A Greedy Homotopy Method for Regression with Nonconvex Constraints

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

Abstract

The goal of this paper is to estimate sparse linear regression models, where for a given partition $\mathcal{G}$ of input variables, the selected variables are chosen from a *diverse* set of groups in $\mathcal{G}$. We consider a novel class of nonconvex constraint functions, and develop RepLasso, a greedy homotopy method that exploits geometrical properties of the constraint functions to build a sequence of suitably adapted convex surrogate problems. We prove that in some situations RepLasso recovers the global minima path of the nonconvex problem. Moreover, even if it does not recover the global minima, we prove that it will often do no worse than the Lasso in terms of (signed) support recovery, while in practice outperforming it. We show empirically that the strategy can also be used to improve over various other Lasso-style algorithms. Finally, a GWAS of ankylosing spondylitis highlights our method's practical utility.

1 Introduction

We are interested in model sparsity for linear observation models of the form

$$y = X\beta^* + w \qquad w \sim \mathcal{N}(0, \sigma^2 I), \tag{1}$$

where X is an $n \times p$ matrix of covariates, β^* is a regression parameter, w is a noise vector and y is a vector of responses. Given X, y, a constraint function $\Omega(\cdot)$, and constraint parameter $\tau > 0$, constrained least squares regression estimates β^* as

$$\underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2n} \|y - X\beta\|_2^2 \quad \text{s.t.} \quad \Omega(\beta) \leq \tau. \tag{G}$$

A closely related formulation writes (G) in penalized form using penalizer $\Omega(\cdot)$ with a penalty parameter $\lambda > 0$. In the following, we will motivate our algorithm using the constrained formulation, but then show that it solves a certain sequence of penalized problems.

Sparse model estimation is an integral part of the statistics toolbox and is especially relevant for the $n \ll p$ case. The Lasso [24] is a well-known instance of (G) which replaces a hard ℓ_0 constraint (i.e., $\|\beta\|_0 \leq \tau$) by an ℓ_1 surrogate (i.e., $\|\beta\|_1 \leq \tau$) that retains sparsity-inducing properties. Since the Lasso regularization path is continuous and piecewise linear [21], it can be easily traced out using the homotopy method [5, 19]. In the following, we will often refer to the Lasso homotopy method simply as the Lasso. The efficiency of the homotopy method is one of the Lasso's key assets and is critical for efficient model selection.

Recently, there has been increased interest in enhancing the Lasso with structured sparsity by replacing the ℓ_1 penalizer with more complex, yet still convex, penalizers [11, 13, 14, 22, 28]. Because the overall objective remains convex in β, efficient algorithms exist to solve these problems. While these methods have many practical applications, the focus on convex formulations has necessarily excluded important inference problems that cannot be phrased in terms of structured convex objectives. In particular, convex formulations do not encourage solutions that are sparser than those of the Lasso. This paper was motivated by a class of applications where this stronger level of sparsity is desired yet impractical to produce using current methods. Specifically, we are interested in situations where: (1) we want to select a small subset of variables for predicting y; and (2) given a partition $\mathcal{G}$ of variables $\{1, \ldots, p\}$, each group contains at most a small number of selected variables. In other words, we seek a sparse solution where the variables are selected from a *diverse* set of groups, each variable acting in some sense as a *representative* of the group. Such a solution is both sparse at the group *and* the within-group level. As an example, in the Genome-Wide Association Study

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

(GWAS) of Section 6 it is reasonable to suppose that only a few Single Nucleotide Polymorphisms (SNPs) (the variables) are truly relevant for predicting the response and that these SNPs lie in a diverse set of genes (the groups).

We note that the desired sparsity behavior is orthogonal to that of the well-known Group Lasso [28]. The problem is also not adequately solved by the Elitist/Exclusive Lasso [14, 30], a convex formulation that selects *at least one* but at most a few variables from each group. Finally, the Sparse Group Lasso (SGL) [22] is another convex formulation that is unsuitable for our purpose. It is effectively a compromise between the Lasso and the Group Lasso and leads to solutions that, while being sparser than those of the Group Lasso, still tend to include multiple variables from the same group (see Figure 1(c)). We seek more parsimonious solutions than the Lasso and for this we must consider nonconvex constraint functions.

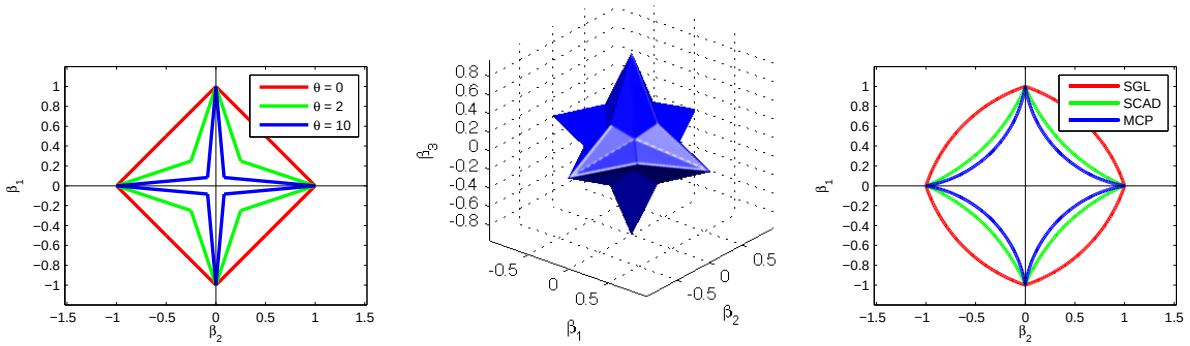
Given a parameter $\theta \in \mathbb{R}^p$ and the partition $\mathcal{G}$, we will encode our constraints as a nonconvex function $\Omega_{\theta,\mathcal{G}}(\cdot)$. There are specialized methods for the nonconvex penalized cousin of (G) for a fixed penalty parameter $\lambda > 0$ [4, 7, 10, 17, 33]. While these methods might be appropriate for finding a local minimum of (G) with $\Omega_{\theta,\mathcal{G}}(\cdot)$ and some τ fixed, they are not ideal for developing a homotopy-like algorithm which allows τ to range over an interval. This complicates their use when model selection over τ must be performed. We propose RepLasso (for "Representative Lasso"), a homotopy-like algorithm that attempts to fill this gap by exploiting certain properties of $\Omega_{\theta,\mathcal{G}}(\cdot)$. Roughly, RepLasso tries to build and solve a sequence of convex surrogates problems so that, as τ is swept out, the boundary of the surrogate constraint ball locally approximates the boundary of the ball induced by $\Omega_{\theta,\mathcal{G}}(\cdot)$. A crucial feature that allows us to do this efficiently is that the nonconvex constraint balls induced by $\Omega_{\theta,\mathcal{G}}(\cdot)$ can be decomposed as unions of convex balls. The sequence of surrogates is chosen so that the induced regularization path is continuous and piecewise linear and can thus be efficiently traced out.

