12. Sparsity and $p >> n$

When $p >> n$ (the "short, fat data problem"), two things go wrong:

- The Curse of Dimensionality is acute.
- There are insufficient degrees of freedom to estimate the full model.

However, there is a substantial body of practical experience which indicates that, in some circumstances, one can actually make good statistical inferences and predictions.

Recently, a great deal of statistical attention has gone into the question of determining under which conditions it is possible to do good statistics, and under what circumstances is it impossible to learn from such data.

The issue of insufficient degrees of freedom has arisen previously. In fractional factorial designs, one does not have sufficient observations to estimate all of the main effects and interactions, and consequently one aliases (or confounds) cleverly chosen sets of terms, estimating their sum rather than the separate components.

The intuition behind that approach is the **bet on sparsity** principle, which says that in high dimensions, it is wise to proceed under the assumption that most of the effects are not signficant.

This principle can be justified in two ways:

- it is often true (Occam's Razor is empirically sound); and
- if the problem is not sparse, you won't be able to do anything useful anyway.

In some cases things aren't strictly sparse, in that all effects are non-zero, but a handful of the effects may explain a very large proportion of the variation in the data.

There are several different kinds of sparsity. The following list is framed in terms of regression, but similar points apply to classification.

- Among the p explanatory variables, only a small number q are relevant. (Classic sparsity.)
- Although all of the p explanatory variables are important, one can partition the space so that in any local region, only a small number are relevant. (Locally-Low Dimension.)
- Although all of the p explanatory variables are important, one can find a small number of linear combinations of those variables that explain most of the variation in the response. (Nearly Dark Representation.)

There may be other kinds, but these are the main ones.

The rest of this unit shall focus on classical sparsity. Locally-low dimension was discussed in the context of CART and q-cubes in p-space. Nearly Dark Representations refer to finding a basis set that enables one to fit a sparse model.

To illustrate the $p >> n$ problem in regression, the book produced samples of size $n = 100$ according to the following procedure.

First, generate p covariates from a Gaussian distribution with pairwise correlation 0.2. Three values of p were used: 20, 100, 1000.

For each set, generate a response according to

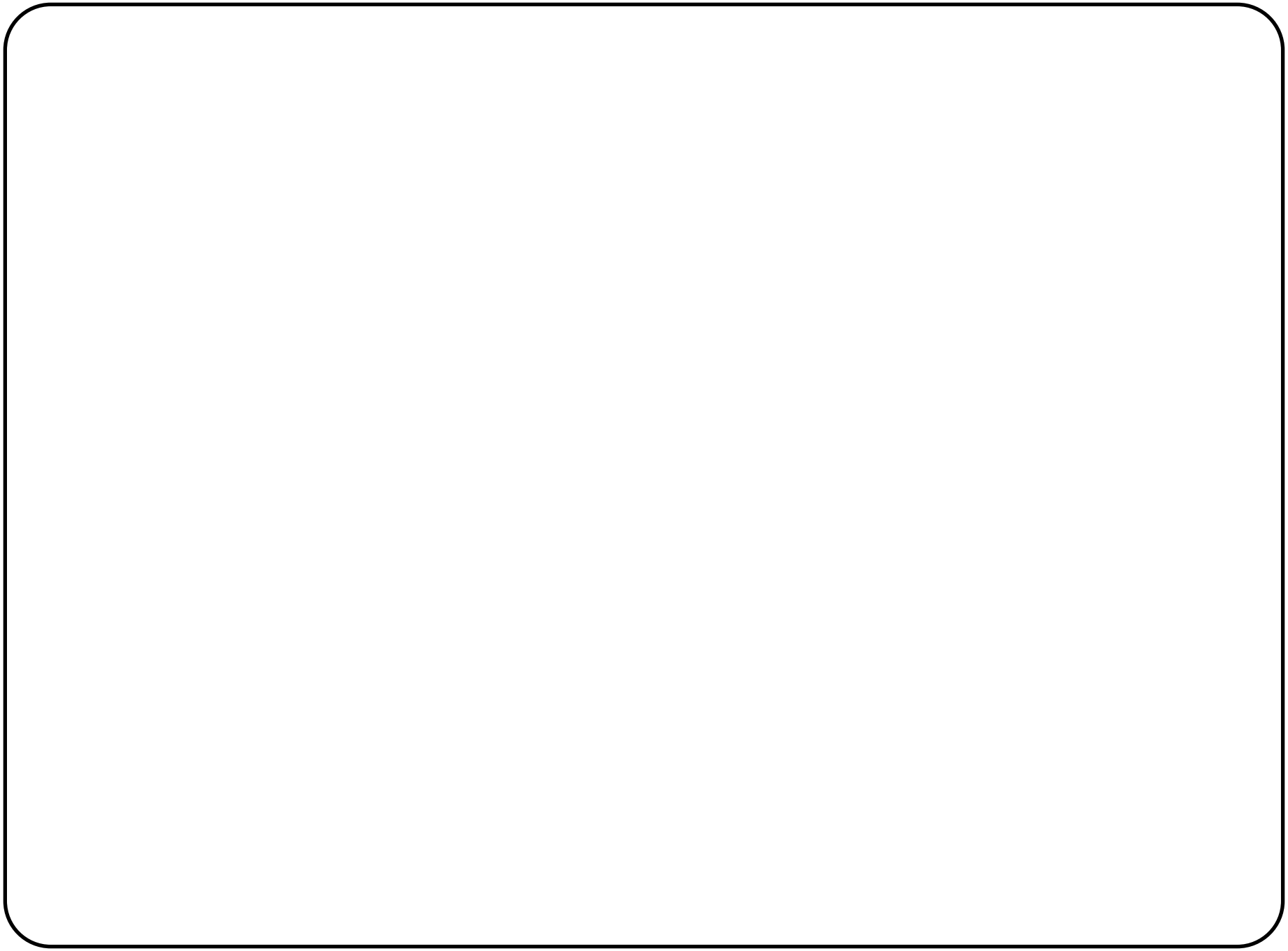
$$Y = \beta_1 X_1 + \cdots + \beta_p X_p + \sigma\epsilon$$

