Algorithms for Fitting the Constrained Lasso

Brian R. Gaines & Hua Zhou
Department of Statistics, North Carolina State University & Department of Biostatistics, University of California, Los Angeles

Introduction

Constrained Lasso (James et al., 2013):

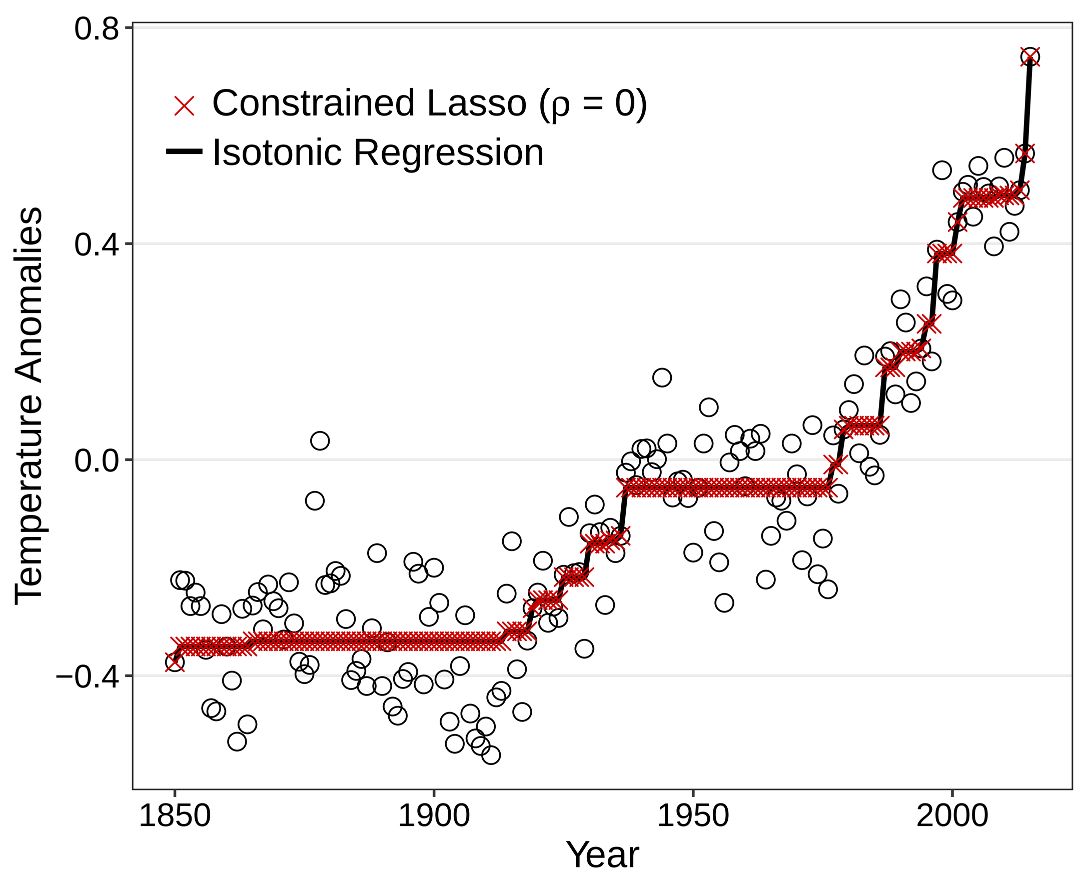
$$\begin{array}{ll}\text{minimize} & \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\rho\|\boldsymbol{\beta}\|_1 \\ \text{subject to} & \boldsymbol{A}\boldsymbol{\beta}=\boldsymbol{b} \text{ and } \boldsymbol{C}\boldsymbol{\beta}\leq \boldsymbol{d}.\end{array} \quad (1)$$

Augments the lasso (Tibshirani, 1996) with linear equality & inequality constraints

