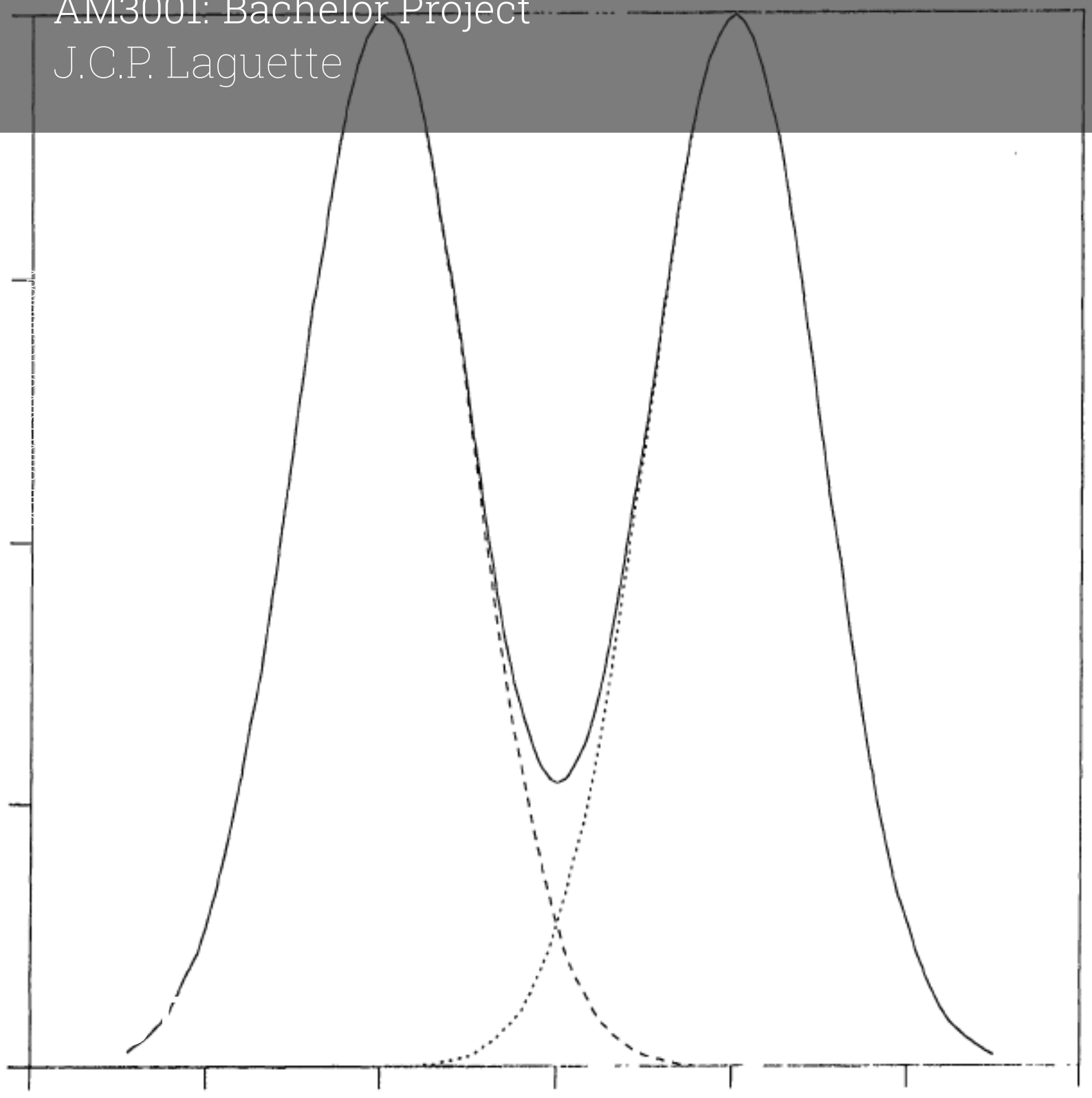
Uniform deconvolution

and its connection to current status

AM3001: Bachelor Project

J.C.P. Laguette



Uniform deconvolution

and its connection to current status

by

J.C.P. Laguette

Instructor:	G. Jongbloed
Project Duration:	April, 2025 - August, 2025
Faculty:	EEMCS, Delft

Style: TU Delft Report Style, with modifications by Daan Zwaneveld

TU Delft

Preface

This thesis is written as final requirement for the degree of Bachelor of Science in Applied Mathematics at Delft University of Technology. This thesis finishes four years of work to obtain this degree. I want to thank Geurt Jongbloed for the opportunity to do this project and the productive meetings. I really enjoyed the meetings with Geurt Jongbloed as he always had sharp and accurate answers on my questions with some fun facts here and there. Before I started this project I had never even heard of the concepts of deconvolution or nonparametric estimation, but in researching this concept it has started to fascinate me more and more. I hope that the reader also finds this concept as interesting as I did when writing this.

J.C.P. Laguette
Delft, August 2025

Abstract

In this abstract, OpenAI, 2025 was used for text references and rewriting parts of the text.

Abstract for non-experts

This report will discuss the uniform deconvolution problem. This a statistical problem where the data one would like to observe is distorted with a noise term which follows a uniform distribution. Although this may seem like a rare situation, the problem occurs often in the mathematical fields of statistics and inverse problems. This problem arises in blurred photos, measurement errors and rounding in survey data. It notably has an unexpected connection to a specific case of interval censoring. This is a problem that occurs in medical studies where one only observes whether a condition (such as an infection) has occurred by a certain time, the testing time, without knowing the exact onset. One only observes the 'current status'. This specific problem will be discussed in this report alongside the general case of the uniform deconvolution problem.

The report begins by formally defining these problems from a mathematical perspective. We then introduce and apply nonparametric methods to estimate the unknown distributions of the underlying data, based on the noisy data. These methods, while not typically covered in standard undergraduate curricula in applied mathematics, offer powerful tools for tackling such problems. We conclude by comparing the different approaches in terms of their effectiveness, assumptions, and practical limitations.

General abstract

In this report the uniform deconvolution problem will be discussed. It is a statistical problem where the observations we would like to make are distorted by an independent additive noise sampled from a standard uniform distribution. The sampling density is therefore the convolution of the distribution function of interest and the standard uniform density. However, we would like to know the distribution of the data without the uniform convolution, or in other words, we want to infer the deconvolution of the distribution of the observables by the uniform density, based on the observable noisy data. We will first define this problem mathematically and find a connection between the distribution of the observed data and the distribution that we would like to know. Having constructed this relation, we aim to estimate the unknown distribution based the noisy observations.

It turns out that this problem is related to the current status problem. This is a problem in which patients are tested for a disease to find out the time of onset of the disease. However, the observations only tell us whether patients are infected or not at the time of the test, but not the time of onset of the disease. We would like to know the times of onset as we can use these to estimate the distribution of the time of onset. To estimate this distribution, we will use the connection to the to the uniform deconvolution problem.

Thus, when the relation between the distribution of the observations and the unknown distribution of the variables of interest is created, we would like to estimate the distribution function only using finitely many noisy observations. To estimate this, estimators like the MLE or MoM cannot be used as we would need to assume that the family of possible distributions can be parameterized smoothly by a Euclidean parameter. In our case, the family of possible distributions is much larger and does not satisfy this assumption, so we have to use nonparametric estimation methods. We will show two different methods of this nonparametric estimation; the nonparametric maximum likelihood estimator and the kernel density estimator.

For the nonparametric maximum likelihood estimator, we first constructs the (log-)likelihood function of the problem and then we maximizes it over all possible distribution functions in the allowed family. This maximizing function is defined as the NPMLE. There is not a general method to obtain this distribution so we will have to use specific properties of our problem. Then we will use the method of isotonic regression to obtain the maximum likelihood estimator for the unknown distribution function.

The other method that will be used to estimate the unknown distribution in a nonparametric way is kernel density estimation. This method constructs a distribution function more directly than the NPMLE. The intuition behind this estimator is simple, if there are a lot of observations near a value, the probability density in that value should be high, and the other way around.

Finally, we will simulate the uniform deconvolution problem and solve it using the two discussed methods. The methods and their respective strengths and limitations will be compared. The kernel density estimator turns out to give a more smooth and accurate estimate but depends on an optimally chosen parameter which, in practice, cannot be calculated exactly and differs from case to case. The nonparametric maximum likelihood estimator does not depend on any parameter and the way of calculating it is the same in every case, but is only able to produce step functions.

Contents

Nomenclature

Abbreviations

Abbreviation	Definition
MLE	Maximum Likelihood Estimator
MoM	Method of moments estimator
NPMLE	Nonparametric Maximum Likelihood Estimator
KDE	Kernel density estimator/estimation
MCM	Maximum convex minorant

Symbols

Symbol	Definition
f_X or g_X	Density function of the random variable X
F_X or G_X	Distribution function of the random variable X
F_0	Unknown distribution function
$(f_X * f_Y)$	Convolution of the densities f_X and f_Y
$\hat{F}$	Estimator of F
$\mathbf{y}$	The vector $\mathbf{y} = (y_1, \ldots, y_n)$
$\mathcal{L}_n$	Likelihood function with n observations
ℓ_n	log-likelihood function with n observations
$K(\cdot)$	Kernel density function
h	bandwidth

1 Introduction

In this introduction, OpenAI, 2025 was used for text references and rewriting parts of the text.

Suppose we want to estimate a distribution of some interesting variable. We measure independent realizations of the of the random variable X. However, our measurements all have some additive noise term following a standard uniform distribution which distort the observations. We would, of course, like to know the distribution of the variables X without the noise. This is the uniform deconvolution problem, we observe data of which its density is a convolution of the distribution of interest and a uniform distribution. Mathematically, we observe instances of Z where $Z = X + U$ with $X \sim F_0$ and $U \sim Unif([0, 1])$. We are interested in the unknown distribution F_0 of X so we want a deconvolution of the sampling density g_z, by the uniform density to estimate F_0.

As a starter, we have to relate the distribution F_0 to the density of the observed data g_Z. More importantly, we would like to solve the inverse problem, expressing the distribution function F_0 in terms of the sampling density g_Z. This is possible, using probability theory we show in Chapter 2 that

$$F_0(z) = \sum_{i=0}^{\infty} g_Z(z - i), \ \forall z \in [0, \infty).$$

The uniform deconvolution problem also has a close relation to a specific instance of the current status problem, which has a direct application in the 'real world'. It is a problem where we test individuals for, for example, a disease. We only know that an individual is infected at the time of the test but not when the individual got infected exactly. We would like to know the distribution of the infection time and to estimate this. We can link this problem to the uniform deconvolution problem. The general current status problem also has some known properties and results that we will use to derive results for the uniform deconvolution problem.

Suppose we know that our observations are sampled from a distribution function F. When F belongs to a parametric family $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^k$, we can use known methods like MLE or MoM to estimate F as this boils down to estimating the k-dimensional parameter θ. This is called 'parametric estimation'. However, in our problem the family of distributions $\mathcal{F}$ cannot be parameterized by some finite dimensional value, our family is a function space which has infinite dimension. This is known as 'nonparametric estimation' and in this report we will discuss two methods of estimating the function F in a nonparametric way.

The first estimator is the nonparametric likelihood estimator (NPMLE) where we will maximize the likelihood function of the uniform deconvolution problem over the nonparametric family of distributions. It turns out that the current status problem and the uniform deconvolution problem both have the same likelihood, so we can use the setting and properties of either of those problems to calculate the NPMLE. This is needed as there is not a general method of finding the NPMLE, we will have to use the property

that in the current status problem, the time of the tests can be ordered chronologically without loss of generality. With this we can use the method of isotonic regression to find the NPMLE.

The other method that will be treated in this thesis is kernel density estimation (KDE). This method builds a density function directly on the data. The idea is that whenever there are a lot of data near a value x, this will lead to a high density of the KDE at x, and vice versa. The KDE does this using a known, symmetric kernel density K. We add n of those kernel densities, each with its center on one of the n observations. Also a bandwidth h is chosen that influences the height and the width of the peaks of the kernel densities. This will influence how detailed and how smooth the KDE will be in the end. The kernel densities will be added and normalized to obtain the KDE.

These methods will be tested in Chapter 6 and we will compare their strengths and weaknesses. Finally, we will discuss what further research is needed in this field and conclusions are drawn.

2

Uniform Deconvolution

In this chapter we will discuss the uniform deconvolution problem. We start by giving some mathematical definitions about convolution in general, then we will focus on the uniform convolution case and finally we will discuss the uniform deconvolution problem. Some derivations will be made such that we can estimate the unknown distribution F_0 based on the noisy data.

2.1. Convolution

Before we can define deconvolution, we will first have to look at what convolution is. Convolution of two functions f and g in general is

$$(f * g)(t) = \int_{-\infty}^{\infty} f(s)g(t-s)ds.$$

This may at first seem like some arbitrary formula, however it is a formula that arises quite naturally in probability theory.

Let X and Y be two random variables with continuous distribution functions F_X and F_Y respectively. Assume the density functions exist and are continuous. They will be denoted by f_X and f_Y. The sum of those two random variables is a new random variable $Z = X + Y$. In this case we only observe instances of Z and thus we want to relate the density f_Z of Z to the distributions of X and Y. This goes as follows.

First we note that the density of Z can be written as

$$f_Z(z) = \frac{d}{dz}\mathbb{P}(Z \leq z)$$

Now we will first show some results of $\mathbb{P}(Z \leq z)$ and return to the density later.

We will use the law of total probability on $\mathbb{P}(Z \leq z)$ and this yields:

$$\begin{aligned}\mathbb{P}(Z \leq z) &= \mathbb{P}(X + Y \leq z) \\ &= \int_{\mathbb{R}} \mathbb{P}(X + Y \leq z | X = x) dF_X(x) \text{ using the partition theorem} \\ &= \int_{\mathbb{R}} \mathbb{P}(x + Y \leq z) dF_X(x) = \int_{\mathbb{R}} \mathbb{P}(Y \leq z - x) dF_X(x) \\ &= \int_{\mathbb{R}} F_Y(z - x) dF_X(x).\end{aligned}$$

Now we return to the density function f_Z.

$$\begin{aligned}f_Z(z) = \frac{d}{dz}\mathbb{P}(Z \leq z) = \frac{d}{dz}\int_{\mathbb{R}} F_Y(z - x) dF_X(x) \text{ By Leibniz rule, under the regularity conditions} \\ = \int_{\mathbb{R}} \frac{d}{dz} F_Y(z - x) dF_X(x) = \int_{\mathbb{R}} f_Y(z - x) dF_X(x)\end{aligned}$$

We therefore get the expression

$$f_Z(z) = \int_{\mathbb{R}} f_Y(z - x) dF_X(x). \tag{2.1}$$

Thus, we have expressed f_Z in terms of f_Y and F_X. As you can see, the formula 2.1 we obtained is the *convolution* of f_X and f_Y and we can write it as $f_Z = (f_X * f_Y)$.

