Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2011)
73, *Part* 3, *pp.* 273–282

Regression shrinkage and selection via the lasso: a retrospective

Robert Tibshirani

Stanford University, USA

[*Presented to* The Royal Statistical Society *at its annual conference in a session organized by the* Research Section *on Wednesday, September 15th, 2010,* Professor D. M. Titterington *in the Chair*]

Summary. In the paper I give a brief review of the basic idea and some history and then discuss some developments since the original paper on regression shrinkage and selection via the lasso.

Keywords: l_1-penalty; Penalization; Regularization

1. The lasso

Given a linear regression with standardized predictors x_{ij} and centred response values y_i for $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, p$, the lasso solves the l_1-penalized regression problem of finding $\beta = \{\beta_j\}$ to minimize

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

This is equivalent to minimizing the sum of squares with a constraint of the form $\Sigma|\beta_j| \leqslant s$. It is similar to *ridge regression*, which has constraint $\Sigma_j \beta_j^2 \leqslant t$. Because of the form of the l_1-penalty, the lasso does variable selection and shrinkage, whereas ridge regression, in contrast, only shrinks. If we consider a more general penalty of the form $(\Sigma_{j=1}^{p} \beta_j^q)^{1/q}$, 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 l_1-norm penalty, and l_1-norms have been around for a long time! My most direct influence was Leo Breiman's *non-negative garrotte* (Breiman, 1995). His idea was to minimize, with respect to $c = \{c_j\}$,

$$\sum_{i=1}^{N}\Big(y_i - \sum_j c_j x_{ij}\hat{\beta}_j\Big)^2 \qquad \text{subject to } c_j \geqslant 0, \quad \sum_{j=1}^{p} c_j \leqslant t,$$

where $\hat{\beta}_j$ are usual least squares estimates. This is undefined when $p > N$ (which was not a hot

Address for correspondence: Robert Tibshirani, Department of Health Research and Policy, and Department of Statistics, Stanford University, Stanford, CA 94305, USA.
E-mail: tibs@stanford.edu

 1369–7412/11/73273

topic in 1995!) so I just combined the two stages into one (as a Canadian I also wanted a gentler name). In other related work around the same time, Frank and Friedman (1993) discussed *bridge regression* using a penalty $\lambda\Sigma|\beta_j|^\gamma$, with both λ and γ estimated from the data, and Chen *et al.* (1998) proposed *basis pursuit*, which uses an l_1-penalty in a signal processing context. Surely there are many other references that I am unaware of. After publication, the paper did not receive much attention until years later. Why?: my guesses are that

(a) the computation in 1996 was slow compared with today,
(b) the algorithms for the lasso were black boxes and not statistically motivated (until the LARS algorithm 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) the 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 does not scale well and is not transparent. The LARS algorithm (Efron *et al.*, 2002) gives an efficient way of solving the lasso and connects the lasso to forward stagewise regression. The same algorithm is contained in the homotopy approach of Osborne *et al.* (2000). *Co-ordinate 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, 2010), Wu and Lange (2008) and Genkin *et al.* (2007). We were made aware of its real potential in the doctoral thesis of Anita van der Kooij (Leiden) working with Jacqueline Meulman. The `glmnet` R language package (Friedman *et al.*, 2010) implements the co-ordinate 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 l_1-like penalties to a variety of problems. Table 1 gives a partial list. There has also been much 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 Bickel, Bühlmann, Candes, Donoho, Johnstone, Meinshausen, van de Geer, Wainwright and Yu. I do not have the qualifications or the space to summarize this work properly, but I hope that Professor Bühlmann will cover this aspect in his discussion.

Lasso methods can also shed light on more traditional techniques. The LARS algorithm, which was 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 log-likelihood plus $\lambda\|\Sigma^{-1}\|_1$, which is an l_1-penalty applied to the inverse covariance matrix. Since a missing edge in the graph corresponds to a zero element of Σ^{-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 *prespecified* edges (corresponding to structural 0s in Σ^{-1}). The details are given in chapter 17 of Hastie *et al.* (2008).

