A Greedy Homotopy Method for Regression with Nonconvex Constraints – Supplementary Material –

Fabian L. Wauthier **Peter Donnelly**
Wellcome Trust Centre for Human Genetics and Department of Statistics
University of Oxford

1 Proofs of Section 3

Recall that given a partition $\mathcal{G} = \{G_1, \ldots, G_g\}$ of $\{1, \ldots, p\}$ without singleton or empty sets, and a vector $\theta = (\theta_1, \ldots, \theta_g) \geq 0$, we defined

$$\Omega_{\theta,\mathcal{G}}(\beta) = \sum_{i<j \in G_{g'} \in \mathcal{G}} \frac{\omega_{\theta_{g'}}(\beta_i, \beta_j)}{|G_{g'}| - 1} \qquad \omega_{\theta_{g'}}(\beta_i, \beta_j) = \min(|\beta_i|, |\beta_j|)(1 + \theta_{g'}) + \max(|\beta_i|, |\beta_j|). \tag{1}$$

Let $B_{\theta,\mathcal{G}}(\tau) = \{\beta \in \mathbb{R}^p : \Omega_{\theta,\mathcal{G}}(\beta) \leq \tau\}$ be the induced constraint balls and $\Gamma(i) \in \{1, \ldots, g\}$ the (unique) group so that $i \in G_{\Gamma(i)}$.

1.1 Proof of Proposition 1

Proposition 1 (Union Decomposition)**.** *Let the partition be $\mathcal{G} = \{G_1, \ldots, G_g\}$ and the parameter $\theta = (\theta_1, \ldots, \theta_g) \geq 0$. There is a finite set $\mathcal{S}_{\theta,\mathcal{G}} \subset \mathbb{R}^p$ of vectors $s \geq \mathbf{1}$, so that for any $\tau > 0$*

$$B_{\theta,\mathcal{G}}(\tau) = \bigcup_{s \in \mathcal{S}_{\theta,\mathcal{G}}} B_s(\tau). \tag{2}$$

Define $\Pi_{g'}$ to be all permutations $\pi_{g'}$ of the elements in $G_{g'}$ and let $\Pi_{\mathcal{G}} = \times_{g'=1}^{g} \Pi_{g'}$ be their cross-product, whose elements $\pi \in \Pi_{\mathcal{G}}$ are g-tuples of permutations $\pi = (\pi_1, \ldots, \pi_g)$. For some $\pi \in \Pi_{\mathcal{G}}$, denote by $\pi_{\Gamma(i)}(i) \in \{1, \ldots, |G_{\Gamma(i)}|\}$ the position of $i \in G_{\Gamma(i)}$ in permutation $\pi_{\Gamma(i)}$. We have

$$\mathcal{S}_{\theta,\mathcal{G}} = \cup_{\pi \in \Pi_{\mathcal{G}}} \{s_\pi\} \tag{3}$$

$$s_{\pi,i} = 1 + (\pi_{\Gamma(i)}(i) - 1) \frac{\theta_{\Gamma(i)}}{|G_{\Gamma(i)}| - 1} \;\; \forall i = 1, \ldots, p. \tag{4}$$

Proof. We first show $B_{\theta,\mathcal{G}}(\tau) \subseteq \bigcup_{s \in \mathcal{S}_{\theta,\mathcal{G}}} B_s(\tau)$. Consider some $\beta \in B_{\theta,\mathcal{G}}(\tau)$ and let $\pi = (\pi_1, \ldots, \pi_g)$ be a tuple of permutations (not necessarily unique) induced by sorting the elements $|\beta_i|$ within each group specified by $\mathcal{G}$ so that for each group index g' we have

$$|\beta_{\pi_{g'}^{-1}(1)}| \geq \ldots \geq |\beta_{\pi_{g'}^{-1}(|G_{g'}|)}|. \tag{5}$$

By construction of $\Omega_{\theta,\mathcal{G}}(\cdot)$, β lies in the set

$$\left\{\beta' \in \mathbb{R}^p : \sum_{g'=1}^{g} \sum_{i=1}^{|G_{g'}|} \left(\frac{(|G_{g'}| - i) + (i-1)(1+\theta_{g'})}{|G_{g'}| - 1}\right) |\beta'_{\pi_{g'}^{-1}(i)}| \leq \tau\right\} \tag{6}$$

$$= \left\{\beta' \in \mathbb{R}^p : \sum_{g'=1}^{g} \sum_{i=1}^{|G_{g'}|} \left(1 + \frac{(i-1)\theta_{g'}}{|G_{g'}| - 1}\right) |\beta'_{\pi_{g'}^{-1}(i)}| \leq \tau\right\} \tag{7}$$

$$= \left\{\beta' \in \mathbb{R}^p : \sum_{g'=1}^{g} \sum_{i \in G_{g'}} \left(1 + (\pi_{g'}(i) - 1)\, \frac{\theta_{g'}}{|G_{g'}| - 1}\right) |\beta'_i| \leq \tau\right\} \tag{8}$$

$$= \left\{\beta' \in \mathbb{R}^p : \sum_{i=1}^{p} \left(1 + \left(\pi_{\Gamma(i)}(i) - 1\right) \frac{\theta_{\Gamma(i)}}{|G_{\Gamma(i)}| - 1}\right) |\beta'_i| \leq \tau\right\} = B_{s_\pi}(\tau) \tag{9}$$

