Published in final edited form as:
J Nonparametr Stat. 2011 ; 23(1): 185–199.

An ordinary differential equation based solution path algorithm

Yichao Wu*
Department of Statistics, North Carolina State University, 2311 Stinson Drive, Raleigh, NC 27695

Abstract

Efron, Hastie, Johnstone and Tibshirani (2004) proposed Least Angle Regression (LAR), a solution path algorithm for the least squares regression. They pointed out that a slight modification of the LAR gives the LASSO (Tibshirani, 1996) solution path. However it is largely unknown how to extend this solution path algorithm to models beyond the least squares regression. In this work, we propose an extension of the LAR for generalized linear models and the quasi-likelihood model by showing that the corresponding solution path is piecewise given by solutions of ordinary differential equation systems. Our contribution is twofold. First, we provide a theoretical understanding on how the corresponding solution path propagates. Second, we propose an ordinary differential equation based algorithm to obtain the whole solution path.

Keywords

generalized linear model; LARS; LASSO; ordinary differential equation; solution path algorithm; QuasiLARS; quasi-likelihood model

1 Introduction

Recently we have seen exploding growth of research in variable selection popularized by Tibshirani (1996), which uses the L_1 penalty to regularize least squares regression. Following this line of research, many other techniques have been proposed. They include the SCAD (Fan and Li, 2001), the LARS (Efron et al., 2004), the elastic net (Zou and Hastie, 2005), the Dantzig selector (Candes and Tao, 2007), the adaptive LASSO (Zou, 2006; Zhang and Lu, 2007), and their related methods.

Computationally, the LASSO, elastic net, and adaptive LASSO can all be solved by any quadratic programming (QP) solver. The Dantzig selector involves a linear programming problem. The SCAD penalty leads to a non-convex optimization problem, for which Fan and Li (2001) proposed a local quadratic approximation (LQA) algorithm and Zou and Li (2008) proposed a local linear approximation (LLA) algorithm. They are two instances of the MM algorithm (Hunter and Li, 2005) and each step of the LQA or LLA involves a QP problem. All these algorithms share one characteristic in common: they solve the corresponding optimization for one regularization parameter at a time.

Efron et al. (2004) proposed the Least Angle Regression (LAR) algorithm and illustrated its close connection to the LASSO and Forward Stagewise linear regression. Together these algorithms are called LARS. By slight modification, their algorithm provides the whole *exact* solution path for the LASSO. The LARS solution paths are piecewise linear. Another algorithm for the LASSO is due to Osborne, Presnell and Turlach (2000) which proposed

*Corresponding author. wu@stat.ncsu.edu .

the homotopy algorithm. Rosset and Zhu (2007) derived a general characterization of the loss-penalty pair which leads to piecewise linear solution paths.

Note that the piecewise quadratic condition of Rosset and Zhu (2007) is not satisfied by generalized linear models (GLMs). The corresponding L_1 regularized solution path is not piecewise linear as demonstrated by Figure 2. To our limited knowledge, it is largely unknown how to extend the LARS to GLMs and more generally to the quasi-likelihood model (QLM) to get an *exact* solution path. Yet some approximate solution path algorithms are available. Madigan and Ridgeway (2004) discussed one possible extension to GLMs. Rosset (2004) suggested a general second-order path-following algorithm to track the curved regularized optimization solution path. Park and Hastie (2007)'s algorithm is based on the predictor-corrector method of convex optimization. To control the overall accuracy, Park and Hastie (2007) pointed out that it is critical to select the step length of the regularization parameter, for which strategies are proposed. These two papers try to approximate the whole regularization solution path by providing a series of solution sets at di erent regularization parameters, but different strategies are proposed to select the set of regularization parameters to control the approximation error. Yuan and Zou (2009) proposed an efficient global approach to approximate nonlinear L_1 regularization solution paths. Their method is based on the approximation of a general loss function by quadratic splines. In this way, the global loss approximation error can be controlled and a generalized LARS-type algorithm is devised to compute the corresponding exact solution path for the approximate quadratic spline loss. This path approximates the original nonlinear regularization solution path and theory is provided to show that the path approximation error is controlled by the global loss approximation error. On one hand, increasing the number of knots in the quadratic spline approximation makes the approximate solution path more accurate. On the other hand, it increases the number of pieces in the corresponding piecewise linear solution path and therefore the computational cost as well (Section 4 of Yuan and Zou, 2009). They further commented that "If the user wants to get the exact solution path from the EGA solution, then it seems not worthy to use a large number of knots."

This urgent need of an exact solution path calls for another algorithm. This is exactly the goal of the current paper. We extend the LAR to the QLM and name our extension QuasiLAR. Piecewise, our QuasiLAR solution path is given by solutions of ordinary differential equation (ODE) systems. We also discuss how the extension QuasiLAR is modified to get the whole solution path of the LASSO regularized quasi-likelihood, and this modified algorithm is called QuasiLASSO. Putting them together, we name our new algorithm QuasiLARS. The QuasiLARS is different from existing algorithms mentioned in the previous paragraph in that they all provide approximate solution paths instead. Our contribution is two-fold. On one hand, the current paper helps us to understand the corresponding optimization problem better by providing an answer to the question: how the general LASSO regularized solution path changes as the regularization parameter varies. On the other hand, we present an ODE based solution path algorithm and it provides a potential way to evaluate how well these existing solution path algorithms approximate the true solution path. Other papers on solution path algorithms include Zhu, Rosset, Hastie and Tibshirani (2004), Hastie, Rosset, Tibshirani and Zhu (2004), Wang and Shen (2006), Yuan and Lin (2007), Li, Liu and Zhu (2007), Wang and Zhu (2007), Wang, Shen and Liu (2008), Li and Zhu (2008), Zou (2008), Rocha, Zhao and Yu (2008), Wu, Shen and Geyer (2009), and references therein. In particular, Friedman, Hastie and Tibshirani (2010) focused on GLMs as well. They proposed a coordinate descent algorithm, which works for a fixed regularization parameter. They get a solution path by obtaining and connecting solutions at a pre-specified (penitentially dense) grid of the regularization parameter.

The rest of the article is organized as follows. Section 2 details the LARS and motivates the QuasiLARS. In Section 3, we present the QuasiLARS . Details for a key step are discussed in Section 4. Section 5 gives some properties of the QuasiLARS path. Numerical examples in Section 6 are used to illustrate how our QuasiLARS works. We conclude with Section 7. All technical proofs are collected in supplementary online material.

2 LARS

Before delving into details, let us see how the LAR works. To facilitate our later discussion, let us consider a general regression with a univariate response $Y \in \mathbb{R}$ and predictor vector $\boldsymbol{X}=\left(X_1, \cdots, X_p\right)^T \in \mathbb{R}^p$, where p denotes the number of predictor variables. The QLM assumes that $\mu(x) \triangleq E(Y|\boldsymbol{X} = x) = g^{-1}(\eta(x))$ with $\eta(x) = \beta_0 + x^T\beta$, and $\mathrm{Var}(Y|\boldsymbol{X} = x) = V(\mu(x))$ for some known monotonic link function $g(\cdot)$ and positive variance function $V(\cdot)$. Define $Q(\mu, y) = \int_y^\mu (y - w)/V(w)\, dw$ and denote our observed data set by $\{(x_i, y_i) : i = 1, \cdots, n\}$ with $\boldsymbol{x}_i=\left(x_{i1}, \cdots, x_{ip}\right)^T \in \mathbb{R}^p$, $y_i \in \mathbb{R}$ and n being the sample size. Predictors have been standardized such that $\sum_{i=1}^n x_{ij}=0$ and $\sum_{i=1}^n x_{ij}^2=1$, $j = 1, \cdots, n$. The QLM estimates β_0 and β by solving

$$\max_{\beta_0,\beta} \sum_{i=1}^n Q\left(g^{-1}\left(\beta_0+\boldsymbol{x}_i^T\beta\right), y_i\right). \tag{1}$$

The QLM includes GLMs as special cases by choosing $g(\cdot)$ and $V(\cdot)$ appropriately.