Another recent example is *nearly isotonic regression* (Fig. 1) (Tibshirani *et al.*, 2010). Given a data sequence $y_1, y_2, \ldots, y_N$ isotonic regression solves the problem of finding $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N$ to minimize

Table 1. A sampling of generalizations of the lasso

Method	*Reference*	*Detail*
Grouped lasso	Yuan and Lin (2007a)	$\Sigma_g \|\boldsymbol{\beta}_g\|_2$
Elastic net	Zou and Hastie (2005)	$\lambda_1 \Sigma \|\beta_j\| + \lambda_2 \Sigma \beta_j^2$
Fused lasso	Tibshirani *et al.* (2005)	$\lambda \Sigma \|\beta_{j+1} - \beta_j\|$
Adaptive lasso	Zou (2006)	$\lambda_1 \Sigma w_j \|\beta_j\|$
Graphical lasso	Yuan and Lin (2007b); Friedman *et al.* (2007)	$\text{loglik} + \lambda \|\Sigma^{-1}\|_1$
Dantzig selector	Candes and Tao (2007)	$\min\{X^{\mathrm{T}}(y - X\beta)\|_\infty\} \|\beta\|_1 < t$
Near isotonic regularization	Tibshirani *et al.* (2010)	$\Sigma(\beta_j - \beta_{j+1})_+$
Matrix completion	Candès and Tao (2009); Mazumder *et al.* (2010)	$\|X - \hat{X}\|^2 + \lambda \|\hat{X}\|_*$
Compressive sensing	Donoho (2004); Candes (2006)	$\min(\|\beta\|_1)$ subject to $y = X\beta$
Multivariate methods	Jolliffe *et al.* (2003); Witten *et al.* (2009)	Sparse principal components analysis, linear discriminant analysis and canonical correlation analysis

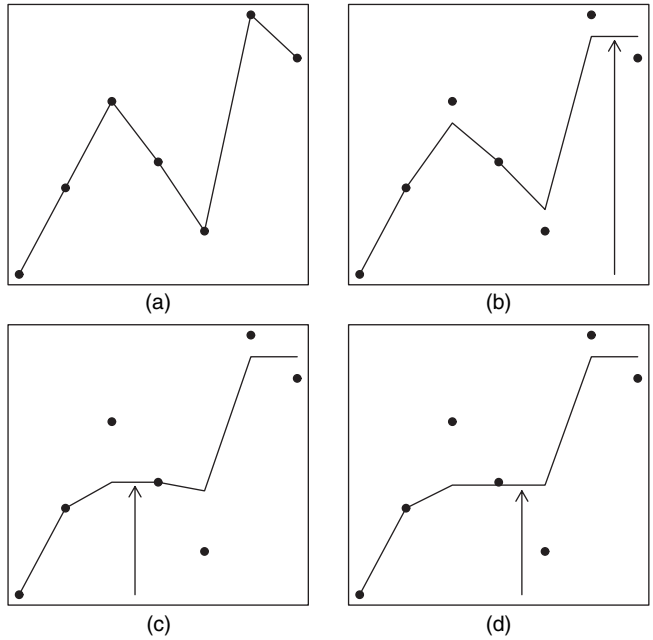

Fig. 1. Illustration of nearly isotonic fits for a toy example: (a) interpolating function ($\lambda = 0$) and three joining events (↑), (b) ($\lambda = 0.25$), (c) ($\lambda = 0.7$) and (d) ($\lambda = 0.77$), with the usual isotonic regression

$$\sum (y_i - \hat{y}_i)^2 \qquad \text{subject to } \hat{y}_1 \leqslant \hat{y}_2 \ldots.$$

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 algorithm (e.g. Barlow *et al.* (1972)). In nearly isotonic regression we minimize, with respect to β,

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

with x_+ indicating the positive part, $x_+ = x \cdot \mathbf{1}(x > 0)$. 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. $(\beta_i - \beta_{i+1})_+$ is 'half' of an l_1-penalty on differences, penalizing dips but not increases in the sequence. This procedure allows us to assess the assumption of monotonicity by comparing nearly monotone approximations with the best monotone approximation. Tibshirani *et al.* (2011) have provided a simple algorithm that computes the entire path of solutions, which is a kind of modified version of the pooled adjacent violators procedure. They also showed that the number of degrees of freedom is the number of unique values of $\hat{y}_i$ in the solution, using results from Tibshirani and Taylor (2011).

5. Discussion

Lasso (l_1-)penalties are useful for fitting a wide variety of models. Newly developed computational algorithms allow application of these models to large data sets, 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 an enjoyable 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.

Acknowledgements

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 thank the Research Section of the Royal Statistical Society for inviting me to present this retrospective paper.

References

Barlow, R. E., Bartholomew, D., Bremner, J. M. and Brunk, H. D. (1972) *Statistical Inference under Order Restrictions; the Theory and Applications of Isotonic Regression*. New York: Wiley.

Breiman, L. (1995) Better subset selection using the non-negative garotte. *Technometrics*, **37**, 738–754.

Candes, E. (2006) Compressive sampling. In *Proc. Int. Congr. Mathematicians, Madrid.* (Available from `www.acm.caltech.edu/emmanuel/papers/CompressiveSampling.pdf`.)

