Regression shrinkage and selection via the lasso *

ROBERT TIBSHIRANI †
Department of Statistics
and
Division of Biostatistics
Stanford University

Abstract

We propose a new method for estimation in linear models. The "lasso" minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. Because of the nature of this constraint it tends to produce some coefficients that are exactly zero and hence gives interpretable models. Our simulation studies suggest that the lasso enjoys some of the favourable properties of both subset selection and ridge regression. It produces interpretable models like subset selection and exhibits the stability of ridge regression. There is also an interesting relationship with recent work in adaptive function estimation by Donoho and Johnstone. The lasso idea is quite general and can be applied in a variety of statistical models: extensions to generalized regression models and tree-based models are briefly described.

Keywords: regression, subset selection, shrinkage, quadratic programming.

1 Introduction

Consider the usual regression situation: we have data: $(\mathbf{x}^i, y_i)$, $i = 1, 2, \ldots N$ where $\mathbf{x}^i = (x_{i1}, \ldots x_{ip})^T$ and y_i are the regressors and response for the *ith* observation. The ordinary least squares (OLS) estimates are obtained by minimizing the residual squared error. There are two reasons why the data analyst is often not satisfied with the OLS estimates. The first is *prediction accuracy*: the OLS estimates often have low bias but large variance; prediction accuracy can sometimes be improved by shrinking or setting to zero some coefficients. By doing so we sacrifice a little bit of bias to reduce the variance of the predicted

*This is a revision of the technical report of the same title, Dept of Statistics, University of Toronto, January 1994

†On sabbatical leave from Department of Preventive Medicine and Biostatistics, and Department of Statistics, University of Toronto

values and hence may improve the overall prediction accuracy. The second reason is *interpretation*. With a large number of predictors, we often would like to determine a smaller subset that exhibit the strongest effects.

The two standard techniques for improving the OLS estimates, subset selection and ridge regression, both have drawbacks. Subset selection provides interpretable models but can be extremely variable because it is a discrete process—regressors are either retained or dropped from the model. Small changes in the data can result in very different models being selected and this can reduce its prediction accuracy. Ridge regression is a continuous process that shrinks coefficients and hence is more stable: however, it doesn't set any coefficients to zero and hence doesn't give an easily interpretable model.

We propose a new technique, called the *lasso*, for "Least Absolute Shrinkage and Selection Operator". It shrinks some coefficients and sets others to zero, and hence tries to retain the good features of both subset selection and ridge regression.

In section 2 we define the lasso, and look at some special cases. A real data example is given in section 3, while in section 4 we discuss methods for estimation of prediction error and the lasso shrinkage parameter. A Bayes model for the lasso is briefly mentioned in section 5. We describe the lasso algorithm in section 6. Simulation studies are described in section 7. Sections 8 and 9 discuss extensions to generalized regression models and other problems. Some results on soft-thresholding and their relation to the lasso is discussed in section 10, while Section 11 contains a summary and some discussion.

2 The lasso

2.1 Definition

Suppose that we have data $(\mathbf{x}^i, y_i)$, $i = 1, 2, \ldots N$ where $\mathbf{x}^i = (x_{i1}, \ldots x_{ip})^T$ are the predictor variables and y_i are the responses. As in the usual regression setup, we assume that either that the observations are independent or that the y_is are conditionally independent given the x_{ij}s. We assume that the x_{ij} are standardized so that $\sum_i x_{ij}/N = 0, \sum_i x_{ij}^2/N = 1$.

Letting $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots \hat{\beta}_p)^T$, the lasso estimate $(\alpha, \hat{\boldsymbol{\beta}})$ is defined by

$$
\begin{aligned}
(\hat{\alpha}, \hat{\boldsymbol{\beta}}) &= \operatorname{argmin} \sum_{i=1}^{N} (y_i - \alpha - \sum_j \beta_j x_{ij})^2 \\
& \quad \text{subject to } \sum_j |\beta_j| \leq t
\end{aligned} \tag{1}
$$

Here $t \geq 0$ is a tuning parameter. Now for all t, the solution for α is $\hat{\alpha} = \bar{y}$. We can assume without loss of generality that $\bar{y} = 0$ and hence omit α.

Computation of the solution to (1) is a quadratic programming problem with linear inequality constraints. We describe some efficient and stable algorithms for this problem in section 6.

The parameter $t \geq 0$ controls the amount of shrinkage that is applied to the estimates. Let $t_0 = \sum |\hat{\beta}_j^{\circ}|$. Values of $t < t_0$ will cause shrinkage of the solutions towards zero, and some coefficients may be exactly equal to zero. For example, if $t = t_0/2$, the effect will be roughly similar to finding the best subset of size $p/2$. Note that the design matrix need not be of full rank. In section 4 we give a number of data-based methods for estimation of t.

This motivation for the lasso came from an interesting proposal of Breiman (1993). Breiman's *non-negative garotte* minimizes

$$\sum_{i=1}^{N}(y_i - \alpha - \sum_j c_j \hat{\beta}_j^{\circ} x_{ij})^2 \quad \text{subject to} \quad c_j \geq 0, \ \ \sum c_j \leq t \tag{2}$$

The garotte starts with the OLS estimates and shrinks them by non-negative factors whose sum is constrained. In extensive simulation studies, Breiman shows that the garotte has consistently lower prediction error than subset selection and is competitive with ridge regression except when the true model has many small non-zero coefficients.

A drawback of the garotte is that its solution depends on both the sign and magnitude of the OLS estimates. In overfit or highly correlated settings where the OLS estimates behave poorly, the garotte may suffer as a result. In contrast, the lasso avoids explicit use of the OLS estimates.

Frank and Friedman (1993) proposed using a bound on the L^q norm of the parameters, where q is some number ≥ 0; the lasso corresponds to $q = 1$. We discuss this briefly in section 10.

2.2 Orthogonal design case

Insight about the nature of the shrinkage can be gleaned from the orthogonal design case. Let $\mathbf{X}$ be the $n \times p$ design matrix with ijth entry x_{ij}, and suppose $\mathbf{X}^T\mathbf{X} = \mathbf{I}$, the identity matrix.

The solutions to (1) are easily shown to be

$$\hat{\beta}_j = \text{sign}(\hat{\beta}_j^{\circ})(|\hat{\beta}_j^{\circ}| - \gamma)^+ \tag{3}$$

where γ is determined by the condition $\sum |\hat{\beta}_j| = t$. Interestingly, this has exactly the same form as the soft shrinkage proposals of Donoho & Johnstone (1994) and Donoho, Johnstone, Kerkyachairan & Picard (1995), applied to wavelet coefficients in the context of function estimation. The connection between soft shrinkage and a minimum L_1 norm penalty was also pointed out by Donoho, Johnstone, Hoch & Stern (1992) for non-negative parameters in the context of signal or image recovery. We elaborate more on this connection in section 10.

In the orthonormal design case, best subset selection of size k reduces to choosing the k largest coefficients in absolute value and setting the rest to zero. For some choice of λ this is equivalent to setting $\hat{\beta}_j = \hat{\beta}_j^o$ if $|\hat{\beta}_j^o| > \lambda$ and zero otherwise. Ridge regression minimizes $\sum_{i=1}^N (y_i - \sum_j \beta_j x_{ij})^2 + \lambda \sum_j \beta_j^2$ or equivalently, minimizes

$$\sum_{i=1}^N (y_i - \sum_j \beta_j x_{ij})^2 \quad \text{subject to } \sum \beta_j^2 \leq t \tag{4}$$

The ridge solutions are

$$\frac{1}{1+\gamma}\hat{\beta}_j^o$$

where γ depends on λ or t. The garotte estimates are

$$\left(1 - \frac{\gamma}{(\hat{\beta}_j^o)^2}\right)^+ \hat{\beta}_j^o$$

Figure 1 shows the form of these functions. Ridge regression scales the coefficients by a constant factor, while the lasso translates by a constant factor, truncating at zero. The garotte function is very similar to the lasso, with less shrinkage for larger coefficients. As our simulations will show, the differences between the lasso and garotte can be large when the design is not orthogonal.

2.3 Geometry of the lasso

It is clear from Figure 1 why the lasso will often produce coefficients that are exactly zero. Why does this happen in the general (non-orthogonal) setting? And why does it not occur with ridge regression, which uses the constraint $\sum \beta_j^2 \leq t$ rather than $\sum |\beta_j| \leq t$? Figure 2 provides some insight for the case $p = 2$.