where the ϵ and all β_j are N(0,1). The value of σ was chosen to that the signal-to-noise ratio $\text{Var}\left[\mathbb{E}(Y|\boldsymbol{X})\right]/\text{Var}\left(\epsilon\right)$ was constant.

Averaging over the simulation runs, there were 9, 33, and 331 significant coefficients, for $p = 20$, 100, and 1000, respectively.

The following boxplots show the result of using ridge regression on these data sets, with regularization parameter $\lambda = 0.001$, 100, and 1000. For a given p, these values give different effective degrees of freedom, shown on the horizontal axes.

One sees that for $p = 20$, $\lambda = 0.001$ gives the smallest relative test error. For $p = 100$, the best is $\lambda = 100$, and for $p = 1000$, the best is $\lambda = 1000$. As p gets large and surpasses n, more regularization (shrinkage) is needed. This shrinkage is a way of enforcing sparsity, and shows that traditional methods fail when $p >> n$.



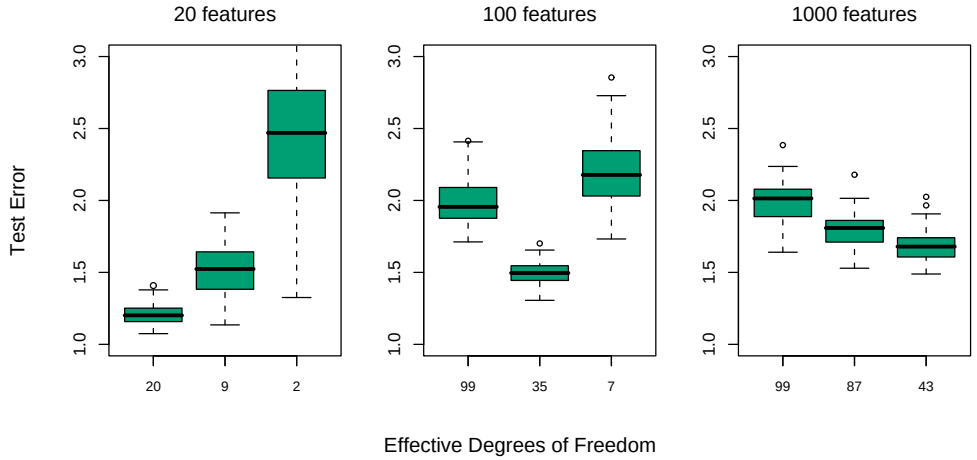


FIGURE 18.1. *Test-error results for simulation experiments. Shown are boxplots of the relative test errors over* 100 *simulations, for three different values of* p*, the number of features. The relative error is the test error divided by the Bayes error,* σ^2*. From left to right, results are shown for ridge regression with three different values of the regularization parameter* λ*:* 0.001, 100 *and* 1000*. The (average) effective degrees of freedom in the fit is indicated below each plot.*

