Problem Statement

Overfitting a machine learning algorithm will lead to well fitting training models but poorly fitting test models. In effect an overfitted model will have a low *squared bias* but a large sample-to-sample *variation*.

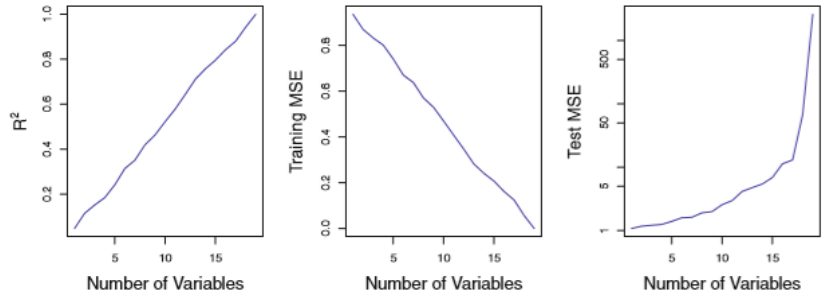

FIGURE 6.23. *On a simulated example with $n = 20$ training observations, features that are completely unrelated to the outcome are added to the model.* Left: *The R^2 increases to 1 as more features are included.* Center: *The training set MSE decreases to 0 as more features are included.* Right: *The test set MSE increases as more features are included.*

Example: Regression analysis with an observation specific dummy variable for each observation will have a perfect fit.

High-Dimensional Data

The number of features $K - 1$ is close to the number of observations n or even larger. If $n < K - 1$ the models cannot be uniquely calibrated.

Examples:

- The possible link of DNA mutations in single nucleotide polymorphisms to specific diseases. There are over 500.000 different SNPs and usually the sample size n is substantially smaller.

- Internet search terms by a set of users are used for marketing purpose. Usually the “bag-of-words” is substantially larger than the number of users.

Subsets of features will be perfectly correlated. This can lead to different but identical sub-set solutions (*one of many possible models problems*):

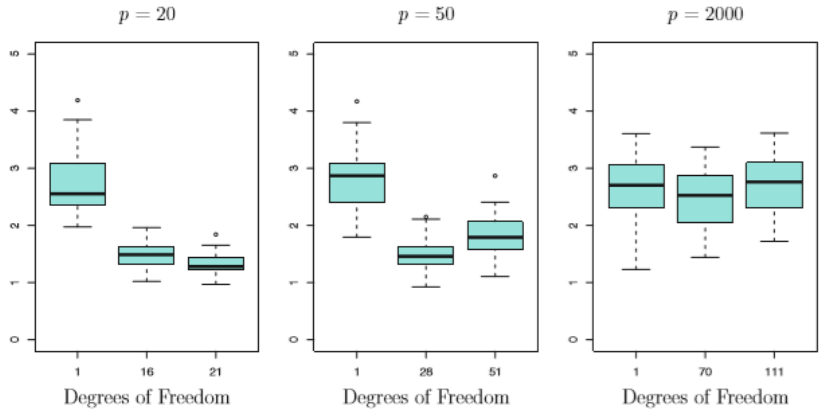

FIGURE 6.24. *The lasso was performed with $n = 100$ observations and three values of p, the number of features. Of the p features, 20 were associated with the response. The boxplots show the test MSEs that result using three different values of the tuning parameter λ in (6.7). For ease of interpretation, rather than reporting λ, the* degrees of freedom *are reported; for the lasso this turns out to be simply the number of estimated non-zero coefficients. When $p = 20$, the lowest test MSE was obtained with the smallest amount of regularization. When $p = 50$, the lowest test MSE was achieved when there is a substantial amount of regularization. When $p = 2{,}000$ the lasso performed poorly regardless of the amount of regularization, due to the fact that only 20 of the 2,000 features truly are associated with the outcome.*

Regularization

Regularization is an umbrella term that comprises of methods which build ***less complex*** models with a low prediction error.

Specific approaches:

- Best subset selection of explanatory variables. This is not practical because $2^{(K-1)}$ different combinations are possible. Approximations by stepwise search strategies are possible.
- Shrinkage methods like Ridge and Lasso regression
- Dimension reduction methods like principal component regression or partial least squares

Regularization are also applied with neutral networks, which directly minimize an objective function.

Shrinkage methods

The tuning parameter λ can be regarded as hyper-parameter.