Sparse additive models in high dimensions with wavelets

Sylvain Sardy1 | Xiaoyu Ma2,3

1Department of Mathematics, University of Geneva, Geneva, Switzerland
2School of Science, National University of Defense Technology, Changsha, China
3Securities Institute for Financial Studies, Shandong University, Jinan, China

Abstract
In multiple regression, when covariates are numerous, it is often reasonable to assume that only a small number of them has predictive information. In some medical applications for instance, it is believed that only a few genes out of thousands are responsible for cancer. In that case, the aim is not only to propose a good fit, but also to select the relevant covariates (genes). We propose to perform model selection with additive models in high dimensions (sample size and number of covariates). Our approach is computationally efficient thanks to fast wavelet transforms, it does not rely on cross validation, and it solves a convex optimization problem for a prescribed penalty parameter, called the quantile universal threshold. We also propose a second rule based on Stein unbiased risk estimation geared toward prediction. We use Monte Carlo simulations and real data to compare various methods based on false discovery rate (FDR), true positive rate (TPR) and mean squared error. Our approach is the only one to handle high dimensions, and has a good FDR–TPR trade-off.

KEYWORDS
LASSO, model selection, quantile universal threshold

1 INTRODUCTION
Regression aims at predicting a scalar output $y$ from an input vector of covariates $x \in \mathbb{R}^p$. After collecting measurements about a scientific phenomenon of interest, a dataset $D^{(n)} = (y^{(n)}, X^{(n)})$ called training set is built, where $n$ is the number of samples, $X^{(n)} \in \mathbb{R}^{n \times p}$ is the matrix of all...
collected row-wise input vectors \( \mathbf{x}_i = (x_{i1}, \ldots, x_{ip}) \) of length \( p \), for \( i = 1, \ldots, n \), and \( y^{(n)} \in \mathbb{R}^n \) are all corresponding outputs.

The common approach in statistical machine learning assumes a multivariate function \( \mu \) maps the input to its output, and that data are realizations from a random pair of variables \((X, Y)\), where the random vector \( X \) comes from a \( p \)-dimensional law and

\[
Y \mid X = \mu(X) + \epsilon, \quad \text{where} \quad \epsilon \sim N(0, \sigma^2).
\]

Using the information in the training set \( D^{(n)} \), the goal is to find an estimate \( \hat{\mu} \). To measure the predictive quality of \( \hat{\mu} \), one relies on an independent dataset \( D^{(n')} = (y^{(n')}, X^{(n')}) \) called test set, and calculates the predictive error of \( \hat{\mu} \) by \( \text{PE}(\hat{\mu}, D^{(n')}) = \sum_{i=1}^{n'} (y_i - \hat{\mu}(x_i))^2 \).

Predicting well on the training set does not necessarily translate to predicting well on the test set, however. To illustrate this so-called over-fitting phenomenon, consider the case where \( p > n \), the rank of \( X^{(n)} \) is \( n \), and the association \( \mu \) is linear, that is

\[
\mu(X) = \mu_{\text{lin}}(\mathbf{x}) := c + \sum_{j=1}^{p} \beta_j x_j,
\]

where \( c \) and \( \beta \) are the \( p + 1 \) parameters. In that case, by solving the least squares problem

\[
\min_{c, \beta} \text{PE}(\mu_{\text{lin}}, D^{(n)}) = \min_{c, \beta} \| y^{(n)} - c - X^{(n)} \beta \|_2^2
\]

one can find parameters \( \hat{c} \) and \( \hat{\beta} \) such that, \( \mu_{\text{lin}}(\mathbf{x}_i) = y_i \) for all measurements \( i = 1, \ldots, n \) in the training set. But although the corresponding error is null for the training set \( \text{PE}(\mu_{\text{lin}}^{\text{lin}}, D^{(n)}) = 0 \), the predictive error \( \text{PE}(\mu_{\text{lin}}^{\text{lin}}, D^{(n')}) \) for the test set is not null, and may in fact be very large, in particular if \( n \) is small and the noise SD \( \sigma \) is large. To decrease the variance (at the cost of allowing some bias with respect to the training set), a common approach consists in adding a constraint on \( \mu \) through its parameters \( \beta \). The amount of constraint is indexed by a parameter often named \( \lambda \) and called a regularization parameter. In some applications for instance, practitioners believe that only a small, but unknown, subset of the \( p \)-long input vector \( \mathbf{x} \) contains predictive information on the output. In that situation, it makes sense to constrain \( \hat{\beta} \) to have a few nonzero entries. Calling

\[
S = \{ j \in \{1, \ldots, p \} : x_j \text{ has predictive information} \},
\]

a secondary goal of regression becomes to identify \( S \), the set of inputs with predictive information; we write \( S \) for the complement of \( S \). For linear models (2), \( S \) is equivalently defined as \( S = \{ j \in \{1, \ldots, p \} : \beta_j \neq 0 \} \). Assuming a small cardinality \( s = |S| \) is often reasonable; in some medical applications for instance, inputs are thousands of genes among which only a few have some effect on the output cancer status. To estimate \( \beta \) and consequently \( S \), owing to its \( \ell_1 \)-sparsity inducing penalty on \( \beta \), LASSO (Tibshirani, 1996) solves

\[
(\hat{c}, \hat{\beta})^{\text{LASSO}} = \arg \min_{c, \beta} \text{PE}(\mu_{\text{lin}}^{\text{lin}}, D^{(n)}) + \lambda \| \beta \|_1.
\]

LASSO identifies potentially important inputs with \( S^{\text{LASSO}}_{\text{lin}} = \{ j \in \{1, \ldots, p \} : \hat{\beta}_j^{\text{LASSO}} \neq 0 \} \). Remarkably, even when \( p > n \), one can have \( S^{\text{LASSO}}_{\text{lin}} = S \) in certain linear regimes depending on
\( X^{(n)} \) and the signal-to-noise ratio (Buehlmann & van de Geer, 2011), for an appropriate choice of \( \lambda \) (Giacobino et al., 2017). In other regimes, one cannot retrieve \( S \) exactly, but one can aim at low false discovery rate (FDR) along with high true positive rate (TPR), defined by

\[
\text{FDR} = \mathbb{E}\left( \frac{|S \cap \hat{S}|}{|\hat{S}| \vee 1} \right) \quad \text{and} \quad \text{TPR} = \mathbb{E}\left( \frac{|S \cap \hat{S}|}{|S|} \right).
\]  

Controlling the FDR is the goal of the knockoffs (Barber & Candès, 2015).

Retrieving \( S \) when \( \mu \) is not linear may be harder. For instance, if \( x_1 \) is useful for prediction, but through \( |x_1| \), the best linear approximation of the absolute value function by a linear association is the constant function, that is, \( c + \beta_1 x_1 \) with \( \beta_1 = 0 \), making the first input impossible to detect with \( \hat{S}_{\text{LASSO}} \). To help detect such input entries, additive models assume that nonlinear associations may occur in all canonical directions by approximating the underlying association with

\[
\mu^{\text{add}}_{c, \mu_1, \ldots, \mu_p}(x) = c + \sum_{j=1}^{p} \mu_j(x_j),
\]

where \( \mu_j \)'s are univariate functions. Although not dense in multivariate function space, additive models provide more flexibility than linear models. To fit a wide range of univariate functions \( \mu_j \), including linear and absolute value, the expansion based approach assumes that each univariate function writes as

\[
\mu_j(x) = \sum_{k=1}^{s_n} \beta_{j,k} \varphi_k(x),
\]

where \( \{ \varphi_k \}_{k=1}^{s_n} \) are chosen basis functions and \( \{ \beta_{j,k} \}_{k=1}^{s_n} \) are their corresponding unknown coefficients that are estimated from the training set. The complexity parameter \( s_n \) grows with \( n \) for nonparametric methods to be dense in function spaces.

