Exploring Large Data Sets

Daniel L Boley
University of Minnesota

Find Reduced Order Representations for Large Unstructured Data Collections to facilitate finding patterns, connections, outliers, and to reduce noise.

NSF Grant 0916750

Exploring Large Data Sets

- Many large unstructured data sets must be analysed
  - Text documents (news, laws, WWW documents).
  - Gene expression profiles
  - Attributes for individual people, transactions, locations, ecosystems, ...

  } tabular

  - Gene-gene or protein-protein interaction networks
  - WWW connectivity graph
  - Computer inter-connect in Internet
  - People-people affinities in Social Media

  } graph

- Many example datasets can easily have up to $O(10^{9+})$ data points.
- Many datasets have much noise or many attributes.
- Many example datasets are sampled, subject to sampling bias.

Tools to Explore

- Dimensionality Reduction
  - Represent each data sample with a reduced set of attribute values
  - Minimize loss of information
  - Implicit assumption: data is subject to some level of noise.
- Graph Properties
  - partitioning
  - identify important nodes or links
  - aggregrate properties
- Sparse Representation
  - Hard to interpret individual components in traditional dimensionality reduction methods.
  - Seek to represent each data sample as a combination of only a few components.
  - Possibly also seek to represent each component as a combination of only a few original attributes.
  - Maintain desire for small approximation error.

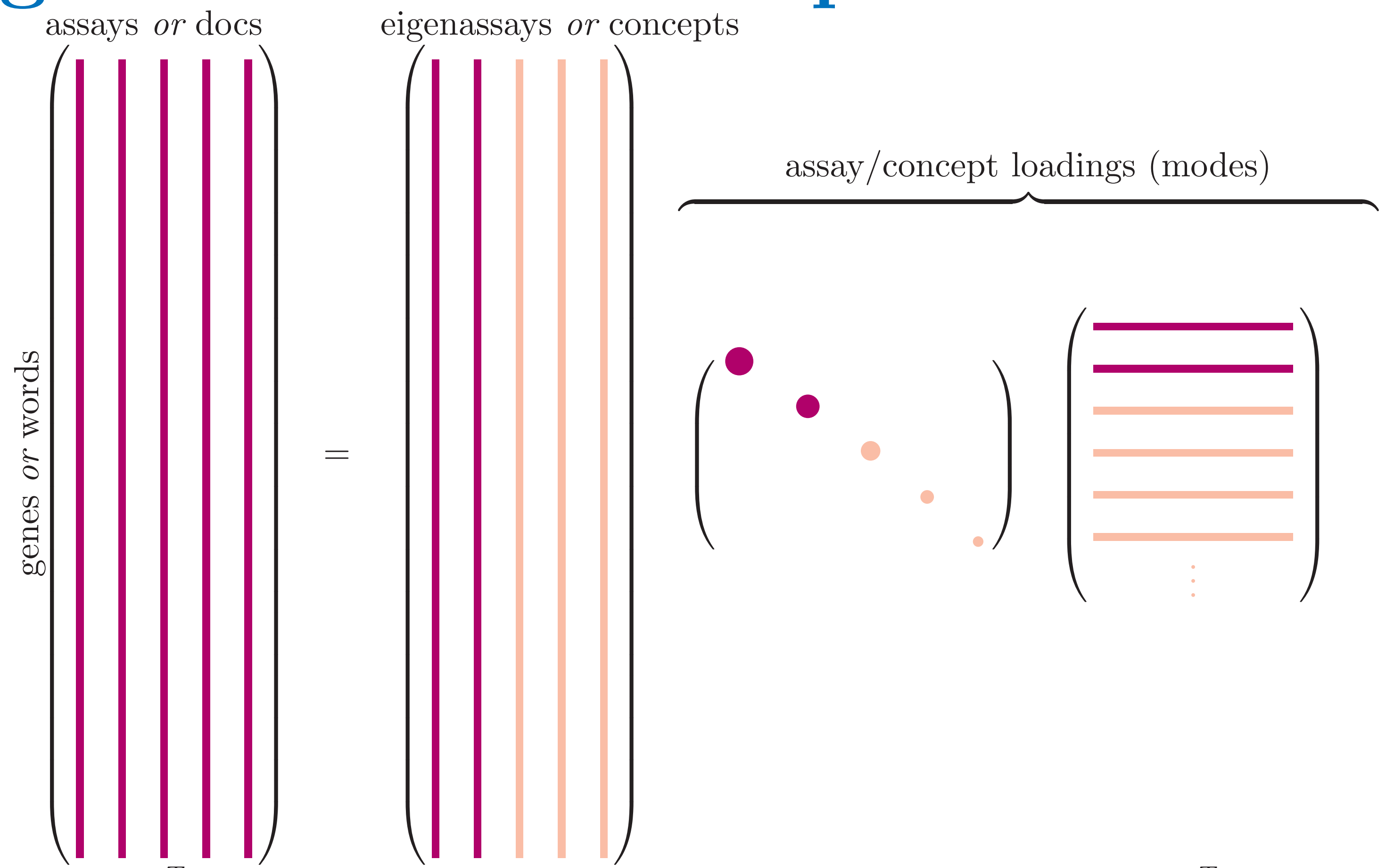
Outline

- Dimensionality Reduction
  - Principal Component Analysis – PCA
  - Latent Semantic Indexing
  - Clustering
- Graph Partitioning
  - Principal Direction Divisive Partitioning
  - Spectral Partitioning
- Sparse Representation – Examples
  - almost shortest path routing.
  - constrained clustering.
  - image/vision,
  - Graph Connection Discovery.
- Finding Sparse Representation

Outline

- Dimensionality Reduction
  - Principal Component Analysis – PCA
  - Latent Semantic Indexing
  - Clustering
- Graph Partitioning
  - Principal Direction Divisive Partitioning
  - Spectral Partitioning
- Sparse Representation – Examples
  - almost shortest path routing.
  - constrained clustering.
  - image/vision,
  - Graph Connection Discovery.
- Finding Sparse Representation

Singular Value Decomposition – SVD



$\mathbf{M} - \boldsymbol{\mu}\mathbf{1}^T$ *or* $\mathbf{M}$ = $\mathbf{U}$ $\mathbf{D}$ $\mathbf{V}^T$

Singular Value Decomposition – SVD

- Eliminate Noise
- Reduce Dimensionality
- Expose Major Components
- Suppose samples are columns of $m \times n$ matrix $\mathbf{M}$.
- Try to find k pseudo-data columns such that all samples can be represented by linear combinations of those k pseudo-data columns.
- Primary criterion: minimize the 2-norm of the discrepancy between the original data and what you can represent using k pseudo-data columns.
- Answer: Singular Value Decomposition.
- Sometimes, for statistical reasons, want to remove uniform signal:
  - $\mathbf{M} \leftarrow \mathbf{M} - \boldsymbol{\mu}\mathbf{1}^T$,
    where $\boldsymbol{\mu} = \mathbf{M} \cdot \mathbf{1}$.
  - Then $\mathbf{M}^T\mathbf{M}$ is the Sample Covariance Matrix.
  - Even without centering, $\mathbf{M}^T\mathbf{M}$ is a “Gram” matrix.

Principal Component Analysis – PCA

- Suppose samples are columns of $m \times n$ matrix $\mathbf{M}$.
  - Optionally center columns of matrix $\mathbf{M} \leftarrow \mathbf{M} - \boldsymbol{\mu}\mathbf{1}^T$.
  - Form sample covariance matrix or Gram matrix: $\mathbf{C} = \mathbf{M}^T\mathbf{M}$, where $\boldsymbol{\mu} = \frac{1}{n}\mathbf{M}\mathbf{1}$ = sample mean, $\mathbf{1}^T = [1, \ldots, 1]$.
  - Diagonalize $\mathbf{C} = \mathbf{V}\mathbf{D}^2\mathbf{V}^T$ to get principal components $\mathbf{V}$, where $\mathbf{D}^2 = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \cdots)$, $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$.

