MLDL

SAND2018-8213C

Machine Learning and Deep Learning Conference 2018

XPCA: Extending PCA for Combinations of Discrete and Continuous Data

- Kina Kincher-Winoto/8762
- Tammy Kolda (8759), Cliff Anderson-Bergman (LLNL)

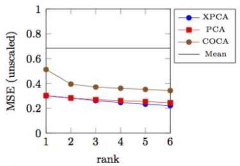
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

The Project

MXD is a project that aims to provide alternatives to principal component analysis (PCA) to MiXeD data types

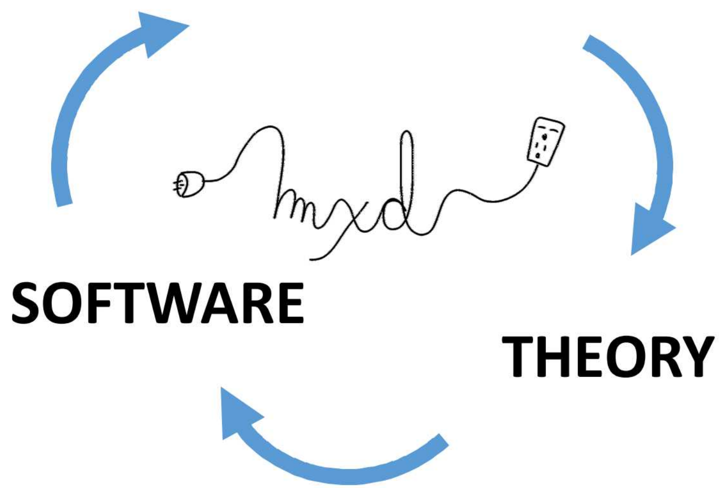
	XPCA	PCA	COCA	Column Mean
Rank 1	**0.466**	0.480	0.489	1.005
Rank 2	**0.350**	0.379	0.364	1.005
Rank 3	**0.330**	0.365	**0.330**	1.005
Rank 4	**0.319**	0.341	0.377	1.005
Rank 5	**0.316**	1.476	0.383	1.005
Rank 6	**0.340**	3.307	0.430	1.005
Rank 7	**0.344**	11.754	0.532	1.005
Rank 8	**0.329**	14.030	0.653	1.005
Rank 9	**0.353**	9.287	0.717	1.005
Rank 10	**0.336**	53.885	0.717	1.005

Table 1: Rescaled MSE by decomposition type



(a) In-sample

APPLICATION



python™

R

Generalized Low Rank Models

Madeleine Udell
Operations Research and Information Engineering
Cornell University
udell@cornell.edu

Corinne Horn
Electrical Engineering

XPCA: Extending PCA for a Combination of Discrete and Continuous Variables

Clifford Anderson-Bergman CIANDE@SANDIA.GOV
Tamara G. Kolda TGKOLDA@SANDIA.GOV
Kina Kincher-Winoto KWINOTO@SANDIA.GOV
Sandia National Labs, Livermore, CA, USA

Abstract

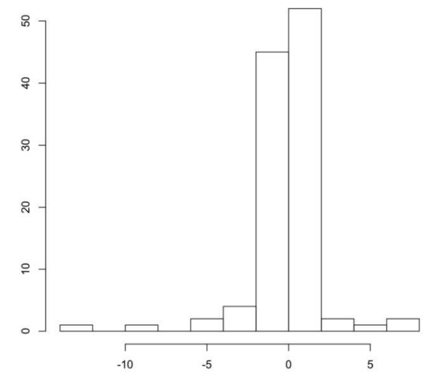
Principal component analysis (PCA) is arguably the most popular tool in multivariate exploratory data analysis. In this paper, we consider the question of how to handle heterogeneous variables that include continuous, binary, and ordinal. In the probabilistic interpretation of low-rank PCA, the data has a normal multivariate distribution and, therefore, normal marginal distributions for each column. If some marginals are continuous but not normal, the semiparametric copula-based principal component analysis (COCA) method is an alternative to PCA that combines a Gaussian copula with nonparametric marginals. If some marginals are discrete or semi-continuous, we propose a new extended PCA (XPCA) method that also uses a Gaussian copula and nonparametric marginals and accounts for discrete variables in the likelihood calculation by integrating over appropriate intervals. Like PCA, the factors produced by XPCA can be used to find latent structure in data, build predictive models, and perform dimensionality reduction. We present the new model, its induced likelihood function, and a fitting algorithm which can be applied in the presence of missing data. We demonstrate how to use XPCA to produce an estimated full conditional distribution for each data point, and use this to produce to provide estimates for missing data that are automatically range respecting. We compare the methods as applied to simulated and real-world data sets that have a mixture of discrete and continuous variables.

