A Review on "Algorithms for Fitting the Constrained Lasso"

Yunran Chen

1 Introduction

A constrained lasso problem is defined as a normal lasso problem with linear equality and inequality constraints (James et al., 2013; Gaines et al., 2018):

$$\begin{aligned} &\texttt{minimize} && \tfrac{1}{2}\|y - X\beta\|_2^2 + \rho\|\beta\|_1 \\ &\texttt{subject to} && A\beta = b \quad \texttt{and} \quad C\beta \leq d, \end{aligned} \tag{1}$$

where $y \in \mathcal{R}^n$ is the response vector, $X \in \mathcal{R}^{n\times p}$ is the design matrix of covariates, $\beta \in \mathcal{R}^p$ is the parameter vector we are interested in, and $\rho \geq 0$ is a tuning parameter controlling the degree of penalty. The constraints can represent the prior knowledge on parameters in practice. For example, we may expect all the coefficients are positive, or all the coefficients are summed up to one, or the coefficients are in an increasing order. These constraints are natural and common in real data analysis(Wu et al., 2001) and thus the constrained lasso problem attracts great attention in recent research (James et al., 2013; Hu et al., 2015; He, 2011). Moreover, constrained lasso can also be considered as an extended version of generalized lasso ($\texttt{minimize} \quad \frac{1}{2}\|y - X\beta\|_2^2 + \rho\|D\beta\|_1$) proposed by Tibshirani et al. (2011) since any generalized lasso problem can be transformed to a constrained lasso problem(Gaines et al., 2018). Therefore, constraint lasso can provide a more flexible framework for data analysis.

Gaines et al. (2018) derive three different algorithms (quadratic programming, ADMM and path algorithms) for the constrained lasso problem and conducted simulations suggesting the path algorithm outperform the other two algorithms in terms of the running time and sensitivity to the tuning parameter. I will mainly focus on adding more details on the path algorithm derivation, commenting on its merits and points of confusion, and discussing possible extensions.

2 Path Algorithm

The solution of the constrained lasso problem depends on the values of tuning parameters. Specifically, the solution path is a piecewise linear function of ρ, with a change happening whenever one of the four events happens: (i) an active coefficient hits 0; (ii) An inactive coefficient becomes active; (iii) A strict inequality constraint hits the boundary; (iv) An inequality constraint escapes the boundary. This behavior allows for a computation ease since we only need to consider the optimization problem on a subset of parameters and interpolated the solution between each kinks along the path. We keep track of two sets $\mathcal{A} = \{j : \beta_j \neq 0\}, \quad \mathcal{Z}_I = \{l : c_l^\top \beta = d_l\}$. The algorithm is as follows:

Initial: $k = 0, \rho^0 = \rho^{max}, \beta^0, \mathcal{A}^0, \mathcal{Z}_I^0$
While $\rho^k > 0$:

- Compute a solution at ρ_k by any optimization method.
- Compute hitting time for event (i) and (iii), denoted as $h_{k+1,i}$ and $h_{k+1,iii}$.
- Compute leaving time for event (ii) and (iv), denote as $l_{k+1,ii}$ and $l_{k+1,iv}$.
- Set $\rho_{k+1} = max\{h_{k+1,i}, l_{k+1,ii}, h_{k+1,iii}, l_{k+1,iv}\}$.
  If $h_{k+1,i} < l_{k+1,ii}$, add the hitting coordinate to $\mathcal{A}$, otherwise, remove it from $\mathcal{A}$.
  If $h_{k+1,ii} < l_{k+1,iv}$, add the hitting coordinate to $\mathcal{Z}_I$, otherwise, remove it from $\mathcal{Z}_I$.
  Set $k = k + 1$.

This algorithm is derived via Karush-kuhn-Tucker (KKT) optimality conditions. We first derive the Lagrangian of the constrained lasso problem:

$$L(\beta, \rho, \lambda, \mu) = \frac{1}{2}\|y - X\beta\|_2^2 + \rho\|\beta\|_1 + \lambda^T(A\beta - b) + \mu^T(C\beta - d)$$

By taking the derivative with respect to β, we obtain the stationary condition

$$-X^T(y - X\beta(\rho)) + \rho s(\rho) + A^T\lambda(\rho) + C^T\mu(\rho) = 0 \quad (2)$$

where $s(\rho)$ is the subgradient $\partial\|\beta\|_1$
Other necessary conditions are:

$$\begin{aligned} (\text{primal feasibility}) \quad & C\beta \le d, \quad A\beta = b \\ (\text{dual feasibility}) \quad & \mu_i \ge 0 \\ (\text{complementary slackness}) \quad & \mu_i(C\beta - d)_i = 0 \end{aligned} \quad (3)$$