We show theoretically that, under certain conditions, RepLasso traces out the global minima of (G) with constraints $\Omega_{\theta,\mathcal{G}}(\cdot)$. More importantly, we prove that, while RepLasso may not exactly solve (G) in general, on relevant instances it will still do at least as well as the Lasso in terms of support subset and signed support recovery. In practice, a strict improvement is observed over Lasso and SGL [22]. A class of Lasso-style algorithms has recently been popularized which pre-process X, y, before solving a Lasso problem (e.g., [9, 12, 20, 31]). As we demonstrate in Section 6, RepLasso can also yield improvements in this setting. Furthermore, RepLasso can be usefully applied to ℓ_1 constrained logistic regression [15], as we show in a GWAS application. Lastly, we prove in the Supplementary Material that, given some mild assumptions, a variant of RepLasso cannot do worse than the Lars algorithm of Efron et al. [5].

The paper is organized as follows: We review related research in Section 2 before introducing $\Omega_{\theta,\mathcal{G}}(\cdot)$ and simplifying (G) in Section 3. In Section 4 we present the RepLasso as a generalization of the Lasso homotopy method and in Section 5 theoretically compare the RepLasso and Lasso. Results on synthetic data and a GWAS application are given in Section 6. Our final remarks are in Section 7. The Supplementary Material contains all proofs and further experiments.

2 Related Research

Methods for optimizing convex loss functions with nonconvex regularizers include, among others, local quadratic approximation [7], minorization-maximization [10], local linear approximation [33] and composite gradient descent [17]. The Adaptive Lasso [4] is also relevant. However, since these methods focus on a fixed penalty parameter, they are not ideal for efficiently minimizing a sequence of problems (G) indexed by τ, as in a homotopy method. The SCAD [7] and MCP [29] are particularly important nonconvex penalties. However, they make no use of the grouping information $\mathcal{G}$ and are thus not useful in our application. Additionally, they do not possess the type of geometrical structure that we will exploit in our greedy homotopy algorithm. While there are other structured, nonconvex penalizers (e.g., [25]), there are also no homotopy algorithms to solve them. Various applications of the homotopy idea have so far focused on other convex problems. Examples include the Elastic Net [32] and the SVM [8]. The literature on gradient-based homotopy methods for convex problems is also growing [27]. The two most salient extensions of homotopy methods to nonconvex least squares problems are due to Zhang [29] and Wang et al. [26]. However, as both assume the penalty to be separable across the p coefficients, they are not useful for the type of structured sparsity we consider in this paper. There has been long-standing interest in convex structured extensions of the Lasso [11, 13, 14, 22, 28], and, as discussed in Section 1, while some methods are at first sight related to our approach (e.g., [14, 22, 28]), all of them rely on a convex penalizer that cannot encapsulate the structural constraints we are interested in. Tools for analyzing nonconvex problems have recently begun to emerge [17, 26], however, since the focus is again on separable penalizers, they are not useful in our case.



(a) (b) (c)

Figure 1: (a) The balls induced by Eq. (2) for $\mathcal{G} = \{\{1,2\}\}$ and varying values of θ. If $\theta = 0$ we recover the ℓ_1 norm as a special case. (b) A ball induced by Eq. (2) for $\mathcal{G} = \{\{1,2,3\}\}, \theta = 2$. (c) Example level sets for SGL [22], SCAD [7] and MCP [29]. For SGL we suppose a single group $G = \{1,2\}$. See [7, 22, 29] for details.

3 Structured Nonconvex Problems

Many situations exist where for some partition $\mathcal{G}$ of $\{1,\dots,p\}$ we know that β^* contains at most a few nonzero elements whose positions cover a diverse set of groups $G \in \mathcal{G}$.[1] Given a partition $\mathcal{G} = \{G_1,\dots,G_g\}$ without singleton or empty sets, and $\theta = (\theta_1,\dots,\theta_g) \geq \mathbf{0}$, the following *union constraint function* targets this situation

$$\Omega_{\theta,\mathcal{G}}(\beta) = \sum_{i<j\in G_{g'}\in\mathcal{G}} \frac{\omega_{\theta_{g'}}(\beta_i,\beta_j)}{|G_{g'}|-1} \tag{2}$$

$$\omega_{\theta_{g'}}(\beta_i,\beta_j) = \min(|\beta_i|,|\beta_j|)(1+\theta_{g'}) + \max(|\beta_i|,|\beta_j|).$$

