5

Lasso and Sparsity in Statistics

Robert J. Tibshirani

Stanford University, Stanford, CA

In this chapter, I discuss the lasso and sparsity, in the area of supervised learning that has been the focus of my research and that of many other statisticians. This area can be described as follows. Many statistical problems involve modeling important variables or outcomes as functions of predictor variables. One objective is to produce models that allow predictions of outcomes that are as accurate as possible. Another is to develop an understanding of which variables in a set of potential predictors are strongly related to the outcome variable. For example, the outcome of interest might be the price of a company's stock in a week's time, and potential predictor variables might include information about the company, the sector of the economy it operates in, recent fluctuations in the stock's price, and other economic factors. With technological development and the advent of huge amounts of data, we are frequently faced with very large numbers of potential predictor variables.

5.1 Sparsity, ℓ_1 Penalties and the Lasso

The most basic statistical method for what is called supervised learning relates an outcome variable Y to a linear predictor variables $x_1, \ldots, x_p$, viz.

$$Y = \beta_0 + \sum_{j=1}^{p} x_j \beta_j + \epsilon, \tag{5.1}$$

where ϵ is an error term that represents the fact that knowing $x_1, \ldots, x_p$ does not normally tell us exactly what Y will be. We often refer to the right-hand side of (5.1), minus ϵ, as the predicted outcome. These are referred to as linear regression models. If we have data on the outcome y_i and the predictor variables x_{ij} for each in a group of N individuals or scenarios, the method of least squares chooses a model by minimizing the sum of squared errors between

the outcome and the predicted outcome, over the parameters (or regression coefficients) $\beta_0, \ldots, \beta_p$.

Linear regression is one of the oldest and most useful tools for data analysis. It provides a simple yet powerful method for modeling the effect of a set of predictors (or features) on an outcome variable. With a moderate or large number of predictors, we don't typically want to include all the predictors in the model. Hence one major challenge in regression is variable selection: choosing the most informative predictors to use in the model. Traditional variable selection methods search through all combinations of predictors and take too long to compute when the number of predictors is roughly larger than 30; see, e.g., Chapter 3 of Hastie et al. (2008) for details.

Penalized regression methods facilitate the application of linear regression to large problems with many predictors. The lasso uses ℓ_1 or absolute value penalties for penalized regression. In particular, it provides a powerful method for doing variable selection with a large number of predictors. In the end it delivers a sparse solution, i.e., a set of estimated regression coefficients in which only a small number are non-zero. Sparsity is important both for predictive accuracy, and for interpretation of the final model.

Given a linear regression with predictors x_{ij} and response values y_i for $i = 1, \ldots, N$ and $j = 1, \ldots, p$, the lasso solves the ℓ_1-penalized regression so as to minimize

$$\frac{1}{2}\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|, \tag{5.2}$$

for the unknown parameters $\beta_0, \ldots, \beta_p$. The second term above is called a penalty function; it balances the fit of the model with its complexity. The non-negative parameter λ governs that tradeoff. The larger λ, the more sparse the final solution vector $\hat{\beta}$. The statistician chooses the value of λ, or uses a method like cross-validation, to estimate it.

The lasso problem (5.2) is equivalent to minimizing the sum of squares with constraint

$$\sum_{j=1}^{p}|\beta_j| \leq s.$$

For every λ in (5.2), there is a bound parameter s yielding the same solution. Note that choosing $\lambda = 0$ or equivalently a sufficiently large value of s, yields the usual least squares solution. Lasso regression is similar to ridge regression, which has constraint

$$\sum_{j=1}^{p}\beta_j^2 \leq s.$$

Because of the form of the ℓ_1 penalty, as shown in Figure 5.1, the lasso does variable selection and shrinkage, while ridge regression, in contrast, only