Letting \( \beta_j = (\beta_{j,1}, \ldots, \beta_{j,s_n}) \in \mathbb{R}^{s_n} \) for \( j = 1, \ldots, p \), expansion-based additive models are a class of functions of the form

\[
\mu^{\text{add}}_{c, \beta_1, \ldots, \beta_p}(x) = c + \sum_{j=1}^{p} \sum_{k=1}^{s_n} \beta_{j,k} \varphi_k(x_j).
\]

A well-known choice of basis functions \( \{ \varphi_k \}_{k=1}^{s_n} \) are splines (Wahba, 1990), as for instance employed by Hastie and Tibshirani (1990) and more recently by Wood et al. (2016) and Wood (2017) with \text{mgcv}. These models suffer from two drawbacks, however: for each of the \( p \) directions, they must build and store an \( n \times s_n \) regression matrix of discretized splines for each \( \mu_j \), and they must select a regularization parameter \( \lambda_j \), for \( j = 1, \ldots, p \). To perform model selection with additive models, Meier et al. (2009) and Ravikumar et al. (2009) use sparsity inducing penalties, the former with two hyperparameters and the latter with one hyperparameter. Both still require storing large matrices. Consequently methods like \text{mgcv} and that of Meier et al. (2009) are computationally prohibitive in high dimensions, unless \( s_n \) in (5) is small, for example, the default value of \( s_n = 3 \) for \text{SAM} (Ravikumar et al., 2009). On the contrary our wavelet-based approach is able to use \( s_n = n \), so as to fit complex associations. Wavelets (Daubechies, 1992) are a rich class of basis functions \( \{ \varphi_k \}_{k=1}^{s_n} \) for \( s_n \) at least as large as \( n \); wavelets are orthonormal for \( s_n = 2^k \) for some \( k \in \mathbb{N} \); wavelets have been employed to fit additive models in low dimension \( p \) (Amato et al., 2022;
Haris et al., 2018; Sardy & Tseng, 2004). As we will see, using wavelets has some advantages: no wavelet matrices are stored hence allowing $s_n = n$ in (5), and a single regularization parameter $\lambda$ indexes the fitting. Consequently wavelet-based additive models can be employed in high dimensions both in $n$ and $p$. Finally the Kernel Knockoffs approach of Dai et al. (2022) controls the FDR while fitting additive models. More generally, the Model-X Knockoffs of Candes et al. (2018) controls the FDR for high-dimensional nonlinear regression, not only additive models.

The paper is organized as follows. In Section 2, we describe the model and our wavelet-based estimator, we show how to solve the corresponding optimization problem in Section 2.1, and we propose two selection rules for the threshold in Sections 2.2 and 2.3. In Section 3, we perform Monte Carlo simulations, then we use three real datasets to compare various methods. Proofs are postponed to the Appendix. The codes are available on https://github.com/StatisticsL/SRAMlet.

2 | SRAMLET

We consider wavelets to write each univariate function $\mu_j$ in (6) as a linear combination of orthonormal basis functions, for $j = 1, \ldots, p$. As a result, the corresponding regression matrix $\mathcal{W}^{(n)}$ is the concatenation of orthonormal matrices. Two key calculations involving this matrix can easily be performed without building and storing this matrix thanks to Mallat’s (1989) “pyramid” algorithm: the analysis operation $(\mathcal{W}^{(n)} \mathbf{r})$ and the synthesis operation $\mathcal{W}^{(n)} \mathbf{b}$ for any $\mathbf{r} \in \mathbb{R}^{pn}$ and $\mathbf{b} \in \mathbb{R}^n$. To define $\mathcal{W}^{(n)}$ for a training set $\mathbf{D}^{(n)} = (\mathbf{y}^{(n)}, \mathbf{X}^{(n)})$, let $P_j$ be the permutation matrices such that $P_j \mathbf{x}_j$ orders the $j$th column of $\mathbf{X}^{(n)}$ and, using isometric wavelets for unequally spaced samples (Kerkyacharian & Picard, 2004; Sardy et al., 1999), let $\Phi$ be an orthonormal wavelet matrix with one father wavelet (i.e., the constant function) and $(n-1)$ mother wavelets. Then we have that each $\mu_j := (\mu_j(x_1), \ldots, \mu_j(x_n)) = P_j^T \Phi \mathbf{b}_j$, for $j = 1, \ldots, p$. So $\mu_{c, \beta} := c \mathbf{1} + \mathbf{\mu}_1 + \ldots + \mathbf{\mu}_p = c \mathbf{1} + \mathcal{W}^{(n)} \mathbf{b}$ with

$$\mathcal{W}^{(n)} = [P_1^T \Phi \ldots P_p^T \Phi] \quad \text{and} \quad \mathbf{b} = (\beta_1, \ldots, \beta_p).$$

The resulting model has $np$ wavelet coefficients plus a constant $c$, while the output signal only has length $n$. Regularization is therefore needed. Owing to the a priori belief only a few of the $p$ variables have predictive power, and owing to the sparse wavelet representation of univariate functions $\mu_j$ (most mother wavelet coefficients are essentially zero to approximate a function), we regularize the least squares with a sparsity inducing penalty. Inspired by square-root LASSO (Belloni et al., 2011) and a selection of $\lambda$ that does not require estimation of the noise variance $\sigma^2$ in (1) (Giacobino et al., 2017), we define the square-root additive models with wavelets (SRAMlet) estimate as

$$(\hat{c}, \hat{\beta})_{SRAMlet} = \arg \min_{c \in \mathbb{R}, \mathbf{b} \in \mathbb{R}^{pn}} \sqrt{\text{PE}(\mu_{c, \beta}^{\text{add}}, \mathbf{D}^{(n)})} + \lambda P(\mathbf{b}),$$

for a positive penalty $\lambda$ and a sparsity inducing penalty $P(\mathbf{b})$. The corresponding estimation of the indexes $S$ of the relevant covariates defined in (3) is given by

$$S_{\text{add}}^{SRAMlet} = \left\{ j \in \{1, \ldots, p\} : \| \hat{\beta}_j^{SRAMlet} \| \neq 0 \right\}.$$

We discuss in the following section the choice of the penalty between $P(\mathbf{b}) = \sum_{j=1}^{p} \| \beta_j \|_2$ of group-LASSO (Yuan & Lin, 2006) and $P(\mathbf{b}) = \| \beta \|_1$ of LASSO. Solving the optimization
problem (9) and selecting \( \lambda \) for (9) is not trivial. First we propose an efficient algorithm to solve (9) in Section 2.1. Second we propose two selection rules for \( \lambda \) in Sections 2.2 and 2.3: one geared towards identification of the indices \( S \) of the relevant inputs, and one towards prediction of the output.

### 2.1 Optimization

Since needle selection is based on group sparsity on the vectors \( \beta_j \), it seems promising to use the group-LASSO penalty. But due to the fact that here \( \mathcal{Y}^{(n)} \) is the concatenation of orthonormal matrices, the following theorem proves that choosing the group-LASSO penalty does not work.

**Theorem 1.** Consider solving (9) for a fixed \( \lambda \geq 0 \), \( P(\beta) = \| \beta \|_2 \) and \( \sqrt{\text{PE} (\beta_0, \beta_n)} = \| y^{(n)} - X^{(n)} \beta \|_2 \). Then when the regression matrix \( X^{(n)} \) is orthonormal, the solution \( \hat{\lambda} \) is the least squares solution for any \( \lambda < 1 \), the null vector for any \( \lambda > 1 \), and any convex combination of the two if \( \lambda = 1 \).

This theorem shows that using the \( \ell_2 \)-norm for both the fit to the data and the penalty leads to a degenerate estimator (with an infinite number of solutions when \( \lambda = 1 \), the fully sparse vector for all \( \lambda > 1 \), and fitting the response exactly for \( \lambda < 1 \)). The theorem shows that SRAMlet is degenerate when \( p = 1 \), which remains true in higher dimension. See Bunea et al. (2014) for a study of the group square-root LASSO.

Consequently we use the LASSO penalty \( P(\beta) = \| \beta \|_1 \) for SRAMlet, in which case the estimator is not degenerate. First we consider square-root soft-waveshrink, that is, the univariate wavelet smoother defined as solution to (9) when \( p = 1 \). Square-root soft-waveshrink is the corner stone of SRAMlet.

**Definition 1.** **Square-root soft-waveshrink.** Given a response vector \( y^{(n)} \) corresponding to \( n \) ordered univariate inputs, and an \( n \times n \) orthonormal wavelet matrix \( \Phi = [\Phi_0, \Psi] \), with \( n_f \) father wavelets \( \Phi_0 \) and \( n_m = (n - n_f) \) mother wavelets \( \Psi \), and corresponding wavelet coefficients \( (\beta_0, \beta) \), then, for a given positive penalty \( \lambda \), the square-root soft-waveshrink estimator of \( (\hat{\beta}_0, \hat{\beta}) \) is defined as a solution to

\[
\min_{\beta_0, \beta} \| y^{(n)} - \Phi_0 \beta_0 - \Psi \beta \|_2^2 + \lambda \| \beta \|_1. \tag{10}
\]

The following theorem states that the solution to (10) has an implicit formulation.

