

Complexity of Inference in Latent Dirichlet Allocation

David Sontag, Daniel Roy
(NYU, Cambridge)

W66

Topic models are powerful tools for exploring large data sets and for making inferences about the content of documents

Documents



Topics

<u>politics</u> .0100
president .0095
obama .0090
washington .0085
religion .0060
...

<u>religion</u> .0500
hindu .0092
judiasm .0080
ethics .0075
buddhism .0016
...

<u>sports</u> .0105
baseball .0100
soccer .0055
basketball .0050
football .0045
...

$$\beta_t = \{ p(w | z = t) \}$$

Almost all uses of topic models (e.g., for unsupervised learning, information retrieval, classification) require **probabilistic inference**:

New document



Words w_1, \dots, w_N



What is this document about?

weather .50
finance .49
sports .01

Distribution of topics θ

Complexity of Inference in Latent Dirichlet Allocation

David Sontag, Daniel Roy
(NYU, Cambridge)

W66

We study the complexity of probabilistic inference in Latent Dirichlet Allocation

Input: new document with words $w_{1:N}$

topic-word distributions $\beta_t, t = 1, 2, \dots, T$ and Dirichlet hyper-parameters $\alpha_{1:T}$

Generative model

- ① $\theta \sim \text{Dirichlet}(\alpha_{1:T})$ Choose a distribution over the T topics
- ② For each word i ,
 $z_i \mid \theta \sim \theta$ Choose a topic for i 'th word
 $w_i \mid z_i \sim \beta_{z_i}$ Sample a word

Popular inference problems

1. Maximize $p(z_{1:N} \mid w_{1:N})$. \leftarrow Discrete. Classification
2. Maximize $p(\theta \mid w_{1:N})$. \leftarrow Dimensionality reduction, IR
3. Sample from $p(\theta \mid w_{1:N})$. \leftarrow Useful for learning

Complexity of Inference in Latent Dirichlet Allocation

David Sontag, Daniel Roy
(NYU, Cambridge)

W66

Main Results

Maximize $p(z_{1:N} | w_{1:N})$

For any α

	# topics in MAP assignment	Complexity	Intuition
Most common setting →	Small	Easy	First choose topic sizes, then match words to topics
	Large	NP-hard	Reduction from set packing

Maximize $p(\theta | w_{1:N})$

	Dirichlet hyper-parameters	Complexity	Intuition
Most common setting →	$\alpha_t \geq 1$	Easy	Maximizing concave function
	$\alpha_t < 1$	NP-hard	Reduction from set cover

Sample from $p(\theta | w_{1:N})$

	Dirichlet hyper-parameters	Complexity	Intuition
	$\alpha_t \geq 1$	Easy	Log-concave distribution
	$\alpha_t \approx 0$	NP-hard	Reduction from set cover

Practical Variational Inference for Neural Networks

Alex Graves

CIFAR Junior Fellow
University of Toronto
Canada

Method

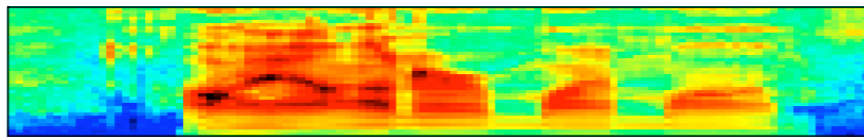
- Instead of learning neural network weights, we learn the mean and variance of a separate Gaussian for each weight: **adaptive weight noise**
- The loss is the number of bits to transmit the errors plus the number of bits to transmit the weights: **optimisation = compression**
- The more information the weights store about the training data, the more they cost to send: **no overfitting**
- Can interpret as **MDL** or stochastic **variational inference**

Advantages

- Applies to **any differentiable log-loss model** (previous variational methods for neural networks were limited to very simple architectures)
- **No validation set** required (as long as the training data is compressed)
- The weight costs tell you how **important** each weight is to the network
- Can **prune** the network by removing weights with high probability at zero

Results

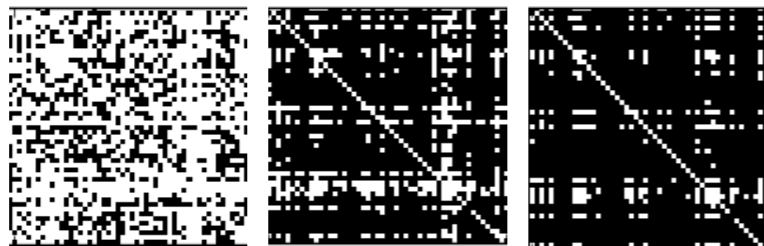
- Outperformed other regularisers for phoneme recognition on TIMIT with a complex neural network



ay aa nx er m ay m aa m

Regulariser	Error Rate
L2	27.4%
L1	26.0%
Weight noise	25.4%
Adaptive weight noise	23.8%

- Allowed many weights to be pruned with little impact (even improvement!) on performance



Weight matrix at different pruning thresholds: black=prune, white=keep

Weights Pruned	Error Rate
22.6%	24.0%
54.8%	23.5%
69.1%	23.7%
88.5%	24.5%

Multilinear Subspace Regression: An Orthogonal Tensor Decomposition Approach

Qibin Zhao¹, Cesar F. Caiafa², Danilo P. Mandic³, Liqing Zhang⁴, Tonio Ball⁵, Andreas Schulze-Bonhage⁵, and Andrzej Cichocki¹