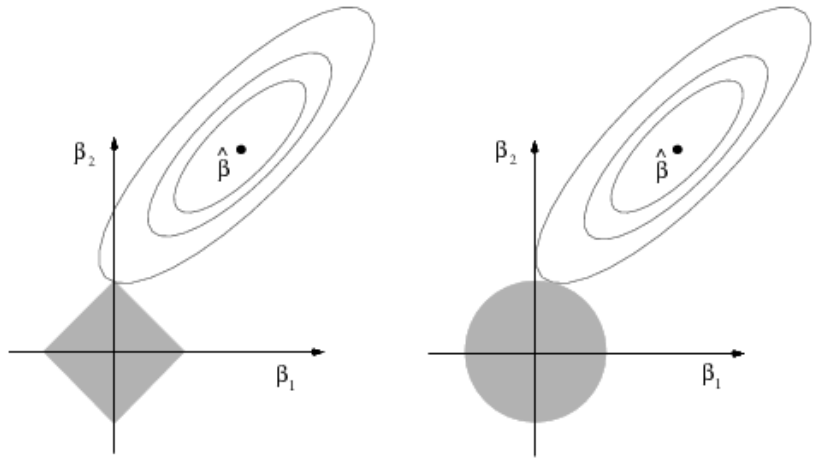

FIGURE 5.1: Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid gray areas are the constraint regions $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t^2$, respectively, while the ellipses are the contours of the least squares error function, centered at the full least squares estimates $\hat{\beta}$. The sharp corners of the constraint region for the lasso regression yield sparse solutions. In high dimensions, sparsity arises from corners and edges of the constraint region.

shrinks. If we consider a more general penalty of the form

$$\left(\sum_{j=1}^{p} \beta_j^q \right)^{1/q},$$

then the lasso uses $q = 1$ and ridge regression has $q = 2$. Subset selection emerges as $q \to 0$, and the lasso corresponds to the smallest value of q (i.e., closest to subset selection) that yields a convex problem. [A convex problem is an optimization of a convex function over a convex set. If a function is strictly convex, the problem is guaranteed to have a unique global solution.]

Figure 5.1 gives a geometric view of the lasso and ridge regression. Figure 5.2 shows an example. The outcome is the value of the log PSA (prostate-specific antigen) for men whose prostate was removed during cancer surgery, modeled as a function of eight measurements such as age, cancer volume, tumor weight, etc. The figure shows the profiles of the lasso coefficients as the shrinkage factor s is varied. This factor is the bound on the total norm $|\hat{\beta}_1| + \cdots + |\hat{\beta}_p|$, and we have scaled it to lie between 0 and 1 for interpretability. The vertical dotted line is the value of s chosen by cross-validation: it yields a model with just three nonzero coefficients, lcavol, svi, and lweight. Recall that s is in one-to-one correspondence to the tuning parameter λ in (5.2): thus λ

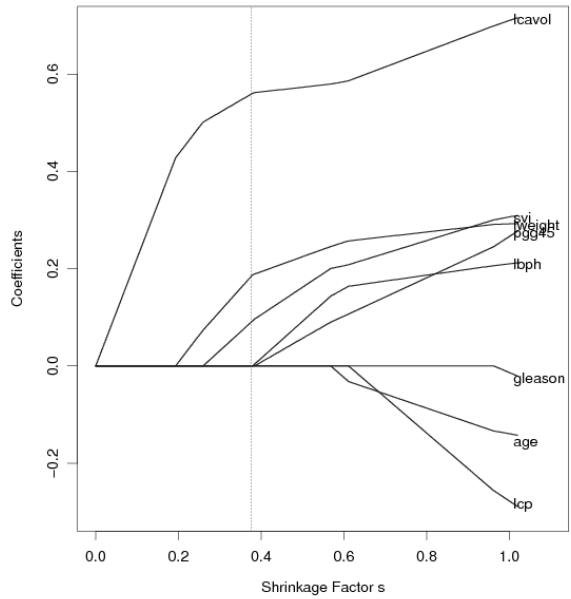

FIGURE 5.2: Profiles of the lasso coefficients for the prostate cancer example.

is large on the left of the plot forcing all estimates to be zero, and is zero on the right, yielding the least squares estimates.

