it. One of the best methods relies on ADMM (see Section 52.8) and is discussed in Section 55.4. The lasso method has some limitations, in particular when the number m of data is smaller than the dimension n of the data. This happens in some applications in genetics and medicine. Fortunately there is a way to combine the best features of ridge regression and lasso, which is to use *two* regularizing terms:

1. An ℓ^2-term $(1/2)K \left\| w \right\|_2^2$ as in ridge regression (with $K > 0$).

2. An ℓ^1-term $\tau \left\| w \right\|_1$ as in lasso.

This method is known as *elastic net regression* and is discussed in Section 55.6. It retains most of the desirable features of ridge regression and lasso, and eliminates some of their weaknesses. Furthermore, it is effectively solved by ADMM.

55.1 Ridge Regression

The problem of solving an overdetermined or underdetermined linear system $Aw = y$, where A is an $m \times n$ matrix, arises as a “learning problem” in which we observe a sequence of data $((a_1, y_1), \ldots, (a_m, y_m))$, viewed as input-output pairs of some unknown function f that we are trying to infer, where the a_i are the *rows* of the matrix A and $y_i \in \mathbb{R}$. The values y_i are sometimes called *labels* or *responses*. The simplest kind of function is a linear function $f(x) = x^\top w$, where $w \in \mathbb{R}^n$ is a vector of coefficients usually called a *weight vector*, or sometimes an *estimator*. In the statistical literature w is often denoted by β. Since the problem is overdetermined and since our observations may be subject to errors, we can't solve for w exactly as the solution of the system $Aw = y$, so instead we solve the least-square problem of minimizing $\left\| Aw - y \right\|_2^2$.

In Section 23.1 we showed that this problem can be solved using the pseudo-inverse. We know that the minimizers w are solutions of the normal equations $A^\top A w = A^\top y$, but when $A^\top A$ is not invertible, such a solution is not unique so some criterion has to be used to choose among these solutions.

One solution is to pick the unique vector w^+ of smallest Euclidean norm $\left\| w^+ \right\|_2$ that minimizes $\left\| Aw - y \right\|_2^2$. The solution w^+ is given by $w^+ = A^+ y$, where A^+ is the pseudo-inverse of A. The matrix A^+ is obtained from an SVD of A, say $A = V \Sigma U^\top$. Namely, $A^+ = U \Sigma^+ V^\top$, where Σ^+ is the matrix obtained from Σ by replacing every nonzero singular value σ_i in Σ by σ_i^{-1}, leaving all zeros in place, and then transposing. The difficulty with this approach is that it requires knowing whether a singular value is zero or very small but nonzero. A very small nonzero singular value σ in Σ yields a very large value σ^{-1} in Σ^+, but $\sigma = 0$ remains 0 in Σ^+.

This discontinuity phenomenon is not desirable and another way is to control the size of w by adding a regularization term to $\left\| Aw - y \right\|^2$, and a natural candidate is $\left\| w \right\|^2$.