PART II

METHODS OF REGRESSION AND CLASSIFICATION:
OVERVIEW OF REGRESSION AND CLASSIFICATION

1.1 REGRESSION

In regression analysis, we are interested in predicting or inferring causal relationships. We try to predict the value of a response variable given the values of explanatory variables or try to discover the causal influence of the explanatory variables on the response variable. The inference of a causal relationship is important when we want to change the values of an explanatory variable in order to get an optimal value for the response variable. For example, we might want to know the influence of education on the employment status of a worker in order to choose the best education. On the other hand, prediction is applied also in 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 $E(Y | X = x)$ of the response variable $Y \in \mathbb{R}$ given the explanatory variables $X \in \mathbb{R}^d$. The choice of the explanatory variables and the method of estimation can depend on the purpose.

Multivariate Nonparametric Regression and Visualization, Per-Anders Ekstrom

Copyright © 2014 John Wiley & Sons, Inc.
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 effects is defined in Section 2.1.3.

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

\[ \hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_d x_d. \]  

(1.1)

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

\[ \hat{f}(x) = \sum_{i=1}^{n} \hat{w}_i(x) Y_i. \]  

(1.2)

for some sequence of weights \( \hat{w}_1(x), \ldots, \hat{w}_n(x) \). In fact, for the linear regression estimator, representations (1.1) and (1.2) hold; see (2.1.1). In the case of local averaging estimators, like regreession, kernel estimation, and nearest-neighbor estimation, we use the notation \( \hat{f}(x) = \sum_{i=1}^{n} \hat{w}_i(x) Y_i \). In the case of local averaging estimator the weights \( \hat{w}_i(x) \) satisfy the properties that \( \hat{w}_i(x) \) is close to zero when \( x_i \) is distant from \( x \) and that \( \hat{w}_i(x) \) is large when \( x_i \) is near \( x \). Local averaging is discussed in Section 3. There exists a regression function estimator that can be written as in (1.2), like the orthogonal series estimator with local thresholding, see (2.7.2).

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 variances, 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.

### 1.1.3 Random Design and Fixed Design

**Random Design Regression** In random design regression the data are a sequence of \( n \) pairs:

\[ (x_1, y_1), \ldots, (x_n, y_n). \]  

(1.3)

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

\[ (X_1, Y_1), \ldots, (X_n, Y_n). \]  

(1.4)
However, sometimes we do not distinguish between a random variable and its realization, and the definition of (1.14) is used also in the place of notation (1.13) to denote a realization of the random variable and not the random variable itself.

In regression analysis we typically want to estimate the conditional expectation

\[ f(x) = \mathbb{E}(Y | X = x), \]

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 expectations, we could estimate conditional medians, conditional variances, conditional quantiles, and so on. Estimation of the conditional means of distribution was 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.

**Linear Regression**

In linear regression analysis, the data are a sequence

\[ (y_1, \ldots, y_n) \]

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

Now the design points are not chosen by a random mechanism, but they are chosen by the constructor 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 linear design regression we typically do not assume that the data are identically distributed. For example, we may assume that

\[ Y_2 = f(x_2) + \epsilon_2, \quad i = 1, \ldots, n, \]

where \(x_i \in [a_i, b_i]\), \(f : [a, b] \rightarrow \mathbb{R}\) is the function we want to estimate, and \(\epsilon_2 \in [0, 1]\). Now the data \(Y_1, \ldots, Y_n\) are not identically distributed, since the observations \(Y_i\) have different expectations.

### 1.1.2 Linear Regression

The regression function is typically defined as a conditional expectation. Besides expectation and conditional expectation also medians and conditional medians can be used to characterize the center of a distribution and thus are present 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 this data
\[(X_1, Y_1), \ldots, (X_n, Y_n)\]
satisfies 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 \(X:\)
\[f(x) = \mathbb{E}(Y | X = x), \quad x \in \mathbb{R}^d, \tag{1.5}\]
where \((X, Y)\) has the same distribution as \((X_i, Y_i), i = 1, \ldots, n,\) and \(X \subseteq \mathbb{R}^d, Y \subseteq \mathbb{R}\). The random variable \(Y\) is called the response variable, and the elements of random vector \(X\) are called the explanatory variables.

The moment regression variables \(Y \subseteq \mathbb{R}\) within continuous distribution can be obtained by
\[e(Y) = \int_{\mathbb{R}} y f_Y(y) \, dy, \tag{1.6}\]
where \(f_Y : \mathbb{R} \to \mathbb{R}\) 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
\[\mathbb{E}(Y | X = x) = \int_{\mathbb{R}} y f_{Y | X}(y | x) \, dy, \tag{1.7}\]
where the conditional density can be defined as
\[f_{Y | X}(y | x) = \frac{f_{X,Y}(x,y)}{f_X(x)}, \quad y \in \mathbb{R}, \tag{1.8}\]
where \(f_X(x) > 0\) and \(f_{Y | X}(y | x) > 0\) everywhere, where \(f_{X,Y} : \mathbb{R}^{d+1} \to \mathbb{R}\) is the joint density of \((X, Y)\) and \(f_X : \mathbb{R}^d \to \mathbb{R}\) is the density of \(X\).

Figure 1.1 illustrates mean regression. Our data contain the daily NASDAQ 100 returns \(X_k := (S_k - S_{k-1}) / S_{k-1}\), where \(S_k\) is the price of the index. There are about 10,000 observations. The NASDAQ 100 index data are described more precisely in Section 1.6.1. We define the explanatory and the response variables as
\[X_k := \log \left( \frac{1}{k} \sum_{i=1}^{k} \Delta S_i \right), \quad Y_k := \log \left( \frac{1}{k} \sum_{i=1}^{k} \Delta S_i \right). \]

Panel (a) shows the scatter plot of \((X_k, Y_k)\), and panels (b) shows the estimated density of \((X_k, Y_k)\) together with the estimated regression functions. The red line shows the linear regression function estimates, and the blue line shows a kernel regression estimate with smoothing parameter \(h := 0.4\). The density is estimated using kernel
Figure 4.11: Kernel 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 $h = 0.3$. Linear regression is discussed in Section 2.11, and kernel methods are discussed in Section 3.2. In the scatter plot we have used histogram smoothing with 1000 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 distributions of a random variable $Y$: as the number $\text{median}(Y) < 10$, satisfying

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

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

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

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

$$\text{median}(Y | X = x) = \inf \{ y : P(Y \leq y | X = x) > 0.5 \}, \quad x \in \mathbb{R}^d. \quad (1.9)$$

The sample median of observations $X_1, \ldots, X_n$ 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,

$$\text{median}(X_1, \ldots, X_n) = Y_{[n/2]} + i,$$

where $Y_1 \leq \cdots \leq Y_n$ is the ordered sample and $[n/2]$ is the largest integer smaller or equal to $n/2$. 

\[ \]
The model is defined as an argument maximizing the density function of a random variable:

\[ \text{mode}(Y) := \arg \max_{y \in \mathbb{R}} f_Y(y), \]

(1.11)

where \( f_Y : \mathbb{R} \to \mathbb{R} \) 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 conditionally mode is defined as an argument maximizing the conditional density:

\[ \text{mode}(Y \mid X = x) := \arg \max_{y \in \mathbb{R}} f_{Y \mid X = x}(y). \]

### 4.1.3 Partial Effects and Derivatives of Prediction

Let us consider a non-linear 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 \( X \) denotes \( x = (x_1, \ldots, x_d) \). 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 \( E(Y \mid X = x) = \alpha + \beta_1 x_1 \) 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:

\[ e(x_1; x_2, \ldots, x_d) := \frac{\partial}{\partial x_1} \log E(Y \mid X = x) = \frac{\partial}{\partial x_1} \log f_Y(Y \mid X = x) \]

\[ = \frac{\partial}{\partial x_1} E(Y \mid X = x) \cdot \frac{E(Y \mid X = x)}{f_Y(Y \mid X = x)} \]
where $x_1 > 0$ and $PS(x' \mid x_2 = \ldots = x_k) > 0$. The partial elasticities describe 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. The partial \textit{max}elasticity of $X_1$ is defined as:

$$
\delta_{\max} \frac{\partial}{\partial x_1} \log \mathbb{E}(y' \mid x_2 = \ldots = x_k) - \frac{1}{\mathbb{E}(y' \mid x_2 = \ldots = x_k)}
$$


where $\mathbb{E}(y' \mid x_2 = \ldots = x_k) > 0$. The partial \textit{max}elasticity 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 these visualizations of partial effects as a tool to visualize regression functions. In Section 4.2 we show how local estimates can be used to visualize the model structure of functions. The model structure of a function means the number of the local maxima and minima, and the location of the local maximum of a function. Analogously, local maxima and minima can be used to visualize the functional structure of a function, whereas the functional structure means the number, the location, and the location of the local maxima 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 model structure or functional 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 closer to a linear function. When we visualize the model structure of the partial effect of variable $X_1$, then we get information about whether a variable $X_1$ is causing the expected values 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 values of the response variable $Y$ (the location of the local maximum of the partial effect), and where the influence of the response variable $X_1$ is largest (the location of the local maximum of the partial effect). Analogous conclusions can be made by visualizing the functional 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 linear regression function estimation, and this method is presented in Section 5.2.1. The second method is to use a local linear estimation, and this method is presented in Section 5.2.2.

4.1.1 \textit{Max}eRgression:

The \textit{max}eRgression gives information about the center of the conditional distribution, and while the \textit{variance} regression we get information about the dispersion and so the

\textit{max}eRgression follows from the approximation:

$$
\log f(x + k) \cdot \log f(x) \approx \left( \frac{f(x + k) - f(x)}{f(x)} \right),
$$

which follows from the approximation $\log(x) \approx x - 1$, when $x \neq 1$. \footnote{This approximation follows from the approximation $\log(x) \approx x - 1$, when $x \neq 1$.}
Variance of the tails of the conditional distribution. Variance is a classical measure of dispersion and risk which is used for examples in the Markovian theory of portfolio selection. Partial moments are risk measures that generalize the variances.

**Variance and Conditional Variance.** The variance of random variable \( Y \) is defined by
\[
\text{Var}(Y) \triangleq \mathbb{E}((Y - \mathbb{E}(Y))^2) = \mathbb{E}Y^2 - (\mathbb{E}Y)^2.
\]  
(1.12)

The standard deviation of \( Y \) is the square root of the variance of \( Y \). The conditional variance of random variable \( Y \) is defined as
\[
\text{Var}(Y \mid X = x) \triangleq \mathbb{E}( (Y - \mathbb{E}(Y \mid X = x))^2 \mid X = x) = \mathbb{E}(Y^2 \mid X = x) - (\mathbb{E}(Y \mid X = x))^2.
\]  
(1.13)

The conditional variance is the square root of the conditional variance. The sample variance is defined by
\[
\widehat{\text{Var}}(Y) \triangleq \frac{1}{m} \sum_{i=1}^{m} (Y_i - \bar{Y})^2 = \sum_{i=1}^{m} Y_i^2 - \bar{Y}^2,
\]  
where \( Y_1, \ldots, Y_m \) is a sample of random variables having identical distribution with \( Y \).

**Conditional Variance Estimation.** Conditional variance \( \text{Var}(Y \mid X = x) \) can be constant and depending on \( x \). Let us assume
\[
Y = f(X) + e,
\]  
where \( f(x) = \mathbb{E}(Y \mid X = x) \) and \( e = Y - f(X) \), so that \( \mathbb{E}(e \mid X = x) \equiv 0 \).

If \( \text{Var}(Y \mid X = x) = \mathbb{E}(e^2 \mid X = x) \) is a constant not depending on \( x \), we say that the noise is homoscedastic. Otherwise the noise is heteroscedastic. If the noise is heteroscedastic, it is of interest to estimate the conditional variance
\[
\text{Var}(Y \mid X = x) = \mathbb{E}(e^2 \mid X = x).
\]

**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 \( f(x) = \mathbb{E}(Y \mid X = x) \) by \( \hat{f}(x) \). Second we calculate the residuals
\[
\hat{e}_i = Y_i - \hat{f}(X_i),
\]
and estimate the conditional variance from the data \( (X_1, \hat{e}_1^2), \ldots, (X_m, \hat{e}_m^2) \).

**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 \( \mathbb{E}(Y^2 \mid X = x) \) using the regression data \( (X_1, Y_1^2), \ldots, (X_m, Y_m^2) \). Second we estimate the conditional expectation \( f(x) = \mathbb{E}(Y \mid X = x) \) using data \( (X_1, Y_1), \ldots, (X_m, Y_m) \).
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 heteroscedastic fixed design regression model:

$$Y_i = f(x_i) + \sigma(x_i) \varepsilon_i, \quad i = 1, \ldots, n, \quad (1.15)$$

where $x_i \in \mathbb{R}^d$, $f: \mathbb{R}^d \to \mathbb{R}$ is the mean function, $\sigma: \mathbb{R}^d \to \mathbb{R}$ is the standard deviation function, and $\varepsilon_i$ are independently distributed with $\mathcal{N}(0, 1)$. Here we want to estimate both the function $f$ and the function $\sigma$. Wasserman (2000), Section 5.5 has proposed making the following transformation. Set $Z_i = \log(Y_i - f(x_i))$. Then we have

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

Let $\hat{f}$ be an estimator of $f$ and define $\hat{Z}_i = \log(Y_i - \hat{f}(x_i))^2$. Let $\hat{\sigma}(x)$ be an estimator of $\log\sigma^2(x)$, obtained using regression data $(x_1, Z_1), \ldots, (x_n, Z_n)$, and define $\hat{\sigma}^2(x) = \exp\{\hat{\sigma}(x)\}$.

A difference-based method for conditional variance estimation has been proposed. Fix $x_2 < \cdots < x_n$. Let $\sigma^2(x)$ be estimated with

$$\hat{\sigma}^2(x) = \exp\{\hat{\sigma}(x)\},$$

where $\hat{\sigma}$ is a regression function estimate obtained with the regression data $(x_1, (Y_i - f(x_i))^2), i = 2, \ldots, n$. This approach has been used in Wang, Brown, Carl & Levine (2008).

Variance Estimation with Heteroscedastic Noise. Let us consider the fixed design regression model:

$$Y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $x_i \in \mathbb{R}^d$, $f: \mathbb{R}^d \to \mathbb{R}$ is the mean function, and $\mathcal{N}(0, \sigma^2)$. In the case of heteroscedastic noise we should estimate

$$\sigma^2 = \mathbb{E}(\varepsilon^2).$$

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

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

These types of estimators were studied by Müller & Staudte (1982), Hall & Wellner (1989), Hall & Brown (1989), and Rosenbaum (1994). Local polynomial estimators were studied by Härdle, Müller, Welln, and Müller (1991), and Li & Yau (1991). A difference-based estimator was studied by Petermann (1991). We need the estimator

$$\hat{\sigma}^2 = \frac{1}{2(n - 1)} \sum_{i=2}^{n} \left( Y_i - Y_{i-1} \right)^2,
where it is assumed that $x_1, \ldots, x_n \in \mathbb{R}$, and $x_1 < \cdots < x_n$. The estimator was developed and modified in various ways in Rice (1994), Last, Reid, & Gasser (1988), Hall, Key, & Thomsen (1990), Hall, Key, & Thomsen (1991), and Honsberg, Key, & Thomsen (2001).

**Conditional Variances for a Time Series Setting** In a time series setting, where we observe $Y_1, Y_2, \ldots$, the conditional heteroscedasticity assumption is that

$$Y_k = \sigma_k, \quad k = 0, 1, 2, \ldots, (1.16)$$

where $\sigma_k$ is an initial sequence, $\sigma_0 \equiv 0$, $\sigma_k^2 = \alpha_k$, and $\sigma_k$ is the volatility process. The volatility process is a predictable random process, that is, $\sigma_k$ is measurable with respect to the sigma-field generated by the variables $Y_k, Y_{k-1}, \ldots$. When we assume that $\sigma_k$ is independent from $Y_{k-1}, Y_{k-2}, \ldots$, then under the conditional heteroscedasticity model,

$$\text{Var}(Y_k | \mathcal{F}_{k-1}) = \text{Var}(\sigma_k^2 | \mathcal{F}_{k-1}) = \sigma_k^2 \text{Var}(\epsilon_k | \mathcal{F}_{k-1}) = \sigma_k^2 \text{Var}(\epsilon_k) = \sigma_k^2, (1.17)$$

where $\mathcal{F}_{k-1}$ is the sigma-algebra generated by variables $Y_{k-1}, Y_{k-2}, \ldots$. In a conditional heteroscedasticity model the main interest is in estimating the value of the random variable $\sigma_k^2$, which is thus related to estimating the conditional variance. The statistical problem is to predict $\sigma_k^2$ using a finite number of past observations $Y_{k-1}, Y_{k-2}, \ldots$. Special cases of conditional heteroscedasticity 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 a random variable $Y \in \mathbb{R}$ is defined as $\text{Var}(Y) = \mathbb{E}\{Y - \mathbb{E}[Y]\}^2$. The variance can be generalized to other ordered moments,

$$\mathbb{E}\{Y - cY\}^k,$$

for $k = 1, 2, \ldots$. The ordered moments take a contribution both from the left and the right tails of the distribution. When we are interested only in the left tail on 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

$$\text{UPPM}_k(Y) := \mathbb{E}\{(Y - c)^k I_{Y \geq c}\}(Y),$$

and the lower partial moment is defined as

$$\text{LPPM}_k(Y) := \mathbb{E}\{(Y - c)^k I_{Y < c}\}(Y),$$

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

$$\text{UPPM}_k(Y) := \int_{c}^{\infty} (y - c)^k f_Y(y) \, dy, \quad \text{LPPM}_k(Y) := \int_{-\infty}^{c} (y - c)^k f_Y(y) \, dy.$$
For example, where \( k \in \mathcal{O}_n \), then:

\[
\Phi_{\text{PMU}}(Y) = \Phi(Y \geq \tau), \quad \Phi_{\text{PMU}}(X) = \Phi(Y \leq \tau),
\]

so that the upper partial moment is equal to the probability that \( Y \) is greater or equal to \( \tau \) and the lower partial moment is equal to the probability that \( Y \) is smaller or equal to \( \tau \). For \( k = X \) and \( \tau = M_Y \) the partial moments are called upper or lower semi-variances of \( Y \). The lower semi-variance is defined as:

\[
\mathbb{E}[(Y - M_Y)^+ X_{< M_Y} | Y]
\]

(1.10)

The square root of the lower semi-variance 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 moment by changing the expectations to conditional expectations.

### 1.1.2 Covariance and Correlation

The covariance of random variables \( Y \) and \( X \) is defined by:

\[
\text{Cov}(Y, X) = \mathbb{E}[(Y - \mathbb{E}[Y])(X - \mathbb{E}[X])] = \text{sc}(Y) \cdot \text{sc}(X).
\]

The sample covariance is defined by:

\[
\hat{\text{Cov}}(Y, X) = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X}) = \frac{1}{n-1} \sum_{i=1}^{n} Y_i X_i - \bar{Y} \bar{X},
\]

where \( Y_1, \ldots, Y_n \) and \( X_1, \ldots, X_n \) are samples of random variables having identical distributions with \( Y \) and \( X \), \( \bar{Y} := \sum_{i=1}^{n} Y_i/n \) and \( \bar{X} := \sum_{i=1}^{n} X_i/n \). 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:

\[
\text{Cov}(Y, X | \mathcal{F}) = \mathbb{E}[((Y - \mathbb{E}[Y])(X - \mathbb{E}[X])) | \mathcal{F}]
\]

and the second method uses:

\[
\text{Cov}(Y, X) = \mathbb{E}(Y | X = x) \cdot \text{sc}(X | X = x)
\]

The correlation is defined by:

\[
\text{Cor}(Y, X) = \frac{\text{Cov}(Y, X)}{\text{sc}(Y) \cdot \text{sc}(X)}
\]

where \( \text{sc}(Y) \) and \( \text{sc}(X) \) are the standard deviations of \( Y \) and \( X \). The conditional correlation is defined by:

\[
\text{Cor}(Y, X | X = x) = \frac{\text{Cov}(Y, X | X = x)}{\text{sc}(Y | X = x) \cdot \text{sc}(X | X = x)}
\]

(1.19)

where:

\[
\text{sc}(Y | X = x) = \sqrt{\text{Var}(Y | X = x)}, \quad \text{sc}(X | X = x) = \sqrt{\text{Var}(X | X = x)}.
\]
We can write

\[ \text{Cov}(Y, Z | X = x) = \text{Cov}(\bar{Y}, \bar{Z} | X = x), \quad (1.19) \]

where:

\[ \bar{Y} = \frac{Y}{\text{std}(Y | X = x)}, \quad \bar{Z} = \frac{Z}{\text{std}(Z | X = x)}. \]

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 \( \text{std}_X(x) \) and \( \text{std}_Z(x) \), and calculate the standardized observations \( \bar{Y}_x = Y_x / \text{std}_Y(X_x) \) and \( \bar{Z}_x = Z_x / \text{std}_Z(X_x) \). Second we estimate the conditional correlation using \( \text{Cov}(X_x, \bar{Y}_x, \bar{Z}_x), x = 1, \ldots, n \).

