Electronic Journal of Statistics
Vol. 14 (2020) 207–238
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1662

Perspective maximum likelihood-type estimation via proximal decomposition

Patrick L. Combettes*

North Carolina State University, Department of Mathematics, Raleigh, USA
e-mail: plc@math.ncsu.edu

and

Christian L. Müller

Center for Computational Mathematics, Flatiron Institute, New York, USA
Institute of Computational Biology, Helmholtz Zentrum, München, Germany
Department of Statistics, Ludwig-Maximilians-Universität, München, Germany
e-mail: cmueller@flatironinstitute.org

Abstract: We introduce a flexible optimization model for maximum likelihood-type estimation (M-estimation) that encompasses and generalizes a large class of existing statistical models, including Huber's concomitant M-estimator, Owen's Huber/Berhu concomitant estimator, the scaled lasso, support vector machine regression, and penalized estimation with structured sparsity. The model, termed perspective M-estimation, leverages the observation that convex M-estimators with concomitant scale as well as various regularizers are instances of perspective functions, a construction that extends a convex function to a jointly convex one in terms of an additional scale variable. These nonsmooth functions are shown to be amenable to proximal analysis, which leads to principled and provably convergent optimization algorithms via proximal splitting. We derive novel proximity operators for several perspective functions of interest via a geometrical approach based on duality. We then devise a new proximal splitting algorithm to solve the proposed M-estimation problem and establish the convergence of both the scale and regression iterates it produces to a solution. Numerical experiments on synthetic and real-world data illustrate the broad applicability of the proposed framework.

MSC 2010 subject classifications: Primary 90C25, 62J02; secondary 46N30, 62P10.
Keywords and phrases: Convex optimization, heteroscedastic model, concomitant M-estimator, perspective function, proximal algorithm, robust regression.

Received April 2019.

Contents

*The work of P. L. Combettes was supported by the National Science Foundation under grant DMS-1818946.

1. Introduction

High-dimensional regression methods play a pivotal role in modern data analysis. A large body of statistical work has focused on estimating regression coefficients under various structural assumptions, such as sparsity of the regression vector [36]. In the standard linear framework, regression coefficients constitute, however, only one aspect of the model. A more fundamental objective in statistical inference is the estimation of both location (i.e., the regression coefficients) and scale (e.g., the standard deviation of the noise) of the statistical model from the data. A common approach is to decouple this estimation process by designing and analyzing individual estimators for scale and location parameters (see, e.g., [21, pp. 140], [41]) because joint estimation often leads to non-convex formulations [14, 34]. One important exception has been proposed in robust statistics in the form of a maximum likelihood-type estimator (M-estimator) for location with concomitant scale [21, pp. 179], which couples both parameters via a convex objective function. To discuss this approach more precisely, we introduce the linear heteroscedastic mean shift regression model. This data formation model will be used throughout the paper.

Model 1.1. The vector $y = (\eta_i)_{1\leqslant i\leqslant n} \in \mathbb{R}^n$ of observations is

$$y = X\overline{b} + \overline{o} + Ce, \tag{1.1}$$

where $X \in \mathbb{R}^{n\times p}$ is a known design matrix with rows $(x_i)_{1\leqslant i\leqslant n}$, $\overline{b} \in \mathbb{R}^p$ is the unknown regression vector (location), $\overline{o} \in \mathbb{R}^n$ is the unknown mean shift vector containing outliers, $e \in \mathbb{R}^n$ is a vector of realizations of i.i.d. zero mean random variables, and $C \in [0, +\infty[^{n\times n}$ is a diagonal matrix the diagonal of which are the (unknown) standard deviations. One obtains the homoscedastic mean shift model when the diagonal entries of C are identical.

The concomitant M-estimator proposed in [21, pp. 179] is based on the objective function

$$(\sigma, b) \mapsto \frac{\sigma}{n}\sum_{i=1}^{n}\left(\mathsf{h}_{\rho_1}\left(\frac{x_i^\top b - \eta_i}{\sigma}\right) + \delta\right), \tag{1.2}$$

where h_{ρ_1} is the Huber function [20] with parameter $\rho_1 \in]0, +\infty[$, $\delta \in [0, +\infty[$, and the scalar σ is a scale. The objective function, which we also refer to as the homoscedastic Huber M-estimator function, is jointly convex in both b and scalar σ, and hence, amenable to global optimization. Under suitable assumptions, this estimator can identify outliers o and can estimate a scale that is proportional to the diagonal entries of C in the homoscedastic case. In [2], it was proposed that joint convex optimization of regression vector and standard deviation may also be advantageous in sparse linear regression. There, the objective function is

$$(\sigma, b) \mapsto \frac{\sigma}{n} \sum_{i=1}^{n} \left(\left| \frac{x_i^\top b - \eta_i}{\sigma} \right|^2 + \delta \right) + \alpha_1 \|b\|_1, \tag{1.3}$$

where the term $\|\cdot\|_1$ promotes sparsity of the regression estimate, $\alpha_1 \in]0, +\infty[$ is a tuning parameter, and σ is an estimate of the standard deviation. This objective function is at the heart of the scaled lasso estimator [35]. The resulting estimator is not robust to outliers but is equivariant, which makes the tuning parameter α_1 independent of the noise level. In [29], an extension of (1.2) was introduced that includes a new penalization function as well as concomitant scale estimation for the regression vector. The objective function is

$$(\sigma, \tau, b) \mapsto \frac{\sigma}{n} \sum_{i=1}^{n} \left(\mathsf{h}_{\rho_1} \left(\frac{x_i^\top b - \eta_i}{\sigma} \right) + \delta_1 \right) + \frac{\alpha_1 \tau}{p} \sum_{i=1}^{p} \left(\mathsf{b}_{\rho_2} \left(\frac{\beta_i}{\tau} \right) + \delta_2 \right), \tag{1.4}$$

where b_{ρ_2} is the reverse Huber (Berhu) function [29] with parameter $\rho_2 \in]0, +\infty[$, constants $\delta_1 \in]0, +\infty[$ and $\delta_2 \in]0, +\infty[$, and tuning parameter $\alpha_1 \in]0, +\infty[$. This objective function is jointly convex in b and the scalar parameters σ and τ. The estimator inherits the equivariance and robustness of the previous estimators. In addition, the Berhu penalty is advantageous when the design matrix comprises correlated rows [23]. In [10], it was observed that these objective functions, turn out to be instances of the class of composite "perspective functions" [8], a powerful construct that extends a convex function of a single variable to a jointly convex one in terms of an additional scale variable (see Section 2.2 for a formal definition). Let us add that perspective functions are also implicitly present in many data analysis models in the form of regularization penalties for structured sparsity [3, 25, 26].

In the present paper, we bring to light the ubiquity of perspective functions in statistical M-estimation and introduce a new statistical optimization model, *perspective M-estimation*. The proposed perspective M-estimation model, put forward in detail in Section 3, uses perspective functions as fundamental building blocks to couple scale and regression variables in a jointly convex fashion. It includes in particular the formulations discussed in [10] as well as the M-estimators discussed above as special cases, and it will be seen to cover a wide range of models beyond those. In [10] an algorithm was proposed to solve a specific formulation involving perspective functions in the context of generalized TREX estimation. To date, however, there exists no provably convergent algorithm to solve composite convex optimization problems involving general

perspective functions. To fill this gap, we construct in Section 4 a new proximal splitting algorithm tailored to the perspective M-estimation problem and rigorously establish the convergence of its iterates. Since the proximity operators of perspective functions are known only in limited cases [10], another important contribution of our work is to derive new ones to broaden the effective scope of the proposed perspective M-estimation framework. Using geometrical insights revealed by the dual problem, we derive in Section 2 new proximity operators for several perspective functions, including the generalized scaled lasso, the generalized Huber, the abstract Vapnik, and the generalized Berhu function. These developments lead to a unifying algorithmic framework for global optimization of the proposed model using modern splitting techniques. The model also allows for the seamless integration of a large class of regularizers for structured sparsity and novel robust heteroscedastic estimators of location and scale. Numerical experiments on synthetic and real-world data illustrate the applicability of the proposed framework in Section 5.

2. Proximity operators of perspective functions

The general perspective M-estimation model to be proposed in Problem 3.1 will hinge on the notion of a perspective function (see (2.15) below). Since perspective functions are nonsmooth, to solve Problem 3.1 we need to bring into play the machinery of proximal methods [4] and must therefore be able to compute the proximity operators of these functions. A few examples of such computations were presented in [10]. In this section, using a novel geometric approach, we derive a number of important new instances. Since these results are of general interest beyond statistical analysis, throughout, $\mathcal{H}$ is a real Hilbert space with scalar product $\langle \cdot \mid \cdot \rangle$ and associated norm $\| \cdot \|$.

2.1. Notation and background on convex analysis

The closed ball with center $x \in \mathcal{H}$ and radius $\rho \in]0, +\infty[$ is denoted by $B(x; \rho)$. Let C be a subset of $\mathcal{H}$. Then

$$\iota_C \colon \mathcal{H} \to \{0, +\infty\} \colon x \mapsto \begin{cases} 0, & \text{if } x \in C; \\ +\infty, & \text{if } x \notin C \end{cases} \tag{2.1}$$

is the indicator function of C,

$$d_C \colon \mathcal{H} \to [0, +\infty] : x \mapsto \inf_{y \in C} \|y - x\| \tag{2.2}$$

is the distance function to C, and

$$\sigma_C \colon \mathcal{H} \to [-\infty, +\infty] : u \mapsto \sup_{x \in C} \langle x \mid u \rangle \tag{2.3}$$

is the support function of C. If C is nonempty, closed, and convex then, for every $x \in \mathcal{H}$, there exists a unique point $\operatorname{proj}_C x \in C$, called the projection of x

onto C, such that $\|x - \operatorname{proj}_C x\| = d_C(x)$. We have

$$(\forall x \in \mathcal{H})(\forall p \in \mathcal{H}) \quad p = \operatorname{proj}_C x \;\Leftrightarrow\; \begin{cases} p \in C \\ (\forall y \in C) \quad \langle y - p \mid x - p \rangle \leqslant 0. \end{cases} \tag{2.4}$$

The normal cone to C is

$$N_C \colon \mathcal{H} \to 2^{\mathcal{H}} \colon x \mapsto \begin{cases} \big\{u \in \mathcal{H} \;\big|\; \sup \langle C - x \mid u \rangle \leqslant 0\big\}, & \text{if } \; x \in C; \\ \varnothing, & \text{otherwise.} \end{cases} \tag{2.5}$$

A function $\varphi \colon \mathcal{H} \to \left]-\infty, +\infty\right]$ is proper if $\operatorname{dom} \varphi = \big\{x \in \mathcal{H} \;\big|\; \varphi(x) < +\infty\big\} \neq \varnothing$ and coercive if $\lim_{\|x\| \to +\infty} \varphi(x) = +\infty$. We denote by $\Gamma_0(\mathcal{H})$ the class of proper lower semicontinuous convex functions from $\mathcal{H}$ to $\left]-\infty, +\infty\right]$. Let $\varphi \in \Gamma_0(\mathcal{H})$. The conjugate of φ is

$$\varphi^* \colon \mathcal{H} \to \left]-\infty, +\infty\right] : u \mapsto \sup_{x \in \mathcal{H}} \big(\langle x \mid u \rangle - \varphi(x)\big). \tag{2.6}$$

It also belongs to $\Gamma_0(\mathcal{H})$ and $\varphi^{**} = \varphi$. The Moreau subdifferential of φ is the set-valued operator

$$\partial\varphi \colon \mathcal{H} \to 2^{\mathcal{H}} \colon x \mapsto \big\{u \in \mathcal{H} \;\big|\; (\forall y \in \operatorname{dom} \varphi) \;\; \langle y - x \mid u \rangle + \varphi(x) \leqslant \varphi(y)\big\}. \tag{2.7}$$

We have

$$(\forall x \in \mathcal{H})(\forall u \in \mathcal{H}) \quad u \in \partial\varphi(x) \quad \Leftrightarrow \quad x \in \partial\varphi^*(u). \tag{2.8}$$

Moreover,

$$(\forall x \in \mathcal{H})(\forall u \in \mathcal{H}) \quad \varphi(x) + \varphi^*(u) \geqslant \langle x \mid u \rangle \tag{2.9}$$

and

$$(\forall x \in \mathcal{H})(\forall u \in \mathcal{H}) \quad u \in \partial\varphi(x) \quad \Leftrightarrow \quad \varphi(x) + \varphi^*(u) = \langle x \mid u \rangle. \tag{2.10}$$

If φ is Gâteaux differentiable at $x \in \operatorname{dom} \varphi$, with gradient $\nabla\varphi(x)$, then

$$\partial\varphi(x) = \{\nabla\varphi(x)\}. \tag{2.11}$$

The infimal convolution of φ and $\psi \in \Gamma_0(\mathcal{H})$ is

$$\varphi \,\square\, \psi \colon \mathcal{H} \to [-\infty, +\infty] : x \mapsto \inf_{y \in \mathcal{H}} \big(\varphi(y) + \psi(x - y)\big). \tag{2.12}$$

Given any $z \in \operatorname{dom} \varphi$, the recession function of φ is

$$(\forall x \in \mathcal{H}) \quad (\operatorname{rec} \varphi)(x) = \sup_{y \in \operatorname{dom} \varphi} \big(\varphi(x + y) - \varphi(y)\big) = \lim_{\alpha \to +\infty} \frac{\varphi(z + \alpha x)}{\alpha}. \tag{2.13}$$

Finally, the proximity operator of φ is [27]

$$\operatorname{prox}_{\varphi} \colon \mathcal{H} \to \mathcal{H} \colon x \mapsto \operatorname*{argmin}_{y \in \mathcal{H}} \left(\varphi(y) + \frac{1}{2}\|x - y\|^2\right). \tag{2.14}$$

For detailed accounts of convex analysis, see [4, 31].

2.2. The geometry of proximity operators of perspective functions

Let $\varphi \in \Gamma_0(\mathcal{H})$. The perspective of φ is

$$\widetilde{\varphi}\colon \mathbb{R} \times \mathcal{H} \to \left]-\infty, +\infty\right] : (\sigma, x) \mapsto \begin{cases} \sigma\varphi(x/\sigma), & \text{if } \sigma > 0; \\ (\operatorname{rec}\varphi)(x), & \text{if } \sigma = 0; \\ +\infty, & \text{otherwise.} \end{cases} \tag{2.15}$$

We have $\widetilde{\varphi} \in \Gamma_0(\mathbb{R} \oplus \mathcal{G})$ [8, Proposition 2.3]. The following property is useful to establish existence results for problems involving perspective functions.

Proposition 2.1. *Let $\varphi \in \Gamma_0(\mathcal{H})$ be such that $\inf \varphi(\mathcal{H}) > 0$ and $0 \in \operatorname{int}\operatorname{dom}\varphi^*$. Then $\widetilde{\varphi}$ is coercive.*

Proof. We have $\varphi^*(0) = -\inf \varphi(\mathcal{H}) < 0$ and $0 \in \operatorname{int}\operatorname{dom}\varphi^*$. Hence, $(0,0) \in \operatorname{int}\operatorname{epi}\varphi^*$. In turn, we derive from [8, Proposition 2.3(iv)] that

$$(0,0) \in \operatorname{int}\big\{(\mu, u) \in \mathbb{R} \oplus \mathcal{H} \;\big|\; \mu + \varphi^*(u) \leqslant 0\big\} = \operatorname{int}\operatorname{dom}(\widetilde{\varphi})^*. \tag{2.16}$$

It therefore follows from [4, Proposition 14.16] that $\widetilde{\varphi}$ is coercive. □

Let us now turn to the proximity operator of $\widetilde{\varphi}$.

