Sparse Learning Reloaded: The Good, the Bad and the Quantile

Raphaël Sala

Internship, Section of Mathematics, University of Geneva
Supervised by Sylvain Sardy

September 1, 2025

Abstract

Under a sparsity hypothesis it is common to penalize feature coefficients through a tuning parameter $\lambda > 0$. We introduce a principled and computationally efficient rule for selecting λ that applies to a broad family of predictive models including neural networks and, as a special case, linear regression. Our framework treats the choice of λ, the loss, and the penalty in a unified manner, with particular emphasis on non-convex penalties that sharpen variable selection beyond the convex ℓ_1-norm [SvCM25].

Keywords— support recovery, penalty, sparsity, non convex optimization, quantile universal threshold (QUT), proximal algorithms, variable selection, duality gap, Φ-convexity, artificial neural networks

1 Context

1.1 From prediction to learners

Machine-learning models, called learners, seek patterns from data to predict future observations. Formally, a learner μ is a map $\mathcal{X} \to \mathcal{Y}$. In regression, $\mathcal{Y} = \mathbb{R}$; in classification it's generally a discrete set. The methodology developed here is entirely inspired from the original article [SvCM25], where both regression and classification are treated. The present work focuses on the regression case.

We begin with the simplest (yet non-trivial) learner: the linear model. Let $X \in \mathbb{R}^{n\times p}$ denote the design matrix whose i-th row $x_i^\top$ collects p features for observation i, and let $y \in \mathbb{R}^n$ be the response vector. A linear learner [1], with $\theta \in \mathbb{R}^p$ its parameters, writes $\mu_\theta : x \mapsto x^\top\theta$, so that the predicted response vector is $X\theta$. We want to find the "best" parameters. Although "best" can be defined in many ways, the literature typically casts it as minimizing a chosen loss to get prediction closest to real data. The Euclidean-norm, ℓ_2-norm, owes its popularity from its mathematical convenience : convexity, differentiability, and closed-form solution-as well as its optimality under the Gaussian noise assumption (i.e., it is the maximum likelihood estimator when errors are i.i.d. Gaussian) [Tib96, HTF09] yielding the least-squares estimator $\mathrm{RSS} : \theta \mapsto \|y - X\theta\|_2^2$. When $n \geq p$ and $X^\top X$ is invertible, analytic geometry gifts us the closed form

$$\theta^{LS} = \arg\min_\theta RSS(\theta) = (X^\top X)^{-1}X^\top y$$

1.2 Sparse hypothesis

Unfortunately, life is rarely so kind due to the multicollinearity inflating variances [2], p may rival or even dwarf n, and for scientific insight we often believe that only a small subset of the p features truly matters [BvdG11]. For instance biologists need to interpret or apply a gene shortlist, not ten thousand noisy coefficients [GE03]. So to get a more insightful and general model, we add the hypothesis that our data are s-sparse, i.e. only $s \leq p$ variables are useful to predict the target y. By denoting the support of vector $\mathrm{supp} : \theta \in \mathbb{R}^p \mapsto \{i \in [\![1, p]\!], \theta_i \neq 0\}$, θ is s-sparse iff $|\operatorname{supp}(\theta)| \leq s$. Regrettably, least squares tells us nothing about which coefficients can be discarded. For instance, if $p > n$ the design matrix X possesses a non-trivial kernel, so any vector added in that null space leaves the residual sum of squares unchanged yielding an entire affine family of equally good but highly variable solutions [BvdG11].

Hence we need to enrich our model to enhance sparsity. We call the estimator minimizing the error the loss function denoted by f, earlier we had $f = RSS$. We add a penalty g and a coefficient $\lambda \geq 0$ to strengthen its effect. This penalized formulation has become standard in the literature for sparse estimation, with g a sparsity inducing penalty (defined later) [BJMO11, Section 1.2].

Nonetheless, we don't want to penalize all parameters ! Typically, we are interested in promoting sparsity only among the coefficients associated with the input features. We split all parameters as $\theta = (\theta_1, \theta_2)$ where θ_1 is attached to features and θ_2 the remaining.

Example 1: Linear model with intercept

Write $x = (x_1, x_2)^\top \in \mathbb{R}^2$ and consider

$$\mu_{\theta=(\theta_1,\theta_2)}(x) = \theta_{1,1}x_1 + \theta_{1,2}x_2 + \theta_2$$

that is, $\theta_1 = (\theta_{1,1}, \theta_{1,2})^\top$ and θ_2 is the intercept. Penalising θ_1 encourages the slope coefficients to vanish when unnecessary, while leaving the intercept free, to the penalty, prevents a simple shift of the response from being falsely interpreted as sparsity because the intercept acts on all features.

Remark 1: Linear learner with intercept in matrix form

Hence, in the first definition of μ_θ we forgot the intercept. Throughout the remainder we adopt the augmented design

$$\tilde{X} := \begin{bmatrix} X \mid \mathbf{1} \end{bmatrix} \in \mathbb{R}^{n\times(p+1)}, \qquad \theta := \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$$

so that only θ_1 is subject to the penalty g with $\theta = (\theta_1, \theta_2) \in \mathbb{R}^p \times \mathbb{R}$ and our learner is defined as,

$$\mu_\theta : X \in \mathbb{R}^{n\times p} \mapsto X\theta_1 + \mathbf{1}\theta_2 = \tilde{X}\theta \in \mathbb{R}^n$$

[1] We define linear learner without intercept, we explain and add it in the next subsection

[2] Denoting σ^2 the variance of the noise in data. The variance of the precedent estimator is given by $\mathrm{Var}(\theta^{\mathrm{LS}}) = \sigma^2(X^\top X)^{-1}$

Our problem become to find, for λ given,

$$\hat{\theta}_\lambda := \underset{\theta=(\theta_1,\theta_2)\in\mathbb{R}^p\times\mathbb{R}}{\arg\min} \; f(\theta) + \lambda g(\theta_1) \tag{1}$$

1.3 Finding λ

Choosing the tuning parameter λ is delicate. Setting to zero and we are back to the unpenalised loss and send $\lambda \to \infty$ and the penalty strangles every slope coefficient (θ_1), leaving at best an intercept only learner. In small dimensions, generally $n \gg p$ the ritual is to estimate the noise variance σ and plug it into Mallows' C_p [Mal73], AIC [Aka74], BIC [Sch78] and pick the λ that minimizes the criterion. A cousin, the universal soft-threshold from wavelet denoising, prescribes $\lambda = \sigma\sqrt{2\log p}$ for a ℓ_1 penalty and an orthogonal design [DJ94]. All those formulas assume that σ is identifiable and use convex penalty g. However, once the design matrix has more columns than rows, $p > n$, the least squares residual can be driven to zero, turning the classical estimator of noise into an optimistic zero... Consequently, the algebraic niceties that convex penalties enjoy (unique minimizer, piece-wise linear solution path, ...) are lost for most non-convex penalties g.

Hence to overcome to this problem, we usually use K-fold Cross-Validation: split the data, train on $K-1$ folds, predict on the holdout fold, and average the resulting errors. Unfortunately, each fold multiplies the computational budget, the final choice of λ may jitter from one random split to the next, and theoretical analyses show a bias toward too many active variables. In very high dimensions, CV tends to overselect variables and still lacks a fully satisfactory theory for non-convex penalties [AC10]. Bootstrap rules are an alternative, but they are even more computationally intensive and provide little guidance on tuning λ [ET93].

In the next section, we suggest a methodology, based on the null model, to select λ. It needs moderate complexity and use specific loss to deal with the permanent noisy data and non convex sharp penalty.

1.4 Alternative point of view

Before digging deeper, we briefly recall what has already been attempted. A natural starting point can be the sparsity hypothesis, then, we could therefore search for the smallest support reproducing the data, for $\varepsilon > 0$

$$\min_\theta \; |\operatorname{supp}\theta| \quad \text{s.t.} \quad \|y - X\theta\|_2 \le \varepsilon$$

which reduces to exact equality when the observations are noise-free. The above objective is the combinatorial ℓ_0 minimization which is NP-hard and numerically solvable until $p \approx 1000$ [BKM16] and ε depends on σ...

A celebrated workaround is to relax ℓ_0 to its convex envelope ℓ_1, yielding Basis Pursuit or Lasso [CRT06, Tib96]. Compressed sensing theory shows that under suitable incoherence conditions this relaxation retains exact support recovery and exhibits a striking phase transition [DT09].

Yet, even for the Lasso no universally reliable high-dimensional rule exists for picking λ without resorting to cross-validation or knowing σ. For sharper selection non convex penalties (SCAD, MCP, HarderLASSO, ...) improve bias but further complicate tuning that we will encounter. Our forthcoming QUT criterion bridges part of this gap, while still leaving open questions we discuss later.

2 QUT Theory

2.1 Assumptions

For each problem we consider a design matrix X and without loss of generality, we assume X is full column rank and it is standardized, i.e.

$$X^\top \mathbf{1} = \mathbf{0} \quad \text{and} \quad \frac{1}{n}\operatorname{diag}(X^\top X) = \mathbf{1}$$

In the following, we focus on a particular class of loss functions and penalties. However, our methodology remains applicable to a much wider class of functions, provided certain mild assumptions hold. We suggest some assumptions on them but it's interesting to think if we can remove them to enlarge this space.

Let parameters split as explained before $\theta = (\theta_1, \theta_2) \in \mathcal{X}_1 \times \mathcal{X}_2 \subset \mathbb{R}^d$ We work in the regression case,

$$f : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathbb{R} \quad \text{locally Lipschitz and coercive}$$

Then, we don't want to penalize the model if it considers all variables useless,

$$g : \mathcal{X}_1 \to \mathbb{R}^+ \quad \text{locally Lipschitz and } g(\mathbf{0}) = 0$$

In the following, we need to analyze local minima and accept non convex penalties. Therefore, we require local Lipschitz continuity to use the Clarke subdifferential, the generalization to subdifferential, exclusive to convex function.

Definition 2.1: Clarke subdifferential [CLSW98]

Let $\mathcal{X} \subset \mathbb{R}^d$ be open and let $f : \mathcal{X} \to \mathbb{R}$ be locally Lipschitz at $x \in \mathcal{X}$. Denote $\mathrm{conv}(\,\cdot\,)$ the closed convex hull. The Clarke subdifferential of f at x is defined as

$$\partial_{\mathrm{C}} f(x) := \mathrm{conv}\Big\{ \lim_{k\to\infty} \partial f(x_k) \;:\; x_k \to x,\ f \text{ is differentiable at } x_k \Big\}$$

f is supposed to be locally Lipschitz by Rademacher's theorem [CLSW98][Chapter 3.] stating it's almost everywhere differentiable. This explains the definition. Hence, $\partial_C f(x)$ is closed, convex, and bounded.

Proposition 2.1: Clarke-Fermat rule [CLSW98]

Let $f : \mathcal{X} \to \mathbb{R}$ be locally Lipschitz on an open set $\mathcal{X} \subset \mathbb{R}^d$ and let $x^\star \in \mathcal{X}$.
If $x^\star$ is a local minimum of f, then

$$\mathbf{0} \in \partial_{\mathrm{C}} f(x^\star)$$

In the following we omit the subscript C to ∂_{C} for clarity.

2.2 Enhance sparsity from the null model

Modern sparse estimators are judged not only by predictive accuracy but also by their ability to avoid false discoveries, i.e. selecting useless features. Inspired by the universal threshold of [DJ94] which guarantees that soft-thresholding outputs the zero vector with high probability under pure gaussian noise, our goal is to calibrate the penalty level λ under the null model $H_0 : \theta = (\mathbf{0}, \theta_2)$. Doing so, places λ exactly on the detection edge that separates noise from meaningful structure for the chosen pair (f, g).

Definition 2.2: Global and local zero threshold [SvCM25]

We define the (global) zero threshold, λ_0 such as

$$\lambda_0 := \inf\big\{\lambda \geq 0, \quad \hat{\theta}_{\tilde{\lambda}} = (\mathbf{0}, \hat{\theta}_{\tilde{\lambda},2}) \quad \forall \tilde{\lambda} > \lambda\big\}$$

That's to say the λ at the frontier before the null model is the solution. Then, we define the local zero threshold, as the infimum on λ creating a local minimum in $(\mathbf{0}, \hat{\theta}_{\lambda,2})$.

$$\lambda_0^{local} := \inf\big\{\lambda \geq 0, \quad \exists \varepsilon > 0,\ \forall x = (x_1, x_2) \in \mathcal{B}\big((\mathbf{0}, \hat{\theta}_{\lambda,2}), \varepsilon\big),\ f(x) + \lambda g(x_1) \geq f\big(\mathbf{0}, \hat{\theta}_{\lambda,2}\big)\big\}$$

Remark 2: Right continuity of λ_0^{local}

The definition of λ_0^{local}, if the object exists, can be improved to look more like λ_0.

$$\lambda_0^{local} = \inf\big\{\lambda \geq 0, \quad (\mathbf{0}, \hat{\theta}_{\tilde{\lambda},2}) \text{ is a local minimum} \quad \forall \tilde{\lambda} \geq \lambda\big\}$$

Indeed, for $\lambda > \lambda_0^{local}$, by definition of the infimum there exists λ^* such as $\lambda_0^{local} \leq \lambda^* < \lambda$ where $(\mathbf{0}, \theta_{\lambda^*,2})$ is a local minimum of F_{λ^*}. So there exists $\varepsilon > 0$ such as,

$$F_{\lambda^*}(x) \geq F_{\lambda^*}(\mathbf{0}, \hat{\theta}_{\lambda^*,2}), \quad \forall x \in \mathcal{B}\big((\mathbf{0}, \hat{\theta}_{\lambda^*,2}), \varepsilon\big)$$

Then, because $g(\mathbf{0}) = 0$ and the optimal intercept doesn't depend on λ (i.e. $\hat{\theta}_{\lambda^*,2} = \hat{\theta}_{\lambda,2}$), we have $(\mathbf{0}, \hat{\theta}_{\lambda,2})$ is a local minimum for F_λ,

$$F_\lambda(x) \geq F_{\lambda^*}(x) \geq F_{\lambda^*}(\mathbf{0}, \hat{\theta}_{\lambda^*,2}) = F_\lambda(\mathbf{0}, \hat{\theta}_{\lambda,2}), \quad \forall x \in \mathcal{B}\big((\mathbf{0}, \hat{\theta}_{\lambda,2}), \varepsilon\big)$$

Ensuring that $\theta = 0$ is at least a local minimum serves for most optimisation routines (ISTA, Coordinate Descent, ...) converging to nearby stationary points. It certifies an attractor whenever the true signal is weak.

Proposition 2.2

Firstly, we see that for all pairs (f, g) such that λ_0 and λ_0^{local} exist,

$$\lambda_0^{local} \leq \lambda_0$$

Secondly, if f and g are convex then $\lambda_0^{\text{local}} = \lambda_0$.

Proof

Let $F_\lambda := f + \lambda g$ with existing $\lambda_0,\ \lambda_0^{local}$.
Let $\lambda > \lambda_0$, because $\mathbf{0}$ is a global minimiser, it is in particular a local minimizer. Hence $\lambda \geq \lambda_0^{\text{local}}$. Letting $\lambda \downarrow \lambda_0$ yields

$$\lambda_0 \ \geq \ \lambda_0^{\text{local}}$$

If f and g are convex then F_λ is convex. For all $\lambda > \lambda_0^{\text{local}}$, by definition, $\mathbf{0}$ is a local minimizer of F_λ and by convexity it's a global minimizer. Therefore, $\lambda_0^{local} \geq \lambda_0$. From the precedent property, $\lambda_0^{local} \leq \lambda_0$, we get,

$$\lambda_0^{local} = \lambda_0$$

□

" The reason for defining a local zero-thresholding function is that, for non-convex costs, the (global) zero-thresholding function may be difficult to derive. On the contrary the local zero-thresholding function sometimes has a closed form expression. "[SvCM25] Nonetheless, we add the assumption on the Proposition 2.2 of existence. We consider for the suit only these specific pair (f, g).

Definition 2.3: Sparsity inducing penalty

The penalty g is said to be sparsity inducing for f if there exists a finite $\lambda_0^{local} > 0$.

In an introductory machine learning course one usually encounters two regularized least-squares models defined with $f = RSS$ and $g = \|\cdot\|_1$ called LASSO [Tib96] allowing sparsity contrary to its cousin with a different penalty $g = \|\cdot\|_2^2$ called Ridge [HK70]. Whereas the last admits a closed solution form for the problem (1), we don't use it in the sparse setting because it doesn't shrink its coefficients all the way to 0.