A time series \( (Y_t)_{t \in \mathbb{N}} \) is weakly stationary if \( E[Y_t] = E[Y], \) and \( E[Y_{t+h} | Y_t] \) depends only on \( h \), for all \( t, h \in \mathbb{N} \). For a weakly stationary time series \( (Y_t)_{t \in \mathbb{N}} \), the autocovariance function is defined by

\[ \gamma(h) = \text{Cov}(Y_t, Y_{t+h}) \]

and the autocorrelation is defined by

\[ \rho(h) = \frac{\gamma(h)}{\gamma(0)}. \]

where \( h = 0, \pm 1, \pm 2, \ldots \).

A vector time series \( (X_t)_{t \in \mathbb{N}}, X_t \in \mathbb{R}^d \) is weakly stationary if \( E[X_t] = E[X], \) and \( E[X_{t+h} | X_t] \) depends only on \( h \), for all \( t, h \in \mathbb{N} \). For a weakly stationary vector time series \( (X_t)_{t \in \mathbb{N}} \), the autocovariance function is defined by

\[ \gamma(h) = \text{Cov}(X_t, X_{t+h}) = \text{E}[(X_t - \mu)(X_{t+h} - \mu)^T], \quad (1.21) \]

for \( h = 0, \pm 1, \pm 2, \ldots \), where \( \mu = E[X] = E[X_t] \). Matrix \( \gamma(h) \) is a \( d \times d \) matrix which is not symmetric. It holds that

\[ \gamma(h) = \gamma(-h)^T. \quad (1.22) \]

### 1.1.5 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 \( p \)th quantile is defined as:

\[
Q_p(Y) = \inf \{ y : F_Y(Y < y) > p \}, \quad x \in \mathbb{R}^d
\]

where \( 0 < p < 1 \). For \( y = 1/2 \), \( Q_p(Y) \) is equal to median \( \text{med}(Y) \), defined in (1.3.1).

In the case of a continuous distribution function we have:

\[
P(Y \leq Q_p(Y)) = p
\]

and thus in locution:

\[
Q_p(Y) = F_Y^{-1}(p),
\]

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

\[
Q_p(Y | X = x) = \inf \{ y : F_Y(Y < y | X = x) > p \}, \quad x \in \mathbb{R}^d,
\]

where \( 0 < p < 1 \). Conditional quantile estimation has been considered in Koenker (2005) and Koenker & Halpern (1975).

Estimation of Quantiles: 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 data \( Y_1, \ldots, Y_n \), as:

\[
\hat{F}(y) = \frac{1}{n} \sum_{i=1}^{n} I(Y_i \leq y), \quad y \in \mathbb{R}.
\]

Now we can define an estimate of the quantile by:

\[
\hat{Q}_p = \inf \{ x : \hat{F}(x) > p \},
\]

where \( 0 < p < 1 \). Now it holds that:

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

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 \( m \) := \lfloor np \rfloor \), where \( \lfloor np \rfloor \) is the greatest integer \( \geq np \).
3. Set \( \hat{Q}_p := Y_{(m)} \).
Standardized Linear Model

We can also use an estimate of the standard deviation to derive an estimator for a quantile. Specifically, consider the linear model

\[ Y = \mu + \epsilon, \]

where \( \mu \in \mathbb{R}^p \) or \( \beta \), and \( \epsilon \) is a random variable with a continuous distribution. Note:

\[ F(Y < y) = \Phi \left( \frac{y - \mu}{\sigma} \right) = \Phi \left( y^* \right), \]

where \( \Phi \) is the distribution function of \( \epsilon \). If \( \epsilon \) is a continuous distribution, then \( \Phi^{-1} \) exists. The quantile function \( \Phi^{-1}(p) \) of \( \epsilon \) satisfies \( F(Y < \Phi^{-1}(p)) = p \), and we can solve this equation to get:

\[ \Phi^{-1}(p) = \frac{y - \mu}{\sigma}. \]

Thus, for a known \( \Phi^{-1} \), we get from the estimator \( \hat{\epsilon} \) of \( \epsilon \) and \( \hat{\mu} \) of \( \mu \), the estimate:

\[ \hat{\Phi}^{-1}(p) = \frac{\hat{y} - \hat{\mu}}{\sigma}. \]  (1.288)

Standardized Nonlinear Model

We can also use a quantile for a conditional quantile in the heteroscedastic normal design model (1.105), we can use:

\[ \hat{\Phi}^{-1} \left( \frac{X}{1 - \alpha} \right) \]

Similarly, for the conditional heteroscedastic model (1.106) we can use:

\[ \hat{\Phi}^{-1} \left( \frac{X}{1 - \alpha} \right) \]

We apply in Section 2.5.1 model in Section 3.11.3 theory quantile estimators which are based in the standard deviation estimates:

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

\[ \hat{\Phi}^{-1}(Y | X) = \hat{\Phi}^{-1}(\hat{y}). \]  (1.301)

where \( \hat{\Phi} \) is the distribution function of the standard normal distribution.

2. Second estimator uses the \( t \) distribution, which gives the quantile estimate:

\[ \hat{\Phi}^{-1}(Y | X) = \hat{\Phi}^{-1}(\hat{y}^t). \]  (1.302)

where \( \hat{\Phi} \) is the distribution function of the \( t \) distribution, with \( \nu \) degrees of freedom. If \( Y = \hat{\Phi} \), then \( \Phi(X) = t(\nu, \cdot, \cdot) \), so that \( \sqrt{\nu} \Phi(Y | X) \)

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

\[ \hat{\Phi}^{-1}(Y | X) = \hat{\Phi}^{-1} \hat{\Phi}^{-1}(\hat{y}^e). \]  (1.303)

Where \( \hat{\Phi}^{-1} \hat{\Phi}^{-1}(\hat{y}^e) \) is the empirical quantile of the residuals \( Y^e | \hat{\Phi}^{-1} \). Empirical quantiles were estimated in (1.25). These estimator were suggested in Moon et al. (2003).
The expected shortfall is a measure of risk which aggregates all quantiles in the right tail (or the left tail). The expected shortfall for the right tail is defined as:

$$\mathbb{ES}_p(Y) := \frac{1}{p} \int_p^1 Q_p(Y) \, dw, \quad 0 < p < 1.$$ 

When $Y$ has a continuous distribution function, then:

$$\mathbb{ES}_p(Y) = F(Y \mid Y > Q_p(Y)) := \frac{1}{p} \int_1^p F(Y \mid Y > Q_p(Y)) \, dw, \quad 0 < p < 1. \quad (1.3)$$

as defined by McNeil, Frey & Embrechts (2005, Lemma 2.16). We have defined the loss in (1.33) as the negative of the change in the value of the portfolio, and thus the risk management want to control the right tails of the loss distribution. However, we can define the expected shortfall for the left tail as

$$\tilde{\mathbb{ES}}_p(Y) := \frac{1}{p} \int_0^p Q_p(Y) \, dw, \quad 0 < p < 1. \quad (1.33)$$

When $Y$ has a continuous distribution function, then:

$$\tilde{\mathbb{ES}}_p(Y) = F(Y \mid Y < Q_p(Y)) = \frac{1}{p} \int_0^p F(Y \mid Y < Q_p(Y)) \, dw.$$ 

This expression shows that in the case of a continuous distribution function, $\tilde{\mathbb{ES}}_p(Y)$ is equal to the expectation which is taken only over the left tail, where 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 by (1.33) by using:

$$\hat{\mathbb{ES}}_p(Y) = \frac{1}{n} \sum_{i=1}^n Y_i,$$

where $Y_1, \ldots, Y_n$ and $n := \lfloor (1 - p) n \rfloor$. When the expected shortfall is given by (1.35), then we estimate:

$$\tilde{\mathbb{ES}}_p(Y) = \frac{1}{n} \sum_{i=1}^n Y_i,$$

where $n := \lfloor p n \rfloor$.

Let us consider the location-scale model:

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

where $\mu \in \mathbb{R}$, or $\sigma > 0$, and $\epsilon$ is a random variable with a continuous distribution. Notice:

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

²Sometimes the expected shortfall for the left tail is defined as $Q_p(Y) := \mathbb{E}[Y \mid Y \leq \epsilon, Q_p(Y)]$ and the absolute shortfall is defined as $\mathbb{E}[\epsilon \mid \epsilon \leq Q_p(Y)]$. 
Thus the cumulative expected shortfall can be obtained as

$$ES_p(X) = \hat{\pi} \cdot \beta \cdot 1_{\mathbb{P}(\beta)}(\pi)$$

where \( \hat{\pi} \) is an estimate of \( \pi \) and \( \beta \) is an estimate of \( \beta \).

Let \( X \sim \mathcal{N}(0, 1) \) and the expected shortfall is obtained for the right tail as in (1.34). Then,

$$ES_p(c) = 1_{\mathbb{P}(\beta)}(\pi)$$

where \( \phi \) is the density function of the standard normal distribution and \( \Phi \) is the distribution function of the standard normal distribution. Let \( X \sim \chi^2_n \), where \( \chi^2_n \) is the \( \chi^2 \) distribution with \( n \) degrees of freedom, and the expected shortfall is obtained for the right tail as in (1.34), then

$$ES_p(c) = g_{\chi^2_n}(\phi^{-1}(\pi)) \cdot 1 - (\phi^{-1}(\pi))^2 \cdot 1_{\mathbb{P}(\chi^2_n)}(\pi)$$

where \( g_{\chi^2_n} \) is the density function of the \( \chi^2 \) distribution with \( n \) degrees of freedom and \( \phi^{-1} \) is the distribution function of the standard normal distribution with \( n \) degrees of freedom.

Expected shortfall is sometimes preferred to the quantiles as the prudent that the expected shortfall satisfies the axioms of subadditivity. Risk measure \( \rho \) is said to be subadditive if \( \rho(X + Y) < \rho(X) + \rho(Y) \), where \( X \) and \( Y \) are random variables interpreted as portfolio losses. Quantiles do not satisfy subadditivity like the expected shortfall. Thus other axioms of a coherent risk measure are the monotonicity: if \( X > Y \), then \( \rho(X) > \rho(Y) \); the positive homogeneity: for \( \lambda > 0 \), \( \rho(\lambda X) = \lambda \rho(X) \); and the translation invariance: for \( \alpha > 0 \), \( \rho(X + \alpha) = \rho(X) + \alpha \). For more about coherent risk measures, see McNeil et al. (2005, Section 6.1).

### 1.1.7 Approximation and Risk Measures

We have defined the regression function in (1.23) as the conditional expectation of the response variables. The conditional expectation can be viewed as an approximation of response variables \( Y \sim \mathcal{R} \), with the help of explanatory random variables \( X_1, \ldots, X_n \). The approximation is a random variable \( \hat{f}(X_1, \ldots, X_n) \sim \mathcal{R} \), where \( \hat{f} : \mathbb{R}^d \to \mathbb{R} \), is a fixed function. This viewpoint leads to generalization. The best approximations of the response variables can be obtained using various loss functions \( \rho : \mathbb{R} \to \mathbb{R} \). The best approximation is \( \hat{f}(X_1, \ldots, X_n) \), where \( f \) is defined as

$$f = \arg\min_{g \in G} \mathbb{E}(\rho(X - g(X)))$$

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

#### Examples of Risk Measures

We give examples of different classes of \( \rho \) and \( G \).
1. When \( p(x) : \mathbb{R}^d \to \mathbb{R}^2 \) and \( f \) is in the class of all measurable functions \( \mathbb{R}^d \to \mathbb{R}^2 
exists f \), then \( f \) s. th. \( \mathbb{E}[Y | X = x] = \arg\min_{g \in \mathcal{G}} \mathbb{E}[Y - g(X)]^2 \).

Indeed,

\[
\mathbb{E}[g(X) - Y]^2 = \mathbb{E}[(g(X) - \mathbb{E}[Y | X])^2] + \mathbb{E}[(\mathbb{E}[Y | X] - Y)^2],
\]

because \( \mathbb{E}[g(X) - Y] = \mathbb{E}[g(X) - \mathbb{E}[g(X) | X]] = 0 \), and thus \( \mathbb{E}[g(X) - Y]^2 \)

is minimized with respect to \( g \) if and only if \( g(x) : = \mathbb{E}[Y | X = x] \).

Note also that the expectation \( \mathbb{E}[Y] \) is the least constant approximation of \( Y \).

That is, if we choose \( \mathcal{G} \) as the class of constant functions

\[
\mathcal{G} = \{ g : \mathbb{R}^d \to \mathbb{R}, \ g(x) = \mu \text{ for all } x \in \mathbb{R}^d, \ \mu \in \mathbb{R} \},
\]

then

\[
\mathbb{E}[Y] = \arg\min_{\mu \in \mathbb{R}} \mathbb{E}[(Y - \mu)^2] = \arg\min_{\mu \in \mathbb{R}} \mathbb{E}[Y - \mu]^2.
\]

Indeed,

\[
\mathbb{E}[(Y - \mu)^2] = \mathbb{E}[(Y - \mathbb{E}[Y] + \mathbb{E}[Y] - \mu)^2] = \mathbb{E}[(Y - \mathbb{E}[Y])^2 + (Y - \mu)^2] - 2 \mathbb{E}[(Y - \mathbb{E}[Y])(Y - \mu)],
\]

and this is minimized with respect to \( \mu \) if \( \mu = \mathbb{E}[Y] \).

2. When \( p(x) : \mathbb{R}^d \to \mathbb{R}^2 \) and \( f \) is in the class of all measurable functions \( \mathbb{R}^d \to \mathbb{R}^2 \), then \( f \) s. th. \( \mathbb{E}[Y | X = x] = \arg\min_{g \in \mathcal{G}} \mathbb{E}[Y - g(X)] \),

where the conditional expectation is defined in (1.39). Equation (1.39) is proven in the next lemma.

3. When \( p \) is defined as

\[
p_p(z) = \begin{cases} p'(z), & \text{if } 0 < p < 1 \\ 0, & \text{if } p = 0 \\ \infty, & \text{if } p = 1 \end{cases},
\]

for \( 0 < p < 1 \) and \( f \) is in the class of all measurable functions, then the best approximation is the conditional quantile. Figure 1.2 shows the best function for \( (1.40) \) with \( p = 0.5 \) (black line) and with \( p = 0.1 \) (red line). We choose that if the distribution function \( F_p \) is strictly increasing, then

\[
Q_p(Y) = \arg\min_{g \in \mathcal{G}} \mathbb{E}[p(Y) - g(Y)],
\]

3Note that the conditional expectation defined as \( f(x) : = \mathbb{E}[Y | X = x] \) is a real-valued function of \( x \), but \( X(Y | Y) \) is a real-valued random variable which can be defined as \( F(X | Y) = f(X) \).
To show (1.41), note that
\[
W_{\rho_p}(y - \theta) = (y - \theta) \int_{-\infty}^{y} (y' - \theta') d\rho_{\rho_p}(y') + y \int_{y}^{\infty} (y' - \theta') d\rho_{\rho_p}(y')
\]
and thus
\[
\frac{\partial}{\partial \theta} W_{\rho_p}(y - \theta) = (1 - y) \int_{-\infty}^{y} d\rho_{\rho_p}(y') - y \int_{y}^{\infty} d\rho_{\rho_p}(y') = \frac{1}{2} \delta_{\rho_p}(\theta - y).
\]
Setting \(\partial W_{\rho_p}(y - \theta) / \partial \theta = 0\), we get (1.41), where \(W_{\rho_p}\) is strictly monotonous.

We can prove similarly the case of conditional quantiles:
\[
Q_p(Y \mid X = x) = \underset{y \in \mathbb{R}}{\text{arg min}} \int_{-\infty}^{y} \delta_{\rho_p}(y' - y) d\rho_{\rho_p}(y'),
\]
where \(\mathcal{G}\) is the class of measurable functions \(\mathbb{R}^d \to \mathbb{R}\). When \(p = 0.5\), then
\[
\rho_{0.5}(y) = \frac{1}{2} \|y\|,
\]
and we have proved the result (1.39).

A fundamental theorem (2nd phase) Let \(\mathcal{G}\) be a minimizer of the 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 3.
For example, conditional expectation \( f(x) \) of \( Y \mid X = x \) can be estimated by minimizing the sum of squared errors:

\[
\hat{f} = \argmin_{f \in \mathcal{F}} \sum_{x \in \mathcal{X}} \mathbb{E}(Y - f(X)) \quad \text{subject to} \quad f \in \mathcal{F},
\]

where \( \mathcal{F} \) is a class of functions \( f: \mathbb{R}^d \to \mathbb{R} \). 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 \( \alpha \)-quantile is

\[
\hat{Q}_{\alpha,Y}(X) := \arg\min_{\alpha \in [0,1]} \sum_{x \in \mathcal{X}} \rho_{\alpha}(Y - f(X))
\]

and the estimator of the \( \alpha \)-conditional quantile \( f(x) \) is \( \hat{Q}_{\alpha,Y}(Y \mid X = x) \) in

\[
\hat{f} = \argmin_{f \in \mathcal{F}} \sum_{x \in \mathcal{X}} \mathbb{E}(Y - f(X))
\]

where \( \mathcal{F} \) is a class of functions \( f: \mathbb{R}^d \to \mathbb{R} \). A further idea which we will discuss in Section 3.2 is to define an estimator for the conditional quantile using local empirical risk:

\[
\hat{f}(x) = \arg\min_{f \in \mathcal{F}} \sum_{x \in \mathcal{X}} \rho_{\alpha}(Y - f(X))
\]

where \( \rho_{\alpha}(x) > 0 \) and \( \sum_{x \in \mathcal{X}} \rho_{\alpha}(x) = 1 \). These weights should have the property that \( \rho_{\alpha}(x) \) is large when \( X \) is close to \( x \) and \( \rho_{\alpha}(x) \) is small when \( X \) is far away from \( x \).

1.1.3 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 Functions. The distribution function of random variable \( Y \in \mathbb{R} \) is defined as

\[
F_Y(y) = P(Y \leq y), \quad y \in \mathbb{R}.
\]

The conditional distribution function is defined as

\[
F_{Y \mid X}(y \mid x) = P(Y \leq y \mid X = x), \quad y \in \mathbb{R}, \quad x \in \mathbb{R}^d.
\]

\[4\text{This definition can be extended to the multivariate case } Y : \{y_1, \ldots, y_n\} \text{ by :}

\[
P_Y(y) = P(Y_1 \leq y_1, \ldots, Y_n \leq y_n), \quad y = (y_1, \ldots, y_n) \in \mathbb{R}^d.
\]
where $Y \in \mathbb{R}$ is a scalar random variable and $X : \mathbb{R}^d \to \mathbb{R}^n$ is a random vector. We have

$$K_Y | \mathbf{x} = \mathbf{y} = K_Y \left[ \mathbf{x} = \mathbf{y} \right] \quad : \quad \mathcal{L}$$

and thus the estimation of the conditional distribution function can be considered as a regression problem, where the conditional expectation of the random variable $f_Y \left( \mathbf{x} = \mathbf{y} \right)$ is estimated. This random variable $f_Y \left( \mathbf{x} = \mathbf{y} \right)$ 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

$$f_Y \left( \mathbf{y} \right) = \frac{1}{n} \sum_{i=1}^{n} I_\mathcal{A} \left( Y_i < \mathbf{y} \right), \quad \mathcal{A} = \left\{ i : Y_i < \mathbf{y}, i = 1, \ldots, n \right\}, \quad : \quad \mathcal{L}$$

where $\mathcal{A}$ means the continuity of set $\mathcal{A}$. The conditional distribution function estimation is considered in Section 3.5, where local averaging estimations are defined.

