ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ
ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ

Χρυσούλα Κιοσσέ

Εκτιμηση της υπο συνθηκη συναρτησης κατανομης και επιλογη ευρους ζωνης για τυχαια δεξια λογοκριμενα δεδομενα

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΑΤΡΙΒΗ

Ιωάννινα, 2023

UNIVERSITY OF IOANNINA

Department of Mathematics

Chrisoula Kiosse

Conditional distribution function estimation and bandwidth selection for randomly right censored data

---

Master's Thesis

Ioannina, 2023

To my family

Η παρούσα Μεταπτυχιακή Διατριβή εκπονήθηκε στο πλαίσιο των σπουδών για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Στατιστική και Επιχειρησιακή Έρευνα που απονέμει το Τμήμα Μαθηματικών του Πανεπιστημίου Ιωαννίνων.

Εγκρίθηκε την 20/11/23 από την εξεταστική επιτροπή:

Ονοματεπώνυμο	**Βαθμίδα**
Δημήτριος Μπάγκαβος	Επίκουρος Καθηγητής
Μπατσίδης Απόστολος	Αναπληρωτής Καθηγητής
Ζωγράφος Κωνσταντίνος	Καθηγητής

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ

"Δηλώνω υπεύθυνα ότι η παρούσα διατριβή εκπονήθηκε κάτω από τους διεθνείς ηθικούς και ακαδημαϊκούς κανόνες δεοντολογίας και προστασίας της πνευματικής ιδιοκτησίας. Σύμφωνα με τους κανόνες αυτούς, δεν έχω προβεί σε ιδιοποίηση ξένου επιστημονικού έργου και έχω πλήρως αναφέρει τις πηγές που χρησιμοποίησα στην εργασία αυτή."

Χρυσούλα Κιοσσέ

Ευχαριστιες

Περιληψη

Η μοντελοποίηση της εκτιμήτριας Kaplan-Meier με τη μέθοδο της τοπικής πολυωνυμικής προσαρμογής για τυχαία λογοκριμένα δεδομένα από δεξιά, για αυθαίρετο αριθμό συνεχών συμμεταβλητών διερευνάται λεπτομερώς. Οι στατιστικές ιδιότητες της εκτιμήτριας συνάρτησης ποσοτικοποιούνται αναλυτικά και η εφαρμογή της στην πράξη γίνεται εφικτή με την ανάπτυξη αντίστοιχης μεθόδου επιλογής εύρους ζώνης που βασίζεται αποκλειστικά και μόνο στα δεδομένα. Οι επιδόσεις της μεθόδου μελετώνται βάσει προσομοιώσεων με δείγματα πεπερασμένου μεγέθους από γνωστές κατανομές. Τέλος, η μέθοδος χρησιμοποιείται στην ανάλυση ενός πραγματικού συνόλου δεδομένων και έτσι αναδεικνύεται η πρακτική χρησιμότητα της μεθόδου.

Abstract

The local polynomial modeling of the Kaplan–Meier estimate for random designs under the right censored data setting in the presence of an arbitrary number of continuous covariates is investigated in detail. The statistical properties of the derived estimate is quantified analytically and its implementation is facilitated by the development of corresponding data driven bandwidth selector. Numerical evidence is also provided on its finite sample performance. A real life data analysis illustrates how the methodological advances proposed herein help to generate additional insights in comparison to existing methods.

Contents

CHAPTER 1

Estimation of Survival Function

1.1 Introduction

Survival analysis is a branch of statistics for analyzing the expected duration of time until one event occurs, such as death in biological organisms and failure in mechanical systems. Survival Analysis is used to estimate the lifespan of a particular population under study. It is also called "Time to Event" Analysis as the goal is to estimate the time for an individual or a group of individuals to experiences an event of interest. While the event of interest is often death (in this case we study the time to death for patients having a specific disease) or recurrence (in this case we study the time to relapse of a certain disease), it is not limited to the fields of medicine and epidemiology. In fact, it can be used in many domains. For example, we may also analyze the time until: finding a new job after a period of unemployment, the failure of a mechanical system or a machine, a bank or a company goes bankrupt, a customer buys a new product or stops its current subscription, a taxi picks you up after having called the taxi company and an employee leaves the company. As is evident, the event of interest does not necessarily have to be a death or a disease, but in all situations we are interested in the time until a specific event occurs. There are unique features of time to event variables. First, time to event is always positive and the distribution of such variables is often skewed. For example, in a study assessing time to relapse in high risk patients, the majority of events (relapses) may occur early in the follow up with very few occurring later. On the other hand, in a study of time to death in a community based sample, the majority of events (deaths) may occur later in the follow up. Standard statistical procedures that assume normality of distributions do not apply. Non-parametric procedures could be invoked except for the fact

that there are additional issues. Specifically, complete data (actual time to event data) is not always available on each participant in a study. In many studies, participants are enrolled over a period of time (months or years) and the study ends on a specific calendar date. Thus, participants who enroll later are followed for a shorter period than participants who enroll early. Some participants may drop out of the study before the end of the follow-up period (e.g., move away, become disinterested) and others may die during the follow-up period (assuming the outcome of interest is not death). In each of these instances, we have incomplete follow up information. These times are called censored times. Additionally, when the event is not yet observed at the end of the study (i.e., the survival duration is greater than the observed duration), this is referred as right-censoring and it is the type of censoring that we will focus on this thesis. Survival analysis attempts to answer certain questions, such as what is the proportion of a population which will survive past a certain time? Of those that survive, at what rate will they die or fail? Can multiple causes of death or failure be taken into account? How do particular circumstances or characteristics increase or decrease the probability of survival? The goal of survival analysis is thus to model and describe time-to-event data in an appropriate way, taking the particularities of this type of data into account.
Firstly, in this thesis, in order to estimate the survival function in the presence of censorship, we use the Maximum Likelihood Estimator, in Section 1.3, for which it is necessary to know the distribution of the data in advance. When we do not know the distribution of the data, we use the Kaplan-Meier estimator, discussed in Section 1.4, which gives us a discontinuous estimate. This discontinuity is addressed by a modification of the estimator using kernel smoothing which is given in Section 2.3. Furthermore, when we do not have censoring we can estimate the cumulative distribution function with kernel estimation. In case we have a multivariate variable with two or more factors we can also estimate the distribution with the multivariate kernel estimation. In both instances the choice of the bandwidth of the estimator is of crucial importance. In Chapter 3 in order to estimate the conditional distribution function we present a new estimator whose asymptotic properties are developed in Chapter 4. To estimate the distribution of the covariates of the estimator we use the kernels where when we have one covariate we deal with the univariate case while otherwise we deal with the multivariate one. In Chapter 5 we give the optimal bandwidth choice, first for the univariate case and then for the multivariate. Finally, in Chapter 6 we provide a real data example to illustrate the performance of our estimator using two covariates.

1.2 Survival analysis and types of data

The important difference in data analysed in survival analysis is the presence of censoring. Censored data is any data for which we do not always know exactly when the event occurred. There are three main types of censored data: right censored, left censored and interval censored, see Turkson et al. (2021). The time T which measures the duration of the event of interest for the population under study is a random variable. T is considered to be left-censored if it is less than the censoring time U, which is again a random variable. In other words, for left-censoring to occur, the event of interest must occur for the subject before that person is observed in the study, i.e. $T < U$. For such subjects, we know that they have experienced the event sometime before time U, but the exact time is not known. The exact time will be known if and only if T is greater than or equal to U. The data from a left-censored sampling scheme can be represented as a pair of random variables (Y, δ), $Y = \max(T, U)$ and

$$\delta = \begin{cases} 0, & \text{if} \quad U \leq T, \\ 1, & \text{if} \quad U > T. \end{cases}$$

When a specific subject is followed for a while, gets lost to follow-up, and returns and continues being studied, is said to be interval-censored. In this case it holds that $U < T < B$, where B is a censoring time, which is again a random variable. In interval-censoring, the observed data consists of intervals $I_1, I_2, \ldots, I_n$, where for each $i = 1, 2, \ldots, n$, the i^{th} response lies in the interval I_i. In this case, an uncensored observation of an observed death corresponds to an observed interval consisting of a single point. Suppose we have the situation where we have performed a test on a subject at time y_1 and the subject tested negative. Subsequently, at time $y_2 > y_1$, the subject was tested positive. In this scenario, we know the subject was exposed to the event of study sometime between y_1 and y_2, but we do not know the exact time of the exposure. For example, if in a clinical trial, the time to remission has been assessed, then if the i^{th} patient is in remission at, say, the 8^{th} week after the trail, but was absent for future check-ups, and resurfaces and was out of remission on the 11^{th} week, then $I_i = [8\ ,\ 11)$ is the i^{th} patient's censoring interval or length of remission.

Right censored data is data for items that have not yet failed. They are considered "still alive" as their failure time has not yet occurred, though it is expected to occur at some point in the future. In particular, for a specific subject under study, if we assume that there is a time T and a censoring

time U, the T's are independent and identically distributed, i.i.d. for short, with probability density function (p.d.f.) $f_T(t)$ and survival function $S_T(t) = 1 - F_T(t)$, where $F_T(t)$ is the cumulative distribution function (c.d.f.). The exact lifetime T of a subject will be known, if and only if T is less than or equal to U; if T is greater than U, the subject is a survivor and the event time is censored at U. The data from this experiment can be represented as a pair of random variables (Y, δ), $Y = \min(T, U)$ and

$$\delta = \begin{cases} 0, & \text{if} \quad T > U, \\ 1, & \text{if} \quad T \leq U. \end{cases}$$

It follows from the above that left-censoring is a special case of right-censoring with the time axis reversed. It is because of this phenomenon that there have been few techniques developed explicitly for left-censored data.
Some reasons for right-censoring include: study ends without subject experiencing the event; the subject is lost to follow up within the study period; subject deliberately withdraws the treatment variable; the subject is obliged to withdraw from the treatment due to reasons beyond their control and subject withdraws from the study due to another reason (i.e., death, if death is not the event of interest).
In addition, double-censoring occurs as a result of a combination of left and right-censoring. In this case, we note that $Y = \max\{\min(T, y), l\}$, where l and y are, respectively, the left and right-censoring times associated with T and $l < y$. In this case, T is only observed if it falls in the interval $[l, y]$. Otherwise, one of the endpoints of the interval is observed and the other endpoint probably remains undisclosed. We should also note that double-censoring is not the same as interval-censoring. The above types of censoring are depicted in Fig. 1.1. In practice, the most common type and the one that we will focus on in this thesis is right censored data.

In addition to the above data types there is Type I and Type II censoring. Type I censoring occurs when a study is designed to end at a fixed time point U. At the end of the study period, any subject that did not experience the event is censored. In type I censoring, the number of uncensored observations is a random variable. This type of censoring is also called "right censored" since the times of failure to the right are missing. Another way to design a study is to assume that for a given sample $Y_1, Y_2, \ldots, Y_n$, only the first $r < n$ lifetimes are observed. The value of r is fixed beforehand. This is Type II censoring scheme. For example, one might employ $n = 100$ units in the study and then test until at least half of them fail. Then $r = 50$, but U is unknown until the 50^{th} failure occurs. Type II censoring has the significant advantage

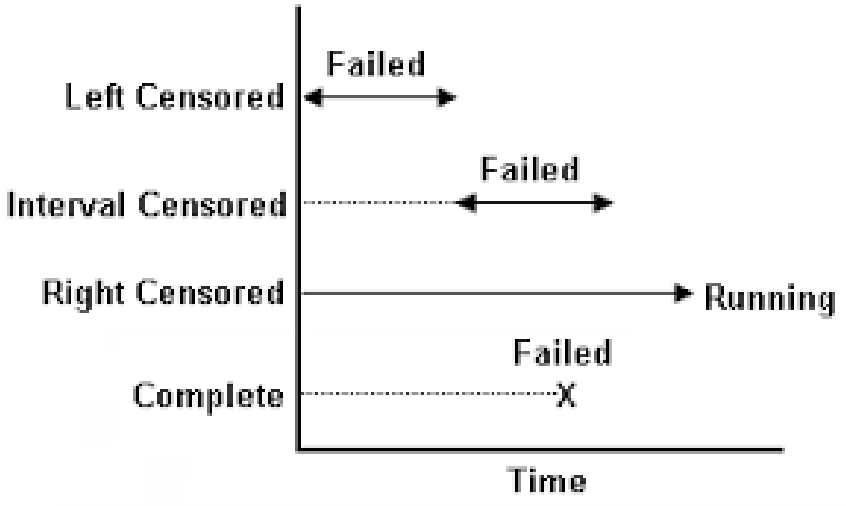

Figure 1.1: Main types of censored data, see `https://help.reliasoft.com/weibull20/index.htm#t=data_types.htm`.

of knowing in advance how many failure times the test will yields. This helps enormously when planning adequate tests. However, an open-ended random test time is generally impractical from a management point of view and this type of testing is rarely seen.
Assume a random sample of i.i.d. observations $T_1, \ldots, T_n$ from c.d.f. F_T and p.d.f f_T. For the survival function $S_T(t)$ of the T population, see Klein and Moeschberger (2003), we have

$$S_T(t) = P(T \geq t) = \int_t^\infty f_T(x)dx,$$

and it expresses the probability that a certain object of interest will survive beyond a certain time t. As presented later on, this definition is frequently met in data analysis.
The random variable T refers to time until the occurrence of an event. For instance, it might refer to the age that an individual passes away. Equivalently, the survival function can be interpreted as the probability that an individual is still alive after age t. For example, if $S_T(60) = 0.8$, it means that there are 80% of the individuals in the population who will still be alive at the age of 60.

The basic properties of S_T are summarized below.

1) $S_T(0) = 1$, i.e. nothing has occurred before time zero.

2) $S_T(\infty) = 0$, i.e. the event of interest will certainly occur in $[0, \infty)$.

3) $S_T(t)$ is non-increasing as a function of $t \in [0, +\infty)$.

4) When T is discrete random variable, then $S_T(t)$ is right-continuous as a function of t.

Often, the p.d.f. of the available sample is unknown and therefore it should be estimated. There are many estimation methods, which are generally divided to the parametric and the non-parametric approaches, respectively. The parametric approach relies on assuming a certain distribution for the available data set in hand. As a result estimation under the parametric framework reduces in estimation of the assumed distribution parameters. The non-parametric approach makes no assumptions about the distribution of the underlying population. The focus here is the non-parametric approach but before elaborating, we briefly discuss the most popular parametric method in the next section.

1.3 Maximum likelihood estimation for censored data

Under the parametric approach we assume that we have available a sample of n observations from the population under study (Y_i, δ_i), $i = 1, \ldots, n$. $Y_i = \min(T_i, U_i)$, where for all i it holds that $T_i \sim f_T$, $U_i \sim f_U$, $T - i$, U_i are independent and we are interested in estimating the p.d.f. f_T or equivalently the survival function $S_T = 1 - F_T$. Obviously, in order to estimate $S_T(y)$ we cannot simply replace the available data in the formula of the survival function as it holds that

$$\begin{aligned} S_Y(y) &= P(Y \geq y) = P(\min(T, U) \geq Y) \\ &= P((T \geq Y) \quad \textit{and} \quad (U \geq Y)) \\ &= P(T \geq Y) P(U \geq Y) \\ &= S_T(y) S_U(y). \end{aligned}$$

The above equation shows that if we apply the definition of the survival function to the available data, we estimate the product of the survival functions of T and U and not the survival function $S_T(y)$ as we would like, see Lawless

(2011).
The available data (Y_i, δ_i), $i = 1, \ldots, n$ actually come from the joint density of (Y, δ). Now, the joint density (Y, δ) is

$$f(Y, \delta) = \lim_{h \to 0} \frac{P(y \leq Y \leq y + h, \delta = \delta^*)}{h}, \quad y \geq 0, \quad \delta^* = 0, 1. \tag{1.1}$$

The p.d.f. $f(Y, \delta)$ can take two forms depending on the value of δ. In the first case $\delta = 1$, which means that $T \leq U$, we conclude that $Y = \min(T, U) = T$. Thus, using P14 from the Appendix we get for $k \in (y, y + h)$

$$\begin{aligned} P(y \leq Y \leq y + h, \delta = 1) &= P(y \leq T \leq y + h, U \geq T) \\ &\simeq P(y \leq T \leq y + h, U \geq y) \\ &\overset{(P14)}{=} P(y \leq T \leq y + h) P(U \geq y) \\ &= f_T(k) h \big\{1 - F_U(y)\big\}. \end{aligned}$$

Hence,

$$\begin{aligned} f(y, \delta = 1) &= \lim_{h \to 0} \frac{f_T(k) h \big\{1 - F_U(y)\big\}}{h} \\ &= f_T(y) \big\{1 - F_U(y)\big\} \\ &= f_T(y) S_U(y). \end{aligned} \tag{1.2}$$

In the second case $\delta = 0$ which means that $T > U$. Hence $Y = \min(T, U) = U$ and thus, using P14 from the Appendix we get for $k \in (y, y + h)$

$$\begin{aligned} P(y \leq Y \leq y + h, \delta = 0) &= P(y \leq U \leq y + h, T > U) \\ &\simeq P(y \leq U \leq y + h, T \geq y) \\ &\overset{(P14)}{=} P(y \leq U \leq y + h) P(T \geq y) \\ &= f_U(k) h \big\{1 - F_T(y)\big\}. \end{aligned}$$

Thus,

$$\begin{aligned} f(y, \delta = 0) &= \lim_{h \to 0} \frac{f_U(k) h \big\{1 - F_T(y)\big\}}{h} \\ &= f_U(y) \big\{1 - F_T(y)\big\} \\ &= f_U(y) S_T(y). \end{aligned} \tag{1.3}$$

Taking into account (1.2) and (1.3) we have from (1.1), that the joint density of (Y, δ) is

$$f(Y, \delta) = \big\{f_T(y) S_U(y)\big\}^{\delta} \big\{f_U(y) S_T(y)\big\}^{1-\delta} = f_T(y)^{\delta} S_T(y)^{1-\delta} f_U(y)^{1-\delta} S_U(y)^{\delta}.$$

As in the case of complete data, here too, every available observation, censored or uncensored, contributes information to the likelihood. In more detail, we have to take into consideration that the p.d.f. of (Y, δ) has a continuous component (Y) and a binary component (δ). So, for fixed i, $(Y_i, \delta_i) = (T_i, 1)$ for T_i uncensored at t and $(Y_i, \delta_i) = (U_i, 0)$ for $T_i > U_i$ censored at t. The corresponding probabilities for these two cases are

$$\begin{aligned} P(Y_i, \delta_i = 1) &= P(Y_i = T_i | \delta_i = 1) P(\delta_i = 1) = P(Y_i = T_i | T_i \leq U_i) P(T_i \leq U_i) \\ &= \frac{f_T(Y_i)}{1 - S_T(Y_i)} \big\{1 - S_T(Y_i)\big\} = f_T(Y_i), \end{aligned}$$

$$P(Y_i, \delta_i = 0) = P(Y_i = U_i | \delta_i = 0) P(\delta_i = 0) = P(T_i \geq U_i) = S_T(Y_i).$$

The Maximum Likelihood Estimation (MLE) method first estimates the scalar parameter θ (or the parameter vector θ) of the underlying p.d.f. f_T by maximizing the likelihood function which is given by

$$L(\theta) = \prod_{i=1}^{n} L_i(\theta) = \Big\{ \prod_{\delta_i = 1} L_i(\theta) \Big\} \Big\{ \prod_{\delta_i = 0} L_i(\theta) \Big\} = \prod_{i=1}^{n} f_T(y_i; \theta)^{\delta_i} S_T(y_i; \theta)^{1 - \delta_i}.$$

The result of the above maximization procedure is substituted to the assumed p.d.f. f_T to yield the parametric estimate of the underlying density.

Example 1.3.1. Suppose we have a sample of censored observations from an exponential distribution. Then the likelihood is

$$L(\theta) = \prod_{i=1}^{n} \big\{\theta \exp(-\theta y_i)\big\}^{\delta_i} \big\{\exp(-\theta y_i)\big\}^{1 - \delta_i}.$$

The log-likelihood is

$$\begin{aligned}
\ln L(\theta) &= \ln\Big[\prod_{i=1}^{n}\big\{\theta\exp(-\theta y_i)\big\}^{\delta_i}\big\{\exp(-\theta y_i)\big\}^{1-\delta_i}\Big] \\
&= \ln\Big[\prod_{i=1}^{n}\big\{\theta\exp(-\theta y_i)\big\}^{\delta_i}\Big] + \ln\Big[\prod_{i=1}^{n}\big\{\exp(-\theta y_i)\big\}^{1-\delta_i}\Big] \\
&= \sum_{i=1}^{n}\delta_i\ln\big\{\theta\exp(-\theta y_i)\big\} + \sum_{i=1}^{n}(1-\delta_i)\ln\big\{\exp(-\theta y_i)\big\} \\
&= \sum_{i=1}^{n}\delta_i\Big[\ln(\theta) + \ln\big\{\exp(-\theta y_i)\big\}\Big] + \sum_{i=1}^{n}(1-\delta_i)\ln\big\{\exp(-\theta y_i)\big\} \\
&= \sum_{i=1}^{n}\delta_i\Big[\ln(\theta) - \theta y_i\Big] - \sum_{i=1}^{n}(1-\delta_i)\theta y_i \\
&= \ln(\theta)\sum_{i=1}^{n}\delta_i - \theta\sum_{i=1}^{n}\delta_i y_i - \theta\sum_{i=1}^{n}(1-\delta_i)y_i.
\end{aligned}$$

Differentiating $\ln L(\theta)$ with respect to θ, so equal to 0 and solving for θ yields that the MLE of θ, say $\hat{\theta}$, is

$$\hat{\theta} = \frac{\sum_{i=1}^{n}\delta_i}{\sum_{i=1}^{n} y_i}. \tag{1.4}$$

Generating 30 observations from the Exponential distribution Exp(7) and 30 censoring times from the Exponential distribution Exp(3) yields 30 right censored observations with 30% censoring. Then the MLE of the underlying survival function is $\exp(-\hat{\theta}y)$ with $\hat{\theta}$ given in (1.4) and its results depicted in Fig. 1.2. As we can see the estimation is very close since the true parameter

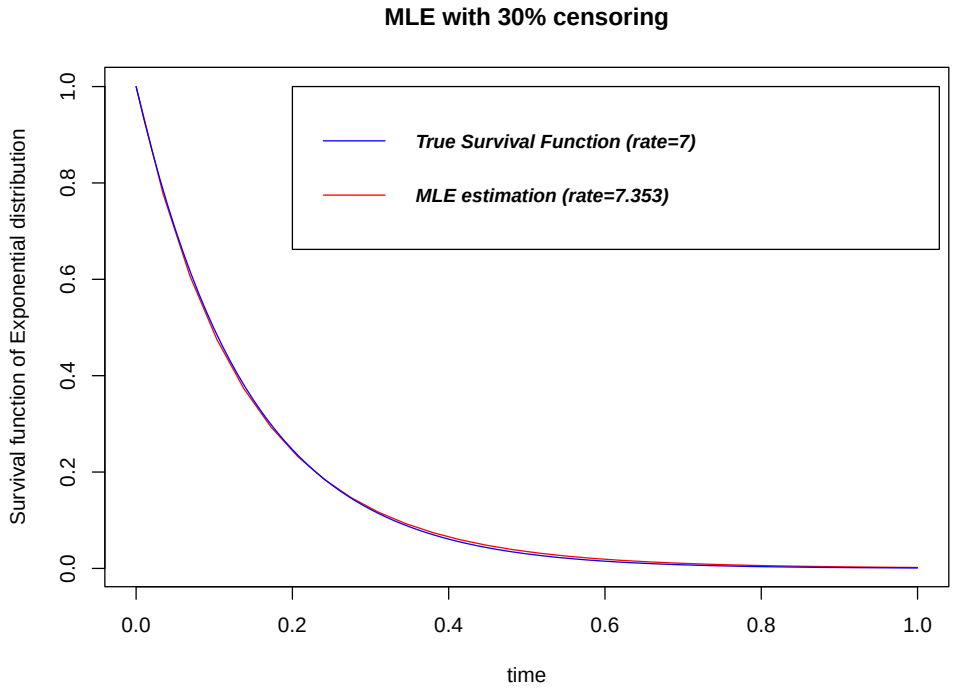

Figure 1.2: MLE estimation for the Survival function of the Exp(7) with 30% censoring

is equal to 7 and the estimated one is equal to 7.353.

Example 1.3.2. Suppose we have a sample of censored observations from a Weibull distribution. Then the likelihood is

$$L(a,b) = \prod_{i=1}^{n}\Bigg(\Big[\frac{b}{a}\Big(\frac{y_i}{a}\Big)^{b-1}\exp\Big\{-\Big(\frac{y_i}{a}\Big)^{b}\Big\}\Big]^{\delta_i}\Big[\exp\Big\{-\Big(\frac{y_i}{a}\Big)^{b}\Big\}\Big]^{1-\delta_i}\Bigg).$$

Then, the log-likelihood is

$$\begin{aligned}
\ln L(a,b) &= \ln\Bigg\{\prod_{i=1}^{n}\Bigg(\Big[\frac{b}{a}\Big(\frac{y_i}{a}\Big)^{b-1}\exp\Big\{-\Big(\frac{y_i}{a}\Big)^{b}\Big\}\Big]^{\delta_i}\Big[\exp\Big\{-\Big(\frac{y_i}{a}\Big)^{b}\Big\}\Big]^{1-\delta_i}\Bigg)\Bigg\}\\
&= \ln\Bigg(\prod_{i=1}^{n}\Big[\frac{b}{a}\Big(\frac{y_i}{a}\Big)^{b-1}\exp\Big\{-\Big(\frac{y_i}{a}\Big)^{b}\Big\}\Big]^{\delta_i}\Bigg)\\
&\quad + \ln\Bigg(\prod_{i=1}^{n}\Big[\exp\Big\{-\Big(\frac{y_i}{a}\Big)^{b}\Big\}\Big]^{1-\delta_i}\Bigg)\\
&= \sum_{i=1}^{n}\ln\Big[\frac{b}{a}\Big(\frac{y_i}{a}\Big)^{b-1}\exp\Big\{-\Big(\frac{y_i}{a}\Big)^{b}\Big\}\Big]^{\delta_i} + \sum_{i=1}^{n}\ln\Big[\exp\Big\{-\Big(\frac{y_i}{a}\Big)^{b}\Big\}\Big]^{1-\delta_i}\\
&= \sum_{i=1}^{n}\delta_i\ln\Big[\frac{b}{a}\Big(\frac{y_i}{a}\Big)^{b-1}\exp\Big\{-\Big(\frac{y_i}{a}\Big)^{b}\Big\}\Big]\\
&\quad + \sum_{i=1}^{n}(1-\delta_i)\ln\Big[\exp\Big\{-\Big(\frac{y_i}{a}\Big)^{b}\Big\}\Big]\\
&= \sum_{i=1}^{n}\delta_i\Big[\ln(b) - \ln(a) + (b-1)\big\{\ln(y_i) - \ln(a)\big\} - \Big(\frac{y_i}{a}\Big)^{b}\Big]\\
&\quad - \sum_{i=1}^{n}(1-\delta_i)\Big(\frac{y_i}{a}\Big)^{b}.
\end{aligned}$$

Differentiating $\ln L(a,b)$ with respect to a and to b, equal to zero and solving the system of equations, no closed form expression is obtained.

Generating 30 observations from the Weibull distribution Weib$(1.1, 1.7)$ and 30 censoring times from the Weibull distribution Weib$(2, 3)$ yields 30 right censored observations with 30% censoring. Then after 6 iterations of the Newton's method we get the results depicted in Fig. 1.3. As we can see the

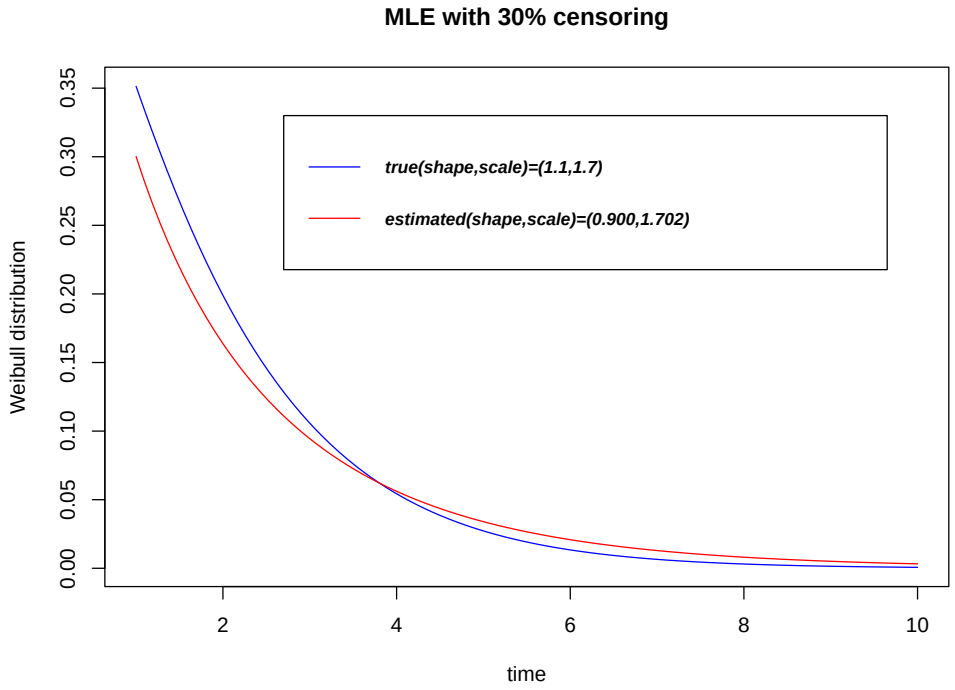

Figure 1.3: MLE estimation for Weibull distribution with 30% censoring

estimation is very close since the true shape and scale parameters are equal to 1.1 and 1.7 respectively and the estimated are equal to 0.900 and 1.702.

The following theorem, see Rao (2017), shows that the MLE estimate in the random right censorship setting is a consistent estimate of the true parameter θ.

Theorem 1.3.1. *Suppose* $\{Y_i\}$ *be i.i.d. random variables with density* $f(Y;\theta)$. *If* θ *is the real parameter and* $\hat{\theta}$ *the MLE then under almost sure uniform convergence of the likelihood we have*

$$\hat{\theta} \overset{a.s.}{\to} \theta.$$

Proof. Define $l(\theta) = \mathrm{E}[\log f(Y;\theta)]$ (the expected log-likelihood). To prove the result we first need to show that the expectation of the maximum likelihood is maximum at the true parameter and that this is the unique maximum. In other words we need to show that $\mathrm{E}\{n^{-1}L(\theta_k) - n^{-1}L(\theta)\} \leq 0$ for all $\theta_k \in \Theta$. To do this, we have

$$\begin{aligned} l(\theta_k) - l(\theta) &= \mathrm{E}\Big\{\frac{1}{n}L(\theta_k)\Big\} - \Big\{\frac{1}{n}L(\theta)\Big\} = \int \log \frac{f(y;\theta_k)}{f(y;\theta)} f(y;\theta)dy \\ &= \mathrm{E}\Big\{\log \frac{f(Y;\theta_k)}{f(Y;\theta)}\Big\} \leq \log \mathrm{E}\Big\{\frac{f(Y;\theta_k)}{f(Y;\theta)}\Big\} \\ &= \log \int \frac{f(y;\theta_k)}{f(y;\theta)} f(y;\theta)dy = \log \int f(y;\theta_k)dy \\ &= \log 1 = 0. \end{aligned}$$

$f(y;\theta_k) = f(y;\theta)$ for all y only when θ and no other function of f gives equality. Hence only when $\theta_k = \theta$ do we have

$$\mathrm{E}\Big\{\log \frac{f(Y;\theta_k)}{f(Y;\theta)}\Big\} = \log \int \frac{f(y;\theta_k)}{f(y;\theta)} f(y;\theta)dy = \log \int f(y;\theta_k)dy = 0,$$

thus, $\mathrm{E}\{n^{-1}L(Y;\theta_k)\}$ has a unique maximum at θ.
Finally, we need to show that $\hat{\theta} \overset{a.s.}{\to} \theta$. To simplify notation for the remainder of this proof we assume the likelihood has been standardized by n i.e

$$L_n(\theta_k) = \frac{1}{n}\sum_{i=1}^{n} \log f(Y_i;\theta_k).$$

We note that since $l(\theta_k)$ is maximum at θ if $|L_n(\hat{\theta}) - l(\theta)| \overset{a.s.}{\to} 0$, then $\hat{\theta} \overset{a.s.}{\to} \theta$. Thus we need to prove θ if $|L_n(\hat{\theta}) - l(\theta)| \overset{a.s.}{\to} 0$. We do this using a sandwich argument. First we note for every MLE $\hat{\theta}$

$$L_n(Y;\theta) \leq L_n(Y;\hat{\theta}) \overset{a.s.}{\to} l(\hat{\theta}) \leq l(\theta), \tag{1.5}$$

where we are treating $\hat{\theta}$ as if it were a non-random fixed value in Θ. Returning to $|\mathrm{E}\{L_n(Y;\theta)\} - L_n(Y;\hat{\theta})|$ we note that the difference can be written as

$$l(\theta) - L_n(Y;\hat{\theta}) = \{l(\theta) - L_n(Y;\theta)\} + \{l(\hat{\theta}) - L_n(Y;\hat{\theta})\} + \{L_n(Y;\theta) - l(\hat{\theta})\}.$$

Now, by using (1.5) we have

$$\begin{aligned} l(\theta) - L_n(Y;\hat{\theta}) &\leq \big\{l(\theta) - L_n(Y;\theta)\big\} + \big\{l(\hat{\theta}) - L_n(Y;\hat{\theta})\big\} + \big\{L_n(Y;\hat{\theta}) - l(\hat{\theta})\big\} \\ &= l(\theta) - L_n(Y;\theta), \end{aligned}$$

and

$$\begin{aligned} l(\theta) - L_n(Y;\hat{\theta}) &\geq \big\{l(\theta) - L_n(Y;\theta)\big\} + \big\{l(\hat{\theta}) - L_n(Y;\hat{\theta})\big\} + \big\{L_n(Y;\theta) - l(\theta)\big\} \\ &= l(\hat{\theta}) - L_n(Y;\hat{\theta}). \end{aligned}$$

Thus,

$$l(\hat{\theta}) - L_n(Y;\hat{\theta}) \leq l(\theta) - L_n(Y;\hat{\theta}) \leq l(\theta) - L_n(Y;\theta).$$

Therefore, we have that

$$|l(\theta) - L_n(Y;\hat{\theta})| \leq \sup_{\theta\in\Theta} |l(\theta_k) - L_n(Y;\theta_k)| \xrightarrow{a.s.} 0.$$

Since $\mathrm{E}\big\{L_n(Y;\theta_k)\big\}$ has a unique maximum at $\mathrm{E}\big\{L_n(Y;\theta)\big\}$ this implies $\hat{\theta} \xrightarrow{a.s.} \theta$. □

Additionally, the MLE estimate $\hat{\theta}$ of θ is asymptotically normal as shown in the next theorem, see Rao (2017).