Example 2: Ridge is NOT sparsity inducing

Let $y \in \mathbb{R}^n,\ X \in \mathbb{R}^{n\times p},\ f : \theta \in \mathbb{R}^p \mapsto \|y - X\theta\|_2^2$ and $g : \theta \in \mathbb{R}^p \mapsto \|\theta\|_2^2$.
$F_\lambda := f + \lambda g$ is strictly convex and coercive so it admits a unique global (so local) minimum $\hat{\theta}_\lambda$.
We have $X^\top X + \lambda I_p$ is symmetric positive definite so it's invertible, then for all $\lambda > 0$, by differentiability,

$$\partial_\theta F_\lambda(\hat{\theta}_\lambda) = 0 \iff -2X^\top\big(y - X\hat{\theta}_\lambda\big) + 2\lambda\hat{\theta}_\lambda = 0 \iff \hat{\theta}_\lambda = (X^\top X + \lambda I_p)^{-1}X^\top y \neq \mathbf{0}$$

2.3 Quantile Universal Threshold (QUT)

Definition 2.4: Quantile Universal Threshold (QUT) [SvCM25]

Let X a design matrix, an output y and a learner μ^a with a penalty g sparsity inducing. Let $\mathbf{Y}$ the random vector under the null model $H_0 : \theta = (\mathbf{0}, \hat{\theta}_2)$. Define the statistic $\Lambda := \lambda_0^{local}(X, \mathbf{Y})$ with cumulative distribution function $\mathcal{F}_\Lambda$. For $0 < \alpha < 1$, the Quantile Universal Threshold is the upper α-quantile of Λ,

$$\lambda_{QUT} := \mathcal{F}_\Lambda^{-1}(1 - \alpha)$$

[a] The learner is implicit in the loss f, which in turn is implicit in λ_0^{local}.

Theorem 2.1: Local sure screening property [MSH$^+$22]

Under the same assumption than Definition 2.4

$$\mathbb{P}_{H_0}\Big((\mathbf{0}, \hat{\theta}_{\lambda_{QUT},2}) \text{ is a local minimum of } F_{\lambda_{QUT}}\Big) \;\geq\; 1-\alpha$$

Proof

From the Remark 2,

$$\lambda \;\geq\; \Lambda = \lambda_0^{\text{local}}(X, \mathbf{Y}) \quad \Longrightarrow \quad (\mathbf{0}, \hat{\theta}_{\lambda,2}) \text{ is a local minimum of } F_\lambda$$

And if $(\mathbf{0}, \hat{\theta}_{\lambda,2})$ is a local minimum ofF_λ then from what is the infimum, we must have $\lambda \geq \Lambda$. Then,

$$\Big\{(\mathbf{0}, \hat{\theta}_{\lambda_{QUT},2}) \text{ is a local minimum of } F_{\lambda_{QUT}}\Big\} = \{\lambda_{QUT} \geq \Lambda\}$$

Hence,

$$\mathbb{P}_{H_0}\Big(\Big\{(\mathbf{0}, \hat{\theta}_{\lambda_{QUT},2}) \text{ is a local minimum of } F_{\lambda_{QUT}}\Big\}\Big) = \mathbb{P}_{H_0}(\lambda_{QUT} \geq \Lambda) = \mathcal{F}_\Lambda(\lambda_{QUT}) \geq 1-\alpha$$

□

Corollary 2.1

If f, g are convex and g is sparsity inducing then,

$$\mathbb{P}_{H_0}\Big((\mathbf{0}, \hat{\theta}_{\lambda_{QUT},2}) \text{ is a global minimum of } F_{\lambda_{QUT}}\Big) \;\geq\; 1-\alpha$$

Proof

The result is immediate from Proposition 2.2 and Theorem 2.1. □

In practice, an explicit form of the statistic Λ, and a fortiori its $(1-\alpha)$ upper quantile, is seldom available. Fortunately, by Monte-Carlo simulation we can approximate it.

At first sight this seems to require knowing the distribution of the noise in $\mathbf{Y}$, which is typically unknown. The following Examples 3 and 4 shows the necessity to restrict our attention to pairs (f, g) for which Λ is pivotal i.e. it doesn't depends on unknown parameters. The simulation becomes feasible without estimating nuisance parameters, yielding a computationally tractable procedure for computing λ_{QUT}.

Example 3: Expression of Λ for $f = RSS$ and $g = \|\cdot\|_1$ and is NOT pivotal

Let $y \in \mathbb{R}^n,\ X \in \mathbb{R}^{n\times p},\ f : \theta \in \mathbb{R}^{(p+1)} \mapsto \|y - \mu_\theta(X)\|_2^2 = \|y - X\theta_1 - \mathbf{1}\theta_2\|_2^2$ and $g : \theta_1 \in \mathbb{R}^p \mapsto \|\theta_1\|_1$
Remember,

$$\partial_{\theta_1} f(\theta) = -2X^\top(y - X\theta_1 - \mathbf{1}\theta_2) \quad \text{and} \quad \partial_{\theta_1} g(\theta_1) = \bigotimes_{i=1}^{p} \begin{cases} \{-1\} & \text{if } \theta_{1,i} < 0 \\ [-1;1] & \text{if } \theta_{1,i} = 0 \\ \{1\} & \text{if } \theta_{1,i} > 0 \end{cases}$$

The penalty g is sparsity inducing because $F_\lambda := f + \lambda g$ is convex, coercive and locally Lipschitz so,

$$\mathbf{0} \in \partial_{\theta_1} F_\lambda(\theta = (\mathbf{0}, \hat{\theta}_2)) \Longleftrightarrow 2X^\top(y - \mathbf{1}\hat{\theta}_2) \in [-\lambda\,;\lambda]^p \Longleftrightarrow \lambda \geq \|2X^\top(y - \mathbf{1}\hat{\theta}_2)\|_\infty \neq +\infty$$

Λ exists, let's to find it ! Taking a linear learner μ_θ leads to assume

$$\mathbf{Y} \;=\; X\theta_1 + \mathbf{1}\theta_2 \;+ e, \qquad e \sim \mathcal{N}\big(m\mathbf{1}, \sigma^2 I_n\big)$$

and under H_0 one has $\mathbf{Y} = \mathbf{1}\theta_2 + e$.
F_λ is convex so from Proposition 2.2 and the precedent calculation,

$$\lambda_0^{local}(X, \mathbf{Y}) = \lambda_0(X, \mathbf{Y}) = \sup_{\hat{\theta}_2 \in \mathbb{R}} \|\partial_{\theta_1} f(\mathbf{0}, \hat{\theta}_2)\|_\infty$$

Let $u := \frac{e}{\sigma} \sim \mathcal{N}(0, I)$, hence $X^\top u \sim \mathcal{N}(0, X^\top X)$. Then because X is standardized (see Section 2.1), and under H_0,

$$\lambda_0^{local}(X, \mathbf{Y}) = 2 \sup_{\hat{\theta}_2 \in \mathbb{R}} \|X^\top \mathbf{Y} - X^\top \mathbf{1}\hat{\theta}_2\|_\infty = 2\|X^\top \mathbf{Y}\|_\infty = 2\|X^\top e\|_\infty = 2\|X^\top (m + \sigma u)\|_\infty = 2\sigma \|X^\top u\|_\infty$$

That's to say Λ no longer depends on m but still on σ. The statistic is NOT pivotal.

We give up RSS to model problems. We suggest to replace it by $\sqrt{RSS}$ which is pivotal and called in the literature SQRT-LASSO or SR-LASSO [BCW11].

Example 4: Expression of Λ for $f = \sqrt{RSS}$ and $g = \|\cdot\|_1$ and is pivotal

Let $y \in \mathbb{R}^n,\ X \in \mathbb{R}^{n \times p},\ f : \theta \in \mathbb{R}^{(p+1)} \mapsto \|y - \mu_\theta(X)\|_2 = \|y - X\theta_1 - \mathbf{1}\theta_2\|_2$ and $g : \theta_1 \in \mathbb{R}^p \mapsto \|\theta_1\|_1$
Recall for $r \in \mathbb{R}^n,\ \|r\|_2 = \sqrt{\langle r, r \rangle} = \sqrt{r^T r}$. Then, by the chain rule,

$$\partial_{\theta_1} f(\theta) = -\frac{X^\top \big(y - X\theta_1 - \mathbf{1}\theta_2\big)}{\big\|y - X\theta_1 - \mathbf{1}\theta_2\big\|_2}$$

Because $F_\lambda := f + \lambda g$ is convex, coercive and locally Lipschitz,

$$\mathbf{0} \in \partial_{\theta_1} F_\lambda\big((\mathbf{0}, \hat{\theta}_2)\big) \iff \frac{X^\top \big(y - \mathbf{1}\hat{\theta}_2\big)}{\|y - \mathbf{1}\hat{\theta}_2\|_2} \in [-\lambda, \lambda]^p \iff \lambda \geq \frac{\big\|X^\top (y - \mathbf{1}\hat{\theta}_2)\big\|_\infty}{\|y - \mathbf{1}\hat{\theta}_2\|_2}$$

Hence g is sparsity inducing for f and in the same way than Example 3,

$$\lambda_0^{local}(X, y) = \sup_{\hat{\theta}_2 \in \mathbb{R}} \frac{\big\|X^\top (y - \mathbf{1}\hat{\theta}_2)\big\|_\infty}{\|y - \mathbf{1}\hat{\theta}_2\|_2}$$

Under H_0, we assume $\mathbf{Y} = \theta_2 \mathbf{1} + e$ with $e \sim \mathcal{N}\big(m\mathbf{1}, \sigma^2 I_n\big)$. Let $\bar{Y} = \frac{1}{n}\mathbf{1}^T \mathbf{Y}$ and note that, because X is standardized (so $X^\top \mathbf{1} = \mathbf{0}$), we want minimize, according to $\hat{\theta}_2$, the denominator. Hence, the supremum is attained at $\hat{\theta}_2 = \bar{Y}$,

$$\Lambda = \lambda_0^{local}(X, \mathbf{Y}) = \frac{\big\|X^\top (\mathbf{Y} - \bar{Y}\mathbf{1})\big\|_\infty}{\|\mathbf{Y} - \bar{Y}\mathbf{1}\|_2}$$

Write $e = m\mathbf{1} + \sigma u$ with $u \sim \mathcal{N}(0, I_n)$ and denote $H := I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$. Then $\mathbf{Y} - \bar{Y}\mathbf{1} = He = \sigma H u$ and,

$$\Lambda = \frac{\big\|X^\top \sigma H u\big\|_\infty}{\|\sigma H u\|_2} = \frac{\big\|X^\top H u\big\|_\infty}{\|H u\|_2}$$

which no longer depends on the unknown parameters m or σ. Therefore the statistic Λ is *pivotal*.

In the original paper [SvCM25], the authors introduce assumptions on the statistic Λ, requiring it to be non-constant in order to be called QUT-compatible. However, this definition is not yet settled. In fact, a new paper is being written, where QUT-compatibility is instead defined through the behavior of the penalty, similar to ℓ_1 near 0. For the purpose of this survey, we may disregard this technical detail.

3 Performance metrics

Consider a prediction problem with p input features and a learner parameterized by θ. Since parameters from different learners typically lie in different spaces, we require a unified method to evaluate which features are effectively utilized for prediction. We introduce a binary feature relevance mask $\Xi : \theta \mapsto m \in \{0, 1\}^p$ where $m_i = 1$ if feature i is deemed useful by the learner, and $m_i = 0$ otherwise. The true signal is encoded by the unknown mask $m^\star$ and the learner's mask is $\hat{m} = \Xi(\theta)$. Define the supports,

$$S = \{i : m_i^\star = 1\} \quad \text{and} \quad \hat{S} = \{i : \hat{m}_i = 1\}$$

For linear learner, $\theta_1 \in \mathbb{R}^p$, we choose

$$\Xi : \theta \in \mathbb{R}^p \mapsto \mathbb{1}_{\{\,|\theta|>0\}} := \big(\mathbb{1}_{\{|\theta_1|>0\}},\ \ldots,\ \mathbb{1}_{\{|\theta_p|>0\}}\big)$$

For a neural network with input dimension p and a hidden layer of size d_{hid}, the first-layer parameters are given by the weight matrix $W_1 \in \mathbb{R}^{p\times d_{\text{hid}}}$. The i-th feature is deemed useful if at least one entry in the i-th row of W_1 is nonzero. Formally, let $w_i^\top \in \mathbb{R}^{1\times d_{\text{hid}}}$ denote the i-th row of W_1. The relevance mask is then defined coordinate-wise. For $i \in [\![1, p]\!]$,

$$\Xi\big(W_1\big)_i \ :=\ \mathbb{1}_{\{\ \|w_i\|_2>0\ \}}$$

In practice, any other learner can define its own mask Ξ. The performance metrics described below become universally applicable.

Definition 3.1: Metrics

$$\text{PESR} := \mathbb{P}(\hat{S} = S) \qquad \text{TPR} := \mathbb{E}\frac{|\hat{S}\cap S|}{|S|} \qquad \text{FDR} := \mathbb{E}\frac{|\hat{S}\cap \overline{S}|}{|\hat{S}|\vee 1} \qquad F_1 := \frac{2\,TP}{2\,TP + FP + FN}$$

PESR stands for Predictive Exact Support Recovery, a stringent criterion that evaluates how often the learner identifies exactly the true set of features without errors. Higher values indicate superior performance.

TPR, True Positive Rate, also known as recall, measures the proportion of truly relevant features that the learner recovers on average. A high TPR indicates that the learner rarely misses relevant features but does not inform us about false positives (incorrectly selected features).

To capture false positives, we use FDR, False Discovery Rate, which measures the expected proportion of selected features that are actually irrelevant. A low FDR means that the selected features are trustworthy. Conversely, a high FDR indicates overselection or false alarms.

To balance recall and false discoveries, we combine TPR and precision into the F_1 score, defined as the harmonic mean of precision and recall. It balances the trade-off between including as many true features as possible and avoiding irrelevant ones.

The notions of True Positives (TP), False Positives (FP), and False Negatives (FN) are explicitly defined as follows,

$$TP := |\hat{S}\cap S|, \quad FP := |\hat{S}\cap \overline{S}|, \quad FN := |\overline{\hat{S}}\cap S|$$

4 Penalties

Ideally, the sparsest explanation would be obtained by directly solving the ℓ_0 minimization problem introduced in Section 1.4. So we proceed by relaxation : preserves the sharpness at the origin that encourages small coefficients to be driven exactly to zero, offers global convexity, or at least local Lipschitz continuity, so that subdifferential calculus remains applicable.

Throughout the paper all numerical experiments are solved with the Iterative Shrinkage Thresholding Algorithm (ISTA), see Section 7.1.1. ISTA exploits the separable form of F_λ by performing a gradient step on the smooth part f followed by the proximal mapping of g. The aim of this section is thus to review penalty classes with explicit and computationally efficient proximal operators.

Although we only address linear models in this report, the class of models can be extended, in particular to neural networks [SvCM25].

We consider for all these examples, the same loss $f = \sqrt{RSS}$ which keeps the same mind compared to the common RSS. Future works will explore different others interesting loss according to hypothesis on data.

Before to see these penalties, we show f verify assumptions required for QUT-Theory and backtracking ISTA, see Remark 9.

Proposition 4.1

$f := \sqrt{RSS}$ is coercive, locally Lipschitz and its gradient is locally Lipschitz.

Proof

Let $y \in \mathbb{R}^n,\ X \in \mathbb{R}^{n\times p}$ and $\tilde{X} := \big[X \mid \mathbf{1}\big] \in \mathbb{R}^{n\times(p+1)}$
Because $\tilde{X}^\top\tilde{X}$ is positive definite, let $\lambda_{\min} > 0$ be its smallest eigenvalue. Then, by Cauchy-Schwarz inequality

and property of the operator norm $\|\cdot\|_2$,

$$f(\theta)^2 = \|y - \tilde{X}\theta\|_2^2 = \|y\|_2^2 + \theta^\top \tilde{X}^\top \tilde{X}\theta - 2\, y^\top \tilde{X}\theta \geq \|y\|_2^2 + \lambda_{\min}\|\theta\|_2^2 - 2\|y\|_2\|\tilde{X}\|_2\|\theta\|_2$$

Hence $f = \Omega_{\|\theta\|_2 \to \infty}\big(\|\theta\|_2\big)$ is coercive.
From the second triangle inequality $|f(\theta_1) - f(\theta_2)| \leq \|\tilde{X}(\theta_1 - \theta_2)\| \leq \|\tilde{X}\|\|\theta_1 - \theta_2\|$. Then f is globally (so locally) Lipschitz.
Finally, f is differentiable with $\partial f(\theta) = -\frac{\tilde{X}^\top(y - \tilde{X}\theta)}{f(\theta)}$. Let $\varepsilon > 0$ and define $\mathcal{D}_\varepsilon = \{\theta \in \mathbb{R}^{p+1} : f(\theta) \geq \varepsilon\}$. For any $\theta_1, \theta_2 \in \mathcal{D}_\varepsilon$,

$$\|\partial f(\theta_1) - \partial f(\theta_2)\|_2 \leq \frac{\|\tilde{X}\|_2^2}{\varepsilon}\, \|\theta_1 - \theta_2\|_2$$

Therefore the gradient is Lipschitz on $\mathcal{D}_\varepsilon$ with constant $L = \|\tilde{X}\|_2^2/\varepsilon$.