The criterion $\sum_{i=1}^N (y_i - \sum_j \beta_j x_{ij})^2$ equals the quadratic function $(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^o)^T \mathbf{X}^T \mathbf{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^o)$ (plus a constant). The elliptical contours of this function are shown by the solid curves in the left panel, there are centered at the OLS estimates; the constraint region is the rotated square indicated by the broken lines. The lasso solution is the first place that the contours touch the square, and this will sometimes occur at a corner, corresponding to a zero coefficient. The picture for ridge regression is shown on the right: there are no corners for the contours to hit and hence zero solutions will rarely result.

An interesting question emerges from this picture: can the signs of the lasso estimates be different from those of the least squares estimates $\hat{\beta}_j^o$? Since the variables are standardized, when $p = 2$ the principal axes of the contours are at $\pm 45^o$ to the coordinate axes, and one can show that the contours must contact the square in the same quadrant that contains $\hat{\boldsymbol{\beta}}^o$. However when $p > 2$ and

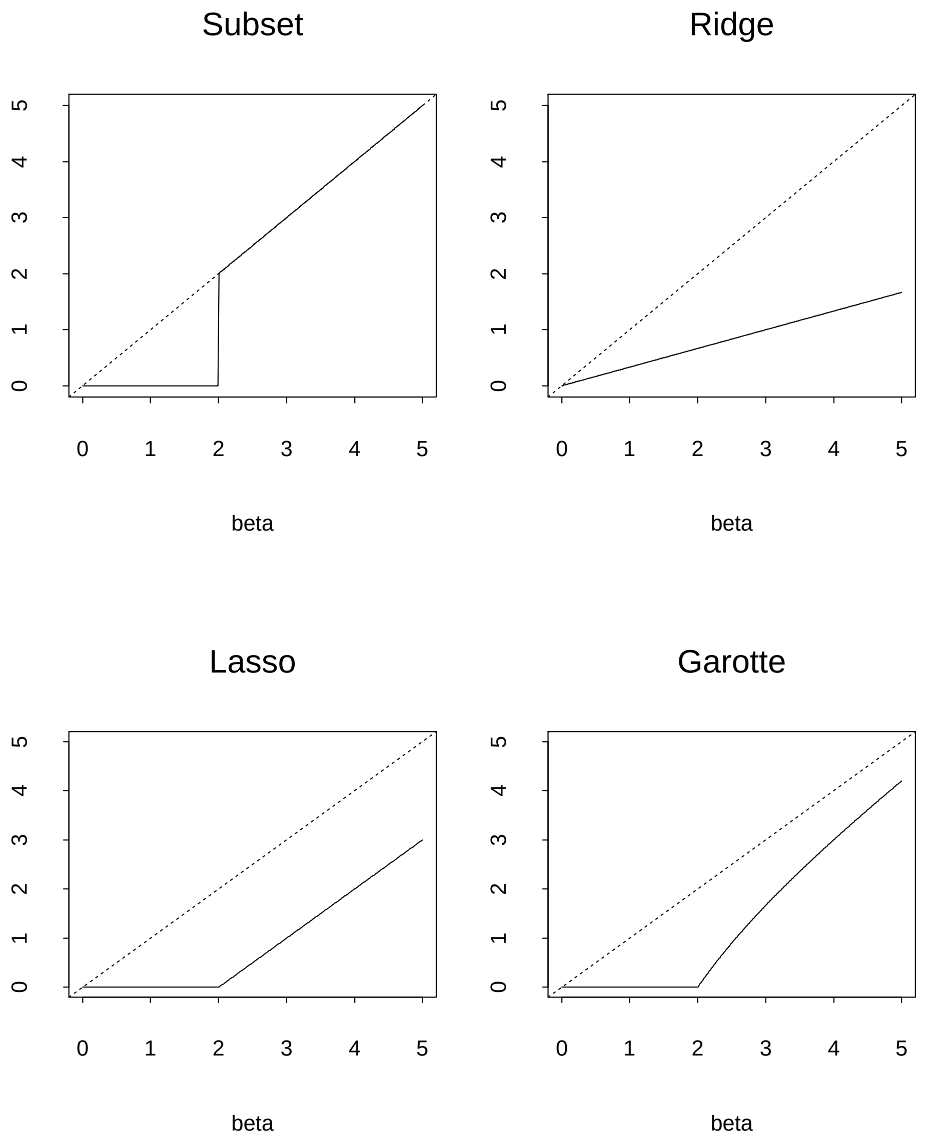

Figure 1: Dark line shows form of coefficient shrinkage from the 45^o line for each technique (orthogonal design case)

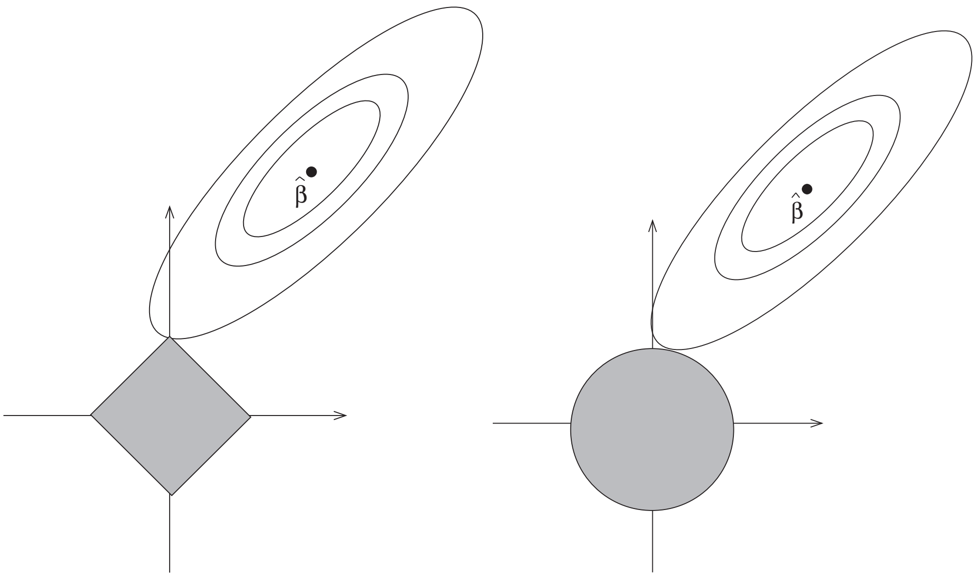

Figure 2: Estimation picture for the lasso (left) and ridge regression (right)

there is at least moderate correlation in the data, this need not be true. Figure 3 shows an example in three dimensions. The view in the right plot confirms that the ellipse touches the constraint region in a different octant than the one in which its center lies.

While the garotte retains the sign of each $\hat{\beta}_j^\circ$, the lasso can change signs. Even in cases where the lasso estimate has the same sign vector as the garotte, the presence of the ordinary least squares estimates in the garotte can make it behave differently. The model $\sum c_j \hat{\beta}_j^\circ x_{ij}$ with constraint $\sum c_j \le t$ can be written as $\sum \beta_j x_{ij}$ with constraint $\sum \beta_j / \hat{\beta}_j^\circ \le t$. If for example $p = 2$ and $\hat{\beta}_1^\circ > \hat{\beta}_2^\circ > 0$ then the effect would be to stretch the square in the left panel of Figure 2 horizontally. As a result, larger values of β_1 and smaller values of β_2 will be favoured by the garotte.

