High-dimensional data and the Lasso

Rajen D. Shah
Statistical Laboratory, University of Cambridge

December 18, 2013

How would you try to solve a linear system of equations with more unknowns than equations? Of course, there are infinitely many solutions, and yet this is the sort of the problem statisticians face with many modern datasets, arising in genetics, imaging, finance and many other fields. What's worse, our equations are often corrupted by noisy measurements! In this article we will introduce a statistical method that has been at the centre of the huge amount of research that has gone into solving these problems. We'll begin by reviewing the classical version of the problems, before moving on to the more modern setting hinted at above.

Regression analysis

Imagine data are available in the form of observations $(Y_i, x_i) \in \mathbb{R} \times \mathbb{R}^p$, $i = 1, \dots, n$, and the aim is to infer a simple *regression function* relating the average value of a *response*, Y_i, and a collection of *predictors* or *variables*, x_i. This is an example of regression analysis, one of the most important tasks in Statistics.

Often, we may assume that the unknown regression function is linear in the predictors, giving the following mathematical formulation of the problem:

$$Y = X\beta + \varepsilon, \tag{0.1}$$

where $Y \in \mathbb{R}^n$ is the vector of responses; $X \in \mathbb{R}^{n \times p}$ is the predictor matrix with i^{th} row x_i^T; $\varepsilon \in \mathbb{R}^n$ represents random error; and $\beta \in \mathbb{R}^p$ is the unknown vector of coefficients that determines the regression function and is to be estimated using the data.

A traditional application of the model (0.1) may have the responses as blood pressure measurements for $n = 100$ patients and the predictors could include height, weight, age and daily calorie intake, for example. In this case, one might estimate β by ordinary least squares (OLS), a technique dating back to Gauss (1795). This yields an estimator $\hat{\beta}^{\text{OLS}}$ with

$$\hat{\beta}^{\text{OLS}} := \underset{b \in \mathbb{R}^p}{\arg\min} \, \|Y - Xb\|_2^2 = (X^T X)^{-1} X^T Y, \tag{0.2}$$

provided X has full column rank. Here $\|\cdot\|_2$ denotes the Euclidean norm. We can analyse the quality of the estimate of the regression function by calculating its *mean-squared prediction error* (MSPE). Under the assumptions that (i) $\mathbb{E}(\varepsilon_i) = 0$ and (ii) $\text{Cov}(\varepsilon_i, \varepsilon_j) = \sigma^2 \mathbb{1}_{\{i=j\}}$, it holds that

$$\text{MSPE}(\hat{\beta}^{\text{OLS}}) := \mathbb{E}\{\tfrac{1}{n}\|X(\beta - \hat{\beta}^{\text{OLS}})\|_2^2\} = \frac{p}{n}\sigma^2. \tag{0.3}$$

We see that provided p/n is small, the MSPE is small. When this is true, and under the assumptions given above, OLS is a very reasonable choice of estimator. Indeed, the Gauss–Markov theorem shows that it has the minimum MSPE among all linear unbiased estimators of the regression function, i.e. among all estimators $\hat{f} := AY$ of $X\beta$, for some $n \times n$ matrix A such that $\mathbb{E}(AY) = X\beta$.

High-dimensional data

One might think that OLS essentially solves the problem of linear regression, at least under assumptions (i) and (ii). However, the field of Statistics must constantly adapt and innovate to develop methods that accommodate the data it is tasked with to study, and today, much of that data is *high-dimensional*: p is very large and often greatly exceeds n. Where in the past only a few carefully chosen variables were measured for each observation, nowadays any variable that might conceivably have an effect on the response tends to be recorded, leading to ratios of $p/n > 1000$ being common in some areas. Out of these many variables, it may be only a few that are really important for predicting the response, but of course which these are would not be known in advance. In the context of our model (0.1), this would translate β being sparse, i.e. many of its components being exactly 0.

