Model Selection and Estimation in High-dimensional Generalized Linear Models

Francisco Rivera   Jiafeng (Kevin) Chen

December 7, 2018

1 Introduction

The workhorse for high-dimensional regressions is the ℓ_1 lasso penalty (Tibshirani, 1996). The original lasso is developed for fitting (normal) linear models with the objective[1]

$$\widehat{\boldsymbol{\beta}}_{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_1, \; \lambda > 0. \tag{1}$$

The first paper about lasso in a GLM setting is Park and Hastie (2007): Consider a scalar GLM with likelihood

$$L(y; \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\},$$

where $g(\mathbb{E}[y]) = \eta = \boldsymbol{x}^\intercal\boldsymbol{\beta}$ for some scalar function g. Thus we may consider a direct extension of (1):

$$\widehat{\boldsymbol{\beta}}_{\text{GLM-lasso}} = \arg\min_{\boldsymbol{\beta}} \underbrace{-\frac{1}{n}\sum_{i=1}^{n}\left[y_i\theta(\boldsymbol{\beta})_i - b\left(\theta(\boldsymbol{\beta})_i\right)\right]}_{\ell_n(\boldsymbol{\beta})} + \lambda \|\boldsymbol{\beta}\|_1, \tag{2}$$

[1] We consider $\boldsymbol{y}$ an n-vector of response variables whose ith element is $\boldsymbol{y}_i$. We consider the covariate matrix $\boldsymbol{X}$ $(n \times p)$. We always assume that $\boldsymbol{y}, \boldsymbol{X}$ are demeaned so the intercept term is zero, as the intercept term is usually not regularized.

where (2) reduces to (1) if the likelihood is Gaussian and g is the canonical link for the Gaussian model, which is the identity function. Note that the objective (2) is convex if we use the canonical link, since the exponential family log-likelihood is concave in θ, and $\theta = \boldsymbol{x}^\intercal \boldsymbol{\beta}$ is linear in $\boldsymbol{\beta}$ for the canonical link.

2 Model Selection with Lasso

3 Estimation

Efron et al. (2004) gives an efficient estimation procedure for the linear lasso (1) called the Least Angle Regression (LAR), which relies on the fact that the *regularization path*—$\widehat{\boldsymbol{\beta}}_i$ as a function of λ—is piecewise linear in (1). Such a structure is often unavailable in applications like (2). The most popular method—proposed by Friedman, Hastie, and Tibshirani (2010) and implemented in R's glmnet package—is *cyclical coordinate descent* with iteratively reweighted least squares. The idea is to approximate $\ell_n(\boldsymbol{\beta})$ in (2) with a second-order Taylor expansion, either globally for all parameters $\boldsymbol{\beta}$ (for scalar-valued GLMs) or locally with a single parameter $\boldsymbol{\beta}_j$ (for vector-valued GLMs, such as the multinomial logistic regression). Such an approximation yields a quadratic function (in the scalar GLM case)

$$\ell_Q(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} w_i (y_i - \boldsymbol{x}_i^\intercal \boldsymbol{\beta})^2.$$

We then solve a local penalized least-squares problem:

$$\boldsymbol{\beta} \leftarrow \arg\min_{\boldsymbol{\beta}} \ell_Q(\boldsymbol{\beta}) + \lambda \left\|\boldsymbol{\beta}\right\|_1. \tag{3}$$

via cyclical coordinate descent, i.e. by iteratively solving

$$\boldsymbol{\beta}_j \leftarrow \arg\min_{\boldsymbol{\beta}_j} \ell_Q(\boldsymbol{\beta}) + \lambda \left\|\boldsymbol{\beta}\right\|_1, \tag{4}$$

holding all other entries $\boldsymbol{\beta}_{-j}$ fixed. (4) has an analytical solution for the lasso penalty[2]

$$\boldsymbol{\beta}_j \leftarrow \frac{S\left(\sum_{i=1}^N w_i x_{ij}\left(y_i - \tilde{y}_i^{(j)}\right), \lambda\right)}{\sum_{i=1}^N w_i x_{ij}^2}, \; S(t,\gamma) = \operatorname{sgn}(t)\left(|z| - \gamma\right)_+, \; \tilde{y}_i^{(j)} = \boldsymbol{x}_i^\intercal \boldsymbol{\beta} - x_{ij}\boldsymbol{\beta}_j. \quad (5)$$

We summarize the procedure described above in Algorithm 1.

