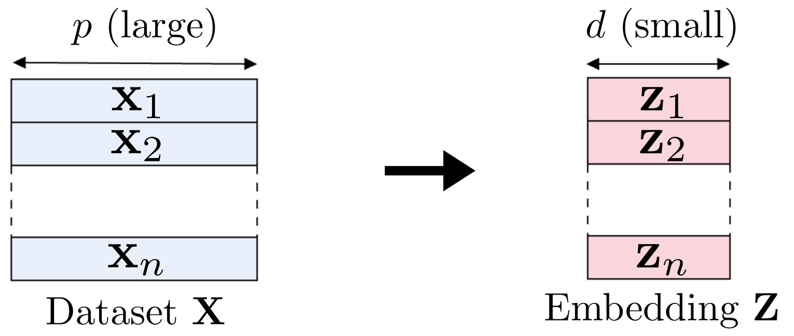
SNEkhorn : Dimension Reduction with Symmetric Entropic Affinities

Hugues Van Assel, Titouan Vayer, Rémi Flamary, Nicolas Courty

Context

Dimension reduction methods rely on **symmetric** affinity matrices.



Starting point | **Cost matrix:** $\mathbf{C} \in \mathbb{R}_+^{n \times n}$ such that $\mathbf{C} = \mathbf{C}^\top$ and $C_{ij} = 0 \iff i = j$.

Example: $C_{ij} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2$.

Doubly Stochastic Affinity

Sinkhorn Algorithm — init $\mathbf{K} = \exp(-\mathbf{C}/\sigma)$

While not converged:

$\mathbf{K} \leftarrow \operatorname{diag}(\mathbf{K}\mathbf{1})^{-1}\mathbf{K}$ # normalize rows

$\mathbf{K} \leftarrow \mathbf{K}\operatorname{diag}(\mathbf{1K})^{-1}$ # normalize columns

converges $\longrightarrow$ $\mathcal{DS} = \{\mathbf{P} \text{ s.t. } \mathbf{P1} = \mathbf{P}^\top\mathbf{1} = \mathbf{1}\}$.

$\mathbf{P}^{\mathrm{ds}}$ solves the optimal transport problem:

$$\min_{\mathbf{P} \in \mathbb{R}_+^{n\times n}} \langle \mathbf{P}, \mathbf{C}\rangle - \sigma \sum_i \mathrm{H}(\mathbf{P}_{i:}) \quad \text{s.t.} \quad \mathbf{P} \in \mathcal{DS}.$$

$\mathrm{H}(\boldsymbol{p}) = -\langle \boldsymbol{p}, \log \boldsymbol{p} - \mathbf{1}\rangle$ is the Shannon entropy.

Entropic Affinity

Controls the entropy in each point with **adaptive bandwidths**.

Definition [Hinton, Roweis 2002]

$$\forall i, \forall j,\ P^{\mathrm{e}}_{ij} = \frac{\exp(-C_{ij}/\varepsilon_i^\star)}{\sum_\ell \exp(-C_{i\ell}/\varepsilon_i^\star)}$$