2.4 More on the two predictor case

Suppose $p = 2$, and assume without loss of generality that the least squares estimates $\hat{\beta}_j^\circ$ are both positive. Then one can show that the lasso estimates are

$$\hat{\beta}_j = (\hat{\beta}_j^\circ - \gamma)^+ \tag{5}$$

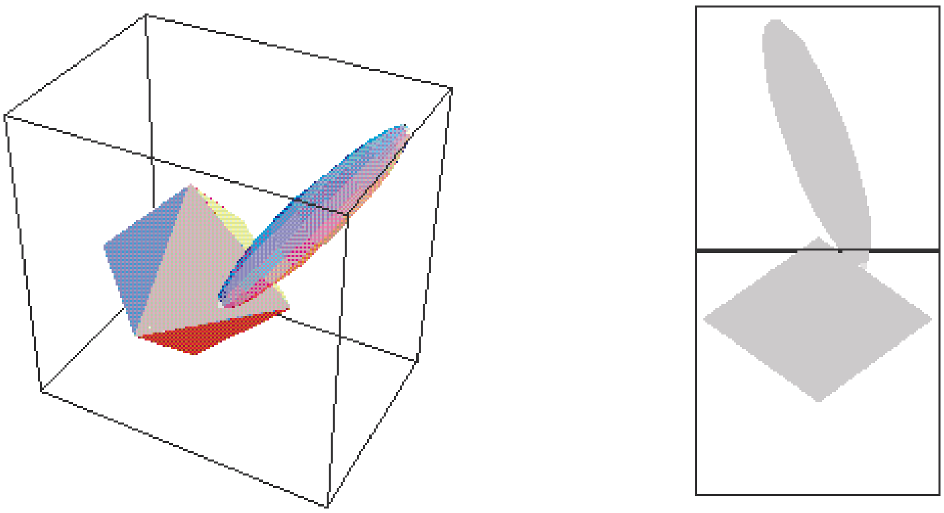

Figure 3: Left panel shows an example in which the lasso estimate falls in a different octant than the overall least squares estimate. Right panel shows an overhead view.

where γ is chosen so that $\hat{\beta}_1 + \hat{\beta}_2 = t$. This formula holds for $t \le \hat{\beta}_1^\circ + \hat{\beta}_2^\circ$, and is valid even if the predictors are correlated. Solving for γ yields

$$\begin{aligned} \hat{\beta}_1 &= [\frac{t}{2} + \frac{\hat{\beta}_1^\circ - \hat{\beta}_2^\circ}{2}]^+ \\ \hat{\beta}_2 &= [\frac{t}{2} - \frac{\hat{\beta}_1^\circ - \hat{\beta}_2^\circ}{2}]^+ \end{aligned} \tag{6}$$

In contrast, the form of ridge regression shrinkage depends on the correlation of the predictors. Figure 4 shows an example. We generated 100 data points from the model $y = 6x_1 + 3x_2$ with no noise. Here x_1 and x_2 are standard normal variates with correlation ρ. The curves in Figure 4 show the ridge and lasso estimates as the bounds on $\beta_1^2 + \beta_2^2$ and $|\beta_1| + |\beta_2|$, respectively, are varied. For all values of ρ the lasso estimates follow the solid curve. The ridge estimates (broken curves) depend on ρ. When $\rho = 0$ ridge regression does proportional shrinkage. However for larger values of ρ the ridge estimates are shrunken differentially and can even increase a little as the bound is decreased. As pointed out by Jerome Friedman, this is due to the tendency of ridge regression to try tomake the coefficients equal in order to minimize their squared norm.

2.5 Standard errors

Since the lasso estimate is a non-linear and non-differentable function of the response values even for a fixed value of t, it is difficult to obtain an accurate estimate of its standard error. One approach is via the bootstrap: either t can be fixed or we may optimize over t for each bootstrap sample. Fixing t is analogous to selecting a best subset, and then using the least squares standard error for that subset.

An approximate closed form estimate may be derived by writing the penalty $\sum |\beta_j|$ as $\sum {\beta_j}^2 / |\beta_j|$. Hence at the lasso estimate $\tilde{\boldsymbol{\beta}}$, we may approximate the solution by a ridge regression of the form $\boldsymbol{\beta}^* = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{W}^-)^{-1}\mathbf{X}^T\mathbf{y}$ where $\mathbf{W}$ is a diagonal matrix with diagonal elements $|\tilde{\beta}_j|$, $\mathbf{W}^-$ denotes the generalized inverse of $\mathbf{W}$ and λ is chosen so that $\sum |\beta_j|^* = t$. The covariance matrix of the estimates may then approximated by

$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{W}^-)^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{W}^-)^{-1}\hat{\sigma}^2, \tag{7}$$

where $\hat{\sigma}^2$ is an estimate of the error variance. A difficulty with this formula is that it gives an estimated variance of zero for predictors with $\hat{\beta}_j = 0$.

This approximation also suggests an iterated ridge regression algorithm for computing the lasso estimate itself, but this turns out to be quite inefficient. However, it does prove to be useful for selection of the lasso parameter t (section 4).

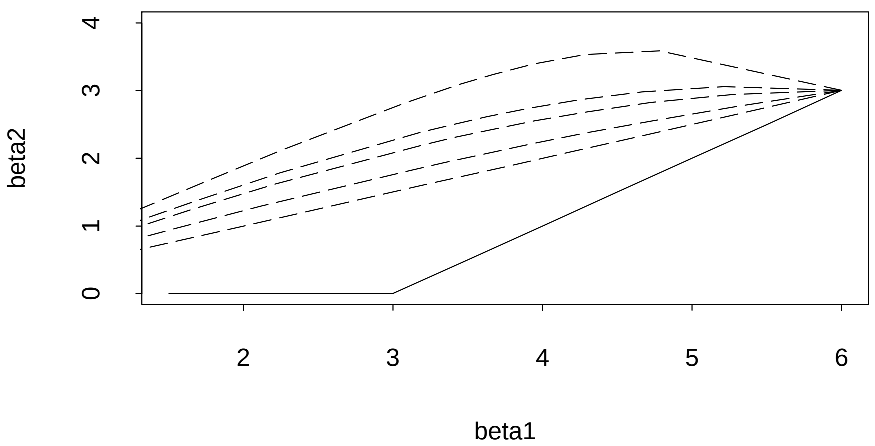

Figure 4: Lasso (solid curve) and ridge (broken curves) for the two predictor example. The curves show the (β_1, β_2) pairs as the bound on the lasso or ridge parameters is varied. Starting with the bottom broken curve and moving upward, the correlation ρ is 0, .23, .45, .68, and .90.

3 Example- prostate cancer data

This data comes from a study by Stamey *et. al.* (1989) that examined the correlation between the level of prostate specific antigen and a number of clinical measures, in men who were about to receive a radical prostatectomy. The factors were log cancer volume (lcavol), log prostate weight (lweight), age, log of benign prostatic hyperplasia amount (lbph), seminal vesicle invasion (svi), log of capsular penetration (lcp), Gleason score (gleason), and percent Gleason scores 4 or 5 (pgg45). We fit a linear model to the log of prostate specific antigen (lpsa) after first standardizing the predictors.

Figure 5 shows the lasso estimates as a function of standardized bound $s = t/\sum|\hat{\beta}_j^o|$. Notice that the absolute value of each coefficient tends to zero as s goes to zero. In this example, the curves decrease in a monotone fashion to zero, but this doesn't always happen in general. This lack of monotonicity is shared by ridge regression and subset regression, where for example the best subset of size 5 may not contain the best subset of size 4. The vertical broken line represents the model for $\hat{s} = .44$, the optimal value as selected by generalized cross-validation. Roughly speaking, this corresponds to keeping just under half of the predictors.

Table 1 shows the results for the full least squares, best subset and lasso procedures. Section 7.1 gives the details of the best subset procedure that was used. The lasso gave non-zero coefficients to lcavol, lweight and svi; subset selection chose the same three predictors. Notice that the coefficients and Z scores for the selected predictors from subset selection tend to be larger than the full model values: this is a common occurence with positively correlated predictors. However the lasso shows the opposite effect, as it shrinks the coefficients and Z scores from their full model values.

The standard errors in the second column from the right were estimated by bootstrap resampling of residuals from the full least squares fit. The standard errors were computed by fixing $\hat{s}$ at its optimal value 0.44 for the original dataset. Table 2 compares the the ridge approximation formula (7) with the fixed t bootstrap, and the bootstrap in which t was re-estimated for each sample. The ridge formula gives a fairly good approximation to the fixed t bootstrap, except for the zero coefficients. Allowing t to vary incorporates an additional source of variation, and hence gives larger standard error estimates. Figure 6 shows boxplots of 200 bootstrap replications of the lasso estimates, with $\hat{s}$ fixed at the estimated value 0.44. The predictors whose estimated coefficient is zero exhibit skewed bootstrap distributions. The central 90% percentile intervals (5th and 95 percentiles of the bootstrap distributions) all contained the value zero, with the exceptions of those for lcavol and svi.

