PART I

METHODS OF REGRESSION AND CLASSIFICATION

CHAPTER 1

OVERVIEW OF REGRESSION AND CLASSIFICATION

1.1 REGRESSION

In regression analysis we are interested in prediction or in inferring causal relationships. We try to predict the value of a response variable given the values of explanatory variables or try to deduce the causal influence of the explanatory variables to the response variable. The inference of a causal relationship is important when we want to to change the values of an explanatory variable in order to get an optimal value for the response variable. For example, we want to know the influence of education to the employment status of a worker in order to choose the best education. On the other hand, prediction is applied also in the cases when we are not able to, or do not wish to, change the values of the response variable. For example, in volatility prediction it is reasonable to use any variables that have a predictive relevance even if these variables do not have any causal relationship to volatility.

Both in prediction and in estimation of causal influence, it is useful to estimate the conditional expectation

$$E(Y \mid X = x)$$

of the response variable $Y \in \mathbf{R}$ given the explanatory variables $X \in \mathbf{R}^d$. The choice of the explanatory variables and the method of estimation can depend on the purpose

Multivariate Nonparametric Regression and Visualization. By Jussi Klemelä

of the research. In prediction an explanatory variable can be any variable that has predictive relevance whereas in the estimation of a causal influence the explanatory variables are determined by the scientific theory about the causal relationship. For the purpose of causal inference, it is reasonable to choose an estimation method that can help to find the partial effect of a given explanatory variable to the response variable. The partial effect is defined in Section 1.1.3.

In linear regression the regression function estimate is a linear function:

$$\hat{f}(x) = \hat{\alpha} + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_d x_d. \tag{1.1}$$

A different type of linearity occurs, if the estimator can be written as

$$\hat{f}(x) = \sum_{i=1}^{n} l_i(x)\, Y_i, \tag{1.2}$$

for some sequence of weights $l_1(x), \ldots, l_n(x)$. In fact, for the linear regression estimate, representations (1.1) and (1.2) hold; see (2.11). In the case of local averaging estimators, like regressogram, kernel estimator, and nearest-neighbor estimator, we use the notation $\hat{f}(x) = \sum_{i=1}^{n} p_i(x)\, Y_i$. In the case of local averaging estimators the weights $p_i(x)$ satisfy the properties that $p_i(x)$ is close to zero when X_i is distant from x and that $p_i(x)$ is large when X_i is near x. Local averaging is discussed in Section 3. There exists regression function estimates that cannot be written as in (1.2), like the orthogonal series estimators with hard thresholding; see (2.72).

In addition to the estimation of the conditional expectation of the response variable given the explanatory variables, we can consider also the estimation of the conditional median of the response variable given the explanatory variables, or the estimation of other conditional quantiles of the response variable given the explanatory variables, which is called quantile regression. Furthermore, we will consider estimation of the conditional variance, as well as estimation of the conditional density and the conditional distribution function of the response variable given the explanatory variables.

In regression analysis the response variable can take any real value or any value in a given interval, but we consider also classification. In classification the response variable can take only a finite number of distinct values and the interest lies in the prediction of the values of the response variable.

1.1.1 Random Design and Fixed Design

Random Design Regression In random design regression the data are a sequence of n pairs

$$(x_1, y_1), \ldots, (x_n, y_n), \tag{1.3}$$

where $x_i \in \mathbf{R}^d$ and $y_i \in \mathbf{R}$ for $i = 1, \ldots, n$. Data are modeled as a realization of a sequence of n random vectors

$$(X_1, Y_1), \ldots, (X_n, Y_n). \tag{1.4}$$

However, sometimes we do not distinguish notationally a random variable and its realization, and the notation of (1.4) is used also in the place of notation (1.3) to denote a realization of the random vectors and not the random vectors themselves.

In regression analysis we typically want to estimate the conditional expectation

$$f(x) = E(Y \mid X = x),$$

and now we assume that the sequence $(X_1, Y_1), \ldots, (X_n, Y_n)$ consists of identically distributed random variables, and (X, Y) has the same distribution as (X_i, Y_i), $i = 1, \ldots, n$. Besides conditional expectation we could estimate conditional mode, conditional variance, conditional quantile, and so on. Estimation of the conditional centers of distribution are discussed in Section 1.1.2 and estimation of conditional risk measures such as variance and quantiles are discussed in Section 1.1.4 and in Section 1.1.6.

Fixed Design Regression In fixed design regression the data are a sequence

$$y_1, \ldots, y_n,$$

where $y_i \in \mathbf{R}$, $i = 1, \ldots, n$. We assume that every observation y_i is associated with a fixed design point $x_i \in \mathbf{R}^d$.

Now the design points are not chosen by a random mechanism, but they are chosen by the conducter of the experiment. Typical examples could be time series data, where x_i is the time when the observation y_i is recorded, and spatial data, where x_i is the location where the observation y_i is made. Time series data are discussed in Section 1.1.9.

We model the data as a sequence of random variables

$$Y_1, \ldots, Y_n.$$

In the fixed design regression we typically do not assume that the data are identically distributed. For example, we may assume that

$$Y_i = f(x_i) + \epsilon_i, \qquad i = 1, \ldots, n,$$

where $x_i = i/n$, $f : [0, 1] \to \mathbf{R}$ is the function we want to estimate, and $E\epsilon_i = 0$. Now the data $Y_1, \ldots, Y_n$ are not identically distributed, since the observations Y_i have different expectations.

1.1.2 Mean Regression

The regression function is typically defined as a conditional expectation. Besides expectation and conditional expectation also median and conditional median can be used to characterize the center of a distribution and thus to predict and explain with the help of explanatory variables. We mention also the mode (maximum of the density function) as a third characterization of the center of a distribution, although the mode is typically not used in regression analysis.

Expectation and Conditional Expectation When the data

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

are a sequence of identically distributed random variables, we can use the data to estimate the regression function, defined as the conditional expectation of Y given X:

$$f(x) = E(Y \mid X = x), \qquad x \in \mathbf{R}^d, \tag{1.5}$$

where (X, Y) has the same distribution as (X_i, Y_i), $i = 1, \ldots, n$, and $X \in \mathbf{R}^d$, $Y \in \mathbf{R}$. The random variable Y is called the response variable, and the elements of random vector X are called the explanatory variables.

The mean of random variable $Y \in \mathbf{R}$ with a continuous distribution can be defined by

$$EY = \int_{-\infty}^{\infty} y f_Y(y)\, dy, \tag{1.6}$$

where $f_Y : \mathbf{R} \to \mathbf{R}$ is the density function of Y. The regression function has been defined in (1.5) as the conditional mean of Y, and the conditional expectation can be defined in terms of the conditional density as

$$E(Y \mid X = x) = \int_{-\infty}^{\infty} y\, f_{Y \mid X=x}(y)\, dy,$$

where the conditional density can be defined as

$$f_{Y \mid X=x}(y) = \frac{f_{X,Y}(x, y)}{f_X(x)}, \qquad y \in \mathbf{R}, \tag{1.7}$$

when $f_X(x) > 0$ and $f_{Y \mid X=x}(y) = 0$ otherwise, where $f_{X,Y} : \mathbf{R}^{d+1} \to \mathbf{R}$ is the joint density of (X, Y) and $f_X : \mathbf{R}^d \to \mathbf{R}$ is the density of X:

$$f_X(x) = \int_{\mathbf{R}} f_{X,Y}(x, y)\, dy, \qquad x \in \mathbf{R}^d.$$

Figure 1.1 illustrates mean regression. Our data consist of the daily S&P 500 returns $R_t = (S_t - S_{t-1})/S_{t-1}$, where S_t is the price of the index. There are about 16,000 observations. The S&P 500 index data are described more precisely in Section 1.6.1. We define the explanatory and the response variables as

$$X_t = \log_e \sqrt{\frac{1}{k} \sum_{i=1}^{k} R_{t-i}^2}, \qquad Y_t = \log_e |R_t|.$$

Panel (a) shows the scatter plot of (X_t, Y_t), and panel (b) shows the estimated density of (X_t, Y_t) together with the estimated regression functions. The red line shows the linear regression function estimate, and the blue line shows a kernel regression estimate with smoothing parameter $h = 0.4$. The density is estimated using kernel

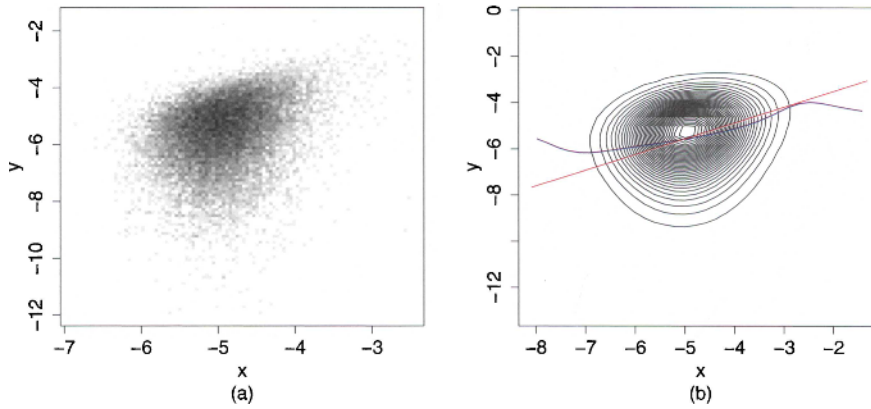

Figure 1.1 *Mean regression.* (a) A scatter plot of regression data. (b) A contour plot of the estimated joint density of the explanatory variable and the response variable. The linear regression function estimate is shown with red and the kernel regression estimate is shown with blue.

density estimation with smoothing parameter $h = 0.6$. Linear regression is discussed in Section 2.1, and kernel methods are discussed in Section 3.2. In the scatter plot we have used histogram smoothing with 100^2 bins, as explained in Section 6.1.1. This example indicates that the daily returns are dependent random variables, although it can be shown that they are nearly uncorrelated.

Median and Conditional Median The median can be defined in the case of continuous distribution function of a random variable $Y \in \mathbf{R}$ as the number $\text{median}(Y) \in \mathbf{R}$ satisfying

$$P(Y \leq \text{median}(Y)) = 0.5.$$

In general, covering also the case of discrete distributions, we can define the median uniquely as the generalized inverse of the distribution function:

$$\text{median}(Y) = \inf\{y : P(Y \leq y) \geq 0.5\}. \tag{1.8}$$

The conditional median is defined using the conditional distribution of Y given X:

$$\text{median}(Y \mid X = x) = \inf\{y : P(Y \leq y \mid X = x) \geq 0.5\}, \qquad x \in \mathbf{R}^d. \tag{1.9}$$

The sample median of observations $Y_1, \ldots, Y_n \in \mathbf{R}$ can be defined as the median of the empirical distribution. The empirical distribution is the discrete distribution with the probability mass function $P(\{Y_i\}) = 1/n$ for $i = 1, \ldots, n$. Then,

$$\text{median}(Y_1, \ldots, Y_n) = Y_{[n/2]+1}, \tag{1.10}$$

where $Y_{(1)} \leq \cdots \leq Y_{(n)}$ is the ordered sample and $[x]$ is the largest integer smaller or equal to x.

Mode and Conditional Mode The mode is defined as an argument maximizing the density function of a random variable:

$$\text{mode}(Y) = \text{argmax}_{y \in \mathbf{R}} f_Y(y), \tag{1.11}$$

where $f_Y : \mathbf{R} \to \mathbf{R}$ is the density function of Y. The density f_Y can have several local maxima, and the use of the mode seems to be interesting only in cases where the density function is unimodal (has one local maximum). The conditional mode is defined as an argument maximizing the conditional density:

$$\text{mode}(Y \mid X = x) = \text{argmax}_{y \in \mathbf{R}} f_{Y|X=x}(y).$$

1.1.3 Partial Effects and Derivative Estimation

Let us consider mean regression, where we are estimating the conditional expectation $E(Y \mid X = x)$, where $X = (X_1, \ldots, X_d)$ is the vector of explanatory variables and we denote $x = (x_1, \ldots, x_d)$. The partial effect of the variable X_1 is defined as the partial derivative

$$p(x_1; x_2, \ldots, x_d) = \frac{\partial}{\partial x_1} E(Y \mid X = x).$$

The partial effect describes how the conditional expectation of Y changes when the value of X_1 is changed, when the values of the other variables are fixed. In general, the partial effect is a function of x_1 that is different for each $x_2, \ldots, x_d$. However, for the linear model $E(Y \mid X = x) = \alpha + \beta' x$ we have

$$p(x_1; x_2, \ldots, x_d) = \beta_1,$$

so that the partial effect is a constant which is the same for all $x_2, \ldots, x_d$. Linear models are studied in Section 2.1. For the additive model $E(Y \mid X = x) = f_1(x_1) + \cdots + f_d(x_d)$ we have

$$p(x_1; x_2, \ldots, x_d) = f'(x_1),$$

so that the partial effect is a function of x_1 which is the same for all $x_2, \ldots, x_d$. Thus additive models provide easily interpretable partial effects. Additive models are studied in Section 4.2. For the single index model $E(Y \mid X = x) = g(\beta' x)$ we have

$$p(x_1; x_2, \ldots, x_d) = g'(\beta' x)\, \beta_1,$$

so that the partial effect is a function of x_1 which is different for each $x_2, \ldots, x_d$. Single index models are studied in Section 4.1.

The partial elasticity of X_1 is defined as

$$\begin{aligned} e(x_1; x_2, \ldots, x_d) &= \frac{\partial}{\partial \log x_1} \log E(Y \mid X = x) \\ &= \frac{\partial}{\partial x_1} E(Y \mid X = x) \cdot \frac{x_1}{E(Y \mid X = x)}, \end{aligned}$$

when $x_1 > 0$ and $E(Y \mid X = x) > 0$. The partial elasticity describes the approximate percentage change of conditional expectation of Y when the value of X_1 is changed by one percent, when the values of the other variables are fixed.[1] The partial semielasticity of X_1 is defined as

$$\begin{aligned} s(x_1; x_2, \ldots, x_d) &= \frac{\partial}{\partial x_1} \log E(Y \mid X = x) \\ &= \frac{\partial}{\partial x_1} E(Y \mid X = x) \cdot \frac{1}{E(Y \mid X = x)}, \end{aligned}$$

when $E(Y \mid X = x) > 0$. The partial semielasticity describes the approximate percentage change of conditional expectation of Y when the value of X_1 is changed by 1 unit, when the values of the other variables are fixed.

We can use the visualization of partial effects as a tool to visualize regression functions. In Section 7.4 we show how level set trees can be used to visualize the mode structure of functions. The mode structure of a function means the number, the largeness, and the location of the local maxima of a function. Analogously, level set trees can be used to visualize the antimode structure of a function, where the antimode structure means the number, the largeness, and the location of the local minima of a function. Local maxima and minima are important characteristics of a regression function. However, we need to know more about a regression function than just the mode structure or antimode structure. Partial effects are a useful tool to convey additional important information about a regression function. If the partial effect is flat for each variable, then we know that the regression function is close to a linear function. When we visualize the mode structure of the partial effect of variable X_1, then we get information about whether a variable X_1 is causing the expected value of the response variable to increase in several locations (the number of local maxima of the partial effect), how much an increase of the value of the variable X_1 increases the expected value of the response variable Y (the largeness of the local maxima of the partial effect), and where the influence of the response variable X_1 is the largest (the location of the local maxima of the partial effect). Analogous conclusions can be made by visualizing the antimode structure of the partial effect.

We present two methods for the estimation of partial effects. The first method is to use the partial derivatives of a kernel regression function estimator, and this method is presented in Section 3.2.9. The second method is to use a local linear estimator, and this method is presented in Section 5.2.1.

1.1.4 Variance Regression

The mean regression gives information about the center of the conditional distribution, and with the variance regression we get information about the dispersion and on the

[1] This interpretation follows from the approximation

$$\log f(x+h) - \log f(x) \approx [f(x+h) - f(x)]/f(x),$$

which follows from the approximation $\log(x) \approx x - 1$, when $x \approx 1$.

heaviness of the tails of the conditional distribution. Variance is a classical measure of dispersion and risk which is used for example in the Markowitz theory of portfolio selection. Partial moments are risk measures that generalize the variance.

Variance and Conditional Variance The variance of random variable Y is defined by

$$\mathrm{Var}(Y) = E(Y - EY)^2 = EY^2 - (EY)^2. \tag{1.12}$$

The standard deviation of Y is the square root of the variance of Y. The conditional variance of random variable Y is equal to

$$\begin{aligned} \mathrm{Var}(Y \mid X = x) &= E\left\{[Y - E(Y \mid X = x)]^2 \mid X = x\right\} && (1.13) \\ &= E(Y^2 \mid X = x) - [E(Y \mid X = x)]^2. && (1.14) \end{aligned}$$

The conditional standard deviation of Y is the square root of the conditional variance. The sample variance is defined by

$$\widehat{\mathrm{Var}}(Y) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \bar{Y}^2,$$

where $Y_1, \ldots, Y_n$ is a sample of random variables having identical distribution with Y.

Conditional Variance Estimation Conditional variance $\mathrm{Var}(Y \mid X = x)$ can be constant not depending on x. Let us write

$$Y = f(X) + \epsilon,$$

where $f(x) = E(Y \mid X = x)$ and $\epsilon = Y - f(X)$, so that $E(\epsilon \mid X = x) = 0$. If $\mathrm{Var}(Y \mid X = x) = E(\epsilon^2)$ is a constant not depending on x, we say that the noise is homoskedastic. Otherwise the noise is heteroskedastic. If the noise is heteroskedastic, it is of interest to estimate the conditional variance

$$\mathrm{Var}(Y \mid X = x) = E(\epsilon^2 \mid X = x).$$

Estimation of the conditional variance can be reduced to the estimation of the conditional expectation by using (1.13). First we estimate the conditional expectation $f(x) = E(Y \mid X = x)$ by $\hat{f}(x)$. Second we calculate the residuals

$$\hat{\epsilon}_i = Y_i - \hat{f}(X_i),$$

and estimate the conditional variance from the data $(X_1, \hat{\epsilon}_1^2), \ldots, (X_n, \hat{\epsilon}_n^2)$.

Estimation of the conditional variance can be reduced to the estimation of the conditional expectation by using (1.14). First we estimate the conditional expectation $E(Y^2 \mid X = x)$ using the regression data $(X_1, Y_1^2), \ldots, (X_n, Y_n^2)$. Second we estimate the conditional expectation $f(x) = E(Y \mid X = x)$ using data $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Theory of variance estimation is often given in the fixed design case, but the results can be extended to the random design regression by conditioning on the design variables. Let us write a heteroskedastic fixed design regression model

$$Y_i = f(x_i) + \sigma(x_i)\,\epsilon_i, \qquad i = 1, \ldots, n, \tag{1.15}$$

where $x_i \in \mathbf{R}^d$, $f : \mathbf{R}^d \to \mathbf{R}$ is the mean function, $\sigma : \mathbf{R}^d \to \mathbf{R}$ is the standard deviation function, and ϵ_i are identically distributed with $E\epsilon_i = 0$. Now we want to estimate both the function f and the function σ. Wasserman (2005, Section 5.6) has proposed making the following transformation. Let $Z_i = \log(Y_i - f(x_i))^2$. Then we have

$$Z_i = \log(\sigma^2(x_i)) + \log \epsilon_i^2.$$

Let $\hat{f}$ be an estimate of f and define $\hat{Z}_i = \log(Y_i - \hat{f}(x_i))^2$. Let $\hat{g}(x)$ be an estimate of $\log \sigma^2(x)$, obtained using regression data $(x_1, \hat{Z}_1), \ldots, (x_n, \hat{Z}_n)$, and define $\hat{\sigma}^2(x) = \exp\{\hat{g}(x)\}$.

A difference-based method for conditional variance estimation has been proposed. Let $x_1 < \cdots < x_n$ be univariate fixed design points. Now $\sigma^2(x)$ is estimated with $2^{-1}\hat{g}(x)$, where $\hat{g}$ is a regression function estimate obtained with the regression data $(x_i, (Y_i - Y_{i-1})^2)$, $i = 2, \ldots, n$. This approach has been used in Wang, Brown, Cai & Levine (2008).

Variance Estimation with Homoskedastic Noise Let us consider the fixed design regression model

$$Y_i = f(x_i) + \epsilon_i, \qquad i = 1, \ldots, n,$$

where $x_i \in \mathbf{R}^d$, $f : \mathbf{R}^d \to \mathbf{R}$ is the mean function, and $E\epsilon_i = 0$. In the case of homoskedastic noise we should estimate

$$\sigma^2 \stackrel{def}{=} E(\epsilon^2).$$

Spokoiny (2002) showed that for twice differentiable regression functions f, the optimal rate for the estimation of σ^2 is $n^{-1/2}$ for $d \leq 8$ and otherwise the optimal rate is $n^{-4/d}$. We can first estimate the mean function f by $\hat{f}$ and then use

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{f}(x_i)\right)^2.$$

These types of estimators were studied by Müller & Stadtmüller (1987), Hall & Carroll (1989), Hall & Marron (1990), and Neumann (1994). Local polynomial estimators were studied by Ruppert, Wand, Holst & Hössjer (1997), and Fan & Yao (1998). A difference-based estimator was studied by von Neumann (1941). He used the estimator

$$\widehat{\sigma^2} = \frac{1}{2(n-1)}\sum_{i=2}^{n}(Y_i - Y_{i-1})^2,$$

where it is assumed that $x_1, \ldots, x_n \in \mathbf{R}$, and $x_1 < \cdots < x_n$. The estimator was studied and modified in various ways in Rice (1984), Gasser, Sroka & Jennen-Steinmetz (1986), Hall, Kay & Titterington (1990), Hall, Kay & Titterington (1991), Thompson, Kay & Titterington (1991), and Munk, Bissantz, Wagner & Freitag (2005).

Conditional Variance in a Time Series Setting In a time series setting, when we observe Y_t, $t = 1, 2, \ldots$, the *conditional heteroskedasticity* assumption is that

$$Y_t = \sigma_t \epsilon_t, \qquad t = 0, \pm 1, \pm 2, \ldots, \tag{1.16}$$

where ϵ_t is an i.i.d. sequence, $E\epsilon_t = 0$, $E\epsilon_t^2 = 1$, and σ_t is the volatility process. The volatility process is a predictable random process, that is, σ_t is measurable with respect to the sigma-field generated by the variables $Y_{t-1}, Y_{t-2}, \ldots$. When we assume that ϵ_t is independent from $Y_{t-1}, Y_{t-2}, \ldots$, then under the conditional heteroskedasticity model,

$$\mathrm{Var}(Y_t \mid \mathcal{F}_{t-1}) = \mathrm{Var}(\sigma_t \epsilon_t \mid \mathcal{F}_{t-1}) = \sigma_t^2 \mathrm{Var}(\epsilon_t \mid \mathcal{F}_{t-1}) = \sigma_t^2 \mathrm{Var}(\epsilon_t) = \sigma_t^2, \tag{1.17}$$

where $\mathcal{F}_{t-1}$ is the sigma-algebra generated by variables $Y_{t-1}, Y_{t-2}, \ldots$. In a conditional heteroskedasticity model the main interest is in predicting the value of the random variable σ_t^2, which is thus related to estimating the conditional variance. The statistical problem is to predict σ_t^2 using a finite number of past observations $Y_1, \ldots, Y_{t-1}$. Special cases of conditional heteroskedasticity models are the ARCH model discussed in Section 2.5.2 and the GARCH model discussed in Section 3.9.2.