¹ *Brain Science Institute, RIKEN, Japan*

² *IAR, CONICET, Argentina*

³ *Imperial College, UK*

⁴ *Shanghai Jiao Tong University, China*

⁵ *Albert-Ludwigs-University, Germany*

NIPS 2011

Presented by Qibin Zhao

POSTER: W043

LABSP: <http://www.bsp.brain.riken.jp/>

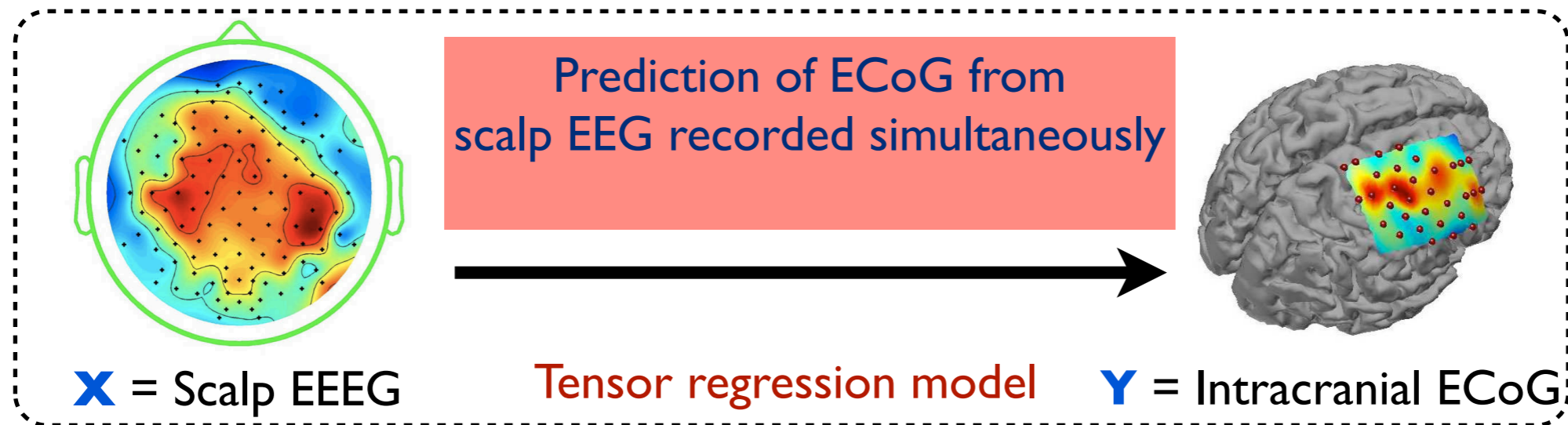


Multilinear regression and applications

► **Tensor** representation of multidimensional data

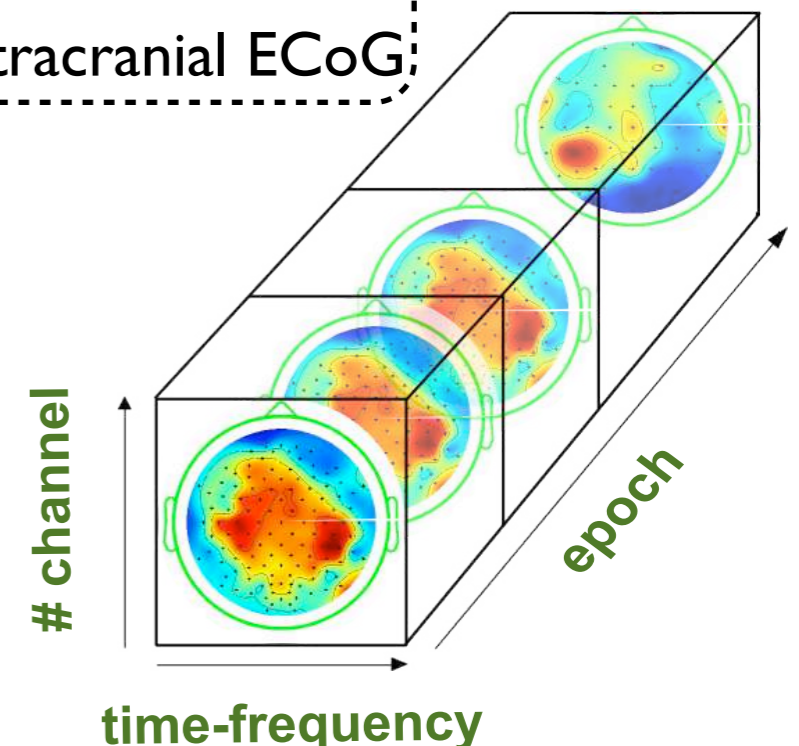
- EEG, ECoG (spatial, temporal, frequency, epoch,...)
- Physical meaning - ease of interpretation

► From multivariate to multi-way array processes - partial least squares (PLS)



► Standard PLS applied on **matricization** of both X and Y

- Small sample size problem
- Overfitting problem (high dimension of subspace basis)
- Lack of physical interpretation for loadings



