Hierarchical Optimal Transport for Multimodal Distribution Alignment

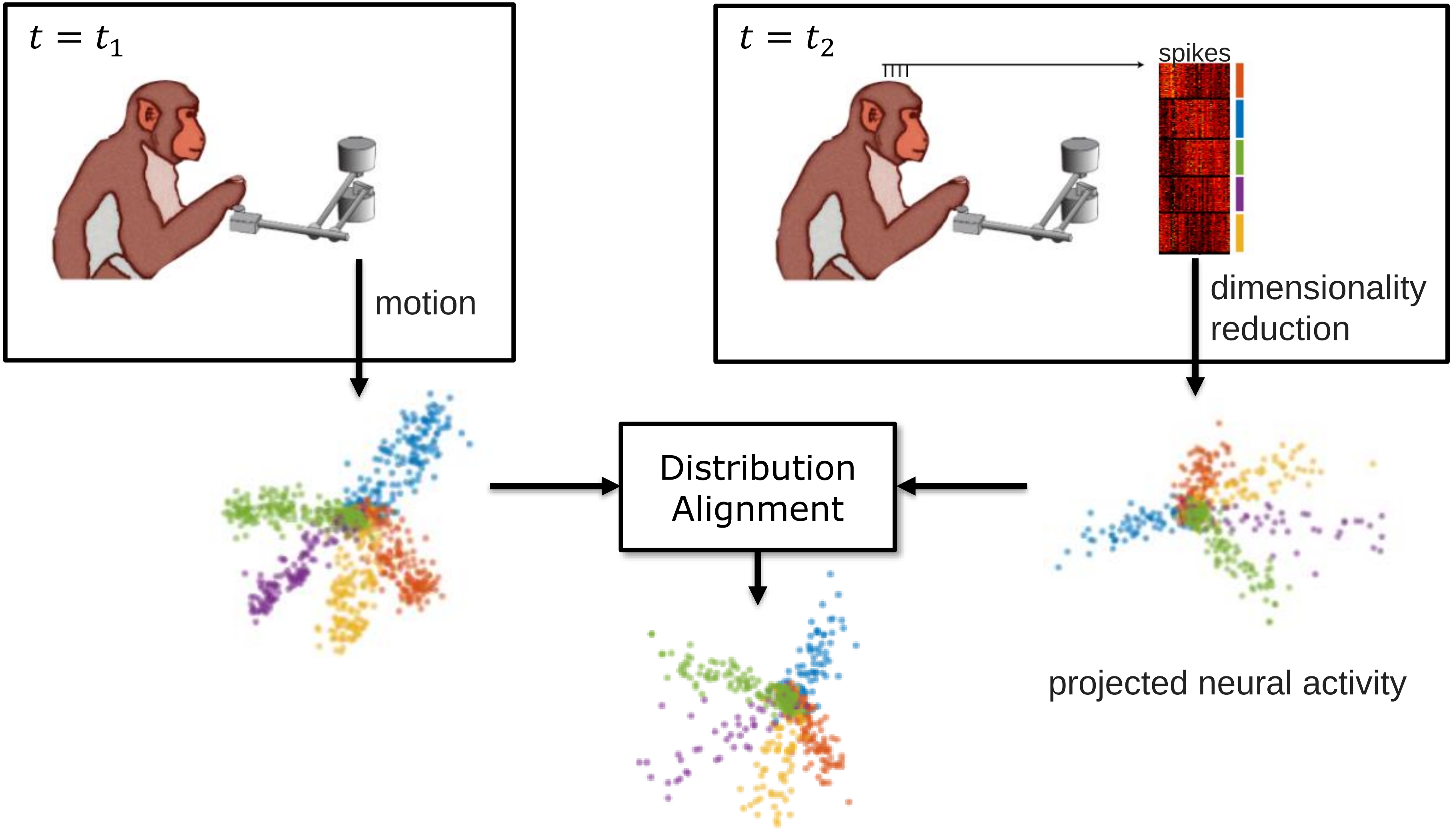
John Lee, Max Dabagia, Eva Dyer and Christopher J. Rozell

Georgia Institute of Technology

ML@GT
Machine Learning at Georgia Tech

Networks: Circuits, Cells, & Genes
Georgia Tech Neuro
Technology: Computation, Devices, & Molecular Tools
Systems: Cognition, Sensing, & Movement

Motivating example: movement decoding

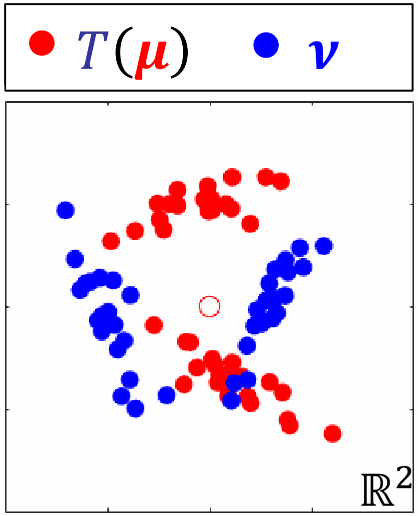


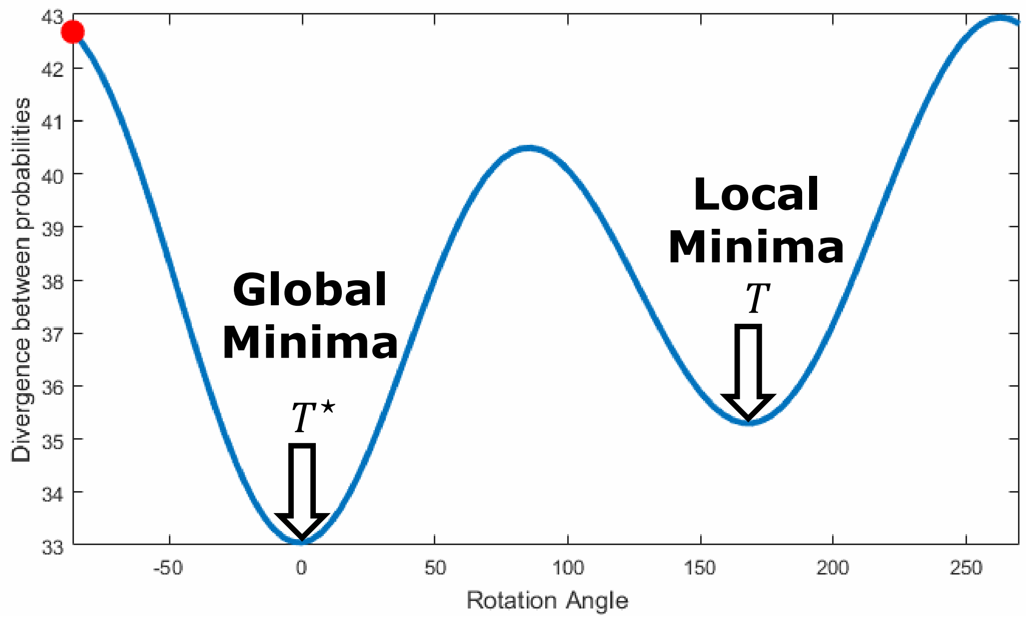
(Dyer et al. 2017)