- Compute above via Singular Value Decomposition

  $\mathbf{M} = \mathbf{U}\mathbf{D}\mathbf{V}^T$

- Top k principal components $\Longrightarrow$ best rank k approximation:

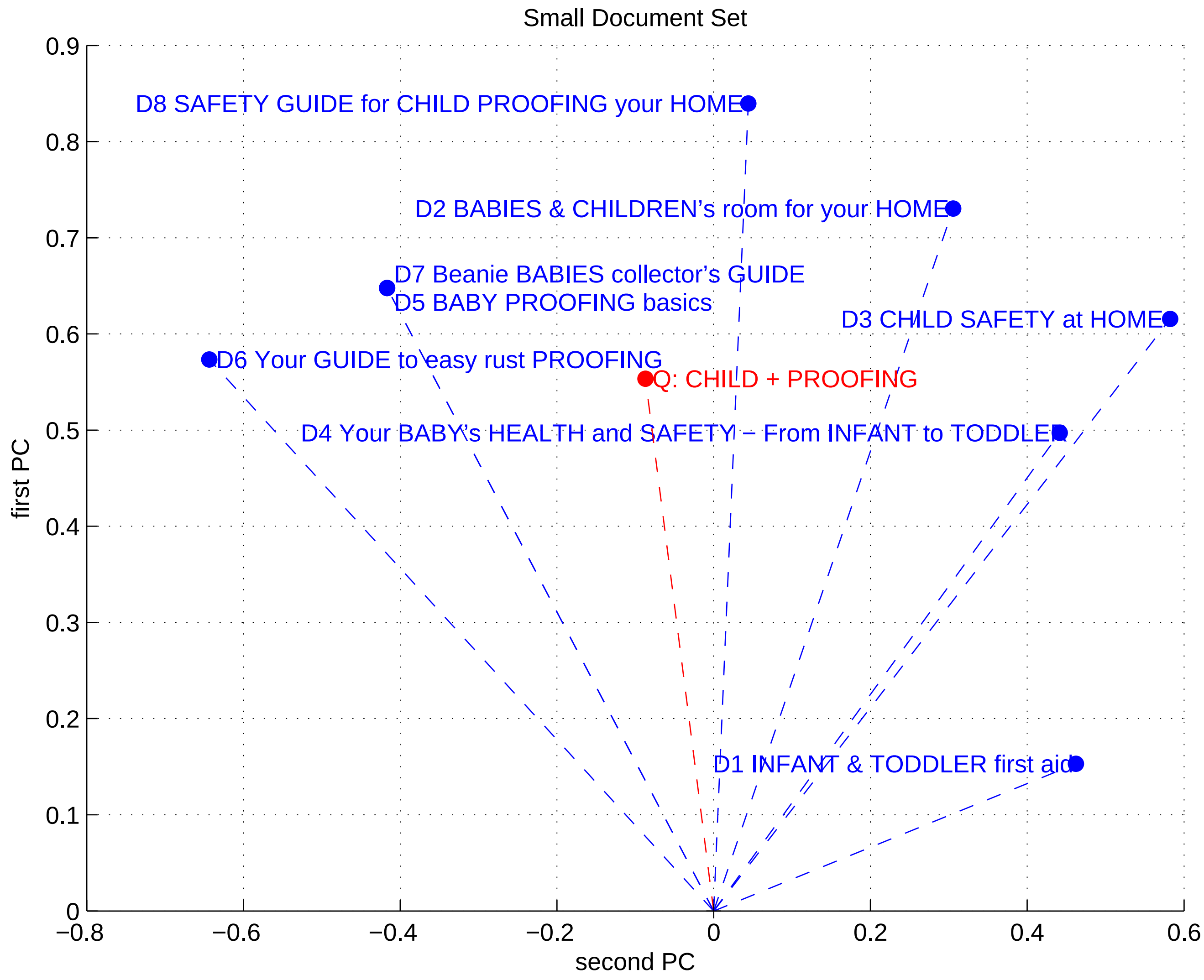
  $\mathbf{U}_{*,1\ldots k} \cdot \mathbf{D}_{1\ldots k,1\ldots k} \cdot \mathbf{V}^T_{*,1\ldots k}$

Text Documents – Data Representation

- Each document represented by n-vector $\mathbf{d}$ of word counts, scaled to unit length.
- Vectors assembled into Term Frequency Matrix $\mathbf{M} = (\, \mathbf{d}_1 \quad \cdots \quad \mathbf{d}_m \,)$.

	D1 INFANT & TODDLER first aid	D2 BABIES & CHILDREN's room for your HOME	D3 CHILD SAFETY at HOME	D4 Your BABY's HEALTH and SAFETY - From INFANT to TODDLER	D5 BABY PROOFING basics	D6 Your GUIDE to easy rust PROOFING	D7 Beanie BABIES collector's GUIDE	D8 SAFETY GUIDE for CHILD PROOFING your HOME
BABY	0	$\sqrt{3}$	0	$\sqrt{5}$	$\sqrt{2}$	0	$\sqrt{2}$	0
CHILD	0	$\sqrt{3}$	$\sqrt{2}$	0	0	0	0	$\sqrt{5}$
GUIDE	0	0	0	0	0	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{5}$
HEALTH	0	0	0	$\sqrt{5}$	0	0	0	0
HOME	0	$\sqrt{3}$	$\sqrt{2}$	0	0	0	0	$\sqrt{5}$
INFANT	$\sqrt{2}$	0	0	$\sqrt{5}$	0	0	0	0
PROOFING	0	0	0	0	$\sqrt{2}$	$\sqrt{2}$	0	$\sqrt{5}$
SAFETY	0	0	$\sqrt{2}$	$\sqrt{5}$	0	0	0	$\sqrt{5}$
TODDLER	$\sqrt{2}$	0	0	$\sqrt{5}$	0	0	0	0

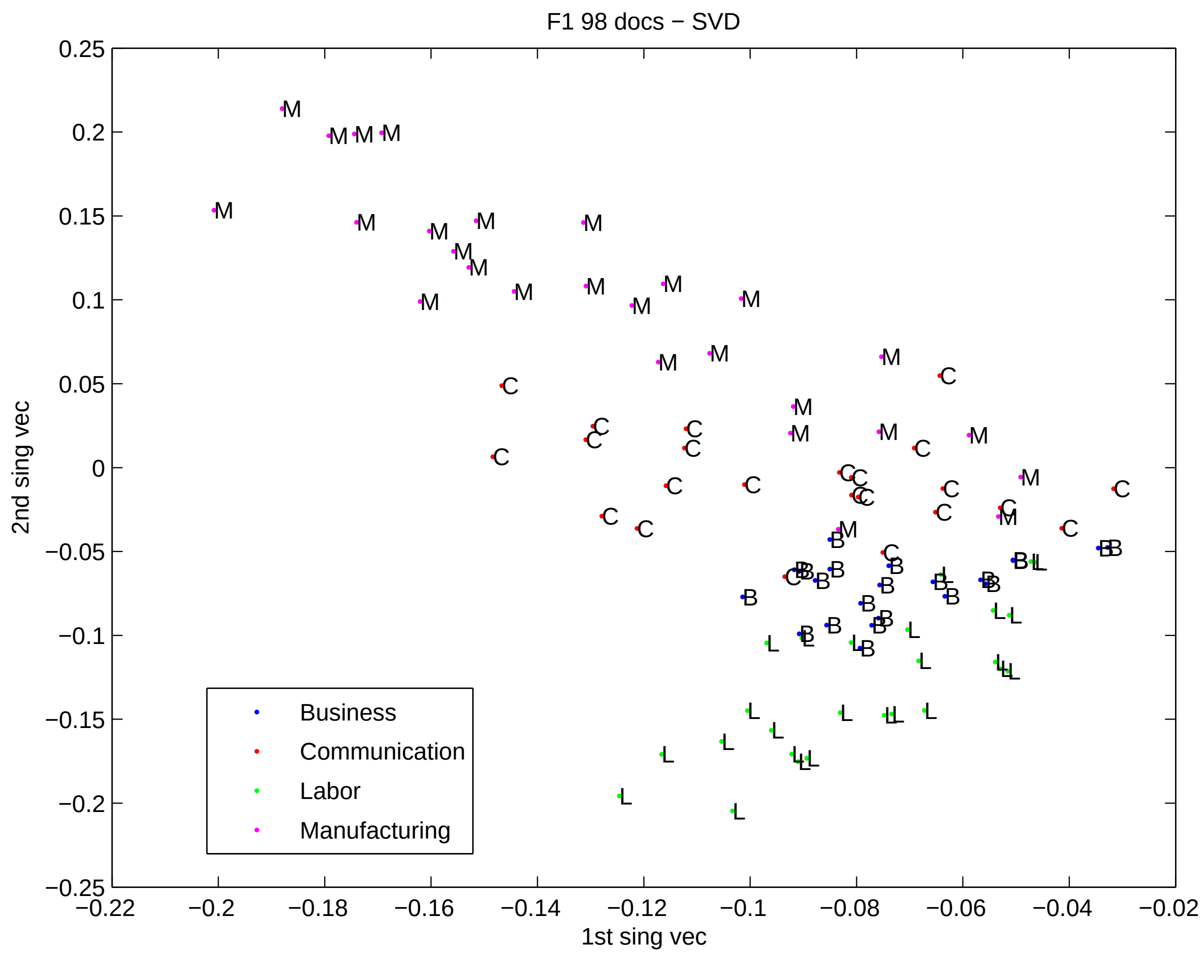
Latent Semantic Indexing – LSI



- Stay length-independent: compare using just angles.