Keywords: principal component analysis (PCA), copula component analysis (COCA), matrix completion, heterogeneous data

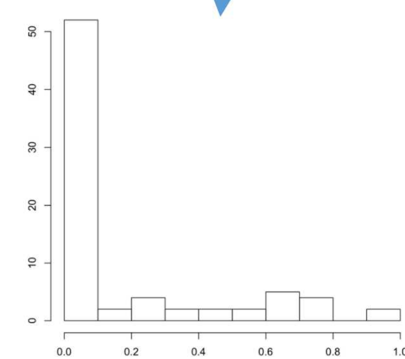
Our Problem

Principal Component Analysis (PCA)

- Standard Statistical Tool
- *Not all column marginals are gaussian*
- *Sensitive to scaling*
- *Sensitive to outliers*



Is this data Gaussian?



Problem:

How can we relax the assumptions of PCA, ultimately to handle data that is binary, ordinal or continuous?

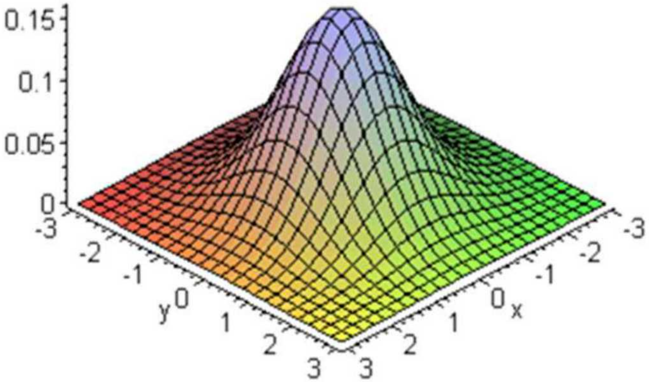
Enter copulas

Copula (kɑ·pu·lə): A connecting word, in particular a form of the verb *be* connecting a subject and complement.

Copula (koʊ·pu·lə): A function that joins univariate distribution functions to form multivariate distribution functions. (Wolfram Mathworld)

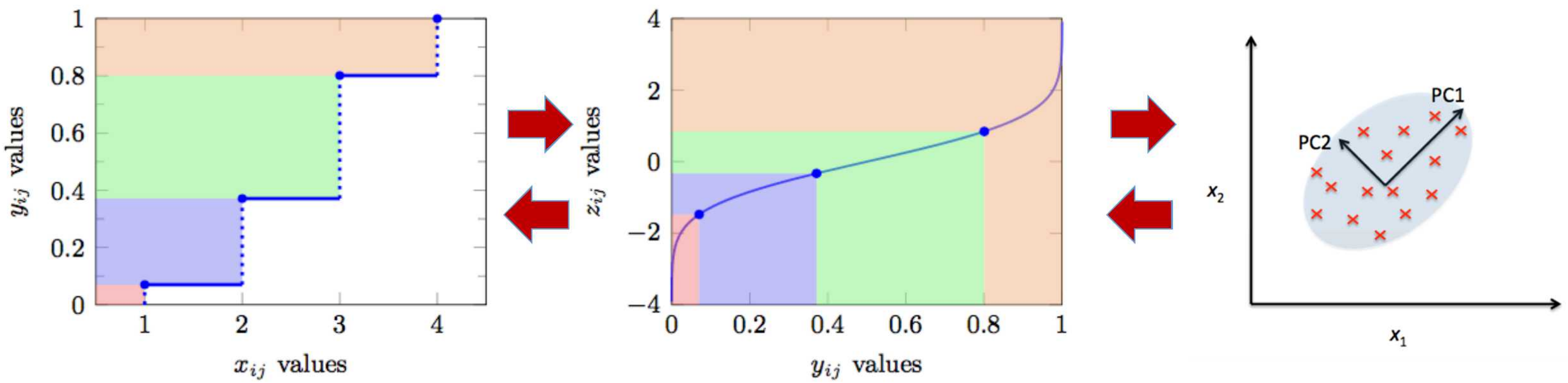
A copula model provides the decomposition of the dependency of the marginal distributions such that the copula contains the dependency structure only…We uncouple variance and dependency structure such that PCA is only influenced by the dependency in the data. (Egger et al., 2016)

Gaussian copula: Assumes multivariate normal dependency.