The ordinary least squares (OLS) regression is a special case with $g(\mu) = \mu$ and $\mathrm{V}(\mu) = \sigma^2$. In this case, by demeaning if necessary to ensure $\sum_{i=1}^n y_i=0$, (1) reduces to

$$\max_{\beta_0,\beta} -\sum_{i=1}^n\left(y_i - \boldsymbol{x}_i^T\beta\right)^2. \tag{2}$$

For OLS (2), the LAR provides a solution path $\beta(t)$ indexed by $t \in [0, \infty)$, with $\beta(0) = 0$. For large enough t, $\beta(t)$ is the same as the full OLS solution to (2). The solution path in between is piecewise linear. Over each piece, it moves along the direction that keeps the correlation between the current residuals and each active predictor equal in absolute value. Define current residuals $e_i(\beta(t)) = y_i - x_i^T\beta(t)$ for $i = 1, \cdots, n$. In terms of the current residual vector $e(\beta(t)) = (e_1(\beta(t)), \cdots, e_n(\beta(t)))^T$ and the jth predictor vector $\boldsymbol{x}^{(j)} = (x_{1j}, \cdots, x_{nj})^T$, the current correlation $e(\beta(t))^T\boldsymbol{x}^{(j)}$ has the same absolute value for each active predictor j. Note that $e(\beta(t))^T\boldsymbol{x}^{(j)}= -\frac{1}{2}\frac{\partial}{\partial\beta_j}\sum_{i=1}^n\left(y_i - \boldsymbol{x}_i^T\beta\right)^2|_{\beta(t)}$. This implies that the absolute values of the objective function's first-order derivatives are equal for each active predictor variable along the LAR solution path, namely $\left|\frac{\partial}{\partial\beta_j}\sum_{i=1}^n\left(y_i - \boldsymbol{x}_i^T\beta\right)^2|_{\beta(t)}\right| = \left|\frac{\partial}{\partial\beta_{j'}}\sum_{i=1}^n\left(y_i - \boldsymbol{x}_i^T\beta\right)^2|_{\beta(t)}\right|$ for any j and j' among the active predictor set at t. In this paper we will take advantage of this observation and extend LARS to the more general QLM. For the diabetes data in the R package LARS, we plot the LAR solution path in the top left panel of Figure 1. The derivatives $\frac{\partial}{\partial\beta_j}\sum_{i=1}^n\left(y_i - \boldsymbol{x}_i^T\beta\right)^2$ along the LAR solution path are shown in the top right

panel of Figure 1. The derivatives in absolute value, namely $\left|\frac{\partial}{\partial\beta_j}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i^T\beta\right)^2\right|$ are given in the bottom panel of Figure 1. It is clearly shows that, at the end of each LAR step, a new predictor variable joins the group of active predictor variables that share the honor of having the same largest absolute value of the first-order derivatives. The LAR algorithm terminates at the full OLS estimate of (2) when all the first-order partial derivatives are exactly zero.

3 QuasiLARS: extension of LARS

Note first that in general β_0 of the QLM cannot be removed from equation (1) by location and scale transformations as in the least squares regression. However for any β, the quasi-likelihood function (1) is concave in β_0. Thus we can define the marginal maximizer of β_0 as a function of β. Namely for any β, define

$$\beta_0(\beta) = \underset{\beta_0}{\operatorname{argmax}} \sum_{i=1}^{n} Q\left(g^{-1}\left(\beta_0 + \boldsymbol{x}_i^T\beta\right), y_i\right). \tag{3}$$

We denote $R(\beta,\beta_0) = \sum_{i=1}^{n} Q\left(g^{-1}\left(\beta_0+\boldsymbol{x}_i^T\beta\right), y_i\right)$ and $R(\beta) = R(\beta, \beta_0(\beta))$.

Based on the above observation that the LAR produces a solution path along which the objective function's first-order partial derivatives have the same absolute value for each active predictor variable, our extension QuasiLAR seeks a solution path $\beta(t)$ such that $\left|\frac{\partial}{\partial\beta_j}R(\beta,\beta_0(\beta))\,|_{\beta(t)}\right| = \left|\frac{\partial}{\partial\beta_{j'}}R(\beta,\beta_0(\beta))\,|_{\beta(t)}\right|$ for j and j' that are active at t. More explicitly, at any t, the solution should move in a special direction $\boldsymbol{\alpha}(\beta(t)) = \frac{d}{dt}\beta(t)$, which is chosen in a way to ensure that the first-order partial derivatives $\frac{\partial}{\partial\beta_j}R(\beta)$ have the same absolute value for each active predictor variable j.

For $R(\beta)$, denote its vector of first-order partial derivatives by $\boldsymbol{b}(\beta) = (b_1(\beta), \cdots, b_p(\beta))^T$ and matrix of second-order partial derivatives by $\boldsymbol{M}(\beta) = (m_{jk}(\beta))_{1\le j,k\le p}$, where $b_j(\beta) = \frac{\partial}{\partial\beta_j}R(\beta)$ and $m_{jk}(\beta) = \frac{\partial^2}{\partial\beta_j\partial\beta_k}R(\beta)$ for $1 \le j,k \le p$.

As in LAR, we use t to index our QuasiLAR solution path $\beta(t)$. Denote the active index set at t by $\mathcal{A}(\beta(t))$ and also by $\mathcal{A}_t$. We will use $\mathcal{A}(\beta(t))$ and $\mathcal{A}_t$ interchangeably.

Note that $\boldsymbol{b}(\beta(t + d_t)) \approx \boldsymbol{b}(\beta(t)) + \boldsymbol{M}(\beta(t))\{\beta(t + d_t) — \beta(t)\}$ for small $d_t > 0$ due to Taylor expansion. Thus in order to keep the absolute values of the first-order partial derivatives with respect to all active predictor variables decrease and be the same, our solution path updating direction $\boldsymbol{\beta}(t+d_t)—\boldsymbol{\beta}(t)$ should satisfy that $b_j(\boldsymbol{\beta}(t+d_t))—b_j(\beta(t))$ has the opposite sign of $b_j(\boldsymbol{\beta}(t))$ and has the same absolute value for each $j \in \mathcal{A}_t$. Here the first requirement guarantees that the first-order partial derivatives of active predictor variables are decreasing in absolute value and the second requirement ensures that they decrease at the same speed. This gives our motivation on how to define an appropriate solution path updating direction. The above discussion can be made rigorous by using differential operators when the above $d_t > 0$ is infinitesimal as presented next.

At any t with solution $\boldsymbol{\beta}(t)$, denote the solution path updating direction by $\boldsymbol{a}(\beta(t)) = \left(a_1(\beta(t)), \cdots, a_p(\beta(t))\right)^T = \frac{d}{dt}\beta(t)$. For any inactive variable $j \notin \mathcal{A}_t$, we keep $\beta_j(t) = 0$ and do not change it; thus $a_j(\boldsymbol{\beta}(t)) = 0$ for $j \notin \mathcal{A}_t$. Consequently we only care about $a_j(\boldsymbol{\beta}(t))$ for active predictor $j \in \mathcal{A}_t$. For any two index sets $\mathcal{A}$ and $\mathcal{B}$, vector $\boldsymbol{a}$, and matrix $\boldsymbol{M}$, denote $\boldsymbol{a}_{\mathcal{A}}$ to be the sub-vector of $\boldsymbol{a}$ consisting of those elements with index in $\mathcal{A}$ and $\boldsymbol{M}_{\mathcal{A},\mathcal{B}}$ to be the sub-matrix of $\boldsymbol{M}$ consisting of those elements with row index in $\mathcal{A}$ and column index in $\mathcal{B}$. When $\mathcal{A} = \{j\}$ and $\mathcal{B} = \{k\}$ are singletons, we also write $\boldsymbol{M}_{j,\mathcal{B}}$ and $\boldsymbol{M}_{\mathcal{A},k}$, which are essentially a row vector and a column vector, respectively. Denote the complement of $\mathcal{A}$ by $\mathcal{A}^c = \{1, \cdots, p\} \setminus \mathcal{A}$. With these notations, our solution path updating direction for active predictor variables should satisfy

$$\boldsymbol{M}_{\mathcal{A}_t,\mathcal{A}_t}(\beta(t))\,\boldsymbol{a}_{\mathcal{A}_t}(\beta(t)) = -\operatorname{sign}\left(\boldsymbol{b}_{\mathcal{A}_t}(\beta(t))\right). \tag{4}$$

The argument is based on the previous paragraph with infinitesimal d_t. Thus our solution path should be updated using $\frac{d}{dt}\beta_{\mathcal{A}_t}(t) = \boldsymbol{a}_{\mathcal{A}_t}(\beta(t))$ and $\frac{d}{dt}\beta_{\mathcal{A}_t^c}(t) = 0$ with

$$\boldsymbol{a}_{\mathcal{A}_t}(\beta(t)) = -\left(\boldsymbol{M}_{\mathcal{A}_t,\mathcal{A}_t}(\beta(t))\right)^{-1}\operatorname{sign}\left(\boldsymbol{b}_{\mathcal{A}_t}(\beta(t))\right) \tag{5}$$

being the solution of (4), where the invertibility of $\boldsymbol{M}_{\mathcal{A}_t,\mathcal{A}_t}(\beta(t))$ is not an issue as long as the quasi-likelihood is well defined. Here we use 0 to denote a column vector of zeros with length depending on the context. Note further that this updating scheme implies that $\frac{d}{dt}\boldsymbol{b}_{\mathcal{A}_t}(\beta(t)) = -\operatorname{sign}\left(\boldsymbol{b}_{\mathcal{A}_t}(\beta(t))\right)$ because $\frac{d}{dt}\boldsymbol{b}_{\mathcal{A}}(\beta(t)) = \boldsymbol{M}_{\mathcal{A}_t,\mathcal{A}_t}(\beta(t))\,\boldsymbol{a}_{\mathcal{A}_t}((\beta(t))$.

