Isometry pursuit

Samson Koelle, Marina Meila

August 6, 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   Background

In this section we give background on both the techniques used in our method, and also the motivating applications on which it is applied.

1.1   Problem

Our goal is, given a matrix $\mathcal{X} \in \mathbb{R}^{D \times P}$, select a subset $\mathcal{S} \subset [P]$ with $|\mathcal{S}| = D_-$ and $D_- \leq D$ such that $X_{\cdot \mathcal{S}}$ is as unitary 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.

1.2   Techniques

1.2.1   Unitary matrices

Definition 1 (Unitary) *A rank D matrix $U \in \mathbb{R}^{D_\alpha \times D_\beta}$ is said to be **unitary** if $U^T U = I_D$.*

    While this definition implies $D_\alpha, D_\beta \geq D$, this paper typically works with equality. Even so, this paper uses the term singular values rather than eigenvalues, and Section **??** contains speculation that a similar set of techniques hold for rank-d unitary rectangular matrices.

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

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

    As a ground truth loss function, we'd like the loss to be minimized uniquely by unitary 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 operator norm, also known as 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. In fact, these losses will behave equivalently over diagonalizable matrices. That is, in a sense, the ground truth loss will be as convex as possible.

    The property of unitary matrices that we use to define our alternative convex loss is slightly different from Proposition 3.

Proposition 2 (Basis vectors of a unitary matrix) *The component vectors $u^1 \ldots u^D \in \mathbb{R}^B$ form a unitary 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 will show that the conditions of the antecedent of Proposition 2 are satisfied by the solutions to the multitask basis pursuit problem applied to matrices consisting of suitably normalized vectors. The vectors are in particular length normalized so that the post-normalization vectors with the longest length have length 1. As we show in Proposition , our main theoretical result, the tendency of the multitask basis pursuit problem to select orthogonal features then ensures that a unitary matrix is deterministically selected, given that one is present.

1.2.2 Basis pursuit and lasso

We use "multitask" instead of "group" to refer to this algorithm, as grouping is entirely determined by the dependent variable.

Basis pursuit programs are of the form

$$\min c(\beta) : d(y, x\beta) \leq o \tag{1}$$

Basis pursuit programs are convex, and solvable via standard interior point methods.

On the other hand, the basis pursuit program admits

1.2.3 Isometries

Another motivating example of this paper is the embedding of non-linear data manifolds by selecting functions from within a dictionary that preserve distances and angles. These sets of functions are called isometries if they do so globally and local isometries if they do so almost everywhere. If more pointwise specificity is required, they are known at isometric at a point. To avoid abstractness, we give the following definition in coordinates.

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 an* **isometry at a point** $p \in \mathcal{M}$ *if*

$$\|v^T D\phi(p)\| = \|v\| \forall v \in \mathbb{R}^D \tag{2}$$

$D\phi$ is the differential implicitly given by $D_F^E \phi = UD\phi V^T$ where U and V are bases for $(T_p\mathcal{M})$ and $T_{\phi(p)}\mathcal{N}$, the tangent spaces of $\mathbb{R}$ and $\mathcal{N}$ at p and $\phi(p)$, respectively.

The reasons that pointwise isometry is interesting are themselves manifold. **Kohli2021-lr** showed how pointwise isometries selected from a dictionary may be stitched together to form global embeddings. Another line of work **Koelle2022-ju**; **Koelle2024-no** has shown how independent functions may be selected globally using group lasso, and that this selection process has implications for the metric properties of the selected embedding **Koelle2022-ju**.

Other approaches like Local Tangent Space Alignment seek to identify isometric coordinates non-parametrically through tangent space estimation prior to stitching, while multidimensional scaling and isomap do so directly through preservation of distances. A classic result in differential geometry equilibrates these two approaches.

Given the metric characterization of isometry in Definition 2, is it appropriate to recall another equivalent characterization of unitary matrices.