From the stationary condition, we can prove the solution path is piecewise linear. Notice we only need to track the coordinate in the set $\mathcal{A}, \mathcal{Z}_I$, we obtain the following stationary condition:

$$-X^T_{:,\mathcal{A}}(y - X\beta(\rho)) + \rho s(\rho) + A^T_{:,\mathcal{A}}\lambda(\rho) + C^T_{\mathcal{Z}_I,\mathcal{A}}\mu(\rho)_{\mathcal{Z}_I} = 0_{|\mathcal{A}|} \quad (4)$$

By applying implicit function theorem, we have

$$\frac{d}{d\rho}\begin{pmatrix} \beta_{\mathcal{A}} \\ \lambda \\ \mu_{\mathcal{Z}_I} \end{pmatrix} = -\begin{pmatrix} X^T:,\mathcal{A}X_{:,\mathcal{A}} & A^T_{:,\mathcal{A}} & C^T_{\mathcal{Z}_I,\mathcal{A}} \\ A_{:,\mathcal{A}} & 0 & 0 \\ C_{\mathcal{Z}_I,\mathcal{A}} & 0 & 0 \end{pmatrix}^{-1} \begin{pmatrix} s_{\mathcal{A}} \\ 0 \\ 0 \end{pmatrix}$$

If the set $\mathcal{A}, \mathcal{Z}_I$ remains unchanged, the derivation is a constant, denoting the piecewise linearity of the solution path. The complementary slackness (equation (3)) defined the set $\mathcal{Z}_I$.

To obtain the next changing time ρ_{k+1}, instead of directly calculating hitting time and leaving time, consider solving a constrained optimization problem:

$$\begin{aligned} &\text{minimize} \quad \Delta\rho \\ &\text{subject to} \quad \text{any of the following constraints :} \\ &\qquad \beta_{\mathcal{A}}^{(k+1)} = \beta_{\mathcal{A}}^{(k)} - \Delta\rho\frac{d}{d\rho}\beta_{\mathcal{A}}^{(k)} = 0_{|\mathcal{A}|} \\ &\qquad \rho^{(k+1)}s_{\mathcal{A}^c}^{(k+1)} = \rho^{(k)}s_{\mathcal{A}^c}^{(k)} - \Delta\rho(\rho s_{\mathcal{A}^c}) = \pm(\rho^{(k)} - \Delta\rho)1_{|\mathcal{A}^c|} \\ &\qquad r_{\mathcal{Z}_I^c}^{(k+1)} = r_{\mathcal{Z}_I^c}^{(k)} - \Delta\rho\frac{d}{d\rho}r_{\mathcal{Z}_I^c} = 0_{|\mathcal{Z}_I^c|} \\ &\qquad \mu_{\mathcal{Z}_I}^{k+1} = \mu_{\mathcal{Z}_I}^{k} - \Delta\rho\frac{d}{d\rho}\mu_{\mathcal{Z}_I} = 0_{|\mathcal{Z}_I|}, \end{aligned}$$

where $\Delta\rho := \rho^{k+1} - \rho^k$, and $r_{\mathcal{Z}_I^c} := C_{\mathcal{Z}_I^c,\mathcal{A}}\beta_{\mathcal{A}} - d_{\mathcal{Z}_I^c}$ represents the inequality residuals. The four constraints correspond to four possible events that may cause kinks in the solution path. It is worth noting that all these derivatives removes linearly with the gradient.

For the initialization, consider $\rho = \infty$, L1-penalty will dominate the objective function of constrained lasso problem as follows:

$$\begin{aligned} &\text{minimize} && \|\beta\|_1 \\ &\text{subject to} && A\beta = b \quad \text{and} \quad C\beta \leq d, \end{aligned}$$

Solve the above problem, we can obtain the initial $\beta^0, \mathcal{A}^0, \mathcal{Z}_I^0$ and Lagrange multiplier λ^0, μ^0. Pluging these values into the stationary condition (2), we can obtain initial value for ρ as follows:

$$\rho^{max} = max|x_j^T(y - X\beta^0 - A_{\cdot,j}^T)\lambda^0 - C_{\mathcal{Z}_I,j}^T\mu_{\mathcal{Z}_I}^0|.$$

3 Discussion

The great advantage of the path algorithm is its computation efficiency due to the piecewise linearity and partial parameter updating. First, the solution path of constrained lasso is piecewise linear of the tuning parameter, allowing for only solving the solution at the kinks and then applying an interpolation between these kinks. The piecewise linearity reduce the continuous space of ρ to a space includes only a few points, which greatly eases the computation. Second, the solution computation at each time k only involves a subset of the data (within the set $\mathcal{A}$ and $\mathcal{Z}_I$), which greatly eases the dimension and is scalable for problems with large dimension for parameter space.

