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] Work conducted outside of Amazon.

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

1 Supplement

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

1.1 Algorithms

We give definitions of the brute and greedy algorithms for the combinatorial problem studied in this paper. The brute force algorithm is computationally intractable for all but the smallest problems, but always finds the global minima.

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

Greedy algorithms are computationally expedient but can get stuck in local optima [**? ?**], even with randomized restarts [**?**].

```
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 $p \in \{1, 2, \dots, P\} \setminus S$ 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($X$, $f$, $S_{\text{best}}$, $d + 1$)
end if
```

1.2 Proofs

1.2.1 Proof of Proposition ??

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

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

Proof:

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

$$= \sum_{p=1}^{P} \|\beta_{p.}\| \tag{2}$$

$$= \|\beta\|_{1,2} \tag{3}$$

□

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

Proposition 2 *Let $U \in \mathbb{R}^{D \times D}$ be unitary. Then $\widehat{\beta}(UX) = \widehat{\beta}(X)U$.*

Proof:

$$\widehat{\beta}(UX) = \arg \min_{\beta \in \mathbb{R}^{P \times D}} \|\beta\|_{1,2} \; : \; I_D = UX\beta \tag{4}$$

$$= \arg \min_{\beta \in \mathbb{R}^{P \times D}} \|\beta\|_{1,2} \; : \; U^{-1}U = U^{-1}UX\beta U \tag{5}$$

$$= \arg \min_{\beta \in \mathbb{R}^{P \times D}} \|\beta\|_{1,2} \; : \; I_D = X\beta U \tag{6}$$

$$= \arg \min_{\beta \in \mathbb{R}^{P \times D}} \|\beta U\|_{1,2} \; : \; I_D = X\beta U \tag{7}$$

$$= \arg \min_{\beta \in \mathbb{R}^{P \times D}} \|\beta\|_{1,2} \; : \; I_D = X\beta. \tag{8}$$

□

1.2.2 Proof of Proposition ??

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

Proof: Without loss of generality, we show that $[I_D 0]$ satisfies the KKT conditions for $X = [I_D X_{-S}]$ The KKT conditions are

- Primal feasibility: $w(X, c)\beta = I_D$
- Stationarity: there exists a dual variable $\nu \in \mathbb{R}^{D \times D}$ such that $0 \in \partial \|\beta\|_{1,2} - w(X, c)^T \nu$ where ∂ is the subdifferential operator.
- Dual feasibility
- Complementary slackness

Stationarity:

$$\|\beta\|_{1,2} = \sum_{p=1}^{P} \|\beta_{p.}\|_2. \tag{9}$$

Then

$$\partial \|\beta\|_{1,2} = \begin{bmatrix} I_D v_{d+1} \\ \dots \\ v_P \end{bmatrix} \tag{10}$$

where $v_p \in \mathbb{R}^D$ satisfies $\|v_p\|_2 \leq 1$.

We therefore must satisfy

$$\begin{bmatrix} I_D \\ w(X_{-S}, c)^T \end{bmatrix} \nu = \begin{bmatrix} I_D \\ V_{-S} \end{bmatrix} \tag{11}$$

where $\nu \in \mathbb{R}^{D \times D}$.

At this point we can see that in fact I_D is an appropriate choice of ν by the boundary on the range of the normalized vector.

□

Proposition 4 *The SCS algorithm identities the minimum 2-norm solution*

Proof: The splitting conic solver algorithm, also known as alternative direction method of multipliers (ADMM), makes use of the follow "split augmented Lagrangian" objective function

$$\mathcal{L}_\rho(X, Z\beta, \lambda_1) = \|\beta\|_{1,2} + \lambda_1^T(I_D - X\beta) + \lambda_2^T(\beta - Z) + \rho\|I_D - X\beta\|_2^2 + \rho\|\beta - Z\|_2^2 \tag{12}$$

This function provably has the same minimizer over β as **??**, but is more amenable to optimization. By differentiating this expression w.r.t. β to write out its update rule, we show that the update requires computing a pseudoinverse. This pseudoinverse is in fact chosen to be a ridge regression pseudoinverse (i.e. $(X^T X + \Lambda)^{-1}$) which, as the closed form solution for ridge regression, for $\|\beta\|_2^2$ minimization.