Theorem 1.3.2. *Assume that for the right censored data* $Y_i = \min(T_i, U_i)$, $i = 1, \ldots, n$, *the censoring times are constant with* $U_1 = \cdots = U_n = c$. *The maximum likelihood estimator* $\hat{\theta} = arg\max\big\{L(\theta)\big\}$ *satisfies*

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, I(\theta)^{-1}),$$

where $I(\theta)$ *is the Fisher's information measure given by*

$$I(\theta) = -\mathrm{E}\bigg\{\frac{1}{n}\sum_{i=1}^{n}\delta_i \frac{\partial \log f_T(Y_i,\theta)}{\partial\theta^2} + \frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)\frac{\partial \log S_T(c_i,\theta)}{\partial\theta^2}\bigg\}.$$

Proof. We recall that Y_i are i.i.d. random variables, then

$$\frac{1}{\sqrt{n}}\frac{\partial L_n(Y;\theta_k)}{\partial\theta_k}\Big|_\theta = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log f(Y_i;\theta_k)}{\partial\theta_k}\Big|_\theta,$$

is the sum of independent random variables. We note that

$$\mathrm{E}\bigg\{\frac{\partial \log f(Y_i;\theta_k)}{\partial\theta_k}\Big|_\theta\bigg\} = \int \frac{\partial \log f(y;\theta_k)}{\partial\theta_k}\Big|_\theta f(y;\theta)dy = 0,$$

thus, $\frac{\partial \log f(Y_i;\theta_k)}{\partial \theta_k}\Big|_\theta$ is a zero mean random variable and its variance is $I(\theta)$. Hence, $\frac{\partial L_n(Y;\theta_k)}{\partial \theta_k}\Big|_\theta$ is the sum of i.i.d. random variables with mean zero and variance $I(\theta)$. Therefore, we have

$$\frac{1}{\sqrt{n}}\frac{\partial L_n(Y;\theta_k)}{\partial \theta_k}\Big|_\theta \xrightarrow{d} N(0, I(\theta)). \tag{1.6}$$

By (1.6) and Taylor's theorem we obtain

$$\frac{1}{n}\frac{\partial L_n(Y;\theta_k)}{\partial \theta_k}\Big|_{\hat\theta} = \frac{1}{n}\frac{\partial L_n(Y;\theta_k)}{\partial \theta_k}\Big|_\theta + (\hat\theta - \theta)\frac{1}{n}\frac{\partial^2 L_n(Y;\theta_k)}{\partial {\theta_k}^2}\Big|_{\tilde\theta}, \tag{1.7}$$

for some real number $\tilde\theta$ between $\hat\theta$ and θ.

$$\begin{aligned} I(\theta) &= -\mathrm{E}\Big\{\frac{\partial^2}{\partial\theta^2}\log f(Y|\theta)\Big\} = -\frac{\partial^2}{\partial\theta^2}\mathrm{E}\big\{n^{-1}\log L(\theta)\big\} \\ &= -\frac{\partial^2}{\partial\theta^2}\frac{1}{n}\mathrm{E}\Big\{\sum_{i=1}^{n}\delta_i \log f_T(y_i;\theta) + \sum_{i=1}^{n}(1-\delta_i)\log S_T(c_i;\theta)\Big\}. \end{aligned}$$

The consistency result in Theorem 1.3.1, (1.3) and (1.7), give

$$\frac{1}{n}\frac{\partial L_n(Y;\theta_k)}{\partial \theta_k}\Big|_\theta - I(\theta)(\hat\theta-\theta) + \left\{\frac{1}{n}\frac{\partial^2 L_n(Y;\theta_k)}{\partial {\theta_k}^2}\Big|_{\tilde\theta} - I(\theta)\right\}(\hat\theta-\theta) = 0. \tag{1.8}$$

Multiplying (1.8) by $\sqrt{n}$ and rearranging gives

$$\sqrt{n}(\hat\theta-\theta) = I(\theta)^{-1}\frac{1}{\sqrt{n}}\frac{\partial L_n(Y;\theta_k)}{\partial \theta_k}\Big|_\theta + \mathrm{o}_p(1). \tag{1.9}$$

By substituting (1.6) into (1.9) finishes the proof. □

Note that for $c = 0$, i.e. when all observations are censored, the Fisher information measure is 0 so the asymptotic variance of the estimator is not finite. This is expected based on the interpretation of the measure. Although, its variance will be greater than the variance of a corresponding estimator with complete data.
Next, we turn our attention to the non-parametric estimation approach. We start with the Kaplan-Meier which is the most popular estimator of the survival function in the right censored data setting.

1.4 Kaplan-Meier estimator

The Kaplan-Meier estimator was introduced in Kaplan and Meier (1958) and it is commonly referred to as the Product-Limit Estimator. Since then, it has been the most commonly-used estimator in medical/public health studies involving failure time data. The estimator may be derived heuristically by partitioning the time axis into a number of small intervals, estimating the conditional survival probability in each interval by the proportion of surviving individuals, and multiplying these estimates together to get an estimate for the unconditional survival probability. Assuming a sample (Y_i, δ_i), $i = 1, \ldots, n$, where $Y_i = \min(T_i, U_i)$ and δ_i the censoring indicator

$$\delta_i = \begin{cases} 0, & \text{if} \quad U_i \leq T_i, \\ 1, & \text{if} \quad U_i > T_i, \end{cases}$$

the Kaplan-Meier estimate, say $\hat{F}_T(y)$, of the unconditional c.d.f. $F_T(y)$, is defined by

$$\hat{F}_T(y) = \begin{cases} 0, & \text{if} \quad 0 \leq y \leq Z_1, \\ 1 - \prod_{i=1}^{k-1} \left(\frac{n-i}{n-i+1}\right)^{\Lambda_i}, & \text{if} \quad Z_{k-1} < y \leq Z_k, \quad k = 2, \ldots, n, \\ 1, & \text{if} \quad y > Z_n, \end{cases} \tag{1.10}$$

where (Z_i, Λ_i) are the ordered $Y_i's$, along with their censoring indicators δ_i, $i = 1, \ldots, n$.
For almost four decades the Kaplan–Meier estimator has been one of the key statistical methods for analyzing censored survival data, and it is discussed in most textbooks on survival analysis. The estimator defined by (1.10) is a step function with jumps at those observations Y_i for which $\delta_i = 1$. We will confirm that later in Example 1.4.1.. If $\delta_i = 1$ for all i (i.e., no censoring occurs), the Kaplan-Meier estimator reduces to a step function with jumps of height $1/n$ at each of the Y_i, which is the usual empirical distribution function.
The Kaplan-Meier estimate has the following properties

1) Since the underlying distribution functions F_T and G may have jumps, some of the Y_i may be identical. In this case, the ordering of the $Y_i's$ into $Z_i's$ is not unique. However, the Kaplan-Meier estimator is.

2) We adopt the convention of defining the Kaplan-Meier estimator $\hat{F}_T$ to be equal to one for $y > Z_n$ if $\delta_i = 0$. Such a convention has the

advantages of definiteness and the fact that it does not matter very much for large n (in the sense that it is asymptotically consistent).

3) When the last non-censored observation is not the biggest observation of the sample, then in the interval of the biggest complete observation and then, the Kaplan-Meier estimation is very unstable.

Furthermore, it is assumed that $y \in [0, T_Y]$, where

$$T_Y = \sup\{y : 1 - F_T(y) > \varepsilon\}, \quad \text{for a small} \quad \varepsilon > 0.$$

The consistency of the Kaplan-Meier was established in Theorem 2.1 in Chen and Lo (1997) according to which for $0 < \beta < 1/2$

$$\sup_{y \in [0, T_Y]} |\hat{F}_T(y) - F_T(y)| = o(n^{-\beta}) \quad \text{a.s..} \tag{1.11}$$

In particular, if $T_y = T_{(n)}$, where $T_{(n)}$ is the largest observation of the ordered sample, is uncensored, i.e. if $\delta_{(n)} = 1$, then $\beta = 1/2$, see Karunamuni and Yang (1991). Otherwise, the rate of convergence of $\hat{F}_T$ is $n^{-1/2}$ for $y \in [0, T_F]$ where T_F is the largest uncensored observation. The rate of convergence of $\hat{F}_T$ is $n^{-\beta}$, $0 < \beta < 1/2$, for $y \in [T_F, T_Y]$. In addition, to ensure the existence of supremum in (1.11) we require the absolute difference to be bounded, i.e. there is M such that $|\hat{F}_T - F_T| \leq M$.
When the biggest observation of the sample is uncensored then the rate of convergence of $\hat{F}_T$ to F_T is $1/2$. Furthermore, the smaller the β is, the fewer the uncensored observations near the endpoint. This, in turn, is reflected in the convergence rate of $\hat{F}_T$: the smaller the β, the slower the rate of convergence. The estimator $\hat{F}_T$ is rather unstable near the endpoint T_Y, thus may converge to F_T at an arbitrarily slow rate near T_Y when the censoring is arbitrarily heavy.
In order to investigate Kaplan-Meier's variance we have to introduce an alternative form of (1.10), which is the following

$$\hat{S}_T(t) = \prod_{i:t_q \leq t} \left(1 - \frac{d_q}{n_q}\right), \quad q = 1, 2, \ldots. \tag{1.12}$$

In (1.12) t_q is a time when at least one event happened, d_q is the number of events that happened at time t_q, and n_q is the number of individuals known to have survived, i.e. have not yet had an event or been censored, up to time t_q.

To establish how this equivalent form is derived, taking into account that T belongs to $[0, T_{(n)})$, where we recall that $T_{(n)}$ is the largest ordered observation, we divide the interval into m sub-intervals $[0, t_1) \dots [t_{m-1}, t_m)$, where $t_m = T_{(n)}$. Suppose that $S(v|u) = S(v)/S(u)$, $v > u$ the conditional probability that an event will occur later than the time v given that it has not occurred until time u.
The survival function at time $t \in [t_j, t_{j+1}]$ where $j \in (0, 1 \dots, m-1)$ is given by

$$\begin{aligned} P(T \geq t) &= \prod_{q=1}^{m} S_T(t_q|t_{q-1}) \\ &= P(T > t_q|T > t_{q-1})P(T > t_{q-1}) \\ &= \big\{1 - P(T \leq t_q|T > t_{q-1})\big\}P(T > t_{q-1}) \\ &= \big\{1 - P(T \leq t_q|T > t_{q-1})\big\}S_T(t_{q-1}) \\ &= \big\{1 - P(T \leq t_q|T > t_{q-1})\big\}\big\{1 - P(T \leq t_{q-1}|T > t_{q-2})\big\}S_T(t_{q-2}) \\ &= \big\{1 - P(T \leq t_q|T > t_{q-1})\big\}\big\{1 - P(T \leq t_{q-1}|T > t_{q-2})\big\}\dots \\ &\quad \big\{1 - P(T \leq t_0|T > t_{-1})\big\}, \end{aligned} \tag{1.13}$$

where in the last line we have that $1 - P(T \leq t_0|T > t_{-1}) = 1 - P(T = t_0)$. In addition, if there is no event in $(t_{q-1}, t_q]$ then $S(t_q|t_{q-1}) = 1$. In order to derive a practically useful version of (1.12) we have to approximate the probability $1 - P(T \leq t_q|T > t_{q-1})$.
The number of events (under study) up to time t_q is

$$d_q = \sum_{i=1}^{n} I_{(t_{q-1} < T_i \leq t_q, \delta_q = 1)}, \quad q = 1, 2, \dots, m.$$

The total number of individuals who survive after time t_q (i.e. have not yet experienced the event and are not censored) is

$$n_q = \sum_{i=1}^{n} I_{(T_i \geq t_{q-1})}, \quad q = 1, 2, \dots, m.$$

Then,

$$1 - P(T \leq t_q|T > t_{q-1}) \approx 1 - \frac{d_q}{n_q}. \tag{1.14}$$

So by repeatedly substituting (1.14) to (1.13) we obtain (1.12).

As is evident from (1.12), the size of the jumps of the Kaplan-Meier estimator depend not only on the number of events observed at each t_q, but also on the pattern of the censored observations prior to t_q. The Product-Limit estimator is based on the assumption of non-informative censoring which means that knowledge of a censoring time for an individual provides no further information about this person's likelihood of survival at a future time. In addition, the estimator is well defined for all time points less than the largest observed study time $t_{\max}$. Supposing that the largest study time corresponds to a death time, then, the estimated survival curve is zero beyond this point. If the largest time point is censored, the value of $S_T(t)$ beyond this point is undetermined as we are unable to predict when this last survivor would have died if the survivor had not been censored. However, several non-parametric suggestions have been made to account for this ambiguity.

Example 1.4.1. We consider the survival times (where + stands for censored observation): 0, 9, 13, 13+, 18, 23, 28+, 31, 34.
The data is already arranged so that the smallest survival is at the beginning and the largest is at the end. Then by using the equivalent form to (1.10) we get

$$\begin{aligned}
&\hat{S}_T(0) = 1,\\
&\hat{S}_T(9) = \hat{S}(0)\frac{8-1}{8} = 0.875,\\
&\hat{S}_T(13) = \hat{S}(9)\frac{7-1}{7} = 0.75,\\
&\hat{S}_T(13+) = \hat{S}(13)\frac{6-0}{6} = 0.75,\\
&\dots\\
&\hat{S}_T(34) = \hat{S}(31)\frac{1-1}{1} = 0.
\end{aligned}$$

The result is shown graphically in Fig. 1.4. There, it is seen that the survival duration of a subject is represented by the length of the horizontal lines along the X-axis of serial times. The plot in Fig. 1.4 also contains the confident intervals of the survival probabilities. The occurrence of the event determines the interval. The discontinuous points correspond to the change in the cumulative probability of surviving a given time as seen in the Y-axis and the dotted lines correspond to confident intervals. The closed form expression is provided in (1.18).

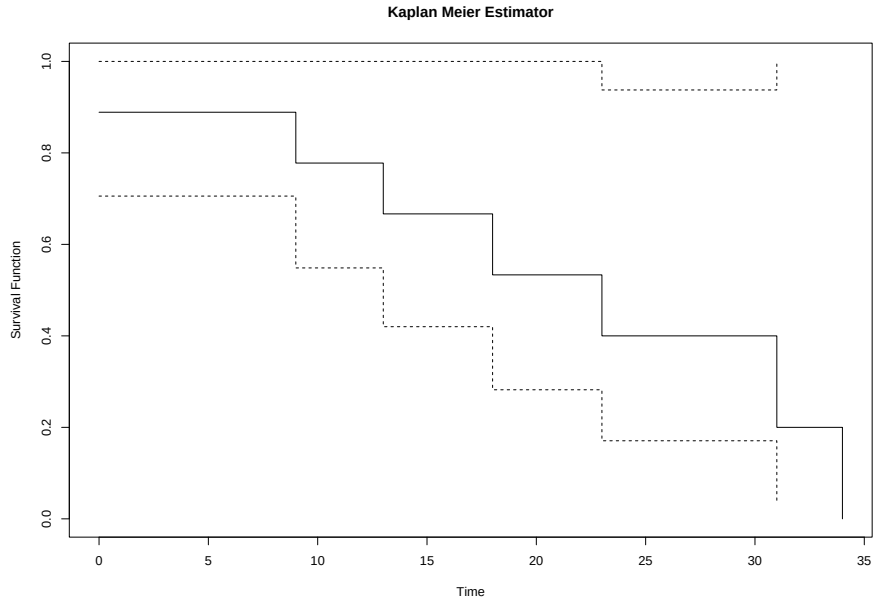

Figure 1.4: The Kaplan-Meier Estimator

As mentioned earlier, the Kaplan-Meier estimator is essentially a maximum likelihood estimator. To see this, we recall first that d_q and n_q is the number of events and the number of individuals at risk up to time t_q, respectively. Then the discrete hazard function is $h_q = d_q/n_q$ and the survival function is

$$S_T(t) = \prod_{q:t_q<t} (1 - h_q).$$

The likelihood for the hazard function is

$$L = \prod_{j=1}^{q} {h_j}^{d_j} (1 - h_j)^{n_j - d_j},$$

so applying logs in both sides we get

$$\log(L) = \sum_{j=1}^{q} \Big\{ d_j \log(h_j) + (n_j - d_j) \log(1 - h_j) \Big\}.$$

In order to find the maximum we set the derivative with respect to h_q equal to zero and we obtain

$$\frac{d_q}{h_q} - \frac{n_q - d_q}{1 - h_q} = 0.$$

Hence,

$$\hat{h}_q = \frac{d_q}{n_q}.$$

So, taking into consideration that the second derivative of $\hat{h}_q$ is negative which implies that it is maximum, the survival function will be estimated from

$$\hat{S}(t) = \prod_{q:t_q<t} (1 - \hat{h}_q) = \prod_{q:t_q<t} \left(1 - \frac{d_q}{n_q}\right) = \hat{S}_T(t).$$

Trying to show that the Kaplan-Meier estimator is a maximum likelihood estimator we referred to the hazard function. The hazard rate function reflect the "approximate" probability of an individual of age t experiencing the event in the next time instant. The hazard rate is defined by

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t | t > 0)}{\Delta t} = \frac{f_T(t)}{1 - F_T(t)}.$$

Now, let $\hat{p}_j = 1 - d_j/n_j$ so

$$\hat{S}_T(t) = \prod_{j:t_j \leq t} \hat{p}_j \Leftrightarrow \log\Big\{\hat{S}_T(t)\Big\} = \sum_{j=1}^{q} \log(\hat{p}_j).$$

Thus,

$$\mathrm{Var}\left[\log\Big\{\hat{S}_T(t)\Big\}\right] = \mathrm{Var}\Big\{\sum_{j=1}^{q} \log(\hat{p}_j)\Big\} = \sum_{j=1}^{q} \mathrm{Var}\Big\{\log(\hat{p}_j)\Big\}.$$

The number of those who survive, $n_j - d_j$, follows binomial distribution with parameters n_j and p_j, for $j = 1, \ldots, q$. Hence, $\mathrm{Var}(n_j - p_j) = n_j p_j (1 - p_j)$. Using this and $P(10)$ from the appendix

$$\begin{aligned}
\mathrm{Var}(\hat{p_j}) &= \mathrm{Var}\left(1 - \frac{d_j}{n_j}\right) = \mathrm{Var}\left(\frac{n_j - d_j}{n_j}\right) \\
&\overset{(P10)}{=} \frac{\mathrm{Var}(n_j - d_j)}{{n_j}^2} = \frac{n_j p_j (1 - p_j)}{{n_j}^2} \\
&= \frac{p_j(1 - p_j)}{n_j}.
\end{aligned}$$

Based on Taylor's Expansion of any real function $g(x)$ at x we have that

$$\mathrm{Var}\left\{g(X)\right\} = \left\{\frac{dg(X)}{dX}\right\}^2 \mathrm{Var}(X).$$

Therefore,

$$\begin{aligned}\mathrm{Var}\left\{\log(\hat{p}_j)\right\} &\simeq \frac{\mathrm{Var}(\hat{p}_j)}{\hat{p}_j^2} = \frac{\frac{\hat{p}_j(1-\hat{p}_j)}{n_j}}{\hat{p}_j^2} = \frac{1-\hat{p}_j}{n_j\hat{p}_j} \\ &= \frac{\left\{1-(1-\frac{d_j}{n_j})\right\}}{n_j(1-\frac{d_j}{n_j})} = \frac{\frac{d_j}{n_j}}{n_j(\frac{n_j-d_j}{n_j})} = \frac{\frac{d_j}{n_j}}{n_j-d_j} = \frac{d_j}{n_j(n_j-d_j)}.\end{aligned}$$

Thus,

$$\mathrm{Var}\Big[\log\big\{\hat{S}_T(t)\big\}\Big] = \sum_{j=1}^{q}\frac{d_j}{n_j(n_j-d_j)}. \tag{1.15}$$

Now, by using the Taylor's expansion one more time we have that

$$\begin{aligned}\mathrm{Var}\Big[\log\big\{\hat{S}_T(t)\big\}\Big] &= \frac{\mathrm{Var}\big\{\hat{S}_T(t)\big\}}{\hat{S}_T(t)^2} \\ &\Leftrightarrow \mathrm{Var}\big\{\hat{S}_T(t)\big\} = \hat{S}_T(t)^2\mathrm{Var}\Big[\log\big\{\hat{S}_T(t)\big\}\Big].\end{aligned} \tag{1.16}$$

By combining (1.15) and (1.16) we get

$$\begin{aligned}\mathrm{Var}\big\{\hat{S}_T(t)\big\} &= \hat{S}_T(t)^2\sum_{j=1}^{q}\frac{d_j}{n_j(n_j-d_j)} \\ &\Leftrightarrow \mathrm{Var}\big\{1-\hat{F}_T(t)\big\} = \big\{1-\hat{F}_T(t)\big\}^2\sum_{j=1}^{q}\frac{d_j}{n_j(n_j-d_j)} \\ &\Leftrightarrow \mathrm{Var}\big\{\hat{F}_T(t)\big\} = \big\{1-\hat{F}_T(t)\big\}^2\sum_{j=1}^{q}\frac{d_j}{n_j(n_j-d_j)}.\end{aligned} \tag{1.17}$$

This is the so-called Greenwood's estimate of the Kaplan-Meier's variance which tends to underestimate the true variance of the Kaplan–Meier estimator for small samples, see Greenwood (1926). However, on average, Greenwood's estimator is closer to the true variance compared to other estimators.
The standard error (SE) of the Product-Limit estimator is given by

$$\mathrm{SE}\big\{\hat{S}_T(t)\big\} = \left[\mathrm{Var}\Big\{\hat{S}_T(t)\Big\}\right]^{1/2} = \left\{\hat{S}_T(t)^2\sum_{j=1}^{q}\frac{d_j}{n_j(n_j-d_j)}\right\}^{1/2}.$$

In the case of no censoring, (1.15) reduces to $\hat{S}_T(t)\{1-S_T(t)\}/n$, the standard binomial variance estimator. In large samples the Kaplan–Meier estimator, evaluated at a given time t, is approximately normally distributed so that a standard $100(1-\alpha)\%$ confidence interval for $S_T(t)$ takes the form

$$\left(\hat{S}_T(t) - z_{1-a/2}\mathrm{Var}\{\hat{S}_T(t)\}, \hat{S}_T(t) + z_{1-a/2}\mathrm{Var}\{\hat{S}_T(t)\}\right), \tag{1.18}$$

with $z_{1-\alpha/2}$ the $1-\alpha/2$ fractile of the standard normal distribution.
In order to avoid the discontinuity of the Kaplan-Meier estimator we will use kernel smoothing. In the next section the main principles for the case of complete data will be introduced and in the end a continuous form of Kaplan-Meier estimator in the presence of censored data will be given.

1.5 Comparing Kaplan-Meier estimator with MLE

In this section we use n observations from the Exponential distribution, Exp(1) which are censored on the right by n observations from Uniform distribution, U$(0, 1.1)$. In total we have 60% censoring in our data. Taking into consideration that we know beforehand the distribution of the data we can use the MLE in order to estimate the parameter of the underlying distribution. Although, we can also use the Kaplan-Meier estimator. Our aim is to produce these two estimates and to compare them.

At first we generate 50 observations from Exp(1) and 50 censoring times from U(0, 1.1). Thus, this yields 50 right censored observations with 60% censoring. The results of the two estimations are presented in Fig. 1.5. The black line is the true distribution, the red is the MLE estimation of the survival function and the blue one is the Kaplan-Meier estimation. The MLE estimation is close enough since the true parameter is equal to 1 and the estimated one is equal to 1.028. Although, the Kaplan-Meier estimation is not that close.

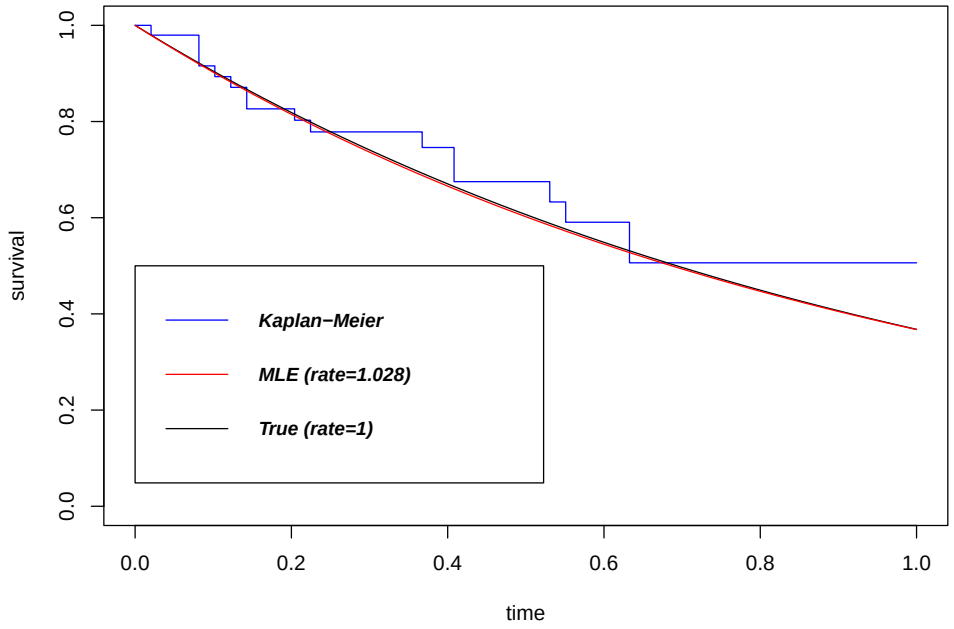

Figure 1.5: MLE and Kaplan-Meier estimations for Exponential distribution with 60% censoring and 50 observations

After that we generate 100 observations from Exp(1) and 100 censoring times from U(0, 1.1). Thus, this yields 100 right censored observations with 60% censoring. The results of the two estimations are presented in Fig. 1.6. We recall that the black line is the true distribution, the red is the MLE estimation of the survival function and the blue one is the Kaplan-Meier estimation. The MLE estimation is closer than previous since the estimated parameter is now equal to 1.011. Also, the Kaplan-Meier estimation is closer than before too.

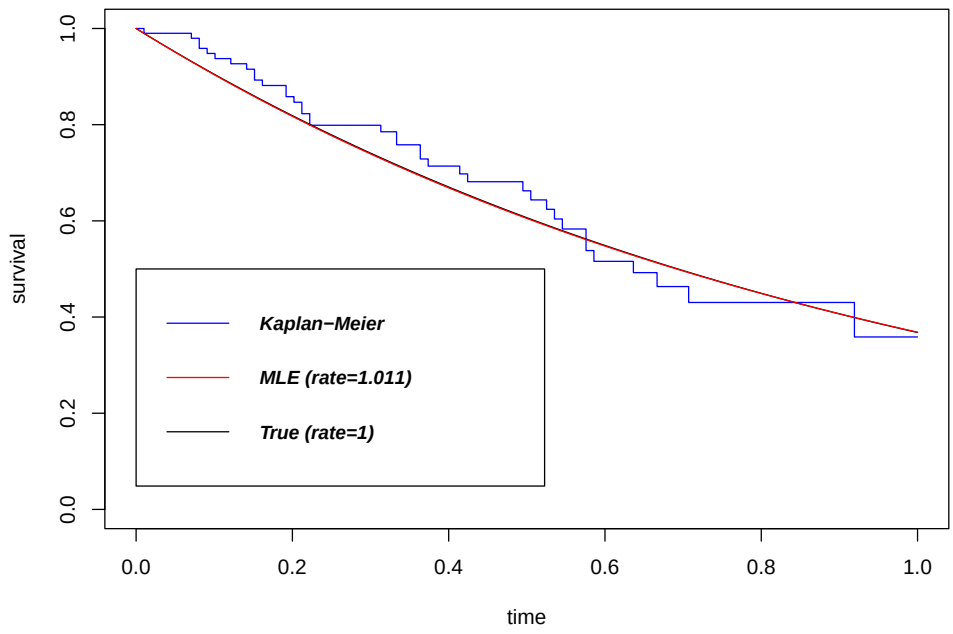

Figure 1.6: MLE and Kaplan-Meier estimations for Exponential distribution with 60% censoring and 100 observations

Then, we generate 200 observations from Exp(1) and 200 censoring times from U(0, 1.1). Thus, this yields 200 right censored observations with 60% censoring. The results of the two estimations are presented in Fig. 1.7. We recall that the black line is the true distribution, the red is the MLE estimation of the survival function and the blue one is the Kaplan-Meier estimation. The MLE estimation is even closer since the estimated parameter is now equal to 1.004. Also, the Kaplan-Meier estimation is even closer too.

So, we conclude that the larger the sample, the closer the two estima-

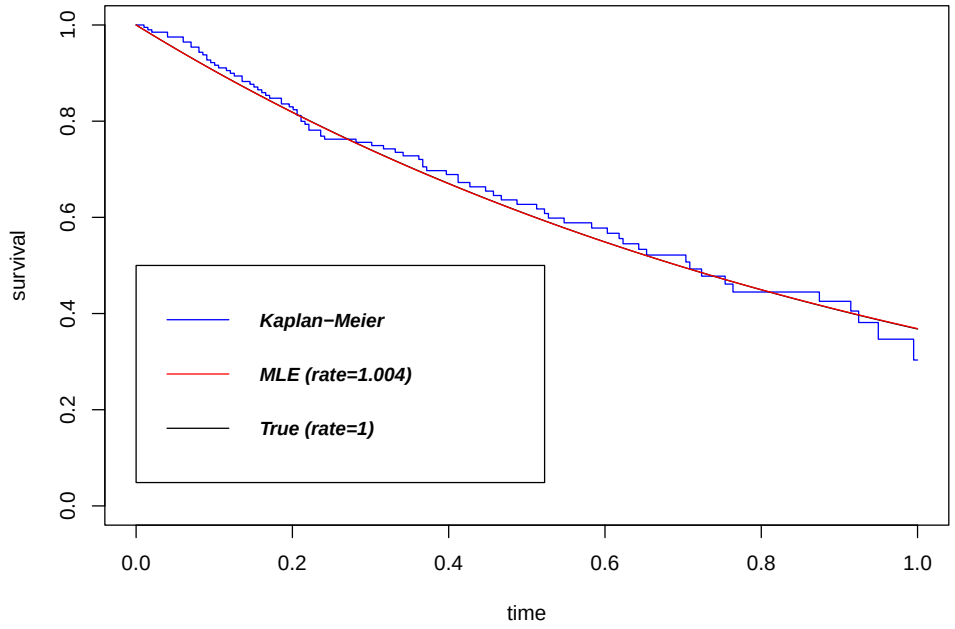

Figure 1.7: MLE and Kaplan-Meier estimations for Exponential distribution with 60% censoring and 200 observations

tions. This, conclusion is very rational taking into consideration that MLE and Kaplan-Meier estimators are asymptotically unbiased.

Now, in order to see how the two estimators are affected by the censorship rate we generate 50 observations from the Exponential distribution and 50 censoring times from the Uniform distribution in three different censorship

rates.

At first, we generate 50 observations from Exp(1) and 50 censoring times from U(1, 2.4). Thus, this yields 50 right censored observations with 20% censoring. The results of the two estimations are presented in Fig. 1.8. The MLE estimation is very close since the true parameter is equal to 1 and the estimated one is equal to 1.021. In addition, the Kaplan-Meier estimation is very close too.

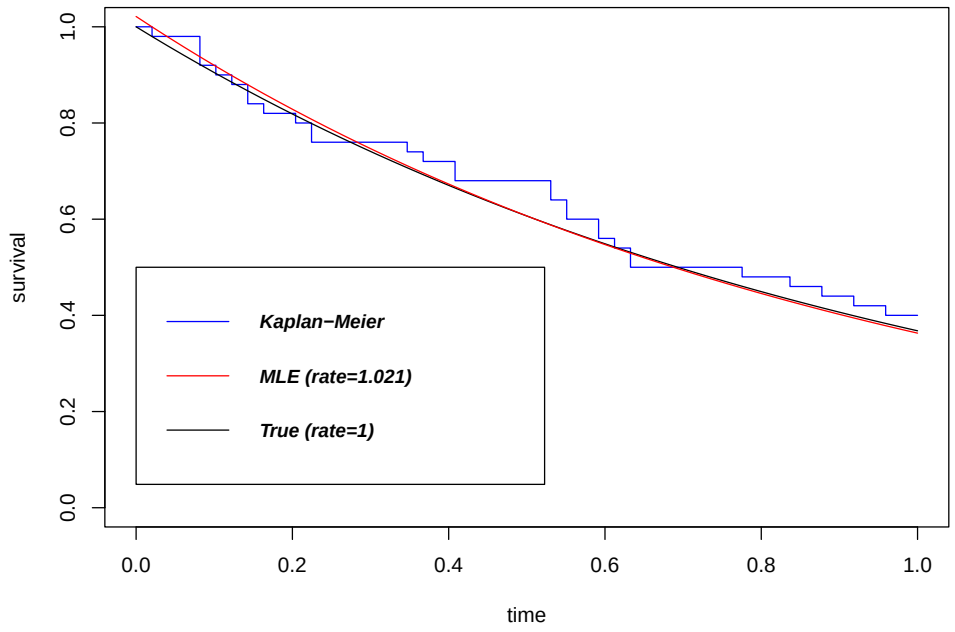

Figure 1.8: MLE and Kaplan-Meier estimations for Exponential distribution with 20% censoring

Then, we generate 50 observations from Exp(1) and 50 censoring times from U$(0, 2.2)$. Thus, this yields 50 right censored observations with 40% censoring. The results of the two estimations are presented in Fig. 1.9. The MLE estimation is close since the estimated parameter is now equal to 1.023. In addition, the Kaplan-Meier estimation is close too. Although, the two estimations are not so close as previously.