In integration format, they become

$$\beta_{\mathcal{A}_t}(t+d_t) = \beta_{\mathcal{A}_t}(t) + \int_t^{t+d_t}\boldsymbol{a}_{\mathcal{A}_t}(\beta(\tau))\,d\tau, \text{ and } \beta_{\mathcal{A}_t^c}(t+d_t) = 0, \tag{6}$$

$$\boldsymbol{b}_{\mathcal{A}_t}(\beta(t+d_t)) = \boldsymbol{b}_{\mathcal{A}_t}(\beta(t)) - \int_t^{t+d_t}\operatorname{sign}\left(\boldsymbol{b}_{\mathcal{A}_t}(\beta(\tau))\,d\tau, \right. \tag{7}$$

where $\boldsymbol{a}_{\mathcal{A}_t}(\beta(t))$ is given by (5). Note that we consider small $d_t > 0$ in all the above discussion and assume that between t and $t + d_t$ the active index has not changed. Consequently, beginning at t we may keep updating the solution path using (6) until the active set changes at some $t' > t$. This happens when another predictor variable $j' \notin \mathcal{A}_t$ joins the active set $\mathcal{A}_t$ to share the honor of having the largest absolute value of the first-order partial derivatives, that is, $|b_{j'}(\boldsymbol{\beta}(t'))| = |b_j(\boldsymbol{\beta}(t'))|$ for any active predictor $j \in \mathcal{A}_t$. At this point, we update the active set by setting $\mathcal{A}_{t'} = \mathcal{A}_t \cup \{j'\}$.

Now we present our extension QuasiLAR for the QLM. We initialize our solution path by identifying the predictor variable j so that the objective function $R(\boldsymbol{\beta})$ changes fastest with respect to β_j beginning at $\boldsymbol{\beta} = 0$. We first set

$t_0 = -\max_{j=1,\cdots,p} \left| \frac{\partial}{\partial \beta_j} R(\beta) \right|_{\beta=0} = -\max_{j=1,\cdots,p} |b_j(0)|$. It will be clear later why we choose t_0 in this way. Our solution path begins with $\boldsymbol{\beta}(t_0) = \mathbf{0}$ and $\beta_0(t_0) = \beta_0(\boldsymbol{\beta} t_0))$ defined in terms of (3). The initial active predictor set is given by

$$\mathcal{A}_{t_0} = \left\{ \underset{1 \le j \le p}{\operatorname{argmax}} \left| \frac{\partial}{\partial \beta_j} R(\beta) \right|_{\beta=0} \right\} = \left\{ \underset{1 \le j \le p}{\operatorname{argmax}} |b_j(0)| \right\}.$$

With t_0, $\boldsymbol{\beta}(t_0)$, and $\mathcal{A}_{t_0}$, we update our solution path using (6) until a new variable joins the active set at some $t_1 (> t_0)$ to be determined. That means the solution for any $t > t_0$ may be temporarily updated by $\tilde{\beta}_{\mathcal{A}_{t_0}}(t) = \beta_{\mathcal{A}_{t_0}}(t_0) + \int_{t_0}^{t} \boldsymbol{a}_{\mathcal{A}_{t_0}} \left(\tilde{\beta}(\tau) \right) d\tau$ and $\tilde{\beta}_{\mathcal{A}_{t_0}^c}(t) = 0$. Here $\tilde{\beta}(t)$ is a temporary solution path defined for any $t > t_0$. For any $j \notin \mathcal{A}_{t_0}$, define $T_j = \min \left\{ t > t_0 : \left| b_j \left(\tilde{\beta}(t) \right) \right| \ge \left| b_m \left(\tilde{\beta}(t) \right) \right| \right\}$, where $m \in \mathcal{A}_{t_0}$. Then t_1 is given by $t_1 = \min_{j \notin \mathcal{A}_{t_0}} T_j$ and we call t_1 a *transition point* in that the set of active predictors changes at $t = t_1$.

Then our QuasiLAR algorithm updates by setting $\beta_{\mathcal{A}_{t_0}}(t) = \beta_{\mathcal{A}_{t_0}}(t_0) + \int_{t_0}^{t} \boldsymbol{a}_{\mathcal{A}_{t_0}}(\beta(\tau))\, d\tau$,

Algorithm 1

QuasiLAR for the QLM

Step 1: Initialize by setting $t_0 = -\max_{j=1,\cdots,p} |b_j(\mathbf{0})|$, $\boldsymbol{\beta}(t_0) = \mathbf{0}$, $\boldsymbol{\beta}_0(t_0) = \beta_0(\boldsymbol{\beta}(t_0))$ ad defined in (3), and $\mathcal{A}_{t_0} = \left\{ \underset{1 \le j \le p}{\operatorname{argmax}} |b_j(0)| \right\}$.

Step 2: For $m = 0, \cdots, p-2$, define a tentative solution path using

$$\tilde{\beta}_{\mathcal{A}_{t_m}}(t) = \beta_{\mathcal{A}_{t_m}}(t_m) + \int_{t_m}^{t} \boldsymbol{a}_{\mathcal{A}_{t_m}} \left(\tilde{\beta}(\tau) \right) d\tau \text{ and } \tilde{\beta}_{\mathcal{A}_{t_m}^c}(t) = 0$$

for $t \ge t_m$. Define a new transition point $t_{m+1} = \min_{j \notin \mathcal{A}_{t_m}} T_j$, where $T_j = \min \left\{ t > t_m : \left| b_j \left(\tilde{\beta}(t) \right) \right| \ge \left| b_k \left(\tilde{\beta}(t) \right) \right| \text{ for some } k \in \mathcal{A}_{t_m} \right\}$ for $j \notin \mathcal{A}_{t_m}$. Update solution path by setting $\beta_{\mathcal{A}_{t_m}}(t) = \beta_{\mathcal{A}_{t_m}}(t_m) + \int_{t_m}^{t} \boldsymbol{a}_{\mathcal{A}_{t_m}}(\beta(\tau))\, d\tau$, $\beta_{\mathcal{A}_{t_m}^c}(t) = 0$, and $\beta_0(t) = \beta_0(\boldsymbol{\beta}(t))$ for $t \in [t_m, t_{m+1}]$. $\mathcal{A}_t = \mathcal{A}_{t_m}$ for $t \in [t_m, t_{m+1})$ and $\mathcal{A}_{t_{m+1}} = \mathcal{A}_{t_m} \cup \left\{ j \notin \mathcal{A}_{t_m} : T_j = t_{m+1} \right\}$.

Step 3: At the end of Step 2, $\mathcal{A}_{t_{p-1}}$ should be exactly $\{1, \cdots, p\}$. Next we update solution path using $\beta(t) = \beta\left(t_{p-1}\right) + \int_{t_{p-1}}^{t} \boldsymbol{a}_{\mathcal{A}_{t_{p-1}}}(\beta(\tau))\, d\tau, \beta_0(t) = \beta_0(\beta(t))$, and $\mathcal{A}_t = \{1, \cdots, p\}$ for t between t_{p-1} and $t_p = 0$.

$\beta_{\mathcal{A}_{t_0}^c}(t) = 0$, and $\beta_0(t) = \beta_0(\boldsymbol{\beta}(t))$ for all $t \in [t_0, t_1]$. The active predictor set stays the same for t $\in [t_0, t_1)$, namely $\mathcal{A}_t = \mathcal{A}_{t_0}$. At t_1, we update the active predictor set by setting $\mathcal{A}_{t_1} = \mathcal{A}_{t_0} \cup \left\{ j \notin \mathcal{A}_{t_0} : T_j = t_1 \right\}$. At $t = t_1$, the number of active predictors is two. Due to (5), and the definitions of $\tilde{\beta}_{\mathcal{A}_{t_0}}(t)$, T_j and t_1, $\beta(t_1)$ satisfies

$|b_j(\beta(t_1))| = |b_{j'}(\beta(t_1))| > |b_k(\beta(t_1))|, \forall k \notin \mathcal{A}_{t_1}$, where $j, j' \in \mathcal{A}_{t_1}$. Note further that (7) and the definition of t_0 ensure that $t_1 = -\left|\frac{\partial}{\partial \beta_j} R(\beta)\Big|_{\beta(t_1)}\right| = -|b_j(\beta(t_1))|$ for any $j \in \mathcal{A}_{t_1}$.