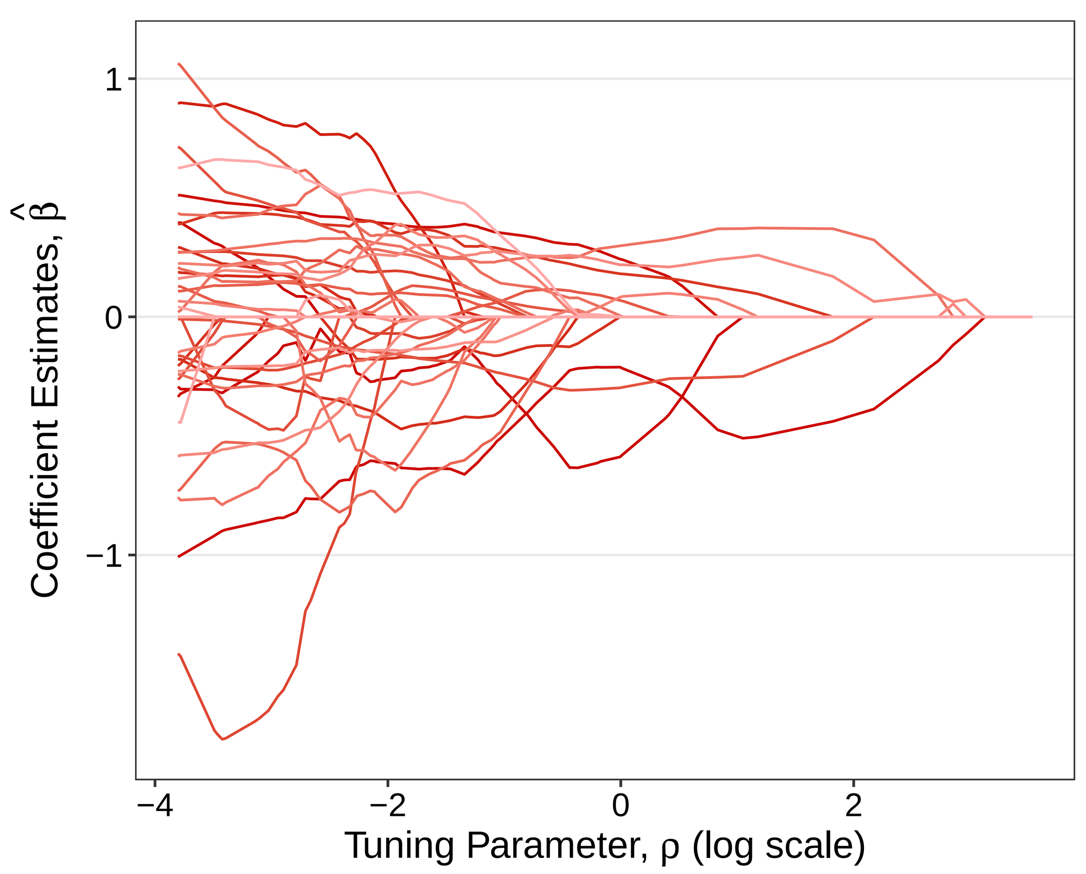
- Can impose prior knowledge on the coefficient estimates
  - Monotonicity, non-negativity, sum-to-zero, etc.

Motivation

Isotonic Lasso: Global Warming Data



Constrained Lasso Solution Path



Algorithms (fixed value of ρ)

Quadratic Programming

Standard QP form:

$$\begin{array}{ll}\text{minimize} & \frac{1}{2}\boldsymbol{x}'\boldsymbol{H}\boldsymbol{x}+\boldsymbol{f}'\boldsymbol{x} \\ \text{subject to} & \boldsymbol{A}\boldsymbol{x}=\boldsymbol{b} \text{ and } \boldsymbol{C}\boldsymbol{x}\leq \boldsymbol{d}.\end{array} \quad (2)$$

Trick: decompose $\beta=\beta^+-\beta^-$, then plug into (1) and massage into the form of (2)