**Theorem 2.** The solution to (10) is \( \hat{\beta}_0 = \Phi_0^T y^{(n)} \) and

\[
\hat{\beta} = \begin{cases} 
\eta^{\text{soft}}_{\varphi(\hat{\beta}; \lambda)}(z); & \frac{1}{\sqrt{\|z\|_0}} < \lambda < \frac{\|z\|_\infty}{\|z\|_2} \\
0; & \lambda \geq \frac{\|z\|_\infty}{\|z\|_2} 
\end{cases}
\]

where \( z = \Psi^T y^{(n)} \), \( \|z\|_0 = |\{ z_j \neq 0, j = 1, \ldots, n_m \}| \), the threshold function is \( \varphi(\cdot; \lambda) = \lambda \| z - \cdot \|_2 \), and the soft-thresholding function (Donoho & Johnstone, 1994) is applied
To determine the implicit threshold $\varphi(\hat{\beta}; \lambda)$, let

$$
\varphi_j = \lambda \sqrt{\frac{\sum_{i=1}^{n} |z_{(i)}|^2}{1 - \lambda^2 (n_m - j)}} \quad \text{for} \quad j \in \{1, \ldots, n_m - 1\};
$$

then $\varphi(\hat{\beta}; \lambda) = \varphi^*_{j^*}$ with $j^* = \max \{ i \in \{1, \ldots, n_m \} : |z_i| \leq \varphi^*_{j^*} \} = n_m - \|\hat{\beta}\|_0$.

To derive the Stein unbiased risk estimate (Stein, 1981) for square-root soft-waveshrink in Section 2.3, we need the following lemma.

**Lemma 1.** The function $\eta_{\varphi(\hat{\beta}; \lambda)}(z)$ of Theorem 2 is Lipschitz continuous with respect to $(z, \beta)$.

A first consequence of Theorem 2 is that the estimate of SD implicitly used by square-root soft-waveshrink is

$$
\hat{\sigma} = \frac{\varphi^*_{j^*}}{\sqrt{n/\lambda}}.
$$

(11)

So since $\hat{\sigma}^2 = \text{RSS}(\lambda)/n$, square-root LASSO (Belloni et al., 2011) and scaled LASSO (Sun & Zhang, 2012), which idea was first introduced by Antoniadis (2010), are equivalent.

A second consequence of Theorem 2 is that, owing to the fact that $\mathcal{W}^{(n)}$ in (8) is the concatenation of orthonormal blocks, the SRAMlet optimization problem (9) can be solved by iteratively employing the solution to (10), as stated in the following theorem.

**Theorem 3.** Block coordinate relaxation which consists in iteratively solving

$$
\min_{\beta \in \mathbb{R}^p} \|y^{(n)} - c1 - \mathcal{W}^{(n)} \beta\|_2 + \lambda \|\beta\|_1, \quad \text{where} \quad \beta = (\beta_1, \ldots, \beta_p),
$$

for $j = 1, \ldots, p$, then solving over $c \in \mathbb{R}$, and repeating until convergence, converges to a solution to (9).

### 2.2 Selection of $\lambda$ by QUT

In the spirit of the universal threshold of Donoho and Johnstone (1994) and Donoho et al. (1995), the first selection rule for $\lambda$ is the quantile universal threshold (QUT) (Giacobino et al., 2017). It is geared toward good identification of $S$, and is based on the property that the SRAMlet estimate (9) is the fully sparse zero-vector $\hat{\beta}_{\text{SRAMlet}} = 0$, given $\lambda$ is larger than a finite value that depends on the data. That specific value of $\lambda$ is given by the zero-thresholding function of Property 1.

**Property 1.** Given the matrix $\mathbf{X}^{(n)}$ and the output vector $y^{(n)}$ of the training set, the smallest $\lambda$ for which $\hat{\beta}_{\text{SRAMlet}}$ solving (9) is the zero-vector is given by the zero-thresholding function

$$
\lambda_0(y^{(n)}, \mathbf{X}^{(n)}) = \frac{\|\mathbf{X}^{(n)T} (y^{(n)} - \bar{y}^{(n)} 1)\|_{\infty}}{\| (y^{(n)} - \bar{y}^{(n)} 1)\|_2}.
$$

(12)

Under the assumption that all input entries carry no information, that is, $S = \emptyset$, the QUT selects $\lambda$ to be large enough to satisfy $\hat{\beta}_{\text{SRAMlet}} = 0$ with probability $1 - \alpha$, for a small $\alpha > 0$, hence
leading to \( \hat{S} = \emptyset \). So the selection of \( \lambda \) is on a probabilistic scale governed by \( \alpha \), in the spirit of hypothesis testing. This selection rule, like the universal threshold of Donoho et al. (1995) is at the detection edge between signal and noise. For a given small \( \alpha \), the QUT selection rule for \( \lambda \) is defined below.

**Definition 2.** Given training inputs \( \lambda^{(n)} \), let \( Y_0 \sim N(c \mathbf{1}, \sigma^2 I_n) \) be the distribution of \( Y \) according to (1) under the null model \( H_0 : \mu(x) = c \), and let \( F_\lambda \) be the c.d.f. of \( \Lambda = \lambda_0(Y_0, \lambda^{(n)}) \). For a small level \( \alpha \in [0, 1] \), the QUT is defined as \( \lambda_{\text{QUT}} = F^{-1}_\lambda(1 - \alpha) \).

Since the zero-thresholding function (12) has both numerator and denominator proportional to \( \sigma \) by multiplication of the response \( y^{(n)} \) by a scalar \( \sigma \), the statistic \( \Lambda \) is independent of \( \sigma \). Moreover subtraction by \( \bar{y}^{(n)} \) in both numerator and denominator makes the statistic \( \Lambda \) independent of \( c \). So the statistic \( \Lambda \) is pivotal. Consequently, estimation of the noise SD \( \sigma \) (which is a hard problem in high dimension \( p \)) is not required for SRAMlet for the selection of \( \lambda \) by QUT. On the contrary, QUT for AMlet (Sardy & Tseng, 2004) has the drawback of requiring an estimation of \( \sigma \) because AMlet’s zero-thresholding function is just the numerator of (12). An attempt for AMlet to circumvent this problem consists in estimating \( \sigma \) while iteratively solving the penalized least squares optimization problem. This approach may lead to slow or even no convergence, because the AMlet optimization, although convex for a fixed \( \lambda \), is no longer convex when the penalty (which depends on \( \hat{\sigma} \)) is regularly updated.

### 2.3 Selection of \( \lambda \) by SURE

To select \( \lambda \) for a good predictive performance measured by the mean squared error (MSE), also called \( \ell^2 \)-risk, one can minimize over \( \lambda \) an unbiased estimate of the risk (Stein, 1981).

**Theorem 4.** The Stein unbiased risk estimate for square-root soft-waveshrink is

\[
\text{SURE}(\lambda) = \text{RSS}(\lambda) + 2\sigma^2 n(\lambda) - n\sigma^2,
\]

where \( \text{RSS}(\lambda) = \|y^{(n)} - \Phi_0 \hat{\beta}_0 - \Psi \hat{\beta}\|_2^2 \) and \( n(\lambda) = n_t + |\{\hat{\beta}_j^\lambda \neq 0, j = 1, \ldots, n_m\}| \).

As for soft-waveshrink, the degree of freedom of its square-root version is the number of non-zero coefficients. To use SURE, the estimation of \( \sigma \) is needed; in dimension \( p = 1 \), we recommend the MAD estimate of Donoho and Johnstone (1995). In higher dimension, this approach does not work, which is one motivation for using the QUT proposed in the previous section.

### 3 MONTE CARLO SIMULATION

#### 3.1 Soft-waveshrink revisited

Since square-root soft-waveshrink is the corner stone of SRAMlet, we first consider the univariate case \( p = 1 \), and investigate the empirical properties of square-root soft-waveshrink in terms of MSE, true positive rate (TPR) and false discovery rate (FDR). We compare square-root soft-waveshrink (in black in the Figures) to the original soft-waveshrink (in red). It amounts to comparing square-root LASSO to LASSO in the orthonormal setting. To have an exactly sparse representation of a function with wavelets, we consider the blocks function (Donoho & Johnstone, 1994) together with the use of piecewise constant Haar wavelets. The number of
father wavelets is \( n_f = 2^3 \) and the number of mother wavelets (that is, the potential needles) is \( n_m = n - 8 \).

