Supplementary materials for this article are available online.
Please go to www.tandfonline.com/r/JCGS

ConvexLAR: An Extension of Least Angle Regression

Wei XIAO, Yichao WU, and Hua ZHOU

The least angle regression (LAR) was proposed by Efron, Hastie, Johnstone and Tibshirani in the year 2004 for continuous model selection in linear regression. It is motivated by a geometric argument and tracks a path along which the predictors enter successively and the active predictors always maintain the same absolute correlation (angle) with the residual vector. Although it gains popularity quickly, its extensions seem rare compared to the penalty methods. In this expository article, we show that the powerful geometric idea of LAR can be generalized in a fruitful way. We propose a ConvexLAR algorithm that works for any convex loss function and naturally extends to group selection and data adaptive variable selection. After simple modification, it also yields new exact path algorithms for certain penalty methods such as a convex loss function with lasso or group lasso penalty. Variable selection in recurrent event and panel count data analysis, Ada-Boost, and Gaussian graphical model is reconsidered from the ConvexLAR angle. Supplementary materials for this article are available online.

Key Words: Group lasso; Lasso; Ordinary differential equation (ODE); Regularization; Solution path.

1. INTRODUCTION

Regularization is a tool to avoid overfitting and obtain parsimonious and interpretable models, especially when the number of parameters exceeds the number of observations. One powerful regularization technique is penalization. In general, a penalty method minimizes the sum of a loss function and a penalty term. The simplest ℓ_1 penalty leads to the popular lasso regression (Donoho and Johnstone 1994; Tibshirani 1996). Various other penalty methods have been developed thereafter. Each one targets on a specific question that arises in applications. For instance, the group penalty (Yuan and Lin 2006; Meier, van de Geer, and Bühlmann 2008) aims to select groups of variables such as in factorial data analysis. The adaptive lasso (Zou 2006) applies a weighted penalty method in a data-driven fashion that improves the asymptotic properties. All these penalty methods are formulated as an optimization problem and the solution is obtained by either optimizing at a grid of tuning

Wei Xiao, is Ph.D. candidate (E-mail: *wxiao@ncsu.edu*), Yichao Wu is corresponding author and Associate Professor (E-mail: *wu@stat.ncsu.edu*), and Hua Zhou is Assistant Professor (E-mail: *hzhou3@ncsu.edu*), Department of Statistics, North Carolina State University, Raleigh, NC 27695.

Journal of Computational and Graphical Statistics, Volume 24, Number 3, Pages 603–626
DOI: 10.1080/10618600.2014.962700
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/jcgs.

parameter values or by a path algorithm that tracks the solution as a function of the tuning parameter.

In contrast to penalty methods, the least angle regression (LAR) proposed by Efron et al. (2004) is purely motivated by a geometric argument rather than optimization. Given a response vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\mathsf{T} \in \mathbb{R}^n$ and its corresponding design matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\mathsf{T} \in \mathbb{R}^{n\times p}$ with $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\mathsf{T}$, let $\mathcal{A}_t$ be the active index set at time t. Following Efron et al. (2004), we assume that the covariates $\boldsymbol{x}^{(j)} = (x_{1j}, x_{2j}, \ldots, x_{nj})^\mathsf{T}$ have mean 0 and unit length, and that the response $\boldsymbol{y}$ has mean 0. At $t = 0$, $\beta(0) = \mathbf{0}_p$ and $\mathcal{A}_0$ contain the predictor that is most correlated with the response vector $\boldsymbol{y}$. At any $t > 0$, regression coefficients of active predictors, namely $\beta_j(t)$ for $j \in \mathcal{A}_t$, move along a direction such that their corresponding predictor vectors $\boldsymbol{x}^{(j)}$ share the same absolute correlation (angle) with the residual vector $\boldsymbol{e}(t) = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}(t)$. Here the correlation is nothing but the scaled score vector of the least squares criterion with entries

$$-\frac{1}{2}\frac{\partial}{\partial\beta_j}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}\right)^2\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}(t)}.$$

LAR has gained wide popularity since its introduction. However, there seem only a few attempts for generalizations, strikingly unparallel to the fast development of penalty methods. Specific versions of group LAR for least squares problem are mentioned in Yuan and Lin (2006) and Park and Hastie (2006). Wu (2011) extended LAR to the generalized linear models and the Cox's proportional hazard model (Wu 2012). In this article, we demonstrate that the powerful geometric idea of LAR can be generalized in a fruitful way leading to potentially many more applications.

The remaining of the article is organized as follows. In Section 2, we derive a basic ConvexLAR algorithm that performs continuous variable selection for any general convex loss functions. For least squares loss, Efron et al. (2004) showed that the LAR solution path is piecewise linear. This leads to their efficient path following algorithm with computational cost of a single ordinary least squares estimation. For a general loss function, the piecewise linearity property is lost. However, it can be shown that the solution path is piecewise smooth and, within each path segment, follows a simple ordinary differential equation (ODE). ConvexLAR tracks the solution path by using the rich numerical resource for solving ODE. Just like the original LAR for least squares, a simple modification of ConvexLAR yields the corresponding lasso solution path.

In Section 3, we show that the geometric idea of LAR can be adapted to various situations. We demonstrate this by incorporating data adaptive weights and group selection into the ConvexLAR algorithm. Comparing to their penalization analogs, these extensions avoid repeated optimizations and are computationally attractive. Moreover, a slight modification of ConvexLAR yields the exact solution path for the corresponding penalization method. ConvexLAR and its extensions are illustrated by various numerical examples in Section 4. To the best of our knowledge, no LAR algorithms have been proposed for these examples in the current literature. Finally we conclude with a brief discussion.

2. CONVEXLAR ALGORITHM AND CONVEXLASSO MODIFICATION

In this section, we derive the ConvexLAR algorithm which forms the basis of various extensions presented in the next section. The algorithm is similar to the LAR algorithms developed for GLM and Cox model (Wu 2011, 2012) but with much more generality and simpler derivation. We consider an arbitrary strictly convex loss function $f(\boldsymbol{\beta})$, where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of parameters subject to regularization. $\nabla f(\boldsymbol{\beta}) = [\nabla_1 f(\boldsymbol{\beta}), \ldots, \nabla_p f(\boldsymbol{\beta})]^\top \in \mathbb{R}^p$ denotes the gradient vector of the loss function and $\boldsymbol{H}(\boldsymbol{\beta}) = d^2 f(\boldsymbol{\beta}) \in \mathbb{R}^{p\times p}$ the Hessian, where $\nabla_j f(\boldsymbol{\beta})$ denotes the partial derivative of $f(\boldsymbol{\beta})$ with respect to β_j for $j = 1, \ldots, p$. When f is a negative log-likelihood function, $\nabla f(\boldsymbol{\beta})$ equals the negative of the score vector and $\boldsymbol{H}(\boldsymbol{\beta})$ is the observed information matrix. We use t to index the LAR solution path, with the solution at any t denoted by $\boldsymbol{\beta}(t)$. The active index set at t is denoted by $\mathcal{A}_t$. For notational simplicity, we will drop the subscript t whenever it is obvious from the context. For instance, $\boldsymbol{\beta}_{\mathcal{A}}(t)$ and $\nabla_{\mathcal{A}} f(\boldsymbol{\beta})$ are the subvectors of $\boldsymbol{\beta}(t)$ and $\nabla f(\boldsymbol{\beta})$ corresponding to active predictors at t, respectively. Similarly, $\boldsymbol{H}_{\mathcal{A}}(\boldsymbol{\beta}(t))$ is the submatrix of the Hessian corresponding to $\mathcal{A}_t$.

The key idea of LAR (Efron et al. 2004) is to move the solution in a direction such that the gradient (score) corresponding to each active predictor variable has the same absolute value. We denote this common value by $s(t)$, where s stands for the score. This prescribes that the active solution vector has to satisfy $|\nabla_j f(\boldsymbol{\beta})| = \operatorname{sgn}(\nabla_j f(\boldsymbol{\beta})) \cdot \nabla_j f(\boldsymbol{\beta}) = s(t)$ or equivalently

$$\nabla_j f(\boldsymbol{\beta}) - \operatorname{sgn}(\nabla_j f(\boldsymbol{\beta}))s(t) = 0, \quad j \in \mathcal{A}_t.$$

Note that, for any active predictor j, $\nabla_j f(\boldsymbol{\beta})$ is nonzero and, therefore, has a constant sign, denoted by $\operatorname{sgn}(\nabla_j f(\boldsymbol{\beta}))$, within a segment. In vector form, we have

$$\nabla_{\mathcal{A}} f(\boldsymbol{\beta}) - \operatorname{sgn}(\nabla_{\mathcal{A}} f(\boldsymbol{\beta}))s(t) = \mathbf{0}. \tag{1}$$

In general $s(t)$ can be any smooth and monotonic function that decreases from $s(0) = \max_j |\nabla_j f(\mathbf{0})|$ to $s(t_{\max}) = 0$ at some finite $t_{\max} > 0$. Intuitively $s(t)$ controls how the common absolute score of active predictors decays with respect to solution index t. Different choices of $s(t)$ lead to different indexing systems yet the same solution path. The classical LAR (Efron et al. 2004) sets $s(t) = s(0) - t$ which implies that $t_{\max} = s(0) = \max_j |\nabla_j f(\mathbf{0})|$.

By construction $s(t)$ is larger than the absolute value of the scores of inactive predictors which are packed at zero. Once the absolute score $|\nabla_j f(\boldsymbol{\beta}(t))|$ of an inactive predictor j coincides with $s(t)$, it joins the club and its score conforms to the ruling $s(t)$ thereafter. Whenever such an event happens, the set of active predictors is updated by adding this new predictor. The index location t corresponding to this event defines a transition point in the sense that the set of active predictors changes (Wu 2011). A path segment is defined as the solution path between two consecutive transition points. The following result follows from the implicit function theorem and provides the path following direction of the

ConvexLAR algorithm within a path segment. See the Appendix for the proof and the following subsection for the definition of path segment operationally.

Theorem 1. For a strictly convex and twice differentiable loss function $f(\boldsymbol{\beta})$, the LAR path solution $\boldsymbol{\beta}(t)$ is continuous and differentiable at t within a path segment. In addition, the solution vector $\boldsymbol{\beta}(t)$ satisfies the ordinary differential equation

$$\frac{d}{dt}\beta_{\mathcal{A}}(t) = s'(t)\boldsymbol{H}_{\mathcal{A}}^{-1}(\beta)\,\mathrm{sgn}(\nabla_{\mathcal{A}} f(\beta(t))) \tag{2}$$

and $\beta_j(t) = 0$ for any $j \notin \mathcal{A}_t$.

2.1 CONVEXLAR ALGORITHM

Theorem 1 suggests that the exact solution path of LAR can be obtained by solving the simple ODE system (2) segment by segment. The size of the ODE system within a segment is equal to the corresponding number of active predictors $|\mathcal{A}|$. The ConvexLAR algorithm is summarized in Algorithm 5. We initialize our solution path with $\boldsymbol{\beta}(0) = \mathbf{0}$ and the beginning active set contains the predictors that change the objective function $f(\mathbf{0})$ fastest. That is

$$\mathcal{A}_0 = \{\mathrm{argmax}_j |\nabla_j f(\mathbf{0})|\}.$$

We then follow the solution path by solving the ODE system (2) until one or more new variables join the active set at some $t_1 > 0$, which is determined by the moment the active score $s(t)$ matches the maximum (absolute) gradient of one or more nonactive predictors. The active predictor set $\mathcal{A}_t$ stays the same within $t \in [t_0, t_1)$. At t_1, it is updated by adding the new predictors that newly join the club. This process continues segment by segment until all the predictors are active. Then, in the final segment, the ConvexLAR solution path moves along a direction such that the absolute values of the first-order partial derivatives decrease at the same speed to zero, which happens at $t_{\max}$. Under assumptions of Theorem 1, the solution $\boldsymbol{\beta}(t_{\max})$ is the global minimizer of the convex loss function, just like the LAR solution ends at the full ordinary least squares estimate. This completes our ConvexLAR solution path algorithm.