- $|\beta|=\beta^++\beta^-$ handles the ℓ_1 penalty term

Alternating Direction Method of Multipliers

A problem of the form

$$\begin{array}{ll}\text{minimize} & f(\boldsymbol{x})+g(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{x}+\boldsymbol{z}=\boldsymbol{c}\end{array}$$

can be solved using the following updates

$$\begin{aligned}\boldsymbol{x}^{(t+1)} &\leftarrow \mathbf{prox}_{\lambda f}(\boldsymbol{z}^{(t)}-\boldsymbol{c}+\boldsymbol{u}^{(t)})\\ \boldsymbol{z}^{(t+1)} &\leftarrow \mathbf{prox}_{\lambda g}(\boldsymbol{x}^{(t+1)}-\boldsymbol{c}+\boldsymbol{u}^{(t)})\\ \boldsymbol{u}^{(t+1)} &\leftarrow \boldsymbol{u}^{(t)}+\boldsymbol{x}^{(t+1)}+\boldsymbol{z}^{(t+1)}-\boldsymbol{c}\end{aligned}$$

Solution Path Algorithm (all values of ρ)

Steps to deriving the path algorithm:

- Derive the KKT stationarity conditions
- Apply the implicit function theorem to view as a function of ρ

$$\frac{d}{d\rho}\begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{\lambda}\\ \boldsymbol{\mu}\end{pmatrix}=-\begin{pmatrix}\boldsymbol{X}^T\boldsymbol{X} & \boldsymbol{A}^T & \boldsymbol{C}^T\\ \boldsymbol{A} & 0 & 0\\ \boldsymbol{C} & 0 & 0\end{pmatrix}^{-1}\begin{pmatrix}\boldsymbol{s}\\ 0\\ 0\end{pmatrix}$$

- Determine what keeps the RHS of this equation constant for the non-zero ("active") coefficients
  - Constant derivatives $\Rightarrow$ piecewise linear path
  - Changes in the derivatives correspond to kinks in the path

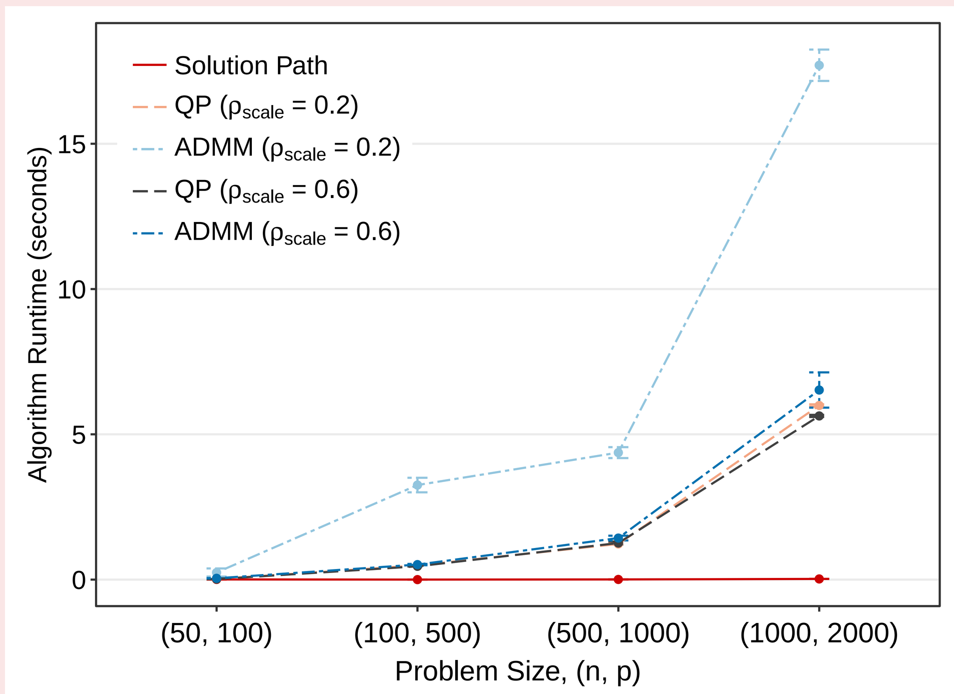
Events to monitor along the path:

- A non-zero ("active") coefficient becomes 0 (or vice versa)
- An inequality constraint hits or escapes the boundary ($\boldsymbol{c}_l^T\boldsymbol{\beta}=d_l$)
- Violations of the subgradient conditions