**Conditional density**

The conditional density function is defined as

$$f_Y | \mathbf{x} = \mathbf{y} : \begin{cases} f_{Y \mid X} (x, y), & \text{if } f_X (x) > 0, \\ 0, & \text{otherwise,} \end{cases}$$

for $y \in \mathbb{R}$, where $f_{X,Y} : \mathbb{R}^{d+1} \to \mathbb{R}$ is the joint density of $(X, Y)$ and $f_X : \mathbb{R}^d \to \mathbb{R}$ 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 estimations $\hat{f}_{X,Y}$ and $\hat{f}_X$ and define

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

for $\hat{f}_X (x) > 0$. This approach is closer to the approach used in Section 3.5, where local averaging estimations of the conditional density are discussed.

Second, empirical risk minimization can be used in the estimation of the conditional density, as explained in Section 3.5.1.3.

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

$$f_{Y \mid X = \mathbf{y}} = f_{Y \mid \theta} (y), \quad : \quad \mathcal{L}$$

where $f_{Y \mid \theta} : \mathcal{A} \in \mathbb{R}_+$, is a family of density functions and $\theta : \mathbb{R}^d \to \mathcal{A}$, where $\mathcal{A} > 1$. Then the estimation of the conditional density reduces to the estimation of the "regression function" $\theta$. The mean regression is a special case of this approach when the distribution of errors is known. Assume that

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

where $\epsilon$ is independent of $X$, $M_{\epsilon} : 0$, and the density of $\epsilon$ is denoted by $f_{\epsilon}$. Then

$$f_{Y \mid X = \mathbf{y}} = f_{\epsilon} (y - f(X)). \quad : \quad \mathcal{L}$$
which is a special case of (1.44), where we take $f(y) : = f_{\theta}(y - \theta)$ and $g(x) : = f(x)$.

The case of heteroscedastic variance is an other example: Now we assume that

$$Y : = f(X) + \sigma(X)c,$$

where $c$ is independent of $X$, $\mathbb{E}c : = 0$, and the density of $c$ is denoted by $f_c$. Then

$$f_Y|X = x(y) : = f(x) f_c((y - f(x))/\sigma(x)),$$

which is a special case of (1.44), where we take $\theta : = \theta_1, \theta_2, f_{\theta}(y) : = \sigma_1 \sigma_2^{-1} f_c((y - \theta_1)/\theta_2)$, and $g(x) : = (f(x), \sigma(x))$. This approach is used in parametric family regression, as explained in Section 1.3.1.

### 1.1.2 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 $X \in \mathbb{R}^d$ is the explanatory variable and $Y \in \mathbb{R}$ is the response variable, as we saw in (1.4). However, we can use regression methods with time series data

$$Z_1, \ldots, Z_n \in \mathbb{R}^d$$

where the observation $Z_k$ is made at time $k$, $k = 1, \ldots, n$. 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 and the explanatory variables. We consider two ways for the choice of the explanatory variables and 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.

#### White Space Prediction

In the state space prediction the auto-regression parameter $k \geq 1$ is chosen and we denote

$$X_k : = X_{k+1} \ldots X_k : = (X_{k+1}, \ldots, X_{k+1}), \quad k \geq 1$$

Let the time series $Z_1, \ldots, Z_n$ in stationary, then the sequence $(Z_1, Z_2), \ldots, (Z_1, Z_2), \ldots, (Z_1, Z_2)$ consists of identically distributed random variables and the sequence $\mathcal{X}$ is a random vector which is identically distributed as $X$. Let the regression function, as previously defined, by

$$f(x) : = f(Y \mid X = x), \quad x \in \mathbb{R}^d.$$  

We can estimate the regression function using data $(X_k, Y_k), \ k = 1, \ldots, n$, estimation of the regression function $f : \mathbb{R}^d \to \mathbb{R}$ can be used to predict or explain the next outcome of the time series using previous observations. For example, let $\hat{f}$ be an estimator of the regression function and times $X$ constructed using data $(X_k, Y_k), \ k = 1, \ldots, n$, the prediction of the next outcome is $\hat{f}(X_n)$, where $X_n : = (X_n, \ldots, X_{n+1})$:  

$$\hat{Z}_1, \ldots, \hat{Z}_n \in \mathbb{R}^d.$$
the use of multivariate vector time series. Definitions (1.47) pertain only to the study of vector time series. Define

\[ Y_t = g(X_{t-1}), \quad X_t = (X_{t-1}, \ldots, X_{t-k+1}) \quad (1.47) \]

\[ x =: k, \ldots, x \in \mathbb{R}^d, \text{ where } g: \mathbb{R}^d \rightarrow \mathbb{R} \text{ is a function with small values. We define the regression function, as previously, by} \]

\[ f(x) = P(Y_t | X_t = x), \quad x \in \mathbb{R}^d. \]

The regression function is hence defined on the higher-dimensional space of dimension \( (k + d) \).

We can predict and explain without any previous regression parameters 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 approaches. Let \( X_1, \ldots, X_T < \mathbb{R}^d \) be a scalar time series and define

\[ Y_t = Y_{t-1}, \quad X_t = (X_{t-1}, \ldots, X_{t-k+1}) \]

\[ x =: k, \ldots, x \in \mathbb{R}^d, \text{ where } Y_t = (Y_{t-1}, \ldots, Y_{t-k+1}) \text{ is not a sequence of identically distributed random vectors. For example, the regression function } f(x) = P(Y_t | X_t = x), x \in \mathbb{R}^d \text{ is defined in a different space for each } t. \]

When \( Y_t \) are time series, the time series prediction is taken as the explanatory variable, in contrast to (1.47), where the previous observations in the time series were taken as the explanatory variables. We denote

\[ Y_t = Y_{t-1}, \quad X_t = x, \quad t = 1, \ldots, T. \quad (1.48) \]

The multivariate regression model is a general multivariate regression model, as described in Section 1.1.1.4.

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

\[ Y_t = \mu_t + c_t \epsilon_t, \quad t = 1, \ldots, T. \quad (1.49) \]

where \( \mu_t \in \mathbb{R} \) is the deterministic signal, \( \epsilon_t > 0 \) are observation values, and the noise \( c_t \) is stationary with mean zero and small variance. For stationary estimation and asymptotic analysis we can use a slightly different model:

\[ Y_{t, \epsilon_t} = \mu_t + c_t \epsilon_t, \quad t = 1, \ldots, T. \quad (1.50) \]

where \( \epsilon_t \in \mathbb{R}^d, \epsilon_t = 0 \) for the deterministic signal, \( \epsilon_t > 0 \) are observation values, and the noise \( c_t \) is stationary with mean zero and small variance. From these we have the theory that the observed time sequence is stationary with a stationary time series process \( \epsilon_t \), \( t \in \mathbb{R} \), and the stationary sequence is stationary time series process in combination with \( Y_{t, \epsilon_t} = \epsilon_t + \epsilon_t, t = 1, \ldots, T. \) The asymptotics are \( \epsilon_t \rightarrow 0 \) for 

}
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 between these stochastic control problems and portfolio selection are explained in Section 1.5.3 and in Section 1.5.4, respectively.

Consider the time series

\[ \mathcal{X}_{t} = \mathcal{X}_{t-1} + \cdots + \mathcal{X}_{1}, \quad t \in \mathbb{N}, \]

and a random variable \( \mathcal{X}_t \in \mathcal{X} \). We now write an objective function \( f(e) \) for \( e = e_{-1}, \ldots, \mathcal{X}_{t} \) and a constraint function \( \mathcal{X}_{t-1} \in \mathcal{X} \). We want to choose these coefficients in such a way that the mean squared error

\[ \text{MSE}(\sigma_o, \beta_0, \ldots, \beta_{r-1}) = E(\sigma_o + \beta_0 \mathcal{X}_{t} + \cdots + \beta_{r-1} \mathcal{X}_{t-r+1})^2 \]

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

\[ (\hat{\sigma}_{t_0}, \hat{\beta}_{t_0}) = \arg \min_{\sigma_o, \beta_0, \ldots, \beta_{r-1}} \text{MSE}(\sigma_o, \beta_0, \ldots, \beta_{r-1}) \quad (1.5.1) \]

where the minimization is done over coefficients \( \beta_k \) at times \( k \) and over coefficient \( \sigma_o \) at time \( t_0 \).

Note that at time \( t_0 \) the coefficients \( \beta_1, \ldots, \beta_{r-1} \) and \( \sigma_o \) are unknown coefficients since they are chosen at later times, and at time \( t_0 \) we can only calculate the optimal values \( \hat{\sigma}_{t_0} \) and \( \hat{\beta}_{t_0} \). When at times \( t_0+1 \), we choose parameters \( \sigma_{t_0+1} \) and \( \beta_{t_0+1} \) and parameters \( \beta_1, \ldots, \beta_{r-1} \) are chosen at times \( t_0 + 2 \).

Note that the problem in the usual least squares problem is identical, but the usual least squares problem is easier to solve the problem

\[ \min_{\sigma_o, \beta_0, \ldots, \beta_{r-1}} \text{MSE}(\sigma_o, \beta_0, \ldots, \beta_{r-1}) \]

at times \( t_0 + 1 \). When \( e_k \) are all unknown at the same time \( \mathcal{X}_t \), \( e_k \) and \( \mathcal{X}_t \) are independent. This problem appears for examples in the linear autoregressive, where we minimize the expected squared error

\[ \mathcal{E}(\sigma_o + \beta_0 \mathcal{X}_{t_0} + \cdots + \beta_{r-1} \mathcal{X}_{t_0-r+1})^2 \]

at times \( t_0 + 1 \). In the one-step case the stochastic control and the usual least squares problem are identical, because in the one-step problem we minimize

\[ \text{MSE}(\sigma_o, \beta_0, \mathcal{X}_{t_0} + \cdots + \mathcal{X}_{t_0-r+1})^2 = \mathcal{E}(\sigma_o + \beta_0 \mathcal{X}_{t_0} + \cdots + \mathcal{X}_{t_0-r+1})^2 \]

at times \( t_0 + 1 \).
where

\[ R_{\text{min}}(\alpha_{0}, \beta_{0}, \ldots, \beta_{m-1}) \]

\[ = \sum \left( \alpha_{0} - \beta_{m} X_{0}^{2} + \cdots + \beta_{2} X_{2} X_{2} + \cdots + \beta_{m-1} X_{m-1}^{2} + \cdots x_{n}^{2} \right)^{2} \]

The connection of this type of stochastic control problem to optimal pricing is explained in Section 1.2.4.

1.1.1.1 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 case where a fixed set of instrumental variables must be used: when there are relevant explanatory variables which are not observed (covariate variables), when the explanatory variable is subject to measurement error, 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 functions \( g : \mathbb{R}^{2} \rightarrow \mathbb{R} \) in the model

\[ Y : = g(X) + U. \]

where \( Y \in \mathbb{R}, X \in \mathbb{R}^{2}, \) and

\[ N\{U \mid X\} = 0. \]

Now \( g(x) \) is not the conditional expectation \( E[Y \mid X = x] \). Extension of \( g \) is possible when we have observations \( (X_{i}, Y_{i}, \tilde{X}_{i}) \), \( i = 1, \ldots, n \), where \( (X_{i}, Y_{i}) \) are
distributed as \((X, Y)\) and \(X_0\) are observations from the distribution of an instrumental variable \(Z \in \mathbb{R}^d\) that satisfies

\[
Y(z) | Z = z \sim P(z).
\]

(1.54)

We give two examples of model (1.53). The first example explains how an omitted variable can lead to (1.53). This 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:

\[
X = \begin{cases} 
0, & \text{if patient receives treatment } A, \\
1, & \text{if patient receives treatment } B,
\end{cases}
\]

and \(Y\) is a variable measuring the health of the patient after receiving the treatment. This example is an extended version taken from Hallgren, Robins & Wasserman (2004). 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, the doctor will most likely give a treatment that is physiological demanding). Under usual regression techniques 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 this estimate would give a pessimistic estimate of the effect of treatment \(A\).

We have three approaches to estimate the causal 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 \(E(Y | X = 0, W = w)\). However, in this example we have not observed \(W\), so the estimation of this conditional expectation is not possible. (3) We can use the instrumental variable technique. In this example the instrumental variable \(Z\) can be chosen as the difference between the shortest distance from a patients house to a hospital giving treatment \(A\) and the shortest distance from a patients house to a hospital giving treatment \(B\). Variables \(X\) 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. Variables \(X\) does not have any influence on the health of patients, so it is otherwise called a confounding variable, influencing only \(X\). These we can use \(Z\) to remove 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 \(Y|X = x, W = w) \sim \theta_0, f_1(X) \sim 0, \) and \(f_2(W) \sim 0\). We have observations \((Y_i, X_i, W_i), i = 1, \ldots, n\), 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 \( X \) on \( Y \).

Denoting \( g(X) = \beta X + U \) and \( O = f_2(Y_i) + \epsilon_i \), we have that

\[ Y_i = g(X) + U_i \]

where \( \beta \neq 0 \), because \( \text{Cov}(X, U_i) = 0 \), and \( \text{Var}(Y_i) = 0 \), because \( X \) is exogenous to the system, bearing influence only on \( X \). Thus we are in the setting of model (1.53):

**Known the Multivariates Linear Model** As an example of model (1.53), consider the case where the linear model:

\[ Y = \alpha + \beta X + \epsilon \]

holds. However, the explanatory variable \( X \) is not observed directly but we observe only pairs \((X_{i1}, X_{i2})\), \( i = 1, \ldots, n \), where:

\[ X_{i1} = X_{i2} + e_i, \quad i = 1, \ldots, n. \]

Then the observed values \( X_{i2} \) may be contaminated with unobserved errors. We assume that:

\[ \text{Cov}(X_{i2}, U_i) = 0, \quad \text{Cov}(U_i, \epsilon_i) = 0 \]  
(1.55)

and

\[ \text{Cov}(X_{i2}, \epsilon_i) = 0. \]  
(1.56)

We can write the observed response variables as:

\[ Y = \alpha + \beta X + \epsilon \]

and the new covariances determined by:

\[ U = U_i + \beta \epsilon_i \]

To get the new linear model:

\[ Y = \alpha + \beta X + U_i \]  
(1.57)

In this new linear model, \( \text{Cov}(Y_i) \neq 0 \). Thus we have the same situation as in (1.53), with \( g(X) = \beta X + U \).

The fact \( \text{Cov}(U_i) \neq 0 \) follows from \( \text{Cov}(X, U_i) \neq 0 \). We have that:

\[ \text{Cov}(X, U_i) = \beta \text{Cov}(X, \epsilon_i) \]

because:

\[ \text{Cov}(X, U_i) = \beta \text{Cov}(X, \epsilon_i) + \text{Cov}(\epsilon_i, \epsilon_i) = 0 \]

by assumption (1.53) and \( \text{Cov}(X, \epsilon_i) = 0 \) by assumption (1.56).


1. Introduction of the Structural Equation Model

We give a linear instrumental variable estimation in (1.24). This estimation can be used to estimate parameters \( \alpha \) and \( \beta \) in (1.27). The linear instrumental variable estimation is:

\[
\tilde{\beta} = \frac{\sum_{i=1}^{n} X_i \cdot Y_i}{\sum_{i=1}^{n} X_i^2}, \quad \tilde{\alpha} = Y \cdot \tilde{\beta} X,
\]

where:

\[
X \cdot \frac{1}{n} \sum_{i=1}^{n} X_{i}, \quad \frac{1}{n} \sum_{i=1}^{n} X_{i} \cdot Y, \quad \frac{1}{n} \sum_{i=1}^{n} X_{i}^2.
\]

Hall & Hsiao's (2015) approach the estimation of \( g(x) \) in the model (1.30) by deriving an operator equation for \( g \). Hence (1.34) we obtain:

\[
K(Y \mid X) = K(g(X) \mid X), \quad K(U \mid X) = K(g)(X),
\]

where the operator \( K \) is defined as:

\[
(Kg)(x) = K(g(X) \mid X = x) = \int_{k} f_{X \mid x} g(x) dx.
\]

This operator \( K \) is an integral operator mapping \( L^2_{\mathbb{R}} \) to \( L^2_{\mathbb{R}} \) with \( K(g)(X) \) < oo. By estimating \( K \) and estimating \( g \) we can find an estimator for \( g \).

1.2 DISCRETE RESPONSE VARIABLES

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 \( Y \sim \text{Bernoulli}(\theta) \), where \( 0 \leq \theta \leq 1 \), then the probability mass function of \( Y \) is:

\[
f_Y(y) = \theta^y(1 - \theta)^{1-y}, \quad y \in \{0, 1\}.
\]

Now we can construct a model for the conditional distribution of \( Y \) given \( X \) as:

\[
f_{Y \mid X}(y) = f(y \mid \theta(x)) = \theta(x)^y(1 - \theta(x))^{1-y}, \quad y \in \{0, 1\}, \quad x \in \mathbb{R}^d.
\]
where $p : \mathbb{R}^d \rightarrow [0, 1]$ is a function. Note that in the Bernoulli model $p_Y \rightarrow p$ and in the conditional Bernoulli model (1.35) 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 $p$. Moreover, it can happen that a regression function estimate $\hat{p}(x) = \hat{g}(x)$ 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
\[ p(x) : = g(x), \]
where $g : \mathbb{R}^d \rightarrow [0, 1]$ 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 simple limit model it is assumed that
\[ p(x) : = \phi(x), \]
where $\phi : \mathbb{R}^d \rightarrow [0, 1]$ is an unknown link function. Simple limit estimators are considered in Section 4.3.1.

3. We can estimate $p$ with the help of a density function estimation. If vector $X$ has a continuous distribution, if vector $X$ has a continuous distribution, we can write
\[ p(Y = 1 \mid X = x) = \int \mathbb{I}(Y = 1) f(x \mid Y = 1) \, dx, \]
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
\[ \hat{p}_1 = \mathbb{E}[Y = 1 \mid x_1, \ldots, x_n, Y = 1]. \]

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
\[ \hat{f}(x) = \frac{1}{n} \sum_{i=1}^n K_h(x - x_i), \]
where $K_h(x) = K((x - x_i)/h)$, $K : \mathbb{R}^d \rightarrow \mathbb{R}$ is the kernel function, and $h > 0$ is the smoothing parameter. See (3.35) for the distribution of the kernel density estimation. Finally, we define the estimator of function $p : \mathbb{R}^d \rightarrow [0, 1]$ as
\[ \hat{p}(x) = \frac{\mathbb{E}[Y \mid x]}{\hat{f}(x)}. \]
Theorem 9. We can estimate the function $y = \text{logistic}(\beta x)$ with a linear averaging:

$$\tilde{y}(x) = \sum_{k=1}^n w_k(x) y_k,$$

where the weights $w_k(x)$ satisfy $w_k(x) \geq 0$ and $\sum_{k=1}^n w_k(x) = 1$. Examples of the linear averaging are given in Chapter 3, where regression weights, kernel weights, and nearest-neighborhood weights are detailed. In the case of kernel regression and kernel density estimation (KDE) (1.60) and (1.61) are equivalent, see (3.3.6).

### 1.3.2.3 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 when the values of the response variable are unordered and the cases when 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 $Y \sim \text{Categorical}(p_1, \ldots, p_K)$, where $0 < p_k < 1$, $\sum_{k=1}^K p_k = 1$, then the probability mass function of $Y$ is:

$$f_Y(y) = \sum_{k=1}^K p_k \delta(y, k), \quad y \in \{0, \ldots, K-1\}.$$  

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

$$f_Y|X=x(y) = \sum_{k=1}^K p_k(x) \delta(y, k), \quad y \in \{0, \ldots, K-1\}, \quad x \in \mathbb{R}^d,$$

where $p_k(x) : \mathbb{R}^d \to [0, 1]$ are functions satisfying $\sum_{k=1}^K p_k(x) = 1$ for each $x \in \mathbb{R}^d$.

