A Sharper Computational Tool for L_2E Regression

Xiaoqian Liu
Department of Statistics, North Carolina State University
Eric C. Chi
Department of Statistics, Rice University
and
Kenneth Lange
Departments of Computational Medicine, Human Genetics, and Statistics
University of California, Los Angeles

Abstract

Building on previous research of Chi and Chi (2022), the current paper revisits estimation in robust structured regression under the L_2E criterion. We adopt the majorization-minimization (MM) principle to design a new algorithm for updating the vector of regression coefficients. Our sharp majorization achieves faster convergence than the previous alternating proximal gradient descent algorithm (Chi and Chi, 2022). In addition, we reparameterize the model by substituting precision for scale and estimate precision via a modified Newton's method. This simplifies and accelerates overall estimation. We also introduce distance-to-set penalties to enable constrained estimation under nonconvex constraint sets. This tactic also improves performance in coefficient estimation and structure recovery. Finally, we demonstrate the merits of our improved tactics through a rich set of simulation examples and a real data application.

Keywords: Integral squared error criterion; MM principle; Newton's method; penalized estimation; distance penalization

1 Introduction

Linear least squares regression quantifies the relationship between a response and a set of predictors. As such, it has been the most popular and productive technique of classical statistics. The growing complexity of modern datasets necessitates special structures on the vector of regression coefficients. A typical example is sparse regression for high-dimensional data, where the number of predictors exceeds the number of responses. In this setting, assuming the coefficient vector is sparse not only improves a regression model's interpretability but also improves its prediction accuracy. The most popular vehicle for dealing with sparse regression is the least absolute shrinkage and selection operator (Lasso) (Tibshirani, 1996). Other examples of structured regression include isotonic regression (Barlow and Brunk, 1972), convex regression (Seijo and Sen, 2011), and ridge regression (Hoerl and Kennard, 1970).

Traditional structured regression estimates parameters by constrained least squares. Unfortunately, least squares estimates are extremely sensitive to outliers. A single outlier can ruin estimation accuracy. Consequently, robust structured regression has gained considerable traction in recent years. Numerous authors have contributed to the current body of techniques. To mention a few, Alvarez and Yohai (2012) propose a family of robust estimates for isotonic regression that replaces the least squares criterion with the M-estimation criterion (Huber, 1992). Blanchet et al. (2019) employ absolute error loss in robust convex regression. This is also an instance of an M-estimator. Nguyen and Tran (2012) suggest an extended Lasso method incorporating a stochastic noise term to account for corrupted observations in robust sparse multiple regression. Alfons et al. (2013) add a

Lasso penalty to the least trimmed squares (LTS) loss to produce a robust sparse estimator that trims outliers by effectively minimizing the sum of squared residuals over a selected subset. Lozano et al. (2016) adopt the minimum distance criterion to design a log-scaled loss function and propose the minimum distance Lasso method for robust sparse regression. Other robust sparse regression methods can be found in Wang et al. (2007); She and Owen (2011); Wang et al. (2013).

The above works investigate robust structured regression on a case-by-case basis. Yang et al. (2018) develop a family of trimmed regularized M-estimators with a wider focus but with the need to select the degree of trimming. Recently, Chi and Chi (2022) derive yet another general framework for robust structured regression that simultaneously estimates regression coefficients as well as a precision parameter, which plays the same role as the trimming parameter in Yang et al. (2018). Chi and Chi (2022) use the L_2E criterion (Scott, 1992) to quantify goodness-of-fit and a convex penalty to enforce structure. Their algorithmic framework solves the corresponding optimization problem by block descent. Although the computational framework presented in Chi and Chi (2022) is general, there is room for some nontrivial improvements. First, the proposed proximal gradient algorithm for updating both the regression coefficients and the precision parameter at each block descent iteration can be slow to converge. Second, the box constraint on the precision parameter introduces two additional hyper-parameters that must be specified. Finally, while Chi and Chi (2022) focused on convex penalties and constraints, the framework that they introduced is not inherently limited to convex options and warrants extension to important nonconvex alternatives that impose desirable structures.

The limitations in Chi and Chi (2022) just discussed motivate the current paper and its new contributions. First, we derive a majorization-minimization algorithm to accelerate the estimation of the regression coefficients. Second, we reparameterize the precision parameter to eliminate the box constraint. A simple one-dimensional approximate Newton's method quickly solves the resulting smooth unconstrained problem for updating precision. Finally, we demonstrate improved statistical performance by imposing nonconvex penalties. Specifically, we adopt distance-to-set penalties to improve estimation accuracy subject to structural constraints. These improvements do not compromise robustness.

The rest of this paper is organized as follows. In Section 2, we review the L_2E criterion, the majorization-minimization (MM) principle, and distance penalization. In Section 3, we set up the optimization problem for robust structured regression under the L_2E criterion. In Section 4, we introduce strategies that improve the estimation techniques of Chi and Chi (2022). In Sections 5 and 6, we provide a rich set of simulation examples and a real data application to demonstrate the empirical performance of our new algorithms. We end with a discussion in Section 7.

2 Background

2.1 The L_2E Criterion

Although traditionally used in nonparametric estimation, the L_2E criterion, also known as the integrated squared error (ISE), can be exploited in parametric settings for robust estimation. Suppose the goal is to estimate a density function $f(\boldsymbol{x} \mid \boldsymbol{\theta})$, where the true

parameter $\boldsymbol{\theta}_*$ is unknown. The L_2E criterion seeks to estimate $\boldsymbol{\theta}$ by minimizing the L_2 distance between $f(\boldsymbol{x} \mid \boldsymbol{\theta})$ and $f(\boldsymbol{x} \mid \boldsymbol{\theta}_*)$; thus

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \operatorname{argmin}_{\boldsymbol{\theta}} \int [f(\boldsymbol{x} \mid \boldsymbol{\theta}) - f(\boldsymbol{x} \mid \boldsymbol{\theta}_*)]^2 \, d\boldsymbol{x} \\ &= \operatorname{argmin}_{\boldsymbol{\theta}} \int f(\boldsymbol{x} \mid \boldsymbol{\theta})^2 \, d\boldsymbol{x} - 2 \int f(\boldsymbol{x} \mid \boldsymbol{\theta}) f(\boldsymbol{x} \mid \boldsymbol{\theta}_*) \, d\boldsymbol{x} + \int f(\boldsymbol{x} \mid \boldsymbol{\theta}_*)^2 \, d\boldsymbol{x}. \end{aligned} \tag{1}$$

The third integral in formula (1) does not depend on $\boldsymbol{\theta}$ and can be excluded from the minimization. The second integral is the expectation of $f(\boldsymbol{x} \mid \boldsymbol{\theta})$ and can be approximated by an unbiased estimate, namely its sample mean. Therefore, an approximate L_2 estimate of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}}_{\mathrm{L_2E}} = \operatorname{argmin}_{\boldsymbol{\theta}} \int f(\boldsymbol{x} \mid \boldsymbol{\theta})^2 \, d\boldsymbol{x} - \frac{2}{n} \sum_{i=1}^{n} f(\boldsymbol{x}_i \mid \boldsymbol{\theta}), \tag{2}$$

where n denotes the sample size. The L_2E represents a trade-off between efficiency and robustness. It is less efficient but more robust than the maximum likelihood estimate (MLE) (Scott, 2001; Warwick and Jones, 2005). Chi and Chi (2022) discuss in detail how the L_2E estimator imparts robustness in structured regression.