Let $B_{\theta,\mathcal{G}}(\tau) = \{\beta \in \mathbb{R}^p : \Omega_{\theta,\mathcal{G}}(\beta) \leq \tau\}$ be the induced constraint balls. We are interested in the following nonconvex instance of (G) with the constrained objective $J_\tau(\beta)$ over β, indexed by τ

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

If $\theta = \mathbf{0}$, then $\Omega_{\theta,\mathcal{G}}(\beta) = \|\beta\|_1$ for all $\mathcal{G}$, and so $(P1)$ recovers the Lasso problem as special case. However, when $\theta \neq \mathbf{0}$, the constraint function is non-separable, nonconvex and, as exemplified in Figure 1, induces star-shaped balls[2]. When $\theta \neq \mathbf{0}$, $\Omega_{\theta,\mathcal{G}}(\beta)$ can be thought of as a nonconvex counterpart to the well-known Group Lasso penalty [28], where the nonconvexity encourages solutions $\beta(\tau)$ of $(P1)$ to select a small number of *representatives* from a diverse set groups in $\mathcal{G}$. The constraint $\Omega_{\theta,\mathcal{G}}(\beta)$ also differs from the Elitist/Exclusive Lasso penalty [14, 30], which effectively encourages each group to select at least one variable, or the SGL penalty [22], which encourages selection of multiple variables from the same group.

For some positive vector s, let $B_s(\tau) = \{\beta \in \mathbb{R}^p : \|\mathrm{diag}(s)\beta\|_1 \leq \tau\}$ be the s-weighted ℓ_1 ball. A key geometric property that we exploit is that $B_{\theta,\mathcal{G}}(\tau)$ can be written as a finite union of weighted ℓ_1 balls and so has planar faces. Let $\Gamma(i) \in \{1,\dots,g\}$ be the (unique) index so that $i \in G_{\Gamma(i)}$.

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

$$B_{\theta,\mathcal{G}}(\tau) = \bigcup_{s\in\mathcal{S}_{\theta,\mathcal{G}}} B_s(\tau). \tag{3}$$

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 $\pi = (\pi_1,\dots,\pi_g)$. For some $\pi \in \Pi_{\mathcal{G}}$, denote by $\pi_{\Gamma(i)}(i) \in \{1,\dots,|G_{\Gamma(i)}|\}$ the position of $i \in G_{\Gamma(i)}$ in permutation $\pi_{\Gamma(i)}$. We have

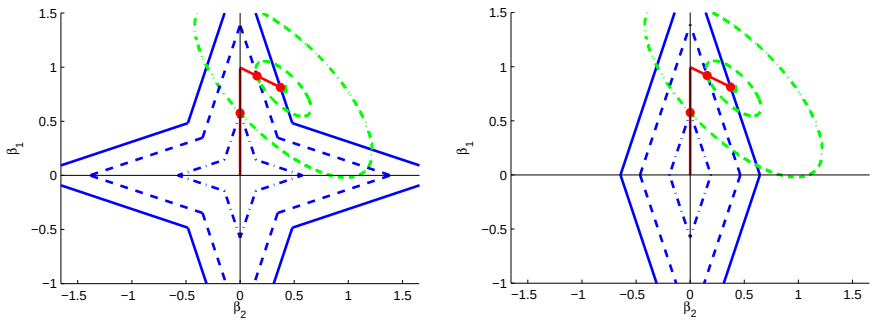
$$\mathcal{S}_{\theta,\mathcal{G}} = \cup_{\pi\in\Pi_{\mathcal{G}}}\{s_\pi\} \tag{4}$$

$$s_{\pi,i} = 1 + (\pi_{\Gamma(i)}(i) - 1)\frac{\theta_{\Gamma(i)}}{|G_{\Gamma(i)}|-1}. \tag{5}$$

Figure 1(c) shows that this property is not shared by other nonconvex penalties, e.g., SCAD [7] or MCP [29]. A common, brute force approach that eliminates the computational issues of $(P1)$ would be to replace $B_{\theta,\mathcal{G}}(\tau)$ by its convex hull, the ℓ_1 ball $B_1(\tau)$, thus recovering the Lasso. We advocate an orthogonal strat-

[1] In the GWAS application in Section 6, $\mathcal{G}$ corresponds to a partition of SNPs by genes and we know that in a diverse set of genes at most a few SNPs are truly relevant for predicting y.

[2] Note that when a subset of components of θ is set to zero, we can effectively treat the variables corresponding to those groups as ungrouped, as they only contribute an ℓ_1 penalty to $\Omega_{\theta,\mathcal{G}}(\beta)$.



(a) (b)

Figure 2: (a) The global minima path for $B_{2,\{\{1,2\}\}}(\tau)$. The dashed-dotted, dashed and solid constraint boundaries correspond to progressively larger τ. The solutions on the regularization path (red) are defined by intersections of the constraint set with the corresponding ellipsoid (red dots). (b) The corresponding regularization path for the weighted ℓ_1 ball $B_{(1,3)^\top}(\tau)$. If the weighting is known the same regularization path can be reproduced.

egy that instead focuses on replacing $B_{\theta,\mathcal{G}}(\tau)$ by a suitable *sequence* of *weighted* ℓ_1 balls, indexed by τ. To achieve this, we will exploit the decompositional structure of $B_{\theta,\mathcal{G}}(\tau)$ highlighted in Proposition 1. Specifically, our method is motivated by the following extension of a well-known result of Rosset and Zhu [21].