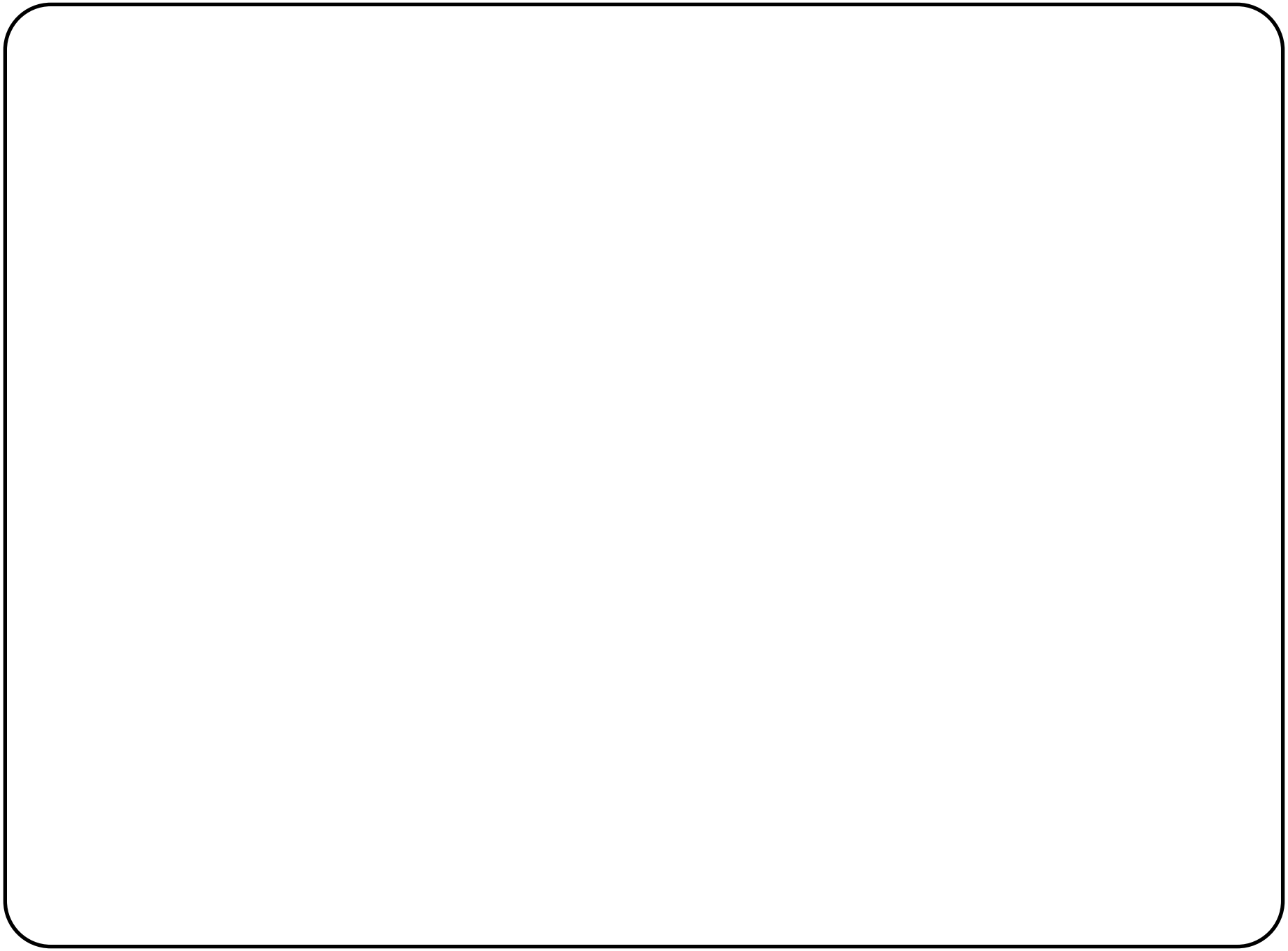
The first serious attempt in the data mining community at the $p >> n$ problem was the LASSO. As you recall, the LASSO seeks to minimize squared error loss (L^2 loss) under a penalty on the sum of the absolute values of the coefficients (an L^1 penalty).

