A Generic Path Algorithm for Regularized Statistical Estimation

Hua ZHOU and Yichao WU

Regularization is widely used in statistics and machine learning to prevent overfitting and gear solution toward prior information. In general, a regularized estimation problem minimizes the sum of a loss function and a penalty term. The penalty term is usually weighted by a tuning parameter and encourages certain constraints on the parameters to be estimated. Particular choices of constraints lead to the popular lasso, fused-lasso, and other generalized ℓ_1 penalized regression methods. In this article we follow a recent idea by Wu and propose an exact path solver based on ordinary differential equations (EPSODE) that works for any convex loss function and can deal with generalized ℓ_1 penalties as well as more complicated regularization such as inequality constraints encountered in shape-restricted regressions and nonparametric density estimation. Nonasymptotic error bounds for the equality regularized estimates are derived. In practice, the EPSODE can be coupled with AIC, BIC, C_p or cross-validation to select an optimal tuning parameter, or provide a convenient model space for performing model averaging or aggregation. Our applications to generalized ℓ_1 regularized generalized linear models, shape-restricted regressions, Gaussian graphical models, and nonparametric density estimation showcase the potential of the EPSODE algorithm. Supplementary materials for this article are available online.

KEY WORDS: Gaussian graphical model; Generalized linear model; Lasso; Log-concave density estimation; Ordinary differential equations; Quasi-likelihoods; Regularization; Shape restricted regression; Solution path.

1. INTRODUCTION

In this article, we consider a general regularization framework

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} f(\boldsymbol{\beta}) + \rho\|\boldsymbol{V\beta}-\boldsymbol{d}\|_1 + \rho\|\boldsymbol{W\beta}-\boldsymbol{e}\|_+, \tag{1}$$

for which we propose an efficient exact path solver based on ordinary differential equations (EPSODE). Here $f:\mathbb{R}^p \mapsto \mathbb{R}$ is a convex, smooth function of $\boldsymbol{\beta}\in\mathbb{R}^p$, where $p>0$ is the dimensionality of the parameters. For any vector $\boldsymbol{v}=(v_i)$, $\|\boldsymbol{v}\|_1=\sum_i |v_i|$ denotes its ℓ_1 norm and $\|\boldsymbol{v}\|_+=\sum_i \max\{v_i,0\}$ is the sum of positive parts of its components. ρ is the regularization tuning parameter and the two regularization terms embodied by the constant matrices $(\boldsymbol{V},\boldsymbol{W})$ and vectors $(\boldsymbol{d},\boldsymbol{e})$ enforce equality and inequality constraints among the parameters, respectively, as explained below. The EPSODE provides the exact solution path to (1) as the tuning parameter ρ varies.

1.1 Generality of (1)

The generality of (1) is two-fold. First, f can by any convex loss or other types of objective functions. For example, it can be the negative log-likelihood function of GLMs, negative quasi-likelihood, the exponential loss function of the AdaBoost (Friedman, Hastie, and Tibshirani 2000), or many other frequently used loss functions in statistics and machine learning. Second, we allow $\boldsymbol{V}$ and $\boldsymbol{W}$ to be any regularization matrices of p columns. This leads to broad applications. In particular, the first regularization term $\rho\|\boldsymbol{V\beta}-\boldsymbol{d}\|_1$ encourages equality constraints $\boldsymbol{V\beta}=\boldsymbol{d}$. When ρ is large enough, the minimizer $\boldsymbol{\beta}(\rho)$ of (1) satisfies $\boldsymbol{V\beta}(\rho)=\boldsymbol{d}$. For instance, when $\boldsymbol{V}$ is the identity matrix and $\boldsymbol{d}=\boldsymbol{0}$, it recovers the well-known lasso regression (Donoho and Johnstone 1994; Tibshirani 1996) which encourages sparsity of the estimates. When

$$\boldsymbol{V}=\begin{pmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{pmatrix}$$

and $\boldsymbol{d}=\boldsymbol{0}$, it corresponds to the fused-lasso penalty (Tibshirani et al. 2005), which leads to smoothness among neighboring regression coefficients. As we will show later, more complicated equality constraints can be incorporated with properly designed $\boldsymbol{V}$ and $\boldsymbol{d}$. On the other hand, the second regularization term $\rho\|\boldsymbol{W\beta}-\boldsymbol{e}\|_+$ enforces regularization by inequality relations among regression coefficients. For large enough ρ, the minimizer $\boldsymbol{\beta}(\rho)$ satisfies $\boldsymbol{W\beta}(\rho)\le\boldsymbol{e}$. For instance, setting $\boldsymbol{W}$ as the negative identity matrix and $\boldsymbol{e}=\boldsymbol{0}$ encourages nonnegativity of the estimates, as required in nonnegative least squares problems (Lawson and Hanson 1987). In the isotonic regression (Robertson, Wright, and Dykstra 1988; Silvapulle and Sen 2005), the estimates have to be nondecreasing. This can be achieved by the regularization matrix

$$\boldsymbol{W}=\begin{pmatrix} 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \end{pmatrix}$$

and $\boldsymbol{e}=\boldsymbol{0}$. More complicated constraints that occur in shape-restricted regression and nonparametric regressions also can be incorporated as we demonstrate in later examples.

In certain applications, both equality and inequality regularizations are required. In that case, as shown in Section 2, at a large but finite ρ, the minimizer $\boldsymbol{\beta}(\rho)$ coincides with the solution to the following linearly constrained optimization problem

$$\min \quad f(\boldsymbol{\beta}) \ \text{ subject to } \ \boldsymbol{V\beta}=\boldsymbol{d} \quad \text{and} \quad \boldsymbol{W\beta}\le\boldsymbol{e}. \tag{2}$$

Hua Zhou is Assistant Professor (E-mail: hua_zhou@ncsu.edu), and Yichao Wu is Associate Professor (E-mail: ywu11@ncsu.edu), Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203. The work was partially supported by NIH grants HG-006139 (Zhou) and CA-149569 (Wu) and NSF grants DMS-1310319 (Zhou), DMS-0905561 (Wu), and DMS-1055210 (Wu). The authors thank the editor, the associate editor, and three referees for their insightful and constructive comments.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/jasa.

Journal of the American Statistical Association
June 2014, Vol. 109, No. 506, Theory and Methods
DOI: 10.1080/01621459.2013.864166

Consequently, EPSODE solves the linearly constrained estimation problem (2) as a byproduct. In this case, path following commences from the unconstrained solution $\operatorname{argmin} f(\boldsymbol{\beta})$ and ends at the constrained solution to (2).

1.2 A Motivating Example

For illustration, we consider a merger and acquisition (M&A) dataset studied in Fan et al. (2013). This dataset constitutes $n = 1371$ U.S. companies with a binary response variable indicating whether the company becomes a leveraged buyout (LBO) target ($y_i = 1$) or not ($y_i = 0$). Seven covariates (1. cash flow; 2. cash; 3. long-term investment; 4. market to book ratio; 5. log market equity; 6. tax; 7. return on S&P 500 index) are recorded for each company. There have been intensive studies on the effects of these factors on the probability of a company being a target for strategic mergers. Exploratory analysis using linear logistic regression shows no significance in most covariates.

To explore the possibly nonlinear effects of these quantitative covariates, the varying-coefficient model (Hastie and Tibshirani 1993) can be adopted here. We discretize each predictor into, say, 10 bins and fit a logistic regression. The first bin of each predictor is used as the reference level and effect coding is applied to each discretized covariate. The circles (o) in Figure 1 denote the estimated coefficients for each bin of each predictor and hint at some interesting nonlinear effects. For instance, the chance of being an LBO target seems to monotonically decease with market-to-book ratio and be quadratic as a function of log-market equity. Regularization can be used to borrow strength between neighboring bins and gear solution toward clearer patterns. To illustrate the flexibility of the regularization scheme (1), we apply cubic trend filtering to five covariates (cash flow, cash, long term investment, tax, return on S&P 500 index), impose the monotonicity (nonincreasing) constraint on the "market-to-book ratio" covariate, and enforce the concavity constraint on the "log market equity" covariate. This can be achieved by minimizing a regularized negative logistic log-likelihood of form

$$-\ell(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_7) + \rho \sum_{j \neq 4,5} \|\boldsymbol{V}_j \boldsymbol{\beta}_j\|_1 + \rho \sum_{j=4,5} \|\boldsymbol{W}_j \boldsymbol{\beta}_j\|_+,$$

where $\boldsymbol{\beta}_j$ is the vector of regression coefficients for the jth discretized covariate. The matrices in the regularization terms are specified as

$$\boldsymbol{V}_j = \begin{pmatrix} -1 & 2 & -1 & & & & & \\ & 1 & -4 & 6 & -4 & 1 & & \\ & & \ddots & \ddots & \ddots & \ddots & & \\ & & 1 & -4 & 6 & -4 & 1 & \\ & & & & & -1 & 2 & -1 \end{pmatrix}$$

for $j = 1, 2, 3, 6, 7$,

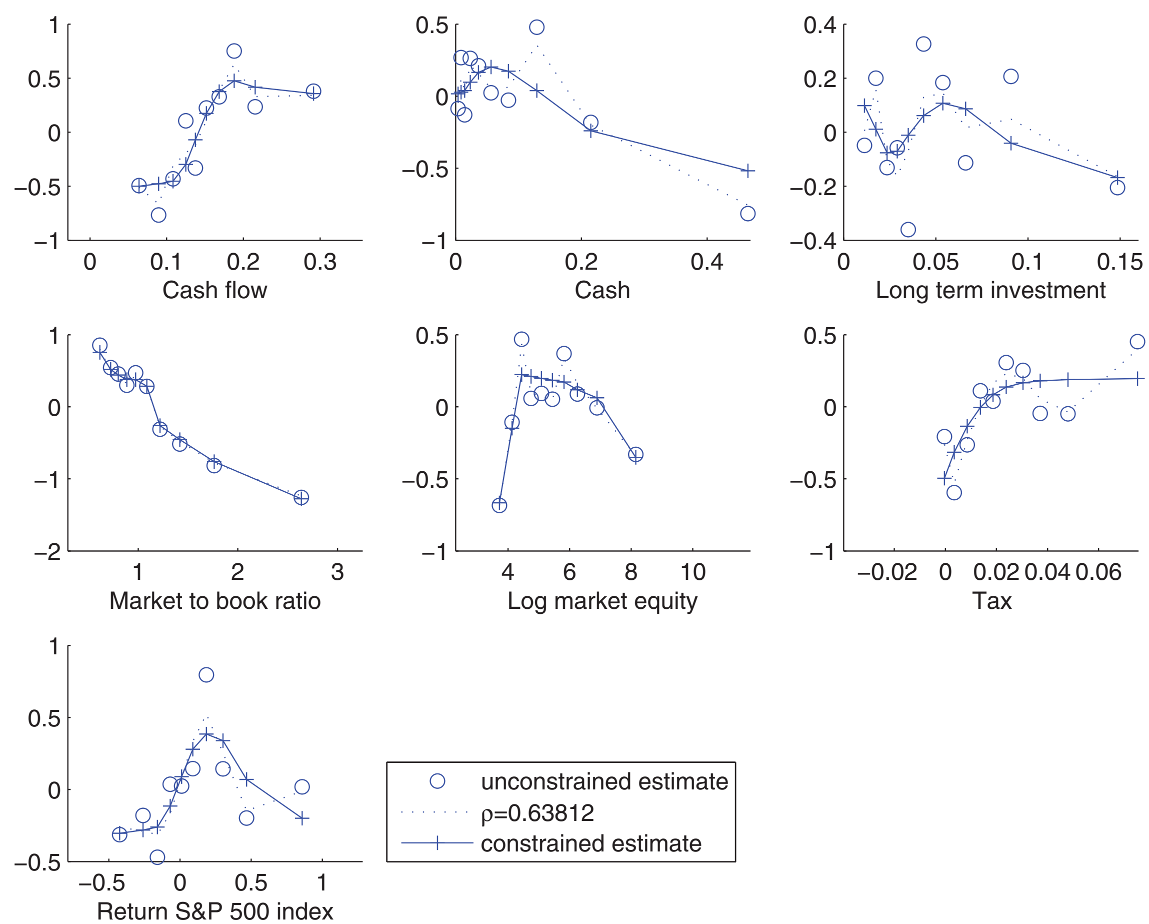

Figure 1. Snapshots of the path solution to the regularized logistic regression on the M&A dataset.

$$\boldsymbol{W}_4 = \begin{pmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{pmatrix},$$

$$\text{and } \boldsymbol{W}_5 = \begin{pmatrix} 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \end{pmatrix}.$$

The equality constraint regularization matrix $\boldsymbol{V}_j$, $j = 1, 2, 3, 6, 7$, penalizes the fourth order finite differences between the bin estimates. Thus, as ρ increases, the coefficient vectors of covariates 1-3,6-7 tend to be piecewise cubic with two ends being linear, mimicking the natural cubic spline. This is one example of the polynomial trend filtering (Kim et al. 2009; Tibshirani and Taylor 2011). Similar to semiparametric regressions, regularizations in polynomial trend filtering "let the data speak for themselves." In contrast, the bandwidth selection in semiparametric regressions is replaced by parameter tuning in regularizations. The number and locations of knots are automatically determined by tuning parameter which is chosen according to model selection criteria. In a similar fashion, the coefficient vector gradually becomes monotone for covariate "market-to-book ratio" and concave for covariate "log market equity." In addition, with ρ large enough, we recover the corresponding constrained solution, which are shown by the crosses (+) on solid lines in Figure 1. As noted above, our exact path algorithm delivers the whole solution path bridging from the unconstrained estimates (denoted by o) to the constrained estimates (denoted by +). For example, the dotted line in Figure 1 is a snapshot of the solution at $\rho = 0.6539$. Availability of the whole solution path renders model selection along the path easy. For instance, the regularization parameter ρ can be chosen by minimizing the cross-validation error or other model selection criteria such as AIC, BIC, or C_p. Figure 2 displays the solution path and the AIC and BIC along the path. It shows that both criteria favor the fully regularized solution, namely the constrained estimates. The whole solution path is obtained within seconds on a laptop using a Matlab implementation of EPSODE.