Lemma 2.2 ([10, Theorem 3.1])**.** *Let $\varphi \in \Gamma_0(\mathcal{H})$, let $\gamma \in \left]0, +\infty\right[$, let $\sigma \in \mathbb{R}$, and let $x \in \mathcal{H}$. Then the following hold:*

(i) *Suppose that $\sigma + \gamma\varphi^*(x/\gamma) \leqslant 0$. Then $\operatorname{prox}_{\gamma\widetilde{\varphi}}(\sigma, x) = (0,0)$.*
(ii) *Suppose that $\operatorname{dom}\varphi^*$ is open and that $\sigma + \gamma\varphi^*(x/\gamma) > 0$. Then*

$$\operatorname{prox}_{\gamma\widetilde{\varphi}}(\sigma, x) = \big(\sigma + \gamma\varphi^*(p), x - \gamma p\big), \tag{2.17}$$

where p is the unique solution to the inclusion $x \in \gamma p + (\sigma + \gamma\varphi^*(p))\partial\varphi^*(p)$. If φ^* is differentiable at p, then p is characterized by $x = \gamma p + (\sigma + \gamma\varphi^*(p))\nabla\varphi^*(p)$.

When $\operatorname{dom}\varphi^*$ is not open, Lemma 2.2 is not applicable. To deal with such cases, we propose a geometric construction that computes $\operatorname{prox}_{\gamma\widetilde{\varphi}}$ via the projection onto a certain convex set. It is based on the following property, which reduces the problem of evaluating the proximity operator of $\widetilde{\varphi}$ to a projection problem in $\mathbb{R}^2$ if φ is radially symmetric.

Proposition 2.3. *Let $\phi \in \Gamma_0(\mathbb{R})$ be an even function, set $\varphi = \phi \circ \|\cdot\| : \mathcal{H} \to \left]-\infty, +\infty\right]$, let $\gamma \in \left]0, +\infty\right[$, let $\sigma \in \mathbb{R}$, and let $x \in \mathcal{H}$. Set*

$$\mathcal{R} = \big\{(\chi, \nu) \in \mathbb{R}^2 \;\big|\; \chi + \phi^*(\nu) \leqslant 0\big\}. \tag{2.18}$$

Then $\mathcal{R}$ is a nonempty closed convex set, and the following hold:

(i) *Suppose that $\sigma + \gamma\phi^*(\|x\|/\gamma) \leqslant 0$. Then $\operatorname{prox}_{\gamma\widetilde{\varphi}}(\sigma, x) = (0,0)$.*
(ii) *Suppose that $\sigma > \gamma\phi(0)$ and $x = 0$. Then $\operatorname{prox}_{\gamma\widetilde{\varphi}}(\sigma, x) = (\sigma - \gamma\phi(0), x)$.*

(iii) *Suppose that* $\sigma + \gamma\phi^*(\|x\|/\gamma) > 0$ *and* $x \neq 0$, *and set*

$$(\chi, \nu) = \operatorname{proj}_{\mathcal{R}}(\sigma/\gamma, \|x\|/\gamma). \tag{2.19}$$

Then

$$\operatorname{prox}_{\gamma\widetilde{\varphi}}(\sigma, x) = \left(\sigma - \gamma\chi, \left(1 - \frac{\gamma\nu}{\|x\|}\right) x\right). \tag{2.20}$$

Proof. The properties of $\mathcal{R}$ follow from the fact that $\phi^* \in \Gamma_0(\mathbb{R})$. Now, let us recall from [10, Remark 3.2] that, if

$$\mathcal{C} = \big\{(\mu, u) \in \mathbb{R} \oplus \mathcal{H} \;\big|\; \mu + \varphi^*(u) \leqslant 0\big\}, \tag{2.21}$$

then

$$\operatorname{prox}_{\gamma\widetilde{\varphi}}(\sigma, x) = (\sigma, x) - \gamma\operatorname{proj}_{\mathcal{C}}\big(\sigma/\gamma, x/\gamma\big). \tag{2.22}$$

In addition, [4, Example 13.8] states that

$$\varphi^* = \phi^* \circ \|\cdot\|. \tag{2.23}$$

(i): This follows from (2.23) and Lemma 2.2(i).

(ii): Let us show that $\operatorname{proj}_{\mathcal{C}}(\sigma/\gamma, 0) = (\phi(0), 0)$, which will establish the claim by virtue of (2.22). Since ϕ is an even function in $\Gamma_0(\mathbb{R})$, $\phi(0) = \inf \phi(\mathbb{R}) = -\phi^*(0)$. Hence $\phi(0)+\varphi^*(0) = \phi(0)+\phi^*(0) = 0$ and $(\phi(0), 0) \in \mathcal{C}$. Now fix $(\eta, y) \in \mathcal{C}$. Then, since φ is an even function in $\Gamma_0(\mathcal{H})$, $\eta \leqslant -\varphi^*(y) \leqslant -\inf \varphi^*(\mathcal{H}) = -\varphi^*(0) = -\phi^*(0) = \phi(0)$ and, since $\sigma > \gamma\phi(0)$, we get

$$\langle(\eta, y) - (\phi(0), 0) \mid (\sigma/\gamma, 0) - (\phi(0), 0)\rangle = \big(\eta - \phi(0)\big)\big(\sigma/\gamma - \phi(0)\big) \leqslant 0. \tag{2.24}$$

Altogether, (2.4) asserts that $\operatorname{proj}_{\mathcal{C}}(\sigma/\gamma, 0) = (\phi(0), 0)$.

(iii): In view of (2.22), it is enough to show that $\operatorname{proj}_{\mathcal{C}}(\sigma/\gamma, x/\gamma) = (\chi, \nu x/\|x\|)$. Since $(\chi, \nu) \in \mathcal{R}$, (2.23) yields $\chi+\varphi^*(\nu x/\|x\|) = \chi+\phi^*(\nu) \leqslant 0$ and, therefore, $(\chi, \nu x/\|x\|) \in \mathcal{C}$. On the other hand, we infer from (2.23) that $\mathcal{C} \subset \mathbb{R} \oplus \mathcal{H}$ is radially symmetric in the $\mathcal{H}$-direction. As a result, $\operatorname{proj}_{\mathcal{C}}\big(\sigma/\gamma, x/\gamma\big) \in V = \mathbb{R} \times \operatorname{span}\{x\}$ and therefore $\operatorname{proj}_{\mathcal{C}}\big(\sigma/\gamma, x/\gamma\big) = \operatorname{proj}_{V\cap\mathcal{C}}\big(\sigma/\gamma, x/\gamma\big)$ [4, Proposition 29.5]. Now fix $(\eta, y) \in V \cap \mathcal{C}$. Then $(\eta, \pm\|y\|) \in \mathcal{R}$ and (2.4) yields

$$\begin{aligned}(\eta - \chi)(\sigma/\gamma - \chi) + (\pm\|y\| - \nu)(\|x\|/\gamma - \nu) = \\ \langle(\eta - \chi, \pm\|y\| - \nu) \mid (\sigma/\gamma - \chi, \|x\|/\gamma - \nu)\rangle_{\mathbb{R}^2} \leqslant 0.\end{aligned} \tag{2.25}$$

Hence, since $y = \pm\|y\|x/\|x\|$,

$$\begin{aligned}&\langle(\eta, y) - (\chi, \nu x/\|x\|) \mid (\sigma/\gamma, x/\gamma) - (\chi, \nu x/\|x\|)\rangle_{\mathbb{R}\oplus\mathcal{H}} \\ &\quad= (\eta - \chi)(\sigma/\gamma - \chi) + \langle y - \nu x/\|x\| \mid x/\gamma - \nu x/\|x\|\rangle \\ &\quad= (\eta - \chi)(\sigma/\gamma - \chi) + \langle\pm\|y\|x/\|x\| - \nu x/\|x\| \mid \|x\|x/(\gamma\|x\|) - \nu x/\|x\|\rangle \\ &\quad= (\eta - \chi)(\sigma/\gamma - \chi) + (\pm\|y\| - \nu)(\|x\|/\gamma - \nu)\langle x \mid x\rangle/\|x\|^2 \\ &\quad= (\eta - \chi)(\sigma/\gamma - \chi) + (\pm\|y\| - \nu)(\|x\|/\gamma - \nu) \\ &\quad\leqslant 0.\end{aligned} \tag{2.26}$$

Altogether, we derive from (2.4) that $(\chi, \nu x/\|x\|) = \operatorname{proj}_{V\cap\mathcal{C}}\big(\sigma/\gamma, x/\gamma\big) = \operatorname{proj}_{\mathcal{C}}\big(\sigma/\gamma, x/\gamma\big)$. □

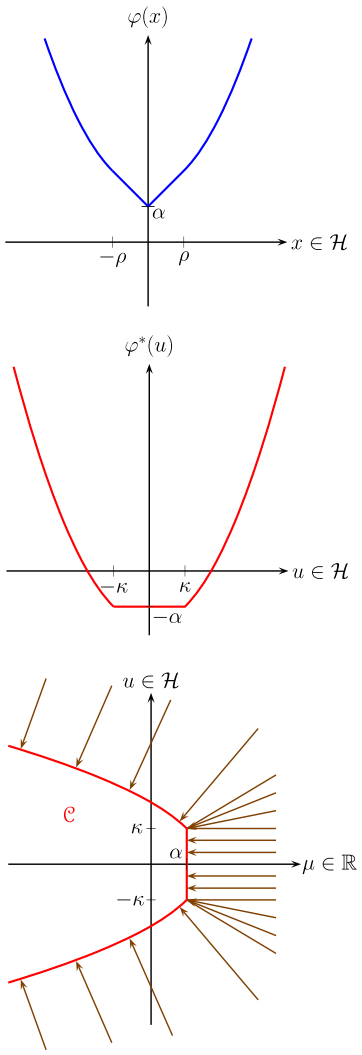

FIG 1. *Geometry of the computation of* $\operatorname{prox}_{\widetilde{\varphi}}$ *in* (2.22). *Top: original function* φ. *Center: conjugate of* φ. *Bottom: action of the projection operator* $\operatorname{proj}_{\mathcal{C}}$ *onto the set* $\mathcal{C}$ *of* (2.21). *The proximity operator of* $\widetilde{\varphi}$ *is* $\operatorname{Id} - \operatorname{proj}_{\mathcal{C}}$. *In the specific example depicted here,* $\mathcal{H} = \mathbb{R}$ *and* φ *is the Berhu function of* (2.58).

2.3. Examples

We provide several examples that are relevant to the statistical problems we have in sight.

Example 2.4 (generalized scaled lasso function)**.** [10, Example 3.7] Let $\alpha \in]0,+\infty[$, $\gamma \in]0,+\infty[$, $\kappa \in]0,+\infty[$, $q \in]1,+\infty[$, $\sigma \in \mathbb{R}$, and $x \in \mathcal{H}$. Set

$\varphi = \alpha + \|\cdot\|^q/\kappa\colon \mathcal{H} \to \mathbb{R}$ and $q^* = q/(q-1)$. Then

$$\widetilde{\varphi}(\sigma, x) = \begin{cases} \alpha\sigma + \dfrac{\|x\|^q}{\kappa\sigma^{q-1}}, & \text{if } \sigma > 0; \\ 0, & \text{if } x = 0 \text{ and } \sigma = 0; \\ +\infty, & \text{otherwise.} \end{cases} \tag{2.27}$$

Now set $\rho = (\kappa/q)^{q^*-1}$. If $q^*\gamma^{q^*-1}\sigma + \rho\|x\|^{q^*} > q^*\gamma^{q^*}\alpha$ and $x \neq 0$, let t be the unique solution in $]0, +\infty[$ to the equation

$$t^{2q^*-1} + \frac{q^*(\sigma - \gamma\alpha)}{\gamma\rho}t^{q^*-1} + \frac{q^*}{\rho^2}t - \frac{q^*\|x\|}{\gamma\rho^2} = 0. \tag{2.28}$$

Set $p = tx/\|x\|$ if $x \neq 0$, and $p = 0$ if $x = 0$. Then (note that [10, Eq. (3.47)] is incorrect when $\alpha \neq 0$)

$$\operatorname{prox}_{\gamma\widetilde{\varphi}}(\sigma, x) = \begin{cases} \big(\sigma + \gamma\big(\rho t^{q^*}/q^* - \alpha\big), x - \gamma p\big), & \text{if } q^*\gamma^{q^*-1}\sigma + \rho\|x\|^{q^*} > q^*\gamma^{q^*}\alpha; \\ (0, 0), & \text{if } q^*\gamma^{q^*-1}\sigma + \rho\|x\|^{q^*} \leqslant q^*\gamma^{q^*}\alpha. \end{cases} \tag{2.29}$$

Given $\rho \in]0, +\infty[$, the classical Huber function is defined as [20]

$$\mathsf{h}_\rho\colon \mathbb{R} \to \mathbb{R}\colon \xi \mapsto \begin{cases} \rho|\xi| - \dfrac{\rho^2}{2}, & \text{if } |\xi| > \rho; \\ \dfrac{|\xi|^2}{2}, & \text{if } |\xi| \leqslant \rho. \end{cases} \tag{2.30}$$

Below, we study the perspective of a generalization of it.

Example 2.5 (generalized Huber function)**.** Let α, γ, and ρ be in $]0, +\infty[$, let $q \in]1, +\infty[$, and set $q^* = q/(q-1)$. Define

$$\varphi\colon \mathcal{H} \to \mathbb{R}\colon x \mapsto \begin{cases} \alpha - \dfrac{\rho^{q^*}}{q^*} + \rho\|x\|, & \text{if } \|x\| > \rho^{q^*/q}; \\ \alpha + \dfrac{\|x\|^q}{q}, & \text{if } \|x\| \leqslant \rho^{q^*/q}. \end{cases} \tag{2.31}$$

Let $\sigma \in \mathbb{R}$ and $x \in \mathcal{H}$. Then

$$\widetilde{\varphi}(\sigma, x) = \begin{cases} \left(\alpha - \dfrac{\rho^{q^*}}{q^*}\right)\sigma + \rho\|x\|, & \text{if } \sigma > 0 \text{ and } \|x\| > \sigma\rho^{q^*/q}; \\ \alpha\sigma + \dfrac{\|x\|^q}{q\sigma^{q-1}}, & \text{if } \sigma > 0 \text{ and } \|x\| \leqslant \sigma\rho^{q^*/q}; \\ \rho\|x\|, & \text{if } \sigma = 0; \\ +\infty, & \text{if } \sigma < 0. \end{cases} \tag{2.32}$$

In addition, the following hold:

(i) Suppose that $\|x\| \leqslant \gamma\rho$ and $\|x\|^{q^*} \leqslant \gamma^{q^*} q^* (\alpha - \sigma/\gamma)$. Then $\operatorname{prox}_{\gamma\widetilde{\varphi}}(\sigma, x) = (0, 0)$.