XPCA: eXtension of PCA

XPCA estimates the marginal distributions of each column and accounts for discrete variables in the likelihood calculation by integrating over appropriate intervals.

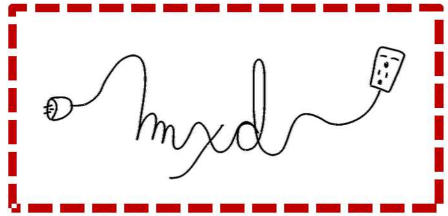


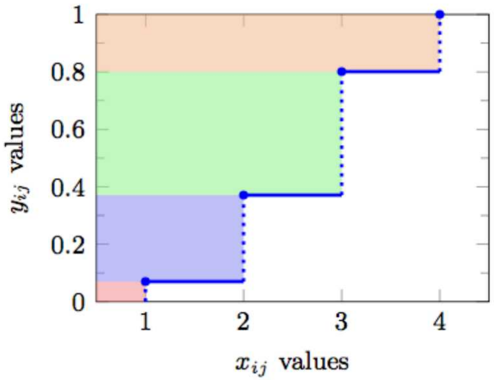
- XPCA uses copulas to better handle discrete values and outliers – both of which PCA struggles with
- XPCA is an improvement of academic research Copula Component Analysis (COCA)[1,2]

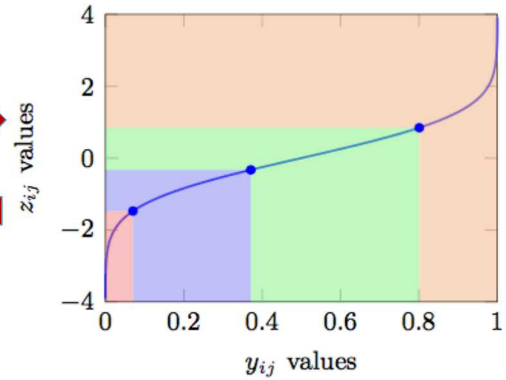
[1] Han & Liu. Semiparametric principal component analysis. 2012

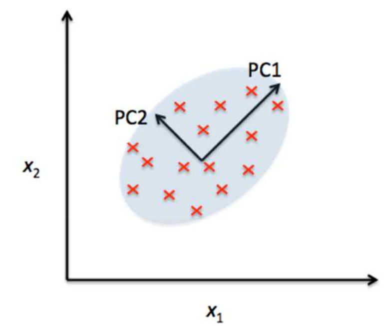
[2] Egger et al. Copula eigenfaces - semiparametric principal component analysis for facial appearance modeling. 2016

XPCA: eXtension of PCA









1

Empirical Distribution Function estimates marginal distribution.

2

A *Gaussian Copula* relates non-Gaussian variables to Gaussian latent variables via Gaussian Cumulative Distribution Function

