42

In praise of sparsity and convexity

Robert J. Tibshirani

Department of Statistics, Stanford, CA

To celebrate the 50th anniversary of COPSS, I discuss some examples of exciting developments of sparsity and convexity, in statistical research and practice.

42.1 Introduction

When asked to reflect on an anniversary of their field, scientists in most fields would sing the praises of their subject. As a statistician, I will do the same. However, here the praise is justified! Statistics is a thriving discipline, more and more an essential part of science, business and societal activities. Class enrollments are up—it seems that everyone wants to be a statistician—and there are jobs everywhere. The field of machine learning, discussed in this volume by my friend Larry Wasserman, has exploded and brought along with it the computational side of statistical research. Hal Varian, Chief Economist at `Google`, said "I keep saying that the sexy job in the next 10 years will be statisticians. And I'm not kidding." Nate Silver, creator of the *New York Times* political forecasting blog "538" was constantly in the news and on talk shows in the runup to the 2012 US election. Using careful statistical modelling, he forecasted the election with near 100% accuracy (in contrast to many others). Although his training is in economics, he (proudly?) calls himself a statistician. When meeting people at a party, the label "Statistician" used to kill one's chances of making a new friend. But no longer!

In the midst of all this excitement about the growing importance of statistics, there are fascinating developments within the field itself. Here I will discuss one that has been the focus my research and that of many other statisticians.

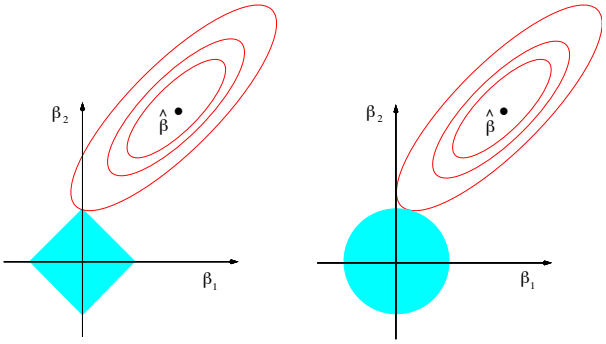


FIGURE 42.1
Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \le t$ and $\beta_1^2 + \beta_2^2 \le t^2$, respectively, while the red ellipses are the contours of the least squares error function. The sharp corners of the constraint region for the lasso yield sparse solutions. In high dimensions, sparsity arises from corners and edges of the constraint region.

42.2 Sparsity, convexity and ℓ_1 penalties

One of the earliest proposals for using ℓ_1 or absolute-value penalties, was the lasso method for penalized regression. Given a linear regression with predictors x_{ij} and response values y_i for $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, p\}$, the lasso solves the ℓ_1-penalized regression

$$\text{minimize}_\beta \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

This is equivalent to minimizing the sum of squares with constraint $|\beta_1| + \cdots + |\beta_p| \le s$. It is similar to *ridge regression*, which has constraint $\beta_1^2 + \cdots + \beta_p^2 \le s$. Because of the form of the ℓ_1 penalty, the lasso does variable selection and shrinkage; while ridge regression, in contrast, only shrinks. If we consider a more general penalty of the form $(\beta_1^q + \cdots + \beta_p^q)^{1/q}$, then the lasso uses $q = 1$ and ridge regression has $q = 2$. Subset selection emerges as $q \to 0$, and the lasso corresponds to the smallest value of q (i.e., closest to subset selection) that yields a convex problem. Figure 42.1 gives a geometric view of the lasso and ridge regression.

The lasso and ℓ_1 penalization have been the focus of a great deal of work recently. Table 42.1, adapted from Tibshirani (2011), gives a sample of this work.

TABLE 42.1
A sampling 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. (2010)
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)

The original motivation for the lasso was interpretability: It is an alternative to subset regression for obtaining a sparse model. Since that time, two unforeseen advantages of convex ℓ_1-penalized approaches have emerged: *Computational* and *statistical* efficiency. On the computational side, convexity of the problem and sparsity of the final solution can be used to great advantage. When most parameter estimates are zero in the solution, those parameters can be handled with minimal cost in the search for the solution. Powerful and scalable techniques for convex optimization can be unleashed on the problem, allowing the solution of very large problems. One particularly promising 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 a wide variety of other ℓ_1-penalized generalized linear models, including Cox's proportional hazards model for survival data. Coordinate descent is implemented in the popular `glmnet` package in the R statistical language, written by Jerome Friedman, Trevor Hastie, and myself, with help in the Cox feature from Noah Simon.

On the statistical side, there has also been a great deal of deep and interesting work on the mathematical aspects of the lasso, examining its ability to produce a model with minimal prediction error, and also to recover the true underlying (sparse) model. Important contributors here include Bühlmann, Candès, Donoho, Greenshtein, Johnstone, Meinshausen, Ritov, Wainwright, Yu, and many others. In describing some of this work, Hastie et al. (2001) coined the informal "Bet on Sparsity" principle. The 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 $p \gg N$, a commonly occurring scenario.