2.1.1. Uniform convolution

Now we will look at the specific case of convolution that we will treat in this report called *Uniform convolution*. In this case, one of the random variables has a Uniform density on $[0, 1]$. So we are in the following setting.

Consider a random variable X, distributed according to a (unknown) distribution function F_0 on $\mathbb{R}_+$. Instead of observing X, we observe the sum $Z = X + U$, where $U \sim \text{Unif}([0, 1])$, independent of X. The random variable Z then has convolution density

$$\begin{aligned}g_Z(z) &= \int_{[0,\infty)} f_U(z - x) dF_0(x) \\ &= \int_{[0,\infty)} \mathbb{1}_{[0,1]}(z - x)\, dF_0(x) = \int_{[z-1,z]} dF_0(x) = F_0(z) - F_0(z - 1).\end{aligned}$$

Here we have a very simple relation between our unknown distribution F_0 and the density of our observations g_Z, namely

$$g_Z(z) = F_0(z) - F_0(z - 1). \tag{2.2}$$

2.2. Uniform deconvolution

We now have g_Z expressed in terms of F_0. However, we want to get an explicit formula for F_0 in terms of g_Z as this is the unknown function we want to estimate using observations of the density g_Z. For this we will use that $F_0(z) = 0, \forall z \leq 0$, we get the following derivation.

$$
\begin{aligned}
z \in [0,1] \Rightarrow g_Z(z) &= F_0(z), && \text{as } F_0(z-1) = 0 \text{ when } z-1 \leq 0 \\
z \in (1,2] \Rightarrow g_Z(z) &= F_0(z) - F_0(z-1) \\
\Rightarrow F_0(z) &= F_0(z-1) + g_Z(z) \\
&= g_Z(z-1) + g_Z(z), && \text{as } z-1 \in [0,1] \text{ so } F_0(z-1) = g_Z(z-1) \\
z \in (2,3] \Rightarrow g_Z(z) &= F_0(z) - F(z-1) \\
\Rightarrow F_0(z) &= F_0(z-1) + g_Z(z) \\
&= g_Z(z-2) + g_Z(z-1) + g_Z(z) \\
&\vdots \\
z \in (j, j+1] \Rightarrow F_0(z) &= \sum_{i=0}^{j} g_Z(z-i)
\end{aligned}
$$

Now we again use the fact that $F_0(z) = 0, \forall z \leq 0$ to get a formula on the whole of $[0, \infty)$.

If we take $j \in \mathbb{N}$ arbitrary, we get for $\forall z \in (j, j+1]$ and $\forall i$ s.t. $i > j$ that $g_Z(z-i) = 0$ as $z - i \leq 0$. So as i becomes larger than j, we will only add zero terms. Therefore we get that

$$
\begin{aligned}
&\forall z \in (j, j+1] \Rightarrow F_0(z) = \sum_{i=0}^{j} g_Z(z-i) = \sum_{i=0}^{j+1} g_Z(z-i) = \sum_{i=0}^{j+2} g_Z(z-i) = \dots \\
&\Rightarrow F_0(z) = \sum_{i=0}^{\infty} g_Z(z-i), \forall z \in [j, j+1)
\end{aligned}
$$

Since j was arbitrary, we get that this formula holds for all j, and therefore also for all $z \in [0, \infty)$. We therefore obtain the result

$$
F_0(z) = \sum_{i=0}^{\infty} g_Z(z-i), \ \forall z \in [0, \infty). \tag{2.3}
$$

With the obtained expressions, we will now propose estimators for the unknown distribution F_0 of X using nonparametric estimation.

We will do this using two different methods of nonparametric estimation. Namely, in Chapter 5 using kernel density estimation and in Chapter 4 using a nonparametric maximum likelihood estimation.

3

Current status

In this section we introduce the current status problem. The current status problem is a problem that has an important medical application and, in a specific case, it has a close relation to the uniform deconvolution problem. The problem comes from survival analysis and is a formulation of a problem we have when testing people on diseases. Namely, when testing patients on a disease over time, we only know the status of that patient at the time of the test, i.e. the current status. This means that, for example, when a patient is tested positive on a disease, we do not know when the patient got infected, only that the infection took place before this inspection time. If we test a lot of patients, we can get a better idea of the time of onset of the disease, but we still have an uncertainty in every test as we only observe whether the patients are infected or not at the times of inspection.

We will treat the specific case of the problem where we only test each patient once and the time of inspection follows a uniform distribution. In addition, we will assume that every individual is guaranteed to get infected at $t = 1$. We choose to look at this specific case as it has a close relation to the uniform deconvolution problem. As it turns out, they have the same likelihood function, this will be needed in the estimation method discussed in Chapter 4 as here we will try to maximize this likelihood function. Due to the two problems having the same likelihood function, we can choose either of the two problems and its setting which will help us when finding the NPMLE. This is the reason why this problem is introduced in this report, it is a nice application, but it also helps us in our estimation in the uniform deconvolution problem when using the NPMLE.

In this chapter we will define the current status problem mathematically and derive some properties that we will use in solving the NPMLE in Chapter 4.

3.1. Defining the current status problem

As mentioned in the introduction, the only thing we observe is whether or not an individual i is infected at the testing time T_i which are i.i.d. The time of infection X_i of individual i is the value of interest and we aim to estimate its distribution function F_0 using the observations. The testing times T_i will be observed, they have a standard uniform distribution and are independent from X_i. We will define the binary variable Δ_i as the indicator of a person being infected at the time of the test T_i. So we get

$$\Delta_i = \begin{cases} 1, & \text{if } X_i \leq T_i \\ 0, & \text{otherwise.} \end{cases}$$

In the problem we assume that F_0 has support in $[0, 1]$. This means that we assume that every individual will be infected at some moment in the time interval $[0, 1]$, according to the distribution function

F_0. This assumption will simplify the problem such that it is easier to solve and it can be interpreted as if nothing is done to prevent the disease such that everyone is guaranteed to get infected at some point in time.

Thus, we observe the pairs $(T_1, \Delta_1), \ldots, (T_n, \Delta_n)$ where Δ_i is defined as above, X_i and T_i are independent random variables and $X_i \sim F_0$ and $T_i \sim Unif([0,1])$. Using the observations, we wish to estimate F_0.

3.2. Connection to the uniform deconvolution problem

We will now show the connection between the two problems. The problems are not entirely equivalent, however, there is a link between the two problems through the likelihood function. When rewriting the uniform deconvolution problem, we can derive a likelihood function which is the same as the likelihood function of the current status model. This is useful as we can use this function to estimate our unknown distribution F_0 for both problems. What the likelihood function exactly is an how we can use this to find an estimate for F_0, is explained in Chapter 4.

3.2.1. Likelihood of current status

In the current status setting, we observe $(T_1, \Delta_1), \ldots, (T_n, \Delta_n)$ with $\Delta_i = \mathbb{1}_{\{X_i \leq T_i\}}$ and $T_i \sim Unif([0,1])$.

We can construct a density function $f_{\Delta,T}$ w.r.t. the product measure of the counting measure on $\{0,1\}$ and the Lebesgue measure on $[0,1]$. It therefore is not a regular joint density function as it is based on a mix of a discrete and continuous measure. However, the following derivations still hold according to Groeneboom and Wellner, 1992.

$$\begin{aligned} f_{\Delta,T}(\delta_i, t_i) &= f_{\Delta|T}(\delta_i \,|t_i) f_T(t_i) \\ &= f_{\Delta|T}(\delta_i \,|t_i) = \mathbb{P}(\Delta_i = \delta_i \,|T_i = t_i) \end{aligned}$$

As $f_T(t_i) = 1$.

Now $\Delta_i = 1$ is equivalent to $X_i \leq T_i$, so:

$$\mathbb{P}_{F_0}(\Delta_i = 1 \,|T_i = t_i) = \mathbb{P}_{F_0}(X_i \leq t_i) = F_0(t_i)$$

And $\Delta_i = 0$ is equivalent to $X_i > T_i$, so:

$$\mathbb{P}_{F_0}(\Delta_i = 0 \,|T_i = t_i) = \mathbb{P}_{F_0}(X_i > t_i) = 1 - F_0(t_i)$$

Note that $\Delta \,|T \sim Ber(F_0(T))$

We now can rewrite $f_{\Delta,T}$ as follows, for $\Delta_i \in \{0,1\}$ and $T_i \in [0,1]$.

$$\begin{aligned} f_{\Delta,T}(\delta_i, t_i) &= \mathbb{P}_{F_0}(\Delta_i = \delta_i \,|T_i = t_i) \\ &= \mathbb{P}_{F_0}(\Delta_i = 1 \,|T_i = t_i)^{\delta_i} \; \mathbb{P}_{F_0}(\Delta_i = 0 \,|T_i = t_i)^{1-\delta_i} = F_0(t_i)^{\delta_i} \; (1 - F_0(t_i))^{1-\delta_i}. \end{aligned}$$

Using independence, we can now construct the likelihood function.

$$\begin{aligned}\mathcal{L}_n((T_1,\Delta_1),\ldots,(T_n,\Delta_n);F_0) &= \prod_{i=1}^{n} f_{\Delta,T}(\Delta_i,T_i) = \\ &= \prod_{i=1}^{n} F_0(T_i)^{\Delta_i}(1-F_0(T_i))^{1-\Delta_i}\end{aligned}$$

and the corresponding log-likelihood

$$\ell_n(F_0) = \sum_{i=1}^{n} [\Delta_i \log(F_0(T_i)) + (1-\Delta_i)\log(1-F_0(T_i))] \tag{3.1}$$

In addition, we note that $\Delta_i \mid T_i \sim Ber(F_0(T_i))$ and $T_i \sim Unif([0,1])$.

Finally, we have obtained the log-likelihood of the current status problem. In Chapter 4 we will show that it is the same log-likelihood as the one of the uniform deconvolution problem and we will use this to find the 'nonparametric maximum likelihood estimator' of F_0 for both problems.

4

Nonparametric maximum likelihood estimation

In this chapter we will discuss a method of estimating a distribution function in a nonparametric way, named the nonparametric maximum likelihood estimator (NPMLE). As the name suggests, it works using a similar idea as the maximum likelihood estimator, however, now it is nonparametric.
Therefore the calculation of the NPMLE and the parametric MLE are not much alike, they only share the same concept of estimation.

First we will give a short recap of the MLE, then we will define the NPMLE and we will discuss the relation and the differences. Finally, the NPMLE will be calculated for the uniform deconvolution problem.

4.1. Defining the Nonparametric MLE

4.1.1. Parametric MLE

Maximum likelihood estimation is a well established concept in courses like Introduction to Statistics. But we will give a short recap of how the MLE works as the NPMLE is built on the same idea.
We consider the following situation in which we will apply the MLE.
Assume $X_1, \ldots, X_n$ are i.i.d. random variables having a probability density belonging to the parametric family of densities $\{f(\cdot;\theta) : \theta \in \Theta\}$ where $\Theta \subseteq \mathbb{R}^k$ is open. This means that all the random variables X_i are i.i.d. with a density $f(\cdot;\theta)$ for some fixed $\theta \in \Theta$. We assume the distribution to be known except for the parameter $\theta \in \Theta$. The parameter θ is the value of interest and we will try to estimate it with the MLE using the data $X_1, \ldots, X_n$ and the density $f(\cdot, \theta)$.

We first define the likelihood function by

$$\mathcal{L}_n(X_1, \ldots, X_n; \theta) := \prod_{i=1}^{n} f(X_i; \theta).$$

The idea behind this is, because all the X_i's are i.i.d., the likelihood of θ factorizes into the product of the marginal densities. So $f(X_1, \ldots, X_n; \theta) \overset{i.i.d.}{=} \prod_{i=1}^{n} f(X_i; \theta)$.

Now we will maximize this function over $\theta \in \Theta$ to get the maximum likelihood estimator $\hat{\theta}_n^{MLE}$ for θ. Since maximizing the likelihood over $\theta \in \Theta$ is equivalent to maximizing the log of the likelihood, this is

often easier to calculate so we introduce the log-likelihood

$$\ell_n(X_1, \ldots, X_n; \theta) = \sum_{i=1}^{n} \log(f(X_i; \theta)).$$

If this function is smooth enough in θ, then we can often maximize this easily by taking the derivative w.r.t. θ and set the expression equal to zero. Then solving for $\theta \in \Theta$ we obtain the parametric maximum likelihood estimator for θ.

4.1.2. The NPMLE

We now are going to look at the nonparametric case. The definition of the log-likelihood, however, stays the same. We get

$$\ell_n(X_1, \ldots, X_n; f) = \sum_{i=1}^{n} \log(f(X_i)). \tag{4.1}$$

But, we now do not want to maximize the log-likelihood over a parametrized family of densities but, for example, over a much larger family of functions like $\mathcal{F} := \{f : \mathbb{R} \to [0, \infty) \mid \int_{\mathbb{R}} f dx = 1\}$. This is a much larger family of possible densities in which we have to find the function $\hat{f}$ that maximizes the log-likelihood. How this is done is described in the next Section 4.3 using the method of isotonic regression.

4.2. Likelihood of uniform deconvolution

This section will be dedicated to obtaining the log-likelihood for the uniform deconvolution problem. When its formula is found, we can start to obtain its NPMLE. To obtain the NPMLE, we have to use certain properties of the uniform deconvolution problem as the method of calculating the NPMLE depends very strongly on the conditions of the density/distribution function over which the likelihood is maximized.

We will focus on the case where $F_0(1) = 1$. This has as a consequence that g_Z has support in $[0, 2]$ as X and U both have support in $[0, 1]$ and $Z = X + U$. Using the equation 2.2, we get that $g_Z(z) = F_0(z)$ for $z \in [0, 1]$ and $g_Z(z) = 1 - F_0(z - 1)$ for $z \in (1, 2]$.

However, we first rewrite the problem a bit such that the likelihood will coincide with the likelihood of the current status problem. The used way of rewriting is given by Groeneboom and Jongbloed, 2025 and is shown to yield the results that we desire. We rewrite this way because when the likelihood functions are the same, we can solve the NPMLE with properties from the current status problem and still obtain the NPMLE for both the uniform deconvolution problem and the current status problem. To get this rewritten version of the uniform deconvolution problem, we introduce the following variable.

