



Mutual Angular Regularization of Latent Variable Models: Theory, Algorithm and Applications

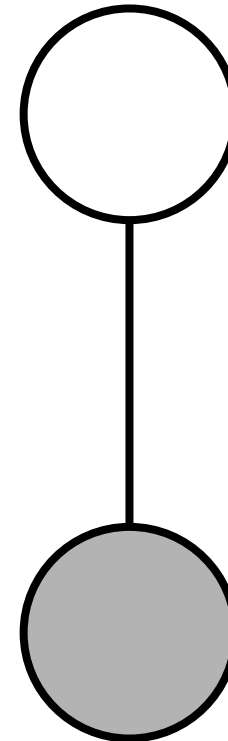
Pengtao Xie

Joint work with Yuntian Deng and Eric Xing
Carnegie Mellon University

Latent Variable Models (LVMs)

Machine Learning

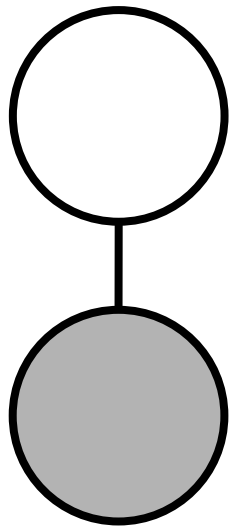
Latent Variable Models



Latent Variable Models

Topic Models

Gaussian Mixture Model



Topics

Groups

Words

Feature vectors

Hidden Markov Model, Kalman Filtering, Restricted Boltzmann Machine, Deep Belief Network, Factor Analysis, etc.

Neural Network, Sparse Coding, Matrix Factorization, Distance Metric learning, Principal Component Analysis, etc.

Latent Variable Models

Latent Factors Behind Data

Topics in Documents

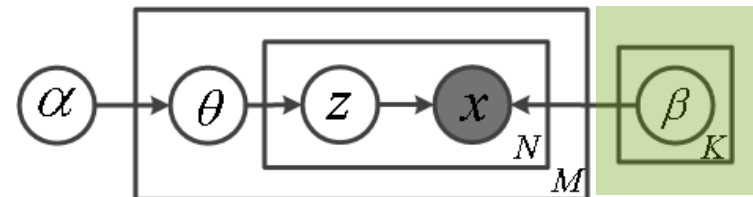
Politics	Economics	Education
Obama Constitution Government	GDP Bank Marketing	University Knowledge Student

Groups in Images

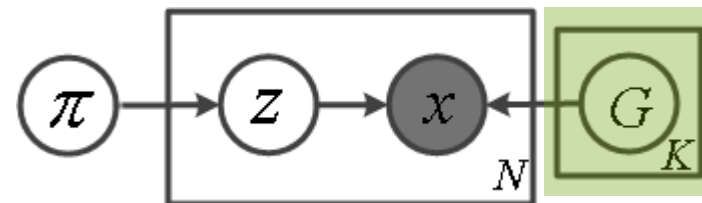


Components in LVMs

Topic Models

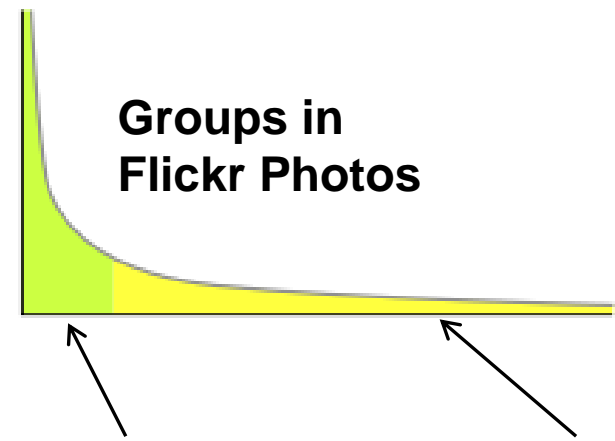
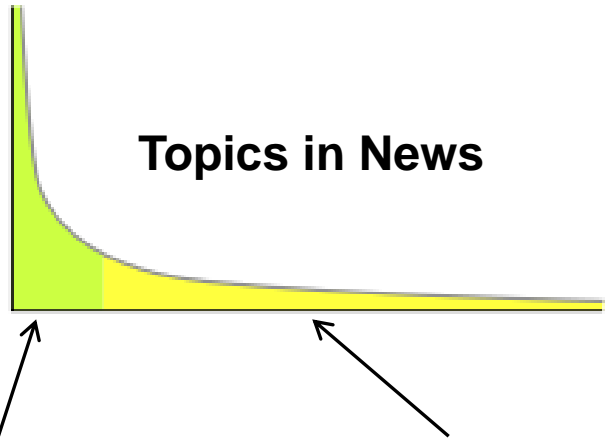