Partial Moments The variance of random variable $Y \in \mathbf{R}$ is defined as $\mathrm{Var}(Y) = E(Y - EY)^2$. The variance can be generalized to other centered moments

$$E|Y - EY|^k,$$

for $k = 1, 2, \ldots$. The centered moments take a contribution both from the left and the right tails of the distribution. When we are interested only in the left tail or in the right tail (losses or gains), then we can use the lower partial moments or the upper partial moments. The upper partial moment is defined as

$$\mathrm{UPM}_{\tau,k}(Y) = E\left[(Y - \tau)^k I_{[\tau,\infty)}(Y)\right]$$

and the lower partial moment is defined as

$$\mathrm{LPM}_{\tau,k}(Y) = E\left[(\tau - Y)^k I_{(-\infty,\tau]}(Y)\right],$$

where $k = 0, 1, 2, \ldots$, and $\tau \in \mathbf{R}$. In risk management τ could be the target rate. When Y has density f_Y, we can write

$$\mathrm{UPM}_{\tau,k}(Y) = \int_\tau^\infty (y - \tau)^k f_Y(y)\, dy, \qquad \mathrm{LPM}_{\tau,k}(Y) = \int_{-\infty}^\tau (\tau - y)^k f_Y(y)\, dy.$$

For example, when $k = 0$, then

$$\mathrm{UPM}_{\tau,0}(Y) = P(Y \geq \tau), \qquad \mathrm{LPM}_{\tau,0}(Y) = P(Y \leq \tau),$$

so that the upper partial moment is equal to the probability that Y is greater or equal to τ and the lower partial moment is equal to the probability that Y is smaller or equal to τ. For $k = 2$ and $\tau = EY$ the partial moments are called upper or lower semivariance of Y. The lower semivariance is defined as

$$E\left[(Y - EY)^2 I_{(-\infty, EY]}(Y)\right]. \tag{1.18}$$

The square root of the lower semivariance can be used to replace the standard deviation in the definition of the Sharpe ratio or in the Markowitz criterion. We can define conditional versions of partial moments by changing the expectations to conditional expectations.

1.1.5 Covariance and Correlation Regression

The covariance of random variables Y and Z is defined by

$$\mathrm{Cov}(Y, Z) = E[(Y - EY)(Z - EZ)] = E(YZ) - EYEZ.$$

The sample covariance is defined by

$$\widehat{\mathrm{Cov}}(Y, Z) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})(Z_i - \bar{Z}) = \frac{1}{n}\sum_{i=1}^{n} Y_i Z_i - \bar{Y}\bar{Z},$$

where $Y_1, \ldots, Y_n$ and $Z_1, \ldots, Z_n$ are samples of random variables having identical distributions with Y and Z, $\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i$, and $\bar{Z} = n^{-1}\sum_{i=1}^{n} Z_i$. The conditional covariance is obtained by changing the expectations to conditional expectations.

We have two methods of estimation of conditional covariance, analogously to two methods of conditional variance estimation based on formulas (1.13) or (1.14). The first method uses $\mathrm{Cov}(Y, Z) = E[(Y - EY)(Z - EZ)]$ and the second method uses $\mathrm{Cov}(Y, Z) = E(YZ) - EYEZ$.

The correlation is defined by

$$\mathrm{Cor}(Y, Z) = \frac{\mathrm{Cov}(Y, Z)}{\mathrm{sd}(Y)\,\mathrm{sd}(Z)},$$

where $\mathrm{sd}(Y)$ and $\mathrm{sd}(Z)$ are the standard deviations of Y and Z. The conditional correlation is defined by

$$\mathrm{Cor}(Y, Z \mid X = x) = \frac{\mathrm{Cov}(Y, Z \mid X = x)}{\mathrm{sd}(Y \mid X = x)\,\mathrm{sd}(Z \mid X = x)}, \tag{1.19}$$

where

$$\mathrm{sd}(Y \mid X = x) = \sqrt{\mathrm{Var}(Y \mid X = x)}, \qquad \mathrm{sd}(Z \mid X = x) = \sqrt{\mathrm{Var}(Z \mid X = x)}.$$

We can write

$$\mathrm{Cor}(Y, Z \mid X = x) = \mathrm{Cov}(\tilde{Y}, \tilde{Z} \mid X = x), \tag{1.20}$$

where

$$\tilde{Y} = \frac{Y}{\mathrm{sd}(Y \mid X = x)}, \qquad \tilde{Z} = \frac{Z}{\mathrm{sd}(Z \mid X = x)}.$$

Thus we have two approaches to the estimation of conditional correlation.

1. We can use (1.19). First we estimate the conditional covariance and the conditional standard deviations. Second we use (1.19) to define the estimator of the conditional correlation.

2. We can use (1.20). First we estimate the conditional standard deviations by $\widehat{\mathrm{sd}}_Y(x)$ and $\widehat{\mathrm{sd}}_Z(x)$, and calculate the standardized observations $\tilde{Y}_i = Y_i/\widehat{\mathrm{sd}}_Y(X_i)$ and $\tilde{Z}_i = Z_i/\widehat{\mathrm{sd}}_Z(X_i)$. Second we estimate the conditional correlation using $(X_i, \tilde{Y}_i, \tilde{Z}_i)$, $i = 1, \ldots, n$.

A time series $(Y_t)_{t\in\mathbf{Z}}$ is weakly stationary if $EY_t = EY_{t+h}$ and EY_tY_{t+h} depends only on h, for all $t, h \in \mathbf{Z}$. For a weakly stationary time series $(Y_t)_{t\in\mathbf{Z}}$, the autocovariance function is defined by

$$\gamma(h) = \mathrm{Cov}(Y_t, Y_{t+h}),$$

and the autocorrelation is defined by

$$\rho(h) = \gamma(h)/\gamma(0),$$

where $h = 0, \pm 1, \ldots$.

A vector time series $(X_t)_{t\in\mathbf{Z}}$, $X_t \in \mathbf{R}^d$, is weakly stationary if $EX_t = EX_{t+h}$ and EX_tX_{t+h}' depends only on h, for all $t, h \in \mathbf{Z}$. For a weakly stationary vector time series $(X_t)_{t\in\mathbf{Z}}$, the autocovariance function is defined by

$$\Gamma(h) = \mathrm{Cov}(X_t, X_{t+h}) = E[(X_t - \mu)(X_{t+h} - \mu)'], \tag{1.21}$$

for $h = 0, \pm 1, \ldots$, where $\mu = EX_t = EX_{t+h}$. Matrix $\Gamma(h)$ is a $d \times d$ matrix which is not symmetric. It holds that

$$\Gamma(h) = \Gamma(-h)'. \tag{1.22}$$

1.1.6 Quantile Regression

A quantile generalizes the median. In quantile regression a conditional quantile is estimated. Quantiles can be used to measure the value at risk (VaR). The expected shortfall is a related measure of dispersion and risk.

Quantile and Conditional Quantile The pth quantile is defined as

$$Q_p(Y) = \inf\{y : P(Y \leq y) \geq p\}, \qquad x \in \mathbf{R}^d, \tag{1.23}$$

where $0 < p < 1$. For $p = 1/2$, $Q_p(Y)$ is equal to median $\text{med}(Y)$, defined in (1.8). In the case of a continuous distribution function we have

$$P\left(Y \leq Q_p(Y)\right) = p$$

and thus it holds that

$$Q_p(Y) = F_Y^{-1}(p),$$

where $F_Y(y) = P(Y \leq y)$ is the distribution function of Y and F_Y^{-1} is the inverse of F_Y. The pth conditional quantile is defined replacing the distribution of Y with the conditional distribution of Y given X:

$$Q_p(Y \mid X = x) = \inf\{y : P(Y \leq y \mid X = x) \geq p\}, \qquad x \in \mathbf{R}^d, \tag{1.24}$$

where $0 < p < 1$. Conditional quantile estimation has been considered in Koenker (2005) and Koenker & Bassett (1978).

Estimation of a Quantile and a Conditional Quantile Estimation of quantiles is closely related to the estimation of the distribution function. It is usually possible to derive a method for the estimation of a quantile or a conditional quantile if we have a method for the estimation of a distribution function or a conditional distribution function.

Empirical Quantile Let us define the empirical distribution function, based on the dta $Y_1, \ldots, Y_n$, as

$$\hat{F}(y) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty, y]}(Y_i), \qquad y \in \mathbf{R}. \tag{1.25}$$

Now we can define an estimate of the quantile by

$$\hat{Q}_p = \inf\{x : \hat{F}(x) \geq p\}, \tag{1.26}$$

where $0 < p < 1$. Now it holds that

$$\hat{Q}_p = \begin{cases} Y_{(1)}, & 0 < p \leq 1/n, \\ Y_{(2)}, & 1/n < p \leq 2/n, \\ \vdots & \\ Y_{(n-1)}, & 1 - 2/n < p \leq 1 - 1/n, \\ Y_{(n)}, & 1 - 1/n < p < 1, \end{cases} \tag{1.27}$$

where the ordered sample is denoted by $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$. A third description of the empirical estimator of the quantile is given by the following steps:

1. Order the sample from the smallest observation to the largest observation: $Y_{(1)} \leq \cdots \leq Y_{(n)}$.
2. Let $m = \lceil pn \rceil$, where $\lceil y \rceil$ is the the smallest integer $\geq y$.
3. Set $\hat{Q}_p = Y_{(m)}$.

Standard Deviation-Based Quantile Estimators We can also use an estimate of the standard deviation to derive an estimate for a quantile. Namely, consider the location-scale model

$$Y = \mu + \sigma\,\epsilon,$$

where $\mu \in \mathbf{R}$, $\sigma > 0$, and ϵ is a random variable with a continuous distribution. Now

$$P(Y \le y) = P\left(\epsilon \le \frac{y-\mu}{\sigma}\right) = F_\epsilon\left(\frac{y-\mu}{\sigma}\right),$$

where F_ϵ is the distribution function of ϵ. If ϵ has a continuous distribution, then F_ϵ is monotone increasing and the inverse function F_ϵ^{-1} exists. The pth quantile $Q_p(Y)$ of Y satisfies $P\,(Y \le Q_p(Y)) = p$, and we can solve this equation to get

$$Q_p(Y) = \mu + \sigma\, F_\epsilon^{-1}(p).$$

Thus, for a known F_ϵ, we get from the estimates $\hat{\mu}$ of μ and $\hat{\sigma}$ of σ the estimate

$$\hat{Q}_p(Y) = \hat{\mu} + \hat{\sigma}\, F_\epsilon^{-1}(p). \tag{1.28}$$

Standard Deviation-Based Conditional Quantile Estimators To get an estimate for a conditional quantile in the heteroskedastic fixed design model (1.15), we can use

$$\hat{Q}_p(Y \mid X = x) = \hat{f}(x) + \hat{\sigma}(x)\, F_\epsilon^{-1}(p). \tag{1.29}$$

Similarly, in the conditional heteroskedasticity model (1.16) we can use

$$\hat{Q}_p(Y_t \mid \mathcal{F}_{t-1}) = \hat{\sigma}_t\, F_{\epsilon_t}^{-1}(p). \tag{1.30}$$

We apply in Section 2.5.1 and in Section 3.11.3 three quantile estimators which are based on the standard deviation estimates.

1. First estimator uses the standard normal distribution, which gives the quantile estimator

$$\hat{Q}_p(Y_t \mid \mathcal{F}_{t-1}) = \hat{\sigma}_t\, \Phi^{-1}(p), \tag{1.31}$$

where Φ is the distribution function of the standard normal distribution.

2. Second estimator uses the t-distribution, which gives the quantile estimator

$$\hat{Q}_p(Y_t \mid \mathcal{F}_{t-1}) = \sqrt{\frac{\nu-2}{\nu}}\; \hat{\sigma}_t\, t_\nu^{-1}(p), \tag{1.32}$$

where t_ν is the distribution function of the t-distribution with ν degrees of freedom. If $X \sim t_\nu$, then $\mathrm{Var}(X) = \nu/(\nu-2)$, so that $\sqrt{(\nu-2)/\nu}\; t_\nu^{-1}(p)$ is the p-quantile of the standardized t-distribution, which has unit variance.

3. Third estimator uses the empirical quantiles of the residuals. Now

$$\hat{Q}_p(Y_t \mid \mathcal{F}_{t-1}) = \hat{\sigma}_t\, \hat{Q}^{res}(p), \tag{1.33}$$

where $\hat{Q}^{res}(p)$ is the empirical quantile of the residuals $Y_t/\hat{\sigma}_t$. Empirical quantiles were defined in (1.26). This estimator was suggested in Fan & Gu (2003).

Expected Shortfall The expected shortfall is a measure of risk which aggregates all quantiles in the right tail (or in the left tail). The expected shortfall for the right tail is defined as

$$\mathrm{ES}_p(Y) = \frac{1}{1-p}\int_p^1 Q_u(Y)\,du, \qquad 0 < p < 1.$$

When Y has a continuous distribution function, then

$$\mathrm{ES}_p(Y) = E\left(Y \mid Y \geq Q_p(Y)\right) = \frac{1}{1-p}\,E\left(Y I_{[Q_p(Y),\infty)}(Y)\right); \tag{1.34}$$

see McNeil, Frey & Embrechts (2005, lemma 2.16). We have defined the loss in (1.86) as the negative of the change in the value of the portfolio, and thus the risk management wants to control the right tails of the loss distribution However, we can define the expected shortfall for the left tail as

$$\mathrm{ES}_p(Y) = \frac{1}{p}\int_0^p Q_u(Y)\,du, \qquad 0 < p < 1. \tag{1.35}$$

When Y has a continuous distribution function, then

$$\mathrm{ES}_p(Y) = E(Y \mid Y \leq Q_p(Y)) = \frac{1}{p}\,E\big(Y I_{(-\infty,Q_p(Y)]}(Y)\big).$$

This expression shows that in the case of a continuous distribution function, $p\mathrm{ES}_p(Y)$ is equal to the expectation which is taken only over the left tail, when the left tail is defined as the region which is to the left of a quantile of the distribution.[2]

The expected shortfall can be estimated from the data $Y_1, \ldots, Y_n$ in the case where the expected shortfall is given in (1.34) by using

$$\hat{\mathrm{ES}}_p = \frac{1}{m}\sum_{i=m}^{n} Y_{(i)},$$

where $Y_{(1)} \leq \cdots \leq Y_{(n)}$ and $m = \lceil (1-p)n \rceil$. When the expected shortfall is given by (1.35), then we define

$$\hat{\mathrm{ES}}_p = \frac{1}{m}\sum_{i=1}^{m} Y_{(i)},$$

where $m = \lceil pn \rceil$.

Let us consider the location-scale model

$$Y = \mu + \sigma\,\epsilon,$$

where $\mu \in \mathbf{R}$, $\sigma > 0$, and ϵ is a random variable with a continuous distribution. Now

$$\mathrm{ES}_p(Y) = \mu + \sigma \mathrm{ES}_p(\epsilon).$$

[2]Sometimes the expected shortfall for the left tail is defined as $Q_p(Y) - E[Y I_{(-\infty,Q_p(Y)]}(Y)]$ and the absolute shortfall is defined as $-E[Y I_{(-\infty,Q_p(Y)]}(Y)]$.

Thus the estimate for the expected shortfall can be obtained as

$$\hat{\mathrm{ES}}_p(Y) = \hat{\mu} + \hat{\sigma}\mathrm{ES}_p(\epsilon),$$

where $\hat{\mu}$ is an estimate of μ and σ is an estimate of σ.

If $\epsilon \sim N(0,1)$ and the expected shortfall is defined for the right tail as in (1.34), then

$$\mathrm{ES}_p(\epsilon) = \frac{\phi(\Phi^{-1}(p))}{1-p},$$

where ϕ is the density function of the standard normal distribution and Φ is the distribution function of the standard normal distribution. If $\epsilon \sim t_\nu$, where t_ν is the t-distribution with ν degrees of freedom, and the expected shortfall is defined for the right tail as in (1.34), then

$$\mathrm{ES}_p(\epsilon) = \frac{g_\nu(t_\nu^{-1}(p))}{1-p}\,\frac{\nu + (t_\nu^{-1}(p))^2}{\nu - 1},$$

where g_ν is the density function of the t-distribution with ν degrees of freedom and t_ν is the distribution function of the t-distribution with ν degrees of freedom.

Expected shortfall is sometimes preferred to the quantiles on the grounds that the expected shortfall satisfies the axiom of subadditivity. Risk measure ϱ is said to be subadditive if $\varrho(X+Y) \le \varrho(X) + \varrho(Y)$, where X and Y are random variables interpreted as portfolio losses. Quantiles do not satisfy subadditivity like the expected shortfall. The other axioms of a coherent risk measure are the monotonicity: if $Y \ge X$, then $\varrho(Y) \ge \varrho(X)$; the positive homogeneity: for $\lambda \ge 0$, $\varrho(\lambda Y) = \lambda\varrho(Y)$; and the translation invariance: for $a \in \mathbf{R}$, $\varrho(Y+a) = \varrho(Y) + a$. For more about coherent risk measures, see McNeil et al. (2005, Section 6.1).

1.1.7 Approximation of the Response Variable

We have defined the regression function in (1.5) as the conditional expectation of the response variable. The conditional expectation can be viewed as an approximation of response variable $Y \in \mathbf{R}$ with the help of explanatory random variables $X_1, \ldots, X_d \in \mathbf{R}$. The approximation is a random variable $f(X_1, \ldots, X_d) \in \mathbf{R}$, where $f : \mathbf{R}^d \to \mathbf{R}$ is a fixed function. This viewpoint leads to generalizations. The best approximation of the response variable can be defined using various loss functions $\rho : \mathbf{R} \to \mathbf{R}$. The best approximation is $f(X_1, \ldots, X_d)$, where f is defined as

$$f = \mathrm{argmin}_{g \in \mathcal{G}} E\rho(Y - g(X)), \qquad X = (X_1, \ldots, X_d), \tag{1.36}$$

where $\mathcal{G}$ is a suitable class of functions $g : \mathbf{R}^d \to \mathbf{R}$. Since f is defined in terms of the unknown distribution of (X, Y), we have to estimate f using statistical data available from the distribution of (X, Y).

Examples of Loss Functions We give examples of different choices of ρ and $\mathcal{G}$.

1. When $\rho(t) = t^2$ and $\mathcal{G}$ is the class of all measurable functions $\mathbf{R}^d \to \mathbf{R}$, then f, defined by (1.36), is equal to the conditional expectation:

$$f(x) = E(Y \mid X = x) = \text{argmin}_{g \in \mathcal{G}} E(Y - g(X))^2.$$

Indeed,

$$E(g(X) - Y)^2 = E(g(X) - E(Y \mid X))^2 + E(E(Y \mid X) - Y)^2, \quad (1.37)$$

because $E[(g(X) - E(Y \mid X))(E(Y \mid X) - Y)] = 0$, and thus $E(g(X) - Y)^2$ is minimized with respect to $g : \mathbf{R}^d \to \mathbf{R}$ by choosing $g(x) = E(Y \mid X = x)$.[3] Note also that the expectation EY is the best constant approximation of Y. That is, if we choose $\mathcal{G}$ as the class of constant functions

$$\mathcal{G} = \{g : \mathbf{R}^d \to \mathbf{R} \mid g(x) = \mu \text{ for all } x \in \mathbf{R}, \mu \in \mathbf{R}\},$$

then

$$EY = \text{argmin}_{g \in \mathcal{G}} E(Y - g(X))^2 = \text{argmin}_{\mu \in \mathbf{R}} E(Y - \mu)^2. \quad (1.38)$$

Indeed,

$$E(Y - \mu)^2 = E(Y - EY)^2 + (EY - \mu)^2,$$

and this is minimized with respect to $\mu \in \mathbf{R}$ by choosing $\mu = EY$.

2. When $\rho(t) = |t|$ and $\mathcal{G}$ is the class of all measurable functions $\mathbf{R}^d \to \mathbf{R}$, then f defined by (1.36) is the conditional median:

$$\text{med}(Y \mid X = x) = \text{argmin}_{g \in \mathcal{G}} E|Y - g(X)|, \quad (1.39)$$

where the conditional median is defined in (1.9). Equation (1.39) is proved in the next item.

3. When ρ is defined as

$$\rho_p(t) = t\,[p - I_{(-\infty,0)}(t))] = \begin{cases} t(p-1), & \text{if } t < 0 \\ tp, & \text{if } t \geq 0, \end{cases} \quad (1.40)$$

for $0 < p < 1$ and $\mathcal{G}$ is the class of all measurable functions, then the best approximation is the conditional quantile. Figure 1.2 shows the loss function in (1.40) with $p = 0.5$ (black line) and with $p = 0.1$ (red line). We show that if the distribution function F_Y is strictly monotonic, then

$$Q_p(Y) = \text{argmin}_{\theta \in \mathbf{R}} E\rho_p(Y - \theta). \quad (1.41)$$

[3]Note that the conditional expectation defined as $f(x) = E(Y \mid X = x)$ is a real-valued function of x, but $E(X \mid Y)$ is a real-valued random variable which can be defined as $E(X \mid Y) = f(X)$.

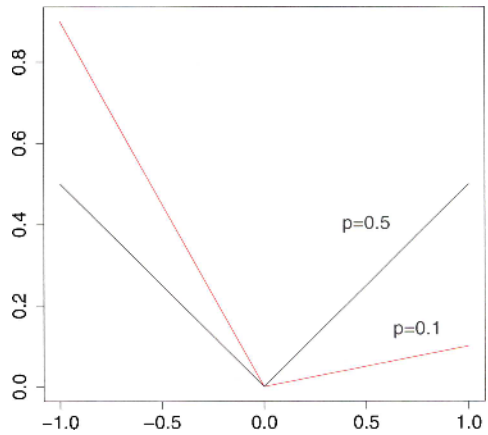

Figure 1.2 *Loss functions for quantile estimation.* Loss function in (1.40) with $p = 0.5$ (black line) and with $p = 0.1$ (red line).

To show (1.41), note that

$$E\rho_p(Y-\theta) = (p-1)\int_{-\infty}^{\theta}(y-\theta)\,dF_Y(y) + p\int_{\theta}^{\infty}(y-\theta)\,dF_Y(y)$$

and thus

$$\frac{\partial}{\partial\theta}E\rho_p(Y-\theta) = (1-p)\int_{-\infty}^{\theta}dF_Y(y) - p\int_{\theta}^{\infty}dF_Y(y) = F_Y(\theta) - p.$$

Setting $\partial E\rho_p(Y-\theta)/\partial\theta = 0$, we get (1.41), when F_Y is strictly monotonic. We can prove similarly the case of conditional quantiles:

$$Q_p(Y \mid X = x) = \text{argmin}_{g\in\mathcal{G}}E\rho_p(Y - g(X)),$$

where $\mathcal{G}$ is the class of measurable functions $\mathbf{R}^d \to \mathbf{R}$. When $p = 1/2$, then

$$\rho_p(t) = \frac{1}{2}|t|,$$

and we have proved the result (1.39).

Estimation Using Loss Function If a regression function can be characterized as a minimizer of a loss function, then we can use empirical risk minimization with this loss function to define an estimator for the regression function. Empirical risk minimization is discussed in Chapter 5.

For example, conditional expectation $f(x) = E(Y \mid X = x)$ can be estimated minimizing the sum of squared errors:

$$\hat{f} = \text{argmin}_{f \in \mathcal{F}} \sum_{i=1}^{n} (Y_i - f(X_i))^2,$$

where $\mathcal{F}$ is a class of functions $f : \mathbf{R}^d \to \mathbf{R}$. For example, $\mathcal{F}$ could be the class of linear functions.