Note that now the conditional probability of $Y$ given $X$ is:

$$p(y | x) = \delta(y - x), \quad x \in \{0, 1, \ldots, K-1\}.$$

There are several reasonable estimations of $p_k(x)$:

1. We can use the parametric form:

$$p_k(x) = \frac{e^{\theta_k^T x}}{1 + \sum_{k=1}^K e^{\theta_k^T x}}.$$
for \( k = 0, \ldots, K - 1 \), and \( p_k(x) : \ldots \). A naive solution is obtained from

\[
p_k(x) = \frac{e^{\beta_k^T x}}{\sum_{k=0}^{K-1} e^{\beta_k^T x}}
\]

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

\[ P_k = \beta_k^T \mathbf{x} + c_k \]

and

\[ \mathbf{X} = \text{average}_{k=0}^{K-1} P_k. \]

Assume that \( c_k \) are independent and identically distributed with the Student

distribution. The distribution function of the Student distribution is \( \mathcal{S}_0(x) = \exp[-e^{-x}] \). Hence \( p_k(x) = \mathcal{S}(Y = k \mid X = x) \) is given by (1.64). This

estimation can be done with the maximum likelihood or with the least squares

method.

2. We can estimate \( \psi \) with the help of any density function estimator. If \( \mathbf{X} \)

has a continuous distribution, then \( \psi \) can solve

\[
L(Y = k \mid X = x) = \frac{1}{f_k(x)} \prod_{i=1}^{M} f_{X_i}(x_{i,k}),
\]

where \( i = 0, \ldots, M - 1 \), \( f_k(x_{i,k}) \) is the density of \( X_i \mid Y = k \) and \( f_k \)

is the density of \( X_i \). The prior probability \( \mathcal{S}(Y = k) \) can be estimated by

\[
\hat{\mathcal{S}}_k = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(X_i = x, Y_i = k),
\]

where \( \mathbb{1}(A) \) denotes the characteristic function of set \( A \). The densities \( f_k(x_{i,k}) \) and \( f_k \)

can be estimated by any density estimation method. Note (5.23) for the definition

of the kernel density estimator. Finally we obtain the estimation of \( \psi \) as

\[
\hat{\psi}(x) = \hat{\mathcal{S}}_k \hat{f}_k(x) \prod_{i=1}^{M} \hat{f}_{X_i}(x_{i,k})
\]

3. Define \( K \) Bernoulli random variables \( Y^{(1)}, \ldots, Y^{(K - 1)} \) with the definition

that \( Y^{(k)} = 1 \) if and only if \( Y = k \). Then

\[
p_k(x) = \mathcal{S}(Y^{(k)} = X = x).
\]

We can estimate, for example, the Bernoulli regression function \( p_k(x) \), using regression

data \( (X_1, Y^{(1)}_1), \ldots, (X_n, Y^{(K - 1)}_n) \), 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 $Y \sim \text{Poisson}(\lambda)$, then

$$X(Y = y) = e^{-\lambda} \frac{\lambda^y}{y!}, \quad y = \{0, 1, 2, \ldots\},$$

where $\lambda > 0$ is the unknown intensity parameter. Note $E(Y) = \lambda$ and $\text{Var}(Y) = \lambda$. In the Poisson regression, the regression function is

$$\mu(x) = \exp(\beta'x),$$

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

$$\mu(x) = \exp(\beta'x),$$

where $\beta \in \mathbb{R}^d$ is the unknown parameter. This choice guarantees that $\mu(x) > 0$. Mullahy, see Lee & Scutipher (2004) makes a comparative simulation study of weighted shrinkage estimators for Poisson counts.

1.3 Parametric Family Regression Model

We obtain the binary response models, discrete choice models, and Poisson count models, introduced in Section 1.2.3, as special cases of parametric family regression, introduced in Section 1.3.1. In fact, these are special cases of exponential family regression, introduced in Section 1.3.3. A different type of parametric family regression is obtained by copula modeling, introduced in Section 1.3.2.

1.3.1 General Parametric Family

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

$$Y \sim P_\mu(x),$$

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

$$Y \mid X = x \sim P_{\mu(x)}.$$  \hspace{1cm} (1.46.1)

The following examples illustrate this model.
1. We obtain a Gaussian mean regression model where \( X \sim \mathcal{N}(\Theta_0, \sigma^2) \), where \( \Theta_0 \in \Theta \subseteq \mathbb{R}^p \). Now
\[
Y \mid X = x \sim \mathcal{N}(f(x), \sigma^2) ;
\]
which follows from
\[
Y = f(X) + c,
\]
where \( c \sim \mathcal{N}(0, \sigma^2) \).

2. We obtain a Gaussian volatility model, where \( Y \sim \mathcal{N}(0, \Theta) \), where \( \Theta \in \Theta \subseteq (0, \infty) \). Now
\[
Y \mid X = x \sim \mathcal{N}(0, f(x)) ,
\]
which follows from
\[
Y = \frac{f(X)}{f_0(x)} c,
\]
where \( c \sim \mathcal{N}(0, 1) \).

3. We obtain a Gaussian heteroscedastic mean regression model, where \( X \sim \mathcal{N}(\Theta_1, \Theta_2) \), where \( \Theta_1 \in \Theta_1 \subseteq (0, \infty) \) and \( \Theta_2 \in \Theta_2 \subseteq \mathbb{R}^p \). Now
\[
Y \mid X = x \sim \mathcal{N}(f_1(x), f_2(x)) ,
\]
which follows from
\[
Y = f_1(X) + f_2(X) \cdot c ,
\]
where \( c \sim \mathcal{N}(0, 1) \), and we denote \( \Theta \times \mathcal{N}(0, 1) \).

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

Let us assume that the probability measures \( \mathbb{P} \) are dominated by a probability measure, and denote the density functions of \( \mathbb{P} \) by \( p(y, \Theta) \). We use the term density function, although \( p(y, \Theta) \) can also be a probability mass function, if \( Y \) has a discrete distribution. In Section 1.3.2, we make the assumption that \( (\Theta, \Theta \subseteq \Theta) \) 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
\[
\prod_{i=1}^{n} \log p(Y_i, f(X_i)) .
\]
This log-likelihood can be maximized over the collection \( F \) of functions, and we denote
\[
\hat{f} = \arg\max_{f \in F} \sum_{i=1}^{n} \log p(Y_i, f(X_i)) .
\]
where \( (\hat{f}, \beta \in \mathbb{R}^q) \) is a large collection of functions, like the collection of linear functions \( f_{B}(x) = -B_0 - B_1 x_1 - \cdots - B_{p} x_p \).
A second possibility is to maximize a local log-likelihood and define
\[
\hat{f}(x) = \arg \max_{f \in \mathcal{F}} \sum_{x_1} \log \left( \sum_{x_2} \cdots \sum_{x_n} \prod_{i=1}^{n} p_{\theta}(X_i | f(x_i)) \right) p_{\theta}(x),
\]
where \( p_{\theta}(x) \) are weights, for example \( p_{\theta}(x) = K_{\theta}(x - x_0) \), where \( K_{\theta}(x) = \frac{1}{\sqrt{2\pi \sigma^2}} \exp \left( -\frac{x^2}{2\sigma^2} \right) \), and \( \sigma \geq 0 \). This is also called \( \hat{f} \) in the context of maximum likelihood estimation. \( f \) is the local likelihood approach, as discussed in (Speckman, 2010).

### 1.3.2 Exponential Families

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 reference measure. In a one-parameter exponential family the density functions have the form
\[
p(y; \theta) = g(y) \exp \{ y \theta - b(\theta) \},
\]
where \( \theta \in \Theta \subset \mathbb{R} \), and \( y \in Y \subset \mathbb{R} \). The functions \( g \) and \( b \) are non-negative. Function \( g \) and function \( y \) are non-negative. In the exponential family with the canonical parameterization, the density functions are
\[
p(y; \theta) = g(y) \exp \{ y \theta - c(\theta) \},
\]
where \( \theta = c(\theta) \) and \( c(\theta) = b(\theta) \). Examples of exponential families include the family of Gaussian, Poisson, and gamma distributions. The expression of exponential families is given by (Busemeyer, 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 \( \theta = f(x) \):
\[
Y | X = x \sim p(y; f(x)),
\]
where \( f : \mathbb{R}^d \rightarrow \Theta \), where we used the natural parameterization in (1.68), and \( \mathcal{Y} \) is the parameter space of the natural parameters.

If the parameterization is natural, and \( c \) is continuously differentiable, then
\[
M_{\theta} \frac{d}{d\theta} \log p(Y; \theta) = 0,
\]
where \( Y = f(y, \theta) \). Indeed
\[
M_{\theta} \frac{d}{d\theta} \log p(Y; \theta) = 0,
\]
On the other hand, for \( Y = f(y, \theta) \),
\[
M_{\theta} \frac{d}{d\theta} \log p(Y; \theta) = 0,
\]
under regularity assumptions. Then, (1.3.2) holds. Under the assumption (1.3.2), we get

$$P(Y \mid X = x) = e^{f(x)}.$$  

If the parameterization is natural, and if one chooses continuously differentiable, then

$$\Psi_{n}(Y) = \psi^T(v).$$  

Indeed,

$$\Psi_{n}(Y) = \sum_{i=1}^{n} (\hat{Y}_i - \hat{Y}(x)) = \sum_{i=1}^{n} \left[ \frac{\partial}{\partial \beta} \log p(Y_i, v) \right]_v = \sum_{i=1}^{n} \frac{\partial^2}{\partial \beta^2} \log p(Y_i, v) = \psi^T(v).$$

Under the assumption (1.3.2), we get

$$\Psi(Y \mid X = x) = e^{f(x)}.$$  

Therefore, Chan & Kalton (2000) suggested a reduction method where the exponential family regression can be transformed to the Gaussian regression by binomial and variance stabilizing transformation.

1.3.8 Copula Modelling:

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

$$F(Y_1, Y_2) = \mathbb{P}(Y_1 \leq y_1, Y_2 \leq y_2),$$

where \(y_1, y_2 \in \mathbb{R}^2\). We can write the distribution function uniquely as

$$F(y_1, y_2) = C(F_1(y_1), F_2(y_2)), \quad (1.3.2)$$

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

$$C(u_1, u_2) = \mathbb{P}(F_1^{-1}(u_1), F_2^{-1}(u_2)), \quad (1.3.2)$$

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

For example, a Gaussian two-dimensional copula is a normally distributed vector with multivariate normal marginals. The family of Gaussian two-dimensional copula \(C_{\theta}\) has the parameter \(\theta \in (0, 1)\), where \(\theta\) is the correlation coefficient between \(Y_1\) and \(Y_2\).

\(^{9}\)Note that \(F_1, F_2, \log p(Y, v) : 1, \frac{\partial}{\partial \theta} \frac{\partial}{\partial \theta} \log p(Y, v) : 1, \int \frac{\partial}{\partial \theta} \frac{\partial}{\partial \theta} \log p(Y, v) \, dv = \frac{\partial^2}{\partial \theta^2} \int \log p(Y, v) \, dv = \frac{\partial^2}{\partial \theta^2} \log p(Y, v) = \frac{\partial^2}{\partial \theta^2} \log p(Y, v) = 0\), if the order of differentiation and integration can be changed.
The copula representation of the distribution as in (1.12) produces a powerful way to construct models and to estimate the unknown parameters of the model. Let \((\mathcal{C}, \Theta) \in \Theta(\mathcal{C})\) be a family of copula densities, where \(\Theta \subset \mathbb{R}^d\). This leads to a semiparametric model with densities

\[
\mathcal{C}(y_1, y_2; \Theta, f_1, f_2) = \mathcal{C}(\mathcal{F}(y_1), \mathcal{F}(y_2)) f_1(y_1) f_2(y_2),
\]

where \(\Theta \subset \mathcal{C}\) and \(f_1, f_2 \subset \mathcal{F}\), where \(\mathcal{F}\) is a nonparametric collection of univariate density functions. The combination of \(\mathcal{C}, f_1, f_2\) 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 \(\Theta\).

Assume that \(X \subset \mathbb{R}^d\) is a vector of explanatory variables and we want to estimate the conditional distribution \((X_1, X_2) | X = x\). We assume that the conditional distribution function is

\[
\mathcal{N}_{x_1, x_2 | x} = (x_1, x_2) \equiv \{ (x_1(x_1 < x_1, x_2 < x_2) | X = x) \in \mathcal{C} \}
\]

where \(\mathcal{C} : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathcal{N}\). The conditional density is

\[
\mathcal{N}_{x_1, x_2 | x} = (x_1, x_2) \equiv \{ \mathcal{C}(\mathcal{F}(x_1), \mathcal{F}(x_2)); f_1(x_1) f_2(x_2) \mathcal{F}(x_1) \mathcal{F}(x_2) \}
\]

In the first stage we estimate nonparametrically the conditional distribution functions and get the estimates \(\hat{F}_{X_1 | x} \equiv \mathcal{F}(x_1)\) and \(\hat{F}_{X_2 | x} \equiv \mathcal{F}(x_2)\). In the second stage we estimate the function \(\hat{\Theta}(x)\). This can be done sequentially in (1.67), and we get the locally constant likelihood estimates as

\[
\hat{\Theta}(x) = \arg \max_{\Theta \in \Theta} \prod_{i=1}^{d} \mathcal{C}(\mathcal{F}(x_i), \mathcal{F}(x_i)); f_1(x_i) f_2(x_i) \mathcal{F}(x_i) \mathcal{F}(x_i)
\]

This method has been studied in, e.g., Abegaz, Czibulka & Varmuza (2012).

We have discussed in (1.12) the standard copula decomposition. Their decomposition can be inconvenient because the copula density \(\mathcal{C}\) is less support inside \([-1, 1]^d\), and the estimation is typically complicated with the likelihood estimates. Alternatively, we can make the copula decomposition

\[
\mathcal{N}(x_1, x_2) \equiv \mathcal{C}^2 \{ \Phi^{-1}(\mathcal{C}(x_1)), \Phi^{-1}(\mathcal{C}(x_2)) \}
\]

where \(\Phi : \mathbb{R} \rightarrow \mathbb{R}\) is the distribution function of the standard Gaussian distribution. Note \(\mathcal{C}\) is a distribution function where its marginals are standard Gaussian, and \(\mathcal{C}\) is defined by

\[
\mathcal{C}(w, v) \equiv \Phi^{-1}(\mathcal{C}_1^{-1}(\Phi(v)), \mathcal{C}_2^{-1}(\Phi(w))), \quad w, v \in \mathbb{R}
\]

1.4 CLASSIFICATION

Let the sequences \((X_1, Y_1), \ldots, (X_m, Y_m)\) consist of identically distributed random vectors. Let \((X, Y)\) be distributed as \((X_1, Y_1)\), for \(i = 1, \ldots, m\). Let the possible
values of $Y$ lie in $\{0, \ldots, K - 1\}$. We want to employ a classification function $g : \mathbb{R}^d \rightarrow \{0, \ldots, K - 1\}$. The classification function is interpreted as such function that if we observe a new random variable $X_{new}$ distributed as $X$, then $g(X_{new})$ guesses the class label of $X_{new}$, that is, we decide that $X_{new}$ comes from the distribution of $X$ if $Y = k$, if $g(X_{new}) = k$.

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

$$f(x) = E(Y \mid X = x) = \mathbb{P}(Y = 1 \mid X = x). \quad (1.3.13)$$

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

We have assumed that classification is 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 $X_{new} : \mathbb{R}^d$ are coming from the $k$th distribution, $k = 0, \ldots, K - 1$. We will consider only the case where the class frequencies are random.

### 1.2.1.2 Bayes Rule

In the model the 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 $J_f = \mathbb{E}\left[ (Y - f(X))^2 \right]$ and Section 1.1.7. Similarly, in the case of classification, we can find a population quantity which minimizes a natural criterion. This quantity is the probability of misclassification, or Bayes risk:

$$J(g) = \mathbb{E}\left[ g(X) \neq Y \right]. \quad (1.3.14)$$

Let

$$g^* = \arg \min_{g} J(g),$$

which gives the minimization in class $f$ for all classification functions $g : \mathbb{R}^d \rightarrow \{0, \ldots, K - 1\}$. The classification function $g^*$ which minimizes the probability of misclassification is called the Bayes rule. It can be proved that

$$g^*(x) = \arg \min_{k} \mathbb{P}(Y = k \mid X = x). \quad (1.3.15)$$
The proof of (1.23) for the case \( X^* = 0 \) can be found in Geyer et al. (2009, p. 40). It holds that

\[
g^*(x) = 1 - \arg\max_{\delta \in \Delta} \int \psi(x) f_{X,Y}(X = \delta, Y = x) \, dx.
\]

where \( f_{X,Y}(\cdot, \cdot) : \mathbb{R}^2 \rightarrow \mathbb{R}^+ \) is the density function of \( X \mid Y = \delta \).

1.4.2 Methodology 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 neighbours.

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 \( \{X_1, Y_1\}, \ldots, \{X_m, Y_m\} \) 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 1.4.2.1, we introduced binary response models, and in Section 1.4.2.2 we introduced discrete choice models.

In a discrete choice model, we estimate the class posterior probabilities

\[
g_k(x) = P(Y = k \mid X = x), \quad k = 0, 1, \ldots, K - 1.
\]

A natural classification function is

\[
g^*(x) = 1 - \arg\max_{\delta \in \Omega} \hat{g}_k(x).
\]

In fact, we note in (1.23) that the classification function \( g^* \) is in essence the optimal classification function. Let us denote the estimates of the class posterior probabilities by \( \hat{g}_k(x) \), and let us define an estimator of the classification function by

\[
\hat{g}(x) = 1 - \arg\max_{\delta \in \Omega} \hat{g}_k(x).
\]

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

\[
y_i^{(k)} = 1_{\{y_i = k\}}, \quad k = 0, 1, \ldots, K - 1.
\]

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

\[
\hat{g}_k(x) = P(Y = k \mid X = x) = \psi(x) f_{X,Y}(X = k, Y = x),
\]

(1.30)

Estimating \( \hat{g}_k(x) \) is constructed using regression data \( \{X_1, Y_1^{(k)}\}, \ldots, \{X_m, Y_m^{(k)}\} \), for \( k = 0, 1, \ldots, K - 1 \).

