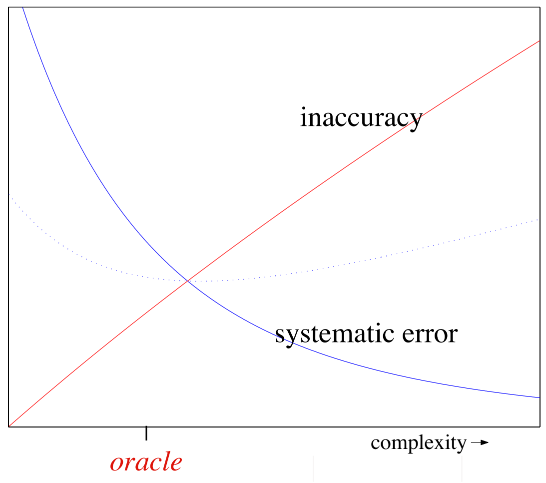

FIGURE 1. The trade off between inaccuracy and systematic error

$\mathcal{P}$ depends on n, and is in fact more "rich" for larger n. This is only natural, since when we have many observations, we may want to use more flexible models and get more information out of the data. Thus, in a parametric model, Θ may depend on n, and in particular its dimension N may depend on n, and in fact grow without limit as $n \to \infty$. This means strictly speaking that we deal with a sequence of parametric models with nonparametric limiting model. We think of such a situation as a nonparametric one.

Parametric models (with N "small") are in a sense less rich than nonparametric models, and there is also a range in the complexity of various nonparametric models. The more complex a model, the larger the inaccuracy will be. On the other hand, too simple models have large systematic error. (Here, we use a generic terminology. we will be more precise in our definitions later on, e.g., in Section 2.3.) Both inaccuracy and systematic error depend on the model, and on the truth P. The optimal model trades off the inaccuracy and systematic error (see Figure 1). However, since P is unknown, it is also not known which model this will be. Only an *oracle* can tell you that. Our aim will be to mimic this oracle.

To evaluate the inaccuracy of a model, we will use empirical process theory. Empirical process theory is about comparing the theoretical distribution P with its empirical counterpart, the empirical distribution P_n, introduced in the next section.

1.2. The empirical distribution

The unknown P can be estimated from the data in the following way. Suppose first that we are interested in the probability that an observation falls in A, where $A \subset \mathcal{X}$ is a certain set chosen by the researcher. We denote this probability by

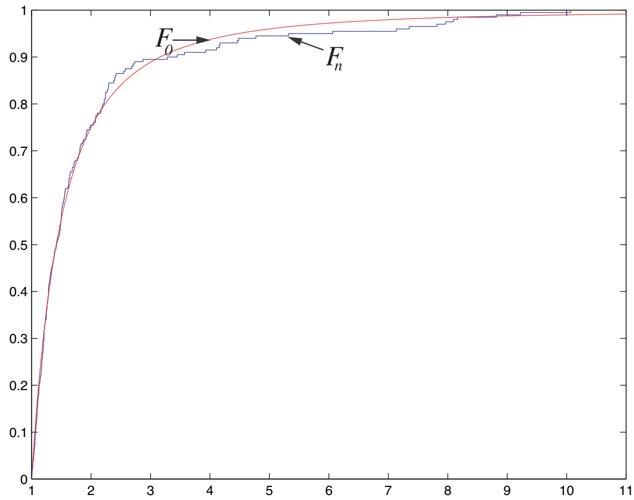

FIGURE 2. Theoretical and empirical distribution function

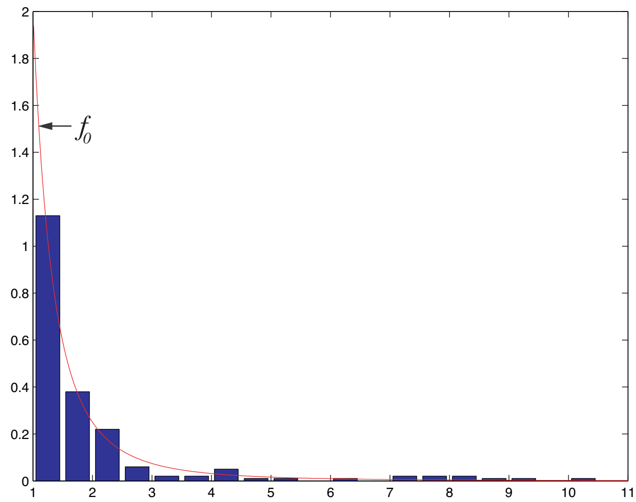

FIGURE 3. True density and a histogram

Figure 3 shows the histogram, with bandwidth $h = 0.25$, for the sample of size $n = 200$ from the Pareto distribution with parameter $\theta_0 = 2$ (i.e., with some abuse of notation, $f_0 = f_{\theta_0}$). The solid line is the density of this distribution.

