A Sharper Computational Tool for L_2E Regression

Xiaoqian Liu

UT MD Anderson Cancer Center

Collaborators:

Eric Chi, Rice University
Kenneth Lange, UCLA

Oct. 5, 2023

Motivation

The classical linear regression:

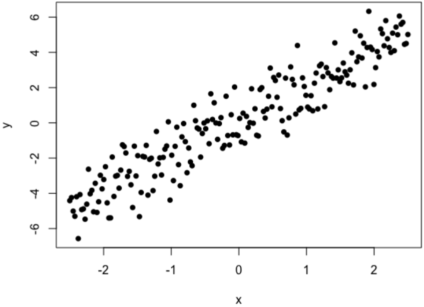
$$\mathrm{y} = \mathrm{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- $\mathrm{y} \in \mathbb{R}^n, \mathrm{X} \in \mathbb{R}^{n \times p}$
- $\boldsymbol{\epsilon} \in \mathbb{R}^n \sim N(0, \tau^{-2}\mathrm{I}_n)$
- Least Square estimator / Maximum Likelihood Estimator (MLE):

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathrm{y} - \mathrm{X}\boldsymbol{\beta}\|_2^2$$

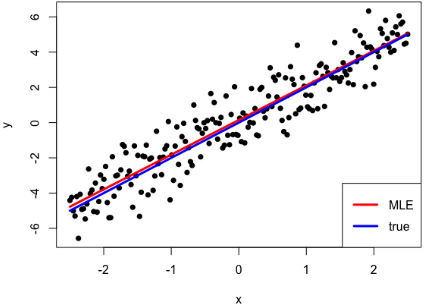
Motivation

An ideal data set:



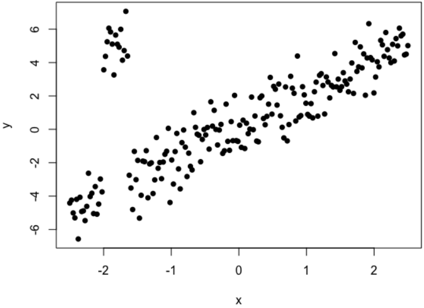
Motivation

An ideal data set:



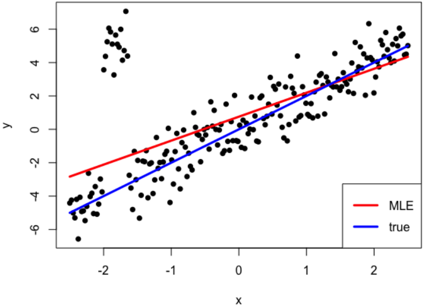
Motivation

In reality, data could be contaminated (outliers!).



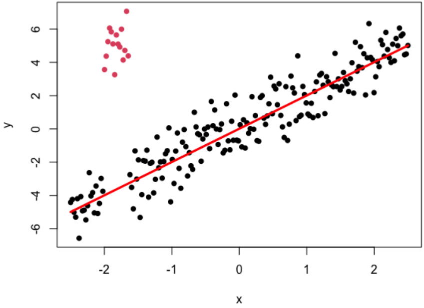
Motivation

In reality, data could be contaminated (outliers!).



Motivation

Our aims: **robust estimation** + **outlier detection** + **structure recovery**



Overview

1. L2E regression
   - L_2E criterion
   - Structured L_2E model
2. Computational framework
   - Updating the vector of coefficients
   - Updating the precision parameter
3. Examples
4. Discussion

L_2E criterion

L_2-distance estimation (L_2E) (Scott, 2001)

Seek a parametric model $f(x \mid \boldsymbol{\theta})$ under a minimum distance criterion (minimum integrated square error)

$$\min_{\boldsymbol{\theta}} \int [f(x \mid \boldsymbol{\theta}) - f(x)]^2 \, dx \tag{1}$$

$$\begin{aligned} & \int [f(x \mid \boldsymbol{\theta}) - f(x)]^2 \, dx \\ = \; & \int f(x \mid \boldsymbol{\theta})^2 \, dx - 2 \int f(x \mid \boldsymbol{\theta}) f(x) \, dx + \int f(x)^2 \, dx \end{aligned}$$

