Chapter 14

Regularization

In several sections of this book we touched on the topic of regularization (see, e.g., §8.2.1.2 and §8.2.3). A variety of statistical procedures and machine learning algorithms employ regularization (under different names) to improve out-of-sample fit. Good out-of-sample fit means generalization from observed data, which, as we've stressed before, is the key problem of statistics. This chapter introduces a number of methods that use regularization and discusses their statistical properties.

14.1 Nonparametric Density Estimation

Nonparametric density estimation is an application of regularization to the problem of recovering distributions from data. It combines the data with a prior belief that probability mass most likely falls in places other than just the sample points observed so far. As well as being of interest from a theoretical perspective, nonparametric density estimation is used in a great variety of applied studies. We begin our analysis with a review of parametric density estimation and then proceed to nonparametric methods.

14.1.1 Introduction

Suppose our data consist of IID observations $\mathbf{x}_1, \ldots, \mathbf{x}_N$ from unknown distribution P on $\mathbb{R}^d$. We assume throughout this section that P is absolutely continuous. Our aim is to estimate the density of P, denoted below by f.

We know how to do this in a parametric setting. For example, let's add the assumption that f belongs to the class of normal densities on $\mathbb{R}$, so that $f = f(\cdot\,; \mu, \sigma) =$ the normal density for distribution $\mathrm{N}(\mu, \sigma^2)$. The MLEs of the parameters are $\hat{\mu}_N :=$

$\bar{x}_N$ and $\hat{\sigma}_N := s_N$ (see page 235). Plugging these back into f gives density estimate $f(\cdot\,; \bar{x}_N, s_N)$. Since $\bar{x}_N$ and s_N are consistent (see §9.2.4), the random density $f(\cdot\,; \bar{x}_N, s_N)$ will be close to $f(\cdot\,; \mu, \sigma)$ with high probability for large N.

We can clarify convergence of the densities themselves if we extend the notion of consistency from vectors to densities. To extend consistency to densities we need a notion of global deviation between densities. Perhaps the most important measures of global distance are the L_p distances. Let's now define and discuss them.

For any $\mathscr{B}$-measurable function $f\colon \mathbb{R}^d \to \mathbb{R}$ and $p \geqslant 1$, we set

$$\|f\|_p := \left\{ \int |f|^p \right\}^{1/p} := \left\{ \int |f(\mathbf{s})|^p \, \mathrm{d}\mathbf{s} \right\}^{1/p} \tag{14.1}$$

where integration is over all of $\mathbb{R}^d$. If this expression is finite, then we write $f \in L_p$. For densities f and g on $\mathbb{R}^d$ the L_p distance between these densities is then defined as

$$d_p(f, g) := \|f - g\|_p \tag{14.2}$$

The norm $\|\cdot\|_p$ satisfies most of the properties of the norms we've met so far. For example, we have the triangle inequality

$$\|f - g\|_p \leqslant \|f - h\|_p + \|h - g\|_p \tag{14.3}$$

for all $p \geqslant 1$ and $f, g, h \in L_p$. See, for example, theorem 5.1.5 of Dudley (2002).

Specializing to $p = 1$ gives the L_1 distance, which sums over absolute deviation. Specializing to $p = 2$ gives the popular L_2 distance, which is a variation of the L_2 distance we used in §5.2.2. See, in particular, (5.28) on page 142, where distance between random variables is evaluated based on expected deviation.

The L_2 distance is popular largely because it's so tractable, and because it's similarity to Euclidean vector distance means that we have L_2 analogies for a variety of fundamental results on vector space. Nevertheless, the L_1 distance is arguably a better choice for studying deviation between densities. For example, the L_1 distance between densities is always well-defined (see ex. 14.4.1), which is essential if one hopes to provide universal consistency results.

Fact 14.1.1 **Scheffés lemma**: If $\{f_n\}$ and f are densities on $\mathbb{R}^d$, then

$$f_n(\mathbf{s}) \to f(\mathbf{s}) \text{ for all } \mathbf{s} \text{ in } \mathbb{R}^d \quad \Longrightarrow \quad \|f_n - f\|_1 \to 0$$

Fact 14.1.2 For any densities f, g and h on $\mathbb{R}^d$ we have

(i) $\|f - g\|_1 \leqslant \sqrt{2D(f, g)}$, where $D(f, g)$ is the KL deviation defined in (8.34), and

(ii) $\|f - g\|_1 = 2 \sup_{B \in \mathscr{B}(\mathbb{R}^d)} \left| \int_B f - \int_B g \right|$.

The bound in (i) is called **Pinsker's inequality**, while (ii) is called **Scheffé's identity**. Scheffé's identity tells us that L_1 distance measures something that we directly care about: when L_1 deviation is small, so is the maximal deviation between probabilities assigned to events.[1]

In what follows, we will say that a sequence $\{\hat{f}_N\}$ of random densities on $\mathbb{R}^d$ is L_p**-consistent** for a density f on $\mathbb{R}^d$ if

$$\|\hat{f}_N - f\|_p \xrightarrow{p} 0 \quad \text{as} \quad N \to \infty$$

Example 14.1.1 Let $\hat{f}_N = f(\cdot\,; \bar{x}_N, s_N)$ be the Nth element of the sequence of normal densities described above, where $x_1, \ldots, x_N$ are independent draws from a normal density $f = f(\cdot\,; \mu, \sigma)$ and $\bar{x}_N$ and s_N the sample mean and standard deviation respectively. This sequence of densities is L_1-consistent for f. See exercise 14.4.3.

14.1.1.1 Failure of Consistency

The risk with the parametric approach is that the parametric assumption is incorrect, in the sense that the parametric class doesn't contain the density generating the data or any good approximation. If this is the case, a parametric approach is typically not consistent. More precisely, if we estimate f with parametric class $\{f_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta} \in \Theta}$, then the L_p deviation between our estimate and f is bounded below by

$$\delta(f) := \inf_{\boldsymbol{\theta} \in \Theta} \|f - f_{\boldsymbol{\theta}}\|_p \tag{14.4}$$

This value will be zero only when f can be attained as the limit of elements of $\{f_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta} \in \Theta}$.

Example 14.1.2 Consider again the setting of 14.1.1 but now suppose that the true density f is not Gaussian. Then either the sequence $\hat{f}_N$ is not L_1-consistent for any density, or, if it is L_1-consistent for some density, then that density is not f. The reason is that $\delta(f)$ in (14.4) is always positive when the parametric class is Gaussian and f is not, since the set of normal densities is closed under the taking of limits in L_1.

14.1.2 Kernel Density Estimation

Sometimes we can make good choices for parametric classes by using descriptive statistics or by appealing to some theory with sharp quantitative implications. At other times this is difficult. In such settings it is best to use a nonparametric approach that is consistent under weaker assumptions.

1. For a proof, see p. 39 of Devroye and Lugosi (2001).

Let's start with an intuitive discussion and then turn to theory. Suppose that we have IID data $\mathbf{x}_1, \ldots, \mathbf{x}_N$ generated from unknown density f on $\mathbb{R}^d$. To estimate f using the data we will employ a **kernel density estimator** (KDE), which takes the form

$$\hat{f}_N(\mathbf{s}) := \frac{1}{Nh^d} \sum_{n=1}^{N} K\left(\frac{\mathbf{s} - \mathbf{x}_n}{h}\right) \tag{14.5}$$

Here K is called the **kernel function** of the estimator, and h is called the **bandwidth**. The kernel function K is required to be a density on $\mathbb{R}^d$ but is otherwise unrestricted. The bandwidth h is any positive number. Exercise 14.4.4 asks you to confirm that $\hat{f}_N$ is always a density.

To get a feeling for the estimate $\hat{f}_N$ in (14.5), let's look at a simple instance created from just three data points x_1, x_2, x_3 on $\mathbb{R}$. For K we take the standard normal density. Since $N = 3$, the function $\hat{f}_N$ is the sum of three individual functions, the nth of which can be written as

$$g_n(s) = \frac{1}{Nh} K\left(\frac{s - x_n}{h}\right) = \frac{1}{Nh} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(s - x_n)^2}{2h^2}\right\}$$

This is $\frac{1}{N}$ times the density of a $\mathrm{N}(x_n, h^2)$ random variable. So the three functions g_n are smooth bumps, each centered on one of the data points and with the degree of concentration around x_n governed by h. The top panel of figure 14.1 shows these three functions along with the data points for when $h = 1$. The black line is the pointwise sum of these functions, which is the density estimate $\hat{f}_N$. By construction, it is large near the data points and small away from data points.

