Isometry pursuit

Samson Koelle
Amazon
koelle@amazon.com

Marina Meila
Department of Statistics
University of Washington
mmp@uw.edu

Abstract

Isometry pursuit is a convex algorithm for identifying orthonormal column-submatrices of wide matrices. It consists of a novel normalization method followed by multitask basis pursuit. Applied to Jacobians of putative coordinate functions, it helps identity isometric embeddings from within interpretable dictionaries. We provide theoretical and experimental results justifying this method. For problems involving coordinate selection and diversification, it offers a synergistic alternative to greedy and brute force search.

1 Introduction

Many real-world problems may be abstracted as selecting a subset of the columns of a matrix representing stochastic observations or analytically exact data. This paper focuses on a simple such problem that appears in interpretable learning and diversification. Given a rank D matrix $X \in \mathbb{R}^{D \times P}$ with $P > D$, select a square submatrix $X_{.S}$ where subset $S \subset P$ satisfies $|S| = D$ that is as orthonormal as possible.

This problem arises in interpretable learning specifically because while the coordinate functions of a given feature space may have no intrinsic meaning, it is sometimes possible to generate a dictionary of interpretable features which may be considered as potential parametrizing coordinates. When this is the case, selection of candidate interpretable features as coordinates can take the above form. While implementations vary across data and algorithmic domains, identification of such coordinates generally aids mechanistic understanding, generative control, and statistical efficiency.

This paper shows that an adapted version of the algorithm in Koelle et al. [1] leads to a convex procedure that can improve upon greedy approaches such as those in Cai and Wang [2], Chen and Meila [3], Kohli et al. [4], Jones et al. [5] for finding isometries. The insight leading to isometry pursuit is that multitask basis pursuit applied to an appropriately normalized X selects orthonormal submatrices. Given vectors in $\mathbb{R}^D$, the normalization log-symmetrizes length and favors those closer to unit length, while basis pursuit favors those which are orthogonal. Our results formalize this intuition within a limited setting, and show the usefulness of isometry pursuit as a trimming procedure prior to brute force search for diversification and interpretable coordinate selection. We also introduce a novel ground truth objective function to measure the success of our algorithm against, and discuss the reasonableness of this trimming procedure.

2 Background

Our algorithm is motivated by spectral and convex analysis.

[1]Work conducted outside of Amazon.
[2]Code is available at `https://github.com/sjkoelle/isometry-pursuit`.

Interpretable AI: Past, Present and Future Workshop at NeurIPS 2024

2.1 Problem

Our goal is, given a matrix $X \in \mathbb{R}^{D \times P}$, to select a subset $S \subset [P]$ with $|S| = D$ such that $X_{.S}$ is as orthonormal as possible in a computationally efficient way. To this end, we define a ground truth loss function that measures orthonormalness, and then introduce a surrogate loss function that convexifies the problem so that it may be efficiently solved.

2.2 Interpretability and isometry

Our motivating example is the selection of data representations from within sets of putative coordinates: the columns of a provided wide matrix. Compared with Sparse PCA [6, 7, 8], we seek a low-dimensional representation from the set of these column vectors rather than their span.

This method applies to interpretability, for which parsimony is at a premium. Interpretability arises through comparison of data with what is known to be important in the domain of the problem. This knowledge often takes the form of a functional dictionary. Evaluation of independence of dictionary features arises in numerous scenarios [9, 10, 11]. The requirement that dictionary features be full rank has been called functional independence [10] or feature decomposability [12], with connection between dictionary rank and independence via the implicit function theorem. Besides independence, the metric properties of such dictionary elements are of natural interest. This is formalized through the notion of differential.

Definition 1 *The* ***differential*** *of a smooth map $\phi : \mathcal{M} \rightarrow \mathcal{N}$ between D dimensional manifolds $\mathcal{M} \subseteq \mathbb{R}^B$ and $\mathcal{N} \subseteq \mathbb{R}^P$ is a map in tangent bases $x_1 \dots x_D$ of $T_\xi \mathcal{M}$ and $y_1 \dots y_D$ of $T_{\phi(\xi)} \mathcal{N}$ consisting of entries*

$$D\phi(\xi) = \begin{bmatrix} \frac{\partial \phi_1}{\partial x_1}(\xi) & \dots & \frac{\partial \phi_1}{\partial x_D}(\xi) \\ \vdots & & \vdots \\ \frac{\partial \phi_D}{\partial x_1}(\xi) & \dots & \frac{\partial \phi_D}{\partial x_D}(\xi) \end{bmatrix}. \tag{1}$$

It is not always necessary to explicitly estimate tangent spaces when applying this definition. The most commonly encountered manifolds are vector spaces for which the tangent spaces are trivial. This is the case for full-rank tabular data, for which isometry has a natural interpretation as a type of diversification, and often for the latent spaces of deep learning models.

Definition 2 *A map ϕ between D dimensional submanifolds with inherited Euclidean metric $\mathcal{M} \subseteq R^{B_\alpha}$ and $\mathcal{N} \subseteq R^{B_\beta}$ ϕ is an* ***isometry at a point*** *$\xi \in \mathcal{M}$ if*

$$D\phi(\xi)^T D\phi(\xi) = I_D. \tag{2}$$

That is, ϕ is an isometry at ξ if $D\phi(\xi)$ is orthonormal.

The applications of pointwise isometry are themselves manifold. Pointwise isometric embeddings faithfully preserve high-dimensional geometry. For example, Local Tangent Space Alignment [13], Multidimensional Scaling [14] and Isomap [15] non-parametrically estimate embeddings that are as isometric as possible. Another approach stitches together pointwise isometries selected from a dictionary to form global embeddings [4]. The method is particularly relevant since it constructs such isometries through greedy search, with putative dictionary features added one at a time.

That $D\phi$ is orthonormal has several equivalent formulations. The one motivating our ground truth loss function comes from spectral analysis.

Proposition 1 *The singular values $\sigma_1 \dots \sigma_D$ are equal to 1 if and only if $U \in \mathbb{R}^{D \times D}$ is orthonormal.*

On the other hand, the formulation that motivates our convex approach is that orthonormal matrices consist of D coordinate features whose gradients are orthogonal and evenly varying.

