Similarity Component Analysis

Soravit Changpinyo*, Kuan Liu*, and Fei Sha

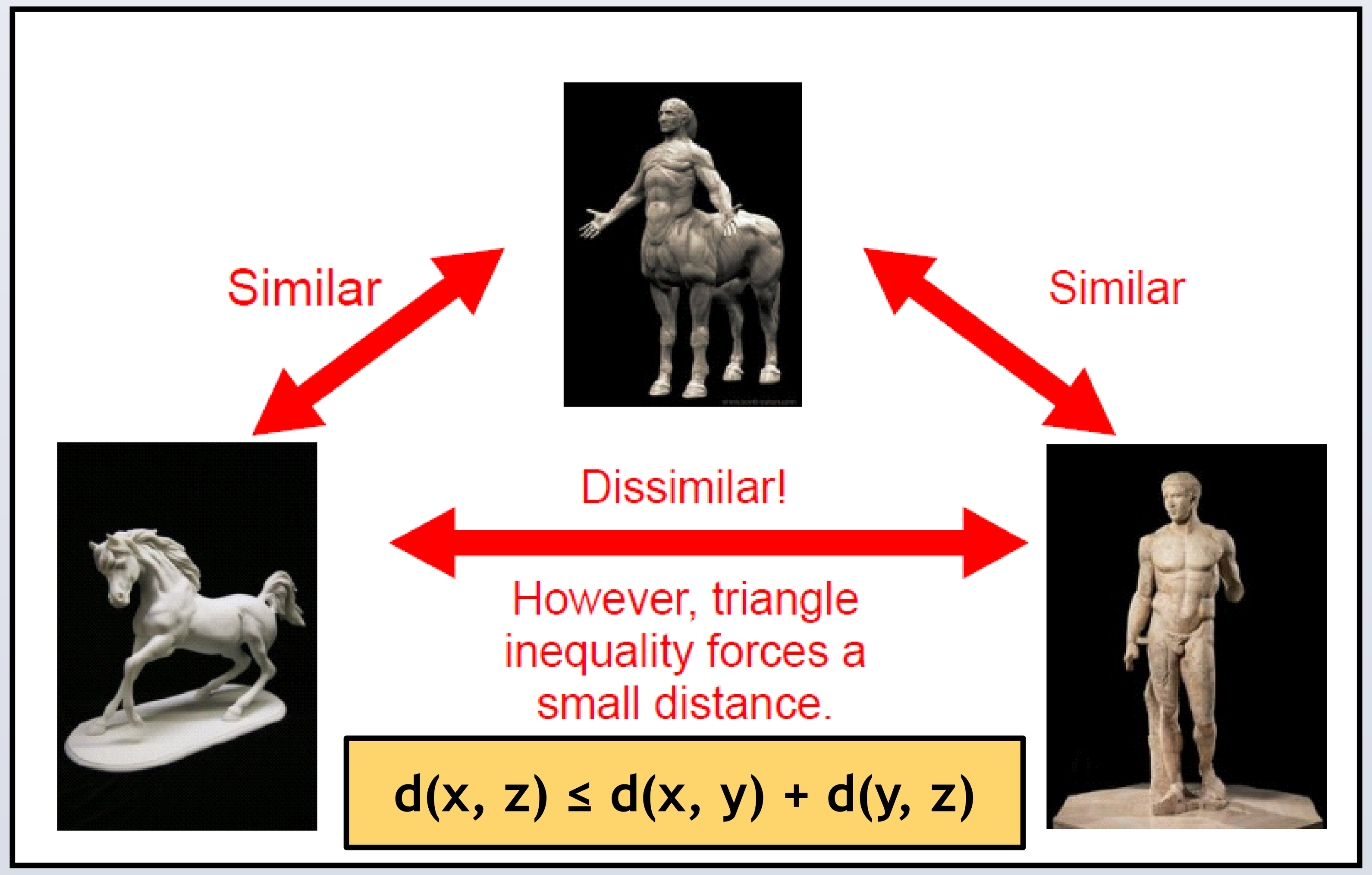
University of Southern California

Highlights

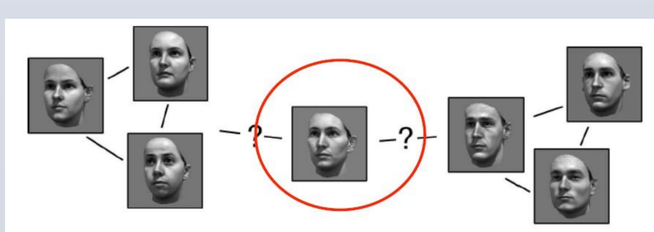
- Learning latent similarities
  - Model *non-metric and noisy* similarity values
  - "Localized" metrics focus on the relevant *subset* of features
- Multiplicative combination of latent components
  - Leads to *tractable* inference
  - Yields *sparse* solutions