Gaussian Mixture Model



Motivation I: Popularity of latent factors is skewed

- Popularity of latent factors follows a power-law distribution



Dominant Topics

Long-Tail Topics

Politics	Economics
Obama Constitution Government	GDP Bank Marketing

Furniture	Flower
Sofa Closet Curtain	Rose Tulip Lily

Dominant Groups

Long-Tail Groups

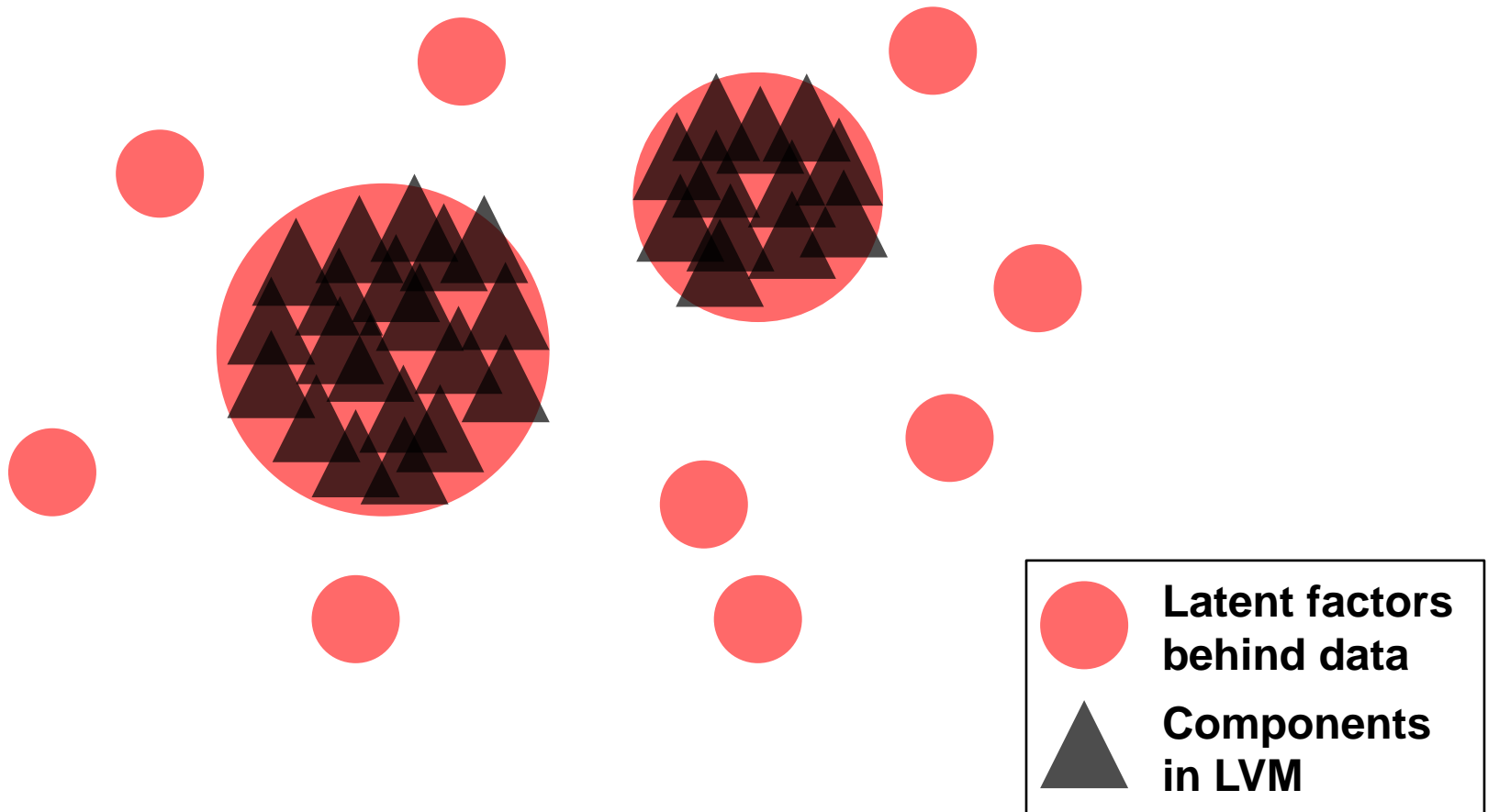


Standard LVMs are insufficient to capture long-tail factors

- Latent Dirichlet Allocation (LDA)
 - “Extremely common words tend to dominate all topics” (Wallach, 2009)
 - Tencent Peacock LDA, “When learning $\geq 10^5$ topics, around 20% ~ 40% topics have duplicates in practice” (Wang, 2015)
- Restricted Boltzmann Machine
 - Ran on 20-Newsgroup dataset
 - Many duplicate topics (e.g., the three exemplar topics are all about politics)
 - Common words occur repeatedly across topics, such as iraq, clinton, united, weapons

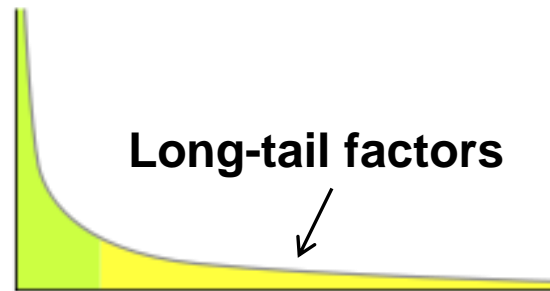
Topic 1	Topic 2	Topic 3
president	iraq	iraq
clinton	united	un
iraq	un	iraqi
united	weapons	lewinsky
spkr	iraqi	saddam
house	nuclear	clinton
people	india	baghdad
lewinsky	minister	inspectors
government	saddam	weapons
white	military	white

Standard LVMs are insufficient to capture long-tail factors



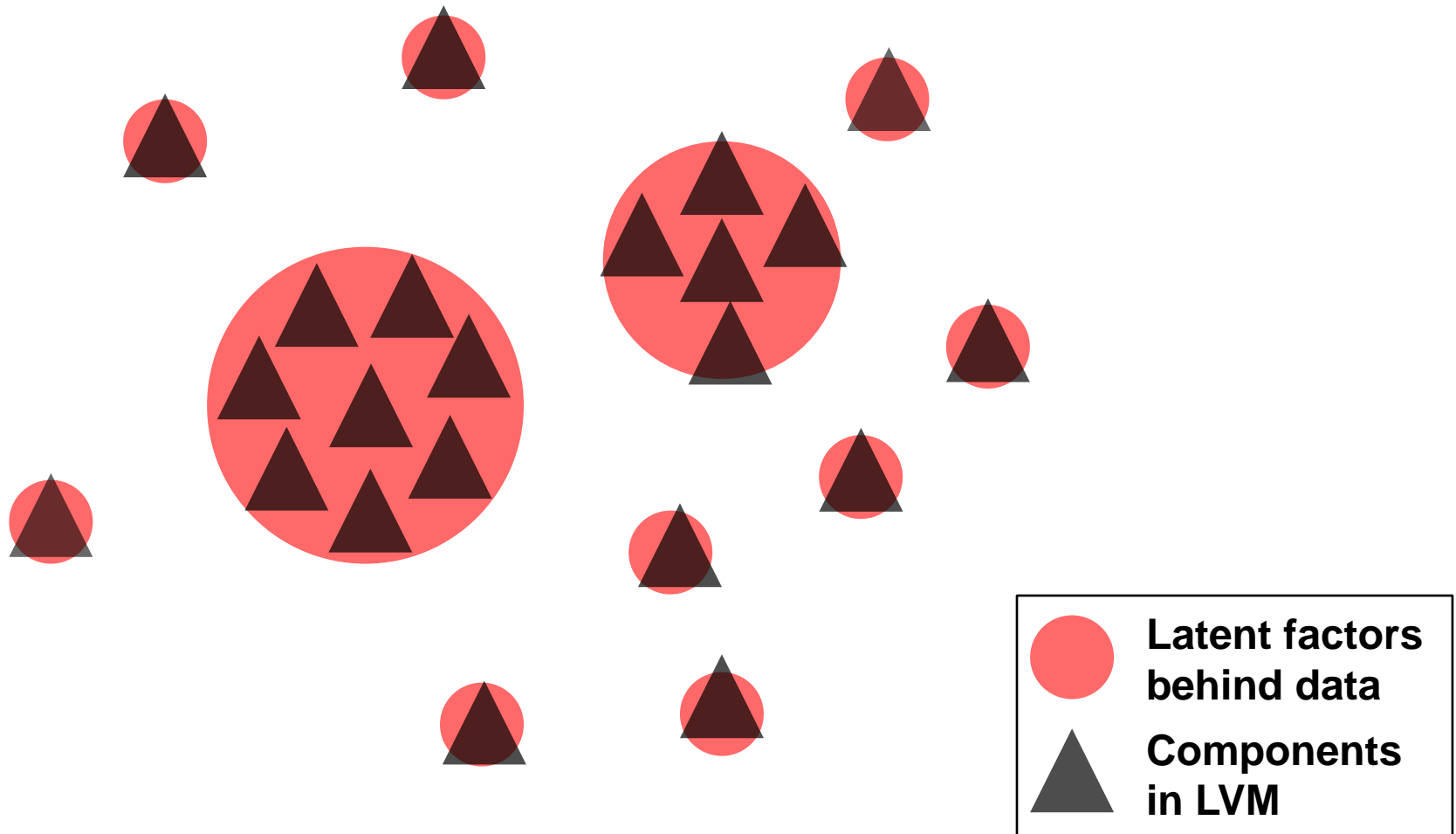
Long-tail factors are important

- The amount of long-tail factors is large



- Long-tail factors are more important than dominant factors in some applications
 - Example: Tencent applied topic models for advertisement and showed that long-tail topics such as “lose weight”, “nursing” improves click-through rate by 40% (Jin, 2015)

Diversification

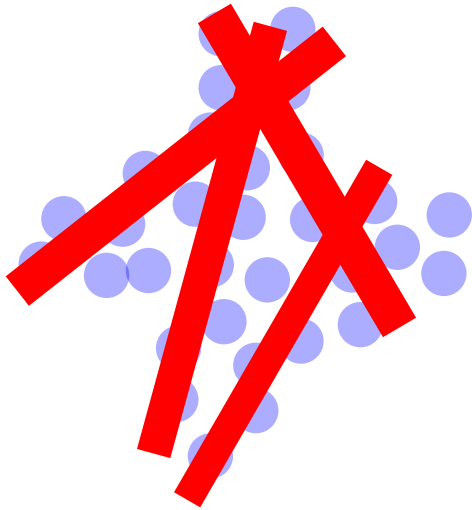


Motivation II: Tradeoff induced by the number of components k

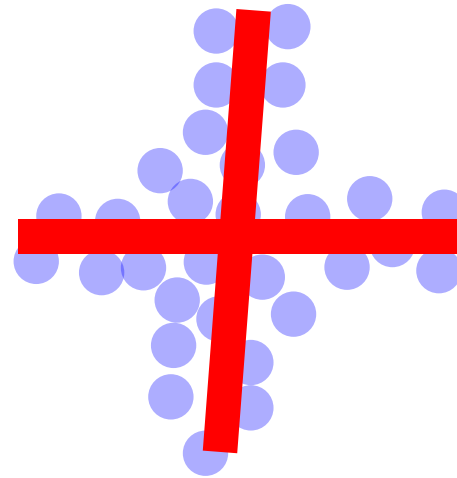
- Tradeoff between Expressiveness and Complexity
 - Small k : low expressiveness, low complexity
 - Large k : high expressiveness, high complexity
- Can we achieve the best of both worlds?
 - Small k : high expressiveness, low complexity