Specifically, one solves:

$$\hat{\boldsymbol{\beta}} = \mathbf{argmin}_{\boldsymbol{\beta}}\{\sum_{I=1}^{n}(y_I - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2\}$$

subject to the constraint that $\sum_{j=1}^{p} |\beta_j| \leq s$.

This solution tends to find estimates of $\boldsymbol{\beta}$ that are mostly zero. From a Bayesian standpoint, it results from using a Laplacian prior on the coefficients, rather than a normal prior, in Bayesian regression.



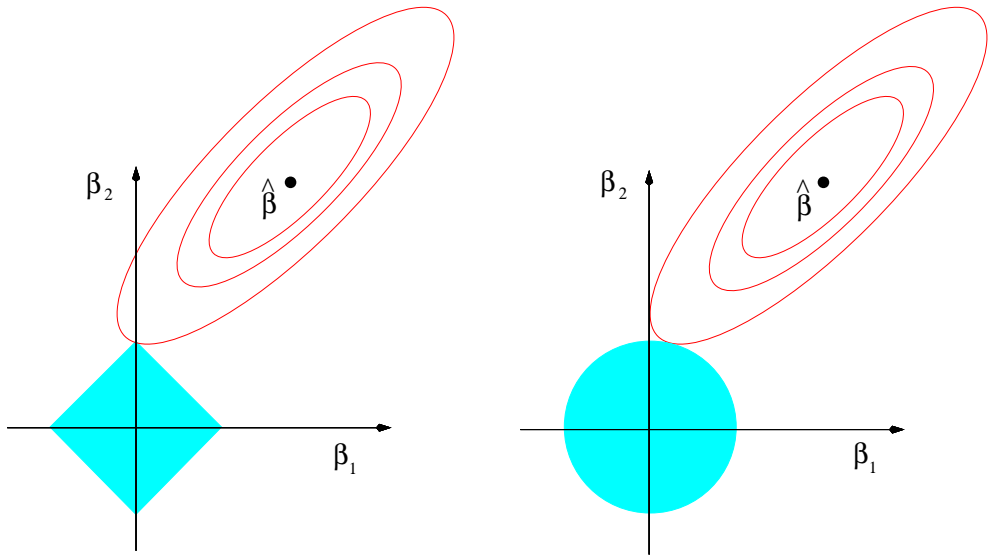


Figure 3.12: *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$, *respectively, while the red ellipses are the contours of the least squares error function.*

The geometry in the previous figure shows why this tends to give estimates that are zero. The same strategy applies to other penalty regions.

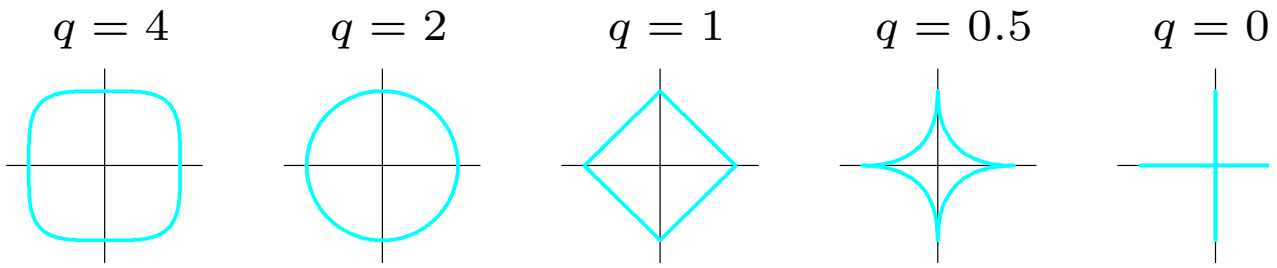


FIGURE 3.12. *Contours of constant value* $\sum_j |\beta_j|^q$ *for given values of* q.

In particular, the L^0 penalty strongly favors sparse solutions. But the penalty is not convex, so solution is difficult and often impossible.

An important variant on the Lasso is the **elastic net**. It is pertinent when there are strong correlations among the explanatory variables.

