Regression shrinkage and selection via the Lasso: a retrospective

ROBERT TIBSHIRANI

Department of Health Research and Policy
and Department of Statistics
Stanford University

September 21, 2010

Presented at the RSS annual meeting 2010, Brighton, U.K.

The work discussed here represents collaborations with many people, especially Bradley Efron, Jerome Friedman, Trevor Hastie, Holger Hoefling, Iain Johnstone, Ryan Tibshirani and Daniela Witten

I would like to thank the research section of the Royal Statistical Society for inviting me to present this retrospective paper. In this paper I give a brief review of the basic idea, some history, and then discuss some developments since the original paper.

1 The lasso

Given a linear regression with predictors $x_{ij}$ and response values $y_i$ for $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, p$, the lasso solves the $\ell_1$-penalized regression

$$
\sum_{i=1}^{N} (y_i - \sum_{j} x_{ij} \hat{\beta}_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|.
$$

This is equivalent to minimizing the sum of squares with constraint $\sum |\beta_j| \leq s$. It is similar to ridge regression, which has constraint $\sum_j \beta_j^2 \leq t$. Because of the form of the $\ell_1$ penalty, the lasso does variable selection and shrinkage; while ridge regression, in contrast, only shrinks. If we consider a more general penalty of the form $[\sum_{j=1}^{p} \beta_j^q]^{1/q}$, then the lasso uses $q = 1$ and ridge regression has $q = 2$. Subset selection emerges as $q \to 0$, and the lasso uses the smallest value of $q$ (i.e. closest to subset selection) that yields a convex problem. Convexity is very attractive for computational purposes.

2 History of the idea

The lasso is just regression with an $\ell_1$ norm penalty, and $\ell_1$ norms have been around for long time! My most direct influence was Leo Breiman’s garotte (Breiman 1995). His idea was to minimize

$$
\sum_{i=1}^{N} (y_i - \sum_{j} c_j x_{ij} \hat{\beta}_j)^2 + c_j \sum_{j=1}^{p} \beta_j^2 \leq t
$$

where $\hat{\beta}_j$ are usual least squares estimates. This is undefined when $p > N$ (not a hot topic in 1995!) so I just combined the two stages into one (as a Canadian I also wanted a gentler name). Other related work around the same time: Frank & Friedman (1993) discussed bridge regression using a penalty $\lambda \sum |\beta_j|^\gamma$, with both $\lambda$ and $\gamma$ estimated from the data. Chen et al. (1998) proposed basis pursuit, which uses an $\ell_1$ norm penalty.
<table>
<thead>
<tr>
<th>Method</th>
<th>Authors</th>
<th>Detail</th>
</tr>
</thead>
<tbody>
<tr>
<td>grouped lasso</td>
<td>Yuan &amp; Lin (2007a)</td>
<td>$\sum_j</td>
</tr>
<tr>
<td>elastic net</td>
<td>Zou &amp; Hastie (2005)</td>
<td>$\lambda_1 \sum_j</td>
</tr>
<tr>
<td>fused lasso</td>
<td>Tibshirani et al. (2005)</td>
<td>$\lambda \sum_j</td>
</tr>
<tr>
<td>adaptive lasso</td>
<td>Zou (2006)</td>
<td>$\lambda_1 \sum_j w_j</td>
</tr>
<tr>
<td>graphical lasso</td>
<td>Yuan &amp; Lin (2007b), Friedman et al. (2007)</td>
<td>loglik $+ \lambda</td>
</tr>
<tr>
<td>Dantzig selector</td>
<td>Candès &amp; Tao (2007)</td>
<td>$\min X^T(y - X\beta)</td>
</tr>
<tr>
<td>near-isotonic reg.</td>
<td>Tibshirani et al. (2010)</td>
<td>$\sum_j (\beta_j - \beta_{j+1})^+$</td>
</tr>
<tr>
<td>matrix completion</td>
<td>Candès &amp; Tao (2009), Mazumder et al. (2010)</td>
<td>$</td>
</tr>
<tr>
<td>compressive sensing</td>
<td>Donoho (2004), Candes (2006)</td>
<td>$\min</td>
</tr>
<tr>
<td>multivariate methods</td>
<td>Joliffe et al. (2003), Witten et al. (2009)</td>
<td>sparse PCA, LDA, CCA</td>
</tr>
</tbody>
</table>