Reduce model complexity without sacrificing expressiveness

Without diversification



With diversification



Use components to capture the principal directions of data point cloud



Mutual Angular Regularization of LVMs

- Goal: encourage the components to diversely spread out to (1) improve the coverage of long-tail latent factors; (2) reduce model complexity without compromising expressiveness
- Approach:
 - Define a score based on mutual angles to measure the diversity of components
 - Use the score to regularize latent variable models and control the geometry of the latent space during learning

Outline

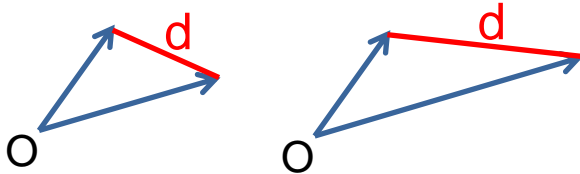
- Mutual Angular Regularizer
- Algorithm
- Applications
- Theory

Mutual Angular Regularizer

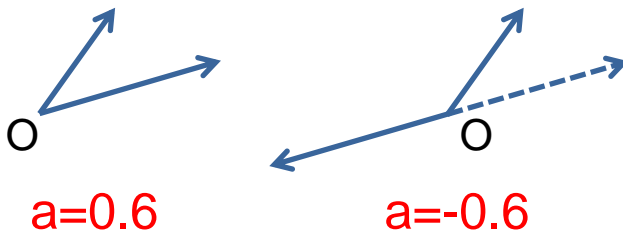
- Components are parametrized by vectors
 - In Latent Dirichlet Allocation, each **topic** has a multinomial vector
 - In Sparse Coding, each **dictionary item** has a real vector
- Measure the dissimilarity between two vectors
- Measure the diversity of a vector set

Dissimilarity between two vectors

- Invariant to scale, translation, rotation and orientation of the two vectors
- Euclidean distance, L1 distance
 - Distance d is variant to scale

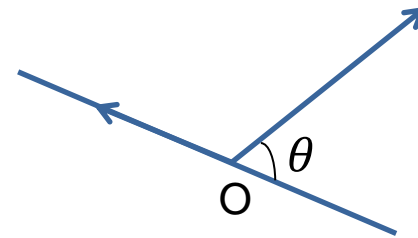
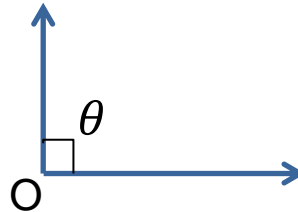
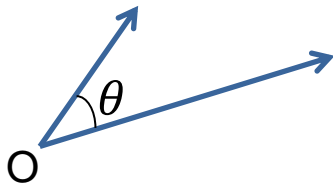


- Negative cosine similarity
 - Negative cosine similarity a is variant to orientation



Dissimilarity between two vectors

- Non-obtuse angle θ



- Invariant to scale, translation, rotation and orientation of the two vectors
- Definition

$$\theta = \arccos\left(\frac{|\mathbf{x} \cdot \mathbf{y}|}{\|\mathbf{x}\| \|\mathbf{y}\|}\right)$$

Measure the diversity of a vector set

- Based on the pairwise dissimilarity measure between vectors
- The diversity of a set of vectors $\mathbf{A} = \{\mathbf{a}_i\}_{i=1}^K$ is defined as

Mutual Angular Regularizer $\rightarrow \Omega(\mathbf{A}) = \underbrace{\frac{1}{K(K-1)} \sum_{i=1}^K \sum_{\substack{j=1 \\ j \neq i}}^K \theta_{ij}}_{\text{Mean of angles}} - \underbrace{\frac{1}{K(K-1)} \sum_{i=1}^K \sum_{\substack{j=1 \\ j \neq i}}^K \left(\theta_{ij} - \frac{1}{K(K-1)} \sum_{\substack{p=1 \\ q=1 \\ q \neq p}}^K \sum_{\substack{p=1 \\ q=1 \\ q \neq p}}^K \theta_{pq} \right)}_{\text{Variance of angles}}^2 \quad \theta_{ij} = \arccos \left(\frac{|\mathbf{a}_i \cdot \mathbf{a}_j|}{\|\mathbf{a}_i\| \|\mathbf{a}_j\|} \right)$

- Mean: summarize how these vectors are different from each other on the whole
- Variance: encourage the vectors to **evenly** spread out



LVM with Mutual Angular Regularization (MAR-LVM)

$$\max_{\mathbf{A}} L(D; \mathbf{A}) + \lambda \Omega(\mathbf{A})$$

$$\Omega(\mathbf{A}) = \frac{1}{K(K-1)} \sum_{i=1}^K \sum_{\substack{j=1 \\ j \neq i}}^K \theta_{ij} - \frac{1}{K(K-1)} \sum_{i=1}^K \sum_{\substack{j=1 \\ j \neq i}}^K \left(\theta_{ij} - \frac{1}{K(K-1)} \sum_{\substack{p=1 \\ q=1 \\ q \neq p}}^K \sum_{q=1}^K \theta_{pq} \right)^2$$

$$\theta_{ij} = \arccos \left(\frac{|\mathbf{a}_i \cdot \mathbf{a}_j|}{\|\mathbf{a}_i\| \|\mathbf{a}_j\|} \right)$$

Algorithm

- Challenge: the mutual angular regularizer is non-smooth and non-convex w.r.t the parameter vectors $\mathbf{A} = \{\mathbf{a}_i\}_{i=1}^K$
- Derive a smooth lower bound
 - The lower bound is easier to derive if the parameter vectors lie on a sphere
 - Decompose the parameter vectors into magnitudes and directions
- Proved that optimizing the lower bound with gradient ascent method can increase the mutual angular regularizer in each iteration

Optimization

Reparametrize

$$\mathbf{a}_i = g_i \tilde{\mathbf{a}}_i \quad g_i = \|\mathbf{a}_i\| \quad \|\tilde{\mathbf{a}}_i\| = 1 \quad \mathbf{A} = \text{diag}(\mathbf{g})\tilde{\mathbf{A}}$$

↑
↑

Magnitude Direction

$$\Omega(\tilde{\mathbf{A}}) = \Omega(\text{diag}(\mathbf{g})\tilde{\mathbf{A}})$$

$$\max_{\mathbf{g}, \tilde{\mathbf{A}}} L(D; \mathbf{A}) + \lambda \Omega(\mathbf{A}) \quad \longrightarrow \quad \max_{\mathbf{g}, \tilde{\mathbf{A}}} L(D; \mathbf{g}\tilde{\mathbf{A}}) + \lambda \Omega(\tilde{\mathbf{A}})$$

s.t. $\forall i, \|\tilde{\mathbf{a}}_i\| = 1, g_i \geq 0$

Alternating Optimization

Fix $\tilde{\mathbf{A}}$, optimize \mathbf{g}

$$\max_{\mathbf{g}} L(D; \mathbf{g}\tilde{\mathbf{A}})$$

s.t. $\forall i, g_i \geq 0$



Fix \mathbf{g} , optimize $\tilde{\mathbf{A}}$

$$\max_{\tilde{\mathbf{A}}} L(D; \mathbf{g}\tilde{\mathbf{A}}) + \lambda \Omega(\tilde{\mathbf{A}})$$

s.t. $\forall i, \|\tilde{\mathbf{a}}_i\| = 1$

Optimize $\tilde{\mathbf{A}}$

$$\begin{aligned} \max_{\tilde{\mathbf{A}}} \quad & L(D; \mathbf{g}\tilde{\mathbf{A}}) + \lambda\Omega(\tilde{\mathbf{A}}) \\ \text{s.t.} \quad & \forall i, \|\tilde{\mathbf{a}}_i\| = 1 \end{aligned}$$

