Partial order embedding with multiple kernels

computer audition laboratory

Brian McFee and Gert Lanckriet
University of California, San Diego

UCSD Jacobs School of Engineering

Introduction

Goal Embed arbitrary objects into a Euclidean space so that pairwise distances correspond to human perception of similarity.

Setup
- Binary similarity measurements may be too coarse for many applications.
- Quantitative similarity varies significantly between people.
- We focus on **relative comparisons**:

$$(i,j,k,\ell) \Leftrightarrow d(i,j) < d(k,\ell)$$

indicates that i and j are **more similar** than k and ℓ.

- Relative comparisons can be combined to form a **directed graph** over pairs of objects.
- Analyzing this graph helps to simplify the embedding procedure.
- Our embedding method integrates **multiple kernels** to form a **unifed space.**

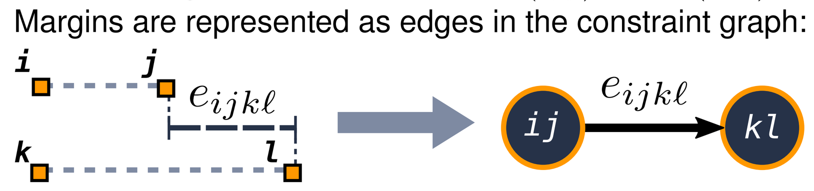
Partial order

Notation

$\mathcal{X}$: Arbitrary set of objects. Not necessarily vectors.

$\mathcal{C}$: Collection of relative comparisons: **(i,j,k,l)**

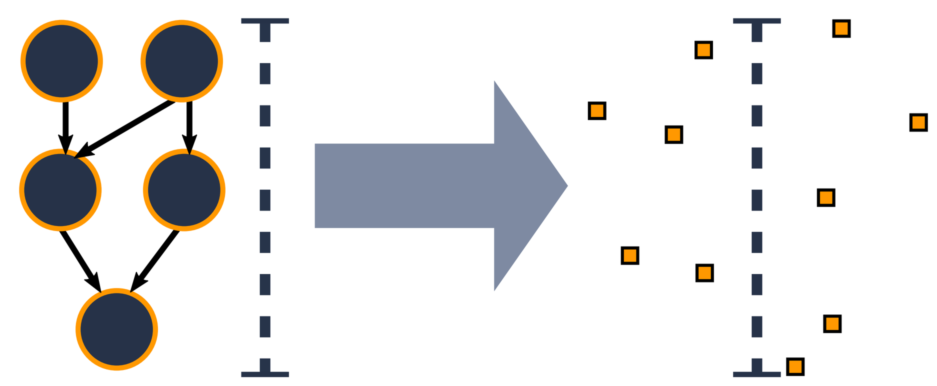
$e_{ijk\ell}$: Desired margin between distances $d(i,j)$ and $d(k,\ell)$. Margins are represented as edges in the constraint graph:



Method
- Process $\mathcal{C}$ to remove **inconsistent** and **redundant** edges
- This results in a **minimal** DAG representation
- The **diameter** of the DAG dictates the diameter of the embedding:

Claim For any margin-weighted constraint DAG, there exists an embedding $g : \mathcal{X} \rightarrow \mathbb{R}^{n-1}$ that satisfies all margins, and for all $i \neq j$:

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

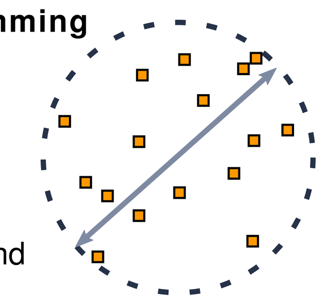


Idea
- Reduce to non-metric MDS with constant-shift embedding [1].

But...
- How should **unconstrained** distances be computed?
- How should we extend to **unseen data**?
- How should we integrate **multiple feature modalities**?

