A Generalized Linear Model for Principal Component Analysis of Binary Data

January 6, 2003

Andrew I. Schein, Lyle H. Ungar and Lawrence K. Saul.
Department of Computer and Information Science
The University Of Pennsylvania. Philadelphia, PA.

Ninth International Workshop on AI and Statistics

PCA and LPCA

Principal Component Analysis is commonly called PCA.

PCA is a widely-used dimensionality reduction technique.

PCA handles real-valued data through a Gaussian assumption.

Today we will explore a different assumption for binary data.

Logistic PCA is to (Linear) PCA
as
Logistic Regression is to Linear Regression

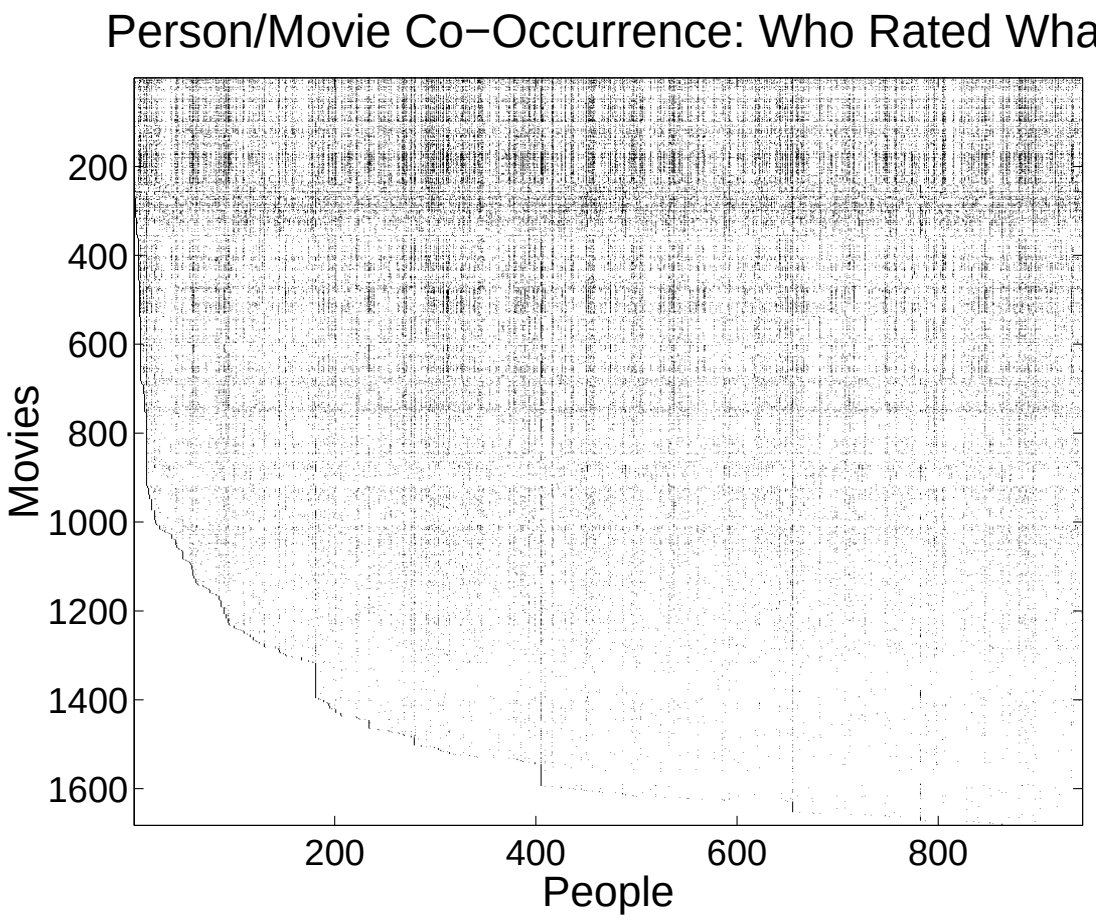
Talk Outline

1. Linear PCA of Real-Valued Data (review)
