Spectral approximation methods

Performance evaluations in clustering and classification

Daniel McDonald

Department of Statistics
Carnegie Mellon University

http://www.stat.cmu.edu/∼danielmc

Classification Society Annual Meetings
June 16, 2011

Joint work with Darren Homrighausen

Metric embeddings

- Spectral connectivity analysis (SCA)
  - Useful tools in classification, clustering, (regression)
  - Linear and nonlinear
  - Dimension reduction or feature creation
  - Examples: PCA and Fisher discriminant analysis, Locally linear embeddings, Hessian maps, Laplacian eigenmaps
- Given data, need positive definite kernel W
  - Slow to compute spectral decompositions of W
  - Often requires approximations

Comparisons

Approximation methods

- Nyström extension
- ‘Gaussian projection’

- Past literature focused on matrix approximation — probably not the right metric
- Theory is unclear
- Applications here: manifold recovery, clustering, classification

Outline

- Given data $V = X_1, \ldots, X_n$, define a graph $G = (V, E, \widetilde{W})$
- Choose weights, e.g.

$$\widetilde{W}_{ij} = \exp\left\{-\frac{||X_i - X_j||^2}{\epsilon}\right\}$$

  for $(i,j) \in E$

- Use normalized version, here

$$W := D^{-1/2}\widetilde{W}D^{-1/2} \qquad \text{or} \qquad W := D^{-1}\widetilde{W}$$

  with $D = \texttt{diag}(\texttt{rowSums}(\widetilde{W}))$.

Diffusion maps

- Optimal p-dimensional embedding uses eigenvectors 2 to $(p+1)$ of W
- Find U such that
$$W = U\Sigma U^T$$
  and keep the columns of U
- Could also use $\mathbb{L} = I - W$: the graph Laplacian (has the same eigenvectors)
- W is $n \times n \longrightarrow$ finding U requires $\mathrm{O}\left(n^3\right)$ computations

Nyström extension

Based on a technique for finding a numerical solution for integral equations

Algorithm (briefly)

1. Subsample W: choose $M \subset \{1, \ldots, n\}$ such that $|M| = m$, $m \ll n$

$$W^{(m)} := W_{M,M}$$

2. Compute eigendecomposition of $W^{(m)} = U^{(m)} \Sigma^{(m)} U^{(m)T}$
3. "Extend" $U^{(m)}$ via simple formula to create U^{nys}

Requires only $O(nm^2)$ computations[1]

[1] Dominant term

'Gaussian projection'

Produces an orthonormal matrix Q which approximates $\mathrm{col}(W)$

Algorithm

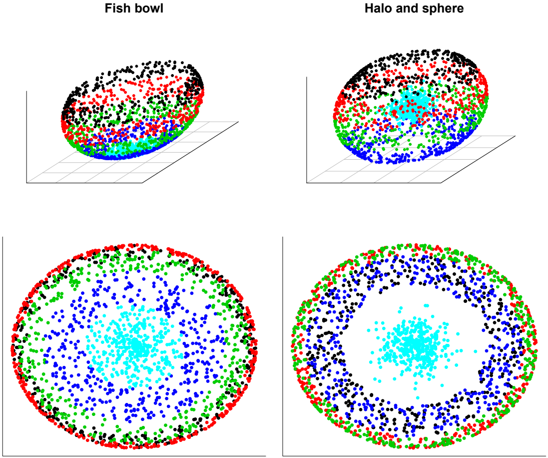
1. Draw $n \times m$ Gaussian random matrix Ω.
2. Form $Y = W\Omega$.
3. Construct Q, an orthonormal matrix such that $\mathrm{col}(Q) = \mathrm{col}(Y)$
4. Form B such that B minimizes $||BQ^\top\Omega - Q^\top Y||_2$
5. Compute the eigenvector decomposition of B, ie: $B = \widehat{U}\widehat{\Sigma}\widehat{U}^\top$
6. Return $U^{gp} = Q\widehat{U}$.