Correlated covariates are common in many important applications. For example, in genomics, genes typically activate in pathways, so expression levels in pathways that affect some phenotype is typically strongly correlated.

The Lasso tends to pick one gene among the correlated ones and put all the weight on it. Ridge regression tends to select all of the correlated variables, and shrink their values towards each other. (If the data are standardized, then the tendency is for all of those coefficients to be equal.)

¿From a sparsity standpoint, one wants to find a small number of sets of correlated explanatory variables. Elastic nets offer that capability. They use the penalty function

$$\sum_{j=1}^{p} \left(\alpha|\beta_j| + \frac{1-\alpha}{\beta_j^2} \right).$$

The second term in the penalty encourages the averaging of coefficients on correlated subsets of explanatory variables. The first term encourages the solution to use only a small number of such subsets.

The following diagram compares the Lasso and the elastic net. The latter finds more non-zero terms, but their coefficients have similar and lower magnitude.

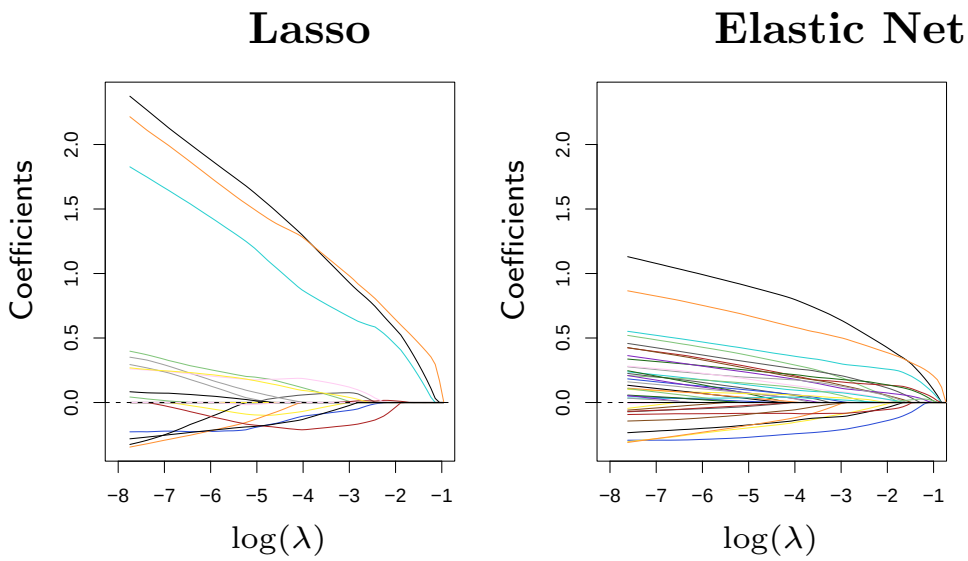


FIGURE 18.5. *Regularized logistic regression paths for the leukemia data. The left panel is the lasso path, the right panel the elastic-net path with* $\alpha = 0.8$. *At the ends of the path (extreme left), there are* 19 *nonzero coefficients for the lasso, and* 39 *for the elastic net. The averaging effect of the elastic net results in more non-zero coefficients than the lasso, but with smaller magnitudes.*

Consider again L^2 minimization with an L^0 penalty. It solves

$$\hat{\boldsymbol{\beta}} = \mathbf{argmin}_{\boldsymbol{\beta}}\{\sum_{i=1}^{n}(y_I - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 - \lambda\|\boldsymbol{\beta}\|_0\}$$

where $\|\boldsymbol{\beta}\|_0$ is number of non-zero components in the $\boldsymbol{\beta}$ vector.

The non-convexity of the penalty region means that one has to perform combinatorial search to find the solution. However, in some situations, solution is possible—in that case it is called the Dantzig selector by Candes and Tao (2007).

The Dantzig estimator $\hat{\boldsymbol{\beta}}_D$ is based on the fact that sometimes (often) the L^1 solution agrees with the L^0 solution.