Distribution alignment

Given point clouds $\boldsymbol{\nu} = \sum_{l=1}^{n_y} \frac{1}{n_y} \delta_{\boldsymbol{y}_l}$, $T(\boldsymbol{\mu}) = \sum_{k=1}^{n_x} \frac{1}{n_x} \delta_{T(\boldsymbol{x}_k)}$

Goal: $\boxed{\min_{T \in \mathcal{T}} D(T(\boldsymbol{\mu}), \boldsymbol{\nu})}$





Problem statement

Given point clouds $\boldsymbol{\nu} = \sum_{l=1}^{n_y} \frac{1}{n_y} \delta_{\boldsymbol{y}_l}$, $T(\boldsymbol{\mu}) = \sum_{k=1}^{n_x} \frac{1}{n_x} \delta_{T(\boldsymbol{x}_k)}$

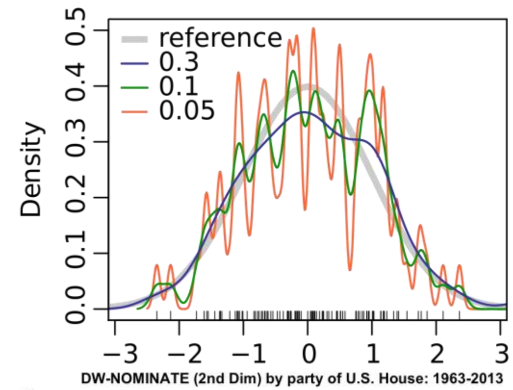
Goal: $\boxed{\min_{T \in \mathcal{T}} D(T(\boldsymbol{\mu}), \boldsymbol{\nu})}$

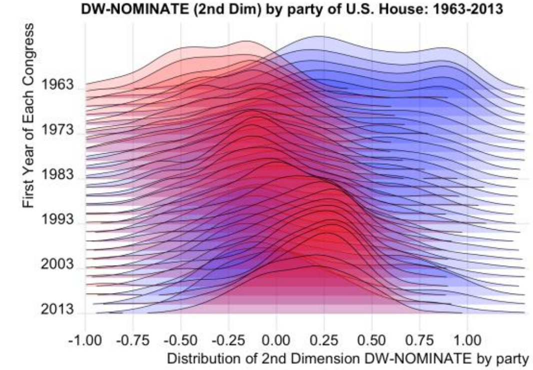
We need:

1. A **divergence** between the point clouds that *exploits* geometry
2. A **transformation** that *preserves* geometry
3. An approach that is computationally **tractable** with **little (or no) supervision**

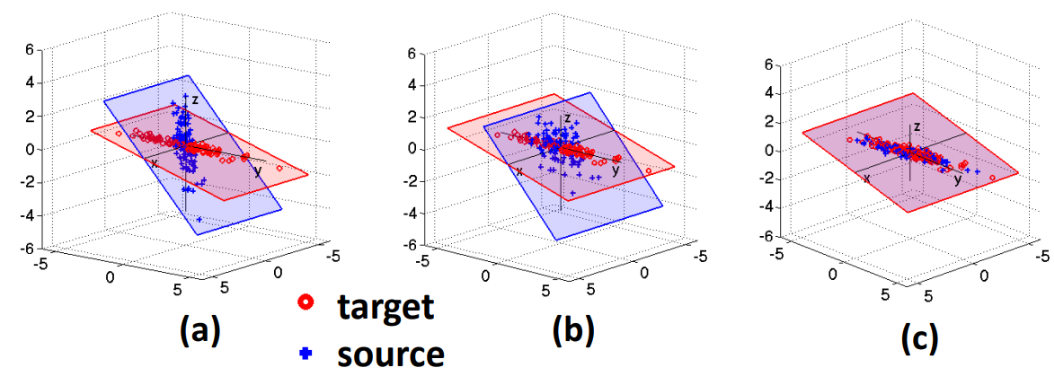
Pointwise divergences

- Kullback-Leibler (KL)
  - Requires kernel estimation that implies dataset geometry
  - Supports need to overlap
  - Brute force in low-D (Dyer et al. 2017)

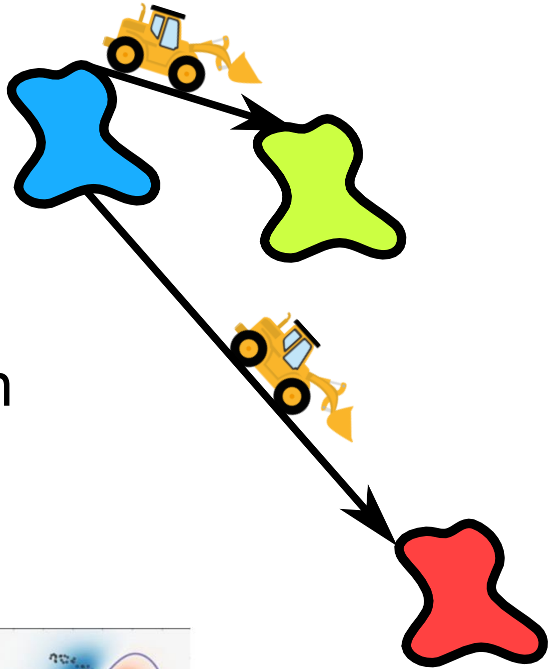




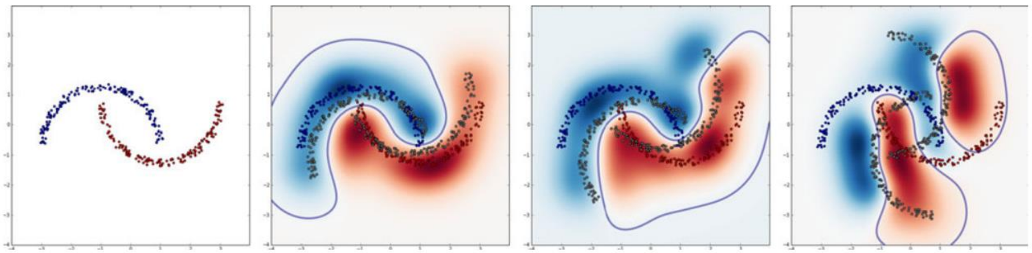
- ℓ_p-norms
  - Sign invariant subspace structure (Fernando et al. 2013)
  - Symmetric covariance structure (Sun et al. 2016)



Today’s focus: optimal transport (OT)



- Classic optimization based on transportation of mass
  - Wasserstein distance (Euclidean)
  - See also: Earth Mover’s Distance
- Used previously for domain adaptation (Courty et al. 2016)
  - Assumes initial alignment is “close”

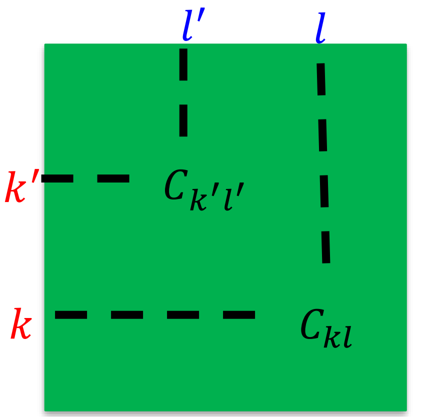


(a) source domain (b) rotation=$20°$ (c) rotation=$40°$ (d) rotation=$90°$

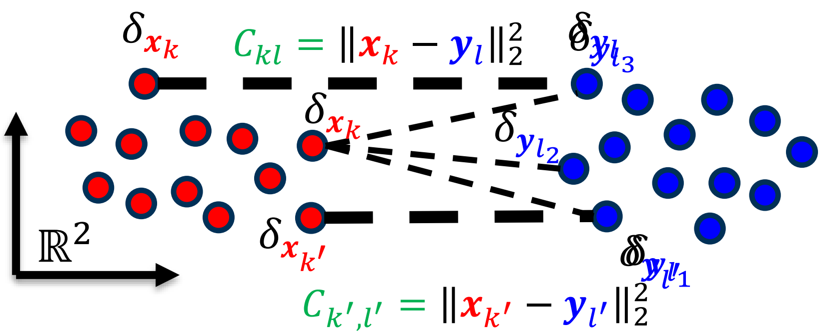
- Results in NP hard formulation
  - Use low-D structure to combat local minima

