This note specifically targets the problem of local isometry estimation using a convex Tangent Space Lasso-type method. It does not directly reproduce the argument that tangent space basis pursuit selects for a "most isometric embedding" from my thesis, nor would I expect this to go in the same paper as tangent space lasso. Other interesting open questions about basis pursuit that are not covered here include can we equate the intrinsic curvature of the data manifold to the basis pursuit loss, or what happens when dictionary = feature space, or what happens if we estimate in the coordinates of the ambient space, or what are the convergence rates. I chose to focus on local isometry estimation because it seemed distinct enough from what everyone else was working on that it could be published as a workshop paper on its own, and because the mathematical components are a little novel, but not disorientingly so, and because it solves an open problem noticed by another research group.

Abstract

The selection of a set of coordinate functions producing an isometry embedding from within a dictionary is an important topic in geometric data analysis **Gal Mishne paper**, 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. Compared with previous greedy approaches **https://openreview.net/pdf?id=GugzbdAoHG**; **chen and meila**; **ldle**, this approach is more computationally expedient.

1 Background

1.1 Notation

Given $X \in \mathbb{R}^{p\times d}$ with $p > d$, define the transform

$$\exp_1 X := \exp(-|\log \|X_{j.}\|_2)|)\frac{X_{j.}}{\|X_{j.}\|}. \tag{1}$$

Define the group basis pursuit penalty function

$$\|X\|_{1,2} := \sum_{j=1}^{p} \|X_{j.}\| \tag{2}$$

Define the loss function

$$l(X,\beta) := \|I_d - X\beta\|^2 + \|\beta\|_{2,1} \tag{3}$$

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.

1.2 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

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

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 **??**, without loss of generality

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

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

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

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 **??**, we can assume without loss of generality (i.e without changing the loss) that

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

Thus

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

and therefore

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

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). □

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).

2 Experiments

Comparison with isometry loss.

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. The Hoeffding bound Dimension estimation, the failure of duality The presence of curvature

Supplement

Proof of isometry (Copy from thesis)

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

3 Discussion

We leave aside the question of patch alignment **https://arxiv.org/pdf/2303.11620.pdf**; **LDLE paper**.

4 Comparison with other approach

We can perform regression in the high dimensional space instead of projecting on span of target variable.

Temporary page!

LaTeX was unable to guess the total number of pages correctly. As there was some unprocessed data that should have been added to the final page this extra page has been added to receive it.

If you rerun the document (without altering it) this surplus page will go away, because LaTeX now knows how many pages to expect for this document.