2.2 The MM Principle

The majorization-minimization principle (Lange et al., 2000; Lange, 2016) for minimizing an objective function $h(\boldsymbol{\theta})$ involves two steps, a) majorization of $h(\boldsymbol{\theta})$ by a surrogate function $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}_k)$ anchored at the current iterate $\boldsymbol{\theta}_k$ and then b) minimization of $\boldsymbol{\theta} \mapsto g(\boldsymbol{\theta} \mid \boldsymbol{\theta}_k)$ to construct $\boldsymbol{\theta}_{k+1}$. The surrogate function $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}_k)$ must satisfy the two requirements:

$$h(\boldsymbol{\theta}_k) = g(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_k), \quad \text{tangency} \tag{3}$$

$$h(\boldsymbol{\theta}) \leq g(\boldsymbol{\theta} \mid \boldsymbol{\theta}_k) \quad \text{for all } \boldsymbol{\theta}, \quad \text{domination.} \tag{4}$$

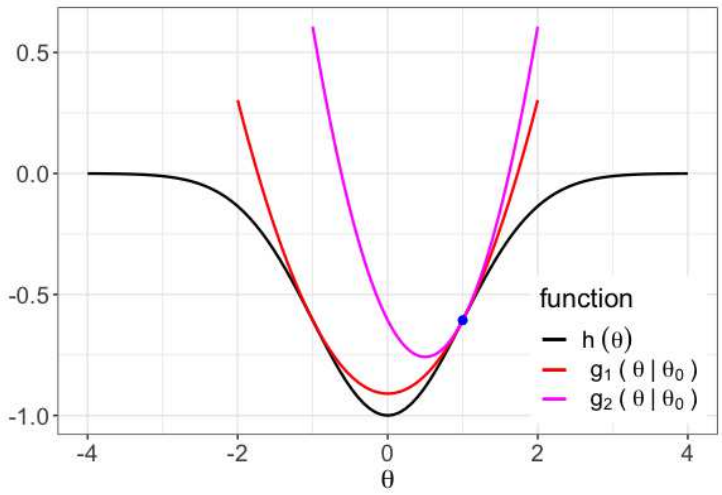

Figure 1: An example of sharp quadratic majorization. The quadratic $g_1(\theta \mid \theta_0)$ offers the sharpest majorization of the loss $h(\theta)$ and falls below every looser quadratic majorization $g_2(\theta \mid \theta_0)$.

Under these conditions, the iterates enjoy the descent property $h(\boldsymbol{\theta}_{k+1}) \leq h(\boldsymbol{\theta}_k)$ as demonstrated by the relations

$$h(\boldsymbol{\theta}_{k+1}) \;\leq\; g(\boldsymbol{\theta}_{k+1} \mid \boldsymbol{\theta}_k) \;\leq\; g(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_k) \;=\; h(\boldsymbol{\theta}_k),$$

reflecting conditions (3) and (4). Ideally, the MM principle converts a hard optimization problem into a sequence of easier ones. The key to success is the construction of a tight majorization that can be easily minimized. In some problems it is possible to construct a sharp majorization within a limited class of majorizers. Figure 1 depicts a sharp quadratic majorization that is best among all quadratic majorizations that share the same tangency point. Sharp majorization accelerates the convergence of a derived MM algorithm (de Leeuw and Lange, 2009). In practice, majorization can be done piecemeal by exploiting the convexity or concavity of the various terms comprising the objective.

2.3 Distance Penalization

To estimate a parameter vector $\boldsymbol{\theta}$ subject to a set constraint $\boldsymbol{\theta} \in C$, it is often convenient to employ a squared Euclidean distance penalty (Chi et al., 2014; Xu et al., 2017). For a closed set C, the penalty is defined as

$$\frac{1}{2}\operatorname{dist}(\boldsymbol{\theta}, C)^2 \quad = \quad \min_{\boldsymbol{\beta} \in C} \frac{1}{2}\|\boldsymbol{\theta} - \boldsymbol{\beta}\|_2^2. \tag{5}$$

The beauty of this penalty is that it is majorized at the current iterate $\boldsymbol{\theta}_k$ by the spherical quadratic

$$\frac{1}{2}\|\boldsymbol{\theta} - \mathcal{P}_C(\boldsymbol{\theta}_k)\|_2^2, \tag{6}$$

where $\mathcal{P}_C(\boldsymbol{\theta})$ denotes the Euclidean projection of $\boldsymbol{\theta}$ onto C (Bauschke and Combettes, 2011). When C is both closed and convex, $\mathcal{P}_C(\boldsymbol{\theta})$ consists of a single point. For nonconvex sets, $\mathcal{P}_C(\boldsymbol{\theta})$ sometimes consists of multiple points. When $\mathcal{P}_C(\boldsymbol{\theta})$ is single valued, the distance penalty (5) has gradient $\boldsymbol{\theta} - \mathcal{P}_C(\boldsymbol{\theta})$

The proximal distance method of constrained optimization minimizes the penalized objective $h(\boldsymbol{\theta}) + \frac{\rho}{2}\operatorname{dist}(\boldsymbol{\theta}, C)^2$ (Xu et al., 2017; Keys et al., 2019). The tuning constant ρ controls the trade-off between minimizing the loss $h(\boldsymbol{\theta})$ and satisfying the constraint $\boldsymbol{\theta} \in C$. Under suitable regularity conditions, the constrained solution can be recovered in the limit as ρ tends towards infinity (Chi et al., 2014; Keys et al., 2019). Therefore, a large value of ρ, say 10^8, is chosen in practice to enforce the constraint. The MM principle suggests majorizing the distance penalty by the spherical quadratic (6) and applying the proximal map $\boldsymbol{\theta}_{k+1} = \operatorname{prox}_{\rho^{-1}h}[\mathcal{P}_C(\boldsymbol{\theta}_k)]$ to generate the next iterate. The proximal distance principle applies to a wide array of models, including sparse regression, nonnegative regression, and

low-rank matrix completion. It is accurate in estimation and avoids the severe shrinkage of Lasso penalization with well-behaved constraint sets (Xu et al., 2017). Landeros et al. (2020) extend distance penalization to fusion constraints of the form $\boldsymbol{D}\boldsymbol{\beta} \in C$ involving a fusion matrix $\boldsymbol{D}$ such as a discrete difference operator. Although the advantages of proximal maps are lost, this extension brings more constrained statistical models under the umbrella of distance penalization.

3 L_2E Robust Structured Regression

Consider the classical linear regression model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}+\tau^{-1}\boldsymbol{\epsilon}$, where $\boldsymbol{y} \in \mathbb{R}^n$ is the response vector, $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is the design matrix of predictors, and $\boldsymbol{\epsilon} \in \mathbb{R}^n$ is the noise vector with independent standard Gaussian components. The regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$ and the precision $\tau \in \mathbb{R}_+$ are the parameters of the model. Collectively, we denote the parameters by $\boldsymbol{\theta} = (\boldsymbol{\beta}^\mathsf{T}, \tau)^\mathsf{T}$. The density of the ith response y_i amounts to

$$f(y_i \mid \boldsymbol{\theta}) \;=\; \frac{\tau}{\sqrt{2\pi}} e^{-\frac{\tau^2 r_i^2}{2}},$$

where $r_i = y_i - \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}$ is the ith residual. A brief calculation shows that equation (2) gives rise to the L_2E loss

$$h(\boldsymbol{\theta}) \;=\; \int f\left(y \mid \boldsymbol{\theta}\right)^2 dy - \frac{2}{n}\sum_{i=1}^n f\left(y_i \mid \boldsymbol{\theta}\right) \;=\; \frac{\tau}{2\sqrt{\pi}} - \frac{\tau}{n}\sqrt{\frac{2}{\pi}}\sum_{i=1}^n e^{-\frac{\tau^2 r_i^2}{2}}. \tag{7}$$

Structured regression introduces set constraints on the regression coefficient vector $\boldsymbol{\beta}$. Consequently, L_2E aims to solve the constrained optimization problem

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p, \tau\in\mathbb{R}_+} h(\boldsymbol{\beta}, \tau), \quad \text{subject to } \boldsymbol{\beta} \in C. \tag{8}$$

For example, $C = \{\boldsymbol{\beta} \in \mathbb{R}^p : \beta_1 \leq \beta_2 \leq \cdots \leq \beta_p\}$ leads to a robust isotonic regression problem. Sparsity can be imposed directly by taking $C = \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}\|_0 \leq k\}$ for some positive integer k or indirectly by taking $C = \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}\|_1 \leq t\}$ for $t > 0$. Alternatively, we can rewrite problem (8) as the non-smooth optimization problem

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p, \tau \in \mathbb{R}_+} h(\boldsymbol{\beta}, \tau) + \phi(\boldsymbol{\beta}), \tag{9}$$

where the penalty $\phi(\boldsymbol{\beta})$ is either the $0/\infty$ indicator of the constraint set C denoted by $\iota_C(\boldsymbol{\beta})$ or a better behaved but still non-smooth substitute such as the Lasso. Although we emphasize structured regression, the formulations (8) and (9) also include unstructured multivariate regression where $C = \mathbb{R}^p$ and $\phi(\boldsymbol{\beta}) \equiv 0$.