□

4.1 ℓ_1

4.1.1 QUT

We study the most common penalty, $g = \|\cdot\|_1$, to do regularization which had the nice property to be convex. From Example 4 we got the statistic Λ pivotal so we can simulate it. Recall, with $u \sim \mathcal{N}(0, I_n)$ and $H := I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$,

$$\Lambda = \lambda_0^{local}(X, \mathbf{Y}) = \frac{\big\|X^\top(\mathbf{Y} - \bar{Y}\mathbf{1})\big\|_\infty}{\|\mathbf{Y} - \bar{Y}\mathbf{1}\|_2} = \frac{\big\|X^\top H u\big\|_\infty}{\|Hu\|_2}$$

4.1.2 Proximal

The global minimizer of $F_\lambda := f + \lambda g$ in (1) is not available in closed form, therefore we approximate it. We use iterative methods, here ISTA, to reach the global minimum because F_λ is convex and coercive. This requires computing the proximal operator, defined in the Definition 7.1.

Theorem 4.1: Soft-thresholding formula

For any $\gamma > 0$ and any $u \in \mathbb{R}^p$ the proximal operator of λg is,

$$\mathrm{prox}_{\gamma, \lambda g}(u) = \mathrm{sign}(u) \odot \max\{|u| - \lambda\gamma, 0\}$$

Proof

We treat each coordinate independently because g is separable. Let $i \in [\![1, p]\!]$ and set

$$x^\star := \mathrm{prox}_{\gamma, \lambda g}(u) = \arg\min_{x \in \mathbb{R}} \Big\{ \lambda|x| + \frac{1}{2\gamma}(x - u_i)^2 \Big\}$$

By Fermat's rule, $x^\star$ satisfies

$$0 \in \partial\Big(\lambda|\cdot| + \frac{1}{2\gamma}(\cdot - u_i)^2\Big)(x^\star) = \lambda\, \partial|\cdot|(x^\star) + \frac{1}{\gamma}\big(x^\star - u_i\big).$$

We distinguish three cases,
(i) $x^\star > 0$. Then, $0 = \lambda \cdot 1 + \frac{1}{\gamma}(x^\star - u_i)$, yields $x^\star = u_i - \gamma\lambda$, valid when $u_i > \gamma\lambda$.
(ii) $x^\star < 0$. Then, $0 = \lambda \cdot (-1) + \frac{1}{\gamma}(x^\star - u_i)$, yields $x^\star = u_i + \gamma\lambda$, valid when $u_i < -\gamma\lambda$.
(iii) $x^\star = 0$. Then, $0 \in \lambda[-1, 1] + \Big\{\frac{1}{\gamma}(0 - u_i)\Big\}$, equivalent to $|u_i| \leq \gamma\lambda$.
Combining the three situations we obtain,

$$x^\star = \begin{cases} u_i - \gamma\lambda & u_i > \gamma\lambda \\ 0 & |u_i| \leq \gamma\lambda \\ u_i + \gamma\lambda & u_i < -\gamma\lambda \end{cases} = \mathrm{sign}(u_i)\, \max\big\{|u_i| - \gamma\lambda, 0\big\}$$

By bringing together all coordinates, gives the vector formula. □

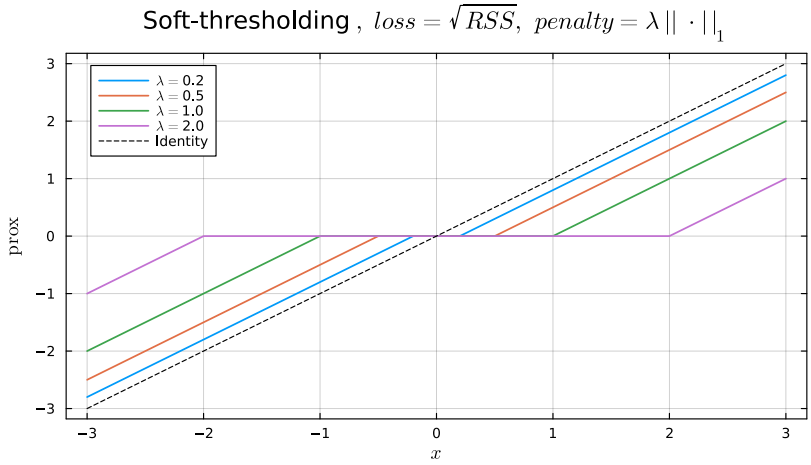

Figure 1: Graph of the proximal operator for the ℓ_1 penalty for different values of λ (with step size $\gamma = 1$). The identity line is shown for reference (dashed). The function is continuous everywhere, with a "dead zone" of width 2λ around zero.

4.1.3 Numerical experiments

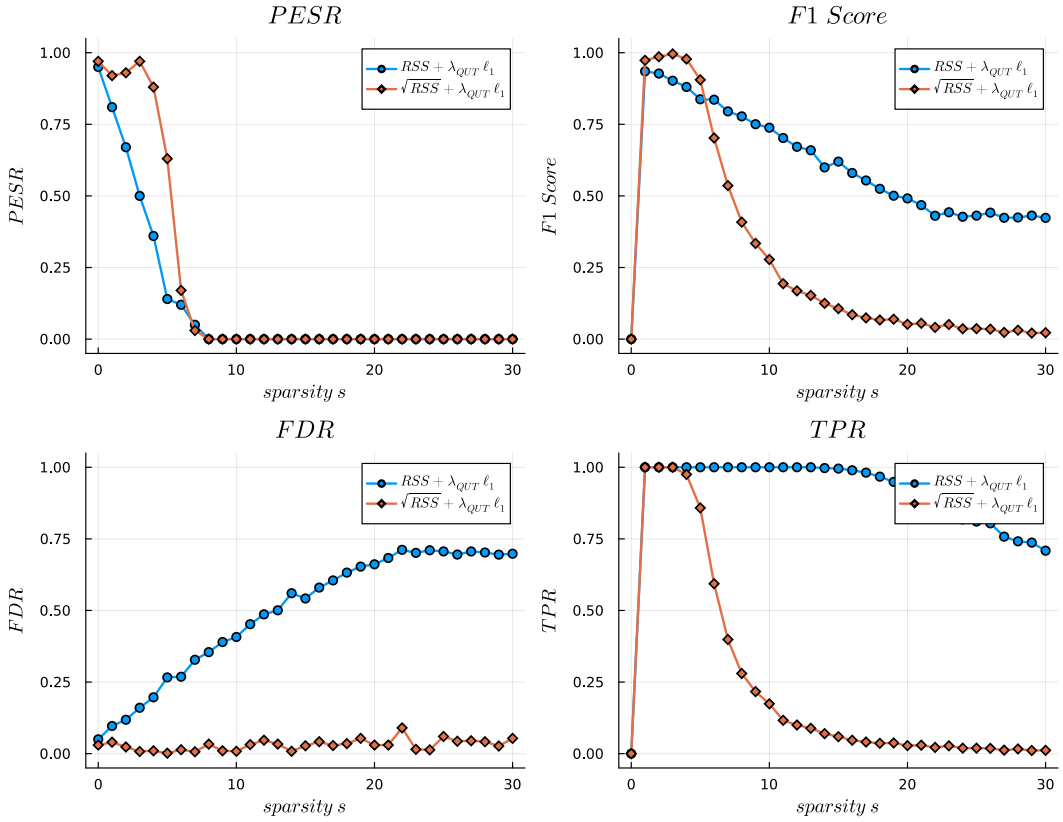

Figure 2: Performance comparison between RSS and $\sqrt{RSS}$ losses combined with ℓ_1 regularization for variable selection in a sparse linear regression model. The simulation parameters are $n = 70$, $p = 250$, noise level $\sigma = 0.1$. Sparsity s ranges from 0 to 30, where each nonzero coefficient is set to 3. Each point in the plot is estimated from 100 Monte Carlo replications, and the Quantile Universal Threshold λ_{QUT} is computed via 1000 simulations under the null model. The RSS criterion operates under oracle knowledge of noise variance.

In this experiment, even with oracle knowledge of the noise variance, the performances using the RSS criterion are notably worse than those based on $\sqrt{RSS}$. The PESR confirms the poor support recovery of RSS. While the TPR

for $\sqrt{RSS}$ decreases sharply as s increases, its FDR remains stable, indicating effective control of false discoveries. The F1 score, which balances true positives and precision, is higher for RSS at small s due to a more gradual TPR decay, despite a slightly higher FDR. This trade-off leads to a moderate, yet not optimal, selection performance.

For the null model ($s = 0$), these results align with Corollary 2.1, ensuring the Type I error remains bounded by $1 - \alpha$. We let $\alpha = 0.05$. Moreover under the null model, TPR and so F1 score are 0 because it doesn't exist positive case and we find the α tolerance in the FDR.

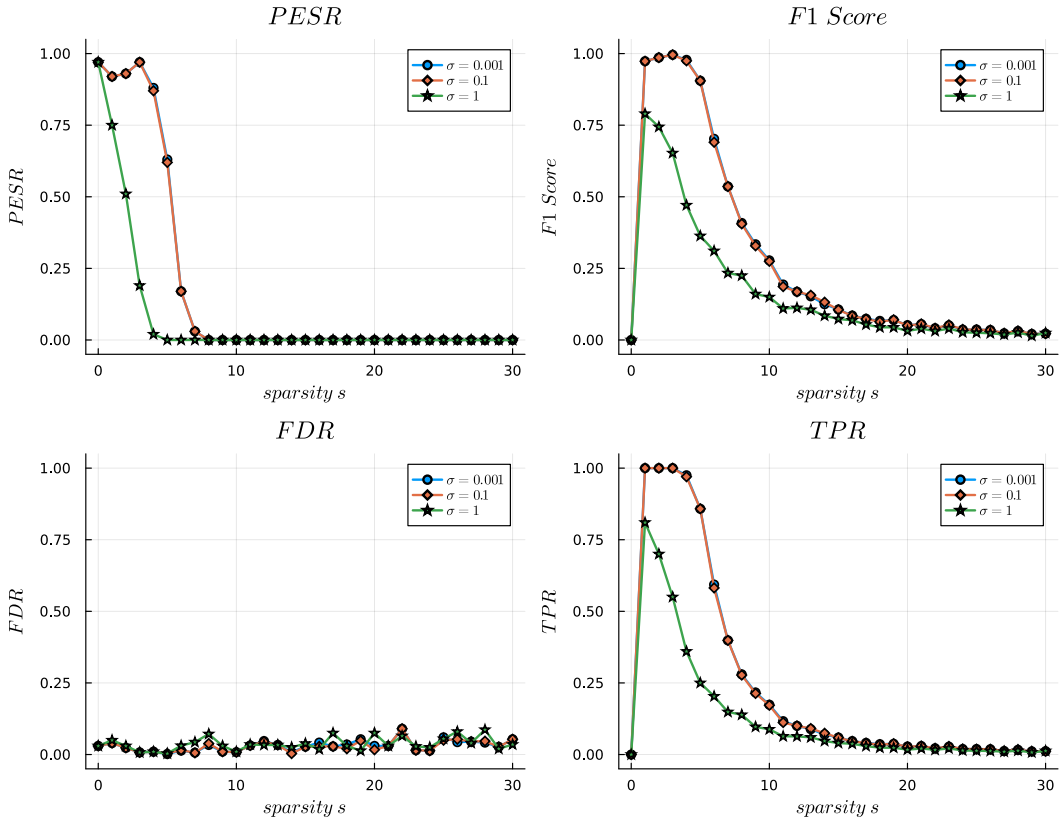

Figure 3: Variable selection performance of the $\sqrt{RSS}$ with ℓ_1 regularization at three noise levels, $\sigma \in \{0.001, 0.1, 1\}$. Parameters match Fig. 2: $n = 70$, $p = 250$, sparsity $s \in [\![0, 30]\!]$ with each non-zero coefficient fixed at 3 and each point averages 100 Monte-Carlo runs, and λ_{QUT} is obtained from 1000 null model simulations.

Even though the QUT statistic Λ is pivotal but performance curves depend on σ because the data distribution changes. Lowering σ shifts the PESR and TPR curves rightward, allowing exact support recovery for larger s. FDR remains almost flat across noise levels, confirming robust false-discovery control. Consequently, the F1 score improves when the noise is lighter, reflecting the broader regime where high power and low error coincide.

4.2 P_ν

Definition 4.1: P_ν

Let $\nu \in]0, 1]$,

$$P_\nu : \theta \in \mathbb{R}^p \mapsto \sum_{i=1}^{p} \rho_\nu(\theta_i) \quad \text{with} \quad \rho_\nu : x \in \mathbb{R} \mapsto \frac{|x|}{1 + |x|^{1-\nu}}$$

This class of penalties had nice properties such as continuity, differentiability except in $\mathbf{0}$, locally Lipschitz [3]. For $\nu = 1$, $P_\nu = \frac{1}{2}\|\cdot\|_1$ and when $\nu \to 0$, it tends to ℓ_0 for large θ. For group variable we can adapt slightly the definition, see [SvCM25] which introduced P_ν.

[3] Even global Lipschitz. Indeed, $\partial\rho_\nu \subset [-1; 1]$ so by finite sum of Lipschitz function, P_ν is also.

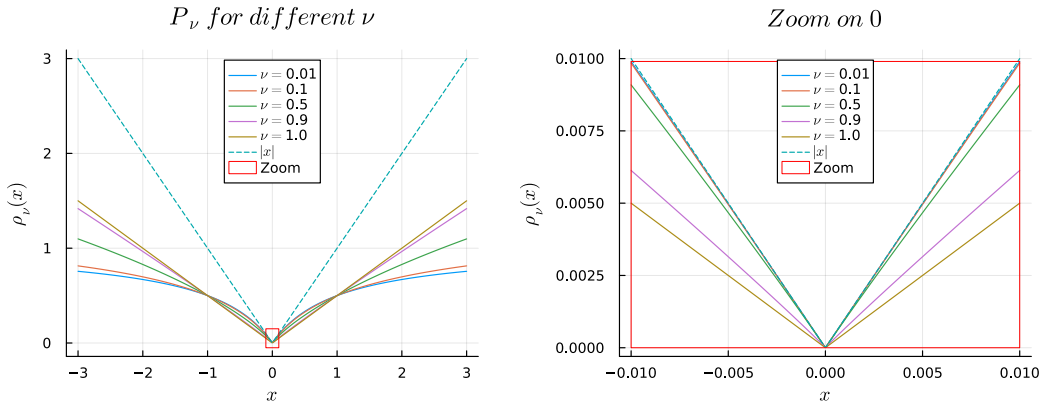

Figure 4: Shape of the P_ν penalty for $p = 1$, for several values of ν, compared to the ℓ_1 norm. The right panel provides a zoom near zero to illustrate how the sharpness at the origin varies with ν. For small ν, the function closely approximates $|x|$, while larger ν leads to increased 'flatness' around zero.

4.2.1 QUT

Proposition 4.2: P_ν is sparsity inducing and expression of Λ

Let $\nu \in]0, 1[$. For $f = \sqrt{RSS}$, P_ν is sparsity inducing.
Let $e = m\mathbf{1} + \sigma u$ with $u \sim \mathcal{N}(\mathbf{0}, I_n)$ and denote $H := I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$,

$$\Lambda = \frac{\left\|X^\top H u\right\|_\infty}{\|Hu\|_2}$$

Proof

For $t > 0$

$$\rho'_\nu(t) \;=\; \frac{1 + |t|^{1-\nu} - |t|\,(1-\nu)|t|^{-\nu}}{\left(1 + |t|^{1-\nu}\right)^2} \;=\; \frac{1 - (1-\nu) + |t|^{1-\nu}}{\left(1 + |t|^{1-\nu}\right)^2} \;\xrightarrow[t\to 0]{}\; 1$$

By symmetry $\rho'_\nu(0^-) = -\rho'_\nu(0^+)$, hence

$$\partial\rho_\nu(0) = \left[-1, 1\right] \quad \text{and} \quad \partial P_\nu(\mathbf{0}) = \left[-1, 1\right]^p$$

Fix $\theta_2 \in \mathbb{R}$ and consider the candidate point $\bar{\theta} := (\mathbf{0}, \theta_2)$. Consider the Clarke-Fermat condition, see Proposition 2.1,

$$\mathbf{0} \;\in\; \partial F_\lambda(\bar{\theta}) \;=\; \partial_{\theta_1} f(\bar{\theta}) + \lambda \partial P_\nu(\mathbf{0}) \iff -\partial_{\theta_1} f(\bar{\theta}) \in \lambda\left[-1, 1\right]^p$$

From this step it's similar to Example 4. So using the chain rule and etc. we get,

$$\lambda \;\geq\; \sup_{\theta_2 \in \mathbb{R}} \frac{\|X^\top (y - 1\theta_2)\|_\infty}{\|y - \mathbf{1}\theta_2\|_2} \;=:\; \lambda_0^{\text{local}}(X, y) < +\infty$$

That's to say,

$$\Lambda = \frac{\left\|X^\top H u\right\|_\infty}{\|Hu\|_2}$$

□

We don't include $\nu = 1$ to lighten the notation. If we want to consider it, when we take the limit to 0 it's not ± 1 but $\pm\frac{1}{2}$. Therefore, the associated statistics Λ is doubled. An interesting fact from this class of penalties is the statistic Λ is independent on $\nu \in]0, 1[$!

