Quantitative bounds for sorting-based permutation-invariant embeddings

Radu Balan

University of Maryland
Department of Mathematics and the Norbert Wiener Center
College Park, Maryland *rvbalan@umd.edu*

January 4, 2026

SIAM Minisymposium on Approximation Theory, II

Joint Mathematics Meeting, January 3-6, 2025.

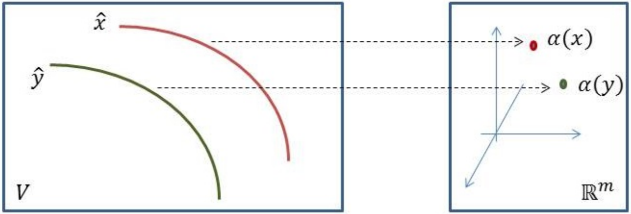
Jointly: Nadav Dym, Matthias Wellershoff, Efstratios Tsoukanis, Daniel Levy

Preprint available at arXiv:2510.22186

Problem Setup

Problem: Construct bi-Lipschitz embeddings of the metric space $\hat{V} = V/\sim$ of orbits, $\alpha : \hat{V} \to \mathbb{R}^m$, where $\mathbf{d}([x],[y]) = \inf_{u\in[x],v\in[y]} \|u-v\|$

$$a_0 \mathbf{d}([x],[y]) \leq \|\alpha([x]) - \alpha([y])\|_2 \leq b_0 \mathbf{d}([x],[y]).$$



Today we focus on the case $V = \mathbb{R}^{n\times d}$, $X \sim Y \Leftrightarrow Y = PX$ for some $P \in \mathcal{S}_n$. Motivation: Graph deep learning, Assignment Problems.

A sorting based embedding [BHS22]

Consider: $\beta_{\mathbf{A}} : \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times D}$,

$$\beta_{\mathbf{A}}(\mathbf{X}) := \begin{pmatrix} | & & | \\ \downarrow(\mathbf{X}\mathbf{a}_1) & \dots & \downarrow(\mathbf{X}\mathbf{a}_D) \\ | & & | \end{pmatrix}, \qquad \mathbf{X} \in \mathbb{R}^{n\times d},$$

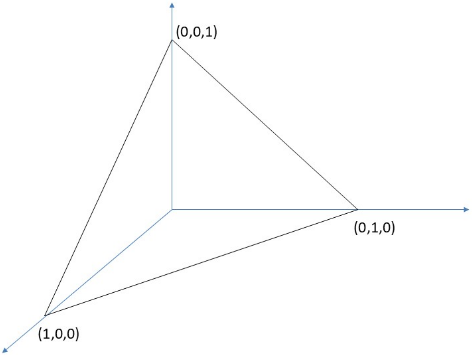
- $\downarrow : \mathbb{R}^n \to \mathbb{R}^n$ denotes sorting vectors in nondecreasing order,
- $(\mathbf{a}_k)_{k=1}^D \in \mathbb{R}^d$ are the columns of $\mathbf{A} \in \mathbb{R}^{d\times D}$.

Note: $\beta_{\mathbf{A}}$ descends through the quotient to $\overline{\beta}_{\mathbf{A}} : \mathbb{R}^{n\times d}/S_n \to \mathbb{R}^{n\times D}$

[BHS22] Radu Balan, Naveed Haghani, and Maneesh Singh. Permutation invariant representations with applications to graph deep learning. March 2022. ACHA 2025.