The $\hat{\boldsymbol{\beta}}_D$ minimizes the L^1 norm of $\boldsymbol{\beta}$ subject to the condition that

$$\sup_{\boldsymbol{\beta}} \|\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\| \leq c_D \sigma \sqrt{2 \ln p}$$

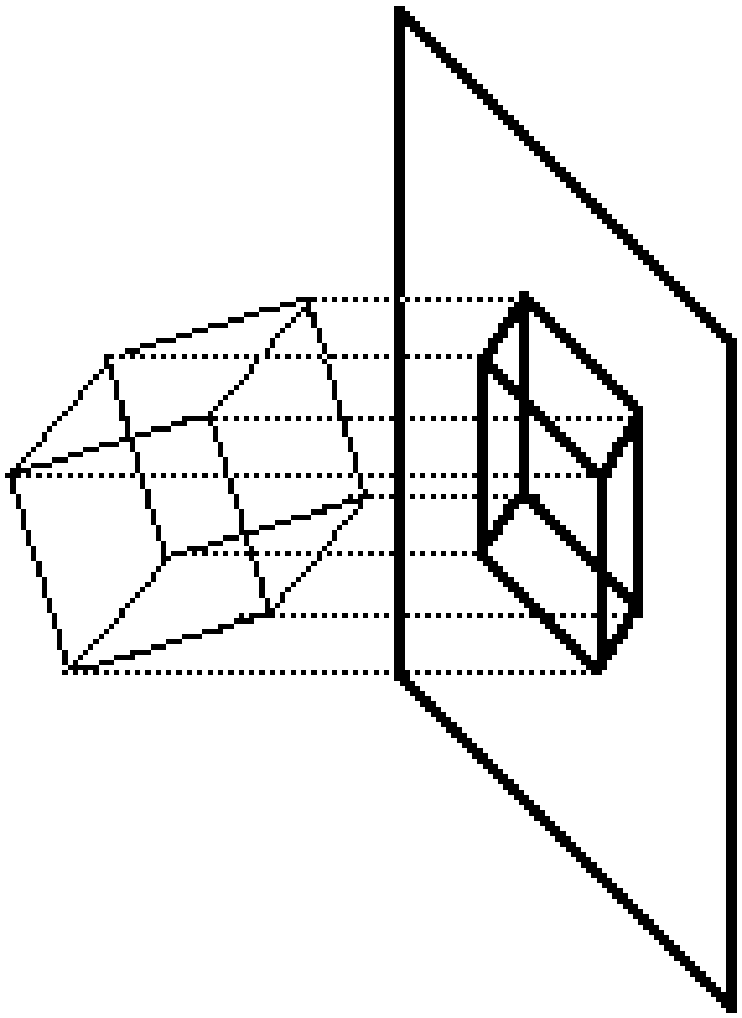
where c_D is a constant that requires some calculation. One uses linear programming to find the minimizing solution.