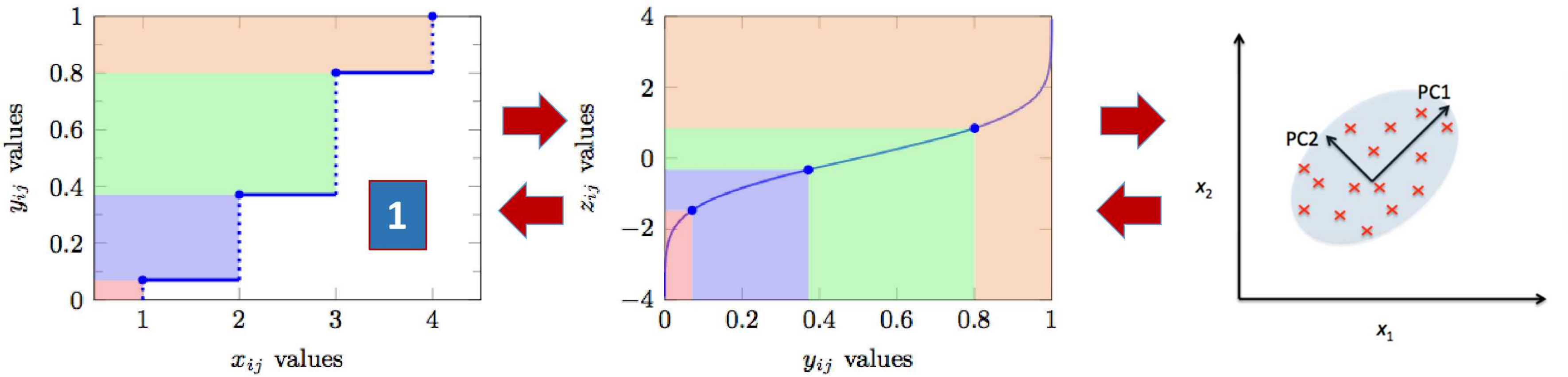
3

Latent variables follow same model as probabilistic PCA BUT subject to a different loss

4

Reverse each of those steps by taking inverse of each step

XPCA: eXtension of PCA



1 Empirical Distribution Function estimates marginal distribution.

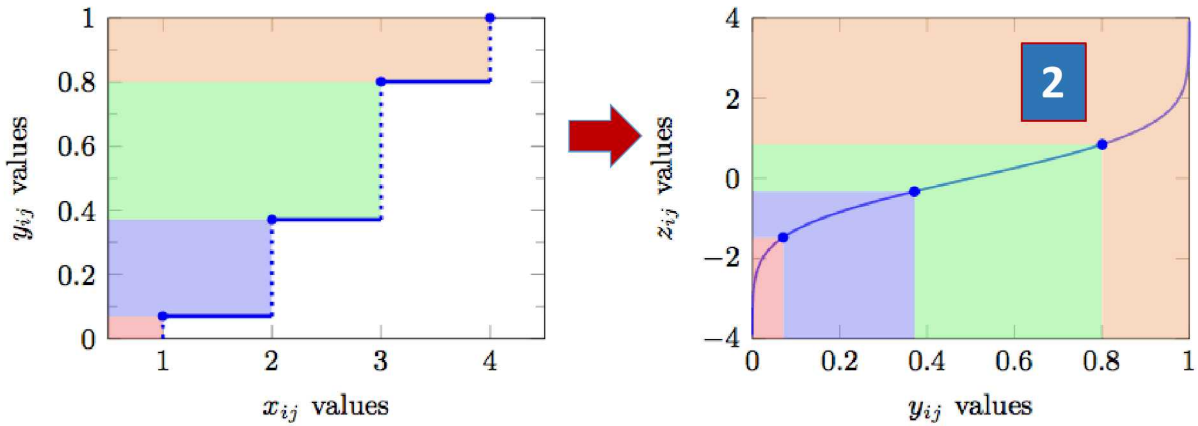
rank of s in column j

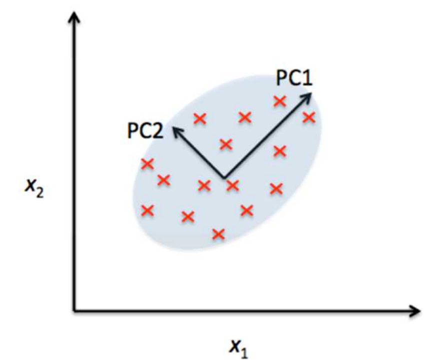
unique values in column j

$$\hat{F}_j(x) = \frac{\max\{\, r_j(s) \mid s \le x \text{ and } s \in \mathcal{C}_j \,\}}{m_j},$$