For each of \( m = 100 \) Monte Carlo run, data \( D_k = \{(x_i, y_i)\}_{i=1}^n \) for \( k = 1, \ldots, m \) are generated according to model (1) with \( \sigma = 1, \mu = \text{blocks} \) (with signal to noise ratio equal to three) sampled at \( n = 2^{10} \) random locations drawn from a uniform distribution on \([0, 1]\). In dimension one, the needles \( S_k \) are defined as the nonzero mother wavelet coefficients obtained by applying the analysis wavelet operator to \( \mu^k = \left( \mu \left( x_{(1)}^k \right), \ldots, \mu \left( x_{(n)}^k \right) \right) \), that is, the true function \( \mu \) evaluated at the ordered sampled locations \( x_{(1)}^k, \ldots, x_{(n)}^k \), for \( k = 1, \ldots, m \). Owing to the randomness of the sampled locations, the number of needles \( s_k = |S_k| \) varies; the median value is of 56 needles from a total of 1016 mother wavelets.

For the selection rule of \( \lambda \), we consider three rules: oracle (that is, the \( \lambda \) with the minimum \( \ell_2 \)-loss), SURE of Section 2.3, and QUT of Section 2.2. Two \( \lambda_{\text{QUT}} \) values are calculated for \( \alpha = 0.05 \), one for soft-waveshrink and one for its square-root version. Soft-waveshrink requires an estimate for \( \sigma \) that we take as the MAD estimate of Donoho and Johnstone (1995).

Figure 1 illustrates the differences between the three rules on a particular sample of the Monte Carlo simulation. The top-left plot shows the \( \ell_2 \)-loss (continuous) and SURE (dotted) as a function of \( \lambda / \lambda_{\text{QUT}} \). The dots show the \( \ell_2 \)-losses for \( \lambda = \lambda_{\text{QUT}} \). We observe that both SURE curves follow well their \( \ell_2 \)-loss, and that \( \lambda_{\text{QUT}} \) is conservatively large, leading to a larger loss than the minimum.

**FIGURE 1** Oracle \( l_2 \)-loss (continuous) and SURE (dotted) for one sample of “blocks” function sampled at \( n = 2^{10} \) points, for square-root (black) and original (red) soft-waveshrink.
FIGURE 2 The selection rule for $\lambda$ based on quantile universal threshold (QUT), SURE, and oracle. Boxplots of 100 Monte Carlo simulation runs for square-root (half left) and original (half right) soft-waveshrink. FDR (top left), TPR (top right), $\ell_2$-loss (bottom left), and estimated $\sigma$’s (bottom right) with true $\sigma = 1$ as the horizontal dotted line. Sample size $n = 2^{10}$.

of the curve. The other three plots show the corresponding estimation of $\mu$, which corroborates that, with $\lambda_{\text{QUT}}$, the estimation is less erratic than with SURE, which will translate to a lower rate of false jump detections.

Figure 2 summarizes the 100 Monte Carlo results with boxplots for FDR (top left), TPR (top right), $\ell_2$-loss (bottom left) and estimation of $\sigma$ (bottom right). The best FDR–TPR trade-off (4) is with QUT, and the best $\ell_2$-loss is with oracle and SURE, which are both comparable. With QUT, the original soft-waveshrink has smaller $\ell_2$-loss than its square-root version thanks to a smaller (and better) estimation of $\sigma$. Indeed, as we can see on the bottom right plot, for square-root soft-waveshrink, the implicit estimate given by (11) overestimates $\sigma$, while, for soft-waveshrink, the MAD estimate of Donoho and Johnstone (1995) is centered around the true value $\sigma = 1$.

3.2 Sparse high-dimensional additive models

We compare SRAMlet to the method of Ravikumar et al. (2009) called Sparse Additive Models (SAM), and, to the LASSO (from the glmnet library with the option “1se”) for a linear benchmark. We also compare our approach to the Kernel Knockoffs (KKO) (Dai et al., 2022) and to the Model-X Knockoffs (KNOCKOFF) (Candes et al., 2018) specifically derived to control the
FDR. Using the kko and knockoff packages, their default FDR control levels are 0.2 and 0.1, respectively, but we set it to 0.5 because their TPR were too low with the default values. Also for KKO, the number of cross-validation folds is \( cv\_folders = 10 \), but we set it to 2 in our high-dimensional settings to prevent lengthy computations. Because mgcv builds spline matrices and selects many regularization parameters by minimizing GCV, it requires lots of memory and is computationally expensive. So for mgcv, with use the default value for \( s_n \) when \( p = 10 \), but set it otherwise to \( s_n = 3 \) for \( p = 100 \). Because Meier et al. (2009) search their hyperparameters on two grids that are not automatically calibrated to the data at hand, we were not able to run their method. Likewise Haris et al. (2018) perform an expensive cross-validation search and do not allow model selection. When \( n \) and \( p \) are getting larger, only SRAMlet and AMlet can be employed since they do not require building and storing a large matrix in each direction.

For the simulation, we choose to have \( s = 4 \) out of \( p \) inputs with predictive information. Their corresponding functions \( \mu_1, \ldots, \mu_4 \) are \textit{blocks}, \textit{bumps}, \textit{heavisine}, and \textit{Doppler} with \( \text{snr} = 3 \) (Donoho & Johnstone, 1994). The \( p \) inputs are drawn from independent uniformly distributed random variable between zero and one. For wavelet-based methods, we use the Daubechies “extremal phase” wavelets with a filter number equal to four and one father wavelet (the constant function that is not penalized).

For the first simulation, the number of samples is fixed to \( n = 2^{10} \) and the number of predictors \( p \in \{10, 100, 1000\} \). Table 1 reports the results in terms of MSE, FDR, and TPR. In terms of MSE, we observe that all additive models (with a reported MSE) perform better than the linear model (last column), as expected since the true association is nonlinear. Second, only SRAMlet has a good low FDR–high TPR trade-off, in the sense that its FDR is significantly lower than that of the other methods, while its TPR is comparable.

For the second simulation, both \( n \) and \( p \) increase with \( (n, p) \in \{(2^j, 2^{j+1}), j \in \{8, 9, 10, 11\}\} \). Because dimension are high (e.g., with \( j = 11 \), the input matrix has \( 8 \, 388 \, 608 \) entries), we can only apply the wavelet-based methods and the KNOCKOFF for \( j \in \{8, 9\}\). Figure 3 summarizes the results. Again, we see that SRAMlet offers a good FDR–TPR trade-off.

| TABLE 1 | Empirical comparison of seven methods based on the estimation of MSE, FDR and TPR by Monte-Carlo simulation. |
|---------|--------------------------------------------------|
| Additive model | Linear model |
| \( p \) | SRAMlet | AMlet \( \delta \) | SAM | MGCV | KKO | KNOCKOFF | LASSO \( _{1se} \) |
| MSE | 10 | 23.8 (0.2) | 22.0 (0.2) | 30.3 (0.2) | 27.5 (0.2) | / | / | 34.4 (0.2) |
| | 100 | 25.7 (0.2) | 22.6 (0.2) | 30.4 (0.2) | 48.3 (3.02) | / | / | 34.3 (0.2) |
| | 1000 | 28.0 (0.2) | 24.0 (0.2) | 31.6 (0.2) | / | / | / | 35.2 (0.2) |
| FDR | 10 | 0.07 (0.01) | 0.17 (0.01) | 0.48 (0.02) | 0.60 (0) | 0.04 (0.12) | 0.18 (0.21) | 0.34 (0.02) |
| | 100 | 0.12 (0.02) | 0.51 (0.02) | 0.79 (0.01) | 0.96 (0) | 0.01 (0.05) | 0.37 (0.28) | 0.66 (0.02) |
| | 1000 | 0.14 (0.02) | 0.73 (0.01) | 0.85 (0.01) | / | / | 0.14 (0.29) | 0.70 (0.03) |
| TPR | 10 | 1 (0) | 1 (0) | 0.99 (0.003) | 1 (0) | 0.94 (0.11) | 0.69 (0.22) | 0.82 (0.02) |
| | 100 | 0.99 (0.005) | 1 (0) | 0.98 (0.007) | 1 (0) | 0.79 (0.15) | 0.55 (0.20) | 0.67 (0.02) |
| | 1000 | 0.94 (0.01) | 1 (0) | 0.89 (0.02) | / | / | 0.09 (0.16) | 0.55 (0.02) |

Note: The size of the haystack is \( p \in \{10, 100, 1000\} \), the number of needles is \( s = 4 \) (blocks, bumps, heaviside, Doppler) and the sample size is fixed to \( n = 2^{10} \).
Focusing on the estimation of $\sigma$, the bottom left plots of Figures 2 and 3 show that the implicit estimation of $\sigma$ used by the square-root versions of soft-waveshrink and AMlet tends to be biased upwards, which can explain the good FDR control of the method.