Remark 3: Only the behavior in 0 interest us

Although Example 4 and Proposition 4.2 have same loss but different penalty, they have the same Λ. It happens because we are interested only on how the sparsity is induce so we look at $\mathbf{0}$ and both penalty act identically.

From the last remark, how to know which penalty is good if only the behavior in the neighborhood of $\mathbf{0}$ matters of us ? It depends on the expression of the proximal.

4.2.2 Proximal

Theorem 4.2: Implicit expression of the proximal of P_ν

Let $0 < \nu < 1$, $\gamma, \lambda > 0$. There exists a unique $\kappa = \kappa(\gamma, \lambda, \nu) > 0$ that solves

$$\kappa^{2-\nu} + 2\kappa + \kappa^{\nu} + 2\gamma\lambda(\nu - 1) = 0$$

A unique threshold

$$\varphi = \varphi(\gamma, \lambda, \nu) = \frac{\kappa}{2} + \frac{\gamma\lambda}{1 + \kappa^{1-\nu}} > 0$$

and let

$$x^\star : z \in \mathbb{R}^+ \mapsto \arg\min_{x \in \mathbb{R}} \Big\{ G\big(x\big) = z \Big\} \quad \text{with} \quad G : x \in \mathbb{R} \mapsto x + \gamma\lambda\rho'_\nu(x)$$

The proximal is given component-wise. For any $z \in \mathbb{R}^p$ and $i \in [\![1, p]\!]$,

$$\Big[\mathrm{prox}_{\gamma, \lambda P_\nu}(z)\Big]_i = \begin{cases} 0 & |z_i| \leq \varphi \\ \mathrm{sign}(z_i)\, x^\star(|z_i|), & |z_i| > \varphi \end{cases}$$

Proof

Recall we search

$$\mathrm{prox}_{\gamma, \lambda P_\nu} = \arg\min_{x \in \mathbb{R}^p} \Big\{ \lambda P_\nu + \frac{1}{2\gamma} \|x - \cdot\|_2^2 \Big\}$$

Because P_ν and $\|\cdot\|_2^2$ are separable, it suffices to minimize each coordinate independently. For $z \in \mathbb{R}$, let

$$m(\cdot\,; z) : x \in \mathbb{R} \mapsto \lambda\rho_\nu(x) + \frac{(x - z)^2}{2\gamma}$$

The function $m(\cdot\,; z)$ is continuous and coercive hence it attains at least one global minimizer. Because ρ_ν is even and for any $x \in \mathbb{R}^*$, $\rho'_\nu(x)\,\mathrm{sign}(x) = \dfrac{1 + \nu|x|^{1-\nu}}{(1 + |x|^{1-\nu})^2} > 0$ we have,

$$\mathrm{sign}\big(m'(x\,; z)\big) = \mathrm{sign}\Big(\gamma\lambda\,\mathrm{sign}(x)\rho'_\nu(x) + (x - z)\Big)$$

Consequently $m'(\cdot\,; z)$ cannot vanish when x and z have opposite signs. Therefore every minimizer has the same sign as z. By symmetry we may assume $z \geq 0$ and minimize over $x \geq 0$. We are restoring the sign at the end.
For $x = 0$, the only point where $m(\cdot\,; z)$ is not differentiable,

$$\partial m(0; z) = \lambda \left[\lim_{x \to 0^-} \rho'_\nu(x; z);\ \lim_{x \to 0^+} \rho'_\nu(x; z) \right] + \left\{ \frac{0 - z}{\gamma} \right\} \iff |z| \leq \lambda\gamma$$

For $x > 0$, we define $G : x \in \mathbb{R} \mapsto x + \gamma\lambda\rho'_\nu(x)$.

$$m'(x; z) = \lambda\rho'_\nu(x) + \frac{x - z}{\gamma} = \frac{G(x) - z}{\gamma},$$

So,

$$m'(x; z) = 0 \iff G(x) = z$$

Analyze G shows,

$$G'(x) = 1 + \lambda\gamma \cdot \left(- \frac{(1 - \nu)\big((2 - \nu)x^{-\nu} + \nu x^{1-2\nu}\big)}{\big(1 + x^{1-\nu}\big)^3} \right)$$

Hence, $\lim_{x\to 0^+} G'(x) = -\infty$, $\lim_{x\to\infty} G'(x) = 1$ and on $\mathbb{R}^+ G'' = \lambda\gamma\rho_\nu''' > 0$. By strict increasing of G', there exists $\eta > 0$ such that $G'(\eta) = 0$. Let $\phi := G(\eta)$. This the variation table to summarize our analysis,

x	0^+		η		$+\infty$
$G'(x)$		$-$	0	$+$	
$G(x)$	$\lambda\gamma$	$\searrow$	ϕ	$\nearrow$	$+\infty$

Hence,

If $z \in [0, \phi[$	:	0 is the minimizer
If $z \in \{\phi\}$	:	0 xor η is the minimizer[a]
If $z \in]\phi, \lambda\gamma]$	:	0 xor $x_1^* \in]0, \eta[$ xor $x_2^* \in]\eta, +\infty[$ is the minimizer
If $z \in]\lambda\gamma, +\infty[$	:	$x^* \in]\eta, +\infty[$ is the minimizer

Let us be interested about solution for $z \in]\phi, \lambda\gamma]$,
Firstly, according to x, $m''(x; z) = \frac{G'(x)}{\gamma}$, therefore $m''(x_1^*, z) < 0$ so x_1^* is a local maximum and $m''(x_2^*, z) > 0$ so x_2^* is a local minimum, we denote now as $x^*(z) := x_2^*$ to show its dependence in z. Secondly, it stays two candidates so we study when one is better than the other.

$$\Delta : z \in]\phi, \lambda\gamma] \mapsto m(x^*(z); z) - m(0; z) = \lambda\rho_\nu\big(x^*(z)\big) + \frac{\big(x^*(z) - z\big)^2}{2\gamma} - \frac{z^2}{2\gamma}$$

Recall the implicit theorem for two variables, let $x_0, z_0 \in \mathbb{R}$ and $F : \mathbb{R}^2 \to \mathbb{R}$ be C^1 function such that $F(x_0, z_0) = 0$ and $\partial_x F(x_0, z_0) \neq 0$, then there exists a neighborhood I of z_0 and a unique C^1 function $h : I \to \mathbb{R}$ verifying for all $z \in I$, $F(h(z), z) = 0$. Moreover, and the most important,

$$h'(z) = -\frac{\partial_2 F(h(z), z)}{\partial_1 F(h(z), z)}$$

Let's apply this theorem with $F : (x, z) \in]\eta, +\infty[\times]\phi, \lambda\gamma] \mapsto G(x) - z$ of class C^1 with $\partial_z F = -1 \neq 0$ and $\partial_x F = G' > 0$ so x^* is C^1 and

$$x^{*\prime}(z) = \frac{1}{G'(x^*(z))}$$

To know the sign of Δ we need to analyze it,

$$\Delta'(z) = x^{*\prime}(z) \cdot \big(\lambda\rho_\nu'(x^*(z)\big) + (x^{*\prime}(z) - 1) \cdot \frac{\big(x^*(z) - z\big)}{\gamma} - \frac{z}{\gamma} = x^{*\prime}(z) \cdot m'\Big(x^*(z); z\Big) - \frac{x^*(z)}{\gamma} = -\frac{x^*(z)}{\gamma} < 0$$

Δ is strictly decreasing. Moreover, we have for $x \in]0, \eta[$, $m'(x; \phi) > 0$ therefore by continuity of m and from strict increasing function $m(\eta; \phi) - m(0; \phi) > 0$, then

$$\lim_{z\to\phi^+} \Delta(z) = m(\eta; \phi) - m(0; \phi) > 0 \tag{2}$$

Then, we show Δ become negative, indicating a switch of minimizer. We extend naturally Δ on $]\phi, +\infty]$. For $z \in]\phi, +\infty]$

$$z = G\big(x^*(z)\big) = x^*(z) + \gamma\lambda\,\rho_\nu'\big(x^*(z)\big) \iff x^*(z) - z = -\gamma\lambda\,\rho_\nu'\big(x^*(z)\big)$$

So,

$$\begin{aligned}
\Delta(z) &= \lambda\rho_\nu\Big(x^*(z)\Big) + \frac{\gamma^2\lambda^2\big(\rho_\nu'(x^*(z))\big)^2}{2\gamma} - \frac{\big(x^*(z) + \gamma\lambda\rho_\nu'(x^*(z))\big)^2}{2\gamma} \\
&= \lambda\Big[\rho_\nu(x^*(z)) - x^*(z)\,\rho_\nu'(x^*(z))\Big] - \frac{(x^*(z))^2}{2\gamma} \\
&= \lambda(1-\nu)\frac{(x^*(z))^{2-\nu}}{\big(1 + (x^*(z))^{1-\nu}\big)^2} - \frac{(x^*(z))^2}{2\gamma}
\end{aligned}$$

We remark for $x > 0$, $x \leq G(x) = x + \gamma\lambda\rho'_\nu(x) \leq x + \gamma\lambda$. Using $z = G(x^+(z))$ we get $z - \lambda\gamma \leq x^*(z) \leq z$, so $\lim_{z\to+\infty} x^*(z) = +\infty$ and therefore,

$$\lim_{z\to+\infty} \Delta(z) \leq \lim_{z\to+\infty} \lambda(1-\nu)\, x^*(z)^\nu - \frac{x^*(z)^2}{2\gamma} = -\infty < 0$$

Consequently by continuity of Δ, $\exists!\varphi \in]\phi, +\infty]$, $\Delta(\varphi) = 0$. We denote $\kappa := x^*(\varphi)$ which has to be a minimizer of m, then we got the system,

$$\begin{cases} m(\kappa;\varphi) = m(0;\varphi) \\ m'(\kappa;\varphi) = 0 \end{cases} \iff \begin{cases} 2\gamma\lambda \frac{\kappa}{1+\kappa^{1-\nu}} = 2\kappa\varphi - \kappa^2 \\ \varphi = G(\kappa) = \kappa + \gamma\lambda \frac{1+\nu\kappa^{1-\nu}}{(1+\kappa^{1-\nu})^2} \end{cases} \iff \begin{cases} \kappa^\nu + 2\kappa + \kappa^{2-\nu} = 2\gamma\lambda(1-\nu) \\ \varphi = \frac{\kappa}{2} + \frac{\gamma\lambda}{1+\kappa^{1-\nu}} \in]\phi, \lambda\gamma] \end{cases}$$

Then, we obtain the implicit proximal rule, where φ acts as a threshold between 0 and $x^*(z)$, with the minimizer given in the statement. □

[a] This case is treated in (2) : 0 is finally the minimizer.

The proximal is not explicit and need to compute κ with Bisection's method or Newton's method, see 7.2 and use a second time one of these method to compute $x^*(\cdot)$. We suggest initialization parameters for these method to compute the proximal of P_ν.

Remark 4: Compute κ with Bisection and Newton's method

Let $T := 2\gamma\lambda(1-\nu)$ and

$$\psi : \kappa \in \mathbb{R}^{+*} \mapsto \kappa^{2-\nu} + 2\kappa + \kappa^\nu - T \in \mathbb{R}$$

This function is C^1 and strictly increasing : $\psi'(\kappa) = (2-\nu)\kappa^{1-\nu} + 2 + \nu\kappa^{\nu-1} > 0$ Moreover, $\psi(0^+) = -T < 0$ and $\psi(\kappa) \xrightarrow[\kappa\to+\infty]{} +\infty$, then ψ has only one simple root and we want to find it.

For **Bisection method** (Algorithm 2). $\psi(a_0) = -T < 0$ and $\psi(b_0) \geq b_0^{2-\nu} - T \geq 0$ so we suggest to define,

$$a_0 := 0 \qquad b_0 := T^{1/(2-\nu)}$$

For **Newton's method** (Algorithm 3), we keep ψ and take also its derivative ψ'. We suggest to consider to start the algorithm from κ_0 defined according to asymptotic consideration.
If the root $\kappa \ll 1$, then close to 0 we have $\kappa^{2-\nu} \ll \kappa \ll \kappa^\nu$ so $\kappa^\nu \approx T$. However, if $\kappa \gg 1$, then $\kappa^{2-\nu} \approx T$. We chose therefore,

$$\kappa_0 := \begin{cases} T^{1/\nu} & \text{if } T \leq 1 \\ T^{1/(2-\nu)} & \text{if } T > 1 \end{cases}$$

This is not a rigorous proof ensuring the convergence of Newton's method. We encourage the reader to consult Kantorovich's theorem.

Remark 5: Compute $x^\star(z)$ with Bisection and Newton's method

Fix $z \geq \varphi$. Recall

$$G : x \in \mathbb{R}^+ \mapsto x + \gamma\lambda\,\rho'_\nu(x) \in \mathbb{R} \quad \text{and} \quad G' : x \in \mathbb{R}^+ \mapsto 1 + \gamma\lambda\,\rho''_\nu(x) \in \mathbb{R}$$

The minimizer of $m(\cdot; z)$ among the nonzero critical points is the large root $x^\star(z) > \eta := G'^{-1}(0)$. For $z \geq \varphi$, we have $x^\star(z) \geq \kappa$ and the root is unique on $[\kappa, +\infty[$.
Define

$$H_z : x \in [\kappa, +\infty[\mapsto G(x) - z \in \mathbb{R} \qquad H'_z : x \in [\kappa, +\infty[\mapsto G'(x) \in \mathbb{R}^+$$

For **Bisection method** (Algorithm 2). H_z is continuous and strictly increasing with

$$H_z(\kappa) = \varphi - z \leq 0 \quad \text{and} \quad H_z(z) = \gamma\lambda\,\rho'_\nu(z) \geq 0$$

so to solve $H_z(x) = 0$ we suggest,

$$a_0 := \kappa \qquad b_0 := z$$

For **Newton's method** (Algorithm 3), convexity and regularity of H_z is supposed to avoid issues. Start the method from the right at,

$$x_0 := z$$

4.2.3 Numerical experiments

The penalty introduced in our model results in a non-convex optimization problem (1), which inherently contains the risk of converging to undesirable local minima. To mitigate this issue, we employ a heuristic known as homotopy optimization, inspired by the methodology originally introduced by [SvCM25]. The core idea behind this approach is that the optimization landscape, and consequently the global minimum, evolves gradually as parameters (ν, λ) vary continuously. Provided these parameters are sufficiently close between successive configurations, their global minima remain proximate.

We refer to this process as a warm start, where the algorithm is iteratively executed across increasingly challenging landscapes. Specifically, the converged solution from the previous step becomes the initial point for the subsequent step. Formally, choose a small positive constant ε and the number of warm start K to do (e.g. $\varepsilon = 10^{-2}, K = 6$). We define a grid of parameters $(\lambda_k, \nu_k)_k$ as follows,

$$\lambda_k = \exp\left(\left(1 - \frac{k}{K}\right)\log(\varepsilon) + \frac{k}{K}\log(\lambda_{QUT})\right) \quad \text{with} \quad k \in [\![1, K]\!]$$

and for the parameter ν, we consider the sequence

$$\nu_k = 1 - \frac{k}{K-1}(1-\nu) \quad \text{with} \quad k \in [\![1, K-1]\!] \quad \text{and} \quad \nu_K = \nu$$

This gradual homotopy-based approach significantly reduces the likelihood of being trapped in suboptimal local minima by ensuring smoother transitions between optimization problems, thereby enhancing the robustness and reliability of the optimization procedure.

