SAND2016-7226PE

Exceptional service in the national interest

Sandia National Laboratories

Mixed Factor Analysis and Model-Comparison Metrics

Jessica Gronski*†, Tamara G Kolda†, Cliff Anderson-Bergman†

Motivation: Mixed Variable Factor Analysis

- Divergence between objects in **"mixed"** feature space
  - Gaussian (real-valued)
  - Bernoulli (binary)
  - Poisson (counts)
  - And others, including non-numeric
- Applications
  - Clustering
  - Inputs to Supervised Learning
  - Matrix Completion
  - Recommender Systems
  - Outlier Detection

	Feature 1	Feature 2	Feature 3	
Object 1	1	0.7	18	⋯
Object 2	1	0.5	15	⋯
Object 3	0	−0.2	3	⋯
Object 4	1	−0.5	7	⋯
	⋮	⋮	⋮	⋱

Binary Counts Real Values

Principal Components Analysis (PCA)

- Fundamental statistical tool for large, noisy datasets
- Useful for detecting **low-dimensional** structure

Given: Observed matrix $\mathrm{X} = m\ x\ n$

Goal: Determine low-rank approximation

$$\mathrm{X} \approx \Theta = AB^T$$

by **minimizing sum squared errors**

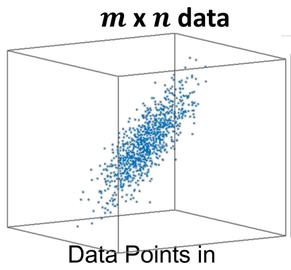
$$\min_{\Theta} \sum_{ij} (x_{ij} - \theta_{ij})^2$$

$A = m\ x\ k$ new representations (weightings)

$B = n\ x\ k$ principal components

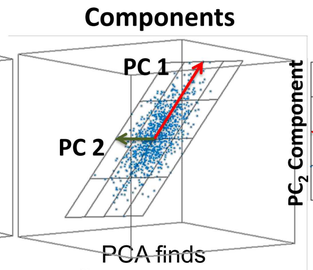
- Common practice to "**center and scale**" each matrix column before looking for principal components.
  - Columns given equal weighting
- **Ideal for normally distributed data!**

$\mathbf{X \approx \Theta = AB^T}$ $\boldsymbol{m}$ **x** $\boldsymbol{n}$ **data**



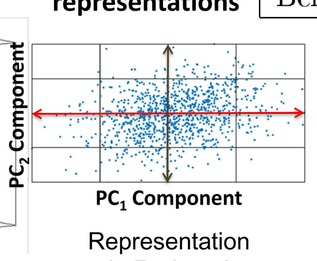
Data Points in Ambient 3D Space (n = 3)

$\boldsymbol{B = n}$ **x** $\boldsymbol{k}$ **Principal Components**



PCA finds Principal Components

$\boldsymbol{A = m}$ **x** $\boldsymbol{k}$ **new representations**



Representation in Reduced Space (k = 2)

Udell et al.'s Generalized Low-Rank Models (GLRM)

Define the *Generalized Low-Rank Model*

$$\min_{\theta} \sum_{ij} L_{ij}(x_{ij}, \theta_{ij})$$

- $L_{ij}(x_{ij}, \theta_{ij}) = (x_{ij} - \theta_{ij})^2$ yields PCA
- Udell et al. [1] introduce **standardized general loss functions**
  - Authors give **popular loss functions** corresponding to exponential family members
    - Gaussian, Poisson, Bernoulli, etc.
  - Similar to PCA, **center and scale** columns **according to loss**
    - Center: Single parameter minimizing column loss
    - Scale: Generalized variance about column "center"

References

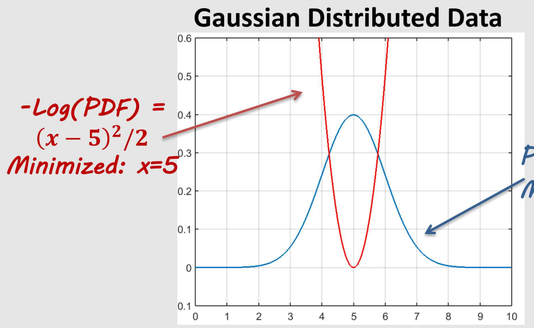
[1] Udell, M., Horn, C., Zadeh, R., & Boyd, S. (2014). Generalized low rank models. *arXiv preprint arXiv:1410.0342*.

[2] Collins, M., Dasgupta, S., & Schapire, R. E. (2001). A generalization of principal components analysis to the exponential family. In *Advances in neural information processing systems* (pp. 617-624).