Solving problem (8), or equivalently solving (9), is challenging for two reasons. First, both problems are nonconvex owing to the nonconvexity of the L_2E loss (7). Second, the penalty term $\phi(\boldsymbol{\beta})$ may be non-differentiable. Fortunately, the block gradients of the L_2E loss with respect to $\boldsymbol{\beta}$ and τ, $\nabla_{\boldsymbol{\beta}} h(\boldsymbol{\beta}, \tau)$ and $\frac{\partial}{\partial \tau} h(\boldsymbol{\beta}, \tau)$, are Lipschitz. This key property motivates a block descent algorithm (Chi and Chi, 2022) that alternates between reducing the objective with respect to $\boldsymbol{\beta}$ and τ, holding the other block fixed. Chi and Chi (2022) also impose the bounds $0 < \tau_{\min} \leq \tau \leq \tau_{\max} < \infty$ on τ.

An appealing property of block descent is that the objective function is guaranteed to decrease at each iteration. Chi and Chi (2022) apply proximal gradient descent to decrease the objective in each block update. Because the proximal gradient updates are based on a loose loss majorization, the algorithm is slow to converge. To ameliorate this fault, we propose new strategies for updating $\boldsymbol{\beta}$ and τ in the next section.

4 Computational Methods

4.1 Updating the Regression Coefficients

Consider the problem of updating the regression coefficients $\boldsymbol{\beta}$. Because the contribution $-e^{-\tau^2 r_i^2/2}$ to the L_2E loss (7) is differentiable and concave with respective to r_i^2, we can exploit the concave majorization

$$f(u) \;\leq\; f(u_k) + f'(u_k)(u - u_k)$$

in the form

$$-e^{-\tau^2 r_i^2/2} \;\leq\; -e^{-\tau^2 r_{ki}^2/2} + \frac{\tau^2}{2} e^{-\tau^2 r_{ki}^2/2} \left(r_i^2 - r_{ki}^2\right) \tag{10}$$

around the tangency point r_{ki}^2. By omitting irrelevant multiplicative and additive terms, this produces the surrogate function

$$f(\boldsymbol{\beta} \mid \boldsymbol{\beta}_k, \tau) \;=\; \frac{1}{2} \sum_{i=1}^{n} e^{-\tau^2 r_{ki}^2/2} (y_i - \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta})^2 \;=\; \frac{1}{2} \|\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}} \boldsymbol{\beta}\|_2^2 \tag{11}$$

for the L_2E loss (7), where $r_{ki} = y_i - \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta}_k$ is the ith residual at iteration k, $\tilde{\boldsymbol{y}} = \sqrt{\boldsymbol{W}_k} \boldsymbol{y}$, $\tilde{\boldsymbol{X}} = \sqrt{\boldsymbol{W}_k} \boldsymbol{X}$, and $\boldsymbol{W}_k \in \mathbb{R}^{n \times n}$ is a diagonal weight matrix with the ith diagonal entry $e^{-\tau^2 r_{ki}^2/2}$.

The next proposition demonstrates that the surrogate (11) is the sharpest quadratic majorization in the residual variables r_i. It does not claim that the majorization (11) is the sharpest multivariate quadratic majorization in the full variable $\boldsymbol{\beta}$. Despite this fact, the majorization yields substantial gains in computational efficiency over the looser proximal gradient majorization pursued by Chi and Chi (2022).

Proposition 4.1. *Let $f(r) = -e^{-ar^2}$ with $a > 0$. Then the symmetric quadratic function $g(r) = -e^{-ar_k^2} + ae^{-ar^2}(r^2 - r_k^2)$ is the sharp quadratic majorizer of $f(r)$.*

Proof. Van Ruitenburg (2005) proves that a univariate quadratic function $g(r)$ majorizing a univariate differentiable function $f(r)$ and touching it at two points is sharp. In the present case, $g(r)$ touches $f(r)$ at the points $r = \pm r_k$. □

For an L_2E loss with penalty $\phi(\boldsymbol{\beta})$, the next MM iterate is

$$\boldsymbol{\beta}_{k+1} \;=\; \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \; \frac{1}{2}\|\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}\|_2^2 + \phi(\boldsymbol{\beta}).$$

In the setting of distance penalization with a fusion penalty, the surrogate reduces to the least squares criterion

$$\frac{1}{2}\left\| \begin{pmatrix} \tilde{\boldsymbol{y}} \\ \sqrt{\rho}\mathcal{P}_C(\boldsymbol{D}\boldsymbol{\beta}_k) \end{pmatrix} - \begin{pmatrix} \tilde{\boldsymbol{X}} \\ \sqrt{\rho}\boldsymbol{D} \end{pmatrix} \boldsymbol{\beta} \right\|_2^2,$$

which is amenable to minimization by the QR algorithm or the conjugate gradient algorithm. The computational complexity of the $\boldsymbol{\beta}$ update is dominated by this least squares problem. Indeed, computation of the current residuals, the matrix $\boldsymbol{W}_k$, the product $\tilde{\boldsymbol{y}}$, and the product $\tilde{\boldsymbol{X}}$ require, respectively, operation counts of $\mathcal{O}(np)$, $\mathcal{O}(n)$, $\mathcal{O}(n)$, and $\mathcal{O}(np)$. Updating $\boldsymbol{\beta}$ using proximal gradient descent requires similar steps. Evaluation of the proximal map of $\phi(\boldsymbol{\beta})$ reduces to penalized least squares with an identity design matrix. Hence, with a diagonal design matrix $\boldsymbol{X}$, the computational cost per iteration of the current MM algorithm is essentially the same as that of the proximal gradient descent algorithm in Chi and Chi (2022). The numbers of iterations until convergence of the two algorithms are vastly different however. Additionally, the distance penalized MM algorithm is more flexible in allowing nonconvex and fusion constraints.

4.2 Updating the Precision Parameter

There are two concerns in updating τ, namely the slow convergence of proximal gradient descent and the presence of box constraints on τ. To attack the latter concern, we reparameterize by setting $\tau = e^{\eta}$ for any real valued η. Because the stationary condition for minimizing the loss $h(\boldsymbol{\beta}, e^{\eta})$ with respect to η is intractable, we turn to a variant of Newton's method. The required first and second derivatives are

$$\begin{aligned}
\frac{\partial}{\partial \eta} h(\boldsymbol{\beta}, e^{\eta}) &= \frac{e^{\eta}}{2\sqrt{\pi}} - \frac{e^{\eta}}{n}\sqrt{\frac{2}{\pi}} \sum_{i=1}^{n} w_i + \frac{e^{3\eta}}{n}\sqrt{\frac{2}{\pi}} \sum_{i=1}^{n} w_i r_i^2 \\
\frac{\partial^2}{\partial \eta^2} h(\boldsymbol{\beta}, e^{\eta}) &= \frac{e^{\eta}}{2\sqrt{\pi}} + \frac{4e^{3\eta}}{n}\sqrt{\frac{2}{\pi}} \sum_{i=1}^{n} w_i r_i^2 - \frac{e^{\eta}}{n}\sqrt{\frac{2}{\pi}} \sum_{i=1}^{n} w_i - \frac{e^{5\eta}}{n}\sqrt{\frac{2}{\pi}} \sum_{i=1}^{n} w_i r_i^4,
\end{aligned}$$

where $w_i = e^{-e^{2\eta} r_i^2/2}$ and r_i is the ith residual. The Newton increment only points downhill when $\frac{\partial^2}{\partial \eta^2} h(\boldsymbol{\beta}, e^{\eta})$ is positive. This prompts discarding the negative contributions and relying on the approximation

$$\frac{\partial^2}{\partial \eta^2} h(\boldsymbol{\beta}, e^{\eta}) \approx d = \frac{e^{\eta}}{2\sqrt{\pi}} + \frac{4e^{3\eta}}{n}\sqrt{\frac{2}{\pi}} \sum_{i=1}^{n} w_i r_i^2.$$

