JDSSV

Journal of Data Science, Statistics, and Visualisation

July 2021, Volume II, Issue II. *doi: 10.52933/jdssv.v2i2.61*

On Generalization and Computation of Tukey's Depth: Part II

Yiyuan She
Florida State University

Shao Tang
Florida State University

Jingze Liu
Florida State University

Abstract

This paper studies how to generalize Tukey's depth to problems defined in a restricted space that may be curved or have boundaries, and to problems with a nondifferentiable objective. First, using a manifold approach, we propose a broad class of Riemannian depth for smooth problems defined on a Riemannian manifold, and showcase its applications in spherical data analysis, principal component analysis, and multivariate orthogonal regression. Moreover, for nonsmooth problems, we introduce additional slack variables and inequality constraints to define a novel slacked data depth, which can perform center-outward rankings of estimators arising from sparse learning and reduced rank regression. Real data examples illustrate the usefulness of some proposed data depths.

Keywords: Riemannian depth, principal component analysis, slacked data depth, reduced rank regression, sparsity-promoting regularizers.

1. Introduction

Tukey's half-space depth (Tukey 1975) can be generalized to a polished subspace depth, as shown in our companion paper (She et al. 2022a). The basic Tukeyfication process there assumes a simple problem structure in the sense that one can directly write down some sample-additive estimating equations. Modern statistical applications however pose new challenges.

First, the parameter space Ω can be curved or have boundaries, so that evaluating the gradient in the ambient Euclidean space may not directly deliver reasonable influences. Second, with a regularizer in use, the objective function is typically non-differentiable. Some specific examples are given as follows.

Example 1. (Watson depth) Assume that all data points lie on an m-dimensional sphere $\boldsymbol{z}_i \in \mathbb{S}^{m-1}$, and $\pm\boldsymbol{z}_i$ are deemed equivalent. This kind of data are typically referred to as axially symmetric data. They have recently received attention in clustering and directional statistics (Dhillon et al. 2003; Bijral et al. 2007; Sra and Karp 2013). To characterize the distribution of such data, a commonly used one is the Watson distribution with density (Watson 1965; Mardia and Jupp 1999)

$$p(\boldsymbol{z}; \boldsymbol{\mu}, \kappa) \propto e^{\kappa(\boldsymbol{\mu}^T \boldsymbol{z})^2}.$$

Here, $\boldsymbol{\mu} \in \mathbb{S}^{m-1}$ gives the mean direction, κ is the so-called concentration parameter, and the normalizing constant does not depend on $\boldsymbol{\mu}$. We require $\kappa \neq 0$ (otherwise $\boldsymbol{\mu}$ is not an effective parameter to introduce depth). When $\kappa > 0$, the data points concentrate around $\boldsymbol{\mu}$, and when $\kappa < 0$, the data spread around the great circle orthogonal to $\boldsymbol{\mu}$. How to "Tukeyfy" more complex distributions defined on a sphere (such as the *Fisher-Bingham distribution*) is nontrivial, but could give rise to more useful spherical data depths. We will see that depth-enhanced principal component analysis to be introduced in Section 2.2 poses a similar manifold challenge.

Example 2. (Nonnegative regression depth) As an extension of the celebrated regression depth (Rousseeuw and Hubert 1999), let us consider data depth in a setting where all coefficients are nonnegative. This corresponds to the nonnegative least squares problem:

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \text{ s.t. } \beta_j \geq 0, 1 \leq j \leq p,$$

where we can denote the constraints by $\boldsymbol{\beta} \in \mathbb{R}^p_+$ with $\mathbb{R}_+ = [0, \infty)$. Clearly, the closed parameter space $\mathbb{R}^p_+$ has boundary points. Regression depth can be simply applied if $\boldsymbol{\beta}^\circ$ is an interior point, but if $\boldsymbol{\beta}^\circ$ lies on the boundary, i.e., $\beta_j^\circ = 0$ for some j, which is of practical interest in significance tests, regular depth does not apply, and the normal-equation based influences must be corrected—but how?

Example 3. (Sparsity depth) Consider a sparse learning problem

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} l_0(\boldsymbol{x}_i^T \boldsymbol{\beta}) + \sum_{j=1}^{p} P(|\beta_j|; \lambda),$$

where l_0 is a loss function defined on the systematic component $\boldsymbol{x}_i^T\boldsymbol{\beta}$ and P is a penalty function to promote sparsity in $\boldsymbol{\beta}$. Examples of P include ℓ_1, ℓ_0, SCAD (Fan and Li 2001), and MCP (Zhang 2010), among many others that are popularly used in high dimensional statistics for building a parsimonious model. We assume that the regularization parameter λ is given, either by theory—see, e.g., Cai et al. (2009), or by tuning, like cross-validation (She and Tran 2019), so that the criterion is fully specified. A new class of depths like ℓ_1-depth or ℓ_0-depth would be helpful for high-dimensional robust inference, but the nondifferentiability and nonconvexity of P make it difficult to obtain sample-additive estimation equations.

To tackle the challenges, we propose two approaches based on manifolds and slack variables, respectively, to extend Tukey's depth to Riemannian depth and slacked depth. The rest of the paper is organized as follows. Section 2 studies how to handle a smooth problem defined on a Riemannian manifold. The resulting Riemannian depth finds applications in spherical data analysis, principal component analysis, and multivariate orthogonal regression. Section 3 uses slack variables to cope with parameter spaces with boundaries and nondifferentiable objectives. A novel class of slacked data depth can perform center-outward rankings of estimators arising from sparse learning and reduced rank regression. Section 4 performs computer experiments on some real data examples. We conclude the whole work in Section 5.

Notation. We use bold symbols to denote vectors and matrices. A matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is frequently partitioned into rows $\boldsymbol{X} = [\boldsymbol{x}_1 \dots \boldsymbol{x}_n]^T$ with $\boldsymbol{x}_i \in \mathbb{R}^p$. The vectorization of $\boldsymbol{X}$ is denoted by $\mathrm{vec}(\boldsymbol{X}) \in \mathbb{R}^{np}$. Let $\mathbb{R}_+ = [0, +\infty]$. We use $\boldsymbol{X}[\mathcal{I}, \mathcal{J}]$ to denote a submatrix of $\boldsymbol{X}$ with rows and columns indexed by $\mathcal{I}$ and $\mathcal{J}$, respectively, and occasionally abbreviate $\boldsymbol{X}[, \mathcal{J}]$ to $\boldsymbol{X}_{\mathcal{J}}$ by selecting the corresponding columns. Given $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\|\boldsymbol{X}\|_F$ and $\|\boldsymbol{X}\|_2$ denote its Frobenius norm and spectral norm, respectively, $\|\boldsymbol{X}\|_{\max} \triangleq \max_{1\le i\le n, 1\le j\le p} |x_{ij}|$, and $\mathrm{rank}(\boldsymbol{X})$ denotes its rank. The Moore-Penrose inverse of $\boldsymbol{X}$ is denoted by $\boldsymbol{X}^+$. The inner product of two matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ (of the same size) is defined as $\langle \boldsymbol{X}, \boldsymbol{Y}\rangle = \mathrm{Tr}(\boldsymbol{X}^T\boldsymbol{Y})$ and their element-wise product (Hadamard product) is $\boldsymbol{X} \circ \boldsymbol{Y}$. The Kronecker product is denoted by $\boldsymbol{X} \otimes \boldsymbol{Y}$ (where $\boldsymbol{X}$ and $\boldsymbol{Y}$ need not have the same dimensions). Given a set $\mathcal{A} \subset \mathbb{R}^{p\times m}$ and a matrix $\boldsymbol{T} \in \mathbb{R}^{n\times p}$, $\boldsymbol{T} \circ \mathcal{A} = \{\boldsymbol{TA} : \boldsymbol{A} \in \mathcal{A}\}$. We use $\mathbb{O}^{m\times r}$ to represent the set of all $m \times r$ matrices $\boldsymbol{V}$ satisfying the orthogonality constraint $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$. For a vector $\boldsymbol{a} = [a_1, \dots, a_n]^T \in \mathbb{R}^n$, $\mathrm{diag}\{\boldsymbol{a}\}$ is defined as an $n \times n$ diagonal matrix with diagonal entries given by $a_1, \dots, a_n$, and for a square matrix $\boldsymbol{A} = [a_{ij}]_{n\times n}$, $\mathrm{diag}(\boldsymbol{A}) := \mathrm{diag}\{a_{11}, \dots, a_{nn}\}$. The indicator function $1_{\mathcal{A}}(t)$ means $1_{\mathcal{A}}(t) = 1$ if $t \in \mathcal{A}$ and 0 otherwise. Given $f : \mathbb{R}^{n\times p} \to \mathbb{R}$, $f \in \mathcal{C}^1$ means that its Euclidean gradient $\nabla f(\boldsymbol{X})$, an $n \times p$ matrix with the (i, j) element $\partial f/\partial x_{ij}$, exists and is continuous for any $\boldsymbol{X} \in \mathbb{R}^{n\times p}$. Given two vectors $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^p$, $\boldsymbol{\alpha} \succeq \boldsymbol{\beta}$ means $\alpha_j \ge \beta_j, 1 \le j \le p$ and $\boldsymbol{\alpha} \succ \boldsymbol{\beta}$ means $\alpha_j > \beta_j, 1 \le j \le p$. Finally, $a \wedge b = \min\{a, b\}$.

2. A Manifold Approach

When the problem is defined on a Riemannian manifold (without boundaries), we can introduce Riemannian influences, along with defining a proper influence space to complete the definition of Riemannian depth. In contrast, the commonly used methods to deal with constraints (such as the elimination approach for (13) below) may be infeasible in higher dimensions. We will see the important role of the influence space $\mathcal{G}$ introduced in Section 2 of She et al. (2022a), since a Riemannian gradient always lies in a tangent space.

2.1. Riemannian depth

We begin with Example 1 to motivate the main idea. Starting from such an example, the Watson depth will be introduced. Useful for the analysis of axial data, it is defined

on a Riemmanian manifold and it will be a special case of a more generic Riemannian depth.

For the MLE problem

$$\min_{\boldsymbol{\mu}} -\kappa \sum_i \langle \boldsymbol{\mu}, \boldsymbol{z}_i \rangle^2 + c(\kappa; m) \text{ s.t. } \|\boldsymbol{\mu}\|_2^2 = 1, \tag{1}$$

Lagrange multiplier or eigenvalue decomposition can be used to solve for $\boldsymbol{\mu}$, but they do not yield a simple set of estimating equations like (1) in She et al. (2022a) to be conveniently used for the purpose of data depth.