Estimation of quantiles and conditional quantiles can also be done using empirical risk minimization. The estimator of the pth quantile is

$$\hat{Q}_p(Y) = \text{argmin}_{\theta \in \mathbf{R}} \sum_{i=1}^{n} \rho_p(Y_i - \theta)$$

and the estimator of the pth conditional quantile $f(x) = Q_p(Y \mid X = x)$ is

$$\hat{f} = \text{argmin}_{f \in \mathcal{F}} \sum_{i=1}^{n} \rho_p(Y_i - f(X_i)),$$

where $\mathcal{F}$ is a class of functions $f : \mathbf{R}^d \to \mathbf{R}$. A further idea which we will discuss in Section 5.2 is to define an estimator for the conditional quantile using local empirical risk:

$$\hat{f}(x) = \text{argmin}_{\theta \in \mathbf{R}} \sum_{i=1}^{n} p_i(x)\, \rho_p(Y_i - \theta),$$

where $p_i(x) \geq 0$ and $\sum_{i=1}^{n} p_i(x) = 1$. These weights should have the property that $p_i(x)$ is large when X_i is close to x and $p_i(x)$ is small when X_i is far away from x.

1.1.8 Conditional Distribution and Density

Instead of estimating only conditional expectation, conditional variance, or conditional quantile, we can try to estimate the complete conditional distribution by estimating the conditional distribution function or the conditional density function.

Conditional Distribution Function The distribution function of random variable $Y \in \mathbf{R}$ is defined as[4]

$$F_Y(y) = P(Y \leq y), \qquad y \in \mathbf{R}.$$

The conditional distribution function is defined as

$$F_{Y \mid X=x}(y) = P(Y \leq y \mid X = x), \qquad y \in \mathbf{R},\ x \in \mathbf{R}^d,$$

[4]This definition can be extended to the multivariate case $Y = (Y_1, \ldots, Y_d)$ by

$$F_Y(y) = P(Y_1 \leq y_1, \ldots, Y_d \leq y_d), \qquad y = (y_1, \ldots, y_d) \in \mathbf{R}^d.$$

where $Y \in \mathbf{R}$ is a scalar random variable and $X \in \mathbf{R}^d$ is a random vector. We have

$$F_{Y \mid X=x}(y) = E\left[I_{(-\infty,y]}(Y) \mid X = x\right] \tag{1.42}$$

and thus the estimation of the conditional distribution function can be considered as a regression problem, where the conditional expectation of the random variable $I_{(-\infty,y]}(Y)$ is estimated. The random variable $I_{(-\infty,y]}(Y)$ takes only values 0 or 1. The unconditional distribution function can be estimated with the empirical distribution function, which is defined for the data $Y_1, \ldots, Y_n$ as

$$\hat{F}_Y(y) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,y]}(Y_i) = n^{-1}\#\{i : Y_i \le y, i = 1, \ldots, n\}, \tag{1.43}$$

where $\#A$ means the cardinality of set A. The conditional distribution function estimation is considered in Section 3.7, where local averaging estimators are defined.

Conditional Density Conditional density function is defined as

$$f_{Y \mid X=x}(y) = \begin{cases} \frac{f_{X,Y}(x,y)}{f_X(x)}, & \text{when } f_X(x) > 0, \\ 0, & \text{otherwise,} \end{cases}$$

for $y \in \mathbf{R}$, where $f_{X,Y} : \mathbf{R}^{d+1} \to \mathbf{R}$ is the joint density of (X, Y) and $f_X : \mathbf{R}^d \to \mathbf{R}$ is the density of X. We mention three ways to estimate the conditional density.

First, we can replace the density of (X, Y) and the density of X with their estimators $\hat{f}_{X,Y}$ and $\hat{f}_X$ and define

$$\hat{f}_{Y \mid X=x}(y) = \frac{\hat{f}_{X,Y}(x,y)}{\hat{f}_X(x)},$$

for $\hat{f}_X(x) > 0$. This approach is close to the approach used in Section 3.6, where local averaging estimators of the conditional density are defined.

Second, empirical risk minimization can be used in the estimation of the conditional density, as explained in Section 5.1.3.

Third, sometimes it is reasonable to assume that the conditional density has the form

$$f_{Y \mid X=x}(y) = f_{g(x)}(y), \tag{1.44}$$

where f_θ, $\theta \in A \subset \mathbf{R}^k$, is a family of density functions and $g : \mathbf{R}^d \to A$, where $k \ge 1$. Then the estimation of the conditional density reduces to the estimation of the "regression function" g. The mean regression is a special case of this approach when the distribution of errors is known: Assume that

$$Y = f(X) + \epsilon,$$

where ϵ is independent of X, $E\epsilon = 0$, and the density of ϵ is denoted by f_ϵ. Then

$$f_{Y \mid X=x}(y) = f_\epsilon(y - f(x)),$$

which is a special case of (1.44), when we take $f_\theta(y) = f_\epsilon(y - \theta)$ and $g(x) = f(x)$. The case of heteroskedastic variance is an other example: Now we assume that

$$Y = f(X) + \sigma(X)\,\epsilon,$$

where ϵ is independent of X, $E\epsilon = 0$, and the density of ϵ is denoted by f_ϵ. Then

$$f_{Y \mid X=x}(y) = \sigma(x)^{-1} f_\epsilon((y - f(x))/\sigma(x)),$$

which is a special case of (1.44), when we take $\theta = (\theta_1, \theta_2)$, $f_\theta(y) = \theta_2^{-1} f_\epsilon((y - \theta_1)/\theta_2)$, and $g(x) = (f(x), \sigma(x))$. This approach is used in parametric family regression, explained in Section 1.3.1.

1.1.9 Time Series Data

Regression data are a sequence $(X_1, Y_1), \ldots, (X_n, Y_n)$ of identically distributed copies of (X, Y), where $X \in \mathbf{R}^d$ is the explanatory variable and $Y \in \mathbf{R}$ is the response variable, as we wrote in (1.4). However, we can use regression methods with time series data

$$Z_1, \ldots, Z_T \in \mathbf{R},$$

where the observation Z_t is made at time t, $t = 1, \ldots, T$. In order to apply regression methods we identify the response variable and the explanatory variables. We consider two ways for the choice of the explanatory variables. In the first case the state space of the time series is used as the space of the explanatory variables, and in the second case the time space is used as the space of the explanatory variables.

State–Space Prediction In the state–space prediction an autoregression parameter $k \geq 1$ is chosen and we denote

$$Y_i = Z_{i+1}, \qquad X_i = (Z_i, \ldots, Z_{i-k+1}), \tag{1.45}$$

$i = k, \ldots, T-1$. When the time series $Z_1, \ldots, Z_T$ is stationary, then the sequence (X_i, Y_i), $i = k, \ldots, T-1$, consists of identically distributed random variables and we can denote by (X, Y) a random vector which is identically distributed as (X_i, Y_i).

We define the regression function, as previously, by

$$f(x) = E(Y \mid X = x), \qquad x \in \mathbf{R}^k. \tag{1.46}$$

We can estimate this regression function using data (X_i, Y_i), $i = k, \ldots, T-1$. Estimator of the regression function $f : \mathbf{R}^k \to \mathbf{R}$ can be used to predict or explain the next outcome of the time series using k previous observations. For example, let $\hat{f}_T$ be an estimator of the regression function at time T, constructed using data (X_i, Y_i), $i = k, \ldots, T-1$. The prediction of the next outcome is $\hat{f}_T(X_T)$, where $X_T = (Z_T, \ldots, Z_{T-k+1})$.

Let

$$Z_1, \ldots, Z_T \in \mathbf{R}^d$$

be a d-dimensional vector time series. Definition (1.45) generalizes to the setting of vector time series. Define

$$Y_i = g(Z_{i+1}), \qquad X_i = (Z_i, \ldots, Z_{i-k+1}), \tag{1.47}$$

$i = k, \ldots, T-1$, where $g : \mathbf{R}^d \to \mathbf{R}$ is a function with real values. We define the regression function, as previously, by

$$f(x) = E(Y_i \mid X_i = x), \qquad x \in \mathbf{R}^{dk}.$$

The regression function is now defined on the higher-dimensional space of dimension kd.

We can predict and explain without autoregression parameter k and take into account all the previous observations and not just the k last observations. However, this approach does not fit into the standard regression approach. Let $Z_1, \ldots, Z_T \in \mathbf{R}$ be a scalar time series and define

$$Y_i = Z_{i+1}, \qquad X_i = (Z_i, \ldots, Z_1),$$

$i = 1, \ldots, T-1$. The sequence of observations $(Y_1, X_1), \ldots, (Y_{T-1}, X_{T-1})$ is not a sequence of identically distributed random vectors. For example, the regression function $f_i(x) = E(Y_i \mid X_i = x)$, $x \in \mathbf{R}^{id}$, is defined in a different space for each i.

Time–Space Prediction In time–space prediction the time parameter is taken as the explanatory variable, in contrast to (1.45), where the previous observations in the time series are taken as the explanatory variables. We denote

$$Y_i = Z_i, \qquad X_i = i, \qquad i = 1, \ldots, T. \tag{1.48}$$

The obtained regression model is a fixed design regression model, as described in Section 1.1.1.

Time–space prediction can be used when the time series can be modeled as a nonstationary time series of signal with additive noise:

$$Y_i = \mu_i + \sigma_i \epsilon_i, \qquad i = 1, \ldots, T, \tag{1.49}$$

where $\mu_i \in \mathbf{R}$ is the deterministic signal, $\sigma_i > 0$ are nonrandom values, and the noise ϵ_i is stationary with mean zero and unit variance. For statistical estimation and asymptotic analysis we can use a slightly different model

$$Y_{i,T} = \mu(t_{i,T}) + \sigma(t_{i,T})\, \epsilon_{i,T}, \qquad i = 1, \ldots, T, \tag{1.50}$$

where $t_{i,T} = i/T$, $\mu : [0,1] \to \mathbf{R}$, $\sigma : [0,1] \to (0, \infty)$, and $\epsilon_{i,T}$ is stationary with mean zero and unit variance. Now it can be thought that the observations are coming from a continuous time process $Y(t)$, $t \in [0,1]$, and the sampled discrete time process is obtained as $Y_{i,T} = Y(i/T)$, $i = 1, \ldots, T$. The asymptotics as $T \to \infty$ is called in-fill asymptotics, because points $t_{i,T}$ are filling the interval $[0,1]$ as $T \to \infty$.

1.1.10 Stochastic Control

We consider two types of stochastic control problems. The first type of stochastic control problem appears in option pricing and hedging and the second type of stochastic control problem appears in portfolio selection. The connection of these stochastic control problems to portfolio selection and to option pricing and hedging are explained in Section 1.5.3 and in Section 1.5.4, respectively.

Option-Pricing-Type Stochastic Control Consider the time series

$$X_{t_0}, X_{t_0+1}, \ldots, X_{T-1} \in \mathbf{R}$$

and a random variable $Y_T \in \mathbf{R}$. We are able to choose coefficient $\beta_t \in \mathbf{R}$ at time t, for $t = t_0, \ldots, T-1$ and a constant term $\alpha_{t_0} \in \mathbf{R}$ at time t_0. We want to choose these coefficients in such a way that the mean squared error

$$\mathrm{MSE}\left(\alpha_{t_0}, \beta_{t_0}, \ldots, \beta_{T-1}\right) = E\left(\alpha_{t_0} + \beta_{t_0} X_{t_0} + \cdots + \beta_{T-1} X_{T-1} - Y_T\right)^2$$

is minimized. The optimal coefficients at time t_0 are defined by

$$\left(\alpha_{t_0}^o, \beta_{t_0}^o\right) = \mathrm{argmin}_{\alpha_{t_0}, \beta_{t_0}} \min_{\beta_{t_0+1}, \ldots, \beta_{T-1}} \mathrm{MSE}\left(\alpha_{t_0}, \beta_{t_0}, \ldots, \beta_{T-1}\right), \qquad (1.51)$$

where the minimization is done over coefficients β_t at time t and over coefficient α_{t_0} at time t_0.

Note that at time t_0 the coefficients $\beta_{t_0+1}, \ldots, \beta_{T-1}$ are nuisance coefficients since they are chosen at later times, and at time t_0 we use them only to calculate the optimal values $\alpha_{t_0}^o$ and $\beta_{t_0}^o$. Then, at time $t_0 + 1$ we choose parameters α_{t_0+1} and β_{t_0+1} and parameters $\beta_{t_0+2}, \ldots, \beta_{T-1}$ are nuisance parameters at time $t_0 + 1$.

Note the difference to the usual least squares problem. Namely, in the usual least squares problem we solve the problem

$$\min_{\alpha_{t_0}, \beta_{t_0}, \ldots, \beta_{T-1}} \mathrm{MSE}\left(\alpha_{t_0}, \beta_{t_0}, \ldots, \beta_{T-1}\right)$$

at time $T-1$. That is, all coefficients are chosen at the same time $T-1$ and at that time all values $X_{t_0}, \ldots X_{T-1}$ are known. This problem appears for example in the linear autoregression, where we minimize the expected squared error

$$E\left(\alpha_{t_0} + \beta_{t_0} X_{t_0} + \cdots + \beta_{T-1} X_{T-1} - X_T\right)^2$$

at time $T-1$. In the one-step case the stochastic control and the usual least squares problem are identical, because in the one-step problem we minimize

$$E\left(\alpha_{t_0} + \beta_{t_0} X_{t_0} - Y_{t_0+1}\right)^2 = E\left(\alpha_{T-1} + \beta_{T-1} X_{T-1} - Y_T\right)^2$$

at time $t_0 = T-1$.

If we have n realizations $(X_1^i, \ldots, X_{T-t_0}^i, Y_{T-t_0+1}^i)$, $i = 1, \ldots, n$, which have the same distribution as $(X_{t_0}, \ldots, X_{T-1}, Y_T)$, then we can find data-based coefficients as

$$\left(\alpha_{t_0}^o, \beta_{t_0}^o\right) = \mathrm{argmin}_{\alpha_{t_0}, \beta_{t_0}} \min_{b_{t_0+1}, \ldots, b_{T-1}} \mathrm{MSE}_n\left(\alpha_{t_0}, \beta_{t_0}, \ldots, \beta_{T-1}\right),$$

where

$$\begin{aligned}&\mathrm{MSE}_n\left(\alpha_{t_0}, \beta_{t_0}, \ldots, \beta_{T-1}\right)\\ &= \sum_{i=1}^{n}\left(\alpha_{t_0} + \beta_{t_0} X_1^i + \cdots + \beta_{T-1} X_{T-t_0}^i - Y_{T-t_0+1}^i\right)^2.\end{aligned}$$

The connection of this type of stochastic control problem to option pricing is explained in Section 1.5.4.

Portfolio-Selection-Type Stochastic Control Consider the time series

$$X_{t_0+1}, X_{t_0+1}, \ldots, X_T \in \mathbf{R}^d.$$

We are able to choose coefficient $\beta_t \in \mathbf{R}^d$ at time t, for $t = t_0, \ldots, T-1$. We want to choose these coefficients in such a way that

$$\mathrm{W}\left(\beta_{t_0}, \ldots, \beta_{T-1}\right) = Eu\left(\prod_{t=t_0}^{T-1} \beta_t' X_{t+1}\right)$$

is maximized, where $u : \mathbf{R} \to \mathbf{R}$. The optimal coefficients at time t_0 are defined by

$$\beta_{t_0}^o = \mathrm{argmax}_{\beta_{t_0}} \max_{\beta_{t_0+1}, \ldots, \beta_{T-1}} \mathrm{W}\left(\beta_{t_0}, \ldots, \beta_{T-1}\right), \tag{1.52}$$

where the maximization is done over vector β_t at time t. The connection of this type of stochastic control problem to portfolio selection is explained in Section 1.5.3, see (1.97).

1.1.11 Instrumental Variables

The method of instrumental variables is used to estimate causal relationships when it is not possible to make controlled experiments. There are three classical examples of the cases where a need for instrumental variables arises: when there are relevant explanatory variables which are not observed (omitted variables), when the explanatory variables are subject to measurement errors, or when the response variable has a causal influence on one of the explanatory variables (reverse causation).

The method of instrumental variables can be used when we want to estimate structural function $g : \mathbf{R}^d \to \mathbf{R}$ in the model

$$Y = g(X) + U, \tag{1.53}$$

where $Y \in \mathbf{R}$, $X \in \mathbf{R}^d$, and

$$E(U \mid X) \neq 0.$$

Now $g(x)$ is not the conditional expectation $E[Y \mid X = x]$. Estimation of g is possible when we have observations (X_i, Y_i, Z_i), $i = 1, \ldots, n$, where (X_i, Y_i) are

distributed as (X, Y) and Z_i are observations from the distribution of an instrumental variable $Z \in \mathbf{R}^d$ that satisfies

$$E(U \mid Z) = 0. \tag{1.54}$$

We give two examples of model (1.53). The first example explains how an omitted variable can lead to (1.53). The second example explains how an error in the explanatory variable can lead to (1.53).

Omitted Variable As an example of a case where model (1.53) can arise, consider the case where X is a variable indicating the type of the treatment a patient receives:

$$X = \begin{cases} 0, & \text{when patient receives treatment A,} \\ 1, & \text{when patient receives treatment B,} \end{cases}$$

and Y is a variable measuring the health of the patient after receiving the treatment. This example is modeled after McClellan, McNeil & Newhouse (1994). We want to estimate the causal influence of X on Y. Let us denote with W the random variable measuring the health of a patient at the time the patient receives the treatment. Also the variable W is influencing Y. In this example W is also affecting X, because the decision about the treatment a patient receives is partially based on the health condition of the patient (if the patient is weak, he will not receive a treatment that is physiologically demanding). Using usual regression methods and observations of X and Y would give a biased estimate of the causal influence of X on Y. (If patients with a weak condition receive treatment A more often, then the estimate would give a pessimistic estimate of the effect of treatment A.)

We have three approaches to estimate the casual influence of X on Y: (1) We can use randomization, so that the value of X is determined by coin tossing, and the influence of W on X is removed. However, in this example this is not possible for ethical reasons. (2) We can estimate the conditional expectation $E(Y \mid X = x,\ W = z)$. However, in this example we have not observed W, so the estimation of this conditional expectation is not possible. (3) We can use the method of instrumental variables. In this example the instrumental variable Z can be chosen as the difference between the shortest distance from a patients home to a hospital giving treatment A and the shortest distance from a patients home to a hospital giving treatment B. Variable Z has an influence on X, because patients had an influence on choice of the treatment they received, and they tended to choose a treatment that was given in the nearest hospital. Variable Z does not have any influence on the health of patients, so it is otherwise external variable, influencing only X. Thus we can use Z to make a pseudo randomization even when a proper randomization was not possible.

We assume an additive model

$$Y = \alpha + f_1(X) + f_2(W) + \epsilon,$$

where $E(\epsilon \mid X = x,\ W = w) = 0$, $Ef_1(X) = 0$, and $Ef_2(W) = 0$. We have observations (Y_i, X_i, Z_i), $i = 1, \ldots, n$, but no observations of W. Using these

observations, we can estimate f_1, but not f_2. Estimation of f_1 is enough to give information on the causal influence of X to Y.

Denoting $g(X) = \alpha + f_1(X)$ and $U = f_2(W) + \epsilon$, we have that

$$Y = g(X) + U,$$

where $E(U \mid X) \neq 0$, because $\text{Cov}(X, W) \neq 0$, and $E(U \mid Z) = 0$, because Z is external to the system, having influence only on X. Thus we are in the setting of model (1.53).

Errors in Variables in a Linear Model As an example of model (1.53), consider the case where the linear model

$$Y = \alpha + \beta X^* + U^*$$

holds. However, the explanatory variable X^* is not observed directly but we observe only pairs (Y_i, X_i), $i = 1, \ldots, n$, where

$$X_i = X_i^* + \epsilon_i, \qquad i = 1, \ldots, n.$$

Thus the observed values X_i are contaminated with additive errors. We assume that

$$\text{Cov}(X^*, U^*) = 0, \qquad \text{Cov}(U^*, \epsilon) = 0 \tag{1.55}$$

and

$$\text{Cov}(X^*, \epsilon) = 0. \tag{1.56}$$

We can write the observed response variables as

$$Y = \alpha + \beta X + U^* - \beta\epsilon$$

and the new error term is denoted by

$$U = U^* - \beta\epsilon$$

to get the new linear model

$$Y = \alpha + \beta X + U. \tag{1.57}$$

In this new linear model $E(U \mid X) \neq 0$. Thus we have the same situation as in (1.53), with $g(X) = \alpha + \beta X$.

The fact $E(U \mid X) \neq 0$, follows from $\text{Cov}(X, U) \neq 0$. We have that

$$\begin{aligned} \text{Cov}(X, U) &= \text{Cov}(X, U^*) - \beta\text{Cov}(X, \epsilon) \\ &= -\beta\left[\text{Cov}(X^*, \epsilon) + \text{Cov}(\epsilon, \epsilon)\right] \\ &= -\beta\text{Var}(\epsilon) \\ &\neq 0, \end{aligned}$$

because

$$\text{Cov}(X, U^*) = \text{Cov}(X^*, U^*) + \text{Cov}(\epsilon, U^*) = 0$$

by assumption (1.55) and $\text{Cov}(X^*, \epsilon) = 0$ by assumption (1.56).

Estimation of the Structural Function We give a linear instrumental variables estimator in (2.24). This estimator can be used to estimate parameters α and β in (1.57). The linear instrumental variable estimator is

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \qquad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X},$$

where

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \qquad \bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i.$$

Hall & Horowitz (2005) approach the estimation of $g(x)$ in the model (1.53) by deriving an operator equation for g. From (1.54) we obtain

$$E(Y \mid Z) = E(g(X) \mid Z) + E(U \mid Z) = (Kg)(Z),$$

where the operator K is defined as

$$(Kg)(z) = E(g(X) \mid Z = z) = \int f_{X \mid Z=z}(x) g(x)\, dx.$$

The operator K is an integral operator mapping $L_X^2 = \{g : \mathbf{R}^d \to \mathbf{R} \mid E(g^2(X)) < \infty\}$ to $L_Z^2 = \{h : \mathbf{R}^d \to \mathbf{R} \mid E(h^2(Z)) < \infty\}$. By estimating K and estimating $E(Y \mid Z)$ we can find an estimator for g.

1.2 DISCRETE RESPONSE VARIABLE

We introduce first binary response models, where the response variable is a Bernoulli random variable, second we introduce discrete choice models, where the response variable is a categorical random variable, and third we introduce count data models, where the response variable is a Poisson random variable. In Section 1.3 we introduce more general exponential family models which contain as special cases the binary response models, discrete choice models, and Poisson count models.

1.2.1 Binary Response Models

In a binary response model the response variable Y is a Bernoulli distributed random variable, so that it takes only values 0 and 1. When $Y \sim \text{Bernoulli}(p)$, where $0 \le p \le 1$, then the probability mass function of Y is

$$f_Y(y) = p^y(1-p)^{1-y}, \qquad y \in \{0, 1\}. \tag{1.58}$$

Now we can construct a model for the conditional distribution of Y given X as

$$f_{Y \mid X=x}(y) = p(x)^y(1-p(x))^{1-y}, \qquad y \in \{0, 1\},\ x \in \mathbf{R}^d, \tag{1.59}$$

where $p : \mathbf{R}^d \to [0,1]$ is a function. Note that in the Bernoulli model $EY = p$ and in the conditional Bernoulli model (1.59) the conditional expectation of Y given X is

$$E[Y \mid X = x] = P(Y = 1 \mid X = x) = p(x).$$

Since function p is a conditional expectation, we can use any regression method to estimate p. However, it can happen that a regression function estimate is such that it takes values outside the interval $[0,1]$. For example, a linear regression function estimate takes values outside the range $[0,1]$ for large or small enough values of the explanatory variables. There are several natural estimators for function p:

1. In a generalized linear model it is assumed that

$$p(x) = G(\alpha + \beta' x),$$

where $G : \mathbf{R}^d \to [0,1]$ is a known link function. Generalized linear models in the case of a binary response model are considered in Section 2.3.2.