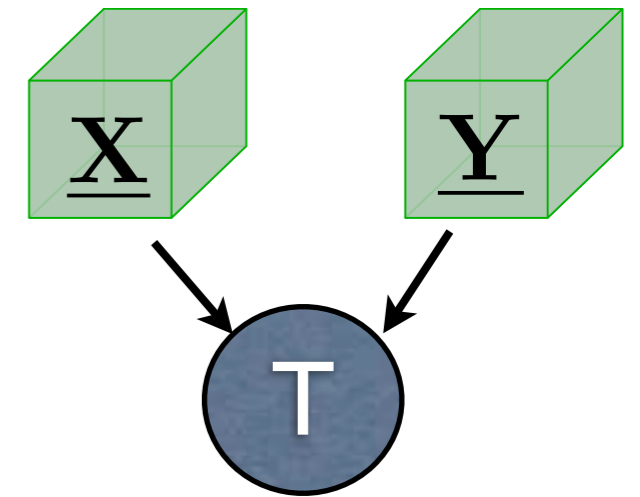
Proposed approach

Objective function

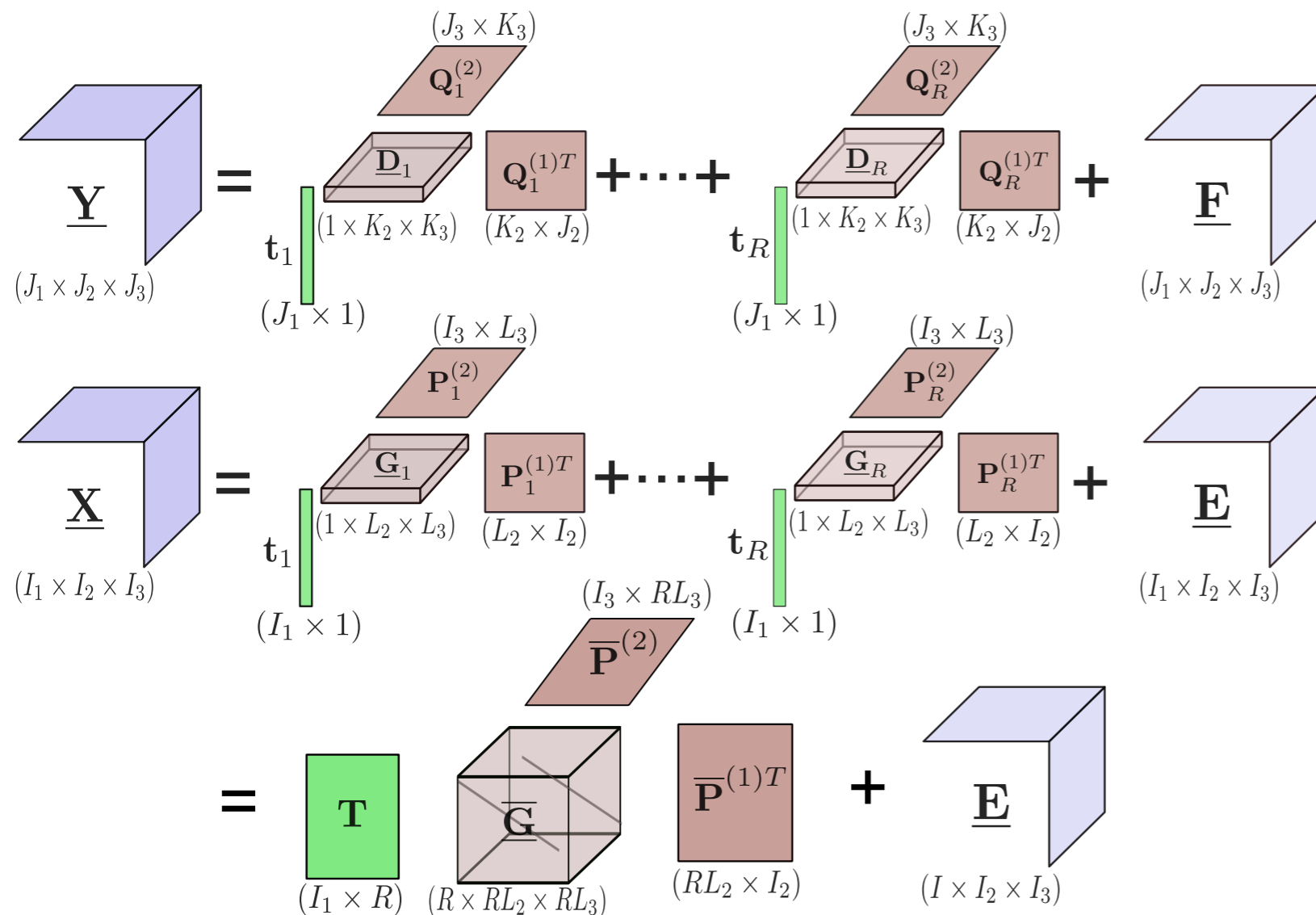
$$\min_{\{\mathbf{P}^{(n)}, \mathbf{Q}^{(m)}\}} \left\| \underline{\mathbf{X}} - \llbracket \underline{\mathbf{G}}; \mathbf{t}, \mathbf{P}^{(1)}, \dots, \mathbf{P}^{(N-1)} \rrbracket \right\|^2 + \left\| \underline{\mathbf{Y}} - \llbracket \underline{\mathbf{D}}; \mathbf{t}, \mathbf{Q}^{(1)}, \dots, \mathbf{Q}^{(M-1)} \rrbracket \right\|^2$$

$$\text{s. t. } \{\mathbf{P}^{(n)T} \mathbf{P}^{(n)}\} = \mathbf{I}_{L_{n+1}}, \quad \{\mathbf{Q}^{(m)T} \mathbf{Q}^{(m)}\} = \mathbf{I}_{K_{m+1}},$$