total number of entries in column j

XPCA: eXtension of PCA





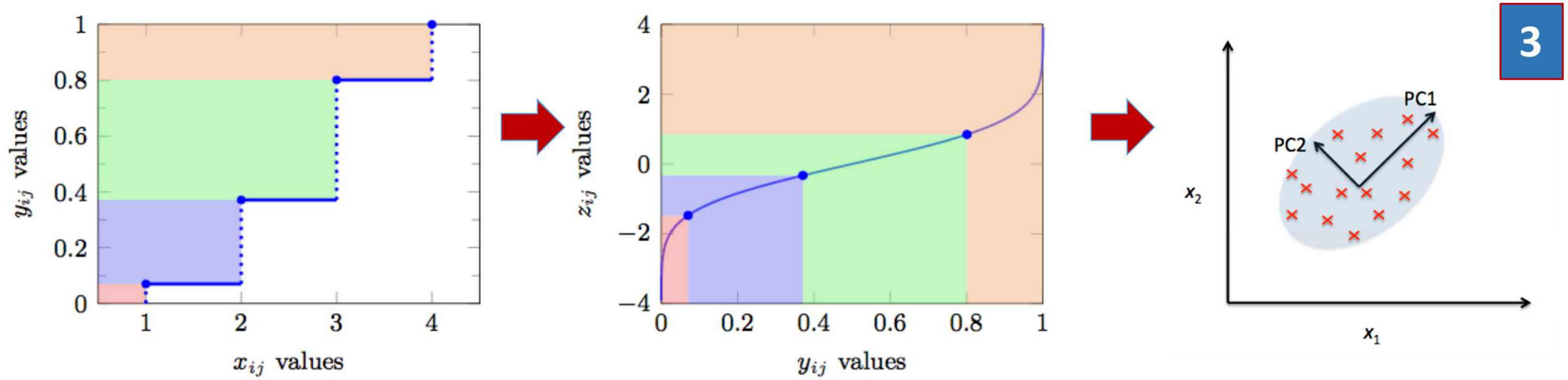
2

A *Gaussian Copula* relates non-Gaussian variables to Gaussian latent variables via Gaussian Cumulative Distribution Function

$$z_{ij} \in \left(\Phi^{-1}(\hat{F}_j(x_{ij} - \epsilon)), \Phi^{-1}(\hat{F}_j(x_{ij}))\right)$$

left side of interval right side of interval

XPCA: eXtension of PCA



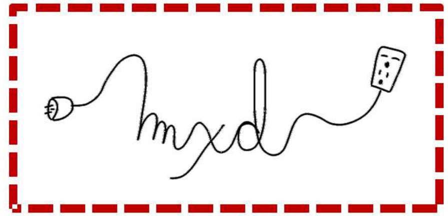
3

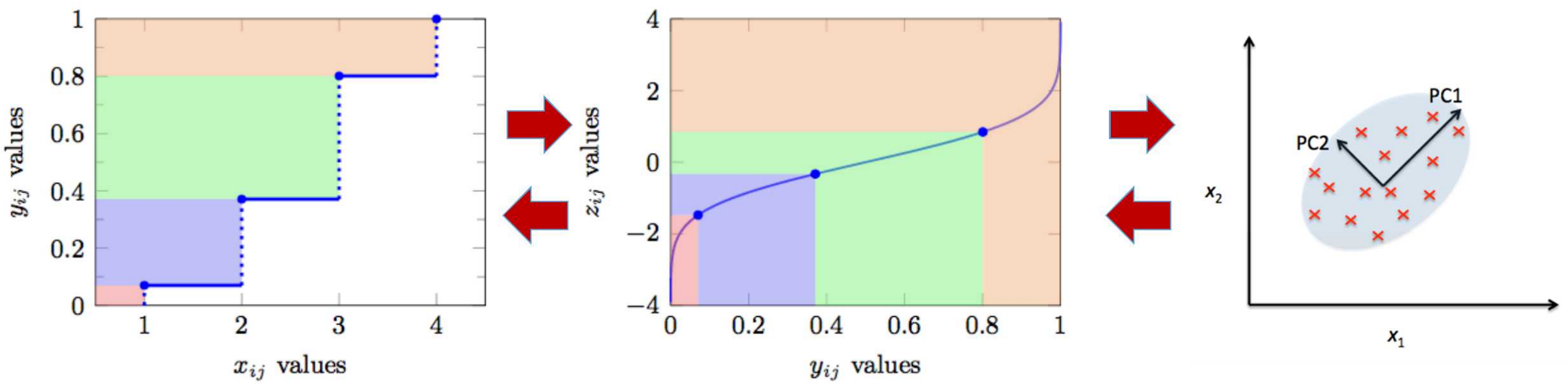
Latent variables follow same model as probabilistic PCA BUT subject to a different loss.

$$\mathbf{Z} \sim \mathrm{MVN}\left(\boldsymbol{\Theta}, \sigma^2\mathbf{I}\right) \quad \text{where} \quad \boldsymbol{\Theta} = \mathbf{U}\mathbf{V}^T$$

σ captures the magnitude of variance on top of the low rank structure

XPCA: eXtension of PCA





4 Reverse each of those steps by taking inverse of each step

Application: the data

1 Senator Voting Data from January 1989 – 2017

- 271 senators and their votes over 101st – 114th congressional sessions
  - 1 : Yay | -1: Nay | 0: abstain
- 271 rows (senators) x 9044 columns (bills)
- 63% of data is missing