1 Initialize: $s(0) = \max_j |\nabla_j f(\mathbf{0})|$, $\boldsymbol{\beta}(0) = \mathbf{0}$, and $\mathcal{A} = \mathrm{argmax}_j |\nabla_j f(\mathbf{0})|$;
2 **repeat**
3 | Solve ODE
$$\frac{d}{dt}\boldsymbol{\beta}_{\mathcal{A}}(t) = s'(t)\boldsymbol{H}_{\mathcal{A}}^{-1}(\boldsymbol{\beta}(t))\mathrm{sgn}(\nabla_{\mathcal{A}} f(\boldsymbol{\beta}(t))$$
| until $|\nabla_j f(\boldsymbol{\beta}(t))| = s(t)$ for some $j \notin \mathcal{A}$;
4 | Update set $\mathcal{A} \leftarrow \mathcal{A} \cup \{j : |\nabla_j f(\boldsymbol{\beta}(t))| = s(t), j \notin \mathcal{A}\}$
5 **until** $s(t) = 0$ *or* $\boldsymbol{H}_{\mathcal{A}}(\boldsymbol{\beta}(t))$ *is singular* ;

Algorithm 1: ConvexLAR.

2.2 REMARK

In this section, we make the following four remarks on the ConvexLAR algorithm.

Remark 1. For the specific choice of $s(t) = s(0) - t$,

$$t_1 = s(0) - s(t_1) = \left(\max_{j'} |\nabla_{j'} f(\mathbf{0})| \right) - |\nabla_j f(\boldsymbol{\beta}(t_1))|$$

for any $j \in \mathcal{A}_{t_1}$. This holds analogously for later transition points. At the end of the mth LAR segment, the transition point

$$t_m = \left(\max_{j'} |\nabla_{j'} f(\mathbf{0})| \right) - |\nabla_j f(\boldsymbol{\beta}(t_m))|$$

for any $j \in \mathcal{A}_{t_m}$.

Remark 2. For least squares problems, the loss is a quadratic function $f(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2/2$ with a constant Hessian matrix $\boldsymbol{H}(\boldsymbol{\beta}) = \boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}$. Since $\operatorname{sgn}(\nabla_j f(\boldsymbol{\beta}(t)))$ is constant for all $j \in \mathcal{A}_t$ in a neighborhood of t, $\boldsymbol{H}_{\mathcal{A}}^{-1}(\boldsymbol{\beta}) \operatorname{sgn}(\nabla_{\mathcal{A}} f(\boldsymbol{\beta}(t)))$ in (2) is piecewise constant. This leads to the piecewise linear solution path of the original LAR (Efron et al. 2004).

Remark 3 (nonstrictly convex losses). The strict convexity assumption on the loss f precludes applications with nonconvex losses or the $n < p$ least squares case. However, from the proof in Appendix, we observe that the only essential ingredient is the positive definiteness of $\boldsymbol{H}_{\mathcal{A}}(\boldsymbol{\beta}(t))$. Therefore, in nonstrictly convex cases, we terminate path following as soon as $\boldsymbol{H}_{\mathcal{A}}(\boldsymbol{\beta}(t))$ becomes singular.

Remark 4 (partial regularization). So far we have assumed that the full set of parameters $\boldsymbol{\beta}$ are subject to regularization. In many applications, only a subset of parameters are regularized. Assume that the loss takes the form $f(\boldsymbol{\beta}_0, \boldsymbol{\beta})$, where $\boldsymbol{\beta}_0 \in \mathbb{R}^{p_0}$ is the vector of parameters exempt from regularization. Depending on the objective function, it may not be easy to remove $\boldsymbol{\beta}_0$ from the objective. However, we can always define the marginal minimizer of $\boldsymbol{\beta}_0$ as a function of $\boldsymbol{\beta}$

$$\boldsymbol{\beta}_0(\boldsymbol{\beta}) = \operatorname{argmin}_{\boldsymbol{\beta}_0} f(\boldsymbol{\beta}_0, \boldsymbol{\beta}). \tag{3}$$

Assume that f is strictly convex and thrice differentiable, then this mapping is uniquely defined and twice differentiable by the implicit function theorem. Denote the Jacobian and Hessian of this mapping by $D\boldsymbol{\beta}_0(\boldsymbol{\beta}) \in \mathbb{R}^{p_0 \times p}$ and $H\boldsymbol{\beta}_0(\boldsymbol{\beta}) \in \mathbb{R}^{p_0 p \times p}$, respectively. Then the first two derivatives with respect to $\boldsymbol{\beta}$ required in Theorem 1 can be obtained by the chain rule

$$\nabla f(\boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}} f(\boldsymbol{\beta}_0, \boldsymbol{\beta}) + \nabla_{\boldsymbol{\beta}_0} f(\boldsymbol{\beta}_0, \boldsymbol{\beta}) \cdot D\boldsymbol{\beta}_0(\boldsymbol{\beta}),$$
$$Hf(\boldsymbol{\beta}) = d^2_{\boldsymbol{\beta},\boldsymbol{\beta}} f(\boldsymbol{\beta}_0, \boldsymbol{\beta}) + D\boldsymbol{\beta}_0(\boldsymbol{\beta})^{\mathsf{T}} \cdot d^2_{\boldsymbol{\beta}_0,\boldsymbol{\beta}_0} f(\boldsymbol{\beta}_0, \boldsymbol{\beta}) \cdot D\boldsymbol{\beta}_0(\boldsymbol{\beta}) + [d_{\boldsymbol{\beta}_0} f(\boldsymbol{\beta}_0, \boldsymbol{\beta}) \otimes \boldsymbol{I}_p] H\boldsymbol{\beta}_0(\boldsymbol{\beta}).$$

The Ada-Boost and Gaussian graphical model examples in Section 4 illustrate this strategy.

2.3 ConvexLASSO Modification

Efron et al. (2004) showed that in the least squares case the lasso solution path can be obtained by a slight modification of the LAR. Same extension applies to ConvexLAR. Consider the lasso regularized problem

$$\min_{\boldsymbol{\beta}} \quad f(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{4}$$

With $\lambda = s(t) = s(0) - t$, the optimality condition for lasso solution is

$$\begin{aligned} \nabla_j f(\boldsymbol{\beta}) + \lambda \operatorname{sgn}(\beta_j) = 0, \quad & \beta_j \neq 0 \\ |\nabla_j f(\boldsymbol{\beta})| \leq \lambda, \quad & \beta_j = 0. \end{aligned} \tag{5}$$

A proof similar to that for Theorem 1 shows that the lasso solution path moves along the direction

$$\begin{aligned} \frac{d}{dt}\boldsymbol{\beta}_{\mathcal{A}}(t) &= -s'(t)\boldsymbol{H}_{\mathcal{A}}(\boldsymbol{\beta}(t)) \operatorname{sgn}(\boldsymbol{\beta}_{\mathcal{A}}(t)) \\ &= s'(t)\boldsymbol{H}_{\mathcal{A}}(\boldsymbol{\beta}(t)) \operatorname{sgn}(\nabla_{\mathcal{A}} f(\boldsymbol{\beta}(t))) \end{aligned}$$

until either (i) $\beta_j(t)$ hits zero for an active predictor j or (ii) $|\nabla_j f(\boldsymbol{\beta}(t))|$ hits boundary $\lambda = s(t)$ for some inactive predictor j. Both events change the active set and redefine the direction. The second equation is based on the fact that by (5) a lasso regularized solution satisfies $\operatorname{sgn}(\boldsymbol{\beta}_{\mathcal{A}}(t)) = -\operatorname{sgn}(\nabla_{\mathcal{A}} f(\boldsymbol{\beta}(t)))$. Similarly, the ConvexLAR algorithm can be modified to obtain the lasso solution path $\boldsymbol{\beta}(\lambda)$ with $\lambda = s(t)$. Observe that the event defining the LAR segment, that is, the gradient of an inactive predictor hits $\lambda = s(t)$, is exactly the same as the second type of event for lasso path. However, the first type of event, that is, the coefficient of an active predictor hitting zero, is not tracked in ConvexLAR. We call the modified algorithm, which tracks both types of events, by ConvexLASSO and the pseudocode is listed in Algorithm 9. The same argument as in Efron et al. (2004) and Wu (2011) shows that, under the assumption that, at each transition point, only one single event can happen, namely either one inactive predictor variable becomes active or one currently active predictor variable becomes inactive, ConvexLASSO algorithm yields the lasso solution path.

Initialize: $s(0) = \max_j |\nabla_j f(\mathbf{0})|$, $\boldsymbol{\beta}(0) = \mathbf{0}$, and $\mathcal{A} = \text{argmax}_j |\nabla_j f(\mathbf{0})|$
repeat
  Solve ODE

$$\frac{d}{dt}\boldsymbol{\beta}_{\mathcal{A}}(t) = s'(t)\boldsymbol{H}_{\mathcal{A}}^{-1}(\boldsymbol{\beta}(t))\text{sgn}(\nabla_{\mathcal{A}} f(\boldsymbol{\beta}(t))$$

  until (i) $\beta_j(t) = 0$ for some $j \in \mathcal{A}$ or (ii) $|\nabla_j f(\boldsymbol{\beta}(t))| = s(t)$ for some $j \notin \mathcal{A}$;
  if *event (i)* **then**
    Update set $\mathcal{A} \leftarrow \mathcal{A} \setminus \{j : \beta_j(t) = 0, j \in \mathcal{A}\}$
  else if *event (ii)* **then**
    Update set $\mathcal{A} \leftarrow \mathcal{A} \cup \{j : |\nabla_j f(\boldsymbol{\beta}(t))| = s(t), j \notin \mathcal{A}\}$
  end
until $s(t) = 0$ *or* $\boldsymbol{H}_{\mathcal{A}}(\boldsymbol{\beta}(t))$ *is singular* ;

Algorithm 2: ConvexLASSO.

3. GENERALIZATIONS

In this section we first summarize a few essential features of ConvexLAR and then demonstrate how these aspects lead to various generalizations.

1. The "influence" of each predictor on the loss function is measured by the magnitude of its score (gradient). Indeed it is these gradient (score) functions that ConvexLAR operates on. For the ConvexLAR algorithm to work properly, we require the influence function to be a monotone $\mathbb{R}^p \to \mathbb{R}^p$ mapping (Ortega and Rheinboldt 2000). For instance, convexity of a loss function guarantees that its gradient (score) function is a monotone mapping.

2. Certain form of "democratic voting" is enforced among the active predictors. Both the original LAR and ConvexLAR force the influences of individual active predictors to be equal. This equality constraint can be generalized when we want to favor certain predictors over others or to impose group structure on predictors.

3. The influences of active predictors continuously decrease along the path so that those of inactive predictors can catch up. The assumption in L1 that the gradient (score) function must be a monotone mapping guarantees that all influences change continuously.

4. The inactive predictors keep parked at zero until their influence meets that of active ones, at which point they join the club.

5. The influences of all active predictors gradually decrease at the same rate and hit zero at the same time, which declares the end of path following.