Latent Semantic Indexing – LSI

- Loadings of top two concepts on set of 98 documents with 5623 words. (Berry et al., 1995; Boley, 1998)

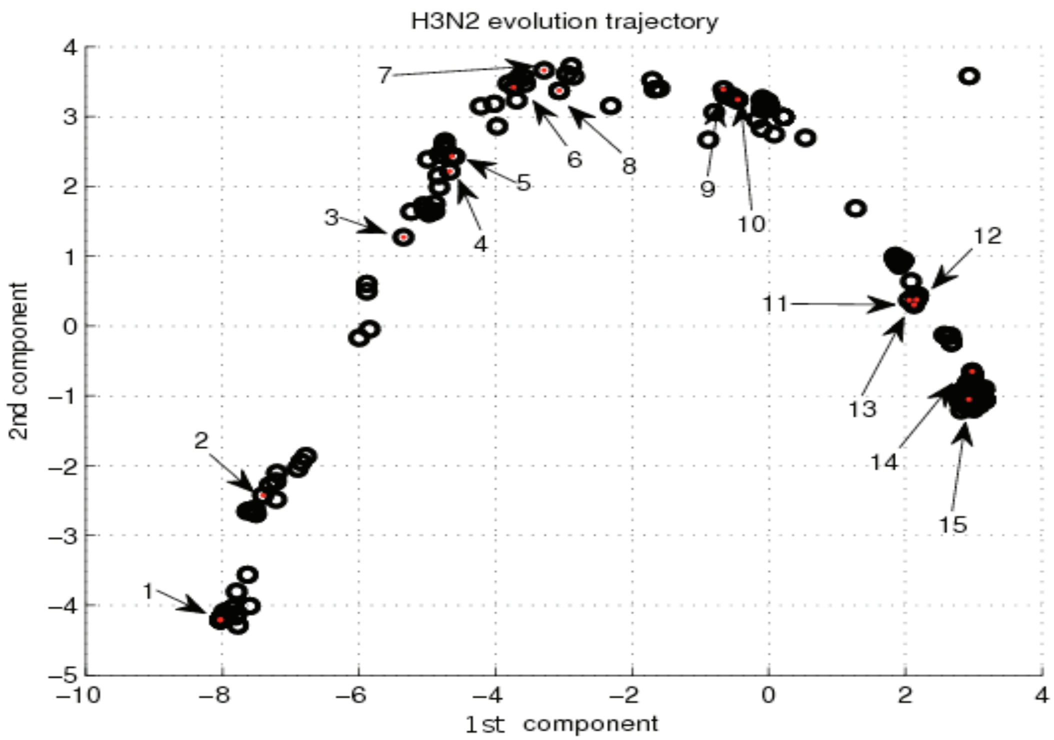


Five Concepts

PC 1	*PC 2*		*PC 3*	
plus end	*minus end*	*plus end*	*minus end*	*plus end*
-----------	----------	----------	----------	-----------
manufactur	manufactur	pipe	edi	behavior
system	employ	seam	employ	chronolog
develop	engin	convert	manufactur	wherev
process	servic	processor	busi	ink
inform	employe	transmitt	electron	incomplet
applic	mean	waste	standard	height
technologi	integr	chip	action	slightli
integr	action	clock	job	pump
standard	affirm	chicago	compani	label
engin	system	scheme	engin	clerk
program	job	highli	mean	french
employ	technologi	phd	affirm	embassi
edi	process	robin	capit	mainli
design	public	reprogramm	data	thirti
servic	law	serc	employe	interv

- Words in concepts are somewhat informative.
- But high degree of overlap.

Model Avian Influenza Virus



from Lam&Boley 2011

Number	Vaccine strain
1	A/Aichi/1968
2	A/Port Chalmers/1/1973
3	A/Philippines/2/1982
4	A/leningrad/360/1986
5	A/Shanghai/11/1987
6	A/Beijing/353/1989
7	A/Shangdong/9/1993
8	A/Johannesburg/33/199
9	A/Sydney/5/1997
10	A/Moscow/10/1999
11	A/Fujian/411/2002
12	A/California/7/2004
13	A/Wisconsin/67/2005
14	A/Brisbane/10/2007
15	A/Perth/16/2009

- Evolution is a flow, naturally falls in chronological order.
- Without vaccine, picture is more a random cloud of points.
- Suggests vaccine use does affect evolution of virus.

Model Avian Influenza Virus

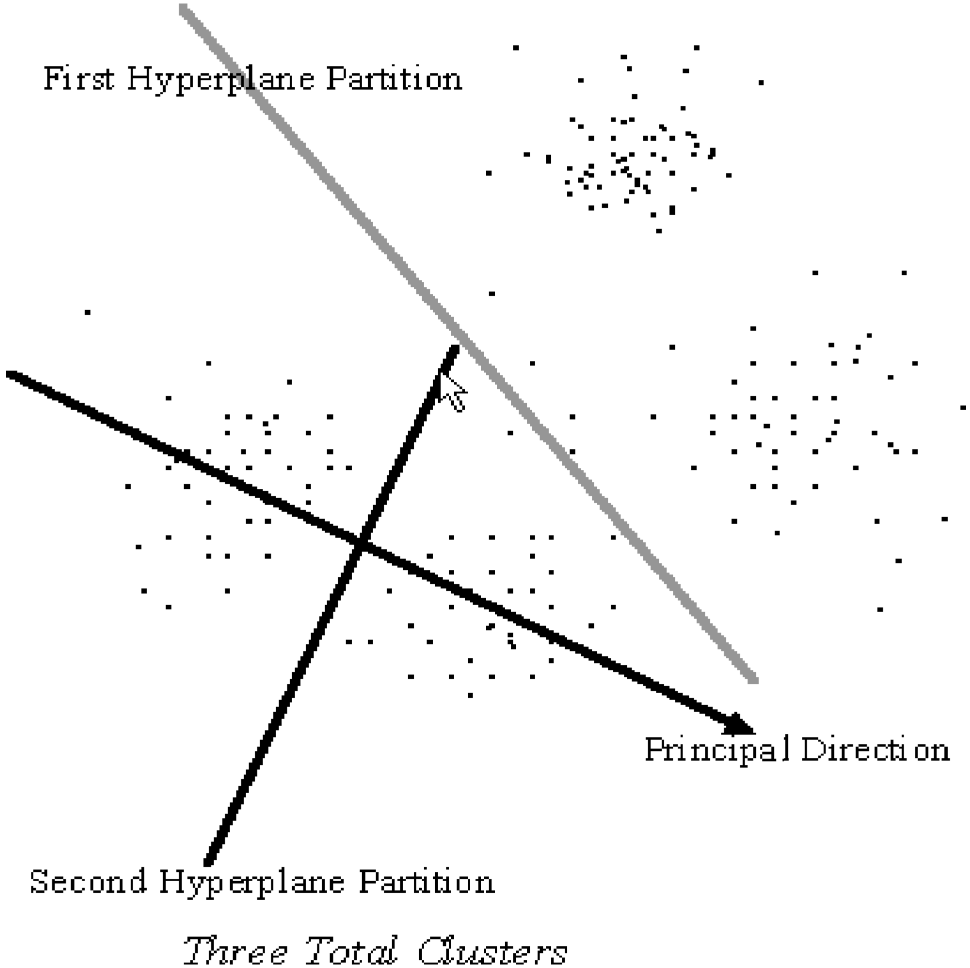
(Lam et al., 2012)

- Avian Flu Virus characterized by the HA protein, which the virus uses to penetrate the cell.
- The protein is described by a string of 566 symbols, each representing one of 20 Amino Acids.
- Embed in high dimensional Euclidean space by replacing each Amino Acid with a string of 20 bits:
  - E.g. 3rd amino acid $= \rightarrow$ 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
- Result is a vector of length $566 \cdot 20 = 11230$.
- Use PCA to reduce dimensions from 11320 to 6.
- Use first 2 components to track evolution of this protein in a simple visual way.

Outline

- Dimensionality Reduction
  - Principal Component Analysis – PCA
  - Latent Semantic Indexing
  - Clustering
- Graph Partitioning
  - Principal Direction Divisive Partitioning
  - Spectral Partitioning
- Sparse Representation – Examples
  - almost shortest path routing.
  - constrained clustering.
  - image/vision,
  - Graph Connection Discovery.
- Finding Sparse Representation

