Isometry pursuit

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Isometry pursuit is an algorithm for identifying unitary column-submatrices of wide matrices. It achieves sparsity via use of the multitask basis pursuit penalty. Applied to Jacobians of putative coordinate functions, it helps identity isometric embeddings from within dictionaries. It therefore has relevance to interpretability of learned representations.

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. Given a rank D matrix $\mathcal{X} \in \mathbb{R}^{D \times P}$ with $P > D$, select a square submatrix $\mathcal{X}_{.\mathcal{S}}$ where subset $\mathcal{S} \subset P$ satisfies $|\mathcal{S}| = D$ that is as isometric as possible.

This problem arises in interpretable learning because while the coordinate functions of a learned latent 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 data representation may 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 found in Chen and Meila [2], Kohli et al. [3], Jones et al. [4] for finding isometries. The insight that leads to isometry pursuit is that D function solutions multitask basis pursuit applied to an appropriately normalized $\mathcal{X}$ selects unitary submatrices. In particular, this normalization log-symmetrizes length in the column-space of $\mathcal{X}$ and favors vectors closer to unit length. The main advantage of this basis pursuit formulation is that it is convex and therefore computationally expedient.

2 Background

Variable selection in unsupervised learning is comparably less studied than in supervised learning. Thus, we explain the geometric criteria used to evaluate selection as well as the algorithmic considerations when designing such an algorithm.

2.1 Problem

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

Submitted to Interpretable AI: Past, Present and Future Workshop @ NeurIPS 2024

2.2 Interpretability and isometry

Our motivating example is the selection of data representations from within sets of putative coordinates. These putative coordinates are simply consist of columns of a matrix. The proposed method is thus even simpler than Sparse PCA [5, 6, 7], in which column-covariance is used to select low-dimensional projections from within the span of such a subset.

This method is specifically applicable with respect 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 a priori knowledge often takes the form of a functional dictionary. Regardless of implementation details such as whether this dictionary is given or learned, core concepts like evaluation of independentness of dictionary features arise in numerous scenarios [8, 9, 10]. After functional independence [9], also known as feature decomposability [**?**], which only requires that the differential of sets of dictionary features be full rank, metric properties of such sets are of natural interest.

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

$$D\phi(\xi) = \begin{bmatrix} \frac{\partial \phi^1}{\partial x^1}(\xi) & \dots & \frac{\partial \phi^1}{\partial x^D}(\xi) \\ \vdots & & \vdots \\ \frac{\partial \phi^D}{\partial x^1}(\xi) & \dots & \frac{\partial \phi^D}{\partial x^D}(\xi) \end{bmatrix}. \tag{1}$$

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

In other words, isometries of a D dimensional data distribution consist of D coordinate features whose gradients are orthogonal and evenly varying. Formulations of this quality within an objective function like deformation [3] and nuclear norm [**?**] use only a subset of the differential's information and are not uniquely minimized at unitarity, respectively. We therefore will introduce a new alternative objective based upon the entire spectrum which is uniquely minimized at isometry.

The applications of pointwise isometry are themselves manifold. Approaches that seek isometric embeddings non-parametrically include Local Tangent Space Alignment, multidimensional scaling and isomap. On the other hand, Kohli et al. [3] showed how pointwise isometries selected from a dictionary may be stitched together to form global embeddings. This approach constructs isometries through greedy search, with putative dictionary features added one at a time. Other work [9, 1] shows how group lasso may be used to select independent dictionary elements through a convex (i.e. non-greedy) algorithm. Will will show how this paradigm can be useful for selection of pointwise isometries as well.

Note that it is not necessary to explicitly estimate tangent spaces when applying the definition of isometry. The most commonly encountered manifolds are simply vector spaces, in which case the tangent spaces are trivial. This is the case for full-rank tabular data, as well as latent spaces of deep learning models. For example, the transformer residual stream at different tokens are analogous to tangent spaces of a non-linear manifold in the sense that the relative directions of dictionary vectors are not consistent between tokens.