The lower panel of figure 14.1 shows the same functions when the bandwidth is increased to 1.4. Each g_n becomes more spread out, and the sum $\hat{f}_N$ is smoother.

As discussed in more detail below, the role of the bandwidth is to add smoothing to the empirical distribution, so we can generalize from a finite sample. But what is the right amount of smoothing? Figure 14.2 illustrates the trade-off associated with smoothing. The shaded distribution is the density f, which we imagine to be unknown. 40 observations are drawn from this distribution and represented as dots. The black line in each panel is the KDE $\hat{f}_N$ built from these observations. The kernel is the standard normal density again, while the bandwidth varies across the panels.

As the bandwidth goes to zero, the kernel density becomes similar to the empirical distribution, with all probability mass in very small regions around the sample points. Hence, when we use a very small bandwidth , we make the mistake of treating the empirical distribution as the true distribution. This is overfitting. At the same time, excessive smoothing adds too much bias, hiding the features of the true distribution.[2]

2. Figure 14.2 is produced using `scikit-learn`. See `johnstachurski.net/emet.html` for code. In R, KDEs

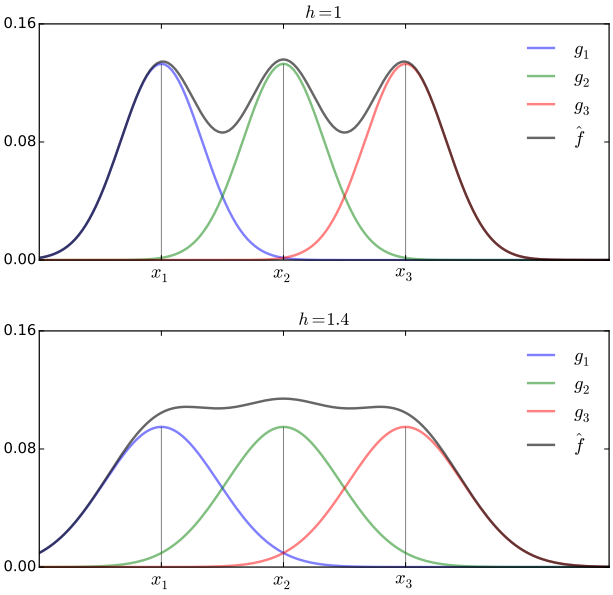

Figure 14.1 Nonparametric KDE, different bandwidths

The optimal bandwidth in terms of minimizing L_p deviation depends on the the unknown density f. For example, if f is smooth, then a relatively large bandwidth should be used. Of course f is unknown. As such there are two standard approaches. One is to make assumptions on f and choose the bandwidth accordingly. Another is to use cross-validation. For a review of both procedures see Scott (2015). We'll say more about cross-validation in the context of ridge regression below.

14.1.3 Theory

A natural way to study kernel density estimates is via the notion of convolutions. The convolution of an arbitrary distribution Q and a density K on $\mathbb{R}^d$ is the density on $\mathbb{R}^d$ defined by

$$(K \star Q)(\mathbf{s}') = \int K(\mathbf{s}' - \mathbf{s}) Q(\mathrm{d}\mathbf{s}) \qquad (\mathbf{s}' \in \mathbb{R}^d) \tag{14.6}$$

We already encountered this concept in the discrete case in exercise 5.4.10 on page 155, where you were asked to prove a version of

---

can be constructed from the `density` function. Try, for example, `plot(density(runif(200)))`.

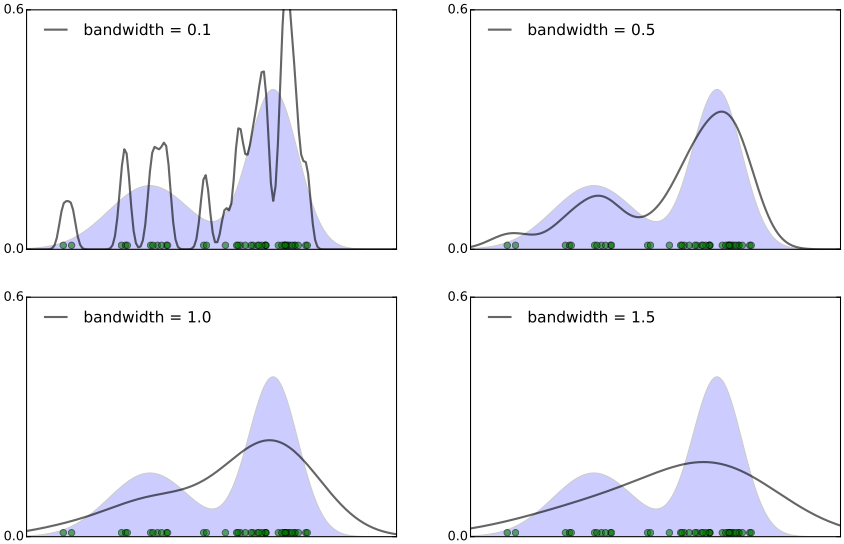

Figure 14.2 Effect of changing the bandwidth

Fact 14.1.3 For any density K and arbitrary distribution Q on $\mathbb{R}^d$, the density $K \star Q$ equals $\mathcal{L}(\mathbf{x} + \mathbf{y})$ when $\mathbf{x}$ and $\mathbf{y}$ are independent with $\mathcal{L}(\mathbf{x}) = K$ and $\mathcal{L}(\mathbf{y}) = Q$.

Example 14.1.3 If $K = \mathrm{N}(0, \sigma^2)$ for some $\sigma > 0$ and Q is a distribution on $\mathbb{R}$ that puts mass q_n on points $s_1, \ldots, s_N$, then by (14.6) and the rule for integrating over discrete distributions (see (5.14) on page 134),

$$(K \star Q)(s') = \sum_{n=1}^{N} K(s' - s_n) q_n \tag{14.7}$$

This distribution is a mixture of normals. Exercise 14.4.5 illustrates the connection between (14.7) and fact 14.1.3.

We will be particularly interested in a certain class of convolutions, induced by densities of the form

$$K_h(\mathbf{s}) := \frac{1}{h^d} K\left(\frac{\mathbf{s}}{h}\right) \tag{14.8}$$

where K is any density and $h > 0$ is a parameter. The density K_h in (14.8) is the density of $h\mathbf{x}$ when $\mathbf{x}$ is a random vector on $\mathbb{R}^d$ with density K. (See ex. 14.4.6.)

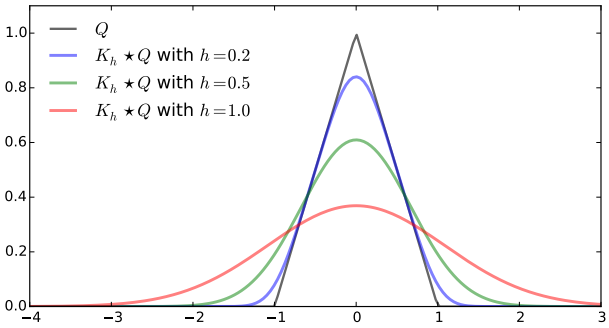

Figure 14.3 Smoothing induced by convolutions

Example 14.1.4 Figure 14.3 shows the convolution of a tent-shaped distribution Q and K_h when K is the standard normal density and h takes different values.

Note how in figure 14.3 the distribution of $K_h \star Q$ is close to Q when h is near zero. This makes sense, since $K_h \star Q$ is the distribution of $x + hy$ where x is drawn from Q and y is standard normal. As we take $h \downarrow 0$, the distribution of $x + hy$ converges to the distribution of x. In fact this is always true:

Fact 14.1.4 Let f and K be any densities on $\mathbb{R}^d$, and let K_h be defined from K via (14.8). If $f \in L_p$, then $K_h \star f \in L_p$, and

$$\lim_{h \downarrow 0} \|K_h \star f - f\|_p = 0$$

For a proof, see theorem 9.1 of Devroye and Lugosi (2001).