4 | APPLICATION TO REAL DATA

We consider three datasets of sizes $(n, p) = (215, 100), (166, 235), (166, 235)$. The first set aims at predicting fat content from meat spectroscopy, the second and the third sets at predicting the concentration of ethanol and glucose from near infrared spectroscopy, respectively. We train the models on $128 = 2^7$ observations (the power of two is chosen by convenience for wavelet-based estimators) for all three datasets, and test on the remaining data. We randomly sample from the original data 20 times and compute an average predictive error and average model size. We believe the underlying association is quite smooth so we use the Daubechies “extremal phase” wavelets with a filter number equal to four with periodic boundary correction. But as opposed to the Monte Carlo simulations, the underlying associations are not necessarily periodic. To account for that, we use a model that concatenates three additive terms (smooth wavelets, a linear trend and Haar wavelets) that do not suffer from the nonperiodic issues. The parameters of each term are treated blockwise by the block coordinate algorithm of Theorem 3. The role of the linear and Haar wavelets is to take care of a possible nonperiodicity of the univariate functions $\mu_j, \ j = 1, \ldots, p$. Table 2 summarizes the results. We observe that SRAMlet selects the least inputs, but only with
FIGURE 4  Ethanol data: SRAMlet selects five variables, with variable 14 pointing to a strong nonlinear trend.

TABLE 2  Results of accuracy versus mode size for seven estimators based on three real datasets.

| Methods         | Meatspec $n = 215, p = 100$ | Ethanol $n = 166, p = 235$ | Glucose $n = 166, p = 235$ |
|-----------------|-----------------------------|-----------------------------|-----------------------------|
| Model size      |                             |                             |                             |
| SRAMlet         | 3 (0.1)                     | 5 (0.2)                     | 11 (0.6)                    |
| AMlet$_g$       | 12 (2)                      | 12 (0.4)                    | 16 (0.4)                    |
| SAM             | 53 (3)                      | 41 (2)                      | 58 (7)                      |
| KKO             | 7.8 (5.2)                   | 28.5 (11.6)                 | 4.8 (7.4)                   |
| KNOCKOFF        | 24.4 (11.2)                 | 15.1 (11.8)                 | 19.2 (10.1)                 |
| LASSO$_{1se}$   | 13 (1)                      | 10 (0.5)                    | 31 (0.7)                    |
| $\hat{y}_{\text{training}}$ | 0                           | 0                           | 0                           |
| Predictive error|                             |                             |                             |
| SRAMlet         | 133 (2)                     | 2.7 (0.2)                   | 119 (5)                     |
| AMlet$_g$       | 143 (3)                     | 26 (2)                      | 68 (5)                      |
| SAM             | 37 (4)                      | 15 (4)                      | 46 (7)                      |
| LASSO$_{1se}$   | 14 (0.4)                    | 3.2 (0.1)                   | 35 (2)                      |
| $\hat{y}_{\text{training}}$ | 169 (4)                    | 500 (11)                    | 196 (7)                     |
the ethanol data (second column) does it achieve the best predictive error. A closer look at the Ethanol data reveals a strong nonlinear trend with respect to variable 14 (see Figure 4). Because here $n$ is fairly small and $p$ is larger than $n$, the LASSO linear model tends to perform quite well in terms of predictive error, though with a fairly high number of selected inputs.

5 CONCLUSION

The wavelet-based SRAMlet method allows to fit sparse additive models in high dimensions, when both sample size $n$ and number of covariate $p$ are high. With the selection of its regularization parameter $\lambda$ by the QUT method, SRAMlet does not require estimation of the variance $\sigma^2$, which is difficult in high dimension. SRAMlet offers a good complexity-predictive performance trade-off, as we observed on Monte-Carlo simulations based on the false discovery rate and true positive rate results. SRAMlet can easily be robustified by Huberizing the least squares loss, which amounts to concatenating an additional identity matrix to the wavelet regression matrix (Sardy et al., 2001).

ACKNOWLEDGMENTS

We thank two anonymous referees and the Associate Editor for helpful comments that improved the quality of the article.

ORCID
Sylvain Sardy © https://orcid.org/0000-0003-4935-2595
Xiaoyu Ma © https://orcid.org/0000-0002-1674-3074

REFERENCES

Amato, U., Antoniadis, A., De Feis, I., & Gibels, I. (2022). Wavelet-based robust estimation and variable selection in nonparametric additive models. Statistics and Computing, 32, 11.
Antoniadis, A. (2010). Comments on: $l_1$-penalization for mixture regression models. TEST, 19, 257–258.
Bach, F. R., Jenatton, R., Mairal, J., & Obozinski, G. (2011). Optimization with sparsity-inducing penalties. CoRR. abs/1108.0775.
Barber, R. F., & Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. The Annals of Statistics, 43(5), 2055–2085.
Belloni, A., Chernozhukov, V., & Wang, L. (2011). Square-root lasso: Pivotal recovery of sparse signals via conic programming. Biometrika, 98(4), 791–806.
Buehlmann, P., & van de Geer, S. (2011). Statistics for high-dimensional data Methods theory and applications. Springer.
Bunea, F., Lederer, J., & She, Y. (2014). The group square-root lasso: Theoretical properties and fast algorithms. IEEE Transactions on Information Theory, 60(2), 1313–1325.
Candes, E., Fan, Y., Janson, L., & Lv, J. (2018). Panning for gold: Model-x knockoffs for high dimensional controlled variable selection. Journal of the Royal Statistical Society: Series B, 80(3), 551–577.
Dai, X., Lyu, X., & Li, L. (2022). Kernel knockoffs selection for nonparametric additive models. Journal of the American Statistical Association, 1–13. doi: 10.1080/01621459.2022.2039671
Daubechies, I. (1992). Ten lectures on wavelets. Cambridge University Press.
Donoho, D. L., & Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. Biometrika, 81(3), 425–455.
Donoho, D. L., & Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. Journal of the American Statistical Association, 90, 1200–1224.
Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., & Picard, D. (1995). Wavelet shrinkage: Asymptopia? Journal of the Royal Statistical Society: Series B, 57(2), 301–369.
Giacobino, C., Sardy, S., Diaz Rodriguez, J., & Hengartner, N. (2017). Quantile universal threshold for model selection. Electronic Journal of Statistics, 11(2), 4701–4722.
APPENDIX A. PROOF OF THEOREM 1

Consider the cost function $L(\alpha) = L_1(\alpha) + L_2(\alpha)$ with $L_1(\alpha) = \|y - X\alpha\|_2$ and $L_2(\alpha) = \lambda \|\alpha\|_2$, where $\lambda > 0$, $X$ is orthonormal. This function $L$ is convex and, for $\alpha \neq 0$ and $y - X\alpha \neq 0$, the first-order optimality conditions are

$$
\nabla_\alpha L(\alpha) = \frac{-X^T(y - X\alpha)}{\|y - X\alpha\|_2} + \frac{\lambda \alpha}{\|\alpha\|_2} = 0
$$

$$
\iff \alpha = k(\alpha)X^Ty \quad \text{and} \quad k(\alpha) = \frac{\|\alpha\|_2}{\lambda \|y - X\alpha\|_2 + \|\alpha\|_2}
$$

$$
\iff \alpha = k(\alpha)X^Ty \quad \text{and} \quad k(\alpha) = \frac{k(\alpha)\|y\|_2}{\lambda \|y - k(\alpha)y\|_2 + k(\alpha)\|y\|_2}
$$
\( \Rightarrow \alpha = k(\alpha)X^T y, (\lambda - 1)(1 - k(\alpha)) = 0, 0 < k(\alpha) < 1 \)
\( \Rightarrow \alpha = kX^T y, \lambda = 1, 0 < k < 1. \)