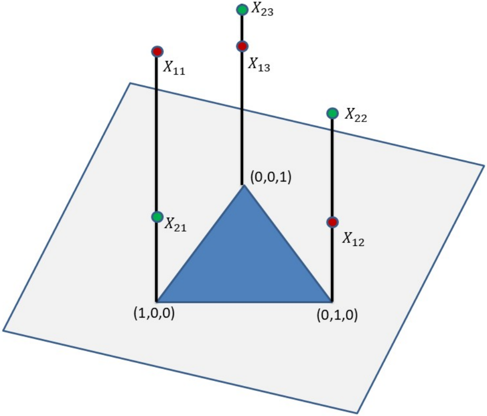
Geometric Interpretation

Consider the case: $d = 3$. Construct the 2-dim simplex:



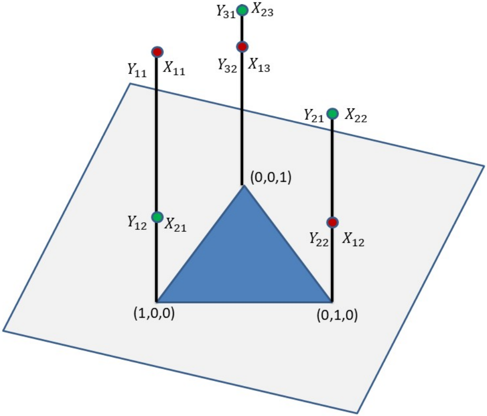
Geometric Interpretation

$d = 3$. Rotate the simplex and place the columns of X (here $n = 2$):



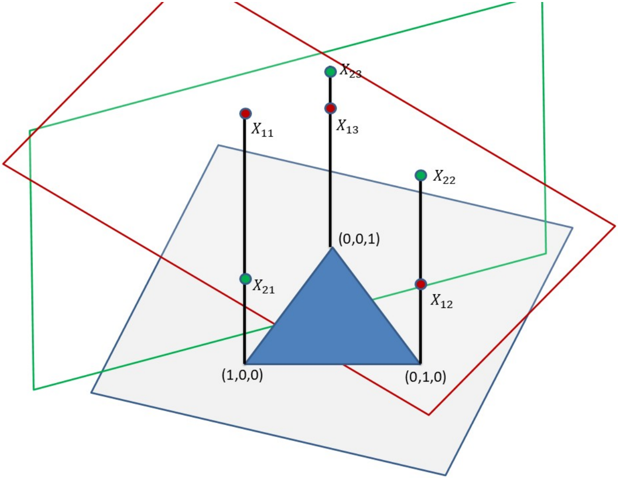
Geometric Interpretation

Place the columns of a non-equivalent matrix Y:



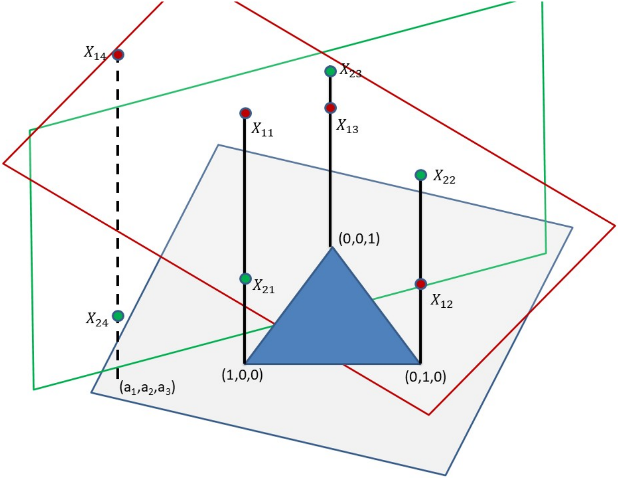
Geometric Interpretation

Construct the $d-1=2$-dim hyperplanes generated by each row of X:



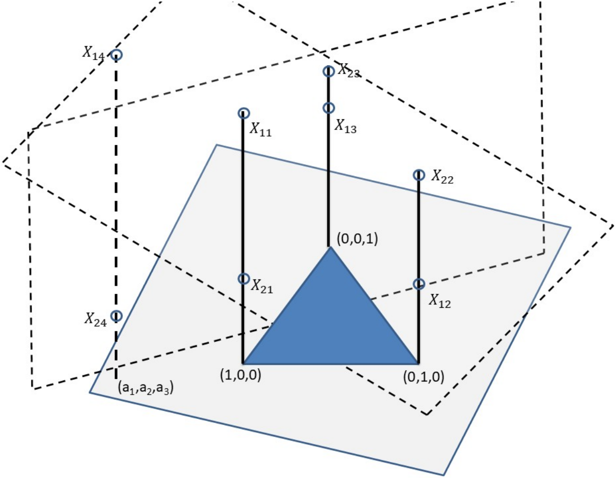
Geometric Interpretation

Sample the hyperplanes at $a = (a_1, a_2, a_3)$ and sort the values.:

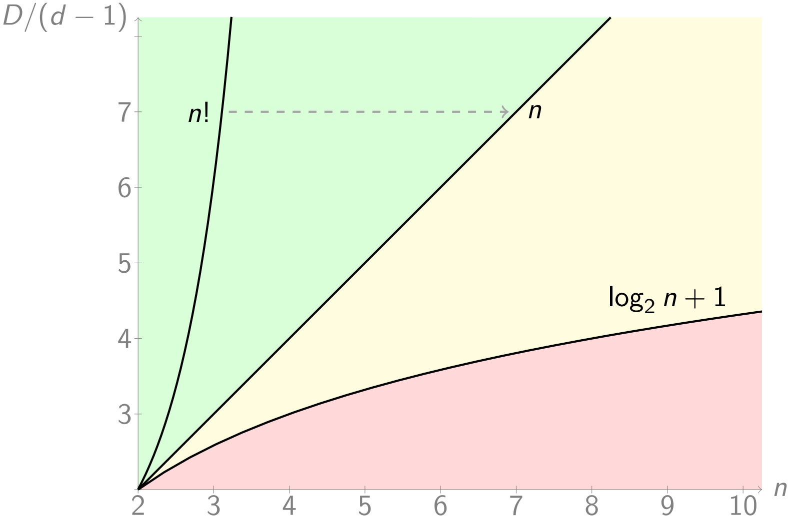


Geometric Interpretation

Can you recover the two (hyper)planes just from the uncolored $n(d+1) = 8$ points? Generically, yes. But $D \geq 2d - 1 = 5$, i.e., another column, for every $[X]$.



Overview of our results



Main results on injectivity

Theorem A:

$$\left\lceil \frac{D}{d-1} \right\rceil \leq \log_2 n + 1 \implies \overline{\beta}_{\mathbf{A}} \textbf{ not injective}$$

(independent of the choice of $\mathbf{A}$).

Theorem B:

$$\frac{D}{d-1} > n \implies \overline{\beta}_{\mathbf{A}} \text{ injective.}$$

($\mathbf{A}$ full spark).

Remark: If $n = 2$, $\overline{\beta}_{\mathbf{A}}$ is injective $\iff$ $(\mathbf{a}_k)_{k=1}^{D}$ form a **phase retrievable** frame for $\mathbb{R}^d$ [BT23]. Hence $D \geq 2d - 1$ is a necessary (and generically sufficient) condition.

[BT23] Radu Balan and Efstratios Tsoukanis. Relationships between the phase retrieval problem and permutation invariant embeddings. In *2023 International Conference on Sampling Theory and Applications (SampTA)*, New Haven, CT, USA, July 2023.

Upper Lipschitz constant

The upper Lipschitz constant of $\overline{\beta}_{\mathbf{A}}$ is equal to the largest singular value of $\mathbf{A}$: $b_0 = \sigma_1(\mathbf{A})$:

$$\|\overline{\beta}_{\mathbf{A}}([x]) - \overline{\beta}_{\mathbf{A}}([y])\|_2 \leq \sigma_1(\mathbf{A})\mathbf{d}([x],[y]).$$

For Gaussian random matrices, with standard i.i.d. entries, $b_0 = \sigma_1(\mathbf{A}) \leq \sqrt{D} + \sqrt{d} + t$ with probability greater than or equal to $1 - 2exp(-c_1 t^2)$.
Hence, we high probability we have $b_0 \sim \sqrt{D} + \sqrt{d}$.