2. Logistic PCA of Binary Data

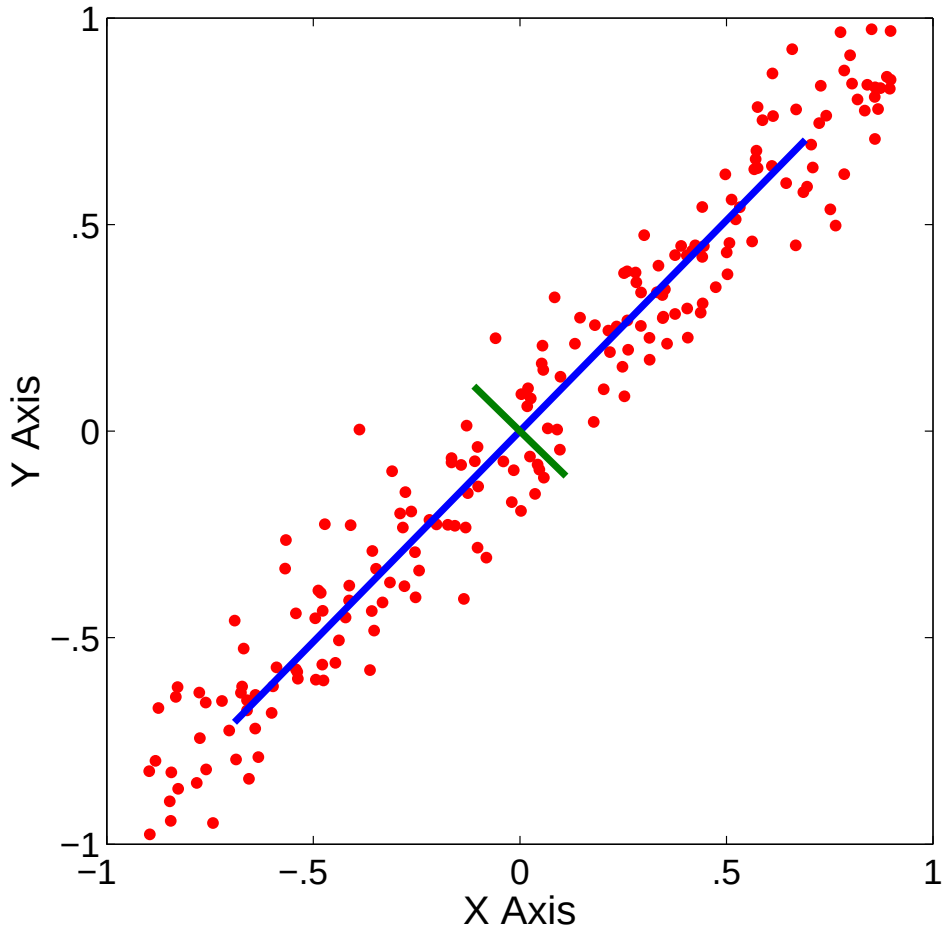
   Our Contributions Consist of:

3. Model Fitting by Alternating Least Squares
4. Experimental Results on 4 Natural Data Sets

Multivariate Binary Data



Visualizing PCA



Applications of PCA

- Noise Removal
- Dimensionality Reduction
- Data Compression
- Visualization
- Exploratory Data Analysis
- Feature Extraction

PCA as Least-Squares Decomposition

$$\text{Error}(V^L, U) \quad = \quad \sum_{n=1}^{N} ||X_n - U_n V^L||^2$$

$X =$ The Data: $N \times D$

$U =$ Latent Coordinates: $N \times L$

$V =$ Orthogonal Latent Axes: $L \times D$

$L << D$, $\quad L$ is the dimensionality of the latent space.