Proposition 3 (Metric properties of a unitary matrix) $U \in \mathbb{R}^{D_\alpha \times D_\beta}$ *is unitary if and only if* $\|v^T U$ *for all* $v \in \mathbb{R}^{D_\beta}$.

Thus, in coordinates, $D\phi(p)$ is unitary if ϕ is isometric at a point p.

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

1.3 Applications

The motivating example for this research is the computation of embeddings by selection from within sets of embedding coordinates. However, as the general approach has usefulness outside of non-linear dimension reduction, we first give examples in the relatively simpler setting of diversification.

1.3.1 Diversification

Say you would like to select wine and flowers for an elegant dinner party.

1.3.2 Embedding

Selection of functions from within a dictionary is a standard way to impose interpretability.

2 Method

Recall that our objective is to, 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. Thus, we first will define a function that is uniquely minimized by unitary matrices and some favorable properties for optimization that will be the ground truth we evaluate the success of our method against. We then define the combination of normalization and multitask basis pursuit that approximates this ground truth loss function. We include claims that ground truth and convex loss values are the same for all diagonalizable matrices, and that the convex basis pursuit program recovers to optimum in a deterministic manner should it exist; proofs are given in Section 3.1. We finally define the lasso dual to the basis pursuit program and a post processing method for ensuring that the solution is D sparse. Experimental results using these methods will then be given in Section **??**

2.1 Ground truth

The main goal of isometry pursuit is to expedite the selection of unitary submatrices. More traditional measures of unitariness which use the singular values of a matrix like the log operator norm (i.e. log deformation) and nuclear norm are poorly suited for optimization since they use a subset of the matrix's information and are not uniquely minimized at unitarity, respectively. Thus, we define the loss

$$l_c : \mathbb{R}^{D \times P} \to \mathbb{R}^+ \tag{3}$$

$$\mathcal{X} \mapsto \sum_{d=1}^{D} g(\sigma^d(\mathcal{X}), c) \tag{4}$$

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

$$g : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+ \tag{5}$$

$$t, c \mapsto e^{t^c} + e^{t^{-c}}. \tag{6}$$

Plainly, g is uniquely maximized by unitary 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 1. Most importantly, this loss enables comparison with produced after normalization as in Section 2.2.

The overall algorithm we seek to improve upon is

$$\widehat{\mathcal{S}}_{GT} = \arg \min_{\mathcal{S} \subseteq [P]:|\mathcal{S}|=D} l_c(\mathcal{X}_{\mathcal{S}}) \tag{7}$$

In practice, non-convexity occurs in two places, but only one is essential. The inessential non-convexity is in the computation of l_c. While this function is in fact convex, computation of the individual singular values prior to summation is not, and our experiments rely on such piecemeal computation rather than implementing an end-to-end method. However, the combinatorial search over $[P]$ is inherently non-convex, and requires combinatorial search over all combinations.

2.2   Normalization

We will propose a basis pursuit method which approximates the results of Program 7. 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 in the subsequent regression. This calls for a "normalization" method that differs from other forms in its requirements, and we can't yet prove that these conditions relate it to any sort of norm, even on an appropriately chosen space. This normalization is Now establish some basic conditions for normalization of vectors $v \in \mathbb{R}^D$.

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

$$\arg \max_{v \in \mathbb{R}^D} \ q(v) = \{v : \|v\| = 1\} \tag{8}$$

$$q(v) = q(\frac{v}{\|v\|^2}) \tag{9}$$

$$q(v^1) = q(v^2) \ \forall \ v^1, v^2 : \|v^1\| = \|v^2\| \tag{10}$$

Note that requiring the full structure of a multiplicative norm here is unnecessary for basic success of the algorithm, but certain characteristics such as $q(v^{-1}) = q(v)$ seem desirable, provided one can give a reasonable way to compute v^{-1}, such as by considering each vector as a scaled rotation subgroup of the general linear group. Mindful of this opportunity, and also of the desire to compare with the ground truth and provide computational expediency, consider the normalization by

$$q : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+ \tag{11}$$

$$t, c \mapsto \frac{e^{t^c} + e^{t^{-c}}}{2e}, \tag{12}$$

and use this to define the vector normalization

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

$$n, c \mapsto \frac{n}{q(\|n\|_2, c)} \tag{14}$$

and matrix normalization

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

$$\mathcal{X}_{\cdot p}, c \mapsto n(\mathcal{X}_{\cdot p}, c) \; \forall \; p \in [P]. \tag{16}$$