Our QuasiLAR algorithm continues with t_1, $\boldsymbol{\beta}(t_1)$, and $\mathcal{A}_{t_1}$. The full algorithm is given by Algorithm 1. Note that at the end of the mth QuasiLAR step, t_m, $\boldsymbol{\beta}(t_m)$, and $\mathcal{A}_{t_m}$ satisfy

$$t_m = -\left|\frac{\partial}{\partial \beta_j} R(\beta)\Big|_{\beta(t_m)}\right| = -|b_j(\beta(t_m))| \text{ for any } j \in \mathcal{A}_{t_m} \text{ and}$$
$$|b_j(\beta(t_m)| = |b_{j'}(\beta(t_m)| > |b_k(\beta(t_m)|, \quad \forall k \notin \mathcal{A}_{t_m}, \text{ for any } j, j' \in \mathcal{A}_{t_m}.$$

Note that at the end of the $(p - 1)$th QuasiLAR step in Step 2 of Algorithm 1, all predictors are active. Then, in Step 3, the QuasiLAR path moves along a direction such that the absolute values of the first-order partial derivatives decrease at the same speed until all the first-order partial derivatives are exactly zero, which happens at $t = 0$. The solution at $t = 0$ exactly corresponds to the full solution of the QLM by solving (1) just like the LAR ends at the full OLS estimate. This completes our QuasiLAR algorithm.

Remark 1

Note that the QuasiLAR instantaneous path updating direction is given by $-\left(\boldsymbol{M}_{\mathcal{A}_t,\mathcal{A}_t}(\beta(t))\right)^{-1} \text{sign}\left(\boldsymbol{b}_{\mathcal{A}_t}(\beta(t))\right)$. For least squares regression, the objective function is exactly quadratic and thus $\boldsymbol{M}_{\mathcal{A}_t,\mathcal{A}_t}$ depends only on the active predictor set $\mathcal{A}_t$, but not on the current solution values $\beta_{\mathcal{A}_t}(t)$. Note that $\text{sign}\left(\boldsymbol{b}_{\mathcal{A}_t}(\beta(t))\right)$ does not change in a small neighborhood of t. This implies that, within a small neighborhood of t, the instantaneous path updating direction is the same for least squares regression. This leads to the piecewise linear solution path of the LAR and Rosset and Zhu (2007) in general.

3.1 Quasi-LASSO modification

Efron et al. (2004) discovered that the LASSO solution path can be obtained by a slight modification of the LAR. Next we make a parallel extension by showing that the Quasi-LAR can be modified to get the whole LASSO regularized quasi-likelihood solution path.

Now consider the LASSO regularized quasi-likelihood in two different formats

$$\min_{\beta_0,\beta} -R(\beta, \beta_0(\beta)) + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{8}$$

$$\min_{\beta_0,\beta} -R(\beta, \beta_0(\beta)) \text{ subject to } \sum_{j=1}^{p} |\beta_j| \le s, \tag{9}$$

which are equivalent with one-to-one correspondence between $\lambda \ge 0$ and $s \ge 0$.

Let $\widehat{\beta}$ be a LASSO solution to (8). Then we can show that the sign of any nonzero component $\widehat{\beta}_j$ must agree with the sign of the current first-order partial derivative $b_j(\widehat{\beta})$. It is given by Lemma 2 in Section 5.

Suppose $t = t^*$ at the end of a QuasiLAR step with a new active set $\mathcal{A}^*$. At the next QuasiLAR step with $t \in [t^*, T]$ for some T to be determined, the QuasiLAR solution path moves along the following tentative solution path

$$\tilde{\beta}_{\mathcal{A}^*}(t) = \beta_{\mathcal{A}^*}(t^*) + \int_{t^*}^{t} \boldsymbol{a}_{\mathcal{A}^*}\left(\tilde{\beta}(\tau)\right) d\tau \text{ and } \tilde{\beta}_{(\mathcal{A}^*)^c}(t) = 0 \tag{10}$$

for $t \geq t^*$. Denote $T_j = \min\left\{t > t^* : \left|b_j\left(\tilde{\beta}(t)\right)\right| \geq \left|b_k\left(\tilde{\beta}(t)\right)\right| \text{ for some } k \in \mathcal{A}^*\right\}$ for $j \notin \mathcal{A}^*$ for $j \notin \mathcal{A}^*$. Then the end point T is given by $\min_{j \notin \mathcal{A}^*} T_j$,

However $\tilde{\beta}_j(t)$ may have changed sign at some point between t^* and T for some $j \in \mathcal{A}^*$, in which case the sign restriction in Lemma 2 must have been violated. We define $S_j = \min\left\{t \in (t^*, \infty) : \tilde{\beta}_j(t) = 0\right\}$ for $j \in \mathcal{A}^*$, where $\tilde{\beta}_j(t)$ is the jth component of $\tilde{\beta}(t)$ defined by (10). If $S = \min_{j \in \mathcal{A}^*} S_j < T, \tilde{\beta}(T)$ defined by (10) cannot be a LASSO quasi-likelihood solution since the sign restriction in Lemma 2 has already been violated. The following Quasi-LASSO modification can be applied to ensure that we can get the LASSO regularized quasi-likelihood solution path.

Quasi-LASSO modification—If $S < T$, stop the ongoing QuasiLAR step at S and remove $\tilde{j}$ from the active set $\mathcal{A}^*$ by set $\mathcal{A}_S = \mathcal{A}_{t^*} \setminus \{\tilde{j}\}$, where $\tilde{j}$ is chosen such that $S_{\tilde{j}} = S$. At the new transition point S, the new path updating direction is calculated based on the new active predictor set $\mathcal{A}^* \setminus \{\tilde{j}\}$.

We have the following theorem to guarantee that the Quasi-LASSO modification leads to the LASSO regularized quasi-likelihood solution path. We name the modified algorithm by QuasiLASSO and use QuasiLARS to refer to both QuasiLAR and QuasiLASSO.

Note that at each transition point of our QuasiLARS solution path, two kinds of event can happen: either an inactive predictor joins the active predictor set or an active predictor is removed from the active predictor set. As in Efron et al. (2004), we assume that a "one at a time" condition holds. With the "one at a time" condition, at each transition point t^*, only one single event can happen, namely either one inactive predictor variable becomes active or one currently active predictor variable becomes inactive.

Theorem 1: With the Quasi-LASSO modification, and assuming the "one at a time" condition, the QuasiLARS algorithm yields the LASSO quasi-likelihood solution path.

Remark 2: Here we make the "one at a time" assumption. However, even when the "one at a time" condition does not hold, a QuasiLASSO solution path is still available. The same discussion in Efron et al. (2004) applies. For practical applications, some slight jittering may simply be applied, if necessary, to ensure the "one at a time" condition.

3.2 Updating via ODE

Our solution path algorithm QuasiLARS involves an essential piecewise updating step $\tilde{\beta}_{\mathcal{A}_{t^*}}(t) = \beta_{\mathcal{A}_{t^*}}(t^*) + \int_{t^*}^{t} \boldsymbol{a}_{\mathcal{A}_{t^*}}\left(\tilde{\beta}(\tau)\right) d\tau$ and $\tilde{\beta}_{\mathcal{A}_{t^*}^c}(t) = 0$ beginning at a transition point t^* with solution $\beta(t^*)$ and active predictor set $\mathcal{A}_{t^*}$. Note that the piecewise updating can be easily achieved by setting $\tilde{\beta}_j(t) = 0$ for $j \notin \mathcal{A}_{t^*}$ and $t > t^*$ and solving the following ODE system

$\frac{d}{dt}\tilde{\beta}_{\mathcal{A}_{t^*}}(t) = \boldsymbol{a}_{\mathcal{A}_{t^*}}\left(\tilde{\beta}(t)\right)$ with initial value condition $\tilde{\beta}_{\mathcal{A}_{t^*}}(t)|_{t=t^*} = \beta_{\mathcal{A}_{t^*}}(t^*)$. This is a standard initial-value ODE system, for which there are many efficient solvers available. We have implemented our QuasiLARS using Matlab ODE solver "ODE45."