2. In the single index model it is assumed that

$$p(x) = g(\alpha + \beta' x),$$

where $g : \mathbf{R}^d \to [0,1]$ is an unknown link function. Single link estimators are considered in Section 4.1.

3. We can estimate p with the help of a density function estimator, if vector X has a continuous distribution. If vector X has a continuous distribution, we can write

$$P(Y = 1 \mid X = x) = \frac{P(Y = 1) f_{X \mid Y=1}(x)}{f_X(x)},$$

where $f_{X \mid Y=1}$ is the density of $X \mid Y = 1$ and f_X is the density of X. The prior probability $P(Y = 1)$ can be estimated by

$$\hat{p}_1 = \frac{1}{n} \#\{i = 1, \ldots, n : Y_i = 1\}.$$

The densities $f_{X \mid Y=1}$ and f_X can be estimated by any density estimation method. For example, in kernel density estimation we take

$$\hat{f}_X(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i), \qquad \hat{f}_{X \mid Y=1}(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i) I_{\{1\}}(Y_i),$$

where $K_h(x) = K(x/h)/h^d$, $K : \mathbf{R}^d \to \mathbf{R}$ is the kernel function, and $h > 0$ is the smoothing parameter. See (3.39) for the definition of the kernel density estimator. Finally, we define the estimator of function $p : \mathbf{R}^d \to [0,1]$ as

$$\hat{p}(x) = \frac{\hat{p}_1 \hat{f}_{X \mid Y=1}(x)}{\hat{f}_X(x)}. \tag{1.60}$$

4. We can estimate function $p : \mathbf{R}^d \to [0,1]$ with a local averaging

$$\hat{p}(x) = \sum_{i=1}^{n} p_i(x)\, Y_i, \tag{1.61}$$

where the weights $p_i(x)$ satisfy $p_i(x) \geq 0$ and $\sum_{i=1}^{n} p_i(x) = 1$. Examples of the local averaging are given in Chapter 3, where regressogram weights, kernel weights, and nearest neighborhood weights are defined. In the case of kernel regression and kernel density estimation formulas (1.60) and (1.61) are equivalent, see (3.37).

1.2.2 Discrete Choice Models

In discrete choice models the response variable is a discrete random variable taking only a finite number of values. We can distinguish the cases where the values of the response variable are unordered and the cases where they are ordered. The random variables whose values are unordered are called nominal or categorical random variables and the random variables whose values are ordered are called ordinal random variables.

Let us consider a discrete choice model with a categorical response variable. A categorical response variable Y has a categorical distribution, taking K distinct values $0, 1, \ldots, K-1$, say. The categorical distribution family generalizes the Bernoulli distribution family, where the variable takes only values 0 and 1. When $Y \sim \text{Categorical}(p_0, \ldots, p_{K-1})$, where $0 \leq p_k \leq 1$, $\sum_{k=0}^{K-1} p_k = 1$, then the probability mass function of Y is

$$f_Y(y) = \sum_{k=0}^{K-1} p_i I_{\{k\}}(y), \qquad y \in \{0, \ldots, K-1\}. \tag{1.62}$$

Now we can construct a model for the conditional distribution of Y given X as

$$f_{Y \mid X=x}(y) = \sum_{k=0}^{K-1} p_k(x) I_{\{k\}}(y), \qquad y \in \{0, \ldots, K-1\},\ x \in \mathbf{R}^d, \tag{1.63}$$

where $p_k : \mathbf{R}^d \to [0,1]$ are functions satisfying $\sum_{k=0}^{K-1} p_k(x) = 1$ for each $x \in \mathbf{R}^d$. Note that now the conditional probability of Y given X is

$$P(Y = k \mid X = x) = p_k(x), \qquad k = 0, \ldots, K-1.$$

There are several reasonable estimators of $p_k(x)$.

1. We can use the parametric form

$$p_k(x) = \frac{e^{\beta_k' x}}{1 + \sum_{i=1}^{K-1} e^{\beta_i' x}},$$

for $k = 1, \ldots, K-1$ and $p_0(x) = 1 - \sum_{i=1}^{K-1} p_i(x)$. A more restrictive form is

$$p_k(x) = \frac{e^{\beta' x}}{\sum_{i=0}^{K-1} e^{\beta' x}}, \tag{1.64}$$

where the conditional probability is the same for all classes. This form is obtained by defining

$$U_i = \beta' X + \epsilon_i$$

and

$$Y = \text{argmax}_{i=0,\ldots,K-1} U_i.$$

Assume that ϵ_i are independent and identically distributed with the Weibull distribution. The distribution function of the Weibull distribution is $F_{\epsilon_i}(x) = \exp\{-e^{-x}\}$. Now $p_k(x) = P(Y = k \mid X = x)$ is given by (1.64). The estimation can be done with the maximum likelihood or with the least squares method.

2. We can estimate p with the help of any density function estimator. If vector X has a continuous distribution, then we can write,

$$P(Y = k \mid X = x) = \frac{P(Y = k) f_{X \mid Y=k}(x)}{f_X(x)},$$

where $k = 0, \ldots, K-1$, $f_{X \mid Y=1}$ is the density of $X \mid Y = 1$ and f_X is the density of X. The prior probability $P(Y = k)$ can be estimated by

$$\hat{p}_k = \frac{1}{n} \#\{i = 1, \ldots, n : Y_i = k\},$$

where $\#A$ denotes the cardinality of set A. The densities $f_{X \mid Y=k}$ and f_X can be estimated by any density estimation method. See (3.39) for the definition of the kernel density estimator. Finally we define the estimator of p as

$$\hat{p}(x) = \frac{\hat{p}_k \hat{f}_{X \mid Y=k}(x)}{\hat{f}_X(x)}. \tag{1.65}$$

3. Define K Bernoulli random variables $Y^{(0)}, \ldots, Y^{(K-1)}$ with the definition that $Y^{(k)} = 1$ if and only if $Y = k$. Then

$$p_k(x) = E(Y^{(k)} \mid X = x).$$

We can use, for example, kernel regression to estimate $p_k(x)$ using regression data $(X_1, Y_1^{(k)}), \ldots, (X_n, Y_n^{(k)})$, for $k = 0, \ldots, K-1$.

1.2.3 Count Data

Count data occurs when the response variable Y gives the number of occurrences of an event. For instance, Y could give the annual number of bank failures. The count data is such that Y takes values $\{0, 1, 2, \ldots\}$. Count data can be modeled with the Poisson distribution. If $Y \sim \text{Poisson}(\nu)$, then

$$P(Y = y) = e^{-\nu} \frac{\nu^y}{y!}, \qquad y = 0, 1, 2, \ldots,$$

where $\nu > 0$ is the unknown intensity parameter. Now $EY = \nu$ and $\text{Var}(Y) = \nu$. In the Poisson regression the regression function is

$$\nu(x) = E(Y \mid X = x),$$

where $X \in \mathbf{R}^d$ is the vector of explanatory variables. The Poisson regression is a heteroskedastic regression model. A parametric Poisson regression model is obtained if

$$\nu(x) = \exp\{x'\beta\},$$

where $\beta \in \mathbf{R}^d$ is the unknown parameter. This choice guarantees that $\nu(x) > 0$. Besbeas, de Feis & Sapatinas (2004) make a comparative simulation study of wavelet shrinkage estimators for Poisson counts.

1.3 PARAMETRIC FAMILY REGRESSION

We obtain the binary response models, discrete choice models, and Poisson count models, introduced in Section 1.2, as special cases of parametric family regression, introduced in Section 1.3.1. In fact, these are a special cases of exponential family regression, introduced in Section 1.3.2. A different type of parametric family regression is obtained by copula modeling, introduced in Section 1.3.3.

1.3.1 General Parametric Family

Let us consider a family $(P_\theta, \theta \in \Theta)$ of probability measures, where $\Theta \subset \mathbf{R}^p$. Let $Y \in \mathbf{R}$ be a response variable and let $X \in \mathbf{R}^d$ be a vector of explanatory variables that satisfy

$$Y \sim P_{f(X)},$$

where $f : \mathbf{R}^d \to \Theta$ is an unknown function to be estimated. The function f is estimated using identically distributed observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ from the distribution of (X, Y). After estimating function f, we have an estimator of the conditional distribution $Y \mid X = x$, because

$$Y \mid X = x \sim P_{f(x)}. \tag{1.66}$$

The following examples illustrate the model.

1. We obtain a Gaussian mean regression model when $P_\theta = N(\theta, \sigma^2)$, where $\theta \in \Theta = \mathbf{R}$. Now
$$Y \mid X = x \sim N\left(f(x), \sigma^2\right),$$
which follows from
$$Y = f(X) + \epsilon,$$
where $\epsilon \sim N(0, \sigma^2)$.

2. We obtain a Gaussian volatility model, when $P_\theta = N(0, \theta)$, where $\theta \in \Theta = (0, \infty)$. Now
$$Y \mid X = x \sim N(0, f(x)),$$
which follows from
$$Y = f(X)^{1/2} \epsilon,$$
where $\epsilon \sim N(0, 1)$.

3. We obtain a Gaussian heteroskedastic mean regression model, when $P_\theta = N(\theta_1, \theta_2)$, where $\theta = (\theta_1, \theta_2)$, and $\Theta = \mathbf{R} \times (0, \infty)$. Now
$$Y \mid X = x \sim N(f_1(x), f_2(x)),$$
which follows from
$$Y = f_1(X) + f_2(X)^{1/2} \epsilon,$$
where $\epsilon \sim N(0, 1)$, and we denote $f = (f_1, f_2)$.

4. We obtain the binary choice model, when $P_\theta = \text{Bernoulli}(\theta)$, where $\theta \in \Theta = [0, 1]$. Then $P(Y = 1) = f(X)$ and $P(Y = 0) = 1 - f(X)$.

Let us assume that the probability measures P_θ are dominated by a σ-finite measure, and denote the density functions of P_θ by $p(y, \theta)$. We use the term density function, although $p(y, \theta)$ can be also a probability mass function, if Y has a discrete distribution. In Section 1.3.2 we make the assumption that $(P_\theta, \theta \in \Theta)$ is an exponential family.

Under the assumption that $(X_1, Y_1), \ldots, (X_n Y_n)$ are i.i.d., the log-likelihood of the sample is
$$\sum_{i=1}^{n} \log p(Y_i, f(X_i)).$$
The log-likelihood can be maximized over collection $\mathcal{F}$ of functions, and we denote $\mathcal{F} = (f_\beta, \beta \in \mathcal{B})$. We have two general approaches.

1. The first possibility is to define
$$\hat{f} = \text{argmax}_{\beta \in \mathcal{B}} \sum_{i=1}^{n} \log p(Y_i, f_\beta(X_i)),$$
where $(f_\beta, \beta \in \mathcal{B})$ is a large collection of functions, like the collection of linear functions: $f_\beta(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d$.

2. A second possibility is to maximize a local log-likelihood and define

$$\hat{f}(x) = \text{argmax}_{f \in \mathcal{F}} \sum_{i=1}^{n} \log p(Y_i, f(X_i))\, p_i(x), \tag{1.67}$$

where $p_i(x)$ are weights, for example $p_i(x) = K_h(x - X_i)$, where $K_h(x) = K(x/h)/h^d$, $K : \mathbf{R}^d \to \mathbf{R}$, and $h > 0$. Now we can take f_β to be a constant function: $f_\beta(x) = \beta$, where $\beta \in \mathbf{R}$. The local likelihood approach has been covered in Spokoiny (2010).

1.3.2 Exponential Family Regression

An exponential family is a collection $\mathcal{P} = (P_\theta, \theta \in \Theta)$ of probability measures. The probability measures in $\mathcal{P}$ are dominated by a σ-finite measure. In a one-parameter exponential family the density functions have the form

$$p(y, \theta) = p(y) \exp\{yc(\theta) - b(\theta)\},$$

where $\theta \in \Theta \subset \mathbf{R}$, and $y \in \mathcal{Y} \subset \mathbf{R}$. The functions c and b are nondecreasing functions on Θ and function $p : \mathcal{Y} \to \mathbf{R}$ is nonnegative. In the exponential family with the canonical parameterization the density functions are

$$p(y, v) = p(y) \exp\{yv - d(v)\}. \tag{1.68}$$

The canonical parameterization is obtained by putting $v = c(\theta)$ and $d(v) = b(\theta)$. Examples of exponential families include the family of Gaussian, Bernoulli, Poisson, and gamma distributions. An exposition of exponential families is given by Brown (1986).

We use the modeling approach in (1.66), and assume that the conditional distribution of Y given $X = x$ belongs to an exponential family and the parameter of the conditional distribution is $v = f(x)$:

$$Y \mid X = x \sim p(y, f(x)), \tag{1.69}$$

for a function $f : \mathbf{R}^d \to \mathcal{V}$, where we use the natural parameterization in (1.68), and $\mathcal{V}$ is the parameter space of the natural parameter.

If the parameterization is natural, and d is continuously differentiable, then

$$E_v Y = d'(v), \tag{1.70}$$

where $Y \sim f(y, v)$. Indeed,

$$\frac{\partial}{\partial v} \log p(y, v) = y - d'(v).$$

On the other hand,

$$E_v \frac{\partial}{\partial v} \log p(Y, v) = 0,$$

under regularity assumptions.[5] Thus, (1.70) holds. Under the assumption (1.69), we get

$$E(Y \mid X = x) = d'(f(x)).$$

If the parameterization is natural, and d is two times continuously differentiable, then

$$\mathrm{Var}_v(Y) = d''(v). \tag{1.71}$$

Indeed,

$$\begin{aligned} \mathrm{Var}_v(Y) &= E\left(Y - d'(v)\right)^2 = E\left[\frac{\partial}{\partial v}\log p(Y,v)\right]^2 = -E\,\frac{\partial^2}{\partial v^2}\log p(Y,v) \\ &= d''(v). \end{aligned}$$

Under the assumption (1.69), we get

$$\mathrm{Var}(Y \mid X = x) = d''(f(x)).$$

Brown, Cai & Zhou (2010) suggest a reduction method where the exponential family regression can be transformed to the Gaussian regression by binning and variance stabilizing transform.

1.3.3 Copula Modeling

Let (Y_1, Y_2) be a random vector with a continuous distribution function

$$F(y_1, y_2) = P(Y_1 \le y_1, Y_2 \le y_2),$$

where $y_1, y_2 \in \mathbf{R}$. We can write the distribution function uniquely as

$$F(y_1, y_2) = C(F_1(y_1), F_2(y_2)), \tag{1.72}$$

where $F_1(y_1) = P(Y_1 \le y_1)$ and $F_2(y_2) = P(Y_2 \le y_2)$ are the distribution functions of Y_1 and Y_2. The function $C : [0,1]^2 \to \mathbf{R}$ is the copula of the distribution of (Y_1, Y_2). Function C is a distribution function whose marginals are uniform on $[0,1]$. The copula is defined by

$$C(u_1, u_2) = F\left(F_1^{-1}(u_1), F_2^{-1}(u_2)\right),$$

where $u_1, u_2 \in [0,1]^2$. These facts were proved in Sklar (1959). See also Nelsen (1999).

For example, a Gaussian two dimensional copula is a normal distribution with unit marginal standard deviations. The family of Gaussian two dimensional copulas C_θ has the parameter $\theta \in (-1, 1)$, where θ is the correlation coefficient between Y_1 and Y_2.

[5] We have that $E_v \frac{\partial}{\partial v}\log p(Y,v) = E_v \frac{\partial p(Y,v)/\partial v}{p(Y,v)} = \int \frac{\partial}{\partial v} p(y,v)\,dy = \frac{\partial}{\partial v}\int p(y,v)\,dy = \frac{\partial}{\partial v} 1 = 0$, if the order of derivation and integration can be changed.

The copula representation of the distribution as in (1.72) gives a useful way to construct models and to estimate the unknown parameters of the model. Let $(c_\theta, \theta \in \Theta)$ be a family of copula densities, where $\Theta \subset \mathbf{R}^p$. This leads to a semiparametric model with densities

$$f(y_1, y_2; \theta, f_1, f_2) = c_\theta\left(F_1(y_1), F_2(y_2)\right)\, f_1(y_1)\, f_2(y_2),$$

where $\theta \in \Theta$ and $f_1, f_2 \in \mathcal{F}$, where $\mathcal{F}$ is a nonparametric collection of univariate density functions. The estimation of θ, f_1, f_2 can be done with the two stage approach. In the first stage we estimate nonparametrically the marginal distributions f_1 and f_2. In the second stage we estimate the copula parameter θ.

Assume that $X \in \mathbf{R}^d$ is a vector of explanatory variables and we want to estimate the conditional distribution $(Y_1, Y_2) \,|\, X = x$. We assume that the conditional distribution function is

$$\begin{aligned} F_{Y_1,Y_2|X=x}(y_1, y_2) &= P(Y_1 \le y_1, Y_2 \le y_2 \,|\, X = x) \\ &= C_{\theta(x)}\left(F_{Y_1|X=x}(y_1), F_{Y_2|X=x}(y_2)\right), \end{aligned}$$

where $\theta : \mathbf{R}^d \to \mathbf{R}$. The conditional density is

$$f_{Y_1,Y_2|X=x}(y_1, y_2) = c_{\theta(x)}\left(F_{Y_1|X=x}(y_1), F_{Y_2|X=x}(y_2)\right)\, f_{Y_1|X=x}(y_1)\, f_{Y_2|X=x}(y_2).$$

In the first stage we estimate nonparametrically the conditional distribution functions and get the estimates $\hat{F}_{Y_1|X=x}(y_1)$ and $\hat{F}_{Y_2|X=x}(y_2)$. In the second stage we estimate the function $\theta(x)$. This can be done analogously to (1.67), and we get the locally constant likelihood estimator as

$$\hat{\theta}(x) = \operatorname{argmin}_{\theta\in\Theta} \sum_{i=1}^{n} p_i(x) \log c_\theta\left(\hat{F}_{Y_1|X=x}(y_1), \hat{F}_{Y_2|X=x}(y_2)\right).$$

This method has been studied in Abegaz, Gijbels & Veraverbeke (2012).

We have defined in (1.72) the standard copula decomposition. This decomposition can be inconvenient because the copula density c has the support inside $[0,1]^2$, and the estimation is typically complicated with the boundary effects. Alternatively, we can make the copula decomposition

$$F(x_1, x_2) = C\left(\Phi^{-1}(F_1(x_1)), \Phi^{-1}(F_2(x_2))\right),$$

where $\Phi : \mathbf{R} \to \mathbf{R}$ is the distribution function of the standard Gaussian distribution. Now C is a distribution function whose marginals are standard Gaussian, and C is defined by

$$C(u, v) = F\left(F_1^{-1}\left(\Phi(u)\right), F_2^{-1}\left(\Phi(v)\right)\right), \qquad u, v \in \mathbf{R}.$$

1.4 CLASSIFICATION

Let the sequence $(X_1, Y_1), \ldots, (X_n, Y_n)$ consist of identically distributed random vectors. Let (X, Y) be distributed as (X_i, Y_i), for $i = 1, \ldots, n$. Let the possible

values of Y be $\{0, \ldots, K-1\}$. We want to find a classification function $g : \mathbf{R}^d \to \{0, \ldots, K-1\}$. The classification function is interpreted as such function that if we observe a new random variable X_{n+1}, distributed as X, then $g(X_{n+1})$ guesses the class label of X_{n+1}, that is, we decide that X_{n+1} comes from the distribution of $X \mid Y = k$, if $g(X_{n+1}) = k$.

In the case of classification Y can take only a finite number of values (as many values as there are classes), since the values of the response variable Y indicate the class label. In the case of regression analysis the response variable Y can in many cases take as values any real number. However, in Section 1.2.1 we have considered binary response models, where the response variable takes only two values and in Section 1.2.2 we have considered discrete choice models, where the response variable takes a finite number of values. In binary response models and in discrete choice models we are, however, interested in estimating the conditional expectation $f(x) = E(Y \mid X = x)$, $f : \mathbf{R}^d \to \mathbf{R}$, whereas in the case of classification, we want to estimate the classification function $g : \mathbf{R}^d \to \{0, \ldots, K-1\}$, which predicts the class label of a future observation. As an example, consider the case where $K = 2$, so that there are two classes, and thus Y is a Bernoulli distributed random variable. Now the regression function is

$$f(x) = E(Y \mid X = x) = P(Y = 1 \mid X = x). \tag{1.73}$$

Thus $f(X_{n+1}) \in [0, 1]$, but we would like to find a classification function g such that $g(X_{n+1}) \in \{0, 1\}$.

We have explained the classification in the case where the number of observations in each class is a random number. There also exist cases where the observation number in each class can be chosen by the designer of the experiment. Then we have fixed numbers $n_0, \ldots, n_{K-1}$ and observations $X_{k1}, \ldots, X_{kn_k} \in \mathbf{R}^d$ are coming from the kth distribution, $k = 0, \ldots, K-1$. We will consider only the case where the class frequencies are random.

1.4.1 Bayes Risk

In the random design regression we can motivate the estimation of conditional expectation $f(x) = E(Y \mid X = x)$ by noting that the conditional expectation minimizes the mean squared error: $f = \operatorname{argmin}_g E(Y - g(X))^2$; see Section 1.1.7. Similarly, in the case of classification, we can find a population quantity which minimizes a natural criterion. This criterion is the probability of misclassification, or Bayes risk,

$$R(g) = P(g(X) \neq Y). \tag{1.74}$$

Let

$$g^* = \operatorname{argmin}_g R(g),$$

where the minimization is done over all classification functions $g : \mathbf{R}^d \to \{0, \ldots, K-1\}$. The classification function g^* which minimizes the probability of misclassification is called the Bayes rule. It can be proved that

$$g^*(x) = \operatorname{argmax}_{k=0,\ldots,K-1} P(Y = k \mid X = x). \tag{1.75}$$

The proof of (1.75) for the case $K = 2$ can be found in Györfi et al. (2002, Lemma 1.1, p. 6). It holds that

$$g^*(x) = \text{argmax}_{k=0,\dots,K-1} P(Y = k)\, f_{X|Y=k}(x), \tag{1.76}$$

where $f_{X|Y=k} : \mathbf{R}^d \to \mathbf{R}$ is the density function of $X \mid Y = k$.

1.4.2 Methods of Classification

We shall mention four principles to construct classification functions: classification using regression function estimates, classification using density estimates, classification using empirical risk minimization, and classification using nearest neighbors.

Classification by Regression Function Estimation A classification function can be constructed from a regression function estimate. In the two class case we can take the data $(X_1, Y_1), \dots, (X_n, Y_n)$ as if it would originate from the binary response model, and in the multiclass case we can take the data as if it would originate from the discrete choice model with a categorical response variable. In Section 1.2.1 we introduced binary response models, and in Section 1.2.2 we introduced discrete choice models.

In a discrete choice model we estimate the class posterior probabilities

$$p_k(x) = P(Y = k \mid X = x), \qquad k = 0, \dots, K-1.$$

A natural classification function is

$$g^*(x) = \text{argmax}_{k=0,\dots,K-1} p_k(x). \tag{1.77}$$

In fact, we note in (1.75) that the classification function g^* is in a sense the optimal classification function. Let us denote the estimators of the class posterior probabilities by $\hat{p}_k(x)$, and let us define an estimator of the classification function by

$$\hat{g}(x) = \text{argmax}_{k=0,\dots,K-1} \hat{p}_k(x). \tag{1.78}$$

We can find the estimators $\hat{p}_k(x)$ in the following way. We define K response variables, that are the indicators of the class labels:

$$Y_i^{(k)} = I_{\{k\}}(Y_i), \qquad i = 1, \dots, n, \;\; k = 0, \dots, K-1. \tag{1.79}$$

Let $\hat{p}_k(x)$ be a regression function estimator of the posterior probability

$$p_k(x) = E(Y^{(k)} \mid X = x) = P(Y = k \mid X = x). \tag{1.80}$$

Estimator $\hat{p}_k(x)$ is constructed using regression data $(X_1, Y_1^{(k)}), \dots, (X_n, Y_n^{(k)})$, for $k = 0, \dots, K-1$.