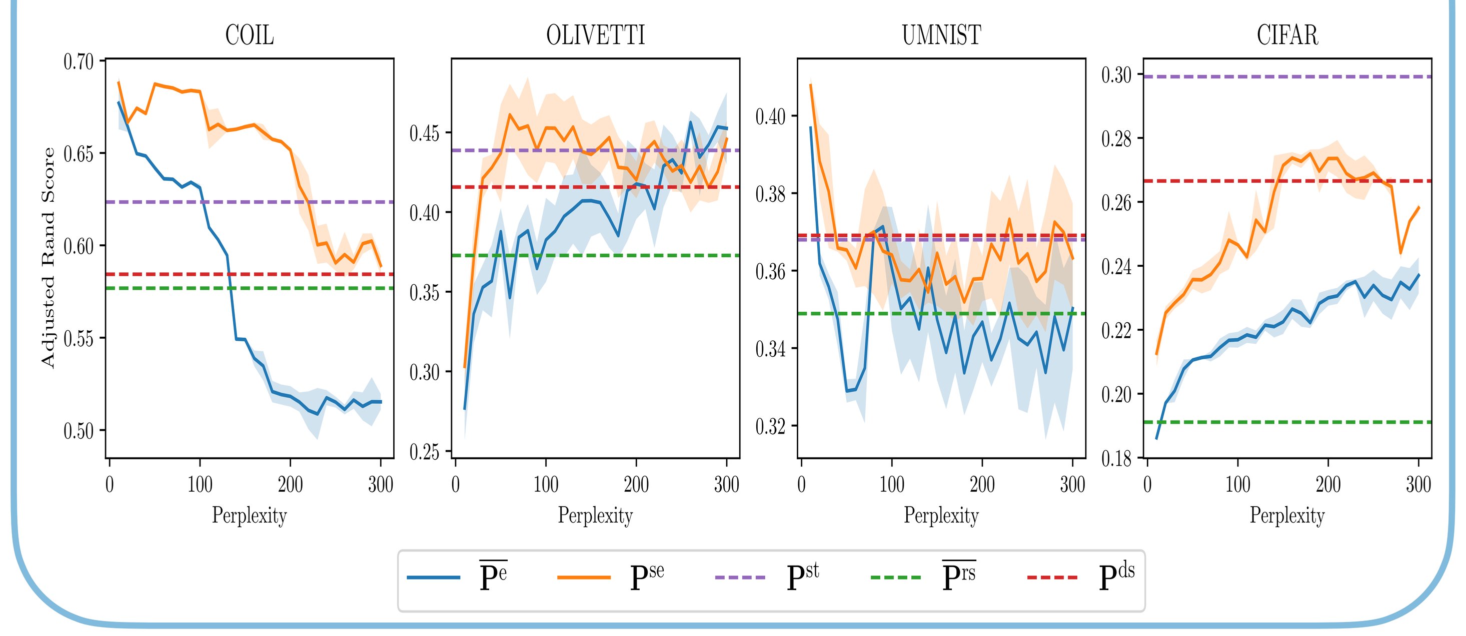
with $\varepsilon_i^\star \in \mathbb{R}_+^*$ s.t. $\mathrm{H}(\mathbf{P}^{\mathrm{e}}_{i:}) = \log\xi + 1$.

$\xi \in [\![1, n]\!]$ is the **perplexity** parameter. **t-SNE algorithm** [Van der Maaten and Hinton, 2008]

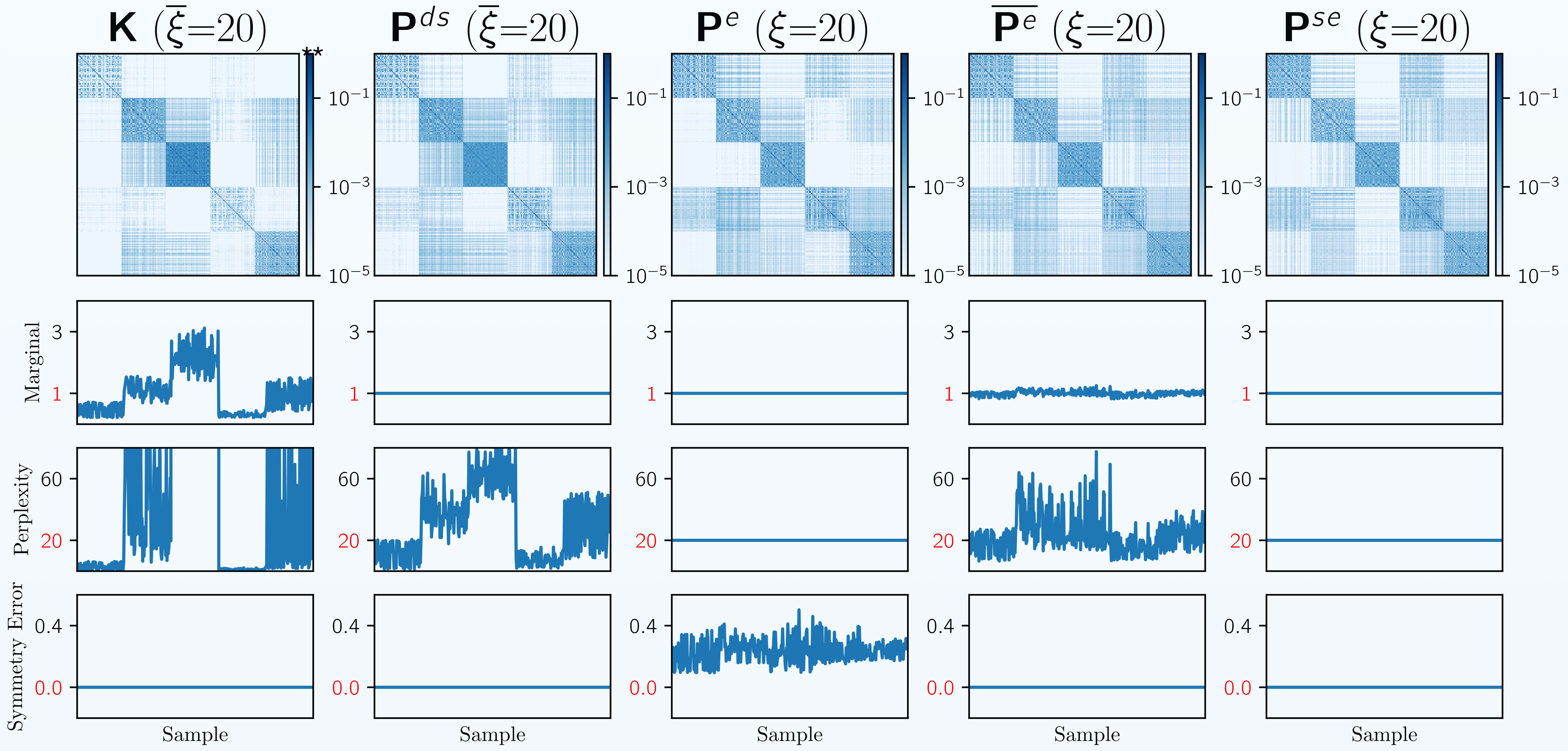
$\mathbf{P}^{\mathrm{e}}$ **is not symmetric.**