Candes, E. and Tao, T. (2007) The dantzig selector statistical estimation when p is much larger than n. *Ann. Statist.*, **35**, 2313–2351.

Candès, E. J. and Tao, T. (2009) The power of convex relaxation: near-optimal matrix completion. Stanford University, Stanford. (Available from `http://www.citebase.org/abstract?id=oai:arXiv.org:0903.1476`.)

Chen, S. S., Donoho, D. L. and Saunders, M. A. (1998) Atomic decomposition by basis pursuit. *SIAM J. Scient. Comput.*, **43**, 33–61.

Donoho, D. (2004) Compressed sensing. *Technical Report*. Department of Statistics, Stanford University, Stanford. (Available from `www.stat.stanford.edu/donoho/Reports/2004/CompressedSensing091604.pdf`.)

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2002) Least angle regression. *Technical Report.* Stanford University, Stanford.

Frank, I. and Friedman, J. (1993) A statistical view of some chemometrics regression tools (with discussion). *Technometrics*, **35**, 109–148.

Friedman, J., Hastie, T., Hoefling, H. and Tibshirani, R. (2007) Pathwise coordinate optimization. *Ann. Appl. Statist.*, **2**, 302–332.

Friedman, J., Hastie, T. and Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. *J. Statist. Sofwr.*, **33**, no. 1.

Fu, W. (1998) Penalized regressions: the bridge vs. the lasso. *J. Computnl Graph. Statist.*, **7**, 397–416.
Genkin, A., Lewis, D. and Madigan, D. (2007) Large-scale Bayesian logistic regression for text categorization. *Technometrics*, **49**, 291–304.
Hastie, T., Tibshirani, R. and Friedman, J. (2008) *The Elements of Statistical Learning; Data Mining, Inference and Prediction*, 2nd edn. New York: Springer.
Jolliffe, I. T., Trendafilov, N. T. and Uddin, M. (2003) A modified principal.component technique based on the lasso. *J. Computnl Graph. Statist.*, **12**, 531–547.
Mazumder, R., Hastie, T. and Tibshirani, R. (2010) Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.*, **11**, 2287–2322.
Osborne, M., Presnell, B. and Turlach, B. (2000) On the lasso and its dual. *J. Computnl Graph. Statist.*, **9**, 319–337.
Tibshirani, R., Hoefling, H. and Tibshirani, R. (2011) Nearly isotonic regression. *Technometrics*, **53**, 54–61.
Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005) Sparsity and smoothness via the fused lasso. *J. R. Statist. Soc.* B, **67**, 91–108.
Tibshirani, R. and Taylor, J. (2011) The solution path of the generalized lasso. *Ann. Statist.*, to be published.
Witten, D., Tibshirani, R. and Hastie, T. (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biometrika*, **10**, 515–534.
Wu, T. and Lange, K. (2008) Coordinate descent procedures for lasso penalized regression. *Ann. Appl. Statist.*, **2**, 224–244.
Yuan, M. and Lin, Y. (2007a) Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc.* B, **68**, 49–67.
Yuan, M. and Lin, Y. (2007b) Model selection and estimation in the Gaussian graphical model. *Biometrika*, **94**, 19–35.
Zou, H. (2006) The adaptive lasso and its oracle properties. *J. Am. Statist. Ass.*, **101**, 1418–1429.
Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *J. R. Statist. Soc.* B, **67**, 301–320.

Comments on the presentation

Peter Bühlmann (*Eidgenössiche Technische Hochschule, Zürich*)

I congratulate Rob Tibshirani for his excellent retrospective view of the lasso. It is of great interest to the whole community in statistics (and beyond), ranging from methodology and computation to applications: nice to read and of wide appeal!

The original paper (Tibshirani, 1996) has had an enormous influence. Fig. 2 shows that its frequency of citation continues to be in the exponential growth regime, together with the false discovery rate paper from Benjamini and Hochberg (1995): both of these works are crucial for high dimensional statistical inference.

The lasso was a real achievement 15 years ago: it enabled estimation and variable selection simultaneously in one stage, in the non-orthogonal setting. The novelty has been the second 's' in lasso (least absolute shrinkage and selection operator). More recently, progress has been made in understanding the selection property of the lasso.