The patterns revealed by the regularized estimates match some existing finance theories. For instance, a company with low cash flow is unlikely to be an LBO target because low cash flow is hard to meet the heavy debt burden associated with the LBO. On the other hand, company carrying a high cash flow is likely to possess a new technology. It is risky to acquire such firms because it is hard to predict their profitability. The tax reason is obvious from the regularized estimates. The more tax the company is paying, the more tax benefits from an LBO. Log of market equity is a measure of company size. Smaller companies are unpredictable in their profitability and extremely large companies are unlikely to be an LBO target because LBOs are typically financed with a large proportion of external debts.

This illustrative example demonstrates the flexibility of our novel path algorithm. First, it can be applied to any convex loss function. In this example, the loss function is the negative log-likelihood of a logistic model. Second, it works for complicated regularizations like polynomial trend filtering (equality constraints), monotonicity constraint, and concavity constraint. More applications will be presented in Section 7 to illustrate the potential of EPSODE.

1.3 Previous Work and Our Contributions

Path algorithms have been devised for some special cases of the general regularization problem (1). Most notably the homotopy method (Osborne, Presnell, and Turlach 2000) and the least angle regression (LARS) procedure (Efron et al. 2004) handle lasso penalized least squares problem. The solution path generated is piecewise linear and illustrates the tradeoffs between goodness-of-fit and sparsity. Rosset and Zhu (2007) gave sufficient conditions for a solution path to be piecewise linear and expand its applications to a wider range of loss and penalty functions. Recently Tibshirani and Taylor (2011) devised a dual path algorithm for generalized ℓ_1 penalized least squares problems, which is problem (1) with f quadratic but without the second inequality regularization term. Zhou and Lange (2013) considered (1) in full generality for quadratic f. All these works concern regularized linear regression for which the solution path is piecewise linear. Several attempts have been made to derive a path following for regularized GLMs for which the solution

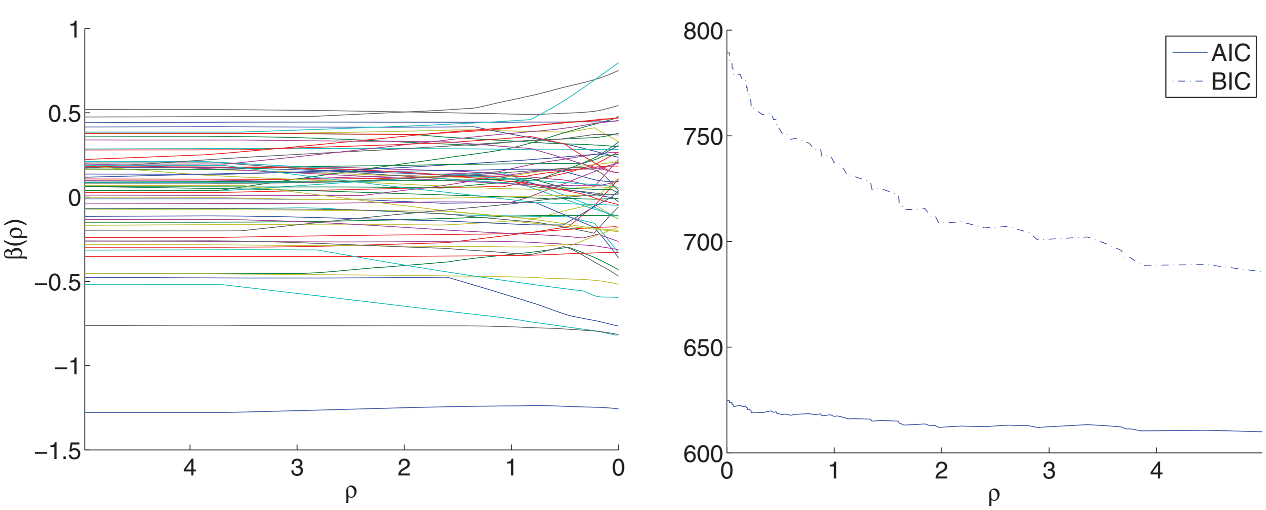

Figure 2. Solution and AIC/BIC paths of the regularized logistic regression on the M&A dataset.

path is no longer piecewise linear. Park and Hastie (2007) proposed a predictor-corrector approach to approximate the lasso path for GLMs. Friedman (2008) derived an approximate path algorithm for any convex loss regularized by a separable, but not necessarily for convex penalty. Here a penalty function is called separable if its Hessian matrix is diagonal.

In two pioneering papers, Wu (2011, 2012) presented an ODE-based LARS path algorithm for GLMs, quasilikelihoods, and Cox model, a modification of which is able to deliver the *exact* path for a lasso solution path. The ODE approach naturally fits problems with piecewise smooth solution paths and is the strategy we adopt in this article. Unfortunately, the separability restriction on the penalty term in (Friedman 2008; Wu 2011, 2012) excludes many important problems encountered in real applications.

Our proposed approach generalizes previous work in several aspects. First, it works for any convex loss (or criterion) function. Second, it allows for any type of regularization in terms of linear functions of parameters, equality or inequality. Equality constrained regularizations include lasso, fused-lasso, and generalized ℓ_1 penalty for example. Inequality constrained regularizations are required in shape-restricted regression and nonparametric log-concave density estimation. Last but not least, it is an *exact* path algorithm. Availability of the exact solution path has certain features that are appealing to statisticians. Compared to individual optimizations over a prespecified grid of tuning parameter values, it gives a more complete picture, capturing all model changes along the path. Moreover, it greatly eases some of the adaptive estimation procedure based on the regularization path. For instance, Bayesian model averaging over the ℓ_1 regularization path has been shown to generate superior prediction, classification, and model selection performances (Ghosh and Yuan 2009; Fraley and Percival 2010). In each MCMC iteration, a new model on regularization path is proposed and the associated (approximate) model likelihood needs to be evaluated. The exact solution path is certainly welcome here as regularized model at any ρ is likely to be sampled.

The rest of the article is organized as follows. Section 2 reviews the exact penalty method for optimization. Here the connections between constrained optimization and regularization in statistics are made clear. Section 3 derives in detail the EPSODE algorithm for strictly convex loss function f. Its implementation via the sweep operator and ordinary differential equations are described in Section 4. An extension of EPSODE for f convex, but not necessarily strictly convex, is discussed in Section 5. Nonasymptotic error bounds for equality regularized estimates are derived in Section 6. Section 7 presents various applications of EPSODE. Finally, Section 8 discusses the limitations of the path algorithm and hints at future generalizations.

2. EXACT PENALTY METHOD FOR CONVEX CONSTRAINED OPTIMIZATION

Consider the convex program

$$\min \quad f(\boldsymbol{x}) \quad \text{subject to} \quad \begin{aligned} &g_i(\boldsymbol{x}) = 0,\, 1 \le i \le r, \text{ and} \\ &h_j(\boldsymbol{x}) \le 0,\, 1 \le j \le s, \end{aligned} \tag{3}$$

where the objective function f is convex, equality constraint functions g_i are affine, and the inequality constraint functions h_j are convex. We further assume that f and h_j are smooth. Specifically we require that f and h_j are continuously twice differentiable. To fix the notation, differential $df(\boldsymbol{x})$ is the row vector of partial derivatives of f at $\boldsymbol{x}$ and the gradient $\nabla f(\boldsymbol{x})$ is the transpose of $df(\boldsymbol{x})$. The Hessian matrix of $f(\cdot)$ is denoted by $d^2 f(\boldsymbol{x})$.

Exact penalty method minimizes the function

$$\mathcal{E}_\rho(\boldsymbol{x}) = f(\boldsymbol{x}) + \rho \sum_{i=1}^{r} |g_i(\boldsymbol{x})| + \rho \sum_{j=1}^{s} \max\{0, h_j(\boldsymbol{x})\} \tag{4}$$

for $\rho \ge 0$. Classical results (Ruszczyński 2006, Theorems 6.9 and 7.21) state that for ρ large enough, the solution to the optimization problem (4) coincides with the solution to the original constrained convex program (3). This justifies the exact penalty method as one way to solve constrained optimization problems.

According to convex calculus (Ruszczyński 2006, Theorem 3.5), the optimal point $\boldsymbol{x}(\rho)$ of the function $\mathcal{E}_\rho(\boldsymbol{x})$ is characterized by the necessary and sufficient condition

$$\mathbf{0} = \nabla f(\boldsymbol{x}) + \rho \sum_{i=1}^{r} s_i \nabla g_i(\boldsymbol{x}) + \rho \sum_{j=1}^{s} t_j \nabla h_j(\boldsymbol{x}) \tag{5}$$

with coefficients satisfying

$$s_i \in \begin{cases} \{-1\} & g_i(\boldsymbol{x}) < 0 \\ [-1, 1] & g_i(\boldsymbol{x}) = 0 \\ \{1\} & g_i(\boldsymbol{x}) > 0 \end{cases}, \quad \text{and} \quad t_j \in \begin{cases} \{0\} & h_j(\boldsymbol{x}) < 0 \\ [0, 1] & h_j(\boldsymbol{x}) = 0 \\ \{1\} & h_j(\boldsymbol{x}) > 0 \end{cases}. \tag{6}$$

The sets defining possible values of s_i and t_j are the subdifferentials of the functions $|x|$ and $x_+ = \max\{x, 0\}$. For the path following to make sense, we require uniqueness and continuity of the solution $\boldsymbol{x}(\rho)$ to (4) as ρ varies. The following lemma concerns the continuity of the solution path and is the foundation of our path algorithm.

Lemma 2.1.

1. (Uniqueness) If $\mathcal{E}_\rho$ is strictly convex, then its minimizer $\boldsymbol{x}(\rho)$ is unique.
2. (Continuity) If $\mathcal{E}_\rho$ is strictly convex and coercive over an open neighborhood of ρ, that is, $\{\boldsymbol{x} : \mathcal{E}_\rho(\boldsymbol{x}) \le \mathcal{E}_\rho(\boldsymbol{z})\}$ is compact for all $\boldsymbol{z}$, then the minimizer $\boldsymbol{x}(\rho)$ is continuous at ρ.
3. (Continuity of s_i and t_j) Furthermore, if the gradients $\{\nabla g_i(\boldsymbol{x}) : g_i(\boldsymbol{x}) = 0\} \cup \{\nabla h_j(\boldsymbol{x}) : h_j(\boldsymbol{x}) = 0\}$ of active constraints are linearly independent at the solution $\boldsymbol{x}(\rho)$ over an open neighborhood of ρ, then the coefficient paths $s_i(\rho)$ and $t_j(\rho)$ are unique and continuous at ρ.

We remark that strict convexity only gives an easy-to-check sufficient condition for uniqueness and continuity; it is not necessary. A convex but not strictly convex function can still have a unique minimum. The absolute value function $|x|$ offers such an example. When the loss function f is strictly convex, then $\mathcal{E}_\rho$ is strictly convex for all $\rho \ge 0$ and by Lemma 2.1 there exists a unique, continuous solution path $\{\boldsymbol{x}(\rho) : \rho \ge 0\}$. In Sections 3 and 4, we derive the path algorithm assuming that f is strictly convex. When f is convex but not strictly convex, for example, when $n < p$ in the least squares problems, the solutions at smaller ρ may not be unique. In that case, it is still possible to obtain a solution path over the region of large ρ where the minimum of $\mathcal{E}_\rho$ is unique. In Section 5, we extend EPSODE to the case f is convex but may not be strictly convex. The third statement of Lemma 2.1 implies that the active constraints ($g_i(\boldsymbol{x}) = 0$ or $h_j(\boldsymbol{x}) = 0$) with interior coefficients must stay active until the coefficients hit the end points of the permissible range, which in turn implies that the solution path is piecewise smooth. This allows us to develop a path following algorithm based on ODE.

3. THE PATH FOLLOWING ALGORITHM

In this article, we specialize to the case where the constraint functions g_i and h_j are affine, that is, the gradient vectors $\nabla g_i(\boldsymbol{x})$ and $\nabla h_j(\boldsymbol{x})$ are constant. This leads to the regularized optimization problem formulated as (1) by defining g_i and h_j as constraint residuals $g_i(\boldsymbol{x}) = \boldsymbol{v}_i^t\boldsymbol{x} - d_i$ and $h_j(\boldsymbol{x}) = \boldsymbol{w}_j^t\boldsymbol{x} - e_j$. In principle, a similar path algorithm can be developed for the general convex program where the inequality constraint functions h_j are relaxed to be convex. But that is beyond the scope of the current article. In Sections 3 and 4, we assume that the loss function f is strictly convex. This assumption is relaxed in Section 5.