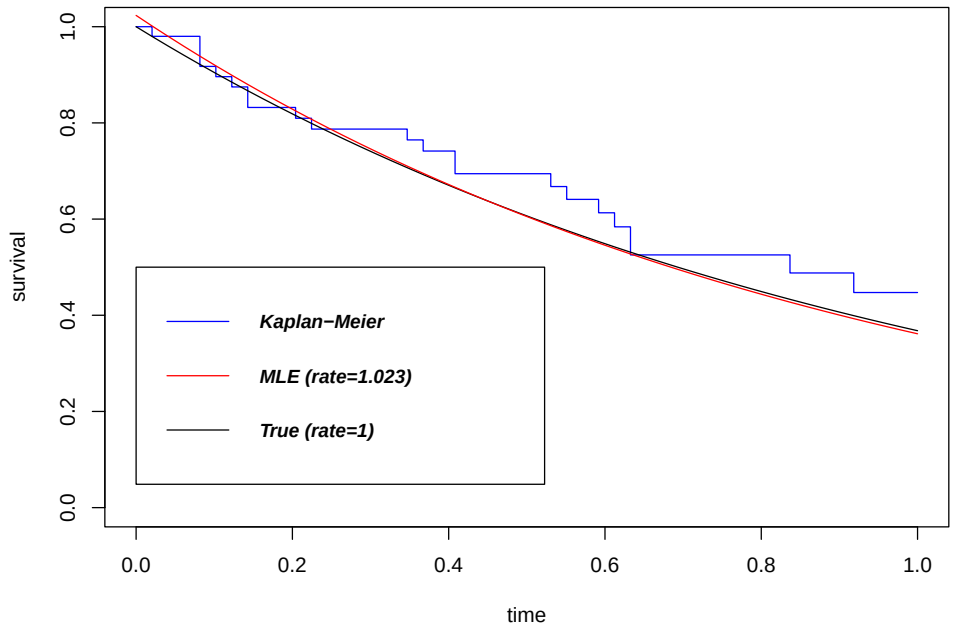

Figure 1.9: MLE and Kaplan-Meier estimations for Exponential distribution with 40% censoring

Then, we generate 50 observations from Exp(1) and 50 censoring times from U(0, 1.1). Thus, this yields 50 right censored observations with 60% censoring. The results of the two estimations are presented in Fig. 1.10. The MLE estimation is close but not so close as previously since the estimated parameter is now equal to 1.028. Also, the Kaplan-Meier estimation is not so close as previously too.

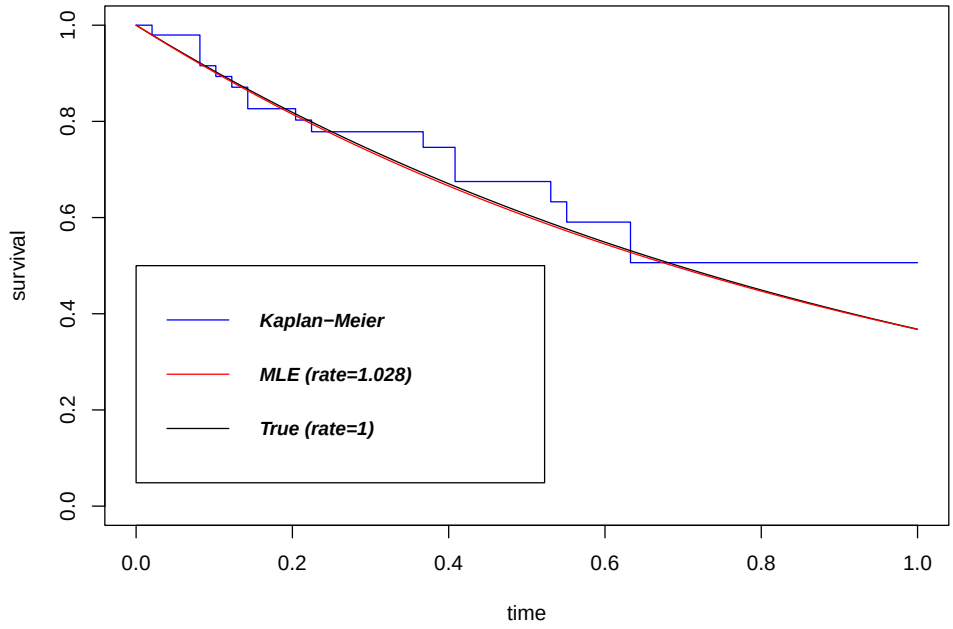

Figure 1.10: MLE and Kaplan-Meier estimations for Exponential distribution with 60% censoring

So, we conclude that the lower the censorship the better the estimations.

CHAPTER 2

Continues Estimation of Distribution Function

2.1 Univariate kernel density estimation

Even thought survival analysis is usually about censored data, when we have available the complete data i.e. data for which we know the exact failure time we can use a simpler estimator. This is the case when all items in the analysis have their exact failure times recorded. For example if we have 10 components under test, all of these components fail during the test, and the exact failure time is recorded then we have complete data.
In the case of complete data, the oldest and most widely used non-parametric way is to review the density of observations in the random sample with a simple histogram. The histogram was the first non-parametric density estimator, though its exact date of invention is not known for certain, but likely to have been in the 17^{th} century. Its chronological primacy is due to its simple computation. From the histogram, we might be able to identify a common and well-understood probability distribution that can be used, such as the normal distribution. If not, we may have to fit a model to estimate the distribution.
The classical frequency histogram is formed by constructing a complete set of non-overlapping intervals, called bins, and counting the number of points in each bin. In order for the bin counts to be comparable, the bins should all have the same width. The counts, or frequencies of observations, in each bin are then plotted as a bar graph with the bins on the X-axis and the frequencies on the Y-axis. If so, then the histogram is completely determined by two parameters, the bin width, h, and the bin origin, x_0, which is any conveniently chosen bin interval endpoint. Often the bin origin is chosen to be $x_0 = 0$. The choice of the number of bins is important as it controls the coarseness of the distribution (number of bars) and, in turn, how well the

density of the observations is plotted. It is a good idea to experiment with different bin sizes for a given data sample to get multiple perspectives or views on the same data. Reviewing a histogram of a data sample with a range of different numbers of bins will help to identify whether the density looks like a common probability distribution or not. A histogram density estimator is a step function with a constant value within each of the bins, where the constant is given by the proportion of data points X_i, $i = 1, \ldots, n$ which fall in the bin divided by the bin volume.
In case the shape of a histogram matches a well-known probability distribution, we can attempt to estimate the density in a parametric way. Although, in some cases, a data sample may not resemble a common probability distribution or cannot be easily made to fit the distribution. This is often the case when the underlying density is multimodal. In this case, the parametric density estimation is inefficient.
The definition of the histogram, with bin width h, is given by

$$\hat{f}(x) = \frac{1}{2hn} \sum_{i=1}^{n} \mathbf{1}_{\{|x-X_i|<h\}}.$$

Equivalently, the non-parametric estimator of $F(x)$ in the case of complete data is the empirical estimation of $F(x) = P(X \leq x)$ given by

$$\hat{F}(x) = n^{-1} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}}. \tag{2.1}$$

The empirical distribution, or empirical distribution function, can be used to describe a sample of observations of a given variable. Its value at a given point is equal to the proportion of observations from the sample that are less than or equal to that point. In other words, the value of the empirical distribution function at a given point is obtained by

1) Counting the number of observations that are less than or equal to x.

2) Dividing the number that obtained by the total number of observations, so as to obtain the proportion of observations that is less than or equal to x.

From (2.1) also follows the very useful conclusion that $\hat{F}(x)$ can be viewed as c.d.f. of a discrete random variable, taking values $X_1, \ldots, X_n$, with probability $1/n$ for each. Moreover, the empirical distribution is the non-parametric MLE

of c.d.f.. The strong convergence of $\hat{F}(x)$ to the real $F(x)$ is provided by the Glivenko–Cantelli theorem which asserts the following, see Glivenko (1933).

Theorem 2.1.1. *Let X_i, $i = 1, \ldots, n$ be an i.i.d. sequence of random variables with distribution function F on $\mathbb{R}$. Then,*

$$\sup_{x \in \mathbb{R}} |\hat{F}(x) - F(x)| \to 0 \quad a.s..$$

This result is perhaps the oldest and most well known result in the very large field of empirical process theory. It was proved by Birkhoff (1931) in the case of continuous $F(x)$ and Borovkov (1984) in the general case.

Proof. Let $\epsilon > 0$. Then fix $k > 1/\epsilon$, and then consider the points $k_0, \ldots, k_k$ such that

$$-\infty = k_0 < k_1 \leq k_2 \leq \cdots \leq k_{k-1} < k_k = \infty.$$

that define a partition of $\mathbb{R}$ into k disjoint intervals such that

$$F(k_j{}^-) \leq j/k \leq F(k_j), \quad j = 1, \ldots, k-1,$$

where, for each j,

$$F(k_j{}^-) = P(T_j < k_j) = F(k_j) - P(X = k_j).$$

Then, by construction, if $k_{j-1} < k_j$,

$$F(k_j{}^-) - F(k_{j-1}) \leq \frac{j}{k} - \frac{(j-1)}{k} = \frac{1}{k} < \epsilon.$$

In the following recall that $\hat{F}(x)$ is a random variable. Now, by the Strong Law of Large Numbers, the estimator converges pointwise to the true c.d.f.. That is, as $n \to \infty$, for $j = 1, \ldots, k-1$,

$$\hat{F}(k_j) \overset{a.s.}{\to} F(k_j) \quad and \quad \hat{F}(k_j{}^-) \overset{a.s.}{\to} F(k_j{}^-).$$

It immediately follows that, for each j, $j = 1, \ldots, k-1$,

$$|\hat{F}(k_j) - F(k_j)| \overset{a.s.}{\to} 0 \quad and \quad |\hat{F}(k_j{}^-) - F(k_j{}^-)| \overset{a.s.}{\to} 0,$$

as $n \to \infty$. Taking the maximum over all j,

$$\Delta_n = \max_j \{|\hat{F}(k_j) - F(k_j)|, |\hat{F}(k_j{}^-) - F(k_j{}^-)|\} \overset{a.s.}{\to} 0 \quad as \quad n \to \infty.$$

For any x identify j such that

$$k_{j-1} \leq t < k_j.$$

Then, for a small $\epsilon > 0$ we have

$$\hat{F}(x) - F(x) \leq \hat{F}(k_j^-) - F(k_{j-1}) \leq \hat{F}(k_j^-) - F(k_j^-) + \epsilon,$$
$$\hat{F}(x) - F(x) \geq \hat{F}(k_{j-1}) - F(k_j^-) \geq \hat{F}(k_{j-1}) - F(k_{j-1}) - \epsilon,$$

and thus for any x,

$$\hat{F}(k_{j-1}) - F(k_{j-1}) - \epsilon \leq \hat{F}(x) - F(x) \leq \hat{F}(k_j^-) - F(k_j^-) + \epsilon,$$

and thus,

$$|\hat{F}(x) - F(x)| \leq \Delta_n + \epsilon \overset{a.s.}{\to} \epsilon \quad as \quad n \to \infty.$$

Hence, as this holds for arbitrary x, it follows that

$$\sup_{x \in \mathbb{R}} |\hat{F}(x) - F(x)| \overset{a.s.}{\to} \epsilon \quad as \quad n \to \infty.$$

This holds for every $\epsilon > 0$; that is, if A_ϵ denotes the set of ω on which this convergence is observed, then $P(A_\epsilon) = 1$, and then by definition

$$A \equiv \cup_{\epsilon > 0} A_\epsilon \equiv \lim_{\epsilon \to 0} A_\epsilon \Rightarrow P(A) = P(\lim_{\epsilon \to 0} A_\epsilon) = \lim_{\epsilon \to 0} P(A_\epsilon) = 1,$$

and it follows that

$$P[\lim_{n \to \infty} \sup_{x \in \mathbb{R}} |\hat{F}(x) - F(x)| = 0] = 1.$$

□

Nevertheless, the problem that arises is that empirical distribution is not a continuous function and so it is not further used. Therefore, a non-parametric alternative to the well known MLE estimate of the p.d.f, which produces continues estimate of the density is the kernel approach

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

where K is a kernel function and h its bandwidth. Equivalently, the distribution function estimator is given by

$$\hat{F}(x) = \frac{1}{nh} \sum_{i=1}^{n} \int_{-\infty}^{x} K\left(\frac{w - X_i}{h}\right) dw.$$

This process is called kernel smoothing, or Kernel Density Estimation (KDE) and perhaps it is the most common non-parametric approach for estimating the probability density function and the distribution function of a continuous random variable, see Wand and Jones (1994), Silverman (1986) and Scott (2015). The first appearance of kernel estimators is likely to be in Fix Hodges in 1951: as the original technical report is difficult to find, it has been re-published in 1989, see Fix and Hodges (1989). KDE is a non-parametric method for using a data-set in order to estimate probabilities for new points. In this case, a kernel K is a mathematical function that returns a probability for a given value of a random variable. Usually K is chosen to be a unimodal probability density function that is also symmetric about zero and the most common types and shapes are given in table 2.1 and Fig. 2.1 respectively.

Kernels	Kernel Function $K(u)$
Gaussian	$K(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}u^2)$
Epanechnikov	$K(u) = \frac{3}{4}(1-u^2)\mathbf{1}_{\{\lvert u\rvert \leq 1\}}$
Uniform	$K(u) = \frac{1}{2}\mathbf{1}_{\{\lvert u\rvert \leq 1\}}$
Triangular	$K(u) = (1-\lvert u\rvert)\mathbf{1}_{\{\lvert u\rvert \leq 1\}}$
Triweight	$K(u) = \frac{35}{32}(1-u^2)^3\mathbf{1}_{\{\lvert u\rvert \leq 1\}}$
Tricube	$K(u) = \frac{70}{81}(1-\lvert u\rvert^3)^3\mathbf{1}_{\{\lvert u\rvert \leq 1\}}$
Biweight	$K(u) = \frac{15}{16}(1-u^2)^2\mathbf{1}_{\{\lvert u\rvert \leq 1\}}$
Cosine	$K(u) = \frac{\pi}{4}\cos(\frac{\pi}{2}u)\mathbf{1}_{\{\lvert u\rvert \leq 1\}}$
Silverman	$K(u) = \frac{1}{2}\exp(-\frac{\lvert u\rvert}{\sqrt{2}}\sin\left(\frac{\lvert u\rvert}{\sqrt{2}} + \frac{\pi}{4}\right)$

Table 2.1: Types of common used kernels

This ensures that $\hat{f}(x)$ is itself also a density. Combining the properties of the kernels and the definition of $\hat{f}(x)$, we can conclude to this assumption. In order to see this, at first we observe that $K > 0$ which implies that $\hat{f}(x) > 0$. Also,

$$\int_{-\infty}^{+\infty} \hat{f}(x)dx = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{+\infty} K\left(\frac{x-X_i}{h}\right)dx = \frac{1}{n}n\int_{-\infty}^{+\infty} K(u)du = 1.$$

Furthermore, $\hat{f}$ will inherit all the continuity and differentiability properties of the kernel K, so that if, for example, K is the normal density function, then $\hat{f}$ will be a smooth curve having derivatives of all orders. However, kernels

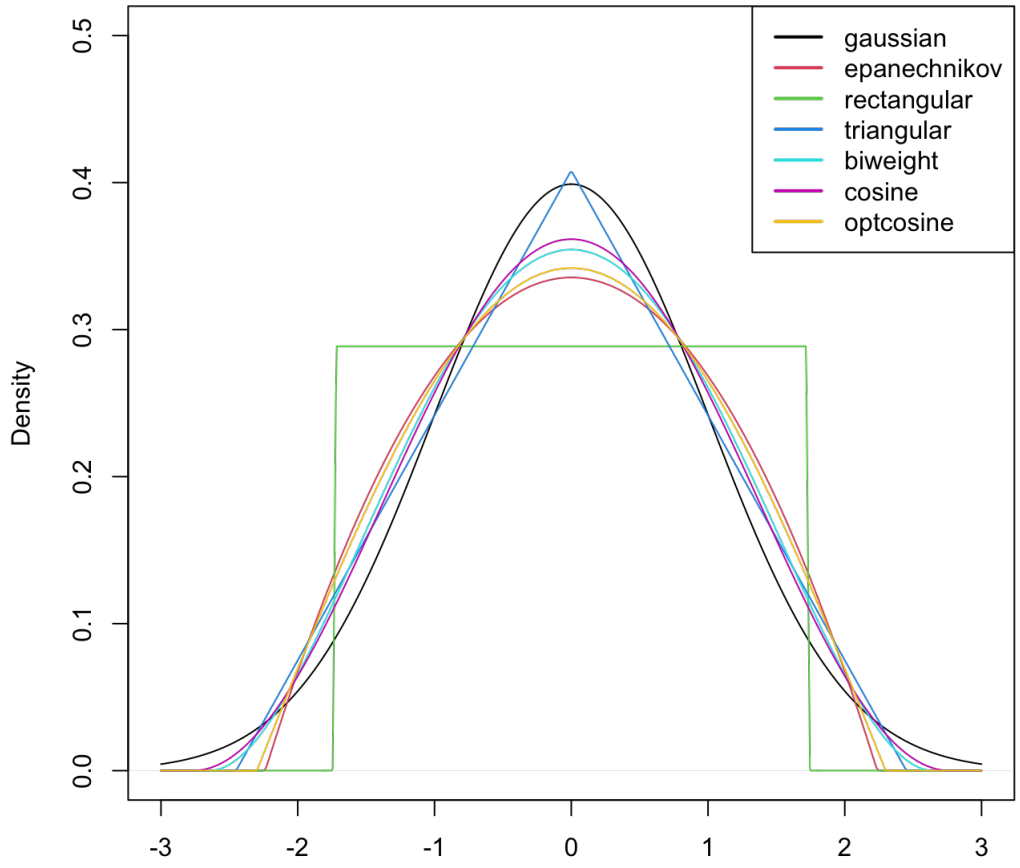

Figure 2.1: Shapes of commonly used kernels

that are not densities are also sometimes used. The kernel effectively smooths or interpolates the probabilities across the range of outcomes for a random variable such that the sum of probabilities equals one, a requirement of well-behaved probabilities which is ensured by the division with n. The kernel function weights the contribution of observations from a data sample based on their relationship or distance to a given sample for which the probability is requested. Specifically, this estimation is constructed by centring a scaled kernel at each observation. The value of the kernel estimate at the point x is simply the average of the n kernel ordinates at that point. One can think of the kernel as spreading a "probability mass" of size $1/n$ associated with each data point about its neighbourhood. Combining contributions from each data point means that in regions where there are many observations, and it is expected that the true density has a relatively large value, the kernel estimate should also assume a relatively large value. The opposite should occur in regions where there are relatively few observations.

As the kernels are placed on each data point, the anchor point placement problem that the histogram suffers from is thus eliminated. The increased smoothness of kernel estimators in comparison to histograms is not solely an aesthetic improvement, as it also leads to improved statistical properties. The choice of the shape of the kernel function is not a particularly important one. The most unimodal densities perform about the same as each other when used as a kernel. Thus, it follows that the choice between kernels can be made on other grounds such as computational efficiency. In addition, we should mention that it is possible to obtain better rates of convergence by relaxing the restriction that the kernel is a density function. When K is constrained to be a probability density function then it is necessarily true that $\mu_2(K) > 0$, where $\mu_2(K) = \int x^2 K(x)dx$. However, without this restriction, it is possible to construct K so that $\mu_2(K) = 0$ which will have the effect of reducing the bias. However, the choice of h is very important as it controls the range of the data that continues to estimate the underlying p.d.f. at a given point. To sum up, the kernel estimator is a sum of "bumps" placed at the observations where the kernel function determines the shape of the bumps while the bandwidth parameter determines their width.

We take the precip data which are publicly available in R and give the average amount of precipitation (rainfall) in inches for each of 70 United States (and Puerto Rico) cities in order to understand how the estimation is affected from the bandwidth. With the black line we have the optimal bandwidth which is equal to 3.348. Then, we take the kernel estimations for three higher values of bandwidth. Specifically, with the red line is depicted the estimation with bandwidth equal to 4, with the green line the one equal to 6 and with the blue line the one equal to 8. As presented in Fig. 2.2 the higher the h, the smoother the estimator. This is a problem because it smooths features of the curve that we are interested in.

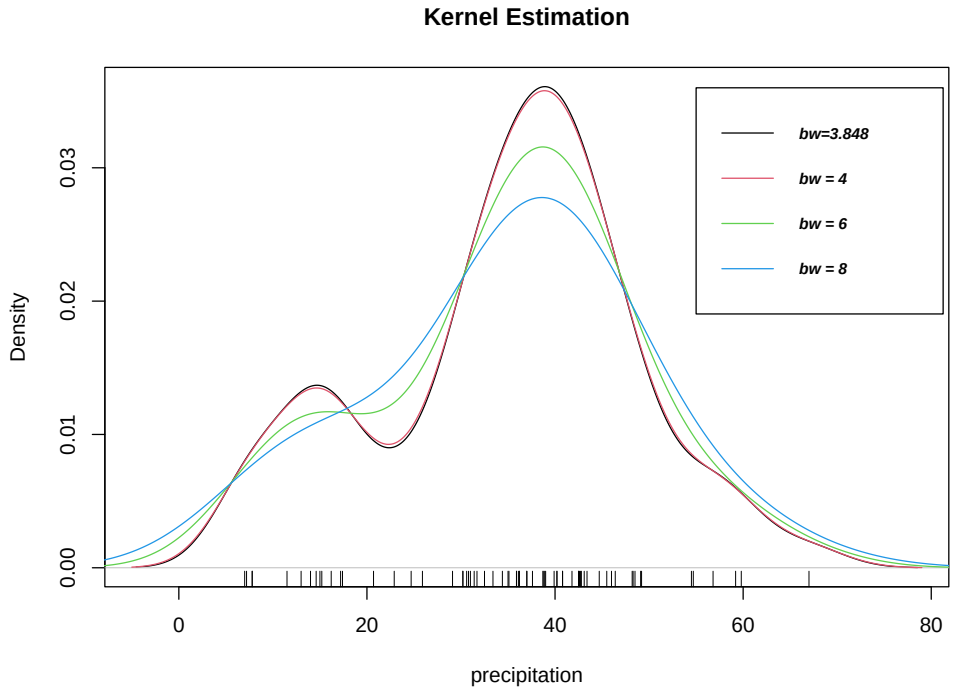

Figure 2.2: Kernel estimation with different bandwidths

Then, working with the same data set we recall that with the black line we have the optimal bandwidth which is equal to 3.348. Then, we take the kernel estimations for three lower values of bandwidth. Specifically, with the red line is depicted the estimation with bandwidth equal to 3, with the green line the one equal to 2 and with the blue line the one equal to 1. As presented in Fig. 2.3 the lower the h, the more the estimator tends to take the form of the data. This is a problem because the variation is greatly increased.

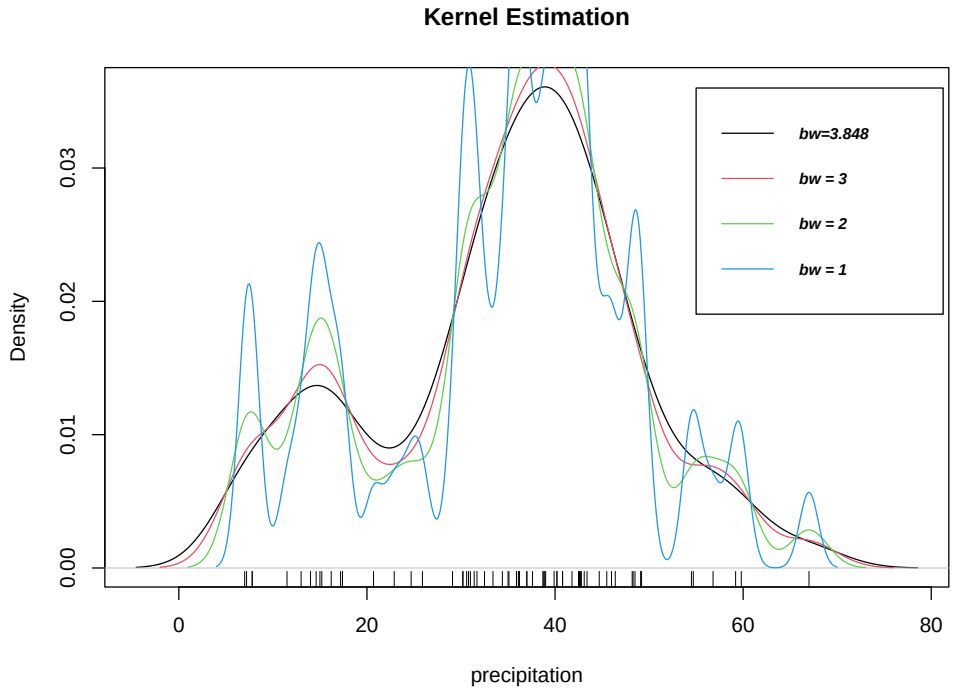

Figure 2.3: Kernel estimation with different bandwidths

Recalling the same data and optimal bandwidth we take again three different estimations. The first one which is depicted with the red line has a small enough bandwidth equal to 0.8, the green line has bandwidth very close to the optimal and equal to 3 and the blue line has a quite large bandwidth equal to 20. As presented in Fig. 2.4 depending on the size of the bandwidth our estimation can vary a lot. Therefore, there has to be a balance in this choice because we want neither small nor large values for the bandwidth. In Chapter 5 we will investigate in more detail how this can be achieved.

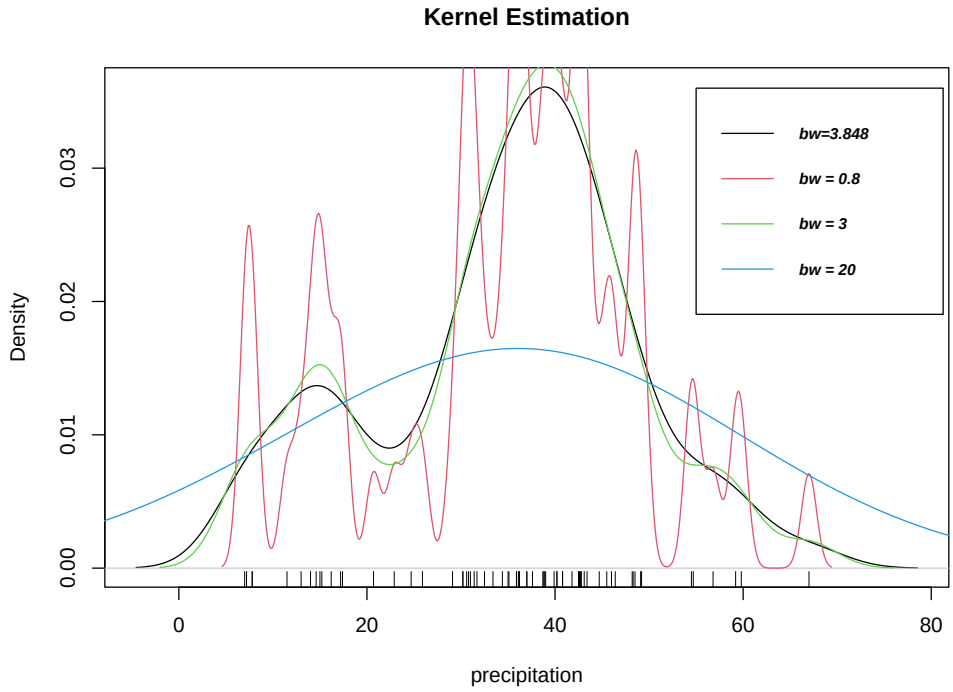

Figure 2.4: Kernel estimation with different bandwidths

Looking at the definition of the histogram and that of the kernel estimator we observe that, by taking K to be the uniform kernel, the two estimators are identical. In addition, some important properties of kernels are given in A1-A5 of the next Chapter.
For the purpose of the next theorem we provide the following definition. Let a_n and b_n each be sequences of real numbers. We say that a_n is of small order b_n (or a_n is "small oh" b_n), and we write

$$a_n = \mathrm{o}(b_n) \quad as \quad n \to \infty \quad \Leftrightarrow \lim_{n\to\infty} |a_n/b_n| = 0.$$

Theorem 2.1.2. *The asymptotic properties of the kernel estimator $\hat{f}(x)$, see Silverman (1986), are given by*

$$E\{\hat{f}(x)\} = f(x) + \frac{h^2}{2} f''(x) \int u^2 K(u)du + o(h^2),$$
$$Var\{\hat{f}(x)\} = \frac{1}{nh} f(x) \int K^2(u)du + \mathrm{o}\{(nh)^{-1}\}.$$

As we can see, $\hat{f}$ is a biased estimator, however asymptotically $h \to 0$ as $n \to \infty$ and thus the bias shrinks to zero.

Proof. For the purpose of the proof we use the $P4$, $P9$, $P11$, $P12$ and $P13$ properties from the appendix. Fixing i in second step below

$$\begin{aligned}
\mathrm{E}\{\hat{f}(x)\} &= \mathrm{E}\left\{ \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h} \right) \right\} \\
&\overset{(P9)}{=} \mathrm{E}\left\{ \frac{1}{h} K\left(\frac{x - X_i}{h} \right) \right\} \\
&\overset{(P4)}{=} \int \frac{1}{h} K\left(\frac{x - z}{h} \right) f(z)dz.
\end{aligned}$$

Using the change of variable $x - z = hu$ so that $dz = hdu$ gives

$$\begin{aligned}
\mathrm{E}\{\hat{f}(x)\} &= \int \frac{1}{h} K(u) f(x - hu) h du \\
&= \int K(u) f(x - hu) du. \qquad (2.2)
\end{aligned}$$

By a Taylor expansion of $f(x - hu)$ around x we get

$$f(x - hu) = f(x) - huf'(x) + \frac{1}{2} h^2 u^2 f''(x) + \mathrm{o}(h^2). \qquad (2.3)$$

By substituting (2.3) in (2.2)

$$\begin{aligned}
\mathrm{E}\{\hat{f}(x)\} &= \int K(u)\{f(x) - huf'(x) + \frac{1}{2}h^2u^2f''(x) + \mathrm{o}(h^2)\}du \\
&= f(x)\int K(u)du - hf'(x)\int uK(u)du \\
&\quad + \frac{1}{2}h^2f''(x)\int u^2K(u)du + \mathrm{o}(h^2) \\
&= f(x) + \frac{h^2}{2}f''(x)\int u^2K(u)du + \mathrm{o}(h^2).
\end{aligned}$$

Regarding the variance expression of the theorem, fixing i in third step below since $X_1, \ldots, X_n$ are i.i.d.

$$\begin{aligned}
\mathrm{Var}\{\hat{f}(x)\} &= \mathrm{Var}\left\{\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)\right\} \\
&\overset{(P12)}{=} \frac{1}{(nh)^2}\mathrm{Var}\left\{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)\right\} \\
&\overset{(P13)}{=} \frac{1}{nh^2}\mathrm{Var}\left\{K\left(\frac{x - X_i}{h}\right)\right\} \\
&\overset{(P11)}{=} \frac{1}{nh^2}\left(\mathrm{E}\left\{K\left(\frac{x - X_i}{h}\right)\right\}^2 - \left[\mathrm{E}\left\{K\left(\frac{x - X_i}{h}\right)\right\}\right]^2\right) \\
&\overset{(P4)}{=} \frac{1}{nh^2}\left[\int K\left(\frac{x - z}{h}\right)^2 f(z)dz - \left\{\int K\left(\frac{x - z}{h}\right)f(z)dz\right\}^2\right].
\end{aligned}$$

Using the change of variable $x - z = hu$ so that $dz = hdu$ gives

$$\begin{aligned}
\mathrm{Var}\{\hat{f}(x)\} &= \frac{1}{nh^2}\left[\int K(u)^2 f(x - hu)hdu - \left\{\int K(u)f(x - hu)hdu\right\}^2\right] \\
&= \frac{1}{nh}\int K(u)^2 f(x - hu)du - \frac{1}{n}\left\{\int K(u)f(x - hu)du\right\}^2. \quad (2.4)
\end{aligned}$$

By substituting (2.3) in (2.4)

$$\begin{aligned}
\mathrm{Var}\{\hat{f}(x)\} &= \frac{1}{nh}\int K(u)^2\Big\{f(x) + \mathrm{o}(1)\Big\}du - \frac{1}{n}\Big\{f(x) + \mathrm{o}(1)\Big\}^2 \\
&= \frac{1}{nh}f(x)\int K(u)^2du + \mathrm{o}\{(nh)^{-1}\},
\end{aligned}$$

thus finishes the proof of the theorem. □

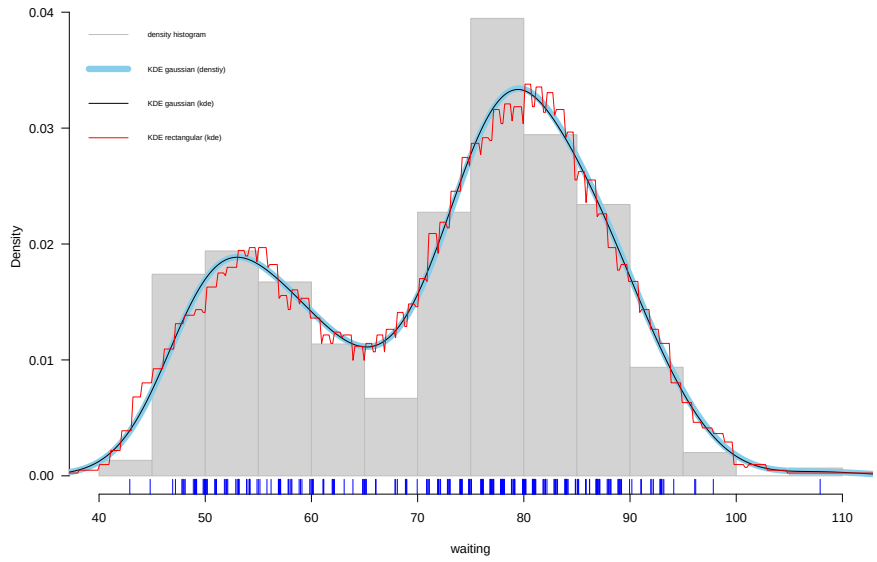

Figure 2.5: Kernel density estimation

Example 2.1.1. In this example we use the geyser data which are publicly available in R. It is about a data frame with 299 observations on 2 variables. The first one, the "duration", is a numeric variable and contains the eruption time for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA, in minutes. The second variable which is the one that we use in this example, the "waiting", is also a numeric variable and contains the waiting time for this eruption. As presented in Fig. 2.5 with grey color is depicted the histogram of the "waiting" data and with the blue line its density. Then, the black line is the kernel density estimation with the Gaussian kernel and the red line is the estimation with the rectangular kernel. In addition, the bandwidth of the kernel estimator is selected automatically and we will discuss in more detail how it is calculated in Chapter 5.

