Proposing the vote of thanks: Regression shrinkage and selection via the Lasso: a retrospective by Robert Tibshirani

Peter Bühlmann

Seminar für Statistik, ETH Zürich, CH-8092 Zürich, Switzerland

I congratulate Rob Tibshirani for his excellent retrospective view on 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 Lasso paper (Tibshirani, 1996) has an enormous impact. Figure 1 shows that its citation frequency 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.

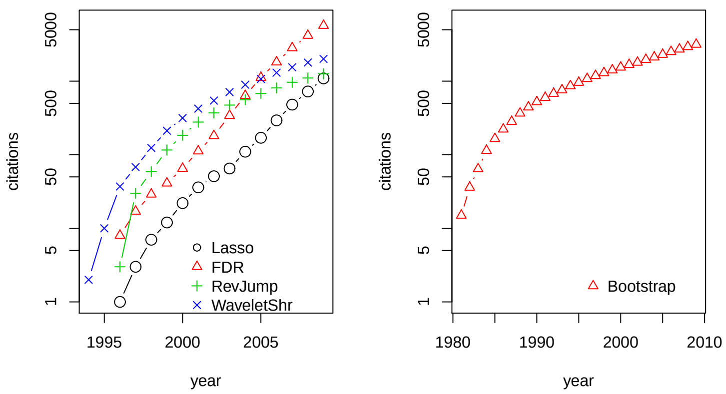

Fig. 1. Cumulative citation counts (y-axis with log-scale) from ISI Web of Knowledge (largest abscissa on x-axis corresponds to August 31, 2010). Left: Lasso (Tibshirani, 1996), False discovery rate (Benjamini and Hochberg, 1995), Reversible jump MCMC (Green, 1995), Wavelet shrinkage (Donoho and Johnstone, 1994), published between 1994 and 1996. Right: Bootstrap (Efron, 1979), published earlier.

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 **S**election Operator). More recently, progress has been made in understanding the selection property of 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 as $p = p_n \gg n$ as $n \to \infty$; but in fact, most of the developed theory is non-asymptotic). It is about 15 lines of proof to show that under *no conditions* on the design $\mathbf{X}$ (assuming fixed design) and rather mild assumptions on the error:

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

cf. Bühlmann and van de Geer (2011, Ch.6), which essentially recovers an early result by Greenshtein and Ritov (2004). And 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 convergence rate 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 (which we assume to be bounded away from zero):

$$\begin{aligned} &\|\mathbf{X}(\hat{\beta} - \beta_0)\|_2^2/n \le s_0/\phi_0^2 O_P(\log(p)/n), \\ &\|\hat{\beta} - \beta_0\|_1 \le s_0/\phi_0^2 O_P(\sqrt{\log(p)/n}), \end{aligned} \qquad (1)$$

cf. 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 one as $p \gg n \to \infty$, we need rather restrictive assumptions which are sufficient and (essentially) *necessary*: the neighborhood 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 $\min_{j \in S_0} |\beta_{0,j}| \ge C s_0/\phi_0^2 \sqrt{\log(p)/n}$ 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 (1) that

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

holds with high probability. The underlying assumption is again a restricted eigenvalue condition on the design: in sparse problems, it is not overly restrictive (van de Geer and Bühlmann, 2009; Bühlmann and van de Geer, 2011)[Cor.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:

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

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

Once we have the screening property, the task is to get rid of 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 re- or sub-sampling for very general problems, and related multi sample-splitting procedures yield p-values in high-dimensional linear or generalised linear models (Meinshausen et al., 2009).

These re-sampling techniques are feasible since computation is efficient: as pointed out by Rob, (block-) coordinatewise algorithms are often extremely fast. Besides Fu (1998), the idea was transferred to statistics (among others) by Paul Tseng, Werner Stuetzle and Sylvain Sardy (former PhD student of Stuetzle), cf. Sardy et al. (2000) or Sardy and Tseng (2004). A key work is from Tseng (2001), and also Tseng and Yun (2009) is crucial for extending the computation to e.g. 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 penalisation to regularise 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., 2010), linear mixed-effects models (Bondell et al., 2010; Schelldorfer et al., 2010) or missing data problems (Allen and Tibshirani, 2010; Städler and Bühlmann, 2009). The beauty of convex optimisation and convex analysis is (partially) lost and further research in this direction seems worthwhile.

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

References

Allen, G. and R. Tibshirani (2010). Transposable regularized covariance models with an application to missing data imputation. *The Annals of Applied Statistics 4*, 764–790.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B 57*, 289–300.

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

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

Bühlmann, P. and S. van de Geer (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer. To appear.

Bunea, F., A. Tsybakov, and M. Wegkamp (2007). Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics 1*, 169–194.

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

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

Efron, B. (1979). Bootstrap methods: another look at the Jackknife. *The Annals of Statistics 7*, 1–26.

Fan, J. and J. Lv (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*, 1348–1360.

Fu, W. (1998). Penalized regressions: The Bridge versus the Lasso. *Journal of Computational and Graphical Statistics 7*, 397–416.

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

Greenshtein, E. and Y. Ritov (2004). Persistence in high-dimensional predictor selection and the virtue of over-parametrization. *Bernoulli 10*(6), 971–988.

Khalili, A. and J. Chen (2007). Variable selection in finite mixture of regression models. *Journal of the American Statistical Association 102*, 1025–1038.

Meier, L., S. van de Geer, and P. Bühlmann (2008). The Group Lasso for logistic regression. *Journal of the Royal Statistical Society Series B 70*, 53–71.

Meinshausen, N. (2007). Relaxed Lasso. *Computational Statistics & Data Analysis 52*, 374–393.

Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics 34*, 1436–1462.

Meinshausen, N. and P. Bühlmann (2010). Stability selection (with discussion). *Journal of the Royal Statistical Society Series B 72*, 417–473.

Meinshausen, N., L. Meier, and P. Bühlmann (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association 104*, 1671–1681.

Sardy, S., A. Bruce, and P. Tseng (2000). Block coordinate relaxation methods for nonparametric wavelet denoising. *Journal of Computational and Graphical Statistics, 9*, 361–379.

Sardy, S. and P. Tseng (2004). On the statistical analysis of smoothing by maximizing dirty Markov random field posterior distributions. *Journal of the American Statistical Association 99*, 191–204.

Schelldorfer, J., P. Bühlmann, and S. van de Geer (2010). Estimation for high-dimensional linear mixed-effects models using l1-penalization. *arXiv:1002.3784v1* , 1–19.

Städler, N. and P. Bühlmann (2009). Missing values: sparse inverse covariance estimation and an extension to sparse regression. *arXiv:0903.5463v2* , 1–25.

Städler, N., P. Bühlmann, and S. van de Geer (2010). ℓ_1-penalization for mixture regression models (with discussion). *Test 19*, 209–285.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B 58*, 267–288.

Tseng, P. (2001). Convergence of a block coordinate descent method for nonsmooth separable minimization. *Journal of Optimization Theory and Applications 109*, 475–494.

Tseng, P. and S. Yun (2009). A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming, Series B 117*, 387–423.

van de Geer, S. (2007). The deterministic lasso. In *JSM proceedings, 2007, 140*. American Statitistical Association.

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

van de Geer, S. and P. Bühlmann (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics 3*, 1360–1392.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics 38*, 894–942.

Zhao, P. and B. Yu (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research 7*, 2541–2563.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association 101*, 1418–1429.