EPSODE works in a segment-by-segment fashion. Along the path we keep track of the following index sets determined by signs of constraint residuals

$$\begin{aligned}
\mathcal{N}_{\mathrm{E}} &= \left\{i : g_i(\boldsymbol{x}) = \boldsymbol{v}_i^t\boldsymbol{x} - d_i < 0\right\},\\
\mathcal{N}_{\mathrm{I}} &= \left\{j : h_j(\boldsymbol{x}) = \boldsymbol{w}_j^t\boldsymbol{x} - e_j < 0\right\}\\
\mathcal{Z}_{\mathrm{E}} &= \left\{i : g_i(\boldsymbol{x}) = \boldsymbol{v}_i^t\boldsymbol{x} - d_i = 0\right\},\\
\mathcal{Z}_{\mathrm{I}} &= \left\{j : h_j(\boldsymbol{x}) = \boldsymbol{w}_j^t\boldsymbol{x} - e_j = 0\right\}\\
\mathcal{P}_{\mathrm{E}} &= \left\{i : g_i(\boldsymbol{x}) = \boldsymbol{v}_i^t\boldsymbol{x} - d_i > 0\right\},\\
\mathcal{P}_{\mathrm{I}} &= \left\{j : h_j(\boldsymbol{x}) = \boldsymbol{w}_j^t\boldsymbol{x} - e_j > 0\right\}.
\end{aligned} \tag{7}$$

Along each segment of the path, the set configuration is fixed. This is implied by the continuity of both the solution and coefficient paths established in Lemma 2.1. Throughout this article, we call the constraints in $\mathcal{Z}_{\mathrm{E}}$ or $\mathcal{Z}_{\mathrm{I}}$ active and others inactive.

Next we derive the ODE for the solution $\boldsymbol{x}(\rho)$ on a fixed segment. Suppose, we are in the interior of a segment. Let $\boldsymbol{x}(\rho)$ be the solution of (4) indexed by the penalty parameter ρ and $\boldsymbol{x}(\rho + \Delta\rho)$ the solution when the penalty is increased by an infinitesimal amount $\Delta\rho > 0$. Then the difference $\Delta\boldsymbol{x}(\rho) = \boldsymbol{x}(\rho + \Delta\rho) - \boldsymbol{x}(\rho)$ should minimize the increase in optimal objective value. That is, to the second order, $\Delta\boldsymbol{x}$ is the solution to

$$\begin{aligned}
\min_{\Delta\boldsymbol{x}} \quad & \mathcal{E}_{\rho+\Delta\rho}(\boldsymbol{x} + \Delta\boldsymbol{x}) - \mathcal{E}_\rho(\boldsymbol{x})\\
&\approx df(\boldsymbol{x}) \cdot \Delta\boldsymbol{x} + \frac{1}{2}\Delta\boldsymbol{x}^t \cdot d^2 f(\boldsymbol{x}) \cdot \Delta\boldsymbol{x}\\
&\quad + (\rho + \Delta\rho) \cdot \left[-\sum_{i\in\mathcal{N}_{\mathrm{E}}} \boldsymbol{v}_i + \sum_{i\in\mathcal{P}_{\mathrm{E}}} \boldsymbol{v}_i + \sum_{j\in\mathcal{P}_{\mathrm{I}}} \boldsymbol{w}_j \right] \cdot \Delta\boldsymbol{x}\\
&\quad + \Delta\rho \cdot \left[-\sum_{i\in\mathcal{N}_{\mathrm{E}}} g_i(\boldsymbol{x}) + \sum_{i\in\mathcal{P}_{\mathrm{E}}} g_i(\boldsymbol{x}) + \sum_{j\in\mathcal{P}_{\mathrm{I}}} h_j(\boldsymbol{x}) \right]\\
\text{s.t.} \quad & \boldsymbol{v}_i^t \cdot \Delta\boldsymbol{x} = 0, i \in \mathcal{Z}_{\mathrm{E}}, \text{ and } \boldsymbol{w}_j^t \cdot \Delta\boldsymbol{x} = 0, j \in \mathcal{Z}_{\mathrm{I}}.
\end{aligned} \tag{8}$$

Note that the active constraints have to be kept active since the set configuration is fixed along this segment by Lemma 2.1. This is why we have these two sets of equality constraints. To ease notational burden, we define

$$\boldsymbol{H}(\boldsymbol{x}) = d^2 f(\boldsymbol{x}) \quad \text{and} \quad \boldsymbol{u}_{\bar{\mathcal{Z}}} = -\sum_{i\in\mathcal{N}_{\mathrm{E}}} \boldsymbol{v}_i + \sum_{i\in\mathcal{P}_{\mathrm{E}}} \boldsymbol{v}_i + \sum_{j\in\mathcal{P}_{\mathrm{I}}} \boldsymbol{w}_j. \tag{9}$$

This leads to the corresponding Lagrange multiplier problem

$$\begin{pmatrix} \boldsymbol{H}(\boldsymbol{x}) & \boldsymbol{U}_{\mathcal{Z}}^t \\ \boldsymbol{U}_{\mathcal{Z}} & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \Delta\boldsymbol{x} \\ \boldsymbol{\lambda}_{\mathcal{Z}} \end{pmatrix} = \begin{pmatrix} -\nabla f(\boldsymbol{x}) - (\rho + \Delta\rho)\boldsymbol{u}_{\bar{\mathcal{Z}}} \\ \boldsymbol{0} \end{pmatrix},$$

where the rows of the matrix $\boldsymbol{U}_{\mathcal{Z}}$ are the constant differentials, $\boldsymbol{v}_i^t, i \in \mathcal{Z}_{\mathrm{E}}$, and $\boldsymbol{w}_j^t, j \in \mathcal{Z}_{\mathrm{I}}(\boldsymbol{x})$, of the active constraint functions. Denoting the inverse of matrix as

$$\begin{pmatrix} \boldsymbol{H}(\boldsymbol{x}) & \boldsymbol{U}_{\mathcal{Z}}^t \\ \boldsymbol{U}_{\mathcal{Z}} & \boldsymbol{0} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{P}(\boldsymbol{x}) & \boldsymbol{Q}(\boldsymbol{x}) \\ \boldsymbol{Q}^t(\boldsymbol{x}) & \boldsymbol{R}(\boldsymbol{x}) \end{pmatrix},$$

where

$$\begin{aligned}
\boldsymbol{P}(\boldsymbol{x}) &= \boldsymbol{H}^{-1}(\boldsymbol{x}) - \boldsymbol{H}^{-1}(\boldsymbol{x})\boldsymbol{U}_{\mathcal{Z}}^t \left[\boldsymbol{U}_{\mathcal{Z}}\boldsymbol{H}^{-1}(\boldsymbol{x})\boldsymbol{U}_{\mathcal{Z}}^t\right]^{-1} \boldsymbol{U}_{\mathcal{Z}}\boldsymbol{H}^{-1}(\boldsymbol{x})\\
\boldsymbol{Q}(\boldsymbol{x}) &= \boldsymbol{H}^{-1}(\boldsymbol{x})\boldsymbol{U}_{\mathcal{Z}}^t \left[\boldsymbol{U}_{\mathcal{Z}}\boldsymbol{H}^{-1}(\boldsymbol{x})\boldsymbol{U}_{\mathcal{Z}}^t\right]^{-1}\\
\boldsymbol{R}(\boldsymbol{x}) &= -\left[\boldsymbol{U}_{\mathcal{Z}}\boldsymbol{H}^{-1}(\boldsymbol{x})\boldsymbol{U}_{\mathcal{Z}}^t(\boldsymbol{x})\right]^{-1},
\end{aligned} \tag{10}$$

the solution of the difference vector $\Delta\boldsymbol{x}$ is

$$\begin{aligned}
\Delta\boldsymbol{x} &= -\boldsymbol{P}(\boldsymbol{x})[\nabla f(\boldsymbol{x}) + (\rho + \Delta\rho)\boldsymbol{u}_{\bar{\mathcal{Z}}}]\\
&= -\boldsymbol{P}(\boldsymbol{x})[\nabla f(\boldsymbol{x}) + \rho\boldsymbol{u}_{\bar{\mathcal{Z}}}(\boldsymbol{x}) + \Delta\rho \cdot \boldsymbol{u}_{\bar{\mathcal{Z}}}]\\
&= -\boldsymbol{P}(\boldsymbol{x})\left[-\rho\boldsymbol{U}_{\mathcal{Z}}^t\boldsymbol{r}_{\mathcal{Z}} + \Delta\rho \cdot \boldsymbol{u}_{\bar{\mathcal{Z}}}\right].
\end{aligned}$$

Note $\boldsymbol{P}(\boldsymbol{x})\boldsymbol{U}_{\mathcal{Z}}^t = \boldsymbol{0}$. Therefore, $\Delta\boldsymbol{x} = -\Delta\rho \cdot \boldsymbol{P}(\boldsymbol{x})\boldsymbol{u}_{\bar{\mathcal{Z}}}$. This gives the direction for the infinitesimal update of solution vector $\boldsymbol{x}(\rho)$. Taking limit in $\Delta\rho$ leads to the following key result for developing the path algorithm.

Proposition 3.1. Within interior of a path segment with set configuration (7), the solution $\boldsymbol{x}(\rho)$ satisfies an ordinary differential equation (ODE)

$$\frac{d\boldsymbol{x}(\rho)}{d\rho} = -\boldsymbol{P}(\boldsymbol{x})\boldsymbol{u}_{\bar{\mathcal{Z}}}, \tag{11}$$

where the matrix $\boldsymbol{P}(\boldsymbol{x})$ and vector $\boldsymbol{u}_{\bar{\mathcal{Z}}}$ are defined by (10) and (9).

Note that the right-hand side of (11) is a constant vector in $\boldsymbol{x}$ when f is quadratic and g_i and h_j are affine. Thus the corresponding solution path is piecewise linear. This recovers the case studied in Zhou and Lange (2013). The differential equation (11) holds on the current segment until one of two types of events happens: an inactive constraint becomes active or vice versa. The first type of event is easy to detect—whenever a constraint function, $g_i(\boldsymbol{x}), i \in \mathcal{N}_{\mathrm{E}} \cup \mathcal{P}_{\mathrm{E}}$, or $h_j(\boldsymbol{x}), j \in \mathcal{N}_{\mathrm{I}} \cup \mathcal{P}_{\mathrm{I}}$, hits zero, we move that constraint to the active set $\mathcal{Z}_{\mathrm{E}}$ or $\mathcal{Z}_{\mathrm{I}}$ and start solving a new system of differential equations. To detect when the second type of event happens, we need to keep track of the coefficients $s_i(\boldsymbol{x})$ and $t_j(\boldsymbol{x})$ for active constraints. Whenever the coefficient of an active constraint hits the boundary of its permissible range in (6), the constraint has to be relaxed from being active in next segment. It turns out the coefficients for active constraints admit a simple representation in terms of current solution vector.

Proposition 3.2. On a path segment with set configuration (7), the coefficients s_i and t_j for active constraints are

$$\boldsymbol{r}_{\mathcal{Z}}(\rho) = \begin{pmatrix} \boldsymbol{s}_{\mathcal{Z}_{\mathrm{E}}}(\rho) \\ \boldsymbol{t}_{\mathcal{Z}_{\mathrm{I}}}(\rho) \end{pmatrix} = -\boldsymbol{Q}^t(\boldsymbol{x})\left[\frac{1}{\rho}\nabla f(\boldsymbol{x}) + \boldsymbol{u}_{\bar{\mathcal{Z}}}\right], \quad (12)$$

where $\boldsymbol{x} = \boldsymbol{x}(\rho)$ is the solution at ρ and the matrix $\boldsymbol{Q}(\boldsymbol{x})$ is defined by (10).

Given current solution vector $\boldsymbol{x}(\rho)$, the coefficients of the active constraints are readily obtained from (12). Once a coefficient hits the end points, we move that constraint from the active set to the inactive set that matches the endpoint being hit. In next section, we detail the implementation of the path algorithm.

4. IMPLEMENTATION: ODE AND SWEEPING OPERATOR

Algorithm 1 EPSODE: Solution path for regularization problem (1) with strictly convex f.

Initialize $\rho = 0$, $\boldsymbol{\beta}(0) = \operatorname{argmin} f(\boldsymbol{\beta})$, and its set configuration (7).
repeat
  Solve ODE (11) until an inactive constraint becomes active or the coefficient (12) of an active constraint hits boundary.
  Update the set configuration (7).
until $\mathcal{N}_{\mathrm{E}} = \mathcal{P}_{\mathrm{E}} = \mathcal{P}_{\mathrm{I}} = \emptyset$

Algorithm 1 summarizes EPSODE based on Propositions 3.1 and 3.2. It involves solving ODEs segment by segment and is extremely simple to implement using softwares with a reliable ODE solver such as the `ode45` function in Matlab and the `deSolve` package (Soetaert, Petzoldt, and Setzer 2010) in R. There has been extensive research in applied mathematics on numerical methods for solving ODEs, notably the Runge-Kutta, Richardson extrapolation and predictor-corrector methods. Some path following algorithms developed for specific statistical problems (Park and Hastie 2007; Friedman 2008) turn out to be approximate methods for solving the corresponding ODE. Wu (2011) first explicitly uses ODE to derive an exact solution path for the lasso penalized GLM. The connection of the path following to ODE relieves statisticians from the burden of developing specific path algorithms for a variety of regularization problems. For instance, the rich numerical resources of Matlab include ODE solvers that control the tolerance for the accuracy of solution (10^{-6} by default) and alert the user when certain events such as constraint hitting and escape occur.

Any ODE solver repeatedly evaluates the derivative. Suppose the number of parameters is p. Computation of the matrix-vector multiplications in (11) and (12) has computation cost of order $O(p^2) + O(p|\mathcal{Z}|) + O(|\mathcal{Z}|^3)$ if the inverse H^{-1} of Hessian matrix of loss function f is readily available, where $\mathcal{Z} = \mathcal{Z}_{\mathrm{E}} \cup \mathcal{Z}_{\mathrm{I}}$ and $|\mathcal{Z}|$ denote its cardinality. Otherwise the computation cost is $O(p^3) + O(p|\mathcal{Z}|) + O(|\mathcal{Z}|^3)$.