2.2 Multivariate kernel density estimation

Up to now, we have concentrated on the estimation of density underling a set of univariate observations. But, for the purpose of this thesis we need to establish the analysis of multivariate data. In the multivariate case, the dis-

tinction between different possible applications of density estimation becomes more important than for the univariate case. It is easy to comprehend a plot of a two-dimensional density function. However, presentational difficulties make it unlikely that density estimates will be useful, directly, for exploratory purpose for more than two dimensions. An experienced user with access to sophisticated graphics facilities might be able to inspect a three-dimensional density function. On the other hand, if the intention is not to look at the density function but instead to use it as an ingredient in some other statistical techniques, it may be necessary to estimate densities in higher-dimensional space.

Thus, we will now investigate the extension of the kernel density estimator to the multivariate setting. The need for non-parametric density estimates for recovering structure in multivariate data is, perhaps, greater since parametric modelling is more difficult than in the univariate case. The most general smoothing parametrization of the kernel estimator in higher dimensions requires the specification of many more bandwidth parameters than in the univariate setting. This leads us to consider simpler smoothing parametrizations as well. Also, the sparseness of data in higher-dimensional space makes kernel smoothing difficult unless the sample size is very large. This phenomenon, usually called the curse of dimensionality, means that, with practical sample sizes, reasonable non-parametric density estimation is very difficult in more than about five dimensions.

Nevertheless, there have been several studies where the kernel density estimator has been an effective tool for displaying structure in bivariate samples. The multivariate kernel density estimator that we will introduce is a direct extension of the univariate estimator. We should note that there are also approaches to multivariate smoothing which attempt to alleviate the curse of dimensionality by assuming that the multivariate function has some simplifying structure. In its most general form, the d-dimensional kernel density estimator is

$$\hat{f}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{K}_{\boldsymbol{H}}(\boldsymbol{X}_i - \boldsymbol{x}),$$

where $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ denote a d-variate random sample with $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{id})^\top$ and $\boldsymbol{x} \in \mathbb{R}^d$ a generic vector with representation $\boldsymbol{x} = (x_1, \ldots, x_d)^\top$. In addition, the crucial tuning parameter $\boldsymbol{H}$ is a symmetric positive definite $d \times d$ matrix called the *bandwidth matrix* and $\boldsymbol{K}$ is a d-variate probability density function, i.e.

$$\boldsymbol{K}_{\boldsymbol{H}}(\boldsymbol{x}) = |\boldsymbol{H}|^{-1/2}\boldsymbol{K}(\boldsymbol{H}^{-1/2}\boldsymbol{x}),$$

where

$$\boldsymbol{K}(\boldsymbol{x}) = \prod_{j=1}^{d} K(x_j).$$

The bandwidth controls the orientation and the extent of the smoothing applied via the kernel. Multivariate kernel functions can be constructed from univariate ones in various ways. There are two popular ways of doing so when we wish the kernel to be a multivariate density itself. In the previous expression we use the product kernel $\boldsymbol{K}$ which is a common technique for generating multivariate kernels from a symmetric univariate kernel K.
A popular choice for $\boldsymbol{K}$ is the standard d-variate normal density

$$\boldsymbol{K}(\boldsymbol{x}) = (2\pi)^{-d/2} \exp\left(-\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{x}\right),$$

in which case $\boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\}$ is the $N(\boldsymbol{X}_i, \boldsymbol{H})$ (normal) density in the vector $\boldsymbol{x}$. The normal kernel can be constructed from the univariate standard normal density using either the product or spherically symmetric extensions. The scaled, translated normal kernel is

$$\boldsymbol{K}_{\boldsymbol{H}}(\boldsymbol{x} - \boldsymbol{X}_i) = (2\pi)^{-d/2} |\boldsymbol{H}|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{X}_i)^\top \boldsymbol{H}^{-1}(\boldsymbol{x} - \boldsymbol{X}_i)\right\},$$

which is a normal density centred at $\boldsymbol{X}_i$ with covariance matrix $\boldsymbol{H}$. This is one of the main reasons that we parametrize $\boldsymbol{H}$ as a variance matrix. For this, and other reasons, the normal kernel is almost universally preferred for multivariate data, in contrast to the univariate case where other kernels can be preferred.
For the purpose of the next theorem we provide the following definition. Let a_n and b_n each be sequences of real numbers. We say that a_n is of order b_n (or a_n is "big oh" b_n), and write

$$a_n = \mathrm{O}(b_n) \quad as \quad n \to \infty \quad \Leftrightarrow \lim_{n\to\infty} \sup |a_n/b_n| < \infty.$$

Also, according to the multivariate Taylor's expansion for any d-variate real function f and a sequence of d-dimensional vectors a_n with all components tending to zero, $\mathcal{D}_f(\boldsymbol{x})$ denotes a vector of first order partial derivatives of f and $\mathcal{H}_f(\boldsymbol{x})$ the Hessian matrix of f, i.e. the $d \times d$ matrix with the (i, j) entry equal to

$$\frac{\partial^2}{\partial x_i \partial x_j} f(\boldsymbol{x}).$$

Then, assuming that all entries of $\mathcal{H}_f(\boldsymbol{x})$ are continuous in a neighborhood of $\boldsymbol{x}$,

$$f(\boldsymbol{x}+\alpha_n) = f(\boldsymbol{x}) + {\alpha_n}^\top \mathcal{D}_f(\boldsymbol{x}) + \frac{1}{2}{\alpha_n}^\top \mathcal{H}_f(\boldsymbol{x})\alpha_n + \mathrm{o}({\alpha_n}^\top\alpha_n). \tag{2.5}$$

Theorem 2.2.1. *In order to derive the asymptotic properties except for A3 and A5 assumptions of the next Chapter we will also assume that each entry of $\mathcal{H}_{f_x}$ is piecewise continuous and square integrable (see Silverman (1986)).*

$$\mathrm{E}\big\{\hat{f}(\boldsymbol{x})\big\} = f(\boldsymbol{x}) + \frac{1}{2}\mathrm{tr}\big\{\boldsymbol{H}^{1/2}\mathcal{H}_{f_x}\boldsymbol{H}^{1/2}\int \boldsymbol{u}^2\boldsymbol{K}(\boldsymbol{u})d\boldsymbol{u}\big\} + \mathrm{o}\big\{\mathrm{tr}(\boldsymbol{H})\big\},$$
$$\mathrm{Var}\big\{\hat{f}(\boldsymbol{x})\big\} = n^{-1}|\boldsymbol{H}|^{-1/2}f(\boldsymbol{x})\int \boldsymbol{K}(\boldsymbol{u})^2 d\boldsymbol{u} + \mathrm{o}(n^{-1}|\boldsymbol{H}|^{-1/2}).$$

Proof. For the purpose of the proof we use the $P1$, $P4$, $P9$, $P11$, $P12$ and $P13$ properties from the appendix. Fixing i in the second step below

$$\begin{aligned}
\mathrm{E}\big\{\hat{f}(\boldsymbol{x})\big\} &= \mathrm{E}\Big\{n^{-1}\sum_{i=1}^{n}\boldsymbol{K}_{\boldsymbol{H}}(\boldsymbol{x}-\boldsymbol{X}_i)\Big\} \\
&\overset{(P1),P(9)}{=} \mathrm{E}\Big\{\boldsymbol{K}_{\boldsymbol{H}}(\boldsymbol{x}-\boldsymbol{X}_i)\Big\} \\
&\overset{(P4)}{=} \int \boldsymbol{K}_{\boldsymbol{H}}(\boldsymbol{x}-\boldsymbol{z})f(\boldsymbol{z})d\boldsymbol{z} \\
&= \int \boldsymbol{H}^{-1/2}\boldsymbol{K}\big\{\boldsymbol{H}^{1/2}(\boldsymbol{x}-\boldsymbol{z})\big\}f(\boldsymbol{z})d\boldsymbol{z}.
\end{aligned}$$

Using the change of variable $\boldsymbol{x}-\boldsymbol{z} = \boldsymbol{H}^{1/2}\boldsymbol{u}$ so that $d\boldsymbol{z} = \boldsymbol{H}^{1/2}d\boldsymbol{u}$ gives

$$\begin{aligned}
\mathrm{E}\big\{\hat{f}(\boldsymbol{x})\big\} &= \int \boldsymbol{H}^{-1/2}\boldsymbol{K}(\boldsymbol{u})f(\boldsymbol{x}-\boldsymbol{H}^{1/2}\boldsymbol{u})\boldsymbol{H}^{1/2}d\boldsymbol{u} \\
&= \int \boldsymbol{K}(\boldsymbol{u})f(\boldsymbol{x}-\boldsymbol{H}^{1/2}\boldsymbol{u})d\boldsymbol{u}.
\end{aligned}$$

By a Taylor expansion of $f(\boldsymbol{x}-\boldsymbol{H}^{1/2}\boldsymbol{u})$ around $\boldsymbol{x}$ we get

$$\begin{aligned}
\mathrm{E}\big\{\hat{f}(\boldsymbol{x})\big\} &= \int \boldsymbol{K}(\boldsymbol{u})\Big\{f(\boldsymbol{x}) - (\boldsymbol{H}^{1/2}\boldsymbol{u})^\top\mathcal{D}_{f_x} + \frac{1}{2}(\boldsymbol{H}^{1/2}\boldsymbol{u})^\top\mathcal{H}_{f_x}(\boldsymbol{H}^{1/2}\boldsymbol{u})\Big\} \\
&\quad + \mathrm{o}\big\{\mathrm{tr}(\boldsymbol{H})\big\} \\
&= f(\boldsymbol{x}) + \frac{1}{2}\mathrm{tr}\Big\{\boldsymbol{H}^{1/2}\mathcal{H}_{f_x}\boldsymbol{H}^{1/2}\int \boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{K}(\boldsymbol{u})d\boldsymbol{u}\Big\} + \mathrm{o}\big\{\mathrm{tr}(\boldsymbol{H})\big\}.
\end{aligned}$$

Regarding the variance we have, fixing i in second step below

$$\begin{aligned}\mathrm{Var}\big\{\hat{f}(\boldsymbol{x})\big\} &= \mathrm{Var}\big\{n^{-1}\sum_{i=1}^{n}\boldsymbol{K}_{\boldsymbol{H}}(\boldsymbol{x}-\boldsymbol{X}_i)\big\}\\ &\overset{(P12),P(13)}{=} n^{-1}\mathrm{Var}\big\{\boldsymbol{K}_{\boldsymbol{H}}(\boldsymbol{x}-\boldsymbol{X}_i)\big\}\\ &\overset{(P11)}{=} n^{-1}\left(\mathrm{E}\Big\{\boldsymbol{K}_{\boldsymbol{H}}(\boldsymbol{x}-\boldsymbol{X}_i)^2\Big\}-\Big[\mathrm{E}\Big\{\boldsymbol{K}_{\boldsymbol{H}}(\boldsymbol{x}-\boldsymbol{X}_i)\Big\}\Big]^2\right).\end{aligned}$$

Using the change of variable $\boldsymbol{x}-\boldsymbol{z}=\boldsymbol{H}^{1/2}\boldsymbol{u}$ so that $d\boldsymbol{z}=\boldsymbol{H}^{1/2}d\boldsymbol{u}$ gives

$$\begin{aligned}\mathrm{Var}\big\{\hat{f}(\boldsymbol{x})\big\} = n^{-1}\Bigg[&|\boldsymbol{H}|^{-1/2}\int \boldsymbol{K}(\boldsymbol{u})^2 f(\boldsymbol{x}-\boldsymbol{H}^{1/2}\boldsymbol{u})d\boldsymbol{u}\\ &-\Big\{\int \boldsymbol{K}(\boldsymbol{u})f(\boldsymbol{x}-\boldsymbol{H}^{1/2}\boldsymbol{u})\Big\}^2\Bigg].\end{aligned}$$

Then, by (2.5) we have

$$\mathrm{Var}\big\{\hat{f}(\boldsymbol{x})\big\}= n^{-1}|\boldsymbol{H}|^{-1/2}f(\boldsymbol{x})\int \boldsymbol{K}(\boldsymbol{u})^2 d\boldsymbol{u}+\mathrm{o}(n^{-1}|\boldsymbol{H}|^{-1/2}).$$

□

Just as in the univariate case, some important choices have to be made when constructing a multivariate kernel density estimator. First of all, the d-variate kernel has to be selected. Secondly, one has to decide on the particular smoothing parametrization. A full bandwidth matrix allows for more flexibility; however, it also introduces more complexity into the estimator since more parameters need to be chosen. A simplification of (2.2) can be obtained by imposing the restriction that $\boldsymbol{H}$ is a diagonal positive definite $d\times d$ matrix. Then for $\boldsymbol{H}=\mathrm{diag}(h_1,\ldots,h_d)$ the kernel estimator can be written (see Epanechnikov (1969))

$$\hat{f}(\boldsymbol{x}) = n^{-1}\Big(\prod_{l=1}^{d}h_l\Big)^{-1}\sum_{i=1}^{n}K\Big(\frac{x_1-X_{i1}}{h_1},\ldots,\frac{x_d-X_{id}}{h_d}\Big),$$

A further simplification follows when we assume that $\boldsymbol{H}=h^2I$ with $h>0$. This choice leads to the single bandwidth kernel estimator (see Cacoullos (1964))

$$\hat{f}(\boldsymbol{x}) = n^{-1}h^{-d}\sum_{i=1}^{n}\boldsymbol{K}\big\{(\boldsymbol{x}-\boldsymbol{X}_i)/h\big\}.$$

Thus, we see that there is a hierarchical class of smoothing parametrizations from which to choose when using a multivariate kernel estimator. We will discuss the implications of this choice in detail in Chapter 5.

Example 2.2.1. In this example we simulate data from a bivariate normal distribution. In particular, we produce a 2-dimensional normal variable with $\mu = (0,0)^\top$ and $\Sigma = \begin{pmatrix} 1.50 & 0.25 \\ 0.25 & 0.50 \end{pmatrix}$. In addition, we take $\boldsymbol{H} = \begin{pmatrix} 1.25 & 0 \\ 0 & 0.75 \end{pmatrix}$, a diagonal and positive defined bandwidth matrix. Then, a plot using the KDE object structure is depicted in Fig. 2.6. To conclude with, a perspective plot of the previous estimation is given in Fig. 2.7.

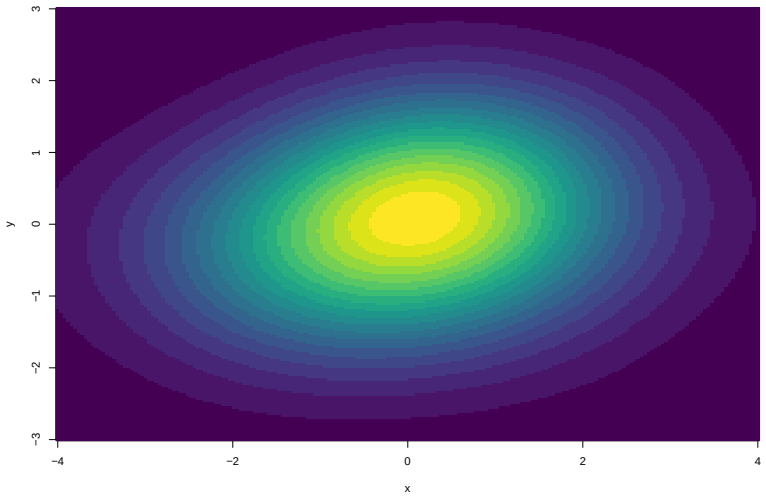

Figure 2.6: Multivariate kernel density estimation

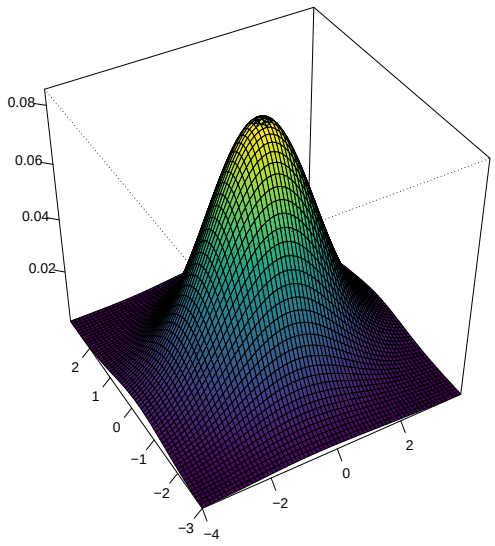

Figure 2.7: Perspective plot of multivariate kernel density estimation

2.3 A continuous non-parametric survival function estimator

Through this section we assume that the available data are the pairs (Y_i, δ_i), $i = 1, \ldots, n$, where $Y_i = \min(T_i, U_i)$ and $\delta_i = \mathbf{1}_{\{T_i \leq U_i\}}$ with $\mathbf{1}_{\{\cdot\}}$ the indicator function of $\{\cdot\}$. Recall that F_T is the c.d.f. of T and G is the c.d.f. of U. The c.d.f. of Y denoted by F, satisfies $1 - F(y) = \big\{1 - F_T(y)\big\}\big\{1 - G(y)\big\}$. An estimate of the unknown survival function can be defined by $\hat{S}_T(y) = 1 - \hat{F}_T(y)$ where

$$\hat{F}_T(y) = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i}{1 - G(Y_i)} W\left(\frac{y - Y_i}{h}\right),$$

where

$$W(y) = \int_{-\infty}^{y} K(u)du.$$

We recall that the real-valued function K is called kernel and integrates to 1, while h is called bandwidth and controls the amount of smoothing applied to the estimate. Estimator $\hat{S}_T(y)$ cannot be used directly in practice as it involves the unknown censoring distribution $G(y)$. An estimate of the unknown censoring distribution $G(y)$ is obtained by reversing the intuitive role played by T_i and U_i and estimate $1 - G(y)$ by the (sightly modified) Kaplan-Meier estimator,

$$1 - \hat{G}(y) = \begin{cases} 1, & 0 \leq y \leq Z_1, \\ \prod_{i=1}^{k-1} \left(\frac{n-i+1}{n-i+2}\right)^{1-\Lambda_i}, & Z_{k-1} < y \leq Z_k, \quad k = 2, \ldots, n, \\ \prod_{i=1}^{n} \left(\frac{n-i+1}{n-i+2}\right)^{1-\Lambda_i}, & y > Z_n, \end{cases}$$

where (Z_i, Λ_i) are the ordered Y_i's, along with their censoring indicators δ_i, $i = 1, \ldots, n$. This gives rise to the practically useful estimator

$$\hat{S}_n(y) = 1 - \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i}{1 - \hat{G}(Y_i)} W\left(\frac{y - Y_i}{h}\right).$$

The estimator $\hat{S}_T(y)$ was employed in the context of survival function estimation in Gulati and Padgett (1996) where its asymptotic mean square error, optimal (with respect to MSE) bandwidth, strong uniform consistency and convergence to a mean zero Gaussian process were established. In addition, Kulasekera et al. (2001) and Lemdani and Ould-Saıd (2003) respectively provided the conditions under which the estimate has superior Mean Square Error

(MSE) and Mean Integrated Absolute Error compared to the Kaplan-Meier Estimate. Applications of $\hat{S}_T(y)$ include the works of Claeskens and Hall (2002) and Kim et al. (2005) on kernel hazard rate estimation. Moreover, the distribution estimate $1 - \hat{S}_T(y)$ has been applied extensively on quantile estimation under random right censorship, see Lio and Padgett (1992) and the references therein. To sum up, this estimator can be used when we have censored data and we want to avoid the discontinuity of the Kaplan-Meier estimator.

CHAPTER 3

Estimation of Conditional Distribution Function

3.1 Conditional c.d.f. estimation under random right censoring

Throughout this section the available data are the pairs of random variables (Y, δ) where $Y = \min(T, U)$ and $\delta = \mathbf{1}_{\{T \leq U\}}$. Recall that $T \sim f_T$ and $U \sim f_U$. Usually there are more than one factors that affect survival at each given time point. Covariate information is typically assumed real valued and denoted as a d-dimensional vector, say $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$, coming from a d-variate p.d.f. $f_{\boldsymbol{X}}$ and c.d.f. $F_{\boldsymbol{X}}$, with support, $\text{supp}(f_{\boldsymbol{X}}) \equiv S = [M_1, T_1] \times \cdots \times [M_d, T_d] \subset \mathbb{R}^d$. Conditioning on covariate information it is common to assume that survival and censoring times, i.e. $T|\boldsymbol{X}$ and $U|\boldsymbol{X}$ are independent from each other. Hence, the way that covariates $X_1, \ldots, X_d$ affect survival times is independent from the way that affect the censoring variable.
Let $(\boldsymbol{X}_i, Y_i, \delta_i)$ be an i.i.d. random sample from the population $(\boldsymbol{X}, Y, \delta)$, with $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{id})^\top$. Given $\boldsymbol{X} = \boldsymbol{x}$, the variables Y and U have conditional distribution functions $F(\cdot|\boldsymbol{X})$ and $G(\cdot|\boldsymbol{X})$ respectively, while $F_T(\cdot|\boldsymbol{X})$ is the conditional distribution function of T given $\boldsymbol{X}$. From the independence of T and U given $\boldsymbol{X}$ it holds that

$$\begin{aligned} F(Y|\boldsymbol{X}) &= P(Y \leq y|\boldsymbol{X} = \boldsymbol{x}) \\ &= P(\min(T, U) \leq y|\boldsymbol{X} = \boldsymbol{x}) \\ &= P\big\{(T \leq y|\boldsymbol{X} = \boldsymbol{x}) \cap (U \leq y|\boldsymbol{X} = \boldsymbol{x})\big\} \\ &= P(T \leq y|\boldsymbol{X} = \boldsymbol{x})P(U \leq y|\boldsymbol{X} = \boldsymbol{x}). \end{aligned}$$

The objective is to estimate the conditional distribution function $F_T(Y|\boldsymbol{X} = \boldsymbol{x}) = P(Y \leq y|\boldsymbol{X} = \boldsymbol{x})$ which implicitly also provides an estimate of the

conditional survival function

$$S_T(Y|\boldsymbol{X}=\boldsymbol{x}) = P(Y > y|\boldsymbol{X}=\boldsymbol{x}) = 1 - F_T(Y|\boldsymbol{X}=\boldsymbol{x}),$$

based on the data $(\boldsymbol{X}_i, Y_i, \delta_i)$, $i = 1, \ldots, n$.
For any conditional p.d.f. or c.d.f. function, say $g(y|\boldsymbol{x})$, $D_d(y|\boldsymbol{x})$ denotes the gradient vector

$$\mathcal{D}_d(y|\boldsymbol{x}) = \dot{g}(y|\boldsymbol{x}) = \left(\frac{\partial g(y|\boldsymbol{x})}{\partial x_1}, \ldots, \frac{\partial g(y|\boldsymbol{x})}{\partial x_d}\right)^\top,$$

and $H_d(y|\boldsymbol{x})$ the Hessian matrix

$$\mathcal{H}_d(y|\boldsymbol{x}) = \ddot{g}(y|\boldsymbol{x}) = \begin{pmatrix} \frac{\partial^2 g(y|\boldsymbol{x})}{\partial {x_1}^2} & \cdots & \frac{\partial^2 g(y|\boldsymbol{x})}{\partial x_1 x_d} \\ \cdots & \cdots & \cdots \\ \frac{\partial^2 g(y|\boldsymbol{x})}{\partial x_d x_1} & \cdots & \frac{\partial^2 g(y|\boldsymbol{x})}{\partial {x_d}^2} \end{pmatrix}.$$

Let $\hat{S}_T = 1 - \hat{F}_T$ be the Kaplan-Meier estimate for the survival function S_T. In the absence of censoring, the Kaplan-Meier estimator is in fact unbiased. Now, if the last survivor in the sample actually dies, then $\hat{S}_T = 0$ and therefore the Kaplan-Meier estimator remains unbiased. However, the possibility that the last remaining survivor will become censored rather than die is increased, if censoring exists. In this case $\hat{S}_T$ will never drop to zero. Therefore, the possibility of the last survivor becoming censored is the source of bias in the Kaplan-Meier estimator. However, Kaplan-Meier estimator is an asymptotically unbiased estimator, i.e.,

$$\mathrm{E}\{\hat{F}_T(Y)|\boldsymbol{X}=\boldsymbol{x}\} = F_T(y|\boldsymbol{x}).$$

Then, it is reasonable to regard the data

$$\{\hat{F}_T(Y_i), \boldsymbol{X}_i, \delta_i = 1\}, \quad i = 1, \ldots, n,$$

as of coming from the non-parametric regression

$$\hat{F}_T(Y_i) = F_T(y|\boldsymbol{X}_i) + \varepsilon_i(y), \quad i = 1, \ldots, n, \tag{3.1}$$

where the $\varepsilon_i(y)'s$ denote conditionally mean zero error terms since,

$$\begin{aligned} \mathrm{E}(\varepsilon) &= \mathrm{E}\Big\{\mathrm{E}(\varepsilon|\boldsymbol{X})\Big\} \\ &= \mathrm{E}\Big(\mathrm{E}\Big[\Big\{\hat{F}_T(Y) - F_T(Y|\boldsymbol{X})\Big\}\Big|\boldsymbol{X}\Big]\Big) \\ &= \mathrm{E}\Big\{F_T(Y|\boldsymbol{X}) - F_T(Y|\boldsymbol{X})\Big\} \\ &= 0. \end{aligned}$$

This motivates a local linear approach in estimating $F_T(y|\boldsymbol{x})$: in parallel to Fan and Gijbels (1996), a multivariate Taylor expansion of $F_T(y|\boldsymbol{X}=\boldsymbol{z})$ around $\boldsymbol{x}=(x_1,\ldots,x_d)^\top$ reveals that each coefficient of the expansion corresponds to a vector of partial derivatives of the conditional distribution function. Provided that $F_T(y|\boldsymbol{z})$ is a smooth function (a function that has continuous derivatives up to some desired order over some domain) it can be accurately approximated locally by d-dimensional hyperplanes with coefficients β_ν, $\nu=0,1$, i.e.

$$\begin{aligned}\mathrm{E}\{\hat{F}_T(Y)|\boldsymbol{X}=\boldsymbol{z}\} = F_T(y|\boldsymbol{z}) &\simeq F_T(y|\boldsymbol{x}) + \dot{F}_T(y|\boldsymbol{x})^\top(\boldsymbol{z}-\boldsymbol{x}) \\ &\equiv \beta_0 + \beta_1^\top(\boldsymbol{z}-\boldsymbol{x}). \end{aligned} \tag{3.2}$$

By (3.2) it is immediately established that the scalar β_0 corresponds to $F_T(y|\boldsymbol{x})$ and the d-variate vector $\beta_1^\top=(\beta_{1,1},\ldots,\beta_{1,d})^\top$ corresponds to the vector of first order partial derivatives

$$\dot{F}_T(y|\boldsymbol{x}) = \left(\frac{\partial F_T(y|\boldsymbol{x})}{\partial x_1},\ldots,\frac{\partial F_T(y|\boldsymbol{x})}{\partial x_d}\right)^\top_{d\times 1}.$$

Thus estimation of β_0 essentially corresponds to estimation of $F_T(y|\boldsymbol{x})$. By (3.2)

$$F_T(y|\boldsymbol{X}_i) \simeq F_T(y|\boldsymbol{x}) + \dot{F}_T(y|\boldsymbol{x})^\top(\boldsymbol{X}_i-\boldsymbol{x}) \equiv \beta_0 + \beta_1^\top(\boldsymbol{X}_i-\boldsymbol{x}). \tag{3.3}$$

By (3.1) and (3.3)

$$\varepsilon_i(y) = \hat{F}_T(Y_i) - F_T(y|\boldsymbol{X}_i) = \hat{F}_T(Y_i) - \beta_0 - {\beta_1}^\top(\boldsymbol{X}_i-\boldsymbol{x}).$$

Thus, the $\hat{\beta}_\nu$, $\nu=0,1$, are obtained by solving the least squares optimization problem where we can incorporate a weight scheme to down-weigh the contributions of a data point away from $\boldsymbol{x}$. We can assign a weight $\boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i-\boldsymbol{x})\}$ to the point $(\boldsymbol{X}_i,Y_i,\delta_i)$ and by taking the square of $\varepsilon_i(y)$, it leads to the following weighted least squares problem

$$\min_{\beta_0,\beta_1}\sum_{i=1}^{n}\{\hat{F}_T(y)-\beta_0-{\beta_1}^\top(\boldsymbol{X}_i-\boldsymbol{x})\}^2\boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i-\boldsymbol{x})\}, \tag{3.4}$$

where $\boldsymbol{H}$ denotes the bandwidth matrix

$$\boldsymbol{H} = \mathrm{diag}\{h_1,\ldots,h_d\},$$

where we recall that h_i is a positive number, usually called the bandwidth or window width, i.e. is the amount of smoothing applied to the data in the

i^{th}, $i = 1, \ldots, d$ direction and $\boldsymbol{K}$ denotes the d-variate kernel introduced in Section 2.2. The choice of the shape of the kernel function is not a particularly important one. However, the choice of the value for the bandwidth is very important.
We can write (3.4) in matrix notation as

$$\min_{B}(Y - \boldsymbol{X}B)^{\top}\boldsymbol{W}(Y - \boldsymbol{X}B),$$

where,

$$\boldsymbol{X} = \begin{pmatrix} 1 & (\boldsymbol{X}_1 - \boldsymbol{x})^{\top} \\ \ldots & \ldots \\ 1 & (\boldsymbol{X}_n - \boldsymbol{x})^{\top} \end{pmatrix}_{n\times(d+1)}, \quad Y = \left(\hat{F}_T(y_1), \ldots, \hat{F}_T(y_n)\right)^{\top}_{n\times 1},$$

$$\boldsymbol{W} = \operatorname{diag}\Big[\boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_1 - \boldsymbol{x})\}, \ldots, \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_n - \boldsymbol{x})\}\Big]_{n\times n}, \quad B = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}_{(d+1)\times 1}.$$

Hence, we have to minimize S with respect to B, where S is

$$\begin{aligned} S &= (Y - \boldsymbol{X}B)^{\top}\boldsymbol{W}(Y - \boldsymbol{X}B) && (3.5)\\ &= (Y^{\top} - B^{\top}\boldsymbol{X}^{\top})(\boldsymbol{W}Y - \boldsymbol{W}\boldsymbol{X}B) \\ &= Y^{\top}\boldsymbol{W}Y - Y^{\top}\boldsymbol{W}\boldsymbol{X}B - B^{\top}\boldsymbol{X}^{\top}\boldsymbol{W}Y + B^{\top}\boldsymbol{X}^{\top}\boldsymbol{W}\boldsymbol{X}B. && (3.6) \end{aligned}$$

By taking the derivative of (3.6) with respect to B so equally to 0 and solving for B yields

$$\hat{B} = (\boldsymbol{X}^{\top}\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^{T}\boldsymbol{W}Y = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}. \tag{3.7}$$

Thus,

$$\hat{\beta}_0 = \boldsymbol{e}_1(\mathbf{X}^{\top}\boldsymbol{W}\mathbf{X})^{-1}\mathbf{X}^{T}\boldsymbol{W}Y, \tag{3.8}$$

where,

$$\boldsymbol{e}_1 = (1, 0, \ldots, 0)_{1\times(d+1)}.$$

In order to find $\hat{\beta}_0$ we will first calculate the quantity $\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X}$.

$$\mathbf{X}^\top \boldsymbol{W}\mathbf{X} = \begin{pmatrix} 1 & \dots & 1 \\ (X_{11}-x_1) & \dots & (X_{n1}-x_1) \\ \vdots & \ddots & \vdots \\ (X_{1d}-x_d) & \dots & (X_{nd}-x_d) \end{pmatrix}_{(d+1)\times n}$$
$$\begin{pmatrix} \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_1-\boldsymbol{x})\} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_n-\boldsymbol{x})\} \end{pmatrix}_{n\times n} \times$$
$$\begin{pmatrix} 1 & (X_{11}-x_1) & \dots & (X_{1d}-x_d) \\ \vdots & \ddots & \ddots & \vdots \\ 1 & (X_{n1}-x_1) & \dots & (X_{nd}-x_d) \end{pmatrix}_{n\times(d+1)}$$
$$= \begin{pmatrix} S_{n,0}(\boldsymbol{x}) & S_{n,1}(\boldsymbol{x})^\top \\ S_{n,1}(\boldsymbol{x}) & S_{n,2}(\boldsymbol{x}) \end{pmatrix}_{(d+1)\times(d+1)}.$$

The calculation of $(\mathbf{X}^\top \boldsymbol{W}\mathbf{X})^{-1}$ is in the Appendix.
Now, we have to calculate the quantity $\boldsymbol{X}^\top \boldsymbol{W}Y$ which is given by

$$\begin{pmatrix} 1 & \dots & 1 \\ (\boldsymbol{X}_1-\boldsymbol{x})^\top & \dots & (\boldsymbol{X}_n-\boldsymbol{x})^\top \end{pmatrix}$$
$$\begin{pmatrix} \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_1-\boldsymbol{x})\} & \dots & 0 \\ \vdots & \ddots & \\ 0 & \dots & \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_n-\boldsymbol{x})\} \end{pmatrix} \begin{pmatrix} \hat{F}_T(y) \\ \dots \\ \hat{F}_T(y) \end{pmatrix}$$
$$= \begin{pmatrix} \sum_{i=1}^n \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i-\boldsymbol{x})\}\hat{F}_T(y) \\ \sum_{i=1}^n \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i-\boldsymbol{x})\}(\boldsymbol{X}_i-\boldsymbol{x})\hat{F}_T(y) \end{pmatrix}. \quad (3.9)$$

For $d=1$ an equivalent form of (3.8), say $\tilde{F}_L(y|\boldsymbol{x})$, is given by

$$\tilde{F}_L(y|\boldsymbol{x}) \equiv \hat{\beta}_0 = \sum_{i=1}^n \frac{S_{n,2}(\boldsymbol{x}) - S_{n,1}(\boldsymbol{x})(\boldsymbol{X}_i-\boldsymbol{x})}{S_{n,2}(\boldsymbol{x})S_{n,0}(\boldsymbol{x}) - S_{n,1}(\boldsymbol{x})S_{n,1}(\boldsymbol{x})} \times$$
$$\boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i-\boldsymbol{x})\}\hat{F}_T(y). \quad (3.10)$$

For $d > 1$ note that we only need the first row of the $(\mathbf{X}^\top \boldsymbol{W}\mathbf{X})^{-1}$. Suppose that,

$$\begin{aligned} p_1 &= (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1} \\ &= \left\{S_{n,0}(\boldsymbol{x}) - S_{n,1}(\boldsymbol{x})^\top S_{n,2}(\boldsymbol{x})^{-1} S_{n,1}(\boldsymbol{x})\right\}^{-1}, \\ \boldsymbol{p}_2 &= -\boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1} \\ &= -S_{n,0}(\boldsymbol{x})^{-1} S_{n,1}(\boldsymbol{x})^\top \left\{S_{n,2}(\boldsymbol{x}) - S_{n,1}(\boldsymbol{x}) S_{n,0}(\boldsymbol{x})^{-1} S_{n,1}(\boldsymbol{x})^\top\right\}^{-1}, \end{aligned}$$

with $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ and $\boldsymbol{D}$ given in Appendix, gives,

$$\begin{aligned} \tilde{F}_L(y|\boldsymbol{x}) \equiv \hat{\beta}_0 &= \begin{pmatrix} p_1 & \boldsymbol{p}_2 \end{pmatrix}_{1\times(d+1)} \begin{pmatrix} \sum_{i=1}^n \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\}\hat{F}_T(y) \\ \sum_{i=1}^n \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\}(\boldsymbol{X}_i - x)\hat{F}_T(y) \end{pmatrix}_{(d+1)\times 1} \\ &= p_1 \sum_{i=1}^n \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\}\hat{F}_T(y) \qquad (3.11) \\ &\quad + \boldsymbol{p}_2 \sum_{i=1}^n \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\}(\boldsymbol{X}_i - x)\hat{F}_T(y) \\ &= \sum_{i=1}^n \Big\{p_1 + \boldsymbol{p}_2(\boldsymbol{X}_i - x)\Big\}\boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\}\hat{F}_T(y). \qquad (3.12) \end{aligned}$$

where,

$$\begin{aligned} S_{n,0}(\boldsymbol{x}) &= \sum_{i=1}^n \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\}, \\ S_{n,1}(\boldsymbol{x}) &= \sum_{i=1}^n \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\}(\boldsymbol{X}_i - \boldsymbol{x}), \\ S_{n,2}(\boldsymbol{x}) &= \sum_{i=1}^n \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\}(\boldsymbol{X}_i - \boldsymbol{x})(\boldsymbol{X}_i - \boldsymbol{x})^\top. \qquad (3.13) \end{aligned}$$

Note that $\tilde{S}_L(y|\boldsymbol{x})$ is continuous only in $\boldsymbol{x}$. Now, let $\boldsymbol{K}^*_{\boldsymbol{H}}(\cdot) = |\boldsymbol{H}|^{-1}\boldsymbol{K}(\boldsymbol{H}^{-1}\cdot)$ and for any real function g set $R(g) = \int g^2$. The following assumptions and definitions are used throughout the thesis.