Introduction

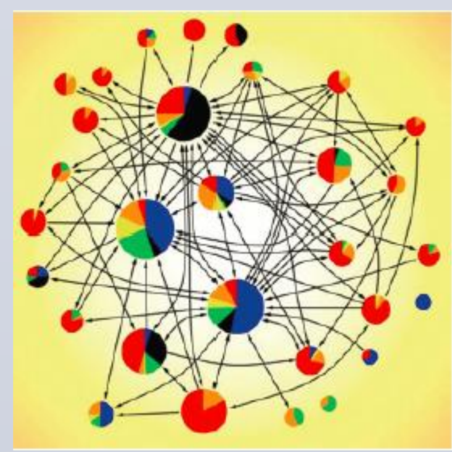
- Metric learning is insufficient for modeling similarity



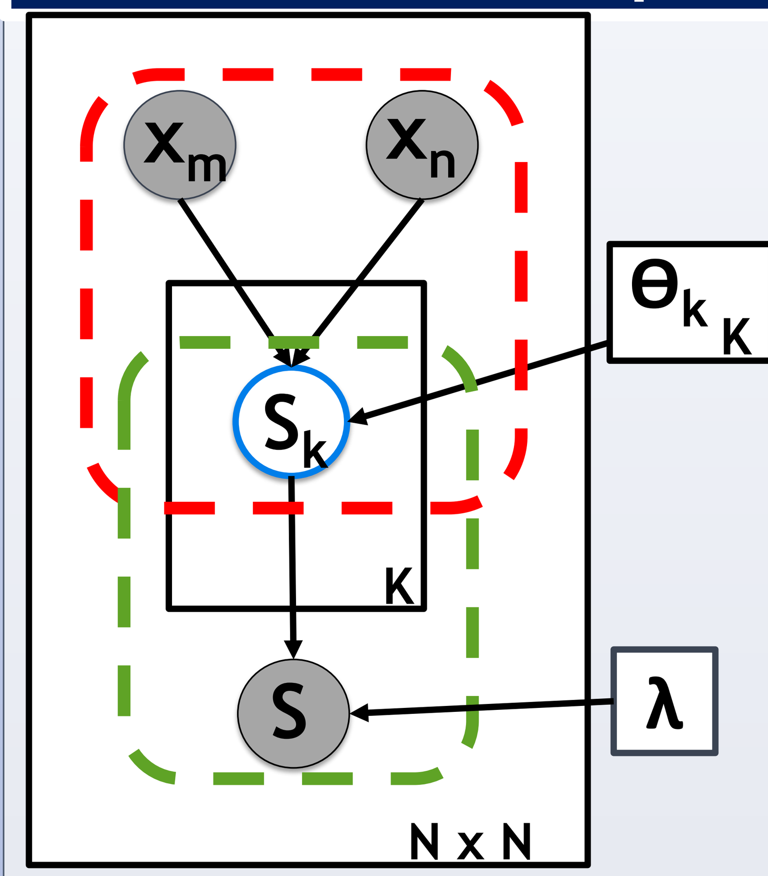
- Non-metric similarity is common
  - Human perception of face [1]



  - "Multiplex" social networks [2]
    - Links are formed for different reasons: same school, religion, zip code, hobbies, political views, etc.



Proposed Approach



SCA

- Observed data (x_m, x_n, s)
- Latent components

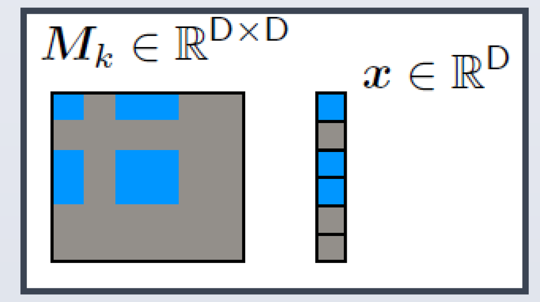
  $s_k \sim P(s_k \mid x_m, x_n; \theta_k)$
- Similarity