Proposition 2 *The component vectors $u_1 \dots u_D \in \mathbb{R}^B$ form a orthonormal matrix if and only if, for all $d_1, d_2 \in [D]$, $\langle u_{d_1}, u_{d_2} \rangle = \begin{cases} 1 & d_1 = d_2 \\ 0 & d_1 \neq d_2 \end{cases}$.*

2.3 Subset selection

Given a matrix $X \in \mathbb{R}^{D \times P}$, we compare algorithmic paradigms for solving problems of the form

$$\arg \min_{S \in \binom{[P]}{d}} l(X_{.S}) \tag{3}$$

where $\binom{[P]}{d} = \{A \subseteq [P] : |A| = d\}$. Brute force algorithms consider all possible solutions. These algorithms are conceptually simple, but have the often prohibitive time complexity $O(C_l P^D)$ where C_l is the cost of evaluating l. Greedy algorithms consist of iteratively adding one element at a time to S. This algorithms have time complexity $O(C_l PD)$ and so are computationally more efficient than brute force algorithms, but can get stuck in local minima. Formal definitions are given in Section 6.1.

Sometimes, it is possible to introduce an objective which convexifies problems of the above form. Solutions

$$\arg\min f(\beta) : Y = X\beta \tag{4}$$

to the overcomplete regression problem $Y = X\beta$ are a classic example [16]. When $f(\beta) = \|\beta\|_0$, this problem is non-convex, and must be solved via greedy or brute algorithms, but when $f(\beta) = \|\beta\|_1$, the problem is convex, and may be solved efficiently via interior-point methods. When the equality constraint is relaxed, Lagrangian duality may be used to reformulate as a so-called Lasso problem, which leads to an even richer set of optimization algorithms.

The form of basis pursuit that we apply is inspired by the group basis pursuit approach in Koelle et al. [10]. In group basis pursuit (which we call multitask basis pursuit when grouping is dependent only on the structure of matrix-valued response variable y) the objective function is $f(\beta) = \|\beta\|_{1,2} := \sum_{p=1}^{P} \|\beta_{p.}\|_2$ [17, 18, 19] This objective creates joint sparsity across entire rows of $\beta_{p.}$ and was used in [10] to select between sets of interpretable features.

3 Method

We adapt the group lasso paradigm used to select independent dictionary elements in Koelle et al. [10, 1] to select pointwise isometries from a dictionary. We first define a ground truth objective computable via brute and greedy algorithms that is uniquely minimized by orthonormal matrices. We then define the combination of normalization and multitask basis pursuit that approximates this ground truth loss function. We finally give a brute post-processing method for ensuring that the solution is D sparse.

3.1 Ground truth

We'd like a ground truth objective to be minimized uniquely by orthonormal matrices, invariant under rotation, and depend on all changes in the matrix. Deformation [4] and nuclear norm [20] use only a subset of the differential's information and are not uniquely minimized at unitarity, respectively. We therefore introduce an alternative ground truth objective that satisfies the above desiderata and has convenient connections to isometry pursuit.

This objective is

$$l_c : \mathbb{R}^{D \times P} \to \mathbb{R}^+ \tag{5}$$

$$X \mapsto \sum_{d=1}^{D} g(\sigma_d(X), c) \tag{6}$$

where $\sigma_d(X)$ is the d-th singular value of X and

$$g : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+ \tag{7}$$

$$t, c \mapsto \frac{e^{t^c} + e^{t^{-c}}}{2e}. \tag{8}$$

Using Proposition 1, we can check that g is uniquely maximized by orthonormal matrices. Moreover, $g(X^{-1}) = g(X)$ and g is convex. Figure 1 gives a graph of l_c when $D = 1$ and compares it with that produced by basis pursuit after normalization as in Section 3.2.

Our ground truth program is therefore

$$\widehat{S}_{GT} = \arg \min_{S \in \binom{[P]}{d}} l_c(X_{.S}). \tag{9}$$

Regardless of the convexity of l_c, brute combinatorial search over $[P]$ is inherently non-convex.

3.2 Normalization

Since basis pursuit methods tend to select longer vectors, selection of orthonormal submatrices requires normalization such that both long and short candidate basis vectors are penalized in the subsequent regression. We introduce the following definition.

Definition 3 (Symmetric normalization) *A function $q : \mathbb{R}^D \to \mathbb{R}^+$ is a symmetric normalization if*

$$\arg \max_{v \in \mathbb{R}^D} q(v) = \{v : \|v\|_2 = 1\} \tag{10}$$

$$q(v) = q(\frac{v}{\|v\|_2^2}) \tag{11}$$

$$q(v_1) = q(v_2) \; \forall \; v_1, v_2 \in \mathbb{R}^D : \|v_1\|_2 = \|v_2\|_2. \tag{12}$$

We use such functions to normalize vector length in such a way that vectors of length 1 prior to normalization have longest length after normalization, vectors in general are shrunk proportionately to their deviation from 1. That is, we normalize vectors by

$$n : \mathbb{R}^D \to \mathbb{R}^D \tag{13}$$

$$v \mapsto q(v)v \tag{14}$$

and matrices by

$$w : \mathbb{R}^{D \times P} \to \mathbb{R}^D \tag{15}$$

$$X_{.p} \mapsto n(X_{.p}) \; \forall \; p \in [P]. \tag{16}$$

In particular, given $c > 0$, we choose q as follows.

$$q_c : \mathbb{R}^D \to \mathbb{R}^+ \tag{17}$$

$$v \mapsto \frac{e^{\|v\|_2^c} + e^{\|v\|_2^{-c}}}{2e}. \tag{18}$$

Besides satisfying the conditions in Definition 3, this normalization has some additional nice properties. First, q is convex. Second, it grows asymptotically log-linearly. Third, while $\exp(-|\log t|) = \exp(-\max(t, 1/t))$ is a seemingly natural choice for normalization, it is non smooth, and the LogSumExp [20] replacement of $\max(t, 1/t)$ with $\log(\exp(t) + \exp(1/t))$ simplifies to 17 upon exponentiation. Finally, the parameter c grants control over the width of the basin, which is important in avoiding numerical issues arising close to 0 and ∞.