Given the observations $Z_1, \ldots, Z_n$, define

$$D_i = \mathbb{1}_{\{Z_i \leq 1\}}$$

and with this, define Y_i such that

$$Y_i = \begin{cases} Z_i, & \text{if } D_i = 1 \\ Z_i - 1, & \text{if } D_i = 0. \end{cases}$$

We now have translated Z_i into the pairs (Y_i, D_i). Since $Z_i \in [0,2]$, we have that $Y_i \in [0,1]$ and therefore we get for a single observation (like $i = 1$) that

$$\begin{aligned}\mathbb{P}(Y \leq y) &= \mathbb{P}(Y \leq y, D = 0) + \mathbb{P}(Y \leq y, D = 1) \\ &= \mathbb{P}(Z - 1 \leq y, D = 0) + \mathbb{P}(Z \leq y, D = 1) = \mathbb{P}(Z \leq y + 1, Z > 1) + \mathbb{P}(Z \leq y) \\ &= \int_1^{y+1} g_Z(z)dz + \int_0^y g_Z(z)dz.\end{aligned}$$

Now using 2.2, the fact that $F_0(z+1) = 1$ whenever $z > 0$ (since $F_0(1) = 1$) and that $F_0(z-1) = 0$ for $z \leq 1$, we get that

$$\begin{aligned}&\int_1^{y+1} g_Z(z)dz + \int_0^y g_Z(z)dz = \int_0^y g_Z(z+1) + g_z(z)dz \\ &= \int_0^y F_0(z+1) - F_0(z) + F_0(z) - F_0(z-1)dz = \int_0^y 1dz = y\end{aligned}$$

So now we see that $\mathbb{P}(Y \leq y) = y$ and thus $Y \sim Unif([0,1])$.
Notice that in the specific case of the current status model from Chapter 3, we had pairs (T_i, Δ_i) with $T_i \sim Unif(0,1)$. Now we have pairs (Y_i, D_i) where $Y_i \sim Unif([0,1])$, so we already see a similarity between the two problems. However, when we look at the definitions of Δ_i and D_i, they are both indicators but of different quantities. This means that we can not yet relate (T_i, Δ_i) to (Y_i, D_i) directly. However, we will show that there is a similarity between the two variables.

First we notice that the conditional distribution of D given $Y = y$ is Bernoulli with success parameter

$$\mathbb{P}(D = 1|Y = y) = \frac{g_Z(y)}{g_Z(y) + g_Z(y+1)} \overset{2.2}{=} \frac{F_0(y) - F_0(y-1)}{F_0(y) - F_0(y-1) + F_0(y+1) - F_0(y)} = F_0(y)$$

Where we again use that $F_0(z+1) = 1$ whenever $z \geq 0$ and that $F_0(z-1) = 0$ for $z \leq 1$. Therefore, we have shown that $D|Y = y \sim Ber(F_0(y))$. Recall that in 3.1, we saw $\Delta|T = t \sim Ber(F_0(t))$, so also here we see that $D\ |Y$ and $\Delta\ |T$ have the same distribution.

Now lastly, we have to derive the formula for the log-likelihood in terms of (Y, D). This will be done in a similar fashion to the derivation is subsection 3.2.1.

We observe $(Y_1, D_1), \ldots, (Y_n, D_n)$ with $D|Y = y \sim Ber(F_0(y))$ and Y uniform.

With this we get the following derivation.

$$f_{D,Y}(d, y) = f_{D|Y}(d\ |y) f_Y(y) = \mathbb{P}(D = d|Y = y) = F_0(y)^d\ (1 - F_0(y))^{1-d}$$

We therefore get the likelihood function

$$\mathcal{L}_n((Y_1, D_1), \ldots, (Y_n, D_n); F_0) = \prod_{i=1}^n f_{D,Y}(D_i, Y_i) = \prod_{i=1}^n F_0(Y_i)^{D_i}(1 - F_0(Y_i))^{1-D_i}$$

and the corresponding log-likelihood

$$\ell_n(F_0) = \sum_{i=1}^n \left[D_i \log(F_0(Y_i)) + (1 - D_i)\log(1 - F_0(Y_i))\right] \tag{4.2}$$

We now see that the log-likelihood function for the uniform deconvolution problem is equivalent to the log-likelihood of the current status problem 3.1. This means that when we find the NPMLE of either of the two problems, we also find it for the other one, thus solving the two problems simultaneously. In addition, for computational matters, we can without loss of generality relabel the observations from the current status problem in such a way that $T_1 \leq \cdots \leq T_n$. Using this, we will be able to calculate the NPMLE and still obtain an estimation for the uniform deconvolution problem where we do not necessarily have such a property. This property will allow us to use the method of isotonic regression to maximize the log-likelihood.

In the next section, the NPMLE will be obtained using this method of isotonic regression.

4.3. Isotonic regression

This section will describe the method we will use to obtain the function $\hat{F}$ that will maximize the log-likelihood 4.2 over all function F in the family of distributions
$\mathcal{F} = \{F\,|F(0) = 0, F(1) = 1$ and F non-decreasing$\}$. We will do this using a method called isotonic regression.

First, in subsection 4.3.1, it will be explained how the method of isotonic regression works and later, in subsection 4.3.2, it will be explained how it can be applied to compute the NPMLE in the current status problem.

4.3.1. Method of isotonic regression

The method of isotonic regression works as follows.

Consider the set $\mathcal{C}_n = \{\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n | x_1 \leq x_2 \leq \cdots \leq x_n\}$. This is a *convex cone* in $\mathbb{R}^n$ meaning that for any two vectors in $\mathcal{C}_\mathrm{n}$, all positive linear combinations of those vectors are also in $\mathcal{C}_n$. Isotonic regression is the projection of a vector $\mathbf{y} \in \mathbb{R}^n$ onto this cone. This is a vector $\hat{\mathbf{y}} \in \mathcal{C}_n$ which minimizes the quadratic distance to $\mathbf{y}$ over all elements of the cone. Define, as a function of $\mathbf{x} \in \mathcal{C}_\mathrm{n}$

$Q(\mathbf{x}) = ||\mathbf{x} - \mathbf{y}||_2^2$

Where $||\mathbf{x}||_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$.

Then the projection $\hat{\mathbf{y}}$ of $\mathbf{y}$ on $\mathcal{C}_\mathrm{n}$ can be defined as

$$\hat{\mathbf{y}} := \underset{\mathbf{x} \in \mathcal{C}_\mathrm{n}}{\arg\min}\, Q(\mathbf{x}).$$

If this value $\hat{\mathbf{y}}$ minimizes the distance from $\mathbf{y}$ to $\mathcal{C}_\mathrm{n}$, then the directional derivative of Q in all allowed directions within $\mathcal{C}_\mathrm{n}$ must be nonnegative. We obtain that $\forall \mathbf{x} \in \mathcal{C}_\mathrm{n}$,

$$\lim_{\epsilon \downarrow 0} \frac{Q(\hat{\mathbf{y}} + \epsilon(\mathbf{x} - \hat{\mathbf{y}})) - Q(\hat{\mathbf{y}})}{\epsilon} \geq 0.$$

From this we get the following (in)equalities that are proven to be equivalent in Lemma 2.1 in Groeneboom and Jongbloed, 2014.

$$\sum_{j=1}^{i} \hat{y}_j = \begin{cases} = \sum_{j=1}^{i} y_j, \text{ if } i \in I \\ \leq \sum_{j=1}^{i} y_j, \text{ if } i \notin I. \end{cases}$$

Where I is the set of 'jump locations' of $\hat{\mathbf{y}}$

$$I = \{1 \leq i \leq n \ | \hat{y}_{i+1} > \hat{y}_i\}.$$

These (in)equalities can be used to construct the projection $\hat{\mathbf{y}}$. To do this, we make a cumulative sum (cusum) diagram. This is a collection of $n+1$ points in $\mathbb{R}^2$, $P_0 = (0,0)$ and

$$P_i = \left(i, \ \sum_{j=1}^{i} y_j \right), 1 \leq i \leq n.$$

We now construct the maximum convex minorant of the point diagram. This is a convex function that ensures that all the points of the cusum diagram are either on the function or above it. Such a functions will have as a property that for every i, the value of the left derivative of the function at i will satisfy the inequalities in 4.3.1. Therefore obtaining the desired projection $\hat{\mathbf{y}}$

To clarify we will give an example of a case and show graphically what is explained above.

Let

$$\mathbf{y} = (-0.210 \ ; -0.135 \ ; -0.072 \ ; 0.041 \ ; 0.036 \ ; 0.080 \ ; 0.075 \ ; 0.099 \ ; 0.084).$$

This vector is clearly not in $\mathcal{C}_n$. Therefore we make the cusum diagram

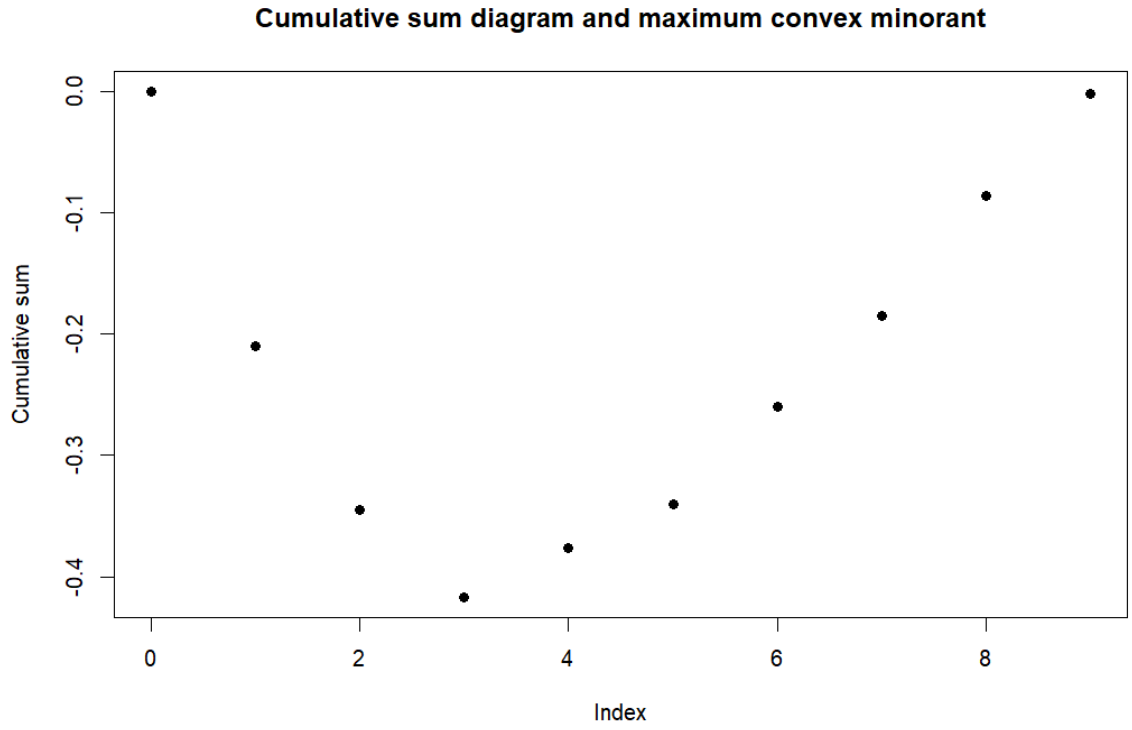


Figure 4.1: Cumulative sum diagram

Now we add the maximum convex minorant using the Pool Adjacent Violators Algorithm (PAVA) as described by Shivam, 2024. This algorithm ensures that the function is convex. It does this by taking the points in the cusum diagram that will result in a non-convex function, so the points which violate the convexity which are not in I. Those points will be replaced with new points that are exactly on a straight line from the previous point to the next point that does not need replacing. To give an intuition, the points in the cusum diagram can be seen as fixed nails, and the minorant as a tightly stretched string with endpoints at P_0 and P_n which is pulled below all the points in the cusum diagram.

This is shown in the following graph.

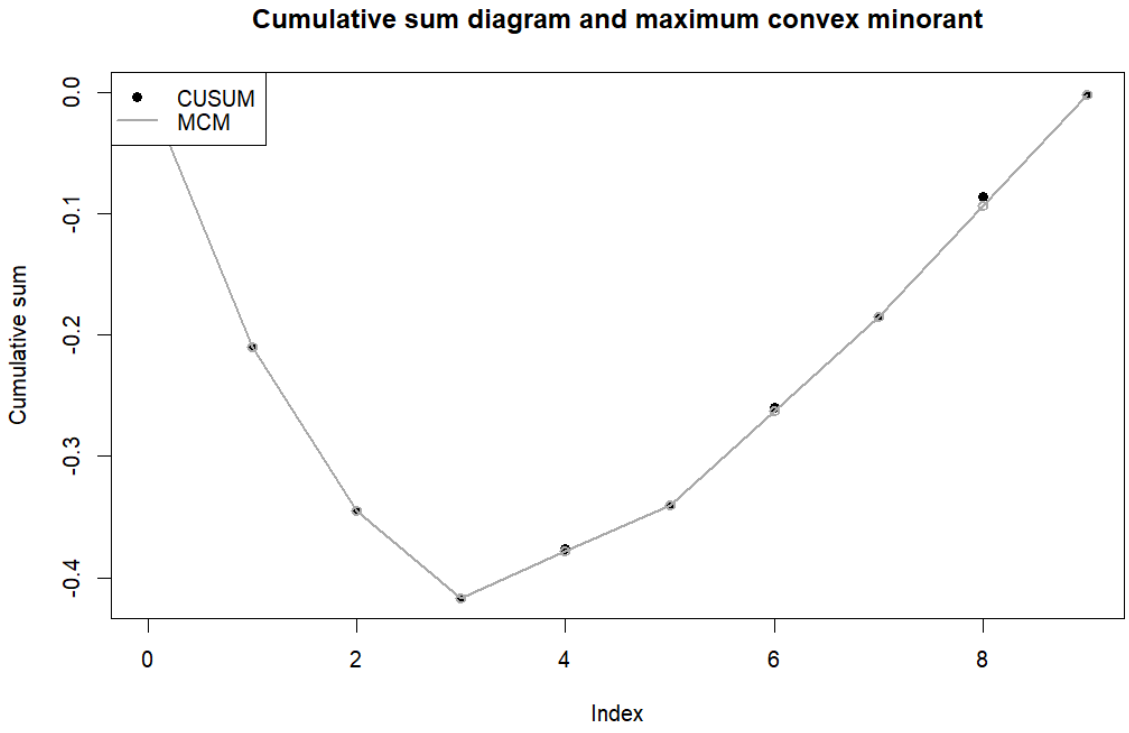


Figure 4.2: Cumulative sum diagram with maximum convex minorant

As can be seen, the convex function goes through the grey circles and those sometimes do not align with the cusum points. This is because these points would violate the convexity and are therefore not considered in the MCM. Like in the nail and stretched string analogy, these nails would not change the line of the string.

When we calculate the left-derivative of the convex minorant, we finally obtain the projection $\hat{y}$. In this example it would compute to
$\hat{\mathbf{y}}$ = (-0.2100 ; -0.1350 ; -0.0720 ; 0.0385 ; 0.0385 ; 0.0775 ; 0.0775 ; 0.0915 ; 0.0915).