Lower bound

$$\Omega(\tilde{\mathbf{A}}) \geq \Gamma(\tilde{\mathbf{A}}) = \arcsin\left(\sqrt{\det(\tilde{\mathbf{A}}^T \tilde{\mathbf{A}})}\right) - \left(\frac{\pi}{2} - \arcsin\left(\sqrt{\det(\tilde{\mathbf{A}}^T \tilde{\mathbf{A}})}\right)\right)^2$$

Intuition of the lower bound: $\det(\tilde{\mathbf{A}}^T \tilde{\mathbf{A}})$ is the volume of the paralleliped formed by the vectors in $\tilde{\mathbf{A}}$. The larger $\det(\tilde{\mathbf{A}}^T \tilde{\mathbf{A}})$ is, the more likely that the vectors in $\tilde{\mathbf{A}}$ have larger angles (not surely). $\Gamma(\tilde{\mathbf{A}})$ is an increasing function w.r.t $\det(\tilde{\mathbf{A}}^T \tilde{\mathbf{A}})$. Hence larger $\Gamma(\tilde{\mathbf{A}})$ is likely to yield larger $\Omega(\tilde{\mathbf{A}})$.

Optimize the lower bound, which is smooth and much more amenable for optimization

$$\begin{aligned} \max_{\tilde{\mathbf{A}}} \quad & L(D; \mathbf{g}\tilde{\mathbf{A}}) + \lambda\Gamma(\tilde{\mathbf{A}}) \\ \text{s.t.} \quad & \forall i, \|\tilde{\mathbf{a}}_i\| = 1 \end{aligned}$$

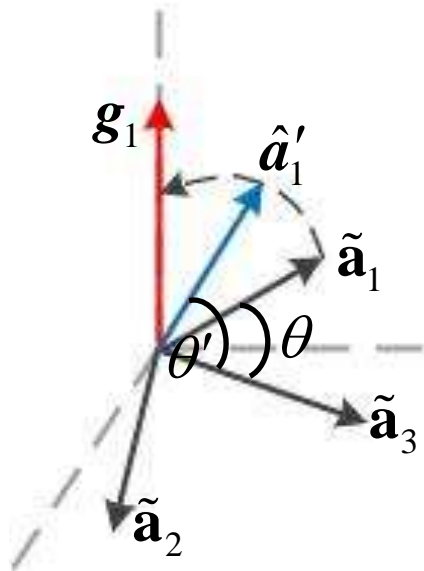
Close Alignment between the Regularizer and its Lower Bound

- If the lower bound is optimized with projected gradient ascent (PGA), the mutual angular regularizer can be increased in each iteration of the PGA procedure
 - Optimizing the lower bound with PGA can **increase the mean of the angles** in each iteration
 - Optimizing the lower bound with PGA can **decrease the variance of the angles** in each iteration

$$\begin{aligned}
 \uparrow \Omega(\mathbf{A}) = & \underbrace{\frac{1}{K(K-1)} \sum_{i=1}^K \sum_{\substack{j=1 \\ j \neq i}}^K \theta_{ij}}_{\text{Mean} \uparrow} - \underbrace{\frac{1}{K(K-1)} \sum_{i=1}^K \sum_{\substack{j=1 \\ j \neq i}}^K \left(\theta_{ij} - \frac{1}{K(K-1)} \sum_{\substack{p=1 \\ p \neq i}}^K \sum_{\substack{q=1 \\ q \neq p}}^K \theta_{pq} \right)^2}_{\text{Variance} \downarrow}
 \end{aligned}$$

Geometry Interpretation of the Close Alignment

- The gradient of the lower bound w.r.t $\tilde{\mathbf{a}}_i$ is orthogonal to all other vectors $\{\tilde{\mathbf{a}}_1, \tilde{\mathbf{a}}_2, \dots, \tilde{\mathbf{a}}_K\} / \{\tilde{\mathbf{a}}_i\}$
- Move $\tilde{\mathbf{a}}_i$ along its gradient direction would enlarge its angle with other vectors



$\tilde{\mathbf{a}}_1 \tilde{\mathbf{a}}_2 \tilde{\mathbf{a}}_3$ are parameter vectors

\mathbf{g}_1 is the gradient of $\tilde{\mathbf{a}}_1$ and are orthogonal to $\tilde{\mathbf{a}}_2 \tilde{\mathbf{a}}_3$

$$\hat{\mathbf{a}}'_1 = \tilde{\mathbf{a}}_1 + \eta \mathbf{g}_1$$

The angle θ' between $\hat{\mathbf{a}}'_1$ and $\tilde{\mathbf{a}}_3$ is greater than θ between $\tilde{\mathbf{a}}_1$ and $\tilde{\mathbf{a}}_3$

Summary of Algorithm for MAR-LVM

$$\begin{aligned} \max_{\mathbf{g}, \tilde{\mathbf{A}}} \quad & L(D; \mathbf{g}\tilde{\mathbf{A}}) + \lambda\Omega(\tilde{\mathbf{A}}) \\ \text{s.t.} \quad & \forall i, \|\tilde{\mathbf{a}}_i\| = 1, g_i \geq 0 \end{aligned}$$

While Not Converge

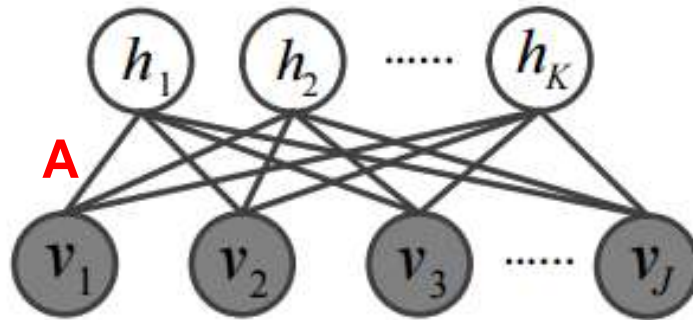
1. Fixing $\tilde{\mathbf{A}}$, solving the following sub-problem with projected gradient ascent or other methods

$$\begin{aligned} \max_{\mathbf{g}} \quad & L(D; \mathbf{g}\tilde{\mathbf{A}}) \\ \text{s.t.} \quad & \forall i, g_i \geq 0 \end{aligned}$$

2. Fixing \mathbf{g} , solving the following sub-problem with projected gradient ascent

$$\begin{aligned} \max_{\tilde{\mathbf{A}}} \quad & L(D; \mathbf{g}\tilde{\mathbf{A}}) + \lambda\Gamma(\tilde{\mathbf{A}}) \\ \text{s.t.} \quad & \forall i, \|\tilde{\mathbf{a}}_i\| = 1 \end{aligned}$$