A singular value-based lower Lipschitz bound

Theorem 1

If $D \geq kd((n-1)^2+1)$ for some $k \in \mathbb{N}$, then the lower Lipschitz constant of $\overline{\beta}_{\mathbf{A}}$ is greater than or equal to

$$a_0 \geq \min_{\substack{I \subset [D] \\ |I| = kd}} \sigma_d(\mathbf{A}(I)).$$

Note: k is an integer that can be optimized by user. For Gaussian matrices, using Gordon's theorem we obtained that, for n large enough

$$\mathbb{E}[a_0] \geq \sqrt{\frac{\pi}{8}} \frac{\sqrt{D}}{((n-1)^2+1)^{3/2}} - \sqrt{d} \quad , \quad \mathbb{E}[b_0] \leq \sqrt{D} + \sqrt{d}$$

For $D \gg n^4$, the $\sqrt{d}$ term is neglected and distortion is bounded by

$$\frac{\mathbb{E}[b_0]}{\mathbb{E}[a_0]} \leq \sqrt{\frac{\pi}{8}} n^3$$

Lower Lipschitz constant based on projective uniformity

Definition Matrix $\mathbf{A} \in \mathbb{R}^{d\times D}$ satisfies (m,δ)-*projective uniformity* [CIMP24] if $\forall \mathbf{x} \in S^{d-1}$, $\downarrow(|\mathbf{A}^\top \mathbf{x}|)_{D-m+1} \geq \delta$, where $m \in [D]$ and $\delta > 0$; i.e., for every unit vector x, the m-th smallest entry of $\mathbf{A}^T\mathbf{x}$ exceeds δ.

Theorem 2

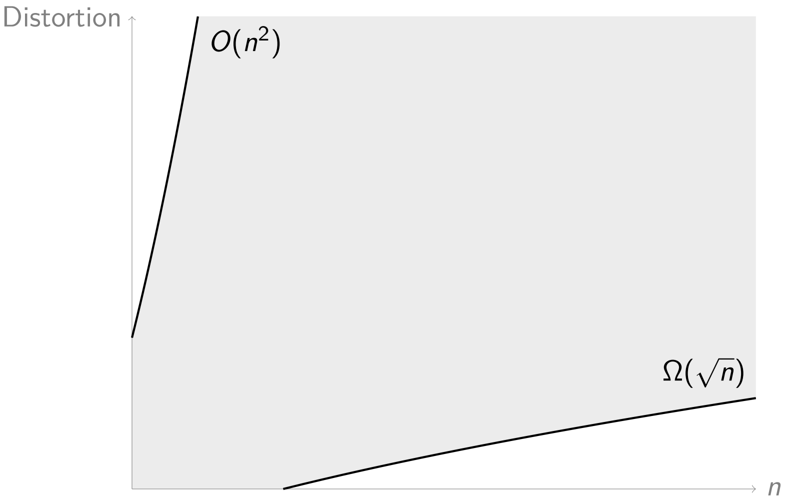
Let $\mathbf{A} \in \mathbb{R}^{d\times D}$ satisfy (m,δ)-projective uniformity with $\delta > 0$ and $m \in [D]$ such that $n^2(m-1) \leq D$. Then, the lower Lipschitz constant of $\overline{\beta}_{\mathbf{A}}$ is greater or equal than $\delta\sqrt{D - n^2(m-1)}$.

Theorem 3

Let $\mathbf{A} \in \mathbb{R}^{d\times D}$ be a matrix with independent standard normal entries. Then, the lower Lipschitz constant $a_0 \geq \frac{\sqrt{2\pi}}{9\sqrt{3}}\frac{\sqrt{D}}{n^2}$ and the distortion of $\overline{\beta}_{\mathbf{A}}$ is in $O(n^2)$ with probability greater or equal than $1 - 2\exp(-c_1 D) - \exp(-c_2 n^{-4} D)$, where $c_1, c_2 > 0$ are universal constants, provided that $D \gtrsim n^4 d$.