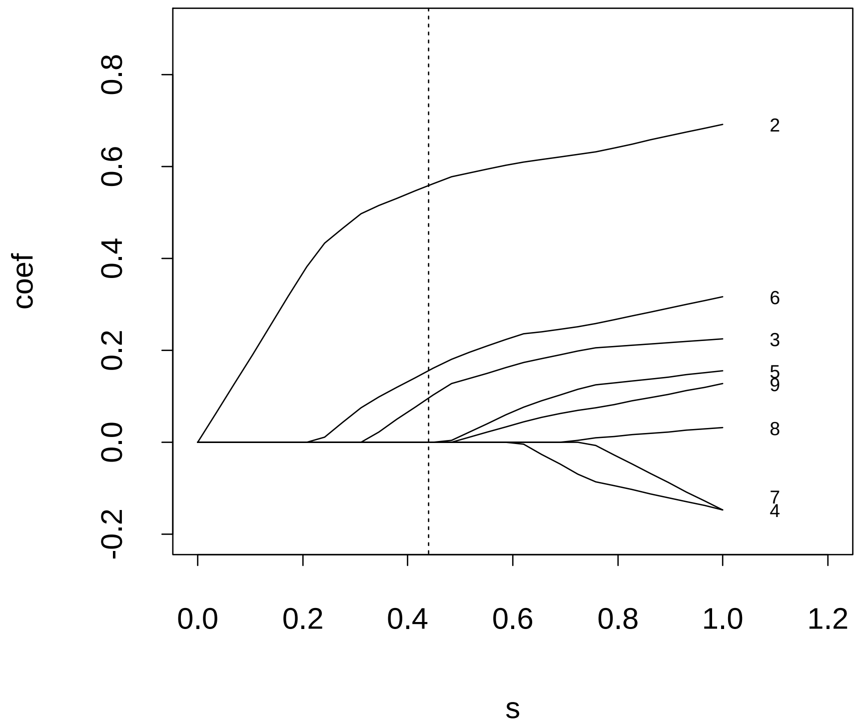

Figure 5: Lasso shrinkage of coefficients in prostate cancer example. Each curve represents a coefficient (labelled on the right) as function of the (scaled) lasso parameter $s = t / \sum |\hat{\beta}_j^\circ|$. The intercept is not plotted. The vertical broken line represents the model for $\hat{s} = .44$, selected by generalized cross-validation.

Table 1: Results for prostate cancer example

Predictor	Least Squares			Subset			Lasso		
	Coef	Se	Z score	Coef	Se	Z score	Coef	Se	Z score
1. Intcpt	2.48	0.07	34.46	2.48	0.07	34.05	2.48	0.07	35.43
2. lcavol	0.69	0.10	6.68	0.65	0.09	7.39	0.56	0.09	6.22
3. lweight	0.23	0.08	2.67	0.25	0.07	3.39	0.10	0.07	1.43
4. age	-0.15	0.08	-1.76	0.00	0.00	-	0.00	0.01	0.00
5. lbph	0.16	0.08	1.83	0.00	0.00	0.00	0.00	0.04	0.00
6. svi	0.32	0.10	3.14	1.02	0.28	0.09	0.16	0.09	1.78
7. lcp	-0.15	0.13	-1.16	0.00	0.00	-	0.00	0.03	0.00
8. gleason	0.03	0.11	0.29	0.00	0.00	-	0.00	0.02	0.00
9. pgg45	0.13	0.12	1.02	0.00	0.00	0.00	0.00	0.03	0.00

Table 2: Standard error estimates for prostate cancer example

		Bootstrap SE		
Predictor	Coefficient	Fixed t	Varying t	SE Approximation (7)
1. Intcpt	2.48	0.07	0.07	0.07
2. lcavol	0.56	0.08	0.10	0.09
3. lweight	0.10	0.06	0.08	0.06
4. age	0.00	0.04	0.05	0.00
5. lbph	0.00	0.04	0.07	0.00
6. svi	0.16	0.09	0.09	0.07
7. lcp	0.00	0.03	0.07	0.00
8. gleason	0.00	0.02	0.05	0.00
9. pgg45	0.00	0.03	0.06	0.00

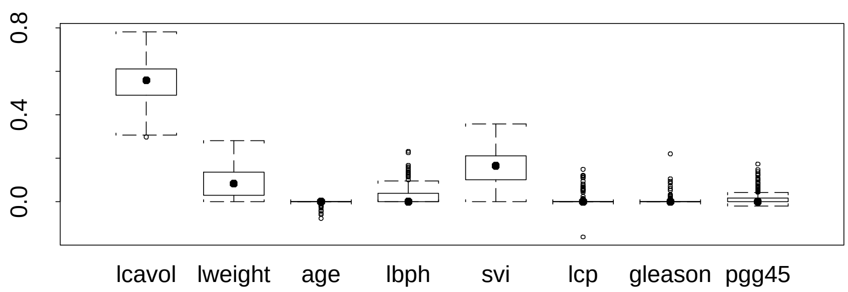

Figure 6: Boxplots of 200 bootstrap values of the lasso coefficient estimates for the 8 predictors in the prostate cancer example.

4 Prediction error and estimation of t

In this section we describe three methods for the estimation of the lasso parameter t: cross-validation, generalized cross-validation and an analytic unbiassed estimate of risk. Strictly speaking the first two methods are applicable in the "X-random" case, where it is assumed that the observations $(\mathbf{X}, Y)$ are drawn from some unknown distribution, and the third method applies to the the X-fixed case. However in real problems there is often no clear distinction between the two scenarios and one might simply choose the most convenient method.

Suppose

$$Y = \eta(\mathbf{X}) + \epsilon$$

where $\mathrm{E}(\epsilon) = 0$, $\mathrm{var}(\epsilon) = \sigma^2$. The mean squared error of an estimate $\hat{\eta}(\mathbf{X})$ is defined by

$$\mathrm{ME} = \mathrm{E}(\hat{\eta}(\mathbf{X}) - \eta(\mathbf{X}))^2,$$

the expected value taken over the joint distribution of $\mathbf{X}$ and Y, with $\hat{\eta}(\mathbf{X})$ fixed. A similar measure is the prediction error of $\hat{\eta}(\mathbf{X})$ given by

$$\mathrm{PE} = \mathrm{E}(Y - \hat{\eta}(\mathbf{X}))^2 \quad = \quad \mathrm{ME} + \sigma^2 \tag{8}$$

We estimate the prediction error for the lasso procedure by five-fold cross-validation as described (for example) in Chapter 17 of Efron & Tibshirani (1993). The lasso is indexed in terms of the normalized parameter $s = t/\sum \hat{\beta}_j^o$, and the prediction error is estimated over a grid of values of s from 0 to 1 inclusive. The value $\hat{s}$ yielding the lowest estimated PE is selected.

Simulation results are reported in terms of ME rather PE. For the linear models $\eta(\mathbf{X}) = \mathbf{X}\hat{\boldsymbol{\beta}}$ considered in this paper, mean-squared error has the simple form

$$\mathrm{ME} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T V(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

where V is the population covariance matrix of $\mathbf{X}$

A second method for estimating t may be derived from a linear approximation to the lasso estimate. We write the constraint $\sum|\beta_j| \le t$ as $\sum \beta_j^2/|\beta_j| \le t$. This latter constraint is equivalent to adding a Lagrangian penalty $\lambda \sum \beta_j^2/|\beta_j|$ to the residual sum of squares, with λ depending on t. Thus we may write constrained solution $\tilde{\boldsymbol{\beta}}$ as the ridge regression estimator

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{W}^-)^{-1}\mathbf{X}^T\mathbf{y} \tag{9}$$

where $\mathbf{W} = \mathrm{diag}(|\tilde{\beta}_j|)$ and $\mathbf{W}^-$ denotes a generalized inverse. Therefore the number of effective parameters in the constrained fit $\tilde{\boldsymbol{\beta}}$ may be approximated by

$$p(t) = \mathrm{tr}[\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{W}^-)^{-1}\mathbf{X}^T]$$

Letting $\mathrm{rss}(t)$ be the residual sum of squares for the constrained fit with constraint t, we construct the GCV-style statistic:

$$\mathrm{GCV}(t) = \frac{1}{N}\frac{\mathrm{rss}(t)}{[1 - p(t)/N]^2}. \tag{10}$$