5.2 Some Background

Lasso regression and ℓ_1 penalization have been the focus of a great deal of work in recent years. Table 5.1, adapted from Tibshirani (2011), gives examples of some this work.

My original lasso paper was motivated by an idea of Leo Breiman's called the garotte (Breiman, 1995). The garotte chooses $c_1, \ldots, c_p$ to minimize

$$\sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} c_j x_{ij} \hat{\beta}_j \right)^2 + \lambda \sum_{j=1}^{p} c_j, \tag{5.3}$$

where $\hat{\beta}_1, \ldots, \hat{\beta}_p$ are the usual least squares estimates and $c_j \geq 0$ for all $j \in \{1, \ldots, p\}$. Thus Leo's idea was to scale the least squares estimates by

TABLE 5.1: Some examples of generalizations of the lasso.

Method	Authors
Adaptive lasso	Zou (2006)
Compressive sensing	Donoho (2004), Candès (2006)
Dantzig selector	Candès and Tao (2007)
Elastic net	Zou and Hastie (2005)
Fused lasso	Tibshirani et al. (2005)
Generalized lasso	Tibshirani and Taylor (2011)
Graphical lasso	Yuan and Lin (2007b), Friedman et al. (2007)
Grouped lasso	Yuan and Lin (2007a)
Hierarchical interaction models	Bien et al. (2013)
Matrix completion	Candès and Tao (2009), Mazumder et al. (2010)
Multivariate methods	Joliffe et al. (2003), Witten et al. (2009)
Near-Isotonic regression	Tibshirani et al. (2011)

nonnegative constants, some of which might be zero. I noticed that the garotte wouldn't be defined for $p > N$, since the least squares estimates are not defined in that case. Hence I just simplified the method by removing the "middle man."

Not surprisingly, it turned out that using absolute value constraints in regression was not a completely new idea at the time. Around the same time, Chen, Donoho and Saunders proposed "Basis Pursuit" (Chen et al., 1998), which used absolute value constraints for signal processing. Earlier, Frank and Friedman (1993) had (briefly) discussed the "bridge" estimate, which proposed a family of penalties of the form $\sum |\beta_p|^q$ for some q.

The lasso paper was published in 1996, but did not get much attention at the time. This may have been in part the due to the relatively limited computational power that was available to the average statistician, and also the relatively small size of datasets (compared to today). Now the lasso and ℓ_1 constraint-based methods are a hot area, not only in statistics but in machine learning, engineering and computer science.

The original motivation for the lasso was for interpretability: it is an alternative to subset regression for obtaining a sparse (or parsimonious) model. In the past 20 years some unforeseen advantages of convex ℓ_1-penalized approaches emerged: statistical and computational efficiency.

On the statistical side, there has also been a great deal of interesting work on the mathematical aspects of the lasso, examining its ability recover a true underlying (sparse) model and to produce a model with minimal prediction error. Many researchers have contributed to this work, including Peter Bühlmann, Emmanuel Candès, David Donoho, Eitan Greenshtein, Iain Johnstone, Nicolai Meinshausen, Ya'acov Ritov, Martin Wainwright, Bin Yu, and many others. In describing some of this work, Hastie et al. (2001) coined the informal "Bet on Sparsity" principle. ℓ_1 methods assume that the truth is sparse, in some basis. If the assumption holds true, then the parameters can be efficiently estimated using ℓ_1 penalties. If the assumption does not hold

— so that the truth is dense — then no method will be able to recover the underlying model without a large amount of data per parameter. This is typically not the case when the number of predictors, p, is much larger than the sample size, N, a commonly occurring scenario.