Brain data Behavior data



Latent variable



Raw Data Latent variables Loadings Residuals

Extension of PLS to higher-order tensor data - HOPLS

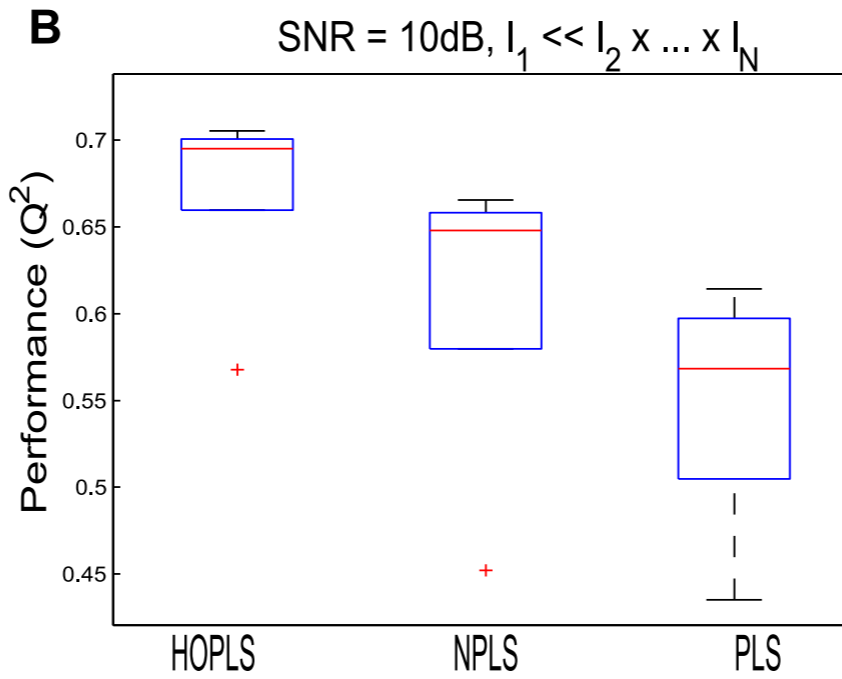
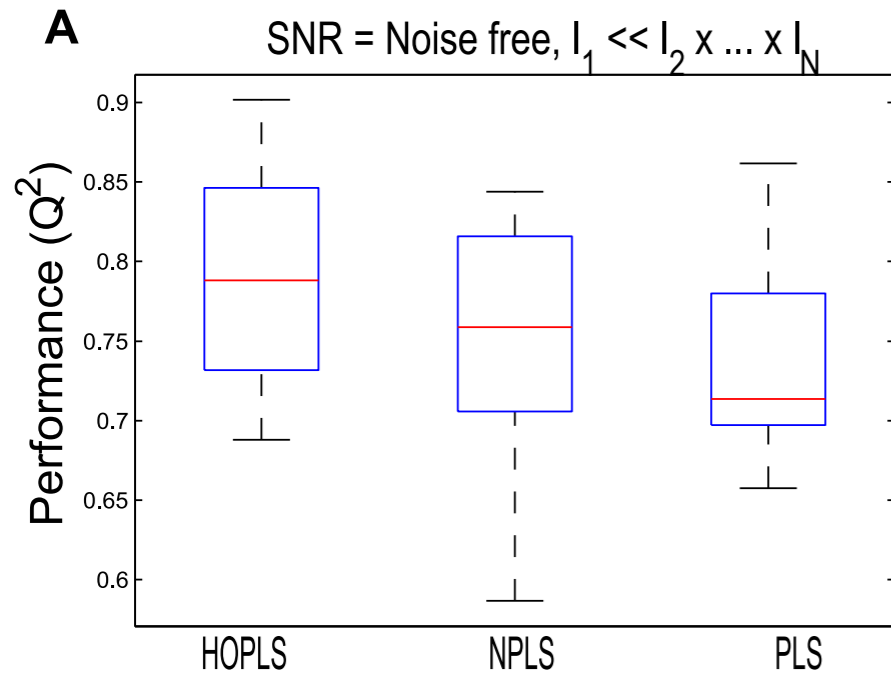
- Goal: to predict a tensor \mathbf{Y} from a tensor \mathbf{X}
- Approach: to extract the common latent variables

Properties:

- Flexible multilinear regression framework
- Projection on tensor subspace basis
- Efficient optimization algorithm using HOOI on the n -mode cross-covariance tensor