An alternative implementation avoids repeated matrix inversions by solving an ODE for the matrices $\boldsymbol{P}$, $\boldsymbol{Q}$, and $\boldsymbol{R}$ themselves. The computations can be conveniently organized around the classical sweep and inverse sweep operators of regression analysis (Dempster 1969; Jennrich 1977; Goodnight 1979; Little and Rubin 2002; Lange 2010). Suppose $\boldsymbol{A}$ is an $m \times m$ symmetric matrix. Sweeping on the kth diagonal entry $a_{kk} \neq 0$ of $\boldsymbol{A}$ yields a new symmetric matrix $\widehat{\boldsymbol{A}}$ with entries

$$\hat{a}_{kk} = -\frac{1}{a_{kk}}, \quad \hat{a}_{ik} = \frac{a_{ik}}{a_{kk}}, \quad i \neq k,$$
$$\hat{a}_{kj} = \frac{a_{kj}}{a_{kk}}, \quad j \neq k, \quad \hat{a}_{ij} = a_{ij} - \frac{a_{ik}a_{kj}}{a_{kk}}, \quad i, j \neq k.$$

These arithmetic operations can be undone by inverse sweeping on the same diagonal entry. Inverse sweeping on the kth diagonal entry sends the symmetric matrix $\boldsymbol{A}$ into the symmetric matrix $\check{\boldsymbol{A}}$ with entries

$$\check{a}_{kk} = -\frac{1}{a_{kk}}, \quad \check{a}_{ik} = -\frac{a_{ik}}{a_{kk}}, \quad i \neq k,$$
$$\check{a}_{kj} = -\frac{a_{kj}}{a_{kk}}, \quad j \neq k, \quad \check{a}_{ij} = a_{ij} - \frac{a_{ik}a_{kj}}{a_{kk}}, \quad i, j \neq k.$$

Both sweeping and inverse sweeping preserve symmetry. Thus, all operations can be carried out on either the lower or upper triangle of $\boldsymbol{A}$ alone, saving both computational time and storage. When several sweeps or inverse sweeps are performed, their order is irrelevant.

At beginning ($\rho = 0$) of the path following, we initialize a sweeping tableau as

$$\left(\begin{array}{c|c} \boldsymbol{H}^{-1}(\boldsymbol{x}) & \boldsymbol{H}^{-1}(\boldsymbol{x})\boldsymbol{U}^t \\ \hline * & \boldsymbol{U}\boldsymbol{H}^{-1}(\boldsymbol{x})\boldsymbol{U}^t \end{array}\right),$$

where the matrix $\boldsymbol{U} \in \mathbb{R}^{(r+s)\times p}$ holds all constraint differentials $\boldsymbol{v}_i^t$ and $\boldsymbol{w}_j^t$ in rows. Further sweeping of diagonal entries corresponding to the active constraints yields

$$\left(\begin{array}{c|cc} \boldsymbol{P}(\boldsymbol{x}) & \boldsymbol{Q}(\boldsymbol{x}) & \boldsymbol{P}(\boldsymbol{x})\boldsymbol{U}_{\bar{\mathcal{Z}}}^t \\ \hline * & \boldsymbol{R}(\boldsymbol{x}) & \boldsymbol{Q}^t(\boldsymbol{x})\boldsymbol{U}_{\bar{\mathcal{Z}}}^t \\ * & * & \boldsymbol{U}_{\bar{\mathcal{Z}}}\boldsymbol{P}(\boldsymbol{x})\boldsymbol{U}_{\bar{\mathcal{Z}}}^t \end{array}\right). \quad (13)$$

Here we conveniently organized the columns of the swept active constraints before those of un-swept ones. In practice the sweep tableau is not necessary as in (13) and it is enough to keep an indicator vector recording the columns being swept. The key elements for the path algorithm can be easily retrieved from the sweep tableau (13) as

$$\frac{d\boldsymbol{x}(\rho)}{d\rho} = -\boldsymbol{P}(\boldsymbol{x})\boldsymbol{U}_{\bar{\mathcal{Z}}}^t\boldsymbol{r}_{\bar{\mathcal{Z}}}$$
$$\boldsymbol{r}_{\mathcal{Z}}(\rho) = -\boldsymbol{Q}^t(\boldsymbol{x})\boldsymbol{U}_{\bar{\mathcal{Z}}}^t\boldsymbol{r}_{\bar{\mathcal{Z}}} - \frac{1}{\rho}\boldsymbol{Q}^t(\boldsymbol{x})\nabla f(\boldsymbol{x}),$$

where $\boldsymbol{r}_{\bar{\mathcal{Z}}}$ denotes the coefficient vector for the inactive constraints, with entries -1 for constraints in $\mathcal{N}_{\mathrm{E}}$, 0 for constraints in $\mathcal{N}_{\mathrm{I}}$, and 1 for constraints in $\mathcal{P}_{\mathrm{E}} \cup \mathcal{P}_{\mathrm{I}}$. Therefore, the path following procedure only involves solving ODE for the whole sweep tableau (13) with sweeping or inverse sweeping at kinks between successive segments. For this purpose, we derive the ODE for the sweep tableau (13). For a matrix function $F(\boldsymbol{X}) : \mathbb{R}^{n\times q} \to \mathbb{R}^{m\times p}$,

$$DF(\boldsymbol{X}) = \frac{\partial \operatorname{vec} F(\boldsymbol{X})}{\partial (\operatorname{vec} \boldsymbol{X})^t}$$

denotes the $mp \times nq$ Jacobian matrix (Magnus and Neudecker 1999). As a special case, Proposition 3.1 states $D\boldsymbol{x}(\rho) = -\boldsymbol{P}(\boldsymbol{x})\boldsymbol{u}_{\bar{\mathcal{Z}}}$ where $m = p = 1$.

Proposition 4.1 (ODE for Sweep Tableau). On a segment of path with fixed set configuration, the matrices $\boldsymbol{P}(\rho)$, $\boldsymbol{Q}(\rho)$, and $\boldsymbol{R}(\rho)$ satisfy the ordinary differential equations (ODE)

$$\begin{aligned} D\boldsymbol{P}(\rho) &= [\boldsymbol{P}(\boldsymbol{x}) \otimes \boldsymbol{P}(\boldsymbol{x})] \cdot [D\boldsymbol{H}(\boldsymbol{x})] \cdot \boldsymbol{P}(\boldsymbol{x})\boldsymbol{u}_{\bar{\mathcal{Z}}} \\ D\boldsymbol{Q}(\rho) &= [\boldsymbol{Q}^t(\boldsymbol{x}) \otimes \boldsymbol{P}(\boldsymbol{x})] \cdot [D\boldsymbol{H}(\boldsymbol{x})] \cdot \boldsymbol{P}(\boldsymbol{x})\boldsymbol{u}_{\bar{\mathcal{Z}}} \\ D\boldsymbol{R}(\rho) &= [\boldsymbol{Q}^t(\boldsymbol{x}) \otimes \boldsymbol{Q}^t(\boldsymbol{x})] \cdot [D\boldsymbol{H}(\boldsymbol{x})] \cdot \boldsymbol{P}(\boldsymbol{x})\boldsymbol{u}_{\bar{\mathcal{Z}}}. \end{aligned}$$

Solving ODE for these matrices requires the p^2-by-p Jacobian matrix of the Hessian matrix $\boldsymbol{H}(\boldsymbol{x}) = d^2 f(\boldsymbol{x})$,

$$D\boldsymbol{H}(\boldsymbol{x}) = \frac{\partial[\mathrm{vec}\boldsymbol{H}(\boldsymbol{x})]}{\partial \mathrm{vec}(\boldsymbol{x})^t} = \frac{\partial \mathrm{vec}[df^2(\boldsymbol{x})]}{\partial \mathrm{vec}(\boldsymbol{x})^t},$$

which we provide for each example in Section 7 for convenience. When the number of parameter p is large, $D\boldsymbol{H}$ is a large matrix. However, there is no need to compute and store $D\boldsymbol{H}$ and only the matrix vector multiplication $D\boldsymbol{H} \cdot \boldsymbol{v}$ for any vector $\boldsymbol{v}$ is needed. In light of the useful identity $(\boldsymbol{B}^t \otimes \boldsymbol{A})\mathrm{vec}(\boldsymbol{C}) = \mathrm{vec}(\boldsymbol{ACB})$, evaluating the derivative for the whole tableau only involves multiplying three matrices and incurs computational cost $O(p^3) + O(p^2|\mathcal{Z}|) + O(p|\mathcal{Z}|^2)$.

Although we have presented the path algorithm as moving from $\rho = 0$ to large ρ, it can be applied in either direction. Lasso and fused-lasso usually start from the constrained solution, while in presence of general equality constraints, for example, polynomial trend filtering, and/or inequality constraints, the constrained solution is not readily available and the path algorithm must be initiated at $\rho = 0$.

5. EXTENSION OF EPSODE

So far we have assumed strict convexity of the loss function f. This unfortunately excludes many interesting applications, especially $p > n$ case of the regression problems. In this section we briefly indicate an extension of EPSODE to the case f is convex but not necessarily strictly convex. In the proof of Proposition 3.1, the infinitesimal change of solution $\Delta\boldsymbol{x}$ is derived via minimizing the equality-constrained quadratic program (8), the solution to which requires inverse of Hessian $\boldsymbol{H}^{-1}$ and thus strict convexity of f. Alternatively we may solve (8) via reparameterization. Let $\boldsymbol{U}_{\mathcal{Z}}$ hold the active constraint vectors and $\boldsymbol{Y} \in \mathbb{R}^{p\times(p-|\mathcal{Z}|)}$ be a null space matrix of $\boldsymbol{U}_{\mathcal{Z}}$, that is, the columns of $\boldsymbol{Y}$ are orthogonal to the rows of $\boldsymbol{U}_{\mathcal{Z}}$. Then the infinitesimal change can be represented as $\Delta\boldsymbol{x} = \boldsymbol{Y}\Delta\boldsymbol{y}$ for some vector $\Delta\boldsymbol{y} \in \mathbb{R}^{p-|\mathcal{Z}|}$. Under this reparameterization, the quadratic program (8) is equivalent to

$$\min_{\Delta\boldsymbol{y}} \frac{1}{2}\Delta\boldsymbol{y}^t[\boldsymbol{Y}^t\boldsymbol{H}(\boldsymbol{x})\boldsymbol{Y}]\Delta\boldsymbol{y} + \left[df(\boldsymbol{x}) + (\rho + \Delta\rho)\boldsymbol{u}^t_{\bar{\mathcal{Z}}}\right]\boldsymbol{Y} \cdot \Delta\boldsymbol{y}$$

with explicit solution

$$\Delta\boldsymbol{y} = -[\boldsymbol{Y}^t\boldsymbol{H}(\boldsymbol{x})\boldsymbol{Y}]^{-1}\boldsymbol{Y}^t[\nabla f(\boldsymbol{x}) + (\rho + \Delta\rho)\boldsymbol{u}_{\bar{\mathcal{Z}}}].$$

Hence, the infinitesimal change in $\boldsymbol{x}(\rho)$ is

$$\begin{aligned} \Delta\boldsymbol{x} &= -\boldsymbol{Y}[\boldsymbol{Y}^t\boldsymbol{H}(\boldsymbol{x})\boldsymbol{Y}]^{-1}\boldsymbol{Y}^t[\nabla f(\boldsymbol{x}) + (\rho + \Delta\rho)\boldsymbol{u}_{\bar{\mathcal{Z}}}] \\ &= -\Delta\rho \cdot \boldsymbol{Y}[\boldsymbol{Y}^t\boldsymbol{H}(\boldsymbol{x})\boldsymbol{Y}]^{-1}\boldsymbol{Y}^t\boldsymbol{u}_{\bar{\mathcal{Z}}}]. \end{aligned}$$

Again taking limit gives the following result in parallel to Proposition 3.1.

Proposition 5.1. Within interior of a path segment with set configuration (7), the solution $\boldsymbol{x}(\rho)$ satisfies an ordinary differential equation (ODE)

$$\frac{d\boldsymbol{x}(\rho)}{d\rho} = -\boldsymbol{Y}[\boldsymbol{Y}^t\boldsymbol{H}(\boldsymbol{x})\boldsymbol{Y}]^{-1}\boldsymbol{Y}^t\boldsymbol{u}_{\bar{\mathcal{Z}}}, \tag{14}$$

where $\boldsymbol{Y}$ is a null space matrix of $\boldsymbol{U}_{\mathcal{Z}}$.

An advantage of (14) is that only nonsingularity of the matrix $\boldsymbol{Y}^t\boldsymbol{H}(\boldsymbol{x})\boldsymbol{Y}$ is required which is much weaker than the nonsingularity of $\boldsymbol{H}$. The computational cost of calculating the derivative in (14) is $O((p - |\mathcal{Z}|)^3) + O(p(p - |\mathcal{Z}|))$, which is more efficient than (11) when $p - |\mathcal{Z}|$ is small. However, it requires the null space matrix $\boldsymbol{Y}$, which is nonunique and may be expensive to compute. Fortunately, the null space matrix $\boldsymbol{Y}$ is constant over each path segment and in practice can be calculated by QR decomposition of the active constraint matrix $\boldsymbol{U}_{\mathcal{Z}}$. At each kink either one constraint leaves $\mathcal{Z}$ or one enters $\mathcal{Z}$. Therefore, $\boldsymbol{Y}$ can be sequentially updated (Lawson and Hanson 1987) and need not to be calculated anew for each segment. Which version of (11) and (14) to use depends on specific application. When the loss function f is not strictly convex, for example, $p > n$ case in regression analysis, only (14) applies.

6. STATISTICAL PROPERTIES

In this section, we derive error bounds for the regularized estimates produced by EPSODE using the regularized M-estimation framework (Negahban et al. 2012). We restrict to the equality constraint regularization

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} f(\boldsymbol{\beta}) + \rho\|\boldsymbol{V}\boldsymbol{\beta}\|_1.,$$

where $\boldsymbol{V} \in \mathbb{R}^{r\times p}$. The case with inequality regularization is outside the scope of this article and will be pursued elsewhere.