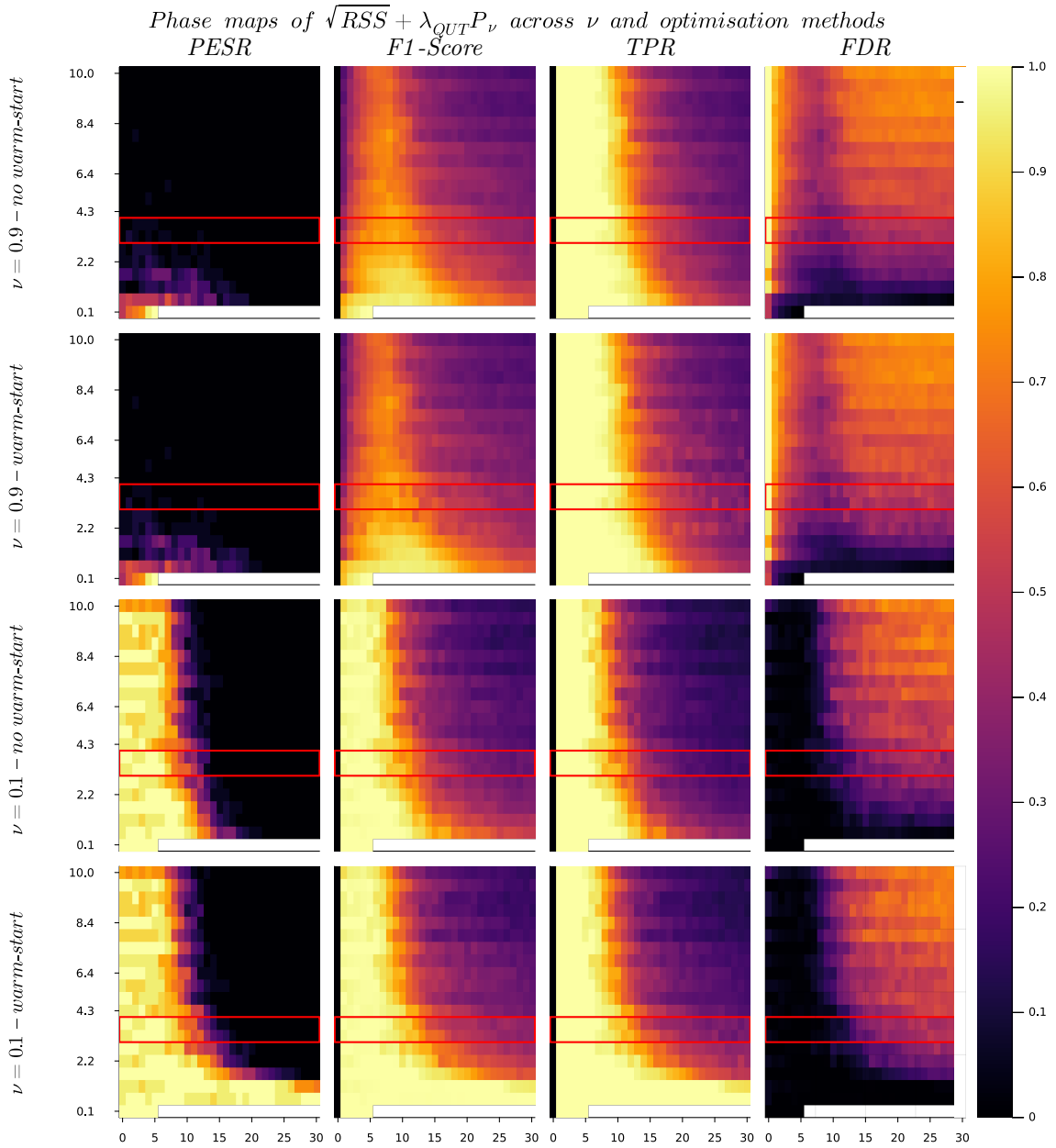

Figure 5: Variable selection performance using the $\sqrt{RSS}$ loss combined with the P_ν penalty at fixed sample size $n = 50$ and varying feature dimensions p (length of grid = 20). This is the vertical axis with ratio p/n and on horizontal axis the sparsity of the true vector. Each point in the heatmap is estimated from 20 Monte-Carlo replications, with the Quantile Universal Threshold (λ_{QUT}) computed from 1000 null model simulations. The bottom white lines in the heatmaps correspond to NaN values, indicating cases where the number of non-zero coefficients available is insufficient to achieve the specified sparsity level s, specifically when $s > p$, which only occurs at the lowest ratio. A warm start procedure, as detailed previously, is employed to enhance optimization stability. This approach incrementally adjusts parameters across increasingly challenging penalty landscapes.

The red rectangles correspond closely to those observed in the reference experiment (Figure 3), offering a direct comparison between ℓ_1 and P_ν penalties. For $\nu = 0.9$, the warm start strategy shows limited effectiveness due to minimal variation across the grid of ν values. However, the warm start considerably improves results for lower p/n ratios.

An interesting counterexample to Corollary 2.1 arises when the hypothesis $\lambda_0^{local} = \lambda_0$ (valid in convex scenarios per Equation (1)) is relaxed. Notably, at sparsity $s = 0$ and $\nu = 0.9$, PESR values fall below $1 - \alpha$ and even approach zero, suggesting convergence to local minima. Despite intuition suggesting a more stringent penalty regime as ν

approaches 0, the scenario with $\nu = 0.1$ demonstrates superior performance. One possible explanation is that, for larger values of ν, the penalty function converges too slowly toward zero, as illustrated in the right panel of Figure 6.

We see two distinct regimes clearly emerging from the heatmap results. Notably, the P_ν penalty with $\nu = 0.1$ consistently demonstrates strong performance, suggesting it is a particularly effective choice for variable selection.

4.3 P_ε^Σ (log sum)

We can experiment with a variety of sparsity promoting penalties and evaluate their performance. In particular, we introduce the widely used log-sum penalty, popularized in [CWB08]. Although its theoretical rationale may be limited, implementing it from scratch provides valuable insight into the method and serves as a useful exercise[4], and this one is corrected !

Definition 4.2: P_ε^Σ

Let $\varepsilon > 0$,

$$P_\varepsilon^\Sigma : \theta \in \mathbb{R}^p \mapsto \sum_{i=1}^{p} \rho^\Sigma(\theta_i)_\varepsilon \quad \text{with} \quad \rho_\varepsilon^\Sigma : x \in \mathbb{R} \mapsto \ln\left(1 + \frac{|x|}{\varepsilon}\right)$$

This function is as both precedents, continuous, even, differentiable except in 0, not convex, locally lipschitz. It belongs to the folded concave penalties class.

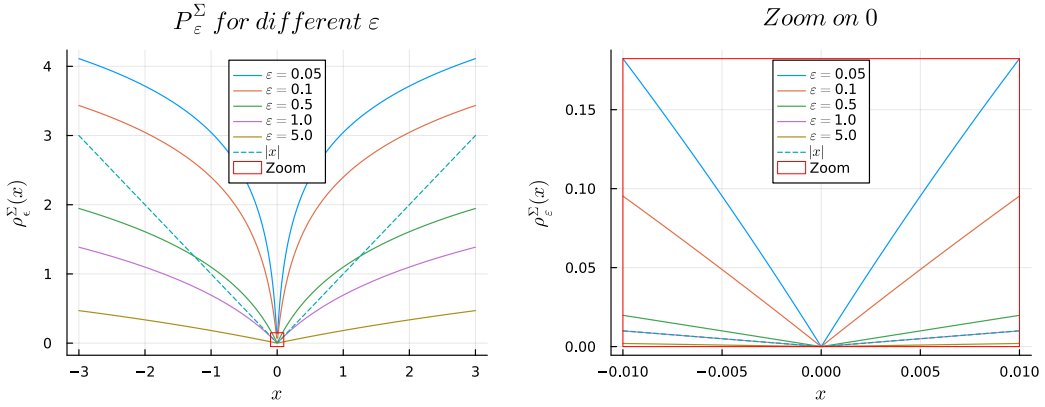

Figure 6: Shape of the P_ε^Σ penalty for $p = 1$, for several values of ε, compared to the ℓ_1 norm. The right panel provides a zoom near zero to illustrate how the sharpness at the origin varies with ε. For small ε, the function closely approximates $c \cdot \ell_0$ according to a large coefficient $c > 0$, while larger ε leads to increased flatness.

4.3.1 QUT

Proposition 4.3: P_ε^Σ **is sparsity inducing and expression of** Λ

Let $\varepsilon > 0$. For $f = \sqrt{RSS}$, the penalty P_ε^Σ is sparsity inducing.
Let $e = m\mathbf{1} + \sigma u$ with $u \sim \mathcal{N}(\mathbf{0}, I_n)$ and denote $H := I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$,

$$\Lambda = \varepsilon \frac{\left\|X^\top H u\right\|_\infty}{\|Hu\|_2}$$

Proof

For $t > 0$

$$\rho_\varepsilon^{\Sigma\prime}(t) = \frac{\operatorname{sign}(t)}{\varepsilon + |t|} \xrightarrow[t \to 0]{} \frac{1}{\varepsilon}$$

[4] **Exercise :** Write all proofs of this section and reproduce the phase transition find in Figure 8

By symmetry ${\rho_\varepsilon^\Sigma}'(0^-) = -{\rho_\varepsilon^\Sigma}'(0^+)$, hence

$$\partial\rho_\varepsilon^\Sigma(0) = \big[-\frac{1}{\varepsilon}, \frac{1}{\varepsilon}\big] \quad \text{and} \quad \partial P_\varepsilon^\Sigma(\mathbf{0}) = \big[-\frac{1}{\varepsilon}, \frac{1}{\varepsilon}\big]^p$$

Fix $\theta_2 \in \mathbb{R}$ and consider the candidate point $\bar{\theta} := (\mathbf{0}, \theta_2)$. Using the Clarke-Fermat condition (see Proposition 2.1),

$$\mathbf{0} \in \partial F_\lambda(\bar{\theta}) = \partial_{\theta_1} f(\bar{\theta}) + \lambda \partial P_\varepsilon^\Sigma(\mathbf{0}) \iff -\partial_{\theta_1} f(\bar{\theta}) \in \lambda\big[-\frac{1}{\varepsilon}, \frac{1}{\varepsilon}\big]^p$$

Proceeding exactly as in Example 4 and applying the chain rule, we get

$$\lambda \geq \varepsilon \sup_{\theta_2 \in \mathbb{R}} \frac{\|X^\top(y - \mathbf{1}\theta_2)\|_\infty}{\|y - \mathbf{1}\theta_2\|_2} =: \lambda_0^{\text{local}}(X, y) < +\infty$$

That is,

$$\Lambda = \varepsilon \frac{\big\|X^\top H u\big\|_\infty}{\|Hu\|_2}$$

□

4.3.2 Proximal

Theorem 4.3: Implicit expression of the proximal of P_ε^Σ

Let $\varepsilon > 0$ and $\gamma, \lambda > 0$. Let,

$$x^\star : s \in [2\sqrt{\lambda\gamma} - \varepsilon; +\infty[\mapsto \frac{s - \varepsilon + \sqrt{(s+\varepsilon)^2 - 4\gamma\lambda}}{2}$$

There exists a unique

$$\varphi = \varphi(\gamma, \lambda, \varepsilon) \in \big]2\sqrt{\gamma\lambda} - \varepsilon,\ \frac{\lambda\gamma}{\varepsilon}\big[$$

characterised by

$$\lambda \ln\Big(1 + \frac{x_\varphi}{\varepsilon}\Big) = \frac{x_\varphi(2\varphi - x_\varphi)}{2\gamma} \quad \text{with} \quad x_\varphi := \frac{\varphi - \varepsilon + \sqrt{(\varphi+\varepsilon)^2 - 4\gamma\lambda}}{2}$$

For any $z \in \mathbb{R}^p$ and $i \in [\![1, p]\!]$,

$$\big[\text{prox}_{\gamma, \lambda P_\varepsilon^\Sigma}(z)\big]_i = \begin{cases} 0 & |z_i| \leq \varphi \\ \text{sign}(z_i)\ x^\star\big(|z_i|\big) & |z_i| > \varphi \end{cases}$$

Proof

Recall, we seek an *explicit* expression of $\text{prox}_{\gamma, \lambda P_\varepsilon^\Sigma}, z \in \mathbb{R}^p \longmapsto \arg\min_{x \in \mathbb{R}^p}\Big\{\lambda P^\Sigma(x) + \frac{1}{2\gamma}\|x - z\|_2^2\Big\}$
By separability, for every $z \in \mathbb{R}^p$ and $i \in [\![1, p]\!]$

$$[\text{prox}_{\gamma, \lambda P^\Sigma}(z)]_i = \arg\min_{x \in \mathbb{R}} \underbrace{\Big\{\lambda\rho^\Sigma(x) + \frac{1}{2\gamma}(x - z_i)^2\Big\}}_{m(x; z_i)}$$

m is $C^1(\mathbb{R}^*)$ and coercive so it has at least one global minimizer.
We have for $x \neq 0$,

$$m'(x; z) = \frac{x - z}{\gamma} + \frac{\lambda}{\varepsilon + |x|}\text{sgn}(x) = \begin{cases} \dfrac{x - z}{\gamma} + \dfrac{\lambda}{\varepsilon + x} & x > 0 \\ \dfrac{x - z}{\gamma} - \dfrac{\lambda}{\varepsilon - x} & x < 0 \end{cases}$$

So $m'(x; z)$ require $sign(x) = sign(z)$ otherwise it's still negative or positive.
If $z > 0$ ($z < 0$ is analogous), then $x > 0$,

$$m'(x;z)=0 \iff \frac{x-z}{\gamma}+\frac{\lambda}{\varepsilon+x}=0 \iff x^2+(\varepsilon-z)\,x-\varepsilon z+\gamma\lambda=0$$

Solution exists if and only if,

$$(\varepsilon-z)^2-4(\lambda\gamma-z\varepsilon)\geq 0 \iff (z+\varepsilon)^2-4\gamma\lambda\geq 0 \iff z\geq 2\sqrt{\lambda\gamma}-\varepsilon$$

So m' vanishes for $z\geq 2\sqrt{\lambda\gamma}-\varepsilon$ on a unique point. Indeed the other root, if discriminant is strictly positive, become negative and we assumed $x>0$.

$$x_z^*=\frac{z-\varepsilon+\sqrt{(z+\varepsilon)^2-4\gamma\lambda}}{2}$$

Then in 0, $\partial\rho_\varepsilon^\Sigma(0)=[-1/\varepsilon,\,1/\varepsilon]$, we have

$$0\in\partial m(0;z_i) \iff -\frac{z_i}{\gamma}\in\lambda\big[-\frac{1}{\varepsilon},\frac{1}{\varepsilon}\big] \iff |z_i|\leq\frac{\lambda\gamma}{\varepsilon}$$

To summarize for positive z ,(by symmetry we get the same interval but negative)

$$\textbf{On } [0,\,2\sqrt{\gamma\lambda}-\varepsilon[\quad \textbf{On } [2\sqrt{\gamma\lambda}-\varepsilon,\,\tfrac{\lambda\gamma}{\varepsilon}] \quad \textbf{On }]\tfrac{\lambda\gamma}{\varepsilon},\,\infty[$$

$$\text{Candidate: } 0 \qquad \text{Candidates: } 0 \text{ or } x^* \qquad \text{Candidate: } x_z^*$$

Let's find the good candidate. Define the differentiable function

$$H: z\in\Big[2\sqrt{\gamma\lambda}-\varepsilon;\frac{\lambda\gamma}{\varepsilon}\Big]\mapsto m(x_z^*;z)-m(0;z)$$

$$H'(z)=\partial_z m(x_z^*;z)-\partial_z m(0;z)=\big[m'(x_z^*;z)\partial_z x_z^*+\partial_z m(x_z^*;z)\big]-\partial_z m(0;z)=0+\frac{z-x_z^*}{\gamma}-\Big(-\frac{0-z}{\gamma}\Big)=-\frac{x_z^*}{\gamma}$$

$H'<0$, so H is stricly decreasing. We have for $z=2\sqrt{\lambda\gamma}-\varepsilon$, $x_z^*=\sqrt{\lambda\gamma}-\varepsilon$

$$H(z)=\lambda\ln\Big(\frac{\sqrt{\lambda\gamma}}{\varepsilon}\Big)+\frac{\Big(\sqrt{\lambda\gamma}-\varepsilon-2\sqrt{\lambda\gamma}+\varepsilon\Big)^2}{2\gamma}-\frac{(2\sqrt{\lambda\gamma}-\varepsilon)^2}{2\gamma}=\lambda\ln\Big(\frac{\sqrt{\lambda\gamma}}{\varepsilon}\Big)+\frac{\lambda}{2}-\frac{(2\sqrt{\lambda\gamma}-\varepsilon)^2}{2\gamma}$$

$$H(z)=\ln\Big(\frac{\sqrt{\gamma\lambda}}{\varepsilon}\Big)-\frac{3\lambda}{2}+\frac{2v\varepsilon\sqrt{\gamma\lambda}}{\gamma}-\frac{\varepsilon^2}{2\gamma}$$

We recall $z,x>0$ so $\sqrt{\lambda\gamma}>\varepsilon$, then,

$$H(z)=\lambda h\Big(\frac{\varepsilon}{\sqrt{\lambda\gamma}}\Big) \quad \text{with} \quad h: w\mapsto\ln\Big(\frac{1}{w}\Big)-\frac{3}{2}+2w-\frac{w^2}{2}$$

We have $h'(w)=-\frac{1}{r}+2-r<0$ on $]0,1[$ hence h strictly decreasing. Moreover, $h(1)=0$ so h is positive on $]0,1[$, then

$$H(z=2\sqrt{\lambda\gamma}-\varepsilon)>0$$

For $z=\frac{\lambda\gamma}{\varepsilon}$, then $x_z^*=\frac{\lambda\gamma}{\varepsilon}-\varepsilon$

$$H(z)=\lambda\ln\Big(\frac{\lambda\gamma}{\varepsilon^2}\Big)+\frac{\varepsilon^2}{2\gamma}-\frac{\lambda^2\gamma}{2\varepsilon^2}=\lambda g\Big(\frac{\sqrt{\lambda\gamma}}{\varepsilon}\Big)$$

with $g: w\mapsto -2\ln(w)+\frac{w^2}{2}-\frac{1}{2w^2}$ and $g'(w)=\frac{-2}{w}+w+\frac{1}{w^3}=\frac{(w^2-1)^2}{w^3}>0$. We remark $g(1)=0$ and $0<\frac{\sqrt{\lambda\gamma}}{\varepsilon}<1$ so

$$H(z=\frac{\lambda\gamma}{\varepsilon})<0$$

Perfect ! By Intermediate Value Theorem we know there exists $\varphi\in]2\sqrt{\gamma\lambda}-\varepsilon;\frac{\lambda\gamma}{\varepsilon}[$ such as $H(\varphi)=0$

$$H(\varphi)=0\iff m(x_\varphi^*;\varphi)=m(0;\varphi)\iff\lambda\ln\Big(1+\frac{x_\varphi^*}{\varepsilon}\Big)=\frac{x_\varphi^*(2\varphi-x_\varphi^*)}{2\gamma}$$

Hence, until $z=\varphi$, 0 is the minimizer and beyond, it switches to the other, x_z^*. □

Although it is referred to as an *explicit* expression, as in [PST22], the proximal operator is in fact only *implicitly*

defined, since it requires solving a nonlinear equation. Nonetheless, once ε is fixed and λ_{QUT} is chosen, this proximal operator is easier to compute than the one in Theorem 4.2, as it needs to be evaluated only once, rather than at each iteration of the proximal algorithm, such as ISTA.

