Partial order embedding with multiple kernels

Brian McFee and Gert Lanckriet

University of California, San Diego

computer
audition
laboratory

UCSD Jacobs | School of Engineering

Goal

Embed a set of objects into a Euclidean space such that:

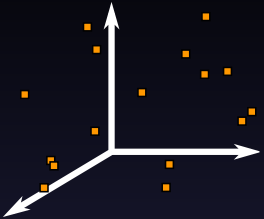
1. Distances conform to human perception
2. Multiple feature modalities are integrated coherently
3. We can extend to unseen data

Motivation: leverage existing technologies for Euclidean data

Example

Musicians

Target space

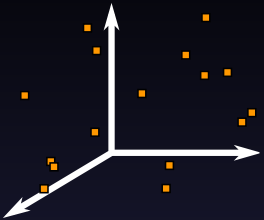


Example

Musicians

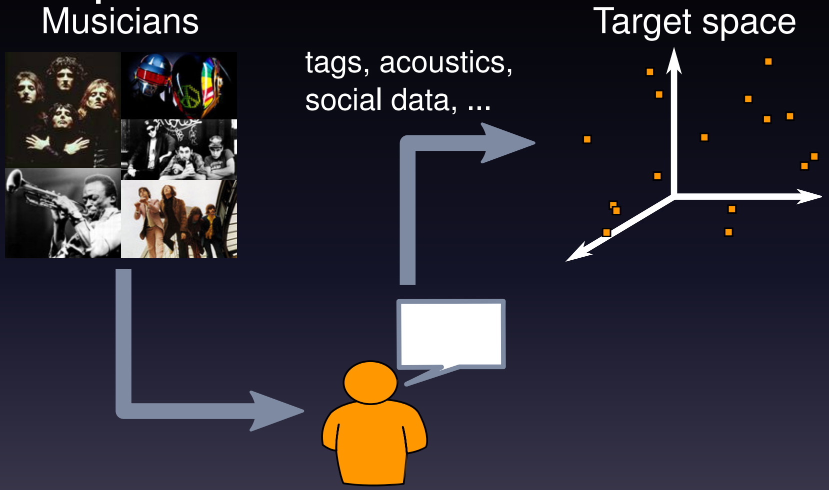
tags, acoustics, social data, ...

Target space



- Features may not match human perception

Example



- Features may not match human perception
- Use human input to guide the embedding

Human input

- Binary similarity can be *ambiguous* in multi-media data
- Example:

Is Oasis similar to The Beatles, or not?

- *Quantifying* similarity may also be difficult... how similar are they?

Relative comparisons

[Schultz and Joachims, 2004, Agarwal et al., 2007]

- Instead, we ask which of two *pairs* is more similar:

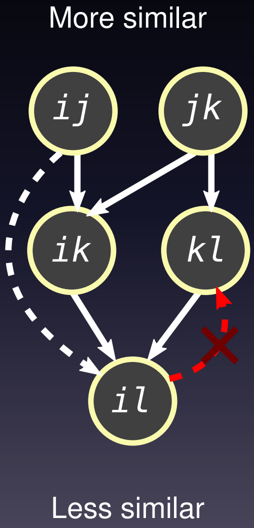
$$(i, j) \text{ or } (k, \ell)?$$

(Oasis, Beatles, Oasis, Metallica)

- Learn a map g from the data set $\mathcal{X}$ to a Euclidean space
- For each (i, j, k, ℓ),

$$\|g(i) - g(j)\| \;<\; \|g(k) - g(\ell)\|$$

Partial order

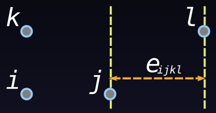


- Relative comparisons should exhibit *global structure*.
- Collect comparisons into a directed graph $\mathcal{C}$
- *Cycles* must be broken by any embedding
  - Comparisons should describe a partial order over $\mathcal{X} \times \mathcal{X}$.