3.3 Isometry pursuit

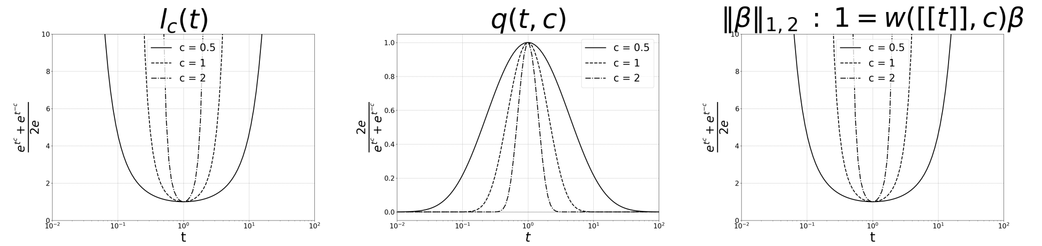
Isometry pursuit is the application of multitask basis pursuit to the normalized design matrix $w(X, c)$ to identify submatrices of X that are as orthonormal as possible. Define the multitask basis pursuit penalty

$$\| \cdot \|_{1,2} : \mathbb{R}^{P \times D} \to \mathbb{R}^+ \tag{19}$$

$$\beta \mapsto \sum_{p=1}^{P} \|\beta_{p.}\|_2. \tag{20}$$

Given a matrix $Y \in \mathbb{R}^{D \times E}$, the multitask basis pursuit solution is

$$\widehat{\beta}_{MBP}(X, Y) := \arg \min_{\beta \in \mathbb{R}^{P \times E}} \|\beta\|_{1,2} \; : \; Y = X\beta. \tag{21}$$



(a) Ground truth loss scaling function g as a function of t.

(b) Length as a function of t

(c) Basis pursuit loss as a function of t

Figure 1: Plots of ground truth loss, normalized length, and basis pursuit loss for different values of c in the one-dimensional case $D = 1$. The two losses are equivalent in the one-dimensional case.

Isometry pursuit is then given by

$$\widehat{\beta}_c(X) := \widehat{\beta}_{MBP}(w(X, c), I_D) \tag{22}$$

where I_D is the D dimensional identity matrix and recovered functions are the indices of the dictionary elements with non-zero coefficients. That is, they are given by $S(\beta)$ where

$$S : \mathbb{R}^{p \times d} \to \binom{[P]}{d} \tag{23}$$

$$\beta \mapsto \{p \in [P] : \|\beta_{p.}\| > 0\}. \tag{24}$$

ISOMETRYPURSUIT(Matrix $X \in \mathbb{R}^{D \times P}$, scaling constant c)

1: Normalize $X_c = w(X, c)$
2: Optimize $\widehat{\beta} = \widehat{\beta}_{MBP}(X_c, I_D)$
3: **Output** $\widehat{S} = S(\widehat{\beta})$

3.4 Theory

The intuition behind our application of multitask basis pursuit in our setting is that submatrices consisting of vectors which are closer to 1 in length and more orthogonal will have smaller loss. A key initial theoretical assertion is that ISOMETRYPURSUIT is invariant to choice of basis for X.

Proposition 3 *Let $U \in \mathbb{R}^{D \times D}$ be orthonormal. Then $S(\widehat{\beta}(UX)) = S(\widehat{\beta}(X))$.*

A proof is given in Section 6.2.1. This has as an immediate corollary that we may replace I_D in the constraint by any orthonormal $D \times D$ matrix.

We also claim that the conditions of the consequent of Proposition 2 are satisfied by minimizers of the multitask basis pursuit objective applied to suitably normalized matrices in the special case where both such a submatrix exists and $|S| = D$.

Proposition 4 *Let w_c be a normalization satisfying the conditions in Definition 3. Then $\arg\min_{X_{.S} \in \mathbb{R}^{D \times D}} \widehat{\beta}_c^D(X)$ is orthonormal. Moreover when X is orthonormal, $(\min_{\beta \in \mathbb{R}^{P \times D}} \|\beta\|_{1,2} : I_D = w(X, c)\beta) = D$.*

While this Proposition falls short of showing that an orthonormal submatrix will be selected should one be present, it provides intuition justifying the preferential efficacy of ISOMETRYPURSUIT on real data. A proof is given in Section 6.2.2.

3.5 Two-stage isometry pursuit

Since cannot in general ensure either that $|\widehat{S}| = D$ or that a orthonormal submatrix $X_{.S}$ exists, we first use the convex problem to prune and then apply brute search upon the substantially reduced feature set.

TwoStageIsometryPursuit(Matrix $X \in \mathbb{R}^{D \times P}$, scaling constant c)

$\widehat{S_{IP}} = \text{IsometryPursuit}(X, c)$
$\widehat{S} = \text{BruteSearch}(X_{.\widehat{S_{IP}}}, l_c)$
Output $\widehat{S}$

Similar two-stage approaches are standard in the Lasso literature [21]. This method forms our practical isometry estimator, is discussed further in Section 5 and Section 6.4.

4 Experiments

Say you are hosting an elegant dinner party, and wish to select a balanced set of wines for drinking and flowers for decoration. We demonstrate TwoStageIsometryPursuit and GreedySearch on the Iris and Wine datasets [22, 23, 24]. This has an intuitive interpretation as selecting diverse elements that reflects the peculiar structure of the diversification problem. Features like *petal width* are rows in X. They are features on the basis of which we may select among the flowers those which are most distinct from another. Thus, in diversification, $P = n$.

We also analyze the Ethanol dataset from Chmiela et al. [25], Koelle et al. [10], but rather than selecting between bourbon and scotch we evaluate a dictionary of interpretable features - bond torsions - for their ability to parameterize the molecular configuration space. In this interpretability use case, columns denote gradients of informative features. We compute Jacoban matrices of putative parametrization functions and project them onto estimated tangent spaces (see Koelle et al. [10] for preprocessing details). Rather than selecting between data points, we are selecting between functions which parameterize the data.