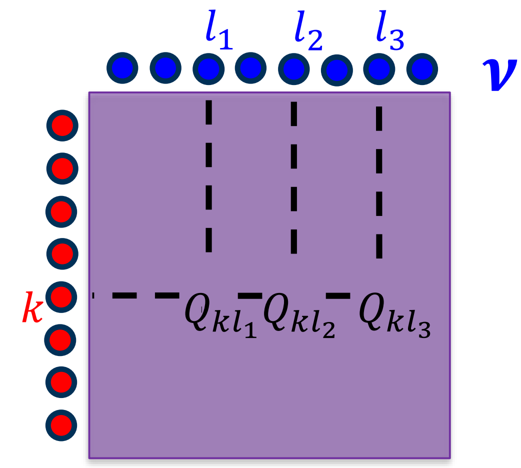
Optimal transport := minimize transport effort

$$\boldsymbol{\mu} = \sum_{k=1}^{n_x} \frac{1}{n_x} \delta_{\boldsymbol{x}_k} \qquad \boldsymbol{\nu} = \sum_{l=1}^{n_y} \frac{1}{n_y} \delta_{\boldsymbol{y}_l}$$



Cost matrix, $\boldsymbol{C}$





$\boldsymbol{\mu}$ Coupling, $\boldsymbol{Q}$

$$\textstyle\sum_l Q_{kl} = 1/n_x, \ \forall k$$
$$\textstyle\sum_k Q_{kl} = 1/n_y, \ \forall l$$

$$W_2^2(\boldsymbol{\mu}, \boldsymbol{\nu}) \coloneqq \text{minimize}(\text{Mass} \times \ell_2^2 \text{ costs}) = \min_{\boldsymbol{Q}} \textstyle\sum_{kl} Q_{kl} \|\boldsymbol{x}_k - \boldsymbol{y}_l\|_2^2$$

(Monge 1781; Kantorovich 1942)

Problem statement

Given point clouds $\boldsymbol{\nu} = \sum_{l=1}^{n_y} \frac{1}{n_y} \delta_{\boldsymbol{y}_l}$, $T(\boldsymbol{\mu}) = \sum_{k=1}^{n_x} \frac{1}{n_x} \delta_{T(\boldsymbol{x}_k)}$

Goal: $\boxed{\min_{T \in \mathcal{T}} D(T(\boldsymbol{\mu}), \boldsymbol{\nu})}$

We need:

1. A **divergence** between the point clouds that *exploits* geometry
2. A **transformation** that *preserves* geometry
3. An approach that is computationally **tractable** with **little (or no) supervision**

Transformation classes

$$\boxed{T(\boldsymbol{\mu}) = \sum_{k=1}^{n_x} \frac{1}{n_x} \delta_{\boldsymbol{R}\boldsymbol{x}_k} \text{ s.t. } \boldsymbol{R}^\top \boldsymbol{R} = \boldsymbol{I}}$$

- Affine
  - May arbitrarily warp geometry
  - (Fernando et al. 2013; Sun et al. 2016; Courty et al. 2016)
- OT's Barycentric Mapping
  - Inferred from the OT's coupling $\boldsymbol{Q}$ (i.e., correspondence)
  - Convex but transform is tied to transport cost (assumes proximity in representations)
  - (Courty et al. 2016)
- Orthogonal
  - Preserves geometry
  - non-convex constraints
  - Routinely used in point set registration literature
  - (Besl & McKay 1992; Gold et al. 1998; Myronenko & Song 2010)

Problem statement

Given point clouds $\boldsymbol{\nu} = \sum_{l=1}^{n_y} \frac{1}{n_y} \delta_{\boldsymbol{y}_l}$, $T(\boldsymbol{\mu}) = \sum_{k=1}^{n_x} \frac{1}{n_x} \delta_{T(\boldsymbol{x}_k)}$

Goal: $\boxed{\min_{T \in \mathcal{T}} D(T(\boldsymbol{\mu}), \boldsymbol{\nu})}$

We need:

1. A **divergence** between the point clouds that *exploits* geometry
2. A **transformation** that *preserves* geometry
3. An approach that is computationally **tractable** with **little (or no) supervision**

Problem formulation

$$\min_{T\in\mathcal{T}} D(T(\boldsymbol{\mu}),\boldsymbol{\nu}) = \min_{T\in\mathcal{T}} W_2^2(T(\boldsymbol{\mu}),\boldsymbol{\nu}) = \min_{\boldsymbol{Q},\boldsymbol{R}} \sum_{k=1}^{n_x}\sum_{l=1}^{n_y} Q_{kl}\|\boldsymbol{R}\boldsymbol{x}_k - \boldsymbol{y}_l\|_2^2$$

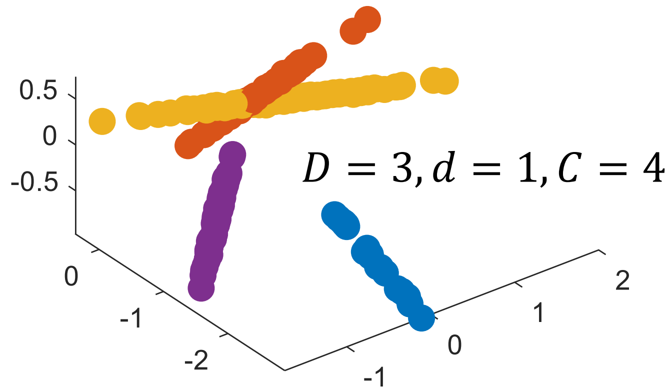
$$\boldsymbol{Q} \in \mathcal{U}(n_x,n_y) \coloneqq \left\{\boldsymbol{Q}\in\mathbb{R}_+^{n_x\times n_y} : \boldsymbol{Q}\mathbf{1} = \frac{\mathbf{1}}{n_x}, \boldsymbol{Q}^\top\mathbf{1} = \frac{\mathbf{1}}{n_y}\right\}, \qquad \boldsymbol{R}\in\mathcal{V}_d \coloneqq \{\boldsymbol{R}\in\mathbb{R}^{d\times d} : \boldsymbol{R}^\top\boldsymbol{R} = \boldsymbol{I}\}$$

- **Divergence**: Squared 2-Wasserstein distance
  - explicitly uses geometry of embedding
- **Transformation**: Orthogonal transformation
  - isometric, i.e., preserves geometry