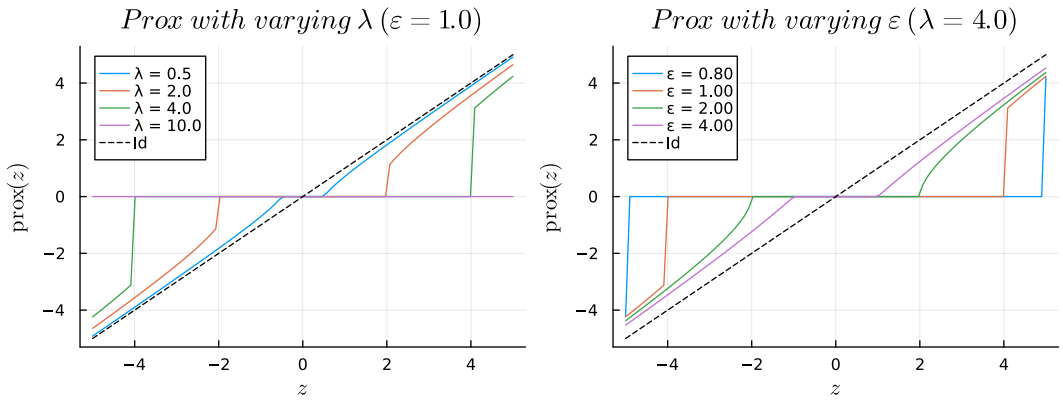

Figure 7: Graph of the proximal operator for the P_ε^Σ penalty for different values of λ and ε, with step size $\gamma = 1$. The identity line is shown for reference (dashed).

4.3.3 Numerical experiments

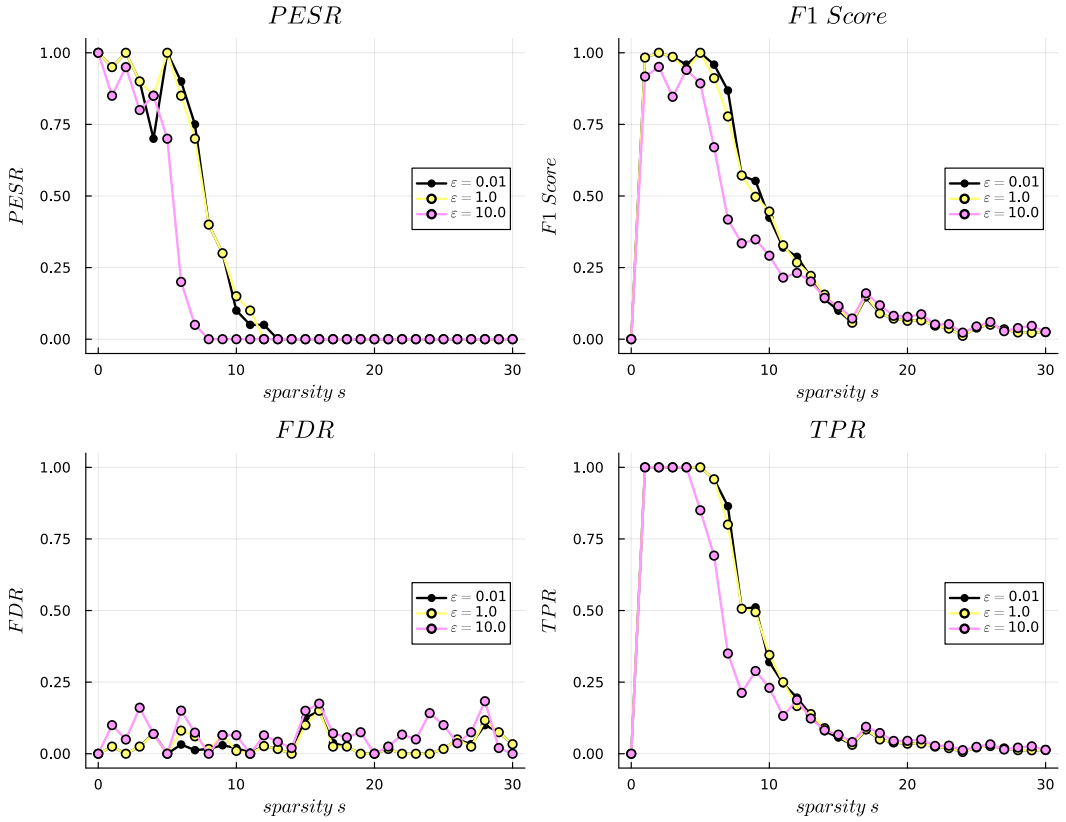

Figure 8: Performance comparison between different $\varepsilon > 0$ with $\sqrt{RSS}$ as loss function combined with P_ε^Σ in a sparse linear regression model. The simulation parameters are the same as 2 : $n = 70$, $p = 250, \sigma = 0.1, s \in [\![0, 30]\!]$, where each nonzero coefficient is set to 3. Each point in the plot is estimated from 100 Monte Carlo replications, and the Quantile Universal Threshold λ_{QUT} is computed via 1000 simulations under the null model.

We repeat the same experiment as in Figure 3, but without using a warm start in order to keep the setup simple and we obtain even better results. However, as previously mentioned, this performance does not match that achieved with the penalty P_ν (see Figure 5).

Moreover, when $\varepsilon \to 0$ does not improve the results as much as one might expect despite getting "closer" to the ℓ_0 penalty to within a coefficient factor.

5 Another perspective : Duality

Until now, we have tackled problem (1) using numerical optimization techniques, in particular proximal methods, aiming to reach a local minimum. However, due to the non-convexity of the penalty term, the algorithm may get trapped in undesirable local minima. To mitigate this, heuristic strategies such as homotopy optimization (see Figure 5) have been employed to escape these poor solutions.

In practice, especially in applications of this kind, theoretical guarantees are often sacrificed in favor of empirical performance. The priority shifts toward achieving good predictive scores, rather than pursuing rigorous mathematical understanding or aesthetic theoretical properties.

In this section, we present an alternative approach that we explored—based on duality—even though **it ultimately failed**. Our aim in documenting this attempt is twofold: to prevent others from following the same unproductive direction and to provide insight that may help refine or correct the method in future work.

5.1 Duality in Non-Convex Problems

Consider an optimization problem of the form where $f : \mathcal{X} \mapsto \bar{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$, $g : \mathcal{X} \mapsto \mathbb{R}^m$, and $\mathcal{X}$ is a Hilbert space,

$$\min_{x \in \mathcal{X}} f(x) \quad \text{s.t.} \quad g(x) \leq 0 \tag{3}$$

When the optimization problem (3) satisfies Slater's condition (thus making it a suitable convex problem, see [BV04]), a key property arises : zero duality gap. The optimal value of the primal problem matches exactly that of the dual problem, and each primal solution can be associated with a dual solution. Hence, solving the dual provides as much information as solving the primal.

But why would one prefer the dual approach? Typically, this is motivated by dimensional considerations. In high-dimensional settings, especially when $n < p$, the dual optimization problem operates in a smaller space $\mathbb{R}^n$ compared to the primal space $\mathbb{R}^p$. Additionally, the dual formulation often offers valuable geometric insights. For instance, as shown by [GVR11] in the case of LASSO, the dual perspective helps identify which features will vanish at optimality before actually solving the optimization problem.

In our study, we are interested in exploring new types of penalties, such as P_ν. However, under these conditions, strong duality generally fails to hold.

To deal with such situations, we will rely on the theoretical framework developed by Bednarczuk and Syga [BS20], which generalizes classical duality to non-convex optimization problems using abstract convex functions, Φ-convex function, where $\Phi \subset \mathbb{R}^{\mathcal{X}}$.

Below, we introduce the key definitions and propositions necessary to apply their framework to our context.

Definition 5.1: Φ-convex function

A proper function $f : \mathcal{X} \to \bar{\mathbb{R}}$ is said to be Φ-convex if

$$\forall x \in \mathcal{X},\ f(x) \ = \ \sup\{\varphi(x) \mid \varphi \in \Phi,\ \varphi \leq f\}$$

Proposition 5.1: Affine Φ-convexity $\iff$ Classical convexity

Let a proper function $f : \mathcal{X} \to \bar{\mathbb{R}}$ and define $\Phi_{Aff} := \{x \in \mathcal{X} \mapsto \langle a, x\rangle + b \in \mathbb{R} \mid a \in \mathcal{X}^*, b \in \mathbb{R}\}$.

$$f \text{ is classically convex} \iff f \text{ is } \Phi_{Aff}\text{-convex}$$

Proof

($\Rightarrow$) If f is Φ_{Aff}-convex, each $\varphi \in \Phi_{Aff}$ is convex, and the pointwise supremum of convex functions is convex. Hence f is convex. Geometrically,

$$\text{epi}(f) = \bigcap_{\varphi \in \Phi_{\text{Aff}}} \text{epi}(\varphi)$$

an intersection of convex sets, so epi f (and thus f) is convex.
($\Leftarrow$) If f is convex, let $x_0 \in \operatorname{dom} f$ and let $(x_0, f(x_0)) \in \operatorname{epi}(f)$. By the Hahn–Banach separation theorem, there exist $(a, -1) \in \mathcal{X}^* \times \mathbb{R}$ and $b \in \mathbb{R}$ such that

$$\langle a, x\rangle + (-1)\cdot r + b \leq 0 \quad \forall (x, r) \in \operatorname{epi}(f) \quad \text{and} \quad \langle a, x_0\rangle - f(x_0) + b = 0$$

Define $\varphi_{x_0}(x) := \langle a, x\rangle + b$. Then $\varphi_{x_0} \leq f$ and

$$f(x_0) = \varphi_{x_0}(x_0) \leq \sup\{\varphi(x_0) \mid \varphi \in \Phi_{Aff},\, \varphi \leq f\} \leq f(x_0)$$

Since x_0 was arbitrary, f is Φ_{Aff}-convex. □

Definition 5.2: Φ-subgradient [BS20, Definition 2]

A function $\varphi \in \Phi$ is called a Φ-subgradient of $f : \mathcal{X} \mapsto \bar{\mathbb{R}}$ at $\bar{x} \in \operatorname{dom}(f)$ if

$$\forall x \in \mathcal{X}, \quad f(x) - f(\bar{x}) \geq \varphi(x) - \varphi(\bar{x})$$

The set of all Φ-subgradients of f at $\bar{x}$ is denoted by,

$$\partial_\Phi f(\bar{x}) := \{\varphi \in \Phi : \varphi \text{ is a } \Phi\text{-subgradient of } f \text{ at } \bar{x}\}$$

Definition 5.3: Φ-conjugate

Let $f : \mathcal{X} \to \bar{\mathbb{R}}$ be a proper function. We call the Φ-conjugate of f,

$$\begin{aligned} f^*_\Phi : X &\longrightarrow \bar{\mathbb{R}} \\ \varphi &\longmapsto \sup_{x\in\mathcal{X}} \big\{\varphi(x) \;-\; f(x)\big\} \end{aligned}$$

Remark 6

When $\Phi = \Phi_{Aff}$, f^*_Φ is the classical Fenchel conjugate.
f doesn't need to be Φ-convex, but we have always f^*_Φ Φ-convex [PR97, Property 1.2.3]
Then, from [BS20, Theorem 1],

$$f \text{ is } \Phi\text{-convex} \iff f = (f^*_\Phi)^*_\Phi$$

Definition 5.4: Perturbation function

Let $\mathcal{Y}$ be a topological vector space. A perturbation function for the objective $f : \mathcal{X} \to \bar{\mathbb{R}}$ is any proper function $p : \mathcal{X} \times \mathcal{Y} \longrightarrow \bar{\mathbb{R}}$ such that there exists $y_0 \in \mathcal{Y}$ (typically $y_0 = \mathbf{0}$),

$$\forall x \in \mathcal{X},\; p(x, y_0) \;=\; f(x)$$

Remark 7: Recasting the constrained problem as unconstrained

Consider the constrained programme (3). Introduce the set-valued constraint operator

$$\begin{aligned} \mathcal{A} : \mathcal{Y} &\rightrightarrows \mathcal{X} \\ y &\longmapsto \big\{\, x \in \mathcal{X} \;\big|\; \forall i \in [\![1, m]\!]\; g_i(x) \leq y_i \big\} \end{aligned}$$

With the distinguished parameter $y_0 \in \mathcal{Y}$ the problem reads

$$\min_{x\in\mathcal{X}} f(x) \quad \text{s.t.} \quad x \in \mathcal{A}(y_0)$$

Define a perturbation map as

$$p : \mathcal{X} \times \mathcal{Y} \longrightarrow \bar{\mathbb{R}}$$
$$(x, y) \longmapsto \begin{cases} f(x) & x \in \mathcal{A}(y) \\ +\infty & \text{otherwise} \end{cases}$$

Thus the original constrained problem is equivalent to the unconstrained minimization

$$\min_{x \in \mathcal{X}} p(x, y_0) \tag{4}$$

Definition 5.5: Lagrangian; Primal, and dual problems [BS20]

Let $f : \mathcal{X} \to \bar{\mathbb{R}}$ the function to minimize and $p : \mathcal{X} \times \mathcal{Y} \to \bar{\mathbb{R}}$ a perturbation of f with distinguished point $y_0 \in \mathcal{Y}$, i.e. $p(x, y_0) = f(x),\ \forall x \in \mathcal{X}$. Fix a family of elementary functions $\Psi \subset \mathbb{R}^{\mathcal{Y}}$ (often we consider $\Psi \equiv \Phi$, the same class of functions applied to different underlying spaces) and set for $x \in \mathcal{X},\ p_x := p(x, \cdot)$
Define the Lagrangian,

$$\mathcal{L} : \mathcal{X} \times \Psi \longrightarrow \bar{\mathbb{R}}$$
$$(x, \psi) \longmapsto \psi(y_0) - p_x{}^*_\Psi(\psi) = \psi(y_0) - \sup_{y \in \mathcal{Y}} \{\psi(y) - p(x, y)\}$$

We state the primal problem,

$$\inf_{x \in \mathcal{X}} \sup_{\psi \in \Psi} \mathcal{L}(x, \psi) \tag{5}$$

And the dual problem,

$$\sup_{\psi \in \Psi} \inf_{x \in \mathcal{X}} \mathcal{L}(x, \psi) \tag{6}$$

Remark 8: When the abstract primal collapses to the original problem

For all $x \in \mathcal{X}$ set $p_x = p(x, \cdot)$. By definition of the Lagrangian,

$$\sup_{\psi \in \Psi} \mathcal{L}(x, \psi) = \sup_{\psi \in \Psi} \{\psi(y_0) - p_x{}^*_\Psi(\psi)\} = (p_x{}^*_\Psi)^*_\Psi(y_0)$$

From Remark 6, if p_x is Ψ-convex for all $x \in \mathcal{X}$, then

$$\sup_{\psi \in \Psi} \mathcal{L}(x, \psi) = p_x(y_0) = p(x, y_0) = f(x)$$

Consequently, the primal problem (5) reduces exactly to the original problem (3)

$$\inf_{x \in \mathcal{X}} \sup_{\psi \in \Psi} \mathcal{L}(x, \psi) = \inf_{\substack{x \in \mathcal{X} \\ g(x) \leq 0}} f(x)$$

Usually, the classical Lagrangian is constructed using affine functionals, Φ_{Aff}. We will enlarger this function set to lower semi continuous functions,

$$\Phi_{lsc} = \{\phi : \mathcal{X} \mapsto \mathbb{R},\ \phi(x) = -a\|x\|^2 + \langle l, x \rangle + c, \quad l \in \mathcal{X}^*,\ a \geq 0,\ c \in \mathbb{R}\}$$

Theorem 5.1: Strong duality (Zero duality gap) [BS20, Theorem 8]

Let a perturbation map $p : \mathcal{X} \times \mathcal{Y} \to \bar{\mathbb{R}}$. If for all $x \in \mathcal{X}$, $p(x, \cdot)$ is Ψ_{lsc}-convex at the reference point $y_0 \in \mathcal{Y}$ and if the value function

$$V : \mathcal{Y} \longrightarrow \bar{\mathbb{R}}$$
$$y \longmapsto \inf_{x \in \mathcal{X}} p(x, y)$$

is proper, lower semicontinuous and notably paraconvex[a], and $y_0 \in \mathring{\text{dom}}\, V$.
Then, strong duality holds, that's to say,

$$(5) \ = \ (6)$$

Moreover the dual is solvable, i.e. there exists at least one

$$\psi^\star \in \arg\max_{\psi\in\Psi_{lsc}} \inf_{x\in\mathcal{X}} \mathcal{L}(x,\psi),$$

and every such maximizer satisfies

$$\psi^\star \in \partial_{\Psi_{lsc}} V(y_0)$$

[a]Meaning $V + c\|\cdot\|^2$ is convex for some $c > 0$.