We therefore conclude that $B_{\theta,\mathcal{G}}(\tau) \subseteq \bigcup_{s \in \mathcal{S}_{\theta,\mathcal{G}}} B_s(\tau)$, with

$$\mathcal{S}_{\theta,\mathcal{G}} = \cup_{\pi \in \Pi_{\mathcal{G}}} \{s_\pi\} \tag{10}$$

$$s_{\pi,i} = 1 + (\pi_{\Gamma(i)}(i) - 1)\frac{\theta_{\Gamma(i)}}{|G_{\Gamma(i)}| - 1} \ \forall i = 1, \ldots, p. \tag{11}$$

For the other direction, suppose that $\beta \in B_{s_{\tilde{\pi}}}(\tau)$ for some arbitrary tuple of permutations $\tilde{\pi} \in \Pi_{\mathcal{G}}$, which means that

$$\sum_{i=1}^{p} \left(1 + \left(\tilde{\pi}_{\Gamma(i)}(i) - 1\right) \frac{\theta_{\Gamma(i)}}{|G_{\Gamma(i)}| - 1}\right) |\beta_i| \leq \tau. \tag{12}$$

Then notice that if π is a (not necessarily unique) tuple of permutations induced by ordering elements $|\beta_i|$ within groups, we have, by arguing from pairwise swaps within groups that take $\tilde{\pi}$ to π, that

$$\sum_{i=1}^{p} \left(1 + \left(\pi_{\Gamma(i)}(i) - 1\right) \frac{\theta_{\Gamma(i)}}{|G_{\Gamma(i)}| - 1}\right) |\beta_i| \leq \sum_{i=1}^{p} \left(1 + \left(\tilde{\pi}_{\Gamma(i)}(i) - 1\right) \frac{\theta_{\Gamma(i)}}{|G_{\Gamma(i)}| - 1}\right) |\beta_i| \leq \tau, \tag{13}$$

and so $\beta \in B_{\theta,\mathcal{G}}(\tau)$. It follows that $\bigcup_{s \in \mathcal{S}_{\theta,\mathcal{G}}} B_s(\tau) \subseteq B_{\theta,\mathcal{G}}(\tau)$ and so $B_{\theta,\mathcal{G}}(\tau) = \bigcup_{s \in \mathcal{S}_{\theta,\mathcal{G}}} B_s(\tau)$.

□

1.2 Proof of Proposition 2

Recall that we are considering the nonconvex optimization problem

$$\begin{aligned}\beta(\tau) &\in \underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}} J_\tau(\beta) \\ &= \underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}} \begin{cases} \frac{1}{2n}\|y - X\beta\|_2^2 & \text{if } \beta \in B_{\theta,\mathcal{G}}(\tau) \\ \infty & \text{o.w.} \end{cases}.\end{aligned} \tag{P1}$$

Proposition 2 (Local Piecewise Linearity)**.** *Suppose X has absolutely continuous distribution and that $\exists\tau' > 0$ s.t. $\nexists\beta \in B_{\theta,\mathcal{G}}(\tau')$ which is a minimum of $\|y - X\beta\|_2^2$. Let $\tau_{\max}$ be the supremum over these τ'. The set of local minima of $J_\tau(\cdot)$ in (P1) with $\tau \in (0, \tau_{\max})$ is w.p. 1 a finite union of piecewise linear paths, each path indexed by τ and lying in the boundary of a ball, $bd(B_s(\tau)), s \in \mathcal{S}_{\theta,\mathcal{G}}$.*

Proof. Given the assumptions, for all $\tau \in (0, \tau_{\max})$, the elements β on the boundary of $B_{\theta,\mathcal{G}}(\tau)$ satisfy $(y - X\beta)^\top X \neq 0$. For each $\tau \in (0, \tau_{\max})$, let $\mathcal{M}_{\theta,\mathcal{G}}(\tau)$ be the set of local minima of $J_\tau(\cdot)$. Let the set $\mathcal{S}_{\theta,\mathcal{G}}$ be defined as in Proposition 1: For $\Pi_\mathcal{G}$ the set of g-tuples of permutations induced by $\mathcal{G}$,

$$\mathcal{S}_{\theta,\mathcal{G}} = \cup_{\pi\in\Pi_\mathcal{G}}\{s_\pi\} \tag{14}$$

$$s_{\pi,i} = 1 + (\pi_{\Gamma(i)}(i) - 1)\frac{\theta_{\Gamma(i)}}{|G_{\Gamma(i)}| - 1} \;\forall i = 1, \ldots, p. \tag{15}$$