Principal Direction Divisive Partitioning



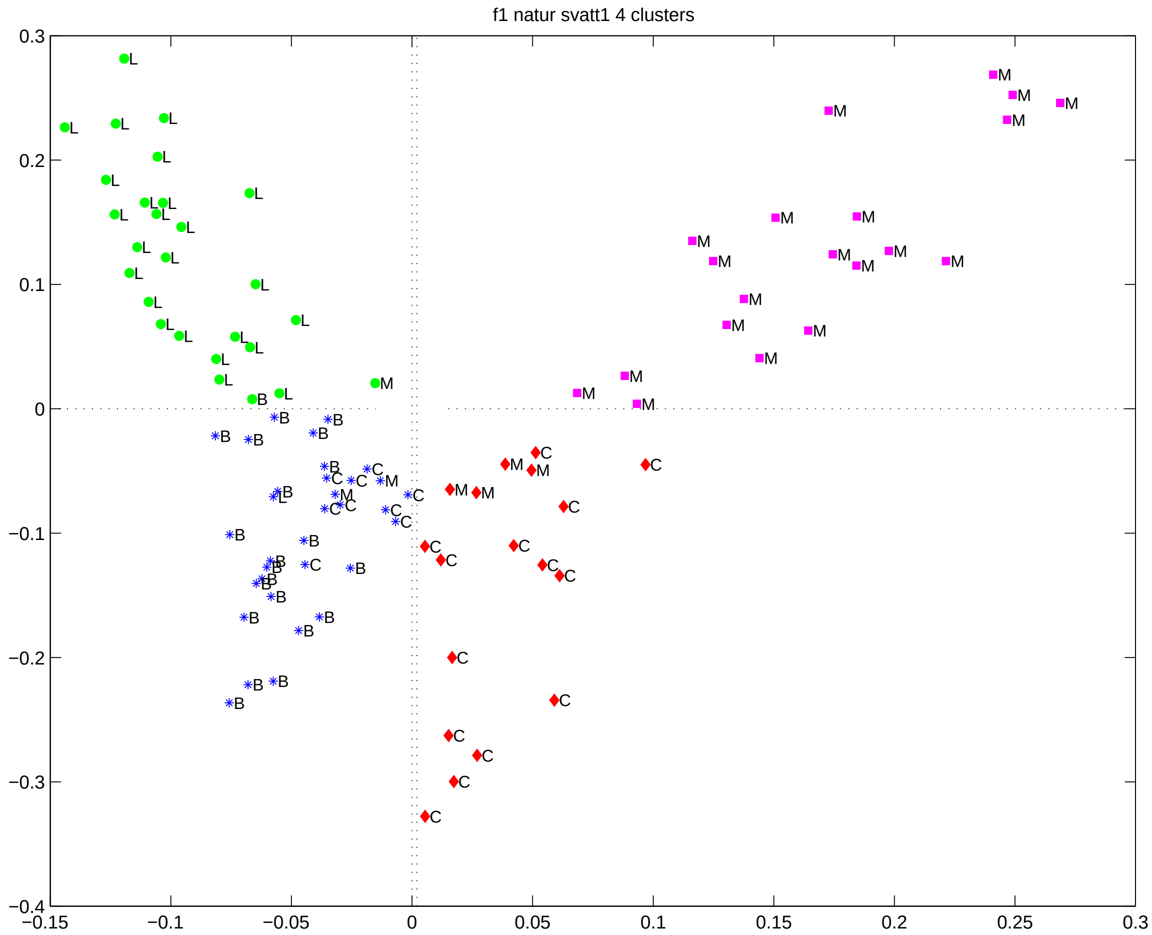
(Boley, 1998)

Divisive Partitioning for Unsupervised Clustering

- Unsupervised, as opposed to Supervised:
  - No predefined categories;
  - No previously classified training data;
  - No a-priori assumptions on the number of clusters.
- Top-down Hierarchical:
  - Imposes a tree hierarchy on unstructured data;
  - Tree is source for some taxomonic information for dataset;
  - Tree is generated from the root down.
  - Result is Principal Direction Divisive Partitioning. (Boley, 1998)
- Multiway Clustering.
  - Project onto first k principal directions. Result: each data sample is represented by k components.
  - Apply classical k-means clustering to projected data.
  - Used for both Graph Partitioning and Data Clustering. (Dhillon, 2001)
- Empirically Best Approach: a hybrid method:
  - Use Divisive Partitioning first (deterministic).
  - Refine with K-means (avoids initialization issues). (Savaresi & Boley, 2004)

PDDP on 98 Document Set

- Loadings of top two concepts on set of 98 documents with 5623 words.



Top distinctive words in top 3 clusters

PC 1 minus end	PC 1 plus end	PC 2 minus end	PC 2 plus end	PC 3 minus end	PC 3 plus end
employ	manufactur	busi	employ	edi	manufactur
action	engin	capit	mean	electron	engin
employe	system	fund	job	standard	design
affirm	integr	credit	servic	busi	project
servic	process	invest	employe	map	tool
mean	technologi	corpor	act	commerc	process
law	develop	investor	action	data	integr
job	project	debt	feder	messag	technologi
right	tool	source	train	paperfre	research
public	design	compani	osha	network	plan
feder	industri	offer	individu	secur	product
act	product	stock	public	compani	sme
copyright	research	click	affirm	interchang	machin
osha	machin	tax	labor	translat	educ
person	data	lease	applic	exchang	univers
labor	*manufacturing*	*business*	*labor*	*communication*	*manufacturing*

Spectral Graph Partitioning

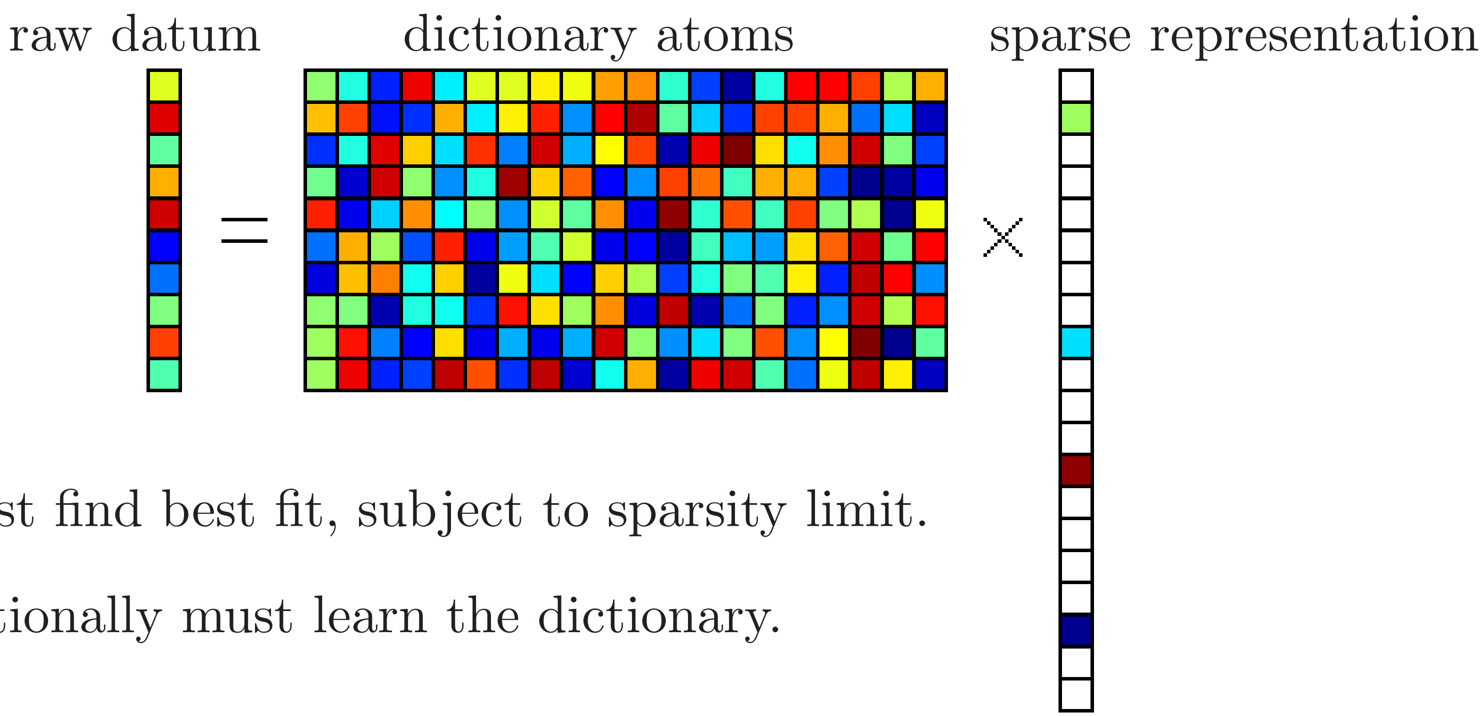
- Model an undirected graph by a random walk.
- Measure distance between two nodes by average round-trip commute time (average number of steps to go from node i to j and back again.)
- Vertices of an undirected connected graph can be embedded in high-dimensional Euclidean space.
- Embedding preserves distances between the vertices.
- Principal Direction splitting on embedding is equivalent to two-way Spectral Graph Partitioning.
- Much more popular in graph setting.
- Can be extended to directed graphs (e.g., commute times still a metric). (Boley et al., 2011)

Outline

- Dimensionality Reduction
  - Principal Component Analysis – PCA
  - Latent Semantic Indexing
  - Clustering
- Graph Partitioning
  - Principal Direction Divisive Partitioning
  - Spectral Partitioning
- Sparse Representation – Examples
  - almost shortest path routing.
  - constrained clustering.
  - image/vision,
  - Graph Connection Discovery.
- Finding Sparse Representation

