The Lasso: Variable selection, prediction and estimation.

Rajen Shah

14th March 2012

High-dimensional statistics deals with models in which the number of parameters may greatly exceed the number of observations — an increasingly common situation across many scientific disciplines. The Lasso (Tibshirani, 1996) estimator has been the cornerstone of much of the development in this area. Of all the methods in high-dimensional statistics, the Lasso has the largest body of theoretical work supporting it, and in this note we aim to give a flavour of the sorts of results one can obtain.

1 Introduction

We consider the setting where we have observed data $(y_1, x_1), \ldots, (y_n, x_n)$ with each y_i a realisation of a scalar random variable Y_i, and each $x_i = (x_{i1}, \ldots, x_{ip})^T$ a p-vector of explanatory variables. Let X be a matrix whose i^{th} row is given by x_i^T. Without loss of generality, we shall require that the columns of X are centred. We assume that

$$Y_i = \mu + (X\beta)_i + \epsilon_i,$$

where each ϵ_i is i.i.d. $N(0, \sigma^2)$.

In the classical linear model, we would assume X has full column rank, and so $p < n$. However, here we consider the high-dimensional setting where we may even have $p \gg n$. In this situation, the standard least squares estimator of β simply does not work as $X^T X$ is singular. In fact, in this case the parameter β is not even identifiable. To make progress, we shall assume that the true model is sparse. Letting $S = \{j : \beta_j \neq 0\}$, this means that $|S| =: s \ll n$.

In view of this, it may seem sensible to estimate β by $\hat{\beta}^{\text{BS}}(\lambda)$, where

$$(\hat{\mu}, \hat{\beta}^{\text{BS}}(\lambda)) = \underset{m,b}{\arg\min}\{\tfrac{1}{2n} \|Y - m - Xb\|^2 + \lambda |\{j : b_j \neq 0\}|\}.$$

The tuning parameter λ controls the sparsity of the estimate, with large values of λ resulting in estimates with many components set to 0. Unfortunately, this optimisation problem is NP hard, and to the best of our knowledge, it is computationally intractable for $p > 50$.

The Lasso (Tibshirani, 1996) solves the related problem:

$$(\hat{\mu}, \hat{\beta}(\lambda)) = \arg\min_{m,b} \{ \tfrac{1}{2n} \|Y - m - Xb\|^2 + \lambda \|b\|_1 \}. \tag{1.1}$$

The non-differentiability of the ℓ_1 norm at 0 ensures that the resulting estimator is sparse, and its convexity makes the overall optimisation problem convex. There exist very efficient algorithms for solving this problem, even when $p > 10^5$ (see for example the `R` package `glmnet` of Friedman *et al.*).

2 Theoretical properties

2.1 Variable selection

In this section we give some necessary and sufficient conditions for the Lasso estimator to correctly estimate the sign of β. We do this for the noiseless case, where

$$y = \mu + X\beta.$$

The case with noise is similar. For convenience we define $N = \{1, \ldots, p\} \setminus S$, and for a set of variables J, we let X_J denote the matrix formed from the columns of X indexed by J. We shall assume that X_S has full column rank.

Theorem 1. *Let $\lambda > 0$, and*

$$\theta = X_N^T X_S (X_S^T X_S)^{-1} \operatorname{sgn}(\beta_S).$$

If $\|\theta\|_\infty \leq 1$, and for $j \in S$

$$|\beta_j| > \lambda |\operatorname{sgn}(\beta_S)^T \{ (\tfrac{1}{n} X_S^T X_S)^{-1} \}^{(j)}|,$$

then there exists a Lasso solution with $\operatorname{sgn}(\hat{\beta}(\lambda)) = \operatorname{sgn}(\beta)$. As a partial converse, if there exists a Lasso solution with $\operatorname{sgn} \hat{\beta}(\lambda) = \operatorname{sgn} \beta$, then $\|\theta\|_\infty \leq 1$.