For some $s_\pi \in \mathcal{S}_{\theta,\mathcal{G}}$, define $\mathcal{M}_{s_\pi}(\tau)$ to be the solution to $(P1)$ with $B_{\theta,\mathcal{G}}(\tau)$ replaced by $B_{s_\pi}(\tau)$. For each $s_\pi \in \mathcal{S}_{\theta,\mathcal{G}}$ the ball $B_{s_\pi}(\tau)$ corresponds to a weighted ℓ_1 norm, and if X is drawn from an absolutely continuous distribution, then the solution $\mathcal{M}_{s_\pi}(\tau)$ is with probability 1 unique on $(0, \tau_{\max})$ [6]. Additionally, the result of Rosset and Zhu [5] shows that the resulting regularization path $\mathcal{M}_{s_\pi}(\tau)$ is piecewise linear on $(0, \tau_{\max})$. Due to the union decomposition of Proposition 1, it follows immediately that $\mathcal{M}_{\theta,\mathcal{G}}(\tau) \subseteq \bigcup_{s_\pi\in\mathcal{S}_{\theta,\mathcal{G}}} \mathcal{M}_{s_\pi}(\tau)$ for $\tau \in (0, \tau_{\max})$. However, we seek not a superset of $\mathcal{M}_{\theta,\mathcal{G}}(\tau)$, but a characterization as a union of paths on the boundaries of weighted ℓ_1 balls. That, is we seek a set $P \subseteq \Pi_\mathcal{G}$ so that $\mathcal{M}_{\theta,\mathcal{G}}(\tau) = \bigcup_{\pi\in P} \mathcal{M}_{s_\pi}(\tau)$ for $\tau \in (0, \tau_{\max})$. The existence of such a set P can be guaranteed if for any $s_\pi \in \mathcal{S}_{\theta,\mathcal{G}}$, $\mathcal{M}_{s_\pi}(\tau)$ either lies $\forall\tau \in (0, \tau_{\max})$ in the interior of $B_{\theta,\mathcal{G}}(\tau)$ or it lies $\forall\tau \in (0, \tau_{\max})$ on the boundary of $B_{\theta,\mathcal{G}}(\tau)$. To show this, we show that for $\tau \in (0, \tau_{\max})$ no local minimum in $\mathcal{M}_{\theta,\mathcal{G}}(\tau)$ lies at a concave kink of $B_{\theta,\mathcal{G}}(\tau)$ (which are the points where a path would switch from being in the interior to being on the boundary or vice versa).

Suppose then (for the purpose of deriving a contradiction) that for some $\tau \in (0, \tau_{\max})$, we have that β is a local minimum in $\mathcal{M}_{\theta,\mathcal{G}}(\tau)$ that lies at one of the concave kinks of $B_{\theta,\mathcal{G}}(\tau)$. If β lies at a concave kink, then since $\tau_{\max} > 0$, we know that for at least two elements $i \neq j \in G \in \mathcal{G}$, $\beta_i \neq 0, \beta_j \neq 0$. For if only a single element $\neq 0$, then we lie at one of the points of $B_{\theta,\mathcal{G}}(\tau)$ and if the only two nonzero elements lie in different groups, β cannot lie at a concave kink. Specifically, the concave kink is identified by sets of indices i in a group $G \in \mathcal{G}$ so that the corresponding $\beta_i \neq 0$ have identical magnitude. The vector β induces a set $\Sigma \subseteq \Pi_\mathcal{G}$ of g-tuples of permutations σ by sorting $|\beta_i|$ by their magnitudes within each group $G \in \mathcal{G} = \{G_1, \ldots, G_g\}$ (with tie-breaking). We know that for each $\sigma \in \Sigma$, $\|\mathrm{diag}(s_\sigma)\beta\|_1 = \tau$, that is, β lies on the boundary of $B_{s_\sigma}(\tau)$. Each σ thus corresponds to an active constraint on β. Since we can think of β as a local minimum of $\|y - X\beta\|_2^2$, subject to either of these (convex) constraints, we have by convexity for any $\sigma \in \Sigma$ a subgradient vector $z_\sigma \in \partial\|\beta\|_1$ and a constant λ_σ so that

$$(y - X\beta)^\top X = \lambda_\sigma \mathrm{diag}(z_\sigma)s_\sigma. \tag{16}$$

Because there are at least two elements $i \neq j \in G \subseteq \mathcal{G}$ with $|\beta_i| = |\beta_j| \neq 0$ we know that $\forall\sigma \in \Sigma$, $z_{\sigma,i} = \mathrm{sgn}(\beta_i), z_{\sigma,j} = \mathrm{sgn}(\beta_j)$, which implies that $z_{\sigma,i}s_{\sigma,i} \neq 0, z_{\sigma,j}s_{\sigma,j} \neq 0$. Additionally, by construction $(y - X\beta)^\top X \neq 0$ and so we know $\lambda_\sigma \neq 0$. By the construction of s_π in Eq. (15), we know that $\exists \sigma_1 \neq \sigma_2 \in \Sigma$, so that s_{σ_1} and s_{σ_2} differ only on elements i, j. However Eq. (16) then cannot simultaneously hold unless $\lambda_\sigma = 0$ and $(y - X\beta)^\top X = 0$ which we ruled out earlier. Thus we have a contradiction and so the assumption that β lies at a concave kink must be wrong.