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, $bd(B_s(\tau)), s \in \mathcal{S}_{\theta,\mathcal{G}}$.*

Proposition 2 emphasizes that for the range of interesting values of $\tau \in (0, \tau_{\max})$, the local minima of $J_\tau(\cdot)$ in $(P1)$ can be grouped into a set of local minima *paths*, each indexed by τ. Moreover, any such local minimum path lies on some weighted ℓ_1 ball $B_s(\tau)$, with $s \in \mathcal{S}_{\theta,\mathcal{G}}$ appropriately chosen. With the aid of Proposition 2, it is possible to re-express $(P1)$ as a special set of penalized optimization problems, indexed by τ. This change of representation will be useful for the homotopy-like algorithm we present shortly. By Proposition 2 and convexity [2], for any solution $\beta(\tau)$ of $(P1)$ with $\tau \in (0, \tau_{\max})$ (i.e., a global minimum of $J_\tau(\cdot)$), $\exists \lambda^*(\tau), s^*(\tau) \in \mathcal{S}_{\theta,\mathcal{G}}$ so that

$$\beta(\tau) \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda^*(\tau)\|\mathrm{diag}(s^*(\tau))\beta\|_1. \quad (P2)$$

Thus, modulo uniqueness issues, there exist $\lambda^*(\tau)$, $s^*(\tau)$ so that $(P2)$ is in some sense equivalent to $(P1)$. Figure 2 shows a motivating example of this. In this case, the global minimum path of Figure 2(a) could be reproduced using the $B_{(1,3)^\top}(\tau)$ balls of Figure 2(b). Of course, knowledge of the vector-valued function $s^*(\tau) \in \mathcal{S}_{\theta,\mathcal{G}}$ would imply knowing for each τ roughly where on $B_{\theta,\mathcal{G}}(\tau)$ the global minimum of $J_\tau(\cdot)$ in $(P1)$ lies, which is hard in general. We thus cannot expect to be able to efficiently produce the *entire* regularization path of $(P1)$ for all $\tau \in (0, \tau_{\max})$ using the equivalence between $(P2)$ and $(P1)$.

A Simplifying Assumption. The formulation in $(P2)$ replicates the regularizing effect of $B_{\theta,\mathcal{G}}(\tau)$ in $(P1)$ using a sequence of weighted ℓ_1 balls that depend on τ (characterized by $s^*(\tau)$). This dependence is necessary as the global minimum of $J_\tau(\cdot)$ in $(P1)$ can "jump" from one weighted ℓ_1 ball to another as we vary $\tau \in (0, \tau_{\max})$. If we let $S(\beta)$ be the support of some vector β, we can simplify the problem of finding sequences $\lambda^*(\tau), s^*(\tau)$ for $(P2)$, by assuming that

A0: $\exists s^* \in \mathcal{S}_{\theta,\mathcal{G}}$ s.t. $\forall \tau \in (0, \tau_{\max})$ $(P1)$ has a unique solution $\in \mathrm{bd}(B_{s^*}(\tau))$. $\forall 0 < \tau_1 < \tau_2 < \tau_{\max}$, the solutions of $(P1)$ satisfy $S(\beta(\tau_1)) \subseteq S(\beta(\tau_2))$.

An example where **A0** holds with $s^* = (1, 3)^\top$ was shown in Figure 2. Under **A0**, the problem reduces to finding a sequence $\lambda^*(\tau)$ and a single vector $s^* \in \mathcal{S}_{\theta,\mathcal{G}}$. In fact, it is not even necessary to know the precise function $\lambda^*(\tau)$: For any $\lambda > 0$, so long as the solution $\bar{\beta}(\lambda)$ to $(P2)$ with $\lambda^*(\tau)$ replaced by λ and $s^*(\tau) = s^*$, satisfies for $\tau \triangleq \left\|\mathrm{diag}(s^*)\bar{\beta}(\lambda)\right\|_1$ that $\tau \in (0, \tau_{\max})$, we know that $\lambda = \lambda^*(\tau)$. Thus, under **A0** we only seek to find the vector $s^* \in \mathcal{S}_{\theta,\mathcal{G}}$ so that solving $(P1)$ is for some λ equivalent to solving the problem

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

This paper makes two main contributions. The first contribution in Section 4 proves that if **A0** holds, then there is an algorithm, RepLasso, which (effectively) greedily estimates the vector s^* making (S) and $(P1)$

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 $\lambda > 0$

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

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

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

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

return $\bar{\beta}$

equivalent while sweeping out $\lambda > 0$ and producing solutions $\bar{\beta}(\lambda)$ in a homotopy-like fashion. Of course, if **A0** does not hold, there may not be an equivalence between (S) and $(P1)$. In that case, we may think of (S) as a convex surrogate for $(P1)$ for some vector s^* that is greedily constructed by RepLasso. The second contribution of this paper is to prove in Section 5 that, whether **A0** holds or not, RepLasso will in relevant regression problems still perform at least as well as the Lasso in terms of (signed) support recovery. Empirical evidence in Section 6 and the Supplementary Material shows that a strict improvement can be achieved.

4 RepLasso: A Homotopy Method

A key observation for the development of RepLasso is the following proposition, which shows that it is in principle sufficient to incrementally estimate s^* while simultaneously sweeping out a regularization path. For a vector $b > \mathbf{0}$, let $\bar{\beta}_b(\lambda)$ be a solution to (S) with penalty $\lambda \|\text{diag}(b)\beta\|_1$.

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 $\bar{\beta}_a(\lambda) = \bar{\beta}_b(\lambda)$.*

Thus, if $\bar{\beta}_b(\lambda)$ has zero coefficients, it doesn't matter if on those coefficients b underestimates the value of a, so long as b matches a on the remaining coefficients.