Gaussian Interpretation of PCA

When σ is known there is an equivalent model:

$$X_{nd} \sim \mathcal{N}((UV^L)_{nd}, \sigma^2)$$

Maximum Likelihood Objective = Least Squares Loss

So PCA assumes a Gaussian distribution on X.

Generalized Principal Component Analysis (GPCA)

Collins et al. (2001) propose a generalized scheme for PCA.

Define a constrained decomposition of the *natural parameter*

$$\Theta_{nd} = (UV)_{nd}, \quad \dim(U) = N \times L, \;\; \dim(V) = L \times D.$$

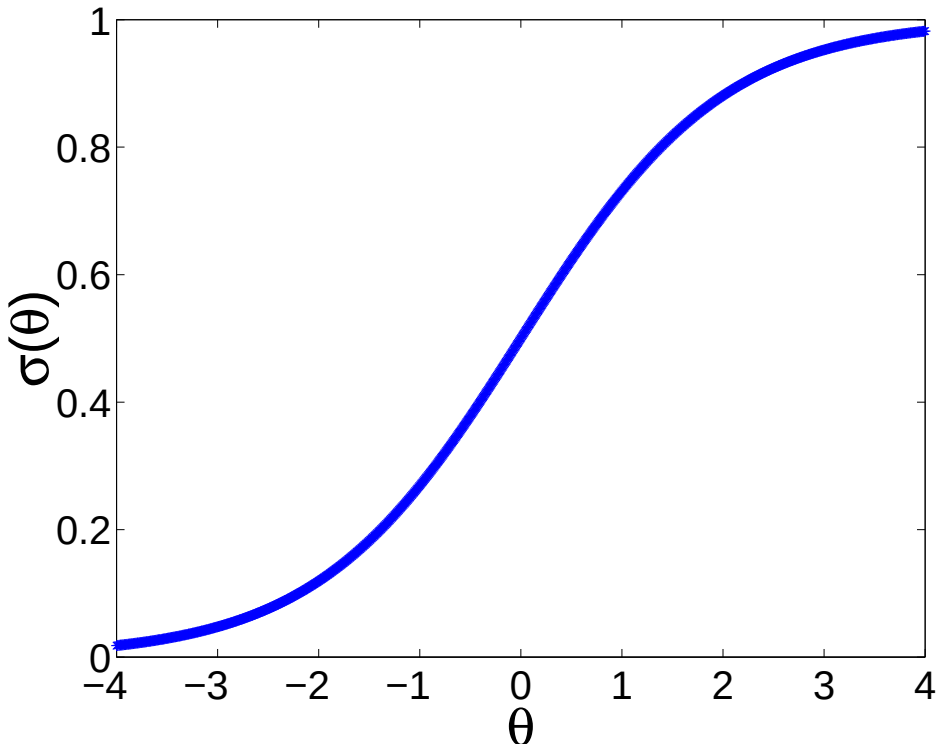
$N =$ Number of Observations
$D =$ Dimensionality of Data
$L =$ Dimensionality of Latent Space

$$\mathcal{L}(V, U) = -\sum_n \sum_d \log \mathsf{P}(X_{nd} | \Theta_{nd})$$

Insert your favorite *exponential family* distribution to instantiate P.

The Logistic Function

$$\sigma(\theta) = \frac{1}{1 + \exp(-\theta)}$$



Logistic PCA Model

Inserting Bernoulli Distribution We Get Log-Likelihood:

$$\mathcal{L} = \sum_{n,d} \left[X_{nd} \log \sigma(\Theta_{nd}) + (1 - X_{nd}) \log \sigma(-\Theta_{nd}) \right]$$

subject to constraint

$$\Theta_{nd} = \sum_{l} U_{nl} V_{ld}$$

$N =$ Number of Observations
$D =$ Dimensionality of Data
$L =$ Dimensionality of Latent Space

How to Fit LPCA?

Collins et al. (2001) propose a general strategy for fitting GPCA.

Applying this framework to the LPCA case looks hard if not intractable.