4.3.2. Isotonic regression for current status

In the current status problem we relabeled the observed pairs (T_i, Δ_i) such that $T_1 \leq T_2 \leq \cdots \leq T_n$. We also want F_0 to be non-decreasing with $F_0(1) = 1$. So we want to maximize the likelihood 3.1 such that $F(T_1) \leq F(T_2) \leq \cdots \leq F(T_n)$ and $F(T_n) = 1$. According to Robertson et al., 1988, maximizing

the likelihood 3.1 is equivalent to minimizing the sum of squares

$$Q(F) = \sum_{i=1}^{n} (F(T_i) - \Delta_i)^2$$
$$\text{subject to } F(T_1) \leq F(T_2) \leq \cdots \leq F(T_n),\ F(T_n) = 1$$

This means that we can see the vector $(F(T_1), \ldots, F(T_n))$ as a member of the cone $\mathcal{C}_\text{n}$ and the vector$(\Delta_1, \ldots, \Delta_n)$ as the vector which we would like to project on the cone. We can therefore use the method of isotonic regression to compute $\hat{F}(T_i)$ which projects $(\Delta_1, \ldots, \Delta_n)$ on $\mathcal{C}_\text{n}$. When we found the projection, these $\hat{F}(T_i)$ will correspond to the distribution function, evaluated at the T_i's, which maximizes the log-likelihood. So, the step function which takes the value $\hat{F}(T_i)$ on $[T_{i-1}, T_i]$ will be the NPMLE. We will therefore search for points $F(T_i)$ on the cone such that it minimizes the distance to the values δ_i.

Now we construct the cusum diagram. Let the cumulative sums be

$$Y_0 = 0$$
$$Y_k = \sum_{i=1}^{k} \Delta_i$$

and the corresponding coordinates (k, Y_k) where $k = 0, 1, \ldots, n$. Now calculating the left-hand slopes of the line segments of the maximum convex minorant between $k-1$ and k gives the NPMLE $\hat{F}(T_k)$.

To visualize how this works precisely, we again give a simple example. Assume that we have 20 observations of the currents status problem where $X \sim Beta(\frac{1}{2}, \frac{1}{2})$ and $T \sim Unif([0, 1])$. We first make the cusum diagram where we plot the points $(k, \sum_{i=1}^{k} \Delta_i)$. We get the following graph.

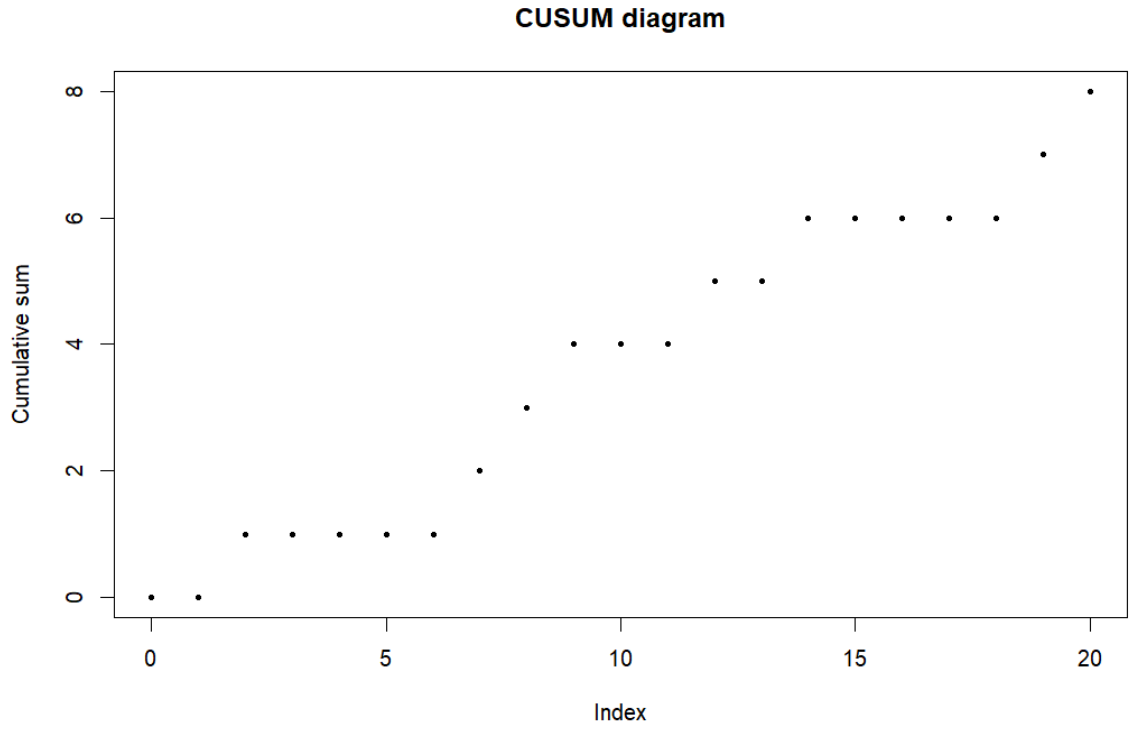


Figure 4.3: Cumulative sum diagram of δ_i

Now again using the PAVA algorithm, we compute the MCM. We get the following figure.

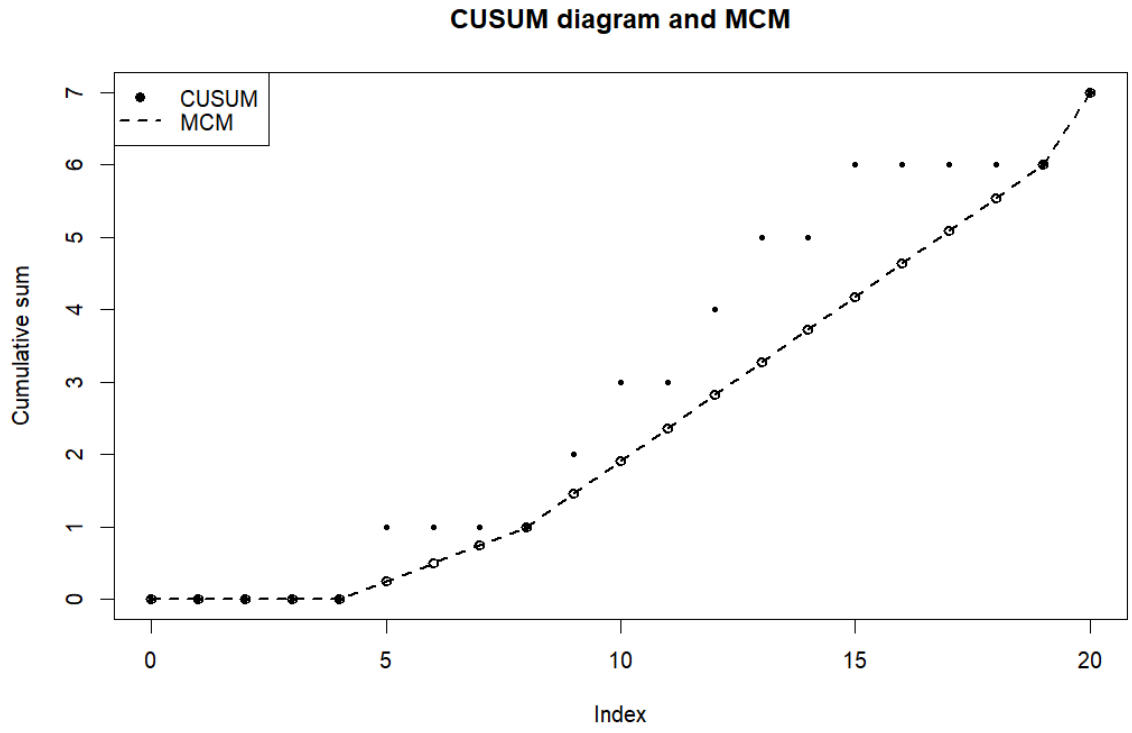


Figure 4.4: Cumulative sum diagram with maximum convex minorant

When now taking the left-derivative of the maximum convex minorant shown in the graph, we obtain the function $\hat{F}_0$ which minimizes the NPMLE. This yields

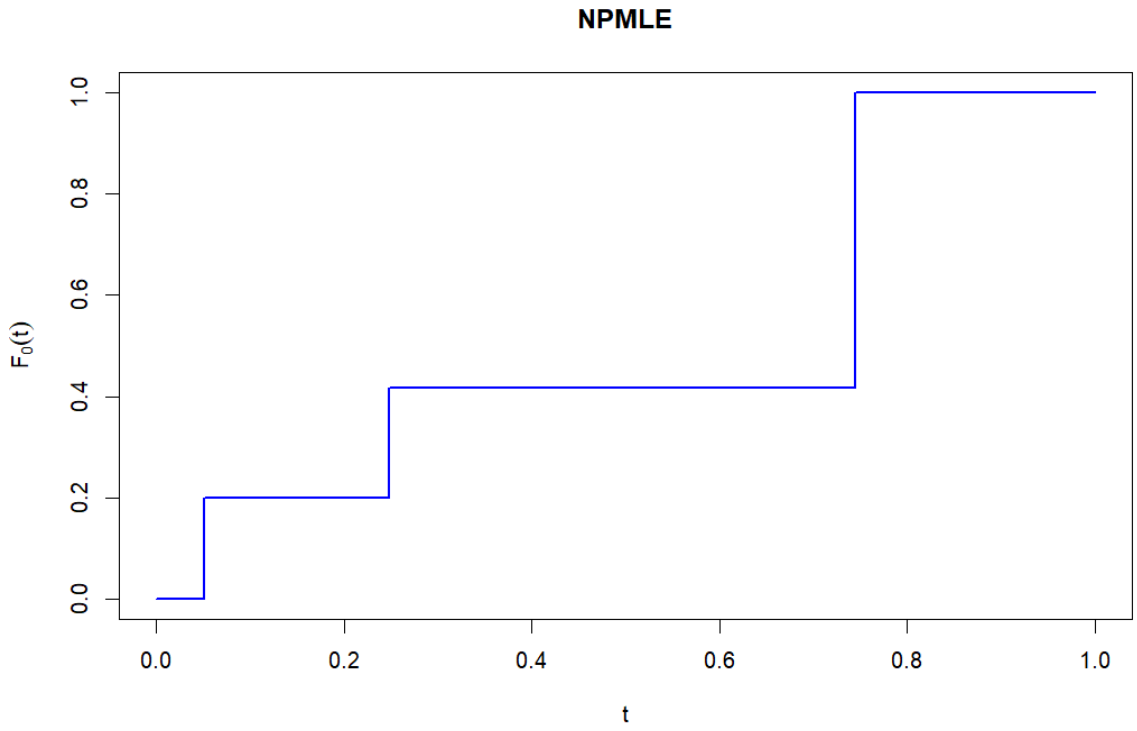


Figure 4.5: NPMLE

Now also adding the true distribution function F_0 as a reference to see how well the NPMLE has approximated F_0 gives us the figure.

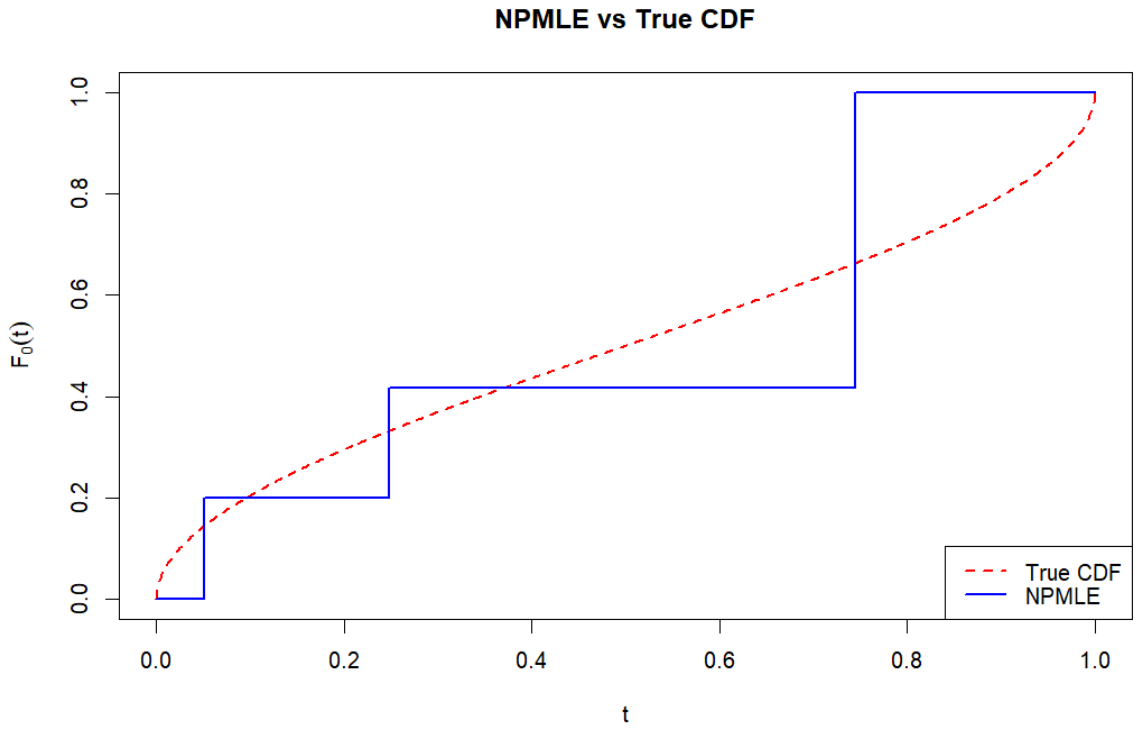


Figure 4.6: NPMLE and the true CDF

In this simulation, we used relatively few observations in order to visualize the MCM. However, the resulting estimation is not very accurate due to these few observations. When taking $n = 10.000$ observations, we obtain the following figure.