While this normalization satisfies **??**, it also 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 11 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 ∞. This completes the deterministic data preprocessing.
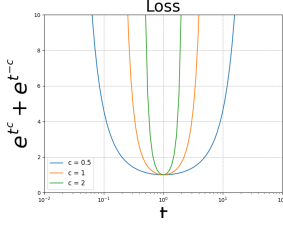

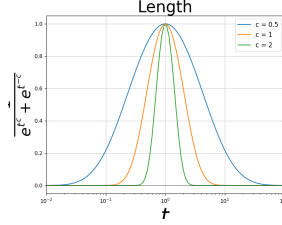Figure 1: Ground truth loss scaling function g as a function of t
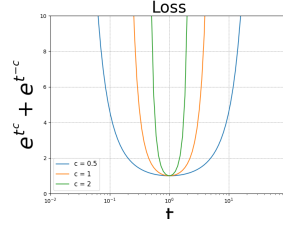
Figure 2: Length as a function of t

Figure 3: Basis pursuit losses as a function of t

Figure 4: 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.

2.3 Isometry pursuit

We will show how to use an appropriate normalized matrix $w(\mathcal{X})$ in multitask basis pursuit to identify submatrices of $\mathcal{X}$ that are as unitary as possible. Multitask basis pursuit is a method for identifying sparse signals from overcomplete dictionaries, and the intuition behind its application in our setting is that submatrices consisting of vectors which are closer to 1 in length and more orthogonal will have smaller loss. In contrast to the typical statistical setting, these features correspond to individual observations in our diversification example, and basis vectors of data manifold tangent spaces in our non-linear dimension reduction example.

Define the multitask basis pursuit penalty

$$\| \cdot \|_{1,2} : \mathbb{R}^{P \times D} \to \mathbb{R}^+ \tag{17}$$

$$\beta \mapsto \sum_{p=1}^{P} \|\beta_{p\cdot}\|_2. \tag{18}$$

7

The isometry pursuit program is then

$$\widehat{\beta}_c^D(\mathcal{X}) := \arg\min_{\beta \in \mathbb{R}^{P \times D}} \|\beta\|_{1,2} \; : \; I_D = w(\mathcal{X}, c)\beta. \tag{19}$$

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

$$S : \mathbb{R}^{p \times d} \to \binom{[P]}{d} \tag{20}$$

$$\beta \mapsto \{p \in [P] : \|\beta_{p.}\| > 0\} \tag{21}$$

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

Proposition 4 (Basis pursuit selection invariance) *Let $U \in \mathbb{R}^{D \times D}$ be unitary. Then $S(\widehat{\beta}(U\mathcal{X})) = S(\widehat{\beta}(\mathcal{X}))$.*

A proof is given in Section 3.1.1 This fact has as an immediate corollary that we may replace I_D in the constraint by any unitary $D \times D$ matrix.

With these preliminaries, we may state our main result. From an intuitive perspective, this is akin to saying selected columns are not only complementary in that they are orthogonal, but also well-balanced in that their columns are of a length 1 rather than length 0.