On the computational side, the convexity of the problem and sparsity of the final solution can be used to great advantage. Parameters whose estimates are zero in the solution can be handled with minimal cost in the search. Powerful and scalable techniques for convex optimization can be applied to the problem, allowing the solution of very large problems. One particularly effective approach is coordinate descent (Fu, 1998; Friedman et al., 2007, 2010), a simple one-at-a-time method that is well-suited to the separable lasso penalty. This method is simple and flexible, and can also be applied to many other ℓ_1- penalized generalized linear models, including multinomial, Poisson and Cox's proportional hazards model for survival data. Coordinate descent is implemented in the `glmnet` package in the R statistical language, written by Jerome Friedman, Trevor Hastie, Noah Simon and myself.

Here is the basic idea of coordinate descent. Suppose we had only one predictor and wished to solve the lasso problem, i.e., minimize

$$\sum_{i=1}^{N}(y_i - x_i\beta)^2 + \lambda|\beta|.$$

Then the solution is easily shown to be the soft-thresholded estimate

$$\text{sign}(\hat{\beta})(|\hat{\beta}| - \lambda)_+,$$

where $\hat{\beta}$ is usual least squares estimate, and the $+$ indicates positive part. The idea then, with multiple predictors, is to cycle through each predictor in turn, solving for the each estimate, using this method. We compute residuals

$$r_i = y_i - \sum_{j\neq k} x_{ij}\hat{\beta}_k$$

and apply univariate soft-thresholding, pretending that our data is (x_{ij}, r_i). We cycle through the predictors $j = 1, \ldots, p$ several times until convergence. Coordinate descent is like skiing to the bottom of a hill: but rather than pointing your skis towards the bottom of the hill, you go as far down as you can in the north-south direction, then east-west, then north-south, etc., until you (hopefully) reach the bottom.

5.3 A History for Coordinate Descent for the Lasso

This history is interesting, and shows the haphazard way in which science sometimes progresses. In 1997 Weijiang Fu, a graduate student of mine at the

University of Toronto wrote his thesis on lasso-related topics, and proposed the "shooting method" for computation of the lasso estimates. I read and signed it, but in retrospect apparently didn't understand it very well. And I basically forgot about the work, which was later published. Then by 2002 I had moved to Stanford and Ingrid Daubechies — an applied mathematician at Stanford — discussed a theorem about a coordinate-wise method for computing solutions to convex problems. Trevor Hastie and I went to the talk, took notes, and then programmed the proposed method ourselves in the S language. We made a mistake in the implementation: trying to exploit the efficient vector operations in Splus, we changed each parameter not one-at-a-time, but at the end of each loop of p updates. This turned out to be a fatal mistake, as the method did not even converge and so we just "wrote it off."

Then in 2006 our colleague Jerry Friedman was the external examiner at the PhD oral of Anita van der Kooij (in Leiden) who used the coordinate descent idea for the elastic net, a generalization of the lasso. Friedman showed us the idea and together we wondered whether this method would work for the lasso. Jerome, Trevor and I started working on this problem, and using some clever implementation ideas by Friedman, we produced some very fast code (`glmnet` in the R language). It was only then that I realized that Weijiang Fu had the same basic idea almost 10 years earlier! Now coordinate descent is a considered a state-of-the-art method for the lasso — one of the best methods around — and remarkable in its simplicity.

For a long time, even the convex optimization community did not seem to take coordinate descent very seriously. For example, my good friend Stephen Boyd's standard book on the topic (Boyd and Vandenberghe, 2004) does not even mention it. The only work I could find on the coordinate descent for convex problems was that of Paul Tseng, a Canadian at the University of Washington who proved (in the 1980s) some beautiful results showing the convergence a coordinate descent for separable problems (Tseng, 1988). These include the lasso, as a special case. When the problem is not separable, coordinate descent may not converge: this may explain the lack of interest in the method in the convex optimization world.

I never met Paul, but we corresponded by email and he was happy that his work was proving to be so important for the lasso. Sadly, in 2009 he went missing while kayaking in the Yangtze River in China and is now presumed dead. His seminal work provides the underpinning for the application of coordinate descent in the lasso and many related problems.