14.1.3.1 Convolution and KDEs

Let h be a positive number and let K_h be as defined in (14.8). Using this notation, we can rewrite the KDE in (14.5) as $\hat{f}_N(\mathbf{s}) = \frac{1}{N} \sum_{n=1}^N K_h(\mathbf{s} - \mathbf{x}_n)$. Letting $\hat{P}_N$ denote the empirical distribution of the sample and recalling the expression for integrals with respect to $\hat{P}_N$ (see page 219), we can also write this as $\hat{f}_N(\mathbf{s}') = \int K_h(\mathbf{s}' - \mathbf{s}) \hat{P}_N(d\mathbf{s})$, or, more simply,

$$\hat{f}_N = K_h \star \hat{P}_N \tag{14.9}$$

This equation says something interesting. In §8.2.1.2 we first tried to estimate the density by a naive application of the sample analogue principle. The method failed. In (14.9) we have an alternative method that employs smoothing, or regularization.

We apply the "smoothing operator" $K_h \star$ in addition to replacing the target distribution with the empirical distribution. As the next section shows, this method can be successful under very mild assumptions.

14.1.3.2 Consistency

It can be proved that, for *any* density f and kernel K, the nonparametric kernel density estimator $\hat{f}_N$ is L_1-consistent for f. See theorem 9.2 of Devroye and Lugosi (2001). The L_1 proof involves a large amount of measure-theoretic machinery. Here we'll state and prove the corresponding L_2 result instead.

Theorem 14.1.1 *Let f and K be densities on $\mathbb{R}^d$ and elements of L_2. If*

(i) $\{\mathbf{x}_n\}_{n \geqslant 1}$ is an IID *sequence of draws from f, and*

(ii) the bandwidth sequence $\{h_N\}$ satisfies $h_N \to 0$ and $Nh_N^d \to \infty$ as $N \to \infty$,

then the sequence of density estimates $\{\hat{f}_N\}$ defined in (14.5) is L_2-consistent for f.

Let's break this claim down into smaller pieces. For the remainder of this section, $\|\cdot\|$ is the L_2 norm and $h := h_N$. Using the expression for $\hat{f}_N$ in (14.9) and the triangle inequality in (14.3), we have

$$\|\hat{f}_N - f\| \leqslant \|K_h \star \hat{P}_N - K_h \star f\| + \|K_h \star f - f\| \tag{14.10}$$

The first term is called the **estimation error**, and the second is called the **approximation error** or **bias**. The first error is caused by the fact that we only observe the empirical distribution $\hat{P}_N$ rather than the true distribution f. The second error is caused by the smoothing we deliberately added to control the estimation error. (Compare with the error decomposition on page 304.)

The fact that $h_N \to 0$ in condition (ii) of theorem 14.1.1 means that we shrink the amount of smoothing as the sample size increases, thereby reducing the approximation error. The requirement $Nh_N^d \to \infty$ ensures that smoothing is not reduced too quickly, allowing us to control estimation error.

One way to think about the relationship between inference and the sample is that, when viewed as a "generalized density" with all its mass on data points, the empirical distribution is too rough. To generalize we need to add smoothing. At the same time, the need for smoothing decreases as the sample size gets large.

Proof of theorem 14.1.1. By fact 14.1.4, the approximation error converges to zero under the conditions of theorem 14.1.1. Hence it suffices to show that the estimation error converges in probability to zero. Fix $\delta > 0$. By Chebyshev's inequality (see page 96),

we have

$$\mathbb{P}\left\{\|K_h \star \hat{P}_N - K_h \star f\| \geqslant \delta\right\} = \mathbb{P}\left\{\|K_h \star \hat{P}_N - K_h \star f\|^2 \geqslant \delta^2\right\} \leqslant \frac{\xi_N}{\delta^2}$$

where

$$\xi_N := \mathbb{E}\left\{\|K_h \star \hat{P}_N - K_h \star f\|^2\right\}$$

To complete the proof, we need only show that ξ_N converges to zero. Let

$$\bar{K}_N(\mathbf{s}) := (K_h \star \hat{P}_N)(\mathbf{s}) - (K_h \star f)(\mathbf{s}) = \frac{1}{N}\sum_{n=1}^{N}\{K_h(\mathbf{s}-\mathbf{x}_n) - \mathbb{E}[K_h(\mathbf{s}-\mathbf{x}_n)]\}$$

We can then write

$$\xi_N = \mathbb{E}\left\{\int [\bar{K}_N(\mathbf{s})]^2 \,\mathrm{d}\mathbf{s}\right\} = \int \mathbb{E}\left\{[\bar{K}_N(\mathbf{s})]^2\right\}\mathrm{d}\mathbf{s} \tag{14.11}$$

(The interchange of order of expectation and integration is valid for nonnegative integrands—see theorem 4.4.5 of Dudley (2002).) Since $\bar{K}_N(\mathbf{s})$ is the sample mean of N IID zero-mean random variables, we have

$$\mathbb{E}\left\{[\bar{K}_N(\mathbf{s})]^2\right\} = \operatorname{var}[\bar{K}_N(\mathbf{s})] = \frac{1}{N}\operatorname{var}[K_h(\mathbf{s}-\mathbf{x}_n)]$$

Moreover

$$\operatorname{var}[K_h(\mathbf{s}-\mathbf{x}_n)] = \mathbb{E}\left\{[K_h(\mathbf{s}-\mathbf{x}_n)]^2\right\} - \{\mathbb{E}[K_h(\mathbf{s}-\mathbf{x}_n)]\}^2 \leqslant \mathbb{E}\left\{[K_h(\mathbf{s}-\mathbf{x}_n)]^2\right\}$$

In summary,

$$\xi_N \leqslant \frac{1}{N}\int \mathbb{E}\left\{[K_h(\mathbf{s}-\mathbf{x}_n)]^2\right\}\mathrm{d}\mathbf{s} \tag{14.12}$$

Now observe that, switching the order of integration once more,

$$\int \mathbb{E}\left\{[K_h(\mathbf{s}-\mathbf{x}_n)]^2\right\}\mathrm{d}\mathbf{s} = \int\left\{\int [K_h(\mathbf{s}-\mathbf{s}')]^2\,\mathrm{d}\mathbf{s}\right\} f(\mathbf{s}')\,\mathrm{d}\mathbf{s}'$$

From the definition of K_h and a change of variable argument,

$$\int [K_h(\mathbf{s}-\mathbf{s}')]^2\,\mathrm{d}\mathbf{s} = \frac{1}{h^{2d}}\int\left[K\left(\frac{\mathbf{s}-\mathbf{s}'}{h}\right)\right]^2\mathrm{d}\mathbf{s} = \frac{1}{h^d}\int [K(\mathbf{u})]^2\,\mathrm{d}\mathbf{u}$$

Putting this together with (14.12) gives the bound

$$\xi_N \leqslant \int \frac{1}{Nh^d} \|K\|^2 f(\mathbf{s}') \, \mathrm{d}\mathbf{s}' = \frac{1}{Nh^d} \|K\|^2$$

The term $\|K\|^2 := \|K\|_2^2$ is finite by assumption. Recalling that $Nh^d = Nh_N^d \to \infty$, we see that $\xi_N \to 0$ as required. □

14.1.4 Commentary

In some fields of science, researchers have considerable knowledge about parametric classes and specific functional forms. For example, the theory of Brownian motion describes how the location of a tiny particle in liquid is approximately normally distributed. Thus the underlying theory provides an exact parametric class. In this kind of setting the parametric approach excels. It allows us to generalize by combining data with information we have on functional forms.

Unfortunately, the quantitative foundations of economics and other social sciences are looser and more prone to shifting over time. Econometricians usually come to the table with less knowledge of parametric classes and functional forms. These facts make nonparametric techniques attractive.

At the same time, nonparametric methods do not solve all our problems. The theoretical results presented above are purely asymptotic. Finite sample results are available, but it isn't possible to obtain strong results in this direction without correspondingly strict assumptions on the target density. There is no uniform rate of convergence for all target densities (Devroye and Lugosi 2001, p. 85). This is because nonparametric methods have relatively little structure in the form of prior knowledge and hence require abundant data.

14.2 Controlling Complexity

In the last section we discussed flexible estimation methods that perform well asymptotically. In this section we turn our attention to finite sample properties. A common thread will be ridge regression, which is a popular method of estimation in both econometrics and machine learning (see, e.g., Peysakhovich and Naecker 2015, Kim and Swanson 2014, or Varian 2014). Ridge regression also connects to some deep ideas at the heart of finite sample theory, including complexity, prior knowledge and the bias–variance trade-off.