Algorithm

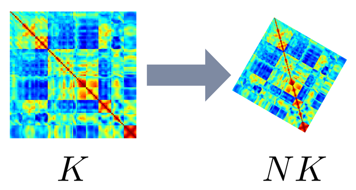
SDP We learn the embedding by **semidefinite programming** over the inner-product matrix of embedded points.



Distances are optimized by **maximum variance unfolding** [2].

All **margins are preserved**, and the diameter bound ensures that an optimum exists.

Out-of-sample To extend the embedding to **unseen data**, we learn a linear projection matrix N.



This generalizes to **non-linear** transformations by using kernels:

$$g(x) = NK_x$$

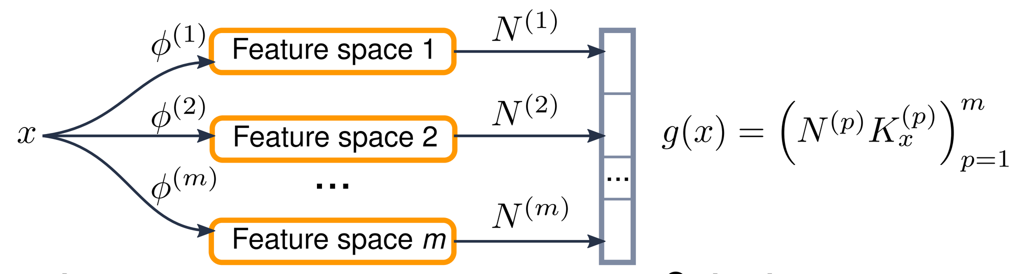
K_x is the column vector containing the kernel function evaluated at x and the training set.

We solve for g by optimizing over a PSD matrix $W = N^{\mathsf{T}}N$.

Adding $-\gamma\,\mathrm{Tr}(WK)$ to the objective controls sparsity and lets us invoke the generalized representer theorem [3].

Constraints are softened and violations are penalized by hinge loss.

Multi-kernel Multiple kernels are combined by learning a separate transformation matrix for each kernel, and **concatenating** the outputs:



Transformation matrices are **jointly optimized** to produce a unified space which optimally integrates all features.

LP **Diagonal constraints** on $W^{(p)}$ simplifies the optimization from SDP to linear programming:

$$W^{(p)} \succeq 0 \quad \Rightarrow \quad W_{ii}^{(p)} \geq 0.$$

Diagonal weights can be interpreted as measuring the utility of a (kernel, training point) pair in defining the embedding.

Multi-Kernel POE

$$\max_{W^{(p)},\xi} \quad \sum_{i,j} d(i,j) - \beta \sum_{\mathcal{C}} \xi_{ijk\ell} - \gamma \sum_{p} \mathrm{Tr}\left(W^{(p)}K^{(p)}\right)$$

$$\forall i,j \in \mathcal{X} \qquad (4n+1)(\mathrm{diam}(\mathcal{C})+1) \geq d(i,j)$$