Case Study --- Restricted Boltzmann Machine with Mutual Angular Regularization (MAR-RBM)



$$\max_{\mathbf{A}} L(D; \mathbf{A}) + \lambda \Omega(\mathbf{A})$$

Experiments

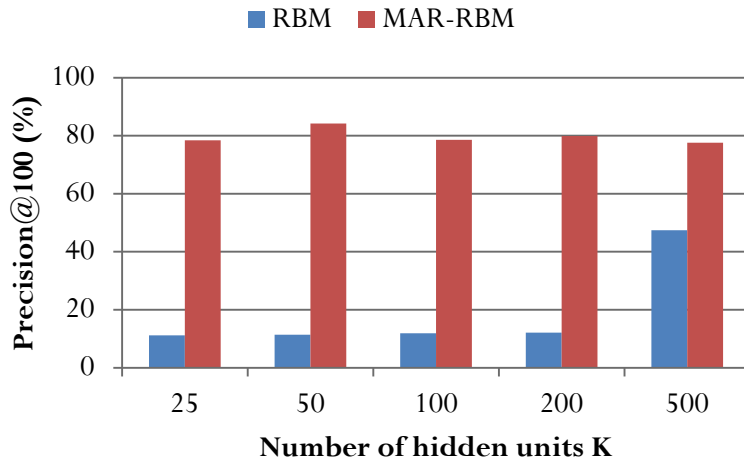
- Task: learn representations for documents
- Datasets

	#categories	#samples	vocab. size
TDT	30	9394	5000
20-News	20	18846	5000
Reuters	9	7195	5000

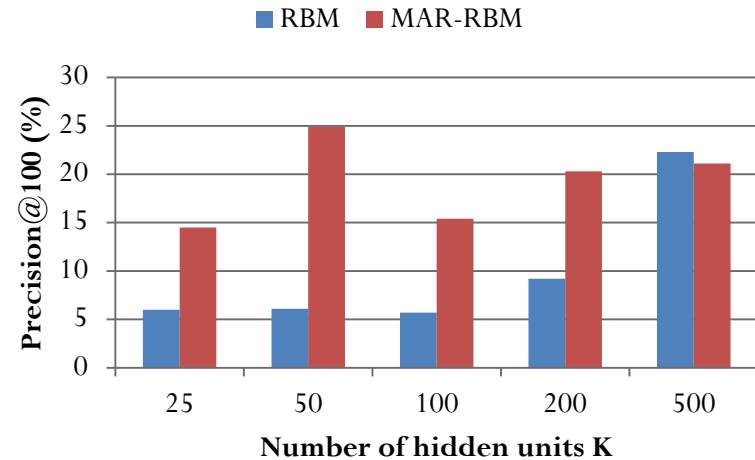
- Baselines
 - Bag-of-Words (BOW); Latent Dirichlet Allocation (LDA); LDA regularized with Determinantal Point Process prior (DPP-LDA); Pitman-Yor Process Topic Model (PYTM); Latent IBP Compound Dirichlet Allocation (LIDA); Neural Autoregressive Topic Model (DocNADE); Paragraph Vector (PV); Restricted Boltzmann Machine
- Evaluation
 - Retrieval: precision@100
 - Clustering: accuracy

Retrieval Precision

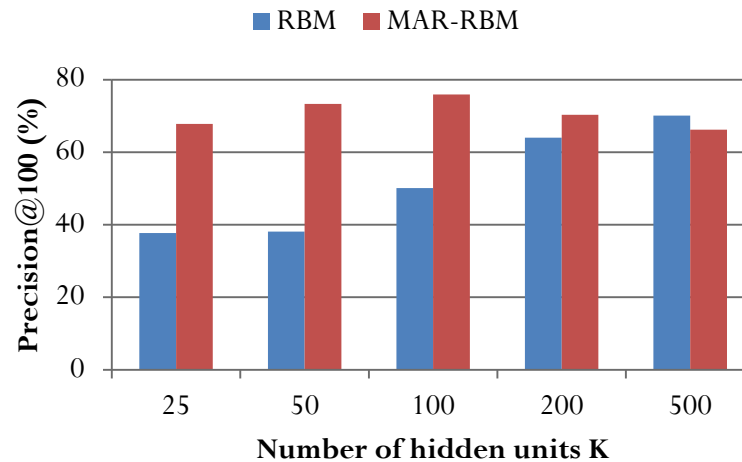
Retrieval Precision on TDT



Retrieval Precision on 20-News



Retrieval Precision on Reuters

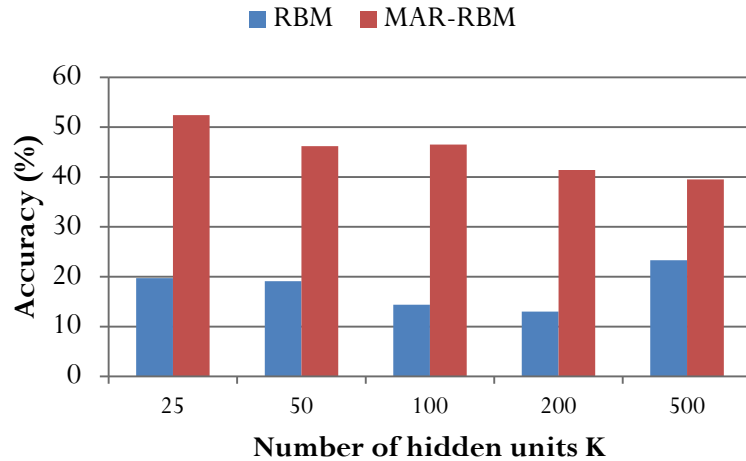


Retrieval Precision

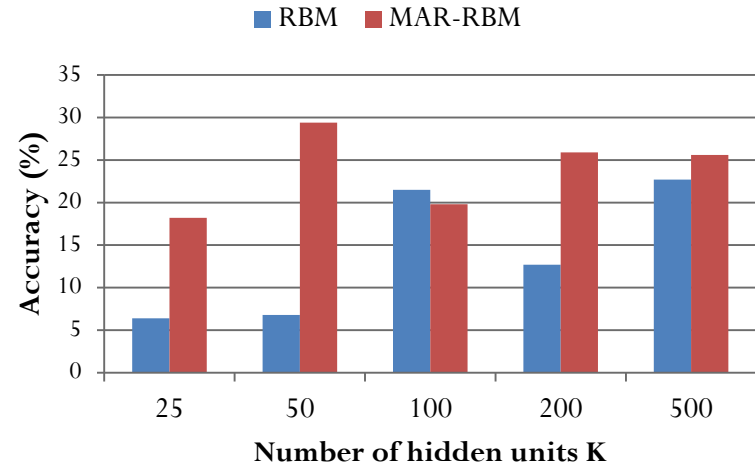
	TDT	20-News	Reuters
BOW	40.9	7.4	69.3
LDA	79.4	19.6	68.5
DPP-LDA	81.9	18.2	69.9
PYTM	78.7	20.1	70.6
LIDA	77.9	21.8	71.4
DocNADE	80.3	16.8	72.6
PV	81.7	19.1	76.9
RBM	47.4	22.3	70.1
MAR-RBM	84.2	24.9	75.9