Because local minima in $\mathcal{M}_{\theta,\mathcal{G}}(\tau)$ never lie at concave kinks of $B_{\theta,\mathcal{G}}(\tau)$ for $\tau \in (0, \tau_{\max})$, we know that for each local minimum path on $(0, \tau_{\max})$, there is a $\pi \in \Pi_{\mathcal{G}}$ so that the path lies on $B_{s_\pi}(\tau)$. That is, there is some nonempty subset $P \subseteq \Pi_{\mathcal{G}}$ so that $\mathcal{M}_{\theta,\mathcal{G}}(\tau) = \bigcup_{\pi \in P} \mathcal{M}_{s_\pi}(\tau)$ is a union of piecewise linear paths. $\square$

2 Proofs of Section 4

2.1 Proof of Proposition 3

Recall that we are considering the surrogate problem

$$\bar{\beta}(\lambda) \in \operatorname{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\operatorname{diag}(s^*)\beta\|_1 . \tag{S}$$

For a positive vector b, let $\bar{\beta}_b(\lambda)$ be a solution to (S) with penalty $\lambda\|\operatorname{diag}(b)\beta\|_1$.

Proposition 3 (Recoverability of (S))**.** *Suppose X has absolutely continuous distribution. For any vectors $a \geq b \geq \mathbf{1}$ and $\lambda > 0$, w.p. 1 $\bar{\beta}_a(\lambda), \bar{\beta}_b(\lambda)$ are unique. If additionally $\|diag(a)\bar{\beta}_b(\lambda)\|_1 = \|diag(b)\bar{\beta}_b(\lambda)\|_1$, then $\bar{\beta}_a(\lambda) = \bar{\beta}_b(\lambda)$.*

Proof. Since X is absolutely continuous, $a \geq b \geq 0$ and $\lambda > 0$, it follows that $\bar{\beta}_a(\lambda), \bar{\beta}_b(\lambda)$ are almost surely unique [6]. Since $a \geq b \geq \mathbf{1}$, we have $\forall \beta \in \mathbb{R}^p$

$$\frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\operatorname{diag}(b)\beta\|_1 \leq \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\operatorname{diag}(a)\beta\|_1 . \tag{17}$$

However, we also know

$$\lambda \left\|\operatorname{diag}(b)\bar{\beta}_b(\lambda)\right\|_1 = \lambda \left\|\operatorname{diag}(a)\bar{\beta}_b(\lambda)\right\|_1 . \tag{18}$$

It follows that we must have $\bar{\beta}_a(\lambda) = \bar{\beta}_b(\lambda)$. $\square$

3 Proofs of Section 5

Section 5 compares the estimator of β^* produced by the RepLasso algorithm, with the estimator of β^* produced by the Lasso. For convenience we reproduce the RepLasso algorithm below.

Algorithm 1: REPLASSO$(X, y, \mathcal{G}, \theta)$

$\bar{y} = 0, A = (), L = 0, \lambda = \|X^\top y\|_\infty, s(\lambda) = \mathbf{1}, \bar{\beta}(\lambda) = 0$

while $\lambda > 0$

$$\text{Stage 1}\begin{cases} \textbf{if } L = 0 \ \ \#\text{ Add a variable} \\ \quad A = (A, i^*), \ \text{where } i^* = \text{argmax}_{j \in A^c} \left| X_j^\top (y - \bar{y}) / s_j(\lambda) \right| \\ \quad s_M(\lambda^-) = s_M(\lambda) + \frac{\theta_{\Gamma(i^*)}}{|G_{\Gamma(i^*)}| - 1} 1_M, \quad s_{M^c}(\lambda^-) = s_{M^c}(\lambda), \quad \text{with} \quad M = \left\{ A^c \cap G_{\Gamma(i^*)} \right\} \\ \textbf{if } L = 1 \ \ \#\text{ Delete a variable} \\ \quad A = A \backslash i^*, \ \text{where } i^* = \text{arg}_{i \in A} [\![\bar{\beta}_i(\lambda) = 0]\!] \end{cases}$$

$$\text{Stage 2}\left\{ \bar{w}_A = A_A \left(X_A^\top X_A \right)^{-1} \text{diag}(\text{sgn}\left(X_A^\top (y - \bar{y}) \right)) s_A(\lambda), \ \text{with } A_A \text{ s.t. } \|X_A \bar{w}_A\|_2^2 = 1 \right.$$

$$\text{Stage 3}\begin{cases} \text{Find smallest } \rho > 0 \text{ s.t.} \\ \quad \bullet \ \exists j \in A^c \text{ s.t. } |X_j^\top (y - \bar{y} - \rho X_A \bar{w}_A) / s_j(\lambda)| = \lambda - \rho\text{: set } L = 0 \\ \quad \bullet \ \exists i \in A \text{ s.t. } \bar{\beta}_i(\lambda) \neq 0 \text{ and } \bar{\beta}_i(\lambda) + \rho w_i = 0\text{: set } L = 1 \end{cases}$$

$$\text{Stage 4}\begin{cases} \bar{\beta}_A(\lambda - \rho) = \bar{\beta}_A(\lambda) + \rho \bar{w}_A, \quad \bar{\beta}_{A^c}(\lambda - \rho) = 0, \quad \bar{y} = X\bar{\beta}(\lambda - \rho) \\ \lambda = \lambda - \rho \end{cases}$$