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

$$\xi_{ijk\ell} \geq 0$$

$$\forall p \in 1,2,\ldots,m \qquad W^{(p)} \succeq 0$$

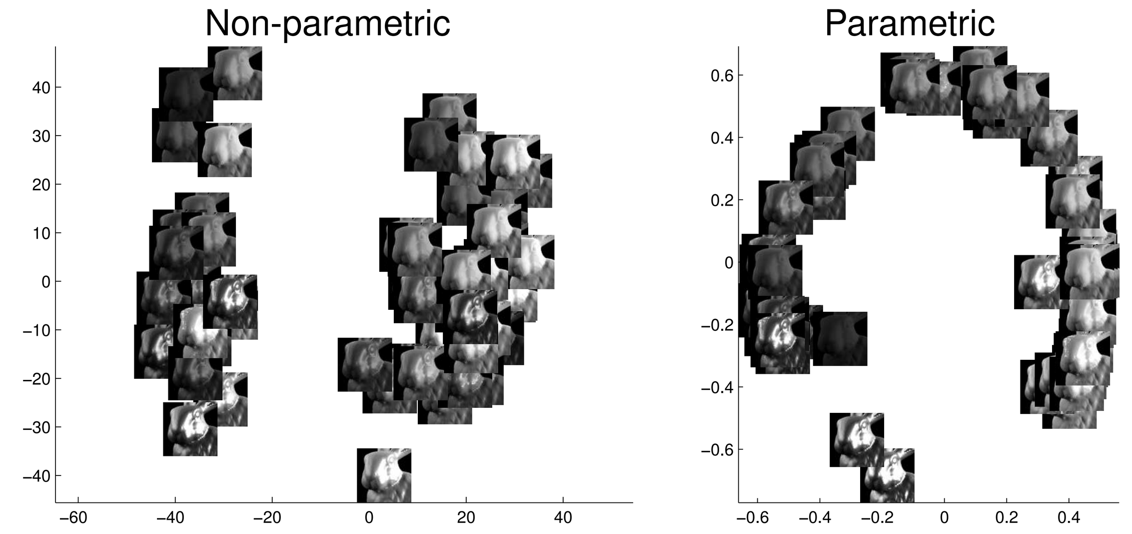
$$d(i,j) \doteq \sum_{p=1}^{m} \left(K_i^{(p)} - K_j^{(p)}\right)^{\mathsf{T}} W^{(p)} \left(K_i^{(p)} - K_j^{(p)}\right)$$

Experiments

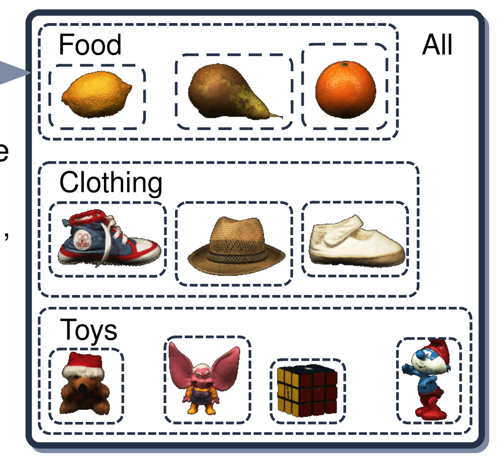
Data
- 55 images of 3D-rendered rabbits with varying surface reflectance
- 13049 relative comparisons obtained from human test subjects [4]

Setup
- 8765 edges remain in the constraint graph after pruning
- Each pair receives unit margins
- Constraint graph has diameter 55

Left: Non-parametric embedding (SDP)
Right: Out-of-sample extension via radial-basis kernels over intensity histograms. Learned metric is a diagonal matrix (LP).



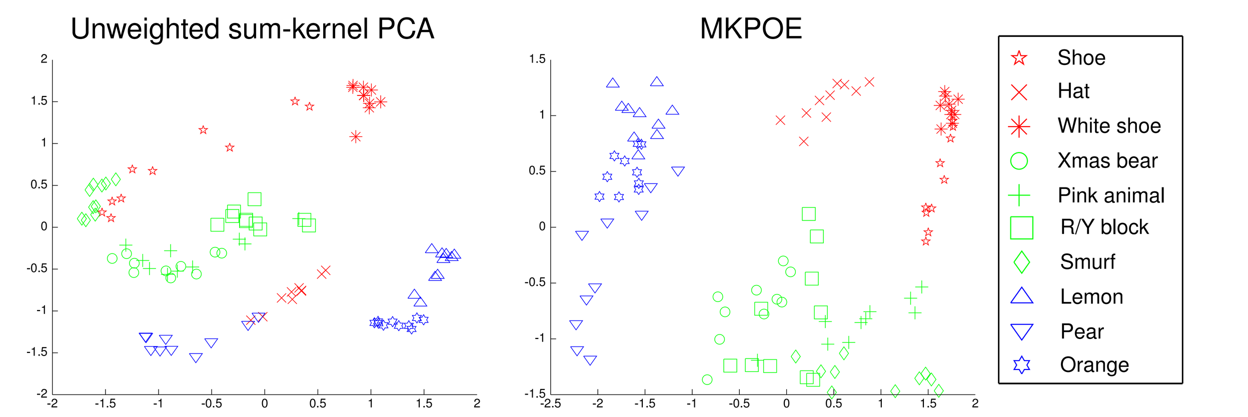
Data
- 100 images from the Amsterdam Library of Object Images (ALOI), varying out-of-plane rotation [5]
- A taxonomy of 10 object classes