Instead, we view (1) as an *unconstrained* problem on the sphere $\mathbb{S}^{m-1}$, which is a Riemannian manifold. Then the Riemannian gradient with respect to $\boldsymbol{\mu}$ can be calculated to define the desired (Riemannian) influence function. We show some detailed derivation to give the reader more intuition.

Concretely, adopting the canonical metric induced by the inner product $G_{\boldsymbol{\mu}}(\boldsymbol{u}_1, \boldsymbol{u}_2) = \boldsymbol{u}_1^T(\boldsymbol{I} - \boldsymbol{\mu}\boldsymbol{\mu}^T/2)\boldsymbol{u}_2$ (Edelman et al. 1998), the Riemannian gradient of $l_i := -\kappa\langle \boldsymbol{\mu}, \boldsymbol{z}_i \rangle^2$ with respect to $\boldsymbol{\mu}$, denoted by $\boldsymbol{g}_i(\boldsymbol{\mu})$, is defined as the unique element in the tangent space

$$\mathcal{T}_{\boldsymbol{\mu}}(\mathbb{S}^{m-1}) \triangleq \{\boldsymbol{u} \in \mathbb{R}^m : \boldsymbol{u}^T\boldsymbol{\mu} = 0\}$$

satisfying

$$G_{\boldsymbol{\mu}}(\boldsymbol{g}_i(\boldsymbol{\mu}), \boldsymbol{u}) = \boldsymbol{u}^T \nabla l_i, \ \forall \boldsymbol{u} \in \mathcal{T}_{\boldsymbol{\mu}}(\mathbb{S}^{m-1}) \tag{2}$$

where ∇l_i is the Euclidean gradient. It follows that

$$\boldsymbol{g}_i(\boldsymbol{\mu}) = [\nabla l_i \boldsymbol{\mu}^T - \boldsymbol{\mu}(\nabla l_i)^T]\boldsymbol{\mu} = -2\kappa\langle \boldsymbol{z}_i, \boldsymbol{\mu}\rangle(\boldsymbol{z}_i - \langle \boldsymbol{z}_i, \boldsymbol{\mu}\rangle\boldsymbol{\mu}).$$

From Boothby (1986) (and $\kappa \neq 0$), the optimal $\boldsymbol{\mu}$ satisfies

$$\sum_i \langle \boldsymbol{z}_i, \boldsymbol{\mu}\rangle(\boldsymbol{z}_i - \langle \boldsymbol{z}_i, \boldsymbol{\mu}\rangle\boldsymbol{\mu}) = \boldsymbol{0}. \tag{3}$$

Given $\boldsymbol{\mu}^\circ \in \mathbb{S}^{m-1}$, the Riemannian influence $\langle \boldsymbol{z}_i, \boldsymbol{\mu}^\circ\rangle(\boldsymbol{z}_i - \langle \boldsymbol{z}_i, \boldsymbol{\mu}^\circ\rangle\boldsymbol{\mu}^\circ)$, denoted by $\boldsymbol{T}^{\mathrm{R}}(\boldsymbol{\mu}^\circ; \boldsymbol{z}_i)$, is no longer $\boldsymbol{z}_i - \boldsymbol{\mu}^\circ$ as in location depth. Notably, $\boldsymbol{T}^{\mathrm{R}}(\boldsymbol{\mu}^\circ; \boldsymbol{z}_i)$ vanishes when $\theta_i = j\pi/2$ $(j = 0, 1, 2, 3)$ with $\cos\theta_i = \langle \boldsymbol{z}_i, \boldsymbol{\mu}^\circ\rangle$, corresponding to various circumstances with $\kappa > 0$ and $\kappa < 0$.

Not only does the manifold perspective provide the desirable estimation equations, but it defines an important influence space $\mathcal{G} = \mathcal{T}_{\boldsymbol{\mu}^\circ}(\mathbb{S}^{m-1})$ to restrict $\boldsymbol{v}$. Accordingly, our Watson depth considers all one-dimensional projections *tangentially* passing through $\boldsymbol{\mu}^\circ$:

$$d_{01}^{\mathrm{W}}(\boldsymbol{\mu}^\circ) = \min_{\boldsymbol{v}} \sum_i 1_{\geq 0}(G_{\boldsymbol{\mu}^\circ}(\boldsymbol{v}, (\boldsymbol{z}_i^T\boldsymbol{\mu}^\circ)[\boldsymbol{z}_i - (\boldsymbol{z}_i^T\boldsymbol{\mu}^\circ)\boldsymbol{\mu}^\circ])) \text{ s.t. } \boldsymbol{v}^T\boldsymbol{\mu}^\circ = 0, \boldsymbol{v}^T\boldsymbol{v} = 1 \tag{4}$$

or equivalently

$$d_{01}^{\mathrm{W}}(\boldsymbol{\mu}^\circ) = \min_{\boldsymbol{v}} \sum_i 1_{\geq 0}(\langle \boldsymbol{v}, (\boldsymbol{z}_i^T\boldsymbol{\mu}^\circ)\boldsymbol{z}_i\rangle) \text{ s.t. } \boldsymbol{v}^T\boldsymbol{\mu}^\circ = 0, \boldsymbol{v}^T\boldsymbol{v} = 1, \tag{5}$$

regardless of the Riemannian metric, as an outcome of (2). The factor $\boldsymbol{z}_i^T\boldsymbol{\mu}^\circ$ in (5), possibly negative, amounts to replacing $\boldsymbol{z}_i$ by $\mathrm{sgn}\langle \boldsymbol{z}_i^T, \boldsymbol{\mu}^\circ\rangle \cdot \boldsymbol{z}_i$. This is in accordance

with the Watson distribution for axially symmetric spherical data. The algorithms in Section 3 of our companion paper can be applied, after a simple reparametrization of $\boldsymbol{v}$ in the orthogonal complement space of $\boldsymbol{\mu}^\circ \boldsymbol{\mu}^{\circ T}$.

The above derivation is standard and can be generalized to introduce a *Riemannian depth* for the Tukeyfication of a differentiable loss l on a Riemannian manifold $\mathcal{M}$ of an Euclidean space: $\min_{\boldsymbol{B}} \sum_i l(\boldsymbol{B}; \boldsymbol{x}_i, \boldsymbol{y}_i)$ s.t. $\boldsymbol{B} \in \mathcal{M}$. Given a point $\boldsymbol{B}^\circ \in \mathcal{M}$ of interest, letting $\boldsymbol{T}_i^\circ = \boldsymbol{T}(\boldsymbol{B}^\circ; \boldsymbol{x}_i, \boldsymbol{y}_i) = \nabla_{\boldsymbol{B}} l(\boldsymbol{B}^\circ; \boldsymbol{x}_i, \boldsymbol{y}_i)$ as before and considering all directional derivatives of l in the directions of $\boldsymbol{V} \in \mathcal{T}_{\boldsymbol{B}^\circ}(\mathcal{M})$, we define

$$\textbf{Riemannian depth: } d_{01}^{\mathrm{R}}(\boldsymbol{B}^\circ) = \min_{\boldsymbol{V}} \sum_i 1_{\geq 0}(\langle \boldsymbol{V}, \boldsymbol{T}_i^\circ \rangle) \quad \text{s.t. } \boldsymbol{V} \in \mathcal{T}_{\boldsymbol{B}^\circ}(\mathcal{M}), \|\boldsymbol{V}\|_F = 1. \tag{6}$$

Eqn. (6) performs location depth of Riemannian influences in the tangent space $\mathcal{T}_{\boldsymbol{B}^\circ}(\mathcal{M})$. Because $\mathcal{T}_{\boldsymbol{B}^\circ}(\mathcal{M})$ is linear, the restricted Procrustes rotation in Section 3 of She et al. (2022a) applies with no difficulty in optimization.

When $\mathcal{M}$ is compact and/or l is nonconvex, it becomes necessary to exclude locally maximal solutions in the estimating equations. We give an "*order-2 Tukeyfication*" as follows. Given $\boldsymbol{B}^\circ \in \mathcal{M}$ and $\boldsymbol{V} \in \mathcal{T}_{\boldsymbol{B}^\circ}(\mathcal{M})$, let γ be the geodesic satisfying $\gamma(0) = \boldsymbol{B}^\circ$ and $\gamma'(0) = \boldsymbol{V}$. The first step is to restrict l to the geodesic and define

$$g_i = \frac{\mathrm{d}}{\mathrm{d}t} l(\gamma(t); \boldsymbol{x}_i, \boldsymbol{y}_i)\Big|_{t=0,}, \quad h_i = \frac{\mathrm{d}^2}{\mathrm{d}t^2} l(\gamma(t); \boldsymbol{x}_i, \boldsymbol{y}_i)\Big|_{t=0,}$$

where g_i simplifies to $\langle \boldsymbol{V}, \boldsymbol{T}_i^\circ \rangle$ and h_i can be calculated via Riemannian Hessian. (Our companion paper mostly considers an Euclidean $\mathcal{M}$, where a line restriction $l(\boldsymbol{B}^\circ + t\boldsymbol{V})$ with $\boldsymbol{V} \neq \boldsymbol{0}$ is used, and g_i and h_i only involve the ordinary gradient and Hessian of l.) The second step robustly measures how well the following two optimality condition are obeyed:

$$\sum_{i=1}^n g_i = 0, \quad \sum_{i=1}^n h_i \geq 0.$$

Concretely, changing the one-dimensional averages to medians motivates us to adopt $\sum (1_{=0} + 2(1_{<0} \wedge 1_{>0}))\,(g_i)$ and $\sum 1_{\geq 0}(h_i)$ to quantify to what extent the two conditions are satisfied, respectively, in the possible occurrence of extreme outliers. Finally, combining the two measures leads to

$$\textbf{Riemannian depth (order 2): } d_{01}^{\mathrm{R2}}(\boldsymbol{B}^\circ) = \min_{\boldsymbol{V}} \sum_i 1_{\gtrsim 0}(g_i) \sum_i 1_{\geq 0}(h_i) \quad \text{s.t. } \boldsymbol{V} \in \mathcal{T}_{\boldsymbol{B}^\circ}(\mathcal{M}), \|\boldsymbol{V}\|_F = 1,$$

where $1_{\gtrsim 0} := 0.5 \cdot 1_{=0} + 1_{>0}$ replaces $1_{=0} + 2(1_{<0} \wedge 1_{>0})$ in the optimization because g_i is linear in $\boldsymbol{V}$, the Riemannian Hessian is a bilinear map, and $T_{\boldsymbol{B}^\circ}(\mathcal{M})$ is a linear space. A more aggressive proposal is to use $\sum_i 1_{\gtrsim 0}(g_i) 1_{\geq 0}(h_i)$ as the criterion. (Notice the mild difference between $1_{\gtrsim 0}$ and $1_{\geq 0}$; the first seems to be more appropriate to deal with equality-type optimality conditions in defining a d_{01}-type data depth.) When l is (geodesically) convex, $h_i \geq 0$ and thus the associated factor with proper scaling, $\sum_i 1_{\geq 0}(h_i)/n$, will not affect the depth.