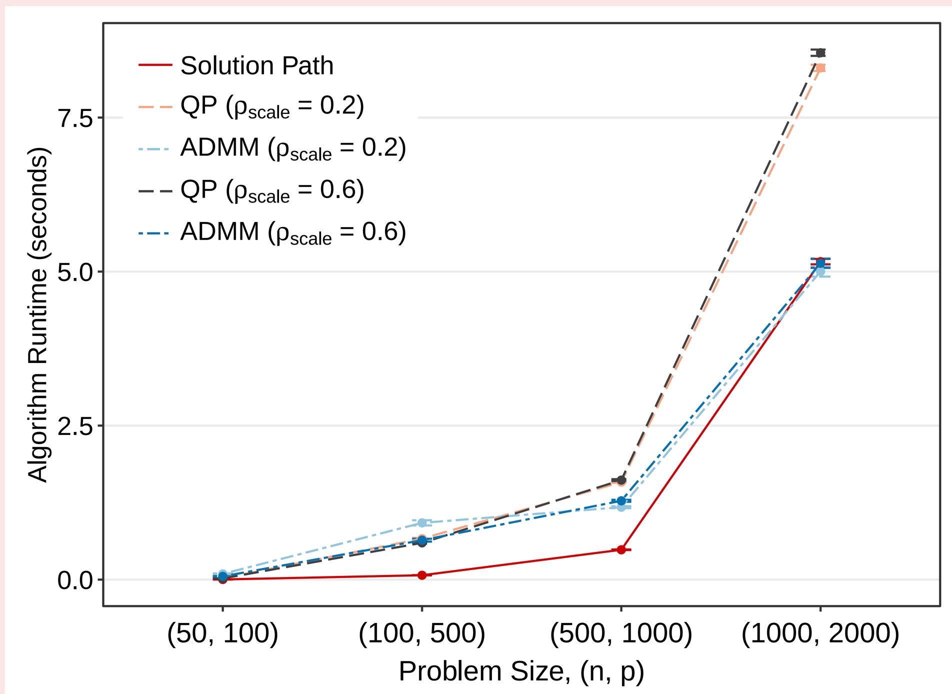
$$\frac{\partial}{\partial\beta_j}\|\boldsymbol{\beta}\|_1=s_j(\rho)=\begin{cases}1 & \beta_j(\rho)>0\\ [-1,1] & \beta_j(\rho)=0\\ -1 & \beta_j(\rho)<0\end{cases}$$

Algorithm Runtime Comparison: Simulation Results

Sum-to-zero Constraint



Non-negative Lasso



Algorithm runtimes from 20 replicates using an Intel i7-4510U 2.0 GHz processor with 8 GB memory. QP was solved using the Gurobi Optimizer. ADMM used a convergence tolerance of 10^{-4}. Solution path algorithm is averaged across the number of kinks in the path. ρ_{scale} is the value of ρ as a fraction of the maximum ρ; that is, $\rho=\rho_{\text{scale}}\cdot\max(\rho)$.

- ADMM is comparable or faster than QP but is more sensitive to ρ
- Solution path is comparable to other methods at 10-15 values of ρ, but gives entire path
- ADMM is usually faster than QP
- Solution path has less of an advantage, but at worst is comparable to other methods

Connection to the Generalized Lasso

Generalized lasso (Tibshirani & Taylor, 2011):

$$\text{minimize} \quad \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\rho\|\boldsymbol{D}\boldsymbol{\beta}\|_1, \quad (3)$$

$\boldsymbol{D}\in\mathbb{R}^{m\times p}$ is a fixed regularization matrix.

