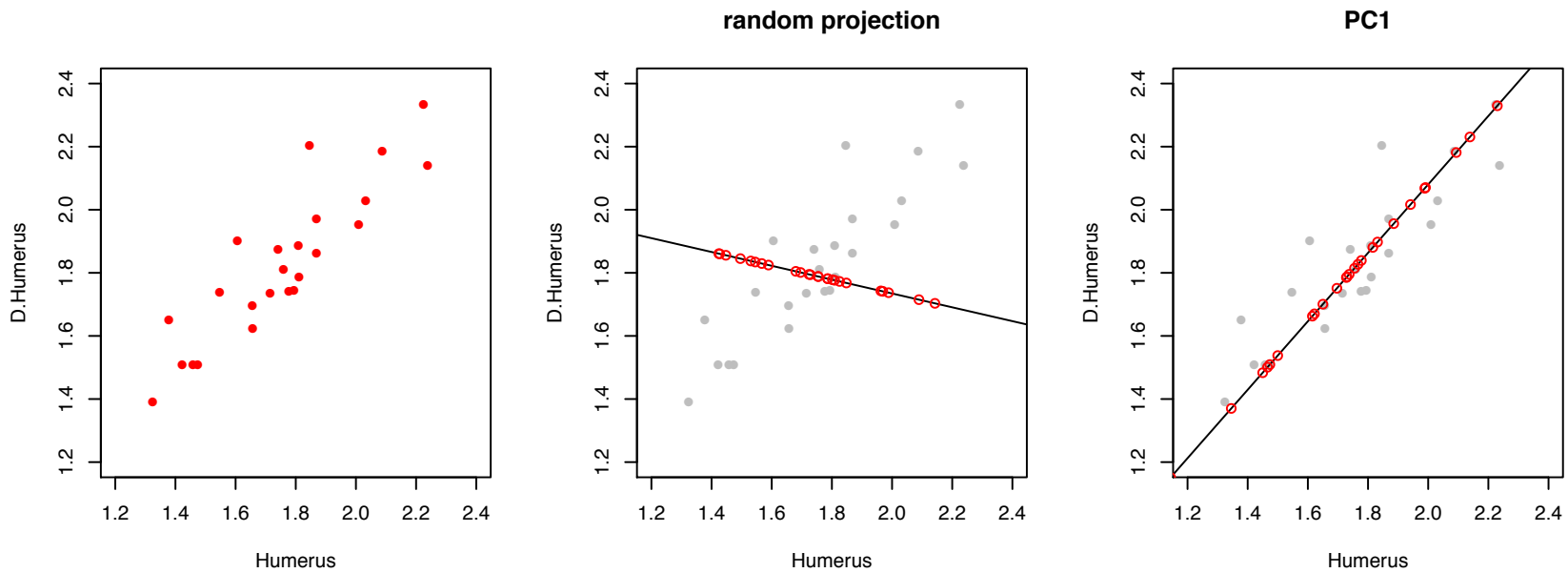


Figure: Data on the mineral content measurements (g/cm) of three bones (humerus, radius and ulna) on the dominant and nondominant sides for 25 old women

Variance Maximization

- Given p correlated variables $X = (X_1, \cdots, X_p)^\top$, consider a linear combination of X_j's,

$$\sum_{j=1}^{p} a_j X_j = a^\top X$$

for $a = (a_1, \ldots, a_p)^\top \in \mathbb{R}^p$ with $\|a\|^2 = 1$.

- The first principal component direction is defined as the vector a that gives the largest sample variance of $a^\top X$ amongst all normalized linear combinations of X_j:

$$\max_{a \in \mathbb{R}^p, \|a\|^2 = 1} a^\top \boldsymbol{S}_n a$$

where $\boldsymbol{S}_n$ is the sample variance-covariance matrix of X.