  $s \sim P(s \mid s_1, \ldots, s_K; \lambda)$

Latent similarities

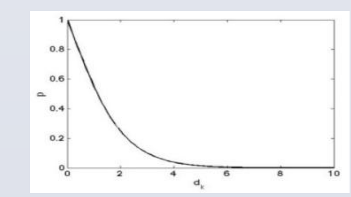
- Latent components
  - Focus on a subset of features

$$d_k = (\boldsymbol{x}_m - \boldsymbol{x}_n)^{\mathrm{T}} \boldsymbol{M}_k (\boldsymbol{x}_m - \boldsymbol{x}_n)$$



- Localized similarity values

$$p(s_k = 1 \mid \boldsymbol{x}_m, \boldsymbol{x}_n) = (1 + e^{-b_k}) \left[1 - \frac{1}{1 + e^{-(d_k - b_k)}} \right]$$



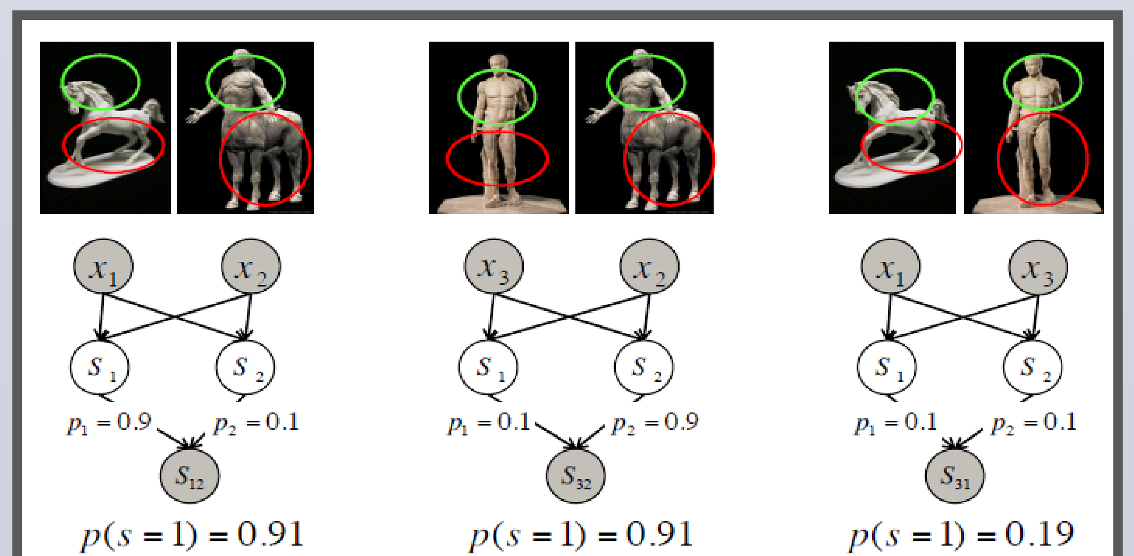
Combining latent components

- Multiplicatively combine with OR gate

$$P(s = 1 \mid s_1, s_2, \cdots, s_K) = 1 - \prod_{k=1}^{K} \mathbb{I}[s_k = 0]$$

Marginalize out all latent components

$$P(s = 1 \mid \boldsymbol{x}_m, \boldsymbol{x}_n) = 1 - \prod_{k} [1 - P(s_k = 1 \mid \boldsymbol{x}_m, \boldsymbol{x}_n)]$$



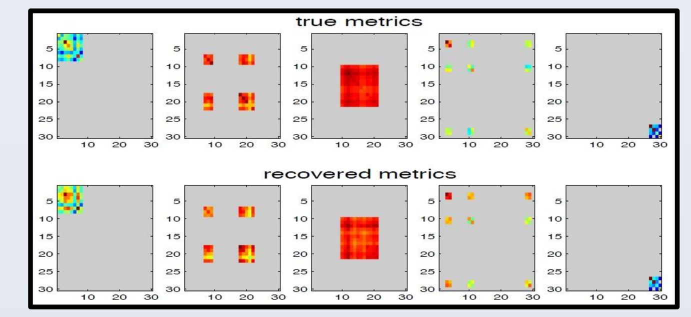
Inference and Learning

- Tractable posterior over latent variables s_k
- EM algorithm
  - Learning each component independently in M step
  - Each component is fit analogously as a softly labeled logistic regression