Algorithm 1 Cyclic coordinate descent algorithm for solving (2) (scalar GLM case) in Friedman, Hastie, and Tibshirani (2010)

```
Initialize β
for λ on regularization path do
    while β has not coverged do
        Approximate ℓ_n(β) by ℓ_Q(β)
        while cyclical descent has not converged do
            for j do
                Update β_j according to (5)
            end for
        end while
    end while
    β̂_λ ← β
    Initialize β for next iteration to β̂_λ
end for
```

The machine learning literature slightly alters Algorithm 1 and changes (5) into

$$\boldsymbol{\beta}_j \leftarrow S\left(\boldsymbol{\beta}_j - \left(\nabla_{\boldsymbol{\beta}} \ell_n(\boldsymbol{\beta})\right)_j \kappa^{-1}, \frac{\lambda}{\kappa}\right)$$

for some *learning rate* $1/\kappa$, in keeping with gradient descent. Moreover, Shalev-Shwartz and Tewari (2011) proves a convergence guarantee for stochastic coordinate descent in this fashion, where, instead of cycling through the coordinates of $\boldsymbol{\beta}$, a coordinate is chosen uniformly at random.

Theorem 1. Let $Q(\boldsymbol{\beta})$ be the objective in (2). At iteration T of the first while-loop in a

[2] Friedman, Hastie, and Tibshirani (2010) show a similar expression for the *elastic net* penalty:

$$\lambda P_\alpha(\boldsymbol{\beta}) = \lambda\left(\alpha \left\|\boldsymbol{\beta}\right\|_1 + (1-\alpha)\left\|\boldsymbol{\beta}\right\|_2\right).$$

verison of Algorithm 1 with stochastic coordinate descent and gradient updates,

$$\mathbb{E}[Q(\boldsymbol{\beta}_T)] - \mathbb{E}[Q(\widehat{\boldsymbol{\beta}}_{\text{GLM-lasso}})] \leq C\frac{p\kappa}{T+1}$$

for constant C a function of the initial starting value $\boldsymbol{\beta}^{(0)}$, assuming that ℓ_n is differentiable with

$$\ell_n(\boldsymbol{\beta} + \eta \boldsymbol{e}_j) \leq \ell_n(\boldsymbol{\beta}) + \eta \left(\nabla \ell_n\right)_j + \frac{\kappa}{2}\eta^2$$

for all $\eta, \boldsymbol{\beta}, j$.[3]

Corollary 2. The runtime to achieve ϵ expected accuracy is bounded by

$$O\left(\frac{np\kappa}{\epsilon}\left\|\widehat{\boldsymbol{\beta}}_{\text{GLM-lasso}}\right\|_2^2\right).$$

Moreover, Bradley et al. (2011) show that a parallel version of the coordinate gradient descent procedure above where at each iteration, P (possibly duplicate) coordinates are updated in parallel. For correlated features, such parallelism is dangerous, since updating two correlated features simulataneously may over or undercompensate for the gradient direction. Bradley et al. (2011) quantifies the interference due to correlated features and shows that efficiency increases linearly in the number of parallel processes P so long as $\mathsf{P} \leq \frac{p}{\rho}$ where ρ is the largest modulus of the eigenvalues of $X^\intercal X$.

Coordinate descent methods described above can also become expensive if n is large. The standard machine learning and optimization answer to this problem is to use *stochastic gradient descent*, replacing $\nabla_{\boldsymbol{\beta}} \ell_n(\boldsymbol{\beta})$ with an unbiased estimate $\boldsymbol{g}_i = \nabla_{\boldsymbol{\beta}} \log L(y_i; \boldsymbol{\beta})$, which is the gradient evaluated on a single observation.[4] Shalev-Shwartz and Tewari (2011) consider a mirror descent algorithm in the lasso context, by running stochastic gradient descent on the dual problem and enforcing sparsity in an intelligent manner. Let $\boldsymbol{\gamma} = f(\boldsymbol{\beta})$ be the dual

[3] This condition restricts the choice of κ as a function of the loss criterion.