Consider a potentially high dimensional linear model: $Y = \mathbf{X}\beta_0 + \varepsilon$ ($p \gg n$), with active set $S_0 = \{j; \beta_{0,j} \neq 0\}$ and sparsity index $s_0 = |S_0|$. The evolution of theory looks roughly as follows (to simplify, I use an asymptotic formulation where the dimension can be thought of as $p = p_n \gg n$ as $n \to \infty$, but, in fact, most of the developed theory is non-asymptotic). It requires about 15 lines of proof to show that, under *no conditions* on the design $\mathbf{X}$ (assuming a fixed design) and rather mild assumptions on the error,

$$\|\mathbf{X}(\hat{\beta} - \beta_0)\|_2^2/n \leqslant \|\beta_0\|_1 \, O_P[\sqrt{\{\log(p)/n\}}];$$

see Bühlmann and van de Geer (2011), chapter 6, which essentially recovers an early result by Greenshtein and Ritov (2004). Hence, the lasso is consistent for prediction if the regression vector is sparse in the l_1-norm $\|\beta_0\|_1 = o[\sqrt{\{n/\log(p)\}}]$. Achieving an optimal rate of convergence for prediction and estimation of the parameter vector requires a design condition such as restricted eigenvalue assumptions (Bickel *et al.*, 2009) or the slightly weaker compatibility condition (van de Geer, 2007; van de Geer and Bühlmann, 2009). Denoting by ϕ_0^2 such a restricted eigenvalue or compatibility constant (which we assume to be bounded away from zero),

$$\begin{aligned} &\|\mathbf{X}(\hat{\beta} - \beta_0)\|_2^2/n \leqslant s_0/\phi_0^2 \, O_P\{\log(p)/n\}, \\ &\|\hat{\beta} - \beta_0\|_q \leqslant s_0^{1/q}/\phi_0^2 \, O_P[\sqrt{\{\log(p)/n\}}], \qquad q \in \{1, 2\}; \end{aligned} \tag{1}$$

see Donoho *et al.* (2006), Bunea *et al.* (2007), van de Geer (2008) and Bickel *et al.* (2009). Finally, for recovering the active set S_0, such that $\mathbb{P}(\hat{S} = S_0)$ is large, tending to 1 as $p \gg n \to \infty$, we need quite

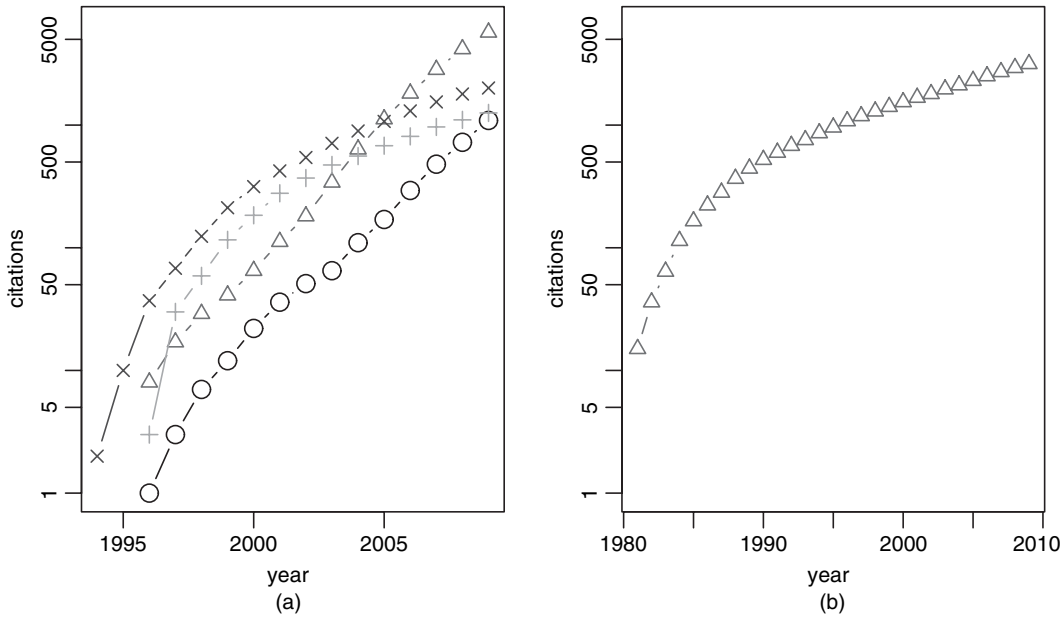

Fig. 2. Cumulative citation counts (on a log-scale) from the Thomson ISI *Web of Knowledge* (the largest abscissa on the *x*-axis corresponds to August 31st, 2010): (a) the lasso (○) (Tibshirani, 1996), false discovery rate (Δ) (Benjamini and Hochberg, 1995), reversible jump Markov chain Monte Carlo sampling (+) (Green, 1995) and wavelet shrinkage (×) (Donoho and Johnstone, 1994), published between 1994 and 1996; (b) the bootstrap (Δ) (Efron, 1979), published earlier

restrictive assumptions which are sufficient and (essentially) *necessary*: the neighbourhood stability condition for $\mathbf{X}$ (Meinshausen and Bühlmann, 2006), which is equivalent to the irrepresentable condition (Zhao and Yu, 2006; Zou, 2006), and a 'beta-min' condition