return $\bar{\beta}$

The RepLasso is a generalization of the Lasso homotopy method, which maintains a set of weights $s(\lambda)$. Indeed, the RepLasso is identical to the Lars algorithm with Lasso modification of Efron et al. [1] if we force $\theta = \mathbf{0}$, which implies that $\forall \lambda, s(\lambda) = \mathbf{1}$ (We note, however, that for notational convenience the definition of $\bar{w}_A$ differs slightly from that in Efron et al. [1] in that case). In the following we will carry out our comparison of RepLasso with the Lasso homotopy method by comparing the RepLasso with $\theta \neq \mathbf{0}$ and the RepLasso with $\theta = \mathbf{0}$. We will denote by $\hat{\beta}(\lambda)$ the estimator resulting from the specialization to the Lasso case. Similarly, we let $\hat{w}_A$ be the vector corresponding to $\bar{w}_A$ for the Lasso specialization. Suppose that the support set of β^* is $S \triangleq S(\beta^*)$. Let X_j be the column j of X and X_A a matrix which consists of the columns indexed by A.

The proofs of Section 5 use the following assumptions.

A1: $\forall G \in \mathcal{G}, |\{i \in G : \beta_i^* \neq 0\}| \leq 1$

A2: $\forall A \subset S$ and u_A the equiangular vector in Eq. (2.6) of [1], $\nexists j \in A^c, |X_A^\top u_A| = |X_j^\top u_A| \mathbf{1}$

3.1 Proof of Theorem 2

Theorem 2 (Lasso Recovery)**.** *Assume that* **A1–2** *hold. Conditioned on* X, y*, we have for any* $\lambda_{\min} > 0$

$$\forall\lambda \geq \lambda_{\min}\ S(\hat{\beta}(\lambda)) \subseteq S \implies \forall\lambda \geq \lambda_{\min}\ \hat{\beta}(\lambda) = \bar{\beta}(\lambda).$$

Proof. Suppose then that $\forall\lambda \geq \lambda_{\min}, S(\hat{\beta}(\lambda)) \subseteq S$. Suppose that $\hat{\beta}(\lambda_{\min})$ corresponds to iteration t of the Lasso. For $t' \leq t$, let $\hat{A}_{t'}$ and $\bar{A}_{t'}$ be the sequence of active sets of the Lasso and RepLasso up to iteration t. With a slight abuse of notation we will temporarily treat an active set as an unordered set. Assumption **A2** guarantees that for the Lasso, any variable that is at some point in the active set is also at some point in the support set. To see this, note that that by **A2**, the vector $\hat{w}_A$ never contains a zero element. If it did, then an equiangular vector u_A of X_A as in Eq. (2.6) of [1] could be constructed using a strict subset of vectors indexed by A, violating assumption **A2**. But if $\hat{w}_A$ does not contain a zero element, then the elements in the active set A cannot indefinitely be assigned a $\hat{\beta}_A(\lambda)$ coefficient of zero as λ is swept out. Finally, because we know $\forall\lambda \geq \lambda_{\min}, S(\hat{\beta}(\lambda)) \subseteq S$, this means that $\forall t' \leq t, \hat{A}_{t'} \subseteq S$. We will now argue by induction that the induced sequence of active sets $\bar{A}_{t'}$ of the RepLasso also satisfies $\forall t' \leq t, \bar{A}_{t'} \subseteq S$.

Base case: Since $s(\|X^\top y\|_\infty) = \mathbf{1}$, the first variable selected by RepLasso and the Lasso method is the same. That is, $\hat{A}_1 = \bar{A}_1 \subseteq S$ at iteration 1.

Inductive step: Assume that $\forall t'' \leq t', \hat{A}_{t''} = \bar{A}_{t''} \subseteq S$. Since $\forall t'' \leq t', \bar{A}_{t''} \subseteq S$, we know by **A1** that for all λ up to iteration t', $s(\lambda)$ did not change on S, i.e. $s_S(\lambda) = \mathbf{1}$. This in particular means that both the Lasso and the RepLasso will have arrived at the same value of λ and intermediate estimate $\hat{\beta}(\lambda) = \bar{\beta}(\lambda)$ of β^* at the end of stage 4 of iteration $t' - 1$ and the same vectors $\hat{w}_A = \bar{w}_A$ at the end of stage 2 of iteration t'. To see this, notice that since $\forall t'' \leq t', \bar{A}_{t''} \subseteq S$, and since for all λ up to iteration t' we had $s_S(\lambda) = \mathbf{1}$, the RepLasso is up to stage 2 of iteration t' equivalent to running Lasso on the subset of variables X_S, y.

At stage 3 of iteration t', the RepLasso algorithm determines whether to add or remove a variable from $\bar{A}_{t'}$ in stage 1 of iteration $t' + 1$. Since the value of λ and the intermediate variables $\hat{\beta}(\lambda) = \bar{\beta}(\lambda)$ and $\hat{w}_A = \bar{w}_A$ are the same at stage 2 of iteration t' we can now use properties of $s(\lambda)$ to show that this implies $\hat{A}_{t'+1} = \bar{A}_{t'+1}$. We consider two cases:

1. The Lasso determines to add a variable in iteration $t' + 1$ (first bullet in stage 3). Since $s_S(\lambda) = \mathbf{1}$ did not change on S, since we always have $s(\lambda) \geq \mathbf{1}$ and since $\hat{A}_{t'+1} \subseteq S$, it follows that the RepLasso will add the same variable.