(ii) Suppose that $\sigma \leqslant \gamma(\alpha - \rho^{q^*}/q^*)$ and $\|x\| > \gamma\rho$. Then

$$\operatorname{prox}_{\gamma\widetilde{\varphi}}(\sigma, x) = \left(0, \left(1 - \frac{\gamma\rho}{\|x\|}\right)x\right). \tag{2.33}$$

(iii) Suppose that $\sigma > \gamma(\alpha - \rho^{q^*}/q^*)$ and $\|x\| \geqslant \gamma\rho^{q^*-1}(\sigma/\gamma + \rho^{2-q^*} + \rho^{q^*}/q^* - \alpha)$. Then

$$\operatorname{prox}_{\gamma\widetilde{\varphi}}(\sigma, x) = \left(\sigma + \gamma\left(\frac{\rho^{q^*}}{q^*} - \alpha\right), \left(1 - \frac{\gamma\rho}{\|x\|}\right)x\right). \tag{2.34}$$

(iv) Suppose that $\|x\|^{q^*} > q^*\gamma^{q^*}(\alpha - \sigma/\gamma)$ and $\|x\| < \gamma\rho^{q^*-1}(\sigma/\gamma + \rho^{2-q^*} + \rho^{q^*}/q^* - \alpha)$. If $x \neq 0$, let t be the unique solution in $]0, +\infty[$ to the equation

$$\gamma t^{2q^*-1} + q^*(\sigma - \gamma\alpha)t^{q^*-1} + \gamma q^* t - q^*\|x\| = 0. \tag{2.35}$$

Set $p = tx/\|x\|$ if $x \neq 0$, and $p = 0$ if $x = 0$. Then

$$\operatorname{prox}_{\gamma\widetilde{\varphi}}(\sigma, x) = \begin{cases} \big(\sigma + \gamma(t^{q^*}/q^* - \alpha), x - \gamma p\big), & \text{if } q^*\gamma^{q^*-1}\sigma + \|x\|^{q^*} > q^*\gamma^{q^*}\alpha; \\ (0, 0), & \text{if } q^*\gamma^{q^*-1}\sigma + \|x\|^{q^*} \leqslant q^*\gamma^{q^*}\alpha. \end{cases} \tag{2.36}$$

Proof. We derive (2.32) from (2.31), (2.15), and the property that $\operatorname{rec}\varphi = \operatorname{rec}(\rho\|\cdot\|) = \rho\|\cdot\|$. Now set

$$\phi\colon \mathbb{R} \to \mathbb{R}\colon \xi \mapsto \begin{cases} \alpha - \dfrac{\rho^{q^*}}{q^*} + \rho|\xi|, & \text{if } |\xi| > \rho^{q^*/q}; \\ \alpha + \dfrac{|\xi|^q}{q}, & \text{if } |\xi| \leqslant \rho^{q^*/q}. \end{cases} \tag{2.37}$$

Then $\phi = (\rho|\cdot|) \,\square\, (|\cdot|^q/q) + \alpha$ is convex and even, and $\varphi = \phi \circ \|\cdot\|$. We derive from [4, Proposition 13.24(i) and Example 13.2(i)] that

$$\phi^* = \left((\rho|\cdot|) \,\square\, \left(\frac{|\cdot|^q}{q}\right)\right)^* - \alpha = \iota_{[-\rho,\rho]} + \frac{|\cdot|^{q^*}}{q^*} - \alpha. \tag{2.38}$$

In turn, (2.38) and (2.18) yield

$$\mathcal{R} = \mathcal{R}_1 \cap \mathcal{R}_2, \quad \text{where} \quad \begin{cases} \mathcal{R}_1 = \mathbb{R} \times [-\rho, \rho] \\ \mathcal{R}_2 = \big\{(\chi, \nu) \in \mathbb{R}^2 \;\big|\; |\nu|^{q^*} \leqslant q^*(\alpha - \chi)\big\}. \end{cases} \tag{2.39}$$

Now set $(\chi, \nu) = \operatorname{proj}_{\mathcal{R}}(\sigma/\gamma, \|x\|/\gamma)$.

(i): This follows from Proposition 2.3(i) and (2.38).

(ii): Since $\sigma/\gamma \leqslant \alpha - \rho^{q^*}/q^*$, we have

$$(\chi, \nu) = \operatorname{proj}_{\mathcal{R}_1}\big(\sigma/\gamma, \|x\|/\gamma\big) = \big(\sigma/\gamma, \operatorname{proj}_{[-\rho,\rho]}(\|x\|/\gamma)\big). \tag{2.40}$$

Thus, since $\|x\|/\gamma > \rho$, $(\chi, \nu) = (\sigma/\gamma, \rho)$ and (2.33) follows from Proposition 2.3(iii).

(iii): The point $\Pi = (\alpha - \rho^{q^*}/q^*, \rho)$ is in the intersection of the boundaries of $\mathcal{R}_1$ and $\mathcal{R}_2$. Therefore, the normal cone to $\mathcal{R}$ at Π is generated by outer normals n_1 to $\mathcal{R}_1$ and n_2 to $\mathcal{R}_2$ at Π. A tangent vector to $\mathcal{R}_2$ at Π is $t(\Pi) = (-(|\cdot|^{q^*}/q^*)'(\rho), 1) = (-\rho^{q^*-1}, 1)$. We can take $n_1 = (0, 1)$ and $n_2 = (1, \rho^{q^*-1}) \perp t(\Pi)$. Thus, the set of points which have projection Π onto $\mathcal{R}$ is

$$\begin{aligned}
&\Pi + N_{\mathcal{R}}\Pi \\
&= \Pi + \operatorname{cone}(n_1, n_2) \\
&= \big(\alpha - \rho^{q^*}/q^*, \rho\big) + \big\{(\tau, \xi) \in \mathbb{R} \times \mathbb{R} \;\big|\; \tau \geqslant 0 \;\text{ and } \xi \geqslant \rho^{q^*-1}\tau\big\} \\
&= \big\{(\tau, \xi) \in \mathbb{R} \times \mathbb{R} \;\big|\; \tau \geqslant \alpha - \rho^{q^*}/q^* \;\text{ and } \xi - \rho \geqslant \rho^{q^*-1}(\tau - \alpha + \rho^{q^*}/q^*)\big\} \\
&= \big\{(\tau, \xi) \in \mathbb{R} \times \mathbb{R} \;\big|\; \tau \geqslant \alpha - \rho^{q^*}/q^* \;\text{ and } \xi \geqslant \rho + \rho^{q^*-1}(\tau - \alpha) + \rho^{2q^*-1}/q^*)\big\},
\end{aligned} \tag{2.41}$$

and therefore

$$(\chi, \nu) = \big(\alpha - \rho^{q^*}/q^*, \rho\big) \quad \Leftrightarrow \quad \begin{cases} \sigma \geqslant \gamma\big(\alpha - \rho^{q^*}/q^*\big) \\ \|x\| \geqslant \gamma\big(\rho + \rho^{q^*-1}(\sigma/\gamma - \alpha) + \rho^{2q^*-1}/q^*\big)\big). \end{cases} \tag{2.42}$$

In view of Proposition 2.3(iii), this yields (2.34).

(iv): Here $(\sigma/\gamma, \|x\|/\gamma) \notin \mathcal{R}_2$ and $(\chi, \nu) = \operatorname{proj}_{\mathcal{R}_2}(\sigma/\gamma, \|x\|/\gamma)$. Since

$$\mathcal{R}_2 = \big\{(\chi, \nu) \in \mathbb{R}^2 \;\big|\; \chi + \big(\alpha + |\cdot|^q/q\big)^*(\nu) \leqslant 0\big\}, \tag{2.43}$$

the expression of $\operatorname{prox}_{\gamma\widetilde{\varphi}}(\sigma, x)$ is computed exactly as though we were dealing with the generalized scaled lasso function $\alpha + \|\cdot\|^q/q$ of Example 2.4 with $\kappa = q$ and the result is given in (2.29). □

Example 2.6 (generalized Berhu function)**.** Let α, γ, ρ, and κ be in $]0, +\infty[$, let $q \in]1, +\infty[$, and set $C = B(0; \rho)$. Define $\varphi\colon \mathcal{H} \to \mathbb{R}$ by

$$\varphi = \alpha + \kappa\|\cdot\| + \frac{d_C^q}{q\rho^{q^*-1}}, \quad \text{where} \quad q^* = \frac{q}{q-1}, \tag{2.44}$$

and let $\sigma \in \mathbb{R}$ and $x \in \mathcal{H}$. Then

$$\widetilde{\varphi}(\sigma, x) = \begin{cases} \alpha\sigma + \kappa\|x\| + \dfrac{\sigma}{q\rho^{q^*-1}}\left(\dfrac{\|x\|}{\sigma} - \rho\right)^q, & \text{if } \sigma > 0 \text{ and } \|x\| > \rho\sigma; \\ \alpha\sigma + \kappa\|x\|, & \text{if } \sigma > 0 \text{ and } \|x\| \leqslant \rho\sigma; \\ 0, & \text{if } \sigma = 0 \text{ and } x = 0; \\ +\infty, & \text{otherwise.} \end{cases} \tag{2.45}$$

Furthermore, set $\Delta\colon \mathbb{R} \to \mathbb{R}\colon \mu \mapsto \max(|\mu| - \kappa, 0) + \max^{q^*}(|\mu| - \kappa, 0)/q^*$. Then the following hold:

(i) Suppose that $\Delta(\|x\|/\gamma) \leqslant (\alpha - \sigma/\gamma)/\rho$. Then $\operatorname{prox}_{\gamma\widetilde{\varphi}}(\sigma, x) = (0, 0)$.

(ii) Suppose that $\Delta(\|x\|/\gamma) > (\alpha - \sigma/\gamma)/\rho$ and that $\|x\| > \gamma\kappa + \rho(\sigma - \gamma\alpha)$. If $x \neq 0$, let t be the unique solution in $]\kappa, +\infty[$ to the polynomial equation

$$\rho\Bigg(\sigma - \gamma\alpha + \gamma\rho\Bigg(t - \kappa + \frac{(t-\kappa)^{q^*}}{q^*}\Bigg)\Bigg)\Big(1 + (t-\kappa)^{q^*-1}\Big) + \gamma t - \|x\| = 0. \tag{2.46}$$

Set $p = tx/\|x\|$ if $x \neq 0$, and set $t = 0$ and $p = 0$ if $x = 0$. Then $\operatorname{prox}_{\gamma\widetilde{\varphi}}(\sigma, x) = \big(\sigma - \gamma\alpha + \gamma\rho\Delta(t), x - \gamma p\big)$.

(iii) Suppose that $\gamma\kappa \leqslant \|x\| \leqslant \gamma\kappa + \rho(\sigma - \gamma\alpha)$. Then

$$\operatorname{prox}_{\gamma\widetilde{\varphi}}(\sigma, x) = \big(\sigma - \gamma\alpha, (1 - \gamma\kappa/\|x\|)x\big). \tag{2.47}$$

(iv) Suppose that $\sigma > \gamma\alpha$ and $\|x\| < \gamma\kappa$. Then $\operatorname{prox}_{\gamma\widetilde{\varphi}}(\sigma, x) = (\sigma - \gamma\alpha, 0)$.

Proof. The geometry underlying the proof is that depicted in Fig. 1, where $q = 2$. Set $R = [-\rho, \rho]$, $D = [-\kappa, \kappa]$, $\phi = \alpha + \kappa|\cdot| + d^q_{[-\rho,\rho]}/(q\rho^{q/q^*})$, $\theta\colon \mathbb{R} \to \mathbb{R}\colon t \mapsto |t|^q/(q\rho^{q/q^*})$, and $\psi\colon \mathbb{R} \to \mathbb{R}\colon t \mapsto \rho(|t| + |t|^{q^*}/q^*)$. Then $\phi\colon \mathbb{R} \to \mathbb{R}$ is convex and even, and it follows from (2.44) and [4, Example 13.8] that

$$\varphi = \phi \circ \|\cdot\| \quad \text{and} \quad \varphi^* = \phi^* \circ \|\cdot\|. \tag{2.48}$$

Furthermore, $\sigma_D = \kappa|\cdot|$ and we derive from [4, Examples 13.26 and 13.2(i)] that

$$\begin{aligned}
\phi^* &= \big(\sigma_D + \theta \circ d_R\big)^* - \alpha \\
&= \sigma_D^* \,\square\, \big(\theta \circ d_R\big)^* - \alpha \\
&= \iota_D \,\square\, \big(\sigma_R + \theta^* \circ |\cdot|\big) - \alpha \\
&= \iota_D \,\square\, \big(\rho|\cdot| + \theta^* \circ |\cdot|\big) - \alpha \\
&= \iota_D \,\square\, \big(\psi \circ |\cdot|\big) - \alpha \\
&= \big(\psi \circ d_D\big) - \alpha \\
&= \rho\Bigg(d_D + \frac{d_D^{q^*}}{q^*}\Bigg) - \alpha.
\end{aligned} \tag{2.49}$$

In turn, [4, Example 17.33] yields

$$(\forall \nu \in \mathbb{R}) \quad \partial\phi^*(\nu) = \begin{cases} \Bigg\{\rho\Bigg(\dfrac{1 + d_D^{q^*-1}(\nu)}{d_D(\nu)}\Bigg)(\nu - \operatorname{proj}_D \nu)\Bigg\}, & \text{if } \nu \notin D; \\ \big(N_D\nu\big) \cap [-\rho, \rho], & \text{if } \nu \in D. \end{cases} \tag{2.50}$$

However, since $D = [-\kappa, \kappa]$, we have $d_D\colon \nu \mapsto \max(|\nu| - \kappa, 0)$. Therefore, (2.49) implies that

$$(\forall \nu \in \mathbb{R}) \quad \phi^*(\nu) = \rho\Delta(\nu) - \alpha$$

$$= \begin{cases} \rho\big(|\nu| - \kappa + (|\nu| - \kappa)^{q^*}/q^*\big) - \alpha, & \text{if } |\nu| > \kappa; \\ -\alpha, & \text{if } |\nu| \leqslant \kappa \end{cases} \tag{2.51}$$

and

$$(\forall \nu \in \mathbb{R}) \quad \partial\phi^*(\nu) = \begin{cases} \left\{\rho\Big(1 + (|\nu| - \kappa)^{q^*-1}\Big)\operatorname{sign}(\nu)\right\}, & \text{if } |\nu| > \kappa; \\ [0, \rho], & \text{if } \nu = \kappa; \\ [-\rho, 0], & \text{if } \nu = -\kappa; \\ \{0\}, & \text{if } |\nu| < \kappa. \end{cases} \tag{2.52}$$

On the other hand, (2.51) and (2.18) yield

$$\mathcal{R} = \big\{(\chi, \nu) \in \mathbb{R}^2 \;\big|\; \rho\Delta(\nu) \leqslant \alpha - \chi\big\}. \tag{2.53}$$

Now set $\Pi = (\alpha, \kappa)$ and $(\chi, \nu) = \operatorname{proj}_{\mathcal{R}}(\sigma/\gamma, \|x\|/\gamma)$. In view of (2.52), the normal cone to $\operatorname{epi}\phi^*$ at $(\kappa, -\alpha)$ is generated by the vectors $(\rho, -1)$ and $(0, -1)$. Hence, the normal cone to $\mathcal{R}$ at Π is generated by $n_1 = (1, \rho)$ and $n_2 = (1, 0)$, that is

$$N_{\mathcal{R}}\Pi = \big\{(\tau, \xi) \in \mathbb{R} \times \mathbb{R} \;\big|\; 0 \leqslant \xi \leqslant \rho\tau\big\}. \tag{2.54}$$

In turn,

$$\operatorname{proj}_{\mathcal{R}}^{-1}\{\Pi\} = \Pi + N_{\mathcal{R}}\Pi = \big\{(\tau, \xi) \in \mathbb{R}^2 \;\big|\; \kappa \leqslant \xi \leqslant \kappa + \rho(\tau - \alpha)\big\}. \tag{2.55}$$

(i): It follows from the assumptions and (2.51) that $\sigma + \gamma\phi^*(\|x\|/\gamma) \leqslant 0$. In turn, Proposition 2.3(i) implies that $\operatorname{prox}_{\gamma\tilde{\varphi}}(\sigma, x) = (0, 0)$.

