Latent Outlier Detection and the Low Precision Problem

Fei Wang, Sanjay Chawla and Didi Surian
University of Sydney
Australia

Outline

- Background
- Low Precision Problem/Multiple Subspace View
- Ambient and Intrinsic Dimensionality
- Non-Negative Matrix Factorization
- Summary

Background

- Outliers often trigger "paradigm shifts"
  - Black swan events
- Outliers associated with low probability events
  - In Information Theory, "surprise" is **-log p**
- Looking for outliers in data → plagued with large variance/ low precision
  - Model underlying latent process

Low Variance → Easy to Explain Outliers [NBA example]

Outlier Rank	Player Name	All Star Team (Y/N)
1	Kevin Durant	Y
2	Kobe Bryant	Y
3	LeBron James	Y
4	Kevin Love	N
5	Russell Westbrook	Y

Applied Kmeans--[SDM13] on 2012 NBA players stats (~20 dim). The Top-5 outliers are easily recognized.

NBA is highly competitive (low variance) – There are no bad players in the NBA, but some are very good!

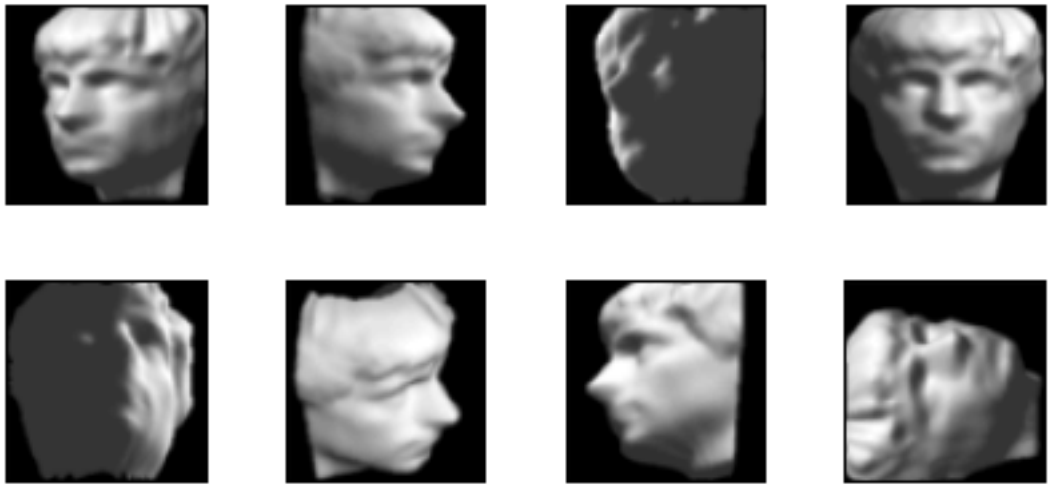
Low Precision Problem [LPP]

- In practice, Outlier Detection techniques (distance/density/subspace/statistical) suffer from LPP
- Hard to distinguish measurement noise from unexpected change in the underlying state
- ~~Curse~~ Blessing of high-dimensionality
  - Features are correlated (information redundancy)
  - Measurement noise is unlikely to be correlated across correlated features

Multiple Subspace View

- High-dimensional features are correlated.
- View data from multiple subspace perspective.
  - Genuine outliers are likely to persist across many subspaces.
- However, need to algorithmically tackle exponential size of space of subspaces
  - Can dim-reduction/projections come to our rescue ?

Ambient vs. Intrinsic Dimensionality



Yale Face Data Set: Ambient Dimension is high (number of pixels). But intrinsic dimensionality is low. Bottom-right image is an "outlier"

Non-Negative Matrix Factorization (NMF)

- Projections, in general, lead to loss of semantic information.
- For example, eigenvectors of SVD are often hard to interpret (mixed signs).
- When data is non-negative, NMF are easier to interpret

NMF Problem

- **Input:** Non-negative n x d data matrix X, k
- **Find:** n x k matrix U >= 0; k x d matrix V >=0
- Such that ||X – UV|| is minimized
- NP-Hard
- Local Search – Alternate minimization
  - Some recent progress on polynomial time algorithms under mild separation conditions.