A1. $f_{\boldsymbol{X}}$ is continuous at $\boldsymbol{x}$, while $S_T(y|\boldsymbol{x})$ is twice continuously partially differentiable in a neighbourhood of $\boldsymbol{x}$.

A2. The point $\boldsymbol{x} = (x_1, ..., x_d)^\top$ is in the support of $f_{\boldsymbol{X}}$. At $\boldsymbol{x}$, F_T is continuously differentiable and all second-order derivatives of F_T are continuous.

A3. The sequence of bandwidth matrices $\boldsymbol{H}$ is such that $n^{-1}|\boldsymbol{H}|$ and each entry of $\boldsymbol{H}$ tends to zero as $n \to \infty$, with $\boldsymbol{H}$ remaining symmetric and positive definite. Also, there is a fixed constant L such that the condition number of $\boldsymbol{H}$ (i.e., the ratio of its largest to its smallest eigenvalue) is at most L for all n.

A4. $|\boldsymbol{H}|$ converges to zero in such a way so that $n\boldsymbol{H} \to \infty$ and $\boldsymbol{H} \log n \to 0$.

A5. The kernel $\boldsymbol{K}$ is a compactly supported, bounded kernel such that

$$\int uu^\top \boldsymbol{K(u)} = \mu_2(\boldsymbol{K})I,$$

where $\mu_2(\boldsymbol{K})$ is a scalar and I is the $d \times d$ identity matrix. In addition, all odd-order moments of $\boldsymbol{K}$ vanish, that is,

$$\int u_1^{l_1} ... u_d^{l_d} \boldsymbol{K(u)} = 0,$$

for all non-negative integers $l_1, ..., l_d$ such that their sum is odd. Note that this last condition is satisfied by spherically symmetric kernels and product kernels based on symmetric univariate kernels. Furthermore, $\boldsymbol{K}$ satisfies

$$R(\boldsymbol{K}) = \int \boldsymbol{K}^2 < \infty, \int |u^2 \boldsymbol{K}| < \infty, \int \boldsymbol{K} = 1 \quad \text{and} \quad \int u\boldsymbol{K} = 0.$$

An important consequence of taking $\text{supp}(f_{\boldsymbol{X}}) \equiv S$ is that S is compact, closed and bounded with smooth boundary, satisfying $cl(S_{int}) = S_{int}$, where $cl(\cdot)$ is the closure of the set $(\cdot)$ and

$$S_{int} = [M_1 + h_1, T_1 - h_1] \times \cdots \times [M_d + h_d, T_d - h_d].$$

Then, since for any p.d.f. $f_{\boldsymbol{X}} \geq 0$ on S it follows that a discontinuity will occur across the boundary of S, denoted by ∂S, while $f_{\boldsymbol{X}}$ is smooth on S.

Since $J = [1, 1]^d$ is a compact and simply connected set which satisfies $cl(J_{int}) = J$, then the effective kernel support or smoothing window of $\boldsymbol{K}$ is $J_n(\boldsymbol{x}) \cap S$, where $J_n(\boldsymbol{x}) = \boldsymbol{x} - \boldsymbol{H}J$. The assumptions on S and the properties of J ensure that $J_n(\boldsymbol{x}) \cap S$ is measurable and has a positive measure for all $\boldsymbol{x} \in S$. Now, let $\boldsymbol{x} \in S$. Transposing and rescaling the support J of the kernel, leads to the set

$$J_{\boldsymbol{x}} = \boldsymbol{H}^{-1}\big\{\boldsymbol{x} - J_n(\boldsymbol{x}) \cap S\big\} = \{z \in \mathbb{R}^d : z = \boldsymbol{H}^{-1}(\boldsymbol{x} - \boldsymbol{y}), \quad y \in J_n(\boldsymbol{x}) \cap S\}.$$

Since only data falling in $J_n(\boldsymbol{x}) \cap S$ contribute to $\tilde{F}_L(y|\boldsymbol{x})$, for $\boldsymbol{x} \in S_{int}$, it holds that $J_n(\boldsymbol{x}) \cap S = J_n(\boldsymbol{x}) \subset S$ which implies that $J_{\boldsymbol{x}} = J$. Consequently, when $\boldsymbol{x} \in S_{int}$, the moment conditions of assumption A5 hold and the estimate behaves as expected. Note that the same is true asymptotically for all $\boldsymbol{x} \in S$ as by assumption A3, $J_{\boldsymbol{x}} = J$, for $n \to \infty$. However, from a finite sample standpoint, when $\boldsymbol{x}$ is in the boundary of S, i.e. for $\boldsymbol{x} \in S_{b,l}$ or $\boldsymbol{x} \in S_{b,r}$, where

$$S_{b,l} = [M_1, M_1+h_1]\times\cdots\times[M_d, M_d+h_d], \quad S_{b,r} = [T_1-h_1, T_1]\times\cdots\times[T_d-h_d, T_d],$$

then $J_n(\boldsymbol{x}) \cap S \neq J_n(\boldsymbol{x})$ or equivalently $J_n(\boldsymbol{x}) \cap S \not\subset S$ and hence part of the smoothing window is devoid of data. For example, in the case where $\boldsymbol{x} \in S_{b,l}$, the effective support of $\boldsymbol{K}$ is actually $J_{\boldsymbol{x}} = [1, c_1] \times \cdots \times [1, c_d] \equiv D$, with $0 < c_i < 1$, for all $i \in 1, \ldots, d$ and consequently the moment conditions for $\boldsymbol{K}$ in assumption A5 no longer hold. For the same reason, the effective kernel support of $\boldsymbol{K}$ is also asymmetric in the right boundary as $J_{\boldsymbol{x}} = [c_1, 1] \times \cdots \times [c_d, 1]$, for $\boldsymbol{x} \in S_{b,r}$. Thus employing $\boldsymbol{K}$ without any modification for kernel estimation of $F_T(y|x)$ is expected to result in inflated bias at any $\boldsymbol{x} \in S_b = S_{b,l} \cap S_{b,r}$ in the usual asymptotic analysis. Focusing on the left boundary, as treatment of the right boundary is similar, the automatic boundary adjustment of $\tilde{F}_L(y|\boldsymbol{x})$ and its asymptotic properties are discussed next.

CHAPTER 4

Asymptotic properties of Conditional Distribution Function

4.1 Asymptotic properties

Now, let $\mathbf{M} = (M_1, \ldots, M_d)$ and $\mathbf{T} = (T_1, \ldots, T_d)$ where T_i, M_i, $i = 1, \ldots, d$ denote positive constants and let

$$S_{b,l} = [M_1, M_1+h_1]\times\cdots\times[M_d, M_d+h_d], \quad S_{b,r} = [T_1-h_1, T_1]\times\cdots\times[T_d-h_d, T_d],$$

so that $S_b = S_{b,l} \cup S_{b,r}$ is the boundary of $\text{supp}(f_{\boldsymbol{X}})$. For a positive d-variate constant $\boldsymbol{c} = (c_1, \ldots, c_d)$ if $(0, \ldots, 0)_{d\times 1} < \boldsymbol{c} < (1, \ldots, 1)_{d\times 1}$ then $\boldsymbol{x} = \mathbf{M} + \boldsymbol{Hc} \in S_{b,l}$ is a left boundary point, while if $(1, \ldots, 1)_{d\times 1} < \boldsymbol{c} < \text{diag}\{M_1/h_1 - 1, \ldots, M_d/h_d - 1\}$, so that $\boldsymbol{x} = \mathbf{M} + \boldsymbol{Hc} \in \text{supp}(f_{\boldsymbol{X}})$ and in the right boundary $\text{diag}\{T_1/h_1 - 1, \ldots, T_d/h_d - 1\} \leq \boldsymbol{c} \leq \text{diag}\{T_1/h_1, \ldots, T_d/h_d\}$ so that $\boldsymbol{x} = \mathbf{T} - \boldsymbol{Hc} \in S_{b,r}$. The analysis focuses on the left boundary as treatment of the right boundary is similar. First denote the bias of $\hat{F}_L(y|\boldsymbol{x})$ by

$$b_{\tilde{F}_{L(,c)}}(y|\boldsymbol{x}) = \begin{cases} b_{\tilde{F}_{L,c}}(y|\boldsymbol{x}), & y \in [0, T_Y], \quad \boldsymbol{x} \in [0, h_j]^d \cup [M_j - h_j, M_j]^d, \\ b_{\tilde{F}_L}(y|\boldsymbol{x}), & y \in [0, T_Y], \quad \boldsymbol{x} \in [M_j - h_j, M_j]^d, \end{cases}$$

and its variance by

$$\sigma^2{}_{\tilde{F}_{L(,c)}}(y|\boldsymbol{x}) = \begin{cases} \sigma^2{}_{\tilde{F}_{L,c}}(y|\boldsymbol{x}), & y \in [0, T_Y], \quad \boldsymbol{x} \in [0, h_j]^d \cup [M_j - h_j, M_j]^d, \\ \sigma^2{}_{\tilde{F}_L}(y|\boldsymbol{x}), & y \in [0, T_Y], \quad \boldsymbol{x} \in [M_j - h_j, M_j]^d, \end{cases}$$

for $j = 1, \ldots, d$. Also let

$$\begin{aligned} S_{0,c} &= \int_{-c}^{\infty} \boldsymbol{K(u)} d\boldsymbol{u}, \\ S_{1,c} &= \int_{-c}^{\infty} \boldsymbol{u K(u)} d\boldsymbol{u}, \\ S_{2,c} &= \int_{-c}^{\infty} \boldsymbol{u}^2 \boldsymbol{K(u)} d\boldsymbol{u}. \end{aligned} \tag{4.1}$$

Lemma 4.1.1. *Assume that* $|\boldsymbol{H}| \to 0$ *and* $n|\boldsymbol{H}| \to \infty$. *Then*

$$S_{n,l} = n|\boldsymbol{H}|^{l+1} f_{\boldsymbol{X}}(\boldsymbol{x}) S_{l,c}\{1 + o_p(1)\}, \quad l = 0, 1, 2.$$

Proof. The proof of lemma is given for $j = 2$. The cases $j = 0, 1$ are proved in an entirely similar manner. We use the $P1$, $P4$ and $P9$-$P13$ properties from the appendix. In addition, first, write

$$\begin{aligned} S_{n,2} &= \sum_{i=1}^{n} \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\}(\boldsymbol{X}_i - \boldsymbol{x})(\boldsymbol{X}_i - \boldsymbol{x})^\top \\ &= \mathrm{E}\big\{S_{n,2}(\boldsymbol{x})\big\} + \mathrm{Var}\big\{S_{n,2}(\boldsymbol{x})\big\}. \end{aligned} \tag{4.2}$$

For fixed i in the third step below

$$\begin{aligned} \mathrm{E}\big\{S_{n,2}(\boldsymbol{x})\big\} &= \mathrm{E}\Big[\sum_{i=1}^{n} \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\}(\boldsymbol{X}_i - \boldsymbol{x})(\boldsymbol{X}_i - \boldsymbol{x})^\top\Big] \\ &\overset{(P1),(P9)}{=} n\mathrm{E}\Big[\boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\}(\boldsymbol{X}_i - \boldsymbol{x})(\boldsymbol{X}_i - \boldsymbol{x})^\top\Big] \\ &\overset{(P4)}{=} \int n\boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{u} - \boldsymbol{x})\}(\boldsymbol{X}_i - \boldsymbol{x})(\boldsymbol{X}_i - \boldsymbol{x})^\top f_{\boldsymbol{X}}(\boldsymbol{u}) d\boldsymbol{u}. \end{aligned}$$

Using the change of variable $\boldsymbol{u} - \boldsymbol{x} = \boldsymbol{Hz}$ so that $d\boldsymbol{u} = \boldsymbol{H}d\boldsymbol{z}$ gives

$$\begin{aligned} \mathrm{E}\big\{S_{n,2}(\boldsymbol{x})\big\} &= \int n\boldsymbol{K}(\boldsymbol{z})(\boldsymbol{Hz})(\boldsymbol{Hz})^\top f_{\boldsymbol{X}}(\boldsymbol{x} + \boldsymbol{Hz}) \boldsymbol{H} d\boldsymbol{z} \\ &= n|\boldsymbol{H}|^3 \int \boldsymbol{z}^\top \boldsymbol{z} \boldsymbol{K}(\boldsymbol{z}) f_{\boldsymbol{X}}(\boldsymbol{x} + \boldsymbol{Hz}) d\boldsymbol{z}. \end{aligned}$$

Now, by using (2.5) we get

$$\begin{aligned}
\mathrm{E}\big\{S_{n,2}(\boldsymbol{x})\big\} &= n|\boldsymbol{H}|^3 \int \boldsymbol{z}^\top \boldsymbol{z}\boldsymbol{K}(\boldsymbol{z})\bigg[f_{\boldsymbol{X}}(\boldsymbol{x}) + (\boldsymbol{H}\boldsymbol{z})^\top \mathcal{D}_{f\boldsymbol{X}}(\boldsymbol{x}) \\
&\quad + \frac{1}{2}(\boldsymbol{H}\boldsymbol{z})^\top \mathcal{H}_{f_{\boldsymbol{X}}}(\boldsymbol{x})\boldsymbol{H}\boldsymbol{z} + \mathrm{o}\big\{\mathrm{tr}(\boldsymbol{H}^2)\big\}\bigg] d\boldsymbol{z} \\
&= n|\boldsymbol{H}|^3 f_{\boldsymbol{X}}(\boldsymbol{x}) \int \boldsymbol{z}^\top \boldsymbol{z}\boldsymbol{K}(\boldsymbol{z}) d\boldsymbol{z}\{1 + \mathrm{o}_p(|\boldsymbol{H}|)\} \\
&= n|\boldsymbol{H}|^3 f_{\boldsymbol{X}}(\boldsymbol{x}) S_{2,c}\{1 + \mathrm{o}_p(|\boldsymbol{H}|)\}. \qquad (4.3)
\end{aligned}$$

Regarding the variance we have

$$\begin{aligned}
\mathrm{Var}\big\{S_{n,2}(\boldsymbol{x})\big\} &= \mathrm{Var}\bigg[n^{-1}\sum_{i=1}^{n} \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\}(\boldsymbol{X}_i - \boldsymbol{x})(\boldsymbol{X}_i - \boldsymbol{x})^\top\bigg] \\
&\overset{(P12),(P13)}{=} n^{-1}\mathrm{Var}\bigg[\boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\}(\boldsymbol{X}_i - \boldsymbol{x})(\boldsymbol{X}_i - \boldsymbol{x})^\top\bigg] \\
&\overset{(P11)}{=} n^{-1}\bigg(\mathrm{E}\Big[\Big\{\boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\}(\boldsymbol{X}_i - \boldsymbol{x})(\boldsymbol{X}_i - \boldsymbol{x})^\top\Big\}^2\Big] \\
&\quad - \Big[\mathrm{E}\Big\{\boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\}(\boldsymbol{X}_i - \boldsymbol{x})(\boldsymbol{X}_i - \boldsymbol{x})^\top\Big\}\Big]^2\bigg).
\end{aligned}$$

For $\boldsymbol{X}_i = \boldsymbol{u}$, using the change of variable $\boldsymbol{u} - \boldsymbol{x} = \boldsymbol{H}\boldsymbol{z}$ so that $d\boldsymbol{u} = \boldsymbol{H}d\boldsymbol{z}$ and fixing i in the first step below

$$\begin{aligned}
\mathrm{Var}\big\{S_{n,2}(\boldsymbol{x})\big\} &\overset{(P4)}{=} n^{-1}\bigg[\int \boldsymbol{K}(\boldsymbol{z})^2(\boldsymbol{H}\boldsymbol{z})(\boldsymbol{H}\boldsymbol{z})^\top(\boldsymbol{H}\boldsymbol{z})(\boldsymbol{H}\boldsymbol{z})^\top f_{\boldsymbol{X}}(\boldsymbol{x} + \boldsymbol{H}\boldsymbol{z})\boldsymbol{H}d\boldsymbol{z} \\
&\quad - \bigg\{\int \boldsymbol{K}(\boldsymbol{z})(\boldsymbol{H}\boldsymbol{z})(\boldsymbol{H}\boldsymbol{z})^\top f_{\boldsymbol{X}}(\boldsymbol{x} + \boldsymbol{H}\boldsymbol{z})\boldsymbol{H}d\boldsymbol{z}\bigg\}^2\bigg] \\
&\overset{(P10)}{=} n^{-1}|\boldsymbol{H}|\bigg[\int \boldsymbol{K}(\boldsymbol{z})^2(\boldsymbol{H}\boldsymbol{z})(\boldsymbol{H}\boldsymbol{z})^\top(\boldsymbol{H}\boldsymbol{z})(\boldsymbol{H}\boldsymbol{z})^\top f_{\boldsymbol{X}}(\boldsymbol{x} + \boldsymbol{H}\boldsymbol{z})d\boldsymbol{z} \\
&\quad - \bigg\{|\boldsymbol{H}|\int \boldsymbol{K}(\boldsymbol{z})(\boldsymbol{H}\boldsymbol{z})(\boldsymbol{H}\boldsymbol{z})^\top f_{\boldsymbol{X}}(\boldsymbol{x} + \boldsymbol{H}\boldsymbol{z})d\boldsymbol{z}\bigg\}^2\bigg] \\
&= \mathcal{O}_p\big(n^{-1}|\boldsymbol{H}|\big). \qquad (4.4)
\end{aligned}$$

Combining (4.2), (4.3) and (4.4) finishes the proof. □

A useful and asymptotically equivalent representation of $\tilde{F}_L(y|\boldsymbol{x})$ by using (3.13), (4.1) and Lemma 4.1.1 is obtained by the relationship

$$\begin{aligned}
\tilde{F}_L(y|\boldsymbol{x}) &= \sum_{i=1}^{n} \frac{S_{n,2}(\boldsymbol{x}) - S_{n,1}(\boldsymbol{x})(\boldsymbol{X}_i - \boldsymbol{x})}{S_{n,2}(\boldsymbol{x})S_{n,0}(\boldsymbol{x}) - S_{n,1}(\boldsymbol{x})S_{n,1}(\boldsymbol{x})} \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\}\hat{F}_T(y) \\
&= \sum_{i=1}^{n} \frac{n|\boldsymbol{H}|^3 f_{\boldsymbol{X}}(\boldsymbol{x})S_{2,\boldsymbol{c}} - n|\boldsymbol{H}|^2 f_{\boldsymbol{X}}(\boldsymbol{x})S_{1,\boldsymbol{c}}(\boldsymbol{X}_i - \boldsymbol{x})}{n|\boldsymbol{H}|^3 f_{\boldsymbol{X}}(\boldsymbol{x})S_{2,\boldsymbol{c}} n|\boldsymbol{H}| f_{\boldsymbol{X}}(\boldsymbol{x})S_{0,\boldsymbol{c}} - \left(n|\boldsymbol{H}|^2 f_{\boldsymbol{X}}(\boldsymbol{x})S_{1,\boldsymbol{c}}\right)^2} \times \\
&\quad \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\}\hat{F}_T(y) \\
&= \frac{1}{n|\boldsymbol{H}| f_{\boldsymbol{X}}(\boldsymbol{x})} \sum_{i=1}^{n} \frac{S_{2,\boldsymbol{c}} - S_{1,\boldsymbol{c}}|\boldsymbol{H}|^{-1}(\boldsymbol{X}_i - \boldsymbol{x})}{S_{2,\boldsymbol{c}}S_{0,\boldsymbol{c}} - {S_{1,\boldsymbol{c}}}^2} \times \\
&\quad \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\}\hat{F}_T(y) \\
&= \frac{1}{n|\boldsymbol{H}| f_{\boldsymbol{X}}(\boldsymbol{x})} \sum_{i=1}^{n} {\mathbf{K}_c}^*\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\}\hat{F}_T(y).
\end{aligned}$$

where,

$$ {\boldsymbol{K}_c}^*(u) = \frac{S_{2,\boldsymbol{c}} - S_{1,\boldsymbol{c}}\boldsymbol{u}}{S_{2,\boldsymbol{c}}S_{0,\boldsymbol{c}} - S_{1,\boldsymbol{c}}^2} \boldsymbol{K}(\boldsymbol{u})\boldsymbol{I}_{\{-\boldsymbol{c},+\infty\}}(\boldsymbol{u}). $$

The case of estimation in the interior is obtained by taking $(1,\ldots,1)_{d\times 1} < \boldsymbol{c} < \mathbf{M}\boldsymbol{H}^{-1} - (1,\ldots,1)_{d\times 1}$ and assumption A1 in Section 3.1. The asymptotic properties of $\hat{F}_L(y|\boldsymbol{x})$ are summarized in the next theorem.

Theorem 4.1.2. *Under assumptions A1 and A5 in Section 3.1, the bias and variance of $\tilde{F}_L(y|\boldsymbol{x})$ are given by*

$$\begin{aligned}
b_{\tilde{F}_{L(,c)}}(y|\boldsymbol{x}) &= \frac{\mu_2({\boldsymbol{K}_c}^*)}{2f_{\boldsymbol{X}}(\boldsymbol{x})}\Big[2tr\big\{\boldsymbol{H}^\top \mathcal{D}_{f_{\boldsymbol{X}}}(\boldsymbol{x})\dot{F}_T(y|\boldsymbol{x})^\top \boldsymbol{H}\big\} + f_{\boldsymbol{X}}(\boldsymbol{x})tr\big\{\boldsymbol{H}^\top \mathcal{H}_{F_T}(y|\boldsymbol{x})\boldsymbol{H}\big\} \\
&\quad + tr\big\{\boldsymbol{H}^\top \mathcal{H}_{f_{\boldsymbol{X}}}(\boldsymbol{x})\boldsymbol{H}\big\}F_T(y|\boldsymbol{x})\Big] + \mathrm{o}(\boldsymbol{H}^\top \boldsymbol{H}). \\
\sigma^2{}_{\tilde{F}_{L(,c)}}(y|\boldsymbol{x}) &= \frac{R({\boldsymbol{K}_c}^*)}{n|\boldsymbol{H}| f_{\boldsymbol{X}}(\boldsymbol{x})}\big\{1 - F_T(y|\boldsymbol{x})\big\}^2 \sum_{i=1}^{n} \frac{d_i}{n_i(n_i - d_i)} + \mathrm{O}(n^{-1}).
\end{aligned}$$

The above results imply that $\tilde{F}_L(y|\boldsymbol{x})$ achieves the same rate of convergence in the boundary of the covariates domain and in the interior and that the derivative order leaves the bias rate of convergence unaffected. However, the second term on the right hand side of the variance expression is negative which implies that kernel smoothing improves the estimate variance by a

second order effect. Although, the presence of the survival function of the censoring distribution in the denominator of the variance expression indicates that $\tilde{F}_L$ are expected to be more variable in practice than their complete sample counterparts. Now, let $N_T \leq M_Y$ denote the largest uncensored observation. For $y \in [0, N_T]$ it is known from Karunamuni and Yang, see Karunamuni and Yang (1991), that $\hat{S}_T$ converges to S_T in probability with rate $n^{-1/2}$. As a consequence, $\tilde{F}_L$ is expected to exhibit a very robust behavior in $[0, N_T]$. On the contrary and in correspondence to $\hat{S}_T$, the finite sample performance of $\tilde{F}_L$ is expected to diminish for $y > N_T$, i.e. beyond the last uncensored observation.

Proof. For the purpose of the proof we use the $P1$- $P6$, $P8$, $P10$-$P13$ and $P17$ properties from the appendix. For fixed i in the last step below

$$\begin{aligned}
\mathrm{E}\{\tilde{F}_L(y|\boldsymbol{x})\} &= \mathrm{E}\left[\frac{1}{n|\boldsymbol{H}|f_{\boldsymbol{X}}(\boldsymbol{x})}\sum_{i=1}^{n}\boldsymbol{K}_c{}^*\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i-\boldsymbol{x})\}\hat{F}_T(y)\right] \\
&\overset{(P1)}{=} \frac{1}{n|\boldsymbol{H}|f_{\boldsymbol{X}}(\boldsymbol{x})}\mathrm{E}\left[\sum_{i=1}^{n}\boldsymbol{K}_c{}^*\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i-\boldsymbol{x})\}\hat{F}_T(y)\right] \\
&\overset{(P2)}{=} \frac{1}{n|\boldsymbol{H}|f_{\boldsymbol{X}}(\boldsymbol{x})}\mathrm{E}\left[\sum_{i=1}^{n}\boldsymbol{K}_c{}^*\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i-\boldsymbol{x})\}\mathrm{E}\{\hat{F}_T(y)|\boldsymbol{X}_i\}\right] \\
&\overset{(P1),(P9)}{=} \frac{n}{n|\boldsymbol{H}|f_{\boldsymbol{X}}(\boldsymbol{x})}\mathrm{E}\left[\boldsymbol{K}_c{}^*\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i-\boldsymbol{x})\}\mathrm{E}\{\hat{F}_T(y)|\boldsymbol{X}_i\}\right] \\
&\overset{(P3)}{=} \frac{1}{|\boldsymbol{H}|f_{\boldsymbol{X}}(\boldsymbol{x})}\mathrm{E}\left[\boldsymbol{K}_c{}^*\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i-\boldsymbol{x})\}F_T(y|\boldsymbol{X}_i)\right] \\
&\overset{(P4)}{=} \frac{1}{|\boldsymbol{H}|f_{\boldsymbol{X}}(\boldsymbol{x})}\int \boldsymbol{K}_c{}^*\{\boldsymbol{H}^{-1}(\boldsymbol{u}-\boldsymbol{x})\}f_{\boldsymbol{X}}(\boldsymbol{u})F_T(y|\boldsymbol{u})d\boldsymbol{u}.
\end{aligned}$$

Using the change of variable $\boldsymbol{u}-\boldsymbol{x} = \boldsymbol{H}\boldsymbol{z}$ so that $d\boldsymbol{u} = \boldsymbol{H}d\boldsymbol{z}$ gives

$$\begin{aligned}
\mathrm{E}\{\tilde{F}_L(y|\boldsymbol{x})\} &= \frac{1}{|\boldsymbol{H}|f_{\boldsymbol{X}}(\boldsymbol{x})}\int \boldsymbol{K}_c{}^*\{\boldsymbol{H}^{-1}(\boldsymbol{H}\boldsymbol{z})\}f_{\boldsymbol{X}}(\boldsymbol{x}+\boldsymbol{H}\boldsymbol{z})F_T(y|\boldsymbol{x}+\boldsymbol{H}\boldsymbol{z})\boldsymbol{H}d\boldsymbol{z} \\
&\overset{(P10)}{=} \frac{|\boldsymbol{H}|}{|\boldsymbol{H}|f_{\boldsymbol{X}}(\boldsymbol{x})}\int \boldsymbol{K}_c{}^*(\boldsymbol{z})f_{\boldsymbol{X}}(\boldsymbol{x}+\boldsymbol{H}\boldsymbol{z})F_T(y|\boldsymbol{x}+\boldsymbol{H}\boldsymbol{z})d\boldsymbol{z} \\
&= \frac{1}{f_{\boldsymbol{X}}(\boldsymbol{x})}\int \boldsymbol{K}_c{}^*(\boldsymbol{z})f_{\boldsymbol{X}}(\boldsymbol{x}+\boldsymbol{H}\boldsymbol{z})F_T(y|\boldsymbol{x}+\boldsymbol{H}\boldsymbol{z})d\boldsymbol{z}. \qquad (4.5)
\end{aligned}$$

Now, we want to replace the unknown $f_{\boldsymbol{X}}(\boldsymbol{x}+\boldsymbol{H}\boldsymbol{z})$ and $F_T(y|\boldsymbol{x}+\boldsymbol{H}\boldsymbol{z})$ in (4.5). In order to accomplish that we will use the multivariate version of Taylor's

theorem. Applying (2.5) with $f(x) = f_{\boldsymbol{X}}(\boldsymbol{x})$, $\alpha_n = \boldsymbol{Hz}$ gives

$$\begin{aligned} f(\boldsymbol{x}+\boldsymbol{Hz}) &= f_{\boldsymbol{X}}(\boldsymbol{x}) + (\boldsymbol{Hz})^\top \mathcal{D}_{f_{\boldsymbol{X}}}(\boldsymbol{x}) + \frac{1}{2}(\boldsymbol{Hz})^\top \mathcal{H}_{f_{\boldsymbol{X}}}(\boldsymbol{x})\boldsymbol{Hz} + \mathrm{o}\big\{(\boldsymbol{Hz})^\top \boldsymbol{Hz}\big\} \\ &= f_{\boldsymbol{X}}(\boldsymbol{x}) + (\boldsymbol{Hz})^\top \mathcal{D}_{f_{\boldsymbol{X}}}(\boldsymbol{x}) + \frac{1}{2}(\boldsymbol{Hz})^\top \mathcal{H}_{f_{\boldsymbol{X}}}(\boldsymbol{x})\boldsymbol{Hz} + \mathrm{o}\big\{\mathrm{tr}(\boldsymbol{H}^2)\big\}. \end{aligned} \tag{4.6}$$

Note that,

$$\begin{aligned} \mathrm{o}\big\{(\boldsymbol{Hz})^\top \boldsymbol{Hz}\big\} = \mathrm{o}(\boldsymbol{z}^\top \boldsymbol{H}^\top \boldsymbol{Hz}) &\overset{(P5)}{=} \mathrm{o}\big\{\mathrm{tr}(\boldsymbol{z}^\top \boldsymbol{H}^\top \boldsymbol{Hz})\big\} \overset{(P8)}{=} \mathrm{o}\big\{\mathrm{tr}(\boldsymbol{zz}^\top \boldsymbol{H}^\top \boldsymbol{H})\big\} \\ &\overset{(P6)}{=} \mathrm{o}\big\{\boldsymbol{zz}^\top \mathrm{tr}(\boldsymbol{H}^\top \boldsymbol{H})\big\} \overset{(A3)}{=} \mathrm{o}\big\{\mathrm{tr}(\boldsymbol{H}^2)\big\}. \end{aligned}$$

Further, (2.5) with $f(x) = F_T(y|\boldsymbol{x}), \alpha_n = \boldsymbol{Hz}$ gives

$$\begin{aligned} F_T(y|\boldsymbol{x}+\boldsymbol{Hz}) &= F_T(y|\boldsymbol{x}) + (\boldsymbol{Hz})^\top \mathcal{D}_{F_T}(y|\boldsymbol{x}) + \frac{1}{2}(\boldsymbol{Hz})^\top \mathcal{H}_{F_T}(y|\boldsymbol{x})\boldsymbol{Hz} \\ &\quad + \mathrm{o}\big\{(\boldsymbol{Hz})^\top \boldsymbol{Hz}\big\} \\ &\overset{(17)}{=} F_T(y|\boldsymbol{x}) + (\boldsymbol{Hz})^\top \mathcal{D}_{F_T}(y|\boldsymbol{x}) + \frac{1}{2}(\boldsymbol{Hz})^\top \mathcal{H}_{F_T}(y|\boldsymbol{x})\boldsymbol{Hz} \\ &\quad + \mathrm{o}\big\{\mathrm{tr}(\boldsymbol{H}^2)\big\}. \end{aligned} \tag{4.7}$$