Proposition 5 (Unitary selection) *Given a matrix $\mathcal{X} \in \mathbb{R}^{D \times P}$ with a rank D submatrix $\mathcal{X}_{.\mathcal{S}} \in \mathbb{R}^{D \times D}$ that is the unique $D \times D$ rank D-unitary submatrix, $\mathcal{S} = S(\widehat{\beta}_c^P(\mathcal{X}))) \forall c \in \mathbb{R}^+$.*

This proof admits two immediate generalizations. First, any normalization function that satisfies the normalization conditions will do. Second, assuming that we do in fact use w for normalization, the ground truth and convex losses are equivalent for diagonalizable matrices. This is summarized in the following proposition, which is slightly stronger than Proposition 5

Proposition 6 (Loss equivalence) *Given a diagonalizable matrix $\mathcal{X} \in \mathbb{R}^{D \times D}$, $\|\widehat{\beta}_c^P(\mathcal{X})\|_{1,2} = l_c(\mathcal{X})$.*

This proposition should be thought of as applying to D column matrices after selection. We know that unitary submatrices minimize loss, and should they be present, they will be selected by either method. This proposition shows that should diagonalizable matrices be selected, they will also have equivalent loss, but not necessarily that they will be selected.

8

2.4 Isometric lasso

The convex loss function 17 and linear constraint in 19 admit a Lagrangian dual which we shall call Isometric Lasso. While we defer full discussion of computational complexity to Section **??**, the general principal which motivates the use of this formulation in our setting is the increased computational expediency in high-dimensions.

The Isometric Lasso loss is

$$l_\lambda(\mathcal{X}, \beta) = \|I_D - \tilde{\mathcal{X}}_c \beta\|_2^2 + \lambda \|\beta\|_{1,2} \tag{22}$$

which can be optimized as

$$\hat{\beta}_\lambda(\mathcal{X}) = \arg \min_{\beta \in \mathbb{R}^{P \times D}} l_\lambda(\mathcal{X}, \beta) \tag{23}$$

Similarly to Section **??**, we assert that $S(\widehat{\beta}_\lambda(\mathcal{X}))$.

Proposition 7 (Lasso selection equivalence) *Let $U \in \mathbb{R}^{D \times D}$ be unitary. Then $S(\widehat{\beta}_\lambda(U\mathcal{X})) = S(\widehat{\beta}_\lambda(\mathcal{X}))$.*

This also covers changing the target variable.

Note that it also may be possible to argue the basis pursuit invariance from the lasso ones plus Lagranian duality, but to avoid taking the limit we prove both propositions independently.

2.5 Extension to non-linear spaces

Proposition 8 (Isometry at a point selection) *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 $\arg \min_\beta$.*

A proof is given in Section 3.1.2.

2.6 Two-stage isometry pursuit

A standard approach in the lasso literature is to first use the lasso to prune, prior to a final feature selection step **Hesterberg2008-iy** This avoids issues from shrinkage caused by the lasso estimator at high values of λ. For example, we cannot as of yet prove analogs of Proposition 5 for the Lasso formulation, and similar lasso-specific conditions such as those found in **Koelle2024-no** are not satisfied by overcomplete dictionaries.

2.7 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.

2.8 Computational complexity

3 Supplement

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

3.1 Proofs

3.1.1 Proof of Proposition 4

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

Proposition 9 *Loss equivalence Let $U \in \mathbb{R}^{D \times D}$ be unitary. Then $\|\beta\|_{1,2} = \|\beta U\|$.*

Proof:

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

$$= \sum_{p=1}^{P} \|\beta_{p.}\| \qquad (25)$$

$$= \|\beta\|_{1,2} \qquad (26)$$

$$\square$$

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

Proposition 10 *Let $U \in \mathbb{R}^{D \times D}$ be unitary. Then $\widehat{\beta}(U\mathcal{X}) = \widehat{\beta}(\mathcal{X})U$.*