Our modified Newton's iterates are defined by

$$\eta_{k+1} = \eta_k - t_k d_k^{-1} \frac{\partial}{\partial \eta} h(\boldsymbol{\beta}, e^{\eta_k}),$$

where t_k is a positive stepsize parameter chosen via Armijo backtracking started at $t_k = 1$. Little backtracking is needed because replacing $\frac{\partial^2}{\partial \eta^2} h(\boldsymbol{\beta}, e^{\eta})$ by the larger value d diminishes the chances of overshooting the minimum of $h(\boldsymbol{\beta}, e^{\eta})$.

Our modified Newton's method enjoys the same computational complexity as proximal gradient descent. The dominant computational expense in updating η in both algorithms comes from computing the residuals r_i. This step requires $\mathcal{O}(np)$ operations. Once all r_i

are updated, computing the derivatives only requires an additional $\mathcal{O}(n)$ operations. In summary, our new strategy converges in fewer iterations, removes the box constraint on τ, and enjoys the same computational cost per iteration as proximal gradient descent.

Algorithm 1 summarizes our algorithm for minimizing the penalized loss (9). As in Chi and Chi (2022), we set the maximum numbers of inner iterations for updating $\boldsymbol{\beta}$ and η to be $N_{\boldsymbol{\beta}}$ and N_η, respectively, at each outer iteration. Extreme values $N_{\boldsymbol{\beta}}$ and N_η tend to slow overall convergence. In our simulation studies, we set $N_{\boldsymbol{\beta}} = N_\eta = 100$. In the algorithm the notation $\boldsymbol{W}_+$ signifies that $\boldsymbol{W}$ depends on the previous inner iterate $\boldsymbol{\beta}_+$.

Algorithm 1 Block descent with MM and approximate Newton for problem (9)

Initialize: $\boldsymbol{\beta}_0 \in \mathbb{R}^p$, $\tau_0 \in \mathbb{R}_+$, $N_{\boldsymbol{\beta}}$, and N_η.

for $k = 1, 2, \cdots$ **do**
  $\boldsymbol{\beta}^+ \leftarrow \boldsymbol{\beta}_{k-1}$
  for $i = 1, \cdots, N_{\boldsymbol{\beta}}$ **do**
    $\tilde{\boldsymbol{y}} = \sqrt{\boldsymbol{W}_+}\boldsymbol{y}$
    $\tilde{\boldsymbol{X}} = \sqrt{\boldsymbol{W}_+}\boldsymbol{X}$
    $\boldsymbol{\beta}^+ = \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{2}\|\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}\|_2^2 + \phi(\boldsymbol{\beta})$
  end for
  $\boldsymbol{\beta}_k \leftarrow \boldsymbol{\beta}^+$
  $\eta^+ \leftarrow \log(\tau_{k-1})$
  for $i = 1, \cdots, N_\eta$ **do**
    $\eta^+ = \eta^+ - t_i d_i^{-1} \frac{\partial}{\partial \eta} h(\boldsymbol{\beta}_k, e^{\eta^+})$
  end for
  $\tau_k \leftarrow e^{\eta^+}$
end for

We close this section by stressing the importance of the weight matrix $\boldsymbol{W}_+$ in the success of L_2E regression. The diagonal entry $e^{-\tau^2 r_{+i}^2/2}$ of $\boldsymbol{W}_+$ depends on the ith residual from the previous inner iterate $\boldsymbol{\beta}_+$ and downweights case i if its residual is large. The converged weights also conveniently flag outliers. We will exploit this bonus later in Section 6.

5 Numerical Experiments

To compare the estimation accuracy and computational efficiency of Algorithm 1 (abbreviated MM) and proximal gradient descent (abbreviated PG), we consider isotonic regression and convex regression. To highlight the advantages of distance penalization over competing model selection methods, we consider sparse regression and trend filtering. For the sake of brevity, we relegate two of the examples to the supplement. Readers wishing to implement our version of L_2E regression should visit the eponymous `L2E` R package (Liu et al., 2022) on the Comprehensive R Archive Network (CRAN).

5.1 Robust Isotonic Regression

Classical isotonic regression involves minimizing the least squares criterion

$$\|\boldsymbol{y} - \boldsymbol{\beta}\|_2^2 \;=\; \sum_{i=1}^{n} (y_i - \beta_i)^2$$

subject to $\boldsymbol{\beta}$ belonging to the set $C_1 = \{\boldsymbol{\beta} \in \mathbb{R}^n : \beta_1 \leq \cdots \leq \beta_n\}$. Independent standard normal errors are implicit in this formulation. Here the design matrix $\boldsymbol{X} = \boldsymbol{I}_n$, and the mean function of the model is monotonically increasing and piecewise constant. In the L_2E version of the problem, we impose the $0/\infty$ penalty $\phi(\boldsymbol{\beta}) = \iota_{C_1}(\boldsymbol{\beta})$. The MM update of $\boldsymbol{\beta}$ succumbs to the `gpava` function in the `isotone` R package (de Leeuw et al., 2010). As mentioned earlier, the MM $\boldsymbol{\beta}$ update enjoys the same per-iteration computational cost as the PG $\boldsymbol{\beta}$ update (Chi and Chi, 2022).

In our simulation, 1000 responses are generated by sampling points x_i evenly from $[-2.5, 2.5]$ and setting $y_i = x_i^3 + s_i + \varepsilon_i$, where the ε_i are i.i.d. standard normal deviates, and

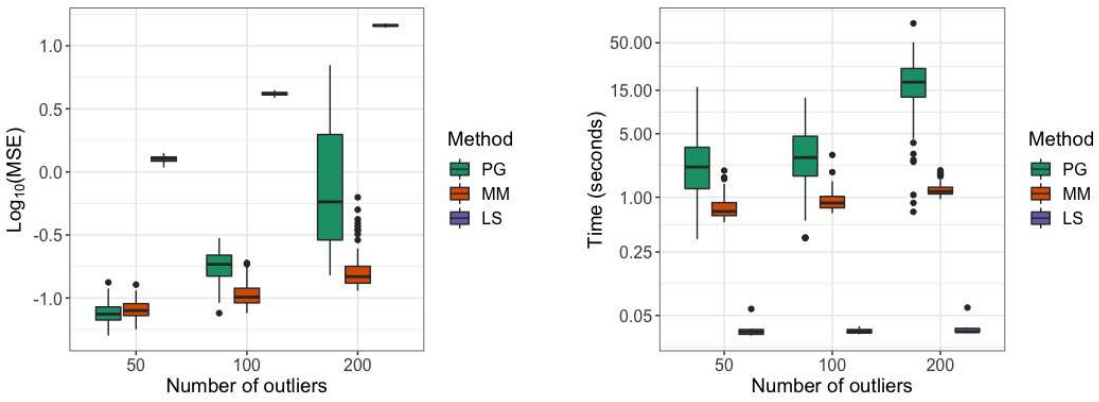

Figure 2: Simulation results for isotonic regression under different numbers of outliers. Boxplots depict the MSE (left panel) and run time (right panel) over 100 replicates.

the s_i shift the underlying cubic signal. The responses define mean vector $\boldsymbol{\beta} \in \mathbb{R}^{1000}$. Outliers are introduced at consecutive responses by setting $s_i = 14$ for $i = 251, 252, \cdots, 250+m$, where m is the number of outliers; all other responses have $s_i = 0$. The shift of 14 makes the contaminated responses match the maximum observed value in the uncontaminated responses. Each method is tested over 100 replicates and initialized by $\boldsymbol{\beta}_0 = \boldsymbol{y}$, $\tau_0 = \text{MAD}(\boldsymbol{y})^{-1}$ for PG, and $\eta_0 = -\log[\text{MAD}(\boldsymbol{y})]$ for MM, where $\text{MAD}(\boldsymbol{y})$ is the reciprocal of the median absolute deviation of the responses.