The RepLasso algorithm (Algorithm 1) is a greedy homotopy method that exploits Proposition 3 to solve (S). If X is absolutely continuous and **A0** holds, then Proposition 3 suggests the existence of a sequence $s(\lambda)$, satisfying $\forall \lambda > 0, s^* \geq s(\lambda) \geq \mathbf{1}$, so that w.p. 1 (S) can $\forall \lambda > 0$ be solved as $\bar{\beta}(\lambda) \triangleq \bar{\beta}_{s^*}(\lambda) = \bar{\beta}_{s(\lambda)}(\lambda)$. As Theorem 1 shows, RepLasso computes such a sequence $s(\lambda)$, while simultaneously producing solutions $\bar{\beta}_{s(\lambda)}(\lambda)$. Notice that RepLasso is identical to the Lasso homotopy method if $\theta = \mathbf{0}$ (which means that $\forall \lambda > 0, s(\lambda) = \mathbf{1}$). The only differences are that $s(\lambda) \neq \mathbf{1}$ when $\theta \neq \mathbf{0}$. We will discuss RepLasso as proof for Theorem 1. Let X_j be the column j of X and X_A a matrix which consists of the columns indexed by A.

Theorem 1 (RepLasso)**.** *Assume that X has absolutely continuous distribution and that* **A0** *holds. Let $s^* \in \mathcal{S}_{\theta,\mathcal{G}}$ be the vector so that $(P1)$ is equivalent to (S). Then w.p. 1, RepLasso produces a sequence $s(\lambda)$ so that $\bar{\beta}_{s^*}(\lambda) = \bar{\beta}_{s(\lambda)}(\lambda)$. By the equivalence of $(P1)$ and (S), it follows that w.p. 1, RepLasso produces the global minima of $(P1)$.*

Proof. Note from Proposition 3 that it is sufficient for RepLasso to estimate sequences $s(\lambda)$ which are *piecewise constant* with changepoints at values λ_t where the support of $\bar{\beta}_{s^*}(\lambda_t)$ changes. By **A0**, we know that the support of $\bar{\beta}_{s^*}(\lambda)$ is monotonically increasing with λ decreasing. Hence, we only need to discuss the variable addition case of RepLasso (case $L = 0$ in stage 1) for this argument. Conceptually, RepLasso first initializes $s(\infty) = \mathbf{1}$ (for practical reasons it suffices to start at $\lambda = \|y^\top X\|_\infty$). Then, while keeping $s(\lambda) = s(\infty)$ constant, RepLasso (conceptually) traces out $\lambda = \infty \downarrow 0$ while solving $\bar{\beta}_{s(\lambda)}(\lambda) = \mathbf{0}$ until reaching $\lambda_1 = \|y^\top X\|_\infty$, where the first variable i_1^* is selected by $\bar{\beta}_{s(\lambda)}(\lambda)$ (the $L = 0$ case in stage 1 of RepLasso). Because $s(\lambda) = \mathbf{1}$ was up to now fixed, RepLasso is up to this point identical to the Lasso homotopy method. Due to Proposition 3, we know that w.p. 1, $\forall \lambda \in [\lambda_1, \infty]$ we have $\bar{\beta}_{s^*}(\lambda) = \bar{\beta}_{s(\lambda)}(\lambda)$. Under **A0**, we know that $\forall 0 < \lambda \leq \lambda_1$, i_1^* will remain selected and that the relative order of i_1^* in the set of variables $G_{\Gamma(i_1^*)}$, as induced by the magnitude of their coefficients in $\bar{\beta}_{s(\lambda_1)}(\lambda_1)$ will not change. Using this and the general form of $s^* \in \mathcal{S}_{\theta,\mathcal{G}}$ given by Proposition 1, we can modify $s(\lambda)$ in a way that is consistent with Proposition 3. Specifically, if we let $t = 1$, then the current active set is $A_t = \{i : \left| X_i^\top (y - X\bar{\beta}_{s(\lambda_t)}(\lambda_t)) \right| / s_i(\lambda_t) =$

$\lambda_t\}$. We may apply the following generic update to $s(\lambda)$ so that at λ_t^- (i.e., for a value of λ infinitesimally smaller than λ_t) it satisfies

$$s_j(\lambda_t^-) = \begin{cases} s_j(\lambda_t) + \frac{\theta_{\Gamma(i_t^*)}}{|G_{\Gamma(i_t^*)}|-1} & j \in \{A_t^c \cap G_{\Gamma(i_t^*)}\} \\ s_j(\lambda_t) & \text{o.w.} \end{cases}. \quad (6)$$

Notice that the change leaves the path $\bar{\beta}_{s(\lambda_t)}(\lambda_t)$ continuous in the neighborhood of λ_t. RepLasso then continues to decrease $\lambda = \lambda_1 \downarrow 0$, again keeping $s(\lambda) = s(\lambda_1^-)$ constant and producing solutions $\bar{\beta}_{s(\lambda)}(\lambda)$ along the way, until a point $\lambda_2 > 0$ is reached when a new variable is selected by $\bar{\beta}_{s(\lambda)}(\lambda)$. Because $s(\lambda)$ was kept constant for $\lambda \in [\lambda_2, \lambda_1^-]$, this can be achieved by a straightforward modification of the Lasso homotopy method[3]. As before, we know from our update of $s(\lambda)$ and Proposition 3 that w.p. 1, $\forall \lambda \in [\lambda_2, \lambda_1]$ we have $\bar{\beta}_{s^*}(\lambda) = \bar{\beta}_{s(\lambda)}(\lambda)$. At this point, **A0** and Proposition 1 again allow us to update $s(\lambda)$ using Eq. (6) with $t = 2$. RepLasso continues sweeping out λ in this fashion until some final value $\lambda_T > 0$ is reached. By the time the algorithm has completed, we know that w.p. 1, $\forall \lambda \in [\lambda_T, \infty]$ we have $\bar{\beta}_{s^*}(\lambda) = \bar{\beta}_{s(\lambda)}(\lambda)$. The final claim follows immediately. □

If **A0** does not hold, we can apply Proposition 3 to $(P2)$ to see that RepLasso will generally still recover global minima of $(P1)$ for large $\lambda > 0$. Indeed, if RepLasso adds variables one by one, the first variable selected by RepLasso is also the first selected by $(P1)$.