General formulation, includes many variants of the lasso as special cases

- Lasso, fused and sparse fused lassos (1d, 2d, and graph versions)
- Smoothing & trend filtering

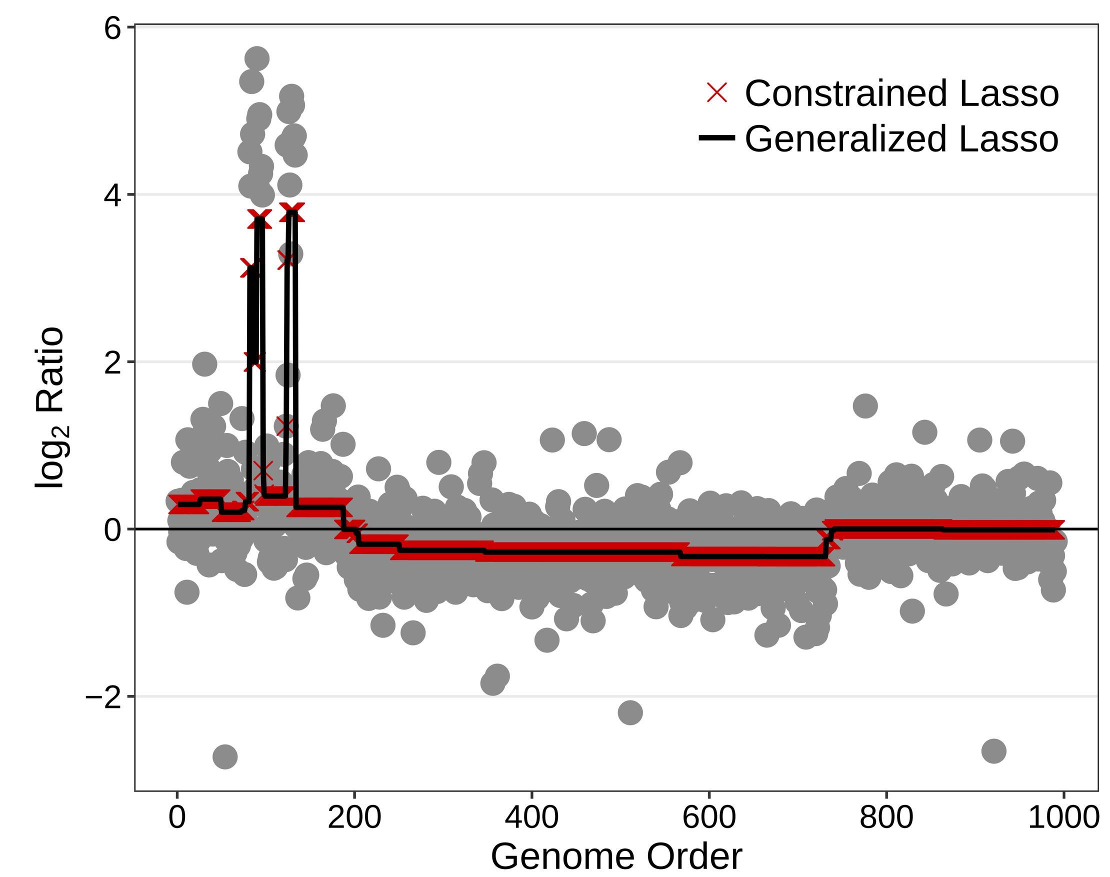
Can it be transformed into a constrained lasso?

Using a change of variables, after simplifying we ultimately end up with a constrained lasso problem

$$\begin{array}{ll}\text{minimize} & \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{D}^+\boldsymbol{\alpha}-\boldsymbol{X}\boldsymbol{V}_2\boldsymbol{\gamma}\|_2^2+\rho\|\boldsymbol{\alpha}\|_1 \\ \text{subject to} & \boldsymbol{U}_2^T\boldsymbol{\alpha}=\boldsymbol{0}.\end{array} \quad (4)$$

- Holds for any regularization matrix $\boldsymbol{D}$
  - Other researchers required special structure on $\boldsymbol{D}$ (James et al., 2013; Tibshirani & Taylor, 2011)

Sparse Fused Lasso: Brain Tumor Data



- Verifies use of the constrained lasso to solve a generalized lasso

Contributions

- Derive & compare various algorithms for solving the constrained lasso
  - General formulation, no need to reinvent the wheel
  - Novel derivation of an efficient solution path algorithm
- Show that any generalized lasso can be formulated as a constrained lasso
  - Algorithms/results applicable to a large class of problems
- Matlab code available in SparseReg toolbox on Github

NC STATE UNIVERSITY

brgaines@ncsu.edu