2.3 Subset selection

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

$$\arg \min_{\mathcal{S} \subseteq [P] : |\mathcal{S}| = D} l(\mathcal{X}_{.\mathcal{S}}) \tag{3}$$

Brute force algorithms consider all possible solutions. These algorithms are conceptually simple, but often have 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 $\mathcal{S}$. This algorithms have time complexity $O(CPD)$ and so are computationally more efficient than brute force algorithms, but can get stuck in local minima.

Sometimes, it is possible to introduce an objective which convexifies problems of the above form. A classic example comes from basis pursuit solutions to the overcomplete regression problem $y = x\beta$ [**?**].

$$\arg\min f(\beta) : y = x\beta \tag{4}$$

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 particular form of basis pursuit that we apply is inspired by the group basis pursuit approach in Koelle et al. [9]. In group basis pursuit (which we call multitask basis pursuit when grouping is dependent only on the structure of response variable y) the objective function is $f(\beta) = \|\beta\|_{1,2} := \sum_{p=1}^{P} \|\beta_{p.}\|_2$ [11, 12, 13] This objective creates joint sparsity across entire rows of $\beta_{p.}$ and was used in [9] to select between sets of interpretable features.

3 Method

We first will 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

The property of orthonormal matrices that motivates our ground truth loss function comes from spectral analysis.

Proposition 1 (Singular values of a orthonormal matrix) *The singular values $\sigma_1 \ldots \sigma_D$ are equal to 1 if and only if $U \in \mathbb{R}^{D \times D}$ is orthonormal.*

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. The first desired property precludes, for example, the log determinant, while the last precludes the the deformation. Finally, to simplify our overall exposition and exemplify the key features of the method, we'd like this ground truth loss to be as similar as possible to the convex loss we introduce.

Thus, we define the loss

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

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

where $\sigma^d((X))$ is the d-th singular value of $\mathcal{X}$ and

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

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

Plainly, g is uniquely maximized by orthonormal matrices, and $g(\mathcal{X}^\dagger) = g(\mathcal{X})^{-1}$. The former condition is necessary for success of the method, while the latter, as well as the convexity of g, are somewhat aesthetic choices. A graph of g is given in Figure **??**. Most importantly, this loss enables comparison with produced after normalization as in Section 3.2.

$$\widehat{\mathcal{S}}_{GT} = \arg \min_{\mathcal{S} \subseteq [P] : |\mathcal{S}| = D} l_c(\mathcal{X}_{.\mathcal{S}}) \tag{9}$$

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 make the following definition.

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

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

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

$$q(v^1) = q(v^2) \; \forall \; v^1, v^2 : \|v^1\| = \|v^2\| \tag{12}$$

Normalization by functions satisfying this definition is sufficient for the application of multitask basis pursuit for finding isometries. The vectors are in particular length normalized so that the post-normalization vectors with the longest length correspond to pre-normalization vectors of length 1. We therefore propose the following normalization.

$$q : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+ \tag{13}$$

$$t, c \mapsto \frac{e^{t^c} + e^{t^{-c}}}{2e}, \tag{14}$$

and use this to define the vector normalization

$$n : \mathbb{R}^D \times \mathbb{R}^+ \to \mathbb{R}^D \tag{15}$$

$$n, c \mapsto \frac{n}{q(\|n\|_2, c)} \tag{16}$$

and matrix normalization

$$w : \mathbb{R}^{D \times P} \times \mathbb{R}^+ \to \mathbb{R}^D \tag{17}$$

$$\mathcal{X}_{.p}, c \mapsto n(\mathcal{X}_{.p}, c) \; \forall \; p \in [P]. \tag{18}$$