We take the approach of fitting LPCA through specialized strategies.

Our methods exploit bounds on the logistic function.

Defining the Auxiliary Function for LPCA

A useful fact:

$$\log \sigma(\theta) = -\log 2 + \theta/2 - \log \cosh(\theta/2)$$

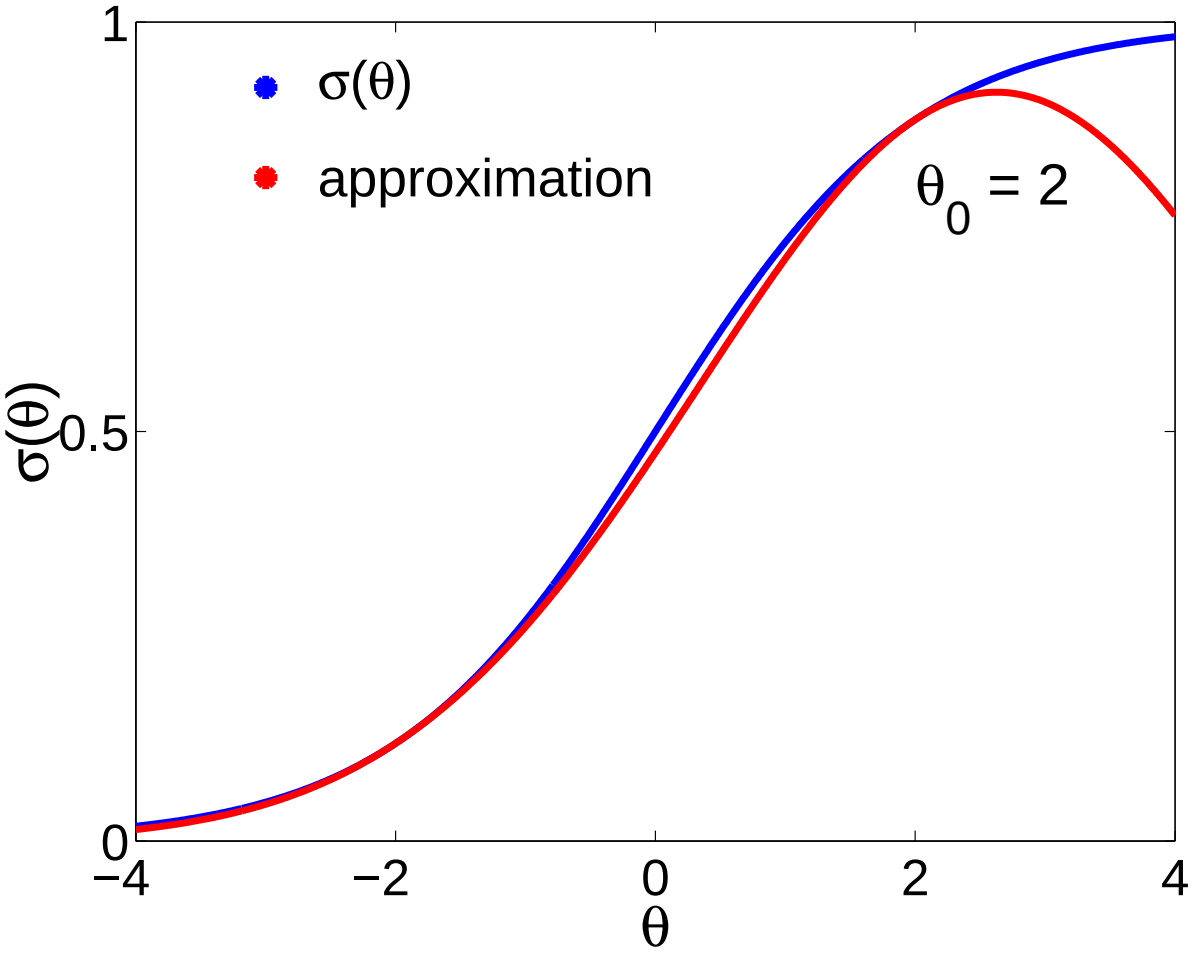
We exploit a bound:

$$\log \cosh(\theta/2) \leq \log \cosh(\theta_0/2) + (\theta^2 - \theta_0^2) \left[\frac{\tanh(\theta_0/2)}{4\theta_0} \right]$$

[Jaakkola and Jordan, 1997. Tipping, 1999.]