[4] We can replace this with *batched gradient descent* as well, where the gradient estimate is averaging over a batch of observations.

parameter for $\boldsymbol{\beta}$ with an invertible link f. We choose an observation i at random, compute $\boldsymbol{g}_i$, and update

$$
\begin{aligned}
\boldsymbol{\gamma} &\leftarrow \boldsymbol{\gamma} - \eta \boldsymbol{g}_i \\
\boldsymbol{\gamma}' &\leftarrow \boldsymbol{\gamma} - \eta\lambda \operatorname{sgn}(\boldsymbol{\gamma}) && \text{(Decrease } \|\boldsymbol{\beta}\|_1\text{)} \\
\boldsymbol{\gamma}_j &\leftarrow \boldsymbol{\gamma}'_j \mathbb{1}\left(\operatorname{sgn}(\boldsymbol{\gamma}_j) = \operatorname{sgn}(\boldsymbol{\gamma}'_j)\right) && \text{(Maintains sparsity)} \\
\boldsymbol{\beta} &\leftarrow f^{-1}(\boldsymbol{\gamma}).
\end{aligned}
$$

The runtime bound for the stochastic mirror descent algorithm in Shalev-Shwartz and Tewari (2011) is

$$
O\left(\frac{p\log p}{\epsilon^2}\left\|\widehat{\boldsymbol{\beta}}_{\text{GLM-lasso}}\right\|_2^2\right).
$$

We pay the price of the $p\log p$ and ϵ^{-2} dependence, as opposed to p and ϵ^{-1}, in order to achieve the benefit of a n-free runtime.

References

Bradley, Joseph K, Aapo Kyrola, Danny Bickson, and Carlos Guestrin. 2011. “Parallel coordinate descent for l1-regularized loss minimization.” *arXiv preprint arXiv:1105.5379* .

Efron, Bradley, Trevor Hastie, Iain Johnstone, Robert Tibshirani et al. 2004. “Least angle regression.” *The Annals of statistics* 32 (2):407–499.

Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. 2010. “Regularization paths for generalized linear models via coordinate descent.” *Journal of statistical software* 33 (1):1.

James, Gareth M, Courtney Paulson, and Paat Rusmevichientong. 2013. “Penalized and constrained regression.” *Unpublished Manuscript, available at http://www-bcf. usc. edu/˜ gareth/research/Research. html* .

Jiang, Wenxin et al. 2007. “Bayesian variable selection for high dimensional generalized linear models: convergence rates of the fitted densities.” *The Annals of Statistics* 35 (4):1487–1511.

Lockhart, Richard, Jonathan Taylor, Ryan J Tibshirani, and Robert Tibshirani. 2014. “A significance test for the lasso.” *Annals of statistics* 42 (2):413.

O’sullivan, Finbarr, Brian S Yandell, and William J Raynor Jr. 1986. “Automatic smoothing of regression functions in generalized linear models.” *Journal of the American Statistical Association* 81 (393):96–103.

Park, Mee Young and Trevor Hastie. 2007. “L1-regularization path algorithm for generalized linear models.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69 (4):659–677.

Shalev-Shwartz, Shai and Ambuj Tewari. 2011. “Stochastic methods for l1-regularized loss minimization.” *Journal of Machine Learning Research* 12 (Jun):1865–1892.

Sørensen, Øystein, Kristoffer Herland Hellton, Arnoldo Frigessi, and Magne Thoresen. 2018. "Covariate selection in high-dimensional generalized linear models with measurement error." *Journal of Computational and Graphical Statistics* (just-accepted).

Tibshirani, Robert. 1996. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* :267–288.

Valdora, Marina, Claudio Agostinelli, and Victor J Yohai. 2017. "Robust Estimation in High Dimensional Generalized Linear Models." *arXiv preprint arXiv:1709.10261* .

Van de Geer, Sara, Peter Bühlmann, Yaacov Ritov, Ruben Dezeure et al. 2014. "On asymptotically optimal confidence regions and tests for high-dimensional models." *The Annals of Statistics* 42 (3):1166–1202.

Van de Geer, Sara A et al. 2008. "High-dimensional generalized linear models and the lasso." *The Annals of Statistics* 36 (2):614–645.

Xu, Jason, Eric Chi, and Kenneth Lange. 2017. "Generalized Linear Model Regression under Distance-to-set Penalties." In *Advances in Neural Information Processing Systems*. 1385–1395.