In the two-class case, when \( Y \in \{0, 1\} \), we did not have to use (1.29), because \( Y \) is already a class indicator. In the two-class case, we can write the empirical decision

\[
g^*(x) = \hat{g}_k(x) = 1 - \hat{g}_0(x),
\]

(1.31)
null hypothesis. Let \( \tilde{g} : \mathbb{R}^d \rightarrow \mathbb{R} \) be any regression function estimation constructed using regression function estimates \( (X_1, Y_1), \ldots, (X_m, Y_m) \), and define
\[
\tilde{g}(x) := \begin{cases} 1, & \text{if } \tilde{g}(x) > 1/\Delta, \\ 0, & \text{otherwise}, \end{cases}
\]
(1.31)
which estimates the natural classification function
\[
g(x) := \begin{cases} 1, & \text{if } K^*(Y = 1 | X = x) > K^*(Y = 0 | X = x), \\ 0, & \text{otherwise}, \end{cases}
\]
(1.32)

Classification by Probability Estimation: A classification function can be constructed from density estimation of the class-conditional densities. We assume now that \( X \) is a random vector with a continuous distribution. If we consider the classification rule
\[
g^*(x) := \arg\max_{k \in \mathbb{K}} P_{X,Y}(x),
\]
defined in (1.17), we can write
\[
p_k(x) := P(Y = k | X = x) := \frac{K^*(Y = k | X = x) \tilde{g}(x)}{\tilde{g}(x)},
\]
where \( k = 0, \ldots, K-1 \), and \( x \in \mathbb{R}^d \). Then,
\[
\arg\max_{k \in \mathbb{K}} p_k(x) := \arg\max_{k \in \mathbb{K}} P(Y = k | X = x) \tilde{g}(x).
\]

An estimator for the classification function, based on data \( (X_1, Y_1), \ldots, (X_m, Y_m) \), is obtained as
\[
\hat{g}(x) := \arg\max_{k \in \mathbb{K}} P_{X,Y}(x), \quad \hat{g}(x) := \frac{K^*(Y = k | X = x) \tilde{g}(x)}{\tilde{g}(x)},
\]
(1.33)
where \( \hat{g}(x) \) is the density estimator of the class density function \( g_k(x) \) and \( \hat{g}(x) \) is an estimator of the class prior probability \( P(Y = k) \). We can define
\[
\hat{\gamma}_K := \frac{1}{m} \sum_{i=1}^{m} I(Y_i = k).
\]

Classification by Empirical Risk Minimization: A classification function can be constructed using empirical risk minimization. In (1.19), classification is reduced to regression function estimation (for binary response models or in discrete choice models). In (1.33), classification is reduced to density estimation. However, according to Vapnik's principles, 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
\[
\tilde{g} := \arg\max_{g \in \mathcal{G}, \gamma_n(g)},
\]
where \( \mathcal{G} \) is a class of functions \( g : \mathbb{R}^d \rightarrow \{0, \ldots, K-1\} \) 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
\[
\gamma_n(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}[y_i \neq g(x_i)].
\] (1.3.5)

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
\[
\gamma_n^k(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}[y_i = k \land g(x_i) \neq k],
\]
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:
\[
\gamma_n(g) = \sum_{k=0}^{K-1} m_k \gamma_n^k(g).
\]

For \( m_k = 1 \) we get the overall classification error (1.3.5).

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 classifier \( h : \mathbb{R}^d \to \mathbb{R} \) and define the classification function \( g(x) = \text{sign}(h(x)) \). The empirical risk is defined as
\[
\gamma_n(g) = \frac{1}{n} \sum_{i=1}^{n} \phi(Y_i h(X_i)),
\]
where \( \phi : \mathbb{R} \to (0, +\infty) \) is a convex non-increasing function with \( \phi(a) \geq \frac{1}{a} \phi(1/a) \) for \( a \in \mathbb{R} \). We can take the hinge loss \( \phi(a) = \max(0, 1 - a) \), the exponential loss \( \phi(a) = -\exp(-a) \) or the logit loss \( \phi(a) = \log(1 + \exp(-a)) \). Support vector machines, mentioned in Section 1.3, use the hinge loss and a penalized empirical risk.

An example for the choice of class \( G \) is given in (2.2.4). In this example the class \( G \) is chosen so that the classification functions are linear.

**Classification by Nearest Neighbors**

The nearest neighbor rule defines the class estimate to be the class label that occurs most often among the \( k \) nearest neighbors. That is, for an integer \( k \), \( k \leq 1, 2, \ldots \), define the \( k \) nearest neighbors, based on observations \( (X_1, Y_1), \ldots, (X_n, Y_n) \), as the set
\[
\mathcal{N}_k(x) = \{ Y_i : \|X_i - x\| \leq r_k(x) \},
\]
where
\[
r_k(x) = \min\{ r > 0 : \# \mathcal{N}_k(x) = k \},
\]
where \( \mathcal{N}_r(x) = \{ z \in \mathbb{R}^d : \|z - x\| < r \} \). Now we can define the classifier:
\[
\hat{g}(x) = \arg\max_{y \in \{-1, 1\}, \ldots, K} |\{ Y_i \in \mathcal{N}_k(x) : Y_i = y \}|
\]

As in all (2001, Section 1.3) use the term "prototype methods" to denote classifications which classify the most observations to the class whose observed values are most similar to the most observations.

\( ^6 \)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.
1.3 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 option contracts.

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 conditional regression can be applied in hedge fund replications and in performance measurement. These were also described in Section 1.6.

1.3.1 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 meet the economic or financial needs of a company in a worst-case scenario. The definition of the economic capital can be understood with the concept of a value at risk. The economic capital is used in portfolio selection to calculate return distributions. The regulatory capital is the capital required by the regulations that financial institutions should maintain. The regulatory capital is often defined in terms of value at risk.

Variance-covariance can be used in speculation, but variance-covariance 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 \( V_{t+1} \) the value of the portfolio at a later time, then the distribution of \( V_{t+1} - V_t \) is called the profit and loss distribution for the time period from \( t+1 \) to \( t+1 \). We define the loss as the negative of the change in the value of the portfolio:

\[
L_t = -(V_{t+1} - V_t).
\]

The upper quantiles of the loss distribution are called the value at risk \( \text{VaR}_p \):

\[
\text{VaR}_p = \text{VaR}_p(L_t),
\]

where \( p \) is equal to 0.05 or 0.01, for example. A larger value of \( \text{VaR}_p \) indicates that the portfolio is more risky, because \( \text{VaR}_p \) is such that \( 1 - p \) is the probability that the loss is larger than \( \text{VaR}_p \), in smaller or equal to \( 1 - p \). We can write:

\[
L_t = -V_t + R_t,
\]

where \( R_t \) is the return of the portfolio.

\[
R_t = \frac{V_{t+1} - V_t}{V_t},
\]
Thus, if we have the quantiles $Q_p(Y_{t,n})$ of the return distribution, the VaR at the $p$th percentile is obtained by the formula:

$$\text{VaR}_p = -\frac{1}{p} Q_p(Y_{t,n}).$$

Quantile at 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.

### Investment Grappling with Covariance Risk

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 same, 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 A-rated companies over a one-year period has been roughly one in three thousand. Thus one would choose the 0.000005th quantile of the profit loss distribution (1/3000 = 0.000005 = 0.03%), which would be for example a loss of 1 million Swiss, 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 Blackman (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.

### Sharpe Ratio

The Sharpe ratio of a portfolio is defined as:

$$\text{Sharpe}(\mathcal{R}) = \frac{\mathbb{E}(\mathcal{R}) - r}{\text{std}(\mathcal{R})},$$

where $\mathcal{R}$ is the return of the portfolio for a given time period, $r$ is the return of a risk-free asset for the same time period, and $\text{std}(\mathcal{R})$ 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
in the return of the portfolio minus the return of a risk-free asset.

In portfolio selection, the risk aversion can be taken into account by using the
black-scholes criterion:

$$\kappa \left[ \ell^2 - v^2 \right] + \frac{\lambda}{2} \left[ \rho \ell - \sigma \right],$$  \hspace{1cm} (1.329)

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

$$\text{expected return} - \frac{\lambda}{2} \cdot \text{risk}.$$  \hspace{1cm}

The Sharpe ratio and the black-scholes criterion use the standard deviation of the
expected return as the risk measure. The standard deviation does not take into account
the possibility of a non-symmetric 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
black-scholes criterion. The partial variance is defined in (1.18).

1.5.3. Variance hedging

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

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

$$V_T \in [K \cdot \mathbb{P}(\mathbb{S}_T < K), K \cdot \mathbb{P}(\mathbb{S}_T > K)],$$

at the expiration date $T$, where $K$ is the delivery price, and $\mathbb{P}(\cdot)$ is the realized variance,
defined by

$$\mathbb{V}_T := \sum_{i=1}^{n} \left[ \log \left( \frac{\mathbb{S}_i}{\mathbb{S}_{i-1}} \right) \right]^{2},$$

where $\mathbb{S}_0$ is the starting day of the contract, and $\mathbb{S}_T$ are the prices of a financial asset.

The volatility swap pays at the expiration:

$$\sqrt{\mathbb{V}_T} \cdot |K|,$$

where $K$ is the delivery price.

Variance and volatility swaps are traded on the exchanges (CME) and the Chicago
Board Options Exchange (CBOE). We estimate variance functions for the variance of the S&P
500 index, calculated with the daily returns of the index.

Variance swaps offer an opportunity to contract hedging, if we have no access to
a variance swap on an index and the variance swaps on the contract, then we consider
an index variance variance swap:

$$\kappa \left[ \rho \ell - \sigma \right] + q \mathbb{V}_q,$$
where \( \lambda_k^{(3)} \) are the log returns of the index constituents and \( p \) and \( q \) are the weights of the constituents. Let us define the realized variance as

\[
\sigma_x \equiv \sum_{k=1}^{K} \lambda_k^{(3)} \lambda_k^{(3) T}
\]

Thus,

\[
\sigma_x = \frac{1}{2pq} \left( \sigma_e^{(1)} - p \sigma_e^{(2)} - q \sigma_e^{(3)} \right),
\]

where \( \sigma_e \) is the realized variance of the index and \( \sigma_e^{(3)} \equiv \sum_{i \in \mathcal{A}_0} (\lambda_i^{(3)})^2 \) and the realized variances of the index constituents.

### 3.3 Portfolio Selection

**Basic Concepts of Portfolio Selection**

Let

\[
S_e = (S_e^1, \ldots, S_e^{N}),
\]

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

\[
\sum_{i=1}^{N} b_e^i = 1. \tag{1.90}
\]

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

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

\[
R_e = \left( \frac{S_e^2}{S_e^1}, \ldots, \frac{S_e^{N+1}}{S_e^N} \right), \quad i = 1, 2, \ldots, T. \tag{1.91}
\]

It is reasonable to assume that the time series \( R_e \) is approximately stationary.\(^8\) In statistical portfolio selection we have available, besides the historical returns \( R_e \) of the assets, also other information \( Z_e \). The variables in vector \( Z_e \) can be macroeconomic variables, like the term premium, default premium, and dividend yield.\(^9\) This problem of portfolio selection can now be described as a problem of choosing a portfolio vector \( b_e \) at time \( T \) using data \((R_e, Z_e), i = 1, \ldots, T\).\(^10\)

\(^7\) 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. Practically, selling short means that an asset is sold before it is borrowed, or before nothing once 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.
Simple Terminal Variations - Let \( W_T > 0 \) be the wealth available at time \( T \). When the portfolio vector is \( \mathbf{x}_T \), then the gross return of the portfolio for the time period from \( T \) to \( T+1 \) is:

\[
W_{T+1} \mathbf{x}_T = \sum_{t=1}^{N} \mathbf{b}_t \mathbf{x}_T s^t \mathbf{s}_T \mathbf{s}_t^{-1}.
\]

In the simple terminal portfolio selection, the optimal portfolio vector can be defined as:

\[
\hat{\mathbf{x}}_T = \arg\max_{\mathbf{x} \in X_T} \mathbb{E}_T \left[ r_T \right] \left( \mathbf{x}_T, \mathbf{r}_T \right)
\]

where \( x : (0, \infty) \rightarrow \mathbb{R} \) is a utility function and

\[
X_T = \left\{ \left( b_1, \ldots, b_N \right) : \sum_{i=1}^{N} b_i = 1 \right\}.
\]

Note that \( 0 < \mathbf{x}_T, \mathbf{r}_T < \infty \). The notation \( \mathbb{E}_T \) means that the expectation is taken at time \( T \), using information available at time \( T \). If the available information is contained in the historical return \( \mathbf{r}_T \) and in the historical values of the variables \( \mathbf{x}_T \), then the expectation \( \mathbb{E}_T \) can be taken as the conditional expectation, conditional on the previous returns and previous values of variables \( \mathbf{x}_T \):

\[
\mathbb{E}_T \left[ r_T \right] \left( \mathbf{x}_T, \mathbf{r}_T \right) = \mathbb{E}_T \left[ \mathbb{E}_T \left[ r_T | \mathbf{x}_T, \mathbf{r}_T, \mathbf{x}_{T-1} \right] \right].
\]

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

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

\[
x_\gamma \left( x \right) = \left\{ \begin{array}{cl} x^\gamma & \text{if } x > 1, \\ \log x & \text{if } x = 1, \\ \theta x & \text{if } x < 1. \end{array} \right.
\]

for \( \gamma > 0 \). The power utility functions are called constant relative risk aversion utility functions (CRRA). A utility function is used instead of the pure return, because although it takes into account and does not optimize the pure return. Government bonds and a 90-day Treasury bill. The default premium is the difference between the interest rate of a long-term bond and a higher grade bond. For example, the default premium can be the difference between the normalized yields of long-term bonds and 90-day T-bills. The default yield is the default payment of a company divided by its market capitalization, when the market capitalization is the value of the stock multiplied by the number of shares.

It does not matter whether we take the utility from the wealth or from the gross return. Instead, for \( \gamma > 0 \),

\[
x_\gamma \left( W_{T+1} \mathbf{x}_T \right) = W_{T+1}^{1-\gamma} x_\gamma \left( \mathbf{x}_T \right)
\]

and for \( \gamma < 0 \),

\[
x_\gamma \left( W_{T+1} \mathbf{x}_T \right) = -x_\gamma \left( \mathbf{x}_T \right).
\]

Thus the optimal portfolio vector is the same regardless of the initial wealth \( W_T \).
Parameter $\alpha \geq 1$ is the risk aversion parameter, and larger $\alpha$ means larger risk aversion. A person returns expected zero wealth return that we know as insurance and, thus, the utility of zero gross return should be equal to minus infinity. Thus, the utility function realizes a severe penalization of returns near zero. Also, the utility of a positive return does not grow linearly but as a concave function of the return.

When we start with wealth $W_0^0$ at time $t^*$ and use portfolio weights $\delta_{t^*}^{0}, \ldots, \delta_{t^*}^{N^*}$, then the wealth at time $X_1$ in

$$W_{X_1} = \left\{ \delta_{t^*}^{0} X_0 + 1 \right\}_{t^*}^{X_1}$$

The gross return on the portfolio for the time period from $t^*$ to $X_1$ is

$$\left\{ \delta_{t^*}^{0} X_0 + 1 \right\}_{t^*}^{X_1}$$

In the multiperiod portfolio selection, assuming that new investment horizon extends from $t^*$ to a future time $X_1$, and we are able to change the portfolio weights at all times $t^*, \ldots, X_1$, the optimal portfolio weights at times $t^*$ are defined by

$$\delta_{t^*}^{0} \cdots \delta_{X_1}^{N^*} = \arg \max_{\delta_{t^*}^{0} \cdots \delta_{X_1}^{N^*}} \left\{ \left( \begin{array}{c} x_{t^*} \cdots x_{X_1} \\ 0_{N^*} \end{array} \right) \right\}$$

In the maximization problem (1.96), we supply utility function as in the multiperiod growth return, given in (1.96). This simple period case is obtained as a special case when $X_1 = X^* + 1$. The optimization problem (1.97) is of the same type as the optimization problem of the stochastic growth (1.52).

In the multiperiod portfolio selection, we introduce the following notation: $\mathbf{X}_d$ is the vector of random variables for the portfolio selection. We consider the multiperiod portfolio selection and want to choose a portfolio vector $\delta_{t^*} = \left( \delta_{t^*}^{0}, \ldots, \delta_{X_1}^{N^*} \right) \in \mathbb{R}^{N^*}$ at times $t^*$ so that the expected utility of the wealth in the maximization problem (1.97), as in the optimization problem (1.97). We can define, for a fixed portfolio vector $\delta \in \mathbb{R}^{N^*}$, where $\sum \delta^i = 1$, the expected utility of the portfolio and the asymptotic variance

$$\mathcal{Y}_{\delta} = \mathbb{E} \left\{ \delta^t \mathbf{X}_{d} \right\}, \quad \mathcal{X}_\delta \in \mathbb{R}^{N^*}$$

where $\mathcal{Y}_{\delta}$ is the expected utility of the portfolio and $\mathcal{X}_\delta$ is the asymptotic variance.
collection of all predictions, for all indices of the portfolio vector \( \tilde{b} \), given a way to choose the optimal portfolio vector. Namely, at time \( T \) we use the data
\[
(X_k, x), \quad k = 1, \ldots, T - 1
\]
to estimate the regression function, let us denote this estimate by
\[
\hat{f}_{b, x} : \mathbb{R}^d \rightarrow \mathbb{R}.
\]
We choose the optimal portfolio vector \( \tilde{b}_T \) at time \( T \) by
\[
\tilde{b}_T = \arg \max_{b \in \mathcal{B}} \hat{f}_{b, X_T}(X_T),
\]
where \( \mathcal{B} \subset \mathbb{R}^{K_T} \) where \( X_T \) is the sphere in \( \mathbb{R}^{K_T} \) defined in (1.94). Thus we choose the portfolio vector for which the prediction of the utility of the return of the portfolio in the highest sphere \( T \) is the current time, we use \( \tilde{b}_T \) to allocate the current wealth, and the portfolio vectors \( \tilde{b}_T, t = 1, \ldots, T - 1 \), can be used to analyze the statistical properties of the portfolio selection method.

We can also describe the procedure by defining function \( \tilde{b} : \mathbb{R}^d \rightarrow \mathcal{B} \) by
\[
\tilde{b}(x) = \arg \max_{b \in \mathcal{B}} \hat{f}_{b, x}(x).
\]
This function is estimated at time \( T \) by
\[
\tilde{b}_T(x) = \arg \max_{b \in \mathcal{B}} \hat{f}_{b, X_T}(x).
\]
At time \( T \), we choose the portfolio vector \( \tilde{b}_T(X_T) \).

We can use the ideas of (1.94) to transform the time series (1.91) to regression data and one can define the explanatory variables
\[
X_k = (x_{k1}, \ldots, x_{kd}^K), \quad k = 1, \ldots, T - 1
\]
and the explanatory variable \( X_0 \) is defined as a vector of length \( K \) of past gross returns. Then 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 one makes some transformation of the past returns. Possible transformations are discussed in Section 1.7. If the time series \( x_{k1}, \ldots, x_{kd} \) is stationary, then \((X_k, x), k = 1, \ldots, T - 1 \), are identically distributed.


**Markov Selection and Classification** We assume to have data \((x_k, X_k), k = 1, \ldots, T\), where \( X_k \subset \mathbb{R}^{K_k} \) is the gross return vector observed in (1.91) and \( X_k \subset \mathbb{R}^d \) is the vector of explanatory variables observed at time \( k \).
Let $X \in \{ \theta_0, \ldots, \theta_N \}$ be a finite class of portfolio vectors. Define the class labels $Y_{\theta}$ by

$$Y_{\theta} : = \theta \in \theta_0, \ldots, \theta_N$$

where $k = 0, \ldots, K - 1$. Then $Y_{\theta} : = \theta_k$ is the portfolio vector chosen at time $t$ that yields the best return at time $t + 1$, among all the portfolio vectors $\theta_k$.

We have now defined classification data $(X_{\theta}, Y_{\theta}, t : = 1, \ldots, T) = 1, 2$, which is used at time $T$ to estimate the classification function. The estimated classification function $\hat{g}$ chooses one of the portfolio vectors in $\theta_k$. Thus we define the portfolio vector which is chosen at time $T$ by

$$\hat{g}(X_T) : = \hat{g}(X_T)$$

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 prescribed. The portfolios obtained by classification corresponds to using the risk aversion parameter $\gamma = 0$.

Amenc and Jeanblanc (2012) use a classification based approach to portfolio selection. They make for each stock in $\mathbb{R}^T$, 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 Black (1962). This method provides an alternative to the use of the minimization of the expected return. The optimal portfolio vector in the mean variance sense maximizes the penalized expected return

$$\mathbb{E}(\theta^T \beta_{t+1}^T) - \frac{\gamma}{2} \text{Var}(\theta^T \beta_{t+1}^T)$$

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

$$\beta_{t+1}^T : = (\beta_{t+1}^1, \ldots, \beta_{t+1}^N)$$

is the vector of the gross returns of the $N$ portfolio components, see (1.92) and (1.93). The minimization is done over a space of portfolio vectors $\theta : \mathcal{P}_m$, where $\mathcal{P}_m$ is the sphere in $\mathbb{R}^N$, defined in (1.93). We have

$$\mathbb{E}(\theta^T \beta_{t+1}^T) : = \theta^T \beta_{t+1}^T, \quad \text{Var}(\theta^T \beta_{t+1}^T) : = \theta^T \text{Var}(\theta^T \beta_{t+1}^T) \theta$$

where $\text{Var}(\theta^T \beta_{t+1}^T)$ is the $\theta^T \times \theta^T$ covariance matrix of $\beta_{t+1}^T$. We have to estimate the vector of expected returns $\theta^T \beta_{t+1}^T$ and the covariance matrix $\text{Var}(\theta^T \beta_{t+1}^T)$.