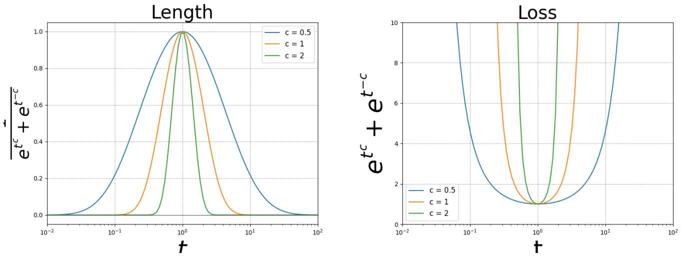
Besides satisfying the conditions in Definition 3, this normalization has some additional nice properties. First, q is convex. Second, it grows asymptotically log-linearly. Third, while $\exp(-|\log t|) = \exp(-\max(t, 1/t))$ is a seemingly natural choice for normalization, it is non smooth, and the LogSumExp replacement of $\max(t, 1/t)$ with $\log(\exp(t) + \exp(1/t))$ simplifies to 13 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 appropriate normalized design matrix $w(\mathcal{X})$ to identify submatrices of $\mathcal{X}$ that are as orthonormal as possible. Define the multitask basis pursuit penalty

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

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



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

(b) Length as a function of t

Figure 1: Plots of Length and Loss for different values of c. Since t is one dimensional and therefore diagonalizable, basis pursuit and ground truth give identical loss values.

The isometry pursuit program is then

$$\widehat{\beta}_c^D(\mathcal{X}) := \arg \min_{\beta \in \mathbb{R}^{P \times D}} \|\beta\|_{1,2} \ : \ I_D = w(\mathcal{X}, c)\beta. \tag{21}$$

The recovered functions are the indices of the dictionary elements with non-zero coefficients. That is, they are given by $S(\beta)$ where

$$S : \mathbb{R}^{p \times d} \to \binom{[P]}{d} \tag{22}$$

$$\beta \mapsto \{p \in [P] : \|\beta_{p.}\| > 0\} \tag{23}$$

and $\binom{[P]}{d} = \{A \subseteq [P] : |A| = d\}$.

```
IsometryPursuit(Matrix $\mathcal{X} \in \mathbb{R}^{D \times P}$, scaling constant $c$)
1: Output $\widehat{S} = S(\widehat{\beta}_P(w_c(\mathcal{X})))$
```

A key initial theoretical assertion for the feasibility of IsometryPursuit is that it is invariant to choice of basis for $\mathcal{X}$.

Proposition 2 (Basis pursuit selection invariance) *Let $U \in \mathbb{R}^{D \times D}$ be orthonormal. Then $S(\widehat{\beta}(U\mathcal{X})) = S(\widehat{\beta}(\mathcal{X}))$.*

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

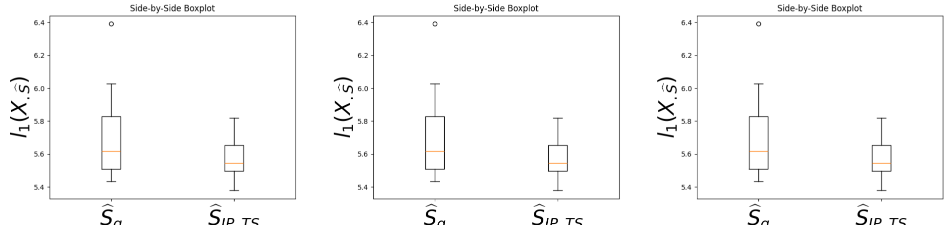
The intuition behind our application of multitask basis pursuit in our setting is that submatrices consisting of vectors which are closer to 1 in length and more orthogonal will have smaller loss. The property of orthonormal matrices that corresponds to this is slightly different from Proposition 1.

Proposition 3 (Basis vectors of a orthonormal matrix) *The component vectors $u^1 \dots u^D \in \mathbb{R}^B$ form a orthonormal matrix if and only if, for all $d_1, d_2 \in [D], u_{d_1} u^{d_2} = \begin{cases} 1 & d_1 = d_2 \\ 0 & d_1 \neq d_2 \end{cases}$.*

We show theoretically that the conditions of the consequent of Proposition 3 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 $|\mathcal{S}| = D$.