So a solution exists at \( \alpha = kX^T y \) for any \( k \in (0, 1) \) if and only if \( \lambda = 1 \), in which case the cost is \( L(kX^T y) = \|y\|_2 \). Moreover \( \alpha = 0 \) is a solution if and only if \( 0 \in \{ \nabla a L_1(0) + \lambda \{ g : \|g\|_2 \leq 1 \} \} \), where \( \{ g : \|g\|_2 \leq 1 \} \) is the subgradient of \( \|y\|_2 \) at \( \alpha = 0 \). But \( \| \nabla a L_1(0) \|_2 = \| X^T y \|_2 = 1 \) since \( X \) is orthonormal. So \( \alpha = 0 \) is a solution if and only if \( \lambda \geq 1 \), in which case the cost is \( L(0) = \|y\|_2 \). Finally \( \alpha = X^T y \) is a solution if and only if \( 0 \in \{ g : \|g\|_2 \leq 1 \} \), where \( \{ g : \|g\|_2 \leq 1 \} \) is the subgradient of \( \|y - X\alpha\|_2 \) at \( \alpha = X^T y \). So \( \alpha = X^T y \) is a solution if and only if \( \lambda \leq 1 \), in which case the cost is \( L(X^T y) = \lambda \|y\|_2 \). In summary, the solution is

\[
\alpha = \begin{cases} 
X^T y & \text{if } \lambda < 1, \\
kX^T y, \ k \in (0, 1) & \text{if } \lambda = 1, \\
0 & \text{if } \lambda > 1.
\end{cases}
\]

**APPENDIX B. PROOF OF THEOREM 2**

The optimization (10) is equivalent to

\[
\min_{\beta_0, \beta} \left\| \left( \Phi_T y^{(n)} \right) - \left( \beta_0 \right) \right\|_2 + \lambda \| \beta \|_1,
\]

(B1)

thanks to the orthonormality of the wavelet matrix \( \Phi \). Clearly the minimum with respect to \( \beta_0 \) is at \( \hat{\beta}_0 = \Phi_T^T y^{(n)} \). Therefore we discuss the solution in \( \beta \) for \( \beta_0 = \Phi_T^T y \).

The cost function (B1) begin convex, we rely on Proposition 1.1 of Bach et al. (2011) to derive the solution in \( \beta \). First, letting \( z = \Psi_T y^{(n)} \), \( \beta = 0 \) is the solution if the condition \( \{0\} \in \left\{ \frac{z}{\|z\|_2} + \lambda[-1, 1] \cdot 1 \right\} \) is satisfied, which means \( -1 \leq \frac{z}{\|z\|_2} \leq 1 \). So only when \( \lambda \geq \frac{\max |z|}{\|z\|_2} = \frac{\|z\|_\infty}{\|z\|_2} \), can the condition \( \{0\} \in \left\{ \frac{z}{\|z\|_2} + \lambda[-1, 1] \cdot 1 \right\} \) hold. This is also consistent with the zero thresholding function (12). Second, \( z \) is the solution if it satisfy \( \{0\} \in \{ \{ g \|g\|_2 \leq 1 \} + \lambda r \} \), where \( r = [r_1, r_2, \ldots, r_m]^T \) and \( r_i = \begin{cases} 
1, & z_i > 0 \\
[-1, 1], & z_i = 0 \\
-1, & z_i < 0
\end{cases} \). This condition means that \( \lambda \leq \frac{1}{\|r\|_2} \). With Gaussian data, we have that \( z_i \neq 0 \) with probability one for \( i = 1, \ldots, n_m \), so \( \|r\|_2 = \sqrt{n_m} \). But more generally, letting \( \|z\|_0 \) be the number of nonzero in \( z \), then \( \|r\|_2 \geq \|z\|_0^{1/2} \). So only when \( \lambda \leq \frac{1}{\|z\|_0^{1/2}} \), can the condition \( \{0\} \in \{ \{ g \|g\|_2 \leq 1 \} + \lambda r \} \) hold. In other words, \( \lambda \|z\|_0^{1/2} \) is the smallest shift of the ball \( \{ g \|g\|_2 \leq 1 \} \) that guarantees the condition to hold. If the shift is smaller or equal one, then the point \( 0 \) still belongs to the ball. Finally, we consider when \( \lambda \in \left( \frac{1}{\|z\|_0^{1/2}}, \frac{\|z\|_\infty}{\|z\|_2} \right) \). In that case, the KKT conditions are

\[
\{0\} \in \frac{\beta_i - z_i}{\|z - \beta\|_2} + \lambda \begin{cases} 
1, & \beta_i > 0 \\
[-1, 1], & \beta_i = 0 \\
-1, & \beta_i < 0
\end{cases} \quad i = 1, \ldots, n_m,
\]
which is equivalent to

\[ \beta = \eta^\text{soft}_{\varphi; \lambda}(z) \quad \text{for} \quad \varphi(\beta; \lambda) = \lambda \| z - \beta \|_2, \]

(B2)

where \( \eta^\text{soft}_{\varphi} \) is the soft thresholding function with threshold \( \varphi \) (Donoho & Johnstone, 1994). This is an implicit definition of the solution since \( \beta \) is in both the left and right hand sides of (B2). The number of zero entries in \( \beta \) is

\[ n_0(\beta) = |\{ i \in \{1, \ldots, n_m \} : \beta_i = 0 \}| = |\{ i \in \{1, \ldots, n_m \} : |z_i| \leq \varphi(\beta; \lambda) \}|, \]

which satisfies \( 0 < n_0(\beta) < n_m \) for the range of \( \lambda \) considered. The threshold is also implicitly defined, but it is easy to see that

\[ \varphi^2(\beta; \lambda) = \lambda^2 \left\{ \varphi^2(\beta; \lambda)(n_m - n_0(\beta)) + \sum_{i=1}^{n_0(\beta)} (|z_i|)^2 \right\}, \]

where \( |z_i| \) is the \( i \)th element for ordered \( |z| \). So among all

\[ \varphi_j = \lambda \sqrt{\frac{\sum_{i=1}^{j-1} (|z_i|)^2}{1 - \lambda^2(n_m - j)}} \quad \text{for} \quad j \in \{1, \ldots, n_m - 1\}, \]

only \( \varphi_j \) satisfying \( j^* = n_0(\hat{\beta}) = |\{ i \in \{1, \ldots, n_m \} : |z_i| \leq \varphi_{j^*} \}| = n_m - \| \hat{\beta} \|_0 \) leads to the solution \( \hat{\beta} = \eta^\text{soft}_{\varphi_{j^*}}(z) \). In practice, one tries all \( j \in \{1, \ldots, n_m - 1\} \) until \( j = n_0(\eta^\text{soft}_{\varphi_j}(z)) \).

### APPENDIX C. PROOF OF LEMMA 1

#### C.1 Lipschitz continuous with respect to \( \beta \)

Given a fixed \( z \in \mathbb{R}^{n_m} \) and \( \lambda > 0 \), \( \eta^\text{soft}_{\lambda \| \beta - z \|_2}(z) \) is Lipschitz continuous with respect to \( \beta \) if there exists a constant \( K_1 \) such that

\[ \left\| \eta^\text{soft}_{\lambda \| \beta_1 - z \|_2}(z) - \eta^\text{soft}_{\lambda \| \beta_2 - z \|_2}(z) \right\|_2 \leq K_1 \| \beta_1 - \beta_2 \|_2, \]

for all vectors \( \beta_1 \) and \( \beta_2 \). Without loss of generality, we assume that \( \lambda \| \beta_1 - z \|_2 \leq \lambda \| \beta_2 - z \|_2 \). For every element \( z_i, i = 1, \ldots, n_m \), we have the following three cases:

1. if \( z_i < -\lambda \| \beta_2 - z \|_2 \), then

\[ \left| \eta^\text{soft}_{\lambda \| \beta_1 - z \|_2}(z_i) - \eta^\text{soft}_{\lambda \| \beta_2 - z \|_2}(z_i) \right| = |z_i + \lambda \| \beta_1 - z \|_2 - (z_i + \lambda \| \beta_2 - z \|_2)| \leq \lambda \| \beta_1 - \beta_2 \|_2 \quad \text{(Reverse triangle inequality)}. \]

Likewise when \( z_i > \lambda \| \beta_2 - z \|_2 \) by symmetry.