(ii): We have $(\sigma/\gamma, \|x\|/\gamma) \notin \mathcal{R} \supset]-\infty, \alpha] \times [-\kappa, \kappa]$ and $\|x\|/\gamma > \kappa + \rho(\sigma/\gamma - \alpha)$. Hence $|\nu| > \kappa$. Now set $(\pi, p) = \operatorname{proj}_{\mathcal{C}}(\sigma/\gamma, x/\gamma)$. Then $\|p\| = |\nu| > \kappa$. Therefore, since it results from (2.48) and (2.49) that $\operatorname{dom}\varphi^* = \operatorname{dom}(\phi^* \circ \|\cdot\|) = \mathcal{H}$, Lemma 2.2(ii), (2.48), and (2.52) yield

$$\begin{aligned} x &= \gamma p + \big(\sigma + \gamma\varphi^*(p)\big)\nabla\varphi^*(p) \\ &= \gamma p + \big(\sigma + \gamma\phi^*(\|p\|)\big)\nabla(\phi^* \circ \|\cdot\|)(p) \\ &= \left(\gamma + \rho\left(\sigma - \gamma\alpha + \gamma\rho\left(\|p\| - \kappa + \frac{(\|p\| - \kappa)^{q^*}}{q^*}\right)\right)\left(\frac{1 + (\|p\| - \kappa)^{q^*-1}}{\|p\|}\right)\right)p. \end{aligned} \tag{2.56}$$

Hence,

$$p = \frac{1}{\gamma + \rho\left(\sigma - \gamma\alpha + \gamma\rho\left(t - \kappa + \dfrac{(t - \kappa)^{q^*}}{q^*}\right)\right)\left(\dfrac{1 + (t - \kappa)^{q^*-1}}{t}\right)}x, \tag{2.57}$$

where $t = \|p\|$ is the unique solution in $]\kappa, +\infty[$ to (2.46), which is obtained by taking the norm of both sides of (2.56). We then get the conclusion by invoking (2.17).

(iii): In view of (2.55), the assumptions imply that $(\sigma/\gamma, \|x\|/\gamma) \in \Pi + N_{\mathcal{R}}\Pi$ and therefore that $(\chi, \nu) = (\alpha, \kappa)$. Consequently, Proposition 2.3(iii) yields $\operatorname{prox}_{\gamma\widetilde{\varphi}}(\sigma, x) = \big(\sigma - \gamma\alpha, (1 - \gamma\kappa/\|x\|)x\big)$.

(iv): Set $H =]-\infty, \alpha] \times \mathbb{R}$. Then $\mathcal{R} = \mathcal{R} \cap H$ and $(\sigma/\gamma, \|x\|/\gamma) \in]\alpha, +\infty[\times [-\kappa, \kappa]$. Hence, $(\chi, \nu) = \operatorname{proj}_H(\sigma/\gamma, \|x\|/\gamma) = (\alpha, \|x\|/\gamma)$. In turn, we derive from Proposition 2.3(iii) that $\operatorname{prox}_{\gamma\widetilde{\varphi}}(\sigma, x) = (\sigma - \gamma\alpha, 0)$. □

Example 2.7 (standard Berhu function)**.** Let α, γ, and ρ be in $]0, +\infty[$. The standard Berhu function of [29] with shift α is obtained by setting $\mathcal{H} = \mathbb{R}$, $\kappa = 1$, and $q = 2$ in (2.44), that is

$$\mathsf{b}_\rho \colon \mathbb{R} \to \mathbb{R} \colon x \mapsto \begin{cases} \alpha + \dfrac{|x|^2 + \rho^2}{2\rho}, & \text{if } |x| > \rho; \\ \alpha + |x|, & \text{if } |x| \leqslant \rho. \end{cases} \tag{2.58}$$

Now let $\sigma \in \mathbb{R}$ and $x \in \mathbb{R}$. Then we derive from Example 2.6 that

$$\widetilde{\mathsf{b}}_\rho(\sigma, x) = \begin{cases} \alpha\sigma + \dfrac{|x|^2 + \sigma^2\rho^2}{2\rho\sigma}, & \text{if } \sigma > 0 \text{ and } |x| > \sigma\rho; \\ \alpha\sigma + |x|, & \text{if } \sigma > 0 \text{ and } |x| \leqslant \sigma\rho; \\ 0, & \text{if } \sigma = 0 \text{ and } x = 0; \\ +\infty, & \text{otherwise}, \end{cases} \tag{2.59}$$

and that $\operatorname{prox}_{\gamma\widetilde{\mathsf{b}}_\rho}(\sigma, x)$ is given by

$$\begin{cases} (0, 0) & \text{if } \max(|x|^2 - \gamma^2, 0) \leqslant 2\gamma(\gamma\alpha - \sigma)/\rho; \\ (\sigma - \gamma\alpha, 0) & \text{if } \sigma > \gamma\alpha \text{ and } |x| \leqslant \gamma; \\ \big(\sigma - \gamma\alpha, (1 - \gamma/|x|)x\big) & \text{if } \sigma > \gamma\alpha \text{ and } \gamma < |x| \leqslant \gamma + \rho(\sigma - \gamma\alpha); \\ \big(\sigma - \gamma\alpha + \gamma\rho(|p|^2 - 1)/2, x - \gamma p\big) & \text{if } |x| > \gamma + \rho(\sigma - \gamma\alpha) \text{ and} \\ & \quad |x| > \sqrt{\gamma^2 + 2\gamma(\gamma\alpha - \sigma)/\rho}, \end{cases} \tag{2.60}$$

with

$$p = \frac{1}{\gamma + \rho\Big(\sigma - \gamma\alpha + \dfrac{\gamma\rho}{2}\big(t^2 - 1\big)\Big)}\, x, \tag{2.61}$$

where t is the unique solution in $]1, +\infty[$ to the reduced third degree equation

$$t^3 + \left(\frac{2\big(\gamma + \rho(\sigma - \gamma\alpha)\big)}{\gamma\rho^2} - 1\right)t - \frac{2\|x\|}{\gamma\rho^2} = 0, \tag{2.62}$$

which can be solved explicitly via Cardano's formula.

Example 2.8 (abstract Vapnik function)**.** Let α, ε, and γ be in $]0, +\infty[$, and define $\varphi \colon \mathcal{H} \to \mathbb{R}$ by $\varphi = \alpha + \max(\|\cdot\| - \varepsilon, 0)$. Then

$$\widetilde{\varphi} \colon \mathcal{H} \to]-\infty, +\infty] : (\sigma, x) \mapsto \begin{cases} \alpha\sigma + \max(\|x\| - \varepsilon\sigma, 0), & \text{if } \sigma \geqslant 0; \\ +\infty, & \text{if } \sigma < 0. \end{cases} \tag{2.63}$$

Now let $\sigma \in \mathbb{R}$ and $x \in \mathcal{H}$. Then the following hold:

(i) Suppose that $\sigma + \varepsilon\|x\| \leqslant \gamma\alpha$ and $\|x\| \leqslant \gamma$. Then $\operatorname{prox}_{\gamma\widetilde{\varphi}}(\sigma, x) = (0, 0)$.
(ii) Suppose that $\sigma \leqslant \gamma(\alpha - \varepsilon)$ and $\|x\| > \gamma$. Then

$$\operatorname{prox}_{\gamma\widetilde{\varphi}}(\sigma, x) = \left(0, \left(1 - \frac{\gamma}{\|x\|}\right)x\right). \tag{2.64}$$

(iii) Suppose that $\sigma > \gamma(\alpha - \varepsilon)$ and $\|x\| \geqslant \varepsilon\sigma + \gamma(1 + \varepsilon(\varepsilon - \alpha))$. Then

$$\operatorname{prox}_{\gamma\widetilde{\varphi}}(\sigma, x) = \left(\sigma + \gamma(\varepsilon - \alpha), \left(1 - \frac{\gamma}{\|x\|}\right)x\right). \tag{2.65}$$

(iv) Suppose that $\sigma + \varepsilon\|x\| > \gamma\alpha$ and $\varepsilon(\sigma - \gamma\alpha) < \|x\| < \varepsilon\sigma + \gamma(1 + \varepsilon(\varepsilon - \alpha))$. Then

$$\operatorname{prox}_{\gamma\widetilde{\varphi}}(\sigma, x) = \frac{\sigma + \varepsilon\|x\| - \gamma\alpha}{1 + \varepsilon^2}\left(1, \frac{\varepsilon}{\|x\|}\, x\right). \tag{2.66}$$

(v) Suppose that $\sigma \geqslant \gamma\alpha$ and $\|x\| \leqslant \varepsilon(\sigma - \gamma\alpha)$. Then $\operatorname{prox}_{\gamma\widetilde{\varphi}}(\sigma, x) = (\sigma - \gamma\alpha, x)$.

Proof. We derive (2.63) at once from (2.15). Set $\phi = \alpha + \max(|\cdot| - \varepsilon, 0)$. Then $\varphi = \phi \circ \|\cdot\|$ and $\phi = \alpha + d_{[-\varepsilon,\varepsilon]} = \alpha + \iota_{[-\varepsilon,\varepsilon]} \,\square\, |\cdot|$. Therefore

$$\phi^* = \varepsilon|\cdot| + \iota_{[-1,1]} - \alpha. \tag{2.67}$$

Thus, (2.18) yields

$$\mathcal{R} = \mathcal{R}_1 \cap \mathcal{R}_2, \quad \text{where} \quad \begin{cases} \mathcal{R}_1 = \left]-\infty, \alpha\right] \times [-1, 1] \\ \mathcal{R}_2 = \big\{(\chi, \nu) \in \mathbb{R}^2 \;\big|\; \varepsilon|\nu| \leqslant \alpha - \chi\big\}. \end{cases} \tag{2.68}$$

Now set $(\chi, \nu) = \operatorname{proj}_{\mathcal{R}}(\sigma/\gamma, \|x\|/\gamma)$.

(i): This follows from (2.67) and Proposition 2.3(i).

(ii): Since $\sigma/\gamma \leqslant \alpha - \varepsilon$ and $\|x/\gamma\| > 1$, it follows from (2.68) that

$$(\chi, \nu) = \operatorname{proj}_{\mathcal{R}_1}\big(\sigma/\gamma, \|x\|/\gamma\big) = \big(\sigma/\gamma, 1\big). \tag{2.69}$$

In turn, we derive (2.64) from Proposition 2.3(iii).

(iii): The point $\Pi = (\alpha - \varepsilon, 1)$ lies in the intersection of the boundaries of $\mathcal{R}_1$ and $\mathcal{R}_2$, which are line segments. Therefore, the normal cone to $\mathcal{R}$ at Π is generated by outer normals n_1 to $\mathcal{R}_1$ and n_2 to $\mathcal{R}_2$ at Π. A tangent vector to $\mathcal{R}_2$ at Π is $t(\Pi) = (-\varepsilon, 1)$. Therefore we take $n_1 = (0, 1)$ and $n_2 = (1, \varepsilon) \perp t(\Pi)$. Consequently, the set of points which have projection Π onto $\mathcal{R}$ is

$$\begin{aligned} \operatorname{proj}_{\mathcal{R}}^{-1}\{\Pi\} &= \Pi + N_{\mathcal{R}}\Pi \\ &= \Pi + \operatorname{cone}(n_1, n_2) \\ &= \big(\alpha - \varepsilon, 1\big) + \big\{(\tau, \xi) \in \mathbb{R}^2 \;\big|\; \tau \geqslant 0 \;\text{ and } \xi \geqslant \varepsilon\tau\big\} \\ &= \big\{(\tau, \xi) \in \mathbb{R}^2 \;\big|\; \tau \geqslant \alpha - \varepsilon \;\text{ and } \xi \geqslant 1 + \varepsilon(\tau - \alpha + \varepsilon)\big\} \\ &= \big\{(\tau, \xi) \in \mathbb{R}^2 \;\big|\; \tau \geqslant \alpha - \varepsilon \;\text{ and } \xi \geqslant \varepsilon\tau + 1 + \varepsilon(\varepsilon - \alpha)\big\}, \end{aligned} \tag{2.70}$$

and it contains $(\sigma/\gamma, \|x\|/\gamma)$. Hence

$$(\chi,\nu) = \big(\alpha-\varepsilon, 1\big) \quad \Leftrightarrow \quad \begin{cases} \sigma \geqslant \gamma(\alpha-\varepsilon) \\ \|x\| \geqslant \varepsilon\sigma + \gamma\big(1+\varepsilon(\varepsilon-\alpha)\big). \end{cases} \tag{2.71}$$

We then use Proposition 2.3(iii) to get (2.65).

(iv): In this case, $(\chi,\nu) = \operatorname{proj}_{\mathcal{R}_2}(\sigma/\gamma, \|x\|/\gamma)$. More precisely, (χ,ν) is the projection of $(\sigma/\gamma, \|x\|/\gamma)$ onto the half-space $\big\{(\tau,\xi)\in\mathbb{R}^2 \;\big|\; \varepsilon\xi \leqslant \alpha-\tau\big\} = \big\{(\tau,\xi)\in\mathbb{R}^2 \;\big|\; \langle (\tau,\xi) \mid n_2\rangle \leqslant \alpha\big\}$, where $n_2 = (1,\varepsilon)$. Thus,

$$\begin{aligned} (\chi,\nu) &= \frac{1}{\gamma}(\sigma,\|x\|) + \frac{\alpha - \langle (\sigma,\|x\|) \mid n_2\rangle/\gamma}{\|n_2\|^2} n_2 \\ &= \left(\frac{\sigma}{\gamma} + \frac{\alpha-(\sigma+\varepsilon\|x\|)/\gamma}{1+\varepsilon^2}, \frac{\|x\|}{\gamma} + \varepsilon\frac{\alpha-(\sigma+\varepsilon\|x\|)/\gamma}{1+\varepsilon^2}\right), \end{aligned} \tag{2.72}$$

and (2.66) follows from Proposition 2.3(iii).

(v): Set $\Pi = (\alpha, 0)$, $n_2 = (1,\varepsilon)$, and $n_3 = (1,-\varepsilon)$. The set of points which have projection Π onto $\mathcal{R}$ is

$$\begin{aligned} \Pi + N_{\mathcal{R}}\Pi &= \Pi + N_{\mathcal{R}_2}\Pi \\ &= \Pi + \operatorname{cone}(n_2, n_3) \\ &= (\alpha,0) + \big\{(\tau,\xi)\in\mathbb{R}^2 \;\big|\; \tau \geqslant 0 \;\text{ and }\; \xi \leqslant \varepsilon\tau\big\} \\ &= \big\{(\tau,\xi)\in\mathbb{R}^2 \;\big|\; \tau \geqslant \alpha \;\text{ and }\; \xi \leqslant \varepsilon(\tau-\alpha)\big\}, \end{aligned} \tag{2.73}$$

and it therefore contains $(\sigma/\gamma, \|x\|/\gamma)$. In turn, $(\chi,\nu) = (\alpha,0)$ and the conclusion follows from Proposition 2.3(iii). □

3. Optimization model and examples

Let us first recall that our data formation model is Model 1.1. We now introduce our perspective M-estimation model, which enables the simultaneous estimation of the regression vector $\overline{b} = (\overline{\beta}_k)_{1\leqslant k\leqslant p} \in \mathbb{R}^p$ as well as scale vectors $\overline{s} = (\overline{\sigma}_i)_{1\leqslant i\leqslant N} \in \mathbb{R}^N$ and $\overline{t} = (\overline{\tau}_i)_{1\leqslant i\leqslant P} \in \mathbb{R}^P$. If robust data fitting functions are used, the outlier vector in Model 1.1 can be identified from the solution of (3.2) below. For instance, if the Huber function is used for data fitting, one can estimate the mean shift vector $\overline{o}$ in (1.1) [1, 33].

The proposed perspective M-estimation optimization problem is as follows.