[CIMP24] Cahill, Iverson, Mixon, Packer. Group-invariant max filtering. FoCM 2024

Overview of our distortion bounds



Upper and lower bounds for distortion

Theorem A: A with independent standard normal entries:

$$\mathrm{dist}(\beta_{\mathbf{A}}) \lesssim n^2$$

with overwhelming probability if $D \gtrsim n^4 d$.

Theorem B: A with columns drawn independently from the uniform distribution on S^{d-1}:

$$\mathrm{dist}(\beta_{\mathbf{A}}) \lesssim n^2 \cdot \sqrt{1 + \frac{\log n}{d}}$$

with overwhelming probability if $D \gtrsim n^2 d \log(n(d + \log n))$.

Theorem C:

$$\mathrm{dist}(\beta_{\mathbf{A}}) \gtrsim \sqrt{n}$$

(independent of the choice of **A**).

Thank you!

Questions?

A Universal Embedding

Measure Space Embedding

Consider the map

$$\mu : \widehat{\mathbb{R}^{n\times d}} \to \mathcal{P}(\mathbb{R}^d) \quad , \quad \mu(X)(x) = \frac{1}{n}\sum_{k=1}^{n} \delta(x - x_k)$$

where $\mathcal{P}(\mathbb{R}^d)$ denotes the convex set of probability measures over $\mathbb{R}^d$, and δ denotes the Dirac measure. x_k is the k^{th} row of X.
Clearly $\mu(X') = \mu(X)$ iff $X' = PX$ for some $P \in \mathcal{S}_n$.
The Wasserstein-2 distance is equivalent to the natural metric:

$$W_2(\mu(X), \mu(Y))^2 := \inf_{q \in J(\mu(X),\mu(Y))} \mathbb{E}_q[\|x - y\|_2^2] = \min_{P \in \mathcal{S}_n} \|Y - PX\|^2$$

By Kantorovich-Rubinstein theorem, the Wasserstein-1 distance (the Earth moving distance) extends to a norm on the space of signed Borel measures.
Main drawback: $\mathcal{P}(\mathbb{R}^d)$ is infinite dimensional!

Finite Dimensional Embeddings

Idea: “Project” the measure onto a finite dimensional space. This is accomplished by *kernel methods*:
Fix a family of functions $f_1, \cdots, f_m$ and consider:

$$\mu(X) \mapsto \int_{\mathbb{R}^d} f_j(x) d\mu(X) = \frac{1}{n} \sum_{k=1}^{n} f_j(x_k) \quad , \quad j \in [m]$$

Finite Dimensional Embeddings

Idea: "Project" the measure onto a finite dimensional space. This is accomplished by *kernel methods*:
Fix a family of functions $f_1, \cdots, f_m$ and consider:

$$\mu(X) \mapsto \int_{\mathbb{R}^d} f_j(x) d\mu(X) = \frac{1}{n} \sum_{k=1}^{n} f_j(x_k) \quad , \quad j \in [m]$$

Possible choices:

1. Polynomial embeddings: $\mathbb{R}[X]^{\mathcal{S}_n}$, ring of invariant polynomials; [Lipman&al.],[Peyré&al.],[Sanay&al.],[Kemper book] ...
2. Gaussian kernels: $f_j(x) = exp(-\|x - a_j\|^2/\sigma_j^2)$; [Gilmer&al.],[Zaheer&al.], [Vinyals&al.],...
3. Fourier kernels (cmplx embd): $f_j(x) = exp(2\pi i \langle x, \omega_j \rangle)$; related to Prony method; [Li&Liao] for bi-Lipschitz estimates.

Main drawback: No global bi-Lipschitz embeddings [Cahill&al.]. Ok on (some) compacts.