Suppose the data are generated from $\boldsymbol{\beta}^* \in \mathbb{R}^p$. Let $\mathcal{S} = \mathrm{supp}(\boldsymbol{V}\boldsymbol{\beta}^*) = \{j : \boldsymbol{v}_j^t\boldsymbol{\beta}^* \neq 0\}$ be the set of violated constraints, and $\boldsymbol{V}_{\mathcal{S}}$ and $\boldsymbol{V}_{\mathcal{S}^c}$ be the submatrices of $\boldsymbol{V}$ with corresponding rows in $\mathcal{S}$ and $\mathcal{S}^c$, respectively. We make two assumptions: (1) $\boldsymbol{V}\boldsymbol{\beta}^*$ is s-sparse with $|\mathcal{S}| = s$ and (2) $\boldsymbol{V}$ has full column rank, that is, $\mathrm{rank}(\boldsymbol{V}) = p$. Note Assumption (2) implies that $r \geq p$. Many popular regularizations with a tall regularization matrix $\boldsymbol{V}$ such as the sparse fused-lasso (Tibshirani et al. 2005) satisfies this assumption. Define spaces

$$\begin{aligned} \mathcal{M} &= \mathrm{null}(\boldsymbol{V}_{\mathcal{S}^c}), \quad & \mathcal{M}^\perp &= \mathrm{row}(\boldsymbol{V}_{\mathcal{S}^c}), \\ \overline{\mathcal{M}} &= \mathrm{row}(\boldsymbol{V}_{\mathcal{S}}), \quad & \overline{\mathcal{M}}^\perp &= \mathrm{null}(\boldsymbol{V}_{\mathcal{S}}), \end{aligned}$$

and the projections

$$\begin{aligned} \boldsymbol{\theta}_{\overline{\mathcal{M}}} &:= \mathrm{Proj}_{\overline{\mathcal{M}}}(\boldsymbol{\theta}) = \boldsymbol{V}^t_{\mathcal{S}}\left(\boldsymbol{V}_{\mathcal{S}}\boldsymbol{V}^t_{\mathcal{S}}\right)^+\boldsymbol{V}_{\mathcal{S}}\boldsymbol{\theta} \\ \boldsymbol{\theta}_{\overline{\mathcal{M}}^\perp} &:= \mathrm{Proj}_{\overline{\mathcal{M}}^\perp}(\boldsymbol{\theta}) = \left[\boldsymbol{I}_p - \boldsymbol{V}^t_{\mathcal{S}}\left(\boldsymbol{V}_{\mathcal{S}}\boldsymbol{V}^t_{\mathcal{S}}\right)^+\boldsymbol{V}_{\mathcal{S}}\right]\boldsymbol{\theta} \end{aligned}$$

of a vector $\boldsymbol{\theta} \in \mathbb{R}^p$ onto spaces $\overline{\mathcal{M}}$ and $\overline{\mathcal{M}}^\perp$, respectively. The regularizer $\|\boldsymbol{V}\boldsymbol{\theta}\|_1$ is decomposable with respect to the model space pair $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$ in the sense that, for any $\boldsymbol{\beta} \in \mathcal{M}$ and $\boldsymbol{\gamma} \in \overline{\mathcal{M}}^\perp$, $\|\boldsymbol{V}(\boldsymbol{\beta} + \boldsymbol{\gamma})\|_1 = \|\boldsymbol{V}\boldsymbol{\beta}\|_1 + \|\boldsymbol{V}\boldsymbol{\gamma}\|_1$. Define a cone $\mathbb{C}$ by

$$\mathbb{C} := \{\boldsymbol{\Delta} \in \mathbb{R}^p : \|\boldsymbol{V}\boldsymbol{\Delta}_{\overline{\mathcal{M}}^\perp}\|_1 = \|\boldsymbol{V}_{\mathcal{S}^c}\boldsymbol{\Delta}_{\overline{\mathcal{M}}^\perp}\|_1 \le 3\|\boldsymbol{V}\boldsymbol{\Delta}_{\overline{\mathcal{M}}}\|_1\} \tag{15}$$

and a compatibility constant $\Psi = \sup_{\boldsymbol{\theta}\in\mathrm{row}(\boldsymbol{V}_\mathcal{S}),\|\boldsymbol{\theta}\|_2\le 1} \|\boldsymbol{V}\boldsymbol{\theta}\|_1$. Then we have the following deterministic error bounds for the discrepancy between the regularized estimate $\widehat{\boldsymbol{\beta}}(\rho)$ and the true parameter value $\boldsymbol{\beta}^*$.

Proposition 6.1. Suppose $\rho \ge 2\|\boldsymbol{V}(\boldsymbol{V}^t\boldsymbol{V})^{-1}\nabla f(\boldsymbol{\beta}^*)\|_\infty$ and f satisfies the restricted strong convexity on $\mathbb{C}$ with parameter $\kappa > 0$, that is, for all $\boldsymbol{\Delta} \in \mathbb{C}$,

$$f(\boldsymbol{\beta}^* + \boldsymbol{\Delta}) - f(\boldsymbol{\beta}^*) - \langle \nabla f(\boldsymbol{\beta}^*), \boldsymbol{\Delta}\rangle \ge \kappa\|\boldsymbol{\Delta}\|_2^2.$$

Then

$$\|\widehat{\boldsymbol{\beta}}(\rho) - \boldsymbol{\beta}^*\|_2 \le \frac{3}{2}\kappa^{-1}\rho\Psi,$$
$$\|\boldsymbol{V}[\widehat{\boldsymbol{\beta}}(\rho) - \boldsymbol{\beta}^*]\|_1 \le 6\kappa^{-1}\rho\Psi^2.$$

Now we specialize to the linear regression case $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$, where $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^t$ are iid mean zero random variables. The loss function under consideration is $f(\boldsymbol{\beta}) = \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$.

Corollary 6.1. Suppose that $\boldsymbol{X}$ satisfies the restricted eigenvalue condition

$$\frac{\|\boldsymbol{X}\boldsymbol{\theta}\|_2^2}{n} \ge \kappa\|\theta\|_2^2$$

for all $\boldsymbol{\theta} \in \mathbb{C}$ and the column normalization condition $\|[\boldsymbol{X}(\boldsymbol{V}^t\boldsymbol{V})^{-1}\boldsymbol{V}^t]_j/\sqrt{n}\|_2 \le 1$ for all $j = 1, \ldots, r$. With $\rho = 4\sigma\sqrt{(\ln r)/n}$ and the errors $\boldsymbol{\epsilon}$ are mean-zero sub-Gaussian random variables with constant σ^2, then with probability at least $1 - 2/r$,

$$\|\widehat{\boldsymbol{\beta}}(\rho) - \boldsymbol{\beta}^*\|_2 \le 6\kappa^{-1}\sigma\sqrt{\frac{\ln r}{n}}\Psi$$
$$\|\widehat{\boldsymbol{V}}[\boldsymbol{\beta}(\rho) - \boldsymbol{\beta}^*]\|_1 \le 24\kappa^{-1}\sigma\sqrt{\frac{\ln r}{n}}\Psi^2.$$

Proposition 6.1 and Corollary 6.1 highlight a few differences with the corresponding error bounds for lasso regularized estimates ($\boldsymbol{V} = \boldsymbol{I}_p$) (Negahban et al. 2012).

- The number of parameter p does not play a role in the error bounds; the number of regularization terms r does.
- The compatibility constant $\Psi = \sup_{\boldsymbol{\theta}\in\mathrm{row}(\boldsymbol{V}_\mathcal{S}),\|\boldsymbol{\theta}\|_2\le 1} \|\boldsymbol{V}\boldsymbol{\theta}\|_1$ emphasizes the effect of the structure of the regularization matrix $\boldsymbol{V}$ on the error bounds. For lasso, $\boldsymbol{V} = \boldsymbol{I}_p$ and $\Psi = \sup_{\|\boldsymbol{\theta}_\mathcal{S}\|_2\le 1}\|\boldsymbol{\theta}_\mathcal{S}\|_1 = \sqrt{s}$ (Negahban et al. 2012). For a general regularization matrix $\boldsymbol{V}$, there is no analytic expression for Ψ but it can be readily computed numerically.
- The lasso error bound requires the column normalization condition on the original design matrix $\boldsymbol{X}$; a general $\boldsymbol{V}$ imposes the same condition on the transformed matrix $\boldsymbol{X}(\boldsymbol{V}^t\boldsymbol{V})^{-1}\boldsymbol{V}^t$.

7. APPLICATIONS

In this section, we collect some representative regularized or constrained estimation problems and demonstrate how they can be solved by path following. For all applications, we list the first three derivatives of the loss function f in (1). In fact, the third derivative is only needed when implementing by solving the ODE for the sweep tableau.

In applications such as regularized GLMs, the tuning parameter ρ in the regularization problem (1) is chosen by a model selection criterion such as AIC, BIC, C_p, or cross-validation. The cross-validation errors can be readily computed using the solution path output by EPSODE. Yet the AIC, BIC, and C_p criteria require an estimate of the degrees of freedom of estimate $\boldsymbol{\beta}(\rho)$. In this article, we use $\mathrm{df}(\boldsymbol{\beta}(\rho)) = p - |\mathcal{Z}_\mathrm{E} \cup \mathcal{Z}_\mathrm{I}|$ as a measure of the degrees of freedom under GLMs. It has previously been shown to be an unbiased estimate of the degrees of freedom for lasso penalized least squares (Efron et al. 2004; Zou et al. 2007), generalized lasso penalized least squares (Tibshirani and Taylor 2011), and the least squares version of the regularized problem (1) (Zhou and Lange 2013). Using the same degrees of freedom formula for GLMs is justified by the local approximation of GLM log-likelihood by weighted least squares (Park and Hastie 2007).

7.1 GLMs and Quasilikelihoods With Generalized ℓ_1 Regularizations

The generalized linear model (GLM) deals with exponential families in which the sufficient statistics is Y and the conditional mean μ of Y completely determines its distribution. Conditional on the covariate vector $\boldsymbol{x} \in \mathbb{R}^p$, the response variable y is modeled as

$$p(y|\boldsymbol{x};\boldsymbol{\beta},\sigma) \propto \exp\left\{\frac{y\langle\boldsymbol{x},\boldsymbol{\beta}\rangle - \psi(\langle\boldsymbol{x},\boldsymbol{\beta}\rangle)}{c(\sigma)}\right\}, \tag{16}$$

where the scalar $\sigma > 0$ is a fixed and known scale parameter and the vector $\boldsymbol{\beta}$ is the parameters to be estimated. The function $\psi : \mathbb{R} \mapsto \mathbb{R}$ is the link function. When $y \in \mathbb{R}$, $\psi(u) = u^2/2$ and $c(\sigma) = \sigma^2$, (16) is the *normal regression model*. When $y \in \{0, 1\}$, $\psi(u) = \ln(1 + \exp(u))$, and $c(\sigma) = 1$, (16) is the *logistic regression model*. When $y \in \mathbb{N}$, $\psi(u) = \exp(u)$, and $c(\sigma) = 1$, (16) is the *Poisson regression model*.

The quasilikelihoods generalize GLM without assuming a specific distribution form of Y. Instead only a function relation between the conditional means μ_i and variances $\sigma_i^2, \sigma_i^2 = V(\mu_i)$ for some variance function V, is needed. Then the integral

$$Q(\mu, y) = \int_y^\mu \frac{y - t}{\sigma^2 V(t)}\,dt$$

behaves like a log-likelihood function under mild conditions and is called the quasilikelihood. The quasilikelihood includes GLMs as special cases with appropriately chosen variance function $V(\cdot)$. Readers are referred to the classical text (McCullagh and Nelder 1983, Table 9.1) for the commonly used quasilikelihoods. By slightly abusing our notation, we assume a known link function between the conditional mean μ_i and linear predictor $\boldsymbol{x}_i^T\boldsymbol{\beta}$, $\mu = \mu(\boldsymbol{x}_i^T\boldsymbol{\beta})$ and denote $Q_i(\boldsymbol{\beta}) = Q(\mu(\boldsymbol{x}_i^T\boldsymbol{\beta}), y_i)$. Then the quasilikelihood with generalized ℓ_1 regularization takes the form

$$-Q(\boldsymbol{\beta}) + \rho\|\boldsymbol{V}\boldsymbol{\beta} - \boldsymbol{d}\|_1 = -\sum_{i=1}^n Q_i(\boldsymbol{\beta}) + \rho\|\boldsymbol{V}\boldsymbol{\beta} - \boldsymbol{d}\|_1, \tag{17}$$

which is a special case of the general form (1). Specific choices of the regularization matrix $\boldsymbol{V}$ and constant vector $\boldsymbol{d}$ lead to lasso, fused-lasso, trend filtering, and many other applications.