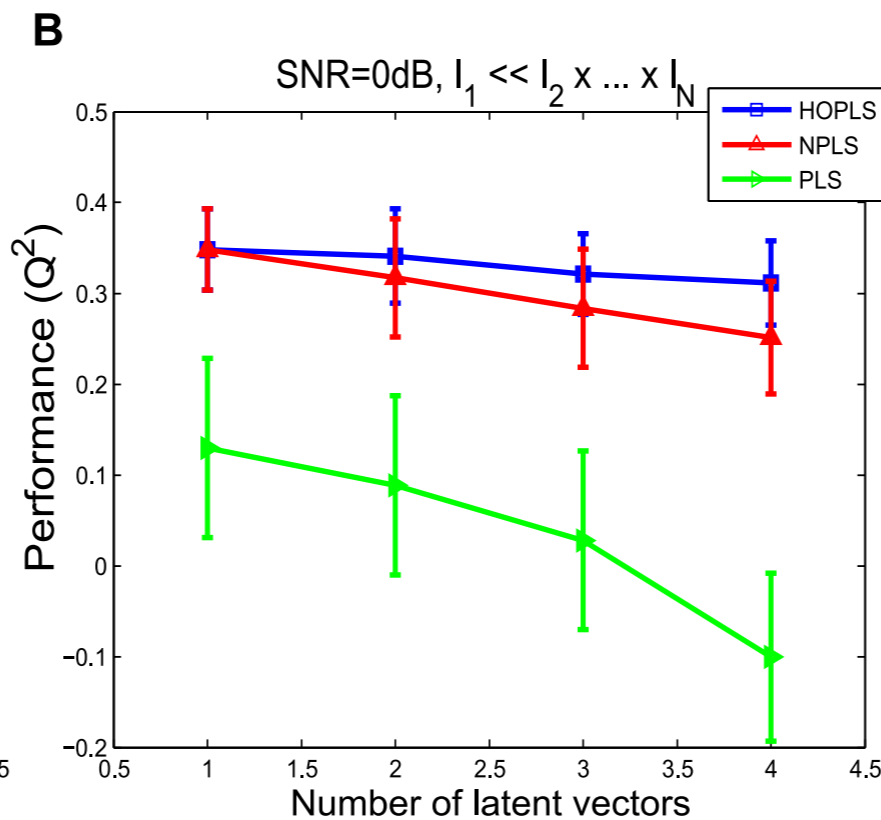
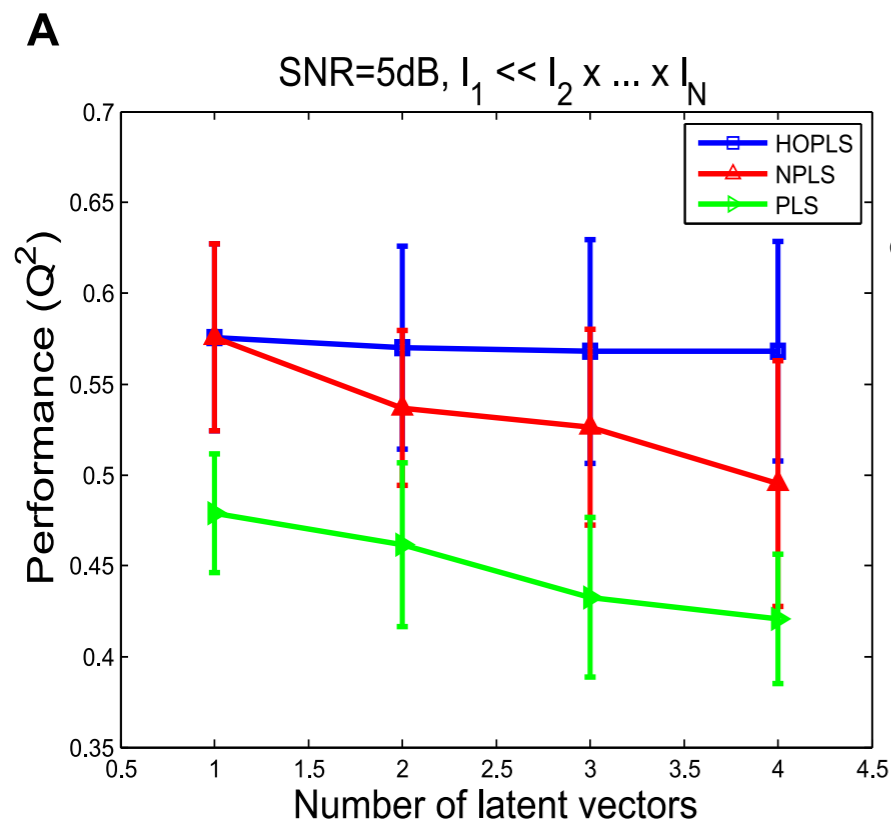
Key advantages