The Embedding Problem

Notations (2)

Definition 5.1

Fix $X \in \mathbb{R}^{n\times d}$. A matrix $A \in \mathbb{R}^{d\times D}$ is called admissible *for X if $\beta_A^{-1}(\beta_A(X)) = \hat{X}$. In other words, if $Y \in \mathbb{R}^{n\times d}$ so that $\downarrow(XA) = \downarrow(YA)$ then there is $\Pi \in \mathcal{S}_n$ sot that $Y = \Pi X$.*

We denote by $\mathcal{A}_{d,D}(X)$ (or $\mathcal{A}(X)$) the set of admissible keys for X.

Definition 5.2

Fix $A \in \mathbb{R}^{d\times D}$. A data matrix $X \in \mathbb{R}^{n\times d}$ is said separated by A *if $A \in \mathcal{A}(X)$.*

We let $\mathcal{S}(A)$ denote the set of data matrices separated by A.
The key A is universal iff $\mathcal{S}(A) = \mathbb{R}^{n\times d}$.
The Problem: Design universal keys.

Genericity Results for $d \geq 2$

Admissible keys

Theorem 5.3

Let $X \in \mathbb{R}^{n\times d}$. For any $D \geq d+1$ the set $\mathcal{A}_{d,D}(X)$ of admissible keys for X is dense in $\mathbb{R}^{d\times D}$ with respect to Euclidean topology, and it is generic with respect to Zariski topology. In particular, $\mathbb{R}^{d\times D} \setminus \mathcal{A}_{d,D}(X)$ has Lebesgue measure 0, i.e., almost every key is admissible for X.

Proof

It is sufficient to consider the case $D = d+1$. Also, it is sufficient to analyze the case $A = [I_d \;\; b]$ and to show that a generic $b \in \mathbb{R}^d$ defines an admissible key. The vector $b \in \mathbb{R}^d$ does **not** define an admissible key if there are $\Xi, \Pi_1, \cdots, \Pi_d \in S_n$ so that for $Y = [\Pi_1 x_1, \cdots, \Pi_d x_d]$,

$$Yb = \Xi X b \ \ \text{but} \ \ Y - \Pi X \neq 0 \ , \ \forall \Pi \in \mathcal{S}_n$$

Define the linear operator

Genericity Results for $d \geq 2$

Admissible keys

Proof - cont'd

Let

$$\mathcal{P} = \left\{ (\Pi_1, \cdots, \Pi_d) \in (\mathcal{S}_n)^d \ \ \forall \Pi \in \mathcal{S}_n, \exists k \in [d] \ s.t. \ (\Pi - \Pi_k)x_k \neq 0 \right\}$$

Then

$$\{b \in \mathbb{R}^d : [I_d \ b] \text{ not admissible for } X\} = \bigcup_{(\Xi;\Pi_1,\cdots,\Pi_d) \in \mathcal{S}_n \times \mathcal{P}} \ker(B(\Xi; \Pi_1, \cdots, \Pi$$

It is now sufficient to show that each null space has dimension less than d. Indeed, the alternative would mean $B(\Xi; \Pi_1, \cdots, \Pi_d) = 0$ but this would imply $(\Pi_1, \cdots, \Pi_d) \notin \mathcal{P}$. □

Non-Universality of vector keys

Insufficiency of a single vector key

The following is a no-go result, which shows that there is no universal single vector key for data matrices tall enough.

Proposition 5.4

If $d \geq 2$ and $n \geq 3$,

$$\bigcup_{X \in \mathbb{R}^{n \times d}} \{b \in \mathbb{R}^d : \; A = [I_d \;\; b] \text{ not admissible for} X\} = \mathbb{R}^d.$$

Consequently,

$$\bigcap_{X \in \mathbb{R}^{n \times d}} \mathcal{A}_{d,d+1}(X) = \emptyset.$$

On the other hand, for $n = 2$, $d = 2$, any vector $b \in \mathbb{R}^2$ with $b_1 b_2 \neq 0$ defines a universal key $A = [I_2 \;\; b]$.