14.2.1 Ridge Regression

Let's start off in the classical OLS setting of §12.1.1, with assumptions 12.1.2–12.1.4 all in force. The usual OLS estimator is $\hat{\boldsymbol{\beta}}$ as given in (12.3). It can also be expressed as

$$\hat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\mathbf{b} \in \mathbb{R}^K} \sum_{n=1}^{N} (y_n - \mathbf{x}_n^\top \mathbf{b})^2$$

As shown in §11.1.2, the OLS estimator minimizes the empirical risk under quadratic loss when the hypothesis space is the set of linear functions. Under our current assumptions, it is also unbiased for $\boldsymbol{\beta}$ (theorem 12.1.1) and, by the Gauss–Markov theorem, it has the lowest variance among all linear unbiased estimators of $\boldsymbol{\beta}$ (theorem 12.1.3).

While the Gauss–Markov theorem has historical importance, a more natural way to evaluate estimators is to consider their mean squared error, which tells us directly how much probability mass the estimator puts around the object it is trying to estimate (see §9.2.3). Recalling (9.9) on page 259, the MSE of an estimator $\hat{\mathbf{b}}$ of $\boldsymbol{\beta}$ is defined as

$$\operatorname{mse}(\hat{\mathbf{b}}, \boldsymbol{\beta}) := \mathbb{E}\left\{ \|\hat{\mathbf{b}} - \boldsymbol{\beta}\|^2 \right\}$$

As exercise 14.4.8 asks you to show, the following representation is also valid:

$$\operatorname{mse}(\hat{\mathbf{b}}, \boldsymbol{\beta}) = \mathbb{E}\left\{ \| \hat{\mathbf{b}} - \mathbb{E}[\hat{\mathbf{b}}] \|^2 \right\} + \| \mathbb{E}[\hat{\mathbf{b}}] - \boldsymbol{\beta} \|^2 \tag{14.13}$$

This equation extends (9.10) on page 259, and tells us that MSE is the sum of a variance and a bias term. Minimization of MSE involves a trade-off between these two terms. Typically, the optimal choice is not at either extreme: MSE is minimized when some amount of bias is admitted.

Applying this idea to the OLS setting, Hoerl and Kennard (1970) showed that there exists a biased linear estimator that has lower mean squared error than $\hat{\boldsymbol{\beta}}$. The estimator is defined as the solution to the modified least squares problem

$$\min_{\mathbf{b} \in \mathbb{R}_K} \left\{ \sum_{n=1}^{N} (y_n - \mathbf{x}_n^\top \mathbf{b})^2 + \lambda \|\mathbf{b}\|^2 \right\} \tag{14.14}$$

where $\lambda \geqslant 0$ is called the **regularization parameter**. In solving (14.14), we are minimizing the empirical risk plus a term that penalizes large values of $\|\mathbf{b}\|$. The effect is to "shrink" the solution relative to the unpenalized solution $\hat{\boldsymbol{\beta}}$. Some calculus shows

that the solution to (14.14) is

$$\hat{\boldsymbol{\beta}}_\lambda := (\mathbf{X}^\intercal \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\intercal \mathbf{y} \tag{14.15}$$

The estimator $\hat{\boldsymbol{\beta}}_\lambda$ is called the **ridge regression estimator**. Note that

(i) $\hat{\boldsymbol{\beta}}_\lambda$ is the OLS estimator when $\lambda = 0$ and

(ii) $\hat{\boldsymbol{\beta}}_\lambda$ is biased whenever $\lambda > 0$ (ex. 14.4.9).

The following result is proved in Hoerl and Kennard (1970).

Theorem 14.2.1 *Under the OLS assumptions 12.1.2–12.1.4, there exists a $\lambda > 0$ such that*

$$\text{mse}\left(\hat{\boldsymbol{\beta}}_\lambda, \boldsymbol{\beta}\right) < \text{mse}\left(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}\right)$$

The reduction in MSE over the least squares estimator occurs because, for some intermediate value of λ, the variance of $\hat{\boldsymbol{\beta}}_\lambda$ falls by more than enough to offset the bias induced by regularization.

Note that, for suitable choice of λ, the ridge regression estimator $\hat{\boldsymbol{\beta}}_\lambda$ outperforms $\hat{\boldsymbol{\beta}}$ even though all of the classical OLS assumptions are valid. At the same time the right choice of λ is a nontrivial problem. This problem falls under the heading of model selection, which is the topic treated in the next few sections.

14.2.1.1 Interpretation

The traditional view of ridge regression runs as follows: The standard OLS assumptions are treated as valid. There may, however, be instances where $\mathbf{X}^\intercal \mathbf{X}$ is almost singular due to strong correlation between regressors. In this case the process of inverting $\mathbf{X}^\intercal \mathbf{X}$ is numerically unstable. We can stabilize the inversion by adding some positive value of λ in (14.15).

Here's another view: The standard OLS assumptions are implausible. Since our loss function is quadratic, we would ideally like to obtain the regression function f^* but recovering this infinite dimensional object with a finite amount of data is ill-posed. We need to control the complexity of the candidate functions we use to approximate the regression function. The regularization term in ridge regression provides a means of managing complexity.

14.2.1.2 Tikhonov Regularization

Let's build some more intuition for the result in theorem 14.2.1 by running simulations. We'll frame the simulations in a general setting, of which ridge regression is a special case.

We know that the least squares estimator is the solution to an overdetermined system of equations. There is an existing theory for solving ill-posed linear systems in high dimensions. The basic idea is that any attempt to back out or infer a complex object by solving a system about which we have limited information requires a degree of regularization.

To illustrate, suppose that

(i) $\mathbf{Ab} = \mathbf{c}$ is an overdetermined system, where $\mathbf{A}$ is $N \times K$ with $N > K$.

(ii) Due to measurement error, we only observe an approximation $\mathbf{c}_0$ of $\mathbf{c}$.

(iii) $\mathbf{b}^*$ is the (unobservable) least squares solution $\operatorname{argmin}_{\mathbf{b}} \|\mathbf{Ab} - \mathbf{c}\|^2$.

In the absence of additional information, a natural approach to approximating $\mathbf{b}^*$ is to solve $\mathbf{Ab} = \mathbf{c}_0$ by least squares. An alternative, less obvious approach, is to minimize

$$m(\lambda) := \|\mathbf{Ab} - \mathbf{c}_0\|^2 + \lambda \|\mathbf{b}\|^2 \tag{14.16}$$

for some small but positive λ. This second approach is called **Tikhonov regularization**. We are minimizing least squares plus a penalty term.

Let's look at a simulation where $\mathbf{A}$ is chosen stochastically but with a tendency towards multicollinearity.[3] To aid clarity, we first set $\mathbf{b}^* := (10, 10, \ldots, 10)^\intercal$, and then set $\mathbf{c} := \mathbf{Ab}^*$. By construction, $\mathbf{b}^*$ is a solution to the system $\mathbf{Ab}^* = \mathbf{c}$, and also the least squares solution.

Measurement of $\mathbf{c}$ is corrupted with a Gaussian shock. In particular, $\mathbf{c}_0$ is drawn from $\mathrm{N}(\mathbf{c}, \sigma^2 \mathbf{I})$ where σ is a small positive number. We then plot the OLS solution based on $\mathbf{c}_0$, which minimizes $\|\mathbf{Ab} - \mathbf{c}_0\|^2$, and the regularized solution, which minimizes $m(\lambda)$ in (14.16). The former are plotted against their index in green, while the latter are plotted in blue. The true solution $\mathbf{b}^*$ is plotted in black. The result is figure 14.4.

The figure shows 10 solutions each for the ordinary and regularized solutions, corresponding to 10 draws of $\mathbf{c}_0$. The regularized solutions are much closer to the true solution on average.

Note that this result is dependent on a reasonable choice for λ. If you experiment with the code on the text website, you will see that for very small values of λ, the regularized solutions are almost the same as the unregularized solutions. Conversely, very large values of λ pull the regularized solutions too close to the zero vector.