The bound is concave and quadratic in the parameter θ.

Visualizing the Approximation of σ



Model Fitting by Alternating Least Squares

We develop a model fitting strategy that alternates between two steps:

- Fix V, find the least squares solution for U rows.
- Fix U, find the least squares solution for V columns.

Each iteration guarantees an improvement in log-likelihood.

The likelihood structure: global vs. local maxima is unknown.

A Related Use of the Bound

Normal Linear Factor Analysis (NLFA) is a generative cousin of PCA.

[Tipping, 1999], uses the bound to fit a logit/normit factor analysis.

Logit/normit factor analysis is a type of factor analysis for binary data.

LPCA is binary PCA
while
logit/normit factor analysis is binary factor analysis

We follow Tipping's factor analysis strategy in fitting LPCA.

Logit/Normit Factor Analysis

Maximize:

$$\mathcal{L} = \sum_{n,d} X_{nd} \log \sigma(\Theta_{nd}) + (1 - X_{nd}) \log \sigma(-\Theta_{nd})$$

subject to constraint

$$\Theta_{nd} = \sum_{l} U_{nl} V_{ld}, \quad \text{where } l \text{ is a latent space dimension}$$

and $U_n \sim \mathcal{N}(0, I)$

Logit/Normit Factor Analysis

$$U_n \sim \mathcal{N}(0, I)$$

Fitting the U_n in logit/normit factor is harder than in LPCA.

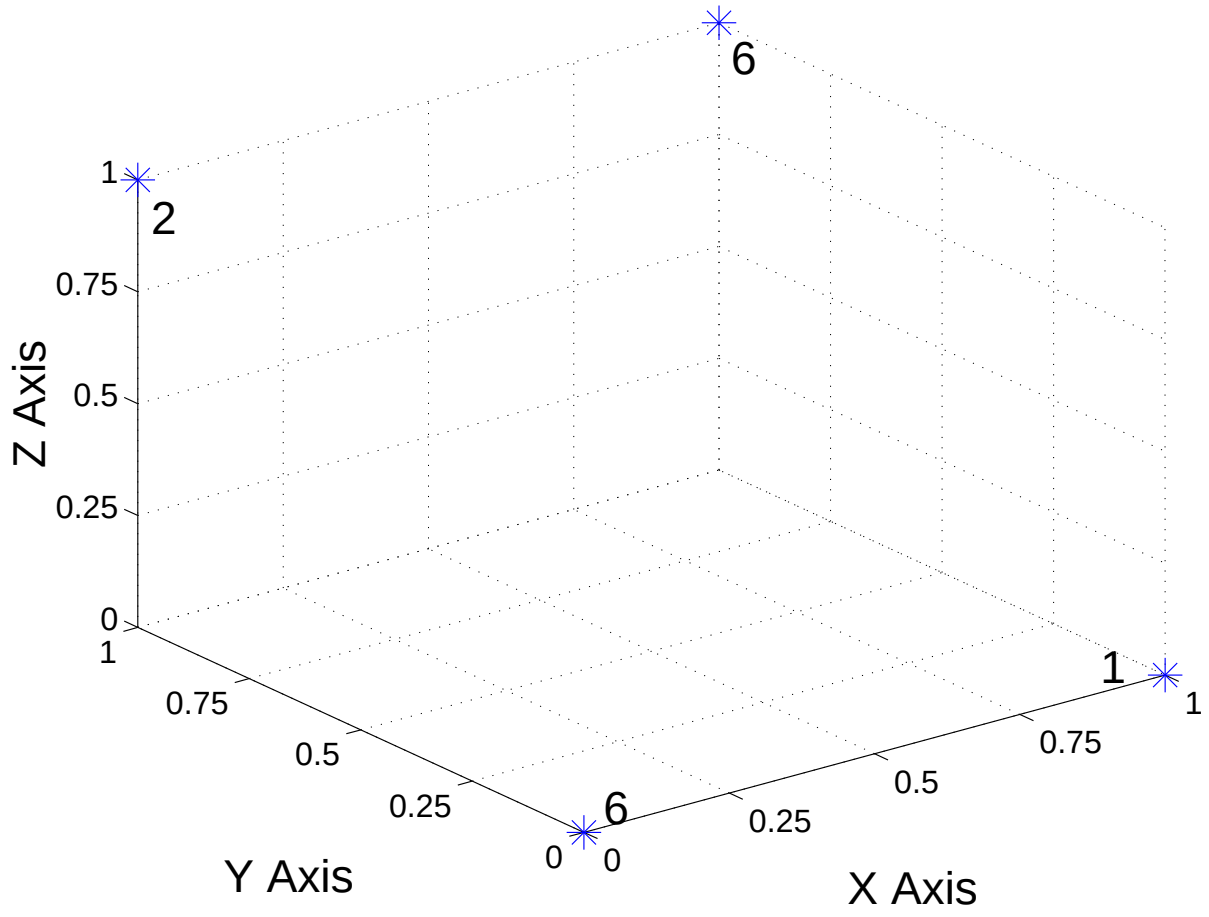
It requires an additional variational approximation and iterative process.