Remark 1. We can interpret $\|\theta\|_\infty$ as the maximum in absolute value over $j \in N$ of the dot product of $\operatorname{sgn} \beta_S$ and the coefficient vector obtain by regressing $X^{(j)}$ on X_S. That is

$$\|\theta\|_\infty = \max_{j \in N} |\operatorname{sgn}(\beta_S)^T (X_S^T X_S)^{-1} X_S^T X^{(j)}|.$$

The condition $\|\theta\|_\infty \leq 1$ is known as (a form of) the irrepresentable condition in the literature.

Proof. By considering subgradients or simply directional derivatives, we have that the Lasso

estimator satisfies

$$\tfrac{1}{n}X^T\{X(\beta-\hat{\beta})+(\mu-\hat{\mu})\mathbf{1}\}=\tfrac{1}{n}X^TX(\beta-\hat{\beta})=\lambda\tau,$$

where $\|\tau\|_\infty \leq 1$, and $\tau_j = \operatorname{sgn}(\hat{\beta}_j)$ for j such that $\hat{\beta}_j \neq 0$ (and we have suppressed the dependence of $\hat{\beta}$ on λ). These are known as the KKT conditions for the Lasso (in the noiseless case) in the literature. We expand this equation into

$$\frac{1}{n}\begin{pmatrix} X_S^TX_S & X_S^TX_N \\ X_N^TX_S & X_N^TX_N \end{pmatrix}\begin{pmatrix} \beta_S-\hat{\beta}_S \\ -\hat{\beta}_N \end{pmatrix}=\lambda\begin{pmatrix}\tau_S\\ \tau_N\end{pmatrix}. \tag{2.1}$$

We prove the converse first. Suppose $\operatorname{sgn}(\beta)=\operatorname{sgn}(\hat{\beta})$ (so $\hat{\beta}_N=0$). Then since X_S has full column rank, the top block of (2.1) can be re-written as

$$\beta_S-\hat{\beta}_S=\lambda(\tfrac{1}{n}X_S^TX_S)^{-1}\tau_S. \tag{2.2}$$

We can substitute this into the second block of equations of (2.1) to get

$$\tfrac{1}{n}X_N^TX_S(\beta_S-\hat{\beta}_S)=\lambda X_N^TX_S(X_S^TX_S)^{-1}\tau_S=\lambda\tau_N. \tag{2.3}$$

But if $\operatorname{sgn}(\beta_S)=\operatorname{sgn}(\hat{\beta}_S)$ then $\tau_S=\operatorname{sgn}(\beta_S)$. Thus observing that $\|\tau_N\|_\infty \leq 1$ completes the proof of the converse.

Now to the positive statement. We claim that taking

$$\begin{aligned}(\hat{\beta}_S,\hat{\beta}_N)&=(\beta_S-\lambda(\tfrac{1}{n}X_S^TX_S)^{-1}\operatorname{sgn}(\beta_S),\,0)\\ (\tau_S,\tau_N)&=(\operatorname{sgn}(\beta_S),\,X_N^TX_S(X_S^TX_S)^{-1}\operatorname{sgn}(\beta_S))\end{aligned}$$

satisfies the KKT conditions (2.1) or equivalently, since we are taking $\hat{\beta}_N=0$, equations (2.2) and (2.3). Indeed, our assumption that

$$|\beta_j|>\lambda|\operatorname{sgn}(\beta_S)^T\{(\tfrac{1}{n}X_S^TX_S)^{-1}\}^{(j)}|$$

for $j\in S$ ensures that $\operatorname{sgn}(\hat{\beta}_S)=\operatorname{sgn}(\beta_S)$, so the condition for τ_S is satisfied. Then checking (2.2) and (2.3) is easy. □