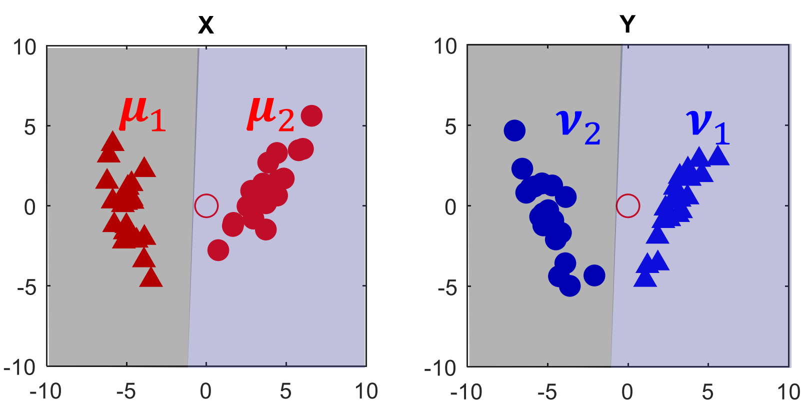
- **Advantages**: transformation is *decoupled from divergence*
  - No implicit proximity assumption (c.f. Courty et al. 2016)
- **Challenges**:
  - non-convex : prone to poor local-minima
  - expensive to compute : requires solving many OT problems in alternating minimization

Cluster assumptions to combat local minima

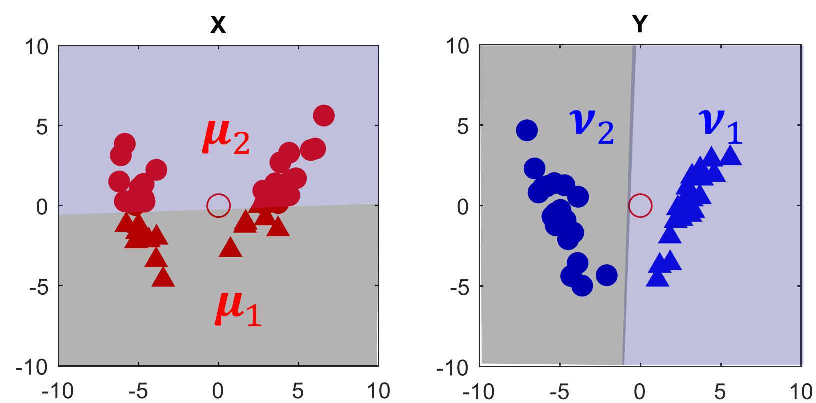
- Data comes from a union of subspaces
  - Notation: D:ambient dim, d:subspace dim, C:# clusters



- Datasets are clustered similarly (regardless of order)
  - Semi-supervised with labels or unsupervised without



Similar clustering



Dissimilar clustering

Hierarchical Wasserstein Alignment (HiWA)

- Given pre-clustered data: $\{\boldsymbol{\mu}_i\}_{i=1}^{C}$ and $\{\boldsymbol{\nu}_j\}_{j=1}^{C}$

$$\min_{\substack{\boldsymbol{P}\in\mathcal{U}(C,C)\\ T,\{T_{ij}\}_{ij}\in\mathcal{T}}} \sum_{ij} P_{ij} \underbrace{W_2^2\left(T_{ij}(\boldsymbol{\mu}_i),\boldsymbol{\nu}_j\right)}_{\text{updated in parallel}} + H_\varepsilon(\boldsymbol{P}) \quad \text{s.t.} \quad \underbrace{T = T_{ij}}_{\substack{\text{consensus}\\ \text{constraint } \forall i,j}}$$

- Strength of cluster correspondence is denoted by P_{ij}
- Cluster-alignment costs are defined by $W_2^2\left(T(\boldsymbol{\mu}_i),\boldsymbol{\nu}_j\right)$
- Interpretation: **nested/block** OT
  - Minimize OT *between* clusters with C_{ij} = OT *within* cluster

- To improve tractability:
  - **Sinkhorn relaxation:** entropic regularization improves computational/sample complexity (Cuturi 2013)
  - **Non-convex distributed ADMM:** complexity improved similar to (Wang et al. 2019)

Distributed HiWA on synthetic data

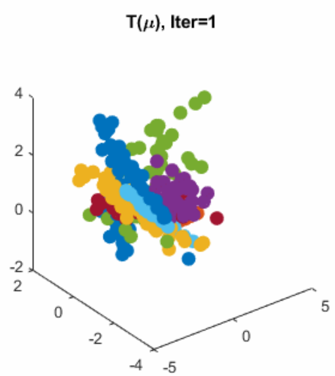
$$\min_{\boldsymbol{P},T,\{T_{ij}\}_{ij}} \sum_{ij} P_{ij}\, \underbrace{W_2^2\big(T_{ij}(\boldsymbol{\mu}_i),\boldsymbol{\nu}_j\big)}_{C_{ij}} + H_\varepsilon(\boldsymbol{P}) \quad \text{s.t.} \quad \underbrace{T = T_{ij}}_{\forall i,j}$$

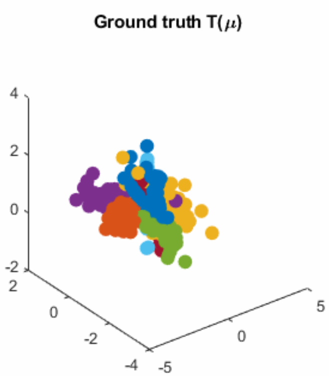
Simulation parameters
ambient dim = 8
subspace dim = 2
clusters = 8
pts/cluster = 50

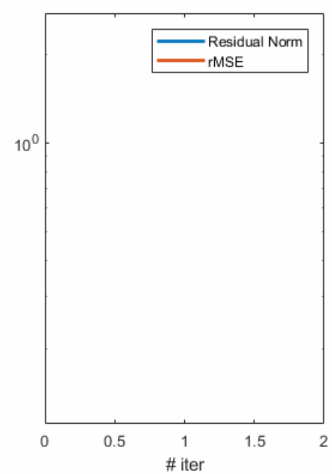
$$\boldsymbol{P}^\star = \boldsymbol{I}$$

Transform error: Residual Norm = $\left\|\boldsymbol{R}^{(t)} - \boldsymbol{R}^{(t-1)}\right\|_F^2$

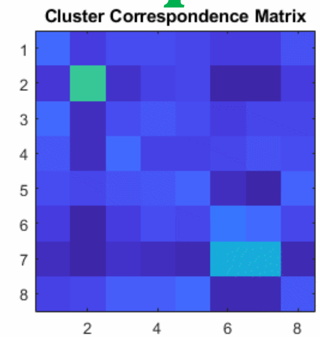
Alignment error: rMSE = $\left\|\boldsymbol{R}^\star\boldsymbol{X} - \boldsymbol{R}^{(t)}\boldsymbol{X}\right\|_F^2 / \left\|\boldsymbol{R}^\star\boldsymbol{X}\right\|_F^2$