For basis pursuit, we use the SCS interior point solver [26] from CVXPY [27, 28], which is able to push sparse values arbitrarily close to 0 [29]. Statistical replicas for Wine and Iris are created by resampling across P. Due to differences in scales between rows, these are first standardized. For the Wine dataset, even BruteSearch on $\widehat{S}_1$ is prohibitive in $D = 13$, and so we truncate our inputs to $D = 6$. For Ethanol, replicas are created by sampling from data points and their corresponding tangent spaces.

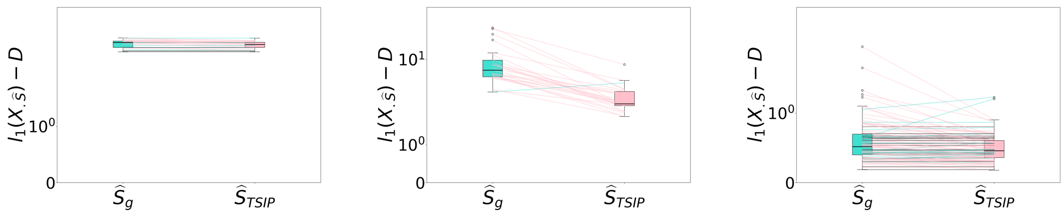
Figure 2 and Table 1 show that the l_1 accrued by the subset $\widehat{S}_G$ estimated using GreedySearch with objective l_1 is higher than that for the subset estimated by TwoStageIsometryPursuit. This effect is statistically significant, but varies across datapoints and datasets. Figure 3 details intermediate support recovery cardinalities from IsometryPursuit. We also evaluated second stage BruteSearch selection after random selection of $\widehat{S}_{IP}$ but do not report it since it often lead to catastrophic failure to satisfy the basis pursuit constraint. Wall-clock runtimes are given in Section 6.5.

Name	D	P	R	c	$l_1(X_{.\widehat{S}_G})$	$\lvert\widehat{S}_{IP}\rvert$	$l_1(X_{.\widehat{S}})$	$P_R(l_1(X_{.\widehat{S}_G}) > l_1(X_{.\widehat{S}}))$	$P_R(l_1(X_{.\widehat{S}_G}) = l_1(X_{.\widehat{S}}))$	$\widehat{P}(\bar{l}_1(X_{.\widehat{S}_G}) > \bar{l}_1(X_{.\widehat{S}}))$
Iris	4	75	25	1	13.8 ± 7.3	7 ± 1	6.9 ± 1.4	0.96	0.	2.4e-05
Wine	6	89	25	1	7.7 ± 0.3	13 ± 2	7.6 ± 0.3	0.64	0.16	6.3e-04
Ethanol	2	756	100	1	2.6 ± 0.3	90 ± 165	2.5 ± 0.2	0.66	0.17	2.1e-05

Table 1: Experimental parameters and results. For Iris and Wine, P is randomly downsampled by a factor of 2 to create R replicates. P_R values are empirical probabilities, while estimated P-values $\widehat{P}$ are computed by paired two-sample T-test on $l_1(X_{.\widehat{S}})$ and $l_1(X_{.\widehat{S}_G})$. For brevity, in this table $\widehat{S} := \widehat{S}_{TSIP}$.

5 Discussion

We have shown that multitask basis pursuit can help select isometric submatrices from appropriately normalized wide-matrices. This approach - isometry pursuit - is convex alternative to greedy methods for selection of orthonormalized features from within a dictionary. Isometry pursuit can be applied to



(a) Wine dataset (b) Iris dataset (c) Ethanol dataset

Figure 2: Isometry losses l_1 for Wine, Iris, and Ethanol datasets across R replicates. Lower brute losses are shown with turquoise, while lower two stage losses are shown with pink. Equal losses are shown with black lines. As detailed in Table 1, losses are generally lower for two-stage isometry pursuit solutions.

diversification and geometrically-faithful coordinate estimation. Our experiments exemplify these applications, but more can be done. One potential application is diversification in recommendation systems [30, 31, 32] and other retrieval systems such as in RAG [33, 34, 35, 36, 37]. Another is decomposing interpretable yet overcomplete dictionaries in transformer residual streams, with each token considered as generating its own tangent space [12, 38].

Compared with the greedy algorithms used in such areas [39, 40, 41, 42, 43, 44, 45, 46, 47, 34], the convex reformulation may add speed and convergence to a global minima. The comparison of greedy [48, 49, 50, 51] and convex [16, 52, 53] basis pursuit formulations has a rich history, and theoretical understanding of the behavior of this approximation is evolving. Diversification problems have been cited as NP-hard, and isometry pursuit can be considered analogous to them in the sense of basis pursuit and the lasso against best subset selection, with the caveat that best subset selection of the basis pursuit loss minimizer isn't totally equivalent to isometry pursuit even though they share the same unique optimum. Characterization of solutions resulting from removal of the restriction $|S| = D$ on the conditions of Proposition 4 may help justify the second selection step. That the solution of a lasso problem can sometimes be a non-singleton convex set containing the sparsest solution is well-known [54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64]. Perhaps surprisingly, it appears empirically that for isometry pursuit that this can occur even when the design matrix is not in general position.

The convergence of SCS algorithm to the 2-norm minimizing solution due to the dual constraint penalty and the convexity of this preimage suggest that a related two stage procedure always succeeds in identifying the brute $\|\|_{1,2}$ minimizer. Related conditions have been discussed in [65, 61], and we examine this topic experimentally in Section 6.4. For this problem, the best subset support estimate - the isometry estimate - appears contained within the loss preimage fiber at the minimum loss.

Algorithmic variants include the multitask lasso [66] extension of our estimator, as well as characterization of D function selection within spaces larger than $\mathbb{R}^D$. Tangent-space specific variants have been studied in more detail in [10, 1] with additional grouping across datapoints, and a corresponding variant of the isometry theorem that missed non-uniqueness was claimed in [67]. Comparison of our loss with curvature - whose presence prohibits D element isometry - could prove fertile, as could comparison with the so-called restricted isometry property used to show guaranteed recovery at fast convergence rates in supervised learning [68, 66].