5 Comparing RepLasso and Lasso

In this section we show several results irrespective of whether **A0** holds, but assuming that $\mathcal{G}$, β^* satisfy some mild conditions. Before continuing, we briefly outline some more notation. Let the support set of β^* be $S \triangleq S(\beta^*)$. Denote the signed support of β^* by $S_\pm = S_\pm(\beta^*)$, where element-wise we have

$$S_\pm(\beta_i) \triangleq \begin{cases} +1 & \text{if } \beta_i > 0 \\ -1 & \text{if } \beta_i < 0 \\ 0 & \text{o.w.} \end{cases}. \quad (7)$$

We rely on 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 [5], $\nexists j \in A^c, |X_A^\top u_A| = |X_j^\top u_A|\mathbf{1}$

Assumption **A1** formalizes that β^* is nonzero on at most a few elements (in this case one) of each group of $\mathcal{G}$. Assumption **A2** ensures that the active set of the Lasso homotopy method in [5] matches the support set. This condition is mild and holds, e.g., w.p. 1 if X has spherical and absolutely continuous distribution.

Many analyses of the Lasso focus on its (signed) support recovery properties. The following theorem, which we prove in the Supplementary Material, will allow us to easily compare RepLasso against Lasso in terms of these measures. Let $\hat{\beta}(\lambda)$ and $\bar{\beta}(\lambda)$ be the Lasso and RepLasso solutions to a regression problem using penalty parameter λ.

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

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

On the other hand, if the conditions of Theorem 2 do not hold, then for large θ, $\exists \lambda \geq \lambda_{\min}, \hat{\beta}(\lambda) \neq \bar{\beta}(\lambda)$ in general. Many consequences for the support recovery behavior of RepLasso can be derived from Theorem 2. The following are two example corollaries that follow. With a slight abuse of notation, let the $\subseteq$ notation applied to signed vectors denote that a correctly signed subset of the signed support is recovered. We have

Corollary 1 (Support Recovery)**.** *Assume that* **A1**–**2** *hold. Conditioned on X, y, we have for any $\lambda_{\min} > 0$*

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

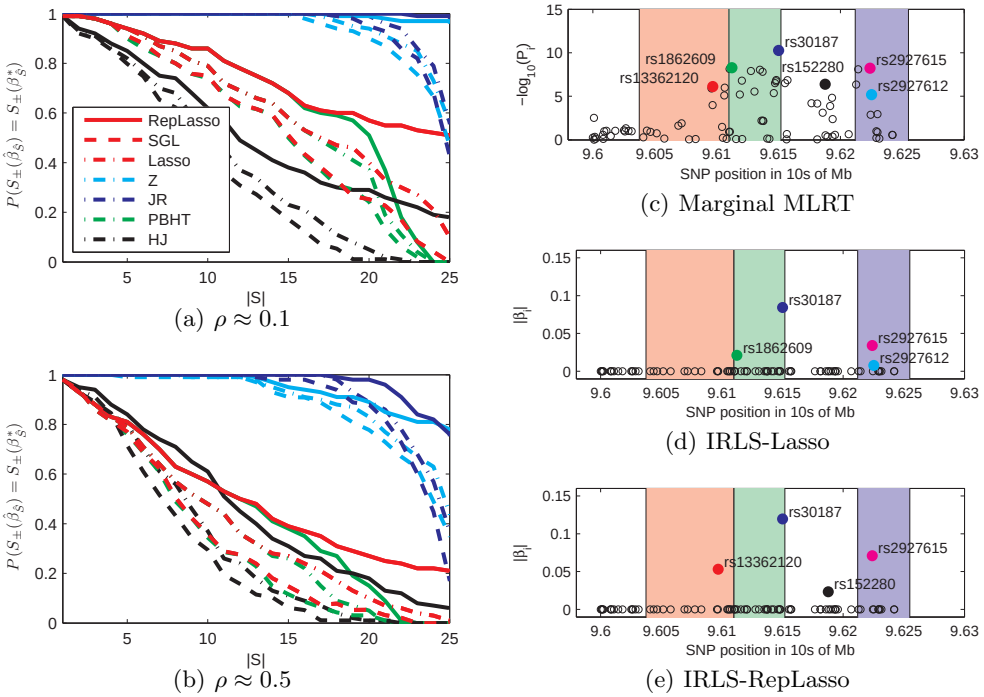
$$\forall \lambda \geq \lambda_{\min} S_\pm(\hat{\beta}(\lambda)) \subseteq S_\pm \implies \forall \lambda \geq \lambda_{\min} S_\pm(\bar{\beta}(\lambda)) \subseteq S_\pm.$$

Corollary 1 shows that whenever the Lasso recovers the (signed) support, so does RepLasso. That is, under **A1**–**2**, RepLasso cannot perform worse than the Lasso in terms of (signed) support recovery. In Section 6 and the Supplementary Material we show empirically that RepLasso often strictly outperforms the Lasso.

Consequences for other Methods. Besides the Lasso, Theorem 2 also applies to many related algorithms that pre-process the data X, y in some way, prior to running the Lasso on the modified data. Instances of these algorithms are, for example, the Adaptive Lasso [31] and various Preconditioned Lasso algorithms [9, 12, 20]. Indeed, if the assumptions hold, the results are even true for ℓ_1 regularized minimization of quadratic approximations to logistic regression as proposed in [15]. We will empirically highlight this property in Section 6 and the Supplementary Material.