How are we to proceed with analysis given such datasets? Clearly OLS is unhelpful when $p \geq n$ if X has full column rank, as then predictions are simply the original responses themselves. Ideally, we would like a sparse estimator that first attempted to detect the relevant variables, and then estimated just *their* coefficients. Even when β is not truly sparse, it can make sense to produce a sparse estimate for it. A final estimate with only a few non-zero coefficients may be much easier to interpret, and given a new observation $x \in \mathbb{R}^p$, computing the estimate of the regression function at x would be much faster.

In view of this, one might consider trying to estimate β by $\hat{\beta}^{\mathrm{BS}}$ (Best Subsets) defined as the minimiser of a penalised least squares objective:

$$\hat{\beta}^{\mathrm{BS}} := \operatorname*{arg\,min}_{b \in \mathbb{R}^p} \{\tfrac{1}{n}\|Y - Xb\|_2^2 + \lambda\|b\|_0\}, \tag{0.4}$$

where $\|b\|_0 := \sum_{k=1}^p \mathbb{1}_{\{b_k \neq 0\}}$. Large values of the *regularisation parameter*, λ, will cause $\hat{\beta}$ to have very few non-zero components, and lower values will produce less sparse models. Several methods are available for choosing λ; we will not go into the details here.

A major problem with the estimator $\hat{\beta}^{\mathrm{BS}}$ is that the optimisation in (0.4) is in general computationally infeasible as one would essentially need to evaluate the objective at all 2^p possible subsets of variables in order to guarantee finding the optimiser. As p runs into the hundreds, the number of computations required to perform the optimisation quickly surpasses the number of atoms in the observable universe!

The Lasso

The key property of the objective in (0.4) which makes the optimisation intractable is that it is non-convex. Convex problems are much easier to solve, one reason for this being that any local optimum is also a global optimum. It is thus sensible to consider convex approximations to the objective in (0.4). One such approximation results from replacing $\|b\|_0$ with $\|b\|_1 := \sum_{k=1}^p |b_k|$, so our estimator $\hat{\beta}$ is given by

$$\hat{\beta} := \operatorname*{arg\,min}_{b \in \mathbb{R}^p} \{\tfrac{1}{n}\|Y - Xb\|_2^2 + \lambda\|b\|_1\}. \tag{0.5}$$

This is the Lasso (Least Absolute Shrinkage and Selection Operator) estimator (Tibshirani, 1996): one of the most popular methods in high-dimensional data analysis. Applications of the Lasso and related methods range from identifying which of our thousands of genes are related to particular diseases, to the click-through rate prediction task that helps optimise web advertising for search engines.

The optimisation in (0.5) can be solved even for very large problems where p is hundreds of thousands. Yet importantly, sparsity of the estimator is retained. This is essentially because the set $\{b : \|b\|_1 \leq r\}$ for $r > 0$ has corners; see Bühlmann and van de Geer (2011) for details.

Sparsity and computational feasibility are attractive properties, but what really makes the Lasso appealing is its remarkable performance as both a selector of important variables, and as a prediction engine. This is perhaps surprising given that the Lasso optimisation only approximates (0.4). A vast amount of work has gone into trying to understand why the Lasso works so well, and also into developing improvements and adapting the method to suit many other problems. We will finish with a theorem that forms part of that work: we show that provided $\|\beta\|_1$ is small, as would be the case when β is sparse, p can be almost exponential in n, and the Lasso can still estimate the regression function well. This is a really rather striking result, considering that OLS requires p to be much smaller than n. In fact much more is true; the interested reader is directed to Bühlmann and van de Geer (2011).

Prediction error of the Lasso

Let $\hat{\beta}$ be the Lasso estimator (0.5) with $\lambda = \sigma\sqrt{2(\log p + t^2)/n}$. Assume that the errors ε_i are independent and normally distributed. Then with probability at least $1 - e^{-t^2}$, we have

$$\tfrac{1}{n}\|X(\beta - \hat{\beta})\|_2^2 \le 2\sigma\|\beta\|_1\sqrt{\frac{2(\log p + t^2)}{n}}.$$

Proof. By the definition of $\hat{\beta}$ we have

$$\tfrac{1}{n}\|Y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \le \tfrac{1}{n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1.$$