2. The Lasso determines to remove a variable in iteration $t'+1$ (second bullet in stage 3). Since $s_S(\lambda) = \mathbf{1}$ did not change on S and since we always have $s(\lambda) \geq \mathbf{1}$, it then follows that the RepLasso will remove the same variable.

Hence, it follows that $\hat{A}_{t'+1} = \bar{A}_{t'+1} \subseteq S$. By the principle of induction, we have shown that $\forall t' \leq t, \bar{A}_{t'} \subseteq S$.

Now notice that the value of λ at the end of stage 4 of iteration t must be the same for RepLasso and Lasso and that for that value we have by definition $\lambda < \lambda_{\min}$. Next, since for all values of $\lambda > \lambda_{\min}$ we have $s_S(\lambda) = \mathbf{1}$, and since $\forall t' \leq t, \bar{A}_{t'} \subseteq S$, RepLasso is up to $\lambda_{\min}$ equivalent to running Lasso on X_S, y, and so $\forall\lambda \geq \lambda_{\min}, \hat{\beta}(\lambda) = \bar{\beta}(\lambda)$. ∎

4 RepLars: A RepLasso variant

As noted in the paper, if $\theta = \mathbf{0}$, and we force $L = 0$ then the RepLasso algorithm reduces to the Lars algorithm of Efron et al. [1]. (We note, however, that the definition of $\bar{w}_A$ differs slightly from that of the Lars algorithm given in [1]). When we force $L = 0$ but allow $\theta \neq \mathbf{0}$ we have a new algorithm, which we call RepLars. In this section we present this algorithm and analyze its behavior.

Algorithm 2: REPLARS$(X, y, \mathcal{G}, \theta)$

$\bar{y} = 0, A = (), \lambda = \|X^\top y\|_\infty, s(\lambda) = \mathbf{1}, \bar{\beta}(\lambda) = 0$

while $\lambda > 0$

$$\text{Stage 1}\begin{cases} A = (A, i^*),\ \text{where } i^* = \operatorname{argmax}_{j \in A^c} \left|X_j^\top (y - \bar{y})/s_j(\lambda)\right| \\ s_M(\lambda^-) = s_M(\lambda) + \frac{\theta_{\Gamma(i^*)}}{|G_{\Gamma(i^*)}|-1}\mathbf{1}, \quad s_{M^c}(\lambda^-) = s_{M^c}(\lambda), \quad \text{with} \quad M = \left\{A^c \cap G_{\Gamma(i^*)}\right\} \end{cases}$$

$$\text{Stage 2}\left\{ \bar{w}_A = A_A \left(X_A^\top X_A\right)^{-1} \operatorname{diag}(\operatorname{sgn}\left(X_A^\top (y - \bar{y})\right)) s_A(\lambda),\ \text{with } A_A \text{ s.t. } \|X_A \bar{w}_A\|_2^2 = 1 \right.$$

$$\text{Stage 3}\left\{ \text{Find smallest } \rho > 0 \text{ s.t. } \exists j \in A^c \text{ s.t. } |X_j^\top (y - \bar{y} - \rho X_A \bar{w}_A)/s_j(\lambda)| = \lambda - \rho \right.$$

$$\text{Stage 4}\begin{cases} \bar{\beta}_A(\lambda - \rho) = \bar{\beta}_A(\lambda) + \rho \bar{w}_A, \quad \bar{\beta}_{A^c}(\lambda - \rho) = 0, \quad \bar{y} = X\bar{\beta}(\lambda - \rho) \\ \lambda = \lambda - \rho \end{cases}$$

return $\bar{\beta}$

In the following we will compare of RepLars with the Lars by comparing RepLars with $\theta > 0$ and RepLars with $\theta = \mathbf{0}$. We will denote by $\hat{\beta}$ the Lars estimate corresponding to $\bar{\beta}$ produced by the algorithm above. Similarly, we let $\hat{w}_A$ be the vector in the Lars specialization corresponding to $\bar{w}_A$ above. We will assume throughout this analysis that variables are added to the active set one by one.