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

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 Isometry Pursuit on real data.



(a) Boxplot of $\widehat{S}_g$.

(b) Boxplot of $\widehat{S}_{TS}$.

(c) Boxplot of $\widehat{S}_{third}$.

Figure 2: Comparison with isometry loss.

3.4 Two-stage isometry pursuit

Since cannot in general ensure either that $|\widehat{S}| = D$ or that a orthonormal submatrix $X_{.S}$ exists, we apply a standard approach in the lasso literature: to first use the convex problem to prune, prior to a final feature selection step [**?**]. For simplicity, we propose using brute selection applied to the estimated feature set $\widehat{S}$. In practice, this

```
TwoStageIsometryPursuit(Matrix $\mathcal{X} \in \mathbb{R}^{D \times P}$, scaling constant $c$, objective $f$)
$\widehat{S_1} =$ IsometryPursuit($\mathcal{X}, c$)
$\widehat{S} =$ Brute($\mathcal{X}_{.\widehat{S_1}}, f$)
**Output** $\widehat{S}$
```

4 Experiments

We apply Two Stage Isometry Pursuit to the standard Iris and Wine datasets, as well as the Ethanol dataset from **?**]. The latter is an interpretability dataset where a dictionary of precomputed interpretable features are evaluated for their ability to parameterize the data manifold. Preprocessing details such as tangent space estimation are shared with **?**].

We use the CVXPY Python package. We use the SCS interior point solver from CVXPY, which is able to push sparse values arbitrarily close to 0 [**?**].

5 Discussion

It could be used in the stiching step of an algorithm like the kohli one We leave aside the question of patch alignment [**? ?**]. The full gradient approach. In this case normalization prior to projection is subsumbmed by the larger coefficients needed to get the tangent space. Good news is tangent space estimation need not be performed. Let's compare the coefficients involved in projecting versus not projecting. We can perform regression in the high dimensional space instead of projecting on span of target variable.

With respect to pseudoinverse estimation, sparse methods have been applied in [**?**] Even though by Lagrangian duality, the basis pursuit solution corresponds to λ approaching 0, the solution is sparse [**?**].

While the sparse PCA problem is non-convex, our approach can be taken as a simpler version in the sense that the loadings are constrained to be the identify matrix. [6] gives a method for solving the sparse-PCA method more efficiently than the original greedy approach. Compared with the FISTA method used in [9, 1], coordinate descent [**? ? ?**] is faster [**? ?**]. Compared with [**? ?**], the sklearn multitask lasso is $2, 1$ rather than $\infty, 1$ regularized. This misses the utility of our normalization for finding unitary matrices since isometry embeddings preserve important properties like distances between points.

Our notion of isometric recovery is distinct from the restricted isometry property [**? ?**], which is used to show guaranteed recovery at fast convergence rates in supervised learning. In particular, our approach does not consider statistical error or the presence of a true underlying model. However, we note that disintegration of performance at high λ values in the lasso formulation may have some relation to these properties, as discussed in [9, 1].

A major area of comparison is in diversification in recommendation systems where greedy algorithms are used [**? ?**], and also in document diversification for Retrieval Augmented Generation.

The most pressing piece of theoretical work which remains on this topic is the removal of the resttriction $|S| = D$ on the conditions of Proposition. The resulting proposition, which seems almost obvious, is in fact more difficult to argue, and is seemingly violated by empirical results. Nevertheless, these violations are subtly non-dispositive since absence of sparsity and improvements of primal loss below D are accompanied by violations of the constraint of a similar magnitude, suggesting that we a more refined approach to optimization, substantial improvements in estimation accuracy may be possible. From a geometric perspective, we note that isometries may not always exist in the presence of curvature, and comparison of our loss with curvature could prove fertile. Finally, the speed increases garnered by the particularly simple form of our algorithm warrants comparison with other pseudoinverse estimators warrant further comparison. An extension of our estimator we omit for brevity is to use the multitask lasso formulation to trim the size of $\hat{P}$. Besides our new normalization, this results in a simpler procedure than in **?**] and Koelle et al. [1] that is amenable to the more performant multitask lasso solver in sklearn.