Small sample size



HOPLS: better prediction performance and enhanced robustness to noise

Robustness against overfitting and noise



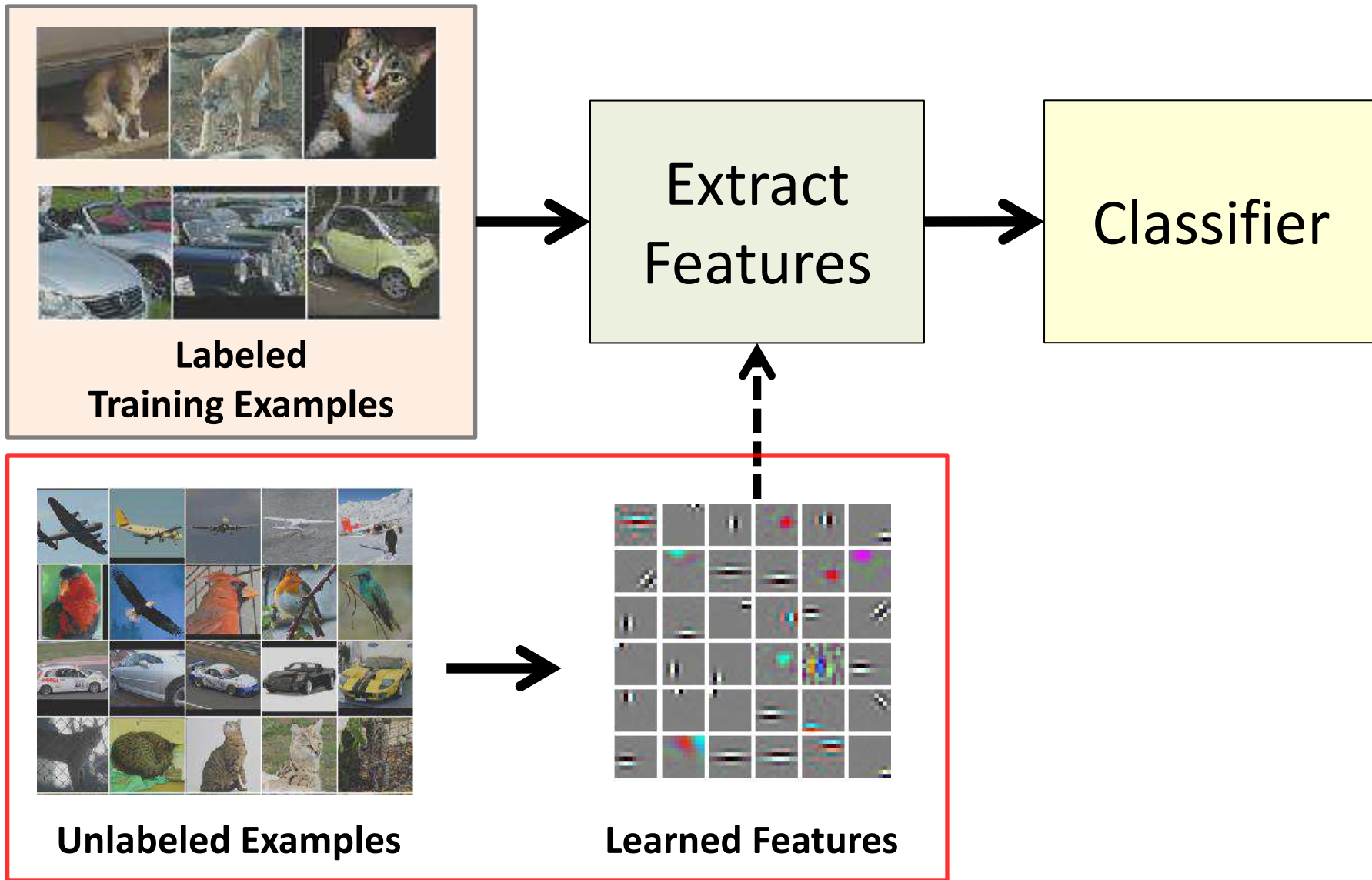
Stability of the performance of HOPLS, NPLS and PLS for a varying number of latent vectors under different noise conditions

POSTER: W043



Sparse Filtering

Jiquan Ngiam, Pang Wei Koh, Zhenghao Chen, Sonia Bhaskar & Andrew Y. Ng



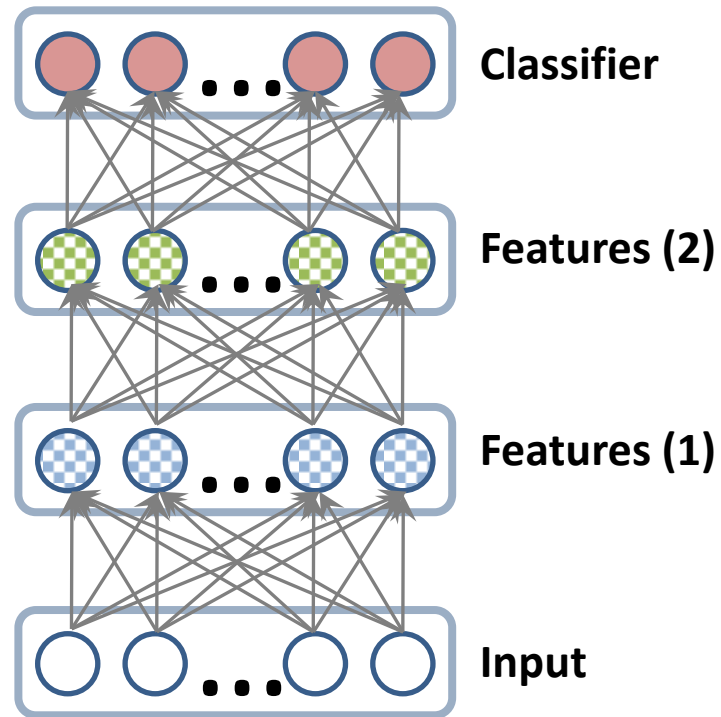


Sparse Filtering

Jiquan Ngiam, Pang Wei Koh, Zhenghao Chen, Sonia Bhaskar & Andrew Y. Ng

Why Sparse Filtering?