5.4 An Example in Medicine

I am currently working on a cancer diagnosis project with coworkers at Stanford. They have collected samples of tissue from a number of patients undergo-

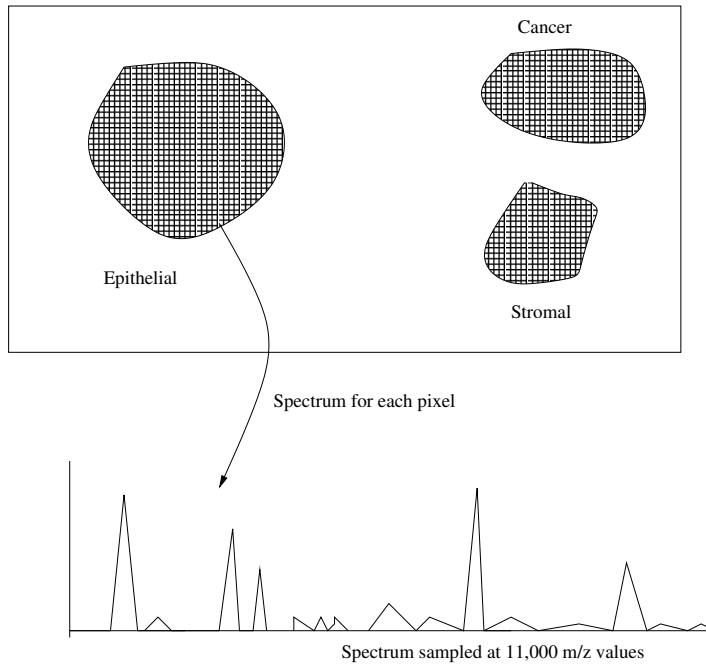

FIGURE 5.3: Schematic of the cancer diagnosis problem. Each pixel in each of the three regions labeled by the pathologist is analyzed by mass spectrometry. This gives a feature vector of 11,000 intensities for each pixel (bottom panel), from which we try to predict the class of that pixel.

ing surgery for cancer. We are working to build a classifier that can distinguish three kinds of tissue: normal epithelial, stromal, and cancer. Such a classifier could be used to assist surgeons in determining, in real time, whether they had successfully removed all of the tumor. It could also yield insights into the cancer process itself. The data are in the form of images, as sketched in Figure 5.3. A pathologist has labeled each region (and hence the pixels inside a region) as epithelial, stromal or cancer. At each pixel in the image, the intensity of metabolites is measured by a type of mass spectrometry, with the peaks in the spectrum representing different metabolites. The spectrum has been finely sampled, with the intensity measured at about $11,000$ sites (frequencies) across the spectrum. Thus, the task is to build a prediction model to classify each pixel into one of the three classes, based on the $11,000$ features. There are about 8000 pixels in all.

For this problem, I have applied an ℓ_1-regularized multinomial model. A multinomial model is one which predicts whether a tissue sample (pixel) is of type 1(epithelial), 2 (stromal) or 3 (cancer). For each class $k \in \{1, 2, 3\}$, the model has a vector of parameters $\beta_{1k}, \ldots, \beta_{pk}$, representing the weight given

to each feature in that class. I used the `glmnet` package for fitting the model: it computes the entire path of solutions for all values of the regularization parameter λ, using cross-validation to estimate the best value of λ (I left one patient out at a time). The entire computation required just a few minutes on a standard Linux server.

The results so far are promising. The classifier shows 90–97% accuracy in the three classes, using only around 100 features. This means that when the model is used to predict the tissue type of a pixel, it is correct 90–97% of the time. These features could yield insights about the metabolites that are important in stomach cancer. The power of this approach, both its prediction accuracy and interpretability, are not shared by competing methods such as support vector machines or decision trees. For example, this method is based on the multinomial probability model and so we obtain not only class predictions but estimated probabilities for each class (unlike support vector machines). Thus for example we can create a "I don't know" category, and assign a pixel to that category if the gap between the two largest class probabilities in small (say 10%). There is much more work to be done — collecting more data, and refining and testing the model on more difficult cases. But this shows the potential of ℓ_1-penalized models in an important and challenging scientific problem.