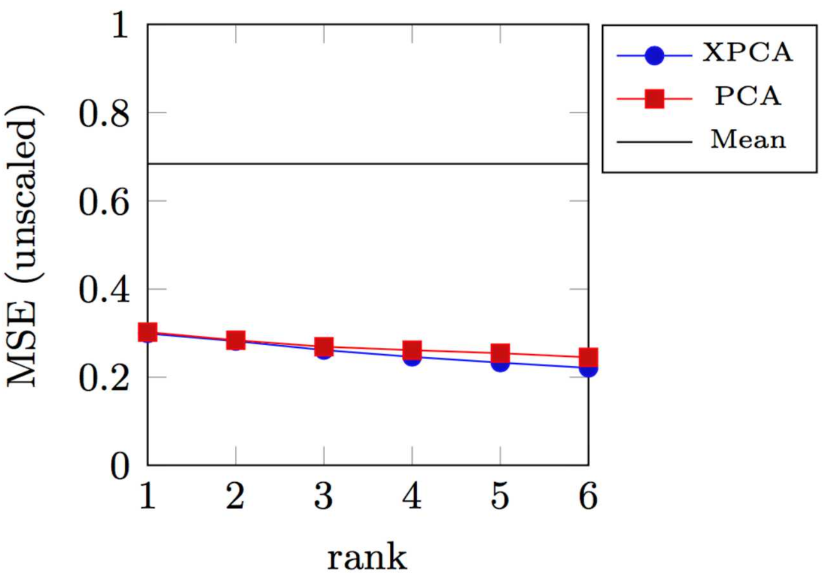
www.senate.gov/legislative/votes.htm

2 NBA Basketball Statistics from 2015-16 season

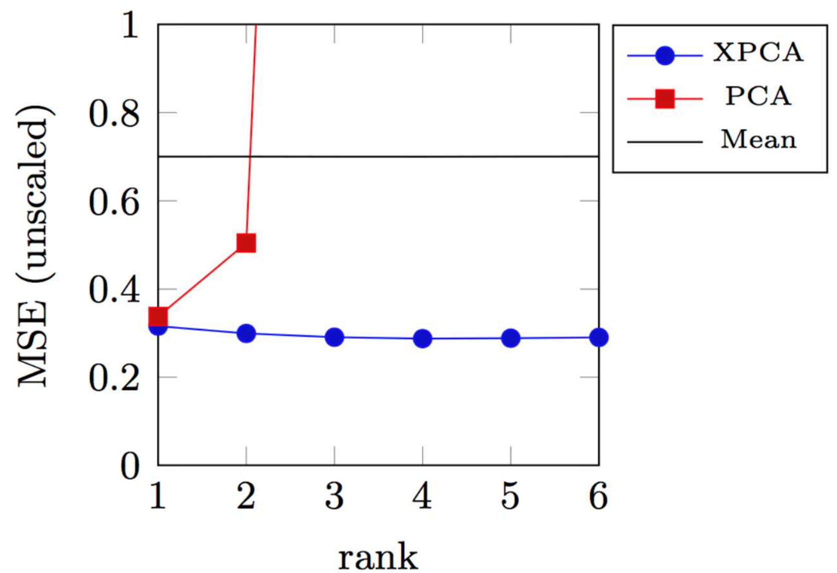
- 476 players and their various statistics
  - e.g. shots scored, number of assists, draft number
- 476 rows (players) x 40 columns (stats)