Our Standardization of GLRMs

- Re-derivation of Udell et al.'s GLRM
- Statistically-derived loss functions
  - Based on probability mass or density function
- Maximizing probability ≡ Minimizing negative log-likelihood

Gaussian Distributed Data



Let L_j be the loss defined over column j, define column

- Center:

$$\bar{\theta}_j = \operatorname{argmin}_{\theta \in \mathbb{R}} \sum_i L_j(x_{ij}, \theta)$$

- Shift:

$$\delta_j(x) = \inf_{\theta \in \mathbb{R}} L_j(x, \theta)$$

- Scale:

$$\eta_j = \frac{1}{m_j} \sum_i (L_j(x_{ij}, \bar{\theta}_j) - \delta_j(x)) \geq 0$$

Our Standardized Loss:

$$\hat{L}_j(x, \theta) = \frac{L_j(x, \theta + \bar{\theta}_j) - \delta_j(x)}{\eta_j}$$

- Every loss function has a *link*, $x = g(\theta)$, relating mean to model parameters

Name	$L(x,\theta)$	$\delta(\theta)$	θ	$g(\theta)$
Gaussian	$(x-\theta)^2$	0	$\bar{x}$	x
Poisson	$\theta - x\log(\theta)$	$x - x\log(x)$	$\bar{x}$	x
Poisson	$e^\theta - x\theta$	$x - x\log(x)$	$\log(\bar{x})$	$\log(x)$
Bernoulli	$\log(1+e^\theta) - x\theta$	0	$\mathrm{logit}(\bar{x})$	$\mathrm{logit}(x)$

Underlying Data Distribution — Loss Function — Shift (Minimum Achievable Loss) — Center of Low-Rank Matrix — Link Function

Models to Compare

- **No centering/scaling** but appropriate losses
- **PCA** (centering/scaling but sum squared error loss)
- **Our standardization** (centering/scaling/shifting appropriate losses)

Suitable "Goodness" Measures

- **Bregman Divergence**
  - Generalized squared Euclidean distance to class of divergences sharing similar properties
  - Measures error in first order approximation of unscaled loss between two quantities

$$\boldsymbol{B_F(p||q) = F(p) - F(q) - F'(q)(p - q)}$$

where $\boldsymbol{F(x)}$ is a minimizer [2]

Squared Error — KL-Divergence — $L(x,\theta) - \delta(\theta)!$

Exponential Family	$B_F(p\|\|q)$	$p \to x, q \to g(\theta)$
Gaussian (MSE)	$(p-q)^2$	$(x-\theta)^2$
Poisson (MKL)	$p\log\frac{p}{q} + q - p$	$e^\theta - x\theta - x + x\log(x)$
Bernoulli (MLL)	$p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$	$\log(1+e^\theta) - x\theta$

Logistic Loss

- **Normalized Residual**

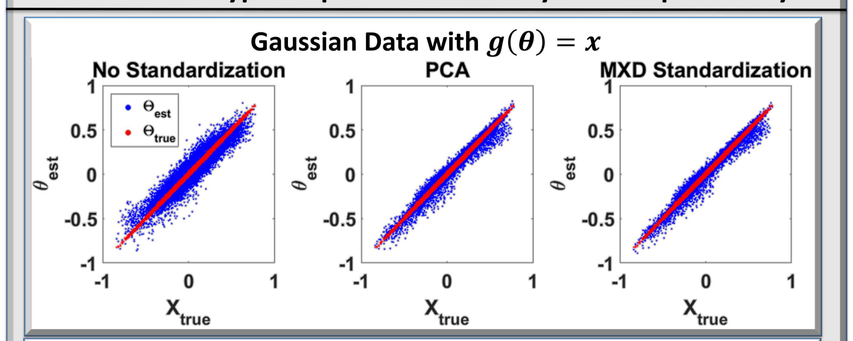
$$NR(\Theta) = \frac{\left\|\Theta - \hat{\Theta}\right\|_F^2}{\left\|\Theta\right\|_F^2}$$ where $||\ .\ ||_F^2$ is squared Frobenius norm