14.2.2 Subset Selection and Ridge Regression

One problem frequently faced in regression problems is which variables to include. For example, if we are comparing crime rates across different cites, we can think of

3. Full details on the choice of $\mathbf{A}$ are given in the code for this chapter at `johnstachurski.net/emet.html`.

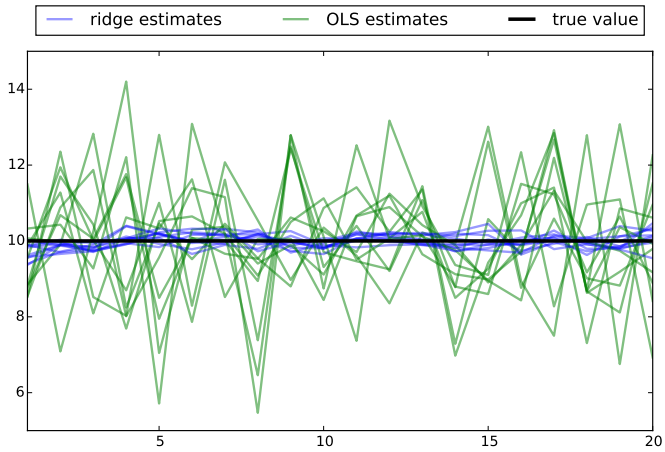

Figure 14.4 Effect of Tikhonov regularization, $\lambda = 1$

any number of variables that might be relevant (median wage, unemployment, police density, etc.). The same is true if we are trying to model credit default rates for some group of individuals, or educational attainment across schools. A similar problem arises in time series models, where we want to know how many lags of the state variables to include. The general problem is known as **subset selection**, since we are trying to choose the right subset of all candidate regressors.

Further dimensions appear when we consider basis functions. Given a set of covariates $\mathbf{x}$, we have the option to map this into a larger vector $\boldsymbol{\phi}(\mathbf{x})$ using basis functions, as discussed in §11.2.1. For example, given a single covariate x, we may consider mapping it into $\boldsymbol{\phi}(x) = (1, x, x^2, \ldots, x^d)$ and regressing y on $\boldsymbol{\phi}(x)$. As we saw in figures 8.5–8.8 (page 231), a good choice of d is crucial. Choosing d is another example of the subset selection problem because we are trying to decide whether to include the regressor x^j for some given j.

Subset selection is a version of the empirical risk minimization problem. Suppose that we have output y and inputs $\mathbf{x} \in \mathbb{R}^K$, in the sense that $\mathbf{x}$ contains K candidate regressors. If we want to include all regressors, then we can minimize empirical risk over $\mathcal{H}_\ell$ in (11.4), the hypothesis space of linear functions from $\mathbb{R}^K$ to $\mathbb{R}$. If we wish to exclude some subset of regressors, we can set $I \subset \{1, \ldots, K\}$ to be the set of indices of the regressors we want to exclude and regress y on the remainder, which is equivalent to minimizing the empirical risk over the hypothesis space

$$\mathcal{H}_{-I} := \{ \text{ all functions } f(\mathbf{x}) = \mathbf{b}^\mathsf{T}\mathbf{x} \text{ with } b_k = 0 \text{ for all } k \in I \}$$

We are back to the problem of choosing a suitable hypothesis space over which to minimize empirical risk.

The subset selection problem has been tackled by many researchers. Well-known approaches include the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), and Mallow's C_p statistic. For example, Mallow's C_p statistic consists of two terms, one increasing in the size of the empirical risk, and the other increasing in $|I|$, the size of the subset selected. The objective is to minimize the statistic, which involves trading off poor fit (large empirical risk) against excess complexity of the hypothesis space (large $|I|$).

One of the problems with subset selection is that K regressors means 2^K subsets to step through.[4] To avoid this computational problem, one alternative is to use ridge regression. With ridge regression, the regularization term leads us to choose an estimate with smaller norm. What this means, in practice, is that the coefficients of less helpful regressors are driven towards zero, thereby "almost excluding" those regressors. While the model selection problem is not solved, it has been reduced to tuning a single parameter.

We can illustrate the idea by reconsidering the regression problem discussed in §8.2.3. Figures 8.5–8.8 (see page 231) showed the fit we obtained by minimizing empirical risk over larger and larger hypothesis spaces. The hypothesis spaces were the sets $\mathscr{P}_d$ of degree d polynomials for different values of d. For each d we minimized the empirical risk over $\mathscr{P}_d$, which translates into solving

$$\min_{\mathbf{b}} \sum_{n=1}^{N} \left[y_n - \mathbf{b}^\mathsf{T}\boldsymbol{\phi}(x_n)\right]^2 \quad \text{where} \quad \boldsymbol{\phi}(x) = (x^0, x^1, \ldots, x^d)$$

As discussed above, choosing the right d is isomorphic to subset selection, since we are deciding what powers of x to include as regressors. Figure 8.4 (page 230) showed that intermediate values of d did best at minimizing risk.

We can do a similar thing using ridge regression. First, let's take $\mathscr{P}_{14}$ as our hypothesis space. This space is certainly large enough to provide a good fit to the data, but with empirical risk minimization the result is overfitting (see figure 8.8 on page 233). Here, instead of using empirical risk minimization, we solve the regularized problem

$$\min_{\mathbf{b}} \sum_{n=1}^{N} \left\{ \left[y_n - \mathbf{b}^\mathsf{T}\boldsymbol{\phi}(x_n)\right]^2 + \lambda\|\mathbf{b}\|^2 \right\}$$

for different values of λ. The data used here are exactly the same data used in the original figures 8.5–8.8 from §8.2.3. The solution for each λ we denote by $\hat{\boldsymbol{\beta}}_\lambda$, which is the ridge regression estimator, and the resulting prediction function we denote by $\hat{f}_\lambda$,

4. Remember Sala-i-Martin (1997) and his two million growth regressions?

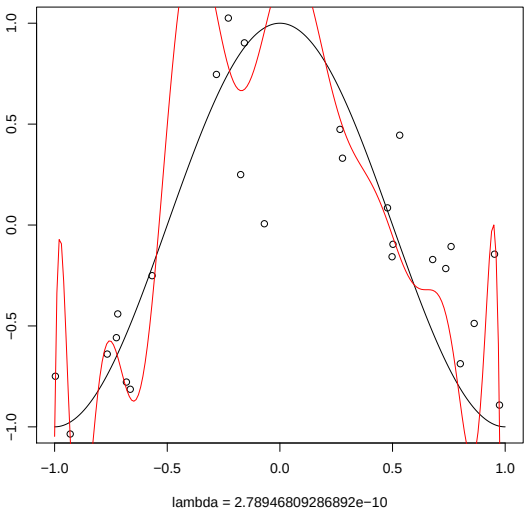

Figure 14.5 Fitted polynomial, $\lambda \approx 0$

so that $\hat{f}_\lambda(x) = \hat{\boldsymbol{\beta}}_\lambda^\mathsf{T} \boldsymbol{\phi}(x)$.

The function $\hat{f}_\lambda$ is plotted in red for increasingly larger values of λ over figures 14.5–14.7. The black line is the risk-minimizing function. In figure 14.5, the value of λ is too small to impose any real restriction, and the procedure overfits. In figure 14.6, the value of λ is larger and the fit is good. In figure 14.7 the value of λ is too large and the estimate is poor.

As in §8.2.3, we can compute the risk of each function $\hat{f}_\lambda$, since we know the underlying model (see (8.25) on page 229). The risk is plotted against λ in figure 14.8. The x-axis is on log-scale. The risk is smallest for small but nonzero values of λ.

14.2.3 Bayesian Methods and Regularization

The ideal case with model selection is that we have clear guidance from economic theory on which regressors to include, which functional forms to use, which values of our regularization parameter to choose, and so on. If theory or prior knowledge provides this information, then every effort should be made to exploit it. One technique for injecting prior information into statistical estimation is via Bayesian analysis. Let's now look at Bayesian linear regression and how it compares to ridge regression.