4 Details for deriving the path updating direction

Note that the path updating direction defined by (5) asks for $\frac{\partial}{\partial\beta_j}\sum_{i=1}^{n} Q\left(g^{-1}\left(\beta_0+\boldsymbol{x}_i^T\beta\right), y_i\right)$ and $\frac{\partial^2}{\partial\beta_j\partial\beta_{j'}}\sum_{i=1}^{n} Q\left(g^{-1}\left(\beta_0+\boldsymbol{x}_i^T\beta\right), y_i\right)$. By the chain rule, we are required to have the implicit partial derivatives $\frac{\partial}{\partial\beta_j}\beta_0(\beta)$ and $\frac{\partial^2}{\partial\beta_j\partial\beta_{j'}}\beta_0(\beta)$. Next we show how to obtain them.

According to its definition (3), $\beta_0(\beta)$ satisfies

$\frac{\partial}{\partial\beta_0}\sum_{i=1}^{n} Q\left(g^{-1}\left(\beta_0+\boldsymbol{x}_i^T\beta\right), y_i\right) = \sum_{i=1}^{n} Q_1\left(g^{-1}\left(\beta_0+\mathbf{x}_i^T\beta\right), y_i\right)\left(g^{-1}\right)'\left(\beta_0+\boldsymbol{x}_i^T\beta\right) = 0$, where $Q_1(\mu, y) = \frac{\partial}{\partial\mu}Q(\mu, y)$. Now treat β_0 as a function of β and take derivative of each term with respect to β_j. We should get

$$\begin{aligned}&\sum_{i=1}^{n} Q_{11}\left(g^{-1}(\beta_0 + \boldsymbol{x}_i^T\beta\right), \\ &\quad y_i\left[\left(g^{-1}\right)'\left(\beta_0+\boldsymbol{x}_i^T\beta\right)\right]^2\left(x_{ij} + \frac{\partial}{\partial\beta_j}\beta_0\right) \\ &+\sum_{i=1}^{n} Q_1\left(g^{-1}(\beta_0 + \boldsymbol{x}_i^T\beta\right), \\ &\quad y_i\left(g^{-1}\right)''(\beta_0 + \boldsymbol{x}_i^T\beta\left(x_{ij} + \frac{\partial}{\partial\beta_j}\beta_0\right) = 0\end{aligned}$$

, where $Q_{11}(\mu, y) = \frac{\partial^2}{\partial\mu^2}Q(\mu, y)$.

Thus by solving for $\frac{\partial}{\partial\beta_j}\beta_0$, we get $\frac{\partial}{\partial\beta_j}\beta_0(\beta) = -\frac{\sum_{i=1}^{n} x_{ij}c_i(\beta)}{\sum_{i=1}^{n} c_i(\beta)}$, where

$c_i(\beta) = Q_{11}\left(g^{-1}\left(\beta_0(\beta)+\boldsymbol{x}_i^T\beta\right), y_i\right)\left[\left(g^{-1}\right)'\left(\beta_0(\beta)+\boldsymbol{x}_i^T\beta\right)\right]^2 + Q_1\left(g^{-1}\left(\beta_0(\beta)+\boldsymbol{x}_i^T\beta\right), y_i\right)\left(g^{-1}\right)''\left(\beta_0(\beta)+\boldsymbol{x}_i^T\beta\right)$

for $i = 1, \cdots, n$. To get the second-order partial derivatives $\frac{\partial^2}{\partial\beta_j\partial\beta_{j'}}\beta_0(\beta)$, we may apply another layer of differential operator $\frac{\partial}{\partial\beta_{j'}}$.

For some particular generalized linear models, it may be much simpler to get those partial derivatives as shown in the following subsections.

4.1 Binomial

For the Binomial distribution, the data set is given by $\{(x_i, y_i) : i = 1, \cdots, n\}$ with $y_i \in \{0, 1\}$. With the canonical logit link $\eta(x) = \log(\mu(x)/(1-\mu(x)))$, the corresponding loglikelihood

function is given by $L(\beta,\beta_0)=\sum_{i=1}^{n}\left(y_i\left(\boldsymbol{x}_i^T\beta+\beta_0\right)-\log\left(1+e^{\boldsymbol{x}_i^T\beta+\beta_0}\right)\right)$. Then for any β, the corresponding optimal $\beta_0(\beta)$ is given by the solution of $\sum_{i=1}^{n}y_i-\sum_{i=1}^{n}\frac{e^{\boldsymbol{x}_i^T\beta+\beta_0}}{1+e^{\boldsymbol{x}_i^T\beta+\beta_0}}=0$ which is equivalent to $\sum_{i=1}^{n}(1-y_i)-\sum_{i=1}^{n}\left(1+e^{\boldsymbol{x}_i^T\beta+\beta_0}\right)^{-1}=0$. We next differentiate both sides with respect to β_j and solve for $\frac{\partial}{\partial\beta_j}\beta_0$ to get $\frac{\partial}{\partial\beta_j}\beta_0=\left(\sum_{i=1}^{n}\frac{x_{ij}e^{\boldsymbol{x}_i^t\beta+\beta_0(\beta)}}{\left(1+e^{\boldsymbol{x}_i^t\beta+\beta_0(\beta)}\right)^2}\right)/\left(\sum_{i=1}^{n}\frac{e^{\boldsymbol{x}_i^t\beta+\beta_0(\beta)}}{\left(1+e^{\boldsymbol{x}_i^t\beta+\beta_0(\beta)}\right)^2}\right)$. We may take another layer of differentiation to get second-order partial derivatives.

4.2 Poisson

In the case of Poisson distribution with the canonical log link $\eta(x) = \log\mu(x)$, the likelihood function is given, up to a constant, by $L(\beta,\beta_0)=\sum_{i=1}^{n}\left[-e^{\boldsymbol{x}_i^T\beta+\beta_0}+y_i\left(\boldsymbol{x}_i^T\beta+\beta_0\right)\right]$. For any β, the maximizer $\beta_0(\beta)$ of is given by $\beta_0(\beta)=\log\left(\sum_{i=1}^{n}y_i\right)-\log\left(\sum_{i=1}^{n}e^{\boldsymbol{x}_i^T\beta}\right)$. We may take differentiation to get partial derivatives $\frac{\partial}{\partial\beta_j}\beta_0(\beta)$ and $\frac{\partial^2}{\partial\beta_j\partial\beta_{j'}}\beta_0(\beta)$.

With the closed-form formula of $\beta_0(\beta)$, we may simply plug it into the likelihood and get $L(\beta)\equiv L(\beta,\beta_0(\beta))=-\sum_{i=1}^{n}y_i+\sum_{i=1}^{n}y_i\boldsymbol{x}_i^T\beta-\left(\sum_{i=1}^{n}y_i\right)\log\left(\sum_{i=1}^{n}e^{\boldsymbol{x}_i^T\beta}\right)+\left(\sum_{i=1}^{n}y_i\right)\log\left(\sum_{i=1}^{n}y_i\right)$, which corresponds to our notation $R(\beta)$.

5 Properties of QuasiLARS

We next establish some properties of QuasiLARS solution path and prove Theorem 1.

With the "one at a time" condition, at each transition point t^*, only one single event can happen, namely either one inactive predictor variable becomes active or one currently active predictor variable becomes inactive. For the first type of event, it means the active set changes from $\mathcal{A}$ to $\mathcal{A}^*=\mathcal{A}\cup\{j^*\}$ for some $j^*\notin\mathcal{A}$. We next show in Lemma 1 that this new active variable j^* joins in a "correct" manner. This is the key result for proving Theorem 1. Lemma 1 applies to QuasiLARS (both QuasiLAR and QuasiLASSO).

Lemma 1

For any transition point t^* during the QuasiLARS solution path, if predictor variable *j^* is the only addition to the active set at t^* with solution $\beta(t^*)$ and active set changing from $\mathcal{A}$ to $\mathcal{A}^*=\mathcal{A}\cup\{j^*\}$, then the path updating direction $\alpha(\beta(t^*))$ at t^* has its j^**th component agreeing in sign with the current first-order partial derivative $b_{j^*}(\beta(t^*))$.

Our next four lemmas concern properties of the LASSO regularized quasi-likelihood solution. These lemmas will lead to the proof of Theorem 1. For any $s \geq 0$, we denote the solution of (9) by $\widehat{\beta}=\widehat{\beta}(s)$, which is unique for each s and continuous in s. The uniqueness is due to the convexity of $\sum_{j=1}^{p}|\beta_j|$ and the strict convexity of $-R(\boldsymbol{\beta},\beta_0(\boldsymbol{\beta}))$. Throughout the paper, the hat notation always refers to the LASSO regularized quasi-likelihood solution. For any $s \geq 0$, let $\mathcal{N}_s\equiv\mathcal{N}\left(\widehat{\beta}(s)\right)\triangleq\left\{j:\widehat{\beta}_j(s)\neq 0\right\}$ denote the index set of nonzero components

of $\widehat{\beta}(s)$. We will show that the nonzero set $\mathcal{N}_s$ is also the active predictor set that determines the QuasiLARS path updating direction.