Figure 2 displays boxplots of the MSEs and run times in seconds in fitting the isotonic regression model under different numbers of outliers. We include the results from ordinary least squares (abbreviated LS) as a baseline. As anticipated, the estimation accuracy of LS degrades as the number of outliers increases. In contrast, both MM and PG exhibit much more modest increases in estimation error, with MM less sensitive to outliers than PG. Note that the optimization problems of PG and MM differ slightly. We put a box constraint on τ for PG but reparameterize τ as $\tau = e^{\eta}$ for MM to eliminate the box constraint on τ. For sufficiently large box constraints, the solutions to the two problems coincide, but differences

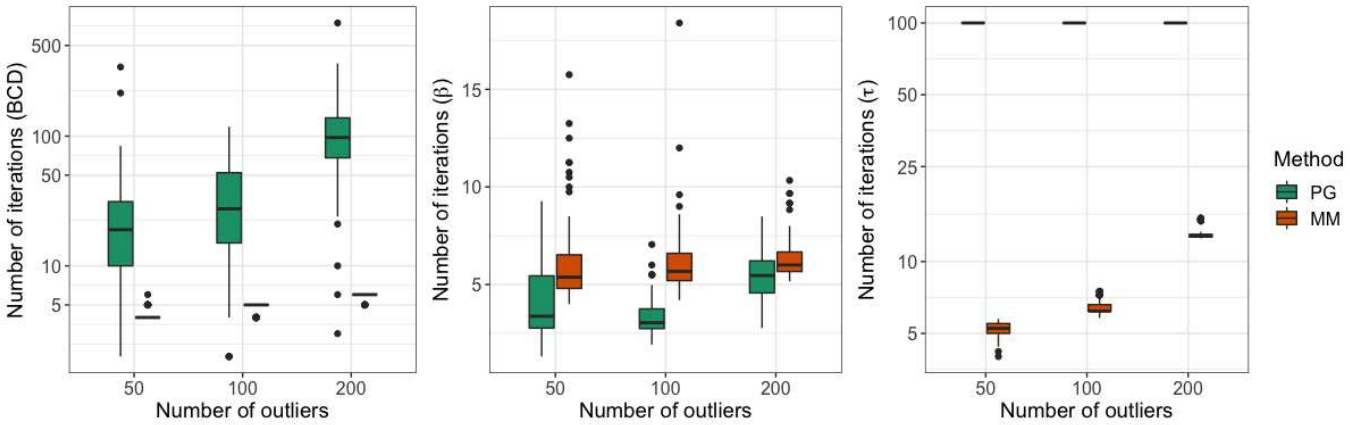

Figure 3: Boxplots of the mean number of outer block descent iterations (left panel), the mean number of inner iterations for updating $\boldsymbol{\beta}$ per outer iteration (middle panel), and the mean number of inner iterations for updating τ per outer iteration (right panel). All plots refer to the experiment summarized in Figure 2.

in the algorithms will still produce different algorithm iterate trajectories. As discussed in Section 3, the L_2E optimization problem is nonconvex and may exhibit multiple local minima. Thus, PG and MM may converge to different minima and produce different MSEs.

The right panel of Figure 2 shows the significant speed advantage of MM over PG. Run times of PG increase rapidly as the number of outliers increases, while run times of MM are far more stable against the number of outliers. MM is less computationally efficient than LS, which avoids computation of case weights. The difference in run time between PG and MM is directly attributable to MM's reduced number of outer iterations until convergence. For the same experiment, Figure 3 depicts boxplots of the mean number of outer block descent iterations, the mean number of inner iterations for updating $\boldsymbol{\beta}$ per outer iteration, and the mean number of inner iterations for updating τ per outer iteration. Note that in our implementation, we terminate the inner iterations for updating $\boldsymbol{\beta}$ and τ

if certain convergence conditions are satisfied. Readers may refer to the `L2E` package for details. It may seem paradoxical that PG takes fewer inner iterations than MM to update $\boldsymbol{\beta}$. However, recall that PG is fitting a less snug surrogate than MM. PG also takes far more inner iterations than MM to update τ. This reflects the speed of our approximate Newton method.

The robust isotonic simulations also illustrate the ability of L_2E regression to handle outliers under various contamination levels. To explore this tendency, we fix the number of outliers at $m = 100$, vary the shifts s_i over the grid $\{2, 5, 8, 14, 20\}$, adopt the same initialization as the previous experiment, and run 100 replicates for each scenario. Figure 4 summarizes the estimation and computation performance of PG, MM, and LS under different contamination levels. When the data are only slightly contaminated ($s_i = 2$), the two robust methods, PG and MM, fail to detect the outliers and achieve estimation accuracy comparable to LS. However, as the level of contamination s_i grows, the MSE of LS increases rapidly, while the MSE of MM behaves robustly and quickly declines. Interestingly, the MSE of PG decreases gradually as the shift grows. These results suggest that both PG and MM need a certain level of contamination to successfully detect outliers. MM is more responsive to the contamination than PG even if the data are modestly contaminated. This is yet another advantage of MM over PG.

The right panel of Figure 4 illustrates how PG's run times increase as the contamination level increases. The run times of MM, however, are stable with contamination level and consistently shorter than those of PG, though longer than those of LS. Figure 5 explains the difference in the computational performance between PG and MM. The numbers of

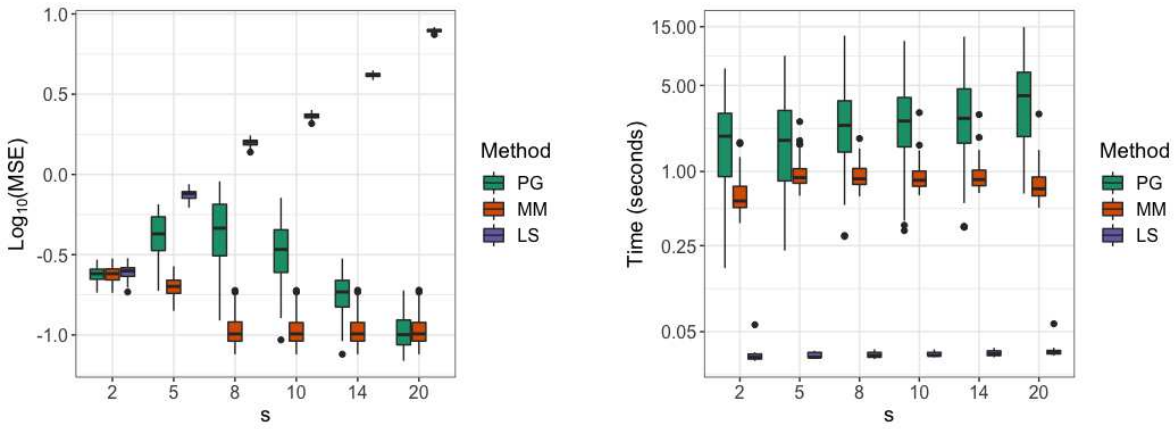

Figure 4: Simulation results for isotonic regression under different contamination levels. Boxplots depict the MSE (left panel) and run time (right panel) over 100 replicates.

inner iterations for updating $\boldsymbol{\beta}$ and τ for both PG and MM are insensitive to contamination level. MM's number of outer block descent iterations is always small, while PG's number of outer iterations increases. This difference explains the speed advantage of MM.