42.3 An example

I am currently involved in a cancer diagnosis project with researchers at Stanford University. They have collected samples of tissue from 10 patients undergoing surgery for stomach cancer. The aim is to build a classifier than 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 42.2. A pathologist has labelled 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 kind of mass spectrometry, with the peaks in the spectrum representing different metabolites. The spectrum has been finely sampled at about 11,000 sites. Thus the task is to build a classifier 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. For each class $k \in \{1, 2, 3\}$, the model has a vector $(\beta_{1k}, \ldots, \beta_{pk})$ of parameters 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 encouraging. The classifier shows 93–97% accuracy in the three classes, using only around 100 features. These features could yield insights about the metabolites that are important in stomach cancer. There is much more work to be done—collecting more data, and refining and testing the model. But this shows the potential of ℓ_1-penalized models in an important and challenging scientific problem.

42.4 The covariance test

So far, most applications of the lasso and ℓ_1 penalties seem to focus on large problems, where traditional methods like all-subsets-regression can't deal with the problem computationally. In this last section, I want to report on some

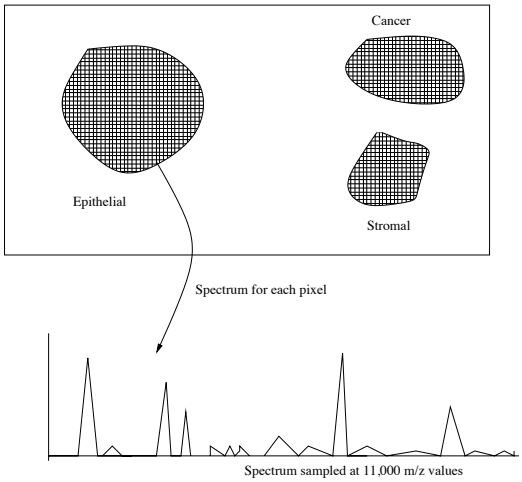


FIGURE 42.2
Schematic of the cancer diagnosis problem. Each pixel in each of the three regions labelled by the pathogolist is analyzed by mass spectometry. This gives a feature vector of $11,000$ intensities (bottom panel), from which we try to predict the class of that pixel.

very recent work that suggest that ℓ_1 penalties may have a more fundamental role in classical mainstream statistical inference.

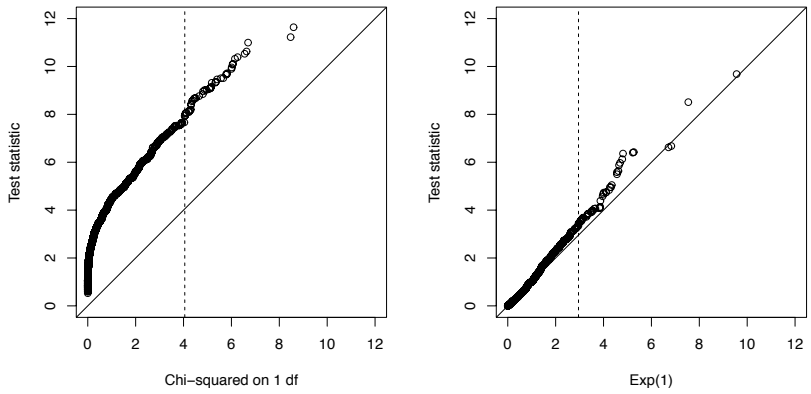
To begin, consider standard forward stepwise regression. This procedure enters predictors one a time, choosing the predictor that most decreases the residual sum of squares at each stage. Defining RSS to be the residual sum of squares for the model containing j predictors and denoting by RSS_{null} the residual sum of squares for the model omitting the predictor $k(j)$, we can form the usual statistic

$$R_j = (\text{RSS}_{\text{null}} - \text{RSS})/\sigma^2$$

(with σ assumed known for now), and compare it to a $\chi^2_{(1)}$ distribution.

Although this test is commonly used, we all know that it is wrong. Figure 42.3 shows an example. There are 100 observations and 10 predictors in a standard Gaussian linear model, in which all coefficients are actually zero. The left panel shows a quantile-quantile plot of 500 realizations of the statistic R_1 versus the quantiles of the $\chi^2_{(1)}$ distribution. The test is far too liberal and the reason is clear: The $\chi^2_{(1)}$ distribution is valid for comparing two *fixed* nested linear models. But here we are *adaptively* choosing the best predictor, and comparing its model fit to the null model.

In fact it is difficult to correct the chi-squared test to account for adaptive selection: Half-sample splitting methods can be used (Meinshausen et al., 2009;



(a) Forward stepwise (b) Lasso

FIGURE 42.3
A simple example with $n = 100$ observations and $p = 10$ orthogonal predictors. All true regression coefficients are zero, $\beta^* = 0$. On the left is a quantile-quantile plot, constructed over 1000 simulations, of the standard chi-squared statistic R_1, measuring the drop in residual sum of squares for the first predictor to enter in forward stepwise regression, versus the χ^2_1 distribution. The dashed vertical line marks the 95% quantile of the $\chi^2_{(1)}$ distribution. The right panel shows a quantile-quantile plot of the covariance test statistic T_1 for the first predictor to enter in the lasso path, versus its asymptotic distribution $\mathcal{E}(1)$. The covariance test explicitly accounts for the adaptive nature of lasso modeling, whereas the usual chi-squared test is not appropriate for adaptively selected models, e.g., those produced by forward stepwise regression.