$$\min_{j \in S_0} |\beta_{0,j}| \geqslant C s_0^{1/2} / \phi_0^2 \sqrt{\{\log(p)/n\}}$$

requiring that the non-zero coefficients are not too small. Both of these conditions are restrictive and rather unlikely to hold in practice! However, it is straightforward to show from the second inequality in expression (l) that

$$\hat{S} \supseteq S_{\mathrm{relev}}, \qquad S_{\mathrm{relev}} = \left\{ j; |\beta_{0,j}| > C \frac{s_0^{1/2}}{\phi_0^2} \sqrt{\left\{ \frac{\log(p)}{n} \right\}} \right\}$$

holds with high probability. The underlying assumption is again a restricted eigenvalue condition on the design: in sparse problems, it is not overly restrictive; see van de Geer and Bühlmann (2009) and Bühlmann and van de Geer (2011) (corollary 6.8). Furthermore, if the beta-min condition holds, then the true active set $S_0 = S_{\mathrm{relev}}$ and we obtain the variable screening property

$$\hat{S} \supseteq S_0 \qquad \text{with high probability.}$$

Regarding the choice of the regularization parameter, we typically use $\hat{\lambda}_{\mathrm{CV}}$ from cross-validation. 'Luckily', empirical and some theoretical indications support that $\hat{S}(\hat{\lambda}_{\mathrm{CV}}) \supseteq S_0$ (or $\hat{S}(\hat{\lambda}_{\mathrm{CV}}) \supseteq S_{\mathrm{relev}}$): this is the relevant property in practice! The lasso is doing variable screening and, hence, I suggest that we interpret the second 's' in lasso as 'screening' rather than 'selection'.

Once we have the screening property, the task is to remove the false positive selections. Two-stage procedures such as the adaptive lasso (Zou, 2006) or the relaxed lasso (Meinshausen, 2007) are very useful. Recently, we have developed methods to control some type I (multiple-testing) error rates, guarding against false positive selections: stability selection (Meinshausen and Bühlmann, 2010) is based on resampling or subsampling for very general problems, and related multiple sample splitting procedures yield *p*-values in high dimensional linear or generalized linear models (Meinshausen *et al.*, 2009).

These resampling techniques are feasible since computation is efficient: as pointed out by Rob, (block) co-ordinatewise algorithms are often extremely fast. Besides Fu (1998), the idea was transferred to statistics (among others) by Paul Tseng, Werner Stuetzle and Sylvain Sardy (who was a former doctoral student of Stuetzle); see Sardy *et al.* (2000) or Sardy and Tseng (2004). A key work is Tseng (2001), and also Tseng and Yun (2009) is crucial for extending the computation to for example group lasso problems for the non-Gaussian, generalized linear model case (Meier *et al.*, 2008).

The issue of assigning uncertainty and variability in high dimensional statistical inference deserves further research. For example, questions about power are largely unanswered. Rob Tibshirani laid out very nicely the various extensions and possibilities when applying convex penalization to regularize empirical risk corresponding to a convex loss function. There is some work arguing why concave penalties have advantages (Fan and Lv, 2001; Zhang, 2010): the latter reference comes up with interesting properties about local minima. The issue of non-convexity is often more severe if the loss function (e.g. negative log-likelihood) is non-convex. Applying a convex penalty to such problems is still useful, yet more challenging in terms of computation and understanding the theoretical phenomena: potential applications are mixture regression models (Khalili and Chen, 2007; Städler *et al.*, 2011), linear mixed effects models (Bondell *et al.*, 2010; Schelldorfer *et al.*, 2011) or missing data problems (Allen and Tibshirani, 2010; Städler and Bühlmann, 2009). The beauty of convex optimization and convex analysis is (partially) lost and further research in this direction seems worthwhile.

The lasso, which was invented by Rob Tibshirani, has stimulated and continues to stimulate exciting research: it is a true success! It is my great pleasure to propose the vote of thanks.

Chris Holmes (*University of Oxford*)

It is both an honour and a great pleasure to have been invited to second the vote of thanks on this ground breaking paper that is as relevant today as it was 14 years ago at the time of publication, although the tradition of the Society for the seconder to offer a more critical appraisal makes this challenging to say the least. There can be few papers which have had such a marked influence on our way of thinking about regression analysis and parameter estimation; and it is one of a select handful of papers that have strongly influenced both Bayesian and non-Bayesian statistics.