5.2 Robust Sparse Regression

Sparse linear regression minimizes the penalized least squares criterion

$$\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \phi(\boldsymbol{\beta}),$$

with $\phi(\boldsymbol{\beta})$ promoting sparsity. Typical choices of $\phi(\boldsymbol{\beta})$ includes the Lasso and the nonconvex MCP penalty (Zhang et al., 2010). In the L_2E framework, each MM update solves a ϕ-penalized least squares problem. The `ncvfit` function in the R package `ncvreg` is ideal for this purpose (Breheny and Huang, 2011). In the distance penalty context, the constraint set is $C_2 = \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}\|_0 \leq k\}$, where the positive integer k encodes the sparsity level. The MM update of $\boldsymbol{\beta}$ relies on the proximal distance principle and reduces to least squares.

To shed light on the statistical performance of L_2E regression with Lasso, MCP, and

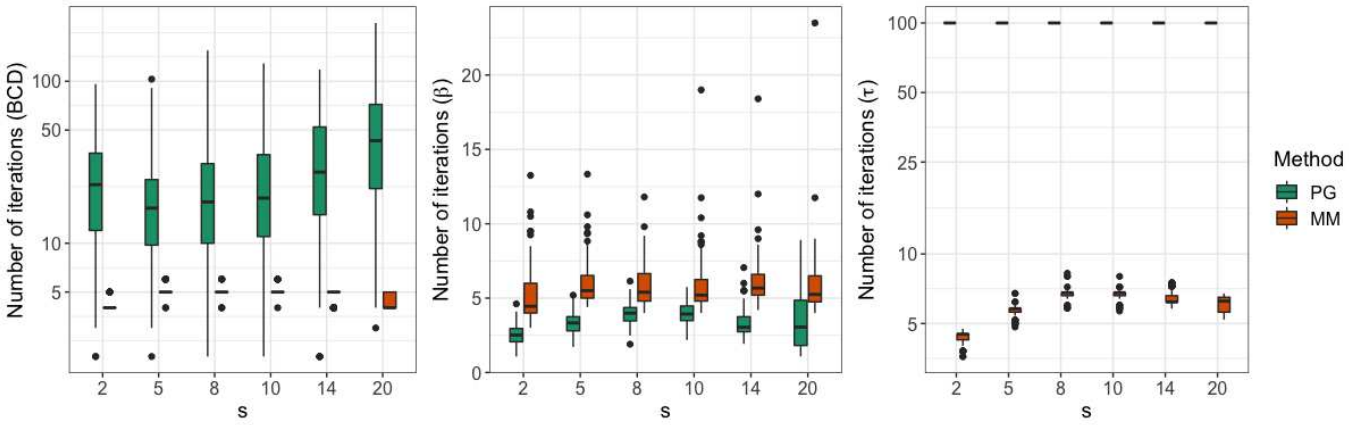

Figure 5: Boxplots of the mean number of outer block descent iterations (left panel), the mean number of inner iterations for updating $\boldsymbol{\beta}$ per outer iteration (middle panel), and the mean number of inner iterations for updating τ per outer iteration (right panel). All plots refer to the experiment summarized in Figure 4.

distance penalties, we undertake a small simulation study involving a sparse coefficient vector $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 0, \cdots, 0)^{\mathsf{T}} \in \mathbb{R}^{50}$ and a design matrix $\boldsymbol{X} \in \mathbb{R}^{200\times 50}$ whose independent entries are standard Gaussian deviates. The response $\boldsymbol{y}$ is simulated as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where components of $\boldsymbol{\epsilon}$ are standard normal noises. We then shift the first m entries of $\boldsymbol{y}$ and the first m rows of $\boldsymbol{X}$ by 5 to produce observations that are outlying with respect to the responses and also high leverage with respect to the predictors. The number of outliers m is chosen from the grid $\{10, 20, 30, 50\}$. For the distance penalization, the ideal choice of the sparsity parameter k is 5. We employ five-fold cross-validation to select the tuning parameters for all three penalties. The sparsity level k for distance penalization is varied over the grid $\{3, 5, 7, 9, 11, 13, 15\}$, and the penalty constant ρ is set to 10^8 to enforce the desired sparsity as discussed in Section 2.3. We initialize L_2E estimation by setting $\boldsymbol{\beta}_0 = \mathbf{0}$ and $\eta_0 = -\log[\mathrm{MAD}(\boldsymbol{y})]$. All performance metrics depend on 100 replicates. These met-

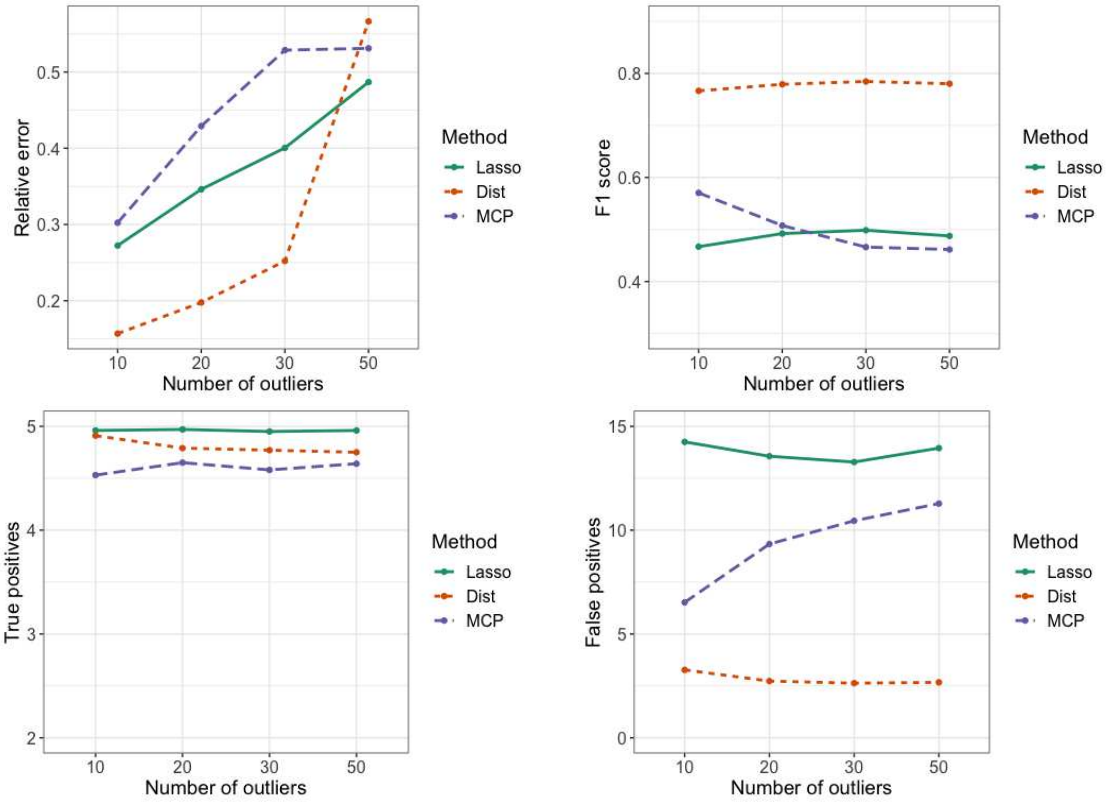

Figure 6: Simulation results for sparse regression under different numbers of outliers. Average performance based on 100 replicates for each method.

rics include: (a) estimation accuracy (measured by the relative error compared to the true $\boldsymbol{\beta}$), (b) support recovery (measured by the F1 score), (c) the number of true positives, and (d) the number of false positives. The F1 score (harmonic mean of precision and recall) accounts for both true and false positives and takes on values in $[0, 1]$, with a higher score indicating better support recovery.