Problem 3.1. Let N and P be strictly positive integers, let $\varsigma \in \Gamma_0(\mathbb{R}^N)$, let $\varpi \in \Gamma_0(\mathbb{R}^P)$, let $\theta \in \Gamma_0(\mathbb{R}^p)$, let $(n_i)_{1\leqslant i\leqslant N}$ be strictly positive integers such that $\sum_{i=1}^N n_i = n$, and let $(p_i)_{1\leqslant i\leqslant P}$ be strictly positive integers. For every $i\in\{1,\ldots,N\}$, let $\varphi_i \in \Gamma_0(\mathbb{R}^{n_i})$, let $X_i \in \mathbb{R}^{n_i\times p}$, and let $y_i \in \mathbb{R}^{n_i}$ be such that

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_N \end{bmatrix} \quad \text{and} \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}. \tag{3.1}$$

Finally, for every $i \in \{1, \dots, P\}$, let $\psi_i \in \Gamma_0(\mathbb{R}^{p_i})$, and let $L_i \in \mathbb{R}^{p_i \times p}$. The objective of perspective M-estimation is to

$$\underset{s \in \mathbb{R}^N,\, t \in \mathbb{R}^P,\, b \in \mathbb{R}^p}{\text{minimize}} \quad \varsigma(s) + \varpi(t) + \theta(b) + \sum_{i=1}^{N} \widetilde{\varphi}_i\big(\sigma_i, X_i b - y_i\big) + \sum_{i=1}^{P} \widetilde{\psi}_i\big(\tau_i, L_i b\big). \tag{3.2}$$

Remark 3.2. Let us make a few observations about Problem 3.1.

(i) In (3.2), $N+P$ perspective functions $(\widetilde{\varphi}_i)_{1 \leqslant i \leqslant N}$ and $(\widetilde{\psi}_i)_{1 \leqslant i \leqslant P}$ are used to penalize affine transformations $(X_i b - y_i)_{1 \leqslant i \leqslant N}$ and $(L_i b)_{1 \leqslant i \leqslant P}$ of b. The operators $(L_i)_{1 \leqslant i \leqslant P}$ can, for instance, select a single coordinate, or blocks of coordinates (as in the group lasso penalty), or can model finite difference operators. Constraints on the scale variables $(\sigma_i)_{1 \leqslant i \leqslant N}$ and $(\tau_i)_{1 \leqslant i \leqslant P}$ of the perspective functions can be enforced via the functions ς and ϖ.

(ii) It is also possible to use "scaleless" non-perspective functions of the variables $(X_i b - y_i)_{1 \leqslant i \leqslant N}$ and $(L_i b)_{1 \leqslant i \leqslant P}$. For instance, given $i \in \{1, \dots, N\}$, the term $\varphi_i(X_i b - y_i)$ is obtained by using $\widetilde{\varphi}_i(\sigma_i, X_i b - y_i)$ and imposing $\sigma_i = 1$ via ς.

(iii) We attach individual scale variables to each of the functions $(\widetilde{\varphi}_i)_{1 \leqslant i \leqslant N}$ and $(\widetilde{\psi}_i)_{1 \leqslant i \leqslant P}$ for flexibility in the case of heteroscedastic models, but also for computational reasons. Indeed, the proximal tools we are proposing in Sections 4 and 5 can handle separable functions better. For instance, it is hard to process the function

$$(\sigma, x_1, x_2) \mapsto \widetilde{\varphi}_1(\sigma, x_1) + \widetilde{\varphi}_2(\sigma, x_2) \tag{3.3}$$

via proximal tools, whereas the equivalent separable function with coupling of the scales

$$(\sigma_1, \sigma_2, x_1, x_2) \mapsto \varsigma(\sigma_1, \sigma_2) + \widetilde{\varphi}_1(\sigma_1, x_1) + \widetilde{\varphi}_2(\sigma_2, x_2),$$
$$\text{where} \quad \varsigma(\sigma_1, \sigma_2) = \begin{cases} 0, & \text{if } \sigma_1 = \sigma_2; \\ +\infty, & \text{if } \sigma_1 \neq \sigma_2, \end{cases} \tag{3.4}$$

will be much easier.

We now present some important instantiations of Problem 3.1.

Example 3.3. Consider the optimization problem

$$\underset{b \in \mathbb{R}^p}{\text{minimize}} \quad \|Xb - y\|_q^q + \alpha_1 \|b\|_1 + \alpha_2 \|b\|_r^r, \tag{3.5}$$

where $\alpha_1 \in [0, +\infty[$, $\alpha_2 \in [0, +\infty[$, $q \in \{1, 2\}$, and $r \in [1, 2]$. For $q = r = 2$, $\alpha_1 > 0$, and $\alpha_2 > 0$, (3.5) is the elastic-net model of [42]; in addition, if $\alpha_1 = 0$ and $\alpha_2 > 0$, we obtain the ridge regression model [19] and, if $\alpha_1 > 0$ and $\alpha_2 = 0$, we obtain the lasso model [36]. On the other hand, taking $q = 1$, $\alpha_1 > 0$, and $\alpha_2 = 0$, leads to the least absolute deviation lasso model of [39]. Finally, taking

$q = 2$, $\alpha_1 = 0$, and $\alpha_2 > 0$ yields to the bridge model [16]. The formulation (3.5) corresponds to the special case of Problem 3.1 in which

$$\begin{cases} N = 1,\ n_1 = n,\ \varphi_1 = \|\cdot\|_q^q \\ P = 1,\ p_1 = p,\ \psi_1 = 0,\ L_1 = 0 \\ \varsigma = \iota_{\{1\}},\ \varpi = 0,\ \theta = \alpha_1\|\cdot\|_1 + \alpha_2\|\cdot\|_r^r. \end{cases} \tag{3.6}$$

Note that our choice of ς imposes that $\sigma_1 = 1$ and therefore that $\widetilde{\varphi}_1(\sigma_1, \cdot) = \|\cdot\|_q^q$. The proximity operator of φ_1 is derived in [7] and that of θ in [11].

Example 3.4. Given α_1 and α_2 in $[0, +\infty[$ and $q \in \{1, 2\}$, consider the model

$$\underset{b\in\mathbb{R}^p}{\text{minimize}}\ \|Xb - y\|_2^2 + \alpha_1 \sum_{i=1}^{p} |\beta_i| + \alpha_2 \sum_{i=1}^{p-1} |\beta_{i+1} - \beta_i|^q. \tag{3.7}$$

It derives from Problem 3.1 by setting

$$\begin{cases} N = 1,\ n_1 = n,\ \varphi_1 = \|\cdot\|_2^2 \\ P = p - 1,\ (\forall i \in \{1, \ldots, P\})\ p_i = 1,\ \psi_i = \alpha_2|\cdot|^q,\ L_i \colon b \mapsto \beta_{i+1} - \beta_i \\ \varsigma = \iota_{\{1\}},\ \varpi = \iota_{\{(1,\ldots,1)\}},\ \theta = \alpha_1\|\cdot\|_1. \end{cases} \tag{3.8}$$

For $q = 1$, we obtain the fused lasso model [38], while $q = 2$ yields the smooth lasso formulation of [18]. Let us note that one obtains alternative formulations such that of [37] by suitably redefining the operators $(L_i)_{1\leqslant i\leqslant P}$ in (3.8).

Example 3.5. Given ρ_1 and ρ_2 in $]0, +\infty[$, the formulation proposed in [29] is

$$\underset{\sigma\in]0,+\infty[,\,\tau\in]0,+\infty[,\,b\in\mathbb{R}^p}{\text{minimize}}\ \sigma \sum_{i=1}^{n} \mathsf{h}_{\rho_1}\left(\frac{x_i^\top b - \eta_i}{\sigma}\right) + n\sigma + \alpha_1\tau \sum_{i=1}^{p} \mathsf{b}_{\rho_2}\left(\frac{\beta_i}{\tau}\right) + p\tau, \tag{3.9}$$

where h_{ρ_1} and b_{ρ_2} are the Huber and Berhu functions of (2.30) and (2.59), respectively. From a convex optimization viewpoint, we reformulate this problem more formally in terms of the lower semicontinuous function of (2.15) to obtain

$$\underset{\sigma\in\mathbb{R},\,\tau\in\mathbb{R},\,b\in\mathbb{R}^p}{\text{minimize}}\ \sum_{i=1}^{n} [\mathsf{h}_{\rho_1} + n]^\sim\big(\sigma, x_i^\top b - \eta_i\big) + \alpha_1 \sum_{i=1}^{p} [\mathsf{b}_{\rho_2} + p]^\sim\big(\tau, \beta_i\big). \tag{3.10}$$

This is a special case of Problem 3.1 with

$$\begin{cases} N = n \ \text{ and } \ (\forall i \in \{1, \ldots, N\})\ n_i = 1,\ \varphi_i = \mathsf{h}_{\rho_1} + n,\ X_i = x_i^\top \\ P = p \ \text{ and } \ (\forall i \in \{1, \ldots, P\})\ p_i = 1,\ \psi_i = \alpha_1\mathsf{b}_{\rho_2} + p,\ L_i \colon b \mapsto \beta_i \\ \varsigma = \iota_D, \ \text{ where } \ D = \big\{(\sigma, \ldots, \sigma) \in \mathbb{R}^n \;\big|\; \sigma \in \mathbb{R}\big\} \\ \varpi = \iota_E, \ \text{ where } \ E = \big\{(\tau, \ldots, \tau) \in \mathbb{R}^p \;\big|\; \tau \in \mathbb{R}\big\} \\ \theta = 0. \end{cases} \tag{3.11}$$

If one omits the right-most summation in (3.10) one recovers Huber's concomitant model [21]. Note that

$$\operatorname{prox}_{\varsigma} = \operatorname{proj}_D \colon (\sigma_i)_{1\leqslant i\leqslant n} \mapsto \left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i, \ldots, \frac{1}{n}\sum_{i=1}^{n}\sigma_i\right). \tag{3.12}$$

The operator $\operatorname{prox}_{\varpi}$ is computed likewise. On the other hand, the proximity operators of h_{ρ_1} and b_{ρ_2} are provided in Examples 2.5 and 2.6, respectively.

Example 3.6. The scaled square-root elastic net formulation of [30] is

$$\underset{\sigma\in]0,+\infty[,\, b\in\mathbb{R}^p}{\operatorname{minimize}} \quad \frac{\|Xb-y\|_2^2}{2\sigma} + \frac{n\sigma}{2} + \alpha_1\|b\|_1 + \alpha_2\|b\|_2^q, \tag{3.13}$$

where $\alpha_1 \in [0,+\infty[$, $\alpha_2 \in [0,+\infty[$, and $q \in \{1,2\}$. Reformulated more formally in terms of lower semicontinuous functions, this model becomes

$$\underset{\sigma\in\mathbb{R},\, b\in\mathbb{R}^p}{\operatorname{minimize}} \quad \left[\frac{\|\cdot\|_2^2+n}{2}\right]^{\sim}(\sigma, Xb-y) + \alpha_1\|b\|_1 + \alpha_2\|b\|_2^q. \tag{3.14}$$

We thus obtain the special case of Problem 3.1 in which

$$\begin{cases} N=1,\ n_1=n,\ \varphi_1 = \big(\|\cdot\|_2^2+n\big)/2 \\ P=p \ \text{ and } \ (\forall i\in\{1,\ldots,P\}) \ \ p_i=1,\ \psi_i=\alpha_1|\cdot|,\ L_i\colon b\mapsto\beta_i \\ \varsigma=0,\ \varpi=0,\ \theta=\alpha_2\|b\|_2^q. \end{cases} \tag{3.15}$$

The proximity operator of θ is given in [13], while that of $\widetilde{\varphi}_1$ is provided in Example 2.4. Note that, when $q=2$, we could also take the functions $(\psi_i)_{1\leqslant i\leqslant P}$ to be zero and $\theta = \alpha_1\|b\|_1+\alpha_2\|b\|_2^2$ since the proximity operator of θ is computable explicitly in this case [11]. When $\alpha_2=0$ in (3.14), we obtain the scaled lasso model [2, 35]. On the other hand, if we use $\alpha_2=0$ and $\varsigma = \iota_{[\varepsilon,+\infty[}$ for some $\varepsilon\in\,]0,+\infty[$ in (3.14), we recover the formulation of [28].

Example 3.7. Given α, ρ_1, ρ_2, and $(\omega_i)_{1\leqslant i\leqslant p}$ in $]0,+\infty[$, the formulation proposed in [23] is

$$\begin{aligned} \underset{\sigma\in\mathbb{R},\,\tau\in\mathbb{R},\, b\in\mathbb{R}^p}{\operatorname{minimize}} \quad & \begin{cases} \sigma\displaystyle\sum_{i=1}^{n}\mathsf{h}_{\rho_1}\left(\frac{x_i^\top b-\eta_i}{\sigma}\right)+n\sigma, & \text{if } \ \sigma>0; \\ \rho_1\displaystyle\sum_{i=1}^{n}\big|x_i^\top b-\eta_i\big|, & \text{if } \ \sigma=0; \\ +\infty, & \text{if } \ \sigma<0 \end{cases} \\ & + \begin{cases} \alpha\tau\displaystyle\sum_{i=1}^{p}\left(\omega_i\mathsf{b}_{\rho_2}\left(\frac{\beta_i}{\tau}\right)+\frac{1}{\omega_i}\right), & \text{if } \ \tau>0; \\ 0, & \text{if } \ b=0 \text{ and } \tau=0; \\ +\infty, & \text{otherwise,} \end{cases} \end{aligned} \tag{3.16}$$

where h_{ρ_1} and b_{ρ_2} are the Huber and Berhu functions of (2.30) and (2.59), respectively. In view of (2.15), we can rewrite (3.16) as

$$\underset{\sigma\in\mathbb{R},\,\tau\in\mathbb{R},\,b\in\mathbb{R}^p}{\operatorname{minimize}} \sum_{i=1}^{n}[\mathsf{h}_{\rho_1}+n]^{\sim}\big(\sigma,x_i^\top b-\eta_i\big)+\alpha\sum_{i=1}^{p}\bigg[\omega_i\mathsf{b}_{\rho_2}+\frac{1}{\omega_i}\bigg]^{\sim}\big(\tau,\beta_i\big). \tag{3.17}$$

This is a special case of Problem 3.1 with

$$\begin{cases} N=n \;\text{ and }\; (\forall i\in\{1,\dots,N\})\;\; n_i=1,\;\varphi_i=\mathsf{h}_{\rho_1}+n,\;X_i=x_i^\top\\ P=p \;\text{ and }\; (\forall i\in\{1,\dots,P\})\;\; p_i=1,\;\psi_i=\alpha\big(\omega_i\mathsf{b}_{\rho_2}+1/\omega_i\big),\;L_i\colon b\mapsto\beta_i\\ \varsigma=\iota_D,\;\text{ where }\; D=\big\{(\sigma,\dots,\sigma)\in\mathbb{R}^n\;\big|\;\sigma\in\mathbb{R}\big\}\\ \varpi=\iota_E,\;\text{ where }\; E=\big\{(\tau,\dots,\tau)\in\mathbb{R}^p\;\big|\;\tau\in\mathbb{R}\big\}\\ \theta=0. \end{cases} \tag{3.18}$$

In the variant studied in [22], the functions $(\psi_i)_{1\leqslant i\leqslant p}$ of (3.18) are replaced by $(\forall i\in\{1,\dots,p\})\;\psi_i=\alpha\omega_i|\beta_i|$.

Example 3.8. Let $\alpha\in\left]0,+\infty\right[$. The formulation

$$\underset{\tau\in\left]0,+\infty\right[,\,b\in\mathbb{R}^p}{\operatorname{minimize}} \;-\frac{\ln\tau}{2}+\frac{\|y\|_2^2\,\tau}{2n}+\frac{\|Xb\|_2^2}{2n\tau}+\alpha\|b\|_1-\frac{y^\top Xb}{n}, \tag{3.19}$$

was proposed in [40] under the name "natural lasso." It can be cast in the framework of Problem 3.1 with

$$\begin{cases} N=1,\;n_1=n,\;\varphi_1=0\\ P=1,\;p_1=p,\;\psi_1=\|\cdot\|_2^2/(2n),\;L_1=X\\ \varsigma=0,\;\theta=\alpha\|\cdot\|_1-\langle X^\top y\mid\cdot\rangle/n \end{cases} \tag{3.20}$$

and

$$\varpi\colon\tau\mapsto\begin{cases}-(\ln\tau)/2+\|y\|_2^2\tau/(2n), & \text{if }\;\tau>0;\\ +\infty, & \text{if }\;\tau\leqslant 0.\end{cases} \tag{3.21}$$

The proximity operators of θ and ϖ are given in [13].