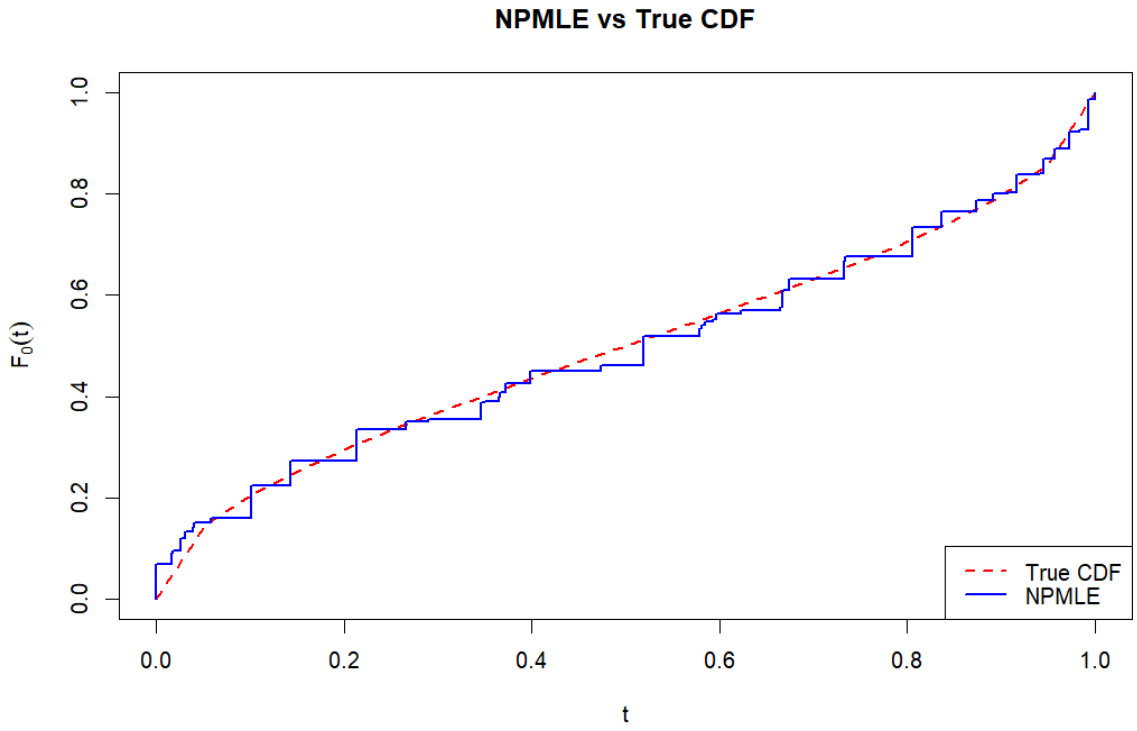


Figure 4.7: NPMLE and the true CDF with $n = 10.000$ observations

Here we see that the NPMLE estimates the desired distribution quite accurately.

We now have shown a way of calculating the nonparametric maximum likelihood estimator of the uniform deconvolution (or the current status) problem. In Chapter 6 we will test this method again using more observations and compare it to the *kernel density estimator* which is another method of nonparametric estimation which will be discussed in Chapter 5.

5

Kernel Density Estimation

This chapter provides an overview of the KDE, using expositions in Eric, 2025 and Akinshin, 2020.[1]

Kernel density estimation (KDE) is a nonparametric method to approximate an unknown density function. The KDE builds a density function based on finitely many observations. It assigns a density to every value x in the support according to how many observations are near that value. When there are a lot of observations near a value x, that value gets a high density $f(x)$ in the estimated density function. When there are a few or no observations near a value, that value gets a low density.

The KDE does this using two ingredients, a kernel K and a bandwidth h. The kernel K is a known symmetric density function centered at 0 and h is a parameter larger than 0 which can be chosen freely. When we observe n data points, n of those kernel densities will be added with their centers shifted onto each of the observed data points X_i. We then divide each kernel density by n. If we would choose a point x in the support, we can calculate the value of all of the n densities. If x is far away form a data point X_i, the value of the density centered at X_i would be low at x and, on the other hand, if x is close to a data point X_j, the value for the density centered at X_j of x would be high.
Adding all those kernels together results in a density function (we divided by n earlier such that the area of the function we now obtained is 1) which will have the property of returning high values in for inputs x that is near a lot of observations and low values for inputs x which are not close to much observations.

The bandwidth h can be used to influence the height and the width of each of the kernel densities. A low bandwidth will have very narrow but high peaks and a large bandwidth will have low and very wide peaks. A low bandwidth will result in that the data are very well represented, but it results in a very rough density, while a large bandwidth will preserve less details from the data but has a smoother result. The choice of bandwidth will boil down to a variance/bias trade-off when the mean squared error criterion is used as a measure of quality as we will see in Section 5.4.

The formula of the kernel density estimator is given by

$$\hat{g}_X(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right). \tag{5.1}$$

In this chapter we will first elaborate further on why the KDE is constructed this way, we will look at the best choice of bandwidth h and we will discuss at what happens when we use different kernel functions K.

[1]These sources were particularly useful for understanding the general framework.

5.1. Simple example of the KDE

In this section we will illustrate how the KDE works and why we use the formula 5.1. We do this with help from an example. In the example we use the following table as the observations that we received.

Observation	X_1	X_2	X_3	X_4	X_5	X_6
Value	-2.5	-1.9	0.1	3.7	4.8	5.6

We plotted the observations on the x-axis as follows.

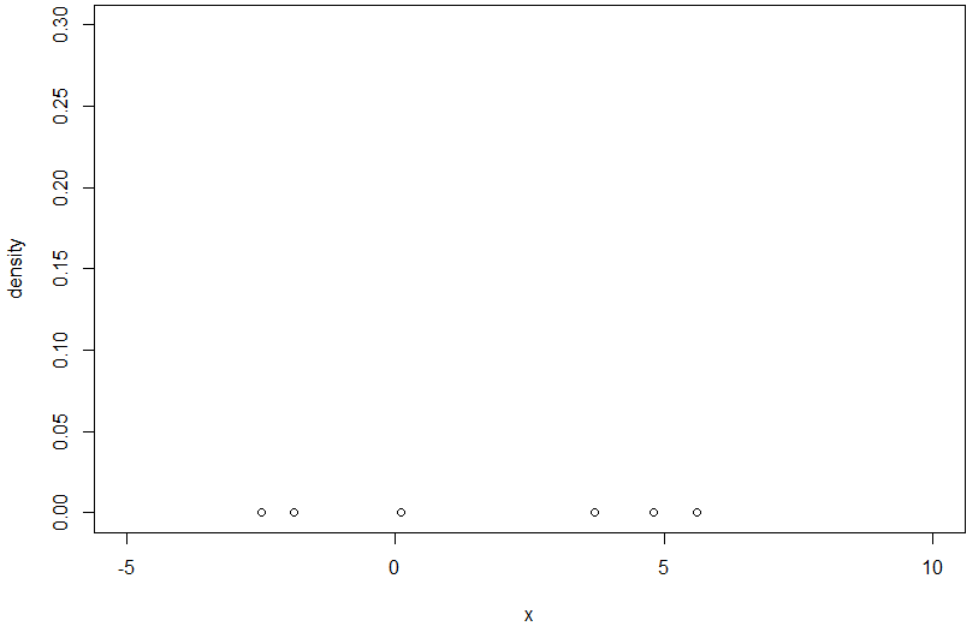


Figure 5.1: Observed data plotted on the X-axis

We are going to use a standard normal density as our kernel, so we use

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}.$$

This density is then centered on every X_i using the translation $K(x - X_i)$. We now get 6 distinct normal densities, one for every data point. For now, we let the bandwidth be $h = 0.5$. We get the formula $K\left(\frac{x-X_i}{h}\right)$. However, this division by h in the argument will have as a result that $K\left(\frac{x-X_i}{h}\right)$ does not have an area of 1 under its curve anymore. We namely know that $\int K(x)dx = 1$ but now have $K\left(\frac{x-X_i}{h}\right)$ which has another area under its curve. According the following calculation.

$$\int K\left(\frac{x - X_i}{h}\right) dx = \int hK(u)du = h$$

Therefore we divide the kernels K by h such that it will be a density.

Lastly, we also divide each density function by n (in our case 6), this will help us with the scaling later. Right now the choice of kernel and bandwidth are not important, what the effects of the bandwidth and the kernel are exactly is covered in the Section 5.2 and 5.3 respectively.

Adding the kernels, each with the formula $\frac{1}{nh}K\left(\frac{x-X_i}{h}\right)$ to the previous plot will result in the following figure.

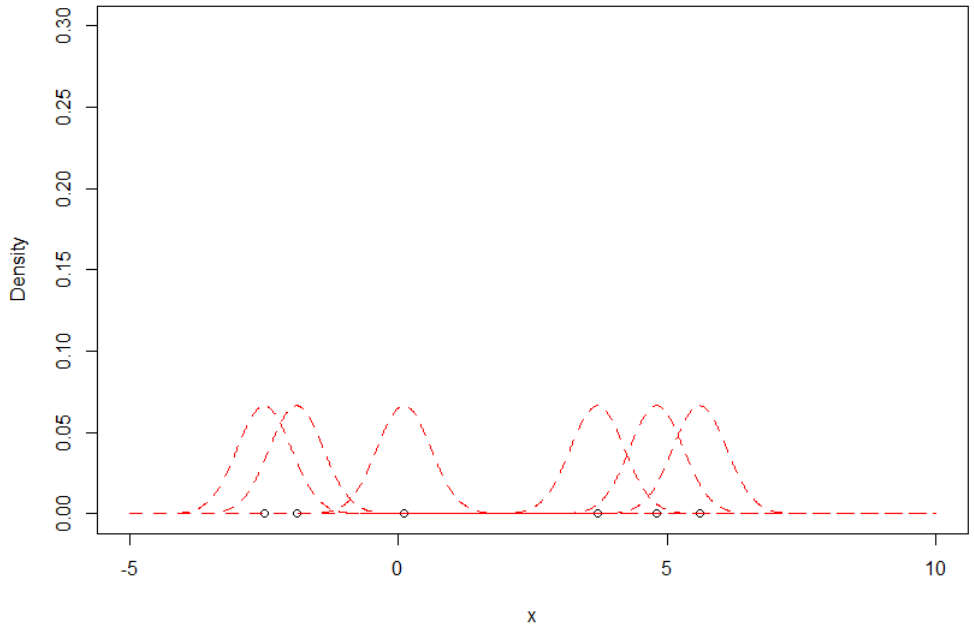


Figure 5.2: Standard normal kernels added

Now if we take some x in our domain, we can plug this value into each of the densities. If x is far away from an observation X_i, the density of the kernel which is centered at X_i will be low. If x is close to some observation X_j, the value in the density centered at X_j will be high.

Thus, now we can calculate the density of x w.r.t. each X_i, however, this is not yet what we want to know.

When adding all of those scaled kernel densities, we get high values for the points x that are close to relatively many observations as we get high values for all of the kernel densities corresponding to the nearby observations. We get low values for points x which are far away from any observations as they have low values for all the kernel densities. In the introduction to this chapter, we said that this is the kind of behavior that we want from our estimate. This results into the following formula.

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

This also clarifies why we scaled the kernel functions with n. This is done because we want our estimate to be a density function, and thus, we want the area under the curve to be 1. If we added all the kernel densities together and let this be our estimate, ten the area under our curve would be 6. With this, we finally have obtained the formula for the KDE, and when adding it to our plot, we get the following figure

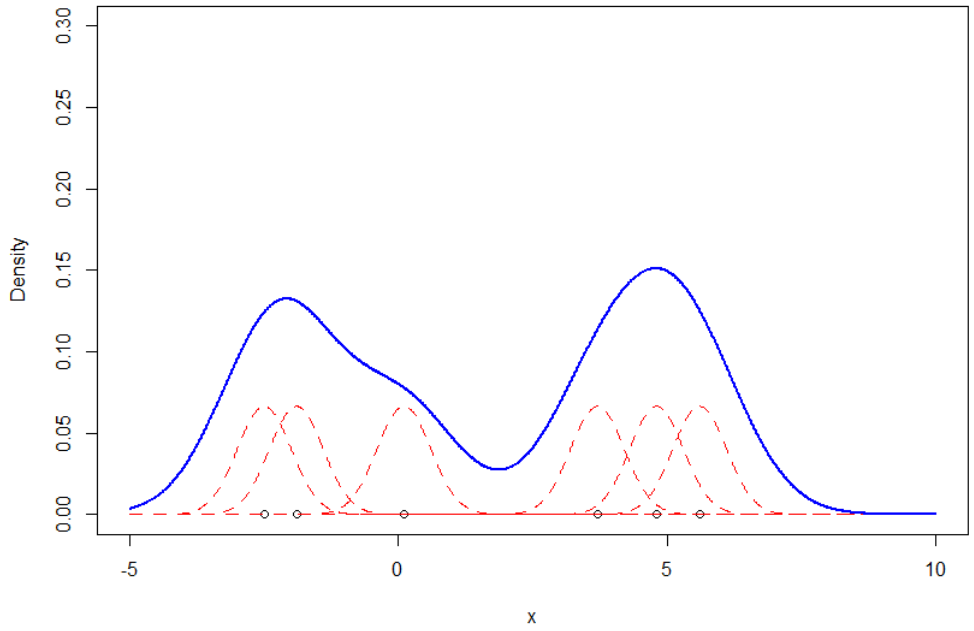


Figure 5.3: Sum of the kernels

We have demonstrated the idea of a KDE. In this, relatively simple, case we have taken the kernel to be the standard normal and the bandwidth $h = 0.5$. In the next Section 5.2 we will discuss differences in our results when using other bandwidths and in Section 5.3 we will dive deeper into consequences of using different kernels.

5.2. Different bandwidths

In this section we will explore the role of the bandwidth h in our KDE. The bandwidth, as said in the introduction of Chapter 5, can be used to influence how high the peaks of the kernel densities. To illustrate this, we use the example from Section 5.1 and make the KDE with different bandwidths.

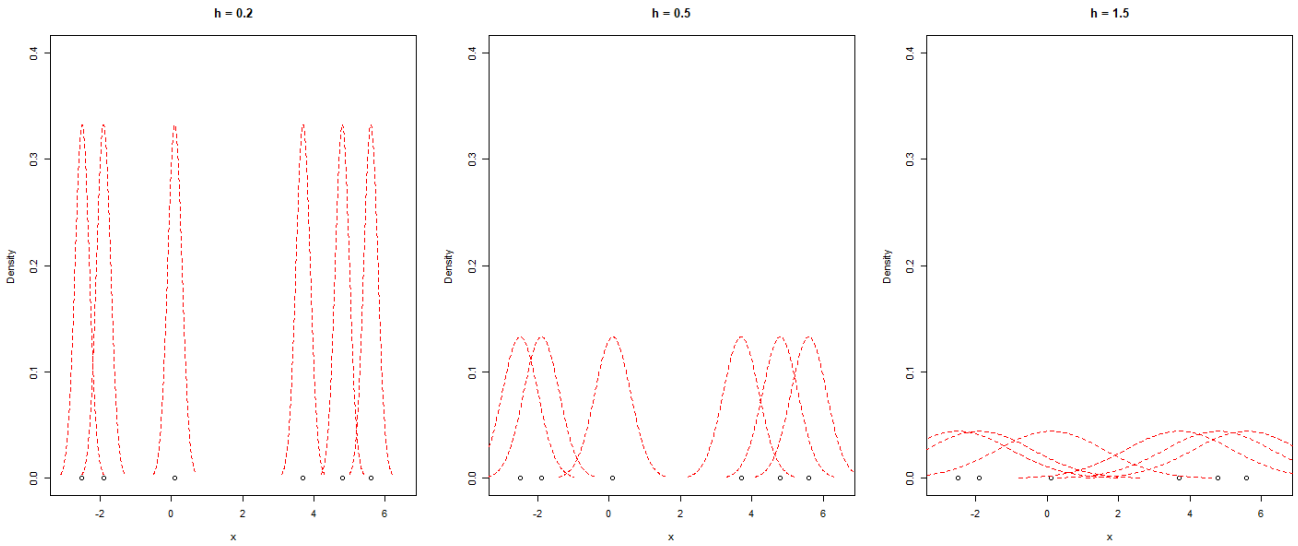


Figure 5.4: Kernels with different bandwidths

In the figures you see that a low bandwidth will have as a result that the kernels will be very narrow and have high peaks at the observed data points. This means that the points x which are close to observations X_i will get very high values for the densities and points x which are not near many X_i will get very low values for its density. There will be a larger contrast in the resulting KDE between the areas where there a lot of observations and where there are hardly any observations.

