Isometry pursuit

Samson Koelle, Marina Meila

July 12, 2024

Abstract

Isometry pursuit is an algorithm for identifying unitary column-submatrices of wide matrices in polynomial time. It achieves sparsity via use of the group lasso norm, and therefore has constrained and penalized formulations. Applied to tabular data, it selects a subset of columns that maximize diversity. Applied to Jacobians of putative coordinate functions, it identifies isometric embeddings from within dictionaries. It therefore has relevance to interpretability of learned representations.

1 Introduction

Many real-life 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 unsupervised 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 unitary as possible.

This problem is motivated by applications in diversification and non-linear dimension reduction. In particular, the name of the method comes from the fact that isometric embeddings have unitary differentials. While variable selection in unsupervised learning is comparably less studied than in supervised learning, substantial literature exists. One method that exemplifies this area is Sparse PCA **Dey2017-mx**, in which a subset of variables are used to generate low-dimensional projections. Within non-linear dimension reduction dictionaries can be either given **Koelle2022-ju**; **Koelle2024-no** or learned **Kohli2021-lr**. In order of specificity, these methods may seek to optimize independent coordinates **Chen2019-km**; **He2023-ch**, low distortion embeddings, or isometric embeddings. Optimization can be global or local. These coordinate selection algorithms can be greedy **NEURIPS2019˙6a10bbd4**; **Kohli2021-lr**; **Jones2007-uc** or convex **Koelle2022-ju**; **Koelle2024-no**.

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. This normalization is log-symmetric length in the column-space of $\mathcal{X}$ and favors vectors closer to unit length. This property is the focus of Section **??**

The basis pursuit formulation is desirable for several reasons. First, it is convex and therefore computationally expedient. Second, while not D-sparse, it is relatively sparse, and so can be used for pruning. Third, it admits a lasso dual problem which is particularly useful in high dimensions.

2 Background

2.1 Problem

As mentioned in Section **??**, our goal is to select a subset $\mathcal{S} \subset [P]$ with $|\mathcal{S}| = D$ such that $X_{.\mathcal{S}}$ is as unitary as possible. To that end, we must define what this means.

2.2 Isometries

Definition 1 *Isometry Given two D dimensional manifolds $\mathcal{M}$ and $\mathcal{N}$ and a map*

$$\phi : \tag{1}$$

Let X could be, for example, the Jacobian matrix dg of a set of candidate coordinate functions $g = [g^1, \dots g^P]$.

3 Method

3.1 Normalization

Since basis pursuit methods tend to select longer vectors, selection of unitary submatrices requires normalization such that long and short candidate basis vectors are penalized equivalently. Thus, let

$$q : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+ \tag{2}$$

$$t, c \mapsto \frac{\exp(t^c) + \exp(t^{-c})}{2e}, \tag{3}$$

and use this to define the normalization

$$n : \mathbb{R}^D \times \mathbb{R}^+ \to \mathbb{R}^D \tag{4}$$

$$n^d, c \mapsto \frac{n^d}{q(\|n\|_2, c)} \forall d \in [D]. \tag{5}$$

This normalization scales lengths down that are far away from 1 is a logarithmically symmetric way. Any rescaling which is maximized at 1 and logarithmically symmetric satisfies Proposition **??**, but n is particularly suitable. 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 **??** upon the 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 ∞.

Using this, define the matrix-wide normalization vector

$$\mathcal{D} : \mathbb{R}^{D \times P} \times \mathbb{R}^{+} \to \mathbb{R}^{P} \tag{6}$$

$$\mathcal{X}_{.p}, c \mapsto n(\mathcal{X}_{.p}, c) \tag{7}$$

and the normalized matrix $\tilde{\mathcal{X}}_c = \mathcal{X}\mathcal{D}(\mathcal{X}, c)$. This completes the data preprocessing.

3.2 Ground truth

The main goal of sparse isometry pursuit is to expediate the selection of unitary submatrices. Typically, unitaryness is measured using the singular values of the matrix. However, measures like the operator norm and deformation compare only the largest an smallest singular values. Compared with the nuclear norm, it is symmetric around 1 **Fazel2001ARM**. That is, they do not account for the whole spectrum.

Define the loss

$$l_{iso} : \mathbb{R}^{D \times P} \to \mathbb{R}^{+} \tag{8}$$

$$(X) \mapsto \sum_{d=1}^{D} g(\sigma^{d}(\mathcal{X})) \tag{9}$$

where $\sigma^d((X))$ is the d-th singular value of $\mathcal{X}$. However, this would not result in a sparse solution.

This loss is an appropriate choice for comparison because it is equal to the basis pursuit loss for orthogonal matrices

Proposition 1

Proof:

$$\tag{10}$$

Then, singular values and regressands are analytically determined. cont. □

3.3 Isometry pursuit

Define the multitask group basis pursuit penalty

$$\| \cdot \|_{1,2} : \mathbb{R}^{P \times D} \to \mathbb{R}^+ \quad (11)$$

$$\beta \mapsto \sum_{p=1}^{P} \|\beta_{p.}\|_2. \quad (12)$$

The isometry pursuit program is then

$$\arg \min_{\beta \in \mathbb{R}^{P \times D}} \|\beta\|_{1,2} \ s.t. \ I_D = \tilde{\mathcal{X}}_c \beta. \quad (13)$$

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

3.4 Isometric lasso

By Lagrangian duality, define an extension of **??** called Isometric Lasso. Isometric Lasso is

$$\arg \min_{\beta \in \mathbb{R}^{P \times D}} \|I_D - \tilde{\mathcal{X}}_c \beta\|_2^2 + \lambda \|\beta\|_{1,2} \quad (14)$$

This extension is a local version of Tangent Space Lasso.