Wasserman and Roeder, 2009), but these may suffer from lower power due to the decrease in sample size.

But the lasso can help us! Specifically, we need the LAR (least angle regression) method for constructing the lasso path of solutions, as the regularization parameter λ is varied. I won't give the details of this construction here, but we just need to know that there are a special set of decreasing knots $\lambda_1 > \cdots > \lambda_k$ at which the active set of solutions (the non-zero parameter estimates) change. When $\lambda > \lambda_1$, the solutions are all zero. At the point $\lambda = \lambda_1$, the variable most correlated with y enters the model. At each successive value λ_j, a variable enters or leaves the model, until we reach λ_k where we obtain the full least squares solution (or one such solution, if $p > N$).

We consider a test statistic analogous to R_j for the lasso. Let $\mathbf{y}$ be the vector of outcome values and $\mathbf{X}$ be the design matrix. Assume for simplicity that the error variance σ^2 is known. Suppose that we have run LAR for $j-1$

steps, yielding the active set of predictors $\mathcal{A}$ at $\lambda = \lambda_j$. Now we take one more step, entering a new predictor $k(j)$, and producing estimates $\hat{\beta}(\lambda_j)$ at λ_{j+1}. We wish to test if the $k(j)$th component $\beta_{k(j)}$ is zero. We refit the lasso, keeping $\lambda = \lambda_{j+1}$ but using just the variables in $\mathcal{A}$. This yields estimates $\hat{\beta}_{\mathcal{A}}(\lambda_{j+1})$. Our proposed *covariance test statistic* is defined by

$$T_j = \frac{1}{\sigma^2}\{\langle \mathbf{y}, \mathbf{X}\hat{\beta}(\lambda_{j+1})\rangle - \langle \mathbf{y}, \mathbf{X}_{\mathcal{A}}\hat{\beta}_{\mathcal{A}}(\lambda_{j+1})\rangle\}. \tag{42.1}$$

Roughly speaking, this statistic measures how much of the covariance between the outcome and the fitted model can be attributed to the $k(j)$th predictor, which has just entered the model.

Now something remarkable happens. Under the null hypothesis that all signal variables are in the model: As $p \to \infty$, T_j converges to an exponential random variable with unit mean, $\mathcal{E}(1)$. The right panel of Figure 42.3 shows the same example, using the covariance statistic. This test works for testing the first variable to enter (as in the example), or for testing noise variables after all of the signal variables have entered. And it works under quite general conditions on the design matrix. This result properly accounts for the adaptive selection: The shrinkage in the ℓ_1 fitting counteracts the inflation due to selection, in just the right way to make the degrees of freedom (mean) of the null distribution exactly equal to 1 asymptotically. This idea can be applied to a wide variety of models, and yields honest p-values that should be useful to statistical practitioners.

In a sense, the covariance test and its exponential distribution generalize the RSS test and its chi-squared distribution, to the adaptive regression setting.

This work is very new, and is summarized in Lockhart et al. (2013). The proofs of the results are difficult, and use extreme-value theory and Gaussian processes. They suggest that the LAR knots λ_k may be fundamental in understanding the effects of adaptivity in regression. On the practical side, regression software can now output honest p-values as predictors enter a model, that properly account for the adaptive nature of the process. And all of this may be a result of the convexity of the ℓ_1-penalized objective.

42.5 Conclusion

In this chapter I hope that I have conveyed my excitement for some recent developments in statistics, both in its theory and practice. I predict that convexity and sparsity will play an increasing important role in the development of statistical methodology.

References

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

Candès, E.J. (2006). Compressive sampling. In *Proceedings of the International Congress of Mathematicians*, Madrid, Spain. `www.acm.caltech.edu/~emmanuel/papers/CompressiveSampling.pdf`

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. `http://www.citebase.org/abstract?id=oai:arXiv.org:0903.1476`

Donoho, D.L. (2004). *Compressed Sensing*. Technical Report, Statistics Department, Stanford University, Stanford, CA. `www-stat.stanford.edu/~donoho/Reports/2004/CompressedSensing091604.pdf`

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

Friedman, J.H., Hastie, T.J., and Tibshirani, R.J. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33:1–22.

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

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

Lockhart, R.A., Taylor, J., Tibshirani, R.J., and Tibshirani, R.J. (2013). A significance test for the lasso. arXiv:1301.7161; submitted.

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

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

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

Wasserman, L.A. and Roeder, K. (2009). High-dimensional variable selection. *Journal of the American Statistical Association*, 37:2178–2201.

Witten, D.M., Tibshirani, R.J., and Hastie, T.J. (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.J. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320.