All this is even more remarkable in view of, or perhaps due to, the simple structure of the lasso estimator,

$$\hat{\beta} = \arg\max_{\beta} \left\{ l(\beta) - \lambda \sum_j |\beta_j|^q \right\} \tag{2}$$

where $l(\beta)$ has the form of a log-likelihood recording fidelity to the data and, with $q = 1$, the lasso penalty encodes *a priori* beliefs about the nature of the unknown regression coefficients. Written in this way we can see that one interpretation of the lasso is as a Bayesian maximum *a posteriori* (MAP) estimate under a double-exponential prior on β. I believe that it is instructive to note an alternative representation of the lasso estimator in the form of a generalized ridge regression (see for example Holmes and Pintore (2007)), where

$$\{\tilde{\beta}, \tilde{\eta}\} = \arg\max_{\beta,\eta} \left\{ l(\beta) - \sum_j \eta_j^{-1} \beta_j^2 \right\} \qquad \text{subject to } \sum_j \eta_j = t, \tag{3}$$

leading to $\hat{\beta} \equiv \tilde{\beta}$ for some t a monotone function of λ. This shows the lasso as a Bayesian MAP estimate under a normal prior on β with individual variance components constrained to sum to some constant t. This sheds light on the essential difference between ridge regression and the lasso. Whereas in ridge regression the variance components are equal and fixed at some constant, in the lasso the variance components can be apportioned in a data-adaptive manner to maximize the likelihood. For variables that are non-relevant to the regression we find $\eta_j \to 0$ and hence $\tilde{\beta}_j = 0$ is 'shrunk' to 0 by the corresponding normal prior which then allows for more variance, and hence less shrinkage, to be placed on the important predictors.

From a frequentist perspective this interpretation also highlights a potential weakness of the lasso when estimating sparse signals. As the sample size becomes large, $n \to \infty$ and (if you are being frequentist) you would clearly wish for some $\eta_j \to 0$ (to achieve sparsity) but also for those predictors that are relevant to the regression to have $\eta_j \to \infty$ (to remove bias), to achieve oracle properties. But the lasso does not give you this freedom without setting $t \to \infty$. In terms of the MAP interpretation of expression (2) we see that the tails of the double-exponential prior are too light. This has led to considerable recent research investigating other forms of penalty that allow for oracle properties (Zou, 2006; Zou and Li, 2008; Candès *et al.*, 2008). However, improved properties of the corresponding estimators come at the expense of computational tractability.

The interpretation of the lasso as an MAP estimate also highlights another issue from a Bayesian perspective—although I fully appreciate that Professor Tibshirani does not refer to, or think of, the lasso as a Bayesian MAP estimator. From the Bayesian perspective, the use or reporting of a parameter estimate must follow from some decision process under an appropriate loss or scoring rule (Bernardo and Smith, 1994). It is then interesting to ask for what class of problems, decisions or loss functions is the lasso the appropriate estimate to use? In other words, *if the answer is the 'lasso', what is the question?* It turns out that there is no natural decision process under which the appropriate answer is to use an MAP estimate. To see this we note that reporting or using the MAP estimate only arises under a 0–1 loss which for continuous parameters is contrived:

$$\text{Loss}(a, \beta) - 1 - \mathbf{1}_{B_\varepsilon(a))}(\beta)$$

where $B_\varepsilon(a)$ is a ball of radius $\varepsilon \to 0$ in Ω_β centred at a Should this worry us at all?: yes if you adhere to the Bayesian school of statistics. It highlights the fact that conditioning on the estimates of a set of unknown parameters is rarely justified from a decision theory perspective and masks a key component of uncertainty in the regression analysis. In the Bayesian approach the reporting of models, predictions and parameter values all typically follow from a process of marginalzation over any unknowns to provide posterior distributions on parameters, models etc. that accommodate the uncertainty arising from finite data. One direct effect of this masking of uncertainty in modern applications, for instance in genomics, is when the nature of the processes leads to highly dependent predictor variables sometimes with cryptic (almost malicious) associations. In the presence of strong correlations between predictors with differing effect sizes, frequentist sparsity approaches, including the lasso, will tend to select a single variable within a group of collinear predictors, discarding the others in the pursuit of sparsity. However, the choice of the particular predictor might be highly variable and by selecting one we may ignore weak (but important) predictors which are highly correlated with stronger predictors. Recent methods have sought to address this in innovative ways (see, for example, Yuan and Lin (2006) and Meinshausen and Bühlmann (2010)), but the Bayesian approach naturally accommodates this uncertainty through reporting of joint distributions on predictor inclusion–exclusion that quantifies the dependence between dependent predictors; see for example, Clyde and George (2004) and Park and Cassella (2008). It seems to me that if you are going to introduce subjective penalty measures then it is much simpler and more natural to do so within a Bayesian paradigm.