Non-Universality of vector keys

Insufficiency of a single vector key - cont'd

Proof

To show the result, it is sufficient to consider a counterexample for $n = 3$, $d = 2$, with key $b = [1, 1]^T$.

$$X = \begin{bmatrix} 1 & -1 \\ -1 & 0 \\ 0 & 1 \end{bmatrix} \quad , \quad Y = \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix}$$

Then $Xb = [0, -1, 1]^T$ and $Yb = [1, 0, -1]^T$, yet $X \not\sim Y$. Thus $[I_2 \ b]$ is not admissible for X.

Then note if $a \in \mathbb{R}^d$ so that $[I_d \ a]$ is admissible for X then for any $P \in S_d$ and L an invertible $d \times d$ diagonal matrix, $L^{-1}P^T A \in \mathcal{A}_{d,1}(XPL)$. This shows how for any $b \in \mathbb{R}^2$, one can construct $X \in \mathbb{R}^{3\times 2}$ so that $b \notin \mathcal{A}_{2,1}(X)$.

For $n > 3$ or $d > 2$, proof follows by embedding this example.

Genericity Results for $d \geq 2$

Admissible Data Matrices

Theorem 5.5

Assume $a \in \mathbb{R}^d$ is a vector with non-vanishing entries, i.e., $a_1 a_2 \cdots a_d \neq 0$. Then for any $n \geq 1$, $\mathcal{S}([I_d \;\; a])$ is dense in $\mathbb{R}^{n \times d}$ and includes an open dense set with respect to Zariski topology. In particular, $\mathbb{R}^{n \times d} \setminus \mathcal{S}([I_d \;\; a])$ has Lebesgue measure 0, i.e., almost every data matrix X is separated by the vector key a.

Genericity Results for $d \geq 2$

Admissible Data Matrices

Theorem 5.5

Assume $a \in \mathbb{R}^d$ is a vector with non-vanishing entries, i.e., $a_1 a_2 \cdots a_d \neq 0$. Then for any $n \geq 1$, $\mathcal{S}([I_d \;\; a])$ is dense in $\mathbb{R}^{n \times d}$ and includes an open dense set with respect to Zariski topology. In particular, $\mathbb{R}^{n \times d} \setminus \mathcal{S}([I_d \;\; a])$ has Lebesgue measure 0, i.e., almost every data matrix X is separated by the vector key a.

Corollary 5.6

Assume $A \in \mathbb{R}^{d \times (D-d)}$ is a matrix such that at least one column has non-vanishing entries. Then for any $n \geq 1$, $\mathcal{S}([I_d \;\; A])$ is dense in $\mathbb{R}^{n \times d}$ and is generic with respect to Zariski topology. In particular, $\mathbb{R}^{n \times d} \setminus \mathcal{S}([I_d \;\; A])$ has Lebesgue measure 0, i.e., almost every data matrix X is separated by the matrix key $[I_d \;\; A]$.

Proof that $\mathcal{S}([I_d\ A])$ is generic

The case $D > d$

Assume $A \in \mathbb{R}^{d\times(D-d)}$ satisfies $A_{1,k}A_{2,k}\cdots A_{d,k} \neq 0$ for some $k \in [D-d]$. The set of non-separated data matrices $X \in \mathbb{R}^{n\times d}$ (i.e., the complement of $\mathcal{S}([I_d\ A])$) factors as follows:

$$\mathbb{R}^{n\times d}\setminus\mathcal{S}([I_d\ A]) = \bigcup_{(\Xi_1,\cdots,\Xi_{D-d};\Pi_1,\cdots,\Pi_d)\in(\mathcal{S}_n)^D} \Big(ker\ L(\Xi_1,\cdots,\Xi_{D-d};\Pi_1,\cdots,\Pi_d;A$$

$$\setminus \bigcup_{\Pi\in\mathcal{S}_n} \ker\ M(\Pi,\Pi_1,\cdots,\Pi_d)\Big) \quad (*)$$