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

L_2E v.s. MLE

Suppose $X \sim N(\mu, 1)$, then

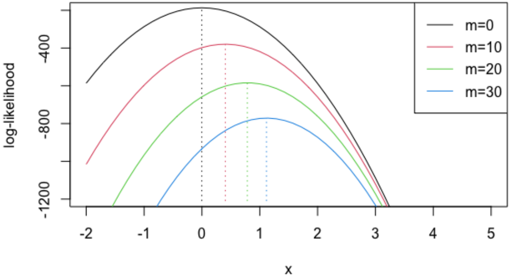
$$\hat{\mu}_{\mathsf{L_2E}} = \mathrm{argmin}_{\mu} \frac{1}{2\sqrt{\pi}} - \frac{2}{n} \sum_{i=1}^{n} f(x_i \mid \mu)$$

$$\hat{\mu}_{\mathsf{MLE}} = \mathrm{argmax}_{\mu} \sum_{i=1}^{n} \log f(x_i \mid \mu)$$

- L_2E maximizes the sum of the densities
- MLE maximizes the product of the densities.

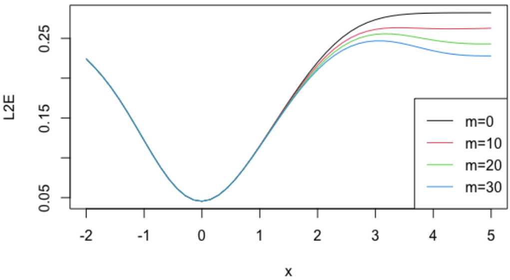
L_2E v.s. MLE

Consider a sample of size 100 from $N(0,1)$ with m additional data points from a contamination desity $N(5,1)$.



L_2E v.s. MLE

Consider sample of size 100 from $N(0, 1)$ with m additional data points from a contamination desity $N(5, 1)$.



L_2E regression

Assume a normal model:

- $y_i \mid X_i = x_i \sim N(x_i^{\mathsf{T}}\boldsymbol{\beta}, \tau^{-2})$
- $\boldsymbol{\theta} = (\boldsymbol{\beta}, \tau)$

$$f(y_i \mid \boldsymbol{\beta}, \tau) \quad = \quad \frac{\tau}{\sqrt{2\pi}} e^{-\frac{\tau^2 r_i^2}{2}} \quad \text{with} \quad r_i = y_i - \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}$$

L_2E loss:

$$h(\boldsymbol{\beta}, \tau) = \frac{\tau}{2\sqrt{\pi}} - \frac{\tau}{n}\sqrt{\frac{2}{\pi}} \sum_{i=1}^{n} e^{-\frac{\tau^2 r_i^2}{2}} \tag{3}$$

Structured L_2E

Structured L_2E regression:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p, \tau \in \mathbb{R}_+} h(\boldsymbol{\beta}, \tau), \quad \text{subject to } \boldsymbol{\beta} \in C \tag{4}$$

Examples of C:

- $C = \{\boldsymbol{\beta} \in \mathbb{R}^p : \beta_1 \leq \cdots \leq \beta_p\}$ (iosotonic regression)
- $C = \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}\|_0 \leq k\}$ (sparse regression)

An alternative formulation of (4):

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p, \tau \in \mathbb{R}_+} h(\boldsymbol{\beta}, \tau) + \psi(\boldsymbol{\beta}), \tag{5}$$

where $\psi(\boldsymbol{\beta})$ is either the indicator function of C or a non-smooth penalty function such as Lasso.