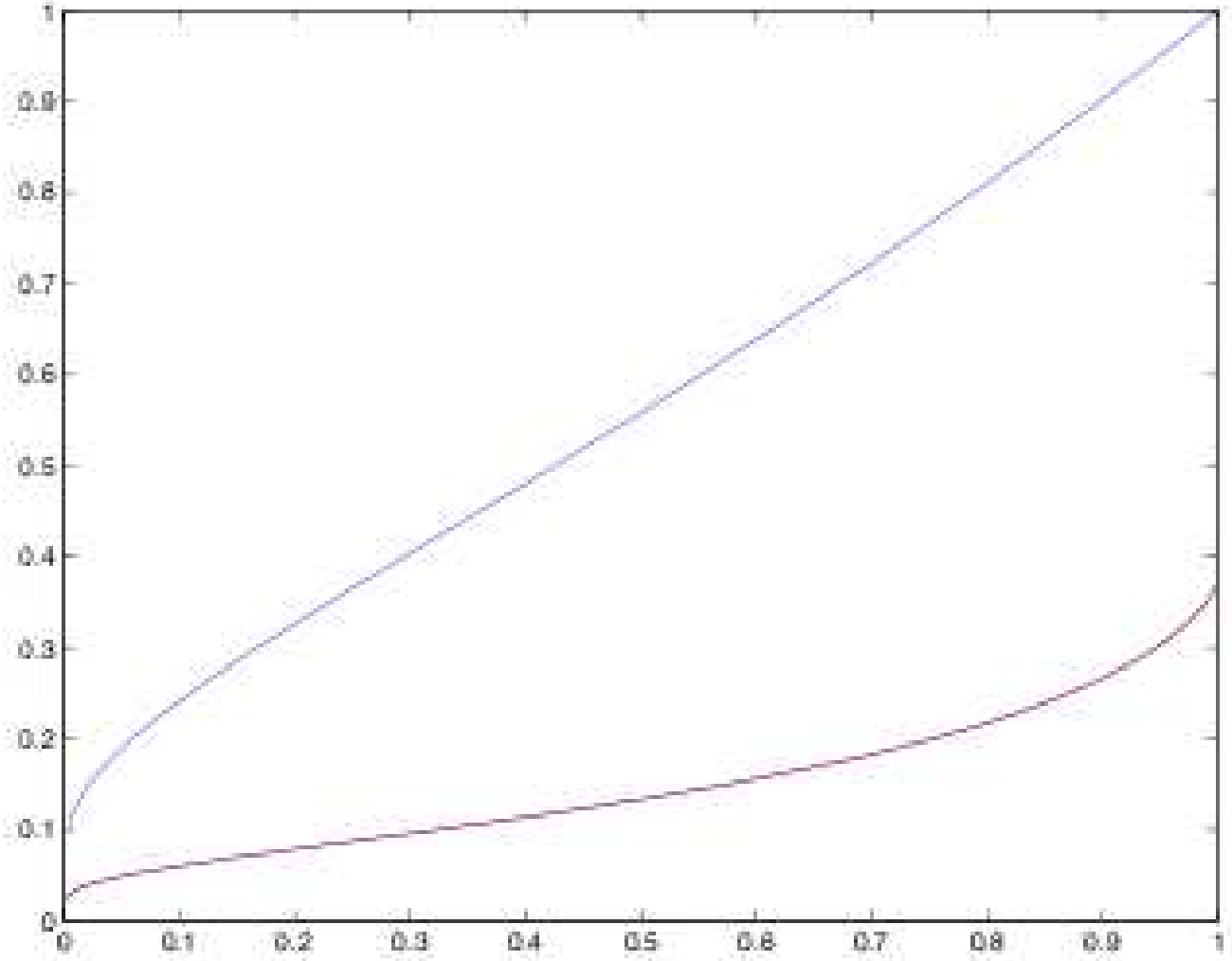
If the number of non-zero elements of the true $\boldsymbol{\beta} = k < n$, then with high probability the Dantzig estimate does asymptotically as well as could be done if the identities of the non-zero components were known to the analyst. (Note: the sense of the asymptotics here is a bit delicate.)

A second approach to this problem is due to Donoho and Tanner (2007). They consider the noise-free case, in which $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}$ but $\boldsymbol{X}$ is $n \times p$ with $p >> n$.

Since the $\boldsymbol{X}'\boldsymbol{X}$ matrix is singular, this system of equations cannot be solved. But, in the following diagram, if the vertical axis is k/n and the horizontal axis is n/p, then there appears to be a phase change, somewhere between the indicated boundaries. Above the upper line, solution is impossible; below the lower line, with high probability the solution can be obtained.

The ability to find the solution depends upon whether to solution to the full system is an interior point or an exterior point when the one projects onto a lower dimensional space.




1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

The third approach to the $p >> n$ problem is due to Wainwright (2008?). He showed that under certain assumptions on $\boldsymbol{X}$, then with high probabilitythe Lasso correctly identifies the non-zero elements of $\boldsymbol{\beta}$.

Specifically, in the limit,

- if $n > 2k \ln(p - k)$, then the probability that the Lasso succeeds tends to 1;
- if $n < \frac{1}{2} k \ln(p - k)$, then the probability that the Lasso succeeds tends to 0.

Note that all three approaches have similar results—under various conditions, sparse linear regression can work, even when $p >> n$.

The results for linear regression probably apply to nonparametric regression. Lafferty and Wasserman (2007) attempt to extend methods for nonparametric regression, based on additive models and generalized additive models, to the $p >> n$ situtation. Their method is called SpAM (Sparse Additive Modeling). A combination of simulation and theory suggests it works fairly well.

Simulation studies suggest that the phase transition found by Donoho and Tanner for the linear model also arises in many nonlinear and nonparametric cases. But there is really no theory to rely upon, other than a general consensus that anything sensible is roughly approximate to the linear model

Regarding $p >> n$ classification problems, less statistical work has been done, but most people believe the problem is similar to regression and perhaps somewhat easier.