Sparse isometry pursuit

Samson Koelle, Marina Meila

July 10, 2024

Abstract

Sparse isometry pursuit is an algorithm for identifying unitary column-submatrices in polynomial time. It achieves sparsity via use of the group lasso norm. It has constrained basis pursuit and penalized group lasso formulations. Applied to Jacobians of putative coordinate functions, it is a useful subroutine for identifying isometric embeddings. This approach has relevance to interpretability of learned representations.

1 Introduction

The problem of computing a sparse inverse matrix.

Within the context of non-linear dimension reduction, selection of coordinate functions of an embedding space from within a dictionary is a core problem in geometric data analysis. In order of specificity, these methods may seek to optimize independent coordinates **NEURIPS2019˙6a10bbd4**; **He2023-ch**, low distortion embeddings, or isometric embeddings. Dictionaries can be either given **Koelle2022-ju**; **Koelle2024-no** or learned **Kohli2021-lr**. Optimization can be global or local.

These coordinate selection algorithms can be greedy **NEURIPS2019˙6a10bbd4**; **Kohli2021-lr** or convex **Koelle2022-ju**; **Koelle2024-no**. In this paper we show that an adapted version of the group lasso type algorithm in **Koelle2024-no** leads to a convex procedure competitive with previous greedy approaches with respect to isometry. This approach relies on a to-our-knowledge novel matrix inversion algorithm that is sparse in the column space of the matrix. This method displays the favorable characteristics of group lasso type problems, including duality of a regularized form with a basis pursuit problem. These problems are solvable with off-the-shelf multitask lasso and interior point solvers, respectively.

since isometry embeddings preserve important properties like distances between points. We describe a convex optimization approach for selection such functions based on the Tangent Space Lasso. This approach combines a strict theoretical criterion and computationally expediency.

2 Background

We are given a rank D matrix $\mathcal{X} \in \mathbb{R}^{D \times P}$ with $P > D$. X could be, for example, the Jacobian matrix $d\mathcal{G}$ of a set of candidate coordinate functions. We assume that we are given a target dimension $T \leq D$

3 Problem

Our goal is to select a subset $\mathcal{S} \subset [P]$ with $|\mathcal{S}| = D$ such that $X_{.\mathcal{S}}$ is unitary.

4 Method

Define the group basis pursuit penalty norm

$$\textbf{??}\|\|_{1,2}\mathbb{R}^{P \times D} \to \mathbb{R}^{+}\beta \mapsto \sum_{p=1}^{P} \|\beta_{p.}\|_2. \tag{1}$$

Direct minimization of Equation **??** will not select for isometry due to the preference for columns with larger norm. Define the transform

$$\exp_1 : \mathbb{R}^{D \times P} \to \mathbb{R}^{D \times P} \tag{2}$$

$$\mathcal{X} \mapsto \exp(-|\log \|\mathcal{X}_{j.}\|_2)|)\frac{\mathcal{X}_{j.}}{\|\mathcal{X}_{j.}\|}. \tag{3}$$

We can also define the basis pursuit loss

$$m(X, \beta) := \|\beta\|_{2,1} : I_d = X\beta \tag{4}$$

Our main interest is in analyzing the properties of $l(\exp_1 X, \beta)$ and $m(\exp_1 X, \beta)$

This is the main loss function whose properties we analyze.

4.1 Tangent Space Lasso

The intuition is that vectors which are closer to 1 in length and more orthogonal will be smaller in loss.

Proposition 1 *Unitary subset selection Given a X that contains a unique subset $X^* \in \mathbb{R}^{d \times d}$ such that X^* is unitary and full rank, then $X^* = \arg\min_\beta l(\exp_1(X), \beta)$.*

Before proceeding, we require the following piece of Lemma **??**.

Proposition 2 *Consider two sets of vector fields X and X^i where $X^i_{i..} = UX_{i..}$, where U is unitary and $X_{i'..} = X^i_{i'..}$ for other values $i' \neq i$. Then $l^*(X) = l^*(X^i)$*

Proof: Without loss of generality, let $i = 1$. We can write

$$l^*(X^i) = l(\beta^i) = \sum_{j=1}^{p} (\sum_{i'=2}^{n} \|\beta_{i'j.}\|_2^2 + \|\beta_{1j.}^i\|_2^2)^{1/2} = \sum_{j=1}^{p} (\sum_{i'=1}^{n} \|\beta_{i'j.} U\|_2^2)^{1/2} = l^*(X) \quad (5)$$

where the second to last equality is because the norm $\|v\|_2^2$ is unitary invariant. □

We first show that vectors which are more orthogonal will be smaller in loss.