where, with $A = [a_1,\cdots,a_{D-d}]$, $X = [x_1,\cdots,x_d]$:

$$L(\Xi_1,\cdots,\Xi_{D-d};\Pi_1,\cdots,\Pi_d;A):\mathbb{R}^{n\times d}\to\mathbb{R}^{n\times D-d}\ ,\ (L((\ldots)X)_k=[(\Xi_k-\Pi_1)x_1,\cdots,(\Xi_k-\Pi_d)x_d]a_k\ ,\ k\in[D-$$

$$M(\Pi,\Pi_1,\cdots,\Pi_d):\mathbb{R}^{n\times d}\to\mathbb{R}^{n\times d}\ ,\ M(\Pi,\Pi_1,\cdots,\Pi_d)X=[(\Pi-\Pi_1)x_1,\cdots,(\Pi-\Pi_d)x_d]$$

Proof that $\mathcal{S}(A)$ is generic

cont'd

1. The outer union can be reduced by noting that on the "diagonal" Δ,

$$\Delta = \{(\Xi_1, \cdots, \Xi_{D-d}; \Pi_1, \cdots, \Pi_d) \in (\mathcal{S}_n)^D \quad , \quad \Pi_1 = \Pi_2 = \cdots = \Pi_d\}$$

$$M(\Pi_1, \Pi_1, \cdots, \Pi_d) = 0 \rightarrow \bigcup_{\Pi \in \mathcal{S}_n} \ker M(\Pi, \Pi_1, \cdots, \Pi_d) = \mathbb{R}^{n \times d}$$

2. If $(\Xi_1, \cdots, \Xi_{D-d}; \Pi_1, \cdots, \Pi_d) \in (\mathcal{S}_n)^D \setminus \Delta$ then for every $k \in [D-d]$ there is $j \in [d]$ such that $\Xi_k - \Pi_j \neq 0$. In particular choose the k column of A that is non-vanishing. Let $x_j \in \mathbb{R}^n$ so that $(\Xi_k - \Pi_j)x_j \neq 0$. Consider the matrix $X = [0, \cdots, 0, x_j, 0, \cdots, 0]$ where x_j is the only non identically 0 column. Claim: $X \notin \ker L(\Xi_1, ..., \Pi_d; A)$. Indeed, the resulting k column of $L()X$ is $A_{j,k}(\Xi_k - \Pi_j)x_j \neq 0$. It follows that

$$\dim \ker L(\Xi_1, \cdots, \Xi_{D-d}; \Pi_1, \cdots, \Pi_d; A) < nd$$

Hence $\mathbb{R}^{n \times d} \setminus \mathcal{S}([I_d \; A])$ is a finite union of subsets of closed linear spaces properly included in $\mathbb{R}^{n \times d}$. This proves the theorem. $\square$

Additional Relations

Note the following relationship and matrix representation of X when matrices are column-stacked:

$$M(\Pi, \Pi_1, \cdots, \Pi_d) = L(\Pi, \cdots, \Pi; \Pi_1, \cdots, \Pi_d; I)$$

$$L \equiv \left[\begin{array}{cccc} A_{1,1}(\Xi_1 - \Pi_1) & A_{2,1}(\Xi_1 - \Pi_2) & \cdots & A_{d,1}(\Xi_1 - \Pi_d) \\ A_{1,2}(\Xi_2 - \Pi_1) & A_{2,2}(\Xi_2 - \Pi_2) & \cdots & A_{d,2}(\Xi_2 - \Pi_d) \\ \vdots & \vdots & \ddots & \vdots \\ A_{1,D-d}(\Xi_{D-d} - \Pi_1) & A_{2,D-d}(\Xi_{D-d} - \Pi_2) & \cdots & A_{d,D-d}(\Xi_{D-d} - \Pi_d \end{array}$$

a $n(D-d) \times nd$ matrix.