Remark 1. *If we Tukeyfy the basic* von Mises-Fisher distribution *(vMF) (Mardia and Jupp 1999), with the density given by* $p(\boldsymbol{z};\boldsymbol{\mu},\kappa) \propto e^{\kappa \boldsymbol{\mu}^T \boldsymbol{z}}$*, where* $\boldsymbol{\mu} : \|\boldsymbol{\mu}\|_2 = 1$ *is the mean direction,* $\kappa > 0$ *and the normalizing constant does not depend on* $\boldsymbol{\mu}$*, (6) yields* $d^R_{01}(\boldsymbol{\mu}^\circ) = \min_{\boldsymbol{v}} \sum_i 1_{\geq 0}(\langle \boldsymbol{v}, \boldsymbol{z}_i\rangle)$ *s.t.* $\boldsymbol{v}^T\boldsymbol{\mu}^\circ = 0, \boldsymbol{v}^T\boldsymbol{v} = 1$*. This is closely related to but different from the angular Tukey's depth that can be defined as* $\min_{\boldsymbol{v}} \sum_i 1_{\geq 0}(\langle \boldsymbol{v}, \boldsymbol{z}_i\rangle)$ *s.t.* $\boldsymbol{v}^T\boldsymbol{\mu}^\circ \geq 0, \boldsymbol{v}^T\boldsymbol{v} = 1$ *for* m*-dimensional spherical data (see Liu and Singh (1992) for some theoretical studies when* $m = 2, 3$*). The order-2 depths involve* $h_i = \langle \boldsymbol{\mu}^\circ, \boldsymbol{z}_i\rangle$ *which are independent of* $\boldsymbol{v}$ *in this case.*

More interesting notions of spherical data depth can be induced by some more flexible distributions through our manifold framework, such as the Kent distribution *and the more general* Fisher-Bingham distribution *whose quadratic exponential form is more powerful than vMF for statistical modeling in bioinformatics, meteorology, and computer vision.*

2.2. Depth-enhanced principal component analysis

This part uses the Riemannian depth introduced in the last subsection to Tukeyfy the well-known principal component analysis (PCA). Let $\boldsymbol{Z} = [\boldsymbol{z}_i, \ldots, \boldsymbol{z}_n]^T \in \mathbb{R}^{n\times m}$ be a data matrix. The PCA model can be stated as

$$\boldsymbol{Z} = \mathbf{1}\boldsymbol{\mu}^{*T} + \boldsymbol{A}^*\boldsymbol{U}^{*T} + \boldsymbol{E}, \tag{7}$$

with $\boldsymbol{\mu}^* \in \mathbb{R}^m$, $\boldsymbol{A}^* \in \mathbb{R}^{n\times r}$, and $\boldsymbol{U}^* \in \mathbb{O}^{n\times r}$ all unknown. Eqn. (7) means that the n data points, after some proper translation, all approximately concentrate in an r-dimensional subspace, and r is typically much lower than m and n. The columns of $\boldsymbol{U}^*$ are often called the principal component (**PC**) loading directions. Assuming that the entries of $\boldsymbol{E}$ are i.i.d. Gaussian, we can estimate the intercept vector and the low-dimensional subspace by

$$\min_{(\boldsymbol{U},\boldsymbol{\mu})} \|(\boldsymbol{Z} - \mathbf{1}\boldsymbol{\mu}^T)(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^T)\|_F^2 \quad \text{s.t.} \quad \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_{r\times r}. \tag{8}$$

The solution is given by standard PCA, which is however sensitive to outliers.

One may want to estimate $\boldsymbol{U}, \boldsymbol{\mu}$ more robustly through a depth enhancement. Here, the orthogonality constraint $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$ may appear more complex than that in the spherical problem (1), but (8) is a smooth problem on a Stiefel manifold. Therefore, we can define a Riemannian depth for any $(\boldsymbol{\mu}^\circ, \boldsymbol{U}^\circ) \in \mathbb{R}^m \times \mathbb{O}^{m\times r}$ based on Section 2.1, which we call the *principal component* (**PC**) depth, as follows

$$\begin{aligned} \textbf{PC-depth}: \quad & \min_{(\boldsymbol{v},\boldsymbol{V})} \sum_i 1_{\geq 0}(\langle \boldsymbol{v}, (\boldsymbol{I} - \boldsymbol{U}^\circ\boldsymbol{U}^{\circ T})(\boldsymbol{\mu}^\circ - \boldsymbol{z}_i)\rangle - \langle \boldsymbol{V}, (\boldsymbol{\mu}^\circ - \boldsymbol{z}_i)(\boldsymbol{\mu}^\circ - \boldsymbol{z}_i)^T\boldsymbol{U}^\circ\rangle) \\ & \text{s.t. } \boldsymbol{V}^T\boldsymbol{U}^\circ + \boldsymbol{U}^{\circ T}\boldsymbol{V} = \mathbf{0}, \ \|\boldsymbol{v}\|_2^2 + \|\boldsymbol{V}\|_F^2 = 1. \end{aligned} \tag{9}$$

All matrix differentiation details are omitted. (9) may need an order-2 modification though, which will be clearly revealed by comparing it to (14) later.

PCA is also helpful when ranking observations in ultra-high dimensions. It is well known that the curse of dimensionality may make every observation look like a corner

point, thus harmful to describing data depth. Fortunately, under (7), the true signals concentrate in the PC subspace determined by $\boldsymbol{U}^*$; so to check a given point's centrality or extremity, it is helpful to project it onto the orthogonal complement (**OC**) subspace to reveal its outlyingness. See She et al. (2016) for more discussions. Specifically, letting $\bar{\boldsymbol{U}}^* \in \mathbb{O}^{m\times \bar{r}}$ ($\bar{r} \le m - r$) that is orthogonal to $\boldsymbol{U}^*$, we can obtain from (7)

$$\boldsymbol{Z}\bar{\boldsymbol{U}}^* = \mathbf{1}\bar{\boldsymbol{\mu}}^{*T} + \bar{\boldsymbol{E}}, \tag{10}$$

where $\bar{\boldsymbol{\mu}}^* = \bar{\boldsymbol{U}}^{*T}\boldsymbol{\mu}^*$, $\bar{\boldsymbol{E}} = \boldsymbol{E}\bar{\boldsymbol{U}}^*$. Eqn. (10) is in the typical location estimation setting except that $\bar{\boldsymbol{U}}^*$ is unknown, which motivates us to consider

$$\min_{(\bar{\boldsymbol{U}},\bar{\boldsymbol{\mu}})} \|\boldsymbol{Z}\bar{\boldsymbol{U}} - \mathbf{1}\bar{\boldsymbol{\mu}}^T\|_F^2 \quad \text{s.t.} \quad \bar{\boldsymbol{U}}^T\bar{\boldsymbol{U}} = \boldsymbol{I}_{\bar{r}\times\bar{r}}. \tag{11}$$

Interestingly, (11) can also be viewed as a *multivariate* extension, of rank $\bar{r}$, of the orthogonal regression due to Mizera (2002):

$$\min_{\mu\in\mathbb{R},\boldsymbol{u}\in\mathbb{R}^{p+1}} \|\,[\boldsymbol{X}\ \boldsymbol{y}]\,\boldsymbol{u} - \mathbf{1}\,\mu\,\|_2^2 \ \text{ s.t. } \|\boldsymbol{u}\|_2^2 = 1. \tag{12}$$

Moreover, when $\boldsymbol{Z} = [\boldsymbol{X}\ \boldsymbol{Y}]$, setting $\bar{\boldsymbol{U}} = [\boldsymbol{B}^T\ \boldsymbol{\Gamma}^T]^T$ gives a model $\boldsymbol{Y}\boldsymbol{\Gamma} + \boldsymbol{X}\boldsymbol{B} - \mathbf{1}\bar{\boldsymbol{\mu}}^T = \boldsymbol{E}$ for canonical correlation analysis.

How to introduce an operational depth for (11) is a meaningful problem. Indeed, with a deep $\bar{\boldsymbol{U}}$ provided, one would be able to rank high-dimensional samples in a lower dimensional subspace.

Restricting to a naive case of (12) with a single predictor and a single response:

$$\min_{(\alpha,\beta,\mu)} \|\beta\boldsymbol{x} + \alpha\boldsymbol{y} - \mathbf{1}\mu\|_F^2 \text{ s.t. } \alpha^2 + \beta^2 = 1, \tag{13}$$

one can *eliminate* the constraint by, say, $\alpha = -\sin t, \beta = \cos t$ with a free parameter t, and then take the Euclidean gradient with respect to (t, μ) to define a tangent depth (Mizera 2002). Nevertheless, the elimination method encounters difficulties when considering multiple predictors, let alone a general $\bar{r}$. As far as we know, there exists no commonly acknowledged multivariate orthogonal regression depth in the literature.

Our manifold approach provides a systematic treatment of (11) for all p, m, and $\bar{r}$. We call the resulting Riemannian depth the *orthogonal complement* (**OC**) depth. It pursues an $\bar{r}$-dimensional subspace in the original input space to rank the observations effectively. The influence space here is $\mathbb{R}^{\bar{r}} \times \mathcal{T}_{\bar{\boldsymbol{U}}}(\mathbb{O}^{m\times\bar{r}})$ with $\mathcal{T}_{\bar{\boldsymbol{U}}}(\mathbb{O}^{m\times\bar{r}}) = \{\boldsymbol{V} : \bar{\boldsymbol{U}}^T\boldsymbol{V} + \boldsymbol{V}^T\bar{\boldsymbol{U}} = \mathbf{0}\}$, and the OC depth for any given $(\bar{\boldsymbol{\mu}}^\circ, \bar{\boldsymbol{U}}^\circ) \in \mathbb{R}^{\bar{r}} \times \mathbb{O}^{m\times\bar{r}}$ is

$$\begin{aligned} \textbf{OC-depth}: \ \min_{(\boldsymbol{v},\boldsymbol{V})} \sum_i 1_{\ge 0}(\langle \boldsymbol{v}, \bar{\boldsymbol{\mu}}^\circ - \bar{\boldsymbol{U}}^{\circ T}\boldsymbol{z}_i\rangle + \langle \boldsymbol{V}, \boldsymbol{z}_i\boldsymbol{z}_i^T\bar{\boldsymbol{U}}^\circ - \boldsymbol{z}_i\bar{\boldsymbol{\mu}}^{\circ T}\rangle) \\ \text{s.t. } \boldsymbol{V}^T\bar{\boldsymbol{U}}^\circ + \bar{\boldsymbol{U}}^{\circ T}\boldsymbol{V} = \mathbf{0},\ \|\boldsymbol{v}\|_2^2 + \|\boldsymbol{V}\|_F^2 = 1. \end{aligned} \tag{14}$$

The derivations are similar to the PC depth and are omitted. Note that the influence space constraint has a multivariate form but is linear. Eqn. (14) also gives a multivariate orthogonal regression depth.