Finally, we outline a third method based on Stein's unbiased estimate of risk. Suppose that $\mathbf{z}$ is a multivariate normal random vector with mean $\boldsymbol{\mu}$ and variance the identity matrix. Let $\hat{\boldsymbol{\mu}}$ be an estimator of $\boldsymbol{\mu}$, and write $\hat{\boldsymbol{\mu}} = \mathbf{z} + \mathbf{g}(\mathbf{z})$ where $\mathbf{g}$ is an almost differential function from R^p to R^p (see definition 1 of Stein, 1981). Then Stein (1981) showed that

$$\mathrm{E}_{\boldsymbol{\mu}}||\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}||^2 = p + \mathrm{E}_{\boldsymbol{\mu}}\Big[||\mathbf{g}(\mathbf{z})||^2 + 2\sum_1^p dg_i/dz_i\Big] \tag{11}$$

We may apply this result to the lasso estimator (3). Denote the estimated standard error of $\hat{\beta}_j^\circ$ by $\hat{\tau} = \hat{\sigma}/\sqrt{N}$, where $\hat{\sigma}^2 = \sum(y_i - \hat{y}_i)^2/(N-p)$. Then the $\hat{\beta}_j^\circ/\hat{\tau}$ are (conditionally on $\mathbf{X}$) approximately independent standard normal variates, and from equation (11) we may derive the formula

$$R[\hat{\boldsymbol{\beta}}(\gamma)] \approx \hat{\tau}^2\Big[p - 2\cdot\#(j; |\hat{\beta}_j^\circ/\hat{\tau}| < \gamma) + \sum_{j=1}^p \max(|\hat{\beta}_j^\circ/\hat{\tau}|, \gamma)^2\Big]$$

as an approximately unbiased estimate of the risk or mean square error $\mathrm{E}(\hat{\boldsymbol{\beta}}(\gamma) - \boldsymbol{\beta})^2$, where $\hat{\beta}_j(\gamma) = \mathrm{sign}(\hat{\beta}_j^\circ)(|\hat{\beta}_j^\circ/\hat{\tau}| - \gamma)^+$. Donoho & Johnstone (1994) give a similar formula in the function estimation setting. Hence an estimate of γ can be obtained as the minimizer of $R[\hat{\boldsymbol{\beta}}(\gamma)]$:

$$\hat{\gamma} = \text{argmin}_{\gamma \geq 0} R[\hat{\boldsymbol{\beta}}(\gamma)].$$

From this we obtain an estimate of the lasso parameter t:

$$\hat{t} = \sum (|\hat{\beta}_j^o| - \hat{\gamma})^+$$

Although the derivation of $\hat{t}$ assumes an orthogonal design, we may still try to use it in the usual non-orthogonal setting. Since the predictors have been standardized, the optimal value of t is roughly a function of the overall signal-to-noise ratio in the data, it should be relatively insensitive to the covariance of $\mathbf{X}$. (On the other hand, the form of the lasso estimator *is* sensitive to the covariance and we need to account for it properly.)

The simulated examples in section 7.2 suggest that this method gives a useful estimate of t. But we can offer only a heuristic argument in favour of it. Suppose $\mathbf{X}^T\mathbf{X} = \mathbf{V}$ and let $\mathbf{Z} = \mathbf{X}\mathbf{V}^{-1/2}$, $\boldsymbol{\theta} = \boldsymbol{\beta}\mathbf{V}^{-1/2}$. Since the columns of $\mathbf{X}$ are standardized, the region $\sum |\theta_j| \leq t$ differs from the the region $\sum |\beta_j| \leq t$ in shape but has roughly the same-sized marginal projections. Therefore the optimal value of $\hat{t}$ should be about the same in each instance.

Finally, note that the Stein method enjoys a significant computational advantage over the cross validation-based estimation of t. In our experiments we optimized over a grid of 15 values of the lasso parameter t and used 5-fold cross-validation. As a result, the cross validation approach required 75 applications of the model optimization procedure of section 6 while the Stein method required only one. The requirements of the GCV approach are intermediate between the two, requiring one application of the optimization procedure per grid point.

5 The lasso as a Bayes estimate

The lasso constraint $\sum |\beta_j| \leq t$ is equivalent to the addition of a penalty term $\lambda \sum |\beta_j|$ to the residual sum of squares (see Murray, Gill and Wright, 1981, chapter 5). Now $|\beta_j|$ is proportional to the (minus) log-density of the double exponential distribution. As a result one can derive the lasso estimate as the Bayes posterior mode under independent double exponential priors for the β_js,

$$f(\beta_j) = \frac{1}{2\tau} \exp\{-\frac{|\beta_j|}{\tau}\}$$

with $\tau = 1/\lambda$.

Figure 7 shows the double exponential density (solid curve) and the normal density (broken curve); the latter is the implicit prior used by ridge regression. Notice how the double exponential density puts more mass near zero and in the tails. This reflects the greater tendency of the lasso to produce estimates that are either large or zero.

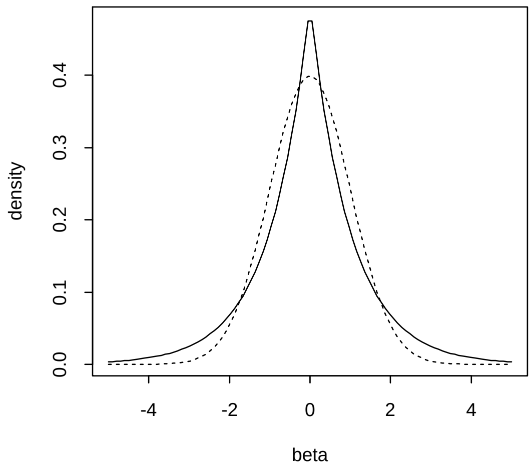

Figure 7: The double exponential density (solid curve) and the normal density (broken curve). The former is the implicit prior used by the lasso; the latter by ridge regression.

6 Algorithms for finding the lasso solutions

We fix $t \geq 0$. Problem (1) can be expressed as a least squares problem with 2^p inequality constraints, corresponding to the 2^p different possible signs for the β_js. Lawson & Hansen (1974) provide the ingredients for a procedure which solves the linear least squares problem subject to a general linear inequality constraint $G\boldsymbol{\beta} \leq \mathbf{h}$. Here G is an $m \times p$ matrix, corresponding to m linear inequality constraints on the p-vector $\boldsymbol{\beta}$. For our problem however, $m = 2^p$ may be very large so that direct application of this procedure is not practical. However the problem can be solved by introducing the inequality constraints sequentially, seeking a feasible solution satisfying the so-called Kuhn-Tucker conditions (Lawson and Hansen, 1974). We outline the procedure below.

Let $g(\boldsymbol{\beta}) = \sum_{i=1}^{N}(y_i - \sum_j \beta_j x_{ij})^2$, and let $\boldsymbol{\delta}_i$, $i = 1, 2, \ldots 2^p$ be the p-tuples of the form $(\pm 1, \pm 1, \ldots \pm 1)$. Then the condition $\sum |\beta_j| \leq t$ is equivalent to $\boldsymbol{\delta}_i^T \boldsymbol{\beta} \leq t$ for all i. For a given $\boldsymbol{\beta}$, let $E = \{i : \boldsymbol{\delta}_i^T \boldsymbol{\beta} = t\}$ and $S = \{i : \boldsymbol{\delta}_i^T \boldsymbol{\beta} < t\}$. The set E is the equality set, corresponding to those constraints which are exactly met, while S is the slack set, corresponding to those constraints for which equality does not hold. Denote by G_E the matrix whose rows are $\boldsymbol{\delta}_i$ for $i \in E$. Let $\mathbf{1}$ be a vector of ones of length equal to the number of rows of G_E.

The algorithm below starts with $E = \{i_0\}$ where $\boldsymbol{\delta}_{i_0} = \mathrm{sign}(\hat{\boldsymbol{\beta}})$, $\hat{\boldsymbol{\beta}}$ being the overall least squares estimate. It solves the least squares problem subject to $\boldsymbol{\delta}_{i_0}^T \boldsymbol{\beta} \leq t$ and then checks if $\sum |\beta_j| \leq t$. If so, the computation is complete; if not, the violated constraint is added to E and the process is continued until $\sum |\beta_j| \leq t$.

Here is an outline of the algorithm:

1. Start with $E = \{i_0\}$ where $\boldsymbol{\delta}_{i_0} = \mathrm{sign}(\hat{\boldsymbol{\beta}}^{\circ})$, $\hat{\boldsymbol{\beta}}^{\circ}$ being the overall least squares estimate.

2. Find $\hat{\boldsymbol{\beta}}$ to minimize $g(\boldsymbol{\beta})$ subject to $G_E \boldsymbol{\beta} \leq t\mathbf{1}$.