A Lars-like Variation. We note at this point that the Lars algorithm [5] is a special case of RepLasso if we set $\theta = \mathbf{0}$ and force $L = 0$. If we only force $L = 0$ but allow $\theta \neq \mathbf{0}$, then the resulting algorithm can be seen as a generalization of Lars. We analyze this method in the Supplementary Material and show similar support recovery behavior.

[3] Specifically, where the Lasso homotopy method traces out *equiangular* directions, the RepLasso follows *skew-angular* directions (given in stage 2), with the angle skew determined by the weights $s_A(\lambda)$.



(a) $\rho \approx 0.1$

(b) $\rho \approx 0.5$

(c) Marginal MLRT

(d) IRLS-Lasso

(e) IRLS-RepLasso

Figure 3: (a), (b) Results on synthetic data with $n = 150, p = 50, \mathcal{G} =$ groups of 2 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). (c) A GWAS Manhattan plot for ankylosing spondylitis on a region of chromosome 5. The x-axis gives SNP position; circles indicate the $-\log_{10}(p)$-values of marginal association tests. The red, green and blue shaded regions indicate the CAST, ERAP1 and ERAP2 genes respectively. Solid circles highlight the SNPs that were chosen by the methods below. (d) The IRLS method of Lee et al. [15] with Lasso. Circles indicate the magnitudes of estimated β coefficients. Solid circles indicate the first 4 SNPs that are chosen by Lasso. The method selects multiple SNPs from the same gene. (e) The RepLasso avoids this if θ is chosen large enough.

Computational Complexity. Under assumption **A0**, RepLasso (and RepLars in the Supplementary Material) strongly resemble the Lars algorithm [5]. The principal difference is that we now maintain a set of weights $s(\lambda)$ in stage 1. This bookkeeping increases the runtime of stage 1 by at most a constant factor, as we already need to compute the residual correlations of all inactive variables. It follows that, under assumption **A0**, RepLasso and RepLars have the same computational complexity as the Lars algorithm, that is, $O(p^3 + np^2)$ if $p < n$ and $O(n^3 + n^2 p)$ if $n \gg p$ (see discussion in [5]). If **A0** does not hold, then RepLasso may drop variables and consequently the runtime may increase. While the Lasso regularization path (i.e., RepLasso with $\theta = \mathbf{0}$) can contain up to $O(3^p)$ linear segments [18], the empirical complexity of the Lasso homotopy (in this case also RepLasso) is often considered to be the same as above [1]. We hope that this behavior carries over to RepLasso more generally.

6 Results

In this section we provide some comparisons of RepLasso against implementations of Lasso [23] and SGL [16], and validate the findings of Section 5. Since SCAD [7] and MCP [29] do not take advantage of the partition $\mathcal{G}$, we will not compare against them.

Synthetic Data. We first focus on a set of experiments which analyzes the probability of correctly recovering a subset of the correct *signed* support. We fix $\mathcal{G}, \beta^*$ so that **A1** holds. Conditioned on $\mathcal{G}$ we also sample X with unit-length columns that are independent between groups $G \in \mathcal{G}$ but exhibit some correlation ρ within groups. Given X, β^*, we generate y according to Eq. (1), with $\sigma^2 = 0.2^2$. In Figures 3(a) and 3(b) we investigate the performance of the RepLasso (solid red), the Lasso (dashed-dotted red) and SGL (dashed red). As there is no homotopy method for SGL we generated the results by probing along the regulari-

sation path until a model of desired complexity was found. Notice that the curve for RepLasso lies above that of Lasso, giving empirical support to Theorem 2 and Corollary 1. Also, the curve for the Sparse Group Lasso (SGL) [22] lies below that of the Lasso and RepLasso. We believe this is because SGL is for almost all parameter settings less sparse than the Lasso (see Figure 1(c)) and because the SGL penalty does not exploit **A1**. In addition to Lasso and SGL, we also 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 [31] (*Z*); (2) the "Whitened" Lasso of Jia and Rohe [12] (*JR*); (3) the Preconditioned Lasso of Paul et al. [20] (*PBHT*); and (4) Correlation Sifting of Huang and Jojic [9] (*HJ*). The results in Figures 3(a) and 3(b) highlight that using RepLasso as a drop-in replacement these algorithms can also be improved, while using SGL as drop-in replacement generally does not help. As we show in the Supplementary Material, results are similar when groups are larger and $n \ll p$.[4]

GWAS application. A second experiment considers the application to a Genome-Wide Association Study (GWAS). A GWAS hopes to find Single Nucleotide Polymorphisms (SNPs) that are associated with disease status. Our focus is on $n = 4000$ cases and controls for the disease ankylosing spondylitis[5] and a region on chromosome 5 spanning $p = 84$ SNPs, where susceptibility SNPs [3] had been previously reported. Mainstream GWAS methodology tests each SNP *marginally* for association using a maximum likelihood ratio test (MLRT) and plots the resulting p-values on a "Manhattan plot", as in Figure 3(c). Due to linkage disequilibrium, many small p-values lie close to each other. Alternatively, a penalized logistic regression could also be used to regress the SNPs onto disease status, which would then highlight interesting SNPs by the magnitudes of the learned regression coefficients. Lee et al. [15] have proposed an IRLS strategy for estimating an ℓ_1 constrained logistic regression by solving a Lasso problem on a quadratic approximation of the logistic objective. The magnitudes of the first four regression coefficients estimated by this method are shown in Figure 3(d). As can be seen, two pairs of selected SNPs lie near each other in two genes. We might wish to discourage the Lasso from choosing multiple SNPs from the same gene. Unlike SGL, the RepLasso is ideally suited to this task. Given a gene partition $\mathcal{G}$ (here by CAST, ERAP1 and ERAP2 genes) we can replace the Lasso in the IRLS algorithm by the RepLasso and produce a different parameter estimate. If θ is large (e.g., 20 for each group), RepLasso avoids selecting multiple SNPs from the same gene, as seen in Figure 3(e). These SNPs may be worthy of further study.