Example 3.9. Given α and ε in $\left]0,+\infty\right[$, define $\mathsf{v}_i\colon\mathbb{R}\to\mathbb{R}\colon\eta\mapsto\alpha+\max(|\eta|-\varepsilon,0)$. Using the perspective function derived in Example 2.8, we can rewrite the linear ν-support vector regression problem of [32] as

$$\underset{\sigma\in\mathbb{R},\,b\in\mathbb{R}^p}{\operatorname{minimize}} \sum_{i=1}^{n}\widetilde{\mathsf{v}}_i\big(\sigma,x_i^\top b-\eta_i\big)+\frac{1}{2}\|b\|_2^2. \tag{3.22}$$

We identify this problem as a special case of Problem 3.1 with

$$\begin{cases} N=n \;\text{ and }\; (\forall i\in\{1,\dots,N\})\;\;\varphi_i=\mathsf{v}_i,\;X_i=x_i^\top\\ P=1,\;p_1=p,\;\psi_1=0,\;L_1=0\\ \varsigma=\iota_D,\;\text{ where }\; D=\big\{(\sigma,\dots,\sigma)\in\mathbb{R}^n\;\big|\;\sigma\in\mathbb{R}\big\}\\ \varpi=0,\;\theta=\|\cdot\|_2^2/2. \end{cases} \tag{3.23}$$

The proximity operator of $\widetilde{\mathsf{v}}_i$ is given in Example 2.8 and that of ς in (3.12). The concomitant parameter σ scales the width of the "tube" in the ν-support vector regression and trades off model complexity and slack variables [32].

The next two examples are novel M-estimators that will be employed in Section 5.

Example 3.10. In connection with (3.1), we introduce a generalized heteroscedastic scaled lasso with N data blocks, which employs the perspective derived in Example 2.4. Recall that n_i is the number of data points in the ith block, let $\alpha_1 \in [0, +\infty[$, and set

$$(\forall i \in \{1, \dots, N\}) \quad \mathsf{c}_{i,q} \colon \mathbb{R}^{n_i} \to \mathbb{R} \colon x \mapsto \|x\|_2^q + \frac{1}{2}. \tag{3.24}$$

The objective is to

$$\underset{s \in \mathbb{R}^N,\, b \in \mathbb{R}^p}{\text{minimize}} \quad \sum_{i=1}^{N} \widetilde{\mathsf{c}}_{i,q}\big(\sigma_i, X_i b - y_i\big) + \alpha_1 \|b\|_1. \tag{3.25}$$

This is a special case of Problem 3.1 with

$$\begin{cases} (\forall i \in \{1, \dots, N\}) \;\; \varphi_i = \mathsf{c}_{i,q} \\ P = 1,\; p_1 = p,\; \psi_1 = 0,\; L_1 = 0 \\ \varpi = 0,\; \varsigma = 0,\; \theta = \alpha_1 \|\cdot\|_1. \end{cases} \tag{3.26}$$

The choice of the exponent $q \in \,]1, +\infty[$ reflects prior distributional assumptions on the noise. This model can handle generalized normal distributions. The proximity operator of $\widetilde{\mathsf{c}}_{i,q}$ is provided in Example 2.4.

Example 3.11. In connection with (3.1), we introduce a generalized heteroscedastic Huber M-estimator, with J scale variables $(\sigma_j)_{1 \leqslant j \leqslant J}$, which employs the perspective derived in Example 2.5. Each scale σ_j is attached to a group of m_j data points, hence $\sum_{j=1}^{J} m_j = n$. Let α_1 and α_2 be in $[0, +\infty[$, let δ, ρ_1, and ρ_2 be in $]0, +\infty[$, and denote by $\mathsf{h}_{\rho_1,q}$ the function in (2.31), where $\mathcal{H} = \mathbb{R}$. The objective is to

$$\underset{s \in \mathbb{R}^J,\, \tau \in \mathbb{R},\, b \in \mathbb{R}^p}{\text{minimize}} \quad \sum_{j=1}^{J} \sum_{i=1}^{m_j} [\mathsf{h}_{\rho_1,q} + \delta]^{\sim}\big(\sigma_j, x_i^\top b - \eta_i\big) + \alpha_1 \|b\|_1 + \alpha_2 \sum_{i=1}^{p} [\mathsf{b}_{\rho_2} + p]^{\sim}\big(\tau, \beta_i\big). \tag{3.27}$$

This statistical model is rewritten in the format of the computational model described in Problem 3.1 by choosing

$$\begin{cases} N = n \;\text{ and }\; (\forall i \in \{1, \dots, N\}) \;\; n_i = 1,\; \varphi_i = \mathsf{h}_{\rho_1,q} + \delta,\; X_i = x_i^\top \\ P = p \;\text{ and }\; (\forall i \in \{1, \dots, P\}) \;\; p_i = 1,\; \psi_i = \alpha_2 \mathsf{b}_{\rho_2} + p,\; L_i \colon b \mapsto \beta_i \\ \varsigma = \iota_D, \;\text{ where }\; D = \big\{(\sigma_1, \dots, \sigma_1, \dots, \sigma_J, \dots, \sigma_J) \in \mathbb{R}^n \;\big|\; (\sigma_j)_{1 \leqslant j \leqslant J} \in \mathbb{R}^J\big\} \\ \varpi = \iota_E, \;\text{ where }\; E = \big\{(\tau, \dots, \tau) \in \mathbb{R}^p \;\big|\; \tau \in \mathbb{R}\big\} \\ \theta = \alpha_1 \|\cdot\|_1. \end{cases} \tag{3.28}$$

The choice of the exponent $q \in]1,+\infty[$ reflects prior distributional assumptions on the noise. This model handles generalized normal distributions and can identify outliers. Note that

$$\operatorname{prox}_{\varsigma} = \operatorname{proj}_{D} \colon (\sigma_i)_{1\leqslant i\leqslant n} \mapsto \Bigg(\underbrace{\frac{1}{m_1}\sum_{i=1}^{m_1}\sigma_i, \dots, \frac{1}{m_1}\sum_{i=1}^{m_1}\sigma_i}_{m_1 \text{ terms}}, \dots, \underbrace{\frac{1}{m_J}\sum_{i=n-m_J+1}^{n}\sigma_i, \dots, \frac{1}{m_J}\sum_{i=n-m_J+1}^{n}\sigma_i}_{m_J \text{ terms}} \Bigg). \tag{3.29}$$

Remark 3.12. Particular instances of perspective M-estimation models come with statistical guarantees. For the scaled lasso, initial theoretical guarantees are given in [35]. In [22, 23] results are provided for the homoscedastic Huber M-estimator with adaptive ℓ^1 penalty and the adaptive Berhu penalty. In [17], explicit bounds for estimation and prediction error for "convex loss lasso" problems are given which cover scaled homoscedastic lasso, the least absolute deviation model, and the homoscedastic Huber model. For the heteroscedastic M-estimators we have presented above, statistical guarantees are, to the best of our knowledge, elusive.

4. Algorithm

Recall from (3.2) that the problem of perspective M-estimation is to

$$\underset{s\in\mathbb{R}^N,\, t\in\mathbb{R}^P,\, b\in\mathbb{R}^p}{\operatorname{minimize}} \quad \varsigma(s)+\varpi(t)+\theta(b)+\sum_{i=1}^{N}\widetilde{\varphi}_i\big(\sigma_i, X_i b - y_i\big) + \sum_{i=1}^{P}\widetilde{\psi}_i\big(\tau_i, L_i b\big). \tag{4.1}$$

This minimization problem is quite complex, as it involves the sum of several terms, compositions with linear operators, as well as perspective functions. In addition, none of the functions present in the model is assumed to have any full domain or smoothness property. In this section, we reformulate (4.1) in a suitable higher dimensional product space through a series of reparametrizations. The resulting reformulation is shown to be solvable by the Douglas-Rachford splitting algorithm. Once reformulated in the original scale/regression space, this algorithm yields a new proximal splitting method which requires only to use separately the proximity operators of the functions ς, ϖ, θ, $(\widetilde{\varphi}_i)_{1\leqslant i\leqslant N}$, and $(\widetilde{\psi}_i)_{1\leqslant i\leqslant P}$, as well as application of simple linear transformations. It will be shown to produce sequences $(s_k)_{k\in\mathbb{N}}$, $(t_k)_{k\in\mathbb{N}}$, and $(b_k)_{k\in\mathbb{N}}$ which converge respectively to vectors s, t, and b that solve (4.1).

Let us set $\varrho \colon \mathbb{R}^N \times \mathbb{R}^P \to]-\infty,+\infty] : (s,t) \mapsto \varsigma(s)+\varpi(t)$, $M = N+P$, and

$$(\forall i \in \{1,\dots,N\}) \begin{cases} \vartheta_i = \varphi_i \\ m_i = n_i \\ w_i = y_i \\ A_i = X_i \end{cases} \quad \text{and} \quad (\forall i \in \{N+1,\dots,M\}) \begin{cases} \vartheta_i = \psi_{i-N} \\ m_i = p_{i-N} \\ w_i = 0 \\ A_i = L_{i-N}. \end{cases} \tag{4.2}$$

Then, upon introducing the variable $v = (s,t) = (\nu_i)_{1\leqslant i\leqslant M} \in \mathbb{R}^M$, we can rewrite (4.1) as

$$\underset{v\in\mathbb{R}^M,\, b\in\mathbb{R}^p}{\text{minimize}} \;\; \varrho(v) + \theta(b) + \sum_{i=1}^{M} \widetilde{\vartheta}_i\big(\nu_i, A_i b - w_i\big). \tag{4.3}$$

Now let us set $m = n + p$ and define

$$A = \begin{bmatrix} A_1 \\ \vdots \\ A_M \end{bmatrix} \tag{4.4}$$

and

$$\begin{cases} \boldsymbol{f}\colon & \mathbb{R}^M \times \mathbb{R}^p & \to &]-\infty,+\infty] \\ & (v,b) & \mapsto & \varrho(v) + \theta(b) \\ \boldsymbol{g}\colon & \mathbb{R}^M \times \mathbb{R}^m & \to &]-\infty,+\infty] \\ & \big(v,z\big) & \mapsto & \displaystyle\sum_{i=1}^{M} \widetilde{\vartheta}_i(\nu_i, z_i - w_i) \\ \boldsymbol{L}\colon & \mathbb{R}^M \times \mathbb{R}^p & \to & \mathbb{R}^M \times \mathbb{R}^m \\ & (v,b) & \mapsto & \big(v, Ab\big). \end{cases} \tag{4.5}$$

Then, upon introducing the variable $\boldsymbol{a} = (v,b) \in \mathbb{R}^M \times \mathbb{R}^p$, (4.3) can be rewritten as

$$\underset{\boldsymbol{a}\in\mathbb{R}^{M+p}}{\text{minimize}} \;\; \boldsymbol{f}(\boldsymbol{a}) + \boldsymbol{g}(\boldsymbol{L}\boldsymbol{a}), \tag{4.6}$$

which we can solve by various algorithms [5, 9]. Following an approach used in [10] and [12], we reformulate (4.6) as a problem involving the sum of two functions $\boldsymbol{F}$ and $\boldsymbol{G}$, and then solve it via the Douglas-Rachford algorithm [4, 15, 24]. To this end, define

$$\boldsymbol{F}\colon \mathbb{R}^{M+p} \times \mathbb{R}^{M+m} \to]-\infty,+\infty] : (\boldsymbol{a},\boldsymbol{c}) \mapsto \boldsymbol{f}(\boldsymbol{a}) + \boldsymbol{g}(\boldsymbol{c}) \tag{4.7}$$

and

$$\boldsymbol{G} = \iota_{\boldsymbol{V}}, \quad \text{where} \quad \boldsymbol{V} = \big\{(\boldsymbol{x},\boldsymbol{h}) \in \mathbb{R}^{M+p} \times \mathbb{R}^{M+m} \;\big|\; \boldsymbol{L}\boldsymbol{x} = \boldsymbol{h}\big\} \tag{4.8}$$

is the graph of $\boldsymbol{L}$. Then, in terms of the variable $\boldsymbol{u} = (\boldsymbol{a},\boldsymbol{c})$, (4.6) is equivalent to

$$\underset{\boldsymbol{u}\in\mathbb{R}^{2M+m+p}}{\text{minimize}} \;\; \boldsymbol{F}(\boldsymbol{u}) + \boldsymbol{G}(\boldsymbol{u}). \tag{4.9}$$

Let $\gamma \in]0,+\infty[$, let $\boldsymbol{v}_0 \in \mathbb{R}^{2M+p+m}$, and let $(\mu_k)_{k\in\mathbb{N}}$ be a sequence in $]0,2[$ such that $\sum_{k\in\mathbb{N}} \mu_k(2-\mu_k) = +\infty$. The Douglas-Rachford algorithm for solving (4.9) is [4, Section 28.3]

$$\begin{array}{l} \text{for } k = 0, 1, \ldots \\ \quad \left\lfloor \begin{array}{l} \boldsymbol{u}_k = \operatorname{prox}_{\gamma\boldsymbol{G}} \boldsymbol{v}_k \\ \boldsymbol{w}_k = \operatorname{prox}_{\gamma\boldsymbol{F}}(2\boldsymbol{u}_k - \boldsymbol{v}_k) \\ \boldsymbol{v}_{k+1} = \boldsymbol{v}_k + \mu_k(\boldsymbol{w}_k - \boldsymbol{u}_k). \end{array} \right. \end{array} \tag{4.10}$$

Under the qualification condition

$$\boldsymbol{V} \cap \operatorname{ri} \operatorname{dom} \boldsymbol{F} \neq \varnothing, \tag{4.11}$$

the sequence $(\boldsymbol{u}_k)_{k\in\mathbb{N}}$ is guaranteed to converge to a solution $\boldsymbol{u}$ to (4.9) [4, Corollary 27.4]. To make this algorithm more explicit, we first use (4.7) and [4, Proposition 24.11] to obtain

$$\operatorname{prox}_{\boldsymbol{F}} \colon (\boldsymbol{a}, \boldsymbol{c}) \mapsto \big(\operatorname{prox}_{\boldsymbol{f}} \boldsymbol{a}, \operatorname{prox}_{\boldsymbol{g}} \boldsymbol{c}\big). \tag{4.12}$$

Next, we derive from (4.8) that $\operatorname{prox}_{\boldsymbol{G}}$ is the projection operator onto $\boldsymbol{V}$, that is [4, Example 29.19(i)],

$$\operatorname{prox}_{\boldsymbol{G}} \colon (\boldsymbol{x}, \boldsymbol{h}) \mapsto (\boldsymbol{a}, \boldsymbol{L}\boldsymbol{a}), \quad \text{where } \boldsymbol{a} = \boldsymbol{x} - \boldsymbol{L}^\top \big(\mathbf{Id} + \boldsymbol{L}\boldsymbol{L}^\top\big)^{-1} (\boldsymbol{L}\boldsymbol{x} - \boldsymbol{h}). \tag{4.13}$$