2.2 Prediction and estimation

In order to understand the sorts of results we should expect for the prediction and estimation properties of the Lasso, let us first imagine that S is known to us. If we knew S, we could simply apply the least squares estimator where we take the design matrix as X_S. If we let $\beta^*=(X_S^TX_S)^{-1}X_S^TY$ and write $\Omega_{jj}=(\frac{1}{n}X_S^TX_S)^{-1}_{jj}$, we have

$$\mathbb{E}\left(\tfrac{1}{n}\left\|X(\beta^*-\beta)\right\|^2\right)=\frac{\sigma^2 s}{n} \tag{2.4}$$

$$\mathbb{E}\left\|\beta^*-\beta\right\|_1=\frac{s\sigma}{\sqrt{n}}\times\frac{1}{s}\sum_{j\in S}\sqrt{\frac{2\Omega_{jj}}{\pi}}. \tag{2.5}$$

We shall show that the Lasso achieves these rates for prediction and estimation up to a $\log(p)$ factor, and subject to some conditions on the design matrix which we now discuss. We shall require that there exists a $\phi>0$ such that for all b satisfying $\left\|b_N\right\|_1\leq 4\left\|b_S\right\|_1$, it holds that

$$\left\|b_S\right\|_1^2\leq\frac{s\left\|Xb\right\|^2}{n\phi^2}. \tag{2.6}$$

This type of condition is known as a compatibility condition. We note that the constant 4 appearing in the definition is quite arbitrary and could be replaced by any constant greater than 1. Furthermore, we will require that the columns of X are scaled such that $\left\|X^{(j)}\right\|^2=n$ for $j=1,\ldots,p$.

Theorem 2. *Let*

$$\lambda=A\sigma\sqrt{\frac{\log(p)}{n}}.$$

Then with probability at least $1-(p^{1-A^2/8}+p^{-5sA^2/(2\phi^2)})$,

$$\tfrac{1}{n}\left\|X(\hat{\beta}-\beta)\right\|^2+\lambda\left\|\hat{\beta}-\beta\right\|_1\leq 25\lambda^2 s/\phi^2=\frac{25A^2}{\phi^2}\frac{\sigma^2 s\log(p)}{n}.$$

Proof. By the definition of $(\hat{\mu},\hat{\beta})$, we have that

$$\begin{aligned}\tfrac{1}{2n}\left\|\mu\mathbf{1}+X\beta+\epsilon-\hat{\mu}\mathbf{1}-X\hat{\beta}\right\|^2+\lambda\left\|\hat{\beta}\right\|_1&\leq\tfrac{1}{2n}\left\|\epsilon\right\|^2+\lambda\left\|\beta\right\|_1\\ \tfrac{1}{2n}\left\|X(\hat{\beta}-\beta)\right\|^2+\lambda\left\|\hat{\beta}\right\|_1&\leq\tfrac{1}{n}\epsilon^T X(\hat{\beta}-\beta)+\tfrac{1}{2}\bar{\epsilon}^2+\lambda\left\|\beta\right\|_1.\end{aligned} \tag{2.7}$$

Define the following events.

$$\begin{aligned}\Omega_1&=\left\{\tfrac{1}{n}\left\|X^T\epsilon\right\|_\infty\leq\lambda/2\right\}\\ \Omega_2&=\{\bar{\epsilon}^2\leq 5\lambda^2 s/\phi^2\}.\end{aligned}$$

It is straightforward to show that $\mathbb{P}(\Omega_1\cap\Omega_2)\geq 1-(p^{1-A^2/8}+p^{-5sA^2/(2\phi^2)})$. In all of the following, we work on $\Omega_1\cap\Omega_2$. Since

$$\tfrac{1}{n}|\epsilon^T X(\hat{\beta}-\beta)|\leq\tfrac{1}{n}\left\|X^T\epsilon\right\|_\infty\left\|\hat{\beta}-\beta\right\|_1\leq\frac{\lambda}{2}\left\|\hat{\beta}-\beta\right\|_1,$$