Figure 6 shows the performance of the Lasso, MCP, and distance penalties in robust sparse regression with the L_2E loss under different numbers of outliers. Estimation degrades for all three methods as the number of outliers increases. Distance penalization consistently achieves a lower relative error than Lasso and MCP, except for $m = 50$, where all methods

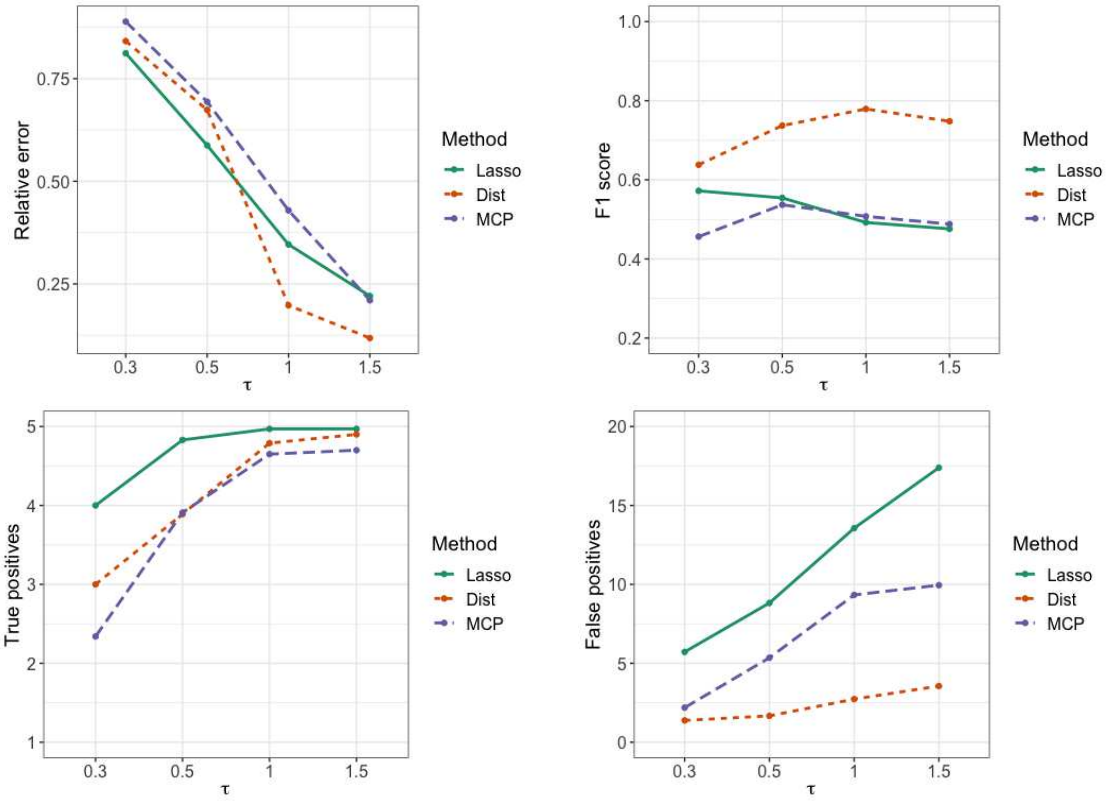

Figure 7: Simulation results for sparse regression under different noise levels. Average performance based on 100 replicates for each method.

produce unacceptable estimates. In support recovery, distance penalization consistently delivers a much higher F1 score than Lasso and MCP. The two plots in the bottom row of Figure 6 highlight the difference in support recovery among the three methods. Lasso identifies the most true positives but suffers from the most false positives in each scenario. MCP selects fewer irrelevant variables compared to Lasso but misses some true positives. In contrast, distance penalization identifies a number of true positives comparable to Lasso while maintaining a much lower false positive rate.

In the second experiment, we compare the performance of the different analysis methods (Lasso, MCP, and distance penalization) under different noise levels. We fix the number

of outliers at $m = 20$ and sample the precision parameter τ over the grid $\{0.3, 0.5, 1, 1.5\}$. A small value of τ represents a high noise level. We use the rules of our first experiment to produce outliers, select tuning parameters, and initialize L_2E estimation. Figure 7 summarizes our analysis results under different noise levels. As expected, the estimation errors of all three methods decrease as the value of τ increases. Distance penalization outperforms Lasso and MCP in estimation accuracy when the noise level is relatively low ($\tau \geq 1$). In addition, distance penalization compares favorably with Lasso and MCP in F1 score across different noise levels. The plots of true and false positives provide detailed insight into the support recovery of the different methods. All methods achieve a larger number of true positives as the value of τ increases, with Lasso leading the others. However, Lasso is plagued by an increasingly large number of false positives as the value of τ increases. Distance penalization achieves a smaller number of false positives, is less sensitive to the noise than Lasso and MCP, and stands out among the three methods in support recovery. This sparse regression example emphasizes the flexibility of L_2E regression in accommodating different penalization methods and the advantages of distance penalization in both estimation accuracy and structure recovery.

6 Real Data Application

To illustrate the application of L_2E regression in unconstrained robust multivariate regression and its effectiveness in detecting outliers, we now turn to the Hertzsprung-Russell diagram data of star cluster CYG OB1 investigated in Rousseeuw and Leroy (2005); Scott (2001); Scott and Wang (2021). This data set includes two variables collected from 47 stars

in the direction of Cygnus. The predictor variable is the logarithm of the temperature at the star's surface, and the response variable is the logarithm of its light intensity. Though small, this data set is commonly used in robust regression owing to its four known outliers – four bright giant stars observed at low temperatures (Vansina and De Greve, 1982).

In this example, the penalty term $\phi(\boldsymbol{\beta}) = 0$. Therefore, the MM update of $\boldsymbol{\beta}$ reduces to a standard least squares problem solvable by many efficient algorithms. In our implementation, we invoke the `lm` function in the R package `stats` (R Core Team, 2020). We initialize $\boldsymbol{\beta}_0 = \mathbf{0}$ and $\eta_0 = -\log[\text{MAD}(\boldsymbol{y})]$. The left panel in Figure 8 displays the fitted L_2E regression model. In comparison with ordinary least squares, L_2E successfully reduces the influence of the four outliers and fits the remaining data points well. The converged weights $w_i = e^{-\tau^2 r_i^2/2}$, where r_i denotes the i-th L_2E residual, serve as a diagnostic tool to detect outliers. As discussed in Section 4, a small weight suggests a potential outlier. The histogram of the logarithm of weights in the right panel in Figure 8 clearly identifies the four outliers. These are colored in red in the scatter plot in the left panel. As a practical matter, we tried different initializations of $\boldsymbol{\beta}$ in L_2E estimation. Different initial values could potentially lead to different estimates. A direct and simple way to compare initializations is to rank their converged L_2E losses (7). In this real data example, the neutral initialization $\boldsymbol{\beta}_0 = \mathbf{0}$ yields the smallest L_2E loss.

7 Discussion

Because robust structured regression is resistant to the undue influence of outliers, it is valuable in many noisy data applications. The L_2E computational framework (Chi and

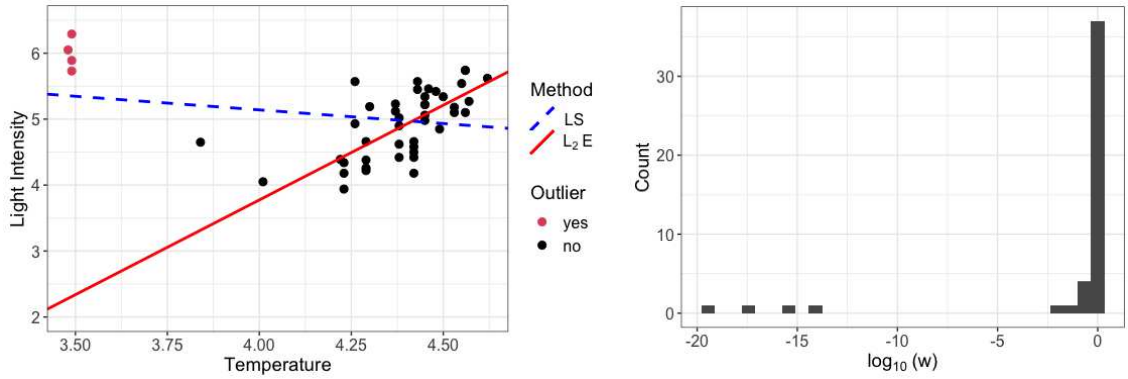

Figure 8: Fitted regression models from L_2E and LS for the Hertzsprung-Russell Diagram Data (left panel). The four known outliers are successfully detected by the L_2E according to the histogram of the resulting weights (right panel).