Theorem 3 (Lars Recovery)**.** *Assume that* **A1–2** *hold. Conditioned on* X, y*, we have for any* $\lambda_{\min} > 0$

$$\forall \lambda \geq \lambda_{\min}\ S(\hat{\beta}(\lambda)) \subseteq S \implies \forall \lambda \geq \lambda_{\min}\ \hat{\beta}(\lambda) = \bar{\beta}(\lambda). \tag{19}$$

Proof. Suppose that $\forall \lambda \geq \lambda_{\min}\ S(\hat{\beta}(\lambda)) \subseteq S$. Suppose further that $\hat{\beta}(\lambda_{\min})$ corresponds to iteration t of Lars. For $t' \leq t$ let $\hat{A}_{t'}$ be the active sets of Lars at iteration t'. With a slight abuse of notation we will temporarily treat an active set as an unordered set. Assumption **A2** guarantees that any variable that is at some point in the active set is also at some point in the support set. To see this, note that that by **A2**, the vector $\hat{w}_A$ never contains a zero element. If it did, then an equiangular vector u_A of X_A as in Eq. (2.6) of [1] could be constructed using a strict subset of vectors indexed by A, violating assumption **A2**. But if $\hat{w}_A$ does not contain a zero element, then the elements in the active set A cannot indefinitely be assigned a $\hat{\beta}_A(\lambda)$ coefficient of zero as λ is swept out. Finally, because we know $\forall \lambda \geq \lambda_{\min}, S(\hat{\beta}(\lambda)) \subseteq S$, this means that $\forall t' \leq t, \hat{A}_{t'} \subseteq S$. We will now argue by induction that the induced sequence of active sets $\bar{A}_{t'}$ of the RepLars also satisfies $\forall t' \leq t, \bar{A}_{t'} \subseteq S$.

Base case: Since $s(\|X^\top y\|_\infty) = \mathbf{1}$, the first variable selected by RepLars and the Lars method is the same. That is, $\hat{A}_1 = \bar{A}_1 \subseteq S$ at iteration 1.

Inductive step: Assume that $\forall t'' \leq t', \hat{A}_{t''} = \bar{A}_{t''} \subseteq S$. Since $\forall t'' \leq t', \bar{A}_{t''} \subseteq S$, we know by **A1** that for all λ up to iteration t', $s(\lambda)$ did not change on S, i.e. $s_S(\lambda) = \mathbf{1}$. This in particular means that both the Lars and the RepLars will have arrived at the same value of λ and intermediate estimate $\hat{\beta}(\lambda) = \bar{\beta}(\lambda)$ of β^* at the end of stage 4 of iteration $t' - 1$ and the same vectors $\hat{w}_A = \bar{w}_A$ at the end of stage 2 of iteration t'. To see this, notice that since $\forall t'' \leq t', \bar{A}_{t''} \subseteq S$, and since for all λ up to iteration t' we had $s_S(\lambda) = \mathbf{1}$, the RepLars is up to stage 2 of iteration t' equivalent to running Lars on the subset of variables X_S, y.

At stage 3 of iteration t', the RepLars algorithm determines which variable to add to $\bar{A}_{t'}$ in stage 1 of iteration $t'+1$. Since the value of λ and the intermediate variables $\hat{\beta}(\lambda) = \bar{\beta}(\lambda)$ and $\hat{w}_A = \bar{w}_A$ are the same at stage 2 of iteration t' we can now use properties of $s(\lambda)$ to show that this implies $\hat{A}_{t'+1} = \bar{A}_{t'+1}$. Specifically, since (1) $s_S(\lambda) = \mathbf{1}$ did not change on S; (2) we always have $s(\lambda) \geq \mathbf{1}$; and (3) $\hat{A}_{t'+1} \subseteq S$, it follows that the RepLars will add the same variable.

Hence, it follows that $\hat{A}_{t'+1} = \bar{A}_{t'+1} \subseteq S$. By the principle of induction, we have shown that $\forall t' \leq t, \bar{A}_{t'} \subseteq S$.

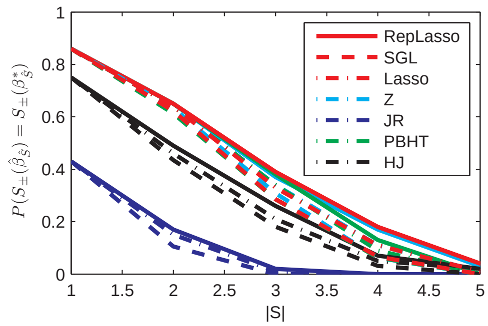
Notice that the value of λ at the end of stage 4 of iteration t must be the same for RepLars and Lars, and that for that value we have by definition $\lambda < \lambda_{\min}$. Since $\forall t' \leq t, \bar{A}_{t'} \subseteq S$, and since for all λ up to iteration t we had $s_S(\lambda) = \mathbf{1}$, the RepLars is up to $\lambda_{\min}$ equivalent to running Lars on the subset of variables X_S, y and so $\forall \lambda \geq \lambda_{\min}, \hat{\beta}(\lambda) = \bar{\beta}(\lambda)$. □

As in the main paper, several corollaries on the support recovery behavior of RepLars and Lars follow immediately. For example, we have

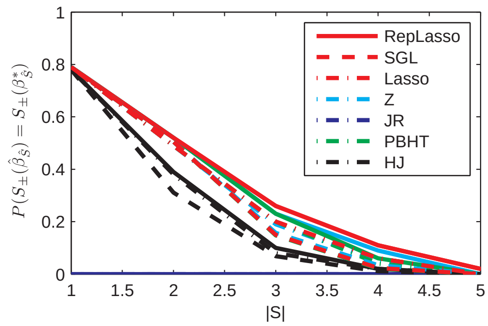
Corollary 2 (Support Recovery)**.** *Assume that* **A1–2** *hold. Conditioned on* X, y, *we have for any* $\lambda_{\min} > 0$

$$\forall \lambda \geq \lambda_{\min}\ S(\hat{\beta}(\lambda)) \subseteq S \implies \forall \lambda \geq \lambda_{\min} S(\bar{\beta}(\lambda)) \subseteq S$$
$$\forall \lambda \geq \lambda_{\min} S_\pm(\hat{\beta}(\lambda)) \subseteq S_\pm \implies \forall \lambda \geq \lambda_{\min} S_\pm(\bar{\beta}(\lambda)) \subseteq S_\pm.$$