Robust NMF

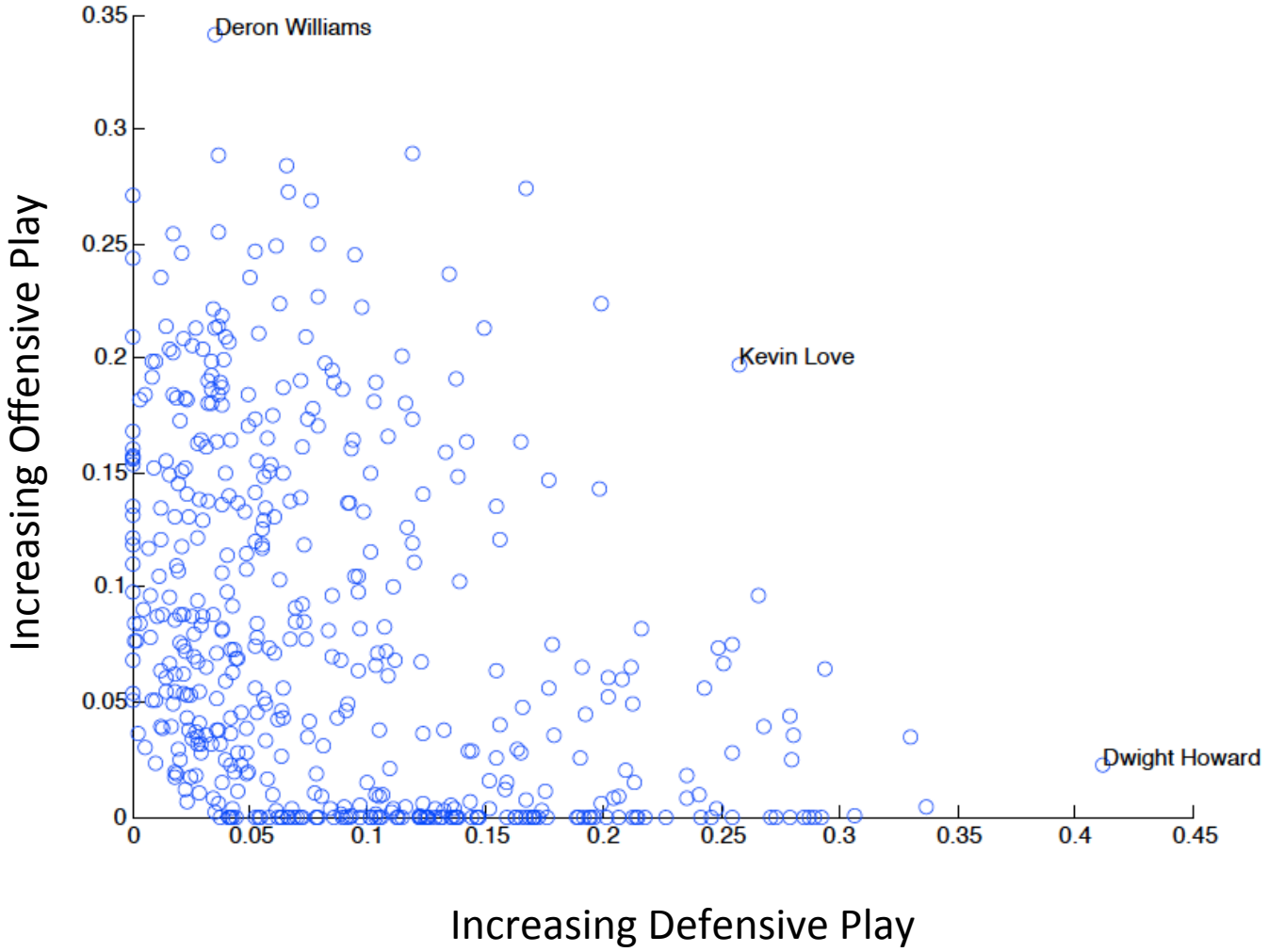
- NMF (like kmeans/SVD) is extremely sensitive to outliers
  - Theoretically, one bad outlier can have unbounded impact
  - Chicken and Egg problem
  - One contribution (of this paper) to "robustify" NMF
  - See results in the paper

R-NMF (k=2) with NBA data set

	1	2	3
1	0.0509	0.0612	'Games Started'
2	0.0673	0.0623	'Minutes Played Per Game'
3	0.0671	0.0613	'Field Goals Per Game'
4	0.0742	0.0546	'Field Goal Attempts Per Game'
5	0.1044	0	'3-Point Field Goals Per Game'
6	0.1063	0	'3-Point Field Goals Attempts Per Game'
7	0.0645	0.0565	'Free Throws Per Game'
8	0.0569	0.0651	'Free Throw Attempts Per Game'
9	0.0013	0.1014	'Offensive Rebounds Per Game'
10	0.0342	0.0871	'Defensive Rebounds Per Game'
11	0.0251	0.0932	'Total Rebounds Per Game'
12	0.0844	0.0245	'Assists Per Game'
13	0.0713	0.0523	'Steels Per Game'
14	0	0.0956	'Blocks Per Game'
15	0.0711	0.0546	'Turnovers Per Game'
16	0.0488	0.0740	'Personal Fouls Per Game'
17	0.0723	0.0563	'Points Per Game'