We shall consider in Section 3.4.2.3 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 $\delta : = (\delta_1, \delta_2) : = (1 \cdot w, w)$, where $w \in \mathbb{R}$. That is, we put proportion $1 - w$ to the first asset and the proportion $w$ to the second asset. Thus

$$\delta^T \beta_{t+1} : = (1 - w) \beta_{t+1}^1 + w \beta_{t+1}^2.$$
Let the expected returns of the stocks be $\mu_{S_{0,t-1}} := \mu_{1}, \mu_{S_{0,t-1}} := \mu_{2}$ and the variances of the returns $\text{Var}(S_{0,t-1}) := \sigma_{1}^{2}$, $\text{Var}(S_{0,t-1}) := \sigma_{2}^{2}$. Assume the covariance of the returns by $\text{Cov}(S_{0,t-1}, S_{0,t-1}) := \sigma_{12}$. We have:

\[\text{Cov}(S_{0,t-1}, S_{0,t-1}) = \sigma_{12} \cdot \text{Var}(S_{0,t-1})^{1/2} \cdot \text{Var}(S_{0,t-1})^{1/2} \cdot \gamma \left[(1 - m)^{2} \sigma_{1}^{2} + m^{2} \sigma_{2}^{2} + \gamma(1 - m)^{2} \sigma_{12}^{2}\right].\]

Setting the derivative with respect to $m$ to zero and solving the two pieces:

\[m = \frac{\gamma \sigma_{1} \sigma_{2} - \mu_{1} \sigma_{2} + \mu_{2} \sigma_{1}}{\gamma \sigma_{1}^{2} + \gamma \sigma_{2}^{2} - 2 \gamma \sigma_{12}}.\]

when $\gamma > 0$. For $\gamma = 0$, as much as possible is invested in the asset for which the expected return $\mu_{i}$ is larger.

### 11.3.4. Optimal Pricing and Hedging

We consider an American option written at time $T_{0}$ (today), whose expiration is at a future time $T$. The option has value $K_{T}$ 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 $K_{T} = \max\{0, S_{T} - K\}$, where $K$ is the strike price. We need to determine a fair price $K_{T_{0}}$ 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 initial wealth equal to zero so that we can take $K_{T_{0}} = 1$ for all $K$. We consider the discrete-time model, where hedging is done at the times points $T_{0}, T_{1}, \ldots, T_{n}$. Let $W_{t}$ be the wealth at time $t$ used in buy stocks and bonds. Let $b_{t}$ be the number of stocks bought at time $t$, $b_{t-1}$, and kept until time $t$, where $a_{t}$ denotes the quantity bought in stocks. Let $c_{t}$ be the number of bonds bought at time $t - 1$ and kept until time $t$. Since the portfolio is self-financing, the quantities $b_{t}$ and $c_{t}$ have to satisfy:

\[W_{t+1} = W_{t} + c_{t} S_{t} - b_{t} B_{t},\]

The wealth at time $t$ is then:

\[W_{t} = c_{t} S_{t} + b_{t} B_{t},\]

which is again distributed among the stocks and the bond by choosing $c_{t+1}$ and $b_{t+1}$. Thus,

\[W_{t+1} = c_{t+1} S_{t+1} + b_{t+1} B_{t+1} = \left(c_{t+1} S_{t+1} + b_{t+1} B_{t+1}\right) + \left(c_{t} S_{t} - b_{t} B_{t}\right) = W_{t} + \left(c_{t} S_{t} - b_{t} B_{t}\right).\]
We get inductively \(^{10}\)

\[
\mathcal{W}_t = \mathcal{W}_{s_0} \cdot \prod_{k=0}^{t-1} \mathcal{L}_{t-k} \left( S_{t-k} \cdot \delta_{k} \right).
\]

We can use two slightly different heuristics to derive the fair price:

1. We consider the fair price to be the initial wealth \(\mathcal{W}_{s_0}\) that minimizes the minimal difference between the final wealth and the payout of the option. That is, we want to minimize:

\[
\rho(\mathcal{W}_{s_0} \cdot \mathcal{M}_t)^{12}
\]

over all initial wealths \(\mathcal{W}_{s_0}\) and over all hedging strategies.

2. The writer of the option maximizes the premium \(\mathcal{M}_{s_0}\) at time \(s_0\), by hedging his position at time points \(s_0, s_1, \ldots, s_t\), which initial wealth \(\mathcal{W}_{s_0} = 0\), and pays \(\mathcal{W}_t\) at the expiration to the holder of the option. Thus the wealth of the option writer at the expiration time \(s_t\) is equal to

\[
\mathcal{W}_{s_t} = \mathcal{M}_{s_0} \cdot \prod_{k=0}^{t-1} \mathcal{L}_{t-k} \left( S_{t-k} \cdot \delta_{k} \right) \cdot \mathcal{W}_{s_0}.
\]

We want to find \(\mathcal{M}_{s_0}\) and \(\delta_{0}, \ldots, \delta_{t}\) so that \(\mathcal{W}_{s_t}\) is as close as possible and the corresponding values for \(\mathcal{M}_{s_0}\) can be considered as a fair value of the option. That is, we want to minimize:

\[
\rho(\mathcal{W}_{s_t})^{14}
\]

where the mean squared error measure closeness is zero.

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

\[
\mathcal{X} = \mathcal{W}_t \quad \left( X_0, \ldots, X_t \right) = \left( S_{t-1} \cdot \delta_{t-1}, \ldots, S_0 \cdot \delta_0 \right)
\]

where \(X_t = \mathcal{W}_t\). We define the fair price and the optimal hedging coefficient at times \(s_0\) as:

\[
\left( \mathcal{M}_{s_0}, \mathcal{L}_{s_0}, \mathcal{K}_{s_0} \right) = \arg\min_{\mathcal{M}_{s_0}, \mathcal{L}_{s_0}, \mathcal{K}_{s_0}} \rho(\mathcal{M}_{s_0} \cdot \mathcal{L}_{s_0} \cdot \mathcal{K}_{s_0} - X_t),
\]

where \(\rho(\cdot) = \frac{\partial^2}{\partial^2 x}\) or \(\rho\) is some quadratic loss function as in (1.169). We have obtained a problem of stochastic control as described in (1.51).

\(^{10}\)When interest rate for one period is \(r \geq 0\), so that \(\mathcal{L}_{s_{k+1}} = \left(1 + r\right)^{r_{s_{k+1}} - r_{s_k}}\), we get the expression:

\[
\mathcal{W}_t = (1 + r)^{t-s_0} \mathcal{W}_{s_0} \cdot \prod_{k=0}^{t-1} \mathcal{L}_{t-k} \left( S_{t-k} \cdot \delta_{k} \right),
\]

where \(X_t = (1 + r)^{t-s_0} \mathcal{W}_{s_0}\).
1.5 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 1.5.1. The second data set is a vector time series of S&P 500 and Nasdaq-100 returns, described in Section 1.5.2.

We use other data sets as examples, such as Section 2.2.1 a vector time series of DAX 30 and DAXindex 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.

1.5.1 Time Series of S&P 500 Index

The S&P 500 index data consists of the daily closing prices of the S&P 500 index during the period from 1950-01-01 until 2015-06-25, which makes 1,59630 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

\[ r_t = \frac{P_t - P_{t-1}}{P_{t-1}} \]

where \( P_t \) is the price of the index at the end of day \( t \).
16.2 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 2003-03-19, which includes 6222 days of observations. The data are provided by Yeh, where the index symbols are "SP500" and "NDX".

Figure 12 shows the S&P 500 and Nasdaq-100 index returns over the observation period. Panel (a) shows the time series of contemporaneous index returns. 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 indices.

16.7 Regressor Transformation

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 in the case where 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 another variable takes values in [0, 100], then the variable with the shorter range would effectively be cancelled out when using spherically symmetric neighborhoods.

First, we define data scaling, 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 cepstral transformation.
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.

12.9.1 Derive Sphehiong

We can impose the scales of variables compatible by normalizing observations so that the sample variance of the variables are equal to one. Let \(X_d : = (X_{id}, \ldots, X_{nd}), i = 1, \ldots, n, d = 1, \ldots, d_d\), be the original observations. The transformed observations are

\[
X_d' : = \begin{pmatrix} X_{i1}' \\ \vdots \\ X_{id}' \end{pmatrix}, \quad i = 1, \ldots, n,
\]

where the sample variances are

\[
s_{ik}^2 : = \frac{1}{n} \sum_{i=1}^{n} (X_{ik}' - \bar{X}_{ik}')^2, \quad k = 1, \ldots, d_d,
\]

with the arithmetic mean \(\bar{X}_{ik}' : = \frac{1}{n} \sum_{i=1}^{n} X_{ik}'\). We can also make the ranges of the variables equal by defining the transformed observations

\[
Z_d : = (Z_{d1}, \ldots, Z_{dd}), \quad i = 1, \ldots, n, \quad d = 1, \ldots, d_d,
\]

where

\[
Z_{dk} : = \frac{X_{ik}' - \min_{1 \leq i \leq n}(X_{ik}')}{\max_{1 \leq i \leq n}(X_{ik}') - \min_{1 \leq i \leq n}(X_{ik}')} - 1, \quad k = 1, \ldots, d_d.
\]

Data sphering is a non-extractive transformation whose just standardizing the sample variances equal to one; we make a null linear transformation of data that the covariance matrix becomes the identity matrix. The sphering is a linear transformation 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 \(X : \mathbb{R}^d\) meaning we transform a linear transformation of \(X\) so that the new random variable has expectation zero and the identity covariance matrix. Let

\[
Y : = A' \{X - \mathbb{E}[X](X - \mathbb{E}[X])'\}
\]

be the covariance matrix and make the spectral representation of \(Y\):

\[
Y : = \Lambda A A' ,
\]

where \(A\) is orthogonal and \(\Lambda\) is diagonal. Then

\[
Z : = A^{-1/2} A' \{X - \mathbb{E}[X]\} ,
\]

is the sphered random vector, having the property\(^{11}\):

\[
\mathbb{Cov}(Z) : = I_d .
\]

\(^{11}\)The orthogonality of \(A\) means that \(A' A = A A' = I_d\). Thus \(A' A = I_d\) and \(\mathbb{Cov}(Z) = A^{-1/2} A \mathbb{Cov}(X) A A'^{-1/2} = I_d\).

2. Data splicing means that the data are transformed so that the arithmetic mean of the observations is removed and the empirical covariance matrix is the unit matrix. Let $X_n$ be the empirical covariance matrix.

$$
X_n := \frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})(X_i - \bar{X})',
$$

where $X_i := \frac{1}{N} \sum_{j=1}^{N} X_{ij}$ is the $i$th column vector of arithmetic means. We find the spectral representation of $X_n$,

$$
X_n = A_n A_n' A_n^{-1},
$$

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

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

The spliced data matrix is the $n \times m$ matrix $Z_n$ defined by

$$
Z_n := A_n^{-1/2} A_n' (X_1, \ldots, X_n)',
$$

where $X_i := (X_{i1}, \ldots, X_{im})'$ is the original $n \times m$ data matrix, and $1_{1 \times m}$ is the

1.1.1.2 Copula transformations

Copula modeling was explained in Section 1.3.3. Copula modeling lends itself to non-full 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 $X := (X_1, \ldots, X_n)$, where $X$ has a continuous distribution, gives random variable $Z := (Z_1, \ldots, Z_m)$ where the marginal variables have the uniform distribution on $[0, 1]$. The random variables are defined as

$$
Z_k := F_{k}(X_k), \quad k = 1, \ldots, m,
$$

where $F_k$ is the distribution function of the $k$th component of $X$. These

$$
Z := (Z_1, \ldots, Z_m)
$$

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

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

12Random variable $F_{X_k}(X_k)$ has the uniform distribution on $[0, 1]$. Because $P(F_{X_k}(X_k) \leq \xi) := F_{X_k}(\xi) = F_{X_k}^{-1}(\xi)$, we have

$$
F_{X_k}(\Phi^{-1}(F_{X_k}(X_k))) = \Phi(F_{X_k}(X_k)).
$$
where $\Phi$ is the distribution function of the standard Gaussian distribution. The components of $X$ have the standard Gaussian distribution.$^{13}$

2. The empirical transformation of data $X_1, \ldots, X_n$ means that the data are transformed so that the marginal distributions are approximately normal, or have approximately some other variable distribution. Let the rank of observation $X_{ij}$, $i = 1, \ldots, n$, $j = 1, \ldots, d$, be:

$$\text{rank}(X_{ij}) = \# \{X_{jk} : X_{jk} < X_{ij}, k = 1, \ldots, n \}.$$  

We normalize the transformation by assigning values to the ranks by uniformly distributed random numbers $[0, 1]$:

$$X_i = \left( \frac{\text{rank}(X_{i1})}{n - 1}, \ldots, \frac{\text{rank}(X_{id})}{n - 1} \right).$$

For $i = 1, \ldots, n$. Obtain the standard Gaussian distribution in a more convenient and use the definition:

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

(1.103)

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

Figure 1.5 shows scatter plots of S&P 500 and NASDAQ-100 empirical transformed normal data. The data is described in Section 1.6.2. Panel (a) shows the case where the marginals are transformed to be approximately normal, Panel (b) shows the case where the marginals are transformed to be approximately exactly uniformly distributed in $[0, 1]$. We have used in scatter plots histogram smoothing with 11% bins, as explained in Section 6.1.1. Univariate marginals make the data concentrate on the lower left and on the upper right corner, which can make the estimation difficult due to the boundary effects. The Gaussian marginals under the distribution of the data have tails which decay smoothly at zero.

1.6.3 Transformations of the Response Variables

The transformation of the response variables can be used to obtain a more normal distribution or to remove heteroscedasticity by stabilizing variance. See D'Agostino (1982).

The power transformations are called the Box-Cox transformations and defined for $\lambda \neq 0, 1$ by:

$$Y_i^{(\lambda)} = \left\{ \begin{array}{ll} X_i^{\lambda - 1} & \lambda \neq 0, \\ \log X_i & \lambda = 0, \end{array} \right.$$  

where we assume that $X_i > 0$. Box-Cox transformations were discussed in Box & Cox (1964). D'Agostino (1982) considered the power transformation $Y_i^{(\lambda)}$ for $\lambda \neq 0, 1$.

13Distribution variable $\Phi^{-1}(U)$, where $U$ has the uniform distribution on $[0, 1]$, has the standard Gaussian distribution because $P(\Phi^{-1}(U) \leq x) = P(U \leq \phi(x)) = \phi(x).$
The natural exponential family was defined in (1.58). In the natural exponential family

\[ \mathcal{N}_\theta(Y) : \mu(y) : \mu(y), \quad \text{Var}_\theta(Y) : \nu(y) : \nu(y), \]

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

\[ \text{Var}_\theta(Y) : \nu(y) : \alpha_1 \mu(y) + \alpha_2 \mu(y)^2, \]

where \( \mu(y) : \mu(y) \). The examples are normal, gamma, Poisson (the natural exponential family generated by the generalized hyperbolic normal distribution); binomial, negative binomial, and Poisson. Denote \( \text{Var}_\theta(Y) : \nu(\mu(y)) \). Define a function \( \Omega : \mathbb{R} \rightarrow \mathbb{R} \) to be such that

\[ (\mu'(y)) \rightarrow \Omega^{-1/2}(\mu). \]

By the central limit theorem, we obtain

\[ n^{1/2}(Y - \mu(y)) \rightarrow^d \mathcal{N}(0, \nu(\mu(y))), \]

as \( n \rightarrow \infty \), where \( Y : \mathbb{R}^n \rightarrow \mathbb{R}^n \). Then, \( Y_1, \ldots, Y_n \) are assumed i.i.d. By the delta method, we have

\[ n^{1/2}(C(Y) - C(\mu(y))) \rightarrow^d \mathcal{N}(0, 1), \]

as \( n \rightarrow \infty \). Thus, we call the transformation \( C(Y) \) a variance-stabilizing transformation.
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 2.1.2.

1.8.1 Independent Observations

Let \( X_1, X_2, \ldots \) be a sequence of real-valued i.i.d. random variables with \( \text{Var}(X_i) = \sigma^2 \), where \( 0 < \sigma^2 < \infty \). According to the central limit theorem, we have

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2),
\]

as \( n \to \infty \). Let \( Y_1, Y_2, \ldots \) be an i.i.d. sequence of random vectors with \( \text{Cov}(Y_i) = \Sigma \), where the diagonal elements of \( \Sigma \) are finite and positive. According to the central limit theorem, we have

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mu) \xrightarrow{d} \mathcal{N}(0, \Sigma),
\]

as \( n \to \infty \).

1.8.2 Dependent Observations

We need a central limit theorem for dependent observations, let \( Y_1, Y_2, \ldots \) be a stationary time series. We define the weak dependence in terms of a condition on the remaining coefficients. Let \( \mathcal{F}_t \) denote the sigma algebra generated by random variables \( Y_1, \ldots, Y_t \). The remaining coefficient is defined as

\[
\omega_0 = \sup_{\mathcal{F}_t \subset \mathcal{F}_{t+1}} \left| \text{Cov}(A_1, A_{-1}) : \text{Cov}(A_1, A_{-1}) \mathcal{F}_t \right|
\]

where \( A_1 = Y_1, A_2 = Y_2, \ldots \). Now we can state the central limit theorem. Let \( \mathbb{E}[|Y_1|^2] < \infty \) and

\[
\sum_{i=1}^{\infty} \omega_i \frac{1}{t^{1/6}} < \infty \quad \text{for some constant } \delta > 0.
\]

Then,

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2),
\]

where

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

\( \gamma(j) := \text{Cov}(X_j, X_{j+1}) \), and we assume that \( \omega_0 > 0 \).

Wanghao and W. Limin (1997). Theorem 18.4.1) gives necessary and sufficient conditions for a central limit theorem under varying conditions. A proof for our statement:
of the central limit theorem in (1.106) can be found in Phillips (1986); see also Hurvich & Yao (2005, Theorem 2.21) and Hud Hondley (2006, Theorem 2.7.2).

Let us state the central limit theorem for the vector time series \( (X_t)_{t \geq 0} \), where \( X_t \in \mathbb{R}^d \). If the time series \( (a' X_t)_{t \geq 0} \) satisfies the conditions for the univariate central limit theorem for all \( a \in \mathbb{R}^d \), then

\[
\frac{1}{\sqrt{n}} \sum_{t=1}^n \langle a', X_t \rangle \rightarrow N(0, \Sigma),
\]

where

\[
X_t = \sum_{j=0}^\infty b_j Y_{t-j}, \quad \Sigma = \text{Cov}(X_t, X_{t+h}),
\]

and the auto-covariance matrix \( \Sigma \) was defined in (1.21) as

\[
\Sigma = \text{Cov}(X_t, X_{t+h}).
\]

Note that we used the property (1.22) \( b_j' \) was defined in (1.22) as

\[
\Sigma = \text{Cov}(Y_t, Y_{t+h}).
\]

Let us explain the expression for the asymptotic variance \( \sigma^2 \) in the univariate central limit theorem in (1.106). We assume that \( \mathbb{E}u_i = 0 \). The variance of the normalized sum is

\[
\text{Var} \left( \frac{1}{\sqrt{n}} \sum_{t=1}^n X_t \right) = \frac{1}{n} \sum_{t=1}^n \text{Var}(X_t) + \frac{1}{n^2} \sum_{t=1}^n \sum_{s=1}^n \text{Cov}(X_t, X_s).
\]

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

\[
\text{Var} \left( \frac{1}{\sqrt{n}} \sum_{t=1}^n X_t \right) = \text{Var}(X_1) = \gamma(0).
\]

For a weakly stationary time series we have

\[
\text{Var} \left( \frac{1}{\sqrt{n}} \sum_{t=1}^n X_t \right) = \frac{1}{n} \sum_{t=1}^n \text{Var}(X_t) + \frac{1}{n^2} \sum_{t=1}^n \sum_{s=1}^n \text{Cov}(X_t, X_s),
\]

\[
= \gamma(0) + \frac{1}{n} \sum_{j=1}^\infty \text{Cov}(X_1, X_{1-j})
\]

\[
= \gamma(0) + \frac{1}{\sqrt{n}} \sum_{j=1}^\infty \gamma(j).
\]

Thus, in order that \( \text{Var} \left( \frac{1}{\sqrt{n}} \sum_{t=1}^n X_t \right) = \gamma(0) \) for a finite positive constant \( \gamma(0) \), we need that \( \gamma(j) \to 0 \) sufficiently fast, say \( j \to \infty \). A sufficient condition is that

\[
\sum_{j=1}^\infty |\gamma(j)| < \infty.
\]

14 Unlike Weil's theorem, we state that \( X_n \rightarrow Y \) if and only if \( a' X_n \rightarrow a' Y \) for all \( a \in \mathbb{R}^d \), as \( n \to \infty \), where \( X_n \) and \( Y \) are random vectors.
Reparameterization and Estimation of the Asymptotic Variance

In the applications we have to estimate the asymptotic variance and the asymptotic covariance matrix. For stationary 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 $\sigma^2$ in (1.1.06) using the observations $Y_1, \ldots, Y_{10}$, and the estimation of the covariance matrix $\Sigma$ in (1.1.07) using the observations $X_1, \ldots, X_{10}$.