Setup
- Similarity constraints were generated according to a bottom-up traversal of the taxonomy tree, e.g.,
  (Orange, Orange, Orange, Lemon),
  (Orange, Lemon, Orange, Smurf).
- Margins shrink exponentially with depth in the taxonomy.

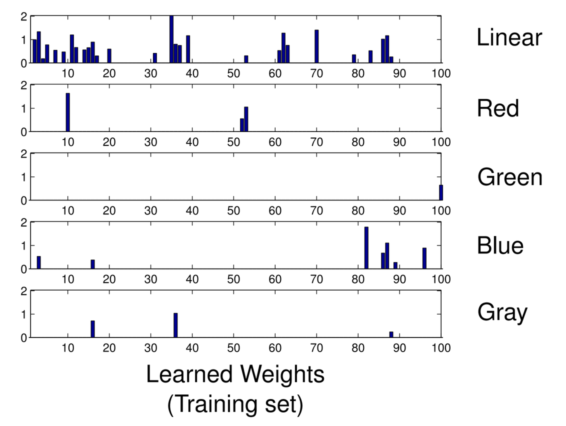
Kernels
- RBF kernels over color histograms in RGB channels, and grayscale intensity
- Linear kernel over grayscale values to capture shape information

The embedding was formed by learning diagonally-constrained matrices, the diagonals are shown below.

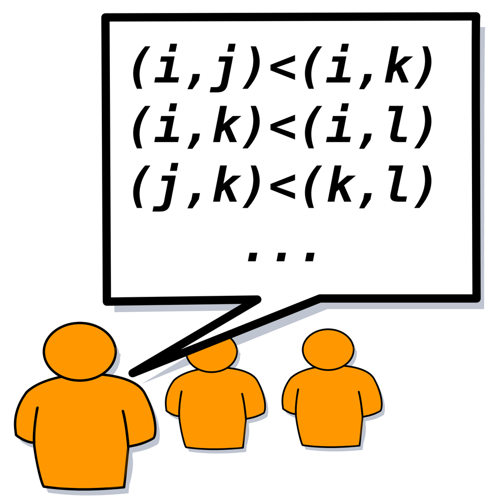


Kernel	Native	Learned
Linear	0.8255	0.8526
RBF Red	0.6303	0.6339
RBF Green	0.6518	0.6746
RBF Blue	0.7717	0.8349
RBF Gray	0.6782	0.6927
Unweighted Sum	0.7628	0.7674
MKPOE		0.9483