Principal Components

- Let $\boldsymbol{S}_n = \sum_{j=1}^{p} \lambda_j v_j v_j^\top$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$, and the corresponding eigenvectors $v_1, \ldots, v_p$.
  Then the first principal component direction is given by v_1.
- The derived variable $Z_1 = v_1^\top X$ is called the first principal component.

- Similarly, the second principal component direction is defined as the vector a that gives the largest sample variance of $a^\top X$ among all normalized a subject to $a^\top X$ being uncorrelated with $v_1^\top X$. It is given by v_2.

- In general, the jth principal component direction is defined successively from $j = 1$ to p.

Pearson's Reconstruction Error Formulation

Pearson, K. (1901), *On Lines and Planes of Closest Fit to Systems of Points in Space*

- Given $x_1, \cdots, x_n \in \mathbb{R}^p$, consider the following data approximation:

$$x_i \approx \mu + vv^\top (x_i - \mu)$$

where $\mu \in \mathbb{R}^p$ and v is a unit vector in $\mathbb{R}^p$ so that $vv^\top$ is a rank-one projection.

- What are μ and $v \in \mathbb{R}^p$ that minimize the reconstruction error?

$$\min_{v^\top v = 1} \sum_{i=1}^{n} \|x_i - \mu - vv^\top (x_i - \mu)\|^2$$

- $\hat{\mu} = \bar{x}$ and $\hat{v} = v_1$ minimize the error.

Minimization of Reconstruction Error

- More generally, consider a rank-k ($< p$) approximation:

$$x_i \approx \mu + VV^\top (x_i - \mu)$$

where $\mu \in \mathbb{R}^p$ and V is a $p \times k$ matrix with orthogonal columns that results in a rank-k projection of $VV^\top$.

- Wish to minimize the reconstruction error:

$$\sum_{i=1}^{n} \|x_i - \mu - VV^\top (x_i - \mu)\|^2$$

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

- $\hat{\mu} = \bar{x}$ and $\hat{V} = [v_1, \cdots, v_k]$ provide the best reconstruction of the data.

PCA for Non-Gaussian Data?

- PCA finds a low rank subspace by implicitly minimizing the reconstruction error under squared error loss, which is linked to Gaussian distribution.
- Binary, count, or non-negative data abound in practice.

  e.g. images, term frequencies for documents, ratings for movies, click-through rates for on-line ads
- How to generalize PCA to non-Gaussian data?

Generalized 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):

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

Generalized 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

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

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

Drawbacks of the Matrix Factorization Formulation

- Involves estimation of both case-specific (or row-specific) factors A and variable-specific (or column-specific) factors B: more of extension of SVD than PCA.
- The number of parameters increases with 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 ($\mu = 0$),

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

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

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

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

New Formulation of Logistic PCA

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

- Given $x_{ij} \sim Bernoulli(p_{ij})$, 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 $D(X; \Theta) = -2\ell(X; \Theta)$ as a loss:

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

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

Logistic PCA vs Logistic SVD

- The previous logistic SVD gives an approximation of logit P:

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

- Alternatively, logistic PCA gives

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

  which has much fewer parameters.

- Computation of PC scores on new data only requires linear combinations of $\tilde{\theta}(x)$ 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.