Let us start with the estimation of $\sigma^2$ in (1.1.06). An application of the sample covariances would lead to the estimator:

$$\hat{\sigma}^2 := \sum_{j=0}^{9} \hat{\gamma}(j),$$

where

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

for $\bar{Y} := \frac{1}{n} \sum_{k=1}^{n} Y_k$. Note that for large $n$ only few observations are used in the estimator $\hat{\gamma}(j)$. For example, when $j = \alpha - 1$ the estimator uses only one observation:

$$\hat{\gamma}(\alpha - 1) := \frac{1}{n-\alpha+1} \sum_{k=1}^{n-\alpha+1} Y_k Y_{k+\alpha-1},$$

which is a very imprecise estimator. We can use weighting to remove the imprecision from the estimator and define:

$$\hat{\sigma}^2 := \sum_{j=1}^{9} \sigma(j) \hat{\gamma}(j),$$

where

$$\sigma(j) := \left( 1 - \frac{j}{n} \right)^{-1},$$

where $1 < \alpha < n - 1$ is a chosen smoothing parameter. We can generalize this estimator to other weights and define:

$$\sigma(j) := \mathcal{K}(j/\alpha),$$

where $\mathcal{K} : [0, 1] \rightarrow [0, 1]$ is a kernel function satisfying $\mathcal{K}(x) := \mathcal{K}(-x)$, $\mathcal{K}(0) := 1$, $|\mathcal{K}(x)| < 1$ for all $x$, and $\mathcal{K}(x) = 0$ for $|x| > 1$.

To estimate $\gamma$ in (1.1.07) we use:

$$\hat{\gamma}(j) := \sum_{k=1}^{n-j} (X_k - \bar{X})(X_{k+j} - \bar{X})' \sigma(j),$$

where:

$$\hat{\gamma}(j) := \sum_{k=1}^{n-j} (X_k - \bar{X})(X_{k+j} - \bar{X})',$$
for \( j = 0, \ldots, n - 1 \). We will apply weights in an estimator of an asymptotic covariance matrix in (2.40).

The weighting was first used in relation to the smoothing in the estimation of the spectral density. The unnormalized spectral density function of a weakly stationary time series, having autocorrelation coefficients \( \gamma(j) \) with \( \gamma_j = O(n) \), \( |\gamma(j)| < c \), is defined by

\[
g(\omega) = \sum_{j=0}^{\infty} \gamma_j e^{j \omega j},
\]

where \( \omega \in \mathbb{R} \), \( \gamma_j \) are the autocorrelation coefficients. The lag window spectral density estimator, based on data \( Y_1, \ldots, Y_n \), is defined by

\[
h(\omega) = \sum_{k=-n^2}^{n^2} K(\omega/k) \gamma(k) e^{j \omega k},
\]

where \( K(\cdot) \) are the kernel functions of the normal distribution. The kernel functions are similar to those in (1.100); see also White (1994, Section 10.4). Now we have

\[
h(0) = \frac{1}{2n^2} \sum_{k=-n^2}^{n^2} K(\omega/k) \gamma(k) e^{j \omega k},
\]

where \( n^2 \) is defined in (1.100) with the weights defined in (1.102).

1.9 MEASURING PERFORMANCE OF REGRESSION ESTIMATES

We discuss measuring the performance of regression function estimates, conditional variances, covariances, and quantitative estimators, estimates 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 \( f(x) \). We define theoretical performance measures, which are used to compare estimators of \( \hat{f} \) under given theoretical assumptions. After that we define empirical performance measures, which are used to estimate the performance of estimates \( \hat{f} \) using the available data.

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 integrated averaged squared error to measure the goodness of regression function estimates \( \hat{f} \) globally, where we want to recover the complete curve and not its values at a single point \( x \in \mathbb{R}^d \).
The prediction error of regression function \( f \) can be measured by

\[ \mathbb{E}(\hat{f}(X) - Y)^2. \]

This measure of prediction is natural since \( f(x) = \mathbb{E}(Y \mid X = x) \) and the conditional expectation minimizes the mean squared error, as shown in (1.101). When we have an estimator \( \hat{f} \) of \( f \), then we can measure the prediction error of the estimator by

\[ \mathbb{E}(\hat{f}(X) - Y)^2. \]

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

\[ \mathbb{E}\left( \left( \hat{f}(X) - Y \right)^2 \right) = \int_{\mathbb{R}^d} \left( \hat{f}(x) - f(x) \right)^2 f_X(x) \, dx \quad \mathbb{E}(\hat{f}(X) - Y)^2. \]

where \( f_X \) is the density function of \( X \). The minimization of expression (1.101) with respect to estimator \( \hat{f} \) is equivalent to the minimization of the expression

\[ \int_{\mathbb{R}^d} \left( \hat{f}(x) - f(x) \right)^2 f_X(x) \, dx. \]

This calculation can be used to justify the mean integrated squared error, denoted \( (1.102) \).

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

\[ \text{MISE}(\hat{f}, f) = \mathbb{E}(\hat{f}(X) - f(X))^2 \]

\[ = \mathbb{E}\left( \left( \hat{f}(X) - f(X) \right)^2 \right) \left( Y_1, X_1 \right), \ldots, \left( Y_n, X_n \right) \]

\[ = \mathbb{E}\int_{\mathbb{R}^d} \left( \hat{f}(x) - f(x) \right)^2 f_X(x) \, dx. \]

where \( X \) is independent of \( \left(Y_1, X_1\right), \ldots, \left(Y_n, X_n\right) \) and \( f_X \) is the density function of \( X \). Using the short-hand notation we write the mean integrated squared error as

\[ \text{MISE}(\hat{f}, f) = \mathbb{E}\left\| \hat{f} - f \right\|_{L^2(X)}^2 \]

where \( \left\| f \right\|_{L^2(X)}^2 = \int_{\mathbb{R}^d} f(x)^2 d\mu_X(x) \), and \( \mu_X \) is the probability distribution of random vector \( X \). We can generalize (1.103) to

\[ \mathbb{E}\int_{\mathbb{R}^d} \left( \hat{f}(x) - f(x) \right)^2 \pi(x) d\mu_X(x), \]
\[ \text{MAE}(\hat{f}, f) = \mathbb{E} \left[ \sqrt{\sum_{i=1}^{n} (\hat{f}(X_i) - f(X_i))^2} \right] \quad \text{for} \quad X_1, \ldots, X_n \]

Using the above formal definitions, we write the mean averaged squared error as:

\[ \text{MAE}(\hat{f}, f) = \mathbb{E} \left[ \sum_{i=1}^{n} (\hat{f}(X_i) - f(X_i))^2 \right] \quad \text{for} \quad X_1, \ldots, X_n \]

where:

\[ \mathbb{E} \left[ \sum_{i=1}^{n} (\hat{f}(X_i) - f(X_i))^2 \right] = \int_{\mathbb{R}^d} f(x)^2 \, d\mathbb{P}(x) - \sum_{i=1}^{n} f(X_i)^2 \quad \text{for} \quad f(X_i) \]

\[ X \sim \mathbb{P} \]

**Mean Square Error (MSE)**

Pointwise performance measures quantify how well the values of \( f \) are measured at a single point \( x \in \mathbb{R}^d \). We can use mean squared error (MSE) either unconditionally or conditionally.

- **The unconditional mean squared error at point \( x \in \mathbb{R}^d \)** is defined as:

\[ \text{MSE}(\hat{f}(x), f(x)) = \mathbb{E} \left[ (\hat{f}(x) - f(x))^2 \right] \]

where \( f \) is the true regression function.

- **The conditional mean squared error at point \( x \in \mathbb{R}^d \)** is defined as:

\[ \text{MSE}(\hat{f}(x); f(x)) = \mathbb{E} \left[ (\hat{f}(x) - f(x))^2 \right] \quad \text{for} \quad X_1, \ldots, X_n \]

where \( f \) is the true regression function.

**The Role of Theoretical Performance Measures**

Theoretical performance measures can be used to compare estimations 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 \( f \) together with the additional assumptions on the distribution i
of \((X, Y)\) and on the distribution of the sample \((X_1, Y_1), \ldots, (X_m, Y_m)\). To compare estimators, we use the sup norm risk
\[
\sup_{f \in F} \| \mathbb{MSE}(\hat{f}, f) \|
\]

We use the sup norm risk because it is necessary to require that an estimator perform uniformly well across a model, because the sup norm of a single regression function \(f\) is easy to define, the best estimator; that is, the regression function \(\hat{f}\) 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 \((X_1, Y_1), \ldots, (X_m, Y_m)\).

### Empirical Performance Measures and Cross-Validation Metrics

The mean integrated squared error
\[
\mathbb{MISE}(\hat{f}, f) = \mathbb{E}\left( \int (\hat{f}(X) - f(X))^2 \right)
\]
defined in (1.1.13), cannot be approximated by \(n^{-1} \sum_{n} \mathbb{E}(\hat{f}(X_k) - Y_k)^2\). 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 this problem using sample splitting or cross-validation.

1. **Sample Splitting**
   - Let \(\hat{f}\) be the regression function estimator constructed from the data \((X_1, Y_1), \ldots, (X_{m^*}, Y_{m^*})\), where \(1 < m^* < n\), and typically \(m^* = \left[ \frac{n}{2} \right] \). Then we use
   \[
   \mathbb{MISE}_{m^*}(\hat{f}) = \frac{1}{m^*} \sum_{k=m^*+1}^{m} \left( \int (\hat{f}(X_k) - Y_k)^2 \right)
   \]  (1.1.13)
   to estimate the mean integrated squared error.

2. **Cross Validation**
   - Let \(\hat{f}\) be a regression function estimator constructed from the other data points but not \((X_k, Y_k)\). Then we use
   \[
   \mathbb{MISE}_{m^*}(\hat{f}) = \frac{1}{m^*} \sum_{k=1}^{m^*} \left( \int (\hat{f}(X_k) - Y_k)^2 \right)
   \]  (1.1.16)
   to estimate the mean integrated squared error.

Cross validation is discussed in Section 1.2.7 in the context of linear estimation.
Proposition 4.1: 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 estimators $\hat{f}_k$ and $\hat{g}_k$ using data $(X_1, Y_1), \ldots, (X_k, Y_k)$ that is observed until time $k$, and define the mean of squared prediction errors by

$$MSE_{REG}(\hat{f}) := \frac{1}{T} \sum_{k=1}^{T-1} \left( \hat{f}_k(X_{k+1}) - Y_{k+1} \right)^2.$$  \hspace{1cm} (4.117)

which is analogous to the estimate of the mean integrated squared error defined in (4.116). We will use later in Section 4.12.1 the mean of absolute prediction errors

$$MAPE_{REG}(\hat{f}) := \frac{1}{T} \sum_{k=1}^{T-1} \left| \hat{f}_k(X_{k+1}) - Y_{k+1} \right|.$$ \hspace{1cm} (4.118)

Diebold & Mariano (1995) proposed a test for testing the equality of forecast accuracy. Let us have two predictors $\hat{f}_k(X_{k+1})$ and $\hat{g}_k(X_{k+1})$ and the corresponding losses

$$L_k := \left( \hat{f}_k(X_{k+1}) - Y_{k+1} \right)^2, \quad C_k := \left( \hat{g}_k(X_{k+1}) - Y_{k+1} \right)^2.$$  

The losses do not have to be squared prediction errors, but we can also use absolute prediction errors, for example. We get the three series of loss differentials

$$d_k := L_k - C_k.$$  

The null hypothesis and the alternative hypothesis are

$$H_0 : E(d) = 0, \quad H_1 : E(d) \neq 0.$$  

We apply the central limit theorem as stated in (4.116). Under the null hypothesis and under the assumptions of the central limit theorem, we have

$$\left( T^{1/2} d_0 + 1 \right) \cdot T^{1/2} \overset{d}{\rightarrow} N(0, \sigma^2),$$

as $T \to \infty$, where

$$\sigma^2 := \sum_{k=1}^{\infty} \gamma(k), \quad \gamma(k) := MSE_{\hat{f}} \cdot MSE_{\hat{g}}.$$  

We can use the estimate

$$\hat{\sigma}^2 := \sum_{k=1}^{T-1} w(k) \gamma(k),$$

where $w(k) = \frac{1}{T-k+1}$.
where $\sigma(x)$ is defined in (1.109). If we choose the test statistic

$$\bar{F} = \bar{C}^{\top} \left( \bar{X}_1 - \bar{e}_0 \cdot \bar{1} \right)^{-1} \bar{1} \bar{Z}_e.$$ 

When we observe $|X| > d_{\alpha/2}$, then the $p$-value is calculated by $X^2(X) > d_{\alpha/2}$ for

$$P(1 - \Phi(d_{\alpha/2})), \text{ where } \Phi \text{ is the distribution function of the standard normal distribution.}$$

1.2.2 Performance of Conditional Variance Estimates

**Theoretical Performance Measures** Theoretical performance measures can be generalized from the case of regression function estimation to the case of conditional variance estimations. For example, when $f(x) = \text{Var}(Y \mid X = x)$ and $\hat{f}(x)$ is an estimator of $f(x)$, then we can measure the performance of $\hat{f}$ by

$$\bar{K} = \int_{\mathbb{R}} \left( f(x) - \hat{f}(x) \right)^2 w(x) \, dx, \quad (1.119)$$

where $w(x) = 1024^2$ if $w(x)$ is a weight function.

**Empirical Performance Measures** Empirical performance measures are based on the cross-sectional data and then for time series data.

1.3 Time-Varying Models

Empirical performance measures of conditional variance estimations can be formed naturally in the case where

$$\mathbb{E}(Y \mid X = x) = 0.$$ 

We then

$$f(x) = \text{Var}(Y \mid X = x) = \mathbb{E}(Y^2 \mid X = x).$$

For example, we can use sample splitting. Let $\hat{f}^* = \text{an estimator of } f$, constructed from the subset $(X_1, Y_1), \ldots, (X_{n^*}, Y_{n^*})$, where $1 < n^* < n$. Then we can use

$$\bar{K}^* = \frac{1}{n^* - n^*} \sum_{i=n^*+1}^{n} \left( f(x_i) - \hat{f}(x_i) \right)^2, \quad (1.120)$$

to measure the performance of the estimation.

**State Space Smoothing** When we have identically distributed time series observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, then we can construct an estimator $\hat{f}$ of the conditional variance using data $(X_1, Y_1), \ldots, (X_n, Y_n)$ and calculate the mean of absolute
prediction error:

\[ \text{MAPE}_{T}(\hat{y}) = \frac{1}{T} \sum_{t=1}^{T} \left| \hat{y}_{t} - y_{t} \right|, \]

(1.1.21)

where \( n_{0} \) is the initial estimation period, \( 0 < n_{0} < T' \). We start to evaluate the performance of the estimator after \( n_{0} \) observations are available, because any estimator can behave arbitrarily when only few observations are available. Mean absolute prediction error is sometimes called the mean absolute deviation error (MADE).

Yousef, 2009(12) proposes to take the square root of the mean square root prediction error criterion as the performance measure:

\[ \text{MSRE}_{T}(\hat{y}) = \frac{1}{T} \sum_{t=1}^{T} \left( \hat{y}_{t} - y_{t} \right)^{2} \]

(1.1.22)

Spall's (2005) proposes to take the square root and use the mean square root prediction error criterion as the performance measure:

\[ \text{MSRE}_{T}(\hat{y}) = \frac{1}{T} \sum_{t=1}^{T} \left( \hat{y}_{t} - y_{t} \right)^{2} \]

(1.1.23)

The mean square root prediction error is such that outliers do not have as strong influence on the estimate. From model (3.1.3) proposes to measure the performance with the mean absolute deviation error:

\[ \text{MADE}_{T}(\hat{y}) = \frac{1}{T} \sum_{t=1}^{T} \left| \hat{y}_{t} - y_{t} \right| \]

(1.1.24)

where the factor \( \sqrt{T} \) comes from the fact that for a standard normal random variable \( Z \sim N(0, 1) \), we have \( \|Z\|_{p} = \sqrt{T/\pi} \).

We can generalize the performance measures (1.1.22) (1.1.24) and define a class of performance measures by

\[ \text{MAPE}_{T}(\hat{y}) = \frac{1}{T} \sum_{t=1}^{T} \left( \hat{y}_{t} - y_{t} \right)^{p}, \]

(1.1.25)

where \( Z \sim N(1, 1) \). For \( p > 1 \), we have

\[ \text{MAE}_{T}(\hat{y}) = \frac{T^{p/2} \Gamma((p + 1)/2)}{\sqrt{T}} \]

(1.1.26)

and the combinations \( (p = 1, q = 1), (p = 1, q = 2) \), \( (p = 1, q = 1) \), 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 $q$, and Figures 3.22 and 3.23. We use $\text{GARCH}^{(p,q)}_{\Delta X}$ with $p = 1$ and $q = 2$ in Section 3.11.1 to compare $\text{GARCH}(1,1)$ and the exponentially weighted moving average.

An additional performance measure is the mean absolute error

$$\text{MAARE}^{(p)}_{\Delta X} : = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{|\xi_t|^{p}} \sqrt{\sum_{t=1}^{T} \left| \frac{\xi_t}{|\xi_t|^{p}} \phi_{\Delta X}^{(p)} \right|^2},$$  

(1.12A)

where $p > 0$ and $\xi_t \sim N(0,1)$. We use $\text{MAARE}^{(p)}_{\Delta X}$ with $p = 2$ in Section 3.11.1 to compare $\text{GARCH}(1,1)$ 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 $k$-step ahead predictions, for $k = 1, 2, \ldots$. However, sometimes we are interested in estimating the realized volatility. Define the $k$-step realized volatility by

$$V_{k,A} : = \sum_{i=1}^{k} \xi_{A+1,i}.$$  

Let $\hat{f}_{k,A}(X_{A+1,i})$ be a prediction of $V_{k,A}$. We can use the mean square root prediction error as in (1.123). We modify (1.123) to obtain

$$\text{MAARE}^{(p)}_{\Delta X} : = \frac{1}{k} \sum_{i=1}^{k} \frac{1}{|\xi_{A+1,i}|^{p}} \sqrt{\sum_{i=1}^{k} \left| \frac{\xi_{A+1,i}}{|\xi_{A+1,i}|^{p}} \phi_{\Delta X}^{(p)}(X_{A+1,i}) \right|^2}. \sqrt{k},$$

We can consider $\hat{f}_{k,A}(X_{A+1,i})$ as an estimate of $E(\sum_{i=1}^{k} \xi_{A+1,i} | X_t)$.  