This framework allows to get the zero gap duality property and hope to find another road to minimize (1) with non convex penalties.

For simplicity, the class Φ_{lsc} will be denoted by Φ throughout the report.

5.2 Dual formulation with $\sqrt{\text{RSS}}$ loss over Φ_{lsc}

We work with the loss function $f = \sqrt{RSS}$ and g a suitable penalty whose dual treatment relies on the elementary family $\Phi = \Psi = \Phi_{lsc}$. To keep the exposition clear, the intercept block θ_2 is omitted, hence $\theta = \theta_1$. Following the "tautological" trick of [GVR11], we embed the equality constraint $X\theta = z$ via an auxiliary variable. Instead of solving $\min_{\theta\in\mathbb{R}^n} \|y_{obs} - X\theta\|_2 + g(\theta)$, we solve,

$$\min_{(\theta,z)\in\mathbb{R}^n\times\mathbb{R}^p} \|y_{obs} - z\|_2 + g(\theta) \quad \text{s.t} \quad X\theta = z \tag{7}$$

Let $\mathcal{X} := \mathbb{R}^{n+p}$, $\mathcal{Y} := \mathbb{R}^{2n}$ and the cost function $F : x = (\theta, z) \in \mathcal{X} \mapsto f(z) + g(\theta) \in \mathbb{R}$ Equality $X\theta = z$ is encoded through two one-sided inequalities. Write a parameter $y \in \mathcal{Y}$ as $y = (y^+, y^-)$ with $y^+, y^- \in \mathbb{R}^n$ and set

$$\begin{aligned} \mathcal{A} : \mathcal{Y} &\rightrightarrows \mathcal{X} \\ y = (y^+, y^-) &\longmapsto \left\{x = (\theta, z) \in \mathcal{X} \mid X\theta - z \le y^+,\ z - X\theta \le y^-\right\} \end{aligned}$$

With the distinguished parameter $y_0 = (\mathbf{0}, \mathbf{0})$ one has $\mathcal{A}(y_0) = \{(\theta, z) \mid X\theta = z\}$, so (7) fits. We perturb F with

$$\begin{aligned} p : \mathcal{X}\times\mathcal{Y} &\longrightarrow \bar{\mathbb{R}} \\ (x, y) &\longmapsto \begin{cases} F(x) & x \in \mathcal{A}(y) \\ +\infty & \text{otherwise} \end{cases} \end{aligned}$$

The first goal is to verify if we verify assumptions of Theorem 5.1. However with this constraint and perturbation function p, when we define the primal value function V, we have

$$\text{dom}(V) = \left\{y = (y^+, y^-) \in \mathcal{Y} \ \mid\ \exists(\theta, z) \in \mathcal{X} \text{ s.t. } -y^- \preceq X\theta - z \preceq y^+\right\} = \{(y^+, y^-) \in \mathcal{Y} : y^+ + y^- \succeq 0\}$$

and unfortunately $y_0 = (\mathbf{0}, \mathbf{0}) \notin \mathring{\text{dom}}(V)$. A way to get around this is to relax the constraint by choosing for $\varepsilon > 0$, $\|X\theta - z\|_2 \le \epsilon$. We set instead $\mathcal{Y} = \mathbb{R}$ and

$$\begin{aligned} \mathcal{A} : \mathcal{Y} &\rightrightarrows \mathcal{X} \\ y &\longmapsto \left\{x = (\theta, z) \in \mathcal{X} \mid \|X\theta - z\|_2 - \varepsilon \le y\right\} \end{aligned}$$

p keep the same definition and using the indicator of $\mathcal{A}(y)$, we can express as

$$p : (x, y) \in \mathcal{X}\times\mathcal{Y} \mapsto F(x) + \iota_{\mathcal{A}(y)}(x) \in \bar{\mathbb{R}}$$

Therefore,

$$\text{dom}(V) = \{\, y \in \mathcal{Y} \mid y \ge -\varepsilon\} \quad \text{and} \quad y_0 = 0 \in \mathring{\text{dom}}(V) =]-\varepsilon; +\infty[$$

Moreover, for $x \in \mathcal{X}$, $p_x := p(x, \cdot)$ is an indicator function of a closed set hence p_x is lsc so p_x is Φ-convex and in particular at y_0. The last assumptions to verify is about

$$V : \ y \in]-\varepsilon\ +\infty[\longmapsto \inf_{x\in\mathcal{X}} p(x, y) = \inf_{\theta\in\mathbb{R}^p, z\in\mathbb{R}^n} \Big\{\ \|z - y_{obs}\|_2 + g(\theta) \quad \Big| \quad \|X\theta - z\|_2 \le \varepsilon + y\ \Big\}$$

V is the infimum of lower semi continuous function proper so its lower semi continuous and proper on $]-\varepsilon; +\infty[$. Let's verify if V is weakly-convex. From projection on Euclidean ball, we have for $y > -\varepsilon$,

$$V(y) \ = \ \inf_{\theta\in\mathbb{R}^p} \Big[\underbrace{\max\{\, 0, \|X\theta - y_{obs}\|_2 - (\varepsilon + y)\,\}}_{:=\ \psi(\theta, y)} \ + g(\theta)\Big] = \inf_{\theta\in\mathbb{R}^p} \big\{\psi(\theta, y) + g(\theta)\big\}$$

ψ is jointly convex on (θ, y) as maximum of convex functions. Hence V is weakly-convex if g is weakly-convex.

Nonetheless, P_ν (and so many in literature : $P_\varepsilon^\Sigma, SCAD, MCP, \dots$) isn't weakly-convex due to their behavior near to 0. Indeed $\rho_\nu''(0^+) = +\infty$. We **can't** use Theorem 5.1.

6 Discussion

6.1 Precisions

We omitted neural networks their inclusion would merely inflate the numerical experiments, contribute no additional theoretical insight, and rely on some heuristics. Moreover, attaining a global minimizer is substantially more challenging in that setting, so the homotopy optimization scheme of Section 4.2.3 would be indispensable. Finally, although classification problems were not examined here, the proposed methodology can still be applied, see again [SvCM25].

Although the algorithm's complexity is theoretically high, its practical runtime is modest. Each subproblem is halted once a fixed error tolerance is met, and the resulting warm-start scheme is inexpensive because every iterate already lies near the next stopping point. Hence the model is fitted only once, avoiding the highly variable and costly repetitions of cross-validation. Warm starts are especially valuable in rugged, non convex landscapes far more demanding than the linear example presented here. In practice, the overhead from proximal updates, warm starts, and the choice of λ_{QUT} is negligible relative to cross-validation and certainly not the bottleneck of the method.

The procedure relies on the $(1-\alpha)$ quantile of the test statistic. In all experiments we adopted the conventional choice $\alpha = 0.05$, but this threshold is ultimately arbitrary. Taking $\alpha = 0$, that is the maximum, would delay the PESR transition to far sparser models (larger s before the PESR drops to 0), yet it simultaneously caps the PESR well below 1, limiting its usefulness when near certainty is required.

Multiple random initializations of the primal problem might be tempting, with the smallest objective value retained. Yet a local minimizer with a larger objective can occasionally be more desirable because it annihilates all irrelevant coefficients, see Figure 9. This raises the question of whether a diagnostic such as the duality (or stationary) gap could reliably distinguish "useful" minima from spurious ones.

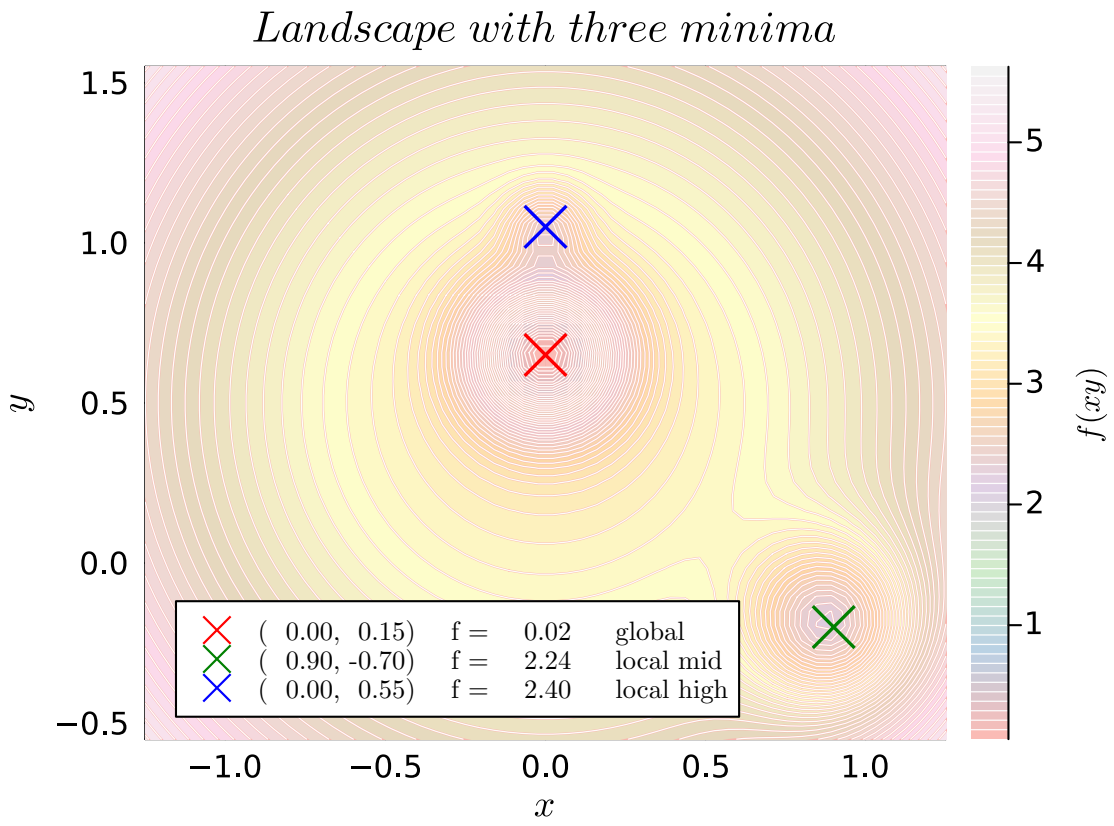

Figure 9: Landscape for $p = 2$. Red cross is the global minimum with x_1 active and $x_2 = 0$. Blue cross is a sparse local minimum (same support, higher cost) while Green cross does not have the same support as the global minimum. Random restarts may land on blue or green, ranking solely by objective value can thus reject the sparser, more useful blue solution.

6.2 Ideas

An immediate research avenue is to equip each stationary point with an a posteriori **certificate**. Is the detected minimum admissible, or even global? In the convex regime (convex loss and penalty) ISTA already converges to

the unique minimizer, rendering certification unnecessary. With non convex penalties, though, the Bednarczuk–Syga framework offers a potential path via Theorem 5.1. The **gap function**, thanks to the point of view of the dual, can consistently **rank local minima** by practical relevance ? See Figure 9.

Assuming the Bednarczuk–Syga certification framework applies, we must still work with a non convex penalty to surpass ℓ_1, so the dual formulation ultimately requires locating a global minimizer of a non convex objective. Although more sophisticated local search routines exist, they are outside this report's scope. We discard Marination–Minimization (MM) and Successive Convex Approximation (SCA) because their surrogate problems lack theoretical guarantees, making proximal strategies preferable. Alternating Direction Method of Multipliers (ADMM) proved similarly unhelpful in preliminary tests, even on ℓ_1 penalty. A promising but underexplored direction is to optimize the **Moreau envelope**. Under weak convexity assumptions it retains the same global minimizer and cost yet is smoother and particularly compatible with proximal methods.

A further route left largely unexplored in the literature, maybe for a good reason, is to exploit the Moreau envelope even though the penalty is not globally weakly convex, here mainly P_ν. Away from the origin, P_ν is locally weakly convex, so one could backtrack on the weak convexity constant, minimizing the corresponding Moreau envelope only while that bound holds. Because the envelope preserves the same global minimizer and objective value under weak-convexity, this phase offers a smoother landscape for proximal descent. Once iterates enter the region where coefficients approach zero, we would switch back to plain ISTA on the original objective. The strategy thus descends a regularized surface until reaching the delicate neighborhood of zero, where the exact penalty is reinstated. The scheme is effectively a **backtracking on the weak-convexity constant**, coupled with an **adaptive swap** between envelope-based and original proximal updates.

A complementary direction is to **select a penalty** that guarantees reliable support recovery and satisfies weak convexity, rendering it compatible with the Moreau envelope and Theorem 5.1. The target profile would include: weak convexity, positivity, continuity, differentiability almost everywhere, coercivity (ideally linear or super-linear at infinity to simplify the dual), even symmetry, a finite slope at the origin (for QUT compatibility), separability (to ease proximal analysis and interpretation), and a finite second derivative.

Returning to the Bednarczuk–Syga framework, an unresolved issue is whether there exists an **intermediate class**,

$$\Phi_{\text{Aff}} \subsetneq \Phi \subsetneq \Phi_{\text{lsc}}$$

strictly larger than affine functions yet strictly smaller than all lower-semi continuous ones. This new class can be not too large to be relevant and handleable.

For linear regression, we likewise wonder whether a **sharper feasibility set** than the tautology $X\theta = z$, or a **more informative relaxation** than $\|X\theta - z\|_2 \leq \varepsilon$, can be formulated.

Throughout this report we have used the practical λ_0^{local}. The **ideal threshold** λ_0 is generally inaccessible known in closed form only in the convex case (see Remark 2.2) but would supersede its local proxy whenever a reliable computation becomes available.

7 Algorithm tools

7.1 Minimizing real function

Definition 7.1: Proximity operator

Let $\gamma > 0$ and function $g : \mathbb{R}^n \to \mathbb{R}$, for $n > 0$, we call proximal operator of g with γ,

$$prox_{\gamma,g} : v \in \mathbb{R}^n \mapsto \arg\min_{x \in \mathbb{R}^n} \left\{ g(x) + \frac{1}{2\gamma}\|x - v\|_2^2 \right\}$$

7.1.1 ISTA

Iterative Soft-Thresholding Algorithm (ISTA) is a proximal gradient method, each iteration (w_k) performs a gradient descent step on the smooth loss followed by a soft-thresholding with a γ as step size.

Definition 7.2: ISTA, [BT09]

To minimize the problem $f + g$ with f convex, where its differential is L-Lipschitz and g convex[a]. We approximate the minimizer w^* by $(w_k)_{k\geq 0}$ defined as, w_0 an initial point and with the update rule, for $\gamma > 0$,

$$w^{(k+1)} = \text{prox}_{\gamma,g}\Big(w_k - \gamma\, \partial f(w_k)\Big)$$

[a] We assume g convex to get a global minimum. If it doesn't, the method is still defined but we don't approximate the global minimum : we will fall in a local minimum.