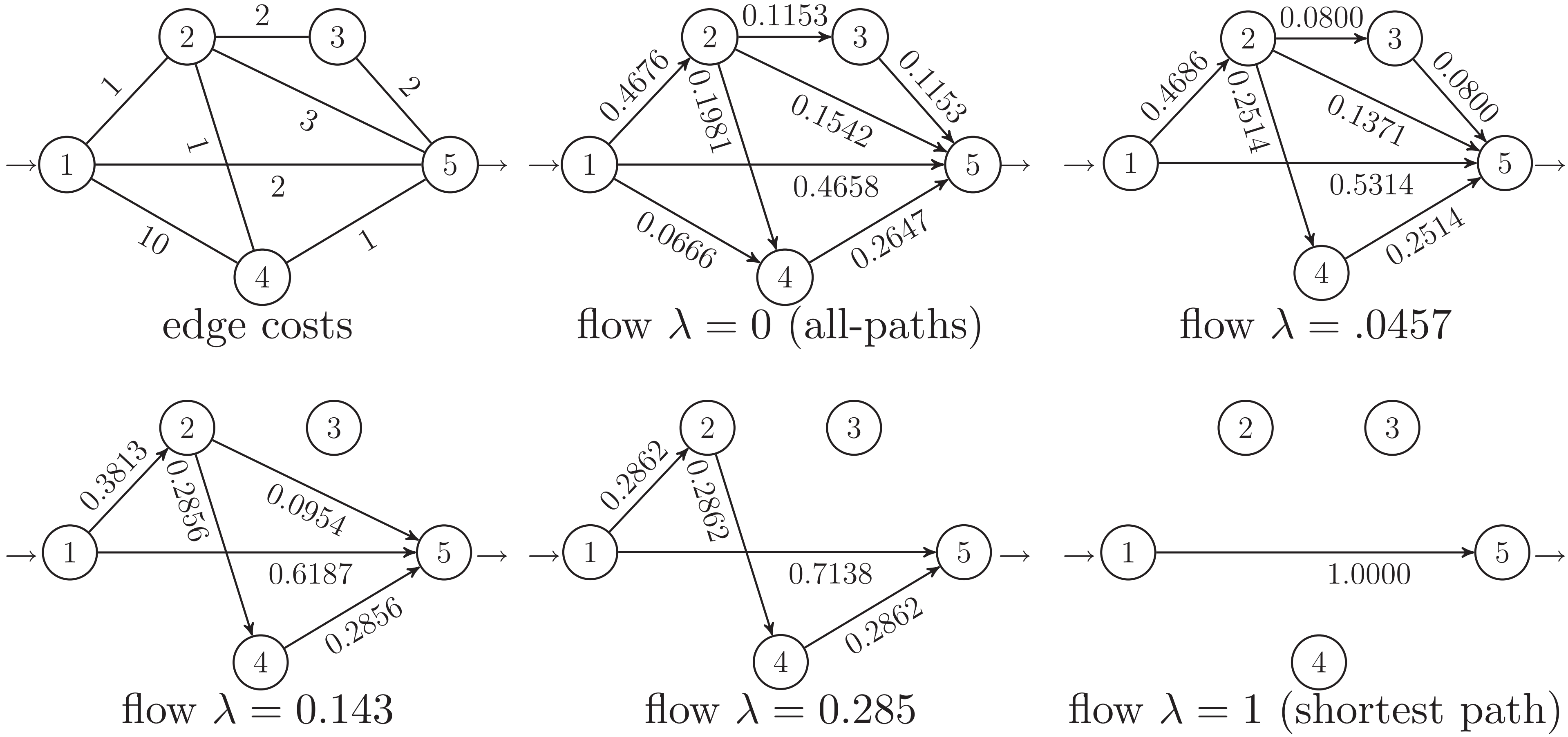
Sparse Representation

- Many machine learning algorithms can explore massive data:
  K-nearest Neighbors, Kernal-SVM, Boosting, Metric Learning, . . .
- All can benefit from denoising by finding a sparse representation:



- Must find best fit, subject to sparsity limit.
- Optionally must learn the dictionary.

Almost Shortest Path Routing



$\min_{\mathbf{x}} \mathbf{x}^T \mathbf{W} \mathbf{x} + \lambda \|\mathbf{x}\|_1 = \sum_{ij \in E} X_{ij}^2 w_{ij} + \lambda |X_{ij}|$ minimize total flow energy

s.t. $\sum_{i:\, ik \in E} X_{ik} = \sum_{j:\, kj \in E} X_{kj} \quad \forall k$ flow in = flow out at every node k

(Li et al., 2011)

Constrained Clustering

- Graph Clustering with *Must-link* and *Cannot-link* constraints.
- Spectral Graph Cut: $= \mathbf{x}^T\mathbf{L}\mathbf{x}$ [where $\mathbf{L}$ = Laplacian].
- Previous approach: minimize $\mathbf{x}^T\mathbf{L}\mathbf{x} + \lambda\mathbf{x}^T\mathbf{L}_{\mathsf{c}}\mathbf{x}$ (Shi et al., 2010).
- Our approach: minimize cut with $L1$ penalty on constraint violations: $\mathbf{x}^T\mathbf{L}\mathbf{x} + \lambda\|\mathbf{C}_{\mathsf{c}}\mathbf{x}\|_1$ [Kawale et al].

Image Descriptors

Image Descriptor

- Pixel Descriptors: for i-th pixel $z_i = \phi(x_i, y_i)$ is a vector of descriptors for the pixel at point (x_i, y_i) in the image.
- Example, could use $z_i = (I_x, I_y, |\mathsf{grad} I|, \angle \mathsf{grad} I, I_{xx}, I_{xy}, I_{yy})$ where I is the intensity value. Could also incorporate color information.

Covariance Descriptor (Tuzel et al., 2006)

- Within each small patch around each pixel compute the covariance C_i of the pixel descriptors.
- Covariance descriptors eliminate differences due to scaling, brightness, large shadows, but enhance local features.
- Use for object detection, tracking, recognition, and more ...
- Each C_i is a small positive semi-definite matrix (7×7 in this example).
- Regularize each C_i by adding a small multiple of the identity.

Covariance Descriptor Example

Raw Image

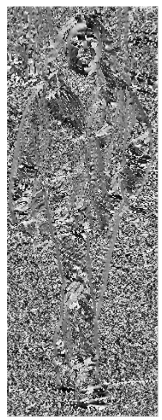
Image

first derivatives

x-grad

grad-mag

y-grad



grad-dir

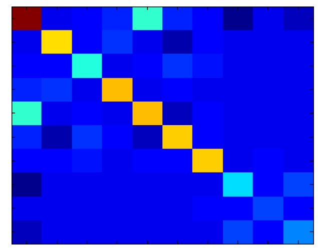
second derivatives

Dxx

Dxy

Dyy

pixel by pixel descriptor



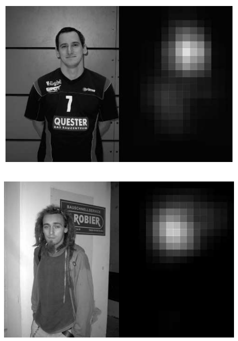
Covariance descriptor

Covariance Descriptor Usage

- **Object Detection and Tracking in Image.**

Object Detection

face



(Opelt et al., 2004; Sivalingam et al., 2011)

license plate

Image from Porikli & Kocak '06

(Porikli & Kocak, 2006)

human

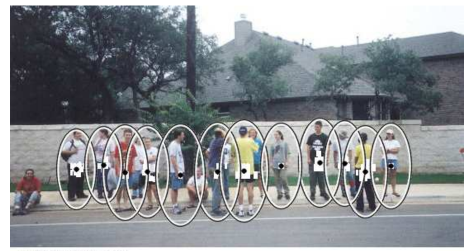


Image from Tuzel et al. '07

(Tuzel et al., 2007)

Object Tracking

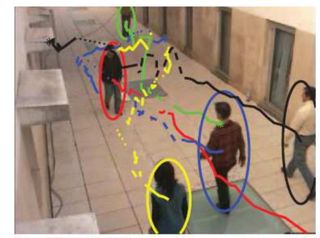


Image from Palaio & Batista '09

(Palaio et al., 2009)

Object Recognition

face

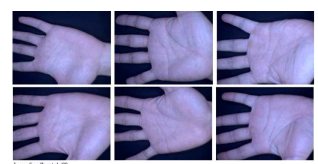
(Pang et al., 2008)

action

Images from the KTH Dataset

KTH dataset

palmprint



(Han et al., 2009)

Optimization Setup for Covariances

Notation: (Sivalingam et al., 2010; Sivalingam et al., 2011)

- S = a raw covariance matrix,
  $\mathbf{x}$ = vector of unknown coefficients.
  $\mathcal{A} = (A_1, A_2, \ldots, A_k)$ = collection of dictionary atoms.
  $\mathbf{x} = (x_1, x_2, \ldots, x_k)$ = vector of unknown coefficients.

- Goal: Approximate $S \approx A_1 x_1 + \cdots + A_k x_k = \mathcal{A} \cdot \mathbf{x}$.

- Use "logdet" divergence as measure of discrepancy:
  $D_{\mathrm{ld}}(\mathcal{A} \cdot \mathbf{x}, S) = tr((\mathcal{A} \cdot \mathbf{x})S^{-1}) - \log\det((\mathcal{A} \cdot \mathbf{x})S^{-1}) - n.$

- Logdet divergence measures relative entropy between two different zero-mean multivariate Gaussians.