4 Theory

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 $l^*(X) = l^*(X^i)$*

Proposition 3 *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 **??**.

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

$$l^*(X^i) = l(\beta^i) = \sum_{j=1}^{p} (\sum_{i'=2}^{n} \|\beta_{i'j.}\|_2^2 + \|\beta^i_{1j.}\|_2^2)^{1/2} = \sum_{j=1}^{p} (\sum_{i'=1}^{n} \|\beta_{i'j.} U\|_2^2)^{1/2} = l^*(X) \quad (15)$$

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

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

Proposition 4 *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

$$\beta^i_{ijk} = \begin{cases} \|\tilde{X}'_{.S_j}\|_2^{-1} & j = k \in \{1 \dots d\} \\ 0 & \text{otherwise} \end{cases}. \tag{16}$$

Therefore,

$$\tilde{l}^*(X') = \sum_{j=1}^{d} \sqrt{\sum_{i=1}^{n} \|\tilde{X}'_{i.S_j}\|_2^{-2}}. \tag{17}$$

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

$$R = \begin{bmatrix} \langle e^1, \tilde{X}_{i.S_1} \rangle & \langle e^1, \tilde{X}_{i.S_2} \rangle & \dots & \langle e^1, \tilde{X}_{i.S_d} \rangle \\ 0 & \langle e^2, \tilde{X}_{i.S_2} \rangle & \dots & \langle e^2, \tilde{X}_{i.S_d} \rangle \\ 0 & 0 & \dots & \dots \\ 0 & 0 & 0 & \langle e^d, \tilde{X}_{i.S_d} \rangle \end{bmatrix}. \tag{18}$$

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 5 *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

$$\tilde{X}_{.S_j} = \begin{bmatrix} c_1 & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & c_d \rangle \end{bmatrix}. \tag{19}$$

Thus

$$\tilde{\exp}_1 X_{.S_j} = \begin{bmatrix} \exp(-|\log \|c_1\|_2)|) & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & \exp(-|\log \|c_d\|_2)|)\rangle \end{bmatrix}. \tag{20}$$

and therefore

$$\tilde{\beta}_{.S_j} = \begin{bmatrix} \exp(-|\log \|c_1\|_2)|)^{-1} & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & \exp(-|\log \|c_d\|_2)|)^{-1}\rangle \end{bmatrix}. \tag{21}$$

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 6 *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).

4.1 Implementation

We use the multitask lasso from sklearn and the cvxpy package for basis pursuit. We use the SCS interior point solver from CVXPY, which is able to push sparse values arbitrarily close to 0 **cvxpy˙sparse˙solution**. Data is IRIS and Wine, as well as flat torus from ldle.

4.2 Computational complexity

5 Experiments

Comparison with isometry loss.

6 Discussion

Tangent space basis pursuit satisfies a similar property **Koelle2022-lp** but the normalization process differs.

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.

With respect to pseudoinverse estimation, sparse methods have been applied in **Sun2012-vp**

Even though by Lagrangian duality, the basis pursuit solution corresponds to λ approaching 0, the solution is sparse **Tropp04-ju**. about the lasso is that all coefficients enter the regularization path. As we see by the correspondence between λ approaching 0 and the basis pursuit problem, some coefficients in fact do not go to 0.

7 Supplement

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

Bertsimas2022-qo gives a method for solving the sparse-PCA method more efficiently than the original greedy approach. Compared with the FISTA method used in **Koelle2022-ju**; **Koelle2024-no**, coordinate descent **Friedman-2007-yb**; **Meier2008-ts**; **Qin2013-tx** is faster **Catalina2018-ek**; **Zhao2023-xn**. Compared with **Liu2009-yo**, the sklearn multitask lasso is $2,1$ rather than $\infty, 1$ regularized.

Compared with Gram-Schmidt It is likely that the transformed singular value loss could be reframed as a semdefinite programming problem, since the composition of two convex functions is convex **Boyd2004-ql**.

Multitask lasso **Obozinski2006-kq**; **Yeung2011-fg** is a form of group lasso **Yuan2006-bt** where coefficients are group by response variable.

See **Obozinski2006-kq** for a comparison of forward and backward selection with lasso.

Our notion of isometric recovery is distinct from the restricted isometry property **Candes2005-dd**; **Hastie2015-qa**, 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 disintigration of performance at high λ values in the lasso formulation may have some relation to these properties, as discussed in **Koelle2022-ju**; **Koelle2024-no**.

A major area of comparison is in diversification in recommendation systems. Greedy algorithms are used **Carbonell2017-gi**; **Wu2019-uk**

Compared with sparse pca **Bertsimas2022-qo**; **Bertsimas2022-dv**, we are not concerned with variability in the dataset, and select. 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. **Tropp06-sg** and **Liu2009-yo** use a $1, \infty$ norm to induce sparsity that misses the utility of our normalization for finding unitary matrices. since isometry embeddings preserve important properties like distances between points.

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.