Experiments

- Synthetic data

True and recovered metrics (K = 5)



Similarity prediction accuracies

Table 1: Similarity prediction accuracies and standard errors (%) on the synthetic dataset

BASELINES		SCA					
ITML	LMNN	K = 1	K = 3	K = 5	K = 7	K = 10	K = 20
72.7±0.0	71.3±0.2	72.8±0.0	82.1±0.1	**91.5±0.1**	**91.7±0.1**	**91.8±0.1**	90.2±0.4

- Link prediction on a network of NIPS proceedings

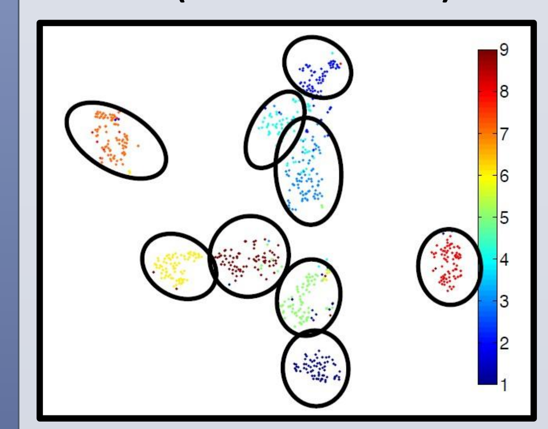
Compare to discriminative methods (SVM, LMNN[3], ITML[4]) for different features and K.

Link prediction accuracies and their standard errors (%) on a network of scientific papers

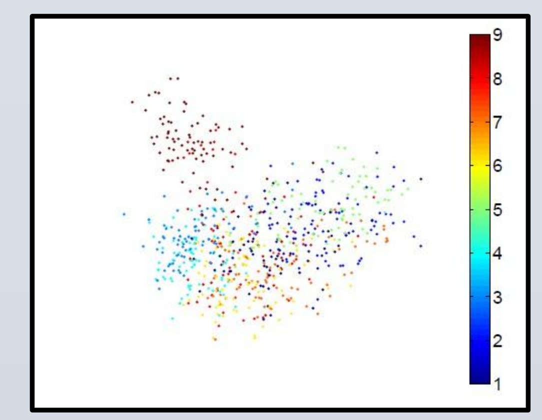
Feature type	BASELINES			SCA-DIAG		SCA	
	SVM	ITML	LMNN	K = 1	K*	K = 1	K*
BoW	73.3±0.0	-	-	64.8 ± 0.1	**87.0 ± 1.2**	-	-
ToW	75.3±0.0	-	-	67.0 ± 0.0	**88.1 ± 1.4**	-	-
ToP	71.2±0.0	81.1±0.1	80.7±0.1	62.6 ± 0.0	81.0 ± 0.8	81.0 ± 0.0	**87.6 ± 1.0**

2-D embedding reveals network structure

Links (local similarities)

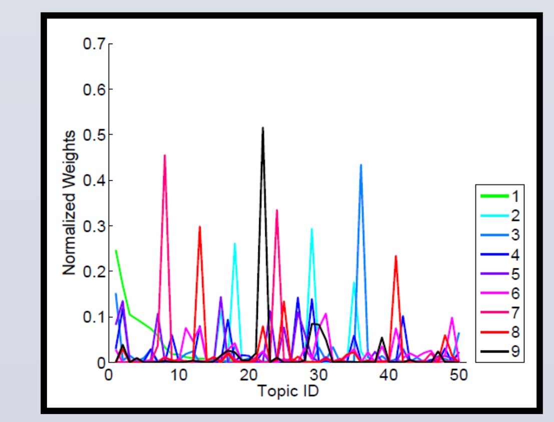


Nodes (original ToP features)



Sparse & disjoint features

Diagonal values of metrics (K = 9)



References

[1] J. Laub, J. Macke, K. R. Müller, and F. Wichmann. *Inducing Metric Violations in Human Similarity Judgements*. NIPS, 2006.

[2] S. E. Fienberg, M. M. Meyer, and S. S.Wasserman. *Statistical Analysis of Multiple Sociometric Relations*. JASA, 1985.

[3] K. Q. Weinberger and L. K. Saul. *Distance Metric Learning for Large Margin Nearest Neighbor Classification*. JMLR, 2009.

[4] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. *Information-theoretic Metric Learning*. ICML, 2007.