In summary, Tibshirani (1996) is a thought-provoking and landmark paper that continues to influence and shape current thinking in regression. It has generated many extensions including branches within domains such as classification, clustering and graphical modelling, and fields such as signal processing and econometrics. I look forward to seeing its continued influence for many years to come.

Author's response

I thank Professor Bühlmann and Professor Holmes for their kind words, both in person and in print. Since this is a retrospective paper published over 10 years ago, it is not surprising that the contributed discussions are uncharacteristically 'tame' for a Royal Statistical Society discussion paper. Hence my rejoinder will be brief.

In my original paper I mentioned that the lasso solution can be viewed as the Bayesian maximum *a posteriori* estimate when the parameters are *a priori* independent, each having a Laplacian (double-exponential) prior distribution. However, the lasso solution is not the posterior mean or median in that setting: these latter solutions will not be sparse. Professor Holmes presents an alternative way to view the lasso as a Bayesian maximum *a posteriori* estimate. His model has a Gaussian prior with unequal variances η_j for each predictor j, and a constraint $\Sigma \eta_j = t$. It interesting that the solution to this problem gives the lasso estimate. But I am not sure what to make of it: is this prior with a constraint something which a statistician might think is reasonable on its own?

Professor Holmes then criticizes maximum *a posteriori* estimates in general and goes on to recommend that a more standard, complete Bayesian analysis be used to assess the instability and uncertainty in the lasso solution due to correlation between features. In this regard he refers to the 'Bayesian lasso' of Park and Casella (2008), which computes the posterior mean and median estimates from the Gaussian regression model with Laplacian prior. But these estimates, as I mentioned above, are not sparse. It seems to me that, if you want to obtain sparse solutions from a standard Bayesian model, you need to specify a prior that puts some mass at zero. One example of this is the spike-and-slab model of George and McCulloch (1993). These kinds of approaches are interesting but lead to non-convex problems that are computationally daunting. And the problem of correlated features that was mentioned by Professor Holmes is an

important one, but I think that a better way to improve the lasso estimates can be through generalized penalties such as the group lasso (Ming and Lin, 2006) (which was mentioned by Professor Holmes), the elastic net (Zou and Hastie, 2005) or the covariance regularization via the 'scout' procedure (Witten and Tibshirani, 2009).

Professor Bühlmann has nicely summarized recent work by many people on the prediction and selection consistency properties of the lasso. He mentions the fact that the lasso is good at finding (asymptotically) a superset of the correct predictors, and that methods that produce even sparser models can be useful. I think that the adaptive lasso is a promising approach for this problem, but the relaxed lasso (which Professor Bühlmann also mentions) only adjusts the non-zero lasso coefficients and so does not generally prune the model. Another way to obtain sparser models is through non-convex penalties such as smoothly clipped absolute deviation (Fan and Li, 2005) or Sparsenet (Mazumder *et al.*, 2010).

Finally, I think that we need better tools for inference with the lasso and related methods. Professor Bühlmann mentions some promising work on multiple testing for lasso models. More basically, we need reliable ways to assess the sampling variability of the lasso estimates. Standard errors would be a start but, since the sampling distributions are mixtures (with some mass at zero), more refined summaries are needed. We might need to fall back on bootstrap methods for this purpose: it would be important to understand how best to apply these methods and to understand their properties.

References in the comments

Allen, G. and Tibshirani, R. (2010) Transposable regularized covariance models with an application to missing data imputation. *Ann. Appl. Statist.*, **4**, 764–790.

Benjamini, Y. and Hochberg, Y. (l995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B, **57**, 289–300.

Bernardo, J. and Smith, A. F. M. (1994) *Bayesian Theory*. Chichester: Wiley.

Bickel, P., Ritov, Y. and Tsybakov, A. (2009) Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, **37**, 1705–1732.

Bondell, H., Krishna, A. and Ghosh, S. (2010) Joint variable selection of fixed and random effects in linear mixed-effects models. *Biometrics*, **66**, in the press.

Bühlmann, P. and van de Geer S. (2011) *Statistics for High-dimensional Data: Methods, Theory and Applications*. New York: Springer. To be published.

Bunea, F., Tsybakov, A. and Wegkamp, M. (2007) Sparsity oracle inequalities for the Lasso. *Electron. J. Statist.*, **1**, 169–194.

Candès, E., Wakin, M. and Boyd, S. (2008) Enhancing sparsity by reweighted l1 minimization. *J. Four. Anal. Appl.*, **14**, 877–905.

Clyde, M. and George, E. (2004) Model uncertainty. *Statist. Sci.*, **19**, 81–94.

Donoho, D., Elad, M. and Temlyakov, V. (2006) Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theor.*, **52**, 6–18.