Structured $\mathrm{L}_2\mathrm{E}$

A computational framework by block descent (Chi and Chi, 2022):

- Update $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}^{(k+1)} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}}\, h(\boldsymbol{\beta}, \tau^{(k)}) + \psi(\boldsymbol{\beta})$$

- Update τ:

$$\tau^{(k+1)} = \underset{\boldsymbol{\tau} \in \mathbb{R}^+}{\operatorname{argmin}}\, h(\boldsymbol{\beta}^{(k+1)}, \tau)$$

Structured L_2E

Our contributions:

	Chi and Chi (2022)	Our work (Liu et al., 2023)
update β	proximal gradient	(sharp) MM
update τ	proximal gradient	reparameterization & Newton
penalization	convex	distance penalization

Structured L_2E — Update β

Majorization-Minimization (Lange et al., 2000; Lange, 2016)

Goal: Minimize a function $f(\mathrm{x})$

A surrogate function $g(\mathrm{x} \mid \tilde{\mathrm{x}})$ majorizes a function $f(\mathrm{x})$ if

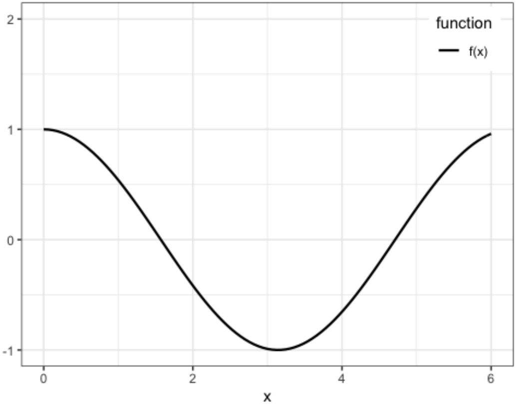
- tangency: $f(\tilde{\mathrm{x}}) = g(\tilde{\mathrm{x}} \mid \tilde{\mathrm{x}})$
- domination: $f(\mathrm{x}) \leq g(\mathrm{x} \mid \tilde{\mathrm{x}})$ for all x

The MM iterate:

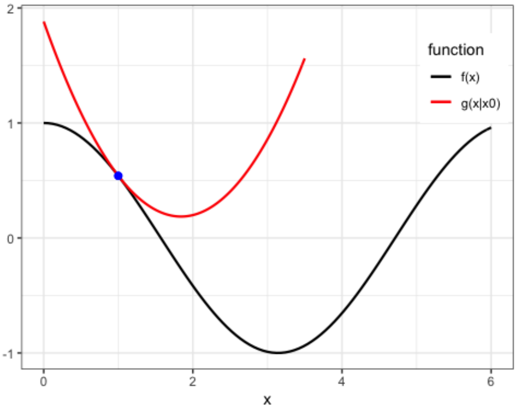
$$\mathrm{x}^{+} = \underset{\mathrm{x}}{\operatorname{argmin}}\, g(\mathrm{x} \mid \tilde{\mathrm{x}})$$

- monotonicity: $f(\mathrm{x}^{+}) \leq g(\mathrm{x}^{+} \mid \tilde{\mathrm{x}}) \leq g(\tilde{\mathrm{x}} \mid \tilde{\mathrm{x}}) = f(\tilde{\mathrm{x}})$