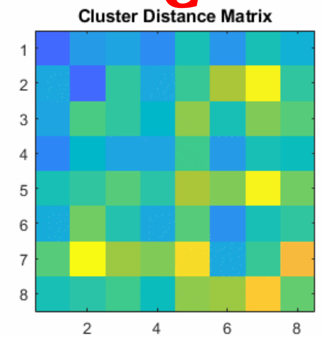




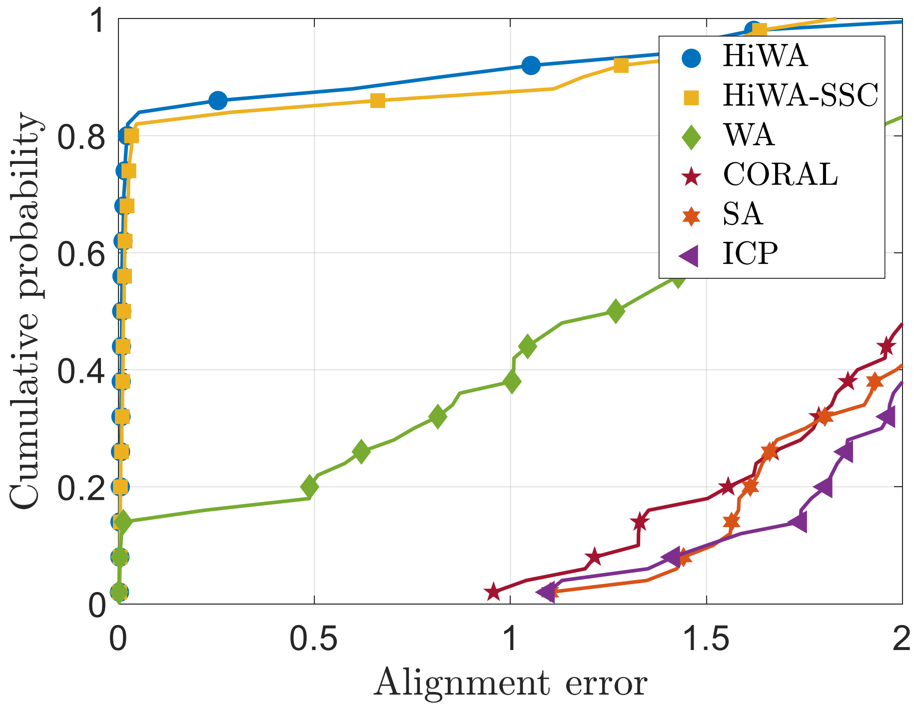
$\boldsymbol{P}$

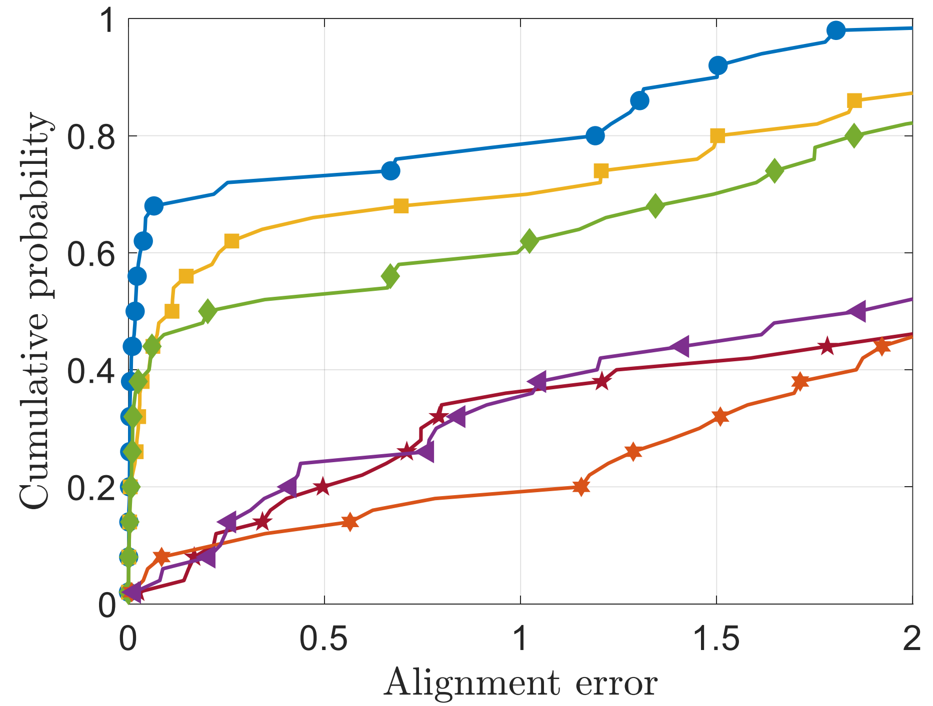


$\boldsymbol{C}$



Synthetic results comparisons






```
# ambient dim = 6
# subspace dim = 2
# clusters = 5
```

```
# ambient dim = 2
# subspace dim = 2
# clusters = 2
```

[HiWA] := oracle cluster labels

[HiWA-SSC] := cluster labels via Sparse Subspace Clustering (Elhamifar & Vidal 2013)

[CORAL] := correlation alignment (Sun et al. 2016)

[SA] := subspace alignment (Fernando et al. 2013)

[ICP] := iterative closest point (Besl et al. 1992)

Alignment guarantees

Can we guarantee alignment from: $\boxed{\min_{\substack{P \in \mathcal{U}(C,C) \\ T \in \mathcal{T}}} \sum_{ij} P_{ij} W_2^2(T(\boldsymbol{\mu}_i), \boldsymbol{\nu}_j)}$?

<u>Theorem (Lee, Dabagia, Dyer, R., 2019):</u>

Let $\hat{C}_{ij} \coloneqq \min_T W_2^2(T(\boldsymbol{\mu}_i), \boldsymbol{\nu}_j)$, $\forall i, j$ then the cluster correspondence matrix will have the correct solution $\boldsymbol{P}^\star = \boldsymbol{I}$ with high probability if

$$\hat{C}_{ij} + \hat{C}_{ji} - (\hat{C}_{ii} + \hat{C}_{jj}) \gtrsim O(n^{-1/d}), \qquad \forall i, j: i \neq j$$

where d is the intrinsic dimension of the clusters.

Local similarities between **matched** $(i = j)$ clusters should be more similar than local similarities between **mismatched** $(i \neq j)$ clusters up to an asymptotic sample complexity dependent on **intrinsic dimensions**.

Error bounds

Theorem (Lee et al., 2019): orthogonal error bounds

Let clusters $\{\boldsymbol{X}_i\}_{i=1}^C, \{\boldsymbol{Y}_i\}_{i=1}^C$ have correspondences $\{\boldsymbol{Q}_{ii}\}_{i=1}^C$, and define

$$\boldsymbol{X} = [\boldsymbol{X}_1\boldsymbol{Q}_{11}, \boldsymbol{X}_2\boldsymbol{Q}_{22}, \dots, \boldsymbol{X}_C\boldsymbol{Q}_{CC}], \qquad \boldsymbol{Y} = [\boldsymbol{Y}_1, \boldsymbol{Y}_2, \dots, \boldsymbol{Y}_C].$$