New Formulation of Generalized PCA

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

$$\min_V D(X; \tilde{\Theta}_X V V^\top)$$

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

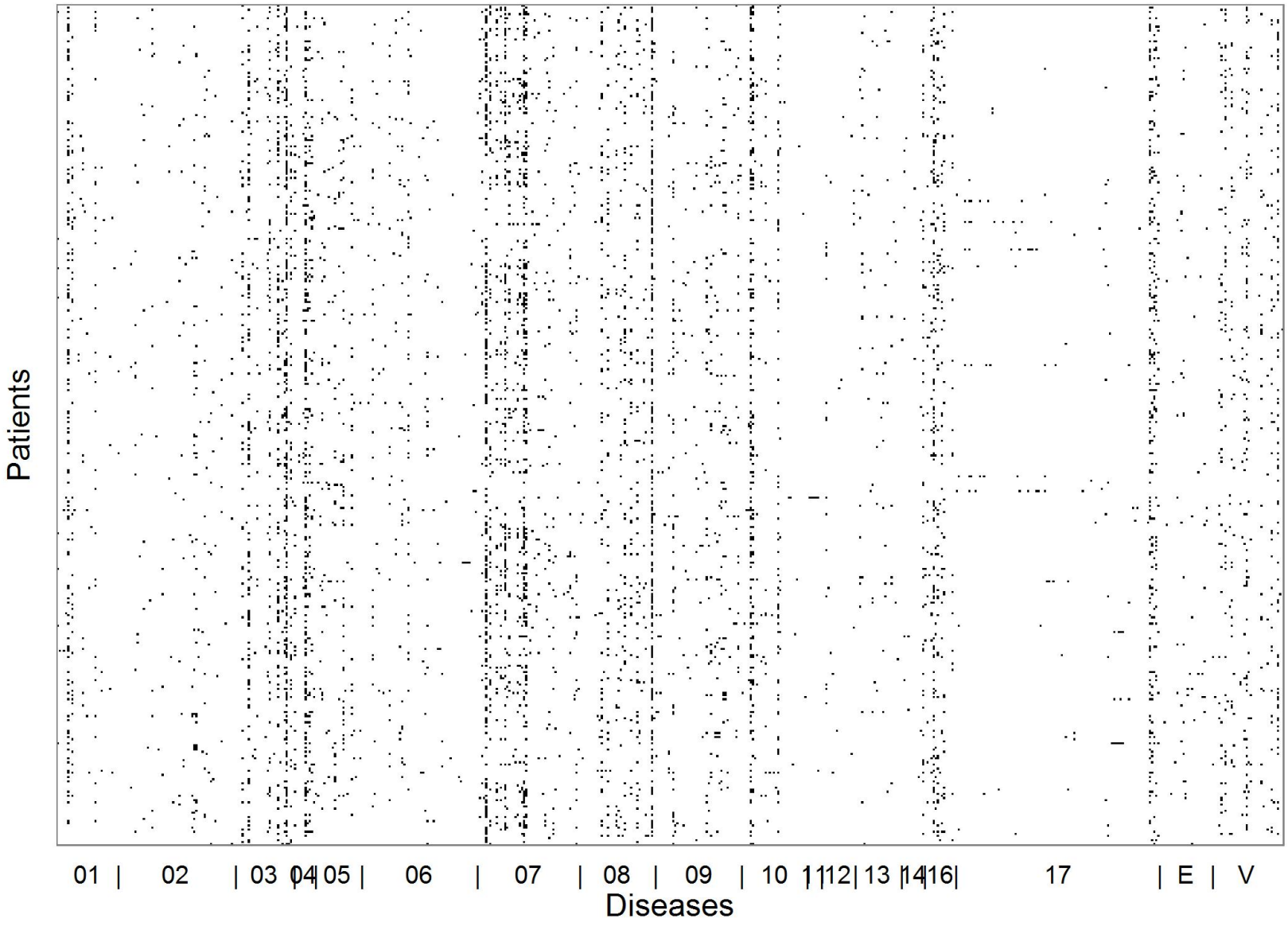
- If desired, main effects μ can be added to the approximation of Θ:

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

Medical Diagnosis Data

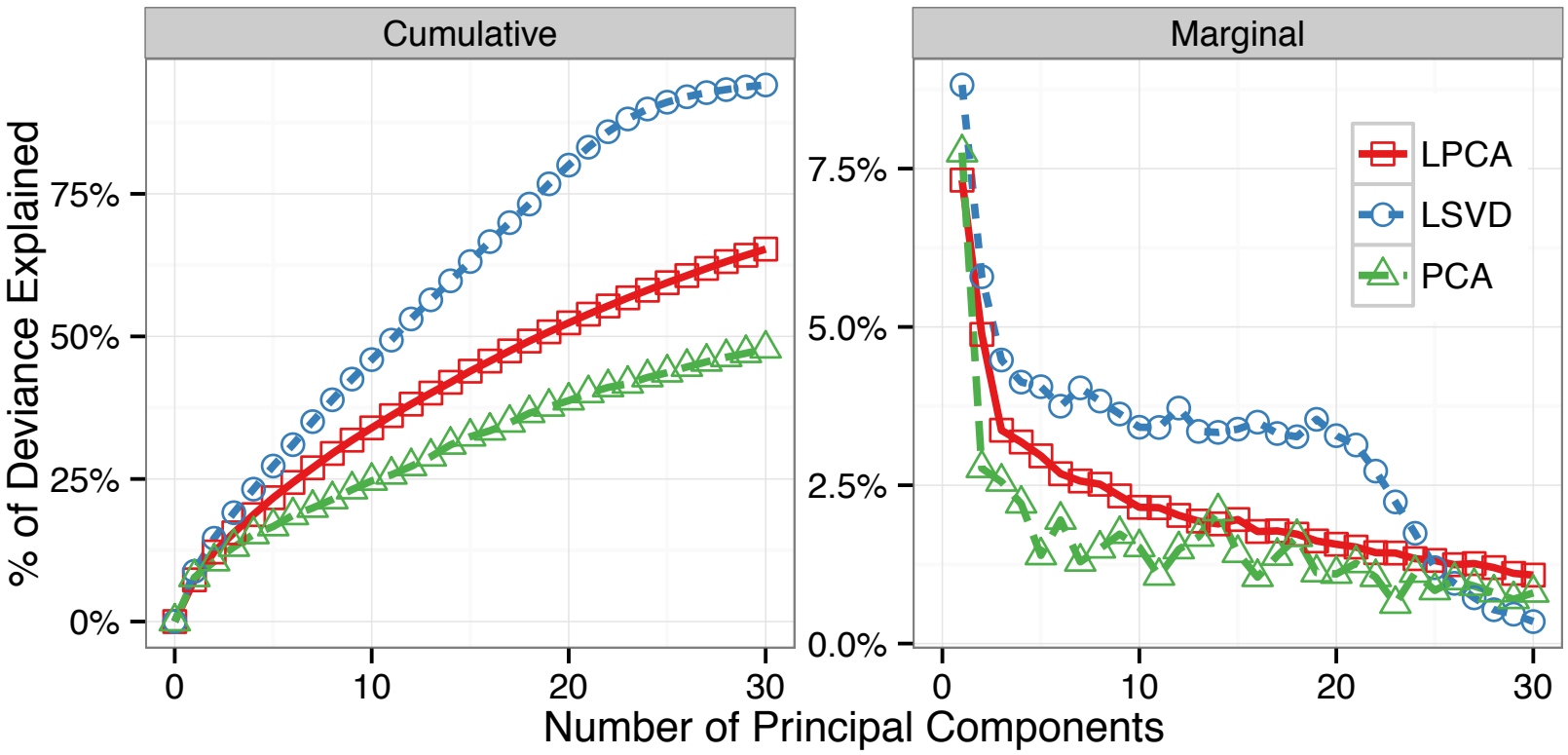
- Part of electronic health record data on 12,000 adult patients admitted to the intensive care units (ICU) in Ohio State University Medical Center from 2007 to 2010 (provided by S. Hyun)
- Patients are classified as having one or more diseases of over 800 disease categories from the International Classification 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.