2. if \( -\lambda \| \beta_2 - z \|_2 \leq z_i < -\lambda \| \beta_1 - z \|_2 \), then

\[ \left| \eta^\text{soft}_{\lambda \| \beta_1 - z \|_2}(z_i) - \eta^\text{soft}_{\lambda \| \beta_2 - z \|_2}(z_i) \right| = |z_i - (-\lambda \| \beta_1 - z \|_2)| \quad \text{(Distance)} \]

\[ \leq |-\lambda \| \beta_2 - z \|_2 - (-\lambda \| \beta_1 - z \|_2)| \]

\[ = \lambda \| \beta_1 - z \|_2 - \| \beta_2 - z \|_2 \leq \lambda \| \beta_1 - \beta_2 \|_2. \]

Likewise when \( \lambda \| \beta_1 - z \|_2 < z_i \leq \lambda \| \beta_2 - z \|_2 \) by symmetry.
3. if $-\lambda \|\beta_1 - z\|_2 \leq z_i \leq \lambda \|\beta_1 - z\|_2$, then
\[
\|\eta_{\lambda\|\beta_1 - z\|_2}^\text{soft}(z_i) - \eta_{\lambda\|\beta_2 - z\|_2}^\text{soft}(z_i)\|_2 = |0 - 0| = 0 \leq \lambda \|\beta_1 - \beta_2\|_2.
\]

Putting the three cases together, we have that
\[
\|\eta_{\lambda\|\beta_1 - z\|_2}^\text{soft}(z) - \eta_{\lambda\|\beta_2 - z\|_2}^\text{soft}(z)\|_2 = \sum_{i=1}^{n_m} \left(\|\eta_{\lambda\|\beta_1 - z\|_2}^\text{soft}(z_i) - \eta_{\lambda\|\beta_2 - z\|_2}^\text{soft}(z_i)\|_2^2\right)^{\frac{1}{2}} \leq \sum_{i=1}^{n_m} \lambda^2 \|\beta_1 - \beta_2\|_2^2 = n_m \lambda^2 \|\beta_1 - \beta_2\|_2^2.
\]

So we get $\|\eta_{\lambda\|\beta_1 - z\|_2}^\text{soft}(z) - \eta_{\lambda\|\beta_2 - z\|_2}^\text{soft}(z)\|_2 \leq K_1 \|\beta_1 - \beta_2\|_2$ for $K_1 = \lambda \sqrt{n_m}$.

C.2 Lipschitz continuous with respect to $z$

Given a fixed $\beta \in \mathbb{R}^{n_m}$ and $\lambda > 0$, $\eta_{\lambda\|\beta - z\|_2}^\text{soft}(z)$ is Lipschitz continuous with respect to $z$ if there exists a constant $K_2$ such that
\[
\|\eta_{\lambda\|\beta - z_1\|_2}^\text{soft}(z_1) - \eta_{\lambda\|\beta - z_2\|_2}^\text{soft}(z_2)\|_2 \leq K_2 \|z_1 - z_2\|_2,
\]
for all vectors $z_1$ and $z_2$. We consider two cases:

1. if $\lambda \|\beta - z\|_2 = \lambda \|\beta - z\|_2$. Consider now a particular $i \in \{1, \ldots, n_m\}$ and, without loss of generality, suppose $z_{i1} < z_{i2}$. There care four subcases:
   (a) if $z_{i1} < z_{i2} < -\lambda \|\beta - z\|_2$, then
   \[
   \|\eta_{\lambda\|\beta - z_1\|_2}^\text{soft}(z_{i1}) - \eta_{\lambda\|\beta - z_2\|_2}^\text{soft}(z_{i2})\|_2 = |z_{i1} + \lambda \|\beta - z\|_2 - (z_{i2} + \lambda \|\beta - z\|_2)|
   \leq |z_{i1} - z_{i2}| \leq |z_{i1} - z_{i2}|.
   \]
   Likewise when $\|\beta - z_1\|_2 < z_{i1} < z_{i2}$ by symmetry.

   (b) if $z_{i1} < -\lambda \|\beta - z\|_2 \leq z_{i2} \leq \lambda \|\beta - z\|_2$, then
   \[
   \|\eta_{\lambda\|\beta - z_1\|_2}^\text{soft}(z_{i1}) - \eta_{\lambda\|\beta - z_2\|_2}^\text{soft}(z_{i2})\|_2 = |z_{i1} - (-\lambda \|\beta - z\|_2)|
   \leq |z_{i1} - z_{i2}| \text{ (Distance)}.
   \]
   Likewise when $-\lambda \|\beta - z\|_2 \leq z_{i1} \leq \lambda \|\beta - z\|_2$ by symmetry.

   (c) if $z_{i1} < -\lambda \|\beta - z\|_2 < z_{i1} < \lambda \|\beta - z\|_2$, then
   \[
   \|\eta_{\lambda\|\beta - z_1\|_2}^\text{soft}(z_{i1}) - \eta_{\lambda\|\beta - z_2\|_2}^\text{soft}(z_{i2})\|_2 = |z_{i1} - (-\lambda \|\beta - z\|_2)|
   \leq |z_{i1} - z_{i2}| \text{ (Distance)}.
   \]
   Likewise when $-\lambda \|\beta - z\|_2 \leq z_{i1} \leq \lambda \|\beta - z\|_2$ by symmetry.

   (d) if $-\lambda \|\beta - z\|_2 \leq z_{i1} < z_{i2} \leq \lambda \|\beta - z\|_2$, then
   \[
   \|\eta_{\lambda\|\beta - z_1\|_2}^\text{soft}(z_{i1}) - \eta_{\lambda\|\beta - z_2\|_2}^\text{soft}(z_{i2})\|_2 = |0 - 0| \leq |z_{i1} - z_{i2}|.
   \]
   So $\|\eta_{\lambda\|\beta - z_1\|_2}^\text{soft}(z_{i1}) - \eta_{\lambda\|\beta - z_2\|_2}^\text{soft}(z_{i2})\|_2 \leq 2|z_{i1} - z_{i2}|$. 


2. if \( \lambda \| \beta - z_1 \|_2 \neq \lambda \| \beta - z_2 \|_2 \). Following similar steps, we get
\[
\left| \eta_{\lambda \| \beta - z_1 \|_2} \left( z_1 \right) - \eta_{\lambda \| \beta - z_2 \|_2} \left( z_2 \right) \right| \leq \max \left\{ 2 |z_{1i}| - |z_{2i}|, |z_{1i}| - |z_{2i}| + \lambda \| z_1 - z_2 \|_2 \right\}.
\]
Putting all results together, we get that
\[
\left\| \eta_{\lambda \| \beta - z_1 \|_2} \left( z_1 \right) - \eta_{\lambda \| \beta - z_2 \|_2} \left( z_2 \right) \right\|_2^2 = \sum_{i=1}^{n_m} \left( \eta_{\lambda \| \beta - z_1 \|_2} \left( z_{1i} \right) - \eta_{\lambda \| \beta - z_2 \|_2} \left( z_{2i} \right) \right)^2 \leq \sum_{i=1}^{n_m} \left( \max \left\{ 2 |z_{1i}| - |z_{2i}|, |z_{1i}| - |z_{2i}| + \lambda \| z_1 - z_2 \|_2 \right\} \right)^2 \leq \sum_{i=1}^{n_m} (2 + \lambda)^2 \| z_1 - z_2 \|_2^2 = (2 + \lambda)^2 n_m \| z_1 - z_2 \|_2^2.
\]
So \( \left\| \eta_{\lambda \| \beta - z_1 \|_2} \left( z_1 \right) - \eta_{\lambda \| \beta - z_2 \|_2} \left( z_2 \right) \right\|_2 \leq K_2 \| z_1 - z_2 \|_2 \) with \( K_2 = (2 + \lambda) \sqrt{n_m} \).

C.3 Lipschitz continuous
Given any vectors \( (z_1^T, \beta_1^T)^T \) and \( (z_2^T, \beta_2^T)^T \), we have
\[
\left\| \eta_{\lambda \| \beta - z_1 \|_2} \left( z_1 \right) - \eta_{\lambda \| \beta - z_2 \|_2} \left( z_2 \right) \right\|_2 = \left\| \eta_{\lambda \| \beta - z_1 \|_2} \left( z_1 \right) - \eta_{\lambda \| \beta - z_1 \|_2} \left( z_1 \right) + \eta_{\lambda \| \beta - z_1 \|_2} \left( z_1 \right) - \eta_{\lambda \| \beta - z_2 \|_2} \left( z_2 \right) \right\|_2 \leq K_1 \| \beta_1 - \beta_2 \|_2 + K_2 \| z_1 - z_2 \|_2 \leq (K_1 + K_2) \left\| \left( \begin{array}{c} z_1 \\ \beta_1 \end{array} \right) - \left( \begin{array}{c} z_2 \\ \beta_2 \end{array} \right) \right\|_2.
\]
So the function \( \eta_{\lambda \| \beta - z_\|_2} \left( z \right) \) is Lipschitz continuous with respect to \( (z, \beta) \).