Model fitting improves the lower bound on the log-likelihood.

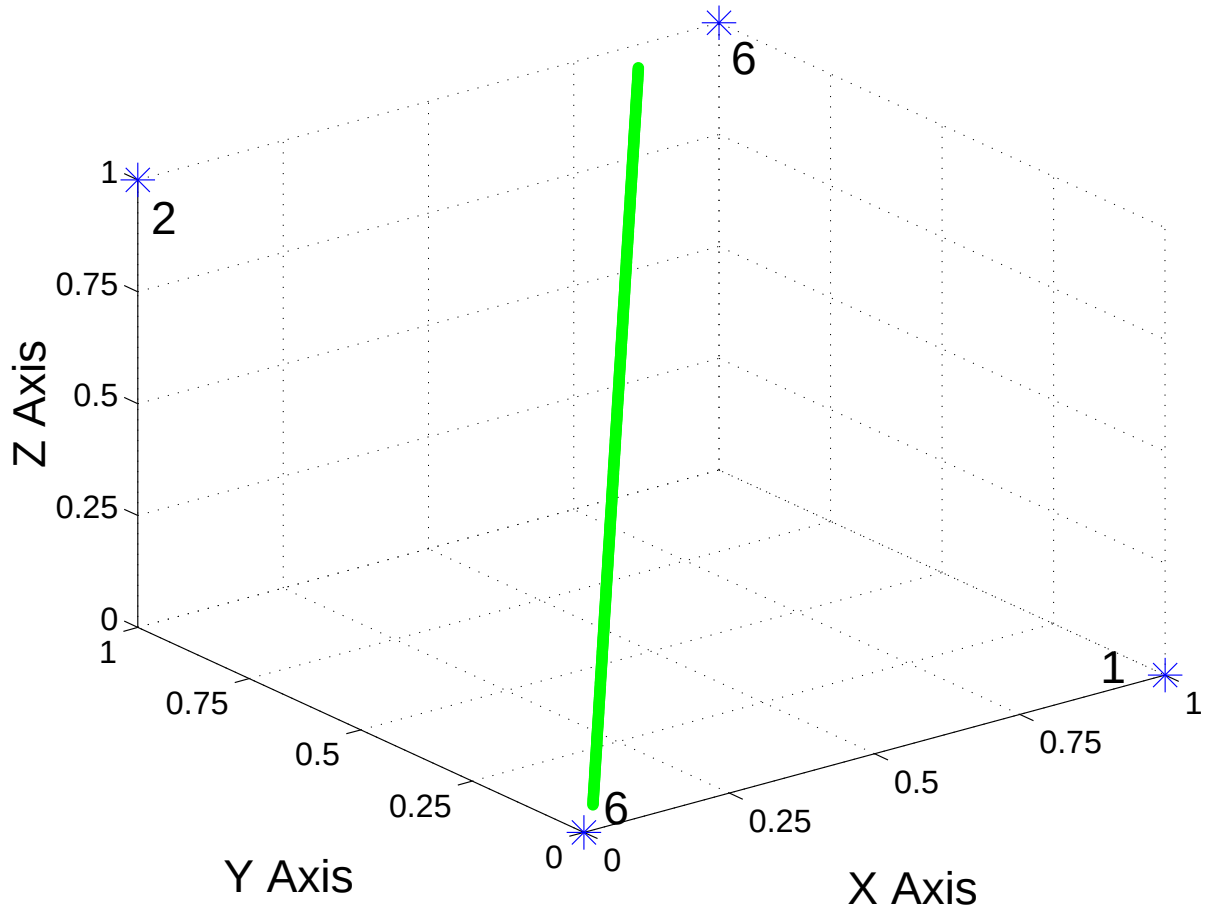
...not necessarily the log-likelihood itself.