Suppose that our regression data take the linear form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$. To simplify the

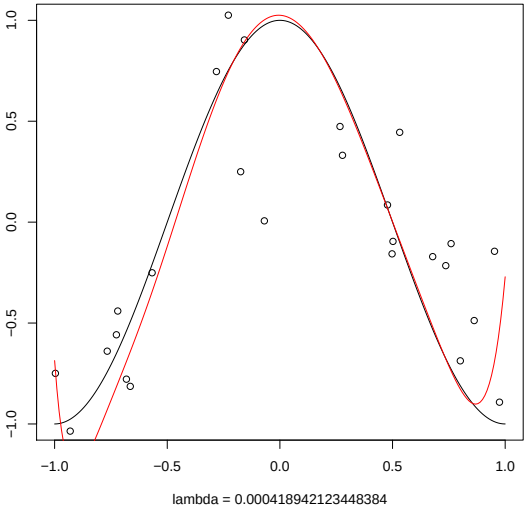

Figure 14.6 Fitted polynomial, $\lambda \approx 0.0004$

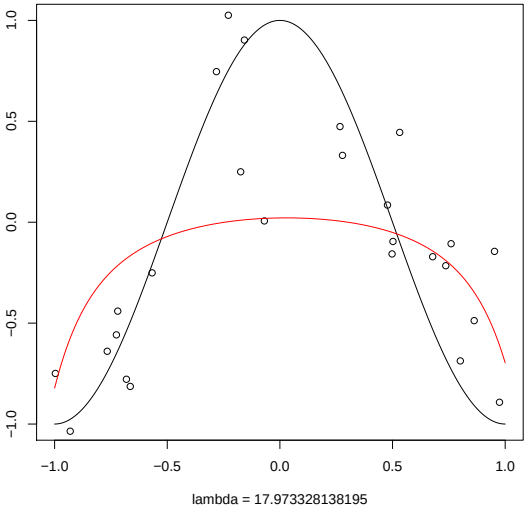

Figure 14.7 Fitted polynomial, $\lambda \approx 18$

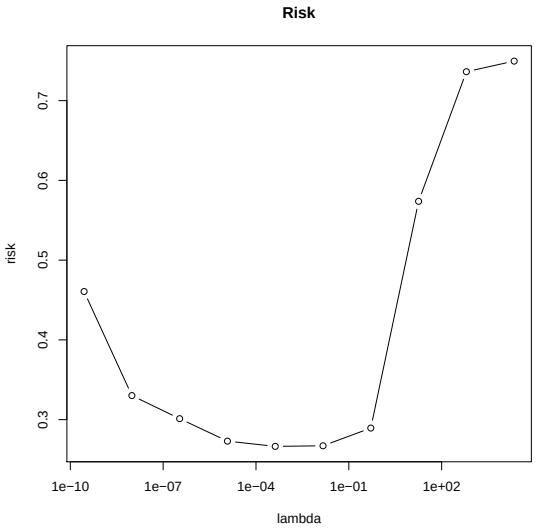

Figure 14.8 Risk of $\hat{f}_\lambda$ plotted against λ

presentation we will assume that $\mathbf{X}$ is nonrandom. (Taking $\mathbf{X}$ to be random leads to the same conclusions but with a longer derivation). As before, $\mathbf{u}$ is random and unobservable. The new feature provided by the Bayesian perspective is that we take $\boldsymbol{\beta}$ to be random and unobservable as well. In addition we are assumed to have subjective prior beliefs regarding likely values for these variates, expressed in the form of probability distributions. Here we take the priors to be $\mathcal{L}(\mathbf{u}) = \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$ and $\mathcal{L}(\boldsymbol{\beta}) = \mathrm{N}(\mathbf{0}, \tau^2\mathbf{I})$.

Given our model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, our prior on $\mathbf{u}$ implies that the density of $\mathbf{y}$ given $\boldsymbol{\beta}$ is $\mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. In generic notation, we can write our distributions as

$$p(\mathbf{y} \mid \boldsymbol{\beta}) = \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}) \quad \text{and} \quad p(\boldsymbol{\beta}) = \mathrm{N}(\mathbf{0}, \tau^2\mathbf{I}) \tag{14.17}$$

Applying Bayes' law (see (5.25)) to the pair $(\mathbf{y}, \boldsymbol{\beta})$, we obtain

$$p(\boldsymbol{\beta} \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \boldsymbol{\beta})p(\boldsymbol{\beta})}{p(\mathbf{y})} \tag{14.18}$$

The left-hand side is the posterior density of $\boldsymbol{\beta}$ given the data $\mathbf{y}$, and represents our new beliefs updated from the prior on the basis of the data $\mathbf{y}$.

One way to summarize the information contained in the posterior is by examining its maximum value. The maximizer of the posterior is called the **maximum a poste-**

riori (MAP) probability estimate. Taking logs of (14.18) and dropping the term that does not contain $\boldsymbol{\beta}$, it can be expressed as

$$\hat{\boldsymbol{\beta}}_M := \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \{\ln p(\mathbf{y} \mid \boldsymbol{\beta}) + \ln p(\boldsymbol{\beta})\} \tag{14.19}$$

Inserting our functional forms into (14.17), dropping constant terms and multiplying by -1, we obtain the expression

$$\hat{\boldsymbol{\beta}}_M = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{n=1}^{N} (y_n - \mathbf{x}_n^\mathsf{T}\boldsymbol{\beta})^2 + \frac{\sigma^2}{\tau^2} \|\boldsymbol{\beta}\|^2 \right\} \tag{14.20}$$

This is precisely the penalized least squares problem (14.14) on page 387, where the regularization parameter λ is equal to $(\sigma/\tau)^2$. In view of (14.15), the solution is

$$\hat{\boldsymbol{\beta}}_M := (\mathbf{X}^\mathsf{T}\mathbf{X} + (\sigma/\tau)^2\mathbf{I})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$$

Thus Bayesian estimation provides a principled derivation of the penalized objective function from ridge regression. Bayesian analysis provides the same effect as Tikhonov regularization, but now regularization arises out of combining prior knowledge with the data. Moreover, at least in principle, the value $(\sigma/\tau)^2$ is part of our prior knowledge, and hence there is no model selection problem.

In practice, one could of course question the assertion that we have so much prior knowledge that the regularization parameter $\lambda := (\sigma/\tau)^2$ is pinned down. If such knowledge is lacking, then we are back at the model selection problem. In the next section we forgo the assumption that this strong prior knowledge is available, and consider a more automated approach to choosing λ.

14.2.4 Cross-Validation

A natural way to think about model selection is to think about minimizing risk. Recall that, given loss function L and a system producing input–output pairs $(\mathbf{x}, y) \in \mathbb{R}^{K+1}$ with joint distribution P, the prediction risk of a function $f\colon \mathbb{R}^K \to \mathbb{R}$ is the expected loss

$$R(f) := \mathbb{E}[L(y, f(\mathbf{x}))] = \int\int L(t, f(\mathbf{s}))P(\mathrm{d}t, \mathrm{d}\mathbf{s})$$

that occurs when we use $f(\mathbf{x})$ to predict y. Now suppose that we observe N IID input–output pairs $\mathbf{z}_\mathcal{D} := \{(\mathbf{x}_1, y_1), \ldots (\mathbf{x}_N, y_N)\}$. Given a selection of models, we would like to find the one that takes this data set and returns a predictor $\hat{f}$ such that $\hat{f}$ has lower prediction risk than the predictors returned by the other models.

We have to be precise in defining risk because, if we define it as $\mathbb{E}[L(y, \hat{f}(\mathbf{x}))]$, then

we are taking expectation over all randomness, including that in $\hat{f}$, which depends on the data set $\mathbf{z}_{\mathcal{D}}$. What we want to do now is take the data set as given, and see how well we can do in terms of predicting new values as evaluated by expected loss. Hence we define the prediction risk of $\hat{f}$ to be the expected loss taking $\mathbf{z}_{\mathcal{D}}$ (and hence $\hat{f}$) as given:

$$R(\hat{f} \mid \mathcal{D}) := \mathbb{E}[L(y, \hat{f}(\mathbf{x})) \mid \mathcal{D}] = \int \int L(t, \hat{f}(\mathbf{s}))P(\mathrm{d}t, \mathrm{d}\mathbf{s})$$

If we have a collection of models M indexed by m, and $\hat{f}_m$ is the predictor produced by fitting model m with data $\mathcal{D}$, then we would like to find the model m^* such that

$$R(\hat{f}_{m^*} \mid \mathcal{D}) \leqslant R(\hat{f}_m \mid \mathcal{D}) \quad \text{for all } m \in M$$

The obvious problem with this idea is that risk is unobservable. If we knew the joint distribution P then we could calculate it; but then again, if we knew P there would be no need to estimate anything in the first place.