Patient-Diagnosis Matrix



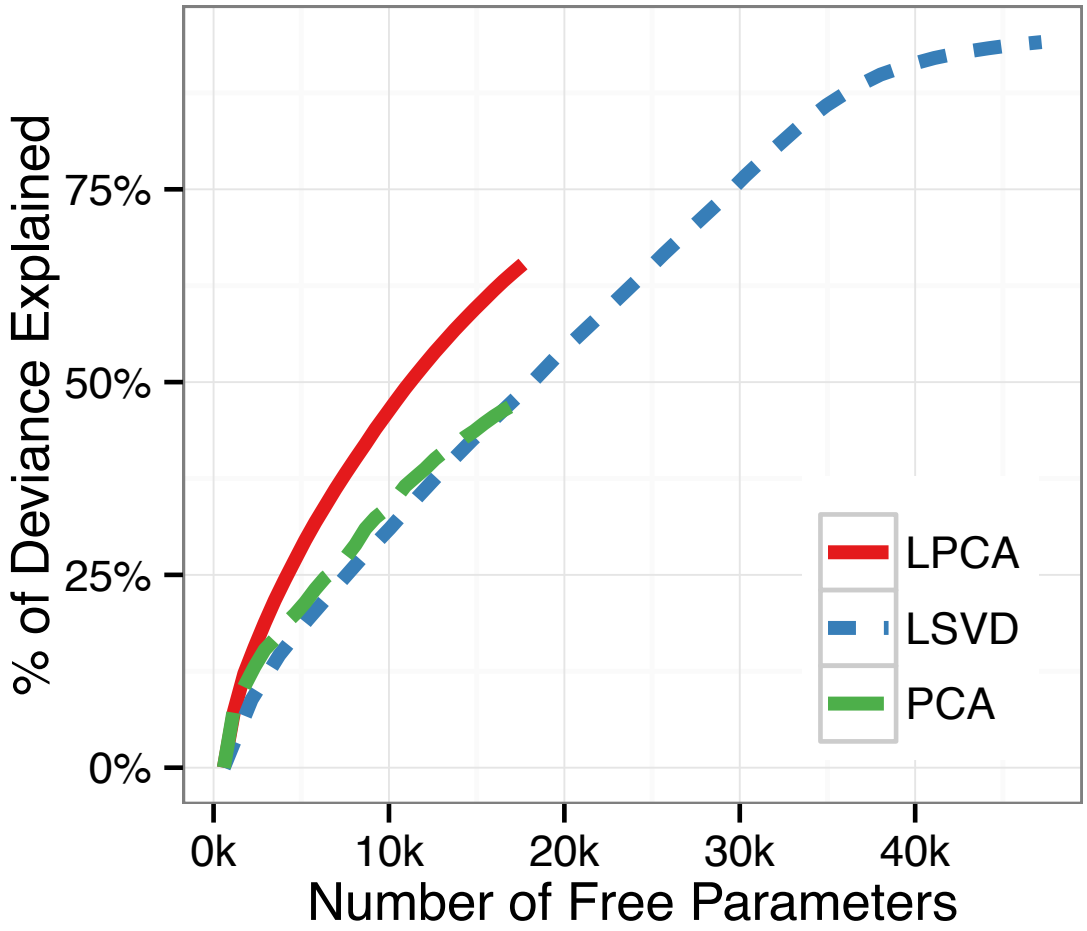
courtesy of A. Landgraf

Deviance Explained by Components



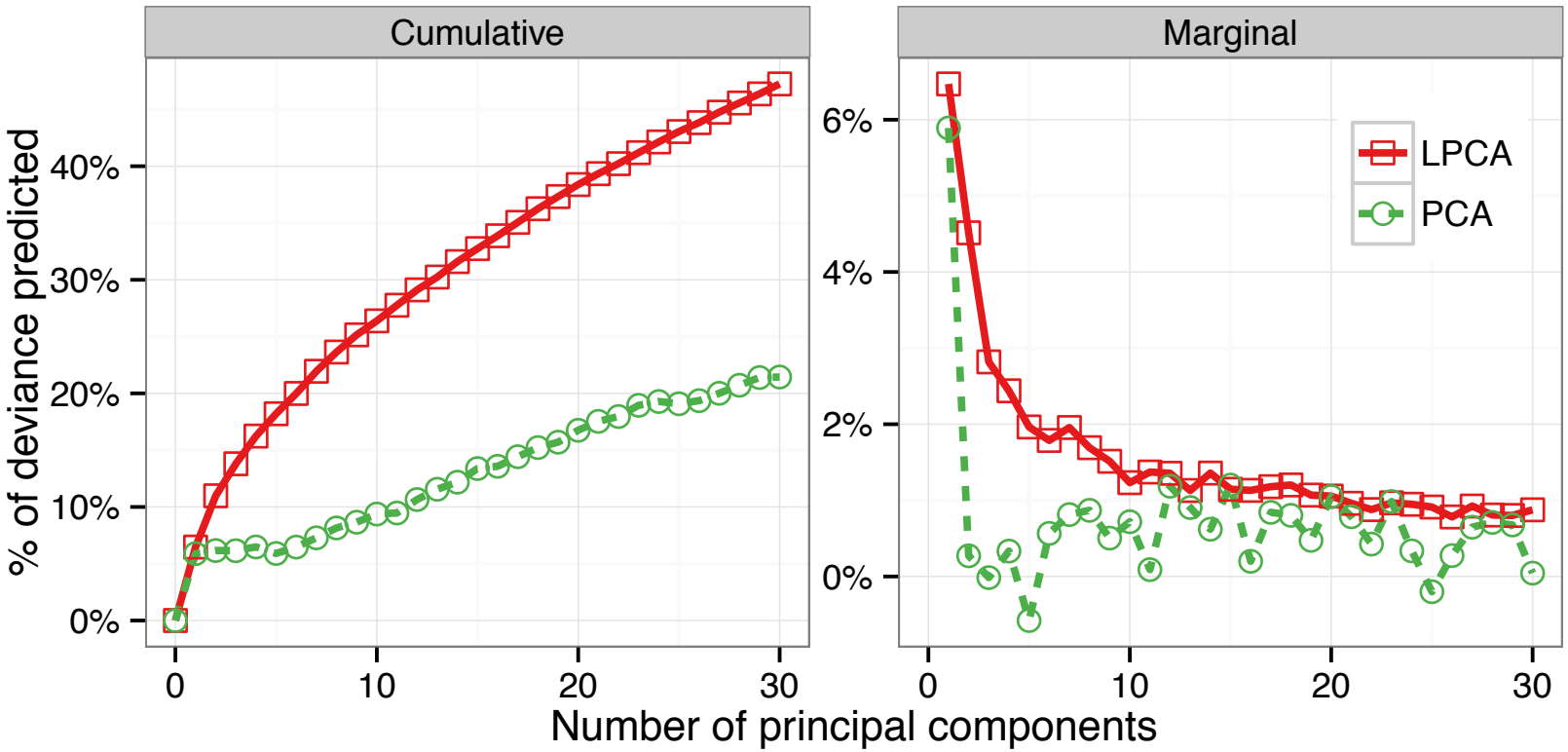
courtesy of A. Landgraf

Deviance Explained by Parameters



courtesy of A. Landgraf

Deviance Predicted



courtesy of A. Landgraf

Interpretation of Loadings

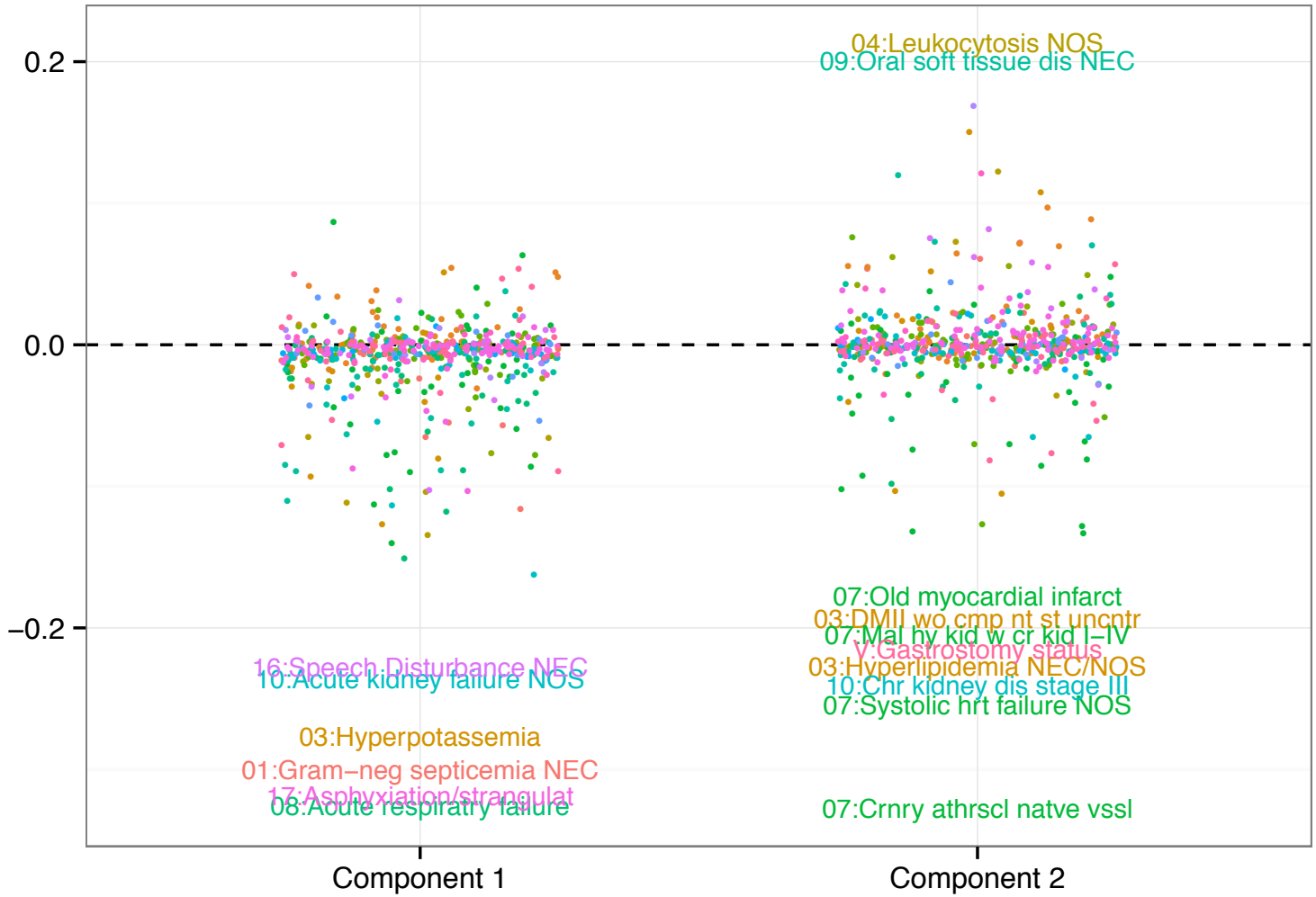


Figure: The first component is characterized by common serious conditions that bring patients to ICU, and the second component is dominated by diseases of the circulatory system (07's). courtesy of A. Landgraf

Acknowledgements

- Grace Wahba and Yi Lin
  (smoothing splines, SVM and regularization)
- Tao Shi (multivariate analysis)
- Rui Wang (comparison of classifiers)
- Andrew Landgraf (logistic PCA)
- Sookyung Hyun and Cheryl Newton (diagnosis data)
- Rogelio Ramos and Johan Van Horebeek @ CIMAT

This short course is largely based on the following references:

T. Hastie, R. Tibshirani, and J. Friedman.

The Elements of Statistical Learning.

Springer Verlag, New York, 2001.

R. A. Johnson and D. W. Wichern.

Applied Multivariate Statistical Analysis.

Pearson Prentice Hall, 6th edition, 2007.

Andrew J. Landgraf and Yoonkyung Lee.

Dimensionality reduction for binary data through the projection of natural parameters.

Technical Report 890, Department of Statistics, The Ohio State University, 2015.