There are three possible extensions that may be interesting. Firstly, we can consider extending the linear constraints to nonlinear constraints. In natural science, decision science and experimental science, the theoretical constraints may take a nonlinear form. If we consider a nonlinear constrained lasso problem, although KKT conditions still hold, the piecewise linearity of the solution path and the linearity of the gradients when solving the changing time may not hold. In order to inherit the merits of the path algorithm, a possible solution can be using a piecewise linear function to approximate the nonlinear function. Probably we can adopt Majority Maximization idea in the approximation to obtain similar convergence result.

Secondly, we can consider extending the regression to a tensor regression, where we extend $\beta \in \mathcal{R}^{p\times 1}$ to $\beta \in \mathcal{R}^{p\times m}, m \geq 1$. For example, researchers may be interested in disease diagnosis from medical imaging data. It may be important to keep several important patterns while denoising an image. Since we still has a linear regression with linear constraints, the piecewise linearity of solution path still hold in this case. If we draw an analogy to the path algorithm for 2D fused lasso proposed by Tibshirani et al. (2011), for such extension, we may only need to trace occurrence of more events which can be derived from KKT conditions.

Thirdly, we may consider a Bayesian constrained lasso model by imposing the constraints on the prior. Park and Casella (2008) and Hans (2010) propose Bayesian version of the lasso by introducing a double exponential prior on the coefficients:

$$\begin{aligned} Y|\alpha, \beta^s, \phi \sim &\quad N(1_n\alpha + X^s\beta^s, I_n/\phi) \\ \beta^s|\alpha, \phi, \tau \sim &\quad N(0, diag(\tau^2)/\phi) \\ \tau_1^2, ..., \tau_p^2 \sim &\quad Exp(\lambda^2/2)(iid) \\ p(\alpha, \phi) \propto &\quad 1/\phi \end{aligned}$$

which induces a double exponential prior for β: $\beta_j|\phi, \lambda \sim DE(\lambda\sqrt{\phi})(iid)$. Other alternative scale mixtures prior are proposed, such as horseshoe prior (Carvalho et al., 2010) and generalized

double Pareto prior (Armagan et al., 2013). An extension allowing for constraints seems natural in Bayesian framework. For example, if we want to restrict all coefficients greater than zero, we may consider a normal prior on β truncated to $(0, \infty)$. If we want to restrict coefficients sum up to 1, we may consider a Dirichlet prior. However, Frequentists' lasso outperforms Bayesian lasso in terms of computation efficiency.

References

Armagan, A., D. B. Dunson, J. Lee, W. U. Bajwa, and N. Strawn (2013). Posterior consistency in linear models under shrinkage priors. *Biometrika 100*(4), 1011–1018.

Carvalho, C. M., N. G. Polson, and J. G. Scott (2010). The horseshoe estimator for sparse signals. *Biometrika 97*(2), 465–480.

Gaines, B. R., J. Kim, and H. Zhou (2018). Algorithms for fitting the constrained lasso. *Journal of Computational and Graphical Statistics 27*(4), 861–871.

Hans, C. (2010). Model uncertainty and variable selection in bayesian lasso regression. *Statistics and Computing 20*(2), 221–229.

He, T. (2011). Lasso and general l1-regularized regression under linear equality and inequality constraints.

Hu, Q., P. Zeng, and L. Lin (2015). The dual and degrees of freedom of linearly constrained generalized lasso. *Computational Statistics & Data Analysis 86*, 13–26.

James, G. M., C. Paulson, and P. Rusmevichientong (2013). Penalized and constrained regression. *Unpublished manuscript, http://www-bcf. usc. edu/˜ gareth/research/Research. html*.

Park, T. and G. Casella (2008). The bayesian lasso. *Journal of the American Statistical Association 103*(482), 681–686.

Tibshirani, R. J., J. Taylor, et al. (2011). The solution path of the generalized lasso. *The Annals of Statistics 39*(3), 1335–1371.

Wu, W. B., M. Woodroofe, and G. Mentz (2001). Isotonic regression: Another look at the change-point problem. *Biometrika 88*(3), 793–804.