$\rightarrow$ $\overline{\mathbf{P}^{\mathrm{e}}} = \frac{1}{2}(\mathbf{P}^{\mathrm{e}} + \mathbf{P}^{\mathrm{e}\top})$ is used in practice. $\rightarrow$ $\overline{\mathbf{P}^{\mathrm{e}}} = \operatorname{Proj}^{\ell_2}_{\mathcal{S}}(\mathbf{P}^{\mathrm{e}})$

Spectral Clustering Experiments



Gibbs kernel	Doubly-Sto	Entropic	Entropic (ℓ_2 Sym)	Sym-Entropic
$\mathbf{K} = \exp(-\mathbf{C}/\sigma)$	$\mathbf{P}^{\mathrm{ds}} = \operatorname{Proj}^{\mathrm{KL}}_{\mathcal{DS}}(\mathbf{K})$	$\mathbf{P}^{\mathrm{e}} = \operatorname{Proj}^{\mathrm{KL}}_{\mathcal{H}_\xi}(\mathbf{K})$	$\overline{\mathbf{P}^{\mathrm{e}}} = \operatorname{Proj}^{\ell_2}_{\mathcal{S}}(\mathbf{P}^{\mathrm{e}})$	$\mathbf{P}^{\mathrm{se}} = \operatorname{Proj}^{\mathrm{KL}}_{\mathcal{H}_\xi \cap \mathcal{S}}(\mathbf{K})$



* On 5 classes of the COIL Dataset [Nene et al., 1996] ** $\bar\xi$ is average perplexity —> same global entropy as with $\xi = \bar\xi$.

Can we build a symmetric affinity with controlled ℓ_1 norm and entropy in each point?

Symmetric Entropic Affinity

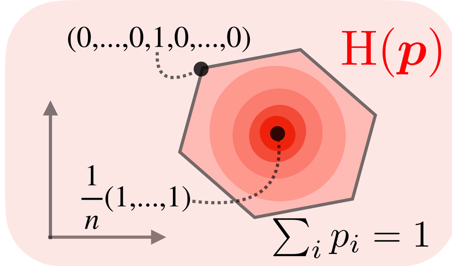
$$\mathcal{H}_\xi := \{\mathbf{P} \in \mathbb{R}_+^{n\times n} \text{ s.t. } \mathbf{P1} = \mathbf{1} \text{ and } \forall i,\ \mathrm{H}(\mathbf{P}_{i:}) \geq \log\xi + 1\}$$

Entropic Affinity as OT

$$\mathbf{P}^{\mathrm{e}} = \operatorname*{arg\,min}_{\mathbf{P} \in \mathcal{H}_\xi} \langle \mathbf{P}, \mathbf{C}\rangle.$$

The constraints in $\mathcal{H}_\xi$ are saturated at the optimum.

Symmetric matrices $\mathcal{S} = \{\mathbf{P} \text{ s.t. } \mathbf{P} = \mathbf{P}^\top\}$.



Definition

$$\mathbf{P}^{\mathrm{se}} := \operatorname*{arg\,min}_{\mathbf{P} \in \mathcal{H}_\xi \cap \mathcal{S}} \langle \mathbf{P}, \mathbf{C}\rangle.$$

Enforce Symmetry

Property

For at least $n-1$ indices $i \in [\![n]\!]$, it holds $\mathrm{H}(\mathbf{P}^{\mathrm{se}}_{i:}) = \log\xi + 1$.

In practice, we have n saturated entropies.