On the other hand, with $r = \bar{r}$ and no intercepts, the (order-1) PC-depth and OC-depth coincide, since in this case the two losses in (8) and (11) only differ by a minus sign

and the influence space is linear in $\boldsymbol{V}$. This means that the most and least informative subspaces will have the same depth. As aforementioned, an order-2 Riemannian depth would be able to distinguish between minimization and maximization problems, which deserves further investigation.

3. The Slack Variable Approach

The nondifferentiability issue in the other two examples in Section 1 is much trickier to cope with. In more detail, Example 2 has a closed parameter space $\mathbb{R}^p_+$ with boundaries, which makes gradient-based influences improper at any boundary point; Example 3 has a nonsmooth regularizer commonly seen in high dimensional statistics, and sometimes regularization can be imposed in a constrained manner.

Following Rousseeuw and Hubert (1999), the first step to define a data depth is to characterize a reasonable "fit", or a class of reasonable estimators, under a given model or method. It turns out that for such nonsmooth problems, we can derive local optimality conditions in form of *inequalities* or obtain some *nonlinear* fixed-point equations by use of a surrogate function, neither of which however results in sample-additive estimating equations directly. The good news is that we can then utilize some "slack variables" subject to proper (convex) inequality and equality constraints to offer a universal solution, which leads to a novel class of slacked data depth.

3.1. Slacked data depth and sparse learning

To begin with, let us consider $\min f(\boldsymbol{\beta}) \triangleq \sum_i l(\boldsymbol{\beta}; \boldsymbol{x}_i, y_i)$ s.t. $\boldsymbol{\beta} \succeq \mathbf{0}$ or $\boldsymbol{\beta} \in \mathbb{R}^p_+$, where l is differentiable in the augmented parameter space $\mathbb{R}^p$ but not necessarily convex. Because f is directionally differentiable in $\mathbb{R}^p_+$, any optimal solution $\hat{\boldsymbol{\beta}}$ must obey

$$\mathrm{D}_{\boldsymbol{u}} f(\hat{\boldsymbol{\beta}}) \geq 0 \text{ for all feasible } \boldsymbol{u}$$

where $\mathrm{D}_{\boldsymbol{u}} f(\boldsymbol{\beta})$ denotes the *one-sided* directional derivative of f at $\boldsymbol{\beta}$ with increment $\boldsymbol{u}$, namely, $\mathrm{D}_{\boldsymbol{u}} f(\boldsymbol{\beta}) = \lim_{\epsilon \to 0+} [f(\boldsymbol{\beta} + \epsilon \boldsymbol{u}) - f(\boldsymbol{\beta})]/\epsilon$. Nevertheless, unlike equalities that are maintained after projection (i.e., $\sum \boldsymbol{T}_i(\boldsymbol{B}) = \mathbf{0} \Rightarrow \sum \langle \boldsymbol{V}, \boldsymbol{T}_i(\boldsymbol{B}) \rangle = \langle \boldsymbol{V}, \sum \boldsymbol{T}_i(\boldsymbol{B}) \rangle = 0, \forall \boldsymbol{V} \in \mathcal{G}$), applying the same operation on inequalities may destroy their meanings totally during the process of Tukeyfication.

Our proposal is to associate each inequality with an additional slack variable, and append a nonnegative constraint when performing projection and error measurement. Let $\boldsymbol{e}_j$ be a vector with the jth component 1 and the remaining 0. In Example 2, taking $\boldsymbol{u} = \pm \boldsymbol{e}_j$ for $j \in \mathcal{J} = \{j : \beta_j \neq 0\}$ and $\boldsymbol{u} = \boldsymbol{e}_j$ for $j \in \mathcal{J}^c$ leads to the following slacked estimating equation:

$$\sum_i (\nabla l(\boldsymbol{\beta}; \boldsymbol{x}_i, y_i) - \boldsymbol{s}/n) = \mathbf{0},$$

where $\boldsymbol{s}_{\mathcal{J}^c} \succeq \mathbf{0}$ and $\boldsymbol{s}_{\mathcal{J}} = \mathbf{0}$. The ordinary Tukeyfication now goes through, and we obtain a depth optimization problem for any $\boldsymbol{\beta}^\circ \succeq \mathbf{0}$:

$$\min_{(\boldsymbol{v}, \boldsymbol{s}) \in \mathbb{R}^p \times \mathbb{R}^p} \sum_i 1_{\geq 0}(\langle \boldsymbol{v}, \nabla l(\boldsymbol{\beta}^\circ; \boldsymbol{x}_i, y_i) - \boldsymbol{s}/n \rangle) \text{ s.t. } \|\boldsymbol{v}\|_2 = 1,\ \boldsymbol{s} \circ \boldsymbol{\beta}^\circ = \mathbf{0},\ \boldsymbol{s} \succeq \mathbf{0}.$$

When $l(\boldsymbol{\beta}; \boldsymbol{x}_i, y_i) = (\boldsymbol{x}_i^T\boldsymbol{\beta} - y_i)^2/2$, we get the nonnegative regression depth

$$\min_{(\boldsymbol{v},\boldsymbol{s})\in\mathbb{R}^p\times\mathbb{R}^p} \sum_i 1_{\geq 0}(\langle \boldsymbol{v}, \boldsymbol{x}_i(\boldsymbol{x}_i^T\boldsymbol{\beta}^\circ - y_i) - \boldsymbol{s}/n\rangle) \text{ s.t. } \|\boldsymbol{v}\|_2 = 1, \boldsymbol{s}\circ\boldsymbol{\beta}^\circ = \mathbf{0}, \boldsymbol{s}\succeq \mathbf{0}. \tag{15}$$

Recall that $\circ$ denotes the elementwise product. When $\boldsymbol{\beta}^\circ \succ \mathbf{0}$, $\boldsymbol{s} = \mathbf{0}$, and (15) becomes the regression depth. In general, the inclusion of $\boldsymbol{s}$ in the minimization, as an outcome of the nonnegativity restriction, often results in a lower depth value.

The slack-variable technique can introduce useful depth notions for sparse learning that is at the core of high dimensional statistics:

$$\min_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) \triangleq \bar{l}(\boldsymbol{X}\boldsymbol{\beta}; \boldsymbol{y}) + \sum_{j=1}^{p} P(|\beta_j|; \lambda), \tag{16}$$

where $\bar{l}(\boldsymbol{X}\boldsymbol{\beta}; \boldsymbol{y}) = \sum_i l_0(\boldsymbol{x}_i^T\boldsymbol{\beta}; \boldsymbol{y}_i)$ with l_0 differentiable. Here, we assume that P is sparsity-promoting in the sense that it is induced by a *thresholding rule* $\Theta(\cdot;\lambda)$ with λ as the threshold (see She (2012) for the rigorous definition and more details): $P(t;\lambda) = P_\Theta(t;\lambda) + q(t;\lambda)$, where

$$P_\Theta(t;\lambda) = \int_0^{|t|} (\Theta^{-1}(u;\lambda) - u)\, \mathrm{d}u \text{ with } \Theta^{-1}(u;\lambda) = \sup\{t : \Theta(t;\lambda) \leq u\}$$

and q is an arbitrary nonnegative function satisfying $q(t;\lambda) = 0$ if $t = \Theta(s;\lambda)$ for some $s \in \mathbb{R}$. Hence if $\Theta(\cdot;\lambda)$ is a continuous function, q must be identical to zero, but if Θ has discontinuities, the mapping from P to Θ is *many-to-one*. The universal Θ-P framework covers many practically used penalties such as ℓ_r $(0 \leq r \leq 1)$, SCAD, MCP, which can be nonconvex. For centered response and predictors, (16) suffices; when centering the response is inappropriate, an intercept α subject to no regularization should often be added in the systematic component. For clarity, we assume $\alpha = 0$ in the following derivation, but the extension to $\boldsymbol{X}\boldsymbol{\beta} + \alpha\mathbf{1}$ is straightforward.

For penalties with $q \equiv 0$ (continuous Θ), like ℓ_1 and SCAD, we can use the directional derivatives along $\pm\boldsymbol{e}_j$ to show that any locally optimal $\hat{\boldsymbol{\beta}}$ satisfies the thresholding equation (She 2016)

$$\boldsymbol{\beta} = \Theta(\boldsymbol{\beta} - \boldsymbol{X}^T\nabla\bar{l}(\boldsymbol{X}\boldsymbol{\beta}); \lambda), \tag{17}$$

under the mild assumption that $\Theta(\cdot;\lambda)$ is continuous at $\hat{\boldsymbol{\beta}} - \boldsymbol{X}^T\nabla\bar{l}(\boldsymbol{X}\hat{\boldsymbol{\beta}})$. But nontrivial q's and discontinuous Θ's constitute an important class of nonsmooth penalties, including, in particular, the discontinuous ℓ_0 penalty

$$\frac{\lambda^2}{2}\|\boldsymbol{\beta}\|_0,$$

for which Θ is the hard-thresholding $\Theta_H(t;\lambda) = t1_{|t|>\lambda}$, and $q(t;\lambda) = (1/2)(\lambda - |t|)^2 1_{0<|t|<\lambda}$. In such scenarios, if $\nabla\bar{l}$ is L-Lipschitz continuous, the solutions can be characterized by the *fixed points* of an iterative optimization algorithm based on a surrogate function g:

$$\beta \in \arg\min g(\cdot, \boldsymbol{\beta}^-)|_{\boldsymbol{\beta}^- = \boldsymbol{\beta}}$$

where $g(\boldsymbol{\beta}, \boldsymbol{\beta}^-) = \bar{l}(\boldsymbol{X}\boldsymbol{\beta}^-) + \langle \nabla \bar{l}(\boldsymbol{X}\boldsymbol{\beta}^-), \boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^-)\rangle + \sum P(|\beta_j|; \lambda) + \rho\|\boldsymbol{\beta} - \boldsymbol{\beta}^-\|_2^2/2$ with $\rho \geq L\|\boldsymbol{X}\|_2^2$, and we can prove that they all enjoy nearly minimax error rate under a proper choice of λ and some regularity conditions (She et al. 2021b). It can be verified that the fixed-point estimators satisfy (17) as well, under the same continuity assumption, as long as $\boldsymbol{X}$ has been properly scaled: $\|\boldsymbol{X}\|_2 \leq 1/\sqrt{L}$. Unfortunately, the thresholding equation does not belong to the estimating equation framework examined in our companion paper—specifically, the nonlinear thresholding effect desired in sparse learning means that (17) is not sample additive.

