GSVD for Geometry-Grounded Dataset Comparison: *An Alignment Angle Is All You Need*

Eduarda de Souza Marques · Arthur Sobrinho · João Rego Paixão
Daniel Sadoc Menasche · Heudson Tosta Mirandola
Federal University of Rio de Janeiro — UFRJ, Brazil

ICLR International Conference On Learning Representations

UFRJ Universidade Federal do Rio de Janeiro
CNPq FAPERJ

Main question. Given a sample z, and two datasets A and B, which dataset better explains z?

Main message. With $Ax = By = z$, GSVD gives a sample-level angle $\theta(z) \in [0, \pi/2]$: small $\theta = A$-aligned, large $\theta = B$-aligned.

Sample-Level Dataset Comparison

State-of-the-art dataset comparison (e.g., CKA)

In: Datasets A, B → **Black Box** → **Out:** **similarity = 65%**

Question. Which patterns drive similarity of A and B?
Our answer. Compute **alignment angle** $\theta(z)$ per sample.

From Co-Span to a GSVD Geometry

We study vectors jointly explained by both datasets:

$$Ax = By = z, \qquad A \in \mathbb{R}^{d\times p},\ B \in \mathbb{R}^{d\times q},\ z \in \mathbb{R}^{d}.$$

Key idea. GSVD provides a coordinate system adapted to the shared geometry of A and B.

$$A = HCU, \qquad B = HSV, \qquad C^\top C + S^\top S = I.$$

- H is the common ambient basis.
- C and S quantify how each direction leans toward A or B.
- This makes shared and dataset-specific structure explicit.

Scoring a Sample with the Alignment Angle

For $z \in \mathrm{col}(A) \cap \mathrm{col}(B)$, define

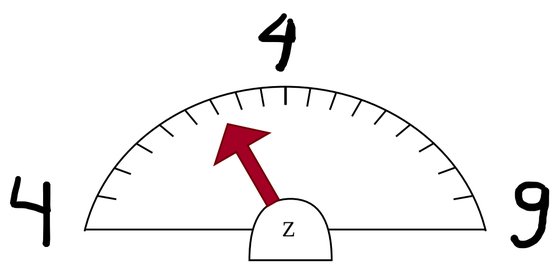
$$\mathcal{R}_{\text{co-span}}(z) = \{(x, y) \in \mathbb{R}^p \times \mathbb{R}^q : Ax = By = z\}.$$

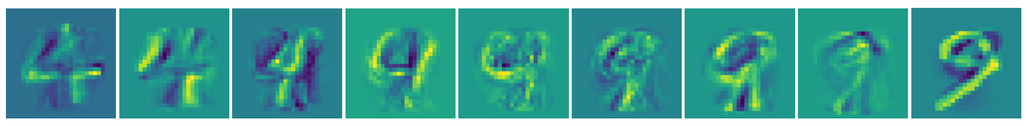
With $x \perp \mathrm{Ker}(A)$ and $y \perp \mathrm{Ker}(B)$,

$$\theta(z) := \arctan\left(\frac{\|x\|_2}{\|y\|_2}\right) \in [0, \pi/2].$$

GSVD formula:

$$c(z) = H^\dagger z, \qquad a(z) = \|C^\dagger c(z)\|_2, \qquad b(z) = \|S^\dagger c(z)\|_2,$$
$$\theta(z) = \arctan\left(\frac{a(z)}{b(z)}\right).$$





Cost per sample. One projection + two rescalings.

Information-Geometric Interpretation

Key idea. The alignment angle $\theta(z)$ induces a simple probabilistic interpretation, enabling an information-geometric view.

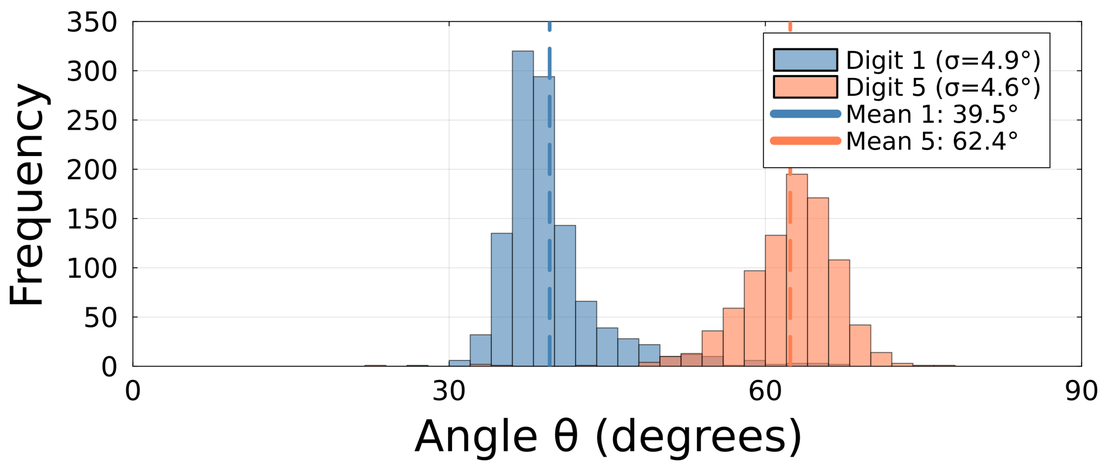
Bernoulli posterior:

$$P(A \mid \theta) = \cos^2\theta, \qquad P(B \mid \theta) = \sin^2\theta.$$

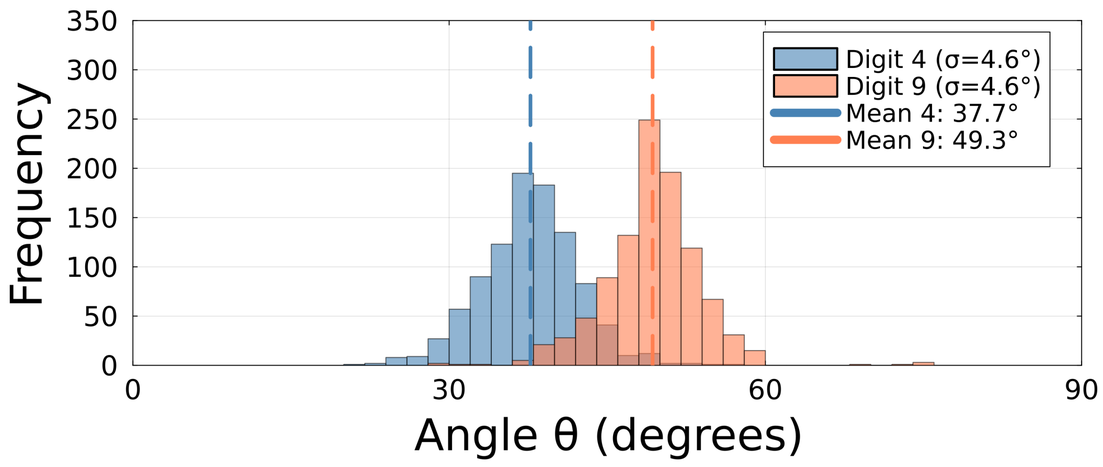
MNIST Evidence: Easy vs Hard Relative Geometry

Protocol. Build A and B from two digit classes, then evaluate $\theta(z)$ on test samples.

Key idea. Separated angle histograms mean distinct geometry; overlap near the middle means ambiguity.



Digits 1 vs 5: clearer separation.



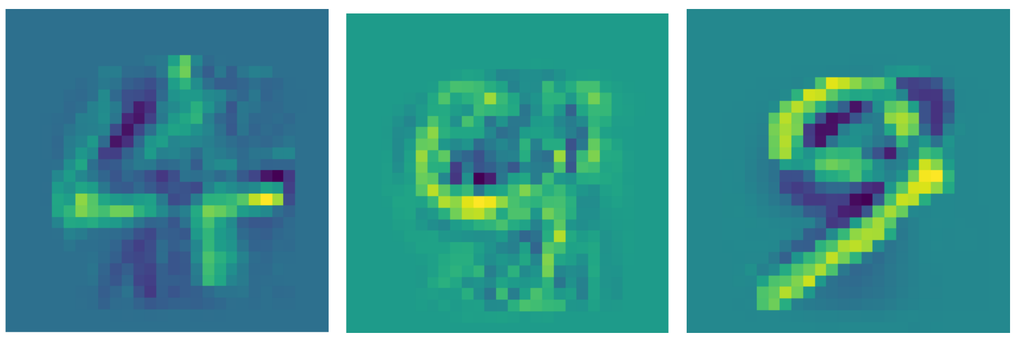
Digits 4 vs 9: stronger overlap.

- Easy pairs: stronger separation in angle space.
- Hard pairs: more mass near the middle.

Representative Extreme Directions

GSVD orders directions: most aligned with A vs with B.

$$z_{\min} \in \arg\min_{z\in\mathcal{Z}} \theta(z), \qquad z_{\max} \in \arg\max_{z\in\mathcal{Z}} \theta(z).$$



more 4-like | shared | more 9-like

Contributions

1. A geometry-grounded co-span view of dataset comparison.
2. A GSVD-based sample-level score $\theta(z)$.
3. Interpretable extreme directions and visual diagnostics.
4. MNIST evidence showing easy vs. hard relative geometry.

Contact

Arthur Sobrinho: arthursfr@ic.ufrj.br
Eduarda Marques: eduardasm@ic.ufrj.br

Scan for paper

References. Van Loan (1976); Paige and Saunders (1981); Edelman and Wang (2020); Björck and Golub (1973); LeCun et al. (1998).

Date: April 2026.