we have from (2.7) that

$$\begin{aligned}
\tfrac{1}{n}\left\|X(\hat\beta-\beta)\right\|^2+2\lambda\left\|\hat\beta\right\|_1 &\le \lambda\left\|\hat\beta-\beta\right\|_1+2\lambda\left\|\beta\right\|_1+5\lambda^2 s/\phi^2\\
\tfrac{1}{n}\left\|X(\hat\beta-\beta)\right\|^2+2\lambda\left\|\hat\beta_N\right\|_1+2\lambda\left\|\hat\beta_S\right\|_1 &\le \lambda\left\|\hat\beta_S-\beta_S\right\|_1+\lambda\left\|\hat\beta_N\right\|_1+2\lambda\left\|\beta_S\right\|_1+5\lambda^2 s/\phi^2\\
\tfrac{1}{n}\left\|X(\hat\beta-\beta)\right\|^2+\lambda\left\|\hat\beta_N\right\|_1 &\le 3\lambda\left\|\hat\beta_S-\beta_S\right\|_1+5\lambda^2 s/\phi^2. \qquad (2.8)
\end{aligned}$$

First suppose that $\left\|\hat\beta_S-\beta_S\right\|_1 \le 5\lambda s/\phi^2$. Then from (2.8),

$$\tfrac{1}{n}\left\|X(\hat\beta-\beta)\right\|^2+\lambda\left\|\hat\beta-\beta\right\|_1=\tfrac{1}{n}\left\|X(\hat\beta-\beta)\right\|^2+\lambda\left\|\hat\beta_N\right\|_1+\lambda\left\|\hat\beta_S-\beta_S\right\|_1\le 25\lambda^2 s/\phi^2.$$

Now suppose instead that $\left\|\hat\beta_S-\beta_S\right\|_1 > 5\lambda s/\phi^2$. Then from (2.8) we have,

$$\tfrac{1}{n}\left\|X(\hat\beta-\beta)\right\|^2+\lambda\left\|\hat\beta_N\right\|_1\le 4\lambda\left\|\hat\beta_S-\beta_S\right\|_1, \qquad (2.9)$$

so in particular $\left\|\hat\beta_N-\beta_N\right\|_1\le 4\left\|\hat\beta_S-\beta_S\right\|_1$. Thus

$$\begin{aligned}
\tfrac{1}{n}\left\|X(\hat\beta-\beta)\right\|^2+\lambda\left\|\hat\beta-\beta\right\|_1&=\tfrac{1}{n}\left\|X(\hat\beta-\beta)\right\|^2+\lambda\left\|\hat\beta_N\right\|_1+\lambda\left\|\hat\beta_S-\beta_S\right\|_1\\
&\le 5\lambda\left\|\hat\beta_S-\beta_S\right\|_1\\
&\le \frac{5\lambda\sqrt{s}\left\|X(\hat\beta-\beta)\right\|}{\phi\sqrt{n}}\le 25\lambda^2 s/\phi^2,
\end{aligned}$$

where in the last line we made use of the compatibility condition (2.6). □

References

[1] Bühlmann, P. and van de Geer, S. (2011) On the Conditions Used to Prove Oracle Results for the Lasso. *J. Electron. Stat.*, **3**, 1360–1392.

[2] Bühlmann, P. and van de Geer, S. (2011) Statistics for High-Dimensional Data: Methods, Theory and Algorithms. Springer, Springer Series in Statistics.

[3] Friedman, J., Hastie, T. and Tibshirani, R. (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, **33**, 1–22.

[4] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *J. Roy. Statist. Soc., Ser. B*, **58**, 267–288.

[5] Wainwright, M. (2009) Sharp Thresholds for High-Dimensional and Noisy Sparsity Re-

covery Using ℓ_1-Constrained Quadratic Programming (Lasso). *IEEE Trans. Inf. Theory*, **55**, 2183–2202.

[6] Zhao, P. and Yu, B. (2006) On Model Selection Sonsistency of the Lasso *J. Machine Learning Research*, **7**, 2541–2563.