Constraint graphs

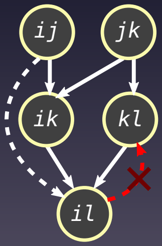
- Force margins between distances:

$$\|g(i) - g(j)\|^2 + e_{ijk\ell} \leq \|g(k) - g(\ell)\|^2$$

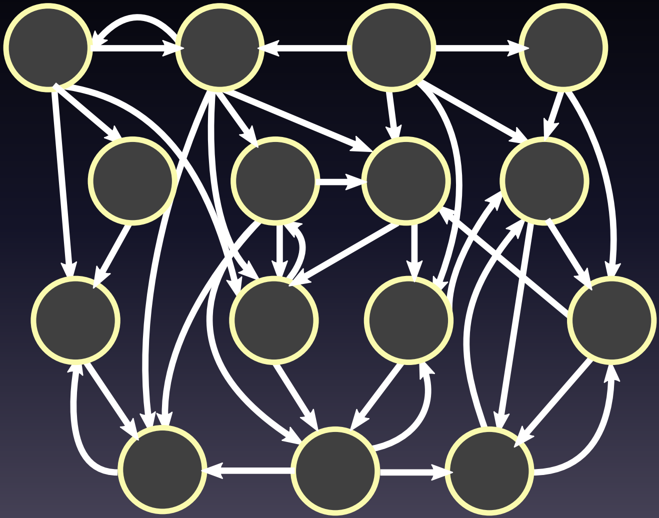


- Represent $e_{ijk\ell}$ as edge weights

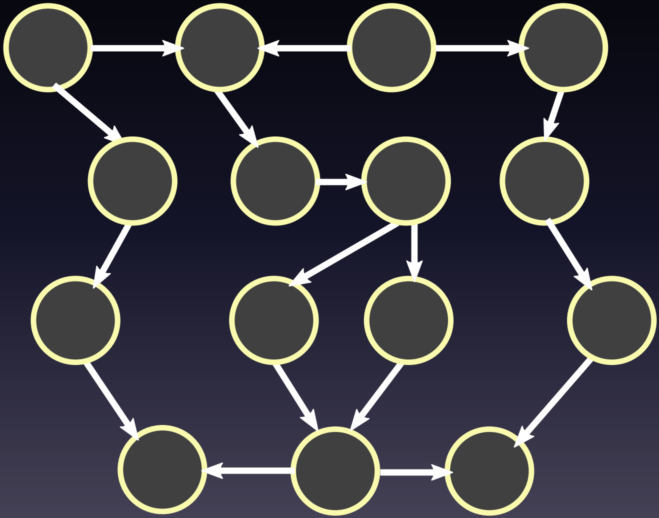
- Graph representation lets us
  - detect inconsistencies (*cycles*)
  - prune redundancies by *transitive reduction*
  - simplify: focus on meaningful constraints



Constraint simplification



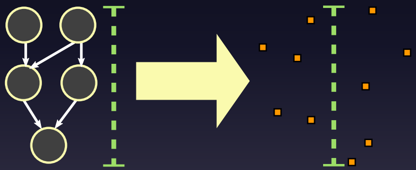
Constraint simplification



Margin-preserving embeddings

- *Claim*: There exists $g : \mathcal{X} \to \mathbb{R}^{n-1}$ such that
  all margins are preserved, and for all $i \neq j$:

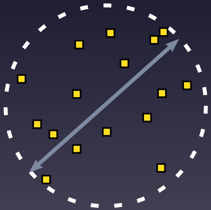
$$1 \leq \|g(i) - g(j)\| \leq \sqrt{(4n+1)(\mathrm{diam}(\mathcal{C})+1)}$$



- Reduction via constant-shift embedding [Roth et al., 2003]
- *Constraint diameter* bounds *embedding diameter*
- May produce artificially *high-dimensional* embeddings

Dimensionality reduction

- We show that it's *NP-hard* to minimize dimensionality for POE
- Instead, optimize a convex objective that prefers low-dimensional solutions
- Assume objects are dissimilar, unless otherwise informed
- Adapt MVU [Weinberger et al., 2004]:
    - Maximize all distances
    - Diameter bound ensures that a solution exists
    - Respect all partial order constraints



Partial Order Embedding (SDP)

- Input: n objects $\mathcal{X}$, margin-weighted constraints $\mathcal{C}$
- Output: $g : \mathcal{X} \to \mathbb{R}^n$

$$
\begin{array}{lll}
\max_{A \succeq 0} & \mathrm{Tr}(A) & \text{(Variance)} \\
& d(i,j) \leq O(n \cdot \mathrm{diam}(\mathcal{C})) & \text{(Diameter)} \\
& d(i,j) + e_{ijk\ell} \leq d(k,\ell) & \text{(Margins)} \\
& \sum_{i,j} A_{ij} = 0 & \text{(Centering)} \\
& d(i,j) \doteq A_{ii} + A_{jj} - 2A_{ij} & \text{(Distance}^2\text{)}
\end{array}
$$