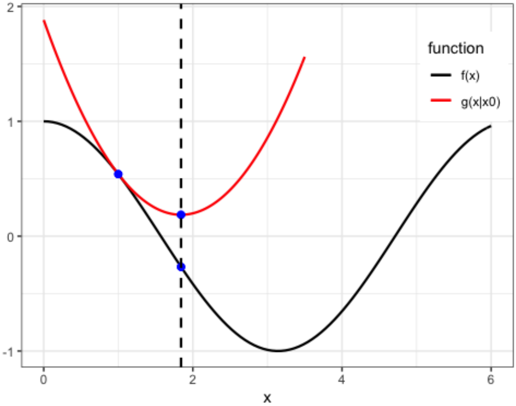
Structured L_2E — Update β



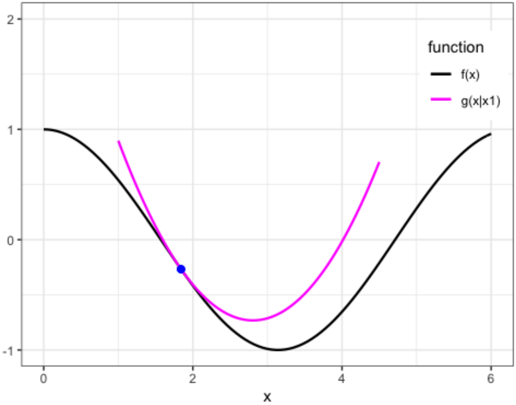
Structured L_2E — Update β



Structured L_2E — Update β

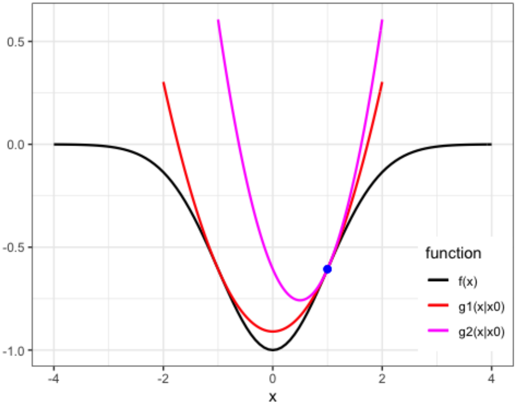


Structured L_2E — Update β



Structured L_2E — Update β

What makes a majorization better than another?



- $g_1(x|x_0) \rightarrow$ a sharp majorization

Structured L_2E — Update β

L_2E loss:

$$h(\beta, \tau) = \frac{\tau}{2\sqrt{\pi}} - \frac{\tau}{n}\sqrt{\frac{2}{\pi}}\sum_{i=1}^{n} e^{-\frac{\tau^2 r_i^2}{2}}$$

- $f(u) = -\exp(-u)$ is concave

A sharp quadratic univariate majorization w.r.t. r^2:

$$-exp(-\frac{\tau^2 r^2}{2}) \leq -\exp(-\frac{\tau^2 \tilde{r}^2}{2}) + \frac{\tau^2}{2}\exp(-\frac{\tau^2 \tilde{r}^2}{2})(r^2 - \tilde{r}^2)$$

Structured L_2E — Update β

Majorization for the L_2E loss:

$$g(\boldsymbol{\beta}|\tilde{\boldsymbol{\beta}}) = \frac{\tau}{2\sqrt{\pi}} + \frac{\tau^3}{\sqrt{2\pi}n} \sum_{i=1}^{n} w_i(y_i - \mathrm{x}_i^{\mathsf{T}}\boldsymbol{\beta})^2$$

- $w_i = \exp(-\frac{\tau^2(y_i - \mathrm{x}_i^{\mathsf{T}}\tilde{\boldsymbol{\beta}})^2}{2})$
- Weights w_i based on residuals from last iterate
- Effect: Downweight points as outliers when $\tilde{r}_i = y_i - \mathrm{x}_i^{\mathsf{T}}\tilde{\boldsymbol{\beta}}$ is large

Structured $\mathrm{L_2E}$ — Update $\boldsymbol{\beta}$

Recall for updating $\boldsymbol{\beta}$:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\, h(\boldsymbol{\beta},\tau)+\psi(\boldsymbol{\beta})$$

MM iterates for updating $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}^{+}=\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\,\frac{1}{2}\|\tilde{\mathrm{y}}-\tilde{\mathrm{X}}\boldsymbol{\beta}\|_2^2+\psi(\boldsymbol{\beta})$$

- $\tilde{\mathrm{y}}=\sqrt{\mathrm{W}}\mathrm{y}$, $\tilde{\mathrm{X}}=\sqrt{\mathrm{W}}\mathrm{X}$, W weight matrix
- A penalized least squares problem
- General, simple, and flexible ("plug-and-play")

Structured L$_2$E — Update τ

Updating τ:

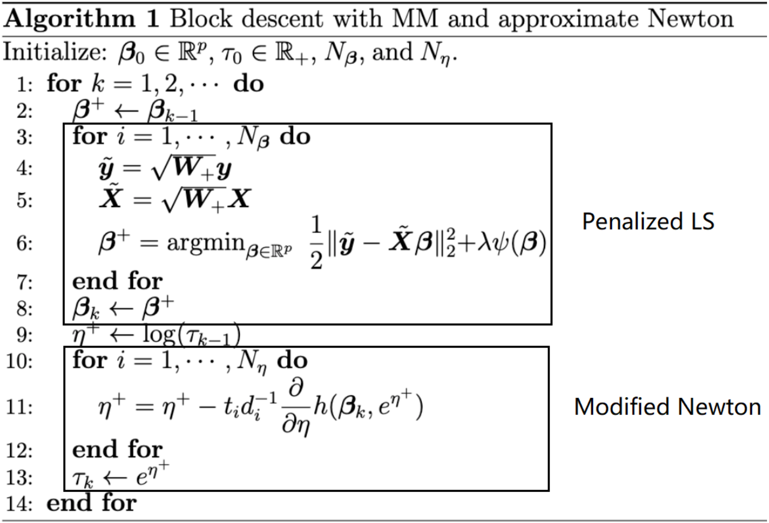
$$\underset{\tau \in \mathbb{R}^+}{\operatorname{argmin}} \frac{\tau}{2\sqrt{\pi}} - \frac{\tau}{n}\sqrt{\frac{2}{\pi}} \sum_{i=1}^{n} e^{-\frac{\tau^2 r_i^2}{2}}$$

- Reparameterize $\tau = e^{\eta} \implies$ no constraint on η
- An approximate Newton method

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

where $t_k > 0$ is a stepsize parameter chosen via backtracking, and d_k is an approximation of the second derivative $\frac{\partial^2}{\partial \eta^2} h(\boldsymbol{\beta}, e^{\eta})$.

Structured L_2E — computational framework



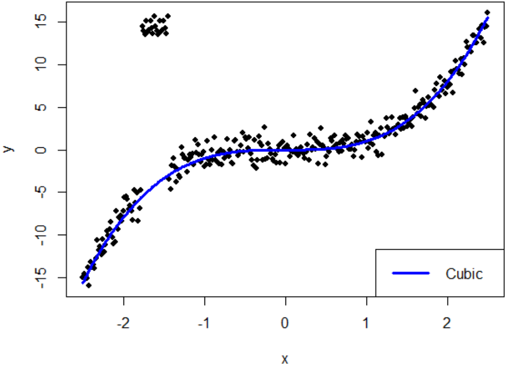
Algorithm 1 Block descent with MM and approximate Newton

```
Initialize: $\beta_0 \in \mathbb{R}^p$, $\tau_0 \in \mathbb{R}_+$, $N_\beta$, and $N_\eta$.
for $k = 1, 2, \cdots$ do
    $\beta^+ \leftarrow \beta_{k-1}$
    for $i = 1, \cdots, N_\beta$ do
        $\tilde{y} = \sqrt{W_+} y$
        $\tilde{X} = \sqrt{W_+} X$
        $\beta^+ = \text{argmin}_{\beta \in \mathbb{R}^p} \ \frac{1}{2} \|\tilde{y} - \tilde{X}\beta\|_2^2 + \lambda\psi(\beta)$
    end for
    $\beta_k \leftarrow \beta^+$
    $\eta^+ \leftarrow \log(\tau_{k-1})$
    for $i = 1, \cdots, N_\eta$ do
        $\eta^+ = \eta^+ - t_i d_i^{-1} \frac{\partial}{\partial \eta} h(\beta_k, e^{\eta^+})$
    end for
    $\tau_k \leftarrow e^{\eta^+}$
end for
```