3. While $\{\sum |\hat{\beta}_j| > t\}$

    4. Add i to the set E where $\boldsymbol{\delta}_i = \mathrm{sign}(\hat{\boldsymbol{\beta}})$. Find $\hat{\boldsymbol{\beta}}$ to minimize $g(\boldsymbol{\beta})$ subject to $G_E \boldsymbol{\beta} \leq t\mathbf{1}$.

This procedure must always converge in a finite number of steps since one element is added to the set E at each step, and there are a total of 2^p elements. The final iterate is a solution to the original problem since the Kuhn-Tucker conditions are satisfied for the sets E and S at convergence.

A modification of the above procedure removes elements from E in step 4 for which the equality constraint is not satisfied. This is more efficient but it is not clear how to establish its convergence.

The fact that the algorithm must stop after at most 2^p iterations is of little comfort if p is large. In practice we have found that the average number of iterations required is in the range $(.5p, .75p)$, and is therefore quite acceptable for practical purposes.

A completely different algorithm for this problem was suggested by David Gay. We write each β_j as $\beta_j^+ - \beta_j^-$, where β_j^+ and β_j^- are non-negative. Then we solve the least squares problem with the constraints $\beta_j^+ \geq 0$, $\beta_j^- \geq 0$, and $\sum \beta_j^+ + \sum_j \beta_j^- \leq t$. In this way we transform the original problem (p variables, 2^p constraints) to a new problem with more variables ($2p$) but fewer constraints ($2p+1$). One can show that this new problem has the same solution as the original one.

Standard quadratic programming techniques can be applied, with the convergence assured in $2p+1$ steps. We have not extensively compared these two algorithms, but in examples have found that the second algorithm is usually (but not always) a little faster than the first algorithm.

7 Simulations

7.1 Outline

In the following examples, we compare the full least squares estimates to the lasso, non-negative garotte, best subset selection, and ridge regression. We used five-fold cross-validation to estimate the regularization parameter in each case. For best subset selection, we used the "leaps" procedure in the S language, with five-fold cross-validation to estimate the best subset size. This procedure is described and studied in Breiman and Spector (1992). Breiman and Spector, recommend five or ten-fold cross-validation for use in practice.

For completeness, here are the details of this procedure. The best subsets of each size are first found for the original dataset: call these $S_0, S_2, \ldots S_p$. (S_0 represents the null model; since $\bar{y} = 0$ the fitted values are zero for this model). Denote the full training set by T, and the cross-validation training and test sets by $T - T^\nu$ and T^ν, for $\nu = 1, 2, \ldots 5$. For each cross-validation fold ν, we find the best subsets of each size for the data $T - T^\nu$: call these $S_0^\nu, S_1^\nu, \ldots S_p^\nu$. Let $\mathrm{PE}^\nu(J)$ be the prediction error when S_J^ν is applied to the test data T^ν, and form the estimate

$$\mathrm{PE}(J) = \frac{1}{5} \sum_{\nu=1}^{5} \mathrm{PE}^\nu(J) \tag{12}$$

We find the $\hat{J}$ that minimizes $\mathrm{PE}(J)$ and our selected model is $S_{\hat{J}}$. Note that this is not the same as estimating the prediction error of the fixed models $S_0, S_1, \ldots S_p$ and then choosing the one with smallest prediction error. This

Table 3: Results for example 1

Method	Median ME (stand. err.)	Ave. # of zero coefs	Ave. $\hat{s}$.
Least squares	2.79(.12)	0.0	-
Lasso (CV)	2.43(.14)	3.3	.63(.01)
Lasso (Stein)	2.07(.10)	2.6	.69(.02)
Lasso (GCV)	1.93(.09)	2.4	.73(.01)
Garotte	2.29(.16)	3.9	-
Best Subset	2.44(.16)	4.8	-
Ridge	3.21(.12)	0.0	-

latter procedure is described in Zhang (1993) and Shao (1992), and can lead to inconsistent model selection unless the cross-validation test set T^{ν} grows at an appropriate asymptotic rate.

7.2 Example 1

In this example we simulated 50 datasets consisting of 20 observations from the model

$$y = \boldsymbol{\beta}^T \mathbf{x} + \sigma \cdot \epsilon,$$

where $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ and ϵ is standard normal. The correlation between x_i and x_j was $\rho^{|i-j|}$ with $\rho = .5$. We set $\sigma = 3$, and this gave a signal to noise ratio of approximately 5.7. Table 3 shows the mean squared errors over 200 simulations from this model. Lasso performs the best, followed by garotte and ridge.

Estimation of the lasso parameter by generalized cross-validation seems to perform best,a trend that we find is consistent through all of our examples. Subset selection picks approximately the correct number of zero coefficients (5), but suffers from too much variability as shown in the boxplots of Figure 8.

Table 4 shows the 5 most frequent models (non-zero coefficients) selected by the lasso (with GCV): although the correct model (1,2,5) was chosen only 2.5% of the time, the selected model contained (1,2,5) 95.5% of the time. The most frequent models selected by subset regression are shown in Table 5. The correct model chosen more often (24% of the time), but subset selection can also underfit: selected model contained (1,2,5) only 53.5% of the time.

7.3 Example 2

This is the same as example 1, but with $\beta_j = .85\ \forall j$ and $\sigma = 3$; the signal to noise ratio was approximately 1.8. The results in the left side Table 6 show that

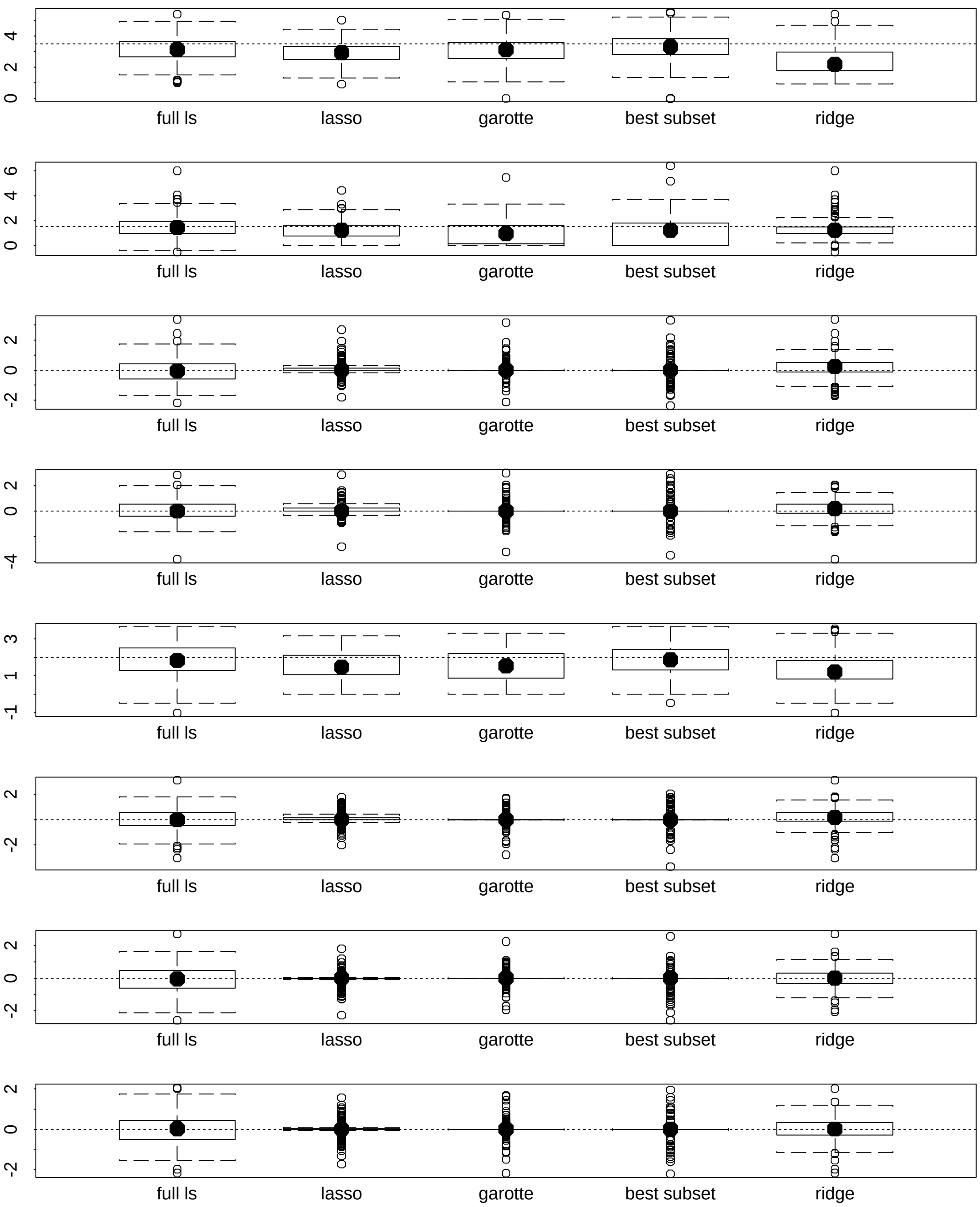

Figure 8: Estimates for the 8 coefficients in example 1, excluding the intercept; horizontal dotted lines indicate the true coefficients