In the two class case, when $Y \in \{0, 1\}$, we do not have to use (1.79), because Y is already a class indicator. In the two class case we can write the empirical decision

rule in a simplified form. Let $\hat{f} : \mathbf{R}^d \to \mathbf{R}$, be any regression function estimator constructed using regression data $(X_1, Y_1), \ldots, (X_n, Y_n)$, and define

$$\hat{g}(x) = \begin{cases} 1, \text{ if } \hat{f}(x) \geq 1/2, \\ 0, \text{ otherwise,} \end{cases} \tag{1.81}$$

which estimates the natural classification function

$$g(x) = \begin{cases} 1, \text{ if } P(Y = 1 \,|\, X = x) \geq P(Y = 0 \,|\, X = x), \\ 0, \text{ otherwise.} \end{cases} \tag{1.82}$$

Classification by Density Estimation A classification function can be constructed from density estimates of the class densities. We assume now that X is a random vector with a continuous distribution. Let us consider the classification rule $g^*(x) = \text{argmax}_{k=0,\ldots,K-1} p_k(x)$, defined in (1.77). We can write

$$p_k(x) = P(Y = k \,|\, X = x) = \frac{P(Y = k) f_{X|Y=k}(x)}{f_X(x)},$$

where $k = 0, \ldots, K-1$, and $x \in \mathbf{R}^d$. Thus,

$$\text{argmax}_{k=0,\ldots,K-1} p_k(x) = \text{argmax}_{k=0,\ldots,K-1} P(Y = k) f_{X|Y=k}(x).$$

An estimator for the classification function, based on data $(X_1, Y_1), \ldots, (X_n, Y_n)$, is obtained as

$$\hat{g}(x) = \text{argmax}_{k=0,\ldots,K-1} \hat{p}_k \hat{f}_{X|Y=k}(x), \tag{1.83}$$

where $\hat{f}_{X|Y=k}$ is a density estimator of the class density function $f_{X|Y=k}$ and $\hat{p}_k$ is an estimator of the class prior probability $P(Y = k)$. We can define

$$\hat{p}_k = \frac{1}{n} \#\{i = 1, \ldots, n : Y_i = k\}.$$

Classification by Empirical Risk Minimization A classification function can be constructed using empirical risk minimization. In (1.78), classification is reduced to regression function estimation (in binary response models or in discrete choice models). In (1.83), classification is reduced to density estimation. However, according to Vapnik's principle, we should not try to estimate more than is needed, and thus we should also consider the direct construction of a classification function, without reducing the problem to regression function estimation or to density function estimation.

We define a classifier by

$$\hat{g} = \text{argmin}_{g \in \mathcal{G}} \gamma_n(g), \tag{1.84}$$

where $\mathcal{G}$ is a class of functions $g : \mathbf{R}^d \to \{0, \ldots, K-1\}$ and $\gamma_n(g)$ is the empirical error of classifier g. We get different classifiers depending on the choice of the empirical error $\gamma_n(g)$ and depending on the choice of class $\mathcal{G}$.

We can define the empirical error of a classifier g by

$$\gamma_n(g) = \#\{i = 1, \ldots, n : g(X_i) \neq Y_i\}. \tag{1.85}$$

Quantity $\gamma_n(g)$ is equal to the number of misclassifications in the learning sample. We can also decompose the number of misclassifications according to the class labels. Let

$$\gamma_n^{(k)}(g) = \#\{i = 1, \ldots, n : g(X_i) \neq k, Y_i = k\},$$

where $k = 0, \ldots, K-1$. Then we can define the empirical error of the classifier as a weighted sum of the single class misclassification errors:

$$\gamma_n(g) = \sum_{k=0}^{K-1} w_k \gamma_n^{(k)}(g).$$

For $w_k \equiv 1$ we get the overall classification error (1.85).

In the two class case it has also been suggested to use class labels $Y \in \{-1, 1\}$ and not the labels $\{0, 1\}$, consider classifiers $h : \mathbf{R}^d \to \mathbf{R}$, and define the classification function $g(x) = \text{sign}(h(x))$. The empirical risk is defined as

$$\gamma_n(g) = \sum_{i=1}^{n} \phi(Y_i h(X_i)),$$

where $\phi : \mathbf{R} \to (0, \infty)$ is a convex nonincreasing function with $\phi(u) \geq I_{(-\infty,0]}(u)$ for $u \in \mathbf{R}$. We can take the hinge loss $\phi(u) = \max\{0, 1-u\}$, the exponential loss $\phi(u) = \exp\{-u\}$, or the logit loss $\phi(u) = \log_2(1 + e^{-u})$. Support vector machines, mentioned in Section 5.3, use the hinge loss and a penalized empirical risk.

An example for the choice of class $\mathcal{G}$ is given in (2.84). In this example the class $\mathcal{G}$ is chosen so that the classification functions are linear.

Classification by Nearest Neighbors The nearest-neighbor rule defines the class estimate to be that class label that occurs most often among the k nearest neighbors. That is, for an integer $k \in \{1, 2, \ldots\}$, define the k nearest neighbors, based on observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, as the set

$$\mathcal{Y}(x) = \{Y_i : \|X_i - x\| \leq r_k(x)\},$$

where

$$r_k(x) = \min\{r > 0 : \#\{X_i \in B_r(x)\} = k\},$$

where $B_r(x) = \{z \in \mathbf{R}^d : \|z - x\| \leq r\}$. Now we can define the classifier[6]

$$\hat{g}(x) = \text{argmax}_{y=0,\ldots,K-1} \#\{Y_i \in \mathcal{Y}(x) : Y_i = y\}.$$

Hastie et al. (2001, Section 13) use the term "prototype methods" to denote classifiers which classify the new observation to the class whose observed values are most similar to the new observation.

[6]We denote now the class labels by $y = 0, \ldots, K-1$, because the symbol k is used to denote the number of nearest neighbors.

1.5 APPLICATIONS IN QUANTITATIVE FINANCE

Portfolio selection, risk management, and option pricing belong to the main branches of quantitative finance. Estimation of conditional variances and conditional quantiles can be applied in risk management. Estimation of conditional expectations can be applied in portfolio selection. Option pricing is related to optimal control.

Other applications are described in later sections. Section 2.1.7 explains how linear regression can be applied to estimate the beta of an asset, the beta of a portfolio, the alpha of a portfolio, and the alpha of a hedge fund. Section 2.2.2 explains how varying coefficient regression can be applied in hedge fund replication and in performance measurement. Data sets are described in Section 1.6.

1.5.1 Risk Management

The process of portfolio selection tries to address the problem of balancing the risk and return, but it is useful to have an independent risk management to make an evaluation of the risk of the portfolios at a daily basis.

The economic capital can roughly be defined to mean the amount of money which is needed to secure survival of a company in a worst case scenario. The definition of the economic capital can be made precise with the concept of a value at risk. The economic capital can be used in portfolio selection to calculate return distributions. The regulatory capital is the capital required by the regulators that financial institutions should maintain. The regulatory capital is often defined in terms of value at risk.

Variance trading can be used in speculation, but variance swaps can also be used in risk management to adjust the overall exposure of a portfolio to the volatility.

Value-at-Risk Quantiles can be used to measure the risk of a portfolio. The distribution of the change in the value of the portfolio is called the profit-and-loss distribution: If we denote by V_t the value of the portfolio at time t and by V_u the value of the portfolio at a later time, then the distribution of $V_u - V_t$ is called the profit-and-loss distribution for the time period from t to u. We define the loss as the negative of the change in the value of the portfolio

$$L_u = -(V_u - V_t). \tag{1.86}$$

The upper quantiles of the loss distribution are called the value at risk or VaR:

$$\text{VaR}_p = Q_p(L_u), \tag{1.87}$$

where p is equal to 0.99 or 0.999, for example. A larger value of VaR_p indicates that the portfolio is more risky, because VaR_p is such threshold that the probability that the loss is larger than VaR_p is smaller or equal to $1 - p$. We can write

$$L_u = -V_t R_u,$$

where R_u is the return of the portfolio,

$$R_u = \frac{V_u - V_t}{V_t}.$$

Thus, if we have the quantile $Q_p(R_u)$ of the return distribution, the VaR_p is obtained by the formula

$$\mathrm{VaR}_p = -V_t\, Q_p(R_u).$$

Quantile as risk measure takes into account the number of exceedances of the VaR threshold, but it does not take into account the largeness of the exceedances. Expected shortfall takes also the largeness of the exceedances into account.

Economic Capital in Portfolio Selection Let us consider a bank which wants to choose among a collection of investment proposals. The investment with the best return distribution will be chosen. The problem is to calculate the return distribution since many investments do not require any initial capital, and we cannot calculate the return by dividing by the initial investment.

First we have to construct a profit–loss distribution for each investment proposal. These profit–loss distributions may be very difficult to estimate, because one has to take into account each possible future state of affairs and its probability. In order to estimate the probabilities of the states, one has to take into account all current investments of the bank and consider the interaction of the new investment with the current investments. For example, when we write a call option, the maximum loss is in general infinite; but if we already own the underlying stock, then the loss is bounded.

We want to set aside enough capital to cover adverse events with a given probability of occurrence. The frequency of default for AA-rated companies over a one-year period has been roughly one in three thousand. Thus one could choose the 0.0003th quantile of the profit–loss distribution ($1/3000 \approx 0.0003 = 0.03\%$), which could be for example a loss of 1 million Euros, and set aside enough capital to cover this loss. This capital is called the economical capital. The return on investment is calculated by dividing by the economical capital. That is, we get the return distribution from the profit–loss distribution by dividing with the economical capital. See Rebonato (2007, Chapter 9).

Finally, we choose the best return distribution by the maximization of the expected utility or by the maximization of the variance penalized expected return.

Variance as a Risk Measure The Sharpe ratio of a portfolio is defined as

$$\frac{E(R-r)}{\mathrm{sd}(R-r)}, \tag{1.88}$$

where R is the return of the portfolio for a given time period, r is the return of a risk-free rate for the same time period, and $\mathrm{sd}(R-r)$ is the standard deviation. The Sharpe ratio belongs to the class of performance measures having the form

$$\frac{\text{expected return}}{\text{risk}}.$$

The basic idea is that in measuring the quality of a portfolio we have to take the risk into account and not only the return. In the definition of the Sharpe ratio the

expected return and the standard deviation is defined using the *excess return*, which is the return of the portfolio minus the return of a risk-free return.

In portfolio selection the risk aversion can be taken into account by using the Markowitz criterion

$$E(R-r) - \frac{\lambda}{2} \cdot \mathrm{sd}(R-r), \tag{1.89}$$

where $\lambda \geq 0$ is the risk aversion parameter. The Markowitz criterion has the general structure of a risk-penalized expected return:

$$\text{expected return} - \frac{\lambda}{2} \cdot \text{risk}.$$

The Sharpe ratio and the Markowitz criterion use the standard deviation of the excess returns as the risk measure. The standard deviation does not take into account the possibility of a nonsymmetric distribution. It penalizes from a positive skewness of the return distribution. Thus, we can consider replacing the standard deviation by the square root of the partial variance in the definition of the Sharpe ratio and the Markowitz criterion. The partial variance is defined in (1.18).

1.5.2 Variance Trading

Variance estimation can be applied in quantile estimation, because standard deviation estimates can be used to construct quantile estimates; see (1.28)–(1.30). Variance estimation can be applied in portfolio performance measurement and in portfolio selection, see (1.88) and (1.89). A third application for variance estimation comes from the volatility trading.

Volatility can be traded with variance and volatility swaps. A variance swap is a forward contract that pays

$$V_T - K$$

at the expiration date T, where K is the delivery price, and V_T is the realized variance, defined by

$$V_T = \sum_{t=t_0+1}^{T} [\log(S_t/S_{t-1})]^2,$$

where t_0 is the starting day of the contract, and S_t are the prices of a financial asset. The volatility swap pays at the expiration

$$\sqrt{V_T} - L,$$

where L is the delivery price.

Variance and volatility swaps are traded over the counter (OTC), but the Chicago Board Options Exchange (CBOE) offers variance futures for the variance of the S&P 500 index, calculated with the daily returns of the index.

Variance swaps open an opportunity to covariance trading if we have an access to a variance swap of an index and to variance swaps of its constituents. Let us consider an index whose returns are

$$R_t = pR_t^{(1)} + qR_t^{(2)},$$

where $R_t^{(i)}$ are the log returns of the index constituents and p and q are the weights of of the constituents. Let us define the realized covariance as

$$C_T = \sum_{t=t_0+1}^{T} R_t^{(1)} R_t^{(2)}.$$

Thus,

$$C_T = \frac{1}{2pq}\left(V_T - p^2 V_T^{(1)} - q^2 V_T^{(2)}\right),$$

where V_T is the realized variance of the index and $V_T^{(i)} = \sum_{t=t_0}^{T} (R_t^{(i)})^2$ are the realized variances of the index constituents.

1.5.3 Portfolio Selection

Basic Concepts of Portfolio Selection Let

$$S_t = (S_t^1, \ldots, S_t^N), \qquad t = 0, 1, 2, \ldots, T,$$

be a vector time series of N asset prices. Asset prices satisfy $0 < S_t^i < \infty$, $i = 1, \ldots, N$. A portfolio vector $b_t = (b_t^1, \ldots, b_t^N) \in \mathbf{R}^N$ determines how the wealth is allocated among the assets at time t. A portfolio vector b_t satisfies

$$\sum_{i=1}^{N} b_t^i = 1. \tag{1.90}$$

When $0 \leq b_t^i \leq 1$ for all $i = 1, \ldots, N$, then the portfolio is called a long only portfolio and the value b_t^i is equal to the proportion of wealth is invested in asset S_t^i at time t. Negative values of b_t^i are interpreted as short selling.[7] One of the assets can be a bank account, and selling a bank account short is interpreted as borrowing. For example, when $N = 2$, $b_t^1 = -1$, and $b_t^2 = 2$, this means that at time t we sell short asset 1 with an amount which equals all our wealth and simultaneously buy asset 2 with all our wealth and with the proceedings obtained from selling short the asset 1.

We define a new vector time series of gross returns (price relatives) by

$$R_t = \frac{S_t}{S_{t-1}} = \left(\frac{S_t^1}{S_{t-1}^1}, \ldots, \frac{S_t^N}{S_{t-1}^N}\right), \qquad t = 1, 2, \ldots, T. \tag{1.91}$$

It is reasonable to assume that the time series $R_1, \ldots, R_T$ is approximately stationary. In statistical portfolio selection we have available, besides the historical returns R_t of the assets, also other information Z_t. The variables in vector Z_t can be macroeconomic variables, like the term premium, default premium, and dividend yield.[8] The problem of portfolio selection can now be described as a problem of choosing a portfolio vector b_T at time T using data (R_t, Z_t), $t = 1, \ldots, T$.

[7] Selling short an asset means that we borrow the asset and then sell it, that is, we sell an asset that we do not own. Naked short selling means that an asset is sold before it is borrowed, or before making sure that it can be borrowed.

[8] The term premium is the difference between the long-term and short-term interest rates. For example, the term premium can be the difference between the annualized yields of a portfolio of 10-year U.S.

Single-Period Portfolio Selection Let $W_T > 0$ be the wealth available at time T. When the portfolio vector is b_T, then the gross return of the portfolio for the time period from T to $T+1$ is

$$\frac{W_{T+1}}{W_T} = \sum_{i=1}^{N} b_T^i \frac{S_{T+1}^i}{S_T^i} = b_T' R_{T+1}. \tag{1.92}$$

In the single-period portfolio selection the optimal portfolio vector can be defined as

$$b_T^o = \text{argmax}_{b_T \in B_N} E_T u\left(b_T' R_{T+1}\right), \tag{1.93}$$

where $u : (0, \infty) \to \mathbf{R}$ is a utility function and

$$B_N = \left\{ (b^1, \ldots, b^N) : \sum_{i=1}^{N} b^i = 1 \right\}. \tag{1.94}$$

Note that $0 < b_T' R_{T+1} < \infty$. The notation E_T means that the expectation is taken at time T, using information available at time T. If the available information is contained in the historical returns R_t and in the historical values of the variables Z_t, then the expectation E_T can be taken as the conditional expectation, conditional on the previous returns and previous values of variables Z_t:

$$E_T u\left(b_T' R_{T+1}\right) = E\left[u\left(b_T' R_{T+1}\right) \mid R_1, Z_1, \ldots, R_T, Z_T\right].$$

In the maximization problem (1.93) we apply utility function u to the one-period gross return, given in (1.92).

A utility function $u : (0, \infty) \to \mathbf{R}$ is an increasing function (the derivative is positive) that is concave (the second derivative is negative). The power utility functions are defined by

$$u_\gamma(t) = \begin{cases} \frac{t^{1-\gamma}}{1-\gamma}, & \text{if } \gamma > 1, \\ \log_e t, & \text{if } \gamma = 1, \end{cases} \tag{1.95}$$

for $t > 0$. The power utility functions are called constant relative risk aversion utility functions (CRRA). A utility function is used instead of the pure return, because through it we take also risk into account and do not optimize the pure return.[9]

government bonds and a 90-day Treasury bill. The default premium is the difference between the interest rate of a lower grade bond and a higher grade bond. For example, the default premium can be the difference between the annualized yields of Moody's Baa and Aaa rated bonds. The dividend yield is the dividend payment of a company divided by its market capitalization. when the market capitalization is the value of the stock multiplied by the number of stocks.

[9]It does not matter whether we take the utility from the wealth or from the gross return. Indeed, for $\gamma > 0$,

$$u\left(W_T b_T' U_{T+1}\right) = W_T^{1-\gamma} \cdot u(b_T' U_{T+1})$$

and for $\gamma = 0$,

$$u\left(W_T b_T' U_{T+1}\right) = u(W_T) + u(b_T' U_{T+1}).$$

Thus the optimal portfolio vector is the same regardless of the initial wealth W_T.

Parameter $\gamma \geq 1$ is the risk aversion parameter, and larger γ means larger risk aversion. A gross return equal to zero would mean that we have made a bankruptcy, and thus the utility of zero gross return should be equal to minus infinity. Thus the utility function makes a severe penalization of returns near zero. Also, the utility of a positive return does not grow linearly but is a concave function of the return.

Multiperiod Portfolio Selection When we start with wealth W_T at time T and use portfolio weights $b_T, \ldots, b_{T_1-1}$, then the wealth at time T_1 is

$$W_T \prod_{t=T}^{T_1-1} b_t' R_{t+1}.$$

The gross return of the portfolio for the time period from T to T_1 is

$$\prod_{t=T}^{T_1-1} b_t' R_{t+1}. \tag{1.96}$$

In the multiperiod portfolio selection, assuming that our investment horizon extends from T to a future time T_1, and we are able to change the portfolio weights at all times $T, \ldots, T_1 - 1$, the optimal portfolio weights at time T are defined by

$$b_T^o = \text{argmax}_{b_T} \max_{b_{T+1}, \ldots, b_{T_1-1}} E_T u\left(\prod_{t=T}^{T_1-1} b_t' R_{t+1} \right). \tag{1.97}$$

In the maximization problem (1.97) we apply utility function u to the multiperiod gross return, given in (1.96). The single period case is obtained as a special case when $T_1 = T + 1$. The optimization problem (1.97) is of the same type as the optimization problem of the stochastic control in (1.52).

Portfolio Selection and Regression Function Estimation We describe how regression function estimation can be used in portfolio selection. We consider the single period portfolio selection and want to choose a portfolio vector $b_T = (b_T^1, \ldots, b_T^N) \in \mathbf{R}^N$ at time T so that the expected utility of the wealth is maximized at time $T + 1$, as in the optimization problem (1.93). We can define, for a fixed portfolio vector $b \in \mathbf{R}^N$, with $\sum_{i=1}^N b^i = 1$, the response and the explanatory variables

$$Y_{b,t} = u\left(b' R_{t+1}\right), \qquad X_t \in \mathbf{R}^d,$$

$t = 1, \ldots, T - 1$. We assume that $(Y_{b,t}, X_t)$, $t = 1, \ldots, T - 1$, are identically distributed, and denote by (Y_b, X) a random vector which has the same distribution as $(Y_{b,t}, X_t)$. The data can be used to estimate the regression function

$$f_b(x) = E(Y_b \,|\, X = x), \qquad x \in \mathbf{R}^d,$$

where b is a fixed portfolio vector. This regression function gives a prediction for the utility of the gross return of the portfolio. The prediction can be inaccurate; but the

collection of all predictions, for all values of the portfolio vector b, gives a way to choose the optimal portfolio vector. Namely, at time T we use the data

$$(Y_{b,t}, X_t), \qquad t = 1, \dots, T-1,$$

to estimate the regression function. Let us denote this estimate by

$$\hat{f}_{b,T} : \mathbf{R}^d \to \mathbf{R}.$$

We choose the optimal portfolio vector $\hat{b}_T$ at time T by

$$\hat{b}_T = \text{argmax}_{b \in B} \hat{f}_{b,T}(X_T), \tag{1.98}$$

where $B \subset B_N$, where B_N is the sphere in $\mathbf{R}^N$, defined in (1.94). Thus we choose the portfolio vector for which the prediction of the utility of the return of the portfolio is the highest. Since T is the current time, we use $\hat{b}_T$ to allocate the current wealth, and the portfolio vectors $\hat{b}_t$, $t = 1, \dots, T-1$, can be used to analyze the statistical properties of the portfolio selection method.

We can also describe the procedure by defining function $b : \mathbf{R}^d \to B$ by

$$b(x) = \text{argmax}_{b \in B} f_b(x).$$

This function is estimated at time T by

$$\hat{b}_T(x) = \text{argmax}_{b \in B} \hat{f}_{b,T}(x).$$

At time T we choose the portfolio vector $\hat{b}_T(X_T)$.

We can use the idea of (1.47) to transform the time series (1.91) to regression data and we can define the explanatory variables

$$X_t = (R_t, \dots, R_{t-k+1}) \in \mathbf{R}^{Nk}, \tag{1.99}$$

$t = k, \dots, T-1$. The explanatory variable X_t is defined as a vector of length k of past gross returns. This choice can be justified if the past returns contain all relevant information available to predict the future returns. Clearly it is possible that the quality of predictions can be improved if we make some transformation of the past returns. Possible transformations are discussed in Section 1.7. If the time series $R_1, \dots, R_T$ is stationary, then $(Y_{b,t}, X_t)$, $t = k, \dots, T-1$, are identically distributed.

An application of regression function estimation in portfolio selection has been made by Brandt (1999), Aït-Sahalia & Brandt (2001), and Györfi, Lugosi & Udina (2006). See also Györfi & Schäfer (2003), Györfi, Urbán & Vajda (2007), Györfi, Udina & Walk (2008), and Györfi, Ottucsák & Walk (2012).

Portfolio Selection and Classification We assume to have data (R_t, X_t), $t = 1, \dots, T$, where $R_t \in \mathbf{R}^N$ is the gross return vector defined in (1.91) and $X_t \in \mathbf{R}^d$ is the the vector of explanatory variables observed at time t.