APPENDIX D. PROOF OF THEOREM 3
The SRAMlet optimization problem (9) writes as
\[
\min_{c, \beta} g(c, \beta) + \lambda \sum_{j=1}^{p} \sum_{i=1}^{n} h(\beta_{ji}),
\]
where \( g(c, \beta) = \| y^{(n)} - c_1 - W^{(n)} \beta \|_2 \) is a differentiable convex function on \( \mathbb{R}^{1+p_n} \setminus \{(c, \beta) : y^{(n)} = c_1 + W^{(n)} \beta \} \) and \( h(u) = |u| \) is a continuous function on \( \mathbb{R} \). Under this separable structure of the nondifferentiable part of the cost function, Tseng (1993) showed that the BCR algorithm using the systematic rule converges.

APPENDIX E. PROOF OF THEOREM 4
Denote \( r = \Phi^T y^{(n)}, \theta = \left( \rho_0^T, \beta^T \right)^T, \hat{\theta} = \left( \hat{\rho}_0^T, \hat{\beta}^T \right)^T \). First, we compute the gradient matrix \( \frac{\partial \hat{\theta}}{\partial r} \). When \( \lambda \leq \frac{1}{\sqrt{\| z_0 \|_0}} \), we get \( \frac{\partial \hat{\theta}}{\partial r} = I_{n \times n} \). When \( \lambda \geq \frac{\| z_0 \|_0}{\| z \|_2} \), we get \( \frac{\partial \hat{\theta}}{\partial r} = \left( I_{n \times n} \times n \right) \left( \begin{array}{c} 0_{n \times n} \\ 0_{n \times n} \end{array} \right) \). For the third case
\[
\frac{1}{\sqrt{\|z\|}} < \lambda < \frac{\|z\|}{\|z\|_2},
\]
the solution \( \beta \) in (B2) has an implicit function form, so we us the implicit function theorem to compute the gradient. To see this, consider the function \( f(z, \beta) = \beta - \eta^{\text{soft}}_{\psi(\beta, \lambda)}(z) \) that goes from \( \mathbb{R}^{2n_m} \) to \( \mathbb{R}^{n_m} \). This function is Lipschitz continuous from Lemma 1. According to the Rademacher’s theorem, the following Jacobian matrices exist almost everywhere. For \( i = 1, \ldots, n_m \), we have
\[
\frac{\partial \eta^{\text{soft}}_{\psi(\beta, \lambda)}(z)}{\partial z_i} = \begin{cases} 
\lambda (z_i - \beta_i); & \text{if } -z_i > \lambda \|z - \beta\|_2 \\
0; & \text{if } |z_i| < \lambda \|z - \beta\|_2 \\
1 - \lambda (z_i - \beta_i); & \text{if } z_i > \lambda \|z - \beta\|_2 
\end{cases}
\]
and for \( j \neq i \),
\[
\frac{\partial \eta^{\text{soft}}_{\psi(\beta, \lambda)}(z)}{\partial z_j} = \begin{cases} 
\lambda (z_j - \beta_j); & \text{if } -z_i > \lambda \|z - \beta\|_2 \\
0; & \text{if } |z_j| < \lambda \|z - \beta\|_2 \\
1 - \lambda (z_j - \beta_j); & \text{if } z_j > \lambda \|z - \beta\|_2 
\end{cases}
\]
And for \( i, j = 1, \ldots, n_m \), we have
\[
\frac{\partial \eta^{\text{soft}}_{\psi(\beta, \lambda)}(z)}{\partial \beta_j} = \begin{cases} 
\lambda (z_i - \beta_i); & \text{if } -z_i > \lambda \|z - \beta\|_2 \\
0; & \text{if } |z_j| < \lambda \|z - \beta\|_2 \\
1 - \lambda (z_j - \beta_j); & \text{if } z_j > \lambda \|z - \beta\|_2 
\end{cases}
\]
Denote
\[
J_{\eta, x} = \begin{pmatrix}
\frac{\partial \eta^{\text{soft}}_{\psi(\beta, \lambda)}(z)}{\partial z_1} & \frac{\partial \eta^{\text{soft}}_{\psi(\beta, \lambda)}(z)}{\partial z_2} & \cdots & \frac{\partial \eta^{\text{soft}}_{\psi(\beta, \lambda)}(z)}{\partial z_{n_m}} \\
\frac{\partial \eta^{\text{soft}}_{\psi(\beta, \lambda)}(z)}{\partial z_1} & \frac{\partial \eta^{\text{soft}}_{\psi(\beta, \lambda)}(z)}{\partial z_2} & \cdots & \frac{\partial \eta^{\text{soft}}_{\psi(\beta, \lambda)}(z)}{\partial z_{n_m}} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial \eta^{\text{soft}}_{\psi(\beta, \lambda)}(z)}{\partial z_1} & \frac{\partial \eta^{\text{soft}}_{\psi(\beta, \lambda)}(z)}{\partial z_2} & \cdots & \frac{\partial \eta^{\text{soft}}_{\psi(\beta, \lambda)}(z)}{\partial z_{n_m}}
\end{pmatrix}
\]
\[
J_{\eta, \beta} = \begin{pmatrix}
\frac{\partial \eta^{\text{soft}}_{\psi(\beta, \lambda)}(z)}{\partial \beta_1} & \frac{\partial \eta^{\text{soft}}_{\psi(\beta, \lambda)}(z)}{\partial \beta_2} & \cdots & \frac{\partial \eta^{\text{soft}}_{\psi(\beta, \lambda)}(z)}{\partial \beta_{n_m}} \\
\frac{\partial \eta^{\text{soft}}_{\psi(\beta, \lambda)}(z)}{\partial \beta_1} & \frac{\partial \eta^{\text{soft}}_{\psi(\beta, \lambda)}(z)}{\partial \beta_2} & \cdots & \frac{\partial \eta^{\text{soft}}_{\psi(\beta, \lambda)}(z)}{\partial \beta_{n_m}} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial \eta^{\text{soft}}_{\psi(\beta, \lambda)}(z)}{\partial \beta_1} & \frac{\partial \eta^{\text{soft}}_{\psi(\beta, \lambda)}(z)}{\partial \beta_2} & \cdots & \frac{\partial \eta^{\text{soft}}_{\psi(\beta, \lambda)}(z)}{\partial \beta_{n_m}}
\end{pmatrix}
\]
So the Jacobian matrix with respect to \( z \) is \( J_{f,z} = -J_{\eta, z} \), the Jacobian matrix with respect to \( \beta \) is 
\[
J_{f, \beta} = I_{nm \times nm} - J_{\eta, \beta}.
\]

From the implicit function theorem, we get
\[
\frac{\partial \hat{\eta}}{\partial z} - \frac{1}{J_{f, z}} \frac{\partial J_{f, z}}{\partial \beta} = \frac{1}{J_{f, \beta} - 1} \frac{\partial J_{f, \beta}}{\partial \beta} = \frac{1}{I_{nm \times nm} - J_{\eta, \beta}} \frac{\partial J_{\eta, \beta}}{\partial \beta}.
\]

We notice that \( \frac{\partial \hat{\eta}}{\partial z} \bigg|_{ij} = 1 \) if \(|z_j| > \lambda \| z - \beta \|_2\), otherwise it is null. So the trace of \( \frac{\partial \hat{\eta}}{\partial z} \) is \(|\{ \hat{\beta}_j \neq 0, j = 1, \ldots, n_m \}|\) when \( \frac{1}{\sqrt{\| z \|_0}} < \lambda < \frac{\| z \|_m}{\| z \|_2} \).

We conclude the derivative of solution with data is

\[
\frac{\partial \hat{\theta}}{\partial r} = \begin{cases} 
I_{n \times n}; & \lambda \leq \frac{1}{\sqrt{\| z \|_0}} \\
(I_{n \times n} - 0_{n \times n_m}); & \lambda \geq \frac{\| z \|_m}{\| z \|_2} \\
0_{n \times n \times n_m} - J_{\eta, \beta}^{-1} J_{f, z} & \frac{1}{\sqrt{\| z \|_0}} < \lambda < \frac{\| z \|_m}{\| z \|_2} 
\end{cases}
\]

Stein’s unbiased risk estimate formula leads to

\[
\text{SURE}(\lambda; \hat{\beta}_0, \hat{\beta}) = \left\| \left( \hat{\beta}_0 - \hat{\beta} \right) - \left( \Phi_0^T y^{(n)} - \Psi_0^T y^{(n)} \right) \right\|_2^2 + 2\sigma^2 \text{tr} \left( \frac{\partial \hat{\theta}}{\partial \hat{r}} \right) - \frac{\| x \|_2}{\| z \|_2} - n\sigma^2.
\]

The first term is \( \text{RSS}(\lambda) \) owing to the orthonormality of \( \Phi \), and the second term is \( n(\lambda) \).