3.1 WEIGHTED/ADAPTIVE CONVEXLAR

In many applications, there exists prior information about the importance of predictors, which can also be obtained in a data-driven fashion. This motivates the development of adaptive lasso (Zou 2006), which enjoys favorable asymptotic properties. To

incorporate such information into ConvexLAR, we weight the "influence" of each predictor differentially and consider the weighted "influence" $w_j|\nabla_j f(\boldsymbol{\beta})|$ of each predictor, where $w_j \geq 0$ are predictor specific weights. A larger w_j implies higher "influence" and vice versa. A zero weight means no regularization for the corresponding predictor. For simplicity, we assume that all weights are positive. In this case the function $s(t)$ is the common value $w_j|\nabla_j f(\boldsymbol{\beta})|$ shared by active predictors. Setting $w_j \equiv 1$ reduces to the ConvexLAR. The stationarity condition for the "democratic voting" reads

$$\nabla_j f(\boldsymbol{\beta}) - w_j^{-1}\text{sgn}(\nabla_j f(\boldsymbol{\beta}))s(t) = 0, \quad j \in \mathcal{A},$$

and the ODE system becomes

$$\frac{d}{dt}\boldsymbol{\beta}_{\mathcal{A}}(t) = s'(t)\boldsymbol{H}_{\mathcal{A}}^{-1}(\boldsymbol{\beta})\text{diag}(\boldsymbol{w}_{\mathcal{A}}^{-1})\text{sgn}(\nabla_{\mathcal{A}} f(\boldsymbol{\beta}(t))),$$

where the vector $\boldsymbol{w}_{\mathcal{A}}^{-1}$ collects the inverse weights w_j^{-1} for active predictors. Segment terminates when $w_j|\nabla_j f(\boldsymbol{\beta}(t))|$ hits $s(t)$ for some inactive predictors $j \notin \mathcal{A}$. The weighted ConvexLAR is summarized in Algorithm 5 and a similar modification can be applied to get the corresponding adaptive ConvexLASSO.

3.2 GROUP CONVEXLAR

In this section we outline a strategy for extending ConvexLAR to incorporate group structure among predictors. Suppose predictors are divided into m groups with group size p_i, $i = 1, \ldots, m$. Slightly abusing notation, we use g to represent both the gth group and the index set of all predictors belonging to the gth group. In a similar manner, we use $\mathcal{G}$ to represent both the set of active groups and the index set of all predictors belonging to current active groups. For an arbitrary matrix $\boldsymbol{H}$, $\boldsymbol{H}_{\mathcal{I},\mathcal{J}}$ denotes the submatrix of $\boldsymbol{H}$ with rows in $\mathcal{I}$ and columns in $\mathcal{J}$.

Initialize $s(0) = \max_j (w_j|\nabla_j f(\mathbf{0})|)$, $\boldsymbol{\beta}(0) = \mathbf{0}$, and $\mathcal{A} = \text{argmax}_j w_j|\nabla_j f(\mathbf{0})|$;
repeat
  Solve ODE

$$\frac{d}{dt}\boldsymbol{\beta}_{\mathcal{A}}(t) = s'(t)\boldsymbol{H}_{\mathcal{A}}^{-1}(\boldsymbol{\beta}(t))\text{diag}(\boldsymbol{w}_{\mathcal{A}}^{-1})\text{sgn}(\nabla_{\mathcal{A}} f(\boldsymbol{\beta}(t))$$

  until $w_j|\nabla_j f(\boldsymbol{\beta}(t))| = s(t)$ for some $j \notin \mathcal{A}$;
  Update set $\mathcal{A} \leftarrow \mathcal{A} \cup \{j : w_j|\nabla_j f(\boldsymbol{\beta}(t))| = s(t), j \notin \mathcal{A}\}$
until $s(t) = 0$ *or* $\boldsymbol{H}_{\mathcal{A}}(\boldsymbol{\beta}(t))$ *is singular* ;

Algorithm 3: Weighted ConvexLAR with predictor specific weights $w_j > 0$.

Group ConvexLAR can be devised based on the following considerations: (i) We need one way to gauge the aggregate "influence," denoted by $I_g(\boldsymbol{\beta})$, of a group g. For instance, we can use either the ℓ_2 norm or ℓ_1 norm of the subvector of the gradient corresponding to a group, that is, $I_g(\boldsymbol{\beta}) = \|\nabla_g f(\boldsymbol{\beta})\|_2 = \sqrt{\sum_{j\in g}[\nabla_j f(\boldsymbol{\beta})]^2}$ or $I_g(\boldsymbol{\beta}) = \|\nabla_g f(\boldsymbol{\beta})\|_1 = \sum_{j\in g}|\nabla_j f(\boldsymbol{\beta})|$. Note that in principle any ℓ_p norm, $p \geq 1$, can be used. (ii) For all groups in the active group set $\mathcal{G}$, we use the function $s(t)$ to control the common value of $e(p_g)^{-1} I_g(\boldsymbol{\beta})$,

$g \in \mathcal{G}$, where p_g denotes the number of predictors in group g and $e(p_g)$ is the effective size of gth group. Common choices for $e(p_g)$ are $\sqrt{p_g}$ or p_g. (iii) To make the ODE well defined, we need to assign the proportions of individual "influences" $c_g(j)$ to each predictor j in group g. We require $\|\boldsymbol{c}_g\|_p = 1$ to satisfy the equation $\|\nabla_g f(\boldsymbol{\beta})\|_p = \|\boldsymbol{c}_g I_g(\boldsymbol{\beta})\|_p$, where $\boldsymbol{c}_g$ is a column vector with $c_g(j)$, $j \in g$, stacked together. These conditions imply the general group LAR identity

$$\nabla_j f(\boldsymbol{\beta}) = c_g(j) I_g(\boldsymbol{\beta}) = c_g(j) e(p_g) s(t), \quad j \in g, g \in \mathcal{G}. \tag{6}$$

Now the implicit function theorem yields the group LAR updating direction (7) in Algorithm 5. Here $\boldsymbol{c}_\mathcal{G}$ is a column vector with $\boldsymbol{c}_g$ for $g \in \mathcal{G}$ stacked together and $\boldsymbol{p}_\mathcal{G}$ is a column vector with the corresponding effective group size $e(p_g)$ for each predictor j in group g stacked together.

Specific choices of the group "influence" $I_g(\boldsymbol{\beta})$, effective group size $e(p_g)$, and individual "influence" $c_g(j)$ lead to different updating directions. We specialize to the following three variants. The first has a close connection to the group lasso. The second and third recover two versions of group LAR algorithms developed in the literature.

1. GroupConvexLAR: We measure the group influence by $I_g(\boldsymbol{\beta}) = \|\nabla_g f(\boldsymbol{\beta})\|_2$, which is distributed to individual predictors within the group according to their effect size $c_g(j) = -\beta_j / \|\boldsymbol{\beta}_g\|_2$. Let $e(p_g) = \sqrt{p_g}$ and assume $\mathcal{G} = \mathcal{G}_0 \cup \mathcal{G}_1$, where $\mathcal{G}_0 = \{g_0, \ldots, g_a\}$ denotes the set of active groups with $\|\boldsymbol{\beta}_g\|_2 = 0$ and $\mathcal{G}_1 = \{g_{a+1}, \ldots, g_{a+b}\}$ denotes the set of active groups with $\|\boldsymbol{\beta}_g\|_2 > 0$.

Initialize $s(0) = \max_g e^{-1}(p_g) I_g(\mathbf{0})$, $\boldsymbol{\beta}(0) = \mathbf{0}$, and $\mathcal{G} = \operatorname{argmax}_g \frac{I_g(\mathbf{0})}{e(p_g)}$;
repeat
  Solve ODE system defined by

$$\frac{d}{dt}\boldsymbol{\beta}_\mathcal{G}(t) = \boldsymbol{H}_\mathcal{G}^{-1}(\boldsymbol{\beta}(t)) \left[s'(t) \operatorname{diag}(\boldsymbol{p}_\mathcal{G}) \boldsymbol{c}_\mathcal{G} + s(t) \operatorname{diag}(\boldsymbol{p}_\mathcal{G}) \frac{d}{dt} \boldsymbol{c}_\mathcal{G}(t) \right] \tag{7}$$

  and $\boldsymbol{\beta}_g = \mathbf{0}$ for any inactive group $g \notin \mathcal{G}$ until $I_{g^*}(\boldsymbol{\beta}) = s(t) e(p_{g^*})$ for some $g^* \notin \mathcal{G}$;
  Update set $\mathcal{G} \leftarrow \mathcal{G} \cup \{g^*\}$
until $s(t) = 0$ *or* $\boldsymbol{H}_\mathcal{G}(\boldsymbol{\beta}(t))$ *is singular* ;

Algorithm 4: A general scheme for Group ConvexLAR.

Theorem 2. For a strictly convex and twice differentiable loss function $f(\boldsymbol{\beta})$, the LAR solution $\boldsymbol{\beta}(t)$ is continuous and differentiable at t within a segment.

(a) If $\mathcal{G}_0$ is an empty set, the active solution vector $\boldsymbol{\beta}_\mathcal{G}(t)$ satisfies the differential equation

$$\begin{aligned} \frac{d}{dt}\boldsymbol{\beta}_\mathcal{G}(t) &= [\boldsymbol{H}_\mathcal{G}(t) + \boldsymbol{D}]^{-1} s'(t) \operatorname{diag}(\boldsymbol{p}_\mathcal{G}) \boldsymbol{c}_\mathcal{G} \\ &= \frac{s'(t)}{s(t)} [\boldsymbol{H}_\mathcal{G}(t) + \boldsymbol{D}]^{-1} \nabla_\mathcal{G} f(\boldsymbol{\beta}), \end{aligned} \tag{8}$$

where $\boldsymbol{D}$ is the block diagonal matrix with blocks

$$\frac{s(t)\sqrt{p_g}}{\|\boldsymbol{\beta}_g(t)\|_2}\left(\boldsymbol{I}_{p_g}-\frac{1}{\|\boldsymbol{\beta}_g(t)\|_2^2}\boldsymbol{\beta}_g(t)\boldsymbol{\beta}_g^{\mathsf{T}}(t)\right),\quad g\in\mathcal{G}.$$

(b) If $\mathcal{G}_0$ is not empty, the solution vector $\boldsymbol{\beta}_{g_i}$ for $g_i \in \mathcal{G}_0$ satisfies

$$\frac{d\boldsymbol{\beta}_{g_i}(t)}{dt}=k_i\nabla_{g_i}f(\boldsymbol{\beta}), \tag{9}$$

and the constants k_i, $i=1,\ldots,a$, and updating direction for the groups in $\mathcal{G}_1$ are jointly determined by

$$\begin{pmatrix} k_1\\ \vdots\\ k_a\\ \frac{d}{dt}\boldsymbol{\beta}_{\mathcal{G}_1}(t)\end{pmatrix}=s(t)\begin{pmatrix}\boldsymbol{A} & \boldsymbol{B}^{\mathsf{T}}\\ \boldsymbol{B} & \boldsymbol{C}\end{pmatrix}^{-1}\begin{pmatrix} p_{g_1}s'(t)\\ \vdots\\ p_{g_a}s'(t)\\ \nabla_{g_{a+1}}f(\boldsymbol{\beta})/s'(t)\\ \vdots\\ \nabla_{g_{a+b}}f(\boldsymbol{\beta})/s'(t)\end{pmatrix}, \tag{10}$$

where $\boldsymbol{A}\in\mathbb{R}^{a\times a}$ has entries $a_{ij}=d_{g_i}f(\boldsymbol{\beta})H_{g_i,g_j}(\boldsymbol{\beta})\nabla_{g_j}f(\boldsymbol{\beta})$, $1\le i,j\le a$,

$$\boldsymbol{B}=\left[\boldsymbol{H}_{\mathcal{G}_1,g_1}\nabla_{g_1}f(\beta\,),\quad\ldots,\quad\boldsymbol{H}_{\mathcal{G}_1,g_a}\nabla_{g_a}f(\beta\,)\right]\in\mathbb{R}^{\sum_{i=a+1}^{a+b}p_i\times a},$$

and

$$\boldsymbol{C}=\boldsymbol{H}_{\mathcal{G}_1,\mathcal{G}_1}+\boldsymbol{D}_{\mathcal{G}_1,\mathcal{G}_1}\in\mathbb{R}^{\sum_{i=a+1}^{a+b}p_i\times\sum_{i=a+1}^{a+b}p_i}.$$

Note that, when $\mathcal{G}_0$ is empty, (10) reduces to (8). The technical proof of Theorem 2 is delegated to the Appendix. Again the strict convexity assumption can be relaxed to the positive definiteness of $\boldsymbol{H}_{\mathcal{G}}(\boldsymbol{\beta}(t))$ along the path, which guarantees the nonsingularity of the matrix involved in (10).