Proof:

$$\widehat{\beta}(U\mathcal{X}) = \arg \min_{\beta \in \mathbb{R}^{P \times D}} \|\beta\|_{1,2} \; : \; I_D = UX\beta \qquad (27)$$

$$= \arg \min_{\beta \in \mathbb{R}^{P \times D}} \|\beta\|_{1,2} \; : \; U^{-1}U = U^{-1}UX\beta U \qquad (28)$$

$$= \arg \min_{\beta \in \mathbb{R}^{P \times D}} \|\beta\|_{1,2} \; : \; I_D = X\beta U \qquad (29)$$

$$= \arg \min_{\beta \in \mathbb{R}^{P \times D}} \|\beta U\|_{1,2} \; : \; I_D = X\beta U \qquad (30)$$

$$= \arg \min_{\beta \in \mathbb{R}^{P \times D}} \|\beta\|_{1,2} \; : \; I_D = X\beta. \qquad (31)$$

$$\square$$

3.1.2 Proof of Proposition 5

In support of the proposition in the main text, we give a more general theorem which illustrates the key features of the method. From there, it is a straightforward step to specify to the chosen normalization.

Proposition 11 (Generalized unitary selection) *Given a matrix $\mathcal{X} \in \mathbb{R}^{D \times P}$ with a rank D submatrix $\mathcal{X}_{.S} \in \mathbb{R}^{D \times D}$ that is the unique $D \times D$ rank D-unitary submatrix, and let n_q be a normalization composed of normalizing functions q that satisfy the conditions of Definition 3, then $\mathcal{S} = S(\widehat{\beta}_q^P(\mathcal{X}))$.*

Proof: This proof consists of showing that replacing an arbitrary D subset $\mathcal{S}$ with a unitary $D \times D$ matrix leads to existence of a solution which reduces loss, and that if such a subset already exists, loss may not be reduced any further.

Proposition 12 (Unitary Preference) *Given a rank D matrix $X \in \mathbb{R}^{D \times P}$, let β be the solution to Program 19. Select an arbitrary subset $\mathcal{S} \subseteq [P]$ with $|\mathcal{S}| = D$. Define a new matrix $X' \in \mathbb{R}^{D \times P}$ as follows:*

$$X'_{.p} = \begin{cases} X_{.p} & \text{if } p \notin \mathcal{S}, \\ U_{.p} & \text{if } p \in \mathcal{S}, \end{cases}$$

where $U \in \mathbb{R}^{D \times D}$ is any unitary matrix.

Proof: By the conditions in Definition 3, the maximum length for vectors $\mathcal{X}_{.p}^q$ is attained when $\|\mathcal{X}_{.p}\|_2 = 1$. By Proposition 2, this is the case for unitary matrices, but it is also true for matrices for which $\mathcal{X}_{.p}^q$ are not orthogonal.

Proposition 13 (Support orthogonalization) *Let β'_S be the solution to*

$$X_S \beta_S = X'_S \beta'_S. \tag{32}$$

Note that this is guaranteed to exist since X'_S is full rank. Then $\|\beta_S\|_{1,2} \geq \|\beta'_S\|_{1,2}$

Proof: Since X'_S is full rank we can define $\gamma_{SS} \in \mathbb{R}^{D \times D}$ such that

$$X_S = X'_S \gamma_{SS}. \tag{33}$$

Then

$$X_S \beta_S = X'_S \gamma_{SS} \beta_S \tag{34}$$

and so

$$\gamma_{SS} \beta_S = \beta'_S. \tag{35}$$

12

Thus, we want to show that

$$\|\gamma_{SS}\beta_S\|_{1,2} \leq \|\beta_S\|_{1,2}. \tag{36}$$

Since $\|X'_S\|$ is unitary, $\|X'_S\gamma_{Sp}\|_2 = \|\gamma_{Sp}\|_2$. Therefore, Equation 33 and $\|X_{.p}\|_2 \leq 1$ imply that $\|\gamma_{Sp}\|_2 \leq 1 \; \forall p \in S$. Thus, showing the proposition amounts to showing that the $\|\|_{1,2}$ norm is submultiplicative.

In lieu of showing directly, we show that $\|\|_{1,2}$ is subadditive and therefore, since it is unitary invariant by Proposition 10 and $\|\mathrm{diag}(1,0,\ldots,0)\|_{1,2} = 1$ satisfies the conditions in **HIAI2004155**, and thus is submultiplicative.