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] 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`.

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

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

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

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

6 Supplement

We give proofs in support of the propositions in the main text and supplemental experimental information to better contextualize the results.

6.1 Algorithms

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

```
Brute(Matrix $\mathcal{X} \in \mathbb{R}^{D \times P}$, objective $f$)
for each combination $S \subseteq \{1, 2, \ldots, P\}$ with $|S| = D$ do
    Evaluate $f(\mathcal{X}_{.S})$
end for
Output the combination $S^*$ that minimizes $f(\mathcal{X}_{.S})$
```

```
Greedy(Matrix $\mathcal{X} \in \mathbb{R}^{D \times P}$, 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 $f(\mathcal{X}_{.(S \cup \{p\})})$
        if $f(\mathcal{X}_{.(S \cup \{p\})}) < f_{\text{best}}$ then
            Update $S_{\text{best}} = S \cup \{p\}$
            Update $f_{\text{best}} = f(\mathcal{X}_{.(S \cup \{p\})})$
        end if
    end for
    Return Greedy($\mathcal{X}$, $f$, $S_{\text{best}}$, $d + 1$)
end if
```

6.2 Proofs

6.2.1 Proof of Proposition 2

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

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

Proof:

$$\|\beta U\|_{1,2} = \sum_{p=1}^{P} \|\beta_{p.} U\| \tag{24}$$

$$= \sum_{p=1}^{P} \|\beta_{p.}\| \tag{25}$$

$$= \|\beta\|_{1,2} \tag{26}$$

□

We then show that this implies that the resultant loss is unchanged by unitary transformation of $\mathcal{X}$.

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

Proof:

$$\begin{aligned}
\widehat{\beta}(U\mathcal{X}) &= \arg \min_{\beta\in\mathbb{R}^{P\times D}} \|\beta\|_{1,2} \ : \ I_D = UX\beta && (27)\\
&= \arg \min_{\beta\in\mathbb{R}^{P\times D}} \|\beta\|_{1,2} \ : \ U^{-1}U = U^{-1}UX\beta U && (28)\\
&= \arg \min_{\beta\in\mathbb{R}^{P\times D}} \|\beta\|_{1,2} \ : \ I_D = X\beta U && (29)\\
&= \arg \min_{\beta\in\mathbb{R}^{P\times D}} \|\beta U\|_{1,2} \ : \ I_D = X\beta U && (30)\\
&= \arg \min_{\beta\in\mathbb{R}^{P\times D}} \|\beta\|_{1,2} \ : \ I_D = X\beta. && (31)
\end{aligned}$$

□

6.2.2 Proof of Proposition 4

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

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

$$\beta_{dd'} = \begin{cases} 1 & d = d' \in \{1 \dots D\} \\ 0 & \text{otherwise} \end{cases}. \qquad (32)$$

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

From the conditions in Definition **??**, 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 [**?**]. Denoting Q to be composed of basis vectors $[e^1 \dots e^d]$, the matrix R has form

$$R = \begin{bmatrix} \langle e^1, X_{.S_1}\rangle & \langle e^1, X_{.S_2}\rangle & \dots & \langle e^1, X_{.S_D}\rangle \\ 0 & \langle e^2, X_{.S_2}\rangle & \dots & \langle e^2, X_{.S_D}\rangle \\ 0 & 0 & \dots & \dots \\ 0 & 0 & \dots & \langle e^d, X_{.S_D}\rangle \end{bmatrix}. \qquad (33)$$

Thus, $|R_{dd}| \leq \|X_{.S_d}\|_2$, with equality obtained across d only by orthonormal matrices. On the other hand, by Proposition **??**, $l(X) = l(R)$ 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.

□