Below we introduce p additional slack variables to find a proper substitute for (17) so that one can define data depth for (16) given an *arbitrary* thresholding Θ. Let $\boldsymbol{\beta}$ be a locally optimal solution to the problem as $P = P_\Theta$, or a fixed-point solution as $P = P_\Theta + q$. Define $\mathcal{J} = \{j : \beta_j \neq 0\}$ and $\mathcal{J}^c = \{j : \beta_j = 0\}$, and denote by $\boldsymbol{X}[, j]$ the jth column vector of $\boldsymbol{X}$. Using the directional derivatives of f when $P = P_\Theta$ (see, e.g., the proof of Theorem 1 in She (2016)), or the directional derivatives of $g - q$ under the continuity assumption when $P = P_\Theta + q$, we get

$$\Theta^{-1}(|\beta_j|; \lambda)\,\mathrm{sgn}(\beta_j) = \beta_j - \boldsymbol{X}[, j]^T \nabla \bar{l}(\boldsymbol{X}\boldsymbol{\beta}),\ \forall j \in \mathcal{J}, \tag{18}$$

which holds even if Θ is not strictly increasing in a neighborhood of $|\beta_j|$ ($j \in \mathcal{J}$), while for $j \in \mathcal{J}^c$, $\beta_j = 0$, and so

$$-\lambda \leq \boldsymbol{X}[, j]^T \nabla \bar{l}(\boldsymbol{X}\boldsymbol{\beta}) \leq \lambda, \forall j \in \mathcal{J}^c. \tag{19}$$

Next, define $\boldsymbol{\gamma}(\boldsymbol{\beta}) = [\gamma_j]$ with

$$\gamma_j = \begin{cases} \Theta^{-1}(|\beta_j|; \lambda)\,\mathrm{sgn}(\beta_j) - \beta_j & \text{if } j \in \mathcal{J} \\ 0 & \text{if } j \in \mathcal{J}^c. \end{cases} \tag{20}$$

It follows from (18), (19) that $\boldsymbol{X}^T \nabla \bar{l}(\boldsymbol{X}\boldsymbol{\beta}) + \boldsymbol{\gamma}(\boldsymbol{\beta}) + \boldsymbol{s} = \boldsymbol{0}$ for some $\boldsymbol{s} \in \mathbb{R}^p$, $\boldsymbol{s}_\mathcal{J} = \boldsymbol{0}$, and $|s_j| \leq \lambda$, $j \in \mathcal{J}^c$.

Now, given a penalty induced by a thresholding rule $\Theta(\cdot; \lambda)$ and a point of interest $\boldsymbol{\beta}^\circ \in \mathbb{R}^p$, the slacked data depth resulting from (16), which we call "Θ-depth", can be cast as a **joint** optimization problem with respect to direction $\boldsymbol{v} \in \mathbb{R}^p$ and slack variables $\boldsymbol{s} = [s_j] \in \mathbb{R}^p$:

$$\begin{aligned} \textbf{Θ-depth:}\quad d_{01}^\Theta(\boldsymbol{\beta}^\circ) = \min_{(\boldsymbol{v}, \boldsymbol{s})} \sum_i 1_{\geq 0}(\langle \boldsymbol{v}, \boldsymbol{x}_i l_0'(\boldsymbol{x}_i^T \boldsymbol{\beta}^\circ; y_i) + (\boldsymbol{\gamma}^\circ + \boldsymbol{s})/n\rangle) \\ \text{s.t. } \|\boldsymbol{v}\|_2 = 1,\ \boldsymbol{s} \circ \boldsymbol{\beta}^\circ = \boldsymbol{0},\ \|\boldsymbol{s}\|_\infty \leq \lambda, \end{aligned} \tag{21}$$

where $\boldsymbol{\gamma}^\circ = \boldsymbol{\gamma}(\boldsymbol{\beta}^\circ)$ and $\|\boldsymbol{s}\|_\infty = \max|s_j|$. When Θ is the hard thresholding Θ_H (corresponding to the class of ℓ_0 penalties), $\boldsymbol{\gamma}^\circ = \boldsymbol{0}$. The user should specify a reasonably small λ (otherwise extremely low depth values are to be expected): a theoretical choice in sparse regression is $\lambda = \sigma\sqrt{cn \log p}$ (with say $c = 2$) where σ is the Orlicz ψ_2-norm of the noise, and a less conservative one can often be obtained via cross-validation. A fascinating fact is that the slack variable approach requires no convexity of either the loss or the penalty.

An important alternative to penalized sparse learning is to directly limit the sparsity level: $\|\boldsymbol{\beta}\|_0 \leq q$, instead of specifying a penalty parameter λ. Due to the lack of nonsmoothness of

$$\min_{\|\boldsymbol{\beta}\|_0 \leq q} \bar{l}(\boldsymbol{X}\boldsymbol{\beta}; \boldsymbol{y}), \tag{22}$$

we take the surrogate route. Statistically accurate estimates can be obtained from the resulting iterative quantile-thresholding algorithm (She et al. 2022b), which all satisfy $\boldsymbol{\beta} = \Theta^{\#}(\boldsymbol{\beta} - (1/\rho)\boldsymbol{X}^T\nabla\bar{l}(\boldsymbol{X}\boldsymbol{\beta}); q)$, assuming no ties occur and ρ is large enough (e.g., $L\|\boldsymbol{X}\|_2^2$). Here, the quantile thresholding $\Theta^{\#}(\boldsymbol{\alpha}; q)$ for any $\boldsymbol{\alpha} \in \mathbb{R}^p$ is a vector $\boldsymbol{\zeta}$ with $\zeta_{(j)} = \alpha_{(j)}$ if $1 \leq j \leq q$, and 0 otherwise, where $\alpha_{(1)}, \ldots, \alpha_{(p)}$ are the order statistics of $\alpha_1, \ldots, \alpha_p$ satisfying $|\alpha_{(1)}| \geq \cdots \geq |\alpha_{(p)}|$. $\Theta^{\#}(\boldsymbol{\alpha}; q)$ can be viewed as a variant of $\Theta_H(\boldsymbol{\alpha}; \lambda)$ (by setting $\lambda = |\alpha_{(q+1)}|$, say), but it uses an adaptive threshold. Again, we suppose that the regularization parameter q is already given.

By use of slack variables to rewrite the $\Theta^{\#}$-equation (details omitted), we can define the q-sparse constrained ℓ_0-depth (which we call "$\Theta^{\#}$-depth") for any $\boldsymbol{\beta}^\circ : \|\boldsymbol{\beta}^\circ\|_0 = q$ as

$$\begin{aligned} \Theta^{\#}\textbf{-depth:} \quad & \min_{(\boldsymbol{v},\boldsymbol{s}) \in \mathbb{R}^p \times \mathbb{R}^p} \sum_i 1_{\geq 0}(\langle \boldsymbol{v}, \boldsymbol{x}_i l_0'(\boldsymbol{x}_i^T\boldsymbol{\beta}^\circ; y_i) + \boldsymbol{s}/n \rangle) \\ & \text{s.t. } \|\boldsymbol{v}\|_2 = 1, \ \boldsymbol{s} \circ \boldsymbol{\beta}^\circ = \boldsymbol{0}, \ \|\boldsymbol{s}\|_\infty \leq \|\boldsymbol{X}^T_{(\mathcal{J}^\circ)^c}\nabla\bar{l}(\boldsymbol{X}\boldsymbol{\beta}^\circ)\|_\infty, \end{aligned} \tag{23}$$

where $\mathcal{J}^\circ = \{j : \beta_j^\circ = 0\}$, $\boldsymbol{X}_{(\mathcal{J}^\circ)^c}$ is a submatrix of $\boldsymbol{X}$ by selecting the columns corresponding to the complement of $\mathcal{J}^\circ$, and both $\boldsymbol{\gamma}^\circ$ and ρ (as long as $\rho > 0$) disappear in the ℓ_0-constrained depth, just like in the ℓ_0-penalized case. Of all the constraints on $\boldsymbol{s}$, the equality ones are affine, and the inequality ones are convex. The deepest q-sparse estimate is defined as the saddle point that maximizes (23) over all $\boldsymbol{\beta}^\circ : \|\boldsymbol{\beta}^\circ\|_0 = q$ (cf. (35) in She et al. (2022a)).

Clearly, in the special case of $q = p$, all slack variables are removed, but as $q < p$, the constrained problem (22) results in more stringent estimating equations that are easier to violate, compared with the plain (non-regularized) problem. This is reflected by the inclusion of $\boldsymbol{s}$ during the minimization, thereby lower depth values. The same conclusion holds for the Θ-depth (21) due to the existence of additional slack variables. On the other hand, sparsity depths may be very **low** for large p and λ in (21) or large p and small q in (23). To alleviate the issue, it is beneficial to change the crude "0-1 loss" to some more elegant φ, as discussed in our companion paper (She et al. 2022a). For example, (23) could be replaced by $\min_{(\boldsymbol{v},\boldsymbol{s})} \sum_i \varphi(\langle \boldsymbol{v}, \boldsymbol{x}_i l_0'(\boldsymbol{x}_i^T\boldsymbol{\beta}^\circ; y_i)/\rho + \boldsymbol{s}/n \rangle)$ s.t. $\|\boldsymbol{v}\|_2 = 1$, $\boldsymbol{s} \circ \boldsymbol{\beta}^\circ = \boldsymbol{0}$, $\|\boldsymbol{s}\|_\infty \leq \|\boldsymbol{X}^T_{(\mathcal{J}^\circ)^c}\nabla\bar{l}(\boldsymbol{X}\boldsymbol{\beta}^\circ)/\rho\|_\infty$, with $\varphi(\cdot)$ nonzero for mild or moderate negative inputs, which warrants further investigation in the future. A tight upper bound of sparsity depths is also worth studying in theory.