Example Data Attributes

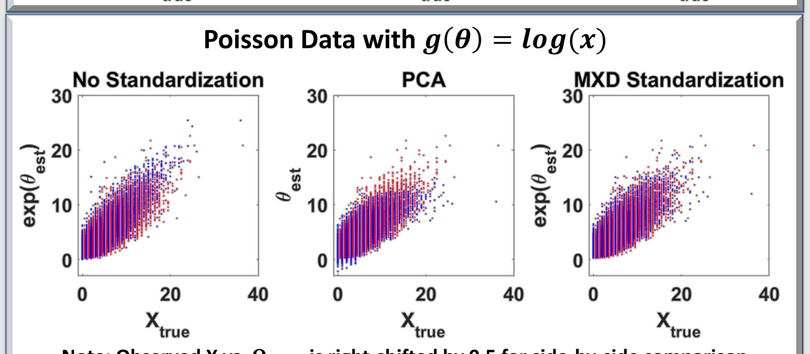
- $m = 1000$ Objects
- $n = 75$ Features: 25 Gaussian, 25 Poisson, 25 Bernoulli
- Rank, $k = 5$
- 5% data withheld to test
- $A \sim U(0,1)$
- $B \sim N(0, 1/5), U(-1,1), N(0, 1.44/5)$
- $x_{ij} \sim N(\theta_{ij}, 0.01), P(\lambda_{ij} = e^{\theta_{ij}}), B(p_{ij} = 1/(1+e^{-\theta_{ij}}))$
- Fit Θ w/ 30 different random starts

Comparing Observed X vs. $\Theta_{true} = AB^T$, Model Θ_{est} where data types separated and analyzed independently

Gaussian Data with $g(\theta) = x$



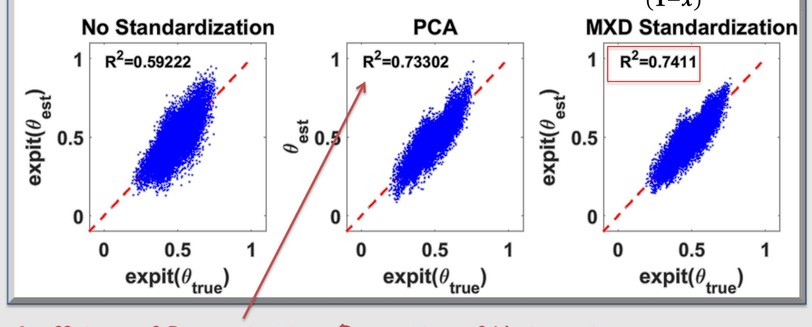
Poisson Data with $g(\theta) = log(x)$



Note: Observed X vs. Θ_{true} is right-shifted by 0.5 for side-by-side comparison

Comparing Θ_{true} vs Θ_{est} with Ideal

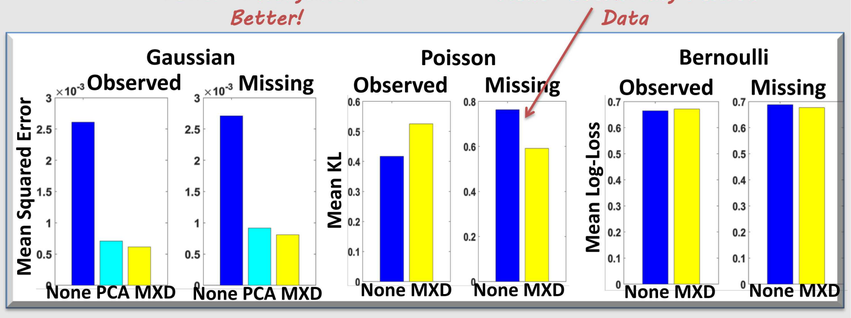
Bernoulli Data with $g(\theta) = logit(x) = log\frac{x}{(1-x)}$



Coefficient of Determination: Proportion of Variance in Dependent Variable Predicted from Independent Variable

Mean Divergence over 30 random starts

Lower Divergence is Better! — 'None' Overfitting Poisson Data



Note: PCA minimizes squared error and may result in negative θ_{ij}. Therefore, applying Bregman Divergence other than squared error does not make sense.

Numerical Comparisons

Description:	None	PCA	MXD
Normalized Residual Θ_{est}	1.098e-01	5.426e+00	**3.610e-02**
MSE Gaussian Obs.	2.610e-03	7.056e-04	**6.130e-04**
MSE Gaussian Miss.	2.708e-03	9.162e-04	**8.066e-04**
MKL Poisson Obs.	**4.169e-01**	N/A	5.254e-01
MKL Poisson Miss.	7.630e-01	N/A	**5.914e-01**
MLL Bernoulli Obs.	**6.645e-01**	N/A	6.710e-01
MLL Bernoulli Miss.	6.876e-01	N/A	**6.766e-01**

U.S. DEPARTMENT OF ENERGY

National Nuclear Security Administration

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.
SAND No. 2011-XXXXP

* Ph.D. Student, University of Colorado – Boulder, Department of Applied Mathematics
† Sandia National Laboratory Org. 8962, Livermore, CA