For the path algorithm, we require the first two or three derivatives of the complete quasilikelihood. Denoting $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$ with $\boldsymbol{X} = (\boldsymbol{x}_1^t, \boldsymbol{x}_2^t, \ldots, \boldsymbol{x}_n^t)^t$, we have

$$\begin{aligned}
\nabla Q(\boldsymbol{\beta}) &= [D\boldsymbol{\mu}(\boldsymbol{\beta})]^t \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})/\sigma^2 \\
&= \boldsymbol{X}^t [D\boldsymbol{\mu}(\boldsymbol{\eta})] \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})/\sigma^2, \\
\boldsymbol{H}(\boldsymbol{\beta}) = d^2 Q(\boldsymbol{\beta}) &= [(\boldsymbol{y} - \boldsymbol{\mu})^t \boldsymbol{V}^{-1} \otimes \boldsymbol{X}^t] \cdot D^2\boldsymbol{\mu}(\boldsymbol{\eta}) \cdot \boldsymbol{X}/\sigma^2, \\
D\boldsymbol{H}(\boldsymbol{\beta}) = d^3 Q(\boldsymbol{\beta}) &= [\boldsymbol{X}^t \otimes (\boldsymbol{y} - \boldsymbol{\mu})^t \boldsymbol{V}^{-1} \otimes \boldsymbol{X}^t] \\
&\quad \cdot D^3\boldsymbol{\mu}(\boldsymbol{\eta}) \cdot \boldsymbol{X}/\sigma^2, \qquad (18)
\end{aligned}$$

where $\boldsymbol{V}$ is a n-by-n diagonal matrix with diagonal entries $V(\mu(\boldsymbol{x}_i^t \boldsymbol{y}))$, $D\boldsymbol{\mu}(\boldsymbol{\eta})$ is a n-by-n diagonal matrix with diagonal entries $\mu'(\boldsymbol{x}_i^t\boldsymbol{\beta})$, $D^2\boldsymbol{\mu}(\boldsymbol{\eta})$ is a n^2-by-n matrix with $(n(i-1)+i, i)$ entry equal to $\mu''(\boldsymbol{x}_i^t\boldsymbol{\beta})$ for $i = 1, \ldots, n$ and 0 otherwise, and $D^3\boldsymbol{\mu}(\boldsymbol{\eta})$ is a n^3-by-n matrix with $(n^2(i-1)+n(i-1)+i, i)$ entry equal to $\mu'''(\boldsymbol{x}_i^t\boldsymbol{\beta})$ for $i = 1, \ldots, n$ and 0 otherwise. These formulas simplify for GLM with canonical link.

The most widely used ℓ_1 regularization is the lasso penalty which imposes sparsity on the regression coefficients. For numerical demonstration, we revisit the M&A example introduced in Section 1 without discretizing each predictor. We standardize each predictor first and consider the lasso penalized linear logistic regression model. Figure 3 shows the lasso solution path for each standardized predictor in the left panel and corresponding AIC and BIC scores in the right panel. The order at which predictors enter the model matches the more detailed patterns revealed by the varying coefficient model in Figure 1. The almost monotone effects of the predictors "market-to-book ratio," "cash flow," "cash," and "tax" can be captured by the usual linear logistic regression and these covariates are picked up by lasso first. The nonlinear effects shown in the other predictors are likely to be missed by the linear logistic regression. For instance, the quadratic effects of "log market equity" shown in the regularized estimates in Figure 1 are missed by both AIC and BIC criteria.

7.2 Shape-Restricted Regressions

Order-constrained regression has been an important modeling tool (Robertson, Wright, and Dykstra 1988; Silvapulle and Sen 2005). If $\boldsymbol{\beta}$ denotes the parameter vector, monotone regression imposes *isotone* constraints $\beta_1 \le \beta_2 \le \cdots \le \beta_p$ or *antitone* constraints $\beta_1 \ge \beta_2 \ge \cdots \ge \beta_p$. In *partially ordered regression*, subsets of the parameters are subject to isotone or antitone constraints. In some other problems, it is sensible to impose *convex* or *concave* constraints. Note that if locations of parameters are at irregularly spaced time points $t_1 \le t_2 \le \cdots \le t_p$, convexity translates into the constraints

$$\frac{\beta_{i+2} - \beta_{i+1}}{t_{i+2} - t_{i+1}} \ge \frac{\beta_{i+1} - \beta_i}{t_{i+1} - t_i}$$

for $1 \le i \le p-2$. When the time intervals are uniform, the constraints simplify to $\beta_{i+2} - \beta_{i+1} \ge \beta_{i+1} - \beta_i, i = 1, 2, \cdots, p-1$. Concavity translates into the opposite set of inequalities.

Most of previous works have focused on the linear regression problems because of the computational and theoretical complexities in the generalized linear model setting. The recent work (Rufibach 2010) proposes an active set algorithm for GLMs with order constraints. The EPSODE algorithm conveniently provides a solution to the linearly constrained estimation problem (2). The relevant derivatives of loss function are listed in (18). It is noteworthy that EPSODE not only provides the constrained estimate but also the whole path bridging the unconstrained estimate to the constrained solution. Availability of the whole solution path renders model selection between the two extremes simple.

In the illustrative M&A example of Section 1, the bin predictors for the "market-to-book ratio" are regularized by the antitone constraint and those for the "log market equity" covariate by the concavity constraint.

7.3 Gaussian Graphical Models

In recent years, several authors (Friedman, Hastie, and Tibshirani 2008; Yuan 2008) proposed to estimate the sparse undirected graphical model by using lasso regularizations to the log-likelihood function of the precision matrix, the

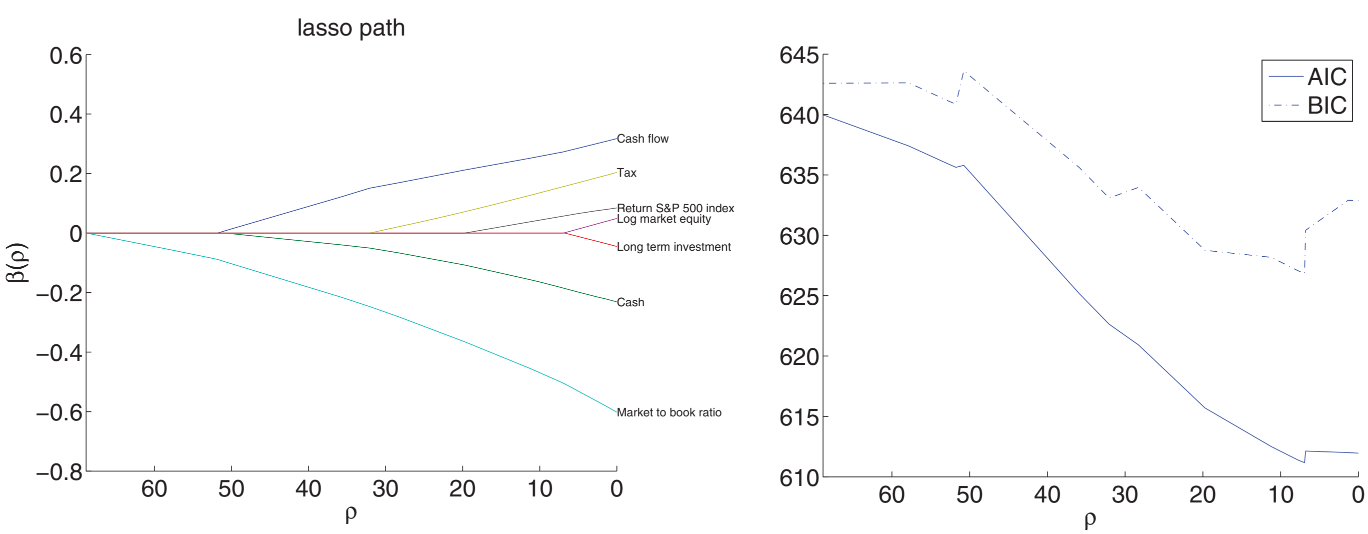

Figure 3. M&A example revisited. Lasso solution path on the seven standardized predictors.

inverse of the variance-covariance matrix. Given an observed variance–covariance matrix $\hat{\boldsymbol{\Sigma}} \in R^{p\times p}$, the negative log-likelihood of the precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ under normal assumption is

$$f(\boldsymbol{\Omega}) = -\log\det\boldsymbol{\Omega} + \mathrm{tr}(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Omega}) \tag{19}$$

with the MLE solution $\hat{\boldsymbol{\Sigma}}^{-1}$ when $\hat{\boldsymbol{\Sigma}}$ is nondegenerate. A zero in the precision matrix implies conditional independence of the corresponding nodes. Graphical lasso proposes to solve

$$f(\boldsymbol{\Omega}) + \rho\sum_{i<j}|\omega_{ij}|, \tag{20}$$

where $\rho \geq 0$ is the tuning constant and ω_{ij} denotes the (i, j)-element of $\boldsymbol{\Omega}$. It is well known that the determinant function is log-concave (Magnus and Neudecker 1999). Therefore, the loss function f (19) is convex and the EPSODE algorithm applies to (20). Friedman, Hastie, and Tibshirani (2008) proposed an efficient coordinate descent procedure for solving (20) at a fixed ρ. A recent attempt to approximate the whole solution path is made by Yuan (2008). Again his path algorithm can be deemed as a primitive predictor-corrector method for approximating the ODE solution.

With symmetry in mind, we parameterize $\boldsymbol{\Omega}$ in terms of its lower triangular part by a $p(p+1)/2$ column vector $\boldsymbol{x}$ and let $D\boldsymbol{\Omega}(\boldsymbol{x}) = \frac{\partial\mathrm{vec}\boldsymbol{\Omega}}{\partial(\mathrm{vec}\boldsymbol{x})^t}$ be the corresponding p^2-by-$p(p+1)/2$ Jacobian matrix. Note $D\boldsymbol{\Omega}(\boldsymbol{x})\cdot\boldsymbol{x} = \mathrm{vec}\boldsymbol{\Omega}(\boldsymbol{x})$ and each row of $D\boldsymbol{\Omega}(\boldsymbol{x})$ has exactly one nonzero entry which equals unity. We list here the first three derivatives of f.

Lemma 7.1.

1. The derivatives for the Gaussian graphical model (19) with respect to $\boldsymbol{\Omega}$ are

$$\begin{aligned} Df(\boldsymbol{\Omega}) &= df(\boldsymbol{\Omega}) = [\mathrm{vec}(-\boldsymbol{\Omega}^{-1} + \boldsymbol{\Sigma})]^t \\ D^2f(\boldsymbol{\Omega}) &= d^2f(\boldsymbol{\Omega}) = \boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}^{-1} \\ D^3f(\boldsymbol{\Omega}) &= -(\boldsymbol{I}_n\otimes\boldsymbol{K}_{nn}\otimes\boldsymbol{I}_n)\cdot[\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}^{-1}\otimes\mathrm{vec}(\boldsymbol{\Omega}^{-1}) \\ &\quad + \mathrm{vec}(\boldsymbol{\Omega}^{-1})\otimes\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}^{-1}], \end{aligned}$$

   where $\boldsymbol{K}_{nn}$ is the commutation matrix (Magnus and Neudecker 1999).

2. The derivatives for the Gaussian graphical model (19) with respect to $\boldsymbol{x}$ are

$$\begin{aligned} Df(\boldsymbol{x}) &= Df(\boldsymbol{\Omega})\cdot D\boldsymbol{\Omega}(\boldsymbol{x}) \\ \boldsymbol{H}(\boldsymbol{x}) &= D^2f(\boldsymbol{x}) = [D\boldsymbol{\Omega}(\boldsymbol{x})]^t\cdot D^2f(\boldsymbol{\Omega})\cdot D\boldsymbol{\Omega}(\boldsymbol{x}) \\ D\boldsymbol{H}(\boldsymbol{x}) &= D^3f(\boldsymbol{x}) = \{[D\boldsymbol{\Omega}(\boldsymbol{x})]^t\otimes[D\boldsymbol{\Omega}(\boldsymbol{x})]^t\} \\ &\quad \cdot D^3f(\boldsymbol{\Omega})\cdot D\boldsymbol{\Omega}(\boldsymbol{x}). \end{aligned}$$

When the covariance matrix $\hat{\boldsymbol{\Sigma}}$ is nonsingular, EPSODE can be initiated either at $\rho = 0$ or $\rho = \infty$. When $\hat{\boldsymbol{\Sigma}}$ is singular, we start from $\rho = \infty$ and the extended version of EPSODE (14) should be used. If starting at $\rho = 0$, the solution is initialized at $\hat{\boldsymbol{\Sigma}}^{-1}$; if starting at $\rho = \infty$, the solution is initialized at $\mathrm{diag}(\hat{\sigma}_{ii}^{-1})$. Minimization of both the unpenalized and penalized objective function has to be performed over the convex cone of symmetric, positive semidefinite matrices, which is not explicitly incorporated in our path following algorithm. The next result ensures the positive definiteness of the path solution.

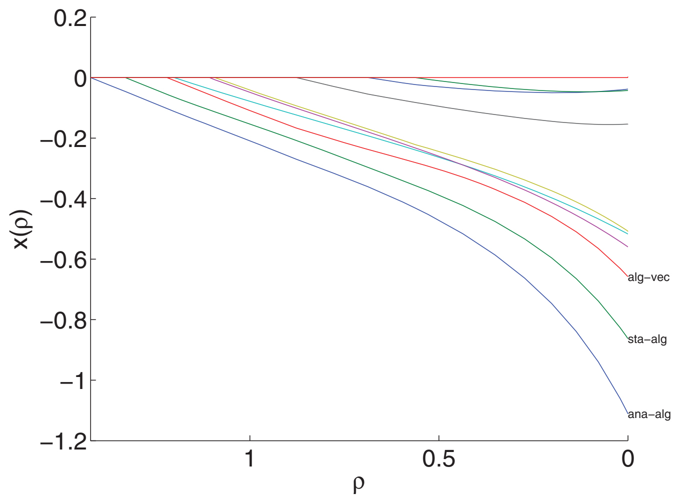

Figure 4. Solution path of the 10 edges in lasso-regularized Gaussian graphical model for the math score data. The top three edges chosen by lasso are labeled.

Lemma 7.2. (Positive definiteness along the path). The path solution $\boldsymbol{\Omega}(\rho)$ minimizes (20) over the convex cone of symmetric, positive semidefinite matrices.