Clustering Accuracy

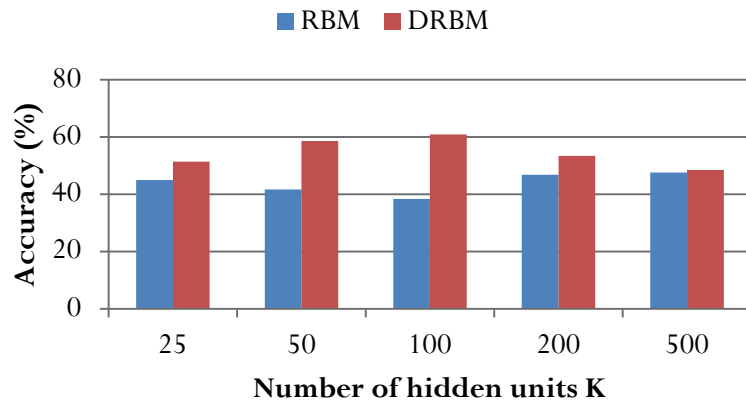
Clustering Accuracy on TDT



Clustering Accuracy on 20-News



Clustering Accuracy on Reuters



$$\text{Accuracy} = \frac{\sum_{i=1}^N \mathbb{I}(t_i = \text{map}(c_i))}{N}$$

t_i -- true label of document i

c_i -- cluster label

map -- Kuhn-Munkres permutation mapping

$\mathbb{I}(\cdot)$ -- indicator function

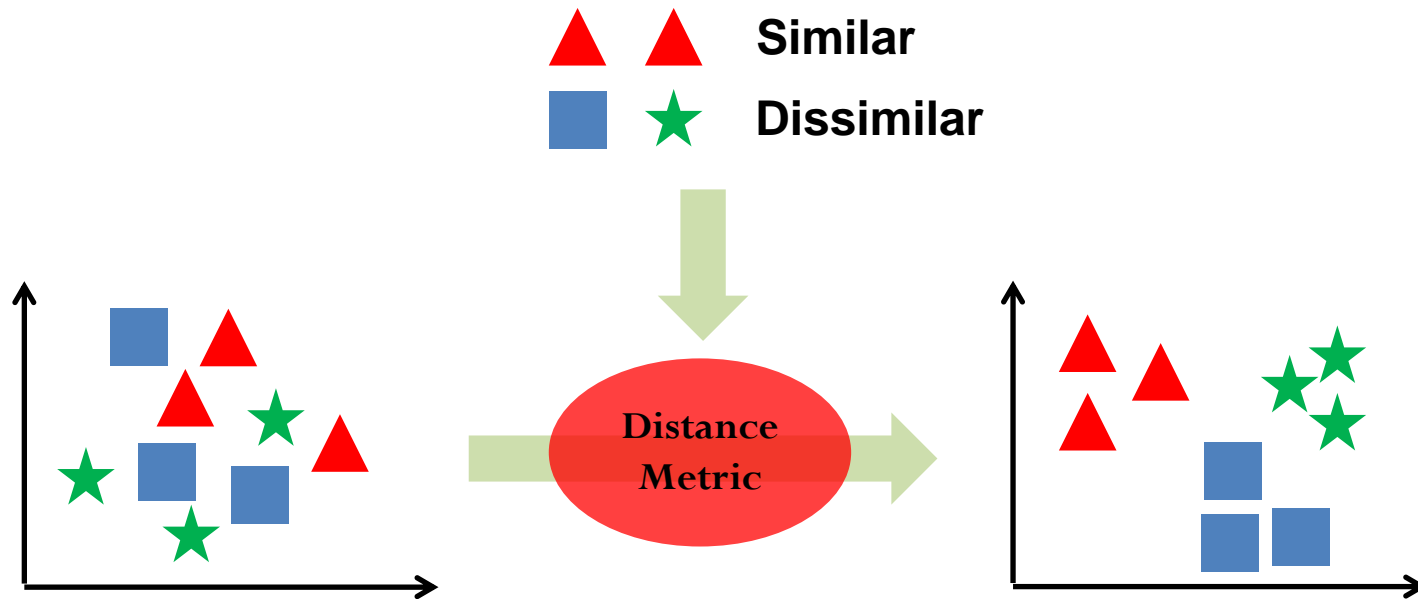
Clustering Accuracy

	TDT	20-News	Reuters
BOW	51.3	21.3	49.7
LDA	45.2	21.9	51.2
DPP-LDA	46.3	10.9	49.3
PYTM	46.9	21.5	51.7
LIDA	47.3	17.4	53.1
DocNADE	45.7	18.7	48.7
PV	48.2	24.3	52.8
RBM	23.3	22.7	47.6
MAR-RBM	52.4	29.4	60.9

Improvement Breakdown --Retrieval on Reuters

Category ID	1	2	3	4	5	6	7	8	9
Number of Documents	3713	2055	321	298	245	197	142	114	110
Precision@100 of RBM	0.69	0.44	0.09	0.10	0.06	0.04	0.04	0.03	0.03
Precision@100 of MAR-RBM	0.90	0.80	0.31	0.40	0.27	0.23	0.09	0.14	0.13
Relative Improvement of MAR-RBM over RBM	31%	81%	245%	289%	324%	421%	148%	366%	397%

Case Study --- Distance Metric Learning with Mutual Angular Regularization (MAR-DML)



- Wide applications in retrieval, clustering and classification

Distance Metric Learning

- Projection matrix



- k incurs tradeoff between model complexity and expressiveness
- Distance Metric Learning

$$\begin{aligned}
 \min_A \quad & \frac{1}{|\mathcal{S}|} \sum_{(x,y) \in \mathcal{S}} \|Ax - Ay\|^2 \\
 \text{s.t.} \quad & \|Ax - Ay\|^2 \geq 1, \forall (x, y) \in \mathcal{D}
 \end{aligned}$$

- Distance Metric Learning with Mutual Angular Regularization

$$\begin{aligned}
 \min_A \quad & \frac{1}{|\mathcal{S}|} \sum_{(x,y) \in \mathcal{S}} \|Ax - Ay\|^2 - \lambda \Omega(A) \\
 \text{s.t.} \quad & \|Ax - Ay\|^2 \geq 1, \forall (x, y) \in \mathcal{D}
 \end{aligned}$$

Experiments

- Datasets

	Feature Dim.	#training data	#data pairs
20-News	5000	11.3K	200K
15-Scenes	1000	3.2K	200K
6-Activities	561	7.4K	200K

- Baselines

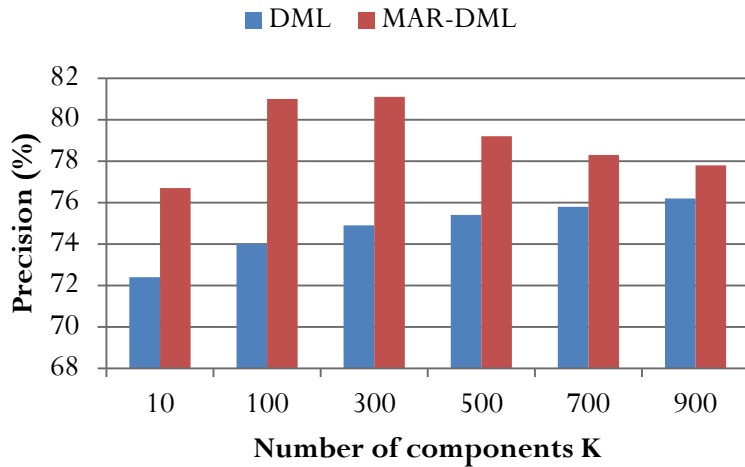
- Euclidean distance (EUC); Distance Metric Learning (DML); Large Margin Nearest Neighbor (LMNN) DML; Information Theoretical Metric Learning (ITML); Distance Metric Learning with Eigenvalue Optimization (DML-eig); Information-theoretic Semi-supervised Metric Learning via Entropy Regularization (Seraph)

- Evaluation

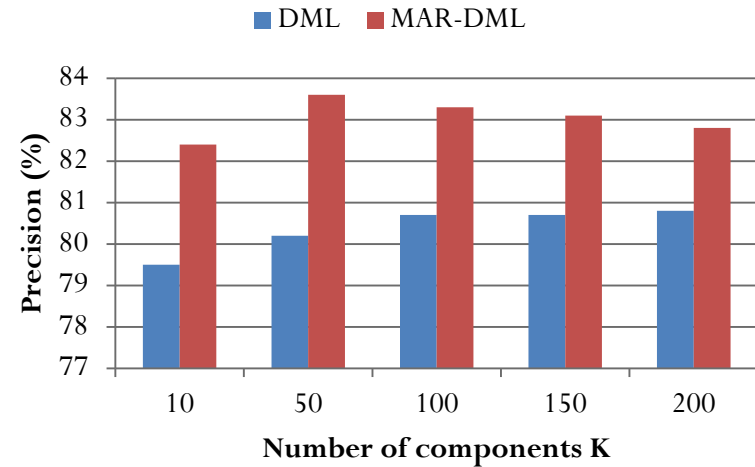
- Retrieval: precision
- Clustering: accuracy