Therefore, using the notation

$$\boldsymbol{R} = \boldsymbol{L}^\top (\mathbf{Id} + \boldsymbol{L}\boldsymbol{L}^\top)^{-1} \quad \text{and} \quad (\forall k \in \mathbb{N}) \quad \begin{cases} \boldsymbol{u}_k = (\boldsymbol{a}_k, \boldsymbol{c}_k) \\ \boldsymbol{v}_k = (\boldsymbol{x}_k, \boldsymbol{h}_k) \\ \boldsymbol{w}_k = (\boldsymbol{z}_k, \boldsymbol{d}_k), \end{cases} \tag{4.14}$$

we see that, given some initial points $\boldsymbol{x}_0 \in \mathbb{R}^{M+p}$ and $\boldsymbol{h}_0 \in \mathbb{R}^{m+M}$, (4.10) amounts to iterating

$$\begin{aligned} &\text{for } k = 0, 1, \ldots \\ &\quad \boldsymbol{q}_k = \boldsymbol{L}\boldsymbol{x}_k - \boldsymbol{h}_k \\ &\quad \boldsymbol{a}_k = \boldsymbol{x}_k - \boldsymbol{R}\boldsymbol{q}_k \\ &\quad \boldsymbol{c}_k = \boldsymbol{L}\boldsymbol{a}_k \\ &\quad \boldsymbol{z}_k = \operatorname{prox}_{\gamma \boldsymbol{f}}(2\boldsymbol{a}_k - \boldsymbol{x}_k) \\ &\quad \boldsymbol{d}_k = \operatorname{prox}_{\gamma \boldsymbol{g}}(2\boldsymbol{c}_k - \boldsymbol{h}_k) \\ &\quad \boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \mu_k(\boldsymbol{z}_k - \boldsymbol{a}_k) \\ &\quad \boldsymbol{h}_{k+1} = \boldsymbol{h}_k + \mu_k(\boldsymbol{d}_k - \boldsymbol{c}_k). \end{aligned} \tag{4.15}$$

In addition, it generates a sequence $(\boldsymbol{a}_k)_{k\in\mathbb{N}}$ that converges to a solution $\boldsymbol{a}$ to (4.6). Now set

$$(\forall k \in \mathbb{N}) \quad \begin{cases} \boldsymbol{a}_k = (s_k, t_k, b_k) \in \mathbb{R}^N \times \mathbb{R}^P \times \mathbb{R}^p \\ \boldsymbol{c}_k = (s_k, t_k, c_{b,k}) \in \mathbb{R}^N \times \mathbb{R}^P \times \mathbb{R}^{n+p} \\ \boldsymbol{x}_k = (x_{s,k}, x_{t,k}, x_{b,k}) \in \mathbb{R}^N \times \mathbb{R}^P \times \mathbb{R}^p \\ \boldsymbol{h}_k = (h_{s,k}, h_{t,k}, h_{b,k}) \in \mathbb{R}^N \times \mathbb{R}^P \times \mathbb{R}^{n+p} \\ \boldsymbol{z}_k = (z_{s,k}, z_{t,k}, z_{b,k}) \in \mathbb{R}^N \times \mathbb{R}^P \times \mathbb{R}^p \\ \boldsymbol{d}_k = (d_{s,k}, d_{t,k}, d_{b,k}) \in \mathbb{R}^N \times \mathbb{R}^P \times \mathbb{R}^{n+p} \\ \boldsymbol{q}_k = (q_{s,k}, q_{t,k}, q_{b,k}) \in \mathbb{R}^N \times \mathbb{R}^P \times \mathbb{R}^{n+p}, \end{cases} \tag{4.16}$$

and observe that (4.5) and (4.14) yield

$$(\forall k \in \mathbb{N}) \quad \boldsymbol{R}\boldsymbol{q}_k = \Big(q_{s,k}/2, q_{t,k}/2, Q q_{b,k}\Big), \quad \text{where} \quad Q = A^\top (\mathrm{Id} + A A^\top)^{-1}. \tag{4.17}$$

Let us further decompose the above vectors as

$$\begin{cases} s_k = (\sigma_{1,k}, \dots, \sigma_{N,k}) \in \mathbb{R}^N \\ t_k = (\tau_{1,k}, \dots, \tau_{P,k}) \in \mathbb{R}^P \\ h_{s,k} = (\eta_{1,k}, \dots, \eta_{N,k}) \in \mathbb{R}^N \\ h_{t,k} = (\eta_{N+1,k}, \dots, \eta_{N+P,k}) \in \mathbb{R}^P \\ h_{b,k} = (h_{1,k}, \dots, h_{N,k}, h_{N+1,k}, \dots, h_{N+P,k}) \\ \qquad\qquad \in \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_N} \times \mathbb{R}^{p_1} \times \cdots \times \mathbb{R}^{p_P} \\ c_{b,k} = (c_{1,k}, \dots, c_{N,k}, c_{N+1,k}, \dots, c_{N+P,k}) \\ \qquad\qquad \in \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_N} \times \mathbb{R}^{p_1} \times \cdots \times \mathbb{R}^{p_P} \\ d_{s,k} = (\delta_{1,k}, \dots, \delta_{N,k}) \in \mathbb{R}^N \\ d_{t,k} = (\delta_{N+1,k}, \dots, \delta_{N+P,k}) \in \mathbb{R}^P \\ d_{b,k} = (d_{1,k}, \dots, d_{N,k}, d_{N+1,k}, \dots, d_{N+P,k}) \\ \qquad\qquad \in \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_N} \times \mathbb{R}^{p_1} \times \cdots \times \mathbb{R}^{p_P}. \end{cases} \tag{4.18}$$

Then, given $x_{s,0} \in \mathbb{R}^N$, $x_{t,0} \in \mathbb{R}^P$, $x_{b,0} \in \mathbb{R}^p$, $h_{s,0} \in \mathbb{R}^N$, $h_{t,0} \in \mathbb{R}^P$, and $h_{b,0} \in \mathbb{R}^m$, (4.15) consists in iterating

$$\begin{array}{l} \text{for } k = 0, 1, \dots \\ \quad q_{s,k} = x_{s,k} - h_{s,k} \\ \quad q_{t,k} = x_{t,k} - h_{t,k} \\ \quad q_{b,k} = Ax_{b,k} - h_{b,k} \\ \quad s_k = x_{s,k} - q_{s,k}/2 \\ \quad t_k = x_{t,k} - q_{t,k}/2 \\ \quad b_k = x_{b,k} - Qq_{b,k} \\ \quad z_{s,k} = \operatorname{prox}_{\gamma\varsigma}(2s_k - x_{s,k}) \\ \quad z_{t,k} = \operatorname{prox}_{\gamma\varpi}(2t_k - x_{t,k}) \\ \quad z_{b,k} = \operatorname{prox}_{\gamma\theta}(2b_k - x_{b,k}) \\ \quad x_{s,k+1} = x_{s,k} + \mu_k(z_{s,k} - s_k) \\ \quad x_{t,k+1} = x_{t,k} + \mu_k(z_{t,k} - t_k) \\ \quad x_{b,k+1} = x_{b,k} + \mu_k(z_{b,k} - b_k) \\ \quad \text{for } i = 1, \dots, N \\ \quad\quad c_{i,k} = X_i b_k \\ \quad\quad (\delta_{i,k}, d_{i,k}) = (0, y_i) + \operatorname{prox}_{\gamma\widetilde{\varphi}_i}(2\sigma_{i,k} - \eta_{i,k}, 2c_{i,k} - h_{i,k} - y_i) \\ \quad \text{for } i = 1, \dots, P \\ \quad\quad c_{N+i,k} = L_i b_k \\ \quad\quad (\delta_{N+i,k}, d_{N+i,k}) = \operatorname{prox}_{\gamma\widetilde{\psi}_i}(2\tau_{i,k} - \eta_{N+i,k}, 2c_{N+i,k} - h_{N+i,k}) \\ \quad h_{s,k+1} = h_{s,k} + \mu_k(d_{s,k} - s_k) \\ \quad h_{t,k+1} = h_{t,k} + \mu_k(d_{t,k} - t_k) \\ \quad h_{b,k+1} = h_{b,k} + \mu_k(d_{b,k} - c_{b,k}). \end{array} \tag{4.19}$$

Using the above mentioned results for the convergence of the sequence $(\boldsymbol{u}_k)_{k\in\mathbb{N}}$ produced by (4.10), we obtain in the setting of Problem 3.1 the convergence of the sequences $(s_k)_{k\in\mathbb{N}}$, $(t_k)_{k\in\mathbb{N}}$, and $(b_k)_{k\in\mathbb{N}}$ generated by (4.19) to vectors s, t, and b, respectively, that solve (4.1).

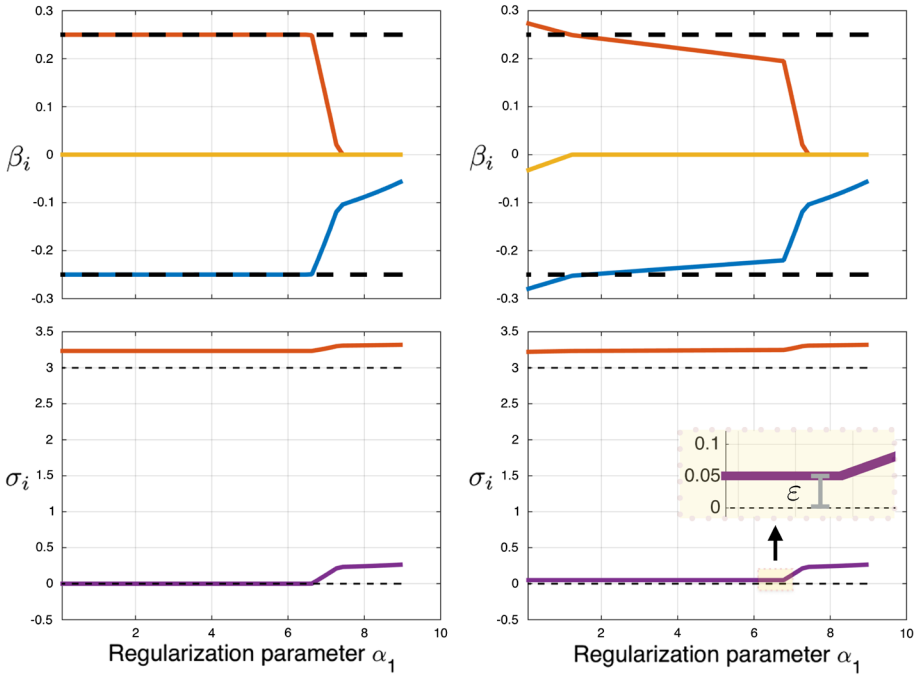

FIG 2. *Generalized heteroscedastic lasso solutions for b (top panel) and s (bottom panel) across the α_1-path. The left panels show the results of fully non-smooth perspective M-estimation. The right panels show the results for a smoothed version with $\sigma_i \in [\varepsilon, +\infty[$ for $\varepsilon = 0.05$. The dashed lines mark the ground truth entries of $\overline{b}$ and $\overline{s}$.*

5. Numerical experiments

We illustrate the versatility of perspective M-estimation for sparse robust regression in a number of numerical experiments. The proposed algorithm (4.19) has been implemented for several important instances in MATLAB and is available at https://github.com/muellsen/PCM. We set $\mu = 1.9$ and $\gamma = 1$ for all model instances. We declare that the algorithm has converged at iteration k if $\|b_{k+1} - b_k\|_2 < \epsilon$, for some $\epsilon \in]0, +\infty[$ to be specified.

5.1. Numerical illustrations on low-dimensional data

Our algorithmic approach to perspective M-estimation can effortlessly handle non-smooth data fitting terms. To illustrate this property, we consider a partially noiseless data formation model in low dimensions. We instantiate the data model (1.1) as follows. We consider the design matrix $X \in \mathbb{R}^{n \times p}$ with $p = 3$ and sample size $n = 18$. Entries in the design matrix and the noise vector $e \in \mathbb{R}^n$ are sampled from a standard normal distribution $\mathcal{N}(0, 1)$. The matrix $C \in \mathbb{R}^{n \times n}$ is a diagonal matrix with $N = 2$ groups. We set $\overline{s} = [\overline{\sigma}_1, \overline{\sigma}_2]^\top = [3, 0]^\top$. The ith

diagonal entry of C is set to $\overline{\sigma}_1$ for $i \in \{1, \dots, 8\}$ and to $\overline{\sigma}_2$ for $i \in \{10, \dots, 18\}$, resulting in noise-free observations for the second group. The mean shift (or outlier) vector is $\overline{o} = [0, \dots, 0]^\top$. The regression vector is $\overline{b} = [0.25, -0.25, 0]^\top$. The goal is to estimate the regression vector $\overline{b} \in \mathbb{R}^3$ as well as the concomitant (or scale) vector $\overline{s} = [\overline{\sigma}_1, \overline{\sigma}_2]^\top$. We consider the generalized heteroscedastic lasso of Example 3.10 with $N = 2$ and $q = 2$. To demonstrate the advantage of our non-smooth approach in this partially noiseless setting, we consider two variations of the model, the standard non-smooth case and a smoothed version with $D = [\varepsilon, +\infty[^2$ for $\varepsilon = 0.05$. The convergence accuracy for the algorithm is set to $\epsilon = 10^{-8}$. Figure 2 shows the estimates $b = [\beta_1, \beta_2, \beta_3]^\top$ and $s = [\sigma_1, \sigma_2]^\top$ across the regularization path, where $\alpha_1 \in \{0.089, \dots, 8.95\}$, with 200 values equally spaced on a log-linear grid for both settings. The results indicate that only the heteroscedastic lasso in the non-smooth setting can recover the ground truth regression vector $\overline{b}$ (top left panel) and $\overline{\sigma}_2 = 0$ (bottom left panel). In both settings the estimate σ_1 is slightly overestimated (due to the finite sample size). The "smoothed" version of the heteroscedastic lasso cannot achieve exact recovery of $\overline{b}$ across the regularization path (top right panel).

5.2. Numerical illustrations for correlated designs and outliers

To illustrate the efficacy of the different M-estimators we instantiate the full data formation model (1.1) as follows. We consider the design matrix $X \in \mathbb{R}^{n \times p}$ with $p = 64$ and sample size $n = 75$ where each row X_i is sampled from a correlated normal distribution $\mathcal{N}(0, \Sigma)$ with off-diagonal entries 0.3 and diagonal entries 1. The entries of $e \in \mathbb{R}^n$ are realizations of i.i.d. zero mean normal variables $\mathcal{N}(0, 1)$. The matrix $C \in \mathbb{R}^{n \times n}$ is a diagonal matrix with three groups. We set $\overline{s} = [\overline{\sigma}_1, \overline{\sigma}_2, \overline{\sigma}_3]^\top = [5, 0.5, 0.05]^\top$. The ith diagonal element of C is set to $\overline{\sigma}_1$ for $i \in \{1, \dots, 25\}$, to $\overline{\sigma}_2$ for $i \in \{26, \dots, 50\}$, and to $\overline{\sigma}_3$ for $i \in \{51, \dots, 75\}$. The mean shift vector $o \in \mathbb{R}^n$ contains $\lceil 0.1n \rceil = 8$ non-zero entries, sampled from $\mathcal{N}(0, 5)$. The entries of the regression vector $\overline{b} \in \mathbb{R}^p$ are set to $\overline{\beta}_i = -1$ for $i \in \{1, 3, 5\}$ and $\overline{\beta}_i = 1$ for $i \in \{2, 4, 6\}$.

The presence of outliers, correlation in the design, and heteroscedasticity provides a considerable challenge for regression estimation and support recovery with standard models such as the lasso. We consider instances of the perspective M-estimation model of increasing complexity that can cope with various aspects of the data formation model. Specifically, we use the models described in Examples 3.10 and 3.11 (with $\alpha_2 = 0$) in homoscedastic and heteroscedastic mode. For all models, we compute the minimally achievable mean absolute error (MAE) $\|Xb - X\overline{b}\|_1/n$ across the α_1-regularization path, where $\alpha_1 \in \{0.254, \dots, 25.42\}$, with 50 values equally spaced on a log-linear grid. The convergence criterion is $\epsilon = 5 \cdot 10^{-4}$.