The advantage of a low h is that it captures the data in more detail, however, it may lead to overfitting and under smoothing, causing the KDE to exhibit high variance, spurious modes, and excessive local fluctuations not present in the true density.
On the other hand, a high bandwidth means that the kernel functions have very low peaks and have very heavy tails. In this case, our resulting KDE will be a very smooth function, but it is possible that the data is not represented well enough and changes in the data will barely be visible.

When calculating the KDE's in our case, we got the following figure.

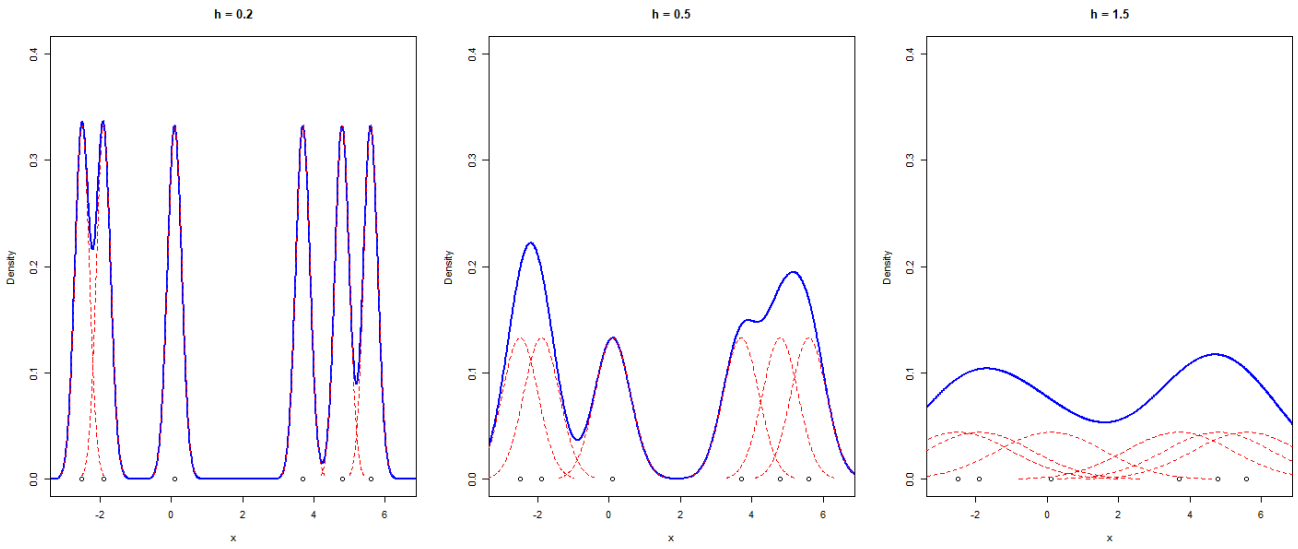


Figure 5.5: KDE's with different bandwidths

Now the question arises, *How to choose the bandwidth h?*. To answer this, we use the concept of the mean integrated squared error ($MISE$). This will be treated in Section 5.4 where we also will find the optimal choice of bandwidth. This turns out to be the equation 5.3. However, we first will look more closely at the role of the kernel.

5.3. Different kernels

The choice of the kernel has impact on the resulting estimation, in last sections we used a standard normal density, but there are many more density functions which can be used as the kernel. In this section we will look at what happens when we use different kernels.

In this report we will look at the next list of kernels. There are many more options, it even is possible to not use a density function, but we will limit us to the following cases.

	Kernel	Density
1	Uniform	$\frac{1}{2}\mathbb{1}\{\lvert x\rvert \leq 1\}$
2	Gaussian	$\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$
3	Epanechnikov	$\frac{3}{4}(1-x^2)\mathbb{1}\{\lvert x\rvert \leq 1\}$
4	Biweight	$\frac{15}{16}(1-x^2)^2\mathbb{1}\{\lvert x\rvert \leq 1\}$

Table 5.1: List of possible Kernels

Now we want to know how the different kernels behave in different settings and which kernels have which benefits. We will do this with a simulation study.

In our first simulation we will draw 100 random numbers from a standard normal distribution. We assume that the knowledge of the data being normal is unknown and we will try to estimate the distribution using KDE with the different kernels from Table 5.1. We will make 4 plots, one for each kernel and in each plot we will also give the graph of the standard normal in light-grey as a reference.

The choice of bandwidth will be explained in Section 5.4 and we will use the formula 5.3 for it. With this we get the following plots.

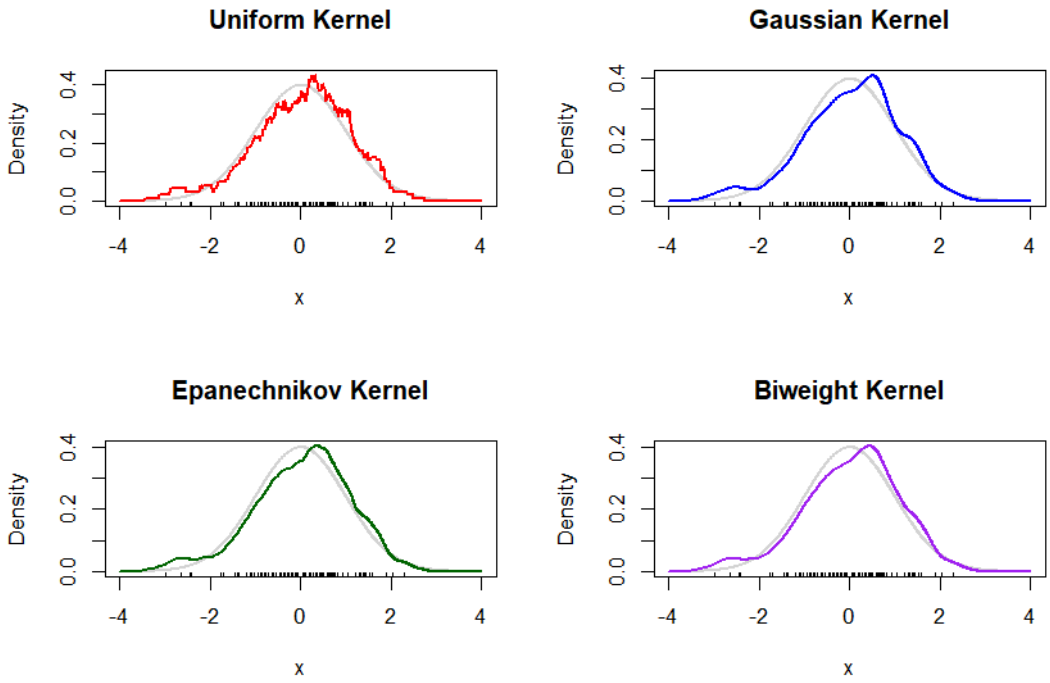


Figure 5.6: Estimation of a normal with different kernels

In the plots, we can see that there are some differences in estimations, especially in smoothness. It now seems like the Epanechnikov kernel gives the best approximation.

However, if we want to estimate a different density, for example a Gamma density with shape $\alpha = 2$ and scale $\theta = 2$, we do a second simulation where we draw $n = 250$ observations from a $\Gamma(2, 2)$ distribution. We get the following figure.

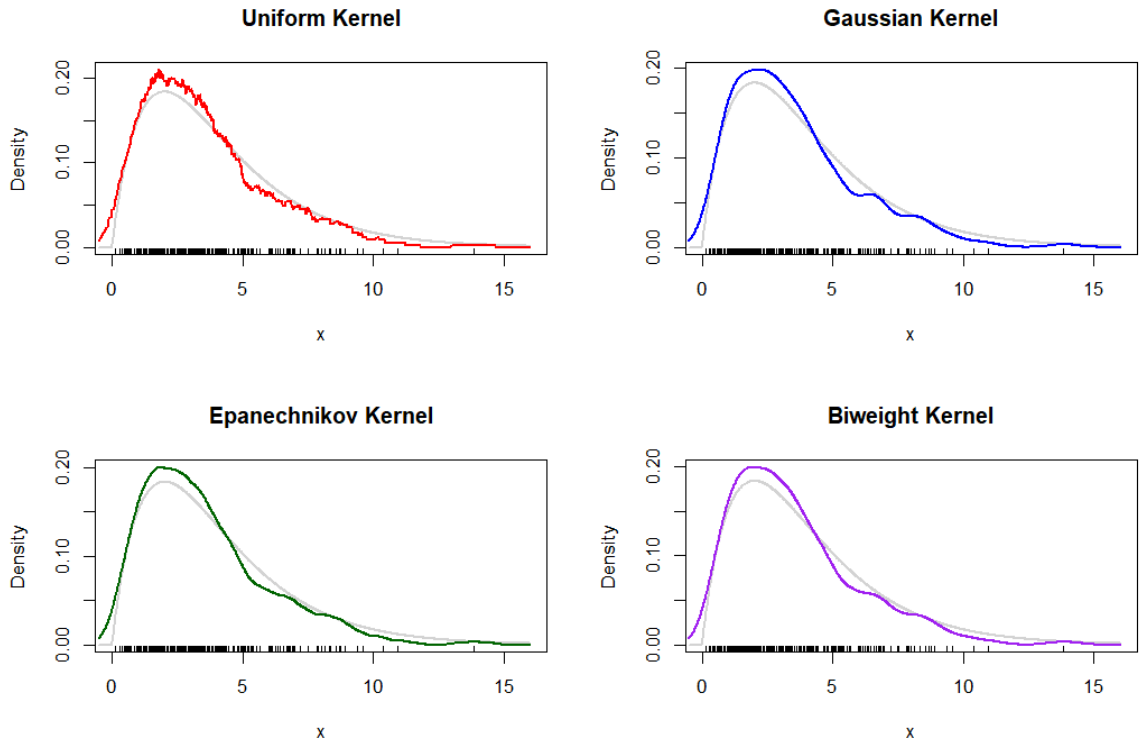


Figure 5.7: Estimation of a gamma with different kernels

Here it once again seems like the Epanechnikov kernel gives the best estimate, while the normal and biweight kernels does not differ by a lot and are more smooth.

Nevertheless, until this point we have only just 'looked' at how the different kernels and bandwidths perform graphically, but this is of course not mathematically rigorous. We therefore desire a more mathematical definition of a 'good' estimation. In the next section we will answer this question and we will determine which bandwidth is optimal.

5.4. Mean Integrated Squared Error and optimal bandwidth

In parametric estimation, the Mean Square Error (MSE) is often used as a measure of the quality of an estimator. The definition is as follows.

Assume we want to know the parameter θ and we use the estimator $\hat{\theta}$ to get an estimation. The MSE is given by

$$MSE_\theta(\hat{\theta}) = \mathbb{E}_\theta[(\hat{\theta} - \theta)^2] = \mathbb{V}ar_\theta(\hat{\theta}) + B_\theta(\hat{\theta})^2$$

Where $B_\theta(\hat{\theta}) = \mathbb{E}_\theta[\hat{\theta} - \theta]$ is the bias of $\hat{\theta}$.

This definition is still useful as with it we can calculate

$$MSE_f(\hat{f}(x)) = \mathbb{V}ar_f(\hat{f}(x)) + B_f(\hat{f}(x))^2$$

However, here we only calculate the MSE for a specific x, but we want to know an aggregated difference between our estimated density function and the true density function. So, here as well, we are dealing

with functions which can be seen as infinite dimensional parameter instead of finite dimensional parameters. Therefore we want to minimize the *integral* of the mean squared difference of those two functions. This gives rise to the definition of the Mean Integrated Squared Error (MISE), defined as follows.

$$\begin{aligned} MISE_f(\hat{f}) &= \mathbb{E}_f||\hat{f} - f||_2^2 = \mathbb{E}_f \int (\hat{f}(x) - f(x))^2 dx \\ &= \int \mathbb{E}_f[(\hat{f}(x) - f(x))^2]dx = \int MSE_f(\hat{f}(x))dx \\ &= \int \mathbb{V}ar_f(\hat{f}(x))dx + \int B_f(\hat{f}(x))^2 dx \end{aligned} \tag{5.2}$$

With the help of this equation, we can finally decide which kernel and which bandwidth are asymptotically optimal.
We cannot, of course, choose a kernel which minimizes the MISE in every case, but we can compute which kernel and which bandwidth will minimize the MISE asymptotically. When $n \to \infty$, $h \downarrow 0$ and $nh \to \infty$, asymptotically for the bias and variance we get the following approximations according to Hansen, 2009.

$$\begin{aligned} \int \mathbb{V}ar(\hat{f}(x))dx &\approx \frac{1}{nh} \int K^2(u)du \\ \int B(\hat{f}(x))^2 dx &\approx \frac{1}{4}h^4 \int u^2 K(u)du \int (f''(x))^2 dx \end{aligned}$$

Substituting this into the MISE gives us, for

$$\begin{aligned} MISE &\approx \frac{R(K)}{nh} + \frac{1}{4}h^4 \mu_2(K)^2 R(f''), \\ \text{Where } R(K) &= \int K^2(u)du, \\ \mu_2(K) &= \int u^2 K(u)du, \\ R(f'') &= \int (f''(x))^2 dx. \end{aligned}$$

This shows us that the choice of bandwidth asymptotically boils down to a variance/bias trade-off. If we take a large h, we get a large bias but a small variance and vice versa. This is consistent with the statement from before that a high bandwidth, the estimator will differ very little when changing between datasets, and the other way around.
In addition, this formula shows us how to obtain the asymptotically MISE optimal bandwidth

$$\begin{aligned} \frac{\partial}{\partial h} MISE(h) &= -\frac{R(K)}{nh^2} + \mu_2(K)^2 h^3 R(f'') = 0 \\ \Rightarrow h_{optimal} &= \frac{R(K)^{1/5}}{\mu_2(K)^{2/5} R(f'')^{1/5}} n^{-1/5}. \end{aligned} \tag{5.3}$$

In addition, the Epanechnikov kernel is considered to be the optimal kernel in terms of MISE. The derivation of this result will is discussed in Hansen, 2009. However, the differences in the accuracy when using different kernels (with appropriate bandwidths) are very small, therefore often other kernels than the Epanechnikov are still chosen as they are more smooth, or the normal is often used because it is a very well known density.