We illustrate the path algorithm by the classical example of 88 students' scores on five math courses—mechanics, vector, algebra, analysis, and statistics (Mardia, Kent, and Bibby 1979, Table 1.2.1). Figure 4 displays the solution path from EPSODE. The top three edges chosen by lasso are analysis-algebra, statistics-algebra, and algebra-vector.

7.4 Nonparametric Density Estimation

The maximum likelihood estimation for nonparametric density estimation often involves a nontrivial, high-dimensional constrained optimization problem. In this section, we briefly demonstrate the applicability of EPSODE to the maximum likelihood estimation of univariate log-concave density. Extensions to multivariate log-concave density estimation (Cule, Gramacy, and Samworth 2009; Cule, Samworth, and Stewart 2010) will be pursued elsewhere. It is noteworthy that, besides providing an alternative solver for log-concave density estimation, EPSODE offers the whole solution path between the unconstrained and constrained solutions. For example, an "almost" log-concave density estimate in the middle of the path can be chosen that minimizes cross-validation or prediction error. This adds another dimension to the flexibility of nonparametric modeling.

The family of log-concave densities is an attractive modeling tool. It includes most of the commonly used parametric distributions as special cases. Examples include normal, gamma with shape parameter ≥ 1, and beta densities with both parameters ≥ 1. The survey article (Walther 2009) gives a recent review. A probability density $g(\cdot)$ on $\mathbb{R}$ is log-concave if its logarithm $\phi(x) = \ln g(x)$ is concave. Given iid observations, from an unknown distribution of density $g(\cdot)$, with support at points

$x_1 < \cdots < x_n$ with corresponding frequencies $p_1, \ldots, p_n$, it is well known (Walther 2002) that the nonparametric MLE of g exists, is unique and takes the form $\hat{g} = \exp(\hat{\phi})$, where $\hat{\phi}$ is continuous and piecewise linear on $[x_1, x_n]$ with the set of knots contained in $\{x_1, \ldots, x_n\}$, and $\hat{\phi} = -\infty$ outside the interval $[x_1, x_n]$. This implies that the MLE is obtained by minimizing the strictly convex function

$$f(\boldsymbol{\phi}) = -\sum_{i=1}^{n} p_i\phi_i + \sum_{k=1}^{n-1}(x_{k+1} - x_k)\int_0^1 e^{(1-t)\phi_k + t\phi_{k+1}}\,dt$$

over $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_n)^t \in \mathbb{R}^n$ subject to constraints

$$\frac{\phi_{i+1} - \phi_i}{x_{i+1} - x_i} \le \frac{\phi_i - \phi_{i-1}}{x_i - x_{i-1}}, \quad i = 2, \ldots, n-1.$$

The consistency of the MLE is proved by Pal, Woodroofe, and Meyer (2007) and the pointwise asymptotic distribution of the MLE studied in Balabdaoui, Rufibach, and Wellner (2009).

Following Duembgen, Rufibach, and Huesler (2007), we use notations

$$\delta_0 = \delta_n = 0,\ \delta_i = x_{i+1} - x_i, \quad i = 1, \ldots, n-1$$

$$J(r,s) = \int_0^1 e^{(1-t)r+ts}\,dt = \begin{cases} \dfrac{e^s - e^r}{s - r} & r \ne s \\ e^r & r = s \end{cases}.$$

Then the objective function becomes

$$f(\boldsymbol{\phi}) = -\sum_{i=1}^{n} p_i\phi_i + \sum_{k=1}^{n-1}\delta_k J(\phi_k, \phi_{k+1}).$$

The path algorithm requires up to the third derivative of the objective function f

$$[\nabla f(\boldsymbol{\phi})]_i = -p_i + \delta_{i-1}J_{01}(\phi_{i-1}, \phi_i) + \delta_i J_{10}(\phi_i, \phi_{i+1})$$

$$[H(\boldsymbol{\phi})]_{ij} = [d^2 f(\boldsymbol{\phi})]_{ij} = \begin{cases} \delta_{i-1}J_{11}(\phi_{i-1}, \phi_i) & j = i-1 \\ \delta_{i-1}J_{02}(\phi_{i-1}, \phi_i) + \delta_i J_{20}(\phi_i, \phi_{i+1}) & j = i \\ \delta_i J_{11}(\phi_i, \phi_{i+1}) & j = i+1 \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial[H(\boldsymbol{\phi})]_{i,i-1}}{\partial\phi_k} = \begin{cases} \delta_{i-1}J_{21}(\phi_{i-1}, \phi_i) & k = i-1 \\ \delta_{i-1}J_{12}(\phi_{i-1}, \phi_i) & k = i \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial[H(\boldsymbol{\phi})]_{i,i}}{\partial\phi_k} = \begin{cases} \delta_{i-1}J_{12}(\phi_{i-1}, \phi_i) & k = i-1 \\ \delta_{i-1}J_{03}(\phi_{i-1}, \phi_i) + \delta_i J_{30}(\phi_i, \phi_{i+1}) & k = i \\ \delta_i J_{21}(\phi_i, \phi_{i+1}) & k = i+1 \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial[H(\boldsymbol{\phi})]_{i,i+1}}{\partial\phi_k} = \begin{cases} \delta_i J_{21}(\phi_i, \phi_{i+1}) & k = i \\ \delta_i J_{12}(\phi_i, \phi_{i+1}) & k = i+1 \\ 0 & \text{otherwise} \end{cases}.$$

Interchanging the derivative and integral operators, justified by the dominated convergence theorem, gives a useful representation for the partial derivatives of J

$$J_{ab}(r,s) = \tfrac{\partial^{a+b}}{\partial r^a \partial s^b}J(r,s) = \textstyle\int_0^1 (1-t)^a t^b e^{(1-t)r+ts}\,dt.$$

We derive a recurrence relation for $J_{ab}(r,s)$ to facilitate its computation.

Lemma 7.3. $J_{ab}(r,s)$ satisfy following recurrence

1. For $r \ne s$,

$$J_{00}(r,s) = \frac{e^s - e^r}{s-r}, \qquad J_{10}(r,s) = -\frac{e^r}{s-r} + \frac{e^s - e^r}{(s-r)^2}$$

$$J_{01}(r,s) = \frac{e^s}{s-r} - \frac{e^s - e^r}{(s-r)^2},$$

$$J_{11}(r,s) = \frac{e^s + e^r}{(s-r)^2} - \frac{2(e^s - e^r)}{(s-r)^3}$$

$$J_{ab}(r,s) = \frac{a+b+s-r}{s-r}J_{a-1,b}(r,s) - \frac{a-1}{s-r}J_{a-2,b}(r,s)$$

$$J_{ab}(r,s) = -\frac{a+b-s+r}{s-r}J_{a,b-1}(r,s) + \frac{b-1}{s-r}J_{a,b-2}(r,s).$$

2. For $r = s$,

$$J_{ab}(r,s) = \frac{e^r a!b!}{(a+b+1)!} = \frac{a}{a+b+1}J_{a-1,b} = \frac{b}{a+b+1}J_{a,b-1}.$$

To illustrate the path algorithm for this problem, we simulate $n = 25$ points from the extremal distribution Gumbel(0,1). Figure 5 displays the constrained and unconstrained estimates of ϕ_i and the solution path bridging the two.

7.5 Kernel Machines

In recent years, kernel methods are widely used in nonlinear regression and classification problems (Scholkopf and Smola 2001). During the review of this article, a referee brought to our attention that various path algorithms in kernel machine methods can be unified in the ODE framework. In this section, we briefly indicate this connection. Given data $(y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$, the responses $y_i \in \mathbb{R}$ are connected to the features $\boldsymbol{x}_i \in \mathbb{R}^p$ through a nonparametric function $h(\boldsymbol{x}_i) = \beta_0 + \sum_{j=1}^n \beta_i k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, where $k : \mathbb{R}^p \times \mathbb{R}^p \mapsto \mathbb{R}$ is a positive definite kernel function and h belongs to the reproducing kernel Hilbert space $\mathcal{H}_k$. The coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$ are estimated by minimizing the criterion $L(\boldsymbol{y}, h(\boldsymbol{x})) + \rho\|h\|^2_{\mathcal{H}_K}$, where L is a loss function and ρ is the regularization parameter. Let $\boldsymbol{K} = [k(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j} \in \mathbb{R}^{p\times p}$ be the kernel matrix based on the observed features $\boldsymbol{x}_i$, $\tilde{\boldsymbol{K}} = [\mathbf{1}_n, \boldsymbol{K}] \in \mathbb{R}^{p\times(p+1)}$, and $\tilde{\boldsymbol{\beta}} = [\beta_0, \boldsymbol{\beta}^t]^t$. This leads to the regularization problem

$$\min_{\boldsymbol{\beta}} \quad L(\boldsymbol{y}, \beta_0\mathbf{1}_n + \boldsymbol{K\beta}) + \rho\boldsymbol{\beta}^t\boldsymbol{K\beta} = L(\boldsymbol{y}, \tilde{\boldsymbol{K}}\tilde{\boldsymbol{\beta}}) + \rho\boldsymbol{\beta}^t\boldsymbol{K\beta},$$

for which the regularization path is sought. Here are a few examples:

- The support vector machines rely on the hinge loss $L(\boldsymbol{y}, \boldsymbol{X}) = \|\mathbf{1}_n - \mathrm{diag}(\boldsymbol{y})\tilde{\boldsymbol{K}}\tilde{\boldsymbol{\beta}}\|_+$. By switching the roles of loss and penalty, we see the criterion belongs to the EPSODE framework (1) with a quadratic $\boldsymbol{\beta}^t\boldsymbol{K\beta}$ loss and

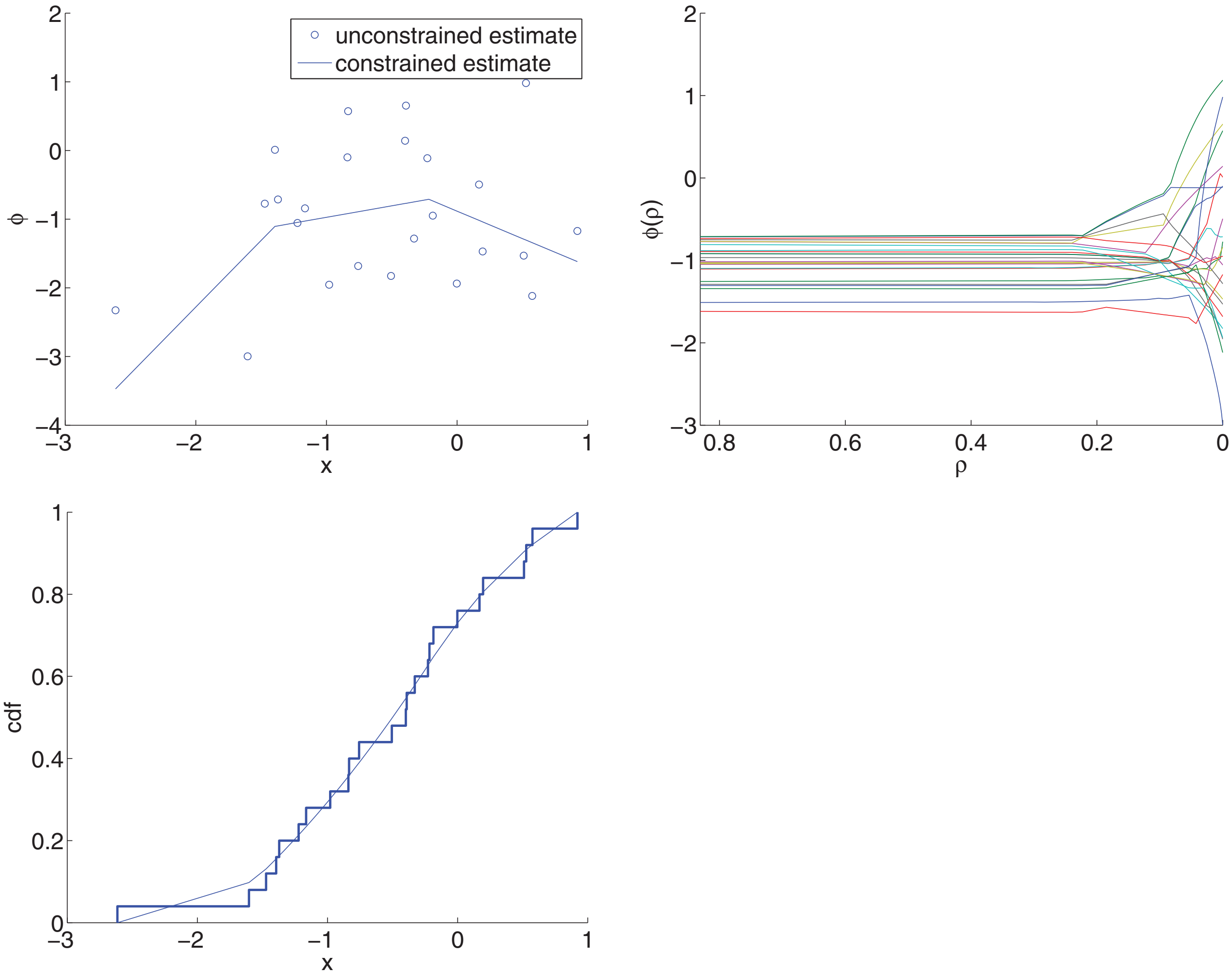

Figure 5. Log-concave density estimation. $n = 25$ points are generated from Gumbel(0,1) distribution. Top left: Unconstrained and concavity-constrained estimates ϕ. Top right: Solution path. Bottom left: Empirical cdf and the cdf of MLE density.

inequality regularization specified by $\boldsymbol{W} = -\mathrm{diag}(\boldsymbol{y})\tilde{\boldsymbol{K}}\tilde{\boldsymbol{\beta}}$ and $\boldsymbol{e} = -\mathbf{1}_n$. Since the Hessian of a quadratic function is constant, the path following directions (11) and (14) are constant, which leads to the piecewise linear solution path originally derived in Hastie et al. (2004). By similar arguments, the kernel quantile regression (Li et al. 2007) also admits piecewise linear solution path and allows fast computation.