Dual Ascent

$$\mathbf{P}^{\mathrm{se}} = \exp\left((\boldsymbol{\lambda}^\star \oplus \boldsymbol{\lambda}^\star - 2\mathbf{C}) \oslash (\boldsymbol{\gamma}^\star \oplus \boldsymbol{\gamma}^\star)\right)$$

where $\boldsymbol{\lambda}^\star$ and $\boldsymbol{\gamma}^\star$ are computed using dual ascent.

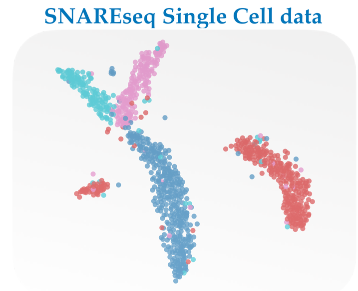
SNE & SNEkhorn

Cost matrix between embeddings: $[\mathbf{C_Z}]_{ij} = \|\mathbf{Z}_{i:} - \mathbf{Z}_{j:}\|_2^2$.

Stochastic Neighbor Embedding (SNE)

$$\min_{\mathbf{Z} \in \mathbb{R}^{n\times d}} \mathrm{KL}(\overline{\mathbf{P}^{\mathrm{e}}} | \widetilde{\mathbf{Q}}_{\mathbf{Z}})$$

where $[\widetilde{\mathbf{Q}}_{\mathbf{Z}}]_{ij} = \exp(-[\mathbf{C_Z}]_{ij}) / \sum_{\ell,t}\exp(-[\mathbf{C_Z}]_{\ell t})$.



SNEkhorn

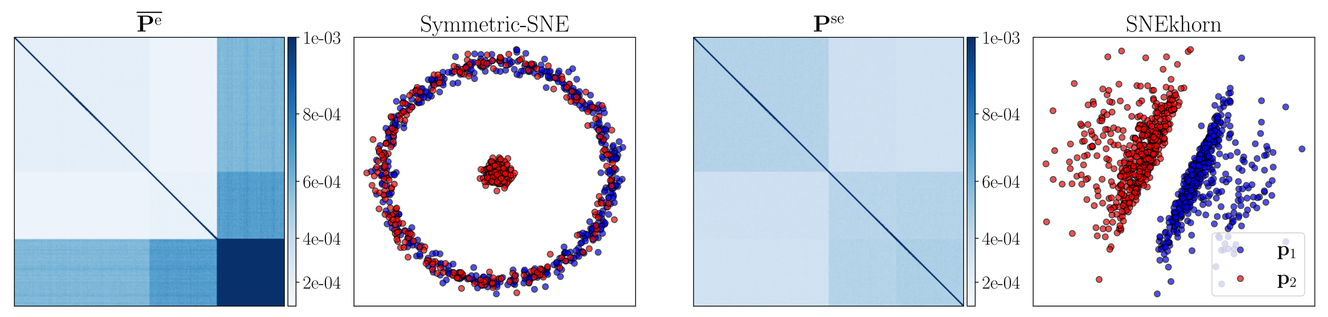
$$\min_{\mathbf{Z} \in \mathbb{R}^{n\times d}} \mathrm{KL}(\mathbf{P}^{\mathrm{se}} | \mathbf{Q}^{\mathrm{ds}}_{\mathbf{Z}})$$

where $\mathbf{Q}^{\mathrm{ds}}_{\mathbf{Z}} = \exp(\mathbf{f_Z} \oplus \mathbf{f_Z} - \mathbf{C_Z})$ is the DS affinity.

Extension to t-SNE / t-SNEkhorn with heavy-tailed kernels: [Van der Maaten and Hinton, 2008]