www.nba.com/stats

Senator Data: model fit



In-Sample Error
the amount of error when fit is made across all available data

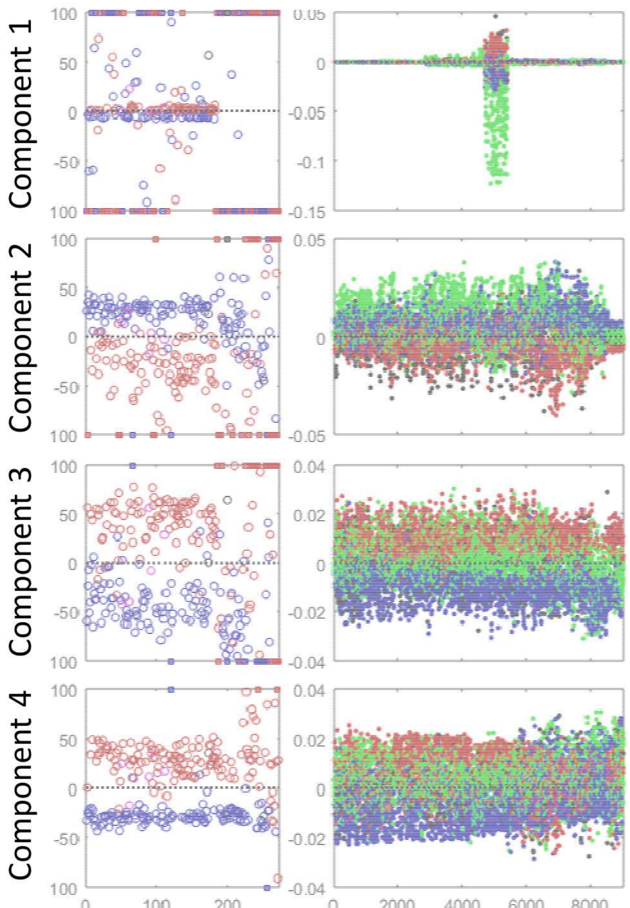


4 Fold Cross-Validation Error
the amount of error when ¼ of data is left out, calculated for each quarter

Senator Data: components

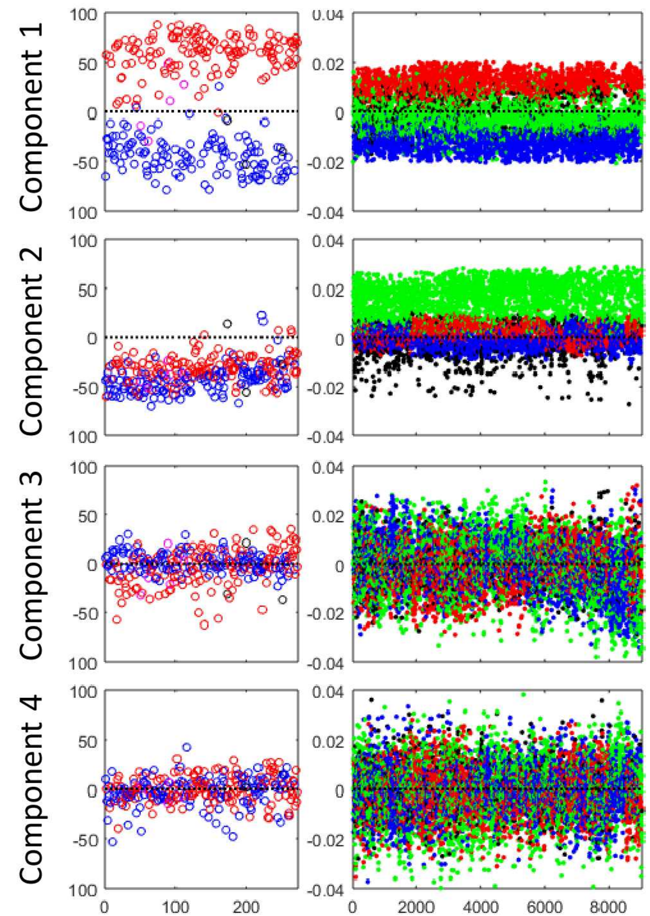
PCA

Senators Bills



XPCA

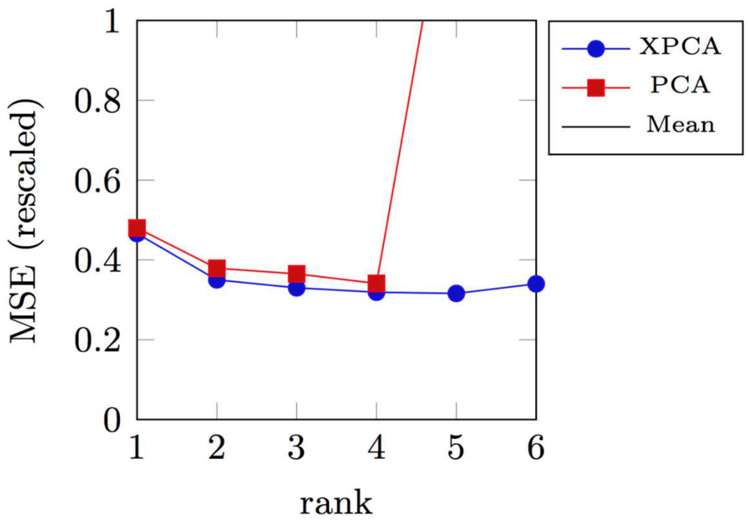
Senators Bills



- ○ Democrat
- ○ Republican
- ○ Independent
- ○ Switched parties

- ● ≥ 50% D/R Yay
- ● ≥ 50% D/R Nay
- ● ≥ 50% D Yay & ≥ 50% R No
- ● ≥ 50% R Yay & ≥ 50% D No
- ● None of above