2. GroupConvexLAR-L1: We choose $I_g(\boldsymbol{\beta})=\|\nabla_g f(\boldsymbol{\beta})\|_1$, $e(p_g)=p_g$, and $c_g(j)=\frac{\nabla_j f(\boldsymbol{\beta}(t_g))}{\|\nabla_g f(\boldsymbol{\beta}(t_g))\|_1}$, where t_g is the time that the group g joins the active set $\mathcal{G}$. Note that in this case $c_g(j)$ is fixed for any $j\in g$ once group g joins the active set. Therefore $\frac{d}{dt}\boldsymbol{c}_{\mathcal{G}}(t)=\boldsymbol{0}$ and the group LAR direction (7) reduces to

$$\frac{d}{dt}\boldsymbol{\beta}_{\mathcal{G}}(t)=s'(t)\boldsymbol{H}_{\mathcal{G}}(\boldsymbol{\beta})^{-1}\mathrm{diag}(\boldsymbol{p}_{\mathcal{G}})\boldsymbol{c}_{\mathcal{G}}. \tag{11}$$

3. GroupConvexLAR-L2: With the choice $I_g(\boldsymbol{\beta})=\|\nabla_g f(\boldsymbol{\beta})\|_2$, $e(p_g)=\sqrt{p_g}$ and $c_g(j)=\frac{\nabla_j f(\boldsymbol{\beta}(t_g))}{\|\nabla_g f(\boldsymbol{\beta}(t_g))\|_2}$, the ODE updating direction is same as (10) with the obvious substitute for $\boldsymbol{c}_{\mathcal{G}}$ and $\boldsymbol{p}_{\mathcal{G}}$.

Note that, when all group sizes p_g are equal to 1,

$$\frac{\nabla_j f(\boldsymbol{\beta}(t_g))}{\|\nabla_g f\left(\boldsymbol{\beta}(t_g)\right)\|_2} = \frac{\nabla_j f(\boldsymbol{\beta}(t_g))}{\|\nabla_g f\left(\boldsymbol{\beta}(t_g)\right)\|_1} = \frac{-\beta_j}{\|\boldsymbol{\beta}_g\|_2} = \operatorname{sgn}(\nabla_j f(\boldsymbol{\beta}))$$

and $\frac{I_g(\boldsymbol{\beta})}{e(p_g)} = |\nabla_g f(\boldsymbol{\beta})|$ for all g. All three variants reduce to the ConvexLAR.

Connection With Previous Group LAR. Consider the variant GroupConvexLAR-L2 in the special case of least squares, that is, $f(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2/2$. In this case, both $\boldsymbol{H}_{\mathcal{G}}(\boldsymbol{\beta}) = \boldsymbol{X}_{\mathcal{G}}^{\mathsf{T}}\boldsymbol{X}_{\mathcal{G}}$ and $\boldsymbol{c}_{\mathcal{G}}(t)$ are constant within a segment. Thus, the group LAR updating direction (11) is constant within each segment, leading to a piecewise linear solution path with segment-wise slope

$$\frac{d}{dt}\boldsymbol{\beta}_{\mathcal{G}}(t) = \frac{s'(t)}{s(t)}\left[\boldsymbol{X}_{\mathcal{G}}^{\mathsf{T}}\boldsymbol{X}_{\mathcal{G}}\right]^{-1}\nabla_{\mathcal{G}} f(\boldsymbol{\beta}).$$

This recovers a version of group LAR proposed by Yuan and Lin (2006) for group selection in least squares. Park and Hastie (2006) argued that this version of group LAR tends to select a large group with only few of its component correlated with the response. To avoid this problem, they proposed another version of group LAR by simply replacing the average squared correlation $(\sum_{j\in g}[\nabla_j f(\boldsymbol{\beta})]^2/p_g)$ with the average absolute correlation $(\sum_{j\in g}|\nabla_j f(\boldsymbol{\beta})|/p_g)$, which is simply GroupConvexLAR-L1 specialized to least squares.

We emphasize that, for a general convex loss f, the solution paths by GroupConvexLAR-L1 and GroupConvexLAR-L2 are both piecewise smooth instead of piecewise linear and ODE solving is necessary.

3.3 GROUP CONVEXLASSO MODIFICATION

We show next that a simple modification of the aforementioned first variant GroupConvexLAR yields the solution path of group lasso penalized problem

$$\min_{\boldsymbol{\beta}} \quad f(\boldsymbol{\beta}) + \lambda \sum_g \sqrt{p_g}\|\boldsymbol{\beta}_g\|_2.$$

Its solution satisfies the following the Karush-Kuhn-Tucker (KKT) conditions

$$\nabla_g f(\boldsymbol{\beta}) + \lambda \frac{\sqrt{p_g}}{\|\boldsymbol{\beta}_g\|_2}\boldsymbol{\beta}_g = 0\,, \quad \boldsymbol{\beta}_g \neq \mathbf{0}; \tag{12}$$

$$\|\nabla_g f(\boldsymbol{\beta})\|_2 \leq \lambda\sqrt{p_g}\,, \quad \boldsymbol{\beta}_g = \mathbf{0}. \tag{13}$$

With the choice $\lambda = s(t)$, the stationarity condition (12) coincides with the group LAR identity (6). This observation together with the above KKT conditions implies that the group lasso solution path moves along the same direction (8) of the GroupConvexLAR until either one of the following two types of event happens. The first type of event occurs when all predictors of an active group hit zero, denoted by event of type (i). The second type of event occurs when $\|\nabla_g f(\boldsymbol{\beta})\|_2$ hits boundary $\lambda\sqrt{p_g}$, denoted by event of type (ii). Both types of event change the active set and redefine the direction. The second type of event is already considered in the GroupConvexLAR algorithm. Thus a simple modification of GroupConvexLAR by tracking the first type of event leads to the group lasso solution path. This modification is summarized in Algorithm 11. This exact path

following algorithm for the group lasso penalized convex loss seems new. On a side note, there is no obvious connection between GroupConvexLAR-L1, L2, and group lasso.

Initialize $s(0) = \max_g I_g(0)/\sqrt{p_g}$, $\boldsymbol{\beta}(0) = \mathbf{0}$, and $\mathcal{A} = \operatorname{argmax}_g I_g(0)/\sqrt{p_g}$;
repeat
  Solve ODE according to (9) and (10) until
  (i) $\boldsymbol{\beta}_g(t) = 0$ for some $g \in \mathcal{A}$ or
  (ii) $I_g(\boldsymbol{\beta}) = s(t)\sqrt{p_g}$ for some $g \notin \mathcal{A}$;
  if *event of type (i)* **then**
    Update set $\mathcal{A} \leftarrow \mathcal{A} \setminus \{g : \boldsymbol{\beta}_g(t) = 0, g \in \mathcal{A}\}$
  else if *event of type (ii)* **then**
    Update set $\mathcal{A} \leftarrow \mathcal{A} \cup \{g : I_g(\boldsymbol{\beta}) = s(t)\sqrt{p_g},\ g \notin \mathcal{A}\}$
  end
until $s(t) = 0$ *or* $\boldsymbol{H}_G(\boldsymbol{\beta}(t))$ *is singular* ;

Algorithm 5: Group ConvexLASSO.

4. EXAMPLES

We illustrate ConvexLAR and its extensions on various statistical problems. To demonstrate the efficiency of the proposed algorithm, we report the running times of all numeric examples on i7 Core 2.93GHz, 8G RAM averaged over 50 independent runs.

4.1 RECURRENT EVENT DATA

Suppose that we have n independent subjects in a recurrent event study. For each subject i, $N_i(t)$ denotes the number of events that occur over the interval $(0, t]$ and $\boldsymbol{x}_i \in \mathbb{R}^p$ is the corresponding covariate vector. Assume that given $\boldsymbol{x}_i$, the counting process $\{N_i(t)\}$ is a nonhomogeneous Poisson process with mean function

$$\mu_i(t) = \mathrm{E}\{N_i(t)|\boldsymbol{x}_i\} = \mu_0(t)\exp\left(\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}\right) \tag{14}$$

for some unknown continuous baseline mean function $\mu_0(t)$ and unknown parameters $\boldsymbol{\beta} \in \mathbb{R}^p$. See Tong, Zhu, and Sun (2009b) for a detailed introduction to recurrent event data.

As typical in survival study, each subject is subject to potential censoring. Let C_i denote the follow-up or dropout time for subject i and $\tilde{N}_i(t) = N_i(\min(t, C_i))$ be a point process for subject i's observed process. The observed data are summarized as $\{(\tilde{N}_i(t), C_i, \boldsymbol{x}_i), i = 1, 2, \ldots, n, 0 \le t \le T\}$, where the constant T denotes the maximum potential follow-up time. The log-partial likelihood function based on model (13) is given by

$$\ell(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\mathsf{T}}\left[\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta} - \ln\left\{\sum_{j=1}^{n} Y_j(t)\exp(\boldsymbol{x}_j^{\mathsf{T}}\boldsymbol{\beta})\right\}\right] d\tilde{N}_i(t). \tag{15}$$

Under mild regularity conditions, the log-partial likelihood $\ell(\boldsymbol{\beta})$ is strictly concave in $\boldsymbol{\beta}$. Thus in this example, our objective function is chosen to be $f(\boldsymbol{\beta}) = -\ell(\boldsymbol{\beta})$. The first two

derivatives of $f(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ are

$$\nabla f(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\left[\boldsymbol{x}_i - \frac{\sum_{j=1}^{n} Y_j(t)\exp\left(\boldsymbol{x}_j^{\mathsf{T}}\boldsymbol{\beta}\right)\boldsymbol{x}_j}{\sum_{j=1}^{n} Y_j(t)\exp\left(\boldsymbol{x}_j^{\mathsf{T}}\boldsymbol{\beta}\right)}\right] d\tilde{N}_i(t),$$

$$d^2 f(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\left[\frac{\sum_{j=1}^{n} Y_j(t)\exp\left(\boldsymbol{x}_j^{\mathsf{T}}\boldsymbol{\beta}\right)\boldsymbol{x}_j\boldsymbol{x}_j^{\mathsf{T}}}{\sum_{j=1}^{n} Y_j(t)\exp\left(\boldsymbol{x}_j^{\mathsf{T}}\boldsymbol{\beta}\right)} - \frac{\left(\sum_{j=1}^{n} Y_j(t)\exp\left(\boldsymbol{x}_j^{\mathsf{T}}\boldsymbol{\beta}\right)\boldsymbol{x}_j\right)\left(\sum_{j=1}^{n} Y_j(t)\exp\left(\boldsymbol{x}_j^{\mathsf{T}}\boldsymbol{\beta}\right)\boldsymbol{x}_j\right)^{\mathsf{T}}}{\left(\sum_{j=1}^{n} Y_j(t)\exp\left(\boldsymbol{x}_j^{\mathsf{T}}\boldsymbol{\beta}\right)\right)^2}\right] d\tilde{N}_i(t).$$

Tong, Zhu, and Sun (2009b) studied a Chronic Granulomatous Disease (CGD) data which were collected from a multicenter placebo-controlled randomized trial of gamma inferon with chronic granulotomous disease. There were 128 patients randomized to two groups, gamma interferon group ($n_1 = 63$) and placebo group ($n_2 = 65$). For each patient the times from the beginning of the study to initial and any recurrent serious infections are available. Eleven covariates are considered: 1. trtmt=treatment (Yes/No), 2. inherit=pattern of inheritance (autosomal/recessive), 3. age, 4. height, 5. weight, 6. cortico=use of corticosteroids (Yes/No), 7. prophy=use of prophylactic antibiotics (Yes/No), 8. gender=female, 9. hosp1=hosp. (category: US/other), 10. hosp2=hosp. (category: Europe-Amsterdam), and 11. hosp3=hosp. (category: Europe-other). We standardize all continuous covariates (age, height, and weight) to have mean 0 and unit length. Furthermore, we also include the six quadratic and interaction terms between three continuous covariates. They are: 12. age×age, 13. height×height, 14. weight×weight, 15. age×height, 16. age×weight, 17. height×weight.