Optimization Problem for Covariances

(Sivalingam et al., 2010; Sivalingam et al., 2011)

- Leads to optimization problem

$$\min_{\mathbf{x}} \quad \underbrace{\sum_i x_i tr(A_i) - \log\det\left[\sum_i x_i A_i\right]}_{\text{Dist}(\mathcal{A}\cdot\mathbf{x},S)} + \lambda \underbrace{\sum_i x_i}_{\text{sparsity}}$$

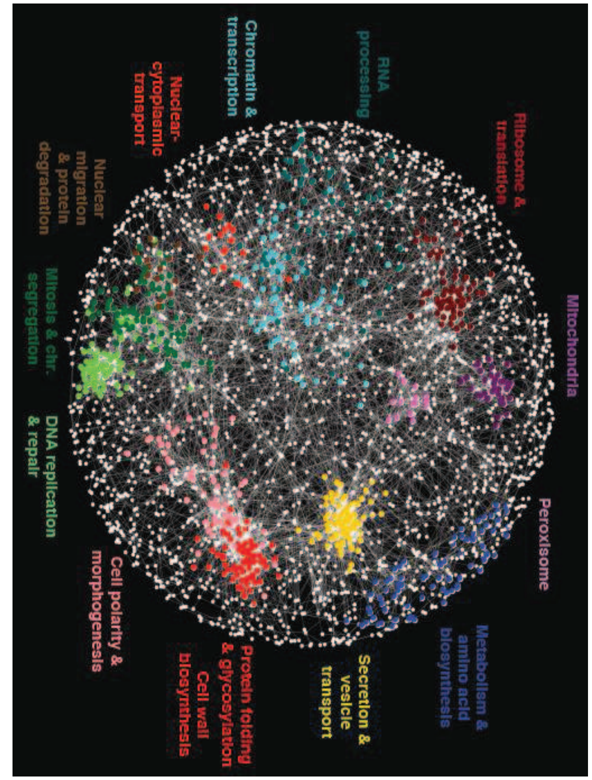
$$\text{s.t.} \quad \mathbf{x} \geq 0$$

$$\textstyle\sum_i x_i A_i \succeq 0 \quad \text{(positive semi-definite)}$$

$$\textstyle\sum_i x_i A_i \preceq S \quad \text{(residual positive semi-def.)}$$

- This is in a standard form for a MaxDet problem.
- The sparsity term is a relaxation of true desired penalty: # nonzeros in $\mathbf{x}$.
- Convex problem solvable by e.g. the `CVX` package (Grant & Boyd, 2010).

Graph Connections Discovery

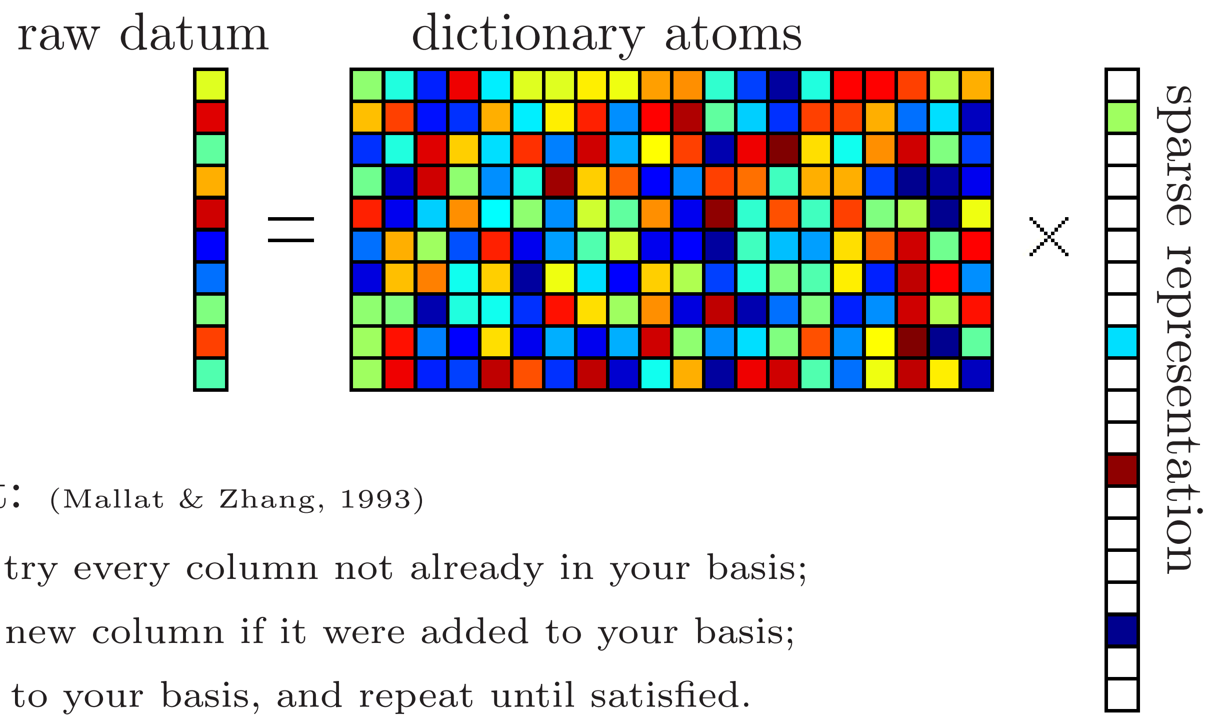


- Signal at node i is gaussian & correlated to neighbors, but conditionally independent of signal at unconnected node j.
- Statistical Theory $\implies (\text{Covariance})^{-1}_{ij} = 0$. $(\text{Covariance})^{-1}$ is called the Precision Matrix.
- If graph is sparse, expect $(\text{Covariance})^{-1}$ to be sparse.
- Problem: Graph connections are unknown.
- Task: Given signals at each node, recover graph edges.
- Applications: biology, climate modelling, social networks.
- Method:
  - Compute sample precision matrix from signals.
  - Find best ***sparse*** approximation to sample precision matrix.
  - Use previous log-det divergence to measure discrepancy between covariance matrices.

Outline

- Dimensionality Reduction
  - Principal Component Analysis – PCA
  - Latent Semantic Indexing
  - Clustering
- Graph Partitioning
  - Principal Direction Divisive Partitioning
  - Spectral Partitioning
- Sparse Representation – Examples
  - almost shortest path routing.
  - constrained clustering.
  - image/vision,
  - Graph Connection Discovery.
- Finding Sparse Representation

Constructing Sparse Basis



- Matching Pursuit: (Mallat & Zhang, 1993)
  - Greedy algorithm: try every column not already in your basis;
  - evaluate quality of new column if it were added to your basis;
  - add “best” column to your basis, and repeat until satisfied.
- Basis Pursuit (Chen et al., 2001)
  - Minimize $\|\mathbf{b} - A\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_0$.
  - Difficulty: this is a NP-hard combinatorial problem.
  - Relax to $\|\mathbf{b} - A\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1$.
  - Relaxed problem is convex, so solvable more efficiently.
  - LASSO: Solve for all λ fast (Tibshirani, 1996).

Convex Relaxation $\Longrightarrow$ LASSO

- Known as Basis Pursuit, Compressed Sensing, "small error + sparse".
- Add penalty for number of nonzeros with weight λ:

$$\min_{\mathbf{x}} ||A\mathbf{x} - \mathbf{b}||_2^2 + \lambda||\mathbf{x}||_0.$$

- Convert hard combinatorial problem into easier convex optimization problem.
- Relax previous $||\mathbf{x}||_0$ to convex problem:

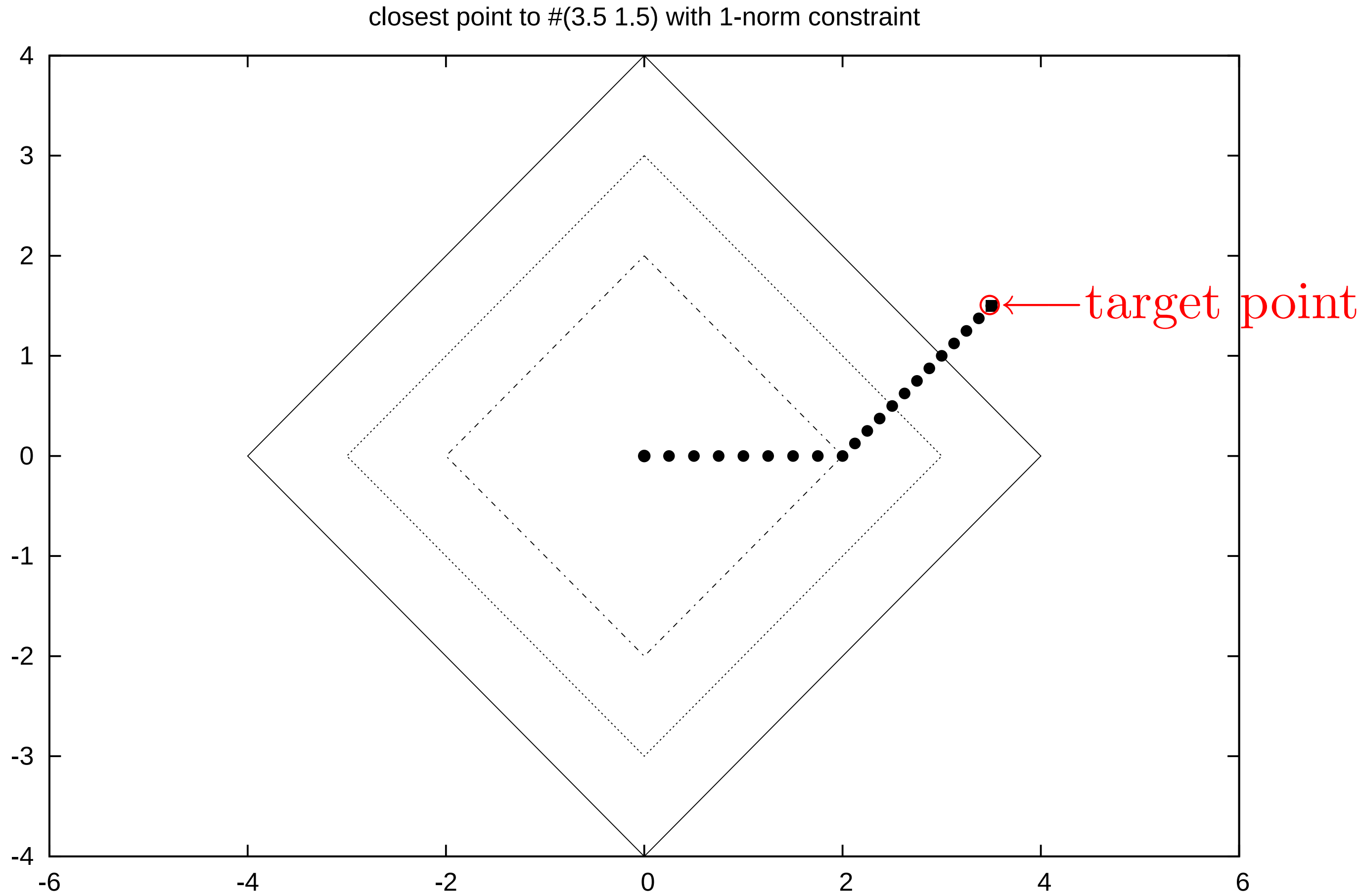
$$\min_{\mathbf{x}} ||A\mathbf{x} - \mathbf{b}||_2^2 + \lambda||\mathbf{x}||_1,$$

- or convert to constrained problem:

$$\min_{\mathbf{x}} ||A\mathbf{x} - \mathbf{b}||_2^2 \quad \text{subject to} \quad ||\mathbf{x}||_1 \le \texttt{tol}.$$

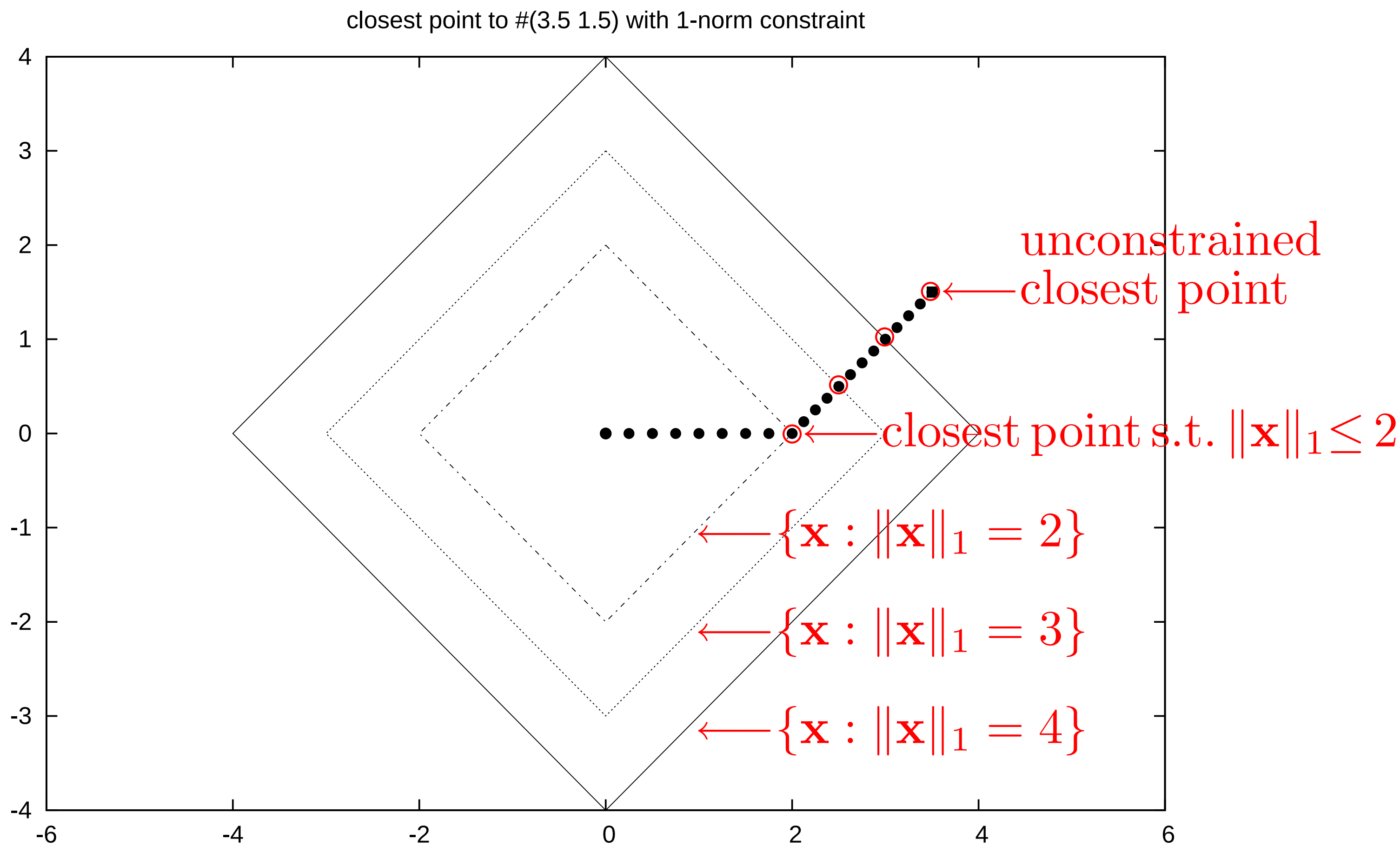
- Vary parameter λ or `tol`, to explore the trade-off between "small error" and "sparse".

Motivation: find closest sparse point



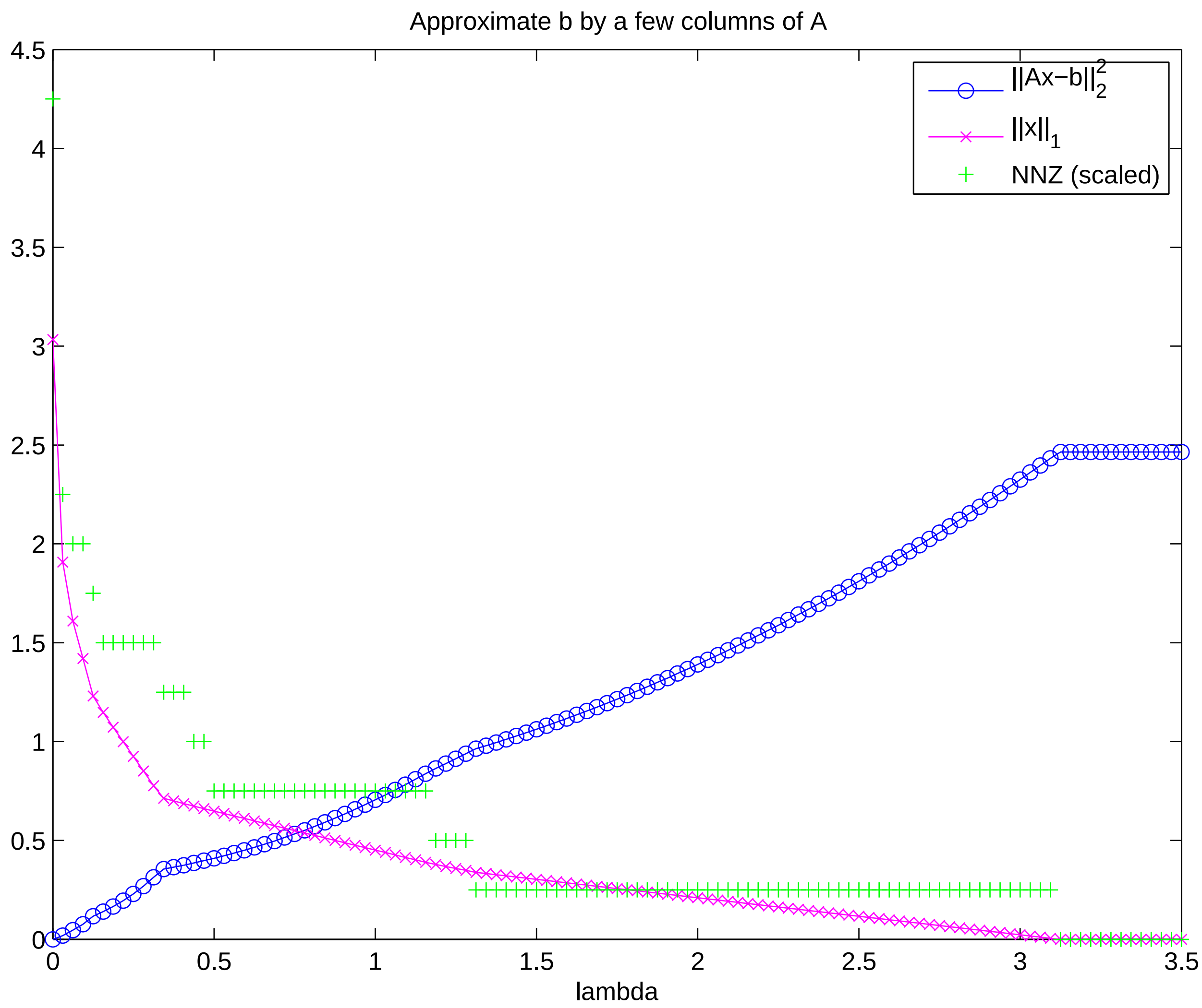
- Find closest point to target . . . subject to ℓ_1 norm constraint.