Chi, 2022) for robust structured regression has the advantage of allowing the simultaneous estimation of regression coefficients and precision. This paper retains the overall strategy of block descent but introduces several non-trivial improvements. We introduce an MM algorithm based on a sharp majorization to accelerate convergence. Each MM update of $\boldsymbol{\beta}$ reduces to penalized least squares and can be readily handled by existing regression solvers. Although this plug-and-play tactic already formed part of the proximal gradient algorithm in Chi and Chi (2022), our tight majorization leads to better results. We also reparameterize precision to avoid box constraint and update the new precision parameter by an approximate Newton's method. The computational cost per iterate remains the same, but again the number of iterations until convergence drops considerably. Finally, we extend penalization to distance and nonconvex penalties. These steps lead to better statistical performance and model selection.

We demonstrate the merits of our refined computational framework through a rich set of simulation examples, including isotonic regression, convex regression, sparse regression,

and trend filtering, and a real data application to unconstrained multivariate regression. Given the same penalties, our simulation results show that the new algorithms outperform the original ones in both computational speed and estimation accuracy. Distance penalties to sparsity sets, in particular, show competitive advantages in both estimation accuracy and model selection. The real data example illustrates the convenience of using the refined framework to identify outliers. Overall, the innovations introduced here make L_2E an attractive tool for robust regression.

Supplementary Material

Supplementary materials and code for this article are available online. The supplement.pdf file contains the two simulation examples of convex regression and trend filtering under the L_2E criterion. The L2E-code.zip file includes code for implementing the L_2E isotonic regression and reproducing Figures 2 and 3 in the paper. To implement other L_2E regression methods in the article, we refer readers to the eponymous `L2E` R package on the CRAN.

Acknowledgement

The authors are grateful to the editor, the associate editor, and the two referees for their helpful comments and suggestions. The authors thank Lisa Lin for her help with the R package.

Funding

Lange's work is supported by the United States Public Health Service (USPHS) grants GM53275 and HG006139. Chi's work is partly supported by the National Science Foundation (NSF) grant DMS-2201136 and National Institutes of Health (NIH) grant R01GM135928.

Conflict of interest

The authors report there are no competing interests to declare.

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*, 226–248.

Alvarez, E. E. and Yohai, V. J. (2012), "M-estimators for isotonic regression," *Journal of Statistical Planning and Inference*, 142, 2351–2368.

Barlow, R. E. and Brunk, H. D. (1972), "The isotonic regression problem and its dual," *Journal of the American Statistical Association*, 67, 140–147.

Bauschke, H. H. and Combettes, P. L. (2011), *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, vol. 408, Springer.

Blanchet, J., Glynn, P. W., Yan, J., and Zhou, Z. (2019), "Multivariate distributionally robust convex regression under absolute error loss," *Advances in Neural Information Processing Systems*, 32, 11817–11826.

Breheny, P. and Huang, J. (2011), "Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection," *Annals of Applied Statistics*, 5, 232–253.

Chi, E. C., Zhou, H., and Lange, K. (2014), "Distance majorization and its applications," *Mathematical Programming*, 146, 409–436.

Chi, J. T. and Chi, E. C. (2022), "A User-Friendly Computational Framework for Robust Structured Regression with the L_2 Criterion," *Journal of Computational and Graphical Statistics*, 1–12.

de Leeuw, J., Hornik, K., and Mair, P. (2010), "Isotone optimization in R: Pool-adjacent-violators algorithm (PAVA) and active set methods," *Journal of Statistical Software*, 32, 1–24.

de Leeuw, J. and Lange, K. (2009), "Sharp quadratic majorization in one dimension," *Computational Statistics & Data Analysis*, 53, 2471–2484.

Hoerl, A. E. and Kennard, R. W. (1970), "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, 12, 55–67.

Huber, P. J. (1992), "Robust estimation of a location parameter," in *Breakthroughs in Statistics*, Springer, pp. 492–518.

Keys, K. L., Zhou, H., and Lange, K. (2019), "Proximal distance algorithms: Theory and practice," *The Journal of Machine Learning Research*, 20, 2384–2421.

Landeros, A., Padilla, O. H. M., Zhou, H., and Lange, K. (2020), "Extensions to the proximal distance method of constrained optimization," *arXiv preprint arXiv:2009.00801*.

Lange, K. (2016), *MM Optimization Algorithms*, SIAM.

Lange, K., Hunter, D. R., and Yang, I. (2000), "Optimization transfer using surrogate objective functions," *Journal of Computational and Graphical Statistics*, 9, 1–20.

Liu, X., Chi, J., Lin, L., Lange, K., and Chi, E. (2022), *L2E: Robust Structured Regression via the L2 Criterion*, R package version 2.0 — For new features, see the 'Changelog' file (in the package source).

Lozano, A. C., Meinshausen, N., and Yang, E. (2016), "Minimum distance lasso for robust high-dimensional regression," *Electronic Journal of Statistics*, 10, 1296–1340.

Nguyen, N. H. and Tran, T. D. (2012), "Robust lasso with missing and grossly corrupted observations," *IEEE Transactions on Information Theory*, 59, 2036–2058.

R Core Team (2020), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Rousseeuw, P. J. and Leroy, A. M. (2005), *Robust Regression and Outlier Detection*, John Wiley & Wons.

Scott, D. W. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization*, John Wiley & Sons.

— (2001), "Parametric statistical modeling by minimum integrated square error," *Technometrics*, 43, 274–285.

Scott, D. W. and Wang, Z. (2021), "Robust multiple regression," *Entropy*, 23, 88.

Seijo, E. and Sen, B. (2011), "Nonparametric least squares estimation of a multivariate convex regression function," *The Annals of Statistics*, 39, 1633–1657.

She, Y. and Owen, A. B. (2011), "Outlier detection using nonconvex penalized regression," *Journal of the American Statistical Association*, 106, 626–639.

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

Van Ruitenburg, J. (2005), "Algorithms for parameter estimation in the Rasch model," *Measurement and Research Department Reports*, 4, 116.

Vansina, F. and De Greve, J. (1982), "Close binary systems before and after mass transfer," *Astrophysics and Space Science*, 87, 377–401.

Wang, H., Li, G., and Jiang, G. (2007), "Robust regression shrinkage and consistent variable selection through the LAD-Lasso," *Journal of Business & Economic Statistics*, 25, 347–355.

Wang, X., Jiang, Y., Huang, M., and Zhang, H. (2013), "Robust variable selection with exponential squared loss," *Journal of the American Statistical Association*, 108, 632–643.

Warwick, J. and Jones, M. (2005), "Choosing a robustness tuning parameter," *Journal of Statistical Computation and Simulation*, 75, 581–588.

Xu, J., Lange, K., and Chi, E. (2017), "Generalized linear model regression under distance-to-set penalties," in *Advances in Neural Information Processing Systems*, pp. 1386–1396.

Yang, E., Lozano, A. C., and Aravkin, A. (2018), "A general family of trimmed estimators for robust high-dimensional data analysis," *Electronic Journal of Statistics*, 12, 3519–3553.

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