References

[1] Samson J Koelle, Hanyu Zhang, Octavian-Vlad Murad, and Marina Meila. Consistency of dictionary-based manifold learning. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 4348–4356. PMLR, 2024.

[2] T. Tony Cai and Lie Wang. Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Transactions on Information Theory*, 57(7):4680–4688, 2011. doi: 10.1109/TIT.2011.2146090.

[3] Yu-Chia Chen and Marina Meila. Selecting the independent coordinates of manifolds with large aspect ratios. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/6a10bbd480e4c5573d8f3af73ae0454b-Paper.pdf`.

[4] Dhruv Kohli, Alexander Cloninger, and Gal Mishne. LDLE: Low distortion local eigenmaps. *J. Mach. Learn. Res.*, 22, 2021.

[5] Peter W Jones, Mauro Maggioni, and Raanan Schul. Universal local parametrizations via heat kernels and eigenfunctions of the laplacian. September 2007.

[6] Santanu S Dey, R Mazumder, M Molinaro, and Guanyi Wang. Sparse principal component analysis and its l_1-relaxation. *arXiv: Optimization and Control*, December 2017.

[7] D Bertsimas and Driss Lahlou Kitane. Sparse PCA: A geometric approach. *J. Mach. Learn. Res.*, 24:32:1–32:33, October 2022.

[8] Dimitris Bertsimas, Ryan Cory-Wright, and Jean Pauphilet. Solving Large-Scale sparse PCA to certifiable (near) optimality. *J. Mach. Learn. Res.*, 23(13):1–35, 2022.

[9] Yu-Chia Chen and M Meilă. Selecting the independent coordinates of manifolds with large aspect ratios. *Adv. Neural Inf. Process. Syst.*, abs/1907.01651, July 2019.

[10] Samson J Koelle, Hanyu Zhang, Marina Meila, and Yu-Chia Chen. Manifold coordinates with physical meaning. *J. Mach. Learn. Res.*, 23(133):1–57, 2022.

[11] Jesse He, Tristan Brugère, and Gal Mishne. Product manifold learning with independent coordinate selection. In *Proceedings of the 2nd Annual Workshop on Topology, Algebra, and Geometry in Machine Learning (TAG-ML) at ICML*, June 2023.

[12] Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL `https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html`.

[13] Zhenyue Zhang and Hongyuan Zha. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM J. Scientific Computing*, 26(1):313–338, 2004.

[14] Lisha Chen and Andreas Buja. Local Multidimensional Scaling for nonlinear dimension reduction, graph drawing and proximity analysis. *Journal of the American Statistical Association*, 104(485):209–219, March 2009.

[15] J.B. Tenenbaum, V. Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[16] Scott Shaobing Chen and David L. Donoho and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM REVIEW*, 43(1):129, February 2001.

[17] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Series B Stat. Methodol.*, 68(1):49–67, February 2006.

[18] G Obozinski, B Taskar, and Michael I Jordan. Multi-task feature selection. 2006.

[19] Dit-Yan Yeung and Yu Zhang. A probabilistic framework for learning task relationships in multi-task learning. 2011.

[20] Stephen P Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, March 2004.

[21] Tim Hesterberg, Nam Hee Choi, Lukas Meier, and Chris Fraley. Least angle and ℓ_1 penalized regression: A review. February 2008.

[22] R. A. Fisher. Iris. UCI Machine Learning Repository, 1988. DOI: https://doi.org/10.24432/C56C76.

[23] Stefan Aeberhard and M. Forina. Wine. UCI Machine Learning Repository, 1991. DOI: https://doi.org/10.24432/C5PC7J.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[25] Stefan Chmiela, Huziel E Sauceda, Klaus-Robert Müller, and Alexandre Tkatchenko. Towards exact molecular dynamics simulations with machine-learned force fields. *Nat. Commun.*, 9(1): 3887, September 2018.

[26] Brendan O'Donoghue, Eric Chu, Neal Parikh, and Stephen Boyd. Conic optimization via operator splitting and homogeneous self-dual embedding. *Journal of Optimization Theory and Applications*, 169(3):1042–1068, June 2016. URL `http://stanford.edu/~boyd/papers/scs.html`.

[27] Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

[28] Akshay Agrawal, Robin Verschueren, Steven Diamond, and Stephen Boyd. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60, 2018.

[29] CVXPY Developers. Sparse solution with cvxpy. `https://www.cvxpy.org/examples/applications/sparse_solution.html`. Accessed: 2024-07-11.

[30] Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. *SIGIR Forum*, 51(2):209–210, August 1998.

[31] Qiong Wu, Yong Liu, Chunyan Miao, Yin Zhao, Lu Guan, and Haihong Tang. Recent advances in diversified recommendation. May 2019.

[32] Select by maximal marginal relevance (MMR). `https://python.langchain.com/docs/how_to/example_selectors_mmr/`. Accessed: 2024-11-22.

[33] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv [cs.CL]*, December 2023.

[34] Marc Pickett, Jeremy Hartman, Ayan Kumar Bhowmick, Raquib-Ul Alam, and Aditya Vempaty. Better RAG using relevant information gain. *arXiv [cs.CL]*, July 2024.

[35] Yeonjun In, Sungchul Kim, Ryan A Rossi, Md Mehrab Tanjim, Tong Yu, Ritwik Sinha, and Chanyoung Park. Diversify-verify-adapt: Efficient and robust retrieval-augmented ambiguous question answering. *arXiv [cs.CL]*, September 2024.

[36] Sam Weiss. Enhancing diversity in RAG document retrieval using projection-based techniques. https://medium.com/@samcarlos_14058/enhancing-diversity-in-rag-document-retrieval-using-projection-based-techniques-9fef5422e043 August 2024. Accessed: 2024-11-22.

[37] Get diverse results and comprehensive summaries with vectara's MMR reranker. URL https://www.vectara.com/blog/get-diverse-results-and-comprehensive-summaries-with-vectaras-mmr-reranker. Accessed: 2024-11-22.

[38] Aleksandar Makelov, George Lange, and Neel Nanda. Towards principled evaluations of sparse autoencoders for interpretability and control. *arXiv [cs.LG]*, May 2024.

[39] Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, August 1998. ACM.