Donoho, D. and Johnstone, I. (1994) Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455.

Efron, B. (1979) Bootstrap methods: another look at the Jackknife. *Ann. Statist.*, **7**, 1–26.

Fan, J. and Li, R. (2005) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Ass.*, **96**, 1348–1360.

Fu, W. (1998) Penalized regressions: the Bridge versus the Lasso. *J. Computnl Graph. Statist.*, **7**, 397–416.

van de Geer, S. (2007) The deterministic lasso. In *Proc. Jt Statist. Meet.*, p. 140. Alexandria: American Statistical Association.

van de Geer, S. (2008) High-dimensional generalized linear models and the Lasso. *Ann. Statist.*, **36**, 614–645.

van de Geer, S. and Bühlmann, P. (2009) On the conditions used to prove oracle results for the Lasso. *Electron. J. Statist.*, **3**, 1360–1392.

George, E. and McCulloch, R. (1993) Variable selection via gibbs sampling. *J. Am. Statist. Ass.*, **88**, 884–889.

Green, P. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Greenshtein, E. and Ritov, Y. (2004) Persistence in high-dimensional predictor selection and the virtue of over-parametrization. *Bernoulli*, **10**, 971–988.

Holmes, C. C. and Pintore, A. (2007) Bayesian relaxation: boosting, the Lasso, and other L_α norms. In *Bayesian Statistics 8* (eds J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West). Oxford: Oxford University Press.

Khalili, A. and Chen, J. (2007) Variable selection in finite mixture of regression models. *J. Am. Statist. Ass.*, **102**, 1025–1038.

Mazumder, R., Friedman, J. and Hastie, T. (2010) Sparsenet: coordinate descent with non-convex penalties. Stanford University, Stanford.

Meier, L., van de Geer, S. and Bühlmann P. (2008) The group lasso for logistic regression. *J. R. Statist. Soc.* B, **70**, 53–71.
Meinshausen, N. (2007) Relaxed Lasso. *Computnl Statist. Data Anal.*, **52**, 374–393.
Meinshausen, N. and Bühlmann, P. (2006) High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, **34**, 1436–1462.
Meinshausen, N. and Bühlmann, P. (2010) Stability selection (with discussion). *J. R. Statist. Soc.* B, **72**, 417–473.
Meinshausen, N., Meier, L. and Bühlmann, P. (2009) P-values for high-dimensional regression. *J. Am. Statist. Ass.*, **104**, 1671–1681.
Ming, Y. and Lin, Y. (2006) Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc.* B, **68**, 49–67.
Park, T. and Cassella, G. (2008) The Bayesian Lasso. *J. Am. Statist. Ass.*, **103**, 681–686.
Sardy, S., Bruce, A. and Tseng, P. (2000) Block coordinate relaxation methods for nonparametric wavelet denoising. *J. Computnl Graph. Statist.*, **9**, 361–379.
Sardy, S. and Tseng, P. (2004) On the statistical analysis of smoothing by maximizing dirty Markov random field posterior distributions. *J. Am. Statist. Ass.*, **99**, 191–204.
Schelldorfer, J., Bühlmann, P. and van de Geer, S. (2011) Estimation for high-dimensional linear mixed-effects models using l_1-penalization. *Scand. J. Statist.*, to be published.
Städler, N. and Bühlmann, P. (2011) Missing values: sparse inverse covariance estimation and an extension to sparse regression. *Statist. Comput.*, to be published.
Städler, N., Bühlmann, P. and van de Geer, S. (2010) l_1-penalization for mixture regression models (with discussion). *Test*, **19**, 209–285.
Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.
Tseng, P. (2001) Convergence of a block coordinate descent method for nonsmooth separable minimization. *J. Optimzn Theor. Appl.*, **109**, 475–494.
Tseng, P. and Yun, S. (2009) A coordinate gradient descent method for nonsmooth separable minimization. *Math. Programing* B, **117**, 387–423.
Witten, D. M. and Tibshirani, R. (2009) Covariance-regularized regression and classification for high dimensional problems. *J. R. Statist. Soc.* B, **71**, 615–636.
Yuan, M. and Lin, Y. (2006) Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc.* B, **68**, 49–67.
Zhang, C.-H. (2010) Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, **38**, 894–942.
Zhao, P. and Yu, B. (2006) On model selection consistency of Lasso. *J. Mach. Learn. Res.*, **7**, 2541–2563.
Zou, H. (2006) The adaptive Lasso and its oracle properties. *J. Am. Statist. Ass.*, **101**, 1418–1429.
Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *J. R. Statist. Soc.* B, **67**, 301–320.
Zou, H. and Li, R. (2008) One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.*, **36**, 1509–1533.