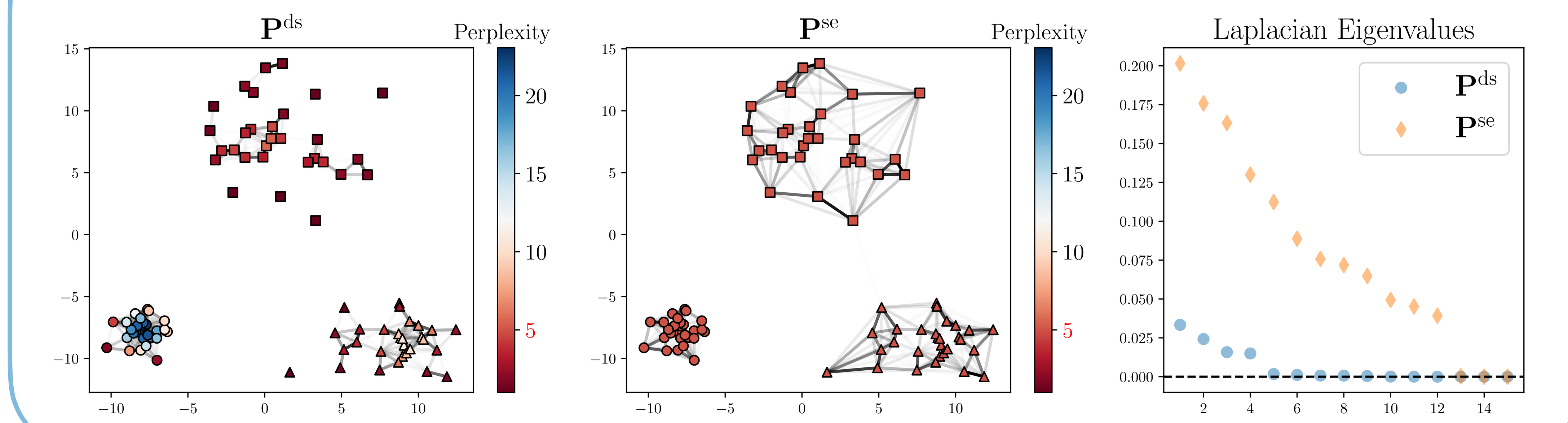
$$[\mathbf{C_Z}]_{ij} = \left(\log(1 + \|\mathbf{Z}_{i:} - \mathbf{Z}_{j:}\|_2^2)\right)_{ij}.$$

Robustness to Noise (1/2)



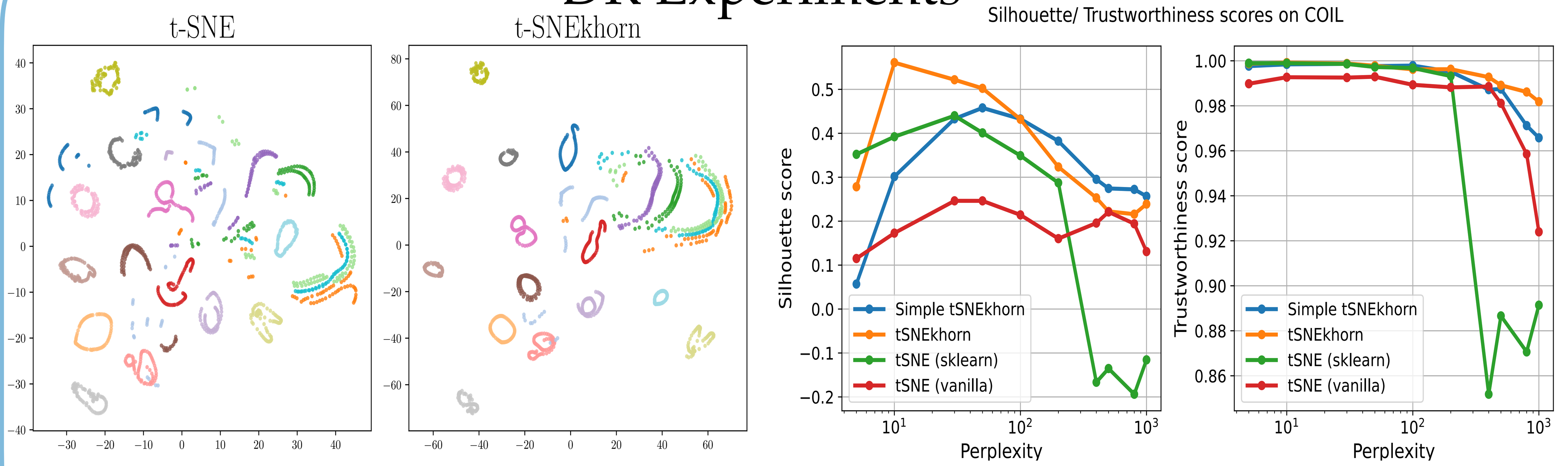
$$\boldsymbol{x}_i = \tilde{\boldsymbol{x}}_i / (\textstyle\sum_j \tilde{x}_{ij}), \quad \tilde{\boldsymbol{x}}_i \sim \begin{cases} \mathcal{M}(1000, \mathbf{p}_1), & 1 \leq i \leq 500 \\ \mathcal{M}(1000, \mathbf{p}_2), & 501 \leq i \leq 750 \\ \mathcal{M}(2000, \mathbf{p}_2), & 751 \leq i \leq 1000. \end{cases}$$

Robustness to Noise (2/2)



DR Experiments

COIL-20 Image data



* Simple t-SNEkhorn —> t-SNEkhorn with same affinity as t-SNE for the embeddings $\mathbf{Z}$ (not doubly stochastic).