[40] Maria C. N. Barioni, Marios Hadjieleftheriou, Marcos R. Vieira, Caetano Traina, Vassilis J. Tsotras, Humberto L. Razente, and Divesh Srivastava. On query result diversification . In *2011 27th IEEE International Conference on Data Engineering (ICDE 2011)*, pages 1163–1174, Los Alamitos, CA, USA, April 2011. IEEE Computer Society. doi: 10.1109/ICDE.2011.5767846. URL https://doi.ieeecomputersociety.org/10.1109/ICDE.2011.5767846.

[41] Marina Drosou and Evaggelia Pitoura. Search result diversification. *SIGMOD Rec.*, 39(1): 41–47, September 2010. ISSN 0163-5808. doi: 10.1145/1860702.1860709. URL https://doi.org/10.1145/1860702.1860709.

[42] Lu Qin, Jeffrey Xu Yu, and Lijun Chang. Diversifying top-k results. *Proceedings VLDB Endowment*, 5(11):1124–1135, July 2012.

[43] Matevž Kunaver and Tomaž Požrl. Diversity in recommender systems – a survey. *Knowledge-Based Systems*, 123:154–162, 2017. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2017.02.009. URL https://www.sciencedirect.com/science/article/pii/S0950705117300680.

[44] Shengbo Guo and Scott Sanner. Probabilistic latent maximal marginal relevance. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, page 833–834, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781450301534. doi: 10.1145/1835449.1835639. URL https://doi.org/10.1145/1835449.1835639.

[45] Mustafa Abdool, Malay Haldar, Prashant Ramanathan, Tyler Sax, Lanbo Zhang, Aamir Manaswala, Lynn Yang, Bradley Turnbull, Qing Zhang, and Thomas Legrand. Managing diversity in airbnb search. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 2952–2960, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3403345. URL https://doi.org/10.1145/3394486.3403345.

[46] Hsiang-Fu Yu, Cho-Jui Hsieh, Qi Lei, and Inderjit S. Dhillon. A greedy approach for budgeted maximum inner product search. In *Neural Information Processing Systems*, 2016. URL https://api.semanticscholar.org/CorpusID:7076785.

[47] Qiang Huang, Yanhao Wang, Yiqun Sun, and Anthony K H Tung. Diversity-aware k-maximum inner product search revisited. *arXiv [cs.IR]*, February 2024.

[48] S.G. Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993. doi: 10.1109/78.258082.

[49] S. Mallat and Z. Zhang. Adaptive time-frequency decomposition with matching pursuits. In *[1992] Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*, pages 7–10, 1992. doi: 10.1109/TFTSA.1992.274245.

[50] Y.C. Pati, R. Rezaiifar, and P.S. Krishnaprasad. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*, pages 40–44 vol.1, 1993. doi: 10.1109/ACSSC.1993.342465.

[51] J.A. Tropp, A.C. Gilbert, and M.J. Strauss. Simultaneous sparse approximation via greedy pursuit. In *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 5, pages v/721–v/724 Vol. 5, 2005. doi: 10.1109/ICASSP.2005.1416405.

[52] Joel A. Tropp. Algorithms for simultaneous sparse approximation. part ii: Convex relaxation. *Signal Processing*, 86(3):589–602, 2006. ISSN 0165-1684. doi: https://doi.org/10.1016/j.sigpro.2005.05.031. URL `https://www.sciencedirect.com/science/article/pii/S0165168405002239`. Sparse Approximations in Signal and Image Processing.

[53] Jie Chen and Xiaoming Huo. Theoretical results on sparse representations of multiple-measurement vectors. *IEEE Transactions on Signal Processing*, 54:4634–4643, 2006. URL `https://api.semanticscholar.org/CorpusID:17333301`.

[54] Michael R. Osborne, Brett Presnell, and Berwin A. Turlach. On the lasso and its dual. *Journal of Computational and Graphical Statistics*, 9:319 – 337, 2000. URL `https://api.semanticscholar.org/CorpusID:14422381`.

[55] Charles Dossal. A necessary and sufficient condition for exact sparse recovery by l1 minimization. *Comptes Rendus Mathematique*, 350(1):117–120, 2012. ISSN 1631-073X. doi: https://doi.org/10.1016/j.crma.2011.12.014. URL `https://www.sciencedirect.com/science/article/pii/S1631073X11003694`.

[56] Stéphane Chrétien and Sébastien Darses. On the generic uniform uniqueness of the lasso estimator. *arXiv: Statistics Theory*, 2011. URL `https://api.semanticscholar.org/CorpusID:88518316`.

[57] Ryan J. Tibshirani. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7: 1456–1490, 2012. URL `https://api.semanticscholar.org/CorpusID:5849668`.

[58] Karl Ewald and Ulrike Schneider. On the distribution, model selection properties and uniqueness of the lasso estimator in low and high dimensions. *Electronic Journal of Statistics*, 2017. URL `https://api.semanticscholar.org/CorpusID:54044415`.

[59] Alnur Ali and Ryan J. Tibshirani. The generalized lasso problem and uniqueness. *Electronic Journal of Statistics*, 2018. URL `https://api.semanticscholar.org/CorpusID:51755233`.

[60] Ulrike Schneider and Patrick Tardivel. The geometry of uniqueness, sparsity and clustering in penalized estimation. *arXiv [math.ST]*, April 2020.

[61] Aaron Mishkin and Mert Pilanci. The solution path of the group lasso. 2022. URL `https://api.semanticscholar.org/CorpusID:259504228`.

[62] Xavier Dupuis and Samuel Vaiter. The geometry of sparse analysis regularization. *SIAM J. Optim.*, 33:842–867, 2019. URL `https://api.semanticscholar.org/CorpusID:195791526`.

[63] Thomas Debarre, Quentin Denoyelle, and Julien Fageot. On the uniqueness of solutions for the basis pursuit in the continuum. *Inverse Problems*, 38, 2020. URL `https://api.semanticscholar.org/CorpusID:246473440`.

[64] Jasper Marijn Everink, Yiqiu Dong, and Martin Skovgaard Andersen. The geometry and well-posedness of sparse regularized linear regression. 2024. URL `https://api.semanticscholar.org/CorpusID:272424099`.