Let $\widehat{\beta}$ be a solution of (8). Next we can show that the sign of any non-zero component $\widehat{\beta}_j$ must agree with the sign of the current first-order partial derivative, namely $\text{sign}\left(\widehat{\beta}_j\right)=\text{sign}\left(b_j\left(\widehat{\beta}\right)\right)$ for $j \in \mathcal{N}\left(\widehat{\beta}\right)$.

Lemma 2

A LASSO regularized quasi-likelihood solution $\widehat{\beta}$ *to* (8) *satisfies* $\text{sign}\left(\widehat{\beta}_j\right)=\text{sign}\left(b_j\left(\widehat{\beta}\right)\right)$ *for any* $j \in \mathcal{N}\left(\widehat{\beta}\right)$.

Let $\mathcal{S}$ be an open interval of the s axis, with infimum $\underline{s}$, within which the nonzero set $\mathcal{N}_s$ of the corresponding LASSO regularized quasi-likelihood solution $\widehat{\beta}(s)$ remains constant, namely, $\mathcal{N}_s=\mathcal{N}$ for $s \in \mathcal{S}$ and some $\mathcal{N}$.

Lemma 3

For $s \in \left\{\underline{s}\right\} \cup \mathcal{S}$, *the LASSO regularized quasi-likelihood estimate* $\widehat{\beta}(s)$ updates along the QuasiLARS path updating direction.

Lemma 4

For an open interval $\mathcal{S}$ with a constant nonzero set $\mathcal{N}$ over the LASSO regularized quasi-likelihood solution path $\widehat{\beta}(s)$, *let* $\underline{s}=\inf(\mathcal{S})$. *Then for* $s \in \mathcal{S} \cap \left\{\underline{s}\right\}$, *the first-order partial derivatives of* $R(\boldsymbol{\beta}, \beta_0(\boldsymbol{\beta}))$ *at* $\widehat{\beta}(s)$ *must satisfy* $|b_j\left(\widehat{\beta}(s)\right)|=\max_{l=1,\cdots,p}|b_1\left(\widehat{\beta}(s)\right)|$ *for* $j \in \mathcal{N}$ *and* $|b_j\left(\widehat{\beta}(s)\right)| \le \max_{l=1,\cdots,p}|b_1\left(\widehat{\beta}(s)\right)|$ *for* $j \notin \mathcal{N}$.

Let $\underline{s}$ denote such a point, $\underline{s}=\inf(\mathcal{S})$ as in Lemma 4, with the LASSO regularized quasi-likelihood solution $\widehat{\beta}$, current derivatives $b_j\left(\widehat{\beta}\right)$, and maximum absolute derivative $\widehat{D}\left(\widehat{\beta}\right)=\max_j|b_j\left(\widehat{\beta}\right)|$. Define $\mathcal{A}_1=\left\{j:\widehat{\beta}_j \neq 0\right\}$, $\mathcal{A}_0=\left\{j:\widehat{\beta}_j=0 \text{and} |b_j\left(\widehat{\beta}\right)|=\widehat{D}\left(\widehat{\beta}(s)\right)\right\}$, and $\mathcal{A}_{10}=\mathcal{A}_1 \cup \mathcal{A}_0$. Define $\beta^{(\gamma)}=\widehat{\beta}+\gamma d$ for some $d \in \mathbb{R}^p$, $T(\gamma) = R(\boldsymbol{\beta}^{(\gamma)}, \beta_0(\boldsymbol{\beta}^{(\gamma)}))$ and $S(\gamma)=\sum_{j=1}^{p}|\beta_j^{(\gamma)}|$. Denote $\dot{S}(\gamma)=\frac{d}{d\gamma}S(\gamma)$, $\dot{T}(\gamma)=\frac{d}{d\gamma}T(\gamma)$, and $\ddot{T}(\gamma)=\frac{d^2}{d\gamma^2}T(\gamma)$.

Lemma 5

At $\underline{s}$, *we have*

$$Z(\boldsymbol{d}) = \dot{T}(0) / \dot{S}(0) \le \widehat{D}\left(\widehat{\beta}\right), \tag{11}$$

with equality only if $d_j = 0$ *for* $j \in \mathcal{A}_{10}^c$ *and* $sign\left(d_j\right)=sign\left(b_j\left(\widehat{\beta}\right)\right)$ *for* $j \in \mathcal{A}_0$. *If so,*

$$\ddot{T}(0)=\boldsymbol{d}_{\mathcal{A}_{10}}^T \boldsymbol{M}_{\mathcal{A}_{10},\mathcal{A}_{10}}\left(\widehat{\beta}\right)\boldsymbol{d}_{\mathcal{A}_{10}}. \tag{12}$$

One implication of Lemma 5 is that, at any transition point, the active predictor set of the LASSO regularized quasi-likelihood solution is a subset of $\mathcal{A}_{10}$. Note that the LASSO regularized quasi-likelihood minimizes $-R(\boldsymbol{\beta}, \beta_0(\boldsymbol{\beta}))$ subject to a constraint on the one norm of $\boldsymbol{\beta}$. Locally around $\widehat{\beta}$, we are maximizing $T(\gamma)$ subject to an upper bound on $S(\gamma)$. The first part of Lemma 5 implies that the instantaneous relative changing rate of $T(\gamma)$ and $S(\gamma)$ is at most $\widehat{D}(\widehat{\beta})$. For $\boldsymbol{\beta}^{(\gamma)}$, its one norm $S(\gamma)$ is increasing in γ as long as $\sum_{j\in\mathcal{A}_1} \operatorname{sign}\left(b_j(\widehat{\beta})\right) d_j + \sum_{j\in\mathcal{A}_0} |d_j| + \sum_{j\in\mathcal{A}_{10}^c} |d_j| > 0$ and the best instantaneous relative changing rate is achieved whenever $d_j = 0$ for $j \in \mathcal{A}_{10}^c$ and $\operatorname{sign}(d_j) = \operatorname{sign}\left(b_j(\widehat{\beta})\right)$ for $j \in \mathcal{A}_0$. Note that $j \in \mathcal{A}_0$ is the same to say that the jth predictor variable is changing from inactive to active. Then, with the "one at a time" condition, the set $\mathcal{A}_0$ is singleton and the requirement $\operatorname{sign}(d_j) = \operatorname{sign}\left(b_j(\widehat{\beta})\right)$ for $j \in \mathcal{A}_0$ is thus guaranteed for our LARS path updating direction due to Lemma 1.

The second part of Lemma 5 provides a closer look at the relative changing rate by checking the second-order derivative $\ddot{T}(0)$. As we only care about direction, we assume that $\dot{S}(0) = \sum_{j\in\mathcal{A}_1} \operatorname{sign}\left(b_j(\widehat{\beta})\right) d_j + \sum_{j\in\mathcal{A}_0} |d_j| = \Delta$ for some fixed $\Delta > 0$. Note that $T(\gamma) = T(0) + \dot{T}(0)\gamma + \frac{1}{2}\ddot{T}(0)\gamma^2 + o(\gamma^2)$. Then we need to find the best direction $\boldsymbol{d}$ to maximize $T(\gamma)$ among all the possible direction $\boldsymbol{d}$ satisfying $\sum_{j\in\mathcal{A}_1} \operatorname{sign}\left(b_j(\widehat{\beta})\right) d_j + \sum_{j\in\mathcal{A}_0} |d_j| = \Delta$ and $\operatorname{sign}(d_j) = \operatorname{sign}\left(b_j(\widehat{\beta})\right)$ for $j \in \mathcal{A}_0$. By taking the second-order information into account, we need to solve

$$\max \quad \boldsymbol{d}_{\mathcal{A}_{10}}^T \boldsymbol{M}_{\mathcal{A}_{10},\mathcal{A}_{10}}(\widehat{\beta})\boldsymbol{d}_{\mathcal{A}_{10}} \tag{13}$$

for some $\Delta > 0$ to select the optimal solution updating direction $\boldsymbol{d}$. As we only care about direction, $\Delta > 0$ can be any number. Our next lemma shows that the optimal direction corresponding to (13) is exactly given by our QuasiLARS path updating direction.

Lemma 6

Our QuasiLARS path updating direction matches the direction corresponding to the solution to (13).