Usually we take $\gamma = \frac{1}{L}$ to exploit the regularity of f and get convergence of order 1.

Proposition 7.1: Convergence of ISTA

Under the hypotheses of Definition 7.2, let $\gamma \in \left]0, \frac{1}{L}\right[$ and we denote $F := f + g$. The algorithm converge sub-linearly,

$$F(w_k) - F(w^*) = \mathcal{O}\Big(\frac{1}{k}\Big) \quad \text{as } k \to +\infty$$

Proof

Since ∂f is L-Lipschitz and $\gamma < \frac{1}{L}$, for $k \geq 0$,

$$f(w_{k+1}) \leq f(w_k) + \partial f(w_k)^\top (w_{k+1} - w_k) + \frac{L}{2} \|w_{k+1} - w_k\|^2$$

Adding $g(w_{k+1})$ and using the optimality of $w_{k+1} = \operatorname{prox}_{\gamma, g}\big(w_k - \gamma\, \partial f(w_k)\big)$, we have

$$g(w_{k+1}) + \partial f(w_k)^\top (w_{k+1} - w_k) + \frac{1}{2\gamma} \|w_{k+1} - w_k\|^2 \leq g(w_k)$$

Summing the two inequalities gives

$$F(w_{k+1}) = f(w_{k+1}) + g(w_{k+1}) \leq f(w_k) + g(w_k) - \frac{1}{2}\Big(\frac{1}{\gamma} - L\Big) \|w_{k+1} - w_k\|^2 \leq F(w_k)$$

so in particular $F(w_{k+1}) \leq F(w_k)$ for any $\gamma \in \left]0, \frac{1}{L}\right]$. Next, we remark optimality condition at w_{k+1} gives,

$$0 \in \partial g(w_{k+1}) + \frac{1}{\gamma}\big(w_{k+1} - w_k + \gamma \partial f(w_k)\big) \iff \frac{w_k - w_{k+1}}{\gamma} - \partial f(w_k) \in \partial g(w_{k+1})$$

Then, we set,

$$s_{k+1} := \frac{w_k - w_{k+1}}{\gamma} + \big(\partial f(w_{k+1}) - \partial f(w_k)\big) \in \partial F(w_{k+1})$$

And because F is convex, we have in particular,

$$F(w^*) \geq F(w_{k+1}) + s_{k+1}^T (w^* - w_{k+1}) \iff F(w_{k+1}) - F(w^*) \leq \frac{1}{\gamma} (w_k - w_{k+1})^\top (w_{k+1} - w^*)$$

Using the three-point identity [a] $(w_k - w_{k+1})^\top (w_{k+1} - w^*) = \frac{1}{2}\Big(\|w_k - w^*\|^2 - \|w_{k+1} - w^*\|^2 - \|w_k - w_{k+1}\|^2\Big)$. Hence

$$F(w_{k+1}) - F(w^*) \leq \frac{1}{2\gamma}\big(\|w_k - w^*\|^2 - \|w_{k+1} - w^*\|^2\big)$$

By non increasing and then by telescoping,

$$k\big(F(w_k) - F(w^*)\big) \leq \sum_{i=0}^{k-1} F(w_{i+1}) - F(w^*) \leq \frac{\|w_0 - w^*\|^2}{2\gamma}$$

Hence, as $k \to \infty$,

$$F(w_k) - F(w^*) = \mathcal{O}\left(\frac{1}{k}\right)$$

□

[a] $\forall\, a, b, c \in \mathbb{R}^n, \quad 2\langle a - b, b - c\rangle = \|a - c\|^2 - \|a - b\|^2 - \|b - c\|^2$

This algorithm is very powerful to reach minimizer [BT09] once we have a way to compute the proximal operator of g and select a correct step size γ, although L can be unknown... We enlarge assumptions on the precedent definition to allow the gradient of f to be locally Lipschitz and use backtracking on this Lipschitz constant.

Definition 7.3: Quadratic surrogate Q

For $y \in \mathbb{R}^n$ and $L > 0$, define the linearization of f at y plus a local quadratic regularization and g

$$Q_L(x,y) \;:=\; f(y) \;+\; \partial f(y)^\top (x-y) \;+\; \frac{L}{2}\|x-y\|^2 \;+\; g(x)$$

Its unique minimizer is

$$p_L(y) \;:=\; \arg\min_x Q_L(x,y) \;=\; \operatorname{prox}_{\frac{1}{L},g}\Big(y - \frac{1}{L}\partial f(y)\Big)$$

We write $L(f;\mathcal{U})$ for any constant such that the gradient is Lipschitz on a set $\mathcal{U} \subset \mathbb{R}^n$ and denote $[y,x] := \{y + t(x-y) \;:\; t \in [0,1]\}$
If $L \;\geq\; L\big(f;\,[y,x]\big)$, then $F(x) \;\leq\; Q_L(x,y)$ for all x and y and thus $F\big(p_L(y)\big) \;\leq\; Q_L\big(p_L(y),y\big)$.

Algorithm 1 ISTA with backtracking [BT09]

Require: $w_0 \in \mathbb{R}^n$, $L_0 > 0$, $\eta > 1$
$k \leftarrow 0$
while not converged **do** ▷ e.g. $\|w_{k+1} - w_k\| \leq \varepsilon$
  $L \leftarrow L_k$
  repeat
    $x \leftarrow \operatorname{prox}_{\frac{1}{L},g}\Big(w_k - \frac{1}{L}\partial f(w_k)\Big)$
    $q \leftarrow f(w_k) + \partial f(w_k)^\top (x - w_k) + \frac{L}{2}\|x - w_k\|^2 + g(x)$
    if $F(x) > q$ **then**
      $L \leftarrow \eta\, L$
    end if
  until $F(x) \leq q$
  $w_{k+1} \leftarrow x$
  $L_{k+1} \leftarrow L$
  $k \leftarrow k+1$
end while

Remark 9: Backtracking keeps ISTA converging

Following the same steps as in the proof of Proposition 7.1, we still have $(F(w_k))_{k\geq 0}$ non increasing. Indeed,

$$F(w_{k+1}) \;\leq\; Q_{L_k}(w_{k+1}, w_k) \;\leq\; F(w_k)$$

If ∂f is globally $L(f;\,\mathbb{R}^n)$-Lipschitz then $L_k \;\leq\; \eta\, L(f;\,\mathbb{R}^n)$ for $\eta > 1$ and $F(w_k) - F(w^*) \;\leq\; \frac{\eta\, L(f)}{2k}\,\|w_0 - w^*\|^2$.
If ∂f is only locally Lipschitz on a bounded neighborhood of the initial point w_0

$$\mathcal{L}_0 \;:=\; \{x \in \mathbb{R}^n :\; F(x) \leq F(w_0)\}$$

then there exists $L_{\mathcal{L}_0} < \infty$ such that ∂f is locally $L_{\mathcal{L}_0}$-Lipschitz on $\mathcal{L}_0$. Consequently, the backtracking accepts $L_k \;\leq\; \eta\, L_{\mathcal{L}_0}$ and

$$F(w_k) - F(w^*) \;\leq\; \frac{\eta\, L_{\mathcal{L}_0}}{2k}\,\|w_0 - w^*\|^2$$

7.2 Find root

To compute proximal operator of P_ν with $0 < \nu < 1$, we will need to solve scalar equations of the form $\phi(x) = 0$. We make explicit the assumptions of two common methods.

7.2.1 Bisection method

Let $\phi : [a_0, b_0] \to \mathbb{R}$ be continuous with $a_0 < b_0$ and $\phi(a_0)\,\phi(b_0) \leq 0$. Then there exists $x^* \in [a_0, b_0]$ such that $\phi(x^*) = 0$. Existence follows from the intermediate value theorem. If in addition ϕ is strictly monotone on $[a_0, b_0]$ then x^* is unique.

Algorithm 2 Bisection method

Require: $a_0 < b_0$, ϕ with $\phi(a_0)\phi(b_0) \leq 0$, $\varepsilon > 0$

$k \leftarrow 0$
while not converged **do** ▷ e.g. $|b_k - a_k| \leq \varepsilon$ or $|\phi(m_k)| \leq \varepsilon$
  $m_k \leftarrow \frac{1}{2}(a_k + b_k)$
  if $\phi(a_k)\,\phi(m_k) \leq 0$ **then**
    $a_{k+1} \leftarrow a_k$, $b_{k+1} \leftarrow m_k$
  else
    $a_{k+1} \leftarrow m_k$, $b_{k+1} \leftarrow b_k$
  end if
  $k \leftarrow k + 1$
end while
return $x_{\text{bis}} \leftarrow m_k$

Proposition 7.2: Linear convergence of bisection method

Let $\phi : [a_0, b_0] \to \mathbb{R}$ be continuous with $a_0 < b_0$ and $\phi(a_0)\,\phi(b_0) \leq 0$. We define the sequence $(m_k)_{k\geq 0}$ as done in Algorithm 2. Then,

$$m_k \xrightarrow[k\to\infty]{} x^* \in [a_0, b_0] \quad \text{where} \quad \phi(x^*) = 0$$

and the convergence is linear.

Proof

Let $I_k := [a_k, b_k]$ and $m_k = \frac{1}{2}(a_k + b_k)$. Either $\phi(a_k)\phi(m_k) \leq 0$ or $\phi(m_k)\phi(b_k) \leq 0$, hence $x^* \in I_{k+1} \subset I_k$ for all k. Therefore (I_k) is a nested sequence of closed intervals and

$$b_{k+1} - a_{k+1} = \frac{1}{2}(b_k - a_k) \quad \Rightarrow \quad b_k - a_k = 2^{-k}(b_0 - a_0)$$

Since $x^* \in I_k$ and m_k is the midpoint of I_k,

$$|m_k - x^*| \;\leq\; \frac{1}{2}(b_k - a_k) \;=\; 2^{-(k+1)}(b_0 - a_0)$$

If there exists $k \geq 0$ such that $m_k = x^*$, we are done. Otherwise, for $k \geq 0$ we have $m_k \neq x^*$ and

$$0 \;\leq\; \frac{|m_{k+1} - x^*|}{|m_k - x^*|} \;\leq\; \frac{2^{-(k+2)}(b_0 - a_0)}{2^{-(k+1)}(b_0 - a_0)} \;=\; \frac{1}{2}$$

Hence

$$\limsup_{k\to\infty} \frac{|m_{k+1} - x^*|}{|m_k - x^*|} \;\leq\; \frac{1}{2} \;<\; 1$$

Therefore $m_k \xrightarrow[k\to\infty]{} x^*$ linearly. □

7.2.2 Newton's method

Bisection is easy to implement but we now accelerate using Newton's method.

Let $\phi : \mathbb{R} \to \mathbb{R}$ be C^2 on an open interval I and suppose there is a simple root $x^* \in I$ (so and so $\phi'(x^*) \neq 0$) with $\phi(x^*) = 0$. The method solves the first order Taylor model at x_k, i.e. $\phi(x_k) + \phi'(x_k)(x - x_k) = 0$, leading to the update $x_{k+1} = x_k - \frac{\phi(x_k)}{\phi'(x_k)}$.

To assume stability, we add also that there exists $r > 0$ such that

$$|\,\phi'(x)\,| \;> 0 \quad \forall x \in B(x^*, r) \subset I$$

Algorithm 3 Newton's method

Require: $x_0 \in I,\ \varepsilon > 0$

$k \leftarrow 0$
while not converged **do** ▷ e.g. $|x_{k+1} - x_k| \leq \varepsilon$ or $|\phi(x_{k+1})| \leq \varepsilon$
  $g_k \leftarrow \phi'(x_k)$
  if $|g_k| = 0$ **then**
    break
  end if
  $x_{k+1} \leftarrow x_k - \frac{\phi(x_k)}{g_k}$
  $k \leftarrow k + 1$
end while
return $x_{\text{newton}} \leftarrow x_k$

Proposition 7.3: Quadratic convergence of Newton's method

Let $\phi : \mathbb{R} \to \mathbb{R}$ be C^2 on an open interval I. Let $x^* \in I$ a simple root of ϕ and there exists $r > 0$ such that $\forall x \in B(x^*, r) \subset I,\ |\,\phi'(x)\,| > 0$. Then, $x_k \xrightarrow[k\to\infty]{} x^*$ quadratically.

Proof

Set for $k \geq 0$, $e_k := x_k - x^*$. By continuity of ϕ' and ϕ'', there exist $m, M > 0$ such that $\forall x \in B(x^*, r),\ |\phi'(x)| \geq m$ and $|\phi''(x)| \leq M$. Let $x_0 \in B(x^*, \min\{r,\ m/M\})$. From mean value theorem, there exist $\xi_k, \eta_k \in [x_k, x^*]^a$

$$\phi(x_k) = \phi'(x^*)\, e_k + \tfrac{1}{2}\,\phi''(\xi_k)\, e_k^2 \quad \text{and} \quad \phi'(x_k) = \phi'(x^*) + \phi''(\eta_k)\, e_k$$

Then,

$$e_{k+1} = e_k - \frac{\phi(x_k)}{\phi'(x_k)} = -\frac{\frac{1}{2}\,\phi''(\xi_k)}{\phi'(x^*) + \phi''(\eta_k)\, e_k}\, e_k^2$$

And from bounds on ϕ' and ϕ'' and because $|e_k| \leq m/M$, for $k \geq 0$,

$$|e_{k+1}| \ \leq\ \frac{M}{2m}\,|e_k|^2 \leq \frac{1}{2}|e_k|$$

We get convergence $e_k \xrightarrow[k\to\infty]{} 0 \iff x_k \xrightarrow[k\to\infty]{} x^*$ and from the precedent inequality we deduce the quadratic convergence,

$$\lim_{k\to\infty} \frac{|e_{k+1}|}{|e_k|^2} = \frac{|\phi''(x^*)|}{2\,|\phi'(x^*)|}$$

□

[a] Regardless of the order, i.e. $[x_k, x^*] = [x^*, x_k]$

The difficulty in this method is to stay near to x^* along iterations with certainty.

References

[AC10] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 2010.

[Aka74] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 1974.

[BCW11] Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

[BJMO11] Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1):1–106, 2011.

[BKM16] Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44(2):813–852, 2016.

[BS20] Ewa M. Bednarczuk and Monika Syga. Lagrangian duality for nonconvex optimization problems with abstract convex functions. *arXiv preprint*, arXiv:2011.09194, November 2020.

[BT09] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004.

[BvdG11] Peter Bühlmann and Sara van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, 2011.

[CLSW98] Frank H. Clarke, Yuri S. Ledyaev, Ronald J. Stern, and Peter R. Wolenski. *Nonsmooth Analysis and Control Theory*, volume 178 of *Graduate Texts in Mathematics*. Springer, New York, 1998.

[CRT06] Emmanuel J. Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 2006.

[CWB08] Emmanuel J. Candès, Michael B. Wakin, and Stephen P. Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier Analysis and Applications*, 14(5-6):877–905, 2008.

[DJ94] David L. Donoho and Iain M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 1994.

[DT09] David L. Donoho and Jared Tanner. Observed universality of phase transitions in high-dimensional geometry, with implications for compressed sensing. *Philosophical Transactions of the Royal Society A*, 2009.

[ET93] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1993.

[GE03] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[GVR11] Laurent El Ghaoui, Vivian Viallon, and Tarek Rabbani. Safe feature elimination for the LASSO and sparse supervised learning problems. *arXiv preprint*, arXiv:1009.4219, 2011.

[HK70] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2009.

[Mal73] Colin L. Mallows. Some comments on c_p. *Technometrics*, 1973.

[MSH+22] Xiaoyu Ma, Sylvain Sardy, Nick Hengartner, Nikolai Bobenko, and Yen Ting Lin. A phase transition for finding needles in nonlinear haystacks with lasso artificial neural networks. *arXiv preprint arXiv:2201.08652*, 2022.

[PR97] Dieter Pallaschke and Stefan Rolewicz. *Foundations of Mathematical Optimization: Convex Analysis Without Linearity*, volume 388 of *Mathematics and Its Applications*. Kluwer Academic Publishers, Dordrecht, 1997.

[PST22] Ashley Prater-Bennette, Lixin Shen, and Erin E. Tripp. The proximity operator of the log-sum penalty. *Journal of Scientific Computing*, 93(3):67, 2022.

[Sch78] Gideon Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 1978.

[SvCM25] Sylvain Sardy, Maxime van Cutsem, and Xiaoyu Ma. Training a neural network for data interpretation and better generalization: towards intelligent artificial intelligence. arXiv preprint arXiv:2411.17180, 2025. Version 3, 17 Apr 2025.

[Tib96] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996.