[4] If $n < p$ we let the Adaptive Lasso scale columns of X by univariate regression coefficients.

[5] See [6]. The EGA accession numbers for the data are EGAD00010000150 and EGAD00000000022.

7 Conclusion

In this paper we presented a greedy homotopy algorithm that approximates an underlying nonconvex problem by a suitable sequence of surrogates that locally approximate $\Omega_{\theta,\mathcal{G}}(\cdot)$ well. As shown by Theorem 1, our method will in certain cases sweep out a global minima path of $(P1)$. Theorem 2 and Corollary 1 showed that even though RepLasso may not exactly solve $(P1)$ in general, in relevant regression problems RepLasso will not do worse than the Lasso in terms of (signed) support recovery. Finally, Section 6 shows that RepLasso often outperforms Lasso.

Several extensions can be considered. Firstly, we defined $\Omega_{\theta,\mathcal{G}}(\cdot)$ as a sum over certain pairs of variables. More flexible constraint functions could potentially be defined if the sum is allowed to range over an arbitrary set of pairs. Secondly, our overall strategy was to decompose the nonconvex constraint balls induced by $\Omega_{\theta,\mathcal{G}}(\cdot)$ as a union of simpler, convex balls. This motivates directly defining nonconvex constraint balls as a union of convex balls. For instance, one could consider unions of weighted ℓ_∞ balls or a mix of weighted ℓ_∞ and weighted ℓ_1 balls. So long as these convex building blocks are consistent with [21] it should still be possible to efficiently compute local minima paths segments as demonstrated in this paper. Thirdly, it would be interesting to see whether results in, e.g., [17] can be extended to argue for consistency of the RepLasso in cases where local minima paths are produced.

Acknowledgments

We thank Francis Bach and Alexander Young for helpful comments and Nebojsa Jojic for early input. This work was supported by the Wellcome Trust [090532/Z/09/Z] and [095552/Z/11/Z]. This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113, 085475 and 090355.

References

[1] F.R. Bach. Bolasso: model consistent Lasso estimation through the bootstrap. In *Proc. Int. Conf. Mach. Learn.*, pages 33–40. ACM, 2008.

[2] S.P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[3] P.R. Burton et al. Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. *Nature genetics*, 39(11):1329–1337, 2007.

[4] E.J. Candès, M.B. Wakin, and S.P. Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *J. Fourier Anal. Appl.*, 14(5-6):877–905, 2008.

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

[6] D.M. Evans et al. Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. *Nature genetics*, 43(8):761–767, 2011.

[7] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96:1348–1360, 2001.

[8] T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu. The entire regularization path for the Support Vector Machine. *J. Mach. Learn. Res.*, 5:1391–1415, 2004.

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

[10] D.R. Hunter and R. Li. Variable selection using MM algorithms. *Ann. Stat.*, 33(4):1617, 2005.

[11] R. Jenatton, J.-Y. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. *J. Mach. Learn. Res.*, 12:2777–2824, 2011.

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

[13] S. Kim and E.P. Xing. Tree-guided Group Lasso for multi-response regression with structured sparsity, with applications to eQTL mapping. *Ann. Appl. Stat.*, 2012.

[14] M. Kowalski and B. Torrésani. Sparsity and persistence: mixed norms provide simple signal models with dependent coefficients. *Signal, Image and Video processing*, 3(3):251–264, 2009.

[15] S.-I. Lee, H. Lee, P. Abbeel, and A.Y. Ng. Efficient ℓ_1 regularized logistic regression. In *Proc. Conf. Artif. Intell.*, volume 21, page 401, 2006.

[16] J. Liu, S. Ji, and J. Ye. *SLEP: Sparse Learning with Efficient Projections*. ASU, 2009. Ver. 4.1.

[17] P.-L. Loh and M.J. Wainwright. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *Adv. Neur. Inf. Process. Syst. 26*, pages 476–484. 2013.

[18] J. Mairal and B. Yu. Complexity analysis of the Lasso regularization path. *arXiv preprint arXiv:1205.0079*, 2012.

[19] M.R. Osborne, B. Presnell, and B.A. Turlach. A new approach to variable selection in least squares problems. *J. Numer. Anal.*, 20(3):389–403, 2000.

[20] 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.

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

[22] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. A Sparse-Group Lasso. *J. Comp. Graph. Stat.*, 22(2):231–245, 2013.

[23] K. Sjöstrand. Matlab implementation of LASSO and LARS, 2005. Ver. 2.0.

[24] R. Tibshirani. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B*, 58(1):267–288, 1994.

[25] L. Wang, G. Chen, and H. Li. Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics*, 23(12):1486–1494, 2007.

[26] Z. Wang, H. Liu, and T. Zhang. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Ann. Stat.*, 42(6):2164–2201, 2014.

[27] L. Xiao and T. Zhang. A proximal-gradient homotopy method for the ℓ_1-regularized least-squares problem. In *Proc. Int. Conf. Mach. Learn.*, pages 839–846, 2012.

[28] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc., Ser. B*, 68(1):49–67, 2006.

[29] C.-H. Zhang. Nearly unbiased variable selection under Minimax Concave Penalty. *Ann. Stat.*, 38(2):894–942, 2010.

[30] Y. Zhou, R. Jin, and S. Hoi. Exclusive Lasso for multi-task feature selection. In *Artificial Intelligence and Statistics*, pages 988–995, 2010.

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

[32] H. Zou and T. Hastie. Regularization and variable selection via the Elastic Net. *J. R. Stat. Soc., Ser. B*, 67:301–320, 2005.

[33] H. Zou and R. Li. One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Stat.*, 36(4):1509, 2008.