5.5 Nearly Isotonic Regression

Another recent example of the use of ℓ_1 constraints is nearly isotonic regression (Tibshirani et al., 2011). Unlike the regression problem, here we have no predictors but just a sequence of outcome values $y_1, \ldots, y_N$ which we wish to approximate. Given this sequence, the method of isotonic regression solves

$$\text{minimize} \sum (y_i - \hat{y}_i)^2 \quad \text{subject to } \hat{y}_1 \leq \hat{y}_2 \leq \cdots$$

This assumes a monotone non-decreasing approximation, with an analogous definition for the monotone non-increasing case. The solution can be computed via the well-known Pool Adjacent Violators (PAVA) algorithm; see, e.g., Barlow et al. (1972). In nearly isotonic regression we solve

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

with x_+ indicating the positive part, $x_+ = x\,\mathbf{1}(x > 0)$. The solutions $\hat{\beta}_i$ are the values $\hat{y}_i$ that we seek. 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. A toy example is given in Figure 5.4.

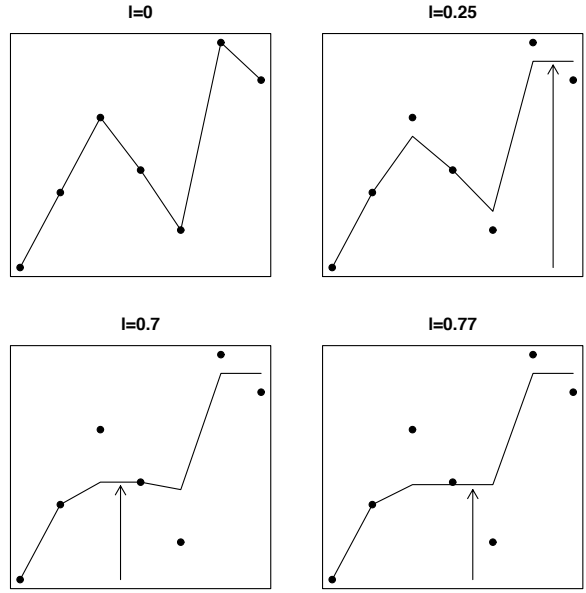

FIGURE 5.4: Illustration of nearly isotonic fits for a toy example. An interpolating function is shown in the top left panel. There are three joining events (indicated by the arrows) shown in the remaining panels, with the usual isotonic regression appearing in the bottom right panel.

Note that $(\beta_i - \beta_{i+1})_+$ is "half" of an ℓ_1 penalty on differences, penalizing dips but not increases in the sequence. This procedure allows one to assess the assumption of monotonicity by comparing nearly monotone approximations to the best monotone approximation. Tibshirani et al. (2011) provide a simple algorithm that computes the entire path of solutions, a kind of modified version of the PAVA procedure. They also show that the number of degrees of freedom is the number of unique values of $\hat{y}_i$ in the solution, using results from Tibshirani and Taylor (2010).

This kind of approach can be extended to higher order differences, and is also known as ℓ_1-trend filtering (Kim et al., 2009). For example a second-order difference penalty (without the positive part) yields a piecewise linear function estimate.

5.6 Conclusion

In this chapter I hope that I have conveyed my excitement for some recent developments in statistics, both in its theory and practice. These methods are already widely used in many areas, including business, finance and numerous scientific areas. The area of medical imaging may be greatly advanced by the advent of compressed sensing, a clever method based on ℓ_1 penalties (Candès and Tao, 2005; Donoho, 2006). I predict that sparsity and convex optimization will play an increasingly important role in the development of statistical methodology and in the applications of statistical methods to challenging problems in science and industry.

One particularly promising area is that of inference, where the covariance test recently proposed by Lockhart et al. (2014) provides a simple way to assess the significance of a predictor, while accounting for the adaptive nature of the fitting. In essence, the exponential distribution that arises in this new work is the analog of the usual chi-squared for the F-test for fixed (non-adaptive) regression. It appears that this new test will have broad applications in other problems such as principal components, clustering and graphical models. See Tibshirani (2014) for a brief overview.