- Decompose $A = V \Lambda V^{\mathsf{T}} \quad \Rightarrow \quad g(i) = \left(\Lambda^{1/2} V^{\mathsf{T}}\right)_i$

Out-of-sample extension: kernels

- How can we extend embeddings to unseen data?
- Learn a linear projection from a feature space

- Parameterization:

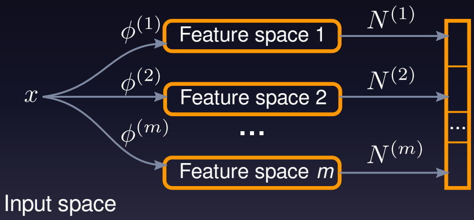
$$g(x) = NK_x \qquad (K_x = x \text{ column of } K)$$

- Learn N by solving an SDP over $W = N^{\mathsf{T}}N \succeq 0$

- PO constraints may be impossible to satisfy:
    - Soften ordering constraints

Multi-kernel embedding

- Concatenate linear projections from *m* feature spaces:



$$g(x) = \begin{bmatrix} N^{(1)} K_x^{(1)} \\ N^{(2)} K_x^{(2)} \\ \dots \\ N^{(m)} K_x^{(m)} \end{bmatrix}$$

Output space

- $N^{(\cdot)}$s are jointly optimized by SDP to form the space

MK-POE

$$\max_{W \succeq 0, \xi \geq 0} \quad \sum_{p=1}^{m} \mathrm{Tr}\left(K^{(p)} W^{(p)} K^{(p)}\right) - \gamma \,\mathrm{Tr}\left(W^{(p)} K^{(p)}\right) - \beta \sum_{\mathcal{C}} \xi_{ijk\ell}$$

s. t.

$$\forall i, j \in \mathcal{X} \qquad d(i,j) \leq O(n \cdot \mathrm{diam}(\mathcal{C}))$$

$$\forall (i,j,k,\ell) \in \mathcal{C} \qquad d(i,j) + e_{ijk\ell} \leq d(k,\ell) + \xi_{ijk\ell}$$

$$d(i,j) \doteq \sum_{p=1}^{m} \left(K_i^{(p)} - K_j^{(p)}\right)^{\mathsf{T}} W^{(p)} \left(K_i^{(p)} - K_j^{(p)}\right)$$

Experiment 1: Human perception

Data [Agarwal et al., 2007]

- 55 images of 3D rabbits with varying surface reflectance
- 13049 human perception measurements: (i, j, i, k)

Constraint processing

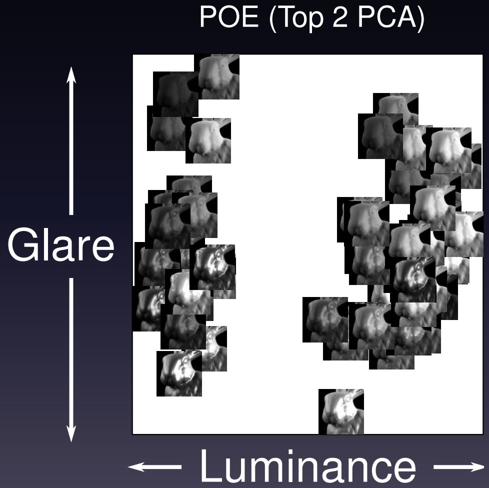
- Random sampling to achieve a *maximal DAG*
- *Transitive reduction* to eliminate redundancies

$13000 \rightarrow 9000$ constraints

Final constraint graph

- Unit margins
- Diameter $= 55$

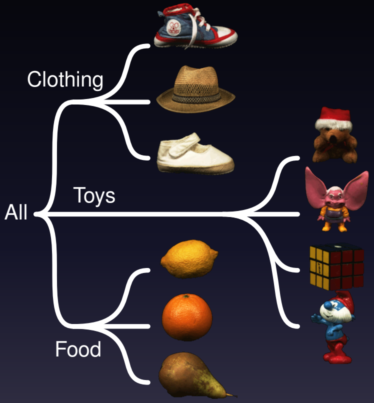
Experiment 1 results



Experiment 2: Multi-kernel

Data [Geusebroek et al., 2005]

- 10 classes from ALOI
- 10 images from each class, varying out-of-plane rotation
- Constraints generated by a label taxonomy