NBA Basketball Data: Model fit



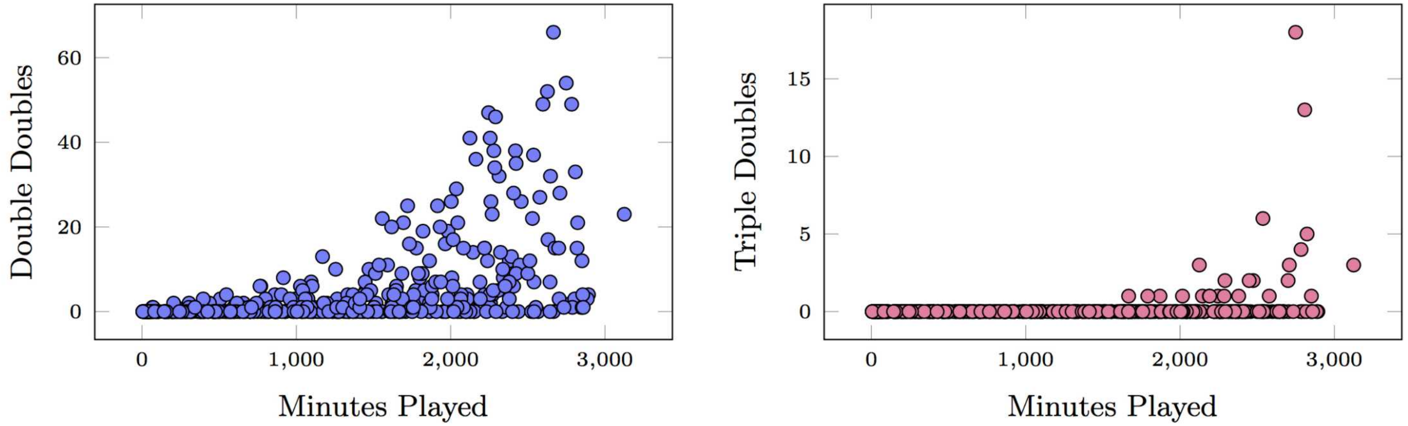
20 Fold Cross-Validation Error
the amount of error when ¼ of data is left out, calculated for each quarter

	Kevin Durant: MVP?	Anthony Morrow: MVP?
True Value	1	0
XPCA estimate at best rank (r=5)	0.3833959	0
PCA estimate at best rank (r=4)	0.1427923	-0.05687824

PCA estimates values outside range

NBA Basketball Data: Components

- The first component of both XPCA and PCA represent the minutes played and the count variables (e.g. shots made or times fouled or double doubles).
- However, XPCA also captures “triple doubles” influence
  - PCA cannot capture this influence possibly due to the heavy atom at 0



These is a clear correlation between minutes played and triple doubles. However, PCA struggles to find this correlation, while XPCA does not.

Top 5 last thoughts

Ongoing Threads:

Application into mission problems

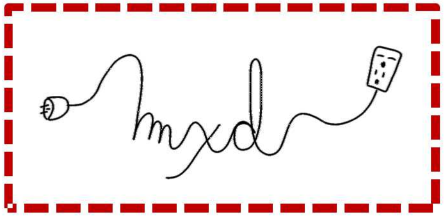
Clustering post-XPCA in the analysis stage!

Scaling up to tens of thousands of rows & columns

Future Work:

Pipeline into other machine learning applications

Scaling up to millions of rows and columns



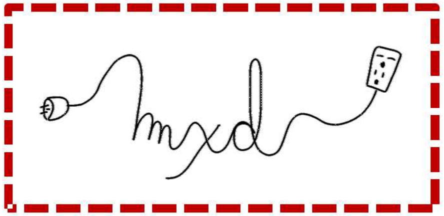
Learn more

To read more:

Paper is written (currently in R&A) and is available to read!

To try it out:

MXDLIB, the software package, is written in both R (stable) and Python (testing)



POC: Kina Kincher-Winoto
kwinoto@sandia.gov