Table 4: Most frequent models selected by the lasso (GCV) in Example 1

Model	Proportion
1245678	.055
123456	.050
1258	.045
1245	.045
13 others	
125 (and 5 others)	.025

Table 5: Most frequent models selected by all subsets regression in Example 1

Model	Proportion
125	.240
15	.200
1	.095
1257	.040

ridge regression does the best by a good margin, with the lasso being the only other method to outperform the full least squares estimate. The right side of the table shows the results when the sample size is increased from 20 to 100. As expected, the performance of most of the procedures improves. A notable exception is the lasso with shrinkage parameter chosen by the Stein method: on the average it shrinks by about 50% when no shrinkage is needed.

7.4 Example 3

Here we chose a setup that should be well-suited for subset selection. The model is the same as example 1, but with $\beta = (5, 0, 0, 0, 0, 0, 0, 0)$, and $\sigma = 2$ so that the signal to noise ratio was about 7.

The results in Table 7 show that the garotte and subset selection perform the best, followed closely by the lasso. Ridge regression does poorly, and has higher mean squared error than the full least squares estimates.

7.5 Example 4

In this example we examine the performance of the lasso in a bigger model. We simulated 50 datasets each having 100 observations and 40 variables (note that best subsets regression is generally considered impractical for $p > 30$). We defined predictors $x_{ij} = z_{ij} + z_i$ where z_{ij} and z_i are independent standard

Table 6: Results for example 2

Method	N=20 Med ME (se.)	N=20 Ave.# of zeroes	N=20 Ave. $\hat{s}$	N=100 Med ME (se.)	N=100 Ave.# of zeroes	N=100 Ave. $\hat{s}$
Least squares	6.50(.64)	0.0	-	.79 (.06)	0.0	-
Lasso (CV)	5.30(.45)	3.0	.50(.03)	.92 (.05)	0.1	.96(.01)
Lasso (Stein)	5.85(.36)	2.7	.55(.03)	4.24 (.44)	1.6	.55(.01)
Lasso (GCV)	4.87(.35)	2.3	.69(.23)	.86 (.06)	0.3	.97 (.01)
Garotte	7.40(.48)	4.3	-	.96 (.07)	0.3	-
Subset	9.05(.78)	5.2	-	1.03 (.08)	0.9	-
Ridge	2.30(.22)	0.0	-	.72 (.04)	0.0	-

Table 7: Results for example 3

Method	Median ME (stand. err.)	Ave.# of zero coefs	Ave. $\hat{s}$.
Least squares	2.89(.04)	0.0	-
Lasso (CV)	0.89(.01)	3.0	.50(.03)
Lasso (Stein)	1.26(.02)	2.6	.70(.01)
Lasso (GCV)	1.02(.02)	3.9	.63(.04)
Garotte	0.52(.01)	5.5	-
Subset	0.64(.02)	6.3	-
Ridge	3.53(.05)	0.0	-

normal variates. This induced a pairwise correlation of 0.5 among the predictors. The coefficient vector was $\boldsymbol{\beta} = (0, 0, \ldots 0, 2, 2, \ldots, 2, 0, 0, \ldots 0, 2, 2, \ldots 2)$, there being 10 repeats in each block. Finally we defined $y = \boldsymbol{\beta}^T \mathbf{x} + 15 \cdot \epsilon$ where ϵ was standard normal. This produced a signal to noise ratio of roughly 9. The results in Table **??** show that the ridge regression performs the best, with the lasso (GCV) a close second. 3

The average value of the lasso coefficients in each of the four blocks of ten were .50(.06), .92(.07), 1.56(.08), and 2.33(.09). While the lasso only produced 14.4 zero coefficients on the average, the average value of $\hat{s}$ (.55) was close to the true proportion of zeroes (.5).

8 The lasso for more general models

The lasso can be applied to a wide variety of models. Consider any model indexed by a vector parameter $\boldsymbol{\beta}$, for which estimation is carried out by maximization of a function $\ell(\boldsymbol{\beta})$; this may be a log-likelihood function or some other measure of fit. To apply the lasso, we standardize the predictors appropriately and then maximize $\ell(\boldsymbol{\beta})$ under the constraint $\sum |\beta_j| \leq t$.

One could carry out this maximization by a general (non-quadratic) programming procedure. Alternatively, consider here models for which a quadratic approximation to $\ell(\boldsymbol{\beta})$ leads to an IRLS (iteratively reweighted least squares) procedure for computation of $\boldsymbol{\beta}$. These models include generalized linear models and other generalized regression models. Using the IRLS approach, we can solve the constrained problem by iterative application of the lasso algorithm for linear models, within an IRLS loop.

Specifically, in the terminology of generalized linear models we define the linear predictor $\eta = \alpha + \sum_1^p X_j \beta_j$, and maximize the log-likelihood under the constraint $\sum |\beta_j| \leq t$. In this model we can no longer eliminate α by centering the response y. If z and w are the adjusted dependent variable and weights for the IRLS step, we center z via $z_i^c = z_i - \sum z_i w_i / \sum w_i$ and similarly for $X_1, X_2, \ldots X_p$. Then we minimize $\sum w_i (z_i^c - \sum_j x_{ij}^c \beta_j)^2$ subject to $\sum |\beta_j| \leq t$. It is simple to modify the algorithms of section 6 to incorporate weights.

Convergence of this procedure is not assured in general, but in our limited experience it has behaved quite well. Tibshirani (1994) applies this idea to the proportional hazards model for survival data. Below we give a brief illustration to logistic regression.

8.1 Logistic regression

For illustration we applied the lasso to the logistic regression model for binary data. We used the kyphosis data, analyzed in Hastie & Tibshirani (1990), chapter 10. The response is kyphosis (0=absent, 1=present); the predictors x_1=age, x_2=number of vertebrae levels, and x_3=starting vertebrae level. There are 83 observations. Since the predictor effects are known to be nonlinear, we included squared terms in the model after centering each of the variables. Finally, the columns of the data matrix were standardized.

The linear logistic fitted model is

$$-2.64 + 0.83x_1 + 0.77x_2 - 2.28x_3 - 1.55x_1^2 + 0.03x_2^2 - 1.17x_3^2$$

Backward stepwise deletion, based on Akaike's information criterion, dropped the x_2^2 term and produced the model

$$-2.64 + 0.84x_1 + 0.80x_2 - 2.28x_3 - 1.54x_1^2 - 1.16x_3^2$$

The lasso chose $\hat{s} = .33$ giving the model

$$-1.51 + 0.01x_1 + 0.37x_2 - 0.61x_3 - 0.39x_1^2$$

Convergence, defined as the $\left\|\hat{\boldsymbol{\beta}}^{new} - \hat{\boldsymbol{\beta}}^{old}\right\|^2 < 10e^{-6}$, was obtained in 5 iterations.

9 Some further extensions

We are currently exploring two quite different applications of the lasso idea. One application is to tree-based models, as reported in LeBlanc and Tibshirani (1994). Rather than prune a large tree as in Breiman *et.al*'s (1984) CART procedure, we use the lasso idea to shrink it. This involves a constrained least squares operation much like the one in this paper, with the parameters being the mean contrasts at each node. A further set of constraints is needed to ensure that the shrunken model is a tree. Results reported in LeBlanc and Tibshirani (1994) suggest that the shrinkage procedure gives more accurate trees than pruning, while still producing interpretable subtrees.