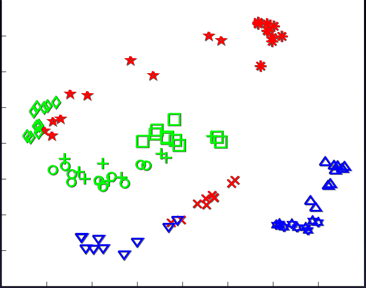
Kernels

- Grayscale dot product
- RBF of R,G,B, and grayscale histograms

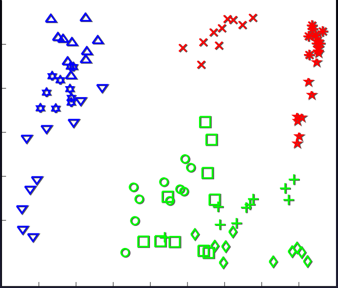
Diagonally-constrained *N*: SDP⇒LP

Experiment 2 results

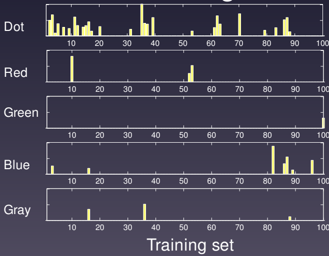
Sum-kernel space



Learned embedding



Learned weights



Experiment 2 kernel comparison

% Constraints satisfied

Kernel	Native	Optimized
Dot product	0.83	0.85
Red	0.63	0.63
Green	0.65	0.67
Blue	0.77	0.83
Gray	0.68	0.69
Unweighted sum	0.76	0.77
Multi	—	0.95

Experiment 3: Out-of-sample

Goal

- Predict comparsions (i, j, i, k) with i out of sample

Data

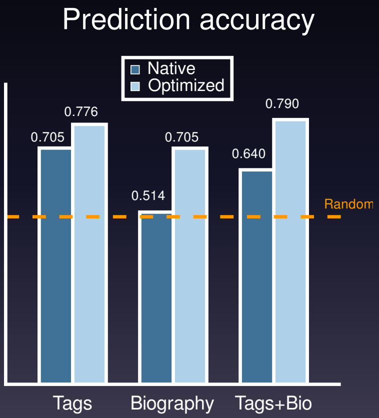
- 412 popular artists (*aset400*)
  [Ellis et al., 2002]
- 10-fold cross-validation
- $\approx$6300 human-derived training constraints
- Mean diameter $\approx$30 (over CV folds)

Features: TFIDF/cosine kernels

- *Tags*: 7737 words
  (e.g., *rock, piano, female vocals*)
- *Biographies*: 16753 words

last.fm

Experiment 3 results



Note: test comparisons are not internally consistent

Conclusion

- We developed the partial order embedding framework
  - Simplifies relative comparison embeddings
  - Enables more careful constraint processing
    - Graph manipulations can increase embedding robustness
- Derived a novel multiple kernel learning technique
  - Widely applicable to metric learning problems

Thanks!

Questions?

- Agarwal, S., Wills, J., Cayton, L., Lanckriet, G., Kriegman, D., and Belongie, S. (2007).
  Generalized non-metric multi-dimensional scaling.
  In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*.
- Ellis, D., Whitman, B., Berenzweig, A., and Lawrence, S. (2002).
  The quest for ground truth in musical artist similarity.
  In *Proeedings of the International Symposium on Music Information Retrieval (ISMIR)*, pages 170–177.
- Geusebroek, J. M., Burghouts, G. J., and Smeulders, A. W. M. (2005).
  The Amsterdam library of object images.
  Int. J. Comput. Vis., 61(1):103–112.
- Roth, V., Laub, J., Buhmann, J. M., and Müller, K.-R. (2003).
  Going metric: denoising pairwise data.

In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in Neural Information Processing Systems 15*, pages 809–816, Cambridge, MA. MIT Press.

- Schultz, M. and Joachims, T. (2004).
  Learning a distance metric from relative comparisons.
  In Thrun, S., Saul, L., and Schölkopf, B., editors, *Advances in Neural Information Processing Systems 16*, Cambridge, MA. MIT Press.

- Weinberger, K. Q., Sha, F., and Saul, L. K. (2004).
  Learning a kernel matrix for nonlinear dimensionality reduction.
  In *Proceedings of the Twenty-first International Conference on Machine Learning*, pages 839–846.