Motivation: find closest sparse point



- As limit on $\|\mathbf{x}\|_1$ is tightened, the coordinates are driven toward zero.
- As soon as one coordinate reaches zero, it is removed, and the remaining coordinates are driven to zero.

Example: 17 signals with 10 time points



- As λ grows, the error grows, fill (#non-zeros) shrinks.

Methods

- All problems are convex.
- Must work exists on software for convex programming problems
- YALMIP is a front end with links to many solver packages (Löfberg, 2004).
- CVX is a free package of convex solvers with easy matlab interface (Grant & Boyd, 2010).
- ADMM is a paradigm for a simple iterative solver especially adapted for very large but separable problems (Boyd et al., 2011).

Conclusions

- Many different types of data, many highly unstructured.
- Extracting patterns or connections in data involves somehow reducing the volume of data one must look at.
- Data Reduction is an old paradigm that has been updated for the modern digital age.
- Methods discussed here started with classical PCA - SVD based approaches (e.g., assuming independent gaussian noise).
- Connections and pair-wise correlations modeled by graphs.
- Graphs modeled by random walks, counting subgraphs, min-cut/max-flow, models, ...
- Sparse representations: wide variety of sparse approximations: low fill, short basis, non-negative basis, non-squared loss function, count violations of some constraints, low rank (nuclear norm = $L1$-norm on the singular values), ...
- Leads to need for scalable solvers for very large convex programs.

THANK YOU!

References

Bamieh, B., Jovanovic, M., Mitra, P., & Patterson, S. (2008). Effect of topological dimension on rigidity of vehicle formations: Fundamental limitations of local feedback. *Proc. CDC* (pp. 369–374). Cancun, Mexico.

Berry, M. W., Dumais, S. T., & O.'Brien, G. W. (1995). Using linear algebra for intelligent information retieval. *SIAM Rev.*, *37*, 573–595.

Boley, D., Ranjan, G., & Zhang, Z.-L. (2011). Commute times for a directed graph using an asymmetric Laplacian. *Linear Algebra and Appl.*, *435*, 224–242.

Boley, D. L. (1998). Principal direction divisive partitioning. *Data Mining and Knowledge Discovery*, *2*, 325–344.

Boyd, S., Parikh, N., Chu, E., Peleato, B., & Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, *3*, 1–122. `http://www.stanford.edu/~boyd/papers/admm/`.

Brualdi, R. A., & Ryser, H. J. (1991). *Combinatorial matrix theory*. Cambridge Univ. Press.

Chebotarev, P., & Shamis, E. (2006). Matrix-forest theorems.

Chen, S. S., Donoho, D. L., & Saunders, M. A. (2001). Atomic decomposition by basis pursuit. *SIAM Rev.*, *43*, 129–159.

Davis, G., Mallat, S., & Avellaneda, M. (1997). Adaptive greedy approximations. *Constructive Approximation*, *13*, 57–98. 10.1007/BF02678430.

Dhillon, I. S. (2001). Co-clustering documents and words using bipartite spectral graph partitioning. *KDD* (pp. 269–274).

Donoho, D., & Stodden, V. (2004). When does non-negative matrix factorization give a correct decomposition into parts? In S. Thrun, L. Saul and B. Schölkopf (Eds.), *Advances in neural information processing systems 16*. Cambridge, MA: MIT Press.

Elad, M., Figueiredo, M., & Ma, Y. (2010). On the role of sparse and redundant representations in image processing. *Proceedings of the IEEE*, *98*, 972 –982.

Grant, M., & Boyd, S. (2010). CVX: Matlab software for disciplined convex programming, version 1.21. `http://cvxr.com/cvx`.

Han, Y., Sun, Z., Tan, T., & Hao, Y. (2009). Palmprint recognition based on regional rank correlation of directional features. In M. Tistarelli and M. Nixon (Eds.), *Advances in biometrics*, vol. 5558 of *Lecture Notes in Computer Science*, 587–596. Springer Berlin / Heidelberg.

Lam, H. C., Sreevatsan, S., & Boley, D. (2012). Analyze influenza virus sequences using binary encoding approach. *Scientific Programming*.

Lee, D. D., & Seung, H. S. (2000). Algorithms for non-negative matrix factorization. In T. Leen, T. Dietterich and V. Tresp (Eds.), *Advances in neural information processing systems 16*, vol. 13, 556–562. Cambridge, MA: MIT Press.

Li, Y., Zhang, Z.-L., & Boley, D. (2011). The routing continuum from shortest-path to all-path: A unifying theory. *The 31st Int'l Conference on Distributed Computing Systems (ICDCS 2011)*. IEEE. to appear.

Liu, J., Ji, S., & Ye, J. (2009). Slep: Sparse learning with efficient projections. http://www.public.asu.edu/~jye02/Software/SLEP. Arizona State University.

Löfberg, J. (2004). YALMIP : A toolbox for modeling and optimization in MATLAB. *Proc. CACSD Conf.*. Taipei, Taiwan. http://users.isy.liu.se/johanl/yalmip .

Mallat, S., & Zhang, Z. (1993). Matching pursuits with time-frequency dictionaries. *Signal Processing, IEEE Transactions on*, *41*, 3397 –3415.

Olfati-Saber, R., Murray, R. M., A, & B (2004). Consensus problems in networks of agents with switching topology and time-delays. *IEEE Trans. Auto. Contr.*, *49*, 1520–1533.

Opelt, A., Fussenegger, M., Pinz, A., & Auer, P. (2004). Weak hypotheses and boosting for generic object detection and recognition. In T. Pajdla and J. Matas (Eds.), *Computer vision - ECCV 2004*, vol. 3022 of *Lecture Notes in Computer Science*, 71–84. Springer Berlin / Heidelberg.

Palaio, H., Maduro, C., Batista, K., & Batista, J. (2009). Ground plane velocity estimation embedding rectification on a particle filter multi-target tracking. *Robotics and Automation, 2009. ICRA '09. IEEE International Conference on* (pp. 825 –830).

Pang, Y., Yuan, Y., & Li, X. (2008). Effective feature extraction in high-dimensional space. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, *38*, 1652 –1656.

Porikli, F., & Kocak, T. (2006). Robust license plate detection using covariance descriptor in a neural network framework. *Video and Signal Based Surveillance, 2006. AVSS '06. IEEE International Conference on* (pp. 107 –107).

Savaresi, S., & Boley, D. (2001). On the performance of bisecting K-means and PDDP. *First SIAM International Conference on Data Mining (SDM'2001)*. Chicago.

Savaresi, S. M., & Boley, D. (2004). A comparative analysis on the bisecting K-means and the PDDP clustering algorithms. *Intelligent Data Analysis*, *8*, 345–362.

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *22*, 888–905.

Shi, X., Fan, W., & Yu, P. S. (2010). Efficient semi-supervised spectral co-clustering with constraints. *ICDM* (pp. 1043–1048).

Sivalingam, R., Boley, D., Morellas, V., & Papanikolopoulos, N. (2010). Tensor sparse coding for region covariances. *European Conf. on Comp. Vision (ECCV 2010)* (pp. 722–735). Springer.

Sivalingam, R., Boley, D., Morellas, V., & Papanikolopoulos, N. (2011). Positive definite dictionary learning for region covariances. *Int'l Conf. on Comp. Vision (ICCV 2011)* (pp. 1013–1019).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, *58*, 267–288.

Tuzel, O., Porikli, F., & Meer, P. (2006). Region covariance: A fast descriptor for detection and classification. In A. Leonardis, H. Bischof and A. Pinz (Eds.), *Computer vision ECCV 2006*, vol. 3952 of *Lecture Notes in Computer Science*, 589–600. Springer Berlin / Heidelberg.

Tuzel, O., Porikli, F., & Meer, P. (2007). Human detection via classification on riemannian manifolds. *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on* (pp. 1 –8).

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, *17*, 395–416.

Young, G. F., Scandovi, L., & Leonard, N. (2010). Robustness of noisy consensus dynamics with directed communication. *Proc. ACC* (pp. 6312–6317).