Retrieval Precision

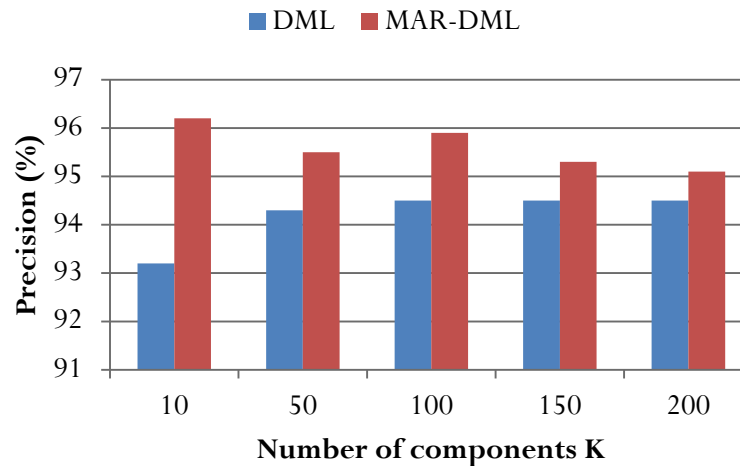
Retrieval Precision on 20-News



Retrieval Precision on 15-Scenes



Retrieval Precision on 6-Activities

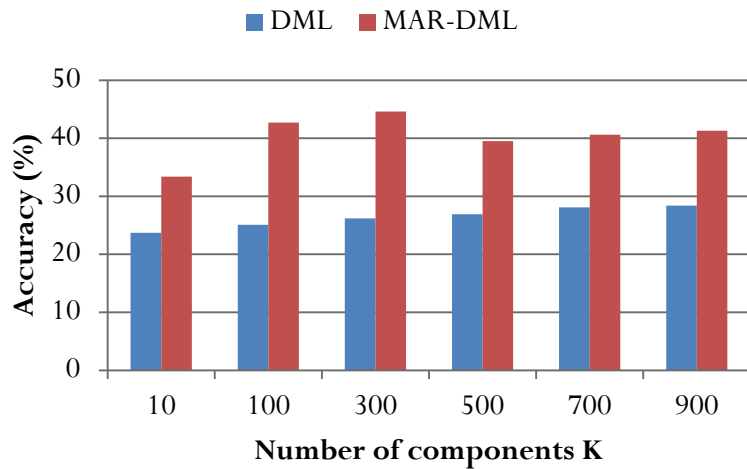


Retrieval Precision

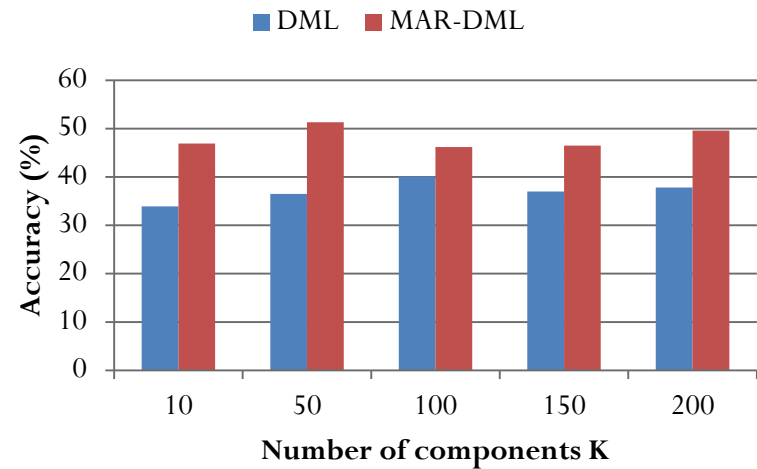
	20-News	15-Scenes	6-Activities
EUC	62.8	65.3	85
DML	76.2	80.8	94.5
LMNN	67	70.3	71.5
ITML	74.7	79.1	94.2
DML-eig	71.2	71.3	86.7
Seraph	75.8	82	89.2
MAR-DML	81.1	83.6	96.2

Clustering Accuracy

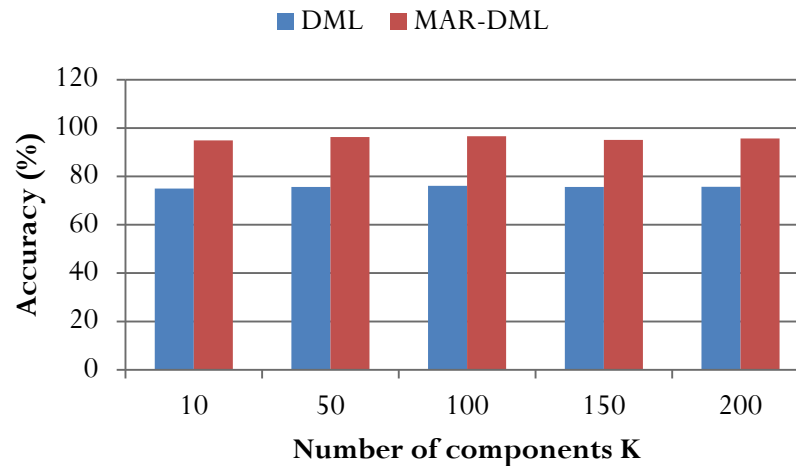
Clustering Accuracy on 20-News



Clustering Accuracy on 15-Scenes



Clustering Accuracy on 6-Activities

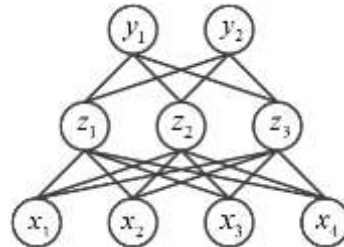


Clustering Accuracy

	20-News	15-Scenes	6-Activities
EUC	36.5	29	61.6
DML	28.4	40.1	76.1
LMNN	32.9	33.6	56.9
ITML	34.5	38.2	93.4
DML-eig	27.3	26.6	63.3
Seraph	48.1	48.2	74.8
MAR-DML	44.6	51.3	96.6

Theoretical Analysis

- Study how the mutual angular regularizer affects generalization error (including estimation error and approximation error) of supervised latent variable models, using the Probably Approximately Correct (PAC) framework
- Use multi-layer perceptron (MLP) as a specific instance
 - MLP is a widely used supervised latent space model



- Rich PAC based analysis for MLP with one hidden layer exists and can be leveraged for our study
- Major results
 - Increasing the mutual angles can reduce estimation error
 - Choosing proper mutual angles can reduce approximation error

Recap of Statistical Learning Theory

- Setup

- Predict an output $y \in \mathcal{Y}$ given an input $x \in \mathcal{X}$
- Let \mathcal{H} be a set of hypotheses
- Let $\ell: (\mathcal{X} \times \mathcal{Y}) \times \mathcal{H} \rightarrow \mathbb{R}$ be a loss function
- Let p^* be the distribution over $\mathcal{X} \times \mathcal{Y}$

- Definitions

- Generalization error $L(h) = \mathbb{E}_{(x,y) \sim p^*} [\ell((x, y), h)]$
- Expected risk minimizer $h^* \in \operatorname{argmin}_{h \in \mathcal{H}} L(h)$
- Empirical risk $\hat{L}(h) = \frac{1}{n} \sum_{i=1}^n \ell((x^{(i)}, y^{(i)}), h)$
- Empirical risk minimizer $\hat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} \hat{L}(h)$

- Generalization error of \hat{h}

$$L(\hat{h}) = \underbrace{L(\hat{h}) - L(h^*)}_{\text{Estimation Error}} + \underbrace{L(h^*)}_{\text{Approximation Error}}$$

- Estimation error is the difference between the generalization error of \hat{h} and h^*
- Approximation error is the best generalization error that can be achieved by the hypotheses set

Setup (Basic Case)