Now, by multiplying (4.6) with (4.7) we have that

$$\begin{aligned} &f(\boldsymbol{x}+\boldsymbol{Hz})F_T(y|\boldsymbol{x}+\boldsymbol{Hz}) = f_{\boldsymbol{X}}(\boldsymbol{x})F_T(y|\boldsymbol{x}) + f_{\boldsymbol{X}}(\boldsymbol{x})(\boldsymbol{Hz})^\top \mathcal{D}_{F_T}(y|\boldsymbol{x}) \\ &+ \frac{1}{2} f_{\boldsymbol{X}}(\boldsymbol{x})(\boldsymbol{Hz})^\top \mathcal{H}_{F_T}(y|\boldsymbol{x})\boldsymbol{Hz} + f_{\boldsymbol{X}}(\boldsymbol{x})\mathrm{o}\big\{\mathrm{tr}(\boldsymbol{H}^2)\big\} + (\boldsymbol{Hz})^\top \mathcal{D}_{f_{\boldsymbol{X}}}(\boldsymbol{x})F_T(y|\boldsymbol{x}) \\ &+ (\boldsymbol{Hz})^\top \mathcal{D}_{f_{\boldsymbol{X}}}(\boldsymbol{x})(\boldsymbol{Hz})^\top \mathcal{D}_{F_T}(y|\boldsymbol{x}) + \frac{1}{2}(\boldsymbol{Hz})^\top \mathcal{D}_{f_{\boldsymbol{X}}}(\boldsymbol{x})(\boldsymbol{Hz})^\top \mathcal{H}_{F_T}(y|\boldsymbol{x})\boldsymbol{Hz} \\ &+ (\boldsymbol{Hz})^\top \mathcal{D}_{f_{\boldsymbol{X}}}(\boldsymbol{x})\mathrm{o}\big\{\mathrm{tr}(\boldsymbol{H}^2)\big\} + \frac{1}{2}(\boldsymbol{Hz})^\top \mathcal{H}_{f_{\boldsymbol{X}}}(\boldsymbol{x})\boldsymbol{Hz}F_T(y|\boldsymbol{x}) \\ &+ \frac{1}{2}(\boldsymbol{Hz})^\top \mathcal{H}_{f_{\boldsymbol{X}}}(\boldsymbol{x})\boldsymbol{Hz}(\boldsymbol{Hz})^\top \mathcal{D}_{F_T}(y|\boldsymbol{x}) + \frac{1}{2}(\boldsymbol{Hz})^\top \mathcal{H}_{f_{\boldsymbol{X}}}(\boldsymbol{x})(\boldsymbol{Hz})\mathrm{o}\big\{\mathrm{tr}(\boldsymbol{H}^2)\big\} \\ &+ \frac{1}{4}(\boldsymbol{Hz})^\top \mathcal{H}_{f_{\boldsymbol{X}}}(\boldsymbol{x})\boldsymbol{Hz}(\boldsymbol{Hz})^\top \mathcal{H}_{F_T}(y|\boldsymbol{x})\boldsymbol{Hz} + \mathrm{o}\big\{\mathrm{tr}(\boldsymbol{H}^2)\big\}\mathrm{o}\big\{\mathrm{tr}(\boldsymbol{H}^2)\big\} \\ &+ \mathrm{o}\big\{\mathrm{tr}(\boldsymbol{H}^2)\big\}\frac{1}{2}(\boldsymbol{Hz})^\top \mathcal{H}_{F_T}(y|\boldsymbol{x})\boldsymbol{Hz} + \mathrm{o}\big\{\mathrm{tr}(\boldsymbol{H}^2)\big\}(\boldsymbol{Hz})^\top \mathcal{D}_{F_T}(y|\boldsymbol{x}) \\ &+ \mathrm{o}\big\{\mathrm{tr}(\boldsymbol{H}^2)\big\}F_T(y|\boldsymbol{x}). \end{aligned} \tag{4.8}$$

Combining (4.8) with the following properties from $A5$ in Section 3.1

$$\int \boldsymbol{z}\boldsymbol{K}_c{}^*(\boldsymbol{z})d\boldsymbol{z} = 0 \quad \textit{and} \quad \int u_1^{l_1}...u_d^{l_d}\boldsymbol{K}(\boldsymbol{u}) = 0,$$

and by including all lower order terms in $\mathrm{o}\big\{\mathrm{tr}(\boldsymbol{H}^2)\big\}$ yields

$$f(\boldsymbol{x}+\boldsymbol{Hz})F_T(y|\boldsymbol{x}+\boldsymbol{Hz}) = f_{\boldsymbol{X}}(\boldsymbol{x})F_T(y|\boldsymbol{x}) + (\boldsymbol{Hz})^\top \mathcal{D}_{f_{\boldsymbol{X}}}(\boldsymbol{x})(\boldsymbol{Hz})^\top \mathcal{D}_{F_T}(y|\boldsymbol{x}) + \frac{1}{2}(\boldsymbol{Hz})^\top \mathcal{H}_{f_{\boldsymbol{X}}}(\boldsymbol{x})\boldsymbol{Hz}F_T(y|\boldsymbol{x}) + \frac{1}{2}f_{\boldsymbol{X}}(\boldsymbol{x})(\boldsymbol{Hz})^\top \mathcal{H}_{F_T}(y|\boldsymbol{x})\boldsymbol{Hz} + \mathrm{o}(\boldsymbol{H}^\top\boldsymbol{H}).$$

Substitute the above equation in (4.5) to obtain

$$\begin{aligned}\mathrm{E}\{\tilde{F}_L(y|\boldsymbol{x})\} = \frac{1}{f_{\boldsymbol{X}}(\boldsymbol{x})}\int \boldsymbol{K}_c{}^*(\boldsymbol{z})\Big\{ & f_{\boldsymbol{X}}(\boldsymbol{x})F_T(y|\boldsymbol{x}) + (\boldsymbol{Hz})^\top \mathcal{D}_{f_{\boldsymbol{X}}}(\boldsymbol{x})(\boldsymbol{Hz})^\top \mathcal{D}_{F_T}(y|\boldsymbol{x}) \\ & + \frac{1}{2}(\boldsymbol{Hz})^\top \mathcal{H}_{f_{\boldsymbol{X}}}(\boldsymbol{x})\boldsymbol{Hz}F_T(y|\boldsymbol{x}) + \frac{1}{2}f_{\boldsymbol{X}}(\boldsymbol{x})(\boldsymbol{Hz})^\top \mathcal{H}_{F_T}(y|\boldsymbol{x})\boldsymbol{Hz} \\ & + \mathrm{o}(\boldsymbol{H}^\top\boldsymbol{H})\Big\}d\boldsymbol{z}. \end{aligned} \tag{4.9}$$

Rearranging (4.8), (4.9) gives

$$\begin{aligned}\mathrm{E}\{\tilde{F}_L(y|\boldsymbol{x})\} &= \int \boldsymbol{K}_c{}^*(\boldsymbol{z})\Big\{F_T(y|\boldsymbol{x}) + \frac{1}{f_{\boldsymbol{X}}(\boldsymbol{x})}(\boldsymbol{Hz})^\top \mathcal{D}_{f_{\boldsymbol{X}}}(\boldsymbol{x})(\boldsymbol{Hz})^\top \mathcal{D}_{F_T}(y|\boldsymbol{x}) \\ &\quad + \frac{1}{2}(\boldsymbol{Hz})^\top \mathcal{H}_{F_T}(y|\boldsymbol{x})\boldsymbol{Hz} + \mathrm{o}(\boldsymbol{H}^\top\boldsymbol{H})\Big\}d\boldsymbol{z} \\ &= \int \boldsymbol{K}_c{}^*(\boldsymbol{z})\Big\{\frac{1}{f_{\boldsymbol{X}}(\boldsymbol{x})}(\boldsymbol{Hz})^\top \mathcal{D}_{f_{\boldsymbol{X}}}(\boldsymbol{x})(\boldsymbol{Hz})^\top \mathcal{D}_{F_T}(y|\boldsymbol{x}) \\ &\quad + \int \boldsymbol{K}_c{}^*(\boldsymbol{z})\Big\{\frac{1}{2}\frac{1}{f_{\boldsymbol{X}}(\boldsymbol{x})}(\boldsymbol{Hz})^\top \mathcal{H}_{f_{\boldsymbol{X}}}(\boldsymbol{x})\boldsymbol{Hz}F_T(y|\boldsymbol{x})\Big\}d\boldsymbol{z} \\ &\quad + \int \boldsymbol{K}_c{}^*(\boldsymbol{z})\Big\{\frac{1}{2}(\boldsymbol{Hz})^T \mathcal{H}_{F_T}(y|\boldsymbol{x})\boldsymbol{Hz}\Big\}d\boldsymbol{z} \\ &\quad + \int \boldsymbol{K}_c{}^*(\boldsymbol{z})F_T(y|\boldsymbol{x})d\boldsymbol{z} + \int \boldsymbol{K}_c{}^*(\boldsymbol{z})\mathrm{o}(\boldsymbol{H}^\top\boldsymbol{H})d\boldsymbol{z}.\end{aligned}$$

By the following conditions from $A5$ in Section 3.1

$$\int \boldsymbol{u}\boldsymbol{u}^\top \boldsymbol{K}(\boldsymbol{u}) = \mu_2(\boldsymbol{K})\boldsymbol{I} \quad \textit{and} \quad \int \boldsymbol{K}_c{}^*(\boldsymbol{z})d\boldsymbol{z} = 1,$$

the above equation becomes

$$\begin{aligned}
\mathrm{E}\{\tilde{F}_L(y|\boldsymbol{x})\} &= \frac{\mu_2(\boldsymbol{K}_c{}^*)}{2}\left[\frac{2}{f_{\boldsymbol{X}}(\boldsymbol{x})}tr\{\boldsymbol{H}^\top\mathcal{D}_{f_{\boldsymbol{X}}}(\boldsymbol{x})\mathcal{D}_{F_T}(y|\boldsymbol{x})^\top\boldsymbol{H}\}+tr\{\boldsymbol{H}^\top\mathcal{H}_{F_T}(y|\boldsymbol{x})\boldsymbol{H}\}\right] \\
&\quad+\left[1+\frac{tr\{\boldsymbol{H}^\top\mathcal{H}_{f_{\boldsymbol{X}}}(\boldsymbol{x})\boldsymbol{H}\}\mu_2(\boldsymbol{K}_c{}^*)}{2f_{\boldsymbol{X}}(\boldsymbol{x})}\right]F_T(y|\boldsymbol{x})+\mathrm{o}(\boldsymbol{H}^\top\boldsymbol{H}) \\
&= \frac{\mu_2(\boldsymbol{K}_c{}^*)}{2}\left[\frac{2}{f_{\boldsymbol{X}}(\boldsymbol{x})}tr\{\boldsymbol{H}^\top\mathcal{D}_{f_{\boldsymbol{X}}}(\boldsymbol{x})\dot{F}_T(y|\boldsymbol{x})^\top\boldsymbol{H}\}+tr\{\boldsymbol{H}^\top\mathcal{H}_{F_T}(y|\boldsymbol{x})\boldsymbol{H}\}\right] \\
&\quad+\left[1+\frac{tr\{\boldsymbol{H}^\top\mathcal{H}_{f_{\boldsymbol{X}}}(\boldsymbol{x})\boldsymbol{H}\}\mu_2(\boldsymbol{K}_c{}^*)}{2f_{\boldsymbol{X}}(\boldsymbol{x})}\right]F_T(y|\boldsymbol{x})+\mathrm{o}(\boldsymbol{H}^\top\boldsymbol{H}).
\end{aligned} \tag{4.10}$$

From definition of bias and (4.10) we have that

$$\begin{aligned}
b_{\tilde{F}_{L(,c)}}(y|\boldsymbol{x}) &= \mathrm{E}\{\tilde{F}_L(y|\boldsymbol{x})\}-F_T(y|\boldsymbol{x}) \\
&= \frac{\mu_2(\boldsymbol{K}_c{}^*)}{2}\left[\frac{2}{f_{\boldsymbol{X}}(\boldsymbol{x})}tr\{\boldsymbol{H}^\top\mathcal{D}_{f_{\boldsymbol{X}}}(\boldsymbol{x})\dot{F}_T(y|\boldsymbol{x})^\top\boldsymbol{H}\}+tr\{\boldsymbol{H}^\top\mathcal{H}_{F_T}(y|\boldsymbol{x})\boldsymbol{H}\}\right] \\
&\quad+\left[\frac{tr\{\boldsymbol{H}^\top\mathcal{H}_{f_{\boldsymbol{X}}}(\boldsymbol{x})\boldsymbol{H}\}\mu_2(\boldsymbol{K}_c{}^*)}{2f_{\boldsymbol{X}}(\boldsymbol{x})}\right]F_T(y|\boldsymbol{x})+\mathrm{o}(\boldsymbol{H}^\top\boldsymbol{H}) \\
&= \frac{\mu_2(\boldsymbol{K}_c{}^*)}{2f_{\boldsymbol{X}}(\boldsymbol{x})}\Big[2tr\{\boldsymbol{H}^\top\mathcal{D}_{f_{\boldsymbol{X}}}(\boldsymbol{x})\dot{F}_T(y|\boldsymbol{x})^\top\boldsymbol{H}\}+f_{\boldsymbol{X}}(\boldsymbol{x})tr\{\boldsymbol{H}^\top\mathcal{H}_{F_T}(y|\boldsymbol{x})\boldsymbol{H}\} \\
&\quad+tr\{\boldsymbol{H}^\top\mathcal{H}_{f_{\boldsymbol{X}}}(\boldsymbol{x})\boldsymbol{H}\}F_T(y|\boldsymbol{x})\Big]+\mathrm{o}(\boldsymbol{H}^\top\boldsymbol{H}).
\end{aligned}$$

which complete the first part of the proof.
In order to prove the variance of $\tilde{F}_L(y|\boldsymbol{x})$ we will use the following bias/variance decomposition

$$\mathrm{Var}\Big\{\tilde{F}_L(y|\boldsymbol{x})\Big\}=\mathrm{Var}\Big[\mathrm{E}\big\{\tilde{F}_L(y|\boldsymbol{x})|\boldsymbol{X}\big\}\Big]+\mathrm{E}\Big[\mathrm{Var}\big\{\tilde{F}_L(y|\boldsymbol{x})|\boldsymbol{X}\big\}\Big]. \tag{4.11}$$

For the first term of the variance above

$$\begin{aligned}
\mathrm{Var}\Big\{\tilde{F}_L(y|\boldsymbol{x})|\boldsymbol{X}\Big\} &= \mathrm{Var}\left[\frac{1}{n|\boldsymbol{H}|f_{\boldsymbol{X}}(\boldsymbol{x})}\sum_{i=1}^{n}\boldsymbol{K}_c{}^*\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i-\boldsymbol{x})\}\hat{F}_T(y)|\boldsymbol{X}_i\right] \\
&\overset{(P12),(P13)}{=} \frac{n}{n^2|\boldsymbol{H}|^2f_{\boldsymbol{X}}(\boldsymbol{x})^2}\boldsymbol{K}_c{}^*\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i-\boldsymbol{x})\}^2\mathrm{Var}\{\hat{F}_T(y)|\boldsymbol{X}_i\} \\
&= \frac{1}{n|\boldsymbol{H}|^2f_{\boldsymbol{X}}(\boldsymbol{x})^2}\boldsymbol{K}_c{}^*\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i-\boldsymbol{x})\}^2\mathrm{Var}\{\hat{F}_T(y)|\boldsymbol{X}_i\}.
\end{aligned} \tag{4.12}$$

Thus, fixing i in the first step below

$$\begin{aligned}\mathrm{E}\Big[\mathrm{Var}\big\{\tilde{F}_L(y|\boldsymbol{x})|\boldsymbol{X}\big\}\Big] &= \mathrm{E}\Bigg[\frac{1}{n|\boldsymbol{H}|^2 f_{\boldsymbol{X}}(\boldsymbol{x})^2}\boldsymbol{K}_c{}^*\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i-\boldsymbol{x})\}^2\mathrm{Var}\{\hat{F}_T(y)|\boldsymbol{X}_i\}\Bigg]\\ &\overset{(P4)}{=} \frac{1}{n|\boldsymbol{H}|^2 f_{\boldsymbol{X}}(\boldsymbol{x})^2}\int \boldsymbol{K}_c{}^*\{\boldsymbol{H}^{-1}(\boldsymbol{u}-\boldsymbol{x})\}^2 f_{\boldsymbol{X}}(\boldsymbol{u})\\ &\qquad \mathrm{Var}\{\hat{F}_T(y)|\boldsymbol{u}\}d\boldsymbol{u}. \qquad (4.13)\end{aligned}$$

Using the change of variable $\boldsymbol{u}-\boldsymbol{x}=\boldsymbol{H}\boldsymbol{z}$ so that $d\boldsymbol{u}=\boldsymbol{H}d\boldsymbol{z}$ (4.13) gives

$$\begin{aligned}\mathrm{E}\Big[\mathrm{Var}\big\{\tilde{F}_L(y|\boldsymbol{x})|\boldsymbol{X}\big\}\Big] &= \frac{1}{n|\boldsymbol{H}|^2 f_{\boldsymbol{X}}(\boldsymbol{x})^2}\int \boldsymbol{K}_c{}^*\{\boldsymbol{z}\}^2 f_{\boldsymbol{X}}(\boldsymbol{x}+\boldsymbol{H}\boldsymbol{z})\times\\ &\qquad \mathrm{Var}\{\hat{F}_T(y|\boldsymbol{x}+\boldsymbol{(Hz)}\}\boldsymbol{H}\,d\boldsymbol{z}\\ &\overset{(P10)}{=} \frac{|\boldsymbol{H}|}{n|\boldsymbol{H}|^2 f_{\boldsymbol{X}}(\boldsymbol{x})^2}\int \boldsymbol{K}_c{}^*\{\boldsymbol{z}\}^2 f_{\boldsymbol{X}}(\boldsymbol{x}+\boldsymbol{H}\boldsymbol{z})\times\\ &\qquad \mathrm{Var}\{\hat{F}_T(y|\boldsymbol{x}+\boldsymbol{H}\boldsymbol{z}\}d\boldsymbol{z}\\ &= \frac{1}{n|\boldsymbol{H}|f_{\boldsymbol{X}}(\boldsymbol{x})^2}\int \boldsymbol{K}_c{}^*\{\boldsymbol{z}\}^2 f_{\boldsymbol{X}}(\boldsymbol{x}+\boldsymbol{H}\boldsymbol{z})\times\\ &\qquad \mathrm{Var}\{\hat{F}_T(y|\boldsymbol{x}+\boldsymbol{H}\boldsymbol{z}\}d\boldsymbol{z}. \qquad (4.14)\end{aligned}$$

By substituting the conditional form of Kaplan-Meier's estimator variance (1.17) in (4.14) we have that

$$\begin{aligned}\mathrm{E}\Big[\mathrm{Var}\big\{\tilde{F}_L(y|\boldsymbol{x})|\boldsymbol{X}\big\}\Big] &= \frac{1}{n|\boldsymbol{H}|f_{\boldsymbol{X}}(\boldsymbol{x})^2}\int \boldsymbol{K}_c{}^*\{\boldsymbol{z}\}^2 f_{\boldsymbol{X}}(\boldsymbol{x}+\boldsymbol{H}\boldsymbol{z})\times\\ &\qquad \big\{1-\hat{F}_T(y|\boldsymbol{x}+\boldsymbol{H}\boldsymbol{z})\big\}^2\times\\ &\qquad \sum_{i=1}^{n}\frac{d_i}{n_i(n_i-d_i)}d\boldsymbol{z}.\end{aligned}$$

From (4.6) and (4.7) we have that

$$f(\boldsymbol{x}+\boldsymbol{H}\boldsymbol{z}) = f_{\boldsymbol{X}}(\boldsymbol{x}) + (\boldsymbol{H}\boldsymbol{z})^\top\mathcal{D}_{f_{\boldsymbol{X}}}(\boldsymbol{x}) + \frac{1}{2}(\boldsymbol{H}\boldsymbol{z})^\top\mathcal{H}_{f_{\boldsymbol{X}}}(\boldsymbol{x})\boldsymbol{H}\boldsymbol{z} + \mathrm{o}\big\{\mathrm{tr}(\boldsymbol{H}^2)\big\}.$$

Similarly,

$$F_T(y|\boldsymbol{x}+\boldsymbol{Hz}) = F_T(y|\boldsymbol{x}) + (\boldsymbol{Hz})^\top \mathcal{D}_{F_T}(y|\boldsymbol{x}) + \frac{1}{2}(\boldsymbol{Hz})^\top \mathcal{H}_{F_T}(y|\boldsymbol{x})\boldsymbol{Hz} + \mathrm{o}\big\{\mathrm{tr}(\boldsymbol{H}^2)\big\}.$$

Now, note that

$$\begin{aligned}
F_T(y|\boldsymbol{x}+\boldsymbol{Hz})^2 =& \Big[F_T(y|\boldsymbol{x}) + (\boldsymbol{Hz})^\top \mathcal{D}_{F_T}(y|\boldsymbol{x}) + \frac{1}{2}(\boldsymbol{Hz})^\top \mathcal{H}_{F_T}(y|\boldsymbol{x})\boldsymbol{Hz} \\
&+ \mathrm{o}\big\{\mathrm{tr}(\boldsymbol{H}^2)\big\}\Big]\Big[F_T(y|\boldsymbol{x}) + (\boldsymbol{Hz})^\top \mathcal{D}_{F_T}(y|\boldsymbol{x}) + \frac{1}{2}(\boldsymbol{Hz})^\top \times \\
&\mathcal{H}_{F_T}(y|\boldsymbol{x})\boldsymbol{Hz} + \mathrm{o}\big\{\mathrm{tr}(\boldsymbol{H}^2)\big\}\Big] \\
=& F_T(y|\boldsymbol{x})^2 + F_T(y|\boldsymbol{x})(\boldsymbol{Hz})^\top \mathcal{D}_{F_T}(y|\boldsymbol{x}) + \frac{1}{2}F_T(y|\boldsymbol{x})(\boldsymbol{Hz})^\top \times \\
&\mathcal{H}_{F_T}(y|\boldsymbol{x})\boldsymbol{Hz} + \big\{(\boldsymbol{Hz})^\top \mathcal{D}_{F_T}(y|\boldsymbol{x})\big\}^2 \\
&+ (\boldsymbol{Hz})^\top \mathcal{D}_{F_T}(y|\boldsymbol{x})F_T(y|\boldsymbol{x}) + (\boldsymbol{Hz})^\top \mathcal{D}_{F_T}(y|\boldsymbol{x})\frac{1}{2}(\boldsymbol{Hz})^\top \times \\
&\mathcal{H}_{F_T}(y|\boldsymbol{x})\boldsymbol{Hz} + \big\{\frac{1}{2}(\boldsymbol{Hz})^\top \mathcal{H}_{F_T}(y|\boldsymbol{x})\boldsymbol{Hz}\big\}^2 + \frac{1}{2}(\boldsymbol{Hz})^\top \times \\
&\mathcal{H}_{F_T}(y|\boldsymbol{x})\boldsymbol{Hz}F_T(y|\boldsymbol{x}) + \frac{1}{2}(\boldsymbol{Hz})^\top \mathcal{H}_{F_T}(y|\boldsymbol{x})\boldsymbol{Hz}(\boldsymbol{Hz})^\top \times \\
&\mathcal{D}_{F_T}(y|\boldsymbol{x}).
\end{aligned}$$

Then,

$$\begin{aligned}
\big\{1 - F_T(y|\boldsymbol{x}+\boldsymbol{Hz})\big\}^2 &= 1 - 2F_T(y|\boldsymbol{x}+\boldsymbol{Hz}) + F_T(y|\boldsymbol{x}+\boldsymbol{Hz})^2 \\
&= \big\{1 - F_T(y|\boldsymbol{x})\big\}^2 - 2(\boldsymbol{Hz})^\top \mathcal{D}_{F_T}(y|\boldsymbol{x}) \\
&\quad - (\boldsymbol{Hz})^\top \mathcal{H}_{F_T}(y|\boldsymbol{x})\boldsymbol{Hz} + F_T(y|\boldsymbol{x})(\boldsymbol{Hz})^\top \mathcal{D}_{F_T}(y|\boldsymbol{x}) \\
&\quad + \frac{1}{2}F_T(y|\boldsymbol{x})(\boldsymbol{Hz})^\top \mathcal{H}_{F_T}(y|\boldsymbol{x})\boldsymbol{Hz} + \big\{(\boldsymbol{Hz})^\top \mathcal{D}_{F_T}(y|\boldsymbol{x})\big\}^2 \\
&\quad + (\boldsymbol{Hz})^\top \mathcal{D}_{F_T}(y|\boldsymbol{x})F_T(y|\boldsymbol{x}) + \frac{1}{2}(\boldsymbol{Hz})^\top \mathcal{D}_{F_T}(y|\boldsymbol{x})(\boldsymbol{Hz})^\top \times \\
&\quad \mathcal{H}_{F_T}(y|\boldsymbol{x})\boldsymbol{Hz} + \big\{\frac{1}{2}(\boldsymbol{Hz})^\top \mathcal{H}_{F_T}(y|\boldsymbol{x})\boldsymbol{Hz}\big\}^2 + \frac{1}{2}(\boldsymbol{Hz})^\top \times \\
&\quad \mathcal{H}_{F_T}(y|\boldsymbol{x})\boldsymbol{Hz}F_T(y|\boldsymbol{x})\frac{1}{2}(\boldsymbol{Hz})^\top \mathcal{H}_{F_T}(y|\boldsymbol{x})\boldsymbol{Hz}(\boldsymbol{Hz})^\top \times \\
&\quad \mathcal{D}_{F_T}(y|\boldsymbol{x}).
\end{aligned}$$

By taking into consideration the above expansions and by ignoring the negligible parts, the first term of (4.11) is

$$\begin{aligned}\mathrm{E}\Big[\mathrm{Var}\big\{\tilde{F}_L(y|\boldsymbol{x})|\boldsymbol{X}\big\}\Big] &= \frac{1}{n|\boldsymbol{H}|f_{\boldsymbol{X}}(\boldsymbol{x})^2}\int {\boldsymbol{K}_c}^*\{\boldsymbol{z}\}^2 f_{\boldsymbol{X}}(\boldsymbol{x})\big\{1-F_T(y|\boldsymbol{x})\big\}^2\times \\ &\qquad\qquad \sum_{i=1}^{n}\frac{d_i}{n_i(n_i-d_i)}d\boldsymbol{z}+\mathrm{O}(n^{-1}) \\ &= \frac{f_{\boldsymbol{X}}(\boldsymbol{x})}{n|\boldsymbol{H}|f_{\boldsymbol{X}}(\boldsymbol{x})^2}\big\{1-F_T(y|\boldsymbol{x})\big\}^2\sum_{i=1}^{n}\frac{d_i}{n_i(n_i-d_i)}\times \\ &\quad \int {\boldsymbol{K}_c}^*\{\boldsymbol{z}\}^2\, d\boldsymbol{z}+\mathrm{O}(n^{-1}) \\ &= \frac{R({\boldsymbol{K}_c}^*)}{n|\boldsymbol{H}|f_{\boldsymbol{X}}(\boldsymbol{x})}\big\{1-F_T(y|\boldsymbol{x})\big\}^2\sum_{i=1}^{n}\frac{d_i}{n_i(n_i-d_i)}+\mathrm{O}(n^{-1}).\end{aligned} \tag{4.15}$$

As regards the other part we have

$$\begin{aligned}\mathrm{E}\big\{\tilde{F}_L(y|\boldsymbol{x})|\boldsymbol{X}\big\} &= \mathrm{E}\bigg[\frac{1}{n|\boldsymbol{H}|f_{\boldsymbol{X}}(\boldsymbol{x})}\sum_{i=1}^{n}{\boldsymbol{K}_c}^*\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i-\boldsymbol{x})\}\hat{F}_T(y)|\boldsymbol{X}_i\bigg] \\ &\overset{(P1),(P9)}{=} \frac{n}{n|\boldsymbol{H}|f_{\boldsymbol{X}}(\boldsymbol{x})}{\boldsymbol{K}_c}^*\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i-\boldsymbol{x})\}E(\hat{F}_T(y)|\boldsymbol{X}_i) \\ &\overset{(P3)}{=} \frac{1}{|\boldsymbol{H}|f_{\boldsymbol{X}}(\boldsymbol{x})}{\boldsymbol{K}_c}^*\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i-\boldsymbol{x})\}F_T(y|\boldsymbol{X}_i).\end{aligned}$$

Thus, fixing i in the third step below

$$\begin{aligned}\mathrm{Var}\Big[\mathrm{E}\big\{\tilde{F}_L(y|\boldsymbol{x})|\boldsymbol{X}\big\}\Big] &= \mathrm{Var}\bigg[\frac{1}{|\boldsymbol{H}|f_{\boldsymbol{X}}(\boldsymbol{x})}{\boldsymbol{K}_c}^*\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i-\boldsymbol{x})\}F_T(y|\boldsymbol{X}_i)\bigg] \\ &\overset{(P11)}{=} \mathrm{E}\bigg[\frac{1}{|\boldsymbol{H}|^2 f_{\boldsymbol{X}}(\boldsymbol{x})^2}{\boldsymbol{K}_c}^*\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i-\boldsymbol{x})\}^2 F_T(y|\boldsymbol{X}_i)^2\bigg] \\ &\quad -\mathrm{E}\bigg[\frac{1}{|\boldsymbol{H}|f_{\boldsymbol{X}}(\boldsymbol{x})}{\boldsymbol{K}_c}^*\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i-\boldsymbol{x})\}F_T(y|\boldsymbol{X}_i)\bigg]^2 \\ &\overset{(P4),(P1)}{=} \frac{1}{|\boldsymbol{H}|^2 f_{\boldsymbol{X}}(\boldsymbol{x})^2}\int {\boldsymbol{K}_c}^*\{\boldsymbol{H}^{-1}(\boldsymbol{u}-\boldsymbol{x})\}^2\times \\ &\quad f_{\boldsymbol{X}}(\boldsymbol{u})F_T(y|\boldsymbol{u})^2 d\boldsymbol{u}-\bigg[\frac{1}{|\boldsymbol{H}|f_{\boldsymbol{X}}(\boldsymbol{x})}\int {\boldsymbol{K}_c}^*\{\boldsymbol{H}^{-1}\times \\ &\quad (\boldsymbol{u}-\boldsymbol{x})\}f_{\boldsymbol{X}}(\boldsymbol{u})F_T(y|\boldsymbol{u})d\boldsymbol{u}\bigg]^2.\end{aligned}$$

Using the change of variable $\boldsymbol{u} - \boldsymbol{x} = \boldsymbol{H}\boldsymbol{z}$ so that $d\boldsymbol{u} = \boldsymbol{H}d\boldsymbol{z}$ we have

$$\begin{aligned}
\mathrm{Var}\Big[\mathrm{E}\big\{\tilde{F}_L(y|\boldsymbol{x})|\boldsymbol{X}\big\}\Big] &= \frac{1}{|\boldsymbol{H}|^2 f_{\boldsymbol{X}}(\boldsymbol{x})^2}\int \boldsymbol{K}_c{}^*\{\boldsymbol{z}\}^2 f_{\boldsymbol{X}}(\boldsymbol{x}+\boldsymbol{H}\boldsymbol{z})F_T(y|\boldsymbol{x}+\boldsymbol{H}\boldsymbol{z})^2\boldsymbol{H}d\boldsymbol{z} \\
&\quad - \frac{1}{|\boldsymbol{H}|^2 f_{\boldsymbol{X}}(\boldsymbol{x})^2}\bigg[\int \boldsymbol{K}_c{}^*\{\boldsymbol{z}\} f_{\boldsymbol{X}}(\boldsymbol{x}+\boldsymbol{H}\boldsymbol{z})F_T(y|\boldsymbol{x}+\boldsymbol{H}\boldsymbol{z})\boldsymbol{H}d\boldsymbol{z}\bigg]^2 \\
&= \frac{1}{|\boldsymbol{H}| f_{\boldsymbol{X}}(\boldsymbol{x})^2}\int \boldsymbol{K}_c{}^*\{\boldsymbol{z}\}^2 f_{\boldsymbol{X}}(\boldsymbol{x}+\boldsymbol{H}\boldsymbol{z})F_T(y|\boldsymbol{x}+\boldsymbol{H}\boldsymbol{z})^2 d\boldsymbol{z} \\
&\quad - \frac{1}{f_{\boldsymbol{X}}(\boldsymbol{x})^2}\bigg[\int \boldsymbol{K}_c{}^*\{\boldsymbol{z}\} f_{\boldsymbol{X}}(\boldsymbol{x}+\boldsymbol{H}\boldsymbol{z})F_T(y|\boldsymbol{x}+\boldsymbol{H}\boldsymbol{z})d\boldsymbol{z}\bigg]^2.
\end{aligned}$$

Now, by applying the Taylor expansions for $f_{\boldsymbol{X}}(\boldsymbol{x}+\boldsymbol{H}\boldsymbol{z})$, $F_T(y|\boldsymbol{x}+\boldsymbol{H}\boldsymbol{z})$, $F_T(y|\boldsymbol{x}+\boldsymbol{H}\boldsymbol{z})^2$ and by ignoring the negligible parts we get

$$\mathrm{Var}\Big[\mathrm{E}\big\{\tilde{F}_L(y|\boldsymbol{x})|\boldsymbol{X}\big\}\Big] = \mathrm{O}(\boldsymbol{H}^\top\boldsymbol{H}). \tag{4.16}$$

Combining (4.15) and (4.16) completes the proof of the theorem. □

CHAPTER 5

Bandwidth Selection

5.1 Univariate bandwidth selection

The practical implementation of the estimator in (3.10) requires the specification of the bandwidth $\boldsymbol{H}$. The problem of choosing the amount of smoothing to apply to the data is of crucial importance in kernel estimation and has become an important topic in recent years. It is still a burgeoning area of research, see Silverman (1986), Wand and Jones (1994), Scott (2015) and Chacón and Duong (2018).
The appropriate choice of smoothing parameter is influenced by the purpose for which the density estimate is to be used. If the purpose is to explore the data in order to suggest possible models, then it will be quite sufficient to choose the bandwidth subjectively by eye. This would involve looking at several density estimates over a range of bandwidths and selecting the density that is the "most pleasing" to the eye. One such strategy is to begin with a large bandwidth and to decrease the amount of smoothing until variations that are more "random" start to appear. This approach is more viable when the user has reasons to believe that there is certain structure in the data. However, there are also many applications that require to select the bandwidth automatically from the data. One reason is that it can be very time consuming to select the bandwidth by eye if there are many density estimates required for a given problem. Another is that, in many cases, the user has no prior knowledge about the structure of the data and would not have any feeling for which bandwidth gives an estimate closest to the true density. When kernel estimators are used for presenting conclusions in larger statistical procedures, automatic bandwidth selection is usually necessary. A method that uses the data to produce a bandwidth h is called a bandwidth selector. The bandwidth selection problem is present in all types of kernel estimation.