6 QuasiLARS in Action

In this section, we apply QuasiLARS to different datasets with different models. In our implementation, we first calculate t_0. Then set $\delta_t = -t_0/K$ with a large positive K. For our numerical examples, we set $K = 2000$. In addition to the transition points t_{ks}, we evaluate the solution over our solution path at a grid of size δ_t. More specifically, for each piece of our solution path over $[t_k, t_{k+1}]$, we calculate our solution $\boldsymbol{\beta}(t)$ at $t = t_k + m\delta_t$ for $m = 1, \cdots, \lfloor \frac{t_{k+1} - t_k}{\delta_t} \rfloor$, where $\lfloor a \rfloor$ denotes the integer part of a.

The first toy example with a Poisson distribution is used to demonstrate that the LASSO regularized quasi-likelihood does have a nonlinear solution path. In Example 2, we consider Diabetes data with Gaussian distribution trying to compare QuasiLARS and LARS. The response of the Diabetes data is actually positive integer valued, and thus can be thought of

coming from some Poisson model. In Example 3, we apply QuasiLARS with Poisson distribution to the Diabetes data. Binomial QuasiLARS is considered in Example 4 with the Wisconsin Diagnostic Breast Cancer (WDBC) Data (available online at http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)).

Example 1 (A Poisson toy example) We set $p = 3$ and $n = 40$. The predictor covariates are generated from $\boldsymbol{X} \sim N(\mathbf{0}, \sum)$, where $\sum$ is the variance-covariance matrix with its (i, j) element being 1 if $i = j$ and 0.9 otherwise. Conditional on $\boldsymbol{X} = (x_1, x_2, x_3)^T$, the response is generated from a Poisson distribution with mean $\exp(4+3x_1-5x_2+x_3)$. We apply our QuasiLARS with the canonical link function $\eta(x) = \log\mu(x)$ and the identity variance function $V(\mu(x)) = \mu(x)$ of the Poisson distribution.

For this toy example, the QuasiLAR and QuasiLASSO lead to the same solution path. In the top panel of Figure 2, we plot our solution path by solid lines. The horizontal axis corresponds to the one norm of $\boldsymbol{\beta}(t)$. If you connect the solutions at different transition points by straight lines, then you get the dashed lines. It clearly demonstrates that the true solution path for the LASSO regularized quasi-likelihood is not piecewise linear. In the bottom panel, the solution $\boldsymbol{\beta}(t)$ is plotted with respect to t.

Example 2 (Gaussian with Diabetes data) In this example, we use the diabetes data (Efron et al., 2004) to compare the solution path of our extension QuasiLARS and that of the original LARS algorithm. In this data set, ten baseline variables, age, sex, body mass index, average blood pressure, and six blood serum measurements were obtained for each of $n = 442$ diabetes patients, as well as the response of interest, a quantitative measure of disease progression one year after baseline. We run our extension QuasiLAR algorithm for this data set. Our QuasiLAR solution path matches the LAR solution path obtained by the R package LARS which is shown in the top left panel of Figure 1. The maximum solution difference at all transition points is very small and in fact bounded from above by 5.0×10^{-7}, namely, $\max_{j=1}^{p}\max_{m=1}^{10}|\beta_j^{LAR}(t_m) - \beta_j^{QuasiLAR}(t_m)|<5.0\times10^{-7}$, where $\boldsymbol{\beta}^{LAR}$ and $\boldsymbol{\beta}^{QuasiLAR}$ denote the LAR and QuasiLAR solution, respectively. Comparing to QuasiLAR, the QuasiLASSO solution path has two more transition points. This is consistent with the result of R package LARS. With LASSO, the maximum solution difference at all transition points is also bounded from above as $\max_{j=1}^{p}\max_{m=1}^{12}|\beta_j^{LASSO}(t_m) - \beta_j^{QuasiLASSO}(t_m)|<9.0\times10^{-7}$. This example confirms that the QuasiLARS matches the LARS in the Gaussian case and works correctly. However to save space, we do not plot our QuasiLAR and QuasiLASSO paths.

Example 3 (Poisson with Diabetes data) The response in the diabetes data is in fact positive integer valued. We apply our QuasiLARS algorithm by choosing Poisson distribution with the canonical log link function and identity variance function, namely, $\eta(x) = \log\mu(x)$ and $V(\mu(x)) = \mu(x)$. Results are shown in Figure 3. As in the Gaussian example, some discrepancy between the QuasiLAR and QuasiLASSO solution paths is observed. The QuasiLASSO has four more transition points than the QuasiLAR does.

Example 4 (Binomial with WDBC Data) The WDBC data is based on $n = 569$ patients. The number of predictors is $p = 30$. The response is binary in that each patient is diagnosed either as malignant ($Y = 1$) or benign ($Y = 0$). We first standardize each predictor variable to have mean zero and variance one. Our QuasiLARS with Binomial distribution is applied to this data set with the logit link $\log\frac{\mu(x)}{1-\mu(x)}=\eta(x)$ and variance function $V(\mu(x)) = \mu(x)(1-\mu(x))$. There are a lot of predictor variables available. To locate an "optimal" solution along the QuasiLARS solution path, we use the Bayesian Information Criterion (BIC) defined by

$\mathrm{BIC}(\beta(t)) = -2\sum_{i=1}^{n}\log L(x_i, y_i, \beta_0(\beta(t))) + (\log n)\, k(\beta(t))$, where $L(x_i, y_i; \boldsymbol{\beta}(t), \beta_0(\boldsymbol{\beta}(t)))$ denotes the Binomial likelihood and $k(\boldsymbol{\beta}(t)) = \#\{1 \le j \le p : |\beta_j(t)| > 0\}$ denotes the number of nonzero coefficients of $\boldsymbol{\beta}(t)$. For the LASSO regularized least squares regression, Zou, Hastie and Tibshirani (2007) proved that the number of nonzero coefficients is an unbiased and asymptotically consistent estimator of the degrees of freedom. Park and Hastie (2007) provided a heuristic proof for the case of generalized linear models. The optimal solution is given by $\boldsymbol{\beta}(t^*)$ with $t^* = \mathrm{argmin}_{t\in[t_0,0]}\,\mathrm{BIC}(\boldsymbol{\beta}(t))$.

For the QuasiLAR, nonzero elements of the optimal solution are given by the second column of Table 1. The BIC score is plotted with respect to the solution's one norm $\sum_{j=1}^{30}|\beta_j(t)|$ in the top left panel of Figure 4 for $t \in [t_0, T]$, where T is a little beyond t^* corresponding to the "optimal" solution. The top right panel of Figure 4 gives the solution path for $t \in [t_0, T]$. Here we truncate the figures at T to make it look more clear.

For the QuasiLASSO, the optimal solution's nonzero elements are shown in the third column of Table 1. The corresponding plots of the BIC and solution path are given in the two bottom panels of Figure 4.

From the QuasiLAR solution path given in the top right panel of Figure 4, we can see that one solution component has changed sign between the second and third transition points. This change causes the violation of the sign constraint of the LASSO regularized quasi-likelihood solution path. Thus in the QuasiLASSO solution path, another transition is added at this point to avoid sign constraint violation.

This example demonstrates that our extension QuasiLARS may be applied to high dimensional data sets. However there is no need to complete the whole solution path. We may design an optimal criterion, say the BIC. This optimal criterion may be used to identify the optimal solution as the QuasiLARS solution path progresses. Thus an earlier termination is possible to save computational effort in that it is computationally expensive to solve the ODE system when the active predictor set is large.

7 Conclusion

In this work, we extend the LARS algorithm to the QLM. Over each piece, the solution path is obtained by solving an initial-value ordinary differential equation system. Several examples are used to demonstrate how it works with real data. In particular, Example 4 uses the BIC to select the "optimal solution" along the solution path to show that the QuasiLARS algorithm may be applied to high dimensional data and an earlier termination is possible. One interesting future research topic is to study how to define degrees of freedom for the QuasiLARS as studied in Zou et al. (2007). This will provide an elegant criterion to select the "optimal solution."

The LARS is attractive because of its super fast speed. This is made possible because the corresponding path is piecewise linear. However the QuasiLARS solution path is not piecewise linear due to the nature of the QLM. Thus we can not expect the QuasiLARS to be as fast as the LARS. We have implemented the primitive version of our algorithm using the Matlab ODE solver "ODE45," which works fairly fast.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The author thanks Jianqing Fan and Chuanshu Ji for mentoring and longtime encouragement. The author also thanks Dennis Boos, Jingfang Huang, Yufeng Liu, John Monahan, and Leonard Stefanski for helpful comments and discussions. This work is supported in part by NSF grant DMS-0905561, NIH/NCI grant R01-CA149569, and NCSU Faculty Research and Professional Development Award.