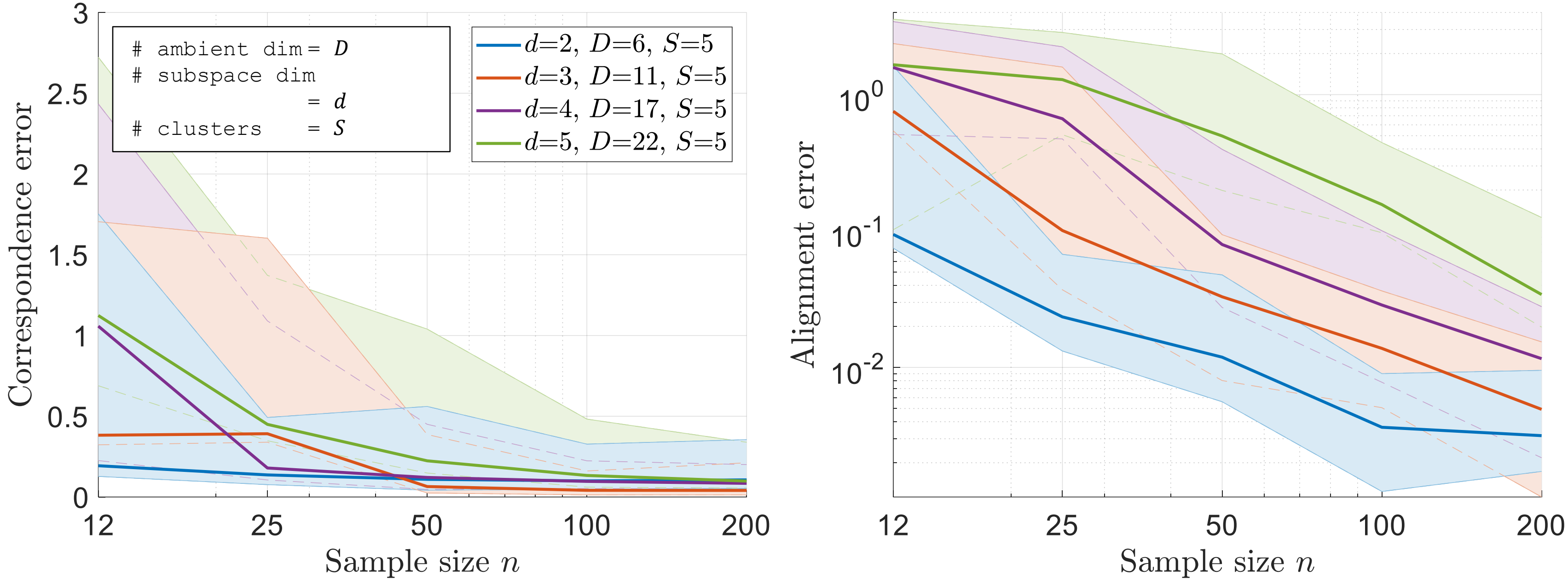
For orthogonal T, and if the dataset is *alignable* (with a few more technical conditions), then

$$\min_{P,T} \sum_{ij} P_{ij} W_2^2\left(T(\boldsymbol{\mu}_i), \boldsymbol{\nu}_j\right) \leq C_1 \|\boldsymbol{Y}^\top \boldsymbol{Y} - \boldsymbol{X}^\top \boldsymbol{X}\|_F^2 + C_2,$$

where C_1, C_2 do not depend on the Grammians of $\boldsymbol{X}, \boldsymbol{Y}$.

Orthogonal alignment is bounded by Grammian differences that capture distortions in global structure.

Synthetic results: sample complexity

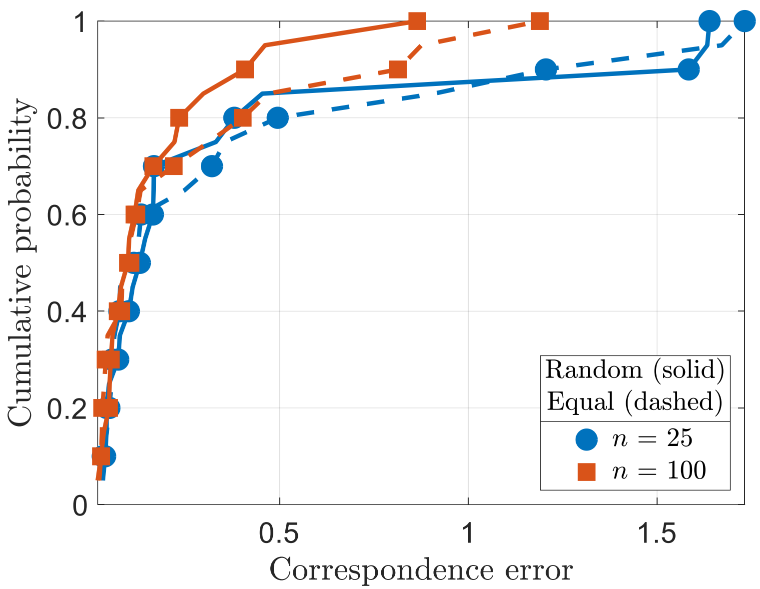


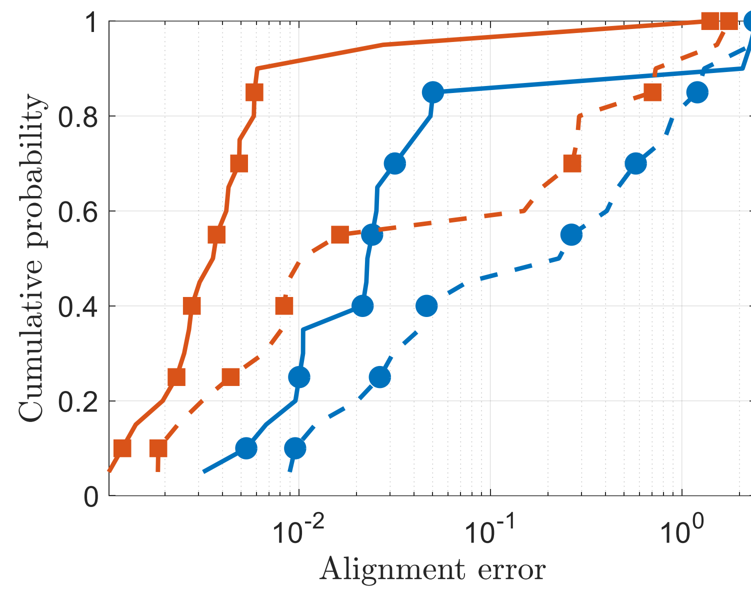
Correspondence Error = $\left\|\widehat{\boldsymbol{P}} - \boldsymbol{P}^{\star}\right\|_1$

Alignment Error = $\dfrac{\left\|\widehat{\boldsymbol{R}}\boldsymbol{X} - \boldsymbol{R}^{\star}\boldsymbol{X}\right\|_F^2}{\left\|\boldsymbol{R}^{\star}\boldsymbol{X}\right\|_F^2}$

Synthetic results: worst case configuration

- See paper for analysis of error rates
- Indicates equally spaced subspaces worst for alignment