Currently available bandwidth selectors can be roughly divided into two classes. The first class consists of simple easily computable formulas which aim to find a bandwidth that is "reasonable" for a wide range of situations, but without out any mathematical guarantees of being close to the optimal bandwidth. We will call such bandwidth selectors quick and simple. Quick and simple bandwidth selectors are motivated by the need to have fast automatically generated kernel estimates for algorithms that require many curve estimation steps as well as providing a reasonable starting point for subjective choice of the smoothing parameter. The second type of bandwidth selector will be labelled as hi-tech since such selection procedures are based on more involved mathematical arguments and require considerably more computational effort, but aim to give a good answer for very general classes of underlying functions. Each of the hi-tech bandwidth selectors that we discuss can be motivated through aiming to minimise MISE and can be shown to attain this goal asymptotically to some extent. Such a bandwidth selector is said to be consistent with respect to MISE. To end with, some high-tech bandwidth selectors are as follows: least squares cross-validation, biased cross validation, plug-in bandwidth selectors etc.
It should be pointed out that there exist approaches to bandwidth selection based on other loss criteria. However, their analysis is more difficult. At the time of writing this thesis in the field of bandwidth selection, new selectors are developed and several unresolved issues. The performance of a kernel estimator requires the specification of appropriate error criteria for measuring the error when estimating the density at a single point as well as the error when estimating the density over the whole real line. In classical parametric statistics it is common to measure the closeness of an estimator to its target parameter by the size of the Mean Squared Error (MSE)

$$\begin{aligned}\mathrm{MSE}\Big\{\tilde{F}_L(y|\boldsymbol{x})\Big\} &= \mathrm{E}\Big\{\tilde{F}_L(y|\boldsymbol{x}) - F_T(y|\boldsymbol{x})\Big\}^2 \\ &= \mathrm{Var}\Big\{\tilde{F}_L(y|\boldsymbol{x})\Big\} + \mathrm{Bias}\Big\{\tilde{F}_L(y|\boldsymbol{x})\Big\}^2. \end{aligned} \tag{5.1}$$

Let us, take into consideration only one covariate. Then we have only one bandwidth $h_1 = h$ and not a bandwidth matrix $\boldsymbol{H}$. Thus, Theorem 4.1.2 gives

$$\begin{aligned} b_{\tilde{F}_{L(,c)}}(y|x) = \frac{h^2\mu_2({\boldsymbol{K}_c}^*)}{2f_{\boldsymbol{X}}(x)}\mathcal{H}_{f_{\boldsymbol{X}}}(x)F_T + \frac{h^2\mu_2({\boldsymbol{K}_c}^*)}{2}\times \\ \left\{2\frac{1}{f_{\boldsymbol{X}}(x)}\mathcal{D}_{f_{\boldsymbol{X}}}(x)\dot{F}_T(y|x) + \mathcal{H}_{F_T}(y|x)\right\} + \mathrm{o}(h^2). \end{aligned} \tag{5.2}$$

$$\sigma^2{}_{\tilde{F}_{L(,c)}}(y|x) = \frac{R(\boldsymbol{K}_c{}^*)}{nhf_{\boldsymbol{X}}(x)}\big\{1 - F_T(y|x)\big\}^2 \sum_{j=1}^{q} \frac{d_j}{n_j(n_j - d_j)} + \mathrm{O}(n^{-1}). \quad (5.3)$$

By substituting (5.2) and (5.3) in (5.1) we get

$$\begin{aligned}\mathrm{MSE}\Big\{\tilde{F}_L(y|x)\Big\} = \frac{R(\boldsymbol{K}_c{}^*)}{nhf_{\boldsymbol{X}}(x)}\big\{1 - F_T(y|x)\big\}^2 \sum_{j=1}^{q} \frac{d_j}{n_j(n_j - d_j)} \\ + \bigg[\bigg\{\frac{h^2\mu_2(\boldsymbol{K}_c{}^*)}{2f_{\boldsymbol{X}}(x)}\mathcal{H}_{f_{\boldsymbol{X}}}(x)\bigg\} F_T(y|x) \\ + \frac{h^2\mu_2(\boldsymbol{K}_c{}^*)}{2}\Big\{2\frac{1}{f_{\boldsymbol{X}}(x)}\mathcal{D}_{f_{\boldsymbol{X}}}(x)\dot{F}_T(y|x) \\ + \mathcal{H}_{F_T}(y|x) + \mathrm{o}(h^2)\Big\}\bigg]^2 + \mathrm{O}(n^{-1}).\end{aligned}$$

The appealing feature of MSE is its simple decomposition into variance and squared bias.
This error criterion is often preferred to other criteria such as mean absolute error since it is mathematically simpler to work with. The variance-bias decomposition allows easier analysis and interpretation of the performance of the kernel density estimator.
Rather than simply estimating the function at a fixed point, it is usually desirable, especially from a data analytic viewpoint, to estimate it over the entire real line. In this case our estimate is the function $\tilde{F}_L(y|\boldsymbol{x})$ so we need to consider an error criterion that globally measures the distance between the functions $\tilde{F}_L(y|\boldsymbol{x})$ and $F_T(y|\boldsymbol{x})$. One such error criterion is the Integrated Squared Error (ISE) given by

$$\mathrm{ISE}\Big\{\tilde{F}_L(y|\boldsymbol{x})\Big\} = \int \Big\{\tilde{F}_L(y|\boldsymbol{x}) - F_T(y|\boldsymbol{x})\Big\}^2 dy.$$

This may be recognized as the L_2 distance between $\tilde{F}_L(y|\boldsymbol{x})$ and $F_T(y|\boldsymbol{x})$. The ISE is appropriate if we are only concerned with the data set at hand, but it does not take into account other possible data sets. Therefore, it will be more appropriate to analyse the expected value of this random quantity, the Mean

Integrated Squared Error (MISE)

$$\begin{aligned}
\mathrm{MISE}\Big\{\tilde{F}_L(y|\boldsymbol{x})\Big\} &= \mathrm{E}\Big[\int\Big\{\tilde{F}_L(y|\boldsymbol{x}) - F_T(y|\boldsymbol{x})\Big\}^2 dy\Big] \\
&= \int \mathrm{Bias}\Big\{\tilde{F}_L(y|\boldsymbol{x})\Big\}^2 dy + \int \mathrm{Var}\Big\{\tilde{F}_L(y|\boldsymbol{x})\Big\}dy \\
&= \int \frac{R(\boldsymbol{K}_c{}^*)}{nhf_{\boldsymbol{X}}(\boldsymbol{x})}\big\{1 - F_T(y|\boldsymbol{x})\big\}^2 \sum_{j=1}^{q}\frac{d_j}{n_j(n_j-d_j)}dy \\
&\quad + \mathrm{O}(n^{-1}) + \int\Bigg[\Bigg\{\frac{h^2\mu_2(\boldsymbol{K}_c{}^*)}{2f_{\boldsymbol{X}}(\boldsymbol{x})}\mathcal{H}_{f_{\boldsymbol{X}}}(\boldsymbol{x})\Bigg\}F_T(y|\boldsymbol{x}) \\
&\quad + \frac{h^2\mu_2(\boldsymbol{K}_c{}^*)}{2}\Bigg\{2\frac{1}{f_{\boldsymbol{X}}(\boldsymbol{x})}\mathcal{D}_{f_{\boldsymbol{X}}}(\boldsymbol{x})\dot{F}_T(y|\boldsymbol{x}) \\
&\quad + \mathcal{H}_{F_T}(y|\boldsymbol{x})\Bigg\}\Bigg]^2 dy + \mathrm{o}(h^4).
\end{aligned}$$

A problem with the MSE and MISE expressions is that they depend on the bandwidth in a complicated way. This makes it difficult to interpret the influence of the bandwidth on the performance of the kernel estimator. Although, as we can see, one way of overcoming this problem involves the derivation of large sample approximations for leading variance and bias terms. These approximations have very simple expressions that allow a deeper appreciation of the role of the bandwidth. They can also be used to obtain the rate of convergence of the kernel estimator and the MISE-optimal bandwidth.
After the integration of the MISE expression we obtain

$$\mathrm{MISE}\Big\{\tilde{F}_L(y|\boldsymbol{x})\Big\} = \mathrm{AMISE}\Big\{\tilde{F}_L(y|\boldsymbol{x})\Big\} + \mathrm{o}(h^4) + \mathrm{O}(n^{-1}),$$

where

$$\begin{aligned}
\text{AMISE}\Big\{\tilde{F}_L(y|x)\Big\} &= \int \frac{R(\boldsymbol{K}_c{}^*)}{nhf_{\boldsymbol{X}}(x)}\big\{1-F_T(y|x)\big\}^2\sum_{j=1}^{q}\frac{d_j}{n_j(n_j-d_j)}dy \\
&\quad + \int\Bigg[\left\{\frac{h^2\mu_2(\boldsymbol{K}_c{}^*)}{2f_{\boldsymbol{X}}(x)}\mathcal{H}_{f_{\boldsymbol{X}}}(x)\right\}F_T(y|x) + \frac{h^2\mu_2(\boldsymbol{K}_c{}^*)}{2}\times \\
&\quad \left\{2\frac{1}{f_{\boldsymbol{X}}(x)}\mathcal{D}_{f_{\boldsymbol{X}}}(x)\dot{F}_T(y|x)+\mathcal{H}_{F_T}(y|x)\right\}\Bigg]^2 dy \\
&= \frac{R(\boldsymbol{K}_c{}^*)}{nhf_{\boldsymbol{X}}(x)}\sum_{j=1}^{q}\frac{d_j}{n_j(n_j-d_j)}\int\big\{1-F_T(y|x)\big\}^2dy \\
&\quad + \frac{h^4\mu_2(\boldsymbol{K}_c{}^*)^2}{4}\int\Bigg\{\frac{\mathcal{H}_{f_{\boldsymbol{X}}}(x)F_T(y|x)}{f_{\boldsymbol{X}}(x)} \\
&\quad + \frac{2}{f_{\boldsymbol{X}}(x)}\mathcal{D}_{f_{\boldsymbol{X}}}(x)\dot{F}_T(y|x)+\mathcal{H}_{F_T}(y|x)\Bigg\}^2 dy. \qquad (5.4)
\end{aligned}$$

We call this the asymptotic MISE since it provides a useful large sample approximation to the MISE. The AMISE is a much simpler expression to comprehend than the expression for the MISE. Notice that the integrated squared bias is asymptotically proportional to h^2, so for this quantity to decrease one needs to take h to be small. However, taking h small means an increase in the leading term of the integrated variance since this quantity is proportional to h^{-1}. Therefore, as n increases h should vary in such a way that each of the components of the MISE becomes smaller. This is known as the variance-bias trade-off and is a mathematical quantification for the critical role of the bandwidth. For very small h, $\tilde{F}_L(y|\boldsymbol{x})$ is very spiky and hence very variable in the sense that, over repeated sampling from $F_T(y|\boldsymbol{x})$, the spikes would appear in different places. There is, however, very little bias. If more smoothing is performed, that is h is increased, then the variability is reduced at the expense of introducing bias: for increasingly large h, there would be large bias because all features are eventually smoothed away, but little variance because the data are essentially ignored.

Another advantage of AMISE is that the optimal bandwidth with respect to this criterion has a closed form expression. This can be easily derived by differentiating (5.4) with respect to h and setting the derivative equal to zero.

$$\begin{aligned}\frac{\partial \text{AMISE}\left\{\tilde{F}_L(y|x)\right\}}{\partial h} &= -\frac{R(\boldsymbol{K}_c{}^*)}{nh^2 f_{\boldsymbol{X}}(x)} \sum_{j=1}^{q} \frac{d_j}{n_j(n_j-d_j)} \int \left\{1 - F_T(y|x)\right\}^2 dy \\ &\quad + h^3 \mu_2(\boldsymbol{K}_c{}^*)^2 \int \left\{ \frac{\mathcal{H}_{f_{\boldsymbol{X}}}(x) F_T(y|x)}{f_{\boldsymbol{X}}(x)} \right. \\ &\quad \left. + \frac{2}{f_{\boldsymbol{X}}(x)} \mathcal{D}_{f_{\boldsymbol{X}}}(x) \dot{F}_T(y|x) + \mathcal{H}_{F_T}(y|x) \right\}^2 dy \\ &= 0.\end{aligned}$$

Thus,

$$\begin{aligned}&\frac{R(\boldsymbol{K}_c{}^*)}{nh^2 f_{\boldsymbol{X}}(x)} \sum_{j=1}^{q} \frac{d_j}{n_j(n_j-d_j)} \int \left\{1 - F_T(y|x)\right\}^2 dy = h^3 \mu_2(\boldsymbol{K}_c{}^*)^2 \times \\ &\int \left\{ \frac{\mathcal{H}_{f_{\boldsymbol{X}}}(x) F_T(y|x)}{f_{\boldsymbol{X}}(x)} + \frac{2}{f_{\boldsymbol{X}}(x)} \mathcal{D}_{f_{\boldsymbol{X}}}(x) \dot{F}_T(y|x) + \mathcal{H}_{F_T}(y|x) \right\}^2 dy \\ &\Leftrightarrow \frac{R(\boldsymbol{K}_c{}^*)}{n f_{\boldsymbol{X}}(x)} \sum_{j=1}^{q} \frac{d_j}{n_j(n_j-d_j)} \int \left\{1 - F_T(y|x)\right\}^2 dy = h^5 \mu_2(\boldsymbol{K}_c{}^*)^2 \times \\ &\int \left\{ \frac{\mathcal{H}_{f_{\boldsymbol{X}}}(x) F_T(y|x)}{f_{\boldsymbol{X}}(x)} + \frac{2}{f_{\boldsymbol{X}}(x)} \mathcal{D}_{f_{\boldsymbol{X}}}(x) \dot{F}_T(y|x) + \mathcal{H}_{F_T}(y|x) \right\}^2 dy \Leftrightarrow \\ &h^5 = \frac{\frac{R(\boldsymbol{K}_c{}^*)}{n f_{\boldsymbol{X}}(x)} \sum_{j=1}^{q} \frac{d_j}{n_j(n_j-d_j)} \int \left\{1 - F_T(y|x)\right\}^2 dy}{\mu_2(\boldsymbol{K}_c{}^*)^2 \int \left\{ \frac{\mathcal{H}_{f_{\boldsymbol{X}}}(x) F_T(y|x)}{f_{\boldsymbol{X}}(x)} + \frac{2}{f_{\boldsymbol{X}}(x)} \mathcal{D}_{f_{\boldsymbol{X}}}(x) \dot{F}_T(y|x) + \mathcal{H}_{F_T}(y|x) \right\}^2 dy}.\end{aligned}$$

For

$$a = \frac{1}{n f_{\boldsymbol{X}}(x)} \sum_{j=1}^{q} \frac{d_j}{n_j(n_j-d_j)} \int \left\{1 - F_T(y|x)\right\}^2 dy,$$

$$b = \int \left\{ \frac{\mathcal{H}_{f_{\boldsymbol{X}}}(x) F_T(y|x)}{f_{\boldsymbol{X}}(x)} + \frac{2}{f_{\boldsymbol{X}}(x)} \mathcal{D}_{f_{\boldsymbol{X}}}(x) \dot{F}_T(y|x) + \mathcal{H}_{F_T}(y|x) \right\}^2 dy,$$

We have

$$h_{opt}{}^5 = \frac{a R(\boldsymbol{K}_c{}^*)}{b \mu_2(\boldsymbol{K}_c{}^*)^2}.$$

Therefore,

$$h_{opt} = \left\{ \frac{aR(\boldsymbol{K}_c{}^*)}{b\mu_2(\boldsymbol{K}_c{}^*)^2} \right\}^{1/5}. \tag{5.5}$$

The formula (5.5) for the optimal window width can not be used in practice since it shows that h_{opt} depends on unknown functions that need to be estimated. Nevertheless, a useful conclusion can be drawn. The ideal bandwidth will converge to zero as the sample size increases, but at a very slow rate. In order to obtain a practically useful expression we can replace the unknown quantities in (5.5) with their data driven estimations. First of all, we can estimate F_T with the Kaplan-Meier estimator. Then we will discuss some methods in order to estimate $f_{\boldsymbol{X}}$.

A very easy and natural approach is to use a standard family of distributions to assign a value to the term $f_{\boldsymbol{X}}$ in the expression (5.5) for the ideal bandwidth. For example, we can suppose the normal distribution with mean μ and variance σ^2. The parameters of the distribution can be estimated from the corresponding MLE estimators by using the data. While this method will work well if the population is normally distributed, it may oversmooth somewhat if the population is multimodal. In order to avoid such situations we can create a histogram of our data. Then, consider that we get a result like Fig. 5.1 our method will be appropriate in order to select the optimal bandwidth. Otherwise, we can choose a more appropriate distribution to assign a value to the $f_{\boldsymbol{X}}$.

Another approach is the plug-in bandwidth selector which is based on the simple idea of "plugging in" estimates of the unknown quantities that appear in formulas for the asymptotically optimal bandwidth. Now, we can estimate then unknown $f_{\boldsymbol{X}}$ by using again kernel density estimation. Unfortunately, this rule is not fully automatic since it depends on the choice of the new bandwidth g. One way of choosing g is to appeal to the formula for the AMSE-optimal bandwidth. However, this rule for choosing g has the same defect as the one for choosing h above: it depends on an unknown density functional. We could estimate this unknown density by using another kernel estimate, but its optimal bandwidth depends again on an amount that we do not know and this problem will not go away. The usual strategy for overcoming this problem is to estimate the unknown quantity in the optimal formula for g with a quick and simple estimate, such as a version of the normal scale rule described in the previous approach. This means that we really have a family of direct plug-in bandwidth selectors that depend on the number of stages of functional estimation before a quick and simple estimate is used.

Also motivated by the formula for the AMISE-optimal bandwidth, solve-the-

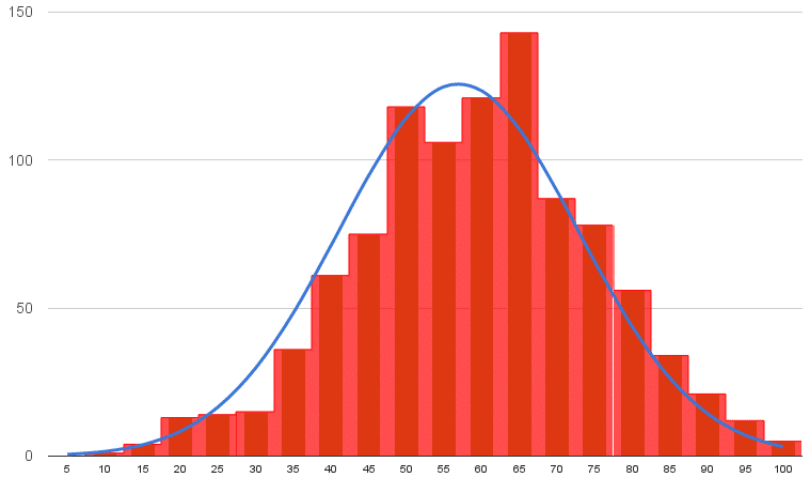

Figure 5.1: Histogram of Normal distribution

equation (STE) rules require that h be chosen to satisfy the relationship (5.5). The only difference with the previous method is that when we use kernel estimation for the $f_{\boldsymbol{X}}$ we consider that its bandwidth g is a function of h. Although this approach also leads to a "stage selection" problem, as in the direct plug-in case.

5.2 Multivariate bandwidth selection

As in the univariate setting we are also able to obtain a simple asymptotic approximation to the MISE of a multivariate kernel estimator under certain smoothness assumptions on the F_T and $f_{\boldsymbol{X}}$. These assumptions are needed to allow us to use a multivariate version of Taylor's theorem which we had previously introduced.

Firstly, we have to obtain the multivariate MSE which is given by

$$\begin{aligned}
\mathrm{MSE}\Big\{\tilde{F}_L(y|\boldsymbol{x})\Big\} &= \mathrm{E}\Big\{\tilde{F}_L(y|\boldsymbol{x}) - F_T(y|\boldsymbol{x})\Big\}^2 \\
&= \mathrm{Var}\Big\{\tilde{F}_L(y|\boldsymbol{x})\Big\} + \mathrm{Bias}\Big\{\tilde{F}_L(y|\boldsymbol{x})\Big\}^2 \\
&= \frac{R(\boldsymbol{K}_c{}^*)}{n|\boldsymbol{H}|f_{\boldsymbol{X}}(\boldsymbol{x})}\big\{1 - F_T(y|\boldsymbol{x})\big\}^2 \sum_{j=1}^{q} \frac{d_j}{n_j(n_j - d_j)} \\
&\quad + \left(\frac{\mu_2(\boldsymbol{K}_c{}^*)}{2f_{\boldsymbol{X}}(\boldsymbol{x})}\Big[2tr\big\{\boldsymbol{H}^\top \mathcal{D}_{f_{\boldsymbol{X}}}(\boldsymbol{x})\dot{F}_T(y|\boldsymbol{x})^\top \boldsymbol{H}\big\}\right. \\
&\quad + f_{\boldsymbol{X}}(\boldsymbol{x})tr\big\{\boldsymbol{H}^\top \mathcal{H}_{F_T}(y|\boldsymbol{x})\boldsymbol{H}\big\} + F_T(y|\boldsymbol{x})\times \\
&\quad \left. tr\big\{\boldsymbol{H}^\top \mathcal{H}_{f_{\boldsymbol{X}}}(\boldsymbol{x})\boldsymbol{H}\big\}\Big] + \mathrm{o}(\boldsymbol{H}^\top \boldsymbol{H})\right)^2 + \mathrm{O}(n^{-1}),
\end{aligned}$$

In addition, the multivariate MISE given by

$$\begin{aligned}
\mathrm{MISE}\Big\{\tilde{F}_L(y|\boldsymbol{x})\Big\} &= \int \mathrm{MSE}\Big\{\tilde{F}_L(y|\boldsymbol{x})\Big\}dy \\
&= \int \mathrm{Bias}\Big\{\tilde{F}_L(y|\boldsymbol{x})\Big\}^2 dy + \int \mathrm{Var}\Big\{\tilde{F}_L(y|\boldsymbol{x})\Big\}dy \\
&= \int \frac{R(\boldsymbol{K}_c{}^*)}{n|\boldsymbol{H}|f_{\boldsymbol{X}}(\boldsymbol{x})}\big\{1 - F_T(y|\boldsymbol{x})\big\}^2 \sum_{j=1}^{q} \frac{d_j}{n_j(n_j - d_j)}dy \\
&\quad + \int \left(\frac{\mu_2(\boldsymbol{K}_c{}^*)}{2f_{\boldsymbol{X}}(\boldsymbol{x})}\Big[2tr\big\{\boldsymbol{H}^\top \mathcal{D}_{f_{\boldsymbol{X}}}(\boldsymbol{x})\dot{F}_T(y|\boldsymbol{x})^\top \boldsymbol{H}\big\}\right. \\
&\quad + f_{\boldsymbol{X}}(\boldsymbol{x})tr\big\{\boldsymbol{H}^\top \mathcal{H}_{F_T}(y|\boldsymbol{x})\boldsymbol{H}\big\} + F_T(y|\boldsymbol{x})\times \\
&\quad \left. tr\big\{\boldsymbol{H}^\top \mathcal{H}_{f_{\boldsymbol{X}}}(\boldsymbol{x})\boldsymbol{H}\big\}\Big] + \mathrm{o}(\boldsymbol{H}^\top \boldsymbol{H})\right)^2 dy + \mathrm{O}(n^{-1}).
\end{aligned}$$

If we integrate the multivariate MISE expression then we obtain the multivariate AMISE

$$\mathrm{MISE}\Big\{\tilde{F}_L(y|\boldsymbol{x})\Big\} = \mathrm{AMISE}\Big\{\tilde{F}_L(y|\boldsymbol{x})\Big\} + \mathrm{o}\Big\{(\boldsymbol{H}^\top \boldsymbol{H})^2\Big\} + \mathrm{O}(n^{-1}),$$

where,

$$\begin{aligned}\text{AMISE}\Big\{\tilde{F}_L(y|\boldsymbol{x})\Big\} = \int \frac{R(\boldsymbol{K}_c{}^*)}{n|\boldsymbol{H}|f_{\boldsymbol{X}}(\boldsymbol{x})}\big\{1-F_T(y|\boldsymbol{x})\big\}^2\sum_{j=1}^{q}\frac{d_j}{n_j(n_j-d_j)}dy \\ + \int\left(\frac{\mu_2(\boldsymbol{K}_c{}^*)}{2f_{\boldsymbol{X}}(\boldsymbol{x})}\Big[2tr\big\{\boldsymbol{H}^\top\mathcal{D}_{f_{\boldsymbol{X}}}(\boldsymbol{x})\dot{F}_T(y|\boldsymbol{x})^\top\boldsymbol{H}\big\}\right. \\ + f_{\boldsymbol{X}}(\boldsymbol{x})tr\big\{\boldsymbol{H}^\top\mathcal{H}_{F_T}(y|\boldsymbol{x})\boldsymbol{H}\big\}+F_T(y|\boldsymbol{x})\times \\ \left. tr\big\{\boldsymbol{H}^\top\mathcal{H}_{f_{\boldsymbol{X}}}(\boldsymbol{x})\boldsymbol{H}\big\}\Big]\right)^2 dy.\end{aligned}$$

At the time of writing, the problem of selecting a bandwidth matrix from the data had received considerably less attention in the literature that its univariate counterpart. However, many of the ideas discussed in the univariate case for selecting h can be extended to the multivariate case. In this section we will briefly discuss some of these ideas, without getting into the more practical issues.

There are several levels of sophistication when specifying the bandwidth matrix $\boldsymbol{H}$. The simplest corresponds to the restriction $\boldsymbol{H} = h\boldsymbol{I}$ for some $h > 0$. The use of a single smoothing parameter h implies that the version of the kernel placed on each data point is scaled equally in all directions. Although, this method is appropriate when the spread of data points is equal in all the coordinate directions. If not, we can pre-scale the data to avoid extreme differences of spread in the covariates. If this is done then there will generally be no need to consider more complicated forms of the kernel density estimate than the one involving a single smoothing parameter. To sum up, this restriction has the advantage that one only has to deal with a single smoothing parameter, but the considerable disadvantage that the amount of smoothing is the same in each coordinate direction.

Now, in order to find the optimal bandwidth we can replace in (5.6) $|\boldsymbol{H}|$ with

h^d and $\boldsymbol{H}$ with $h\boldsymbol{I}$ as we mentioned previously. Then, (5.6) gives

$$\begin{aligned}
\text{AMISE}\Big\{\tilde{F}_L(y|\boldsymbol{x})\Big\} &= \int \frac{R(\boldsymbol{K}_c{}^*)}{nh^d f_{\boldsymbol{X}}(\boldsymbol{x})}\big\{1-F_T(y|\boldsymbol{x})\big\}^2 \sum_{j=1}^{q}\frac{d_j}{n_j(n_j-d_j)}dy \\
&\quad + \int\Bigg(\frac{\mu_2(\boldsymbol{K}_c{}^*)}{2f_{\boldsymbol{X}}(\boldsymbol{x})}\Big[2tr\big\{\boldsymbol{H}^\top \mathcal{D}_{f_{\boldsymbol{X}}}(\boldsymbol{x})\dot{F}_T(y|\boldsymbol{x})^\top \boldsymbol{H}\big\} \\
&\quad + f_{\boldsymbol{X}}(\boldsymbol{x})tr\big\{\boldsymbol{H}^\top \mathcal{H}_{F_T}(y|\boldsymbol{x})\boldsymbol{H}\big\} + F_T(y|\boldsymbol{x})\times \\
&\quad tr\big\{\boldsymbol{H}^\top \mathcal{H}_{f_{\boldsymbol{X}}}(\boldsymbol{x})\boldsymbol{H}\big\}\Big]\Bigg)^2 dy \\
&= \frac{1}{h^d}\int \frac{R(\boldsymbol{K}_c{}^*)}{nf_{\boldsymbol{X}}(\boldsymbol{x})}\big\{1-F_T(y|\boldsymbol{x})\big\}^2 \sum_{j=1}^{q}\frac{d_j}{n_j(n_j-d_j)}dy \\
&\quad + h^4\int\Bigg(\frac{\mu_2(\boldsymbol{K}_c{}^*)}{2f_{\boldsymbol{X}}(\boldsymbol{x})}\Big[2tr\big\{\mathcal{D}_{f_{\boldsymbol{X}}}(\boldsymbol{x})\dot{F}_T(y|\boldsymbol{x})^\top\big\} \\
&\quad + f_{\boldsymbol{X}}(\boldsymbol{x})tr\big\{\mathcal{H}_{F_T}(y|\boldsymbol{x})\big\} + F_T(y|\boldsymbol{x})tr\big\{\mathcal{H}_{f_{\boldsymbol{X}}}(\boldsymbol{x})\big\}\Big]\Bigg)^2 dy.
\end{aligned} \tag{5.6}$$

With

$$a_1 = \frac{1}{h^d}\int \frac{R(\boldsymbol{K}_c{}^*)}{nf_{\boldsymbol{X}}(\boldsymbol{x})}\big\{1-F_T(y|\boldsymbol{x})\big\}^2 \sum_{j=1}^{q}\frac{d_j}{n_j(n_j-d_j)}dy,$$

and

$$\begin{aligned}
b_1 = \int\Bigg(&\frac{\mu_2(\boldsymbol{K}_c{}^*)}{2f_{\boldsymbol{X}}(\boldsymbol{x})}\Big[2tr\big\{\mathcal{D}_{f_{\boldsymbol{X}}}(\boldsymbol{x})\dot{F}_T(y|\boldsymbol{x})^\top\big\} \\
&+ f_{\boldsymbol{X}}(\boldsymbol{x})tr\big\{\mathcal{H}_{F_T}(y|\boldsymbol{x})\big\} + F_T(y|\boldsymbol{x})tr\big\{\mathcal{H}_{f_{\boldsymbol{X}}}(\boldsymbol{x})\big\}\Big]\Bigg)^2 dy.
\end{aligned}$$

Then, (5.6) becomes

$$\text{AMISE}\Big\{\tilde{F}_L(y|\boldsymbol{x})\Big\} = \frac{a_1}{h^d} + b_1 h^4. \tag{5.7}$$

By differentiating (5.7) with respect to h and setting the derivative equal to zero, we get

$$\frac{\partial \text{AMISE}\Big\{\tilde{F}_L(y|\boldsymbol{x})\Big\}}{\partial h} = -d\frac{a_1}{h^{d+1}} + 4b_1 h^3 = 0.$$

Thus,

$$\begin{aligned}
\frac{da_1}{h^{d+1}} &= 4b_1h^3 \\
\frac{da_1}{4b_1} &= h^3h^{d+1} \\
h^{d+4} &= \frac{da_1}{4b_1} \\
h_{opt} &= \left(\frac{da_1}{4b_1}\right)^{1/d+4}.
\end{aligned}$$

Given the more stable performance of plug-in approaches in the univariate setting it seems worthwhile to investigate the performance of their multivariate extensions. Using the asymptotic approximations developed by the previous restriction it is possible to develop multivariate versions of plug-in type bandwidth selectors. From (5.6) and the discussion following we see that $f_{\boldsymbol{X}}$ is unknown quantity. We showed that explicit MISE expressions are available for the univariate kernel density estimator when $f_{\boldsymbol{X}}$ is a normal density and K is a Gaussian kernel. These expressions can perform exact MISE calculations without having to resort to numerical integration. In the multivariate setting the difficulties associated with numerical integration are magnified, so a flexible class of multivariate densities exhibiting explicit MISE expressions is useful to have. Multivariate normal density serve this same purpose in the multivariate setting. Thus, we can take $f_{\boldsymbol{X}}$ to be a standard density such as multivariate normal and $\boldsymbol{K}$ the d-variate normal kernel.
In certain circumstances, it may be more appropriate to use a vector of smoothing parameters or even a matrix of shrinking coefficients. This will be the case, for example, if the spread of the data points is very much greater in one of the coordinates directions than the others. So, at the next level, $\boldsymbol{H} = \text{diag}(h_1, \ldots, h_n)$, where at the expense of introducing $d-1$ additional smoothing parameters, one has the flexibility to smooth by different amounts in each of the d coordinate directions. However, there are situations where one might wish to smooth in directions different to those of the coordinate axes. In this case the full bandwidth matrix, would be appropriate.
The Least Squares Cross-Validation selector (LSCV) can be used for selection of a bandwidth matrix $\boldsymbol{H}$. Where, surprisingly, the relative rate of convergence of LSCV improves for higher dimensions. Least squares cross-validation is the name given to a conceptually simple and appealing bandwidth selector. Its motivation comes from expanding the MISE of the estimator and then ignore term does not depend on $\boldsymbol{H}$. After that, we use an unbiased estimator for

the unknown part thus this is the reason for the term "cross-validation" which refers to the use of part of a sample to obtain information about another part. It therefore seems reasonable to choose $\boldsymbol{H}$ to minimise this unbiased estimator. In addition, biased cross validation also has a straightforward extension to higher dimensions. In this method, instead of the exact MISE formula used by least squares cross-validation, Biased Cross-Validation (BCV) is based on the formula for the asymptotic MISE (AMISE).

CHAPTER 6

Real data analysis

6.1 Real data analysis

Through this section, we will apply a real data example for the estimator given by (3.10). The example is from Statlog (German Credit Data) Data Set which is provided by Prof. Hofmann and describes the profile of 1000 credit recipients of a German bank. The data set is comprised of 20 variables and it is publicly available at `http://archive.ics.uci.edu/ml/datasets/Statlog+(German\\+Credit+D-ata)`. This dataset classifies people described by this set of attributes as good or bad credit risks. Although the initial dataset involves variables which altogether describe the customers profile, such as the level of balance in their checking account (qualitative, measured in four levels representing the range of Deutsche Marks (DM) in which the account balance falls in), the duration of their employment (qualitative, measured in five levels representing periods of years), the level of their balance in their savings account (qualitative as the range of DM in which the balance falls in) etc. for illustrative purpose of our method we only use four attributes. Specifically, the age of the credit recipient ('Age', continuous, in years), its credit amount ('Amount' continuous, in DM), the duration of each credit (in months, variable: 'Duration'), the classification of each applicant as good or bad credit risk, and the variable 'Creditability' which contains the censoring indicator for each record, with $\delta = 1$ denoting the bad/defaulted credits and with $\delta = 0$ the good credits. There are 300 applicants classified as bad credits and 700 as good. The institution regards good credits as censored as it is not known if they are going to default; thus the amount of censoring is 70%.
With this example, we want to model the time to default, i.e. to estimate the probability that the duration of a credit will extend beyond a specific point in time. This information provides very useful guidance to bank managers when deciding whether to approve a loan to a prospective applicant based on

his/her profile. This is because there are two types of risks associated with the bank's decision: if the applicant is a good credit risk, i.e. if it is likely to repay the loan, disapproving the loan results in a loss of business to the bank. On the other hand, if the applicant is a bad credit risk, i.e. if it is not likely to repay the loan, approval of the loan results in a financial loss to the bank. According to the data set description, it is worse to classify a customer as good when he/she is bad, than it is to classify a customer as bad when he/she is good.
In our data the random variable Y which represents survival time corresponds to the duration in order to repay, the censoring indicator δ corresponds to the variable censoring and the covariate information denoted as the d-dimensional vector $\boldsymbol{X}$ corresponds to the variable 'Amount' which is the first column of the vector and the variable 'Age' which is the second column which means that $d = 2$. In addition, for the d-variate kernel $\boldsymbol{K}$ we chose the normal density function. Now, for the bandwidth h we use the function bw.nrd in R which is a bandwidth selector for Gaussian Kernels and the most common variation.
In order to have a better physical interpretation we will estimate the conditional survival function which is given by

$$\tilde{S}_L(y|\boldsymbol{x}) = 1 - \tilde{F}_L(y|\boldsymbol{x}).$$

In accordance to the discussion in Chapter 3 and for ease of presentation and visualization, an indicative situation of a practitioner utilizing only two covariates ('Amount' and 'Age') is exemplified. Thus there are three candidate models

$$\begin{aligned} S_1 &= P(\text{Duration} \geq y|\text{`Amount'} = x_1), \\ S_2 &= P(\text{Duration} \geq y|\text{`Age'} = x_2), \\ B &= P(\text{Duration} \geq y|\text{`Amount'} = x_1, \text{`Age'} = x_2), \end{aligned}$$

and obviously $B = S_1 \cup S_2$. All models (probabilities) above are estimated by $1 - \tilde{F}_L(y|\boldsymbol{x})$.