Requires $\mathrm{O}(n^2m)$ computations

U^{gp} remains orthonormal

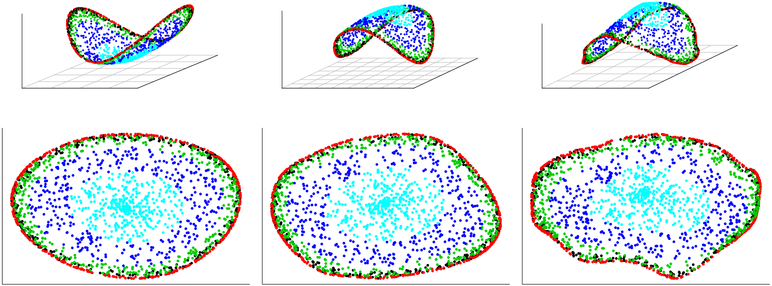
Source: Halko, Martinsson, and Tropp (2009)

MANIFOLD RECOVERY AND CLUSTERING OBJECTS



$n = 2000$

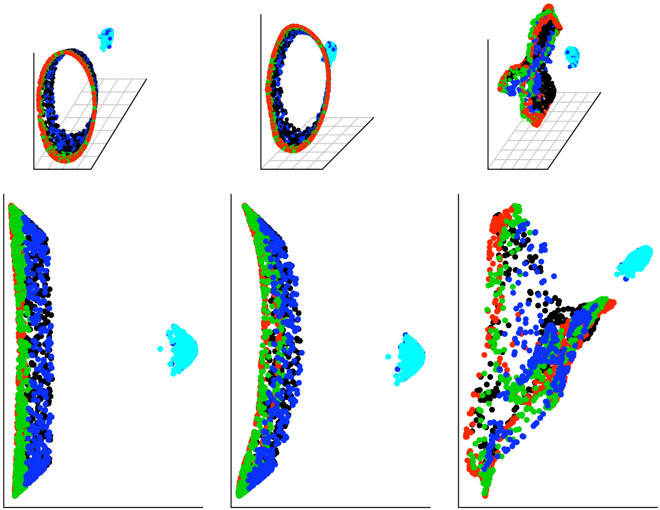
Manifold recovery



Using equivalent computational burden:

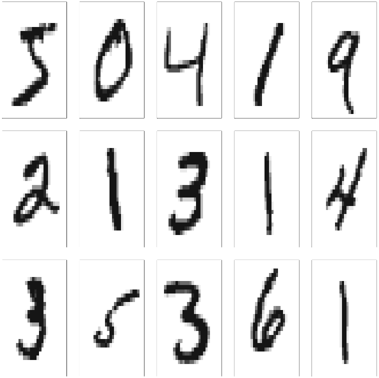
$m_{nys} = 141$ and $m_{gp} = 10$

CLUSTERING



$m_{nys} = 200$ and $m_{gp} = 20$

- Using the MNIST handwritten digit database
- Investigate for both small ($n = 4800$) and large ($n = 18000$) samples
- Semi-supervised approach: need spectral decomposition of the whole dataset
- Train on 80%, test on remainder

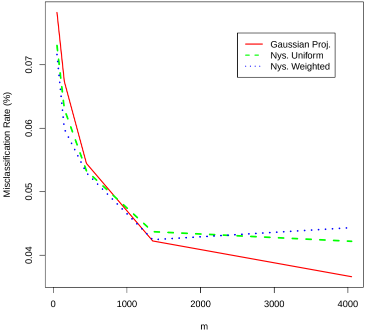


Method	# Correct
True eigenvectors	756
Uniform Nyström	697
Weighted Nyström	701
Gaussian projection	725
$n_{train} = 4000$	
$n_{test} = 800$	

- Use SVMs to perform classification, 10-fold CV to choose tuning parameter
- ϵ is chosen via grid search on test data separately for each method
- Weighted Nyström samples proportional to D
- Here $m = m_{nys} = m_{gp} = 400$

Classification

- 15000 training and 3000 test
- Too big to compute full decomposition (well...)
- Performed for $m \in \{50, 150, 450, 1350, 4050\}$
- Averaged over 10 randomizations of the approximation methods



Conclusions

Performance

- Nyström is faster than Gaussian Projection
- Neither is uniformly better
  - Manifold reconstruction is a toss up
  - Gaussian projection is a mess in the clustering application
  - But it wins handily at classification

Other concerns

- Tuning parameter selection is critical
- Construction of W is important: different similarity measures lead to drastically different solutions, often garbage
- Gaussian projection gives orthogonal features
- Need some targeted theory