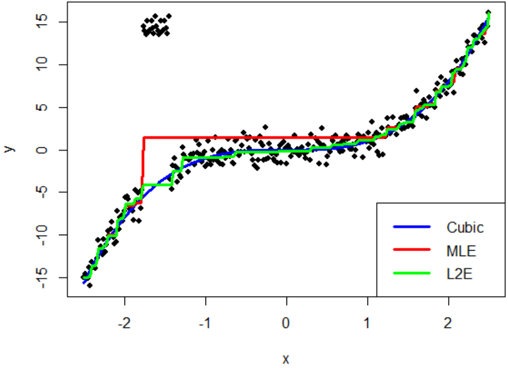
Penalized LS

Modified Newton

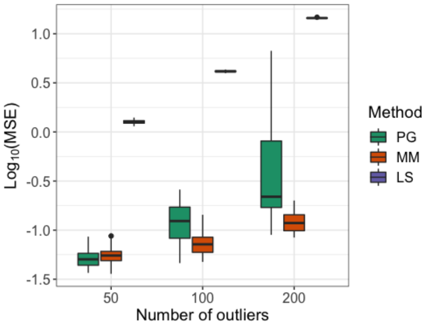
Isotonic regression

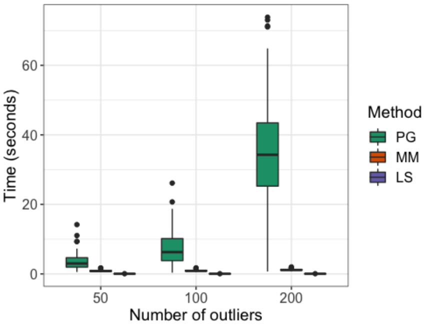


Isotonic regression



Isotonic regression





Distance penalization

For a constrained optimization problem

$$\min_{\boldsymbol{\beta}} \ \ell(\boldsymbol{\beta}) \ \text{ subject to } \boldsymbol{\beta} \in C$$

Distance penalization (Chi et al., 2014; Xu et al., 2017)

$$\psi(\boldsymbol{\beta}) = \frac{1}{2}\text{dist}(\boldsymbol{\beta}, C)^2 = \min_{\mathrm{u} \in C} \frac{1}{2}\|\boldsymbol{\beta} - \mathrm{u}\|_2^2 \tag{6}$$

The resulting optimization problem:

$$\min_{\boldsymbol{\beta}} \ \ell(\boldsymbol{\beta}) + \frac{\rho}{2}\text{dist}(\boldsymbol{\beta}, C)^2$$

- If $\rho \to \infty$, then $\boldsymbol{\beta} \in C$ (recover the constrained solution)
- ρ is assigned a large value in practice

Distance penalization

Advantages of distance penalization:

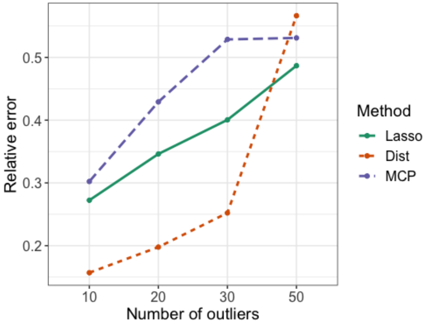
- A general definition
  - diverse structures: sparsity, order constraint, shape constraint
  - multiple constraints: $\frac{1}{2}\sum_{i=1}^{I} w_i \text{dist}(\boldsymbol{\beta}, C_i)^2$
  - fusion constraint: $\mathbf{L}\boldsymbol{\beta} \in C$ (Landeros et al., 2020)
- Only projection onto the constraint set is necessary
  - no requirement that ℓ or C is convex
  - no requirement that ℓ is differentiable
- An efficient proximal distance algorithm (Keys et al., 2019)
  - $\text{dist}(\boldsymbol{\beta}, C)^2 \leq \|\boldsymbol{\beta} - \mathcal{P}_C(\tilde{\boldsymbol{\beta}})\|_2^2$