Recalling that $Y = X\beta + \varepsilon$ and rearranging, we get

$$\tfrac{1}{n}\|X(\beta - \hat{\beta})\|_2^2 \le \tfrac{1}{n}2\varepsilon^T X(\hat{\beta} - \beta) + \lambda\|\beta\|_1 - \lambda\|\hat{\beta}\|_1. \tag{0.6}$$

Denote the k^{th} column of X by $X_k \in \mathbb{R}^n$. Now

$$\tfrac{1}{n}\varepsilon^T X(\hat{\beta} - \beta) \le \tfrac{1}{n}\|\beta - \hat{\beta}\|_1 \max_{1\le k\le p} |X_k^T\varepsilon|,$$

and as $\varepsilon \sim N_n(0, \sigma^2 I)$ and $\|X_k\|_2^2 = n$, we have $X_k^T\varepsilon/n \sim N(0, \sigma^2/n)$. Now let $Z \sim N(0,1)$ so $\sigma Z/\sqrt{n}$ has the same distribution as $X_k^T\varepsilon/n$ for each k. We argue

$$\mathbb{P}\Big(\max_{1\le k\le p} |X_k^T\varepsilon|/n \ge \lambda\Big) = \mathbb{P}\Big(\bigcup_{k=1}^{p}\{|X_k^T\varepsilon|/n \ge \lambda\}\Big) \le \sum_{k=1}^{p}\mathbb{P}(|X_k^T\varepsilon|/n \ge \lambda) = 2p\mathbb{P}\{Z \ge \lambda\sqrt{n}/\sigma\}.$$

A standard tail bound for normal random variables (proved below) gives us that for $\zeta \ge 0$, $1 - \Phi(\zeta) \le e^{-\zeta^2/2}/2$. Applying this to the above, we see that the final term in the last display is bounded above by $p\exp\{-n\lambda^2/(2\sigma^2)\} = e^{-t^2}$. Now working on the event $\{\max_{1\le k\le p}|X_k^T\varepsilon|/n < \lambda\}$, which we have shown has probability at least $1 - e^{-t^2}$, we have from (0.6) that.

$$\tfrac{1}{n}\|X(\beta - \hat{\beta})\|_2^2 \le \lambda\|\beta - \hat{\beta}\|_1 + \lambda\|\beta\|_1 - \lambda\|\hat{\beta}\|_1.$$

Noting that $\|\beta - \hat{\beta}\|_1 - \|\hat{\beta}\|_1 \le \|\beta\|_1$ by the triangle inequality completes the proof. □

Standard normal tail bound

Let $Z \sim N(0,1)$. Then $1 - \Phi(\zeta) := \mathbb{P}(Z \ge \zeta) \le \frac{1}{2}e^{-\zeta^2/2}$ when $\zeta \ge 0$.

Proof. Let $f(\zeta) := 1 - \Phi(\zeta) - \frac{1}{2}e^{-\zeta^2/2}$. Now

$$\mathbb{P}(Z \ge \zeta) = \frac{1}{\sqrt{2\pi\sigma^2}}\int_\zeta^\infty e^{-z^2/2}dz \le \frac{1}{\sqrt{2\pi\sigma^2}}\int_\zeta^\infty \frac{z}{\zeta}e^{-z^2/2}dz = \frac{1}{\zeta\sqrt{2\pi\sigma^2}}e^{-\zeta^2/2}.$$

Thus if $\zeta \ge \sqrt{2/(\pi\sigma^2)}$, then $f(\zeta) \le 0$. Also, $f(0) = 0$. Finally observe that

$$f'(\zeta) = -\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\zeta^2/2} + \frac{\zeta}{2}e^{-\zeta^2/2},$$

so $f'(\zeta) \le 0$ for $\zeta \le \sqrt{2/(\pi\sigma^2)}$. Conclude that $f(\zeta) \le 0$ when $\zeta \ge 0$. □

References

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

Tibshirani, R. (1996) Regression shrinkage and selection via the Lasso *J. Roy. Statist. Soc., Ser. B*, **58**, 267–288.