[65] David L. Donoho. For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59, 2006. URL `https://api.semanticscholar.org/CorpusID:8510060`.

[66] T Hastie, R Tibshirani, and M Wainwright. Statistical learning with sparsity: The lasso and generalizations. May 2015.

[67] Samson Jonathan Koelle. *Geometric algorithms for interpretable manifold learning*. Phd thesis, University of Washington, 2022. URL `http://hdl.handle.net/1773/48559`. Statistics [108].

[68] Emmanuel Candes and Terence Tao. Decoding by linear programming. February 2005.

[69] E Anderson, Z Bai, and J Dongarra. Generalized qr factorization and its applications. *Linear Algebra Appl.*, 162-164:243–271, February 1992.

6 Supplement

This section contains algorithms proofs, and experiments in support of the main text.

6.1 Algorithms

We give definitions of the brute and greedy algorithms used in this paper.

BRUTESEARCH(Matrix $X \in \mathbb{R}^{D \times P}$, objective f)

---

for each combination $S \subseteq \{1, 2, \ldots, P\}$ with $|S| = D$ **do**
    Evaluate $f(X_{.S})$
end for
Output the combination S^* that minimizes $f(X_{.S})$

GREEDYSEARCH(Matrix $X \in \mathbb{R}^{D \times P}$, objective f, selected set $S = \emptyset$, current size $d = 0$)

---

if $d = D$ **then**
    Return S
else
    Initialize $S_{\text{best}} = S$
    Initialize $f_{\text{best}} = \infty$
    for each $p \in \{1, 2, \ldots, P\} \setminus S$ **do**
        Evaluate $f(X_{.(S \cup \{p\})})$
        if $f(X_{.(S \cup \{p\})}) < f_{\text{best}}$ **then**
            Update $S_{\text{best}} = S \cup \{p\}$
            Update $f_{\text{best}} = f(\mathcal{X}_{.(S \cup \{p\})})$
        end if
    end for
    Return GREEDYSEARCH(X, f, S_{best}, $d + 1$)
end if

6.2 Proofs

6.2.1 Proof of Proposition 3

In this proof we first show that the penalty $\|\beta\|_{1,2}$ is unchanged by unitary transformation of β.

Proposition 5 *Let $U \in \mathbb{R}^{D\times D}$ be unitary. Then $\|\beta\|_{1,2} = \|\beta U\|$.*

Proof:

$$\|\beta U\|_{1,2} = \sum_{p=1}^{P} \|\beta_{p.} U\| \tag{25}$$
$$= \sum_{p=1}^{P} \|\beta_{p.}\| \tag{26}$$
$$= \|\beta\|_{1,2} \tag{27}$$

□

We then show that this implies that the resultant loss is unchanged by unitary transformation of X.

Proposition 6 *Let $U \in \mathbb{R}^{D\times D}$ be unitary. Then $\widehat{\beta}(UX) = \widehat{\beta}(X)U$.*

Proof:

$$\widehat{\beta}(UX) = \arg\min_{\beta\in\mathbb{R}^{P\times D}} \|\beta\|_{1,2} \ : \ I_D = UX\beta \tag{28}$$
$$= \arg\min_{\beta\in\mathbb{R}^{P\times D}} \|\beta\|_{1,2} \ : \ U^{-1}U = U^{-1}UX\beta U \tag{29}$$
$$= \arg\min_{\beta\in\mathbb{R}^{P\times D}} \|\beta\|_{1,2} \ : \ I_D = X\beta U \tag{30}$$
$$= \arg\min_{\beta\in\mathbb{R}^{P\times D}} \|\beta U\|_{1,2} \ : \ I_D = X\beta U \tag{31}$$
$$= \arg\min_{\beta\in\mathbb{R}^{P\times D}} \|\beta\|_{1,2} \ : \ I_D = X\beta. \tag{32}$$

□

6.2.2 Proof of Proposition 4

Proposition 7 *Let w_c be a normalization satisfying the conditions in Definition 3. Then $\arg\min_{X_{.S}\in\mathbb{R}^{D\times D}} \widehat{\beta}_c^D(X_{.S})$ is orthonormal. Moreover when X is orthonormal, $\min_{\beta\in\mathbb{R}^{P\times D}} \|\beta\|_{1,2} \ : \ I_D = w(X,c)\beta = D$.*

Proof: The value of D is clearly obtained by β orthonormal, since by Proposition 3, for X orthogonal, without loss of generality

$$\beta_{dd'} = \begin{cases} 1 & d = d' \in \{1 \ldots D\} \\ 0 & \text{otherwise} \end{cases}. \tag{33}$$

Thus, we need to show that this is a lower bound on the obtained loss.

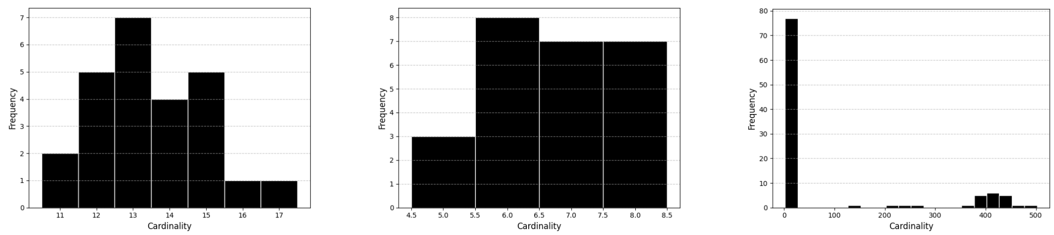
From the conditions in Definition 3, normalized matrices will consist of vectors of maximum length (i.e. 1) if and only if the original matrix also consists of vectors of length 1. Such vectors will clearly result in lower basis pursuit loss, since longer vectors in X require smaller corresponding covectors in β to equal the same result.

Therefore, it remains to show that X consisting of orthogonal vectors of length 1 have lower compared with X consisting of non-orthogonal vectors. Invertible matrices $X_{.S}$ admit QR decompositions $\tilde{X}_{.S} = QR$ where Q and R are orthonormal and upper-triangular matrices, respectively [69].