Let $B = \{b_0, \ldots, b_{K-1}\} \subset \mathbf{R}^N$ be a finite class of portfolio vectors. Define the class labels Y_t by

$$Y_t = k \Leftrightarrow b_k = \text{argmax}_{b \in B} b' R_{t+1}, \tag{1.100}$$

where $k = 0, \ldots, K-1$. Now $b_k \in B$ is the portfolio vector chosen at time t that gave the best return at time $t+1$, among all the portfolio vectors in B.

We have now defined classification data (X_t, Y_t), $t = 1, \ldots, T-1$, which is used at time T to estimate the classification function. The estimated classification function $\hat{g}$ chooses one of the portfolio vectors in B. Thus we define the portfolio vector which is chosen at time T by

$$\hat{b}_T = \hat{g}(X_T).$$

With the classification approach we are not able to introduce a risk aversion parameter, as in the case of regression approach, where a utility transformed return was predicted. The portfolios obtained by classification correspond to using the risk aversion parameter $\gamma = 1$.

Andriyashin, Härdle & Timofeev (2008) use a classification based approach to portfolio selection. They make for each stock in DAX 30 a decision to either buy, sell, or stay neutral, and the final portfolio is an equally weighted portfolio of the individual decisions for each stock.

Mean-Variance Preferences Portfolio choice with mean-variance preferences was proposed by Markowitz (1952) and Markowitz (1959). This method provides an alternative to the use of the maximization of the expected utility. The optimal portfolio vector in the mean-variance sense maximizes the penalized expected return

$$E\,(b' R_{T+1}) - \frac{\gamma}{2} \text{Var}\,(b' R_{T+1}), \tag{1.101}$$

where $\gamma \geq 0$ is the coefficient of risk aversion and

$$R_{T+1} = \left(S^1_{T+1}/S^1_T, \ldots, S^N_{T+1}/S^N_T\right)$$

is the vector of the gross returns of the N portfolio components, see (1.91) and (1.92). The minimization is done over a space of portfolio vectors $B \subset B_N$, where B_N is the sphere in $\mathbf{R}^N$, defined in (1.94). We have

$$E\,(b' R_{T+1}) = b' E R_{T+1}, \qquad \text{Var}\,(b' R_{T+1}) = b' \text{Var}(R_{T+1}) b,$$

where $\text{Var}(R_{T+1})$ is the $N \times N$ covariance matrix of R_{T+1}. We have to estimate the vector of expected returns ER_{T+1} and the covariance matrix $\text{Var}(R_{T+1})$.

We shall consider in Section 3.12.3 an example of portfolio selection with two risky assets. Let us derive the optimal portfolio vector for that case. Let us denote the portfolio vector $b = (b^1, b^2) = (1-w, w)$, where $w \in \mathbf{R}$. That is, we put proportion $1-w$ to the first asset and the proportion w to the second asset. Now

$$b' R_{T+1} = (1-w) R^1_{T+1} + w R^2_{T+1}.$$

Let the expected returns of the stocks be $ER^1_{T+1} = \mu_1$, $ER^2_{T+1} = \mu_2$ and the variances of the returns $\text{Var}(R^1_{T+1}) = \sigma^2_1$, $\text{Var}(R^2_{T+1}) = \sigma^2_2$. Denote the covariance of the returns by $\text{Cov}(R^1_{T+1}, R^2_{T+1}) = \sigma_{12}$. We have

$$\begin{aligned} & E(b'R_{T+1}) - \frac{\gamma}{2}\text{Var}(b'R_{T+1}) \\ &= \mu_1 + w(\mu_2 - \mu_1) - \frac{\gamma}{2}\left[(1-w)^2\sigma^2_1 + w^2\sigma^2_2 + 2(1-w)w\sigma_{12}\right] \\ &= \mu_1 - \frac{\gamma}{2}\sigma^2_1 + w\left[\mu_2 - \mu_1 - \gamma(\sigma_{12} - \sigma^2_1)\right] - w^2\frac{\gamma}{2}(\sigma^2_1 + \sigma^2_2 - 2\sigma_{12}). \end{aligned}$$

Setting the derivative with respect to w to zero and solving for w gives

$$w = \frac{1}{\gamma}\frac{\mu_2 - \mu_1 - \gamma(\sigma_{12} - \sigma^2_1)}{\sigma^2_1 + \sigma^2_2 - 2\sigma_{12}}, \tag{1.102}$$

when $\gamma > 0$. For $\gamma = 0$, as much as possible is invested to the asset for which the expected return μ_i is larger.

1.5.4 Option Pricing and Hedging

We consider an European option written at time t_0 (today), whose expiration is at a future time T. The option has value H_T at the expiration time and this value is a function of the stock price S_T. For example, in the case of a call option $H_T = \max\{0, S_T - K\}$, where K is the strike price. We need to determine a fair price H_{t_0} for the option at the current time t_0.

The price can be determined as the initial wealth needed to finance a hedging of the option. Hedging is done through a self financing trading using the stock S_t and the bond B_t. We take the interest rate equal to zero so that we can take $B_t = 1$ for all t. We consider the discrete time model, where trading is done at the time points $t_0, t_0 + 1, \ldots, T - 1$. Let W_t be the wealth at time t used to buy stocks and bonds. Let ξ_t be the number of stocks bought at time $t - 1$, and kept until time t, where a rebalancing is made. Let a_t be the number of bonds bought at time $t - 1$ and kept until time t. Since the portfolio is self financing, the quantities ξ_t and a_t have to satisfy

$$W_{t-1} = a_t + \xi_t S_{t-1}.$$

The wealth at time t is then

$$W_t = a_t + \xi_t S_t,$$

which is again distributed among the stock and the bond by choosing a_{t+1} and ξ_{t+1}. Thus,

$$\begin{aligned} W_t &= a_t + \xi_t S_t \\ &= (a_t + \xi_t S_{t-1}) + \xi_t(S_t - S_{t-1}) \\ &= W_{t-1} + \xi_t(S_t - S_{t-1}). \end{aligned} \tag{1.103}$$

We get inductively[10]

$$W_T = W_{t_0} + \sum_{t=t_0}^{T-1} \xi_{t+1}(S_{t+1} - S_t).$$

We can use two slightly different heuristics to define the fair price.

1. We consider the fair price to be the initial wealth W_{t_0} that minimizes the minimal difference between the final wealth and the payout of the option. That is, we want to minimize
$$E(W_T - H_T)^2$$
over all initial wealths W_0 and over all hedging strategies.

2. The writer of the option receives the premium H_{t_0} at time t_0, hedges his position at time points $t = t_0, \ldots, T-1$ with initial wealth $W_{t_0} = 0$, and pays H_T at the expiration to the holder of the option. Thus the wealth of the option writer at the expiration time T is equal to
$$\tilde{W}_T = H_{t_0} + \sum_{t=t_0}^{T-1} \xi_{t+1}(S_{t+1} - S_t) - H_T.$$
We want to find H_{t_0} and $\xi_{t_0+1}, \ldots, \xi_T$ so that $\tilde{W}_T$ is as close to zero as possible and the corresponding value for H_{t_0} can be considered as a fair value of the option. That is, we want to minimize
$$E\tilde{W}_T^2,$$
where the mean squared error measures closeness to zero.

Both heuristics lead to the following definition of the fair price and the optimal hedging coefficient. Denote

$$Y = H_T; \qquad (X_1, \ldots, X_d) = (S_{t_0+1} - S_{t_0}, \ldots, S_T - S_{T-1}),$$

where $d = T - t_0$. We define the fair price and the optimal hedging coefficient at time t_0 as

$$(H_{t_0}, \xi_{t_0+1}) = \operatorname{argmin}_{a \geq 0, b_1 \in \mathbf{R}} \min_{b_2, \ldots, b_d \in \mathbf{R}} E\rho(a + b_1 X_1 + \cdots + b_d X_d - Y), \quad (1.104)$$

where $\rho(t) = t^2$ or ρ is some other loss function as in (1.36). We have obtained a problem of stochastic control as described in (1.51).

[10]When interest rate for one period is $r > 0$, so that $B_{t+1} = (1 + r)B_t$, we get the expression

$$W_T = (1 + r)^{T - t_0} \left(W_{t_0} + \sum_{t=t_0}^{T-1} \xi_{t+1}(Z_{t+1} - Z_t) \right),$$

where $Z_t = (1 + r)^{t_0 - t} S_t$.

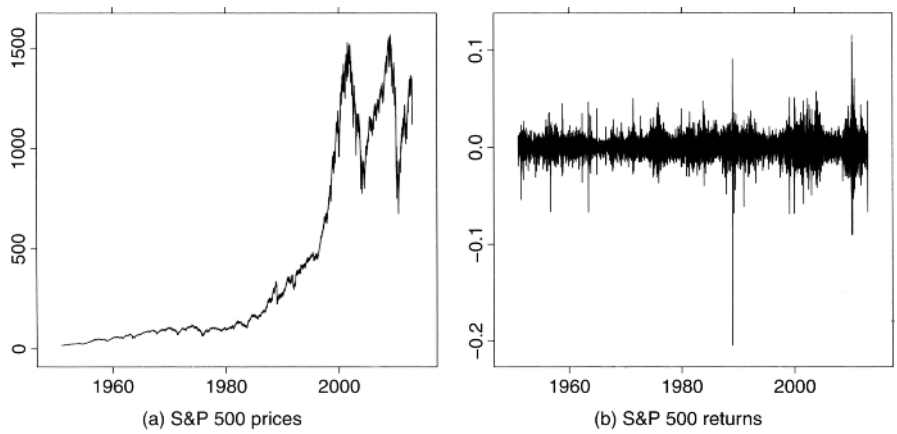

Figure 1.3 *S&P 500 index.* (a) The prices of S&P 500. (b) The net returns of S&P 500.

1.6 DATA EXAMPLES

We use two data sets as the main examples to illustrate the methods of regression and classification. The first data set is a time series of S&P 500 returns, described in Section 1.6.1. The second data set is a vector time series of S&P 500 and Nasdaq-100 returns, described in Section 1.6.2.

We use also other data sets as examples. In Section 2.1.7 a vector time series of DAX 30 and Daimler returns is used to illustrate an application of linear regression to the calculation of the beta of an asset. In Section 2.2.2 a time series of a hedge fund index returns is used to illustrate an application of varying coefficient regression in hedge fund replication. In Section 6.2 density estimation is illustrated with monthly S&P 500 data and U.S. Treasury 10-year bond data. In Section 6.3.2 a time series of DAX 30 returns is used to illustrate multidimensional scaling.

1.6.1 Time Series of S&P 500 Returns

The S&P 500 index data consist of the daily closing prices of the S&P 500 index during the period from 1950-01-03 until 2013-04-25, which makes 15 930 observations. The data are provided by Yahoo, where the index symbol is ^GSPC.

Figure 1.3 shows the prices and the net returns of the S&P 500 index. The net return is defined as

$$Y_t = \frac{P_t - P_{t-1}}{P_{t-1}},$$

where P_t is the price of the index at the end of day t.

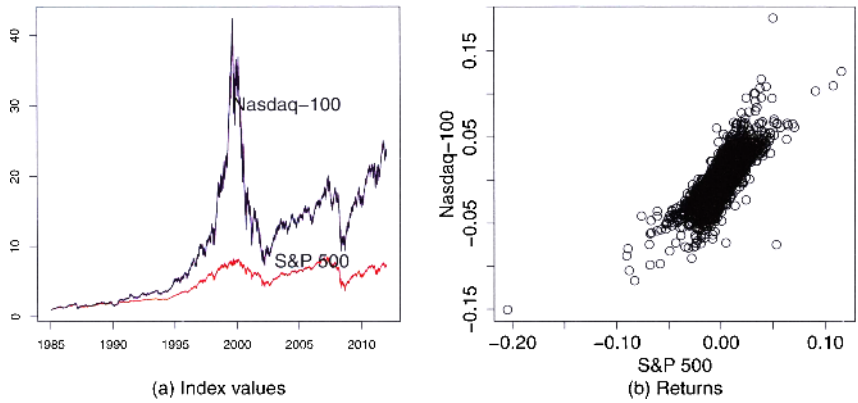

Figure 1.4 *S&P 500 and Nasdaq-100 indexes.* (a) The normalized values of the S&P 500 and Nasdaq-100 indexes. (b) The scatter plot of the net returns.

1.6.2 Vector Time Series of S&P 500 and Nasdaq-100 Returns

The S&P 500 and Nasdaq-100 index data consist of the daily closing prices of the S&P 500 index and the Nasdaq-100 index starting at 1985-10-01 and ending at 2013-03-19, which makes 6925 days of observations. The data are provided by Yahoo, where the index symbols are ^GSPC and ^NDX.

Figure 1.4 shows the S&P 500 and Nasdaq-100 indexes over the observation period. Panel (a) shows the time series of normalized index values. The index values are normalized so that they both have the value one at 1985-10-01. Panel (b) shows the scatter plot of the net returns of the indexes.

1.7 DATA TRANSFORMATIONS

In regression function estimation it is often useful to transform the variables before estimating the regression function. A transformation of the explanatory variables is important when the regression function is estimated with a method of local averaging, defined in Chapter 3. If the local neighborhood of a local averaging estimator is spherically symmetric, as is the case when we use kernel estimation with a spherically symmetric kernel function and with a single smoothing parameter for each variable, then the scales of the explanatory variables should be compatible. For example, if one variable takes values in $[0, 1]$ and an other variable takes values in $[0, 100]$, then the variable with the shorter range would effectively be canceled out when using spherically symmetric neighborhoods.

First, we define data sphering, which is a transformation of the explanatory variables that makes the variances of the explanatory variables equal and the covariance matrix of the explanatory variables diagonal. Second, we define a copula transforma-

tion that makes the marginal distributions of the explanatory variables approximately standard Gaussian, or uniform on $[0, 1]$, but keeps the copula of the explanatory variables unchanged. Third, we define transformations of the response variable.

1.7.1 Data Sphering

We can make the scales of variables compatible by normalizing observations so that the sample variances of the variables are equal to one. Let $X_i = (X_{i1}, \ldots, X_{id})$, $i = 1, \ldots, n$, be the original observations. The transformed observations are

$$Z_i = \left(\frac{X_{i1}}{s_1}, \ldots, \frac{X_{id}}{s_d} \right), \qquad i = 1, \ldots, n,$$

where the sample variances are

$$s_k^2 = \frac{1}{n} \sum_{i=1}^{n} \left(X_{ik} - \bar{X}_k \right)^2, \qquad k = 1, \ldots, d,$$

with the arithmetic mean $\bar{X}_k = n^{-1} \sum_{i=1}^{n} X_{ik}$. We can also make the ranges of the variables equal by defining the transformed observations as $Z_i = (Z_{i1}, \ldots, Z_{id})$, $i = 1, \ldots, n$, where

$$Z_{ik} = \frac{X_{ik} - \min_{i=1,\ldots,n} X_{ik}}{\max_{i=1,\ldots,n} X_{ik} - \min_{i=1,\ldots,n} X_{ik}}, \qquad k = 1, \ldots, d.$$

Data sphering is a more extensive transformation than just standardizing the sample variances equal to one; we make such linear transformation of data that the covariance matrix becomes the identity matrix. The sphering is almost the same as the principal component transformation. In the principal component transformation the covariance matrix is diagonalized but it is not made the identity matrix.

1. Sphering of a random vector $X \in \mathbf{R}^d$ means that we make a linear transform of X so that the new random variable has expectation zero and the identity covariance matrix. Let

   $$\Sigma = E\left[(X - EX)(X - EX)' \right]$$

   be the covariance matrix and make the spectral representation of Σ:

   $$\Sigma = A \Lambda A',$$

   where A is orthogonal and Λ is diagonal. Then

   $$Z = \Lambda^{-1/2} A' (X - EX)$$

   is the sphered random vector, having the property[11]

   $$\mathrm{Cov}(Z) = I_d.$$

[11]The orthogonality of A means that $A'A = AA' = I_d$. Thus $A'\Sigma A = \Lambda$ and $\mathrm{Cov}(Z) = \Lambda^{-1/2} A' \mathrm{Cov}(X) A \Lambda^{-1/2} = \Lambda^{-1/2} A' \Sigma A \Lambda^{-1/2} = I_d$.

2. Data sphering means that the data are transformed so that the arithmetic mean of the observations is zero and the empirical covariance matrix is the unit matrix. Let Σ_n be the empirical covariance matrix,

$$\Sigma_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})',$$

where $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$ is the $d \times 1$ column vector of arithmetic means. We find the spectral representation of Σ_n,

$$\Sigma_n = A_n \Lambda_n A_n',$$

where A_n is orthogonal and Λ_n is diagonal. Define the transformed observations

$$Z_i = \Lambda_n^{-1/2} A_n'(X_i - \bar{X}), \qquad i = 1, \ldots, n.$$

The sphered data matrix is the $n \times d$ matrix $\mathbb{Z}_n$ defined by

$$\mathbb{Z}_n' = \Lambda_n^{-1/2} A_n' \left(\mathbb{X}_n' - \bar{X}_n 1_{1\times n}\right),$$

where $\mathbb{X}_n = (X_1, \ldots, X_n)'$ is the original $n \times d$ data matrix, and $1_{1\times n}$ is the $1 \times n$ row vector of ones.

1.7.2 Copula Transformation

Copula modeling was explained in Section 1.3.3. Copula modeling leads also to useful data transformations. A copula transformation changes the marginal distributions but keeps the copula (the joint distribution) the same.

1. The copula transformation of random vector $X = (X_1, \ldots, X_d)$, when X has a continuous distribution, gives random variable $Z = (Z_1, \ldots, Z_d)$ whose marginals have the uniform distribution on $[0, 1]$, or some other suitable distribution. Let $F_{X_k}(t) = P(X_k \leq t)$, $k = 1, \ldots, d$, be the distribution functions of the components of X. Now

$$Z = (F_{X_1}(X_1), \ldots, F_{X_d}(X_d))$$

is a random vector whose marginal distributions are uniform on $[0, 1]$.[12] The distribution function of this random vector is called the copula of the distribution of $X = (X_1, \ldots X_d)$. Often the copula with uniform marginals is inconvenient due to boundary effects. We may get statistically more tractable distribution by defining

$$Z = \left(\Phi^{-1}(F_{X_1}(X^1)), \ldots, \Phi^{-1}(F_{X_d}(X_d))\right),$$

[12]Random variable $F_{X_k}(X_k)$ has the uniform distribution on $[0, 1]$, because $P(F_{X_k}(X_k) \leq t) = P(X_k \leq F_{X_k}^{-1}(t)) = F_{X_k}(F_{X_k}^{-1}(t)) = t$.

where Φ is the distribution function of the standard Gaussian distribution. The components of Z have the standard Gaussian distribution.[13]

2. The copula transformation of data $X_1, \ldots, X_n$ means that the data are transformed so that the marginal distributions are approximately uniform, or have approximately some other suitable distribution. Let the rank of observation $X_{ik}, i = 1, \ldots, n, k = 1, \ldots, d$, be

$$\text{rank}(X_{ik}) = \#\{X_{jk} : X_{jk} \leq X_{ik}, j = 1, \ldots, n\}.$$

We normalize the ranks to get observations with approximately uniform distribution on $[0, 1]$:

$$Z_i = \left(\frac{\text{rank}(X_{i1})}{n+1}, \ldots, \frac{\text{rank}(X_{id})}{n+1} \right),$$

for $i = 1, \ldots, n$. Often the standard Gaussian distribution is more convenient and we define

$$Z_i = \left(\Phi^{-1}\left(\frac{\text{rank}(X_{i1})}{n+1} \right), \ldots, \Phi^{-1}\left(\frac{\text{rank}(X_{id})}{n+1} \right) \right), \tag{1.105}$$

for $i = 1, \ldots, n$.

Figure 1.5 shows scatter plots of S&P 500 and Nasdaq-100 copula transformed net returns. The data is described in Section 1.6.2. Panel (a) shows the case where the marginals are transformed to be approximately standard Gaussian. Panel (b) shows the case where the marginals are transformed to be approximately uniformly distributed in $[0, 1]$. We have used in scatter plots histogram smoothing with 70^2 bins, as explained in Section 6.1.1. Uniform marginals make the data concentrate on the lower left and on the upper right corners, which can make the estimation difficult due to the boundary effects. The Gaussian marginals make the distribution of the data have tails which decrease smoothly to zero.

1.7.3 Transformations of the Response Variable

The transformation of the response variable can be used to obtain a more normal distribution or to remove heteroskedasticity by stabilizing variance. See Efron (1982).

The power transformations are called the Box–Cox transformations and defined for $\lambda \in \mathbf{R}$ by

$$Z_i^{(\lambda)} = \begin{cases} \frac{Y_i^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \log Y_i, & \lambda = 0, \end{cases}$$

where we assume that $Y_i \geq 0$. Box–Cox transformations were defined in Box & Cox (1962). Tukey (1957) considered the power transformation Y_i^λ for $\lambda \neq 0$.

[13]Random variable $\Phi^{-1}(U)$, where U has the uniform distribution on $[0, 1]$, has the standard Gaussian distribution because $P(\Phi^{-1}(U) \leq t) = P(U \leq \Phi(t)) = \Phi(t)$.

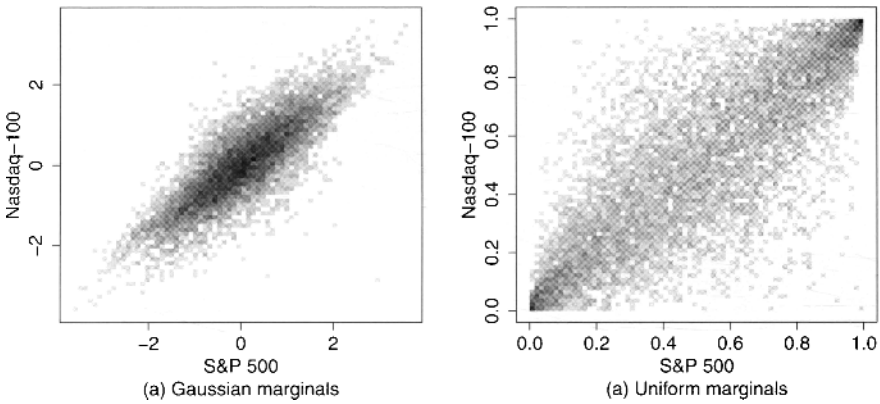

Figure 1.5 *Copula transform.* Scatter plots of S&P 500 and Nasdaq-100 returns are shown. (a) Gaussian marginals. (b) Uniform marginals.

The natural exponential family was defined in (1.68). In the natural exponential family

$$E_v(Y) = \mu(v) = d'(v), \qquad \text{Var}_v(Y) = V(v) = d''(v).$$

A subclass of natural exponential families consists of the families with a quadratic variation function. Now we have

$$\text{Var}_v(Y) = V(v) = a_0 + a_1\mu(v) + a_2\mu(v)^2,$$

where $\mu(v) = d'(v)$. The examples are normal, gamma, NEF-GHS (the natural exponential family generated by the generalized hyperbolic secant distribution), binomial, negative binomial, and Poisson. Denote $\text{Var}_v(Y) = V(\mu(v))$. Define a function $G : \mathbf{R} \to \mathbf{R}$ to be such that

$$G'(\mu) = V^{-1/2}(\mu).$$

By the central limit theorem, we obtain

$$n^{1/2}\left(\bar{Y} - \mu(v)\right) \xrightarrow{d} N(0, V(\mu(v))),$$

as $n \to \infty$, where $\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i$, and $Y_1, \ldots, Y_n$ are assumed i.i.d. By the delta method, we have

$$n^{1/2}\big(G(\bar{Y}) - G(\mu(v))\big) \xrightarrow{d} N(0, 1),$$

as $n \to \infty$. Thus we call the transformation G a variance stabilizing transform.