Proposition 3 *Let $X_{.S} \in \mathbb{R}^{d \times p}$ be defined as above and let $X'_{..S}$ be an array such that $\|X'_{.S_j}\|_2 = \|X_{.S_j}\|_2$ for all $j \in [d]$ and $X'_{.S}$ is column-orthogonal. Then $\tilde{l}^*(X_{..S}) > \tilde{l}^*(X'_{..S})$.*

Proof: By Lemma 2, without loss of generality

$$\beta_{ijk}^i = \begin{cases} \|\tilde{X}'_{.S_j}\|_2^{-1} & j = k \in \{1 \dots d\} \\ 0 & \text{otherwise} \end{cases}. \quad (6)$$

Therefore,

$$\tilde{l}^*(X') = \sum_{j=1}^{d} \sqrt{\sum_{i=1}^{n} \|\tilde{X}'_{i.S_j}\|_2^{-2}}. \quad (7)$$

On the other hand, the invertible matrices $\tilde{X}_{.S}$ admit QR decompositions $\tilde{X}_{.S} = QR$ where Q and R are square unitary and upper-triangular matrices, respectively **Anderson1992-fb**. Since l^* is invariant to unitary transformations, we can without loss of generality, consider $Q = I_d$. Denoting I_d to be composed of basis vectors $[e^1 \dots e^d]$, the matrix R has form

$$R = \begin{bmatrix} \langle e^1, \tilde{X}_{i.S_1} \rangle & \langle e^1, \tilde{X}_{i.S_2} \rangle & \dots & \langle e^1, \tilde{X}_{i.S_d} \rangle \\ 0 & \langle e^2, \tilde{X}_{i.S_2} \rangle & \dots & \langle e^2, \tilde{X}_{i.S_d} \rangle \\ 0 & 0 & \dots & \dots \\ 0 & 0 & 0 & \langle e^d, \tilde{X}_{i.S_d} \rangle \end{bmatrix}. \quad (8)$$

The diagonal entries $R_{jj} = \langle q^j, \tilde{X}_{.S_j} \rangle$ of this matrix have form $\|\tilde{X}_{.S_j} - \sum_{j' \in \{1 \dots j-1\}} \langle \tilde{X}_{.S_j}, e^{j'} \rangle e^{j'} \|$. Thus, $R_j \in (0, \|\tilde{X}_{i.S_j}\|]$. On the other hand $\beta_{iS.} = R^{-1}$, which has diagonal elements $\beta_j = R_j^{-1}$, since R is upper triangular. Thus, $\beta_{jj} \geq \|\tilde{X}_{.S_j}\|^{-1}$, and therefore $\|\beta_{iS_j.}\| \geq \|\beta'_{S_j.}\|$. Since $\|\beta_{S_j.}\| \geq \|\beta'_{S_j.}\|$ for all i, then $\|\beta_{.S_j.}\| \geq \|\beta'_{.S_j.}\|$. □

The above proposition formalizes our intuition that orthogonality of X lowers $l^*(X)$ over non-orthogonality. We now show a similar result for the somewhat less intuitive heuristic that dictionary functions whose gradient fields are length 1 will be favored over those which are non-constant. Since the result on orthogonality holds regardless of length, we need only consider the case where the component vectors in our sets of vector fields are mutually orthogonal at each data point, but not necessarily of norm 1. Note that were they not orthogonal, making them so would also reduce l^*. We then show that vectors which are closer to length 1 are lower in loss. Since vectors which are closer to length 1 are shrunk in length less by $\exp_1$, their corresponding loadings are smaller. This is formalized in the following proposition

Proposition 4 *Let $X''_{.S}$ be a set of vector fields $X''_{.S_j}$ mutually orthogonal at every data point i, and $\|X''_{.S_j}\| = 1$. Then $\tilde{l}^*(X'_{.S}) \geq \tilde{l}^*(X''_{.S})$.*

Proof: Let $\|X''_{i.S_j}\| = c_j$. By Proposition 2, we can assume without loss of generality (i.e without changing the loss) that

$$\tilde{X}_{.S_j} = \begin{bmatrix} c_1 & 0 & 0 \\ 0 & \ldots & 0 \\ 0 & 0 & c_d \rangle \end{bmatrix}. \tag{9}$$

Thus

$$\tilde{\exp}_1 X_{.S_j} = \begin{bmatrix} \exp(-|\log \|c_1\|_2)|) & 0 & 0 \\ 0 & \ldots & 0 \\ 0 & 0 & \exp(-|\log \|c_d\|_2)|)\rangle \end{bmatrix}. \tag{10}$$

and therefore

$$\tilde{\beta}_{.S_j} = \begin{bmatrix} \exp(-|\log \|c_1\|_2)|)^{-1} & 0 & 0 \\ 0 & \ldots & 0 \\ 0 & 0 & \exp(-|\log \|c_d\|_2)|)^{-1}\rangle \end{bmatrix}. \tag{11}$$

The question is therefore what values of c_j minimize $\exp(-|\log\ |c_1\|_2)|)^{-1}$. $|\log\ |c_1\|_2)|$ is minimized (evaluates to 0) when $c_j = 1$, so $-|\log\ |c_1\|_2)|$ is maximized (evaluates to 0, so $\exp(-|\log\ |c_1\|_2)|)$ is maximized (evaluates to 1), so $\exp(-|\log\ |c_1\|_2)|)^{-1}$ is minimized (evaluates to 1). $\square$

Proposition 5 *Local Isometry Given a set of functions G that contains a subset that defines a locally isometric embedding at a point ξ, then these will be selected as* $\arg\min_\beta$.

Algorithm (Local tangent Space basis pursuit)

Algorithm (Local two stage tangent space basis pursuit)

This provides an approach for the problem put forward in (cite) LDLE paper.

Experiments (Loss)

Compare with isometry loss (2 norm of singular values).

5 Experiments

Comparison with isometry loss.

6 Discussion

It could be used in the stiching step of an algorithm like the kohli one We leave aside the question of patch alignment **https://arxiv.org/pdf/2303.11620.pdf**; **LDLE paper**. 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.

7 Supplement

Proof of local isometry (simpler proof since no oscillation game)