Homoscedastic models We first consider homoscedastic instances of Examples 3.10 and 3.11, in which we jointly estimate a regression vector and a single concomitant parameter in the data fitting part. We consider the generalized

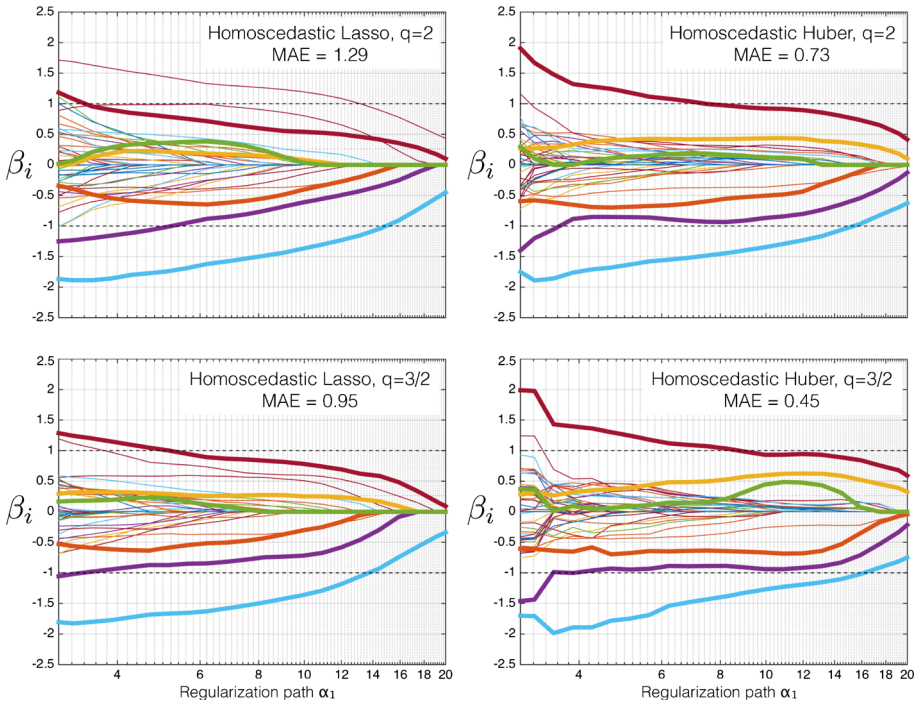

FIG 3. *Generalized homoscedastic lasso and Huber solutions for b when $q = 2$ (top panel) and $q = 3/2$ (bottom panel) across the relevant α_1-path. The minimally achievable mean absolute error (MAE) is shown for all models. The six highlighted β_i trajectories mark the true non-zeros entries of $\overline{b}$. The dashed black lines show the values of the true $\overline{b}$ entries.*

scaled lasso of Example 3.10 and the generalized Huber of Example 3.11 with exponents $q \in \{3/2, 2\}$. Figure 3 presents the estimation results of b over the relevant α_1-path.

Heteroscedastic models We consider the same model instances as previously described but in the heteroscedastic setting. We jointly estimate regression vectors and concomitant scale parameters for each of the three groups. Figure 4 presents the results for heteroscedastic lasso and Huber estimations of b across the relevant α_1-path. The convergence criterion is $\epsilon = 10^{-4}$.

The numerical experiments indicate that only heteroscedastic M-estimators are able to produce convincing b estimates (as captured by lower MAE). The heteroscedastic Huber model with $q = 3/2$ (see Figure 3 lower right panel) achieves the best performance in terms of MAE among all tested models.

5.3. Robust regression for gene expression data

We consider a high-dimensional linear regression problem from genomics [6]. The design matrix X consists of $p = 4088$ highly correlated gene expression

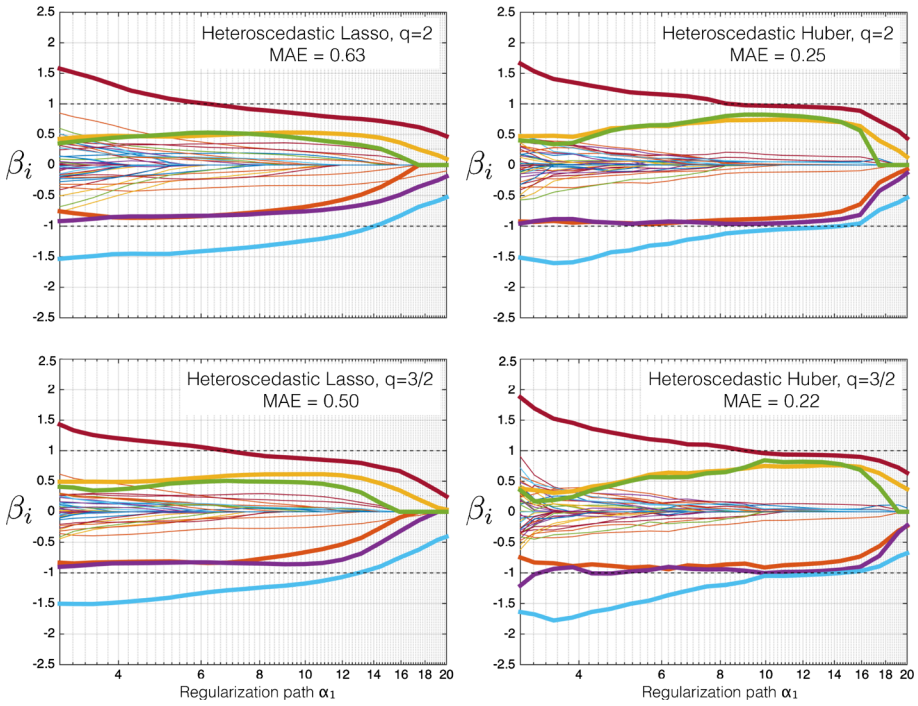

FIG 4. *Generalized heteroscedastic lasso and Huber solutions for* $\overline{b}$ *when* $q = 2$ *(top panel) and* $q = 3/2$ *(bottom panel) across the relevant* α_1*-path. The minimally achievable mean absolute error (MAE) is shown for all models. The six highlighted* β_i *trajectories mark the true non-zeros entries of* $\overline{b}$*. The dashed black lines show the values of the true* $\overline{b}$ *entries.*

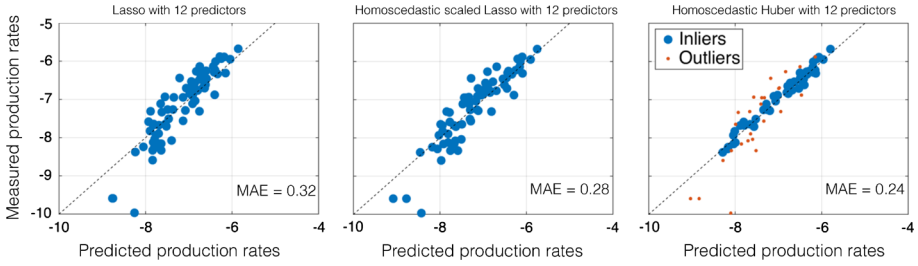

FIG 5. *Standard lasso, homoscedastic lasso, and homoscedastic Huber log-production rate predictions with identical model complexity (every vector* b *comprises twelve non-zero components) and associated MAE (computed across all samples). The Huber M-estimator identifies 26 outliers (marked in red).*

profiles for $n = 71$ different strains of Bacillus subtilis (B. subtilis). The response $y \in \mathbb{R}^{71}$ comprises standardized riboflavin (Vitamin B) log-production rates for each strain. The statistical task is to identify a small set of genes that is highly predictive of the riboflavin production rate. No grouping of the different strain

measurements is available. We thus consider the homoscedastic models from Example 3.6 with $\alpha_2 = 0$ and Example 3.11 with $\alpha_2 = 0$. We optimize the corresponding perspective M-estimation models over the α_1-path where $\alpha_1 = [0.623, \ldots, 6.23]$ with 20 values equally spaced on a log-linear grid. We compare the resulting models with the standard lasso in terms of in-sample prediction performance. Figure 5 summarizes the results for the in-sample prediction of the three different models with identical model complexity (twelve non-zero entries in $\overline{b}$). To assess model quality, we compute the minimally achievable mean absolute error (MAE) $\|Xb - y\|_1/n$ for these three models. The Huber model achieves significantly improved MAE (0.24) compared to lasso (0.32). The Huber models also identifies 26 non-zero components in the outlier vector o (shown in red in the rightmost panel of Figure 5).

References

[1] A. Antoniadis, Wavelet methods in statistics: Some recent developments and their applications, *Stat. Surv.,* vol. 1, pp. 16–55, 2007. MR2520413

[2] A. Antoniadis, Comments on: ℓ_1-penalization for mixture regression models, *TEST,* vol. 19, pp. 257–258, 2010. MR2677723

[3] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, Optimization with sparsity-inducing penalties, *Found. Trends Mach. Learn.,* vol. 4, pp. 1–106, 2011.

[4] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces,* 2nd ed. Springer, New York, 2017. MR3616647

[5] L. M. Briceño-Arias and P. L. Combettes, A monotone+skew splitting model for composite monotone inclusions in duality, *SIAM J. Optim.,* vol. 21, pp. 1230–1250, 2011. MR2854581

[6] P. Bühlmann, M. Kalisch, and L. Meier, High-dimensional statistics with a view toward applications in biology, *Ann. Rev. Stat. Appl.*, vol. 1, pp. 255–278, 2014.

[7] C. Chaux, P. L. Combettes, J.-C. Pesquet, and V. Wajs, A variational formulation for frame-based inverse problems, *Inverse Problems*, vol. 23, pp. 1495–1518, 2007. MR2348078

[8] P. L. Combettes, Perspective functions: Properties, constructions, and examples, *Set-Valued Var. Anal.,* vol. 26, pp. 247–264, 2018. MR3803982

[9] P. L. Combettes and J. Eckstein, Asynchronous block-iterative primal-dual decomposition methods for monotone inclusions, *Math. Program.,* vol. 168, pp. 645–672, 2018. MR3767763

[10] P. L. Combettes and C. L. Müller, Perspective functions: Proximal calculus and applications in high-dimensional statistics, *J. Math. Anal. Appl.,* vol. 457, pp. 1283–1306, 2018. MR3705354

[11] P. L. Combettes and J.-C. Pesquet, Proximal thresholding algorithm for minimization over orthonormal bases, *SIAM J. Optim.,* vol. 18, pp. 1351–1376, 2007. MR2373305

[12] P. L. Combettes and J.-C. Pesquet, Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping, *SIAM J. Optim.,* vol. 25, pp. 1221–1248, 2015. MR3361444

[13] P. L. Combettes and V. R. Wajs, Signal recovery by proximal forward-backward splitting, *Multiscale Model. Simul.,* vol. 4, pp. 1168–1200, 2005. MR2203849

[14] Z. J. Daye, J. Chen, and H. Li, High-dimensional heteroscedastic regression with an application to eQTL data analysis, *Biometrics,* vol. 68, pp. 316–326, 2012. MR2909888

[15] J. Eckstein and D. P. Bertsekas, On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators, *Math. Program.,* vol. 55, pp. 293–318, 1992. MR1168183

[16] I. E. Frank and J. H. Friedman, A statistical view of some chemometrics regression tools, *Technometrics,* vol. 35, pp. 109–135, 1993.

[17] M. Hannay and P.-Y. Deléamont, Error bounds for the convex loss Lasso in linear models, *Electron. J. Stat.,* vol. 11, pp. 2832–2875, 2017. MR3694570

[18] M. Hebiri and S. van de Geer, The Smooth-Lasso and other $\ell_1+\ell_2$-penalized methods, *Electron. J. Stat.,* vol. 5, pp. 1184–1226, 2011. MR2842904

[19] A. E. Hoerl, Application of ridge analysis to regression problems, *Chem. Eng. Progress,* vol. 58, pp. 54–59, 1962.

[20] P. J. Huber, Robust estimation of a location parameter, *Ann. Stat.,* vol. 35, pp. 73–101, 1964. MR0161415

[21] P. J. Huber, *Robust Statistics,* 1st ed. Wiley, New York, 1981. MR0606374

[22] S. Lambert-Lacroix and L. Zwald, Robust regression through the Huber's criterion and adaptive lasso penalty, *Electron. J. Stat.,* vol. 5, pp. 1015–1053, 2011. MR2836768

[23] S. Lambert-Lacroix and L. Zwald, The adaptive BerHu penalty in robust regression, *J. Nonparametr. Stat.*, vol. 28, pp. 487–514, 2016. MR3514507

[24] P.-L. Lions and B. Mercier, Splitting algorithms for the sum of two nonlinear operators, *SIAM J. Numer. Anal.,* vol. 16, pp. 964–979, 1979. MR0551319

[25] A. M. McDonald, M. Pontil, and D. Stamos, New perspectives on k-support and cluster norms, *J. Machine Learn. Res.,* vol. 17, pp. 1–38, 2016. MR3555046

[26] C. A. Micchelli, J. M. Morales, and M. Pontil, Regularizers for structured sparsity, *Adv. Comput. Math.,* vol. 38, pp. 455–489, 2013. MR3037026

[27] J. J. Moreau, Fonctions convexes duales et points proximaux dans un espace hilbertien, *C. R. Acad. Sci. Paris Sér. A Math.,* vol. 255, pp. 2897–2899, 1962. MR0144188

[28] E. Ndiaye, O. Fercoq, A. Gramfort, V. Leclère, and J. Salmon, Efficient smoothed concomitant lasso estimation for high dimensional regression, *J. Phys. Conf. Series,* vol. 904, art. 012006, 2017.

[29] A. B. Owen, A robust hybrid of lasso and ridge regression, *Contemp. Math.,* vol. 443, pp. 59–71, 2007. MR2433285

[30] E. Raninen and E. Ollila, Scaled and square-root elastic net, *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.,* pp. 4336–4340, 2017.

[31] R. T. Rockafellar, *Convex Analysis.* Princeton University Press, Princeton, NJ, 1970. MR0274683

[32] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, New support vector algorithms, *Neural Computation,* vol. 12, pp. 1207–1245, 2000.

[33] Y. She and A. B. Owen, Outlier detection using nonconvex penalized regression, *J. Amer. Statist. Assoc.,* vol. 106, pp. 626–639, 2011. MR2847975

[34] N. Städler, P. Bühlmann, and S. van de Geer, ℓ_1-penalization for mixture regression models, *TEST*, vol. 19, pp. 209–256. 2010. MR2677722

[35] T. Sun and C. Zhang, Scaled sparse linear regression, *Biometrika*, vol. 99, pp. 879–898, 2012. MR2999166

[36] R. Tibshirani, Regression shrinkage and selection via the lasso, *J. Roy. Stat. Soc.,* vol. B58, pp. 267–288, 1996. MR1379242

[37] R. J. Tibshirani, Adaptive piecewise polynomial estimation via trend filtering, *Ann. Stat.,* vol. 42, pp. 285–323, 2014. MR3189487

[38] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, Sparsity and smoothness via the fused lasso, *J. Roy. Stat. Soc.,* vol. B67, pp. 91–108, 2005. MR2136641

[39] J. Xu and Z. Ying, Simultaneous estimation and variable selection in median regression using Lasso-type penalty, *Ann. Inst. Stat. Math.,* vol. 62, pp. 487–514, 2010. MR2608460

[40] G. Yu and J. Bien, Estimating the error variance in a high-dimensional linear model, *Biometrika,* vol. 106, pp. 533–546, 2019. MR3992388

[41] C. Yu and W. Yao, Robust linear regression: A review and comparison, *Comm. Statist. Simulation Comput.,* vol. 46, pp. 6261–6282, 2017. MR3740779

[42] Z. Zou and T. Hastie, Regularization and variable selection via the elastic net, *J. Royal Stat. Soc.,* vol. B67, pp. 301–320, 2005. MR2137327