Proposition 14 (Subadditivity of group lasso norm) *Let $A, B \in \mathbb{R}^{D \times D}$. Then $\|A + B\|_{1,2} \leq \|A\|_{1,2} + \|B\|_{1,2}$*

Proof:

$$\|A + B\|_{1,2} = \sum_{d=1}^{D} \|A_{d.} + B_{d.}\|_2 \tag{37}$$

$$\leq \sum_{d=1}^{D} \|A_{d.}\|_2 + \|B_{d.}\|_2 \tag{38}$$

$$= \|A\|_{1,2} + \|B\|_{1,2} \tag{39}$$

$\square$

Thus, $\|\beta_S\|_{1,2} \geq \|\beta'_S\|_{1,2}$.

$\square$

The next component of the argument deals with elements of $[P]$ which are not contained within $\mathcal{S}$. We show that zeroing components of $\beta'_{p.}$ where $p \notin S$ and compensating for the removal by altering the loadings within $\mathcal{S}$ results in lower loss.

Proposition 15 (Support replacement) *Let β_S" be the solution to*

$$X'_S\beta'_S + X_{.c}\beta_{c.}^T = X'_S\beta_S". \tag{40}$$

Note that this is guaranteed to exist since X'_S is full rank. Then $\|\beta'_S\|_{1,2} + \|\beta_c\|_2 \geq \|\beta_S"\|_{1,2}$

Proof: Since X'_S is full rank we can define $\gamma_{Sc} \in \mathbb{R}^D$ such that $X_{.c} = X'_S\gamma_{Sc}$. Then

$$X'_S\beta_S" = X'_S\beta'_S + X_{.c}\beta_c^T \tag{41}$$

$$= X'_S\beta'_S + X'_S\gamma_{Sc}\beta_c^T \tag{42}$$

$$= X'_S(\beta'_S + \gamma_{Sc}\beta_c^T) \tag{43}$$

13

Thus, $\beta_S" = \beta'_S + \gamma_{Sc}\beta_c^T$.

We therefore want to show that $\|\beta'_S\|_{1,2} + \|\beta_c.\|_2 \geq \|\beta'_{S.} + \gamma_{Sc}\beta_{c.}^T\|_{1,2}$. Another way to say this is $\sum(\|\beta'_p + \gamma_{pc}\beta_c\|_2 - \|\beta'_p\|_2) \leq \|\beta_c.\|$

$\square$

$\square$

$\square$

3.1.3 Proof of Proposition 7

This proof proceeds similarly to the proof of Proposition **??** . Proposition 7 relies on the fact that its loss is invariant under any unitary transformation. As a corollary, this fact gives that the identify matrix which is the "dependent variable" in the regression equation may be replaced by any $d \times d$ unitary matrix. For Proposition **??**, the loss is also invariant under unitary, transformation, but we also check that this transformation. Once again, this also implies that any unitary matrix may replace the identity in the constraint.

Proposition 16 *Loss equivalence Let $U \in \mathbb{R}^{D \times D}$ be unitary. Then $l_\lambda(\mathcal{X}, \beta) = l_\lambda(U\mathcal{X}, \beta U)$.*

Proof: We can write

$$l^*(X^i) = l(\beta^i) = \sum_{j=1}^{p}(\sum_{i'=2}^{n} \|\beta_{i'j.}\|_2^2 + \|\beta_{1j.}^i\|_2^2)^{1/2} = \sum_{j=1}^{p}(\sum_{i'=1}^{n} \|\beta_{i'j.}U\|_2^2)^{1/2} = l^*(X) \quad (44)$$

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

$\square$

Proposition 17 *Programmatic equivalence Let $U \in \mathbb{R}^{D \times D}$ be unitary. Then $\hat{\beta}_\lambda(U\mathcal{X}) = U\hat{\beta}_\lambda(\mathcal{X})$.*

3.2 Experiments