Denoting Q to be composed of basis vectors $[e^1 \dots e^d]$, the matrix R has form

$$R = \begin{bmatrix} \langle e^1, X_{.S_1} \rangle & \langle e^1, X_{.S_2} \rangle & \dots & \langle e^1, X_{.S_D} \rangle \\ 0 & \langle e^2, X_{.S_2} \rangle & \dots & \langle e^2, X_{.S_D} \rangle \\ 0 & 0 & \dots & \dots \\ 0 & 0 & \dots & \langle e^d, X_{.S_D} \rangle \end{bmatrix}. \tag{34}$$

Thus, $|R_{dd}| \leq \|X_{.S_d}\|_2$, with equality obtained across d only by orthonormal matrices. On the other hand, by Proposition 3, $l(X) = l(R)$ and so $\|\beta\|_{1,2} = \|R^{-1}\|_{1,2}$. Since R is upper triangular it has diagonal elements $\beta_{dd} = R_{dd}^{-1}$ and so $\|\beta_{d.}\| \geq \|X_{.S_d}\|^{-1} = 1$. That is, the penalty accrued by a particular covector in β is bounded from below by 1 - the inverse of the length of the corresponding vector in $X_{.S}$ - with equality occurring only when $X_{.S}$ is orthonormal. $\square$

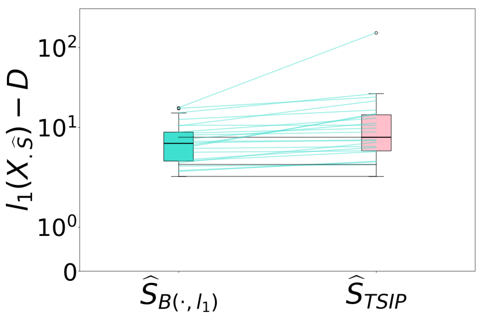


(a) Wine Dataset (b) Iris Dataset (c) Ethanol Dataset

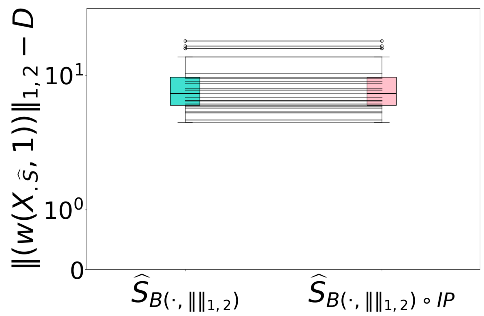
Figure 3: Support Cardinalities for Wine, Iris, and Ethanol datasets

6.3 Support cardinalities

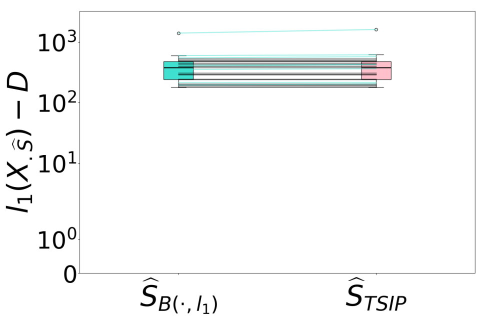
Figure 3 plots the distribution of $|\widehat{S}_{IP}|$ from Table 1 in order to contextualize the reported means. While typically $|\widehat{S}_{IP}| << P$, there are cases for Ethanol where this is not the case that drive up the means.



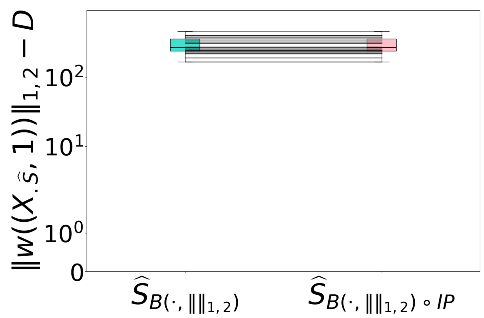
(a) Iris Isometry Losses



(b) Iris Multitask Losses



(c) Wine Isometry Losses



(d) Wine Multitask Losses

Figure 4: Comparison of Isometry and Group Lasso Losses across 25 replicates for randomly downsampled Iris and Wine Datasets with $(P, D) = (4, 15)$ and $(13, 18)$, respectively. Note that this further downsampling compared with Section 4 was necessary to compute global minimizers of BRUTESEARCH. Lower brute losses are shown with turquoise, while lower two stage losses are shown with pink. Equal losses are shown with black lines.

6.4 Proposition 4 deep dive

As mentioned in Section 5, the conditions under which the restriction $S = D$ in Proposition 4 may be relaxed are of theoretical and practical interest. The results in Section 4 show that there are circumstances in which the GREEDYSEARCH performs better than TWOSTAGEISOMETRYPURSUIT, so clearly TWOSTAGEISOMETRYPURSUIT does not always achieve a global optimum. Figure 4 gives results on the line of inquiry about why this is the case based on the reasoning presented in Section 5. In these results a two-stage algorithm achieves the global optimum of a slightly different brute problem, namely brute optimization of the multitask basis pursuit penalty $\|\cdot\|_{1,2}$. That is, brute search on $\|\cdot\|_{1,2}$ gives the same result as the two stage algorithm with brute search on $\|\cdot\|_{1,2}$ subsequent to isometry pursuit. This suggests that failure to select the global optimum by TWOSTAGEISOMETRY-PURSUIT is due to the mismatch between global optimums of brute optimization of the multitask penalty and the isometry loss given certain data. Theoretical formalization, as well as investigation of what data configurations this equivalence holds for, is a logical follow-up.

6.5 Timing

While wall-time of algorithms is a non-theoretical quantity that depends on implementation details, it provides valuable context for practitioners. We therefore report the following runtimes on a 2021 Macbook Pro.

Name	IP	2nd stage brute	Greedy
Iris	1.24 ± 0.02	0.00 ± 0.00	0.02 ± 0.00
Wine	2.32 ± 0.17	0.13 ± 0.12	0.03 ± 0.00
Ethanol	8.38 ± 0.57	0.55 ± 1.08	0.07 ± 0.01

Table 2: Algorithm runtimes in seconds across replicates.