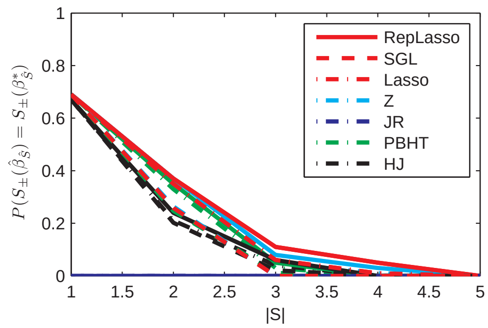
Thus, under **A1–2**, whenever Lars recovers the correct (signed) support, so does RepLars. Hence, we see that RepLars cannot do worse than Lars in terms of (signed) support recovery.



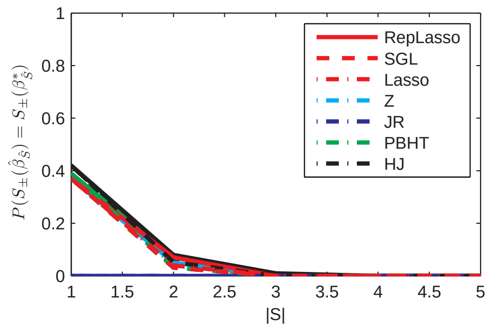
(a) $\rho \approx 0.1$



(b) $\rho \approx 0.2$



(c) $\rho \approx 0.3$



(d) $\rho \approx 0.5$

Figure 1: (a), (b), (c), (d) Results on synthetic data with $n = 172, p = 1000, \mathcal{G} =$ groups of 5 and within-group correlations ρ. We show the empirical probability that a subset of the correct *signed* support is recovered as a function of support size. For each base-method we show three curves, grouped by color. The performance of the existing Lasso variant (i.e., Lasso, Z, JR, PBHT, HJ) is shown as dashed-dotted curve; the performance of the algorithm with the Lasso replaced by RepLasso/SGL is shown in solid/dashed respectively (see text for details).

5 Additional Experimental Results

In this section we present experimental results for the high dimensional setting ($n \ll p$). Figure 1 shows results for a dataset with $n = 172, p = 1000$ and a correct support set of size 5. The setup of is otherwise the same as in Figures 3(a), (b) in the main paper, however, for convenience we will briefly summarize it here. As in the main paper, the figures show the probability of correctly recovering a correctly signed subset of the true support for varying levels of within-group correlation ρ. The red solid, dashed-dotted and dashed lines correspond to RepLasso, Lasso, and Sparse Group Lasso (SGL), as before. Also, we again evaluated the performance of four other methods that solve a standard Lasso problem after pre-processing the data X, y in some way. For each algorithm we show three curves, grouped by colors: the original method is shown as dashed-dotted curve, the method with the Lasso replaced by RepLasso as solid curve, and the method with the Lasso replaced by SGL as dashed curve. The four methods are: (1) the Adaptive Lasso of Zou [7] (*Z*); (2) the "Whitened" Lasso of Jia and Rohe [3] (*JR*); (3) the Preconditioned Lasso of Paul et al. [4] (*PBHT*); and (4) Correlation Sifting of Huang and Jojic [2] (*HJ*). Since $n < p$ we let the Adaptive Lasso scale columns of X by univariate regression coefficients. We notice that while the magnitude of the improvement of RepLasso variants over Lasso variants has decreased relative to those in Figure 3 of the main paper, there is still a strict improvement, in line with Theorem 2 and Corollary 1 of the main paper. Algorithms using SGL as a drop-in replacement for the Lasso still perform worse than the corresponding Lasso or RepLasso variants. Finally, as the problems get harder (i.e., ρ increases), the methods become indistinguishable (except for HJ).

References

[1] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least Angle Regression. *Ann. Stat.*, 32:407–499, 2004.

[2] J.C. Huang and N. Jojic. Variable selection through Correlation Sifting. In *RECOMB*, volume 6577 of *LNCS*, pages 106–123, 2011.

[3] J. Jia and K. Rohe. "Preconditioning" to comply with the irrepresentable condition. 2012.

[4] D. Paul, E. Bair, T. Hastie, and R. Tibshirani. "Preconditioning" for feature selection and regression in high-dimensional problems. *Ann. Stat.*, 36(4):1595–1618, 2008.

[5] S. Rosset and J. Zhu. Piecewise linear regularized solution paths. *Ann. Stat.*, pages 1012–1030, 2007.

[6] R.J. Tibshirani. The Lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490, 2013.

[7] H. Zou. The adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476):1418–1429, 2006.