Figure 1 shows the solution paths from different algorithms. Here the numbers on the right-hand side indicate which variable each path corresponds to. In all plots, the x-axes are in the units of $\|\boldsymbol{\beta}(t)\|_1 / \max_t \|\boldsymbol{\beta}(t)\|_1$ and vertical lines mark the event times for easy comparison between various solution paths. Top row of Figure 1 displays the ConvexLAR (top left panel) and ConvexLASSO (top right panel) solution paths for the CGD data. They are qualitatively different. For instance, in the ConvexLASSO path, β_9 hits zero and then escapes active set at Step 14. In addition we also apply the weighted CovexLAR and ConvexLASSO algorithms with weights set at the maximum likelihood estimator of Equation (15). They differ significantly from the unweighted ones. The running times are 4.84, 5.60, 5.34, and 6.37 sec for the ConvexLAR, ConvexLASSO, weighted ConvexLAR, and weighted ConvexLASSO solution paths, respectively.

4.2 PANEL COUNT DATA

In the above recurrent event example, we assume that the exact time of each event, if not censored, is observed. Unfortunately, this is not the case in many studies. The model for panel count data (Sun and Wei 2000; Tong et al. 2009a) provides a remedy.

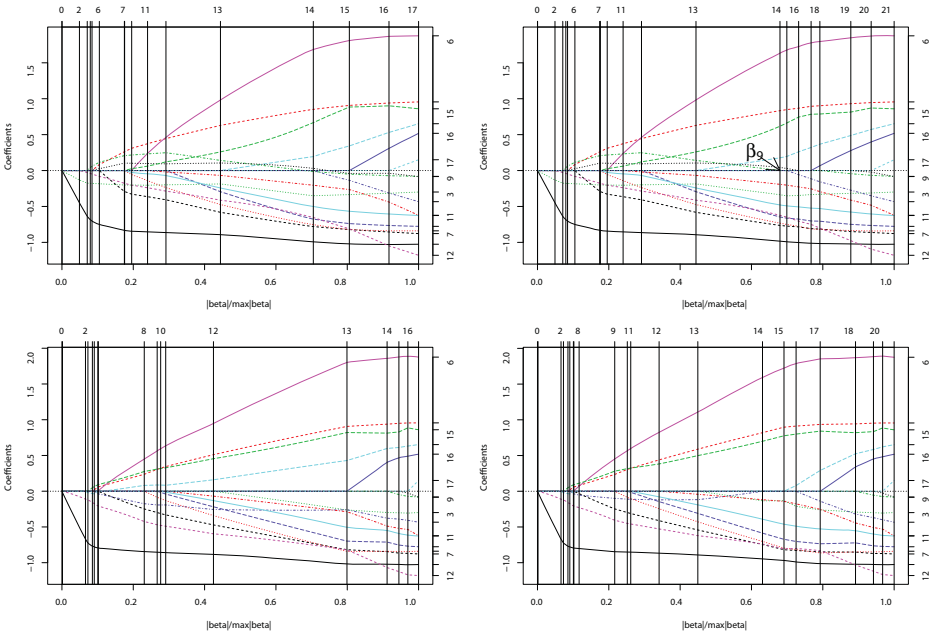

Figure 1. Recurrent event example (CGD data). Top left: ConvexLAR; Top right: ConvexLASSO; Bottom left: weighted ConvexLAR; Bottom right: weighted ConvexLASSO.

Let $T_{i1} < T_{i2} < \cdots < T_{im_i}$ be the potential observation times on process $N_i(t)$ and $H_i(t) = \sum_{j=1}^{m_i} I(T_{ij} \le t)$ denote the observation process. Let $\tilde{H}_i(t) = H(\min(t, C_i))$ be the actual observation times after censoring. Then the observed data for panel count model are

$$\left\{\left(N_i(t)d\tilde{H}_i(t), \tilde{H}_i(t), C_i, \boldsymbol{x}_i^{\mathsf{T}}\right), \quad i = 1, \ldots, n, 0 \le t \le T\right\}.$$

When H_i and C_i are mutually independent and also independent of N_i and $\boldsymbol{x}_i$, Sun and Wei (2000) propose to estimate regression parameters by solving the following estimation equation

$$W(\boldsymbol{\beta}) = \sum_{i=1}^{n} \bar{N}_i e^{-\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}} \boldsymbol{x}_i = \mathbf{0},$$

where $\bar{N}_i = \int_0^{\mathsf{T}} N_i(t)d\tilde{H}_i(t)$. ConvexLAR and its extensions can be applied directly to the influence function

$$-W(\boldsymbol{\beta}) = \sum_{i=1}^{n} \bar{N}_i e^{-\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}} \boldsymbol{x}_i$$

which has a positive definite derivative

$$-DW(\boldsymbol{\beta}) = \sum_{i=1}^{n} \bar{N}_i e^{-\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}} \boldsymbol{x}_i \boldsymbol{x}_i^{\mathsf{T}}.$$

We illustrate with the bladder tumor recurrence data considered in Sun and Wei (2000). A total of 85 bladder tumor patients were randomized into two treatment groups, that is, placebo group and thiotepa treatment group. Most patients visited the hospital several times to have their recurrent tumors removed. The number of new tumors discovered at each visit was recorded and removed after each visit. The binary treatment (trtmt) is one of our explanatory covariates. Furthermore, we consider two additional important baseline covariates, the number of initial tumors (num) and the size of the largest initial tumor (size). All covariates are centered around zero and scaled to have unit length. Figure 2 shows the ConvexLAR and ConvexLASSO solution paths for this example. The two solution paths are identical in this particular example as no active predictors return to zero along the path. The running times are 0.04 and 0.05 sec for the ConvexLAR and ConvexLASSO solution paths, respectively.

4.3 ADA-BOOST

Ada-Boost is considered one of the best off-the-shelf classification methods (Hastie, Tibshirani, and Friedman (2009). In binary classification, we are given a training dataset $\{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, n\}$ with $\boldsymbol{x}_i \in \mathbb{R}^p$ and $y_i \in \{-1, +1\}$. The goal is to estimate a function $f(\boldsymbol{x})$ whose sign will be used as the classification rule. For simplicity we consider the linear classification, in which $f(\boldsymbol{x}) = \beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta}$. The Ada-Boost estimates parameters by solving

$$\min_{\beta_0, \boldsymbol{\beta}} \sum_{i=1}^{n} e^{-y_i\left(\beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)}, \tag{16}$$

which is strictly convex and thus amenable to our ConvexLAR algorithm. Denote $\mathcal{D} = \{i : y_i = -1\}$. Define the marginal minimizer of β_0 as a function of $\boldsymbol{\beta}$

$$\boldsymbol{\beta}_0(\boldsymbol{\beta}) = \text{argmin}_{\boldsymbol{\beta}_0} f(\boldsymbol{\beta}_0, \boldsymbol{\beta}) = \frac{1}{2} \log \left\{ \frac{\sum_{i \in \mathcal{D}^c} e^{-\boldsymbol{x}_i^\top \boldsymbol{\beta}}}{\sum_{i \in \mathcal{D}} e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}} \right\}. \tag{17}$$

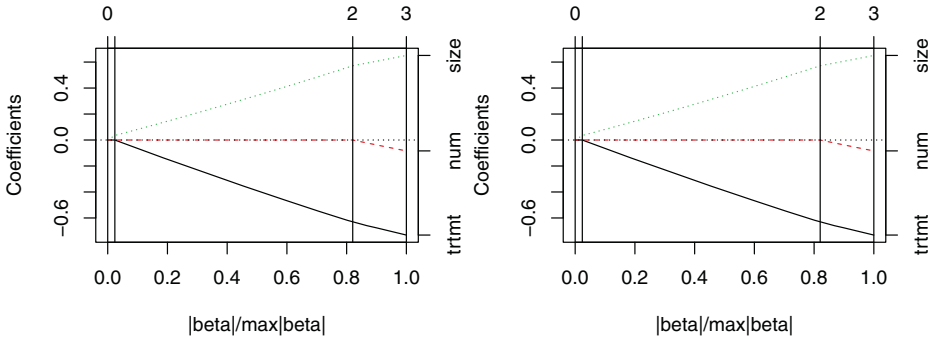

Figure 2. ConvexLAR and ConvexLASSO solution paths for the panel count data (bladder) example.

Thus the first two derivatives of $f(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ are

$$\begin{aligned}
\nabla f(\boldsymbol{\beta}) &= \left(\frac{\partial \beta_0}{\partial \boldsymbol{\beta}}\right)\left[e^{\beta_0}\sum_{\mathcal{D}} e^{\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}} - e^{-\beta_0}\sum_{\mathcal{D}^c} e^{-\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}}\right] \\
&\quad + \left[e^{\beta_0}\sum_{\mathcal{D}} e^{\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}}\boldsymbol{x}_i - e^{-\beta_0}\sum_{\mathcal{D}^c} e^{-\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}}\boldsymbol{x}_i\right], \\
d^2 f(\boldsymbol{\beta}) &= \left(\frac{\partial^2 \beta_0}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^{\mathsf{T}}}\right)\left[e^{\beta_0}\sum_{\mathcal{D}} e^{\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}} - e^{-\beta_0}\sum_{\mathcal{D}^c} e^{-\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}}\right] \\
&\quad + \left(\frac{\partial \beta_0}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \beta_0}{\partial \boldsymbol{\beta}}\right)^{\mathsf{T}}\left[e^{\beta_0}\sum_{\mathcal{D}} e^{\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}} + e^{-\beta_0}\sum_{\mathcal{D}^c} e^{-\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}}\right] \\
&\quad + \left(\frac{\partial \beta_0}{\partial \boldsymbol{\beta}}\right)\left[e^{\beta_0}\sum_{\mathcal{D}} e^{\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}}\boldsymbol{x}_i + e^{-\beta_0}\sum_{\mathcal{D}^c} e^{-\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}}\boldsymbol{x}_i\right]^{\mathsf{T}} \\
&\quad + \left[e^{\beta_0}\sum_{\mathcal{D}} e^{\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}}\boldsymbol{x}_i + e^{-\beta_0}\sum_{\mathcal{D}^c} e^{-\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}}\boldsymbol{x}_i\right]\left(\frac{\partial \beta_0}{\partial \boldsymbol{\beta}}\right)^{\mathsf{T}} \\
&\quad + \left[e^{\beta_0}\sum_{\mathcal{D}} e^{\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathsf{T}} + e^{-\beta_0}\sum_{\mathcal{D}^c} e^{-\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathsf{T}}\right],
\end{aligned}$$

where

$$\begin{aligned}
\frac{\partial \beta_0(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= -\frac{1}{2}\left\{\frac{\sum_{\mathcal{D}^c} e^{-\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}}\boldsymbol{x}_i}{\sum_{\mathcal{D}^c} e^{-\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}}} + \frac{\sum_{\mathcal{D}} e^{\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}}\boldsymbol{x}_i}{\sum_{\mathcal{D}} e^{\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}}}\right\}, \\
\frac{\partial^2 \beta_0(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^{\mathsf{T}}} &= \frac{1}{2}\left\{\frac{\left(\sum_{\mathcal{D}^c} e^{-\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathsf{T}}\right)\left(\sum_{\mathcal{D}^c} e^{-\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}}\right) - \left(\sum_{\mathcal{D}^c} e^{-\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}}\boldsymbol{x}_i\right)\left(\sum_{\mathcal{D}^c} e^{-\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}}\boldsymbol{x}_i\right)^{\mathsf{T}}}{\left(\sum_{\mathcal{D}^c} e^{-\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}}\right)^2}\right\} \\
&\quad - \frac{1}{2}\left\{\frac{\left(\sum_{\mathcal{D}} e^{\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathsf{T}}\right)\left(\sum_{\mathcal{D}} e^{\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}}\right) - \left(\sum_{\mathcal{D}} e^{\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}}\boldsymbol{x}_i\right)\left(\sum_{\mathcal{D}} e^{\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}}\boldsymbol{x}_i\right)^{\mathsf{T}}}{\left(\sum_{\mathcal{D}} e^{\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}}\right)^2}\right\}.
\end{aligned}$$