1.8 CENTRAL LIMIT THEOREMS

A central limit theorem is needed to test the difference between two prediction methods; see Section 1.9.1. A central limit theorem is also needed to derive asymptotic distributions for estimators; see Section 2.1.4.

1.8.1 Independent Observations

Let $Y_1, Y_2, \ldots$ be a sequence of real-valued i.i.d. random variables with $\text{Var}(Y_i) = \sigma^2$, where $0 < \sigma^2 < \infty$. According to the central limit theorem, we have

$$n^{-1/2} \sum_{i=1}^{n} (Y_i - EY_i) \xrightarrow{d} N\left(0, \sigma^2\right),$$

as $n \to \infty$. Let $X_1, X_2, \ldots$ be an i.i.d. sequence of random vectors with $\text{Cov}(X_i) = \Sigma$, where the diagonal elements of Σ are finite and positive. According to the central limit theorem, we have

$$n^{-1/2} \sum_{i=1}^{n} (X_i - EX_i) \xrightarrow{d} N\left(0, \Sigma\right),$$

as $n \to \infty$.

1.8.2 Dependent Observations

We need a central limit theorem for dependent observations. Let $(Y_t)_{t \in \mathbf{Z}}$ be a strictly stationary time series. We define the weak dependence in terms of a condition on the α-mixing coefficients. Let $\mathcal{F}_i^j$ denote the sigma algebra generated by random variables $Y_i, \ldots, Y_j$. The α-mixing coefficient is defined as

$$\alpha_n = \sup_{A \in \mathcal{F}_{-\infty}^{0}, B \in \mathcal{F}_n^{\infty}} |P(A \cap B) - P(A)P(B)|,$$

where $n = 1, 2, \ldots$. Now we can state the central limit theorem. Let $E|Y_t|^\delta < \infty$ and $\sum_{j=1}^{\infty} \alpha_j^{1-2/\delta} < \infty$ for some constant $\delta > 2$. Then,

$$n^{-1/2} \sum_{i=1}^{n} (Y_i - EY_i) \xrightarrow{d} N(0, \sigma^2), \tag{1.106}$$

where

$$\sigma^2 = \sum_{j=-\infty}^{\infty} \gamma(j) = \gamma(0) + 2 \sum_{j=1}^{\infty} \gamma(j),$$

$\gamma(j) = \text{Cov}(X_t, X_{t+j})$, and we assume that $\sigma^2 > 0$.

Ibragimov & Linnik (1971, Theorem 18.4.1) gave necessary and sufficient conditions for a central limit theorem under α-mixing conditions A proof for our statement

of the central limit theorem in (1.106) can be found in Peligrad (1986); see also Fan & Yao (2005, Theorem 2.21) and Billingsley (2005, Theorem 27.4)

Let us state the central limit theorem for the vector time series $(X_t)_{t\in\mathbf{Z}}$, where $X_t \in \mathbf{R}^d$. If the time series $(a'X_t)_{t\in\mathbf{Z}}$ satisfies the conditions for the univariate central limit theorem for all $a \in \mathbf{R}^d$, then[14]

$$n^{-1/2}\sum_{i=1}^{n}(X_i - EX_i) \xrightarrow{d} N(0,\Sigma), \tag{1.107}$$

where

$$\Sigma = \sum_{j=-\infty}^{\infty}\Gamma(j) = \Gamma(0) + \sum_{j=1}^{\infty}(\Gamma(j) + \Gamma(j)'),$$

and the autocovariance matrix $\Gamma(j)$ was defined in (1.21) as

$$\Gamma(j) = \mathrm{Cov}(X_t, X_{t+j}).$$

Note that we used the property (1.22) $\Gamma(j) = \Gamma(-j)'$.

Let us explain the expression for the asymptotic variance σ^2 in the univariate central limit theorem (1.106). Let us assume that $EY_i = 0$. The variance of the normalized sum is

$$\mathrm{Var}\left(n^{-1/2}\sum_{i=1}^{n}Y_i\right) = n^{-1}\sum_{i=1}^{n}\mathrm{Var}(Y_i) + n^{-1}\sum_{i\neq j}\mathrm{Cov}(Y_i, Y_j).$$

Thus, for an i.i.d. time series we have that

$$\mathrm{Var}\left(n^{-1/2}\sum_{i=1}^{n}Y_i\right) = \mathrm{Var}(Y_1) = \gamma(0).$$

For a weakly stationary time series we have

$$\begin{aligned}\mathrm{Var}\left(n^{-1/2}\sum_{i=1}^{n}Y_i\right) &= n^{-1}\sum_{i=1}^{n}\mathrm{Var}(Y_i) + 2n^{-1}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\mathrm{Cov}(Y_i, Y_j)\\ &= \gamma(0) + 2n^{-1}\sum_{i=1}^{n-1}(n-i)\gamma(i)\\ &= \sum_{i=-(n-1)}^{n-1}\left(1 - \frac{|i|}{n}\right)\gamma(i).\end{aligned}$$

Thus, in order that $\mathrm{Var}\left(n^{-1/2}\sum_{i=1}^{n}Y_i\right) \to c$, for a finite positive constant c, we need that $\gamma(n) \to 0$ sufficiently fast, as $n \to \infty$. A sufficient condition is that $\sum_{j=1}^{\infty}|\gamma(j)| < \infty$.

[14]Cramér–Wold theorem states that $Y_n \xrightarrow{d} Y$ if and only if $a'Y_n \xrightarrow{d} a'Y$ for all $a \in \mathbf{R}^d$, as $n \to \infty$, where Y_n and Y are random vectors.

1.8.3 Estimation of the Asymptotic Variance

In the applications we have to estimate the asymptotic variance and the asymptotic covariance matrix. For i.i.d. data we can use the sample variance and the sample covariance matrix. For dependent data the estimation is more complicated. Let us discuss the estimation of the variance σ^2 in (1.106) using the observations $Y_1, \ldots, Y_n$, and the estimation of the covariance matrix Σ in (1.107) using the observations $X_1, \ldots, X_n$.

Let us start with the estimation of σ^2 in (1.106). An application of the sample covariances would lead to the estimator

$$\tilde{\sigma}^2 = \hat{\gamma}(0) + 2\sum_{j=1}^{n-1} \hat{\gamma}(j),$$

where

$$\hat{\gamma}(j) = \frac{1}{n}\sum_{i=1}^{n-j}(Y_i - \bar{Y})(Y_{i+j} - \bar{Y}),$$

for $j = 0, \ldots, n-1$. Note that for large j only few observations are used in the estimator $\hat{\gamma}(j)$. For example, when $j = n-1$ the estimator uses only one observation: $\hat{\gamma}(n-1) = Y_1 Y_n / n$, which is a very imprecise estimator. We can use weighting to remove the imprecise estimators and define

$$\hat{\sigma}^2 = \hat{\gamma}(0) + 2\sum_{j=1}^{n-1} w(j)\hat{\gamma}(j), \tag{1.108}$$

where

$$w(j) = \left(1 - \frac{j}{h}\right)_+,$$

where $1 \leq h \leq n-1$ is a chosen smoothing parameter. We can generalize the estimator to other weights and define

$$w(j) = K(j/h), \tag{1.109}$$

where $K : \mathbf{R} \to \mathbf{R}$ is a kernel function satisfying $K(x) = K(-x)$, $K(0) = 1$, $|K(x)| \leq 1$ for all x, and $K(x) = 0$ for $|x| > 1$.

To estimate Σ in (1.107) we use

$$\hat{\Sigma} = \hat{\Gamma}(0) + \sum_{j=1}^{n-1} w(j)\left(\hat{\Gamma}(j) + \hat{\Gamma}(j)'\right), \tag{1.110}$$

where

$$\hat{\Gamma}(j) = \frac{1}{n}\sum_{i=1}^{n-j}\left(X_i - \bar{X}\right)\left(X_{i+j} - \bar{X}\right)',$$

for $j = 0, \ldots, n-1$. We will apply weights in an estimator of an asymptotic covariance matrix in (2.44).

The weighting we have used is related to the smoothing in the estimation of the spectral density. The unnormalized spectral density function of a weakly stationary time series, having autocorrelation coefficients $\gamma(k)$ with $\sum_{j=-\infty}^{\infty} |\gamma(j)| < \infty$, is defined by

$$g(\omega) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \gamma(j)\, e^{-ij\omega},$$

where $\omega \in [-\pi, \pi]$; see Brockwell & Davis (1991, Section 4.3). The lag window spectral density estimator, based on data $Y_1, \ldots, Y_n$, is defined by

$$\hat{g}(\omega) = \frac{1}{2\pi} \sum_{|j| \le h} K(j/h)\, \hat{\gamma}(j)\, e^{-ij\omega},$$

where $\hat{\gamma}(j)$ are the sample autocorrelation coefficients, $h = 1, 2, \ldots, n-1$, and K is similar as in (1.109); see Brockwell & Davis (1991, Section 10.4). Now we have

$$\hat{g}(0) = \frac{1}{2\pi} \sum_{|j| \le h} K(j/h)\, \hat{\gamma}(j) = \frac{1}{2\pi}\, \hat{\sigma}^2,$$

where $\hat{\sigma}^2$ is defined in (1.108) with the weights defined in (1.109).

1.9 MEASURING THE PERFORMANCE OF ESTIMATORS

We discuss measuring the performance of regression function estimators, conditional variance, covariance, and quantile estimators, estimators of the expected shortfall, and classifiers.

1.9.1 Performance of Regression Function Estimators

We denote by $\hat{f}(x)$ an estimator of the conditional expectation $f(x) = E(Y \mid X = x)$. We define theoretical performance measures, which are used to compare estimators of f under given theoretical assumptions. After that we define empirical performance measures, which try to estimate the performance of estimate $\hat{f}$ using the available data.

Theoretical Performance Measures Theoretical performance measures can be divided into global risk functionals, like the mean integrated squared error, and into pointwise risk functionals, like the mean squared error.

Global Error We can use the mean integrated squared error (MISE) or the mean averaged squared error to measure the goodness of regression function estimators $\hat{f}$ globally, when we want to recover the complete curve and not its value at a single point $x \in \mathbf{R}^d$.

The prediction error of regression function f can be measured by

$$E(f(X) - Y)^2.$$

This measure of prediction is natural since $f(x) = E(Y \mid X = x)$ and the conditional expectation minimizes the mean squared error, as shown in (1.37). When we have an estimator $\hat{f}$ of f, then we can measure the prediction error of the estimator by

$$E\Big(\hat{f}(X) - Y\Big)^2.$$

Now the expectation is with respect to the distribution of

$$(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n),$$

because $\hat{f}$ is a random function depending on the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. We have that

$$\begin{aligned} & E\left[\Big(\hat{f}(X) - Y\Big)^2 \,\middle|\, (Y_1, X_1), \ldots, (Y_n, X_n)\right] \\ &= \int_{\mathbf{R}^d} \Big(\hat{f}(x) - f(x)\Big)^2 f_X(x)\, dx + E(f(X) - Y)^2, \end{aligned} \tag{1.111}$$

where f_X is the density function of X. The minimization of expression (1.111) with respect to estimator $\hat{f}$ is equivalent to the minimization of the expression

$$\int_{\mathbf{R}^d} \Big(\hat{f}(x) - f(x)\Big)^2 f_X(x)\, dx.$$

This calculation can be used to justify the mean integrated error, defined in (1.112).

The Mean Integrated Squared Error The mean integrated squared error is defined as

$$\begin{aligned} \text{MISE}(\hat{f}, f) &= E\left(\hat{f}(X) - f(X)\right)^2 \\ &= EE\left[\Big(\hat{f}(X) - f(X)\Big)^2 \,\middle|\, (Y_1, X_1), \ldots, (Y_n, X_n)\right] \\ &= E\int_{\mathbf{R}^d} \Big(\hat{f}(x) - f(x)\Big)^2 f_X(x)\, dx, \end{aligned} \tag{1.112}$$

where X is independent of $(Y_1, Y_1), \ldots, (Y_n, X_n)$ and f_X is the density function of X. Using the short hand notation we write the mean integrated error as

$$\text{MISE}(\hat{f}, f) = E\left\|\hat{f} - f\right\|_{2,X}^2, \tag{1.113}$$

where $\|f\|_{2,X}^2 = \int_{\mathbf{R}^d} f(x)^2\, dP_X(x)$, and P_X is the probability distribution of random vector X. We can generalize (1.113) to

$$E\int_{\mathbf{R}^d} \Big(\hat{f}(x) - f(x)\Big)^2 w(x)\, dP_X(x),$$

where $w : \mathbf{R}^d \to \mathbf{R}$ is a weight function The weight function could be $w \equiv 1$, to get (1.113). We can choose $w(x) = 1/f_X(x)$, to get the L_2 error with respect to the Lebesgue measure. The weight function $w(x)$ could also be used to trim away boundary effects.

The Mean Averaged Squared Error The mean averaged squared error is defined as

$$\text{MASE}(\hat{f}, f) = E\left[\frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}(X_i) - f(X_i)\right)^2 \,\middle|\, X_1, \ldots, X_n\right]. \tag{1.114}$$

Using the short hand notation we write the mean averaged squared error as

$$\text{MASE}(\hat{f}, f) = E_{X^{(n)}} \left\|\hat{f} - f\right\|^2_{2,X^{(n)}},$$

where

$$\|f\|^2_{2,X^{(n)}} = \int_{\mathbf{R}^d} f(x)^2 \, dP_X^{(n)}(x) = \frac{1}{n}\sum_{i=1}^{n} f(X_i)^2,$$

$P_{X^{(n)}}$ is the empirical probability distribution of the sample $(X_1, \ldots, X_n)$, and $E_{X^{(n)}}$ is the conditional expectation under the condition $(X_1, \ldots, X_n)$. We can generalize the mean averaged squared error by defining $\|f\|^2_{2,X^{(n)}} = n^{-1}\sum_{i=1}^{n} f(X_i)^2 w(X_i)$, where $w : \mathbf{R}^d \to \mathbf{R}$ is a weight function.

Pointwise Error Pointwise performance measures quantify how well the value of f is recovered at a single point $x \in \mathbf{R}^d$. We can use mean squared error (MSE) either unconditionally or conditionally.

- The unconditional mean squared error at point $x \in \mathbf{R}^d$ is defined as

$$\text{MSE}(\hat{f}(x), f(x)) = E\left(\hat{f}(x) - f(x)\right)^2,$$

  where f is the true regression function.

- The conditional mean squared error at point $x \in \mathbf{R}^d$ is defined as

$$\text{MSE}(\hat{f}(x), f(x)) = E\left[\left(\hat{f}(x) - f(x)\right)^2 \,\middle|\, X_1, \ldots, X_n\right],$$

  where f is the true regression function.

The Use of Theoretical Performance Measures Theoretical performance measures can be used to compare estimators in a given model. A model is a collection of probability distributions for the distribution of (X, Y) and on the distribution of the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. We can describe a model also as a collection of regression functions $\mathcal{F}$ together with the additional assumptions on the distribution

of (X, Y) and on the distribution of the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. To compare estimators, we use the supremum risk

$$\sup_{f \in \mathcal{F}} \text{MISE}(\hat{f}, f).$$

We use the supremum risk, because it is necessary to require that an estimator performs uniformly well over a model, because for a single regression function f it is trivial to define the best estimator; this is the regression function f itself: $\hat{f} = f$.

Empirical Performance Measures Empirical performance measures can be used to estimate the performance of an estimator and to compare estimators. Empirical performance measures are calculated using the available regression data $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Empirical Performance Measures for Cross-Sectional Data The mean integrated squared error

$$\text{MISE}(\hat{f}, f) = E\left(\hat{f}(X) - f(X)\right)^2,$$

defined in (1.113), cannot be approximated by $n^{-1}\sum_{i=1}^{n}(\hat{f}(X_i) - Y_i)^2$. This approximation fails, because we are using the same data to construct the estimator and to estimate the prediction error. Using the same learning data and the test data leads to overly optimistic evaluation of the performance. However, we can avoid the problem using sample splitting or cross-validation.

1. *Sample Splitting* Let $\hat{f}^*$ be the regression function estimator constructed from the data $(X_1, Y_1), \ldots, (X_{n^*}, Y_{n^*})$, where $1 \leq n^* < n$, and typically $n^* = [n/2]$. Then we use

$$\text{MISE}_n(\hat{f}) = \frac{1}{n - n^*} \sum_{i=n^*+1}^{n} \left(\hat{f}^*(X_i) - Y_i\right)^2 \tag{1.115}$$

   to estimate the mean integrated squared error.

2. *Cross Validation* Let $\hat{f}_{-i}$ be a regression function estimator constructed from the other data points but not (X_i, Y_i). Then we use

$$\text{MISE}_n(\hat{f}) = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{f}_{-i}(X_i) - Y_i\right)^2 \tag{1.116}$$

   to estimate the mean integrated squared error.

Cross validation is discussed in Section 3.2.7 in the case of kernel estimation.

Empirical Performance Measures in the Time Series Setting In the time series setting we have observations $(X_1, Y_1), \ldots, (X_T, Y_T)$ that are observed at consecutive time instants. We can construct regression function estimator $\hat{f}_t$ using data $(X_1, Y_1), \ldots, (X_t, Y_t)$ that is observed until time t, and define the mean of squared prediction errors by

$$\mathrm{MSPE}_T(\hat{f}) = \frac{1}{T-1} \sum_{t=1}^{T-1} \left(\hat{f}_t(X_t) - Y_{t+1} \right)^2, \tag{1.117}$$

which is analogous to the estimate of the mean integrated squared defined in (1.116). We will use later in Section 3.12.1 the mean of absolute prediction errors

$$\mathrm{MAPE}_T(\hat{f}) = \frac{1}{T-1} \sum_{t=1}^{T-1} \left| \hat{f}_t(X_t) - Y_{t+1} \right|. \tag{1.118}$$

Diebold & Mariano (1995) proposed a test for testing the equality of forecast accuracy. Let us have two predictors $\hat{f}_t(X_{t+1})$ and $\hat{g}_t(X_{t+1})$ and the corresponding losses

$$F_t = \left(\hat{f}_t(X_{t+1}) - Y_{t+1} \right)^2, \qquad G_t = (\hat{g}_t(X_{t+1}) - Y_{t+1})^2.$$

The losses do not have to be squared prediction errors, but we can also use absolute prediction errors, for example. We get the time series of loss differentials

$$d_t = F_t - G_t.$$

The null hypothesis and the alternative hypothesis are

$$H_0 : Ed_t = 0, \qquad H_1 : Ed_t \neq 0.$$

We apply the central limit theorem as stated in (1.106). Under the null hypothesis and under the assumptions of the central limit theorem, we have

$$(T - t_0 + 1)^{-1/2} \sum_{t=t_0}^{T} d_t \xrightarrow{d} N(0, \sigma^2),$$

as $T \to \infty$, where

$$\sigma^2 = \sum_{k=-\infty}^{\infty} \gamma(k), \qquad \gamma(k) = Ed_0 d_k.$$

We can use the estimate

$$\hat{\sigma}^2 = \sum_{k=-(T-1)}^{T-1} w(k) \hat{\gamma}(k),$$

where $w(k)$ is defined in (1.109). Let us choose the test statistics

$$D = \hat{\sigma}^{-1}(T - t_0 + 1)^{-1/2} \sum_{t=t_0}^{T} d_t.$$

When we observe $|D| = d_{obs}$, then the p-value is calculated by $P(|D| > d_{obs}) \approx 2(1 - \Phi(d_{obs}))$, where Φ is the distribution function of the standard normal distribution.

1.9.2 Performance of Conditional Variance Estimators

Theoretical Performance Measures Theoretical performance measures can be generalized from the case of regression function estimation to the case of conditional variance estimators. For example, when $f(x) = \mathrm{Var}(Y \mid X = x)$ and $\hat{f}(x)$ is an estimator of $f(x)$, then we can measure the performance of $\hat{f}$ by

$$E \int_{\mathbf{R}^d} \left(\hat{f}(x) - f(x)\right)^2 w(x)\, dP_X(x), \tag{1.119}$$

where $w : \mathbf{R}^d \to \mathbf{R}$ is a weight function.

Empirical Performance Measures We define the empirical performance measures first for cross-sectional data and then for time series data.

Cross-Sectional Data Empirical performance measures of conditional variance estimators can be found naturally in the case where

$$E(Y \mid X = x) = 0,$$

so that

$$f(x) = \mathrm{Var}(Y \mid X = x) = E(Y^2 \mid X = x).$$

For example, we can use sample splitting. Let $\hat{f}^*$ be an estimator of f, constructed from the data $(X_1, Y_1), \ldots, (X_{n^*}, Y_{n^*})$, where $1 \le n^* < n$. Then we can use

$$\frac{1}{n - n^*} \sum_{i=n^*+1}^{n} \left|\hat{f}^*(X_i) - Y_i^2\right| \tag{1.120}$$

to measure the performance of the estimator.

Time Series Data We use slightly different notation in the case of state space smoothing and in the case of time space smoothing.

State–Space Smoothing When we have identically distributed time series observations $(X_1, Y_1), \ldots, (X_T, Y_T)$, then we can construct an estimator $\hat{f}_t$ of the conditional variance using data $(X_1, Y_1), \ldots, (X_t, Y_t)$ and calculate the mean of absolute

prediction errors

$$\text{MAPE}_T(\hat{f}) = \frac{1}{T-t_0}\sum_{t=t_0}^{T-1}\left|\hat{f}_t(X_{t+1}) - Y_{t+1}^2\right|, \tag{1.121}$$

where t_0 is the initial estimation period, $1 \leq t_0 \leq T-1$. We start to evaluate the performance of the estimator after t_0 observations are available, because any estimator can behave erratically when only few observations are available. Mean absolute prediction error is sometimes called the mean absolute deviation error (MADE).

Time–Space Smoothing In autoregressive time–space smoothing methods, like in the GARCH models studied in Section 3.9.2, the explanatory variables are the previous observations, and the estimate $\hat{\sigma}_t^2$ of $E(Y_t^2 \mid \mathcal{F}_{t-1})$ is calculated using observations $Y_1, \ldots, Y_{t-1}$. Now we have

$$\text{MAPE}_T\left(\hat{\sigma}^2\right) = \frac{1}{T-t_0+1}\sum_{t=t_0}^{T}\left|\hat{\sigma}_t^2 - Y_t^2\right|. \tag{1.122}$$

Spokoiny (2000) proposes to take the square roots and use the mean square root prediction error criterion as the performance measure:

$$\text{MSqPE}_T\left(\hat{\sigma}^2\right) = \frac{1}{T-t_0+1}\sum_{t=t_0}^{T}\left|\hat{\sigma}_t^2 - Y_t^2\right|^{1/2}. \tag{1.123}$$

The mean square root prediction error is such that outliers do not have a strong influence on the results. Fan & Gu (2003) propose to measure the performance with the mean absolute deviation error:

$$\text{MADE}_T\left(\hat{\sigma}^2\right) = \frac{1}{T-t_0+1}\sum_{t=t_0}^{T}\left|\sqrt{\frac{2}{\pi}}\,\hat{\sigma}_t - |Y_t|\right|, \tag{1.124}$$

where the factor $\sqrt{2/\pi}$ comes from the fact that for a standard normal random variable $Z \sim N(0,1)$, we have $E|Z| = \sqrt{2/\pi}$.