The only problem with calculating either the asymptotically optimal bandwidth or the MISE is that one needs to know the true density f, which is unknown. To estimate the optimal bandwidth there is

a variety of data-based methods, but this derivation gives the true asymptotically MISE-optimal h.

We now have presented a basic idea of the KDE. It is a nonparametric way of calculating an unknown density. There is still a lot to learn about the KDE, like how do you deal with densities with a discontinuity (like an exponential), or exactly how the KDE behaves asymptotically. But we have given a first intuition on how the KDE works and what the effects of the kernels and bandwidths are.

6

Simulation study

In the previous chapters, we introduced the uniform deconvolution problem and discussed two methods for addressing it. In this chapter, we simulate a concrete instance of the problem and compare the performance of the different estimation techniques.

In our simulation, we observe samples of $Z = X + U$, where $U \sim \text{Unif}([0, 1])$ and $X \sim \text{Beta}(2, 5)$. This corresponds to the setting where the distribution function F_0 has support on $[0, 1]$ and is non-decreasing, which we assume is known. Other than this, we do not make any assumption about the distribution. For example, we do not assume that the distribution belongs to the class of Beta distributions.

To this end, we apply both the nonparametric maximum likelihood estimator (NPMLE) and kernel density estimation (KDE) to recover the distribution function of X.

6.1. Estimation with NPMLE

As we have shown in Chapter 4, computing the NPMLE of the current status problem is equivalent to computing the NPMLE of the uniform deconvolution problem. Therefore, we consider the current status problem, as it has the convenient property that for a sample of observations $(T_1, \Delta_1), \ldots, (T_n, \Delta_n)$ we can order them such that $T_1 \leq \cdots \leq T_n$. This allows us to use isotonic regression to solve the NPMLE.

We construct a simulation in which we generate $(T_1, \Delta_1), \ldots, (T_n, \Delta_n)$ with $n = 1000$ observations. We draw the T_i's according to the standard uniform distribution and $\Delta = \mathbb{1}_{\{X \leq T\}}$ as in Chapter 3. The random variable X will follow the $Beta(2, 5)$ distribution of which we will estimate the distribution function. We can then directly apply the method of isotonic regression as discussed in Subsection 4.3.2, and the resulting NPMLE yields the following graph.

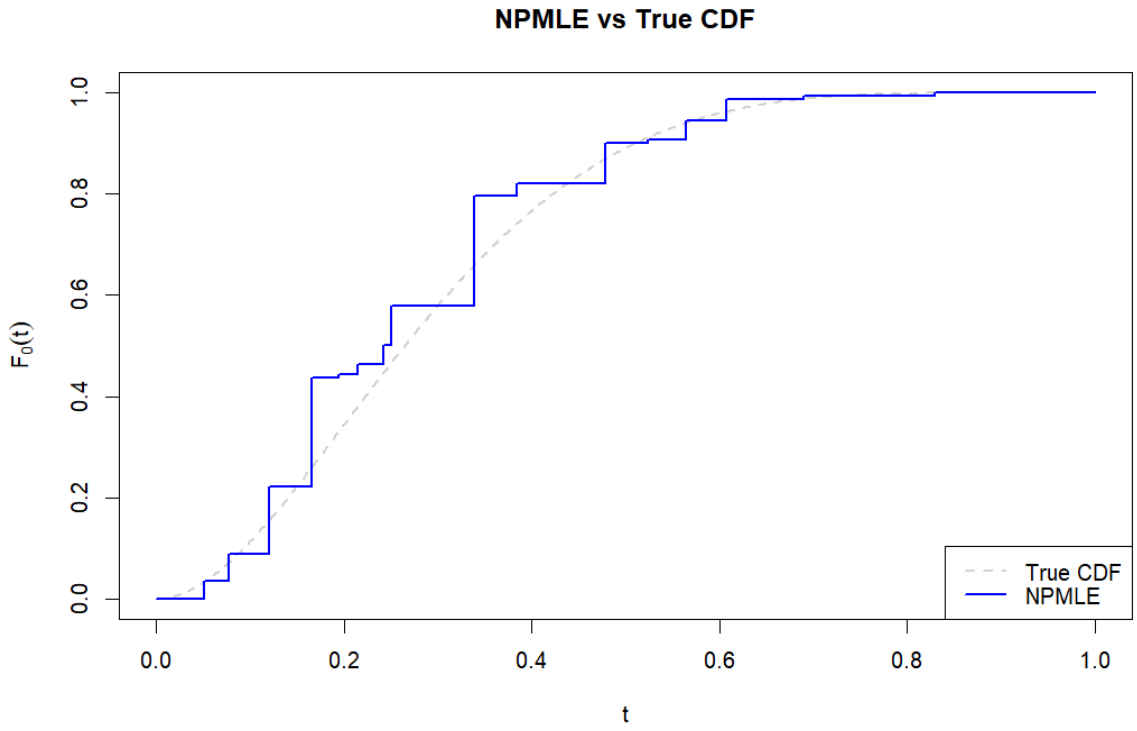


Figure 6.1: NPMLE

To measure the accuracy of the estimation, we calculate the integrated squared error (ISE). The integrated squared error has as its formula

$$\int_{\mathbb{R}} (\hat{f}(x) - f(x))^2 dx$$

Taking the expectation of this quantity will result in the mean integrated squared error as mentioned in Section 5.4.

For this graph, we can calculate the ISE, to get a mathematical rigorous notion of accuracy and it equals $3.317439 * 10^{-3}$.

6.2. Estimation with the KDE

In this section, we estimate the case described in the introduction of this chapter using the kernel density estimation (KDE) method as discussed in Chapter 5. This will be done in two steps. First, we use KDE to estimate the density function g_Z of the observed variable Z. In the second step, we use this estimate of g_Z together with the uniform deconvolution formula from Equation 2.3 to obtain an estimate for the distribution function F_0 of X.

To estimate g_Z, we apply KDE with the Epanechnikov density as the kernel. We select the optimal bandwidth as described in Equation 5.3, which in our case equals $h_{\text{optimal}} = 0.1673477$. It is important to note that in practice, the distribution of X is unknown, and therefore the optimal bandwidth h_{optimal} as defined cannot be computed directly. As discussed in Section 5.4, there exist data based methods that approximate this optimal bandwidth, but these will not be covered in this report. The focus of this report is on exploring and comparing different methods to solve the uniform deconvolution problem, and for this reason, we choose to work under idealized conditions using the theoretical optimal

bandwidth in our simulation.

The simulation has been implemented in `R`, and the resulting graph is shown below.

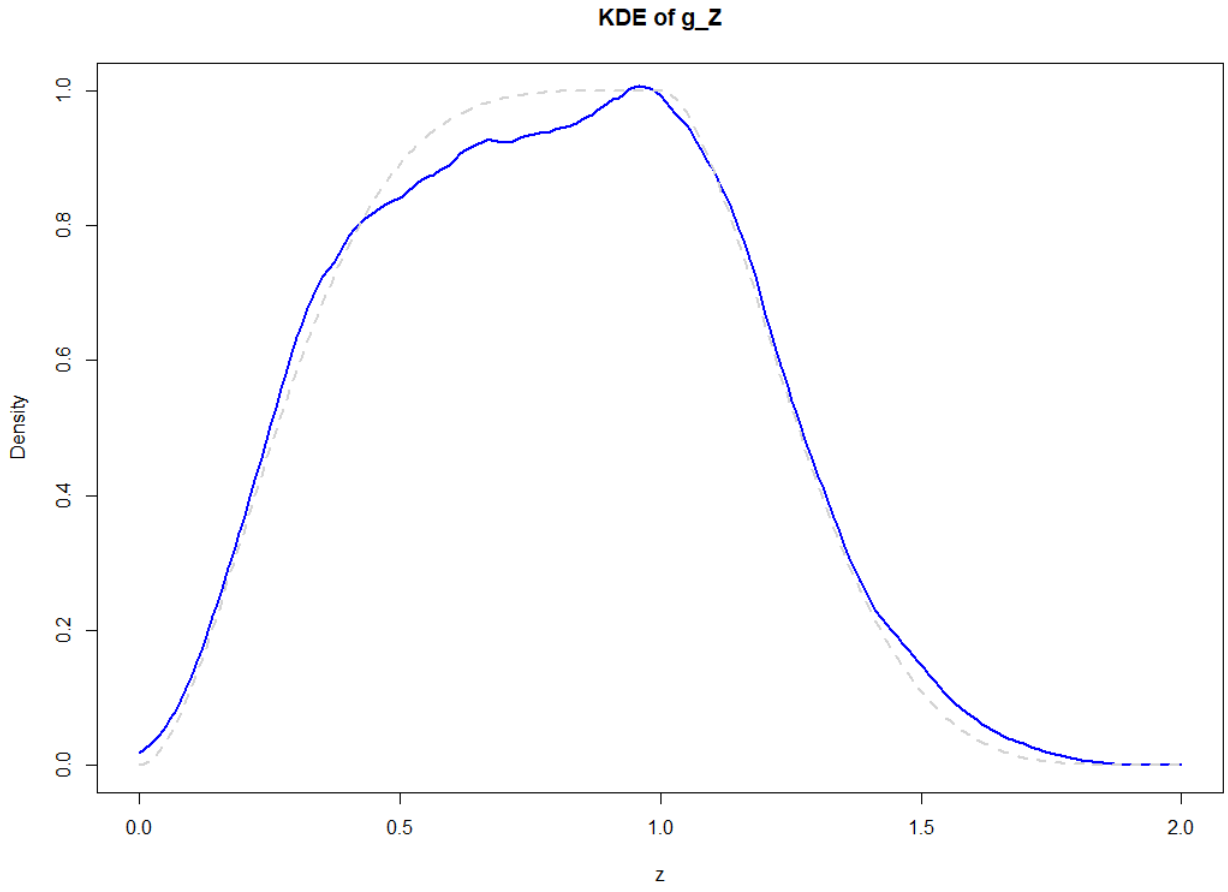


Figure 6.2: KDE of g_Z

The true density of g_Z is shown in light grey as a reference. Recall that g_Z is the convolution of the densities of X and U.

We now proceed to the second step of the estimation. Since we know that $F_0(1) = 1$, Equation 2.3 simplifies to

$$F_0(z) = g_Z(z) + g_Z(z-1), \tag{6.1}$$

as for $i \geq 2$ we have $g_Z(z-i) = 0$.

Furthermore, observe the following:

$$\forall z \in [0,1], \quad F_0(z) = g_Z(z), \qquad \text{since } g_Z(z-1) = 0 \text{ on } [0,1],$$

and for all $z \in (1,2]$,

$$\begin{aligned} g_Z(z) &= F_0(z) - F_0(z-1) && \text{by Equation 2.2} \\ &= 1 - F_0(z-1), && \text{since } F_0(z) = 1 \text{ for all } z \geq 1 \\ &= 1 - g_Z(z-1), && \text{because } g_Z(z-1) = F_0(z-1) \text{ on } [1,2], \end{aligned}$$

which implies:

$$\begin{aligned} & g_Z(z) + g_Z(z-1) = 1, \\ \Rightarrow \quad & F_0(z) = g_Z(z) + g_Z(z-1) = 1, \quad \forall z \in (1,2]. \end{aligned}$$

Thus, we conclude that $F_0(z) = g_Z(z)$ on $[0,1]$, consistent with the assumption $F_0(1) = 1$.

Finally, substituting our estimator $\hat{g}_Z$ for g_Z into Equation 6.1, we obtain an estimate $\hat{F}_0$ for F_0. The result is shown in the following figure.

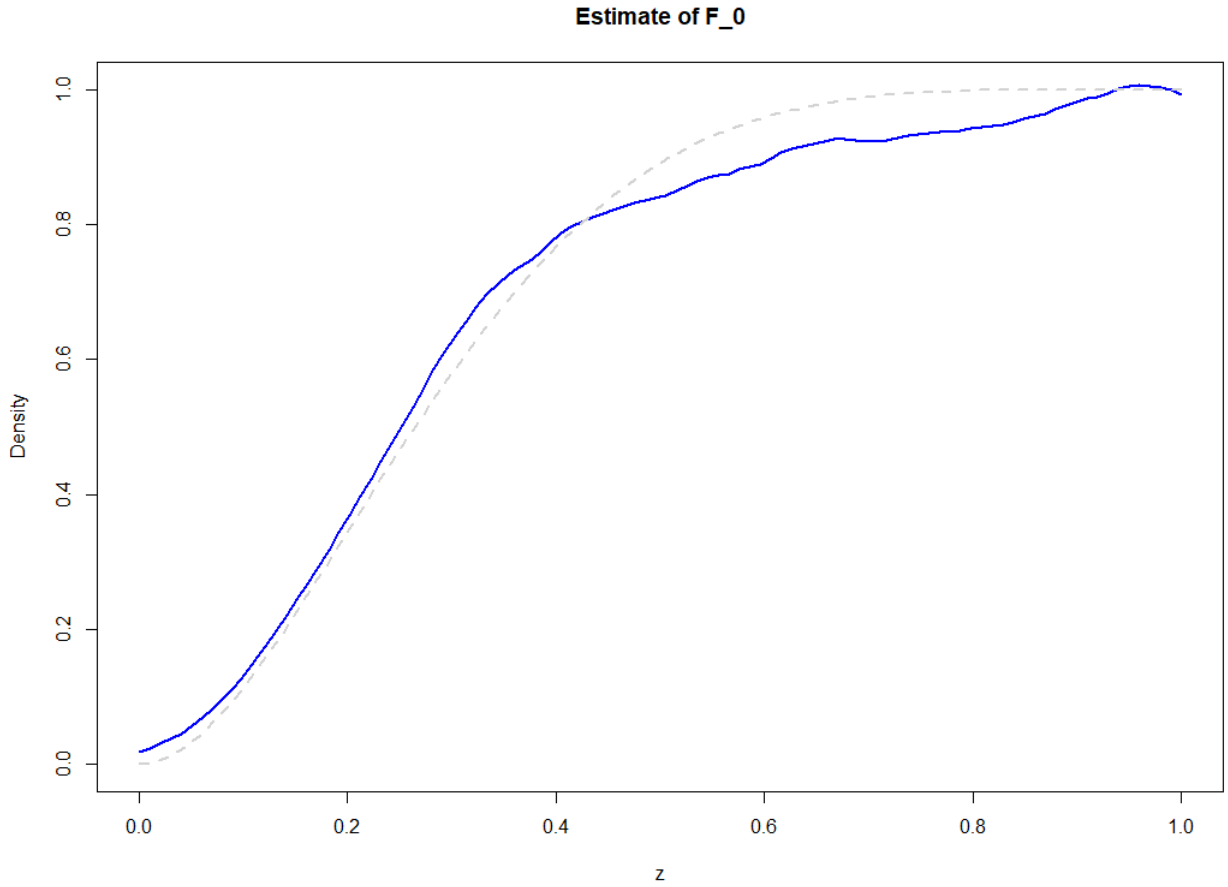


Figure 6.3: Estimate of F_0 using KDE

We again plotted the true distribution of F_0 in light grey for reference on the accuracy. We can calculate the ISE as well to give a more rigorous notion of its accuracy. The ISE equals $1.684332 * 10^{-3}$.