The Wisconsin Diagnostic Breast Cancer (WDBC) data (Frank and Asuncion 2010) are collected on $n = 569$ patients from digitized images of a fine needle aspirate (FNA) of their breast mass. The number of predictors is $p = 10$. The mean, standard error, and "worst" or largest (mean of the three largest values) of these predictors were computed for each patient, resulting in 30 features forming 10 groups each of size 3. The response is binary in that each patient is diagnosed either as malignant ($Y = 1$) or benign ($Y = -1$). Each predictor variable is standardized to have mean zero and variance one. Figure 3 displays the group ConvexLARs solution paths for this example. The x-axis is $\log(1 + s(t))$, where $s(t)$ is the same as the λ in group ConvexLASSO. To have a clear view, only the solution paths where $\log(1 + s(t)) > 3$ are plotted. GroupConvexLAR-L2 and GroupConvexLAR-L1 appear quite different from GroupConvexLAR with larger $\max_{i\in\mathcal{G}} |\beta_i(t)|$ across the same level of $s(t)$. Bottom row of Figure 3 displays the GroupConvexLAR and Group ConvexLASSO

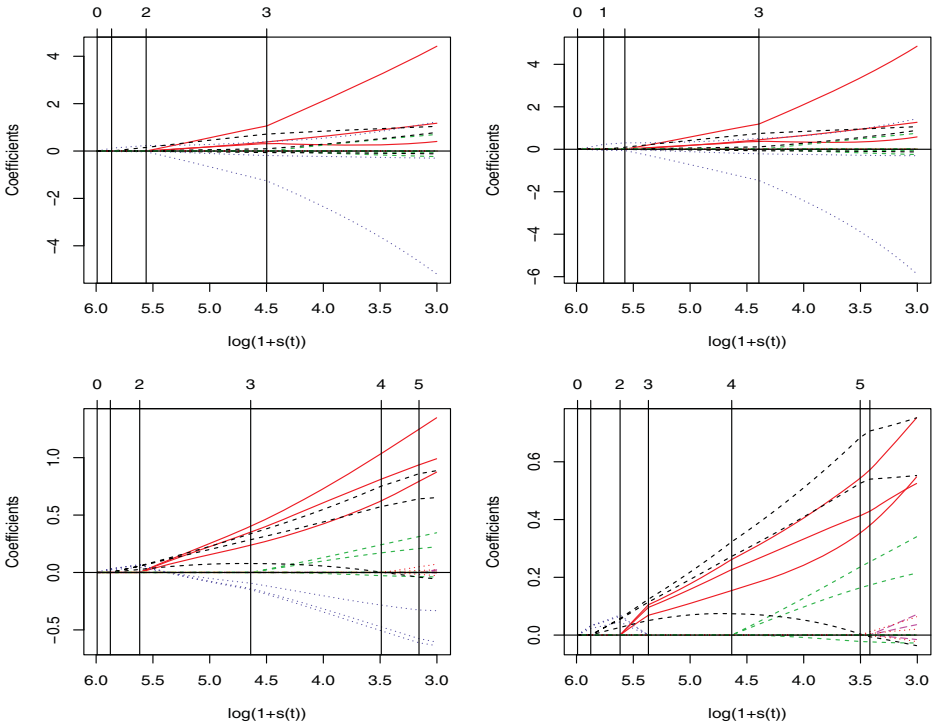

Figure 3. Group LARs solution paths for the Ada-Boost data (WDBC) example. Top left: GroupConvexLAR-L2; Top right: GroupConvexLAR-L1; Bottom left: GroupConvexLAR; Bottom right: Group ConvexLASSO.

solution paths, which are fundamentally different for the WDBC data. The third group (displayed with dotted line) is the first active one and then stay active along the whole GroupConvexLAR solution path. In contrast, the same group hits zero in the event 3 of the Group ConvexLASSO solution path and escapes the active set thereafter. The running times are 1.53, 1.58, 1.49, and 2.13 sec for the GroupConvexLAR-L2, GroupConvexLAR-L1, GroupConvexLAR and GroupConvexLASSO solution paths, respectively.

4.4 GAUSSIAN GRAPHICAL MODEL

Our fourth example concerns the LAR for Gaussian graphical model. Assume we have iid observations $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^p$ from $N(\mathbf{0}, \boldsymbol{\Sigma})$. Denote the sample variance covariance matrix by $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\mathsf{T}$ and precision matrix by $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. Then the negative likelihood for $\boldsymbol{\Omega}$ is given by $-\ell(\boldsymbol{\Omega}) = -\log\det\boldsymbol{\Omega} + \operatorname{tr}(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Omega})$, which is convex. Let ω_{ij} denote the (i, j)-element of $\boldsymbol{\Omega}$. $w_{ij} = 0$, $i \neq j$, implies the conditional independence between variables i and j. We partition the parameters into the sets $\boldsymbol{\Omega}_0 = (\omega_{11}, \omega_{22}, \ldots, \omega_{pp})^\mathsf{T} \in \mathbb{R}^p$ and $\boldsymbol{\Omega}_1 = (\omega_{12}, \omega_{13}, \ldots, \omega_{(p-1)p})^\mathsf{T} \in \mathbb{R}^{p(p-1)/2}$. Only those in $\boldsymbol{\Omega}_1$ are subject to regularization. We may rewrite the negative log-likelihood as $-\ell(\boldsymbol{\Omega}_0, \boldsymbol{\Omega}_1)$. For every fixed $\boldsymbol{\Omega}_1$, we define $\boldsymbol{\Omega}_0(\boldsymbol{\Omega}_1) = \operatorname{argmin}_{\boldsymbol{\Omega}_0} -\ell(\boldsymbol{\Omega}_0, \boldsymbol{\Omega}_1)$. With these notations, we have $f(\boldsymbol{\Omega}_1) = -\ell(\boldsymbol{\Omega}_0(\boldsymbol{\Omega}_1), \boldsymbol{\Omega}_1)$.

To derive the LAR solution path, we need the first two derivatives of the objective function

$$f(\boldsymbol{\Omega}_0, \boldsymbol{\Omega}_1) = -\log\det \boldsymbol{\Omega}(\boldsymbol{\Omega}_0, \boldsymbol{\Omega}_1) + \operatorname{tr}(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Omega}(\boldsymbol{\Omega}_0, \boldsymbol{\Omega}_1)),$$

where $\boldsymbol{\Omega}_0$ is implicitly determined by $\boldsymbol{\Omega}_1$. We first show how to determine $\boldsymbol{\Omega}_0$ given $\boldsymbol{\Omega}_1$. Setting partial derivative of f with respect to $\boldsymbol{\Omega}_0$

$$D_{\boldsymbol{\Omega}_0} f(\boldsymbol{\Omega}_0, \boldsymbol{\Omega}_1) = D_{\boldsymbol{\Omega}_0} f(\boldsymbol{\Omega}) D\boldsymbol{\Omega}(\boldsymbol{\Omega}_0) = -(\operatorname{vec} \boldsymbol{\Omega}^{-1})^{\mathsf{T}} D\boldsymbol{\Omega}(\boldsymbol{\Omega}_0) + (\operatorname{vec} \hat{\boldsymbol{\Sigma}})^{\mathsf{T}} D\boldsymbol{\Omega}(\boldsymbol{\Omega}_0)$$

to $\mathbf{0}$ gives the stationarity condition

$$\operatorname{diag}(\boldsymbol{\Omega}^{-1}(\boldsymbol{\Omega}_0, \boldsymbol{\Omega}_1)) = (\hat{\sigma}_{11}, \ldots, \hat{\sigma}_{pp})^{\mathsf{T}}.$$

In other words, given $\boldsymbol{\Omega}_1$, we need to choose $\boldsymbol{\Omega}_0$ such that the diagonal entries of $\boldsymbol{\Omega}^{-1}$ match those of $\hat{\boldsymbol{\Sigma}}$. In practice, Newton's iteration

$$\omega_0^{(t+1)} = \omega_0^{(t)} + \left\{[D\boldsymbol{\Omega}(\boldsymbol{\Omega}_0)]^{\mathsf{T}}[\boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Omega}^{-1}][D\boldsymbol{\Omega}(\boldsymbol{\Omega}_0)]\right\}^{-1} [\operatorname{diag}(\boldsymbol{\Omega}^{-1} - \hat{\boldsymbol{\Sigma}})]$$

can be applied to solve for $\boldsymbol{\Omega}_0$ given $\boldsymbol{\Omega}_1$. We denote this mapping by $\boldsymbol{\Omega}_0(\boldsymbol{\Omega}_1)$. The gradient $D\boldsymbol{\Omega}_0(\boldsymbol{\Omega}_1) \in \mathbb{R}^{p \times p(p-1)/2}$ will be of use later and is obtained through the implicit function theorem

$$\begin{aligned} D\boldsymbol{\Omega}_0(\boldsymbol{\Omega}_1) = &-\left\{[D\boldsymbol{\Omega}(\boldsymbol{\Omega}_0)]^{\mathsf{T}}(\boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Omega}^{-1})[D\Omega(\boldsymbol{\Omega}_0)]\right\}^{-1} [D\boldsymbol{\Omega}(\boldsymbol{\Omega}_0)]^{\mathsf{T}}(\boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Omega}^{-1}) \\ &\times [D\Omega(\boldsymbol{\Omega}_1)]. \end{aligned} \tag{18}$$

Now the first derivative of objective function f with respect to $\boldsymbol{\Omega}_1$ is

$$D_{\boldsymbol{\Omega}_1} f(\boldsymbol{\Omega}_0(\boldsymbol{\Omega}_1), \boldsymbol{\Omega}_1) = D_{\boldsymbol{\Omega}_1} f(\boldsymbol{\Omega}_0, \boldsymbol{\Omega}_1) + D_{\boldsymbol{\Omega}_0} f(\boldsymbol{\Omega}_0, \boldsymbol{\Omega}_1) D\boldsymbol{\Omega}_0(\boldsymbol{\Omega}_1)$$

but the second term vanishes because $D_{\boldsymbol{\Omega}_0} f(\boldsymbol{\Omega}_0, \boldsymbol{\Omega}_1) = \mathbf{0}$. Hence,

$$D_{\boldsymbol{\Omega}_1} f(\boldsymbol{\Omega}_0(\boldsymbol{\Omega}_1), \boldsymbol{\Omega}_1) = Df(\boldsymbol{\Omega}) D\boldsymbol{\Omega}(\boldsymbol{\Omega}_1) = (-\operatorname{vec} \boldsymbol{\Omega}^{-1} + \operatorname{vec} \hat{\boldsymbol{\Sigma}})^{\mathsf{T}} D\boldsymbol{\Omega}(\boldsymbol{\Omega}_1).$$