- Task: univariate regression
- Network structure: input layer, one hidden layer, output layer
- Activation function: $h(t)$, Lipschitz continuous with constant L , e.g., sigmoid, tanh, rectified linear.
- Let $\mathbf{x} \in \mathbb{R}^d$ be the input feature vector with $\|\mathbf{x}\|_2 \leq C_1$
- Let y be the response value with $|y| \leq C_2$
- Let $\mathbf{w}_j \in \mathbb{R}^d$ be the weight vector of the j th hidden unit, $j = 1, \dots, m$, with $\|\mathbf{w}_j\|_2 \leq C_3$. Further, we assume the angle $\rho(\mathbf{w}_i, \mathbf{w}_j)$ between \mathbf{w}_i and \mathbf{w}_j is lower bounded by a constant θ for all $i \neq j$.
- Let $\boldsymbol{\alpha} \in \mathbb{R}^m$ be the weights connecting the hidden units to the output with $\|\boldsymbol{\alpha}\|_2 \leq C_4$
- Hypothesis: $f(\mathbf{x}) = \sum_{j=1}^m \alpha_j h(\mathbf{w}_j^T \mathbf{x})$, let \mathcal{F} denote the hypothesis set
- Loss function: $l(\mathbf{x}, y, f) = (f(\mathbf{x}) - y)^2$

Estimation Error of Multi-Layer Perceptron under Mutual Angular Regularizer (MAR-MLP)

Theorem 1 With probability at least $1 - \delta$

$$L(\hat{f}) - L(f^*) \leq 8(\sqrt{J} + C_2)(2LC_1C_3C_4 + C_4|h(0)|) \frac{\sqrt{m}}{\sqrt{n}} + (\sqrt{J} + C_2)^2 \sqrt{\frac{2 \log(2/\delta)}{n}}$$

Where

$$J = mC_4^2h^2(0) + L^2C_1^2C_3^2C_4^2((m-1)\cos\theta + 1) + 2\sqrt{m}C_1C_3C_4^2L|h(0)|\sqrt{(m-1)\cos\theta + 1}$$

Larger angle induces lower estimation error bound

- J is a decreasing function w.r.t θ , hence the estimation error bound decreases as θ increases.
- Increasing the mutual angular regularizer increases the mean and decreases the variance of the pairwise angles, hence increases the lower bound θ of these angles
- The analysis has been extended to
 - Multiple output
 - Multiple hidden layers
 - Other losses: hinge loss, logistic loss, cross entropy loss

Approximation Error of MAR-MLP

- Additional Setup

- Let $G = \{g | g = \alpha h(\mathbf{w}^T \mathbf{x}), \|g\| \leq b_g\}$
- The hypothesis function f_m is constructed iteratively

$$f_1 = g_1$$

$$f_m = \beta f_{m-1} + (1 - \beta)g_m$$

$$0 < \beta \leq c < 1$$

- Let $G_m = \{g | g \in G; \forall j < m, \rho(\mathbf{w}, \mathbf{w}_j) \geq \theta\}$, where w and \mathbf{w}_j are the weight vectors of g and g_j respectively; g_j is the function selected in step j when constructing $f_j = \beta f_{j-1} + (1 - \beta)g_j$
- The target function f satisfies $\|f\| \leq b_f, \langle f, g \rangle < \infty$ for all $g \in G_\theta$
- Approximation error: $\|f_m - f\|^2$

Approximation Error of MAR-MLP

Decreasing w.r.t θ

Non-increasing w.r.t θ

Theorem 2 Let e denote

$$b_g^2 + \frac{1+c}{1-c} b_f^2 + \frac{2c}{1-c} b_g b_f + \frac{2c}{1-c} C_1^2 C_3^2 C_4^2 \cos^2\left(\frac{\theta}{2}\right) V + \frac{2}{1-c} s(\theta)$$

where V is the volume of the input L2 ball and $s(\theta)$ is a non-decreasing function w.r.t θ . Suppose f_1 is chosen to satisfy

$$\|f_1 - f\|^2 \leq \inf_{g \in G} \|g - f\|^2 + \epsilon_1$$

and iteratively f_m is chosen to satisfy

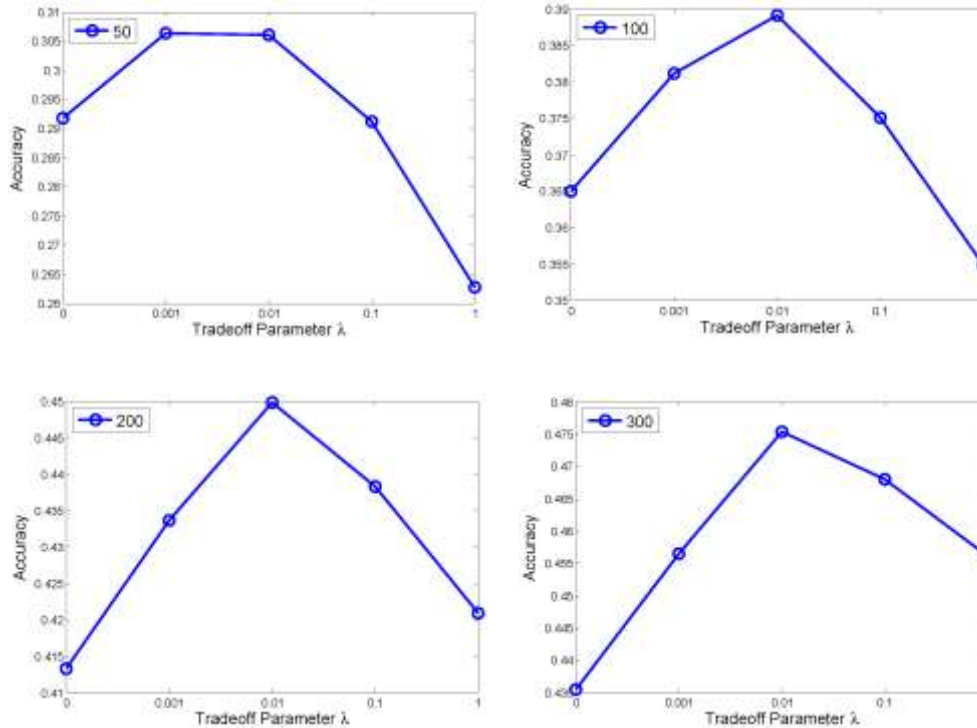
$$\|f_m - f\|^2 \leq \inf_{0 < \beta \leq c} \inf_{g \in G_m} \|\beta f_{m-1} + (1 - \beta)g_m - f\|^2 + \epsilon_m$$

where $\epsilon_m \leq \frac{\rho(\rho+1)e}{m(m+\rho)}$ and ρ is a small constant. Then for every $m \geq 1$,

$$\|f_m - f\|^2 \leq \frac{(\rho + 1)e}{m}$$

- One term in e decreases w.r.t θ , another is non-increasing w.r.t θ , hence a proper chosen θ can reduce approximation error bound

Empirical Corroboration of the Theory



- Classification accuracy versus tradeoff parameter λ in MAR-MLP, on TIMIT speech dataset, under different number of hidden units. Larger λ induces larger angles.
- A proper λ needs to be chosen to achieve the best accuracy, in accordance with the theory that a proper lower bound of the angles yields the lowest generation error

Conclusions

Mutual Angular Regularization (MAR) of Latent Variable Models

- A mutual angle based regularizer
- Capture long-tail factors
- Reduce model complexity while preserving expressiveness

Theory

- Generalization performance of MAR-MLP
- MAR can reduce estimation and approximation errors

Algorithm

- Smooth lower bound of the regularizer
- Optimizing the lower bound with gradient ascent can increase the regularizer in each iteration

Applications

- MAR-RBM and MAR-DML
- Strong empirical performance

Thank you!

Questions?

Papers/slides/code/documents are available at
<http://www.cs.cmu.edu/~pengtaox/projects/dlvm.html>