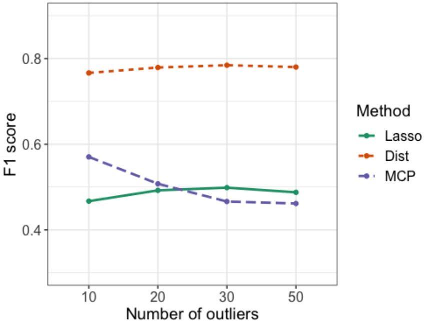
Sparse regression

$$\mathsf{y} = \mathsf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

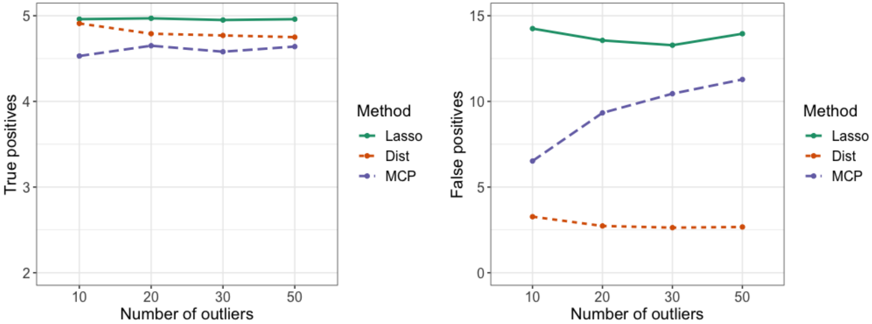
- $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 0, \cdots, 0)^{\mathsf{T}} \in \mathbb{R}^{50}$
- $\mathsf{X} \in \mathbb{R}^{200\times 50}$ from standard normal distribution
- $\boldsymbol{\epsilon}$ standard normal noise
- Shift the first m entries of y and the first m rows of X by 5 to produce outliers

Sparse regression





Sparse regression



Outlier detection

Mutivariate regression:

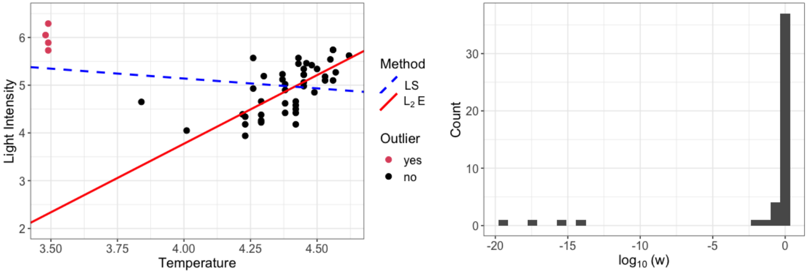


Figure: Fitted regression models from L_2E and LS for the Hertzsprung-Russell Diagram Data (left panel). The four known outliers are successfully detected by the L_2E according to the histogram of the resulting weights (right panel).

Discussion

Take-home message:

- Structured L_2E regression for
  robust estimation + outlier detection + structure recovery
- A sharper computational framework
  - general: various constraints/penalties
  - simple and flexible: "plug-and-play"

Once you have a procedure for solving some structured regression problem, you can use our framework to robustify it!

Thank You!

Reference I

Chi, E. C., Zhou, H., and Lange, K. (2014). Distance majorization and its applications. *Mathematical Programming*, 146(1):409–436.

Chi, J. T. and Chi, E. C. (2022). A user-friendly computational framework for robust structured regression with the l2 criterion. *Journal of Computational and Graphical Statistics*, pages 1–12.

Keys, K. L., Zhou, H., and Lange, K. (2019). Proximal distance algorithms: Theory and practice. *The Journal of Machine Learning Research*, 20(1):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):1–20.

Reference II

Liu, X., Chi, E. C., and Lange, K. (2023). A sharper computational tool for regression. *Technometrics*, 65(1):117–126.

Scott, D. W. (2001). Parametric statistical modeling by minimum integrated square error. *Technometrics*, 43(3):274–285.

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