Feature 1 – Defensive behavior
Feature 2 – Offensive behavior

NBA plot on NMF dimensions

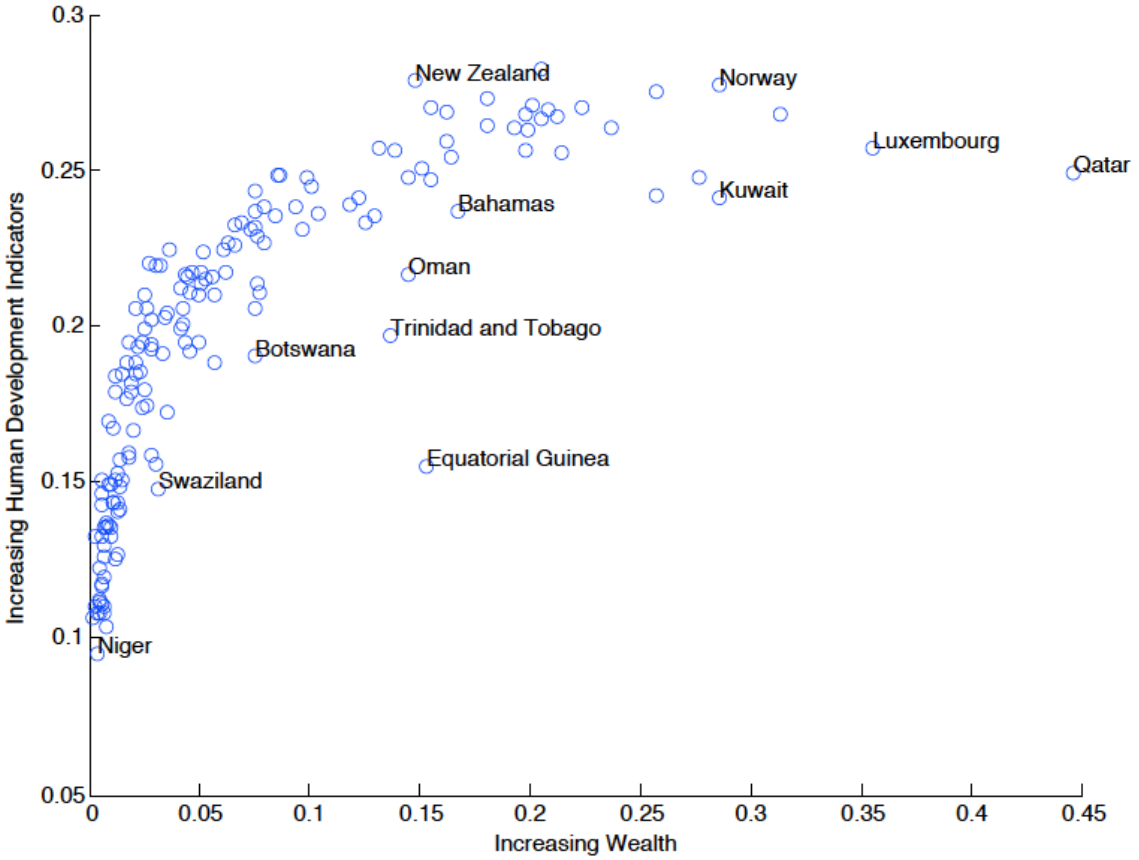


NMF on UN Country dataset

	1	2	3
1	0	0.1499	'Expected Years of Schooling (of children) (years)'
2	0.4961	0	'GDP per capita (2005 PPP $)'
3	0.5039	0	'GNI per capita in PPP terms (constant 2005 international $) (Constant 2005 international $)'
4	0	0.1493	'Health index'
5	0	0.1477	'Human Development Index (HDI) value'
6	0	0.1350	'Income index'
7	0	0.1489	'Life expectancy at birth (years)'
8	0	0.1384	'Mean years of schooling (of adults) (years)'
9	0	0.1307	'Population, urban (%) (% of population)'
10			

Data from United Nation:http://hdrstats.undp.org/en/tables/

World Economic/Well-Being Indicators



Moral: It does not take much wealth to increase well-being !

Summary

- Current Outlier Detection techniques suffers from LPP
- High-dimensionality (info. redundancy) can come to rescue
- Use multiple subspace view in NMF space to balance computational cost/variance-reduction/interpretation