Slacked data depth can be introduced for groupwise variable selection and low-rank matrix estimation (She 2012, 2013) as well; see, e.g., Section 3.2.

Remark 2 (Computation of slacked depth)**.** *A simple alternating optimization or* block coordinate descent *(BCD) algorithm can be used to to compute slacked data depth. Take the φ-form of (21) as an example. Given $\boldsymbol{s}$, the optimization problem for $\boldsymbol{v}$,*

$$\min_{\boldsymbol{v} \in \mathbb{R}^p, \|\boldsymbol{v}\|_2 = 1} \sum_i \varphi(\langle \boldsymbol{v}, r_i\boldsymbol{x}_i + (\boldsymbol{\gamma}^\circ + \boldsymbol{s})/n \rangle)$$

where $\boldsymbol{r} = \nabla_{\boldsymbol{\Theta}} \bar{l}|_{\boldsymbol{\Theta}=\boldsymbol{X}\beta^\circ}$*, has been investigated in She et al. (2022a). Fixing* $\boldsymbol{v}$*, we can rewrite the* $\boldsymbol{s}$*-problem as*

$$\min_{\boldsymbol{s}\in\mathbb{R}^p} \sum_i \varphi(\langle \boldsymbol{v}, r_i \boldsymbol{x}_i + \boldsymbol{\gamma}^\circ/n + \boldsymbol{s}/n \rangle) \quad s.t. \quad \boldsymbol{s}_{\mathcal{J}} = \boldsymbol{0}, \|\boldsymbol{s}\|_\infty \le \lambda.$$

The problem has a differentiable criterion in $\boldsymbol{s}$ *and some simple box constraints, and conventional numerical methods apply, including L-BFGS-B, interior point, and proximal gradient descent algorithms (Byrd et al. 1995; Boyd and Vandenberghe 2004; Parikh and Boyd 2014).*

3.2. Reduced-rank regression depth

Applying ordinary least squares on multiple responses may easily result in a large number of unknowns. Researchers often prefer adding a low-rank constraint in estimating the coefficient matrix, leading to the celebrated reduced-rank regression (RRR) (Anderson 1951)

$$\min_{\boldsymbol{B}\in\mathbb{R}^{p\times m}} f(\boldsymbol{B}; \boldsymbol{X}, \boldsymbol{Y}) \triangleq \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_F^2 \text{ s.t. } \operatorname{rank}(\boldsymbol{B}) \le r, \tag{24}$$

where $\boldsymbol{Y} = [\boldsymbol{y}_1 \dots \boldsymbol{y}_n]^T \in \mathbb{R}^{n\times m}$ and $\boldsymbol{X} = [\boldsymbol{x}_1 \dots \boldsymbol{x}_n]^T \in \mathbb{R}^{n\times p}$ are the (centered) response and predictor matrices. A weighted criterion to account for the dependency between the responses can be given, but the problem can be converted to (24) with a simple reparametrization. If the variables are not centered, an intercept term $\boldsymbol{1}\boldsymbol{\alpha}^T$ should be added in the loss, but the depth derivation below carries over (cf. Section 2.2). We assume that $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ are i.i.d. (or in an approximate sense), and so the data depth in this subsection does not apply to PCA where $\boldsymbol{X} = \boldsymbol{I}$, thus distinct from the PC-depth and OC-depth introduced earlier; see some related discussions in Section 2.3 of our companion paper.

RRR provides a low-dimensional projection space to view and analyze supervised multivariate data, and finds widespread applications in machine learning and econometrics (Reinsel and Velu 1998; Izenman 2008). In fact, once an estimate $\boldsymbol{B}$ of rank r is obtained, we can write $\boldsymbol{B} = \boldsymbol{B}_1\boldsymbol{B}_2^T$ for $\boldsymbol{B}_1 \in \mathbb{R}^{p\times r}$, $\boldsymbol{B}_2 \in \mathbb{R}^{m\times r}$. This suggests that r factors can be constructed by $\boldsymbol{X}\boldsymbol{B}_1$ from p predictors to explain all response variables. The number of factors required in real applications is often much smaller than the number of input x-variables.

Limiting the rank of the matrix estimators at r, how to perform a "center-outward" ranking in high dimensions, or more generally, test

$$H_0 : \boldsymbol{B} \in \Omega_0 \cap \{\operatorname{rank}(\boldsymbol{B}) = r\} \text{ vs. } H_a : \boldsymbol{B} \in \Omega_0^c \cap \{\operatorname{rank}(\boldsymbol{B}) = r\},$$

where the set or event Ω_0 is not necessarily a singleton (cf. Remark 1 in She et al. (2022a)), is an intriguing open question.

In the following, we extend multivariate regression depth (Rousseeuw and Hubert 1999; Bern and Eppstein 2002) to the *reduced-rank regression depth* (27), using the techniques developed in the last subsection. Toward this, we first give a fixed-point formulation of

all RRR estimators. Define a matrix version of the $\Theta^{\#}$ introduced in the last subsection

$$\Theta^{\sigma\#}(\boldsymbol{B};r) \triangleq \boldsymbol{U}\mathrm{diag}\{\Theta^{\#}([\sigma_i(\boldsymbol{B})];r)\}\boldsymbol{V}^T, \quad \forall \boldsymbol{B} \in \mathbb{R}^{p\times m} \tag{25}$$

where $\boldsymbol{U}$, $\boldsymbol{V}$, and $\mathrm{diag}\{\sigma(\boldsymbol{B})_i\}$ are from the SVD of $\boldsymbol{B} = \boldsymbol{U}\mathrm{diag}\{\sigma_i(\boldsymbol{B})\}\boldsymbol{V}$, and $\Theta^{\#}$ is applied to the vector $[\sigma_i(\boldsymbol{B})]$, with $\sigma_i(\boldsymbol{B})$ denoting the ith largest singular value of $\boldsymbol{B}$. Construct a surrogate function

$$g(\boldsymbol{B},\boldsymbol{B}^-) = f(\boldsymbol{B}^-) + \langle \nabla f(\boldsymbol{B}^-), \boldsymbol{B}-\boldsymbol{B}^-\rangle + \rho\|\boldsymbol{B}-\boldsymbol{B}^-\|_F^2/2,$$

where ρ is larger than $\|\boldsymbol{X}\|_2^2$. Let $\hat{\boldsymbol{B}}_{\mathrm{rrr}}$ be an RRR estimator that solves (24). Then for $\tilde{\boldsymbol{B}} \in \arg\min_{\boldsymbol{B}:\mathrm{rank}(\boldsymbol{B})\le r} g(\boldsymbol{B},\hat{\boldsymbol{B}}_{\mathrm{rrr}})$, it follows from the chain inequalities $f(\hat{\boldsymbol{B}}_{\mathrm{rrr}})-f(\tilde{\boldsymbol{B}}) = g(\hat{\boldsymbol{B}}_{\mathrm{rrr}},\hat{\boldsymbol{B}}_{\mathrm{rrr}}) - f(\tilde{\boldsymbol{B}}) \ge g(\tilde{\boldsymbol{B}},\hat{\boldsymbol{B}}_{\mathrm{rrr}}) - f(\tilde{\boldsymbol{B}}) \ge (\rho - \|\boldsymbol{X}\|_2^2)\|\tilde{\boldsymbol{B}}-\hat{\boldsymbol{B}}_{\mathrm{rrr}}\|_F^2/2$ that $\hat{\boldsymbol{B}}_{\mathrm{rrr}} = \tilde{\boldsymbol{B}}$. On the other hand, it is easy to show that $\tilde{\boldsymbol{B}} = \Theta^{\sigma\#}(\hat{\boldsymbol{B}}_{\mathrm{rrr}} - \frac{1}{\rho}\boldsymbol{X}^T(\boldsymbol{X}\hat{\boldsymbol{B}}_{\mathrm{rrr}}-\boldsymbol{Y});r)$ (She 2013), and so $\hat{\boldsymbol{B}}_{\mathrm{rrr}}$ satisfies the matrix thresholding equation

$$\boldsymbol{B} = \Theta^{\sigma\#}(\boldsymbol{B} - \frac{1}{\rho}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{B}-\boldsymbol{Y});r). \tag{26}$$

(In fact, under the mild condition that $\boldsymbol{Y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^+\boldsymbol{X}^T\boldsymbol{Y}$ has distinct eigenvalues, the RRR estimator is unique (Reinsel and Velu 1998), and ρ can be way smaller than $\|\boldsymbol{X}\|_2^2$.) Perform a compact SVD of $\boldsymbol{B}$: $\boldsymbol{B} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{Q}^T$ with $\boldsymbol{P} \in \mathbb{O}^{p\times r}$ and $\boldsymbol{Q} \in \mathbb{O}^{m\times r}$, and denote by $\boldsymbol{P}_\perp \in \mathbb{O}^{p\times(p-r)}$ and $\boldsymbol{Q}_\perp \in \mathbb{O}^{m\times(m-r)}$ their orthogonal complements (which can be obtained from the full SVD of $\boldsymbol{B}$). Like in the $\Theta^{\#}$-case, based on (26), we work on $(1/\rho)\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{B}-\boldsymbol{Y}) + \boldsymbol{S} = \boldsymbol{0}$ for a slack matrix $\boldsymbol{S}$ satisfying

$$\boldsymbol{P}^T\boldsymbol{S} = \boldsymbol{0}, \boldsymbol{S}\boldsymbol{Q} = \boldsymbol{0}, \|\boldsymbol{S}\|_2 \le \|\frac{1}{\rho}\boldsymbol{P}_\perp\boldsymbol{P}_\perp^T\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{B}-\boldsymbol{Y})\boldsymbol{Q}_\perp\boldsymbol{Q}_\perp^T\|_2.$$

