Statistical Learning With Sparsity. The Lasso and Generalizations
Hastie, Tibshirani, Wainwright

Now large quantities of data are collected and mined in nearly every area of
science, entertainment, business, and industry. Medical scientists study the
genomes of patients to choose the best treatments, to learn the underlying
causes of their disease. Online movie and book stores study customer ratings to
recommend or sell them new movies or books. Social networks mine information
about members and their friends to try to enhance their online experience.

Thus the world is awash with data. But as Rutherford D. Roger (and others)
has said: “We are drowning in information and starving for knowledge.”

In this book we study methods that exploit sparsity to help recover the un-
derlying signal in a set of data. The leading example is linear regression, in
which the goal is to predict the outcome from the predictors, both for actual
prediction with future data and also to discover which predictors play an impor-
tant role. A linear regression model assumes that $y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + e_i$ The
method of least squares provides estimates of the parameters by minimization
of the least-squares objective function $\min_{\beta_0, \beta} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$
Typically all of the least-squares estimates from () will be nonzero. Thus there
is a need to constrain, or regularize the estimation process. In the lasso or
$\ell_1$-regularized regression, we estimate the parameters by solving the problem
$\min_{\beta_0, \beta} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$ subject to $\|\beta\|_1 \leq t$
Thus the advantages of sparsity are interpretation of the fitted model and computational convenience.

For all of these reasons, the area of sparse statistical modelling is exciting – for data analysts, computer scientists, and theorists – and practically useful.

As an example, consider the following data set. The data consists of quantitative gene expression measurements of 4718 genes on samples from 349 cancer patients. The cancers have been categorized into 15 different types such as bladder, breast, CNS, etc. For this purpose we applied a lasso-regularized multinomial classifier to these data. This produces a set of 4718 weights or coefficients for each of the 15 classes, for discriminating each class from the rest. Because of the $\ell_1$ penalty, only some of these weights may be nonzero (depending on the choice of the regularization parameter). We used cross-validation to estimate the optimal choice of regularization parameter and found out that only 254 genes have at least one nonzero weight. The cross-validated error rate for this classifier is about 10%, so the procedure correctly predicts the class of about 90% of the samples. By comparison, a standard support vector classifier had a slightly higher error rate (13%) using all of the features. Using sparsity, the lasso procedure has dramatically reduced the number of features without sacrificing accuracy. Sparsity has also brought computational efficiency.