In words, given current $\boldsymbol{\Omega}_1$, calculate $\boldsymbol{\Omega}^{-1}$ at optimal $\boldsymbol{\Omega}_0$; then the off-diagonal entries of $2(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Omega}^{-1})$ form the gradient of f in terms of $\boldsymbol{\Omega}_1$. For the Hessian,

$$\begin{aligned} Hf(\boldsymbol{\Omega}_1) &= D_{\boldsymbol{\Omega}_1}[D_{\boldsymbol{\Omega}_1} f(\boldsymbol{\Omega}_0(\boldsymbol{\Omega}_1), \boldsymbol{\Omega}_1)]^{\mathsf{T}} \\ &= D_{\boldsymbol{\Omega}_1}[D\boldsymbol{\Omega}(\boldsymbol{\Omega}_1)^{\mathsf{T}}(-\operatorname{vec} \boldsymbol{\Omega}^{-1} + \operatorname{vec} \hat{\boldsymbol{\Sigma}})] \\ &= -D_{\boldsymbol{\Omega}_1}[D\boldsymbol{\Omega}(\omega_1)^{\mathsf{T}}(\operatorname{vec} \boldsymbol{\Omega}^{-1})] \\ &= -[D\boldsymbol{\Omega}(\omega_1)]^{\mathsf{T}}(\boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Omega}^{-1})[D\boldsymbol{\Omega}(\boldsymbol{\Omega}_1) + D\boldsymbol{\Omega}(\boldsymbol{\Omega}_0) D\boldsymbol{\Omega}_0(\boldsymbol{\Omega}_1)]. \end{aligned} \tag{19}$$

Now substitute $D\boldsymbol{\Omega}_0(\boldsymbol{\Omega}_1)$ by the expression (18).

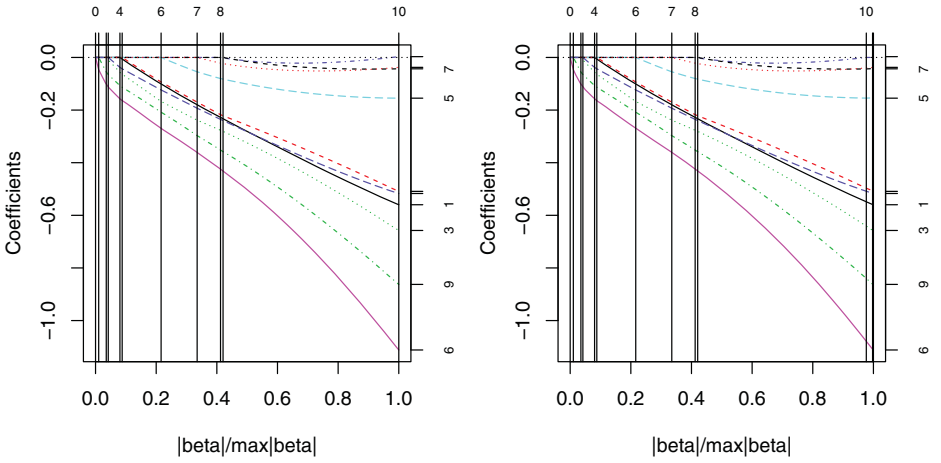

Figure 4. ConvexLAR and ConvexLASSO solution paths for the graphical model (math score) example.

We illustrate our algorithm by two examples. The first dataset contains 88 students' scores on five math courses—mechanics, vector, algebra, analysis, and statistics. See Table 1.2.1 of Mardia, Kent, and Bibby (1979) for more details. Figure 4 displays the ConvexLAR and ConvexLASSO solution paths. The most important three edges are analysis-algebra, statistics-algebra, and algebra-vector by lasso regularization. The ConvexLAR and ConvexLASSO solution paths coincide in this example. The running times are 0.28 and 0.37 sec for the ConvexLAR and ConvexLASSO solution paths, respectively.

Our second example concerns a simulated dataset with $p = 10$ and $n = 200$ and illustrates a case where ConvexLAR and ConvexLASSO yield different paths. The true precision matrix $\mathbf{\Omega} = (\omega_{ij})$ has entries $\omega_{ii} = 1$, $\omega_{i,i-1} = 0.5$, $\omega_{i,i+1} = 0.5$, and $\omega_{ij} = 0$ for $|i - j| > 1$. There are 45 nondiagonal free parameters. Figure 5 displays the ConvexLAR and ConvexLASSO solution paths. The solution paths appear different. The running times are 1.63 and 2.62 sec for the ConvexLAR and ConvexLASSO solution paths, respectively.

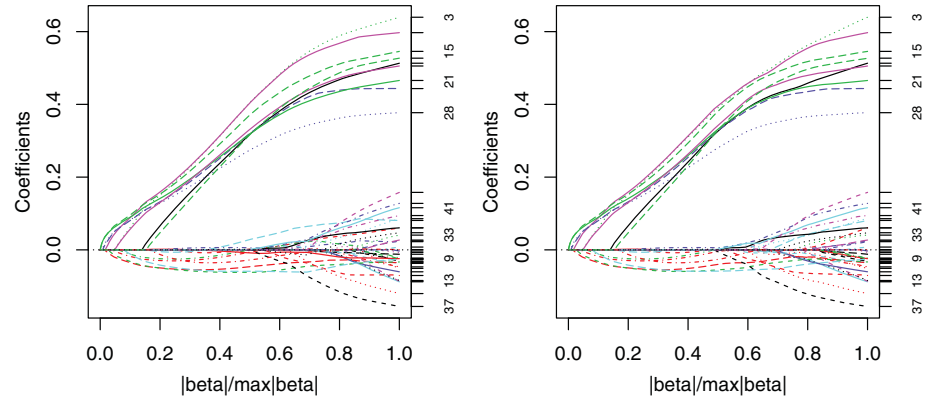

Figure 5. ConvexLAR and ConvexLASSO solution paths for the graphical model example with the simulated data.

5. DISCUSSION

Variable selection has become an essential tool for modern data analysis. So far penalization method such as lasso has been the dominant regularization technique and extended to handle increasingly more applications. In contrast, the original LAR (Efron et al. 2004) has received much less generalizations despite its popularity. In this expository article, we show that the simple geometric idea in LAR can be naturally extended to various situations such as convex loss, group structures in predictors, and data adaptive regularizations. The classical score function plays an essential role throughout the development. The original "least angle" idea translates to the equality of contributions by the active predictors to the score function. In our understanding, this is the fundamental idea in LAR and underpins the various extensions presented in this article.

This illustrative article is meant to whet readers' appetites not satiate them. Much is left undone. For instance, in principle it is the estimation equation that LAR operates on. Therefore, LAR naturally applies many statistical methods without a natural loss function such as generalized estimation equation (GEE), as hinted in the panel count data in Section 4. In this article we focus on LAR algorithmic development and forego the theoretical treatment. There has been intensive study of the asymptotic properties of regularized estimates by penalty methods in recent years. Same study for ConvexLAR is worth pursuing. Especially those that might shed light on the difference between the two regularization methods are much desired.

APPENDIX

Proof of Theorem 1. The LAR fundamental identity (1) for active predictors dictates the vector equation

$$k(\boldsymbol{\beta}_{\mathcal{A}}, t) = \nabla_{\mathcal{A}} f(\boldsymbol{\beta}) - \operatorname{sgn}(\nabla_{\mathcal{A}} f(\boldsymbol{\beta}))s(t) = \mathbf{0}.$$

To solve for $\boldsymbol{\beta}_{\mathcal{A}}$ in terms of t, we apply the implicit function theorem (Lange 2004). This requires calculating the differential of k with respect to the dependent variables $\boldsymbol{\beta}_{\mathcal{A}}$ and the independent variable t

$$\begin{aligned} \partial_{\boldsymbol{\beta}_{\mathcal{A}}} k(\boldsymbol{\beta}_{\mathcal{A}}, t) &= \boldsymbol{H}_{\mathcal{A}}(\boldsymbol{\beta}, t) \\ \partial_t k(\boldsymbol{\beta}_{\mathcal{A}}, t) &= -s'(t)\operatorname{sgn}(\nabla_{\mathcal{A}} f(\boldsymbol{\beta})). \end{aligned}$$

Given the nonsingularity of $\boldsymbol{H}_{\mathcal{A}}(\boldsymbol{\beta}, t)$, the implicit function theorem applies and shows the continuity and differentiability of $\boldsymbol{\beta}_{\mathcal{A}}(t)$ at t. Furthermore, it supplies the derivative (2).

Derivation of GroupConvexLAR directions (8), (9), *and* (10) We use $\mathcal{G}_0$ to denote the set of active groups that equal to $\mathbf{0}$. Let $\mathcal{G}_1 = \mathcal{G} - \mathcal{G}_0$, where $\mathcal{G}$ denotes the set of all active groups. We slightly abuse notation by letting $\mathcal{G}_0$, $\mathcal{G}_1$, and $\mathcal{G}$ to denote both the sets of groups and the set of all predictors belonging to the corresponding groups. Obviously, $\mathcal{G}_0$, $\mathcal{G}_1$, and $\mathcal{G}$ depend on the time index t.

Derivation of updating direction hinges upon the LAR identity (6), which we rewrite here for convenience

$$\nabla_g f(\boldsymbol{\beta}) + \sqrt{p_g} s(t) \frac{\boldsymbol{\beta}_g}{\|\boldsymbol{\beta}_g\|_2} = \mathbf{0}_{p_g}, \quad g \in \mathcal{G}. \tag{A.1}$$

1. When $\mathcal{G}_0$ is empty, differentiating the vector Equation (A.1) with respect to t via chain rule gives

$$\left[\boldsymbol{H}_{g,\mathcal{G}}(\boldsymbol{\beta}) + \sqrt{p_g} \frac{s(t)}{\|\boldsymbol{\beta}_g\|_2} \left(\boldsymbol{I}_{p_g} - \frac{\boldsymbol{\beta}_g \boldsymbol{\beta}_g^{\mathsf{T}}}{\|\boldsymbol{\beta}_g\|_2^2} \right) D_{\boldsymbol{\beta}_{\mathcal{G}}} \boldsymbol{\beta}_g \right] \frac{d\boldsymbol{\beta}_{\mathcal{G}}(t)}{dt}$$
$$+ \sqrt{p_g} s'(t) \frac{\boldsymbol{\beta}_g}{\|\boldsymbol{\beta}_g\|} = \mathbf{0}_{p_g}, \tag{A.2}$$

where $D_{\boldsymbol{\beta}_{\mathcal{G}}} \boldsymbol{\beta}_g = (\mathbf{0}, \boldsymbol{I}_{p_g}, \mathbf{0}) \in \mathbb{R}^{p_g \times \sum_{i=1}^{a+b} g_i}$. Combining all $|\mathcal{G}|$ vector equations and rearranging yields the LAR updating direction

$$\frac{d}{dt} \boldsymbol{\beta}_{\mathcal{G}}(t) = -[\boldsymbol{H}_{\mathcal{G}}(t) + \boldsymbol{D}]^{-1} \sqrt{p_g} s'(t) \frac{\boldsymbol{\beta}_g}{\|\boldsymbol{\beta}_g\|}$$
$$= \frac{s'(t)}{s(t)} [\boldsymbol{H}_{\mathcal{G}}(t) + \boldsymbol{D}]^{-1} \nabla_{\mathcal{G}} f(\boldsymbol{\beta}), \tag{A.3}$$

where $\boldsymbol{D}$ is the block diagonal matrix with blocks

$$\sqrt{p_g} \frac{s(t)}{\|\boldsymbol{\beta}_g\|_2} \left(\boldsymbol{I}_{p_g} - \frac{\boldsymbol{\beta}_g \boldsymbol{\beta}_g^{\mathsf{T}}}{\|\boldsymbol{\beta}_g\|_2^2} \right), \quad g \in \mathcal{G}.$$