Now, given a regularization parameter $r : 1 \le r \le p \wedge m$ and a matrix of interest $\boldsymbol{B}^\circ \in \mathbb{R}^{p\times m} : \mathrm{rank}(\boldsymbol{B}^\circ) = r$, obtain the associated $\boldsymbol{P}_\perp^\circ \in \mathbb{O}^{p\times(p-r)}$, $\boldsymbol{Q}_\perp^\circ \in \mathbb{O}^{m\times(m-r)}$ as above; the rank-r RRR depth of $\boldsymbol{B}^\circ$ is defined by

$$\begin{aligned} d_{01}^{\mathrm{RRR}}(\boldsymbol{B}^\circ) = \min_{(\boldsymbol{V},\boldsymbol{S})} \sum_i 1_{\ge 0}(\langle \boldsymbol{V}, \frac{1}{\rho}\boldsymbol{x}_i(\boldsymbol{x}_i^T\boldsymbol{B}^\circ - \boldsymbol{y}_i^T) + \frac{1}{n}\boldsymbol{S}\rangle) \\ \text{s.t. } \|\boldsymbol{V}\|_F = 1, \boldsymbol{P}^{\circ T}\boldsymbol{S} = \boldsymbol{0}, \boldsymbol{S}\boldsymbol{Q}^\circ = \boldsymbol{0}, \|\boldsymbol{S}\|_2 \le \|\frac{1}{\rho}\boldsymbol{P}_\perp^{\circ T}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{B}^\circ - \boldsymbol{Y})\boldsymbol{Q}_\perp^\circ\|_2, \end{aligned}$$

or equivalently,

$$\begin{aligned} \textbf{RRR-depth: } d_{01}^{\mathrm{RRR}}(\boldsymbol{B}^\circ) = \min_{(\boldsymbol{V},\boldsymbol{L})} \sum_i 1_{\ge 0}(\langle \boldsymbol{V}, \boldsymbol{x}_i(\boldsymbol{x}_i^T\boldsymbol{B}^\circ - \boldsymbol{y}_i^T) + \frac{1}{n}\boldsymbol{P}_\perp^\circ\boldsymbol{L}\boldsymbol{Q}_\perp^{\circ T}\rangle) \\ \text{s.t. } \|\boldsymbol{V}\|_F = 1, \|\boldsymbol{L}\|_2 \le \|\boldsymbol{P}_\perp^{\circ T}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{B}^\circ - \boldsymbol{Y})\boldsymbol{Q}_\perp^\circ\|_2, \end{aligned} \tag{27}$$

where ρ vanishes due to the scale invariance of $1_{\ge 0}$, regardless of how small or large ρ is. Clearly, in the full rank case $r = m \wedge p$, either $\boldsymbol{P}_\perp^\circ$ or $\boldsymbol{Q}_\perp^\circ$ must vanish, and so $\boldsymbol{L} = \boldsymbol{0}$, meaning that (27) reduces to the multivariate regression depth (Bern and Eppstein 2002).

Remark 3 (Combined treatment)**.** *The manifold approach and slack variable approach can be combined together to define data depth for some challenging problems. Consider a* sparse *RRR (one of the variants in She (2017)) that constructs r predictive factors from a subset of predictors*

$$\min_{\boldsymbol{S}\in\mathbb{R}^{p\times r},\boldsymbol{U}\in\mathbb{R}^{m\times r}} \frac{1}{2}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{A}\boldsymbol{U}^T\|_F^2 \ s.t.\ \|\,vec\,(\boldsymbol{A})\|_0 \le q,\ \boldsymbol{U}^T\boldsymbol{U}=\boldsymbol{I}_{r\times r}. \tag{28}$$

The overall coefficient matrix $\boldsymbol{B}=\boldsymbol{A}\boldsymbol{U}^T$ *has rank at most r as in RRR, but sparsity is imposed on the loading matrix* $\boldsymbol{A}$*. By use of a slack matrix* $\boldsymbol{S}$ *for* $\boldsymbol{A}$*, and a Riemannian tangent space for* $\boldsymbol{U}$*, the depth for* $(\boldsymbol{A}^\circ,\boldsymbol{U}^\circ)$*:* $\boldsymbol{A}^\circ\in\mathbb{R}^{p\times r}$, $\|\,vec\,(\boldsymbol{A}^\circ)\|_0=q$, $\boldsymbol{U}^\circ\in\mathbb{O}^{m\times r}$ *is given by* $\min_{\boldsymbol{W}\in\mathbb{R}^{m\times r},\boldsymbol{V}\in\mathbb{R}^{p\times r},\boldsymbol{S}\in\mathbb{R}^{p\times r}}\sum_{i=1}^n 1_{\ge 0}(-\langle\boldsymbol{W},\boldsymbol{y}_i\boldsymbol{x}_i^T\boldsymbol{A}^\circ\rangle+\langle\boldsymbol{V},\boldsymbol{x}_i(\boldsymbol{x}_i^T\boldsymbol{A}^\circ-\boldsymbol{y}_i^T\boldsymbol{U}^\circ)+\boldsymbol{S}/n\rangle)$ *s.t.* $\|\boldsymbol{W}\|_F^2+\|\boldsymbol{V}\|_F^2=1, \boldsymbol{V}^T\boldsymbol{U}^\circ+\boldsymbol{U}^{\circ T}\boldsymbol{V}=\boldsymbol{0}, vec\,(\boldsymbol{A}^\circ)\circ vec\,(\boldsymbol{S})=\boldsymbol{0}, \|\boldsymbol{S}\|_{\max}\le\lambda^\circ$*, with* $\lambda^\circ=\|\,vec\,(\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{A}^\circ-\boldsymbol{Y}\boldsymbol{U}^\circ))[(\mathcal{J}^\circ)^c]\|_\infty=\|(\boldsymbol{I}\otimes\boldsymbol{X}^T)[(\mathcal{J}^\circ)^c,]\,vec\,(\boldsymbol{X}\boldsymbol{A}^\circ-\boldsymbol{Y}\boldsymbol{U}^\circ)\|_\infty$ *and* $\mathcal{J}^\circ=\{j: vec\,(\boldsymbol{A}^\circ)[j]\ne 0, 1\le j\le pr\}$.

4. Experiments

This section performs real data experiments to illustrate the usefulness of some new notions of depth.

4.1. Reduced-rank depth in time series

We consider the 52 weekly stock log-return data for nine of the ten largest American corporations in 2004 (Rothman et al. 2010), with $\boldsymbol{y}_t\in\mathbb{R}^9$ $(t=1,\ldots,T)$ and $T=52$.

For the purpose of constructing market factors that drive general stock movements, a reduced-rank vector autoregressive (VAR) model can be used, i.e., $\boldsymbol{y}_{t+1}=\boldsymbol{B}^{*T}\boldsymbol{y}_t+\boldsymbol{e}_t$, with $\boldsymbol{B}^*$ of low rank. By conditioning on the initial state $\boldsymbol{y}_0$ and assuming the normality of $\boldsymbol{e}_t$, the conditional likelihood function leads to a least squares criterion, so the estimation of $\boldsymbol{B}^*$ can be formulated as a reduced-rank regression problem; see Lütkepohl (2007) for more details. We fit the reduced-rank VAR with $r=6$. The optimization algorithm for (27) (implemented based on Remark 2) however shows that the objective function can reach zero for some feasible $(\boldsymbol{V},\boldsymbol{L})$. Hence, although the standard RRR approach is widely used in multivariate times series and econometrics, our analysis revealed a perhaps surprisingly low data depth on this financial series dataset.

We then considered the Cauchy-based reduced rank regression (Zhao and Palomar 2017; Yang and Zhao 2020) (denoted by C-RRR) and robust reduced rank regression (She and Chen 2017) (denoted by R4, with 5% of data treated as outliers), as well as a deeper estimate obtained by random sampling (denoted by D-RRR). The rank-6 depth values of these estimates are 0.02, 0.08 and 0.12, respectively, suggesting more reliable fitted models than the plain RRR from the perspective of data depth.

To further illustrate the differences between the estimates, we plot the fitted models of Ford $(x_{4,t})$ in response to Walmart $(y_{1,t+1})$ in Figure 1. Notably, the right-most point has high leverage, and the RRR model passes close to that particular observation. In contrast, D-RRR seems to fit better the majority of the sample.

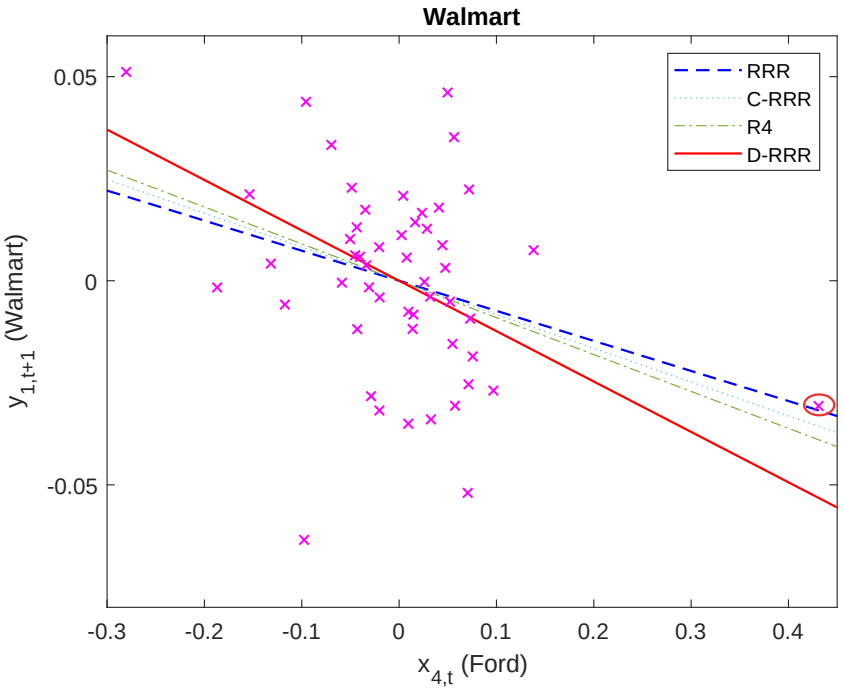


Figure 1: The fitted models $y_{1,t+1} \sim x_{4,t}$ using different methods in the low-rank VAR(1), to demonstrate how the log-return of WALMART is related to that of FORD in the previous week. Notice the right-most point that has a high leverage.