- **Easy, fast** approach to feature learning
- No hyper-parameters that need tuning
- Easy to evaluate objective function
- Minimal data preprocessing required
- Trains well with off-the-shelf optimization toolboxes (e.g., L-BFGS).





Sparse Filtering

Jiquan Ngiam, Pang Wei Koh, Zhenghao Chen, Sonia Bhaskar & Andrew Y. Ng

Examples

Feature Values

	x_1	x_2	x_3	x_4	...	x_m
f_1	0.5	2	0	1.5	...	4
f_2	0	0	2.5	0	...	0
...			...			
f_{99}	3.2	0	1.6	0.3	...	1
...			...			
f_n	4	0.5	0	1	...	3

$$M_{ij} = |w_i^T x_j|$$

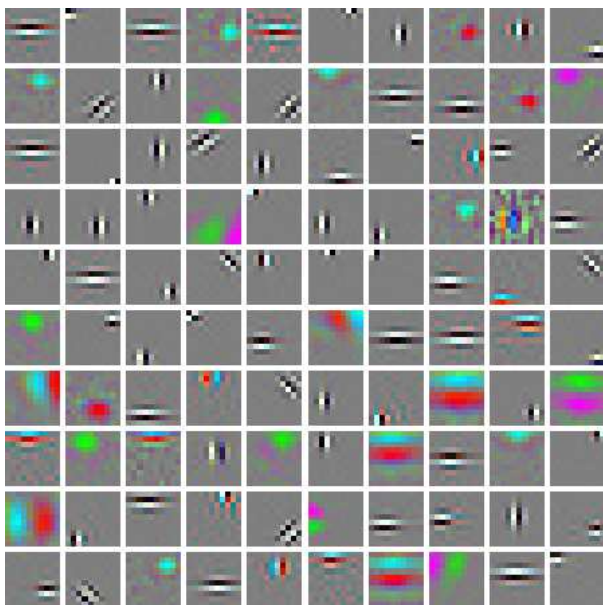
Sparse Filtering Objective Function

1. Normalize across rows
2. Normalize across columns
3. Cost Function =
Sum of the normalized entries

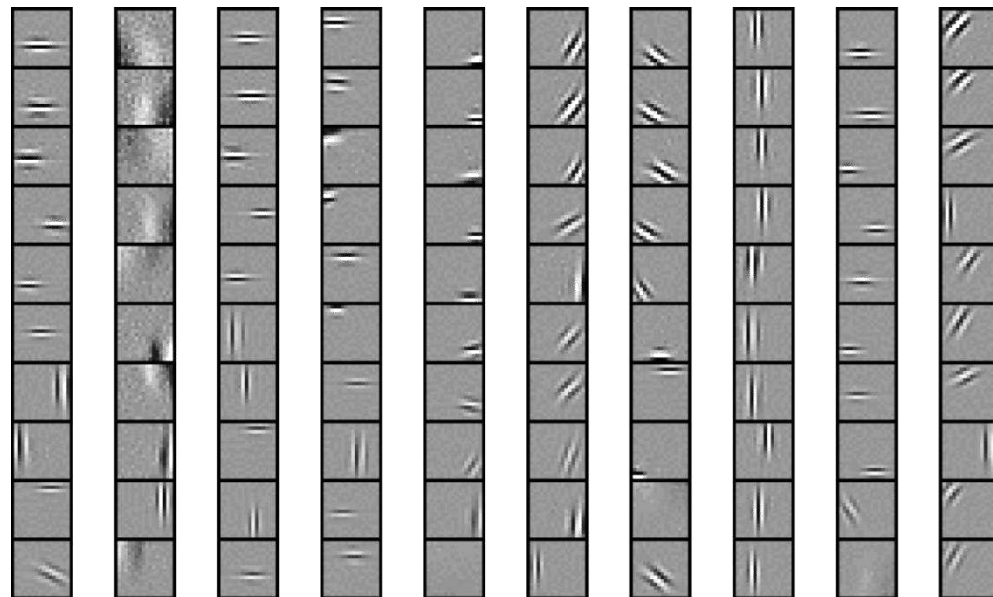


Sparse Filtering

Jiquan Ngiam, Pang Wei Koh, Zhenghao Chen, Sonia Bhaskar & Andrew Y. Ng



1st Layer Visualizations (STL Dataset)



2nd Layer Visualizations (Natural Images)

Evaluated sparse filtering features on natural images, image classification (STL Dataset), audio classification (TIMIT).

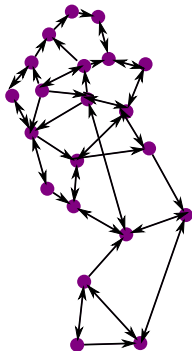
Results comparable to state-of-the-art and fast!

code available at <http://cs.stanford.edu/~jngiam/>

Directed Graph Embedding: an Algorithm based on Continuous Limits of Laplacian-type Operators