Looking at this problem, you might have the following idea: Although we don't know P, we do have the data $\mathbf{z}_{\mathcal{D}}$, which consists of IID draws from P. From the law of large numbers, we know that expectations can be approximated by averages over IID draws, so we could approximate $R(\hat{f} \mid \mathcal{D})$ by

$$\frac{1}{N} \sum_{n=1}^{N} L(y_n, \hat{f}(\mathbf{x}_n))$$

where the pairs $(\mathbf{x}_n, y_n)$ are from the data set $\mathbf{z}_{\mathcal{D}}$.

However, this is just the empirical risk, and the empirical risk is a highly biased estimator of the risk. This point was discussed extensively in §8.2.3. See, in particular, figure 8.4 on page 230. The point that figure made was that complex models tend to overfit, producing low empirical risk, but high risk. In essence, the problem is that we are using the data $\mathbf{z}_{\mathcal{D}}$ twice, for conflicting objectives. First, we are using it to fit the model, producing $\hat{f}$. Second, we are using it to evaluate the predictive ability of $\hat{f}$ on new observations.

So what we really need is fresh data. New data will tell us how $\hat{f}$ performs out-of-sample. If we had J new observations $(y_j^v, \mathbf{x}_j^v)$, then we could estimate the risk by

$$\frac{1}{J} \sum_{j=1}^{J} L(y_j^v, \hat{f}(\mathbf{x}_j^v))$$

Of course, this isn't a genuine solution because we don't have any new data in general. One way to work around this problem is to take $\mathbf{z}_{\mathcal{D}}$ and split it into two disjoint subsets, called the **training set** and the **validation set**. The training set is used to fit

$\hat{f}$ and the validation set is used to estimate the risk of $\hat{f}$. We then repeat this for all models and choose the one with lowest estimated risk.

Since data are scarce, a more common procedure is **cross-validation**, which attempts to use the whole data set for both fitting the model and estimating the risk. To illustrate the idea, suppose that we partition the data set into two subsets $\mathcal{D}_1$ and $\mathcal{D}_2$. First, we use $\mathcal{D}_1$ as the training set and $\mathcal{D}_2$ as the validation set. Next we use $\mathcal{D}_2$ as the training set, and $\mathcal{D}_1$ as the validation set. The estimate of the risk is the average of the estimates of the risk produced in these two steps.

Of course, we could divide the data into more than two sets. The extreme is to partition the data into N subsets. This procedure is called **leave-one-out cross-validation**. Letting $\mathcal{D}_{-n} := \mathbf{z}_{\mathcal{D}} \setminus \{(\mathbf{x}_n, y_n)\}$, the data set with just the nth data point $(\mathbf{x}_n, y_n)$ omitted, the leave-one-out cross validation algorithm can be expressed as follows:

1: **for** $n = 1, \ldots, N$ **do**
2: fit $\hat{f}_{-n}$ using data $\mathcal{D}_{-n}$
3: set $r_n := L(y_n, \hat{f}_{-n}(\mathbf{x}_n))$
4: **end for**
5: return the risk estimate $r := \frac{1}{N}\sum_{n=1}^{N} r_n$

At each step inside the loop, we fit the model using all but the nth data point, and then predict the nth data point using the fitted model. The prediction quality is evaluated in terms of loss. Repeating this n times, we produces estimate of the risk using average loss. On an intuitive level, the procedure is attractive because we are using the available data intensively but still evaluating based on out-of-sample error.

In terms of model selection, the idea is to run each model through the cross-validation procedure, and then select the one that produces the lowest value of r, the estimated risk. Let's illustrate this idea, by considering again the ridge regression procedure used in §14.2.2. In this problem the set of models is indexed by λ, the regularization parameter in the ridge regression. The data set $\mathbf{z}_{\mathcal{D}}$ is the set of points shown in figures 14.5–14.7. For each λ, the fitted function $\hat{f}_\lambda$ is

$$\hat{f}_\lambda(x) = \hat{\boldsymbol{\beta}}_\lambda^{\mathsf{T}} \boldsymbol{\phi}(x) \quad \text{where} \quad \hat{\boldsymbol{\beta}}_\lambda := \operatorname*{argmin}_{\mathbf{b}} \sum_{n=1}^{N} \left\{ (y_n - \mathbf{b}^{\mathsf{T}} \boldsymbol{\phi}(x_n))^2 + \lambda \|\mathbf{b}\|^2 \right\}$$

Recall here that $\boldsymbol{\phi}(x) = (x^0, x^1, \ldots, x^d)$ with d fixed at 14, so we are fitting a polynomial of degree 14 to the data by minimizing regularized least squares error. The amount of regularization is increasing in λ. The resulting functions $\hat{f}_\lambda$ were shown for different values of λ in figures 14.5–14.7. Intermediate values of λ produced the best fit in terms of minimizing risk (see figures 14.6 and 14.8).

In that discussion, we used the fact that we knew the underlying model to evaluate the risk, and hence the values of λ that produce low risk (figure 14.8). In real

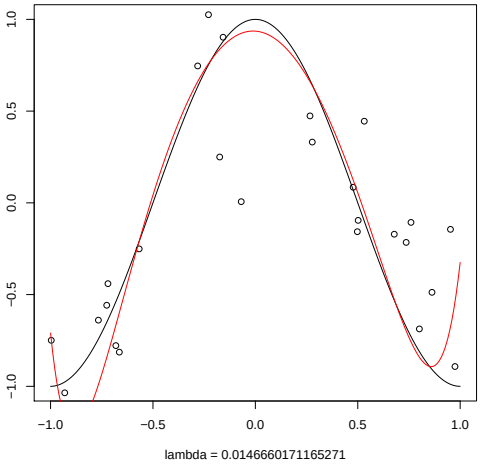

Figure 14.9 Fitted polynomial, $\lambda \approx 0.015$

estimation problems, risk is unobservable, and we need to choose λ on the basis of the data alone (assuming we don't have prior knowledge, as in the Bayesian case—see §14.2.3). Let's see how a data-based procedure such as cross-validation performs in terms of selecting a good value of λ.

In this experiment, for each λ in the grid `exp(seq(-22, 10, length=10))` we perform leave-one-out cross-validation. The fit at each step within the loop is via ridge regression, omitting the nth data point, and the resulting polynomial is used to predict y_n from x_n. The prediction error is measured by squared loss. In other words, for each λ in the grid, we use the following algorithm to estimate the risk:

1: **for** $n = 1, \dots, N$ **do**
2: set $\hat{\boldsymbol{\beta}}_{\lambda,-n} := \operatorname{argmin}_{\mathbf{b}} \sum_{i \neq n} \{(y_i - \mathbf{b}^\intercal \boldsymbol{\phi}(x_i))^2 + \lambda \|\mathbf{b}\|^2\}$
3: set $r_{\lambda,n} := (y_n - \hat{\boldsymbol{\beta}}_{\lambda,-n}^\intercal \boldsymbol{\phi}(x_n))^2$
4: **end for**
5: return $r_\lambda := \frac{1}{N} \sum_{n=1}^{N} r_{\lambda,n}$

The value of λ producing the smallest estimated risk r_λ is around 0.015. This is in fact very close to the value that minimizes the actual risk (see figure 14.8 on page 394). The associated function $\hat{f}_\lambda$ is plotted in red in figure 14.9, and indeed the fit is close. In this instance, our fully automated procedure is successful.[5]

5. The code to run this experiment can be found at `johnstachurski.net/emet.html`.

14.3 Further Reading

Textbook treatments of nonparametric methods in econometrics can be found in Pagan and Ullah (1999), Li and Racine (2006), Ullah (1989), Henderson and Parmeter (2015) and Hansen (2015). Bosq (1996) covers nonparametrics for time series models. Recent studies in the econometrics literature using nonparametric or semiparametric methods include Chen and Hong (2012), Jeong et al. (2012), Henderson et al. (2012), Christensen (2014), Canay et al. (2013), Su and Ullah (2013), Bajari et al. (2013), Newey (2013), Hansen (2014a), Mastromarco and Simar (2015), Hickman and Hubbard (2015), Matzkin (2015), Bhattacharya (2015), and Du and Escanciano (2015). A beautiful overview of nonparametrics and the method of sieves can be found in Geman and Hwang (1982).