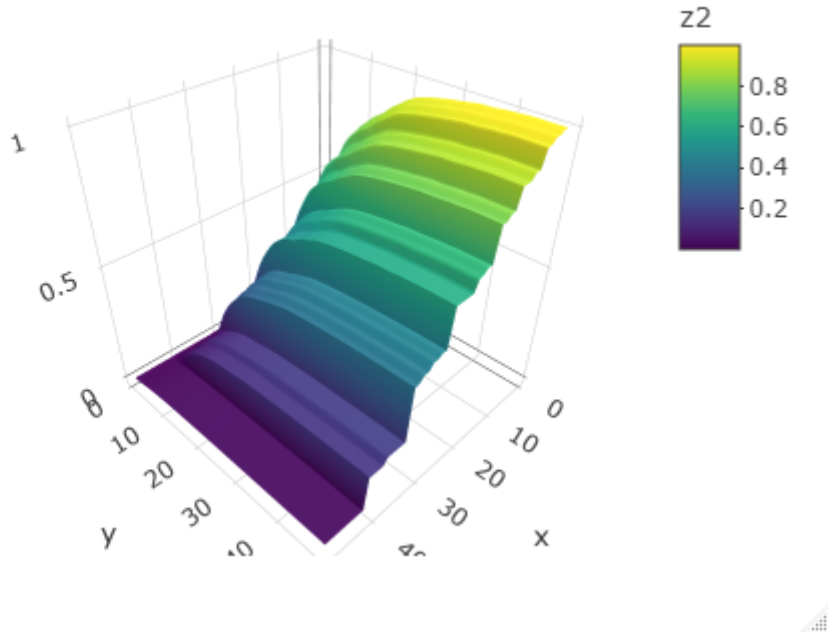

Figure 6.1: $P(\text{Duration} \geq y | \text{`Amount'} = x_1)$

In the first model we only maintain the covariate of credit amount. In order to analyze the present data set, 50 different levels for this covariate and variable Y have been considered by taking a sequence in its domain. Thus there are 2500 different covariate level combinations. Consequently, we plot these estimates and create a graph on $\mathbb{R}^2$ in which the values of the estimator are depicted on the Z-axis. As presented in Fig. 6.1 the higher the credit amount, the higher the probability to survive so the lower the probability to default.

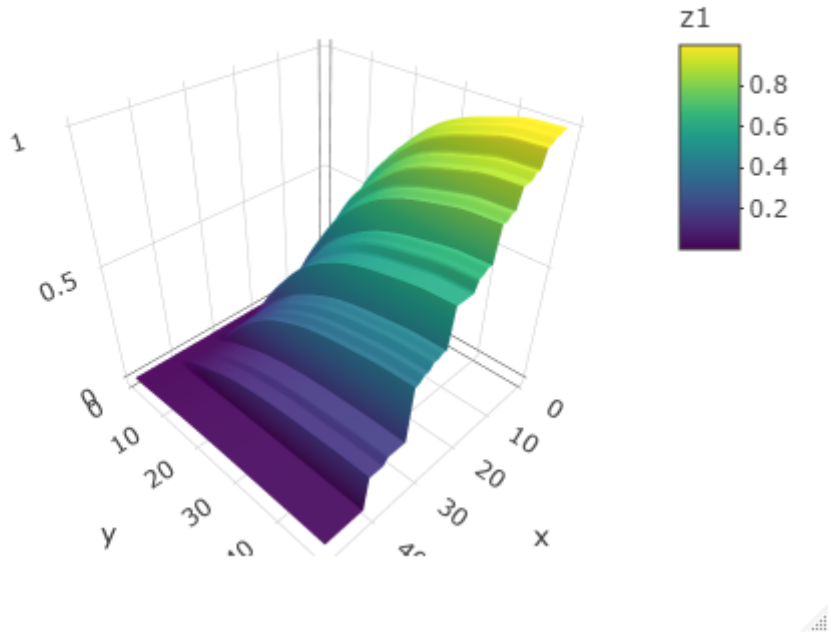

Figure 6.2: $P(\text{Duration} \geq y | \text{`Age'} = x_2)$

In the second model we only maintain the covariate of age. In order to analyze the present data set, 50 different levels for this covariate and variable Y have been considered by taking a sequence in its domain. Thus there are 2500 different covariate level combinations. Consequently, we plot these estimates and create a graph on $\mathbb{R}^2$ in which the values of the estimator are depicted on the Z-axis. As presented in Fig. 6.2 the higher the age, the higher the probability to survive so the lower the probability to default.

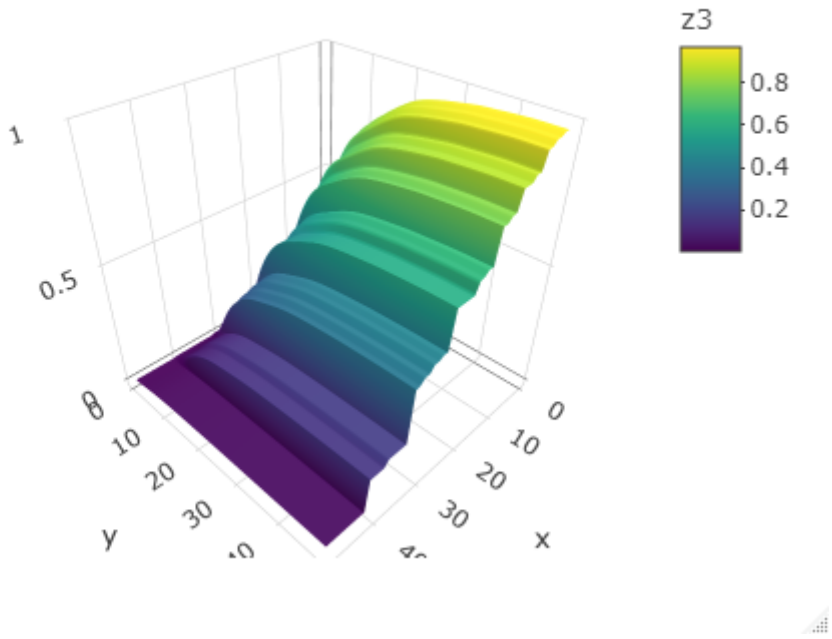

Figure 6.3: $P(\text{Duration} \geq y | \text{`Amount'} = x_1, \text{`Age'} = 63.57)$

In the third model we have both covariates but we maintain one of them constant. Specifically, we consider the age equal to 63.57 years. In order to analyze the present data set, 50 different levels for this covariate and variable Y have been considered by taking a sequence in its domain. Thus there are 2500 different covariate level combinations. Consequently, we plot these estimates and create a graph on $\mathbb{R}^2$ in which the values of the estimator are depicted on the Z-axis. As presented in Fig. 6.3 the higher the credit amount, the higher the probability to survive so the lower the probability to default.

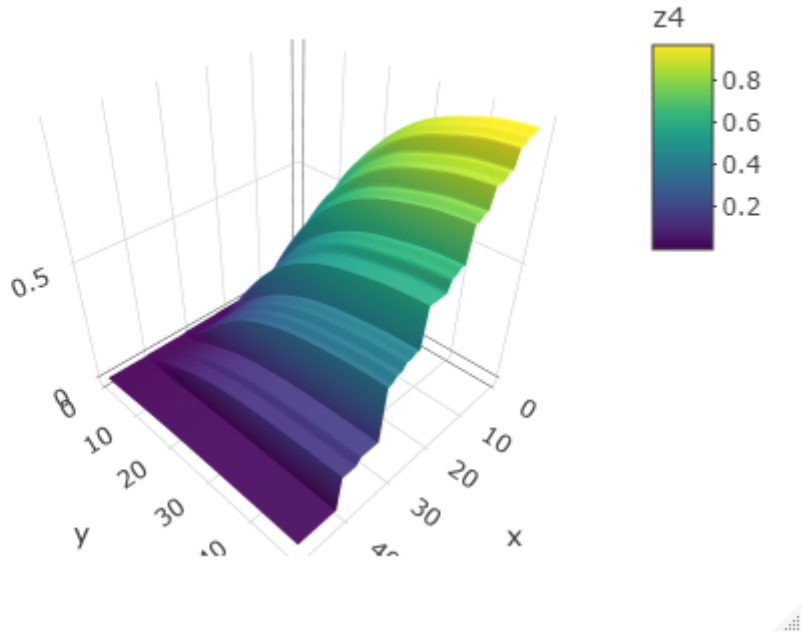

Figure 6.4: $P(\text{Duration} \geq y | \text{`Amount'} = 11006.04, \text{`Age'} = x_2)$

In the fourth model we have both covariates but we maintain one of them constant. Specifically, we consider the amount equal to 11006.04 DM. For the purpose of analyzing the present data set, 50 different levels for this covariate and variable Y have been considered by taking a sequence in its domain. Thus there are 2500 different covariate level combinations. Consequently, we plot these estimates and create a graph on $\mathbb{R}^2$ in which the values of the estimator are depicted on the Z-axis. As presented in Fig. 6.4 the higher the age, the higher the probability to survive so the lower the probability to default.

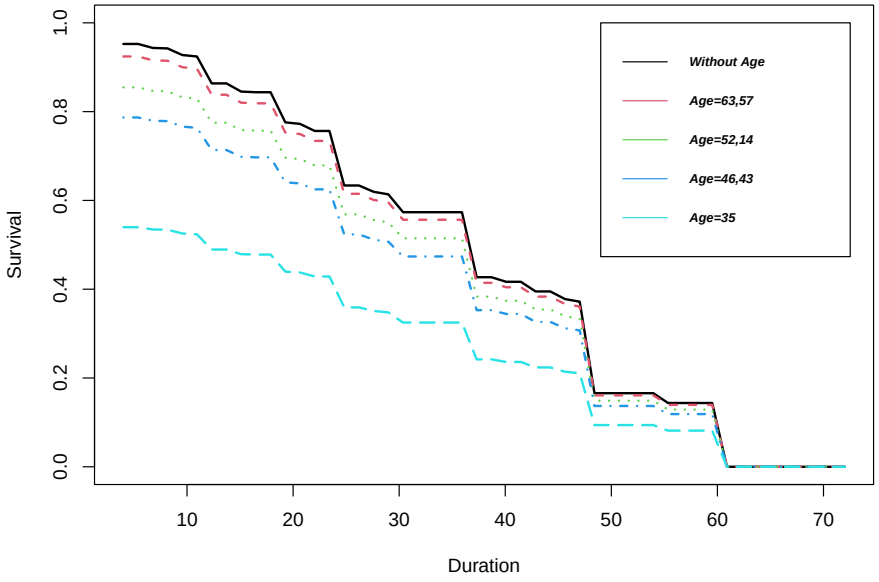

Figure 6.5: $P(\text{Duration}|\text{`Amount'} = 9337, \text{`Age'} = (63.57, 52.14, 46.43 \textit{ and } 35))$

As is evident in Fig. 6.5 by holding the 'Amount' covariate fixed and equal to 9337 DM we can understand how the 'Age' covariate affects the probability to default. In particular, the black line is the estimate without the 'Age' covariate, the red line is the estimate for 'Age'=63.75 years, the green line is for 'Age'=52.14 years, the blue line is for 'Age'=46.43 years and the light blue is for 'Age'=35 years. Thus, we conclude that the higher the age, the higher the probability to survive so the lower the probability to default.

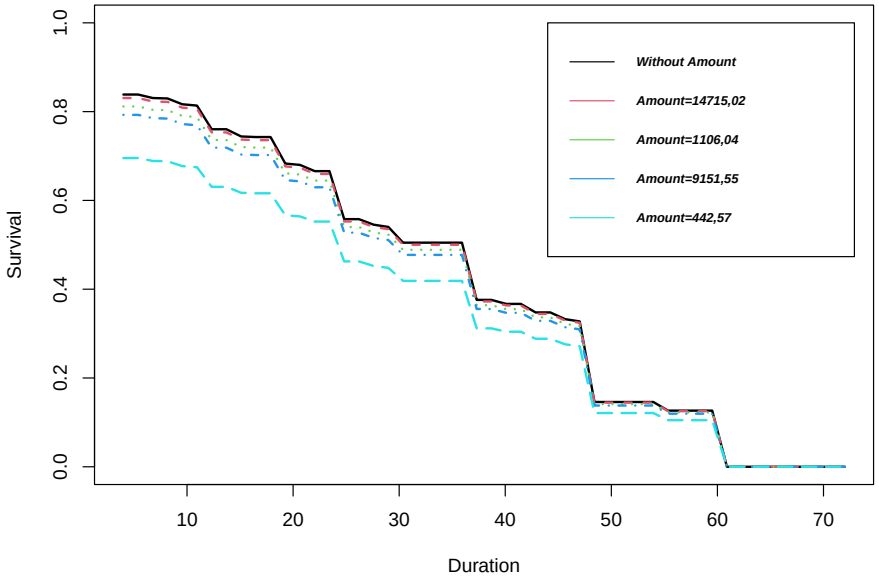

Figure 6.6: $P(\text{Duration} \geq y | \text{`Amount'} = (14715.02, 11006.04, 9151.55 \textit{ and } 5442.57), \text{`Age'} = 47)$

As is evident in Fig. 6.6 by holding the 'Age' covariate fixed and equal to 47 years we can understand how the 'Amount' covariate affects the probability to default. In particular, the black line is the estimate without the 'Amount' covariate, the red line is the estimate for 'Amount'=14715.02 DM, the green line is for 'Amount'=1106.04 DM, the blue line is for 'Amount'=9151.55 DM and the light blue is for 'Amount'=442.57 DM. Thus, we conclude that the higher the amount, the higher the probability to survive so the lower the probability to default.

An in depth investigation of the models behavior can be achieved through the concept of conditional MISE. Consequently it is reasonable, for individuals with the specific attributes, to draw the loan decision based on the probability to default returned by the model with the smallest MISE.
Summarizing the analysis, depending on the attributes of each potential loan recipient, the proposed methodology is capable of suggesting tailor made recommendations to practitioners and thus enable drawing data-driven optimal decisions.

CHAPTER 7

Appendix

The derivation of the asymptotic bias will require some additional matrix results. But first we have to define the trace of a square matrix A, denoted by $\text{tr}(A)$, which is the sum of the diagonal entries of A. Furthermore, in order to show Theorem 4.1.2, we need the following properties:

P1. If α is scalar and W a random variable, $\text{E}(\alpha W) = \alpha\text{E}(W)$.

P2. If W, Z are random variables, $\text{E}(W) = \text{E}\big\{\text{E}(W|Z)\big\}$.

P3. $\text{E}\{\hat{F}_T(Y)|\boldsymbol{X} = \boldsymbol{x}\} = F_T(y|\boldsymbol{x})$.

P4. $\text{E}\big\{g(W)\big\} = \int_{-\infty}^{\infty} g(w)f(w)dw$.

P5. If α is scalar, $\text{tr}(\alpha) = \alpha$.

P6. If α is scalar and A a square matrix, $\text{tr}(\alpha A) = \alpha\text{tr}(A)$.

P7. If W, Z are random variables, $\text{E}(W + Z) = \text{E}(W) + \text{E}(Z)$.

P8. If A is a square matrix and z a real column vector, $\text{tr}(z^\top Az) = \text{tr}(zz^\top A)$.

P9. X_i are identically distributed variables which means that $\text{E}(X_1) = \ldots \text{E}(X_n)$. So, for a fixed i and (P7) $\text{E}(X_1 + \cdots + X_n) = \text{E}(X_1) + \cdots + \text{E}(X_n) = n\text{E}(X_i)$.

P10. $\int g(Ax)dx = |A| \int g(y)dy$, where A represents an invertible $d \times d$ matrix.

P11. $\text{Var}(X) = \text{E}(X^2) - \big\{\text{E}(X)\big\}^2$, where X is a random variable.

P12. If α is scalar and W a random variable, $\text{Var}(\alpha W) = \alpha^2\text{Var}(W)$.

P13. $\text{Var}(W + Z) = \text{Var}(W) + \text{Var}(Z)$, where W and Z are independent random variables.

P14. $P(x < X \leq x + dx) \approx f(x)dx$, where $f(x)$ is the p.d.f. of the random variable X.

For the estimator of the conditional c.d.f. in Chapter 4 we have that $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$ where $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{id})^\top$ and $\boldsymbol{x} = (x_1, \ldots, x_d)^\top$. So,

$$\mathbf{X} = \begin{pmatrix} 1 & (\boldsymbol{X}_1 - \boldsymbol{x})^\top \\ \ldots & \ldots \\ 1 & (\boldsymbol{X}_n - \boldsymbol{x})^\top \end{pmatrix}_{n\times(d+1)} = \begin{pmatrix} 1 & X_{11} - x_1 & \ldots & X_{1d} - x_d \\ \ldots & \ldots & \ldots & \ldots \\ 1 & X_{n1} - x_1 & \ldots & X_{nd} - x_d \end{pmatrix}.$$

Also,

$$Y = \left(\hat{F}_T(y_1), \ldots, \hat{F}_T(y_n)\right)^\top_{n\times 1},$$

$$\boldsymbol{W} = \operatorname{diag}\Big[\boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_1 - \boldsymbol{x})\}, \ldots, \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_n - \boldsymbol{x})\}\Big]_{n\times n},$$

$$B = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}_{(d+1)\times 1}.$$

So, the estimator (3.7) is given by

$$\hat{B} = (\mathbf{X}^\top \boldsymbol{W}\mathbf{X})^{-1}\mathbf{X}^T \boldsymbol{W}Y = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_{11} \\ \ldots \\ \hat{\beta}_{1d} \end{pmatrix}_{(d+1)\times 1}.$$

In order to find $\hat{\beta}_0$, firstly we have to calculate $\mathbf{X}^\top \boldsymbol{W}Y$.

$$\mathbf{X}^\top \boldsymbol{W}Y = \begin{pmatrix} 1 & \ldots & 1 \\ (X_{11} - x_1) & \ldots & (X_{n1} - x_1) \\ \vdots & \ddots & \vdots \\ (X_{1d} - x_d) & \ldots & (X_{nd} - x_d) \end{pmatrix}_{(d+1)\times n}$$

$$\begin{pmatrix} \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_1 - \boldsymbol{x})\} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_n - \boldsymbol{x})\} \end{pmatrix}_{n\times n} \begin{pmatrix} \hat{F}_T(y_1) \\ \ldots \\ \hat{F}_T(y_n) \end{pmatrix}_{n\times 1}$$

$$= \begin{pmatrix} \sum_{i=1}^n \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\}\hat{F}_T(y) \\ \sum_{i=1}^n \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\}(\boldsymbol{X}_{i1} - x_1)\hat{F}_T(y) \\ \ldots \\ \sum_{i=1}^n \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\}(\boldsymbol{X}_{id} - x_d)\hat{F}_T(y) \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{i=1}^n \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\}\hat{F}_T(y) \\ \sum_{i=1}^n \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\}(\boldsymbol{X}_i - x)\hat{F}_T(y) \end{pmatrix}_{(d+1)\times 1}.$$

Now, we have to calculate the quantity $\mathbf{X}^\top \boldsymbol{W}\boldsymbol{X}$.

$$\mathbf{X}^\top \boldsymbol{W}\mathbf{X} = \begin{pmatrix} 1 & \dots & 1 \\ (X_{11}-x_1) & \dots & (X_{n1}-x_1) \\ \vdots & \ddots & \vdots \\ (X_{1d}-x_d) & \dots & (X_{nd}-x_d) \end{pmatrix}_{(d+1)\times n}$$
$$\begin{pmatrix} \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_1-\boldsymbol{x})\} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_n-\boldsymbol{x})\} \end{pmatrix}_{n\times n} \times$$
$$\begin{pmatrix} 1 & (X_{11}-x_1) & \dots & (X_{1d}-x_d) \\ \vdots & \ddots & \ddots & \vdots \\ 1 & (X_{n1}-x_1) & \dots & (X_{nd}-x_d) \end{pmatrix}_{n\times(d+1)} =$$
$$\begin{pmatrix} \sum_{i=1}^n \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i-\boldsymbol{x})\} & \dots & \sum_{i=1}^n \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i-\boldsymbol{x})\}(X_{id}-x_d) \\ \sum_{i=1}^n \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i-\boldsymbol{x})\}(X_{i1}-x_1) & \dots & \sum_{i=1}^n \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i-\boldsymbol{x})\}(X_{i1}-x_1)(X_{id}-x_d) \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^n \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i-\boldsymbol{x})\}(X_{id}-x_d) & \dots & \sum_{i=1}^n \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i-\boldsymbol{x})\}(X_{id}-x_d)^2 \end{pmatrix}$$
$$= \begin{pmatrix} S_{n,0}(\boldsymbol{x}) & S_{n,1}(\boldsymbol{x})^\top \\ S_{n,1}(\boldsymbol{x}) & S_{n,2}(\boldsymbol{x}) \end{pmatrix}_{(d+1)\times(d+1)}.$$

With,

$$S_{n,0}(\boldsymbol{x})_{1\times 1} = \sum_{i=1}^n \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i-\boldsymbol{x})\},$$
$$S_{n,1}(\boldsymbol{x})_{d\times 1} = \sum_{i=1}^n \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i-\boldsymbol{x})\}(\boldsymbol{X}_i-\boldsymbol{x}),$$
$$S_{n,2}(\boldsymbol{x})_{d\times d} = \sum_{i=1}^n \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i-\boldsymbol{x})\}(\boldsymbol{X}_i-\boldsymbol{x})(\boldsymbol{X}_i-\boldsymbol{x})^\top.$$

Now, we want to calculate $(\mathbf{X}^\top \boldsymbol{W}\mathbf{X})^{-1}$. From linear algebra, the inverse, say M^{-1}, of a 2×2 block matrix M where

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix},$$

with $\boldsymbol{A}$ is $m \times m$ invertible matrix, $\boldsymbol{D}$ is $n \times n$ invertible matrix and $\boldsymbol{B}$ and $\boldsymbol{C}$ are conformable with them for partitioning, is

$$\boldsymbol{M}^{-1} = \begin{pmatrix} (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1} & -\boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1} \\ -\boldsymbol{D}^{-1}\boldsymbol{C}(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1} & (\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1} \end{pmatrix}.$$

Note that from (3.7),

$$\hat{\beta}_0 = \boldsymbol{e}_1(\mathbf{X}^\top \boldsymbol{W}\mathbf{X})^{-1}\mathbf{X}^T \boldsymbol{W}Y,$$

where,

$$\boldsymbol{e}_1 = (1, 0, \ldots, 0)_{1\times(d+1)}.$$

Thus, we only need the first row of the $(\mathbf{X}^\top \boldsymbol{W}\mathbf{X})^{-1}$. Suppose that,

$$\begin{aligned} p_1 &= (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1} \\ &= \left\{S_{n,0}(\boldsymbol{x}) - S_{n,1}(\boldsymbol{x})^\top S_{n,2}(\boldsymbol{x})^{-1} S_{n,1}(\boldsymbol{x})\right\}^{-1}, \\ \boldsymbol{p}_2 &= -\boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1} \\ &= -S_{n,0}(\boldsymbol{x})^{-1} S_{n,1}(\boldsymbol{x})^\top \left\{S_{n,2}(\boldsymbol{x}) - S_{n,1}(\boldsymbol{x}) S_{n,0}(\boldsymbol{x})^{-1} S_{n,1}(\boldsymbol{x})^\top\right\}^{-1}, \end{aligned}$$

gives,

$$\boldsymbol{e}_1(\mathbf{X}^\top \boldsymbol{W}\mathbf{X})^{-1} = \begin{pmatrix} p_1 & \boldsymbol{p}_2 \end{pmatrix}.$$

So,

$$\begin{aligned} \hat{\beta}_0 &= \begin{pmatrix} p_1 & \boldsymbol{p}_2 \end{pmatrix}_{1\times(d+1)} \begin{pmatrix} \sum_{i=1}^n \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\}\hat{F}_T(y) \\ \sum_{i=1}^n \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\}(\boldsymbol{X}_i - x)\hat{F}_T(y) \end{pmatrix}_{(d+1)\times 1} \\ &= p_1 \sum_{i=1}^n \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\}\hat{F}_T(y) + \boldsymbol{p}_2 \sum_{i=1}^n \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\}(\boldsymbol{X}_i - x)\hat{F}_T(y) \\ &= \sum_{i=1}^n \Big\{p_1 + \boldsymbol{p}_2(\boldsymbol{X}_i - x)\Big\}\boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\}\hat{F}_T(y). \end{aligned} \tag{7.1}$$

For $d = 1$ an equivalent form of (7.1) is given by

$$\tilde{F}_L(y|\boldsymbol{x}) \equiv \hat{\beta}_0 = \sum_{i=1}^n \frac{S_{n,2}(\boldsymbol{x}) - S_{n,1}(\boldsymbol{x})(\boldsymbol{X}_i - \boldsymbol{x})}{S_{n,2}(\boldsymbol{x})S_{n,0}(\boldsymbol{x}) - S_{n,1}(\boldsymbol{x})S_{n,1}(\boldsymbol{x})} \times \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\}\hat{F}_T(y).$$

Now, by using lemma 4.1.1

$$\begin{aligned}
p_1 &= \{S_{n,0}(\boldsymbol{x}) - S_{n,1}(\boldsymbol{x})^\top S_{n,2}(\boldsymbol{x})^{-1} S_{n,1}(\boldsymbol{x})\}^{-1} \\
&= \big[n|\boldsymbol{H}|f_{\boldsymbol{X}}(\boldsymbol{x})S_{0,c} - n|\boldsymbol{H}|^2 f_{\boldsymbol{X}}(\boldsymbol{x})S_{1,c}^\top \{n|\boldsymbol{H}|^3 f_{\boldsymbol{X}}(\boldsymbol{x})S_{2,c}\}^{-1} n|\boldsymbol{H}|^2 f_{\boldsymbol{X}}(\boldsymbol{x})S_{1,c}\big]^{-1} \\
&= \frac{1}{n|\boldsymbol{H}|f_{\boldsymbol{X}}(\boldsymbol{x})}\{S_{0,c}(\boldsymbol{x}) - S_{1,c}(\boldsymbol{x})^\top S_{2,c}(\boldsymbol{x})^{-1} S_{1,c}(\boldsymbol{x})\}^{-1} \\
&= \frac{1}{n|\boldsymbol{H}|f_{\boldsymbol{X}}(\boldsymbol{x})}\tilde{p}_1. \qquad (7.2)
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{p}_2 &= -S_{n,0}(\boldsymbol{x})^{-1} S_{n,1}(\boldsymbol{x})^\top \{S_{n,2}(\boldsymbol{x}) - S_{n,1}(\boldsymbol{x})S_{n,0}(\boldsymbol{x})^{-1} S_{n,1}(\boldsymbol{x})^\top\}^{-1} \\
&= -\frac{n|\boldsymbol{H}|^2 f_{\boldsymbol{X}}(\boldsymbol{x})S_{1,c}^\top}{n|\boldsymbol{H}|f_{\boldsymbol{X}}(\boldsymbol{x})S_{0,c}}\{n|\boldsymbol{H}|^3 f_{\boldsymbol{X}}(\boldsymbol{x})S_{2,c} \qquad (7.3) \\
&\quad - \frac{n^2|\boldsymbol{H}|^4 f_{\boldsymbol{X}}(\boldsymbol{x})^2}{n|\boldsymbol{H}|f_{\boldsymbol{X}}(\boldsymbol{x})} S_{1,c}(\boldsymbol{x})S_{0,c}(\boldsymbol{x})^{-1} S_{1,c}(\boldsymbol{x})^\top\}^{-1} \\
&= -\frac{1}{n|\boldsymbol{H}^2|f_{\boldsymbol{X}}(\boldsymbol{x})} S_{0,c}(\boldsymbol{x})^{-1} S_{1,c}(\boldsymbol{x})^\top \{S_{2,c}(\boldsymbol{x}) - S_{1,c}(\boldsymbol{x})S_{0,c}(\boldsymbol{x})^{-1} S_{1,c}(\boldsymbol{x})^\top\}^{-1} \\
&= -\frac{1}{n|\boldsymbol{H}^2|f_{\boldsymbol{X}}(\boldsymbol{x})}\tilde{p}_2. \qquad (7.4)
\end{aligned}$$

By substituting (7.2) and (7.4) in (7.1)

$$\begin{aligned}
\tilde{F}_L(y|\boldsymbol{x}) \equiv \hat{\beta}_0 &= \sum_{i=1}^{n}\{p_1 + \boldsymbol{p}_2(\boldsymbol{X}_i - x)\}\boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\}\hat{F}_T(y) \\
&= \sum_{i=1}^{n}\Big\{\frac{1}{n|\boldsymbol{H}|f_{\boldsymbol{X}}(\boldsymbol{x})}\tilde{p}_1 - \frac{1}{n|\boldsymbol{H}^2|f_{\boldsymbol{X}}(\boldsymbol{x})}\tilde{\boldsymbol{p}}_2(\boldsymbol{X}_i - x)\Big\}\times \\
&\quad \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\}\hat{F}_T(y) \\
&= \frac{1}{n|\boldsymbol{H}|f_{\boldsymbol{X}}(\boldsymbol{x})}\sum_{i=1}^{n}\{\tilde{p}_1 - \tilde{\boldsymbol{p}}_2|\boldsymbol{H}|^{-1}(\boldsymbol{X}_i - x)\}\times \\
&\quad \boldsymbol{K}\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\}\hat{F}_T(y) \\
&= \frac{1}{n|\boldsymbol{H}|f_{\boldsymbol{X}}(\boldsymbol{x})}\sum_{i=1}^{n}\mathbf{K}_c{}^*\{\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\}\hat{F}_T(y).
\end{aligned}$$

Where,

$$\boldsymbol{K}_c{}^*(u) = (\tilde{p}_1 - \tilde{\boldsymbol{p}}_2\boldsymbol{u})\boldsymbol{K}(\boldsymbol{u})\boldsymbol{I}_{\{-c,+\infty\}}(\boldsymbol{u}).$$

Bibliography

Birkhoff, G. D. (1931). Proof of the ergodic theorem, *Proceedings of the National Academy of Sciences* **17**(12): 656–660.

Borovkov, A. (1984). Mathematical statistics. parameter estimation, *Nauka, Moscow* .

Cacoullos, T. (1964). Estimation of a multivariate density, *Technical report*, University of Minnesota.

Chacón, J. E. and Duong, T. (2018). *Multivariate kernel smoothing and its applications*, CRC Press.

Chen, K. and Lo, S.-H. (1997). On the rate of uniform convergence of the product-limit estimator: strong and weak laws, *The Annals of Statistics* **25**(3): 1050–1087.

Claeskens, G. and Hall, P. (2002). Theory & methods: Data sharpening for hazard rate estimation, *Australian & New Zealand Journal of Statistics* **44**(3): 277–283.

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density, *Theory of Probability & Its Applications* **14**(1): 153–158.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, Vol. 66, CRC Press.

Fix, E. and Hodges, J. L. (1989). Discriminatory analysis. nonparametric discrimination: Consistency properties, *International Statistical Review/Revue Internationale de Statistique* **57**(3): 238–247.

Glivenko, V. (1933). Sulla determinazione empirica delle leggi di probabilita, *Gion. Ist. Ital. Attauri.* **4**: 92–99.

Greenwood, M. (1926). The" errors of sampling" of the survivorship tables, *Reports on public health and medical subjects* .

Gulati, S. and Padgett, W. (1996). Families of smooth confidence bands for the survival function under the general random censorship model, *Lifetime Data Analysis* **2**: 349–362.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**(282): 457–481.

Karunamuni, R. and Yang, S. (1991). Weak and strong uniform consistency rates of kernel density estimates for randomly censored data, *Canadian Journal of Statistics* **19**(4): 349–359.

Kim, C., Bae, W., Choi, H. and Park, B. U. (2005). Non-parametric hazard function estimation using the kaplan–meier estimator, *Nonparametric Statistics* **17**(8): 937–948.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival analysis: techniques for censored and truncated data*, Vol. 1230, Springer.

Kulasekera, K., Williams, C. L., Coffin, M. and Manatunga, A. (2001). Smooth estimation of the reliability function, *Lifetime Data Analysis* **7**: 415–433.

Lawless, J. F. (2011). *Statistical models and methods for lifetime data*, John Wiley & Sons.

Lemdani, M. and Ould-Saıd, E. (2003). L1-deficiency of the Kaplan–Meier stimator, *Statistics & Probability Letters* **63**(2): 145–155.

Lio, Y. and Padgett, W. (1992). Asymptotically optimal bandwidth for a smooth nonparametric quantile estimator under censoring, *Journal of Nonparametric Statistics* **1**(3): 219–229.

Rao, S. S. (2017). Advanced statistical inference, *Texas A & M University* .

Scott, D. W. (2015). *Multivariate density estimation: theory, practice, and visualization*, John Wiley & Sons.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, Vol. 26, CRC Press.

Turkson, A. J., Ayiah-Mensah, F. and Nimoh, V. (2021). Handling censoring and censored data in survival analysis: a standalone systematic literature review, *International Journal of Mathematics and Mathematical Sciences* **2021**: 1–16.

Wand, M. P. and Jones, M. C. (1994). *Kernel smoothing*, CRC Press.