Dominique Perrault-Joncas, Marina Meilă
University of Washington

Directed Graph Problem

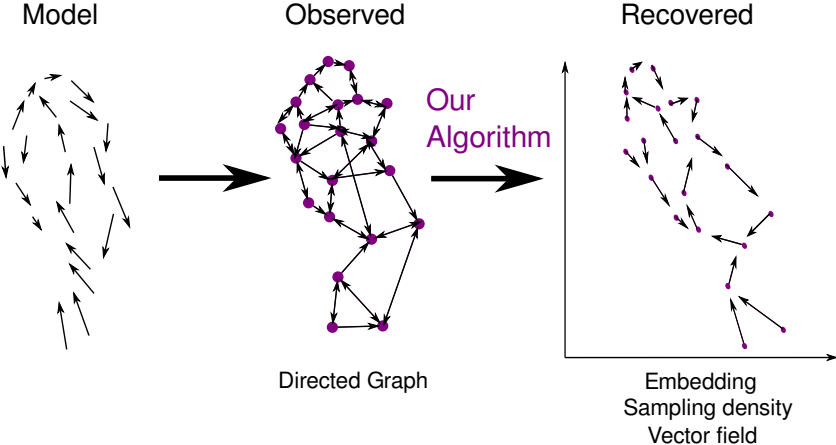


- Embed directed graph in euclidean space

AND

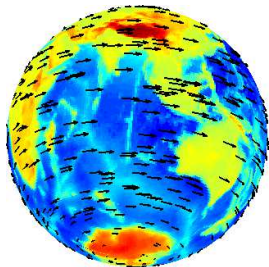
- Capture the directionality of the graph

Model Schematic

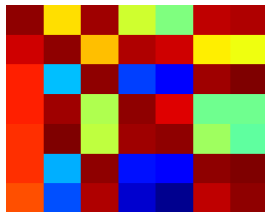


Artificial Data

Model

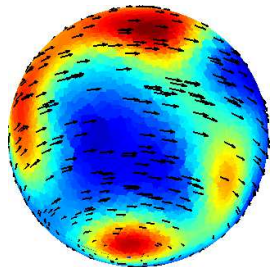


Observed



5000x5000 Asymmetric
adjacency matrix

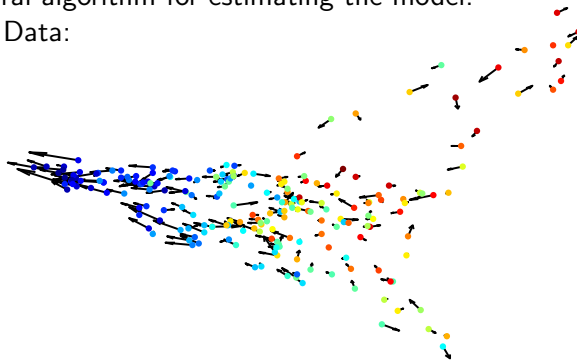
Recovered



Embedding
Sampling density
Vector field

Main Contributions

- 1 Manifold-based generative model for directed graphs with weighted edges.
- 2 Asymptotic results for diffusion operators constructed from the directed graphs.
- 3 Natural algorithm for estimating the model.
- 4 Real Data:



Noise Thresholds for Spectral Clustering

Sivaraman Balakrishnan
Poster: W056



Min Xu



Akshay Krishnamurthy



Aarti Singh



School of Computer Science
Carnegie Mellon University

Traditional analyses of Spectral Clustering

□ k-way spectral clustering

- Compute $\mathbf{L} = \mathbf{D} - \mathbf{W}$, $\mathbf{v}_1, \dots, \mathbf{v}_k \leftarrow$ smallest k eigenvectors of \mathbf{L}
- Embed each data point i into k -dim space $\mathbf{x}(i) = [\mathbf{v}_1(i), \dots, \mathbf{v}_k(i)]$
- Run k -means on embedded data points

High-level justification: Connection to graph cut, random walks on graph, electric network theory, Laplace-Beltrami operator on manifold – **don't translate to cluster recovery guarantees**

Perturbation Analysis: Rohe et. al. (2010) and McSherry (2001) – spectral algorithms for planted partition (structured random graph) model (constant block similarities, low rank)

Jordan, Weiss (2001), Huang, Yan, Jordan, Taft (2009) – eigenvectors are **stable in l_2 -norm** (Davis-Kahan Theorem) under small similarity perturbations

Our contributions – 1/2

- Study hierarchical spectral clustering and traditional k-way spectral clustering
- Characterization of general similarity conditions under which true eigenvectors reflect cluster structure, including **eigenvectors of hierarchically-structured high-rank matrices**
- Stability of eigenvectors in l_∞ -norm under sub-Gaussian perturbation
- Precise characterization of **total clustering error** of k-way and hierarchical spectral clustering
 - As a function of noise variance, number of objects, size of clusters and within v/s between cluster similarity gap

Our contributions – 2/2

- Information theoretic (minimax) optimality of signal-to-noise thresholds

- Minimax lower bound:** No clustering method can succeed if

$$\sigma = \omega \left(\gamma \sqrt{\frac{\log n}{n}} \right)$$

σ - Noise std. dev. of perturbation, n - number of objects

γ - Gap between inter and intra cluster similarity, γ/σ - SNR

- Ratio min-cut** (combinatorial) achieves this rate up to constants

- Spectral clustering** succeeds if

$$\sigma = o \left(\gamma \sqrt[4]{\frac{\log n}{n}} \right)$$

- Remarks:

- Price of computational efficiency: ratio min-cut (combinatorial) outperforms spectral clustering (efficient)

- Conjecture rate can be improved under different conditions on noise