Table 1: A sampling of generalizations of the lasso

penalty in a signal processing context. Surely there are many other references that I am unaware of. After publication, the paper did not get much attention until years later. Why? My guesses are a) the computation in 1996 was slow compared to today, b) the algorithms for the lasso were black boxes, and not statistically motivated (until LARS in 2002), c) the statistical and numerical advantages of sparsity were not immediately appreciated (by me or the community), d) large data problems (in $N, p$ or both) were rare and e) community did not have the R language for fast, easy sharing of new software tools

3 Computational advances

The original lasso paper used an off-the-shelf quadratic program solver. This doesn’t scale well and is not transparent. The LARS algorithm (Efron et al. 2002) gives an efficient way to solve the lasso, and connects the lasso to forward stagewise regression. The same algorithm is contained in the homotopy approach of Osborne et al. (2000). Coordinate descent algorithms are extremely simple and fast, and exploit the assumed sparsity of the model to great advantage. References include Fu (1998), Friedman et al. (2007), Wu & Lange (2008) Genkin et al. (2007) and Friedman et al. (2010). We were made aware of its real potential in the PhD thesis of Anita van der Kooij (Leiden) working with Jacqueline Meulman. The glmnet R language package implements the coordinate descent method for many popular models.

4 Some generalizations and variants of the lasso

There has been much work in recent years, applying and generalizing the lasso and $\ell_1$-like penalties to a variety of problems. Table 1 give a partial list. There has also been a great deal of deep and interesting work on the mathematical aspects of the lasso, examining its ability to produce a model with minimal prediction error, and also to recover the true underlying (sparse) model. Important contributors here include Buhlmann, Candes, Donoho, Johnstone, Meinshausen, Wainwright, and Yu. I do not have the qualifications or the space to properly summarize this work, but hope that Prof. Buhlmann will cover this aspect in his discussion.

Lasso methods can also shed light on more traditional techniques. The LARS algorithm, mentioned above, brings new understanding to forward stepwise selection methods. Another example is the graphical lasso for fitting a sparse Gaussian graph, based on the Gaussian loglikelihood plus $\lambda ||\Sigma^{-1}||_1$, which is an $\ell_1$ penalty applied to the inverse covariance matrix. Since a missing edge in the graph corresponds to a zero element of $\Sigma^{-1}$, this gives a powerful method for graph selection— determining which edges to include. As a bonus, a special case of the graphical lasso gives a new simple method for fitting a graph with pre-specified edges (structural zeroes in $\Sigma^{-1}$). The details are given in Chapter 17 of Hastie et al. (2008).

Another recent example is nearly-isotonic regression (Tibshirani et al. 2010). Given a data sequence $y_1, y_2, \ldots y_N$ isotonic regression solves

\[
\text{minimize} \quad \sum (y_i - \hat{y}_i)^2 \quad \text{subject to} \quad \hat{y}_1 \leq \hat{y}_2, \ldots
\]
This assumes a monotone non-decreasing approximation, with an analogous definition for the monotone non-increasing case. The solution can be computed via the well-known Pool Adjacent Violators (PAVA) algorithm (e.g. Barlow et al. (1972). In nearly-isotonic regression we solve

$$\text{minimize } \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_i)^2 + \lambda \sum_{i=1}^{n-1} (\beta_i - \beta_{i+1})_+,$$

with $x_+$ indicating the positive part, $x_+ = x \cdot 1(x > 0)$. This is a convex problem; with $\hat{\beta}_i = y_i$ at $\lambda = 0$ and culminating in the usual isotonic regression as $\lambda \to \infty$. Along the way it gives nearly monotone approximations. Note that $(\beta_i - \beta_{i+1})_+$ is “half” an of $\ell_1$ penalty on differences, penalizing dips but not increases in the sequence. This procedure allows one to assess the assumption of monotonicity by comparing nearly-monotone approximations to the best monotone approximation. Tibshirani et al. (2010) provide a simple algorithm that computes the entire path of solutions, a kind of modified version of the PAVA procedure. They also show that the number of degrees of freedom is the number of unique values of $\hat{y}_i$ in the solution, using results from Tibshirani & Taylor (2010).

5 Discussion

Lasso ($\ell_1$) penalties are useful for fitting a wide variety of models. Newly-developed computational algorithms allow application of these models to large datasets, exploiting sparsity for both statistical and computation gains. Interesting work on the lasso is being carried out in many fields, including Statistics, Engineering, Mathematics and Computer Science. I conclude with a challenge for statisticians. This is a fun area to work in, but we should not invent new models and algorithms just for the sake of it. We should focus on developing tools and understanding their properties, to help us and our collaborators to solve important scientific problems.
References


URL: www.acm.caltech.edu/ emmanuel/papers/CompressiveSampling.pdf