The bandwidth h is an example of a tuning parameter. Choosing a value for it is a complicated matter, as it leads to considering variance, bias, and related

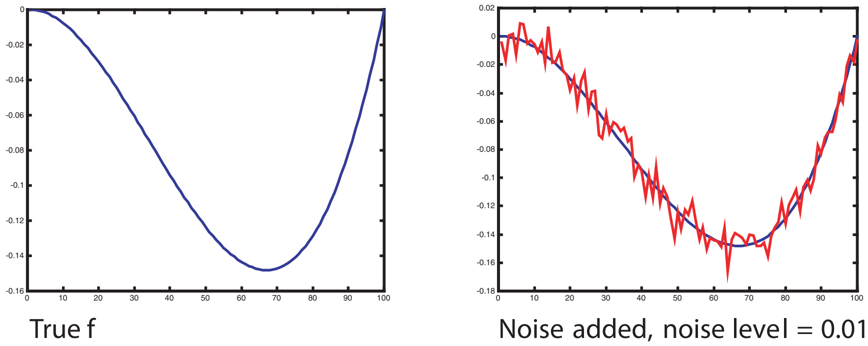

FIGURE 4

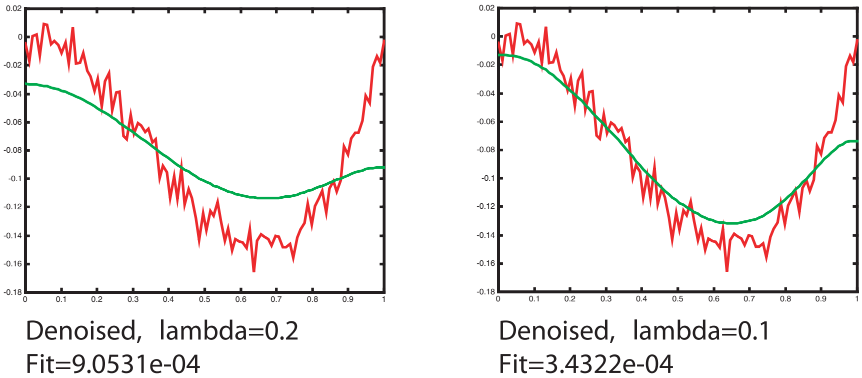

FIGURE 5

Here, "arg" stands for "argument", i.e., the location where the minimum is attained. Moreover, λ is a tuning – or regularization – parameter. If $\lambda = 0$, the estimator $\hat{f}_n$ will just interpolate the data. On the other hand, if $\lambda = \infty$, $\hat{f}_n$ will be a constant function (namely, constantly equal to the average $\sum_{i=1}^n Y_i/n$ of the observations). To the least squares loss function, we have thus added a *penalty* for choosing a too wiggly function. This is called (complexity) *regularization*.

Figure 4 above plots the true f (which is f_0) together with the data (rugged line). The aim is to recover f_0 from the data. Figure 5 shows the estimator $\hat{f}_n$ (smooth curve) for two choices of the tuning parameter λ. The fit of $\hat{f}_n$ is defined as

$$\sum_{i=1}^n |Y_i - \hat{f}_n(x_i)|^2/n.$$

Obviously, the smaller value of λ gives a better fit. Figure 6 plots the estimator $\hat{f}_n$ together with f_0, for two values of λ. The error (or "excess risk", see Chapter 2), which is defined here as

$$\sum_{i=1}^{n} |\hat{f}_n(x_i) - f_0(x_i)|^2/n$$

turns out to be smaller for the smaller value of λ.

Now, in real life situations, it is not possible to make the plots of Figure 6 and/or calculate the error, since the true f is then unknown. Thus, again, we need an oracle to tell us which λ to choose. In Section 4.5, we show that by penalizing small values of λ one may arrive at an oracle inequality.

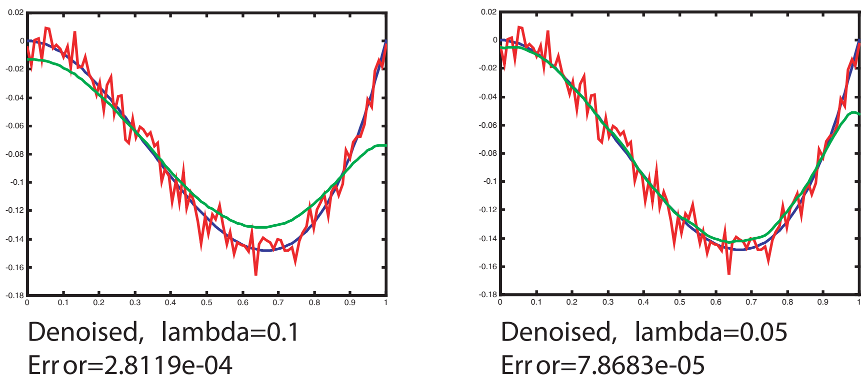

Figure 6

Intermezzo. As a continuous version of the problem studied in Example 1.4, consider

$$\hat{f} = \arg\min_f \left\{ \int_0^1 |y(x) - f(x)|^2 dx + \lambda^2 \int_0^1 |f'(x)|^2 dx \right\}.$$

In fact, let us formulate an extension, namely a continuous version corresponding to the so-called *white noise model*

$$dY(x) = f(x)dx + \sigma dW(x),$$

where W is standard Brownian motion. In that case, the derivative $y(x) = dY(x)/dx$ does not exist, as Brownian motion is nowhere differentiable. We therefore use a formulation avoiding this derivative:

$$\hat{f} = \arg\min_f \left\{ -2\int_0^1 f(x)dY(x) + \int_0^1 f^2(x)dx + \lambda^2 \int_0^1 |f'(x)|^2 dx \right\}.$$

We show in Lemma 1.1 below that the solution $\hat{f}$ can be explicitly calculated (using variational calculus). This solution reveals that the tuning parameter λ plays the role of a *bandwidth* parameter.