About the Author

Robert J. Tibshirani is a professor of statistics, health research and policy at Stanford University; he was affiliated to the University of Toronto from 1985 to 1998. He received a BMath from the University of Waterloo, an MSc from the University of Toronto and a PhD from Stanford. His research interests include statistical theory, statistical learning, and a broad range of scientific areas. He received a Steacie Award, the 1996 COPSS Award, the 2000 CRM–SSC Award, and the 2012 SSC Gold Medal for research. He was elected to the Royal Society of Canada in 2001; he is a fellow of the American Statistical Association and the Institute of Mathematical Statistics.

Bibliography

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

Bien, J., Taylor, J., and Tibshirani, R. J. (2013). A lasso for hierarchical interactions. *The Annals of Statistics*, 41:1111–1141.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, Cambridge.

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

Candès, E. J. (2006). Compressive sampling. In *Proceedings of the International Congress of Mathematicians, Madrid, Spain*.

Candès, E. J. and Tao, T. (2005). Decoding by linear programming. *IEEE Transactions on Information Theory*, 51:4203–4215.

Candès, E. J. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35:2313–2351.

Candès, E. J. and Tao, T. (2009). The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56:2053–2080.

Chen, S., Donoho, D. L., and Saunders, M. (1998). Atomic decomposition for basis pursuit. *SIAM Journal on Scientific Computing*, 20:33–61.

Donoho, D. L. (2004). *Compressed Sensing*. Technical Report, Statistics Department, Stanford University, Stanford, CA.

Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions for Information Theory*, 52:1289–1306.

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

Friedman, J., Hastie, T., and Tibshirani, R. J. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1:302–332.

Friedman, J., Hastie, T., and Tibshirani, R. J. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33:Article 1.

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

Hastie, T., Tibshirani, R. J., and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York.

Hastie, T., Tibshirani, R. J., and Friedman, J. (2008). *The Elements of Statistical Learning: Data Mining, Inference and Prediction,* Second Edition. Springer, New York.

Joliffe, I. T., Trendafilov, N. T., and Uddin, M. (2003). A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, 12:531–547.

Kim, S.-J., Koh, K., Boyd, S., and Gorinevsky, D. (2009). ℓ_1 trend filtering. *SIAM Review, Problems and Techniques Section*, 51:339–360.

Lockhart, R. A., Taylor, J., Tibshirani, R. J., and Tibshirani, R. J. (2014). A significance test for the lasso (with discussion). *The Annals of Statistics*, in press.

Mazumder, R., Hastie, T., and Tibshirani, R. J. (2010). Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11:2287–2322.

Tibshirani, R. J. (2011). Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical Society, Series B*, 73:273–282.

Tibshirani, R. J. (2014). In praise of sparsity and convexity. In *Past, Present, and Future of Statistical Science*, pp. 497–505. Chapman & Hall, London.

Tibshirani, R. J., Hoefling, H., and Tibshirani, R. J. (2011). Nearly-isotonic regression. *Technometrics*, 53:54–61.

Tibshirani, R. J., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society, Series B*, 67:91–108.

Tibshirani, R. J. and Taylor, J. (2010). *The Solution Path of the Generalized Lasso.* Technical Report, Stanford University, Stanford, CA.

Tibshirani, R. J. and Taylor, J. (2011). The solution path of the generalized lasso. *The Annals of Statistics*, 39:1335–1371.

Tseng, P. (1988). Coordinate ascent for maximizing nondifferentiable concave functions. Technical Report LIDS-P; 1840, Massachusetts Institute of Technology, Boston, MA.

Witten, D., Tibshirani, R. J., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biometrika*, 10:515–534.

Yuan, M. and Lin, Y. (2007a). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series 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. *Journal of the American Statistical Association*, 101:1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320.