We can generalize the performance measures (1.122)–(1.124) and define a class of performance measures by

$$\text{MDE}_T^{(p,q)}\left(\hat{\sigma}^2\right) = \frac{1}{T-t_0+1}\sum_{t=t_0}^{T}\left|E|Z|^p\hat{\sigma}_t^p - |Y_t|^p\right|^{1/q}, \tag{1.125}$$

where $Z \sim N(0,1)$. For $p > -1$, we have

$$E|Z|^p = \frac{2^{p/2}\Gamma((p+1)/2)}{\sqrt{\pi}}. \tag{1.126}$$

The combinations $(p=2, q=1)$, $(p=2, q=2)$, $(p=1, q=1)$, and $(p=1, q=2)$ are of special interest. In Section 3.11.1 we illustrate the differences between the

various combinations of p and q; see Figures 3.22 and 3.23. We use $\mathrm{MDE}_T^{(p,q)}$ with $p = 1$ and $q = 2$ in Section 3.11.1 to compare GARCH(1,1) and the exponentially weighted moving average.

Another useful performance measure is the mean of absolute ratio errors

$$\mathrm{MARE}_T^{(p)}(\hat{\sigma}^2) = \frac{1}{T - t_0 + 1} \sum_{t=t_0}^{T} \left| \frac{|Y_t|^p}{E|Z|^p \hat{\sigma}_t^p} - 1 \right|, \tag{1.127}$$

where $p > 0$ and $Z \sim N(0,1)$. We use $\mathrm{MARE}_T^{(p)}$ with $p = 2$ in Section 3.11.1 to compare GARCH(1,1) and the exponentially weighted moving average.

Prediction of Realized Volatility Above we have measured the performance of one step ahead predictions. We can also measure the performance of h-step ahead predictions, for $h = 1, 2, \ldots$. However, sometimes we are interested in estimating the realized volatility. Define the h-step realized volatility by

$$V_{t,h} = Y_{t+1}^2 + \cdots + Y_{t+h}^2.$$

Let $\hat{f}_{t,h}(X_{t+1})$ be a prediction of $V_{t,h}$. We can use the mean square root prediction error as in (1.123). We modify (1.121) to obtain

$$\mathrm{MSqE}_{T,h}(\hat{f}, f) = \frac{1}{T - h - t_0 + 1} \sum_{t=t_0}^{T-h} \left| \hat{f}_{t,h}(X_{t+1}) - V_{t+h} \right|^{1/2}.$$

We can consider $\hat{f}_{t,h}(X_{t+1})$ as an estimate of $E\left(Y_{t+1}^2 + \cdots + Y_{t+h}^2 \,|\, \mathcal{F}_t\right)$.

1.9.3 Performance of Conditional Covariance Estimators

Let us discuss measuring the performance of estimators of conditional covariance $f(x) = \mathrm{Cov}(Y, Z \,|\, X = x)$. Empirical performance measures of conditional covariance estimators can be found naturally in the case where

$$E(Y \,|\, X = x) = 0, \qquad E(Z \,|\, X = x) = 0,$$

so that

$$f(x) = \mathrm{Cov}(Y, Z \,|\, X = x) = E(YZ \,|\, X = x).$$

For example, we can use sample splitting, similarly as in (1.120), where a performance measure for the case of measuring the performance of a conditional variance estimator was given. Let $\hat{f}^*$ be an estimator of f, constructed from the data $(X_1, Y_1, Z_1), \ldots, (X_{n^*}, Y_{n^*}, Z_{n^*})$, where $1 \le n^* < n$. Then we can use

$$\frac{1}{n - n^*} \sum_{i=n^*+1}^{n} \left| \hat{f}^*(X_i) - Y_i Z_i \right|$$

to measure the performance of the estimator.

In autoregressive time–space smoothing methods, like in the MGARCH models and exponential moving average methods studied in Section 3.10.2, the explanatory variables are the previous observations, and the estimate $\hat{\gamma}_t$ of $E(Y_t Z_t \mid \mathcal{F}_{t-1})$ is calculated using observations $(Y_1, Z_1), \ldots, (Y_{t-1}, Z_{t-1})$, and now we define the mean deviation error by

$$\mathrm{MDE}_T^{(q)}(\hat{\gamma}) = \frac{1}{T - t_0 + 1} \sum_{t=t_0}^{T} |\hat{\gamma}_t - Y_t Z_t|^{1/q}, \tag{1.128}$$

where $q > 0$.

1.9.4 Performance of Quantile Function Estimators

Theoretical performance measures for the estimators of the conditional quantile

$$f(x) = \mathrm{Q}_p(Y \mid X = x)$$

can be defined similarly as in the case of conditional variance estimators. For example, using (1.119).

Empirical performance measures can be found in the case of continuous distribution of Y by using the fact

$$\begin{aligned} p &= P\left(Y \le \mathrm{Q}_p(Y \mid X = x) \,\middle|\, X = x\right) \\ &= E\left[I_{(-\infty, \mathrm{Q}_p(Y \mid X = x)]}(Y) \,\middle|\, X = x\right], \end{aligned}$$

where $x \in \mathbf{R}^d$. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be regression data and let

$$\hat{q}_i(x) = \hat{\mathrm{Q}}_{p,-i}(Y \mid X = x)$$

be a conditional quantile estimate constructed using the other data but not the ith observation. Let the cross validation quantity be

$$\hat{p} = \frac{1}{n-1} \sum_{i=1}^{n} I_{(-\infty, \hat{q}_i(X_i)]}(Y_i).$$

Finally, the performance is measured by the difference

$$p - \hat{p}.$$

Let us consider the time series setting, where we have observations $Y_1, \ldots, Y_T$. Then we can construct a conditional quantile estimator

$$\hat{q}_t = \hat{\mathrm{Q}}_p(Y_t \mid Y_{t-1}, \ldots)$$

using data $Y_1, \ldots, Y_{t-1}$, and calculate

$$\hat{p} = \frac{1}{T - t_0} \sum_{t=t_0+1}^{T} I_{(-\infty, \hat{q}_t]}(Y_t), \tag{1.129}$$

where $1 \leq t_0 \leq T-1$. We start to evaluate the performance of the estimator after t_0 observations are available, because any estimator can behave erratically when only a couple of observations are available.

Even when we would know the true quantiles, there is random fluctuation in the numbers $\hat{p}$. The random variables

$$Z_t = I_{(-\infty, q_p]}(Y_{t+1}), \qquad t = t_0, \ldots, T-1,$$

are Bernoulli random variables with $P(Z_t = 1) = p$, where q_p is the true quantile. If random variables Y_t are independent, then random variables Z_t are independent, and

$$M = \sum_{t=t_0}^{T-1} Z_t$$

is a binomial random variable with the distribution $\text{Bin}(n, p)$, where $n = T - t_0$. The probability mass function of M is

$$P(M = i) = \binom{n}{i} p^i (1-p)^{n-i},$$

for $i = 0, \ldots, n$. We can now calculate the numbers c_0 and c_1 such that

$$P(c_0 \leq p - \tilde{p} \leq c_1) \geq 1 - \alpha, \tag{1.130}$$

where $0 < \alpha < 1$ and $\tilde{p} = M/n$. We have

$$c_0 = p - n^{-1} z_{\alpha/2}, \qquad c_1 = p - n^{-1} z_{1-\alpha/2}, \tag{1.131}$$

where $z_{\alpha/2}$ and $z_{1-\alpha/2}$ are such that $P(z_{\alpha/2} \leq M \leq z_{1-\alpha/2}) \geq 1 - \alpha$.

If $\hat{p} > p$, this means that the quantile estimates were in average larger than the true quantiles. When we are estimating the left tail, so that p is close to zero, then the relation $\hat{p} > p$ means that the true distribution has a heavier left tail than the quantile estimates would indicate. When we are estimating the right tail, so that p is close to one, then this relation reverses, and the relation $\hat{p} > p$ means that the true distribution has a lighter left tail than the quantile estimates would indicate.

We will show the performance of quantile estimators by plotting the difference

$$R(p, \hat{p}) = \begin{cases} p - \hat{p}, & \text{when } p \leq 0.5, \\ \hat{p} - p, & \text{when } p > 0.5. \end{cases} \tag{1.132}$$

Thus, the difference $R(p, \hat{p})$ being negative means that the true distribution has a heavier tail than the quantile estimates would indicate. The difference $R(p, \hat{p})$ being positive means that the true distribution has a lighter tail than the quantile estimates would indicate.

Figure 1.6 illustrates the performance measurement of quantile estimators. We estimate the quantiles of the S&P 500 returns Y_t using the S&P 500 index data described in Section 1.6.1. Let $\hat{q}_t^e$ be the empirical quantile, defined in (1.26),

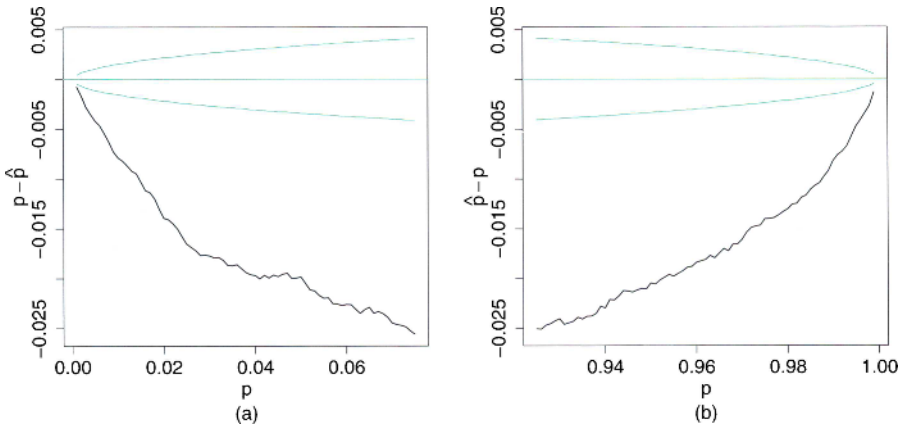

Figure 1.6 *Quantile estimator performance.* Function $p \mapsto R(p, \hat{p})$, defined in (1.132), is plotted with the black curves, when the quantile estimator is the empirical quantile. Panel (a) shows the range $p \in [0.001, 0.075]$, and panel (b) shows the range $p \in [0.925, 0.999]$. The green lines show level $\alpha = 0.05$ fluctuation bands.

and calculated using the data $Y_1, \ldots, Y_t$. We plot the function $p \mapsto R(p, \hat{p})$ in black. Panel (a) shows the range $p \in [0.001, 0.075]$, and panel (b) shows the range $p \in [0.925, 0.999]$. A green line is drawn at level 0, and it is accompanied by the level $\alpha = 0.05$ fluctuation bands, defined in (1.130)–(1.131). Figure 1.6 indicates that the true distribution has heavier tails than the empirical quantile estimates would indicate.

1.9.5 Performance of Estimators of Expected Shortfall

To derive a performance measure for estimators of expected shortfall, we can use the fact that for a continuous distribution of Y, we obtain

$$E\left[(Y - \mathrm{ES}_p(Y))\, I_{(-\infty, q_p]}(Y)\right] = 0.$$

Indeed, for a continuous distribution of Y, we have

$$\mathrm{ES}_p(Y) = \frac{1}{p} E\left[Y\, I_{(-\infty, q_p]}(Y)\right]$$

and

$$E\left[I_{(-\infty, q_p]}(Y)\right] = p.$$

If we are in the time series setting and have identically distributed observations $(X_1, Y_1), \ldots, (X_T, Y_T)$, then we can construct an estimator of the expected shortfall

$$\widehat{\mathrm{ES}}_{p,t}$$

using data $(X_1, Y_1), \ldots, (X_t, Y_t)$ and calculate the performance measure

$$\frac{1}{T-t_0} \sum_{t=t_0}^{T-1} \left(Y_{t+1} - \widehat{\mathrm{ES}}_{p,t} \right)^2 I_{(-\infty, \hat{q}_t]}(Y_{t+1}),$$

where $\hat{q}_t = \hat{Q}_{p,t}$ is a quantile estimator and $1 \leq t_0 \leq T-1$.

1.9.6 Performance of Classifiers

Theoretical Performance Measures Let $g : \mathbf{R}^d \to \{0, \ldots, K-1\}$ be a classification function. The probability of the classification error is

$$R(g) = P(g(X) \neq Y),$$

and this can be used to measure the goodness of g. The goodness of an empirical classification rule $\hat{g}$, calculated from data $(X_1, Y_1), \ldots, (X_n, Y_n)$, is measured by

$$R(\hat{g}) = P\left(\hat{g}(X) \neq Y\right),$$

where P is the probability measure of $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n)$. We can write the probability of the misclassification more transparently. We have that

$$R(\hat{g}) = \sum_{k=0}^{K-1} P(Y = k) \int_{\hat{G}_k^c} f_{X|Y=k},$$

where $f_{X|Y=k} : \mathbf{R}^d \to \mathbf{R}$ is the density function of $X \,|\, Y = k$, and

$$\hat{G}_k = \{x \in \mathbf{R}^d : \hat{g}(x) = k\}, \qquad k = 0, \ldots, K-1,$$

is the subset of the sample space, where the classification function $\hat{g}$ chooses class k.

When we analyze the asymptotic performance of the classification functions, we should note that $R(\hat{g})$ does not converge to zero, but at best we can hope that it converges to the minimal classification error $R(g^*)$, which is the classification error of the Bayes rule g^*, defined in (1.75). Thus we should study the rate of convergence to zero of $R(\hat{g}) - R(g^*)$. Let us consider the two-class case $K = 2$ with the equal class priors $P(Y = 0) = P(Y = 1) = 1/2$. Then,

$$R(g^*) = \frac{1}{2} \int_{\mathbf{R}^d} \min\{f_{X|Y=0}(x), f_{X|Y=1}(x)\}\, dx$$

and

$$R(\hat{g}) - R(g^*) = \frac{1}{2}\, d_{f_{X|Y=0}, f_{X|Y=1}}(\{x : \hat{g}(x) = 1\}, \{x : g^*(x) = 1\}),$$

where

$$d_{g_1, g_2}(G_1, G_2) = \int_{G_1 \Delta G_2} |g_1 - g_2|,$$

with

$$G_1 \Delta G_2 = (G_1^c \cap G_2) \cup (G_1 \cap G_2^c)$$

the symmetric difference of G_1 and G_2. The rate of convergence has been studied in Mammen & Tsybakov (1999).

Empirical Performance Measures The frequency of misclassification can be used as an empirical performance measure of a classification method. We can use sample splitting as in the case of regression function estimation, see (1.115). Let us have classification data $(X_1, Y_1), \ldots, (X_n, Y_n)$ and let us construct classifier $\hat{g}^*$ using the first part $(X_1, Y_1), \ldots, (X_{n_1}, Y_{n_1})$ of data, where $1 \leq n_1 < n$, and typically $n_1 = [n/2]$. We can use

$$\frac{1}{n - n_1} \sum_{i=n_1+1}^{n} I_{\{\hat{g}^*(X_i)\}^c}(Y_i)$$

as an estimator of $P(\hat{g}(X) \neq Y)$, where $\hat{g}$ is constructed from the whole sample.

We can also use cross validation, as in the case of regression function estimation in (1.116). In the time series setting, when we have regression data $(X_1, Y_1), \ldots, (X_T, Y_T)$, it is natural to measure the performance of classification method by

$$\frac{1}{T - t_0} \sum_{t=t_0}^{T-1} I_{\{\hat{g}_t^*(X_{t+1})\}^c}(Y_{t+1}), \tag{1.133}$$

where $\hat{g}_t^*$ is a classifier constructed using the data $(X_1, Y_1), \ldots, (X_t, Y_t)$, and t_0 is chosen so large that the first classifier $\hat{g}_{t_0}^*$ in the sequence is already a reasonable classifier. We can divide the classification error into K components

$$\frac{1}{T - t_0} \sum_{t=t_0}^{T-1} I_{\{\hat{g}_t^*(X_{t+1})\}^c}(k)\, I_{\{k\}}(Y_{t+1}), \tag{1.134}$$

where $k = 0, \ldots, K-1$, which estimate $P(\hat{g}(X) \neq Y \,|\, Y = k)$.

1.10 CONFIDENCE SETS

We give first several definitions of a confidence interval for regression function estimation. Then we define confidence bands.

1.10.1 Pointwise Confidence Intervals

A pointwise confidence interval $[L, U]$ for the estimation of regression function $f : \mathbf{R}^d \to \mathbf{R}$ at point $x \in \mathbf{R}^d$, with the confidence level $1 - \alpha$, is such that for all $P \in \mathcal{P}$, for all x in a suitable subset of $\mathbf{R}^d$, we have

$$P\,(L \leq f(x) \leq U) = 1 - \alpha,$$

where $\mathcal{P}$ is a collection of distributions of (X, Y). Typically we can give asymptotic confidence intervals of type

$$P\,(L_n \leq f(x) \leq U_n) \longrightarrow 1 - \alpha,$$

when $n \to \infty$, where $[L_n, U_n]$, is a sequence of intervals. Asymptotic pointwise confidence intervals can typically be derived from the asymptotic distribution of the estimator. If we have that

$$n^a \left(\hat{f}(x) - f(x) \right) \xrightarrow{d} N(\mu, \sigma^2),$$

where symbol "$\xrightarrow{d}$" denotes the convergence in distribution, then we can choose

$$L_n = \hat{f}(x) - n^{-a}(\mu + z_{1-\alpha/2}\sigma)$$

and

$$U_n = \hat{f}(x) + n^{-a}(\mu + z_{1-\alpha/2}\sigma),$$

where we denote $z_\alpha = \Phi^{-1}(\alpha)$, and Φ is the distribution function of the standard normal distribution. That is, z_p is the p-quantile of the $N(0,1)$ distribution: For $Z \sim N(0,1)$, we have $P(z_{\alpha/2} \le Z \le z_{1-\alpha/2}) = 1 - \alpha$.

More generally, we can use the term "level $1-\alpha$ confidence interval" if the inequality

$$P\left(L \le f(x) \le U\right) \ge 1 - \alpha$$

holds for all $P \in \mathcal{P}$. We can use the term "asymptotic level $1-\alpha$ confidence interval" if

$$\liminf_{n\to\infty} P\left(L_n \le f(x) \le U_n\right) \ge 1 - \alpha$$

for all $P \in \mathcal{P}$. Note that in the asymptotic case is is important to distinguish a uniform asymptotic level $1-\alpha$ confidence interval, which satisfies

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}} P\left(L_n \le f(x) \le U_n\right) \ge 1 - \alpha.$$

As pointed out by Wasserman (2005, p. 6), it is better to have uniform confidence intervals.

We give an example of a confidence interval in Section 3.2.10, for the case of kernel regression. As mentioned in Ruppert et al. (2003, Section 6.2) we can derive an approximate confidence interval for linear estimators under some assumptions. We noted in (1.2) that many estimators can be written as linear estimators

$$\hat{f}(x) = \sum_{i=1}^{n} l_i(x)\, Y_i = l(x)'\mathbf{y},$$

where $l(x) = (l_1(x), \ldots, l_n(x))'$ and $\mathbf{y} = (Y_1, \ldots, Y_n)'$. Let us assume that $\hat{f}(x) \sim N(f(x), \mathrm{Var}(\hat{f}(x)))$. If

$$\mathrm{Cov}(\mathbf{y}) = \sigma^2 I_n,$$

then

$$\mathrm{Var}\left(\hat{f}(x)\right) = l(x)'\mathrm{Cov}(\mathbf{y})l(x) = \sigma^2 \|l(x)\|^2.$$

Estimating σ^2 with $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(Y_i - \hat{f}(X_i))^2$ leads to the confidence interval

$$\left[\hat{f}(x) - \hat{\sigma}\|l(x)\| z_{1-\alpha/2},\ \hat{f}(x) + \hat{\sigma}\|l(x)\| z_{1-\alpha/2}\right],$$

where α is the confidence level, $0 < \alpha < 1$, and $z_{1-\alpha/2}$ is the quantile of the standard normal distribution.

1.10.2 Confidence Bands

A confidence band $(L(x), U(x))$, $x \in A$, for the estimation of regression function $f : \mathbf{R}^d \to \mathbf{R}$, for the set $A \subset \mathbf{R}^d$, with the confidence level $1 - \alpha$, is such that

$$P\left(L(x) \leq f(x) \leq U(x), \text{ for all } x \in A\right) = 1 - \alpha. \tag{1.135}$$

Confidence bands are called also simultaneous confidence bands, confidence envelopes, or variability bands. The confidence statement of the type

$$P\left(\sup_{x \in A} |f(x) - \hat{f}(x)| \leq c_n\right) = 1 - \alpha$$

is equivalent to (1.135) if

$$L(x) = \hat{f}(x) - c_n, \qquad U(x) = \hat{f}(x) + c_n.$$

We can replace the supremum norm with some other function space norm to obtain confidence balls. For example, the L_2 confidence ball with the confidence level $1 - \alpha$ satisfies

$$P\left(\|f(x) - \hat{f}(x)\|_2 \leq c_n\right) = 1 - \alpha.$$

A confidence band in the linear model is mentioned in Section 2.1.5.

1.11 TESTING

In the linear regression model

$$Y = \alpha + \beta_1 X_1 + \cdots + \beta_d X_d + \epsilon$$

the typical tests are the tests of restrictions

$$H_0 : \beta_k = 0, \tag{1.136}$$

for $k = 1, \ldots, d$, and

$$H_0 : \beta_1 = \cdots = \beta_d = 0. \tag{1.137}$$

Testing of these hypothesis is considered in Section 2.1.5. There are several ways to generalize these tests to a nonparametric setting, where

$$Y = f(X) + \epsilon.$$

The hypothesis in (1.137) can be generalized to the hypothesis

$$H_0 : f(x) \equiv 0,$$

when we assume that $EY = 0$. We can use a test statistics $T = \|\hat{f}\|$, where $\hat{f}$ is a nonparametric estimate of f. The norm $\|\cdot\|$ can be the L_2 norm, a weighted

L_2 norm, or some other function space norm. Large values of the test statistics T lead to the rejection of the null hypothesis. For the linear regression function $f(x) = \alpha + \beta_1 x_1 + \cdots + \beta_d x_d$, it holds that

$$\frac{\partial}{\partial x_k} f(x) = \beta_k.$$

Thus we can generalize the parameter restriction hypothesis (1.136) to the nonlinear case by

$$H_0 : \frac{\partial}{\partial x_k} f(x) \equiv 0, \tag{1.138}$$

for $k = 1, \ldots, d$. We can generalize the parameter restriction hypothesis (1.137) to the nonlinear case by

$$H_0 : \frac{\partial}{\partial x_1} f(x) \equiv 0, \ldots, \frac{\partial}{\partial x_d} f(x) \equiv 0.$$

We can test the null hypothesis (1.138) with the test statistics

$$T = \left\| \frac{\partial}{\partial x_k} f(x) \right\|,$$

where $\hat{f}$ is a nonparametric estimator of f and $\| \cdot \|$ is a function space norm.

The distribution of the test statistics can be approximated by bootstrap. Generate first B bootstrap samples from the original sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. Based on a bootstrap sample $(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$, the test statistics T^* is calculated. We obtain a sequence $T_1^*, \ldots, T_B^*$ of values of the test statistics. Let $q_{1-\alpha}$ be the empirical quantile of the sequence of the values of the test statistics. Then we reject the null hypothesis at level $0 < \alpha < 1$, if the observed value t of the test statistics satisfies $t > q_{1-\alpha}$.

Härdle & Mammen (1993) have proposed the wild bootstrap. First the regression function f is estimated with $\hat{f}$ (under the null hypothesis). Then the residuals $\hat{\epsilon}_i = Y_i - \hat{f}(X_i)$ are calculated. Finally, the bootstrap residual ϵ_i^* is generated from a distribution which satisfies $E\epsilon_i^* = 0$, $E(\epsilon_i^*)^2 = \hat{\epsilon}_i^2$, and $E(\epsilon_i^*)^3 = \hat{\epsilon}_i^3$. The bootstrap sample is $(X_1, Y_1^*), \ldots, (X_n, Y_n^*)$, where $Y_i^* = \hat{f}(X_i) + \epsilon_i^*$.