New twists on eigen-analysis (or spectral) learning

Raj Rao Nadakuditi
http://www.eecs.umich.edu/~rajnrao
Role of eigen-analysis in Data Mining

- Principal Component Analysis
- Latent Semantic Indexing
- Canonical Correlation Analysis
- Linear Discriminant Analysis
- Multidimensional Scaling
- Spectral Clustering
- Matrix Completion
- Kernalized variants of above

- Eigen-analysis synonymous with Spectral Dim. Red.
Mechanics of Dim. Reduction

- Many heuristics for picking dimension
  - “Play-it-safe-and-overestimate” heuristic
  - “Gap” heuristic
  - “Percentage-of-explained-variance” heuristic
Motivation for this talk

- Large Matrix Valued Dataset Setting:
  - High-Dimensional Latent Signal Variable + Noise

  "Out intuition in higher dimensions isn’t worth a damn"

George Dantzig, MS Mathematics, 1938 U. of Michigan

Random matrix theory = Science of eigen-analysis
New Twists on Spectral learning

1) All (estimated) subspaces are not created equal

2) Value to judicious dimension reduction

3) Adding more data can degrade performance

Incorporated into next gen. spectral algorithms

- Improved, data-driven performance!
- Match or improve on state-of-the-art non-spectral techniques
Analytical model

\[ \tilde{X}_n = \sum_{i=1}^{k} \theta_i u_i v_i' + X_n \]

- Low dimensional (\(= k\)) latent signal model
- \(X_n\) is an \(n \times m\) Gaussian “noise-only” matrix
- \(c = n/m = \# \text{ rows} / \# \text{ columns of data set}\)
- Theta ~ SNR
1) All estimated subspaces are not equal

\[ |\langle \tilde{u}_i, u_i \rangle|^2 = \frac{\theta_i^4 - c}{\theta_i^4 + c \theta_i^2} + o(1) \]

- \( c = \# \text{ rows} / \# \text{ columns in data set} \)
- \( \Theta \sim \text{SNR} \)
- Subspace estimates are biased (in geometric sense above)
2) Value of judicious dim. reduction

\[
|\langle \tilde{u}_i, u_i \rangle|^2 = \begin{cases} 
\frac{\theta_i^4 - c}{\theta_i^4 + c \theta_i^2} + o(1) & \text{if } \theta_i \geq c^{1/4} \\
\theta_i & \text{otherwise.}
\end{cases}
\]

- “Playing-it-safe” heuristic injects additional noise!
Many heuristics for picking dimension
- “Play-it-safe-and-overestimate” heuristic
- “Gap” heuristic
- “Percentage-of-explained-variance” heuristic
What about the gap heuristic?

(a) Eigenvalue: $\theta > \theta_c$

(c) Eigenvalue: $\theta \leq \theta_c$

- No “gap” at breakdown point.
Percentage-of-variance heuristic?

- $O(1)$ eigenvalues that look “continuous” are noise!
  - Including those dimensions injects noise!
  - Value of judicious dimension reduction!
3) More data can degrade performance

\[
|\langle \tilde{u}_i, u_i \rangle|^2 = \begin{cases} 
\frac{\theta_i^4 - c}{\theta_i^4 + c \theta_i^2} + o(1) & \text{if } \theta_i \geq c^{1/4} \\
o(1) & \text{otherwise.}
\end{cases}
\]

- \( c = n/m = \# \text{ rows} / \# \text{ columns} \)
- Consider \( n = m \) so \( c = 1 \)
  - \( n' = 2n, m' = m \)
  - New critical value = \( 2^{1/4} \) x Old critical value!
  - Weaker latent signals now buried!
  - Value to adding “correlated” data and vice versa!
Role of eigen-analysis in Data Mining

- Principal Component Analysis
- Latent Semantic Indexing
- Canonical Correlation Analysis
- Linear Discriminant Analysis
- Multidimensional Scaling
- Spectral Clustering
- Matrix Completion
- Kernalized variants of above

- Eigen-analysis synonymous with Spectral Dim. Red.
New Twists on Spectral learning

1) All (estimated) subspaces are not created equal

2) Value to judicious dimension reduction

3) Adding more data can degrade performance

Incorporated into next gen. spectral algorithms
- Match or improve on state-of-the-art non-spectral techniques
- Role of random matrix theory in data-driven alg. design

http://www.eecs.umich.edu/~rajinrao/research.html