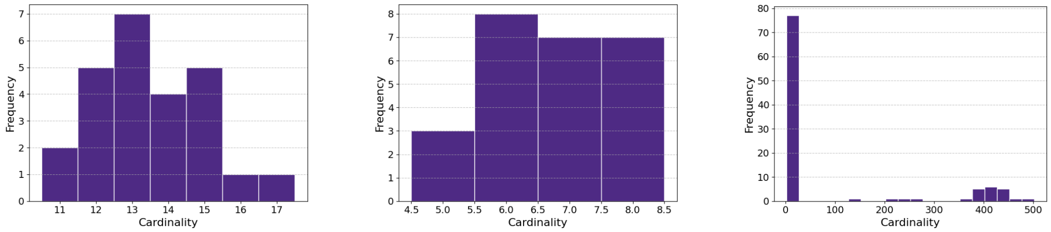
Proposition 5 *Equation* **??** *is strongly convex.*

Proof: □

Since the squared 2-norm is strongly convex, this minimizer is unique. □

Proposition 6 *The minimizing solution is contained within the estimated set*

The uniqueness of the solution bears attention. A well-known result in Lasso literature states that if the columns of X are in so-called general position (meaning that no more than $k + 1$ columns are contained within any k dimensional subspace, with $k < D$), then the standard lasso solution is unique. This result has been generalized to show that the lasso solution is unique if and only if the codimension of the intersection of the row span of X with the P dimensional unit cube does not exceed the dimension of the row span of X. That is, the rowspan of X must not intersect too many corners of the P dimensional cube. In our case, at least the corner among indices S is obtainable, so this codimension is at least one. Since we can rewrite our multitask problem as a block sparse problem with grouping across blocks, we can see that after a point, additional dimensions mandate that the solution be non-unique.



(a) Wine Dataset

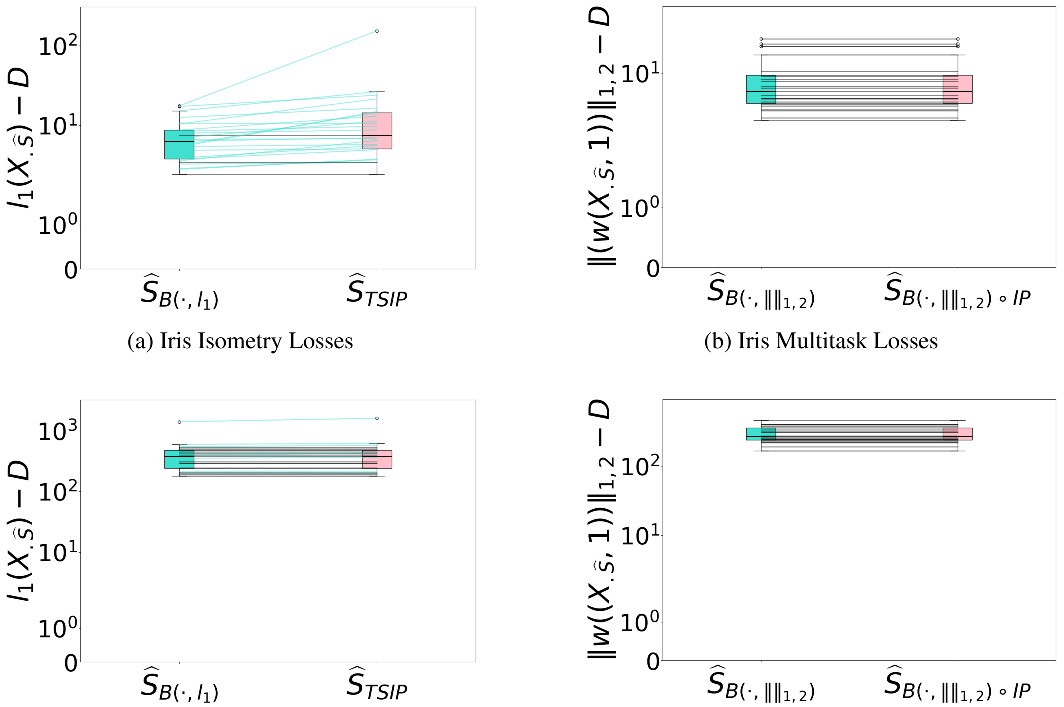
(b) Iris Dataset

(c) Ethanol Dataset

Figure 1: Support Cardinalities for Wine, Iris, and Ethanol datasets

1.3 Support cardinalities

Figure **??** plots the distribution of $|\widehat{S}_{IP}|$ from Table **??** 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 2: 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 **??** 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.

1.4 Proposition ?? deep dive

As mentioned in Section **??**, the conditions under which the restriction $P = D$ in Proposition **??** may be relaxed are of theoretical and practical interest. The results in Section **??** show that there are circumstances in which the GREEDYSEARCH performs better than TWOSTAGEISOMETRYPURSUIT, so clearly TWOSTAGEISOMETRYPURSUIT does not always achieve a global optimum. Figure **??** gives results on the line of inquiry about why this is the case based on the reasoning presented in Section **??**. 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 TWOSTAGEISOMETRYPURSUIT is in fact only 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.

1.5 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. The particularly high variance for brute force search in the second step of TwoStageIsometryPursuit is likely due to the large cardinalities reported in Figure **??**.

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 1: Algorithm runtimes in seconds across replicates.