2. When $\mathcal{G}_0$ not empty, we assume that $\mathcal{G}_0$ contains a groups, $g_1, \ldots, g_a$, and $\mathcal{G}_1$ contains b groups, $g_{a+1}, \ldots, g_{a+b}$. By rearranging the order of groups, we have $\boldsymbol{\beta}_{\mathcal{G}}^{\mathsf{T}} = (\boldsymbol{\beta}_{\mathcal{G}_0}^{\mathsf{T}}, \boldsymbol{\beta}_{\mathcal{G}_1}^{\mathsf{T}})$, where $\boldsymbol{\beta}_{\mathcal{G}_0}^{\mathsf{T}} = (\boldsymbol{\beta}_{g_1}^{\mathsf{T}}, \boldsymbol{\beta}_{g_2}^{\mathsf{T}}, \ldots, \boldsymbol{\beta}_{g_a}^{\mathsf{T}})$ and $\boldsymbol{\beta}_{\mathcal{G}_1}^{\mathsf{T}} = (\boldsymbol{\beta}_{g_{a+1}}^{\mathsf{T}}, \boldsymbol{\beta}_{g_{a+2}}^{\mathsf{T}}, \ldots, \boldsymbol{\beta}_{g_{a+b}}^{\mathsf{T}})$. For any group $g \in \mathcal{G}_1$, that is, $\|\boldsymbol{\beta}_g\|_2 \neq 0$, the vector Equation (A.2) still holds which gives

$$\left[\boldsymbol{H}_{g,\mathcal{G}}(\boldsymbol{\beta}) + \frac{\sqrt{p_g} s(t)}{\|\boldsymbol{\beta}_g\|_2} \left(\boldsymbol{I}_{p_g} - \frac{\boldsymbol{\beta}_g \boldsymbol{\beta}_g^{\mathsf{T}}}{\|\boldsymbol{\beta}_g\|_2} \right) D_{\boldsymbol{\beta}_{\mathcal{G}}} \boldsymbol{\beta}_g \right] \frac{d\beta_{\mathcal{G}}(t)}{dt} = \frac{s(t)}{s'(t)} \nabla_g f(\boldsymbol{\beta}). \tag{A.4}$$

Unfortunately, it does not hold for any $g \in \mathcal{G}_0$ due to the singularity $\|\boldsymbol{\beta}_g\|_2 = 0$. First we show that the updating direction of such groups is proportional to their gradient subvectors. By the LAR identity (A.1) and the fact $\boldsymbol{\beta}_g(t) = \mathbf{0}_{p_g}$,

$$\frac{[\boldsymbol{\beta}_g(t + \delta t) - \boldsymbol{\beta}_g(t)]/\delta t}{\|[\boldsymbol{\beta}_g(t + \delta t) - \boldsymbol{\beta}_g(t)]/\delta t\|_2} = -\frac{1}{\sqrt{p_g} s(t + \delta t)} \nabla_g f(\boldsymbol{\beta}(t + \delta t))$$

for all $\delta t > 0$. Taking limit $\delta t \downarrow 0$ yields

$$\frac{d\boldsymbol{\beta}_g(t)}{dt} = k_g \nabla_g f(\boldsymbol{\beta}), \tag{A.5}$$

where k_g are the constants to be determined.

Equating the norms of the two summand vectors in the LAR identity (A.1) shows $\|\nabla_g f(\boldsymbol{\beta})\|_2^2 = p_g s^2(t)$. Differentiating both sides of this identity with respect to t via chain rule gives

$$d_g f(\boldsymbol{\beta}) \boldsymbol{H}_{g,\mathcal{G}}(\boldsymbol{\beta}) \frac{d\boldsymbol{\beta}_{\mathcal{G}}(t)}{dt} = p_g s(t) s'(t) \tag{A.6}$$

for any $g \in \mathcal{G}_0$. Now substituting (A.5) into the Equations (A.4) and (A.6), we obtain

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B}^\mathsf{T} \\ \boldsymbol{B} & \boldsymbol{C} \end{pmatrix} \begin{pmatrix} k_1 \\ \vdots \\ k_a \\ \frac{d}{dt}\boldsymbol{\beta}_{\mathcal{G}_1}(t) \end{pmatrix} = \begin{pmatrix} p_{g_1} s'(t) \\ \vdots \\ p_{g_a} s'(t) \\ \nabla_{g_{a+1}} f(\boldsymbol{\beta})/s'(t) \\ \vdots \\ \nabla_{g_{a+b}} f(\boldsymbol{\beta})/s'(t) \end{pmatrix} s(t), \tag{A.7}$$

where $\boldsymbol{A} \in \mathbb{R}^{a \times a}$ has entries $a_{ij} = d_{g_i} f(\boldsymbol{\beta}) \boldsymbol{H}_{g_i,g_j}(\boldsymbol{\beta}) \nabla_{g_j} f(\boldsymbol{\beta})$, $1 \le i, j \le a$,

$$\boldsymbol{B} = \left[\boldsymbol{H}_{\mathcal{G}_1,g_1} \nabla_{g_1} f(\boldsymbol{\beta}), \quad \ldots, \quad \boldsymbol{H}_{\mathcal{G}_1,g_a} \nabla_{g_a} f(\boldsymbol{\beta}) \right] \in \mathbb{R}^{\sum_{i=a+1}^{a+b} p_i \times a},$$
$$\boldsymbol{C} = \boldsymbol{H}_{\mathcal{G}_1,\mathcal{G}_1} + \boldsymbol{D}_{\mathcal{G}_1,\mathcal{G}_1} \in \mathbb{R}^{\sum_{i=a+1}^{a+b} p_i \times \sum_{i=a+1}^{a+b} p_i}.$$

Next we show that the linear system is nonsingular and thus admits a unique solution, when $H_{\mathcal{A}}(\boldsymbol{\beta}(t)) = H_{\mathcal{G},G}(\boldsymbol{\beta}(t))$ is positive definite. Rewrite

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B}^\mathsf{T} \\ \boldsymbol{B} & \boldsymbol{C} \end{pmatrix} = \boldsymbol{E}(\boldsymbol{H}_{\mathcal{A}} + \tilde{\boldsymbol{D}})\boldsymbol{E}^\mathsf{T},$$

where

$$\boldsymbol{E} = \begin{pmatrix} d_{g_1} f(\boldsymbol{\beta}) & & & \\ & \ddots & & \\ & & d_{g_a} f(\boldsymbol{\beta}) & \\ & & & \boldsymbol{I}_{\sum_{i=a+1}^{a+b} p_{g_i}} \end{pmatrix} \in \mathbb{R}^{(a+\sum_{i=a+1}^{a+b} p_i) \times \sum_{i=1}^{a+b} p_i},$$
$$\tilde{\boldsymbol{D}} = \begin{pmatrix} \boldsymbol{0}_{\sum_{i=1}^{a} p_{g_i} \times \sum_{i=1}^{a} p_{g_i}} & \\ & \boldsymbol{D}_{\mathcal{G}_1,\mathcal{G}_1} \end{pmatrix} \in \mathbb{R}^{\sum_{i=1}^{a+b} p_i \times \sum_{i=1}^{a+b} p_i}.$$

Combining the facts (i) $\boldsymbol{H}_{\mathcal{A}}$ is positive definite, (ii) $\tilde{\boldsymbol{D}}$ is positive semidefinite, and (iii) $\|\nabla_g f(\boldsymbol{\beta})\|_2^2 = p_g s^2(t) > 0$ for all $g \in \mathcal{G}_0$, $\boldsymbol{E}(\boldsymbol{H} + \tilde{\boldsymbol{D}})\boldsymbol{E}^\mathsf{T}$ is positive definite.

Thus

$$\begin{pmatrix} k_1 \\ \vdots \\ k_a \\ \frac{d}{dt}\boldsymbol{\beta}_{\mathcal{G}_1}(t) \end{pmatrix} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B}^\top \\ \boldsymbol{B} & \boldsymbol{C} \end{pmatrix}^{-1} \begin{pmatrix} p_{g_1} s'(t) \\ \vdots \\ p_{g_a} s'(t) \\ \nabla_{g_{a+1}} f(\boldsymbol{\beta})/s'(t) \\ \vdots \\ \nabla_{g_{a+b}} f(\boldsymbol{\beta})/s'(t) \end{pmatrix} s(t). \tag{A.8}$$

Equation (A.8) coupled with (A.5) yields the LAR updating direction for all active predictors.

SUPPLEMENTARY MATERIALS

Matlab codes used in Section 4 are contained in the zip file matlabcodes.zip available online. Demonstration is provided in main_demo.m for all examples in Section 4.

ACKNOWLEDGMENTS

The authors thank the Editor, the Associate Editor, and two referees for their helpful suggestions that led to significant improvement of the article. The work is partially supported by grants NSF DMS-1055210 (Wu), NSF DMS-1310319 (Zhou), NIH/NCI R01 CA-149569 (Wu and Xiao), and NIH HG006139 (Zhou).

[Received May 2013. Revised August 2014.]

REFERENCES

Donoho, D. L., and Johnstone, I. M. (1994), "Ideal Spatial Adaptation by Wavelet Shrinkage," *Biometrika*, 81, 425–455. [603]

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression" (with discussion and reply), *The Annals of Statistics*, 32, 407–499. [604,605,607,608,622]

Frank, A., and Asuncion, A. (2010), UCI Machine Learning Repository. [618]

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.), Springer Series in Statistics, New York: Springer. [617]

Lange, K. (2004), *Optimization*, Springer Texts in Statistics, New York: Springer-Verlag. [622]

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979), "Multivariate Analysis," in *Probability and Mathematical Statistics: A Series of Monographs and Textbooks*, London: Academic Press (Harcourt Brace Jovanovich Publishers). [621]

Meier, L., van de Geer, S., and Bühlmann, P. (2008), "The Group Lasso for Logistic Regression," *Journal of the Royal Statistical Society,* Series B, 70, 53–71. [603]

Ortega, J. M., and Rheinboldt, W. C. (2000), *Iterative Solution of Nonlinear Equations in Several Variables* (volume 30 of *Classics in Applied Mathematics*), Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM). (Reprint of the 1970 original.) [609]

Park, M. Y., and Hastie, T. (2006), "Regularization Path Algorithms for Detecting Gene Interactions," Technical Report, Stanford University. [604,613]

Sun, J., and Wei, L. J. (2000), "Regression Analysis of Panel Count Data With Covariate-Dependent Observation and Censoring Times," *Journal of The Royal Statistical Society,* Series B, 62, 293–302. [615,616]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [603]

Tong, X., He, X., Sun, L., and Sun, J. (2009a), "Variable Selection for Panel Count Data via Non-Concave Penalized Estimating Function," *Scandinavian Journal of Statistics*, 36, 620–635. [615]

Tong, X., Zhu, L., and Sun, J. (2009b), "Variable Selection for Recurrent Event Data via Nonconcave Penalized Estimating Function," *Lifetime Data Analysis*, 15, 197–215. [614,615]

Wu, Y. (2011), "An Ordinary Differential Equation-based Solution Path Algorithm," *Journal of Nonparametric Statistics*, 23, 185–199. [605,608]

——— (2012), "Elastic Net for Coxs Proportional Hazards Model With a Solution Path Algorithm," *Statistical Sinica*, 22, 271–294. [604]

Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression With Grouped Variables," *Journal of the Royal Statistical Society*, Series B, 68, 49–67. [603,604,613]

Zou, H. (2006), "The Adaptive Lasso and its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [603,609]