A different application is to the Multivariate Adaptive Regression Spline (MARS) proposal of Friedman (1991). MARS is an adaptive procedure that builds a regression surface by sum of products of piecewise linear basis functions of the individual regressors. MARS builds a model that typically includes basis functions representing main effects and interactions of high order. Give the adaptively chosen bases, the MARS fit is simply a linear regression onto these bases. A backward stepwise procedure is then applied to eliminate less important terms.

In ongoing work with Trevor Hastie, we are developing a special lasso-type algorithm to dynamically grow and prune a MARS model. Hopefully this will produce more accurate MARS models and ones that also are interpretable.

The lasso idea can also be applied to ill-posed problems, in which the predictor matrix is not full rank. Chen and Donoho (1994) report some encouraging results for the use of lasso-style constraints in the context of function estimation via wavelets.

10 Results on soft-thresholding

Consider the special case of an orthonormal design $\mathbf{X}^T\mathbf{X} = \mathbf{I}$. Then the lasso estimate has the form

$$\hat{\beta}_j = \text{sign}(\hat{\beta}_j^\circ)(|\hat{\beta}_j^\circ| - \gamma)^+ \tag{13}$$

This is called a "soft-threshold" estimator by Donoho & Johnstone (1994); they apply this estimator to the coefficients of a wavelet transform of a function measured with noise. They then back-transform to obtain a smooth estimate of the function. Donoho and Johnstone prove many optimality results for soft-threshold estimator, and then translate these results into optimality results for function estimation.

Our interest here is not in function estimation but the coefficients themselves. We give one of Donoho and Johnstone's results here. It shows that asymptotically the soft-threshold estimator (lasso) comes as close as subset selection to the performance of an ideal subset selector— one that uses information about the actual parameters.

Suppose

$$y_i = \boldsymbol{\beta}\mathbf{x}^i + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$ and the design matrix is orthonormal. Then we can write

$$\hat{\beta}_j^\circ = \beta_j + \sigma z_j \tag{14}$$

where $z_j \sim N(0, \sigma^2)$.

We consider estimation of $\boldsymbol{\beta}$ under squared error loss, with risk

$$R(\hat{\beta}, \boldsymbol{\beta}) = E||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||^2.$$

Consider the family of diagonal linear projections

$$T_{DP}(\hat{\boldsymbol{\beta}}^\circ, \delta) = (\delta_j \hat{\beta}_j)_{j=1}^p \quad \delta \in \{0, 1\} \tag{15}$$

This estimator either keeps or kills a parameter $\hat{\beta}_j^\circ$, that is, it does subset selection. Now we incur a risk of σ^2 if we use $\hat{\beta}_j^\circ$, and β_j^2 if we use an estimate of zero instead. Hence the ideal choice of δ_j is $I(|\beta_j| > \sigma)$, that is, we keep only those predictors whose true coefficient is larger than the noise level. Call the risk of this estimator R_{DP}: of course this estimator cannot be constructed since the β_j are unknown. Hence R_{DP} is a lower bound on the risk we can hope to attain.

Donoho and Johnstone prove that the hard threshold (subset selection) estimator $\tilde{\beta}_j = \hat{\beta}_j^\circ I(|\hat{\beta}_j^\circ| > \gamma)$ has risk

$$R(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}) \le (2 \log p + 1)(\sigma^2 + R_{DP}) \tag{16}$$

Here γ is chosen as $\sigma(2\log n)^{1/2}$, the choice giving smallest asymptotic risk. They also show that the soft-threshold estimator (13) with $\gamma = \sigma(2\log n)^{1/2}$ achieves the same asymptotic rate.

These results lend some support to the potential utility of the lasso in linear models. However the important differences between the various approaches tend to occur for correlated predictors, and theoretical results such as those given here seem to be more difficult to obtain in that case.

11 Discussion

In this paper we have proposed a new method (the "lasso") for shrinkage and selection for regression and generalized regression problems. The lasso doesn't focus on subsets, but rather defines a continuous shrinking operation that can produce coefficients that are exactly zero. We have presented some evidence in this paper suggests that the lasso is a worthy competitor to subset selection and ridge regression. We examined the relative merits of the methods in three different scenarios:

Small number of large effects: Subset selection does best here, lasso not quite as well. Ridge does quite poorly.

Small to moderate number of moderate-sized effects: Lasso does best, followed by ridge and then subset selection

Large number of small effects: Ridge does best by a good margin, followed by lasso and then subset selection

Breiman's garotte does a little better than lasso in the first scenario, and a little worse in the second two scenarios. These results refer to prediction accuracy. Subset selection, lasso and garotte have the further advantage (vs ridge regression) of producing interpretable submodels.

There are many other ways to carry out subset selection or regularization in least squares regression. The literature is far too fast to attempt to summarize it in this short space so we mention only a few recent developments Computational advances have led to some interesting proposals, such as the Gibbs sampling approach of George & McCulloch (1993). They set up a hierarchical Bayes model and then use the Gibbs sampler to simulate a large collection of subset models from the posterior distribution. This allows the data analyst to examine the subset models with highest posterior probability, and can be carried out in large problems.

Frank and Friedman (1993) discuss a generalization of ridge regression and subset selection, through the addition of a penalty of the form $\lambda \sum_j |\beta_j|^q$ to the residual sum of squares. This is equivalent to a constraint of the form $\sum_j |\beta_j|^q \leq t$; they call this the "bridge". The lasso corresponds to $q = 1$. They suggest that joint estimation of the β_js and q might be an effective strategy, but do not report any results.

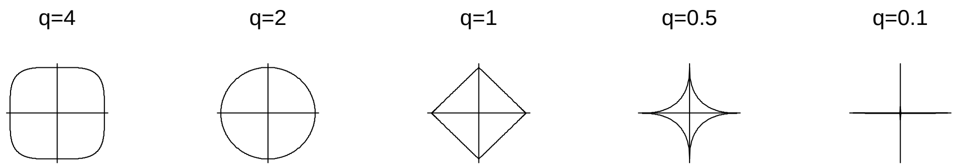

Figure 9: *Contours of constant value of* $\sum_j |\beta_j|^q$ *for given values of* q.

Figure 9 depicts the situation in 2 dimensions. Subset selection corresponds to $q \to 0$. The value $q = 1$ has advantage of being closer to subset selection than ridge regression ($q = 2$), and is also the smallest value of q giving a convex region. Furthermore, the linear boundaries for $q = 1$ are convenient for optimization.

The encouraging results reported here suggest that absolute value constraints might prove to be useful in a wide variety of statistical estimation problems. Further study is needed to investigate these possibilities.

Software

Public domain S/Splus language functions for the lasso are available at the statlib archive at Carnegie-Mellon University. There are functions for linear models, generalized linear models, and the proportional hazards model. To obtain them, ftp to `lib.stat.cmu.edu` and retrieve the file `S/lasso`, or send electronic mail to `statlib@lib.stat.cmu.edu` with the message `send lasso from S`.

Acknowledgements

I would like to thank Leo Breiman for sharing his garotte paper with me before publication, Michael Carter for assistance with the algorithm of section 6, and David Andrews for producing Figure 3 in Mathematica. I would also like to acknowledge enjoyable and fruitful discussions with David Andrews, Shaobeng Chen, Jerome Friedman, David Gay, Trevor Hastie, Geoff Hinton, Iain Johnstone, Stephanie Land, Michael Leblanc, Brenda MacGibbon, Stephen Stigler and Margaret Wright. Comments by editors and a referee led to substantial improvements in the manuscript. This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

References

Breiman, L. (1993), Better subset selection using the non-negative garotte, Technical report, Univ. of Cal., Berkeley.

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984), *Classification and Regression Trees*, Wadsworth.

Efron, B. & Tibshirani, R. (1993), *An Introduction to the Bootstrap*, Chapman and Hall.

Friedman, J. (1991), 'Multivariate adaptive regression splines (with discussion)', *Annals of Statistics* **19**(1), 1–141.

George, E. & McCulloch, R. (1993), 'Variable selection via gibbs sampling', *J. Amer. Statist. Assoc.* **88**, 884–889.

Hastie, T. & Tibshirani, R. (1990), *Generalized Additive Models*, Chapman and Hall.

Lawson, C. & Hansen, R. (1974), *Solving least squares problems*, Prentice-Hall.

Stein, C. (1981), 'Estimation of the mean of a multivariate normal distribution', *Ann. Statist.* **9**, 1135–1151.