References

Candes E, Tao T. The dantzig selector: statistical estimation when *p* is much larger than *n*. The Annals of Statistics. 2007:2313–2351.

Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression (with discussions). Ann. Statist. 2004; 32:409–499.

Fan J, Li R. Variable selection via penalized likelihood. Journal of American Statistical Association. 2001; 96:1348–1360.

Friedman J, Hastie T, Tibshirani R. Regularized paths for generalized linear models via coordinate descent. Journal of Statistical Software. 2010; 33

Hastie T, Rosset S, Tibshirani R, Zhu J. The entire regularization path for the support vector machine. JMLR. 2004; 5:1391–1415.

Hunter DR, Li R. Variable selection using mm algorithm. The Annals of Statistics. 2005; 33:1617–1642.

Li Y, Liu Y, Zhu J. Quantile regression in reproducing kernel hilbert spaces. Journal of the American Statistical Association. 2007; 102:255–268.

Li Y, Zhu J. l_1-norm quantile regressions. Journal of Computational and Graphical Statistics. 2008; 17:163–185.

Madigan D, Ridgeway G. Discussion on “least angle regression”. Annals of Statistics. 2004; 32:465–469.

Osborne M, Presnell B, Turlach B. A new approach to variable selection in least squares problems. IMA Journal of Numerical Analysis. 2000; 20:389–403.

Park MY, Hastie T. l_1-regularization path algorithm for generalized linear models. Journal of the Royal Statistical Society, Series B. 2007; 69:659–677.

Rocha G, Zhao P, Yu B. A path following algorithm for sparse pseudo-likelihood inverse covariance estimation (splice). Technical Report. 2008

Rosset S. Tracking curved regularized opimization solution paths. Advances in Neural Information Processing Systems. 2004; 13

Rosset S, Zhu J. Piecewise linear regularized solution paths. Annals of Statistics. 2007; 35:1012–1030.

Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B. 1996; 58:267–288.

Wang J, Shen X, Liu Y. Probability estimation for large margin classifiers. Biometrika. 2008; 95:149–167.

Wang L, Shen X. Multicategory support vector machines, feature selection and solution path. Statistica Sinica. 2006; 16:617–634.

Wang L, Zhu J. Image denoising via solution paths. Annals of Operations Research (Special issue on data mining). 2007

Wu S, Shen X, Geyer C. Adaptive regularization through entire solution surface. Biometrika. 2009:513–527.

Yuan M, Lin Y. On the nonnegative garrote estimator. Journal of the Royal Statistical Society, Series B. 2007; 69:143–161.

Yuan M, Zou H. Efficient global approximation of generalized nonlinear l1 regularized solution paths and its applications. JASA. 2009:1562–1574.

Zhang HH, Lu W. Adaptive lasso for cox’s proportional hazards model. Biometrika. 2007; 94:691–703.

Zhu J, Rosset S, Hastie T, Tibshirani R. 1-norm support vector machines. Neural Information Processing Systems. 2004; 16

Zou H. The adaptive lasso and its oracle properties. Journal of the American Statistical Association. 2006; 101:1418–1429.

Zou H. A note on path-based variable selection in the penalized proportional hazards model. Biometrika. 2008; 95:241–247.

Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B. 2005; 67:301–320.

Zou H, Hastie T, Tibshirani R. On the degrees of freedom of the lasso. The Annals of Statistics. 2007:2173–2192.

Zou H, Li R. One-step sparse estimates in nonconcave penalized likelihood models (with discussion). Ann. Statist. 2008; 36:1509–1566.

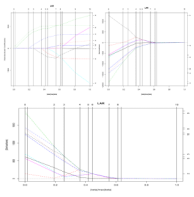

Figure 1.
LAR solution path of the diabetes data: the top left panel gives the solution path of the LAR; the top right panel and the bottom panel plot the derivatives of (2) and their absolute values, respectively, along the LAR solution path.

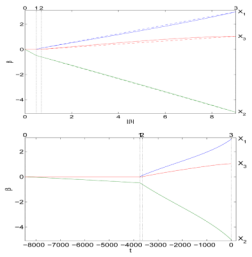

Figure 2.
Poisson QuasiLARS solution path of the toy example: the top panel gives the solution path of the Poisson QuasiLARS with respect to the one norm of $\boldsymbol{\beta}(t)$; the bottom panel plotted with respect to t.

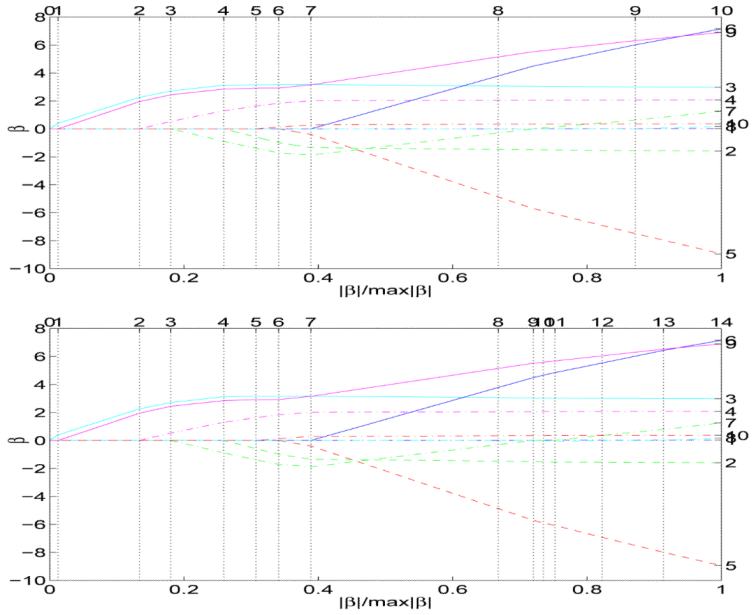

Figure 3.
Poisson QuasiLAR (top) and QuasiLASSO (bottom) solution paths for the Diabetes data.

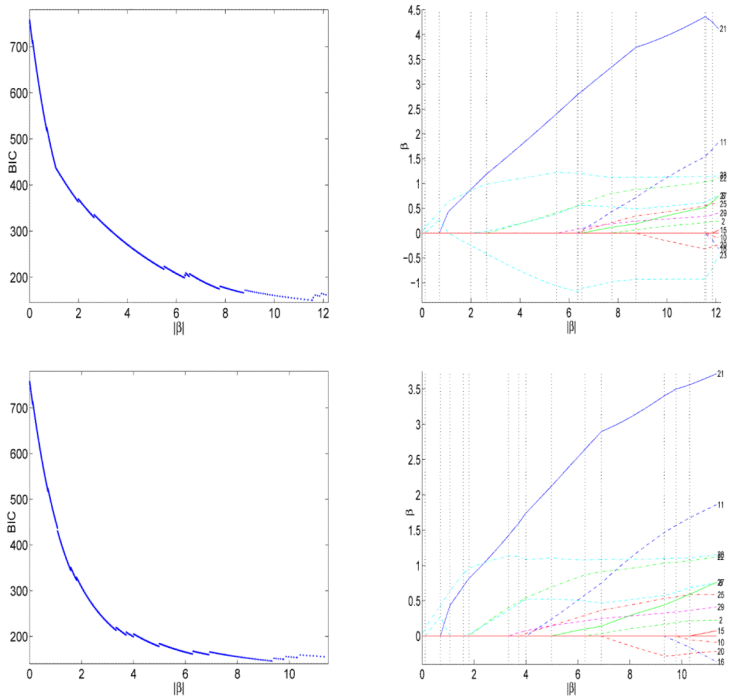

Figure 4.
Binomial QuasiLARS paths for the WDBC data: the top left panel plots the BIC score along the Binomial QuasiLAR solution path; the top right panel gives part of the Binomial QuasiLAR solution path; the bottom left panel plots the BIC score along the Binomial QuasiLASSO solution path; the bottom right panel gives part of the Binomial QuasiLASSO solution path;

Table 1

Nonzero elements of the optimal solution selected by BIC for Example 3

	QuasiLAR	QuasiLASSO
β_2	0.2077	0.1624
β_8	0.6170	0.5767
β_{11}	1.5370	1.4667
β_{20}	−0.3169	−0.2833
β_{21}	4.3576	3.4047
β_{22}	1.0325	1.0343
β_{23}	−0.9287	
β_{25}	0.5470	0.5339
β_{27}	0.5176	0.4395
β_{28}	1.1496	1.0998
β_{29}	0.3378	0.3257