In contrast, ALS for LPCA guarantees an increase in the log-likelihood.

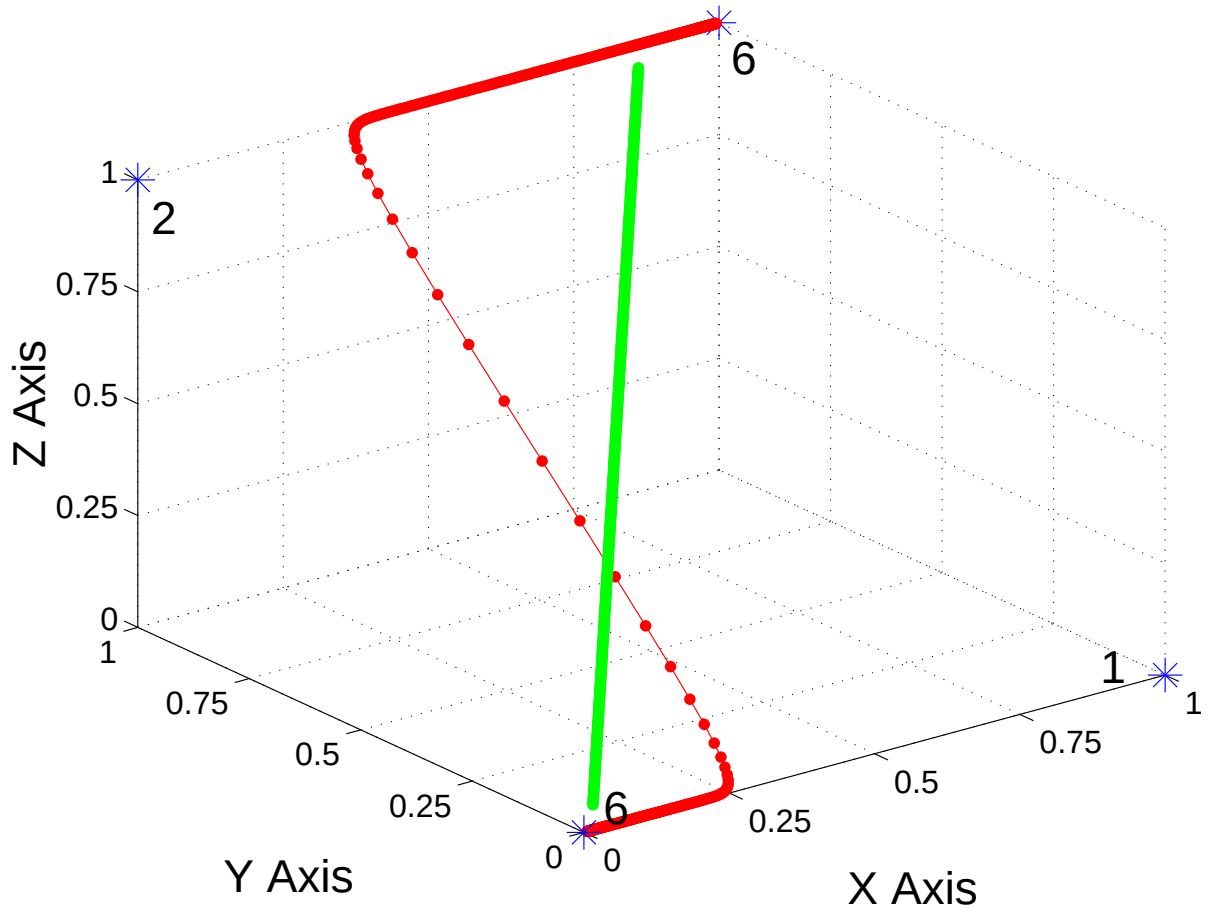
Example in 3 Dimensions



Example in 3 Dimensions



Example in 3 Dimensions



Empirical Evaluation: Data Reconstruction

$X =$ Original Data

$R =$ Reconstructed Data

$N =$ Number of Observations

$D =$ Dimensionality of the Data

$$\text{Error} = \frac{\sum_n^N \sum_d^D |X_{nd} - R_{nd}|}{N * D}$$

Microsoft Web Log Reconstruction Results

Web log shows URL visitation by anonymized users.

Data Set: a matrix of users and URLS clicked on

$N = 32711$ Observations are a session
$D = 285$ URLS Clicked
Density $= 0.011$

Data Set Task: Build a recommender system of URLS.

Our Task: Data Reconstruction

Microsoft Web Log Reconstruction Results

Error Rates (%)

L	Linear PCA	Logistic PCA
1	1.52	1.28
2	1.41	1.15
4	1.36	0.760
8	1.11	0.355

1 LPCA dimension $\simeq$ 6 PCA dimensions

Advertising Data Reconstruction Results

A UC Irvine data set of web linked images and surrounding features.

$N = 3279$ image links
$D = 1555$ context features
density $= 0.072$

Data Set Task: Predict whether an image is an advertisement

Our Task: Data Reconstruction

Features include phrases in the anchor text and around image:

microsoft.com, toyotaofroswell.com, home+page

Advertising Data Reconstruction Results

Error Rates (%)

L	Linear PCA	Logistic PCA
1	2.68	1.97
2	2.39	1.20
4	2.17	0.626
8	1.76	0.268

1 LPCA dimension $\simeq$ 7 PCA dimensions

Other Data Sets (in paper)

- Microarray Gene Expression Data
  - Observations are genes
  - Attributes are environmental conditions
  - Binary values indicate whether genes are expressed or not
- MovieLens Movie Ratings Data
  - Observations are users
  - Attributes are movies
  - Binary values indicate whether a user rated a movie or not

Related Models

Both of these models share a decomposition:

$$\Theta_{nd} = (UV)_{nd}$$

- Factor Analysis: A generative relative of PCA
- Multinomial PCA (MPCA): A multinomial, generative variant of PCA

  MPCA is represented in the proceedings: [Buntine and Perttu, 2003].

Summary

- We derive and implement the ALS algorithm for fitting LPCA.
- In data reconstruction experiments, LPCA outperforms PCA.
- LPCA is well suited for smoothed probability models of binary data:
  - People and URLs they click
  - Phrase features surrounding image links
- Future work will explore LPCA in other traditional PCA tasks.
  - Feature extraction
  - Machine learning