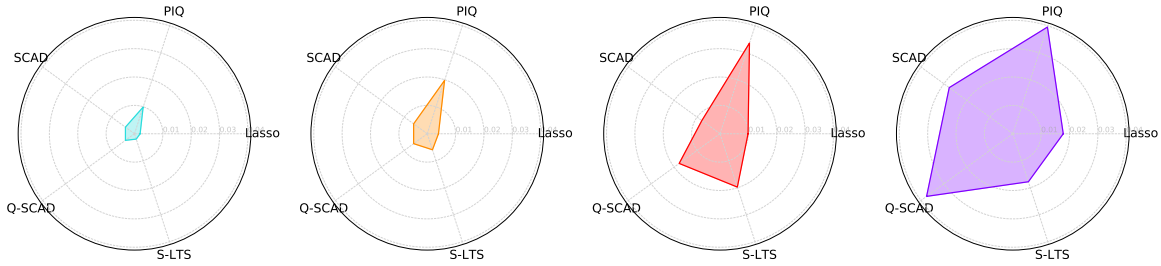
A careful examination of the series shows the point corresponds to the log-return of FORD at week 17, a real major market disturbance attributed to the auto industry. Several other stock returns experienced dramatic short-term changes as well, and we occasionally observe that the slopes obtained from RRR and its robust counterparts can have opposite signs. Financial time series often contain anomalies or demonstrate heavier tails than those of a normal distribution due to extreme market movements. The issue may jeopardize the recovery of common market behaviors and asset return forecasting: the autoregressive structure can make any outlier in the time series also a leverage point in the covariates. Although an elaborate robustification of the low-rank VAR merits further investigation, our depth-based analysis seems to offer an effective fix in this regard.

4.2. Sparsity depth for performance evaluation

Data depth provides a nonparametric means of performance evaluation. In this experiment, we use the sparsity depth defined in (23) to conduct a comparison between some commonly used sparse learning methods on the Boston housing dataset (Harrison and Rubinfeld 1978). The dataset was collected by the U.S. Census Service and consists of 13 predictors regarding socioeconomic and environmental conditions for 506 neighborhoods in the Boston area. The response is the median value of owner-occupied homes in the area.

We compare Lasso (Tibshirani 1996), SCAD (Fan and Li 2001), sparse LTS (S-LTS)

(Alfons et al. 2013), quantile-SCAD (Q-SCAD) (Sherwood and Maidman 2019) and PIQ (She et al. 2022b), in terms of data centrality defined in (23) for the same given support size q. More concretely, assuming that the observations are i.i.d., we split the dataset in halves, fit the methods on the first half, and then evaluate their performance via sparsity depth on the rest half. The whole procedure is repeated 20 times.



(a) $q = 7$ (b) $q = 8$ (c) $q = 9$ (d) $q = 10$

Figure 2: Sparsity depth comparison between Lasso, SCAD, sparse LTS (S-LTS), quantile SCAD (Q-SCAD) and PIQ estimates with respect to the support size q.

Figure 2 shows a series of radar plots for the median $\Theta^{\#}$-depths in respect to the number of selected variables. The depth values are small but present useful ranking information. The estimates are quite different seen from the data depth comparison: Lasso and SCAD exhibit lower depth in most cases, Q-SCAD and PIQ often give deeper estimates, and S-LTS is unstable (and costly) in our experiments. The last three methods all use a more robust loss, as well as a nonconvex regularizer, while Lasso solves a convex optimization problem with the ordinary ℓ_2-loss and ℓ_1-penalty. The depth differences between these sparse learning methods indicate that the data must deviate from Gaussianity and may contain anomalies, and incorporating the desired type of regularization into data depth can provide a helpful tool for robust performance evaluation.

5. Summary

Our work investigated Tukey's notion of depth for robustifying a given optimization criterion, an estimating equation, or an algorithm in statistical inference and estimation. In Part I, we introduced a polished subspace depth framework, where the elements like the influence space constraint, rectified redescending discrepancy measures, and subspace projection are new to the best of our knowledge. In Part II, we proposed two novel approaches based on manifolds and slack variables to extend the concept to problems defined in some restricted parameter spaces or with a nonsmooth regularizer. Our matrix formulation of the problems, together with state-of-the-art optimization techniques (particularly momentum-based acceleration), gave rise to a new class of efficient algorithms that has guaranteed convergence and scales up with problem dimensions. The efficient computation of the deepest point or composite depth (cf. Remark 1 of Part I) is yet more difficult, and recent advances in nonconvex min-max optimization (Razaviyayn et al. 2020) may shed new light on the topic.

The proposed computational inference tool caters to machine learning applications

beyond the standard likelihood setup. For example, given a feedforward neural network, it can be used to evaluate the reliability of a given estimate, or an event concerned with some properties of the unknowns, which only requires the gradient information that can be obtained from back propagation with ease. Moreover, the influence-driven deepest estimation provides a universal means of accommodating distortions and anomalies given any criterion or estimation equations. We hope that the work is helpful to advance the practice of data depth in sophisticated setups and in higher dimensions.

Acknowledgments

We would like to thank the editor, the associate editor and three anonymous referees for their careful comments and useful suggestions that significantly improved the quality of the paper. The first author would also like to thank Peter Rousseeuw for his inspiration and encouragement, as well as his valuable comments on an earlier version of the manuscript. The work is supported by the National Science Foundation.

References

Alfons, A., Croux, C., and Gelper, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *The Annals of Applied Statistics*, 7(1):226–248.

Anderson, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Annals of Mathematical Statistics*, 22:327–351.

Bern and Eppstein (2002). Multivariate regression depth. *Discrete & Computational Geometry*, 28(1):1–17, ISSN: 1432-0444.

Bijral, A. S., Breitenbach, M., and Grudic, G. (2007). Mixture of Watson distributions: A generative model for hyperspherical embeddings. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, pages 35–42, San Juan, Puerto Rico.

Boothby, W. M. (1986). *An introduction to differentiable manifolds and Riemannian geometry*. Academic Press, Orlando, FL.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press, New York, NY, ISBN: 0-521-83378-7.

Byrd, R., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208.

Cai, T. T., Xu, G., and Zhang, J. (2009). On recovery of sparse signals via ℓ_1 minimization. *IEEE Transactions on Information Theory*, 55(7):3388–3397, ISSN: 0018-9448.

Dhillon, I. S., Marcotte, E. M., and Roshan, U. (2003). Diametrical clustering for identifying anti-correlated gene clusters. *Bioinformatics*, 19(13):1612–1619.

Edelman, A., Arias, T. A., and Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Harrison, D. and Rubinfeld, D. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5:81–102.

Izenman, A. (2008). *Modern Multivariate Statistical Techniques: Regression, Classification and Manifold Learning*. Springer, New York.

Liu, R. Y. and Singh, K. (1992). Ordering directional data: concepts of data depth on circles and spheres. *The Annals of Statistics*, 20(3):1468–1484.

Lütkepohl, H. (2007). *New Introduction to Multiple Time Series Analysis*. Springer-Verlag Berlin Heidelberg, ISBN: 3540262393, 9783540262398.

Mardia, K. V. and Jupp, P. E. (1999). *Directional statistics*. John Wiley & Sons, Hoboken, NJ.

Mizera, I. (2002). On depth and deep points: a calculus. *The Annals of Statistics*, 30(6):1681–1736.

Parikh, N. and Boyd, S. (2014). Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, ISSN: 2167-3888.

Razaviyayn, M., Huang, T., Lu, S., Nouiehed, M., Sanjabi, M., and Hong, M. (2020). Nonconvex min-max optimization: Applications, challenges, and recent theoretical advances. *IEEE Signal Processing Magazine*, 37(5):55–66.

Reinsel, G. and Velu, R. (1998). *Multivariate Reduced-Rank Regression: Theory and Applications*. Springer, New York.

Rothman, A. J., Levina, E., and Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, 19(4):947–962.

Rousseeuw, P. J. and Hubert, M. (1999). Regression depth. *Journal of the American Statistical Association*, 94(446):388–402.

She, Y. (2012). An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors. *Computational Statistics & Data Analysis*, 56(10):2976–2990, ISSN: 0167-9473.

She, Y. (2013). Reduced rank vector generalized linear models for feature extraction. *Statistics and Its Interface*, 6:197–209.

She, Y. (2016). On the finite-sample analysis of Θ-estimators. *Electron. J. Statist.*, 10(2):1874–1895.

She, Y. (2017). Selective factor extraction in high dimensions. *Biometrika*, 104(1):97–110.

She, Y. and Chen, K. (2017). Robust reduced-rank regression. *Biometrika*, 104(3):633–647, DOI: 10.1093/biomet/asx032.

She, Y., Li, S., and Wu, D. (2016). Robust orthogonal complement principal component analysis. *Journal of the American Statistical Association*, 111(514):763–771.

She, Y., Tang, S., and Liu, L. (2022a). On Generalization and Computation of Tukey's Depth: Part I. *Journal of Data Science, Statistics, and Visualisation*, 2(1), DOI: 10.52933/jdssv.v2i1.23.

She, Y. and Tran, H. (2019). On cross-validation for sparse reduced rank regression. *Journal of the Royal Statistical Society: Series B*, 81:145–161.

She, Y., Wang, Z., and Jin, J. (2021b). Analysis of Generalized Bregman Surrogate Algorithms for Nonsmooth Nonconvex Statistical Learning. *The Annals of Statistics*, 49(6):3434–3459.

She, Y., Wang, Z., and Shen, J. (2022b). Gaining Outlier Resistance with Progressive Quantiles: Fast Algorithms and Theoretical Studies. *Journal of the American Statistical Association*. To appear.

Sherwood, B. and Maidman, A. (2019). *rqPen: Penalized Quantile Regression*, https://CRAN.R-project.org/package=rqPen. R package version 2.1.

Sra, S. and Karp, D. (2013). The multivariate watson distribution: Maximum-likelihood estimation and other aspects. *Journal of Multivariate Analysis*, 114:256 – 269, ISSN: 0047-259X.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, ISSN: 00359246.

Tukey, J. W. (1975). Mathematics and the picturing of data. In *Proceedings of the international congress of mathematicians*, volume 2.

Watson, G. S. (1965). Equatorial distributions on a sphere. *Biometrika*, 52(1/2):193–201, ISSN: 00063444.

Yang, Y. and Zhao, Z. (2020). *RRRR: Online Robust Reduced-Rank Regression Estimation*, https://CRAN.R-project.org/package=RRRR.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942.

Zhao, Z. and Palomar, D. P. (2017). Robust maximum likelihood estimation of sparse vector error correction model. In *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 913–917. IEEE.

Affiliation:

Yiyuan She
Department of Statistics
Florida State University
117 N, Woodward Ave
E-mail: yshe@stat.fsu.edu

Journal of Data Science, Statistics, and Visualisation https://jdssv.org/
published by the International Association for Statistical Computing
http://iasc-isi.org/

ISSN 2773-0689
July 2021, Volume II, Issue II *Submitted:* 2020-12-22
doi:10.52933/jdssv.v2i2.61 *Accepted:* 2021-07-13