**Performance of Conditional Covariance Estimators**

Let us discuss measuring the performance of estimation of conditional covariance $f(x) : = \text{Cov}(Y, X | X = x)$. Empirical performance measures of conditional covariance estimates can be defined naturally in the case where

$$\text{MAARE}^{(p)}_{\Delta X} : = \text{Cov}(Y, X | X = x) = \text{Cov}(X, X | X = x).$$

so that

$$f(x) : = \text{Cov}(Y, X | X = x) = \text{Cov}(X, X | X = x).$$

For example, we can use sample splitting, similarly as in (1.110), where a performance measure for the case of measuring the performance of a conditional variance estimator was given. Let $\hat{f}_{k}^p$ be an estimator of $f$, constructed from the data $(X_1, \xi_1, X_2), \ldots (X_{m^*}, \xi_{m^*}, X_{m^*+1})$, where $1 < \ell < m^*$, and $X_{m^*+1}$. Then we can use

$$\sqrt{\sum_{i=m+1}^{m^*} \left| \hat{f}_{k}^p(X_i | X_{m^*+1}) \right|^2}$$

to measure the performance of the estimator.
In autoregressive time series smoothing methods, like in the ARMA models and exponential smoothing, average methods such as the simple exponential smoothing method (see Section 3.1.1.2), the explanatory variables are the previous observations, and the estimate \( \hat{y}_t \) of \( y_t \) is calculated using observations \( (y_{t-1}, y_{t-2}), \ldots, (y_{t-n}, y_{t-n-1}) \), and now we define the mean deviation estimate by

\[
MDF(\hat{y}_t)(\gamma) = \frac{1}{N} \sum_{i=1}^{N} \left| y_i - \hat{y}_i \right|^\gamma / \gamma,
\]

where \( \gamma > 0 \).

1.3.1 Performance of Quantile-Binning Estimators

Theoretical performance measures for the estimator of the conditional quantile:

\[
f_\gamma(x) = Q_\gamma(Y \mid X = x)
\]

can be defined similarly as in the case of conditional quantile estimates. For example, using Eq. (1.1.29).

Empirical performance measures can be found in the case of quantile binning distribution of \( Y \) by using the fact

\[
p^\gamma(Y < Q_\gamma(Y \mid X = x) \mid X = x) = p^\gamma[F_{\gamma}(\cdot, Q_\gamma(Y \mid X = x))]
\]

where \( \gamma \in \mathbb{R}^d \). Let \( (X_1, Y_1), \ldots, (X_m, Y_m) \) be regression data and let

\[
\hat{Q}_\gamma(x) = \hat{Q}_\gamma(x) \mid Y \sim x
\]

be a conditional quantile estimate constructed using the whole data but with the ith observation. Let the cross validation quantity be

\[
\hat{\gamma} = \frac{1}{m} \sum_{i=1}^{m} \left| y_i - \hat{Q}_\gamma(x_i) \right| / \gamma
\]

Finally, the performance is measured by the difference

\[
p = \hat{p} - \hat{\gamma}.
\]

Let us consider the time series modeling, where we have observations \( Y_1, \ldots, Y_m \). Then we can construct a conditional quantile estimation

\[
\hat{Q}_\gamma(x) = \hat{Q}_\gamma(Y \mid X_{-1}, \ldots)
\]

using data \( Y_1, \ldots, Y_{m-1} \), and calculate

\[
\hat{p} = \frac{1}{m} \sum_{i=1}^{m} \left| y_i - \hat{Q}_\gamma(Y_i) \right| / \gamma
\]

(1.1.29).
where $1 < s_0 < N - 1$. We start to estimate the performance of the estimation after $s_0$ observations are available, because any estimator can behave erratically when only a small sample of observations are available.

When we would know the true quantiles, there is random fluctuation in the numbers $\hat{q}_i$. The random variables

$$\hat{q}_i = \hat{q}_{(s_0, q)}(Y_{s_0 + 1}), \quad i = s_0, \ldots, s_1 + 1,$$

are Bernoulli random variables with $P(\hat{q}_i = 1) = p$, where $q_{(s)}$ is the true quantile. If random variables $Y_i$ are independent, then random variables $\hat{q}_i$ are independent, and

$$M = \sum_{i=s_0}^{s_1} \hat{q}_i$$

is a binomial random variable with the distribution $\text{Bin}(N, p)$, where $m = Y' - s_0$.

The probability mass function of $M$ is

$$P(M = i) = \binom{N}{i} p^i (1 - p)^{N-i},$$

for $i = 0, \ldots, m$. We can now calculate the numbers $c_0$ and $c_1$ such that

$$P(c_0 < M < c_1) \geq 1 - \alpha,$$

where $0 < \alpha < 1$ and $\bar{c} = m/N$. We have

$$c_0 = p \cdot \bar{c} + z_{\alpha/2} \sqrt{p \cdot (1 - p) \bar{c}}, \quad c_1 = p \cdot \bar{c} + z_{1 - \alpha/2} \sqrt{p \cdot (1 - p) \bar{c}},$$

where $z_{\alpha/2}$ and $z_{1 - \alpha/2}$ are such that $P(z_{\alpha/2} < \hat{q} < z_{1 - \alpha/2}) > 1 - \alpha$.

If $\hat{q} > p$, this means that the quantile estimator $\hat{q}$ is larger than the true quantile. When we are estimating the left tail, this means that $q$ is close to zero, then the relation $\hat{q} > p$ means that the true distribution has a heavier left tail than the quantile estimates would indicate. When we are estimating the right tail, we want $q$ to be close to one, then the relation $\hat{q} > p$ 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

$$\Delta(p, \hat{q}) = \begin{cases} p - \hat{q}, & \text{when } p < 0.5, \\ \hat{q} - p, & \text{when } p > 0.5. \end{cases}$$

Thus, the difference $\Delta(p, \hat{q})$ being negative means that the true distribution has a heavier tail than the quantile estimates would indicate. The difference $\Delta(p, \hat{q})$ 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 $Y_i$ using the S&P 500 index data described in Section 1.6.1. Let $\hat{q}_i$ be the empirical quantile, defined in (1.25),
and calculated using the data $Y_1, \ldots, Y_n$. We plot the function $y \mapsto \mathcal{M}(y, \hat{\beta})$ in black. Panel (a) shows the range $y \in [0.001, 0.075]$, and panel (b) shows the range $y \in [0.925, 0.991]$. The green lines show level $\alpha = 0.016$ simulation bands.

1.3.5 Performance of estimations of expected shortfall

To derive a performance measure for estimations of expected shortfall, we can use the fact that for a continuous distribution of $Y$, we obtain

$$
\mathbb{M}\left(Y \mapsto \mathbb{E}_p(Y) I_{(\cdot, \hat{\alpha})}(Y)\right) = 0.
$$

Indeed, for a continuous distribution of $Y$, we have

$$
\mathbb{E}_p(Y) = \frac{1}{\hat{\alpha}} \mathbb{E}\left[Y / I_{(\cdot, \hat{\alpha})}(Y)\right]
$$

and

$$
\mathbb{E}_p\left[I_{(\cdot, \hat{\alpha})}(Y)\right] = \hat{\alpha}.
$$

If we use in the time series setting and have identically distributed observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, then we can construct an estimation of the expected shortfall.
using data \((X_1, Y_1), \ldots, (X_n, Y_n)\) and calculates the performance measure

\[
\frac{1}{n} \sum_{i=1}^{n} \left( Y_i - \hat{g}_{X,Y}(X_i) \right)^2 \quad \text{or} \quad \frac{1}{n} \sum_{i=1}^{n} \left( Y_i - \hat{g}_{X,Y}(X_i) \right)^2 \cdot \mathbb{I}(Y_i = 1 - Y_i)
\]

where \(\hat{g}_{X,Y} \) is a quadratic estimator and \(1 \leq s_0 \leq T \cdot 1\).

11.9.3 Performance of (Classifiers)

Theoretical Performance Measures

Let \(g : \mathbb{R}^d \to \{0, \ldots, K - 1\} \) be a classification function. The probability of the classification error is:

\[
L(g) = P(Y \neq g(X)),
\]

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

\[
L(\hat{g}) = P(Y \neq \hat{g}(X)),
\]

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 error as

\[
L(\hat{g}) = \sum_{l=0}^{K-1} P(Y = l) \int_{\mathbb{R}^d} f_{x \mid y = l} dx,
\]

where \(f_{x \mid y = l} : \mathbb{R}^d \to \mathbb{R}\) is the density function of \(X \mid Y = l\), and

\[
\mathcal{G}_l = \{x \in \mathbb{R}^d : \hat{g}(x) = l\}, \quad l = 0, \ldots, K - 1,
\]

in the subset of the sample space, where the classification function \(\hat{g}\) chooses class \(l\).

When we analyse the asymptotic performance of the classification functions, we should note that \(L(\hat{g})\) does not converge to zero, but at least we can hope that it converges to the minimal classification error \(L(g^*)\), which is the classification error of the Bayes rule \(g^*\), defined in (1.75). Thus we should study the rate of convergence of \(L(\hat{g})\) to \(L(g^*)\). Let us consider the two-class case \(K = 2\) with the equal class priors \(L(Y = 0) = L(Y = 1) = 1/2\). Then,

\[
L(\hat{g}) = \frac{1}{2} \int_{\mathbb{R}^d} \min\{f_{x \mid y = 0} \cdot g(x) = 0, f_{x \mid y = 1} \cdot g(x) = 1\} dx,
\]

and

\[
L(\hat{g}) = L(g^*) = \frac{1}{2} d_{\text{sym}}(G_1, G_2) = \left\{(x : \hat{g}(x) = 1), (x : g^*(x) = 1)\right\},
\]

where

\[
d_{\text{sym}}(G_1, G_2) = \int_{G_1 \cap G_2} |1 - g_2| d\mu,
\]

with

\[
G_1 = \hat{g}(X), \quad G_2 = \{g^*(x) = 1\}
\]

the symmetric difference of \(G_1\) and \(G_2\). The rate of convergence has been studied in Manninen & Kyyrönen (1999).
1. **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 (1.1.15). Let us have classification data \((X_1, Y_1), \ldots, (X_N, Y_N)\) and let us construct classifier \(\hat{g}^n\) using the first part \((X_1, Y_1), \ldots, (X_{n-1}, Y_{n-1})\) of the data, where \(n \leq N\), \(n\) is typically \(n_1 := \lfloor n/2 \rfloor\). We can use

   \[
   \hat{g}^n(X) = \left\{ \begin{array}{ll}
   1, & \text{if } x < \theta_n, \\
   0, & \text{if } x \geq \theta_n
   \end{array} \right.
   \]

   as an estimator of \(P(\hat{g}(X) \neq Y)\), where \(\theta_n\) is constructed from the whole sample.

   We can also use cross validation, as in the case of regression function estimation in (1.1.16). In the three cases setting, when we have regression data \((X_1, Y_1), \ldots, (X_N, Y_N)\), it is natural to measure the performance of classification method by

   \[
   \sum_{k=1}^{N} \frac{1}{k} \sum_{i=1}^{k} \mathbb{1}(\hat{g}(X_{i+k-1}) \neq Y_{i+k-1}) \quad \text{(1.1.33)}
   \]

   where \(\hat{g}_{k}^n\) is a classifier constructed using the data \((X_1, Y_1), \ldots, (X_{k}, Y_{k})\), and \(k\) is chosen so large that the first classifier \(\hat{g}_{k}^n\) in the sequence is already a reasonable classifier. We can divide the classification error into \(K\) components

   \[
   \sum_{k=1}^{N} \frac{1}{k} \sum_{i=1}^{k} \mathbb{1}(\hat{g}(X_{i+k-1}) \neq Y_{i+k-1}) \quad \text{(1.1.34)}
   \]

   where \(k = 0, \ldots, K - 1\), which estimates \(P(\hat{g}(X) \neq Y | Y = \cdot)\).

1.1.0 **Confidence Intervals**

   We give first several definitions of a confidence interval for regression function estimation. Then we define confidence bands.

1.1.10.1 **Pointwise Confidence Intervals**

   A pointwise confidence interval \([a_0, a_1]\) for the estimation of regression function \(f : \mathbb{R}^d \rightarrow \mathbb{R}\) at point \(x \in \mathbb{R}^d\), with the confidence level \(1 - \alpha\), is such that for all \(x \in \mathbb{R}^d\), \(a_0, a_1\), for all \(a \in \mathbb{R}\), in a suitable subset of \(\mathbb{R}^d\), we have

   \[
   P^\beta(L < f(x) < U) > 1 - \alpha,
   \]

   where \(\beta\) is a collection of distributions of \((X, Y)\). Typically we can give asymptotic confidence intervals of type

   \[
   P^\beta(x_0 < \hat{f}(x) < U) > 1 - \alpha,
   \]
when \( n \to \infty \), we have, \( \lim_{n \to \infty} D_{n} Y_{n} \), is a sequence of independent. Asymptotic pivotal functions confidence intervals can typically be derived from the asymptotic distribution of the estimation. We have that

\[ se_{n} \left( \hat{f}(x) - f(x) \right) \xrightarrow{d} N(0, \sigma_{n}^{2}) \text{.} \]

where symbol, \( \xrightarrow{d} \), denotes the convergence in distribution, then for some choice

\[ X_{n} = \frac{\hat{f}(x) - f(x)}{\sigma_{n} \sqrt{n}} \text{.} \]

and

\[ U_{n} = \frac{\hat{f}(x) - f(x)}{\Phi^{-1}(\alpha)} \text{,} \]

where we denote \( \Phi^{-1}(\alpha) \) is the inverse of the standard normal distribution. That is, \( z_{\alpha} \) is the \( \alpha \) quantile of the \( N(0, 1) \) distribution. For \( X \sim N(0, 1) \), we have \( X \sim F_{x_{1}, x_{2}} \Rightarrow F_{x_{1}, x_{2}} \sim \Phi^{-1}(\alpha) \text{.} \)

More generally, we can use the term "level \( 1 - \alpha \) confidence interval" for the inequality

\[ P\left( X_{n} \leq \frac{\hat{f}(x) - f(x)}{\sigma_{n} \sqrt{n}} \leq U_{n} \right) \geq 1 - \alpha \text{.} \]

holds for all \( x \in D \). We can use the term "asymptotic level \( 1 - \alpha \) confidence interval" of

\[ \lim_{n \to \infty} P\left( X_{n} \leq \frac{\hat{f}(x) - f(x)}{\sigma_{n} \sqrt{n}} \leq U_{n} \right) \geq 1 - \alpha \text{.} \]

for all \( x \in D \). Note that in the asymptotic case it is important to distinguish between asymptotic level \( 1 - \alpha \) confidence intervals, which satisfies:

\[ \lim_{n \to \infty} P\left( \inf_{x \in D} X_{n} \leq \frac{\hat{f}(x) - f(x)}{\sigma_{n} \sqrt{n}} \leq U_{n} \right) \geq 1 - \alpha \text{.} \]

As pointed out by Wasserman (2000, p. 65), 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 linear regression. As mentioned in Wasserman (2000, Section 6.9) we can derive an approximate confidence interval for linear estimation under some assumptions. We noticed in (1.2) that many estimators can be written as linear estimators

\[ \hat{f}(x) = \sum_{k=1}^{n} \beta_{k}(x) Y_{k} = l(x)Y \text{,} \]

where \( l(x) = (l_{1}(x), \ldots, l_{n}(x))^{T} \) and \( y = (Y_{1}, \ldots, Y_{n})^{T} \). Let us assume that \( \hat{f}(x) \sim N(f(x), \text{Var}(\hat{f}(x))) \text{.} \)

\[ \text{Cov}(y) = \sigma^{2}I_{n} \text{.} \]

then

\[ \text{Var}(\hat{f}(x)) = l(x)^{T} \text{Cov}(y) l(x) = \sigma^{2}l(x)^{T} l(x) \text{.} \]

Then

\[ \left[ \hat{f}(x) - \phi l(x)^{T} l(x) \right] \text{ and } \left[ \phi l(x)^{T} l(x) \right] \text{ is the confidence interval} \]

\[ \left[ \hat{f}(x) - \phi l(x)^{T} l(x) \right] \text{ and } \left[ \phi l(x)^{T} l(x) \right] \text{ is the confidence interval} \]

where \( \alpha \) is the confidence level, \( 0 < \alpha < 1 \), and \( x_{1} \alpha/2 \) is the \( \alpha/2 \) quantile of the standard normal distribution.
1.16.2 Confidence Bands

A confidence band \((\hat{f}(x), \hat{V}(x))\), \(x \in A\), for the estimation of regression function \(f: \mathbb{R}^d \rightarrow \mathbb{R}\), for the set \(A \subset \mathbb{R}^d\), with the confidence level \(1 - c\), is such that

\[
P^x \left( \hat{f}(x) - \epsilon \leq f(x) \leq \hat{f}(x) + \epsilon \right) = 1 - c
\]

Confidence bands are called also simultaneous confidence bands, confidence envelopes, or probability bands. The confidence statement of the type

\[
P^x \left( \sup_{x \in A} \| f(x) - \hat{f}(x) \| \leq \epsilon \right) = 1 - c
\]

is equivalent to (1.135) if:

\[
\hat{f}(x) = \hat{f}(x) - c_n, \quad \hat{V}(x) = \hat{f}(x) + c_n.
\]

We can replace the supremum norm with some other function space norm to obtain confidence bands. For example, the \(L^2\) confidence band with the confidence level \(1 - c\) satisfies

\[
P^x \left( \| f(x) - \hat{f}(x) \|_2 \leq \epsilon \right) = 1 - c.
\]

A confidence band for the linear model is mentioned in Section 1.1.3.

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

\[
H_0: \beta_k = 0, \quad k = 1, \ldots, d
\]

for \(k = 1, \ldots, d\), and

\[
H_0: \beta_1 = 0, \ldots, \beta_d = 0.
\]

Testing of these hypothesis is considered in Section 2.0.5. 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

\[
H_0: f(x) = 0,
\]

when we assume that \(f(X) \neq 0\). We can use a test statistic \(W = \| \hat{f} \|\), where \(\hat{f}\)

is a nonparametric estimate of \(f\). The norm \(\| \cdot \|\) can be the \(L_2\) norm, a weighted
$\beta_y$ remains, on average, without functional space norm. Larger values of the test statistic $d(x)$ lead to the rejection of the null hypothesis. Here 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

$$c_k : = \frac{\partial}{\partial x_k} f(x) \in \mathbb{O}_k \quad \text{(1.138)}$$

for $k = 1, \ldots, d$. We can generalize the parameter restriction hypothesis (1.137) to the nonlinear case by

$$c_k : = \frac{\partial}{\partial x_k} f(x) \in \mathbb{O}_k, \quad \cdots, \quad \frac{\partial}{\partial x_d} f(x) \in \mathbb{O}_d.$$

We can test the null hypothesis (1.138) with the test statistics

$$
\chi^2 : = \left\| \frac{\partial}{\partial x_k} f(x) \right\|^2,
$$

where $\hat{f}$ is a nonparametric estimator of $f$ and $\| \cdot \|$ is a function space norm.

The distribution of the test statistic can be approximated by bootstrap. First we generate bootstrap samples from the original sample $(X_1, X_2), \ldots, (X_n, X_n)$. Then for each bootstrap sample $(X_1^*, X_2^*), \ldots, (X_n^*, X_n^*)$, the test statistic $\chi^2$ is calculated. We obtain a sequence $\chi_1^2, \ldots, \chi_n^2$ of values of the test statistic. Now we use the empirical quantile of the sequence of the values of the test statistics. We can reject the null hypothesis at level 0 < $\alpha$ < 1, if the observed value $\hat{\chi}^2$ of the test statistic satisfies $\hat{\chi}^2 \geq \chi_{1-\alpha}^2$.

Fiebig and Kiefer (1993) have proposed the wild bootstrap. Here the regression function $f$ is estimated with $\hat{f}$ (under the null hypothesis). Then the residuals $e_i = Y_i - \hat{f}(X_i)$ are calculated. Finally, the bootstrap residual $e_i^*$ is generated from a distribution of the residuals $e_i^* = 0, 0, (e_i^*)^2 = e_i^2$, and $(e_i^*)^2 = e_i^2$. The bootstrap samples is $(X_1, e_1^*), \ldots, (X_n, e_n^*)$, where $X_i^* = \hat{f}(X_i) + e_i^*$.  

...