6.3. Comparison

In this section, we compare results of several simulations obtained using the two estimation methods: the nonparametric maximum likelihood estimator (NPMLE) and kernel density estimation (KDE). Our aim is to compare the performance of both approaches in the context of the uniform deconvolution problem, and to highlight their strengths and limitations.

As a point of reference, we first present the estimated distribution functions $\hat{F}_0$ obtained in Sections 6.2 and 6.2 by both methods in the figure below.

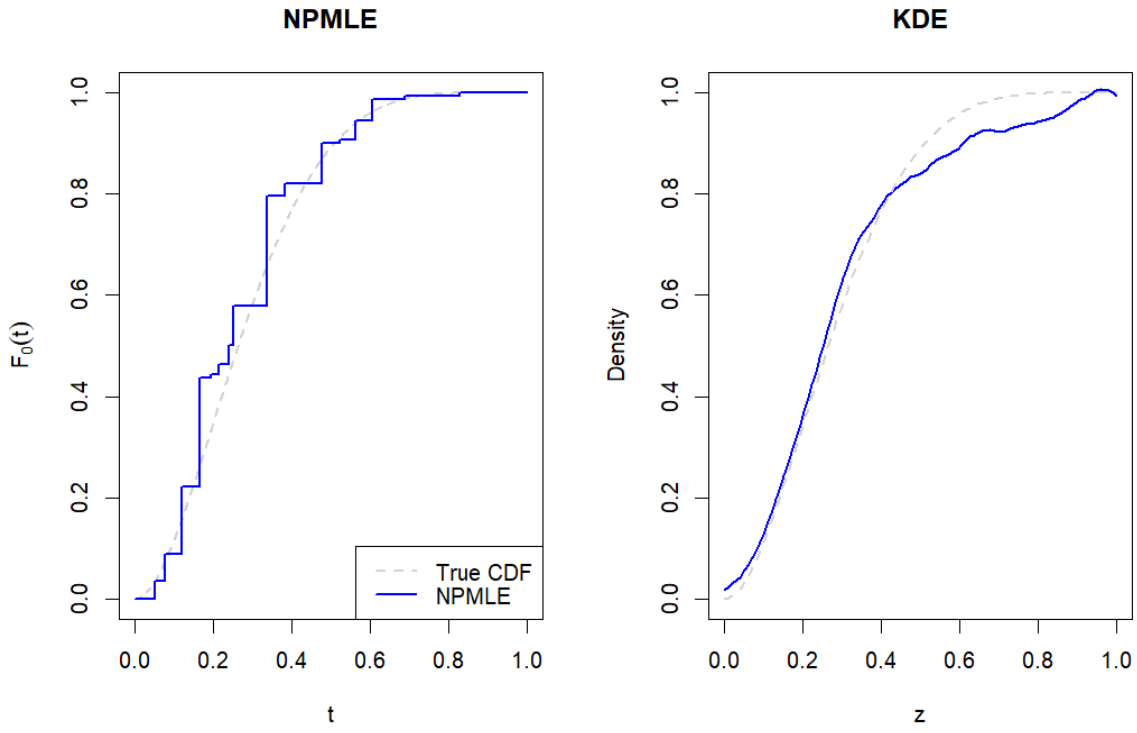


Figure 6.4: Estimates of F_0 using NPMLE and KDE

The estimates have the corresponding values for the ISE.

Method	ISE
NPMLE	$3.317439 * 10^{-3}$
KDE	$1.684332 * 10^{-3}$

Table 6.1: ISE for NPMLE and KDE

However, these are results based on one single simulation. To derive more concrete results we must consider more than this one simulation. This is done by taking 100 different instances of this same case. The ISE's are calculated for each of these cases and we obtain the following box plot of the ISE's.

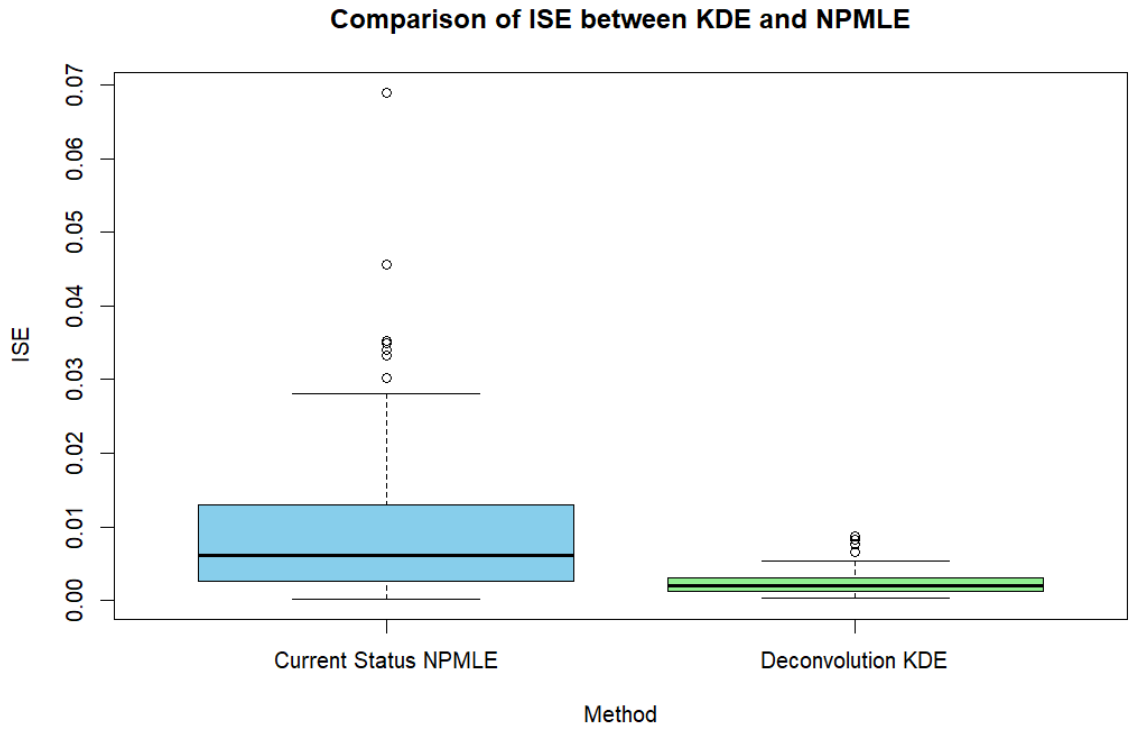


Figure 6.5: ISE's of 100 simulations for NPMLE and KDE

In the box plot, we can see that the ISE's are consistently lower for the KDE than for the NPMLE.

In conclusion, it is hard to say which of the two methods are objectively better. The KDE can be used without the assumption $F_0(1) = 1$, but does depend on the choice of bandwidth. In this case we use the optimal bandwidth but in practice this is unknown which gives some uncertainty. The NPMLE can be calculated consistently when dealing with different underlying distributions but can only be calculated by isotonic regression where we are in the specific case where $F_0(1) = 1$. In addition, it is only able to produce a step function whereas the KDE can produce a smooth function.

At first glance, from Figure 6.4, it is difficult to determine which estimation method performs better. It is immediately apparent, however, that the KDE produces a much smoother estimate, while the NPMLE always returns a step function. When we examine the values for the integrated squared error (ISE) in Table 6.1, we observe that the KDE achieves a lower error, suggesting a more accurate estimation in this particular case.

The KDE method also has some advantages. As discussed in Chapter 5, KDE does not require that $F_0(1) = 1$. In contrast, for the NPMLE to be equivalent to isotonic regression, the assumption $F_0(1) = 1$ is essential. If this condition is not satisfied, isotonic regression can no longer be applied, and an alternative approach to solving the NPMLE (like the iterative convex minorant algorithm) would be required. The KDE is not affected by this constraint and is therefore applicable in a more general case.

However, the KDE method has a significant disadvantage compared to the NPMLE, namely, it depends on the choice of bandwidth. In the simulation shown in Section 6.1, we used the optimal bandwidth for our estimation, but in practice this optimal value is generally unknown. As illustrated in Section 5.2, the KDE's performance can vary significantly when choosing different bandwidths. In contrast, the NPMLE does not require the selection of such a parameter, and thus can be calculated more consistently when the underlying distribution is changed.

It may be interesting to compare the asymptotic properties of the methods, how fast they converge to the true distribution and what kind of distribution the difference between the estimate and the true function converges to, like a nonparametric version of the central limit theorem.

In conclusion, it is difficult to say that one of the two methods is objectively better. The KDE is applicable to a wider class of problems and yields smooth estimates, but its performance is sensitive to bandwidth selection which gives it an additional uncertainty. The NPMLE, while applicable in a more constraint situation and limited to step function estimates. It also does not depend on the choice of some parameter and therefore can be applied more consistently and it does not have the uncertainty whether it can be more accurate when choosing a different parameter.

7

Conclusion

In this conclusion, OpenAI, 2025 was used for text references and rewriting parts of the text.

In this report, we have studied the uniform deconvolution problem. It is a problem in statistics where the observed data has a uniform noise term. We want to estimate the underlying distribution of the variables of interest taking into account that the observations are distorted by noise. The density of the polluted data is the convolution of the densities of the noise and the unpolluted data. Obtaining the CDF of the unpolluted data is called deconvolution and, as the noise is uniformly distributed, we are in the more specific case of *uniform deconvolution*. We have derived a connection between the density function of the observations with the noise, and the CDF of the underlying distribution of the data without the noise. Using this connection, we estimated the CDF of the unknown distribution using nonparametric estimation.

Nonparametric estimation is a form of estimation where the family of possibilities for the CDF is not parameterized by some finite dimensional parameter. This means that the family of possible distributions has an infinite dimension as it consists of every possible distribution function. This is a very large family where we cannot use estimation methods like the MoM. The MLE is able to estimate in a nonparametric way, however it cannot be computed with the method of taking the derivative of the log-likelihood w.r.t. a parameter. These methods, namely, estimate the distribution using a k-dimensional parametrization. We therefore have defined other methods to estimate our CDF in a nonparametric way. Two methods are treated in this report, the nonparametric maximum likelihood estimator (NPMLE) and the kernel density estimator (KDE).

The NPMLE uses the same idea of estimation as the parametric MLE in the sense that it maximizes the log-likelihood function. However, the parametric MLE maximizes the log-likelihood over the parametrization of the family of distributions. This is, of course, not possible in the nonparametric case. To compute the NPMLE, we first have to restrict ourselves to the case where the CDF has support on $[0, 1]$ such that we can make a connection between the uniform deconvolution problem and the current status problem.
The current status problem measures whether an event has occurred at the time of observation (so we only know the "current status") and using this we would like to estimate the distribution of the event time. This can be interpreted as the onset of a disease. When testing patients, we only know whether the patient is infected at the time of the test while we would like to know the onset time of the disease. It turns out that a specific case of this problem has the same log-likelihood function as the uniform deconvolution problem. This is convenient as we now can use properties from the current status problem which allows us to maximize the log-likelihood function using the method of isotonic regression, and hereby obtaining the NPMLE for both the uniform deconvolution and the current status problem.

The KDE is another method of nonparametric estimation with which we estimate the uniform decon-

volution problem directly. In this method we can directly compute a density function based on finitely many observations. When there are a lot of observations with values that are close to each other, the KDE assigns a high value for the density for these observations and we assign a low density to values where there are hardly any observations nearby. With this method, we can obtain an estimate for the density function of our observations. However, this is an estimate based on the observations with the uniform noise so we have an estimate of the convolution of the desired density and the standard uniform density. Using the connection between the convoluted density and the desired CDF, we calculate the estimate for the CDF.

These two methods both have their strengths an weaknesses. Our method of obtaining the NPMLE only works in the case where the CDF has support on $[0, 1]$, when this is not the case, other methods have to be used to compute the NPMLE. In addition, it is only able to produce a step function as the estimate. The KDE works in the case where the CDF has support on any interval and produces a smooth function. Also, when comparing the integrated squared errors in the simulation in Chapter 6, a measure for accuracy of the estimations, the KDE performs better. However, the KDE depends on the choice of a parameter called the bandwidth. There is an asymptotically MISE optimal value for the bandwidth, but the calculation uses the true CDF which, of course, is unknown. Therefore, it is hard to choose an appropriate bandwidth and the asymptotically MISE optimal bandwidth changes when dealing with different situations where we are dealing with an other CDF. In this aspect, the NPMLE may be a better method as it does not depend on any choice of parameter and will consistently give an estimate without needing any knowledge of the unknown CDF.

References

Akinshin, A. (2020). *Aakinshin*. Retrieved July 8, 2025, from https://aakinshin.net/posts/kde-bw/

Eric. (2025). *Aptech*. Retrieved May 8, 2025, from https://www.aptech.com/blog/the-fundamentals-of-kernel-density-estimation/

Groeneboom, P., & Jongbloed, G. (2014). *Nonparametric estimation under shape constraints* [Quadratic distance derivative to (in)equalities]. Cambridge University Press.

Groeneboom, P., & Jongbloed, G. (2025). Nonparametric estimation in uniform deconvolution and interval censoring [Connection between current status and uniform deconvolution].

Groeneboom, P., & Wellner, J. (1992). *Information bounds and nonparametric maximum likelihood estimation* [Connection between current status and uniform deconvolution]. Birkhäuser.

Hansen, B. E. (2009). Lecture notes on nonparametrics [University of Wisconsin].

OpenAI. (2025). Chatgpt (june 4 version). https://chat.openai.com/

Robertson, T., Wright, F., & Dykstra, R. (1988). *Order restricted statistical intference* [Likelihood to quadratic distance].

Shivam. (2024). *Analytics vidhya*. Retrieved July 8, 2025, from https://www.analyticsvidhya.com/blog/author/noobmaster21/