- For regression with squared error loss, $L(\boldsymbol{y}, \boldsymbol{X}) = \|\boldsymbol{y} - \tilde{\boldsymbol{K}}\tilde{\boldsymbol{\beta}}\|_2^2$. At optimal solution, $\beta_0 = \frac{1}{n}\mathbf{1}_n^t(\boldsymbol{y} - \boldsymbol{K}\boldsymbol{\beta})$; thus it suffices to minimize $\|\boldsymbol{y} - (\boldsymbol{I}_n - \mathbf{1}_n\mathbf{1}_n^t)\boldsymbol{K}\boldsymbol{\beta}\|_2^2 + \rho\boldsymbol{\beta}^t\boldsymbol{K}\boldsymbol{\beta}$ for $\boldsymbol{\beta}$. This overall quadratic criterion admits an analytic solution $\widehat{\boldsymbol{\beta}}(\rho) = [\boldsymbol{K}^t(\boldsymbol{I}_n - \mathbf{1}_n\mathbf{1}_n^t)\boldsymbol{K} + \rho\boldsymbol{K}]^{-1}\boldsymbol{K}^t(\boldsymbol{I}_n - \mathbf{1}_n\mathbf{1}_n^t)\boldsymbol{y}$ at each ρ. Suppose, the kernel matrix $\boldsymbol{K}$ is row (column)-centered and admits eigendecomposition $\boldsymbol{K} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^t$, then $\widehat{\boldsymbol{\beta}}(\rho) = \boldsymbol{U}(\boldsymbol{D}^2 + \rho\boldsymbol{D})^{-1}\boldsymbol{D}\boldsymbol{U}^t\boldsymbol{y}$ can be computed efficiently at any ρ and dismisses the need for special path following method.
- Nonlinear logistic regression uses the binomial deviance loss $L(\boldsymbol{y}, \boldsymbol{X}) = \sum_{i=1}^n \{y_i h(\boldsymbol{x}_i) - \ln(1 + e^{h(\boldsymbol{x}_i)})\}$, with $h(\boldsymbol{x}_i) = \beta_0 + \sum_{j=1}^n \beta_i k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Although the regularized criterion $L(\boldsymbol{y}, \boldsymbol{X}) + \rho\boldsymbol{\beta}^t\boldsymbol{K}\boldsymbol{\beta}$ does not belong to the EPSODE formulation, the same argument as the for Proposition 3.1 shows that the path following direction is given by

$$\begin{pmatrix} \dfrac{d\hat{\beta}_0(\rho)}{d\rho} \\ \dfrac{d\hat{\boldsymbol{\beta}}(\rho)}{d\rho} \end{pmatrix} = -2[d^2L(\tilde{\boldsymbol{\beta}}) + 2\rho\tilde{\boldsymbol{K}}_0]^{-1}\boldsymbol{K}\begin{pmatrix} 0 \\ \hat{\boldsymbol{\beta}}(\rho) \end{pmatrix},$$

where $\tilde{\boldsymbol{K}}_0 \in \mathbb{R}^{(p+1)\times(p+1)}$ is the original kernel matrix $\boldsymbol{K}$ augmented by an extra (first) row and (first) column of 0.

8. CONCLUSIONS

In this article, we propose a generic path following algorithm EPSODE that works for any regularization problems of form (1). The advantages are its simplicity and generality. Path following only involves solving ODEs segment by segment and is simple to implement using popular softwares such as R and Matlab. Besides providing the whole regularization path, it also gives a solver for linearly constrained optimization problems that frequently arise in statistics. Our applications to shape-restricted regressions and nonparametric density estimation are special cases in particular.

Several extensions deserve further study. Current algorithm requires sufficient smoothness (twice differentiable) in the loss function. This precludes certain applications with nonsmooth objective function, for example, the Huber loss in robust estimation and the loss function in quantile regression. Generalization of our path algorithm to regularization of these loss functions requires further research. Another restriction in our formulation is the linearity in the regularization terms. In sparse regressions, several authors have proposed nonlinear and nonconvex penalties. The bridge regression (Frank and Friedman 1993) and SCAD penalties (Fan and Li 2001) fall into this category. As observed in (Friedman 2008), when the penalty is not convex, the solution path may not be continuous and poses difficulty in path following, which strongly depends on the continuity and smoothness of the solution path. Fortunately, in these problems, the discontinuities only occur when new variables enter or leave the model. A promising strategy is to initialize the starting point of next segment by solving an equality constrained optimization problem. This again invites further investigation. Lastly, our formulation (1) imposes same penalty parameter on equality and inequality regularization terms. Relaxing to different tuning parameters apparently increases flexibility of the regularization scheme. In this setup, the relevant target will be a "solution surface" instead of "solution path," which is worth further investigation.

SUPPLEMENTARY MATERIALS

Proofs for: Lemmas 2.1, 7.2, and 7.3; and Propositions 3.2, 4.1, and 6.1.

[Received August 2011. Revised June 2013.]

REFERENCES

Balabdaoui, F., Rufibach, K., and Wellner, J. A. (2009), "Limit Distribution Theory for Maximum Likelihood Estimation of a Log-Concave Density," *The Annals of Statistics*, 37, 1299–1331. [696]

Cule, M., Gramacy, R. B., and Samworth, R. (2009), "LogConcDEAD: An R Package for Maximum Likelihood Estimation of a Multivariate Log-Concave Density," *Journal of Statistical Software*, 29, 1–20. [695]

Cule, M., Samworth, R., and Stewart, M. (2010), "Maximum Likelihood Estimation of a Multi-Dimensional Log-Concave Density," *Journal of the Royal Statistical Society,* Series B, 72, 545–607. [695]

Dempster, A. P. (1969), *Elements of Continuous Multivariate Analysis (Addison-Wesley Series in Behavioral Sciences)*, Reading, MA: Addison-Wesley. [691]

Donoho, D. L., and Johnstone, I. M. (1994), "Ideal Spatial Adaptation by Wavelet Shrinkage," *Biometrika*, 81, 425–455. [686]

Duembgen, L., Rufibach, K., and Huesler, A. (2007), "Active Set and EM Algorithms for Log-Concave Densities Based on Complete and Censored Data," Technical Report 61, IMSV, University of Bern, Switzerland. [696]

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression" (with discussion), *The Annals of Statistics*, 32, 407–499. [688,693]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [698]

Fan, J., Maity, A., Wang, Y., and Wu, Y. (2013), "Parametrically Guided Generalized Additive Models With Application to Merger and Acquisition Data," *Journal of Nonparametric Statistics*, 25, 109–128. [687]

Fraley, C., and Percival, D. (2010), "Model-Averaged ℓ_1 Regularization Using Markov Chain Monte Carlo Model Composition," Technical Report 541, Department of Statistics, University of Washington, Seattle, WA. [689]

Frank, I. E., and Friedman, J. H. (1993), "A Statistical View of Some Chemometrics Regression Tools," *Technometrics*, 35, 109–135. [698]

Friedman, J. (2008), "Fast Sparse Regression and Classification," available at *http://www-stat.stanford.edu/jhf/ftp/GPSpaper.pdf*. [689,691,698]

Friedman, J., Hastie, T., and Tibshirani, R. (2000), "Additive Logistic Regression: A Statistical View of Boosting," *The Annals of Statistics*, 28, 337–407. [686]

——— (2008), "Sparse Inverse Covariance Estimation With the Graphical Lasso," *Biostatistics*, 9, 432–441. [694,695]

Ghosh, D., and Yuan, Z. (2009), "An Improved Model Averaging Scheme for Logistic Regression," *Journal of Multivariate Analysis*, 100, 1670–1681. [689]

Goodnight, J. H. (1979), "A Tutorial on the Sweep Operator," *The American Statistician*, 33, 149–158. [691]

Hastie, T., Rosset, S., Tibshirani, R., and Zhu, J. (2004), "The Entire Regularization Path for the Support Vector Machine," *Journal of Machine Learning Research*, 5, 1391–1415. [697]

Hastie, T., and Tibshirani, R. (1993), "Varying-Coefficient Models" (with discussion), *Journal of the Royal Statistical Society,* Series B, 55, 757–796. [687]

Jennrich, R. (1977), "Stepwise Regression," in *Statistical Methods for Digital Computers*, New York: Wiley-Interscience, pp. 58–75. [691]

Kim, S.-J., Koh, K., Boyd, S., and Gorinevsky, D. (2009), "l_1 Trend Filtering," *SIAM Review*, 51, 339–360. [688]

Lange, K. (2010), *Numerical Analysis for Statisticians (Statistics and Computing)* (2nd ed.), New York: Springer. [691]

Lawson, C. L., and Hanson, R. J. (1987), *Solving Least Squares Problems (Classics in Applied Mathematics)*, Philadelphia, PA: Society for Industrial Mathematics. [686,692]

Li, Y., Liu, Y., and Zhu, J. (2007), "Quantile Regression in Reproducing Kernel Hilbert Spaces," *Journal of the American Statistical Association*, 102, 255–268. [697]

Little, R. J. A., and Rubin, D. B. (2002), *Statistical Analysis With Missing Data (Wiley Series in Probability and Statistics)* (2nd ed.), Hoboken, NJ: Wiley-Interscience. [691]

Magnus, J. R., and Neudecker, H. (1999), *Matrix Differential Calculus With Applications in Statistics and Econometrics (Wiley Series in Probability and Statistics)*, Chichester: Wiley. [692,695]

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979), *Multivariate Analysis (Probability and Mathematical Statistics: A Series of Monographs and Textbooks)*, London: Academic Press. [695]

McCullagh, P., and Nelder, J. A. (1983), *Generalized Linear Models (Monographs on Statistics and Applied Probability)*, London: Chapman & Hall. [693]

Negahban, S., Ravikumar, P. D., Wainwright, M. J., and Yu, B. (2012), "A Unified Framework for High-Dimensional Analysis of m-Estimators With Decomposable Regularizers," *Statistical Science*, 27, 538–557. [692,693]

Osborne, M. R., Presnell, B., and Turlach, B. A. (2000), "A New Approach to Variable Selection in Least Squares Problems," *IMA Journal of Numerical Analysis*, 20, 389–403. [688]

Pal, J. K., Woodroofe, M., and Meyer, M. (2007), "Estimating a Polya frequency Function$_2$," in *Complex Datasets and Inverse Problems,* volume 54 of *IMS Lecture Notes Monograph Series*, Beachwood, OH: Institute of Mathematical Statististics, pp. 239–249. [696]

Park, M. Y., and Hastie, T. (2007), "L_1-Regularization Path Algorithm for Generalized Linear Models," *Journal of the Royal Statistical Society,* Series B, 69, 659–677. [689,691,693]

Robertson, T., Wright, F. T., and Dykstra, R. L. (1988), *Order Restricted Statistical Inference (Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics)*, Chichester: Wiley. [686,694]

Rosset, S., and Zhu, J. (2007), "Piecewise Linear Regularized Solution Paths," *The Annals of Statistics*, 35, 1012–1030. [688]

Rufibach, K. (2010), "An Active Set Algorithm to Estimate Parameters in Generalized Linear Models With Ordered Predictors," *Computational Statistics & Data Analysis*, 54, 1442–1456. [694]

Ruszczyński, A. (2006), *Nonlinear Optimization*, Princeton, NJ: Princeton University Press. [689]

Scholkopf, B., and Smola, A. J. (2001), *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, Cambridge, MA: MIT Press. [696]

Silvapulle, M. J., and Sen, P. K. (2005), *Constrained Statistical Inference: Inequality, Order, and Shape Restrictions (Wiley Series in Probability and Statistics)*, Hoboken, NJ: Wiley-Interscience. [686,694]

Soetaert, K., Petzoldt, T., and Setzer, R. W. (2010), "Solving Differential Equations in R: Package deSolve," *Journal of Statistical Software*, 33, 1–25. [691]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society,* Series B, 58, 267–288. [686]

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005), “Sparsity and Smoothness via the Fused Lasso,” *Journal of the Royal Statistical Society,* Series B, 67, 91–108. [686,692]

Tibshirani, R. J., and Taylor, J. (2011), “The Solution Path of the Generalized Lasso,” *The Annals of Statistics*, 39, 1335–1371. [688,693]

Walther, G. (2002), “Detecting the Presence of Mixing With Multiscale Maximum Likelihood,” *Journal of the American Statistical Association*, 97, 508–513. [696]

——— (2009), “Inference and Modeling With Log-Concave Distributions,” *Statistical Science*, 24, 319–327. [695]

Wu, Y. (2011), “An Ordinary Differential Equation-Based Solution Path Algorithm,” *Journal of Nonparametric Statistics*, 23, 185–199. [689,691]

——— (2012), “Elastic Net for Coxs Proportional Hazards Model With a Solution Path Algorithm,” *Statistica Sinica*, 22, 271–294. [689]

Yuan, M. (2008), “Efficient Computation of ℓ_1 Regularized Estimates in Gaussian Graphical Models,” *Journal of Computational and Graphical Statistics*, 17, 809–826. [694,695]

Zhou, H., and Lange, K. (2013), “A Path Algorithm for Constrained Estimation,” *Journal of Computational and Graphical Statistics*, 22, 261–283. [688,690,693]

Zou, H., Hastie, T., and Tibshirani, R. (2007), “On the ‘Degrees of Freedom’ of the Lasso,” *The Annals of Statistics*, 35, 2173–2192. [693]