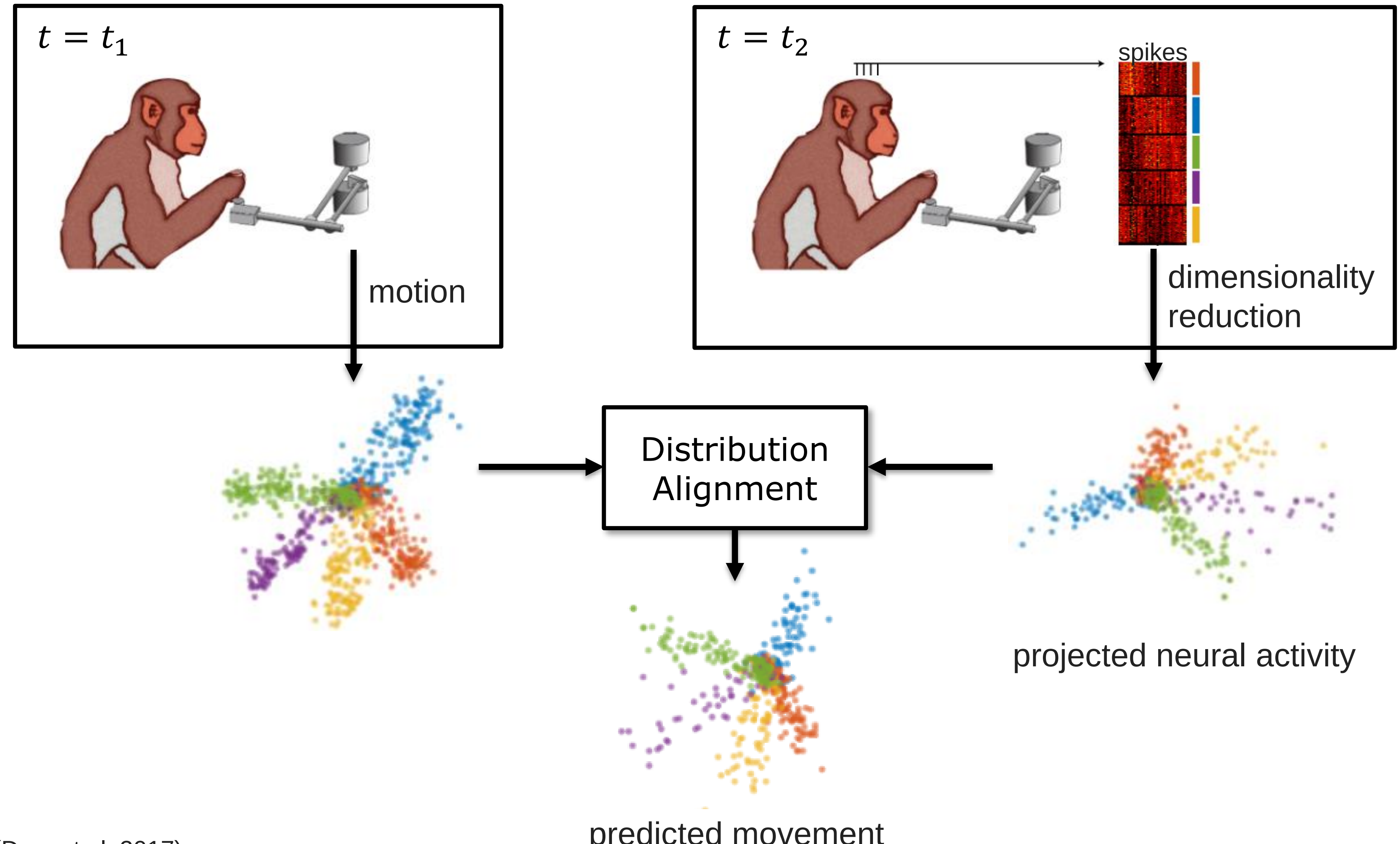




Correspondence Error = $\left\| \widehat{\boldsymbol{P}} - \boldsymbol{P}^{\star} \right\|_1$

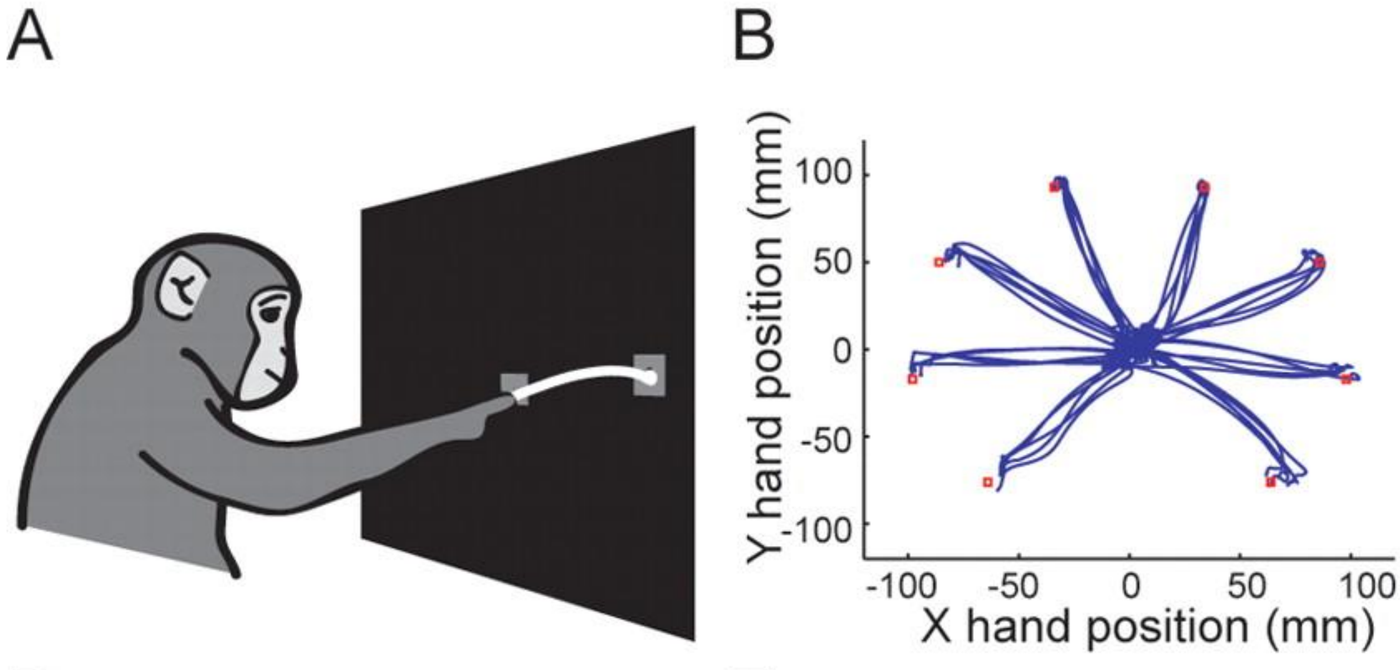
```
# ambient dim  = 5
# subspace dim = 2
# clusters     = 6
```

Recap: movement decoding



(Dyer et al. 2017)