Good discussions of ridge regression and cross-validation can be found in Friedman et al. (2009) and Abu-Mostafa et al. (2012). One alternative to ridge regression is the LASSO method (Tibshirani 1996), where the term $\lambda \sum_{k=1}^{K} b_k^2$ in (14.14) is replaced with an absolute deviation penalty $\lambda \sum_{k=1}^{K} |b_k|$. Also related is penalized maximum likelihood estimation. A recent application can be found in Gentzkow et al. (2015).

An important contribution to the literature on subset selection is the least angle regression of Efron et al. (2004). A discussion of this method in relation to other subset techniques can be found in Hesterberg et al. (2008).

Recent papers treating regularization and related topics within the econometrics literature include Florens (2003), Carrasco et al. (2007), Chen and Reiss (2011), Florens and Simoni (2014), Darolles et al. (2011), Hoderlein and Holzmann (2011), Horowitz (2014), Hautsch et al. (2012), Li et al. (2015), Lunde et al. (forthcoming), Chernozhukov et al. (2015), and Ando and Bai (forthcoming).

Regarding implementations, one powerful and increasingly popular package for Bayesian learning is Stan. See `http://mc-stan.org/`. A popular and high-quality routine for cross validation in the context of ridge regression can be found in the `sklearn.linear_model.RidgeCV` method in Python's `scikit-learn` package.

14.4 Exercises

Ex. 14.4.1 Let $\|\cdot\|_1$ denote the L_1 norm as defined in (14.2). Show that $\|f - g\|_1$ is bounded over the set of all densities.[6]

Ex. 14.4.2 Prove the triangle inequality in the case of L_1 (see (14.3)).

Ex. 14.4.3 Prove the claim in example 14.1.1 on page 379.[7]

6. Hint: It is a basic property of integration that $g \leqslant h$ pointwise implies $\int g \leqslant \int h$.
7. Hint: Use fact 14.1.2 and exercise 8.5.9 on page 245.

Ex. 14.4.4 For the case $d = 1$, show that $\hat{f}$ in (14.5) is a density for every N, every $h > 0$ and every realization of the sample.[8]

Ex. 14.4.5 Let x be a finite random variable taking value s_n in $\{s_n\}_{n=1}^N$ with probability p_n. Let y be independent of x with density g on $\mathbb{R}$. Show that $z = x + y$ has density $f(s') = \sum_{n=1}^N g(s' - s_n)p_n$.

Ex. 14.4.6 Let $\mathbf{x}$ be multivariate Gaussian, with distribution $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and let $h > 0$. Suppose that $\boldsymbol{\Sigma}$ is nonsingular. Let K be the density of $\mathbf{x}$ (see page 130) and let K_h be the density of $h\mathbf{x}$. Show that K and K_h satisfy the relationship shown in (14.8). (The result is true more generally but the Gaussian case is a nice exercise.)

Ex. 14.4.7 Let $x_1, \ldots, x_N$ be IID scalar random variables with common density f and let $\hat{f}$ be as defined in (14.5). Show that if $K(t) = \mathbb{1}\{-1/2 < t < 1/2\}$, then, for any number s, we have

$$\mathbb{E}[\hat{f}(s)] = \frac{1}{h} \int_{s-h/2}^{s+h/2} f(t)\, \mathrm{d}t$$

Ex. 14.4.8 Verify the claim in (14.13).

Ex. 14.4.9 By taking derivatives in (14.14) to find stationary points, derive the expression for the ridge regression estimator $\hat{\boldsymbol{\beta}}_\lambda$. Show that $\hat{\boldsymbol{\beta}}_\lambda$ is a biased estimator of $\boldsymbol{\beta}$ whenever $\lambda > 0$.

Ex. 14.4.10 In deriving $\hat{\boldsymbol{\beta}}_\lambda$ in (14.15), we do not require the full rank assumption (see assumption 11.1.1 on page 302) whenever $\lambda > 0$. Explain why.

Ex. 14.4.11 Verify (14.20) on page 395 using (14.17) and (14.19).

14.4.1 Solutions to Selected Exercises

Solution to Ex. 14.4.1. Let f and g be any two densities on $\mathbb{R}^J$. By the triangle inequality, for any given $\mathbf{s} \in \mathbb{R}^J$, we have

$$|f(\mathbf{s}) - g(\mathbf{s})| \leqslant |f(\mathbf{s})| + |g(\mathbf{s})| = f(\mathbf{s}) + g(\mathbf{s})$$

Integrating both sides of this inequality gives $\|f - g\|_1 \leqslant 2$. □

Solution to Ex. 14.4.2. Fix densities f, g and h on $\mathbb{R}^d$. Pick any $\mathbf{s} \in \mathbb{R}^d$. By the scalar triangle inequality, we have $|f(\mathbf{s}) - g(\mathbf{s})| \leqslant |f(\mathbf{s}) - h(\mathbf{s})| + |h(\mathbf{s}) - g(\mathbf{s})|$. As per the hint in exercise 14.4.1, integration preserves this bound. This gives the triangle inequality in L_1. □

8. Hint: Try a change-of-variables argument.

Solution to Ex. 14.4.3. By Pinsker's inequality and exercise 8.5.9, we have

$$\|f(\cdot\,;\mu,\sigma) - f(\cdot\,;\bar{x}_N, s_N)\|_1 \leqslant \sqrt{\frac{\delta_N}{2}} \quad \text{where } \delta_N := \ln\frac{s_N}{\sigma} + \frac{\sigma^2 + (\mu - \bar{x}_N)^2}{2s_N^2} - \frac{1}{2}$$

Since $\bar{x}_N \xrightarrow{p} \mu$ and $s_N \xrightarrow{p} \sigma$ (see §9.2.4), applying the rules in fact 6.1.1 on page 161, we have $\delta_N \xrightarrow{p} 0$. The claim follows. □

Solution to Ex. 14.4.4. The nonnegativity of $\hat{f}$ is obvious. To show that $\int \hat{f}(s)\,\mathrm{d}s = 1$, it is enough to show that $\int K\left(\frac{s-a}{h}\right)\mathrm{d}s = h$ for any given number a. This equality can be obtained by the change of variable $u := (s-a)/h$, which leads to

$$\int K\left(\frac{s-a}{h}\right)\mathrm{d}s = \int K(u)h\,\mathrm{d}u = h\int K(u)\,\mathrm{d}u$$

Since K is a density, the proof is done. □

Solution to Ex. 14.4.5. Fix $s' \in \mathbb{R}$. By the law of total probability we have

$$\mathbb{P}\{z \leqslant s'\} = \sum_{n=1}^{N} \mathbb{P}\{x + y \leqslant s' \,|\, x = s_n\}\mathbb{P}\{x = s_n\} = \sum_{n=1}^{N} \mathbb{P}\{y \leqslant s' - s_n\}p_n$$

Differentiating with respect to s' gives the stated form for the density of z. □

Solution to Ex. 14.4.6. Adopting the setting and notation of ex. 14.4.6, our aim is to show that K and K_h satisfy (14.8). By our rules for linear transforms of Gaussians (see page 133), the distribution of $h\mathbf{x}$ is $\mathrm{N}(h\boldsymbol{\mu}, h^2\boldsymbol{\Sigma})$. That is,

$$K_h(\mathbf{s}) = (2\pi)^{-d/2}\det(h^2\boldsymbol{\Sigma})^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{s} - h\boldsymbol{\mu})^\mathsf{T}(h^2\boldsymbol{\Sigma})^{-1}(\mathbf{s} - h\boldsymbol{\mu})\right\}$$

Applying scalar multiple rules for inverses and determinants (see page 52 and page 52), we can also write this as

$$K_h(\mathbf{s}) = \frac{1}{h^d}(2\pi)^{-d/2}\det(\boldsymbol{\Sigma})^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{s}/h - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{s}/h - \boldsymbol{\mu})\right\}$$

That is, $K_h(\mathbf{s}) := \frac{1}{h^d}K(\frac{\mathbf{s}}{h})$, where K is the density of $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. □

Solution to Ex. 14.4.10. The full rank assumption is not necessary because the matrix $\mathbf{Z} := \mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I}$ is always invertible. To show this it suffices to show that $\mathbf{Z}$ is positive definite. This is not difficult to verify using the definition. □