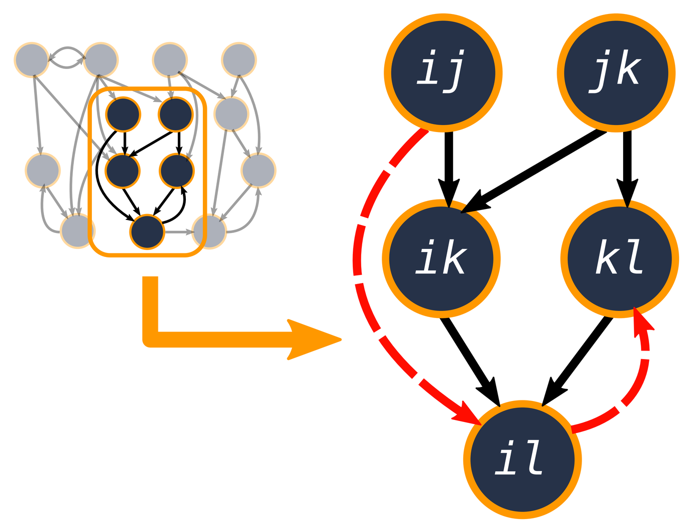
Constraint Satisfaction



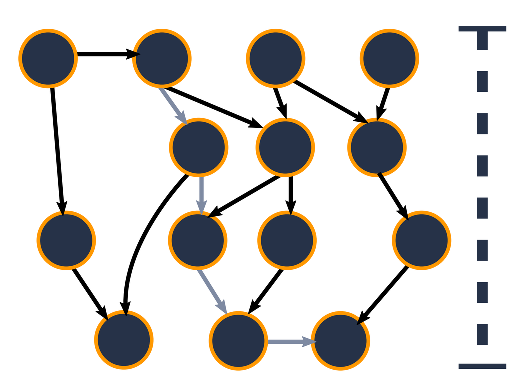
Learned Weights (Training set)



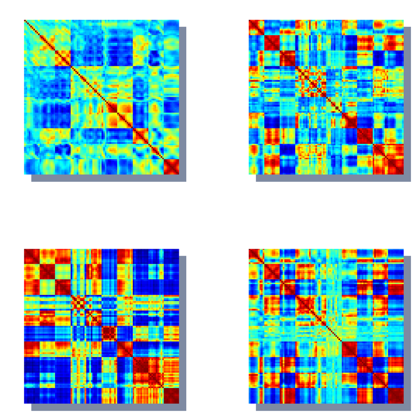
1. Query for similarity measurements.



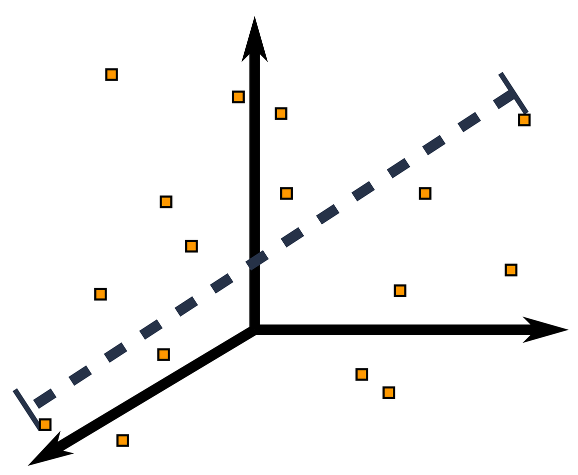
2. Build the constraint graph. Prune inconsistent and redundant edges.



3. Measure the diameter of the constraint DAG. Edge weights define margins between pairwise distances.



4. Compute kernels.



5. Transform kernels into a Euclidean space which conforms to human perception.

References

[1] Roth, V., Laub, J., Buhmann, J.M., and Mueller, K.R. (2003). Going metric: denoising pairwise data. *Advances in Neural Information Processing Systems 15* (pp. 809-816). Cambridge, MA: MIT Press.

[2] Weinberger, K.Q., Blitzer, J., and Saul, L.K. (2006). Learning a kernel matrix for nonlinear dimensionality reduction. *Proceedings of the Twenty-first Internatoinal Conference on Machine Learning* (pp. 839-846).

[3] Scholkopf, B., Herbrich, R., Smola, A.J., and Williamson, R. (2001). A generalized representer theorem. *Proceedings of the 14th Annual Conference on Computational Learning Theory* (pp. 416-426).

[4] Agarwal, S., Wills, J., Cayton, L., Lanckriet, G., Kriegman, D., and Belongie, S. (2007). Generalized non-metric multi-dimensional scaling. *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics.*

[5] Geusebroek, J.M., Burghouts, G.J., and Smeulders, A.W.M. (2005). The Amsterdam library of object images. *International Journal of Computer Vision*, 61, 103-112.