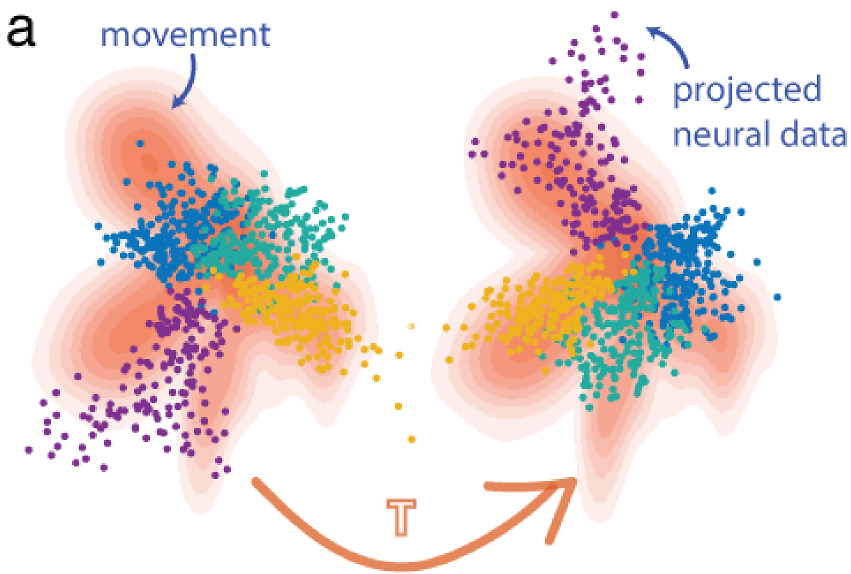
Center-out reaching task

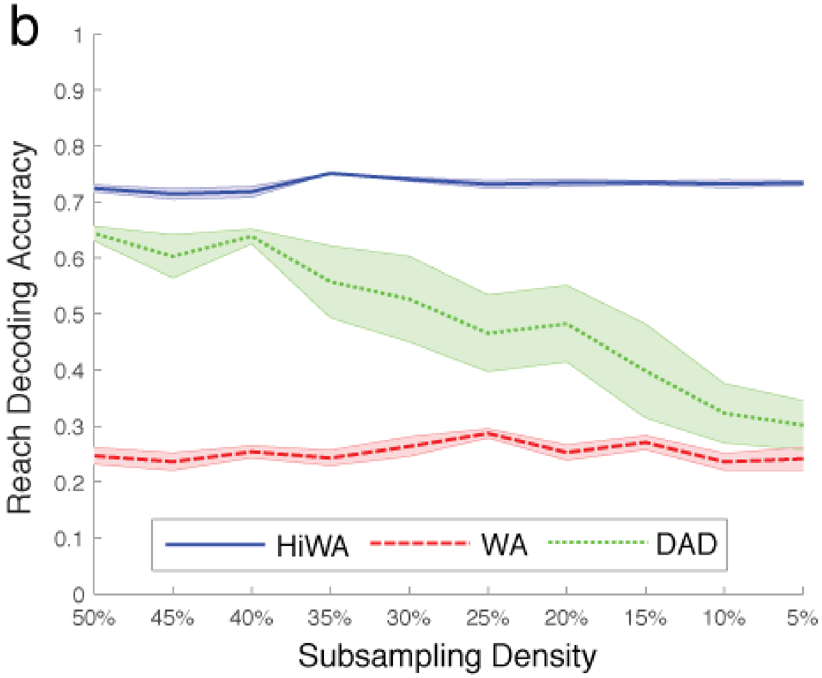


(Chestek, et al. 2007)

- Common and reliable tasks for non-human primates
- Behavior (movement) data and neural population recordings after dimensionality reduction (from L. Miller)
- Eight well-defined clusters (8! possible correspondences)
  - Solved in ~100 iterations with ~50 points/cluster
  - Compute time 1-2 minutes on quad core CPU

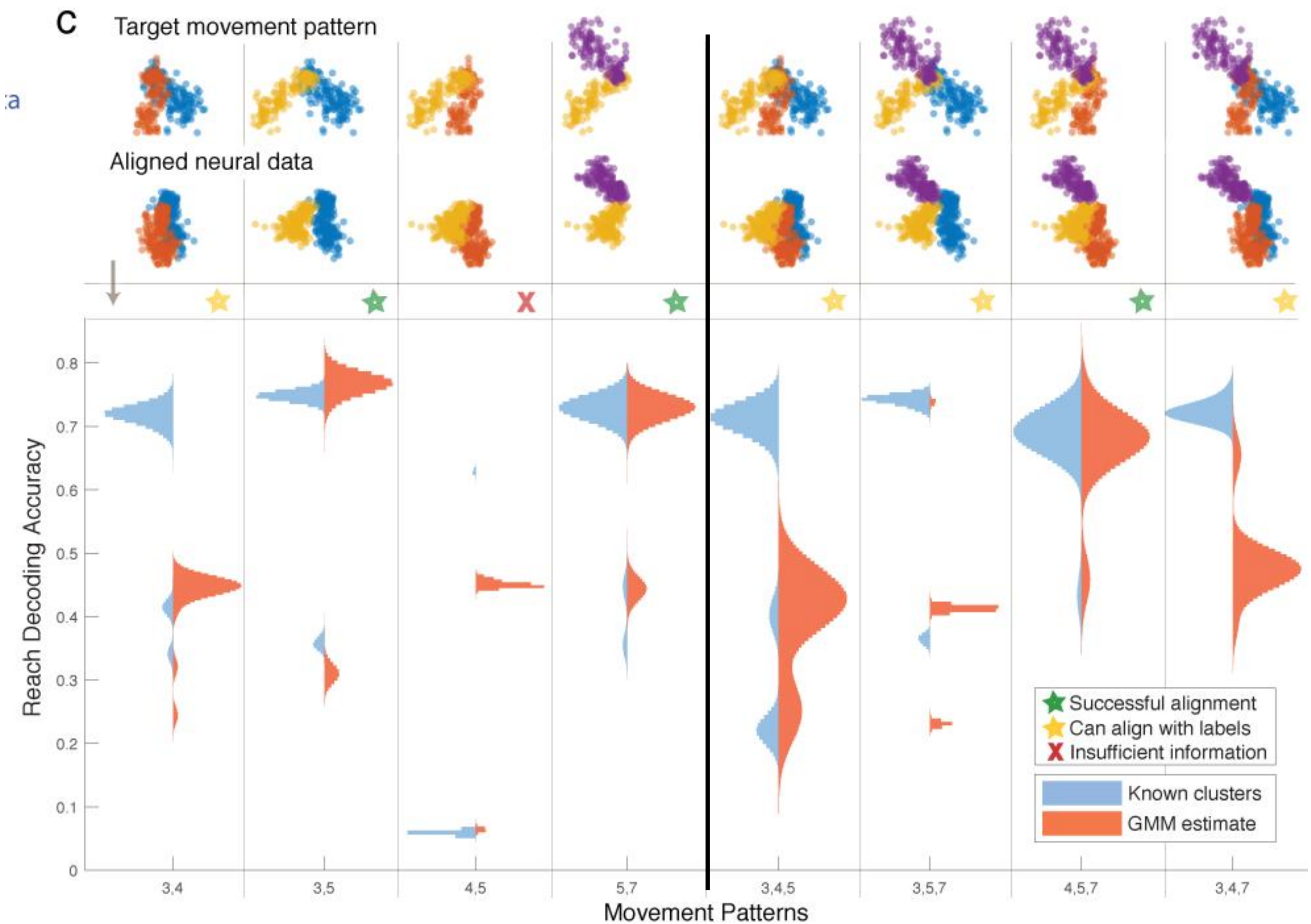
Results on neurophysiology data





[DAD] := Distribution alignment decoding. Brute force KL alignment method developed specifically for ambient dims of d=3 (Dyer et al. 2017)

Variations with movement subsets



2/8 patterns

3/8 patterns

Summary

- Significant improvement in data alignment by combining low-dimensional structure and optimal transport
  - Results in nested OT: correspond clusters using OT where cost is local OT of aligning data within a cluster
  - Unsupervised or semi-supervised (cluster labels) formulation
- Made tractable via:
  - recent OT advances (entropic regularization) and new distributed ADMM algorithm
  - Low-dimensional structure helps avoid local minima
- See paper for details, analysis and more evaluation
  - Includes analysis on alignment conditions and error rates